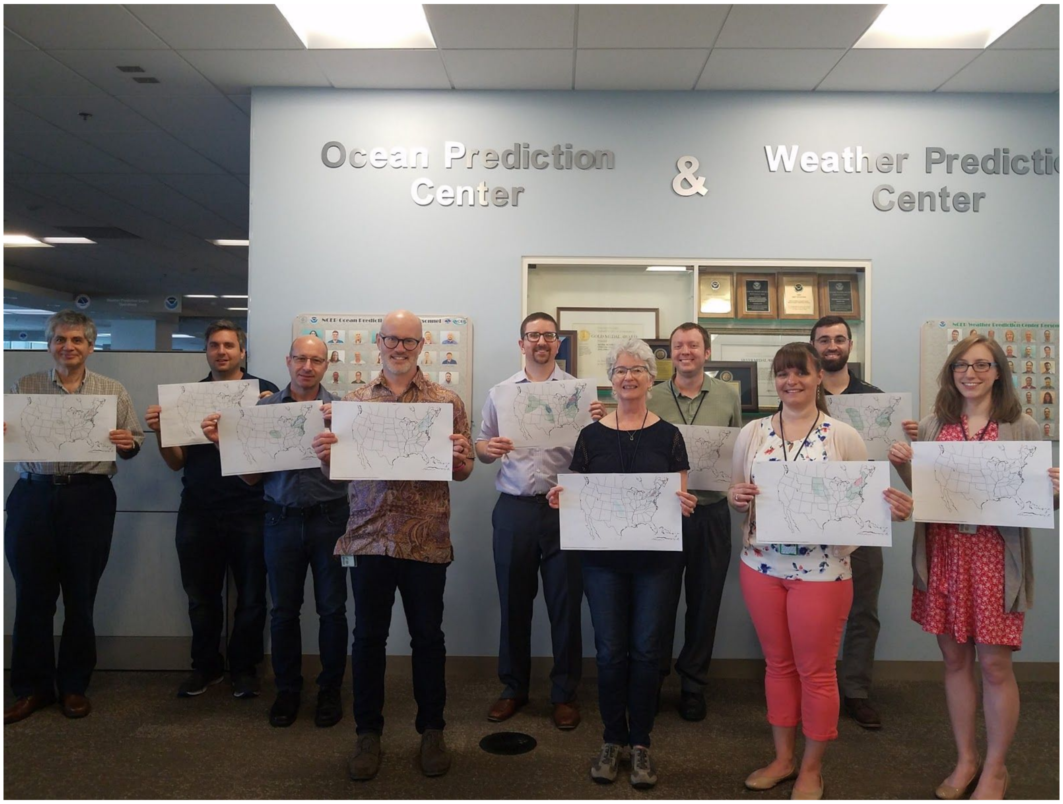
2019 Flash Flood and Intense Rainfall Experiment: *Findings and Results*



June 17 - July 19, 2019

Weather Prediction Center

Sarah Trojniak - Systems Research Group, NOAA/NWS/WPC/HMT
Benjamin Albright - Systems Research Group, NOAA/NWS/WPC/HMT

Hydrometeorology Testbed
College Park, MD
Table of Contents

1. Introduction

The Hydrometeorology Testbed at the Weather Prediction Center (HMT-WPC) has run the Flash Flood and Intense Rainfall (FFaIR) experiment annually since 2012. The experiment brings together researchers and forecasters from across the meteorological community to work towards the common goal to enhance forecasting skill related to flash flooding and intense rainfall events. FFaIR also serves as an important step in the Research to Operations (R2O) process by evaluating a variety of new models and products in a semi-operational environment.

The 2019 FFaIR experiment focused on forecasting flooding and intense rainfall across the continental United States (CONUS) in the Day 1 time period. The experiment was designed to mimic operations on the Day 1 Quantitative Precipitation Forecast (QPF) and MetWatch Desks at WPC. Two of the products that are issued from the desks are: the Excessive Rainfall Outlook (ERO) at Day 1 QPF and the Mesoscale Precipitation Discussion (MPD) at MetWatch. The ERO is a probabilistic product that is issued for the CONUS, highlighting regions where rainfall may exceed flash flood guidance (FFG) within 40 km of a point. The Day 1 ERO is valid from 1200 UTC to 1200 UTC but may be updated as the forecaster sees fit. A MPD's purpose is to highlight areas where there is concern for heavy rainfall and resulting flash flooding over the next six hours or less. The MPD is a short-term, event driven, guidance product that is typically issued over a limited domain and not CONUS wide.

To help identify the usefulness of various experimental model guidance and tools in the Day 1 time period the HMT team partnered with National Weather Service (NWS) meteorologists, hydrologists, and the development and research communities to mimic operations at WPC. Participants were tasked with producing four experimental forecasts each day comprising one ERO and three MPD-like products. These were created and issued almost exclusively off of the analysis of the experimental data that was evaluated during FFaIR. During this process, the participants were encouraged to have lively discussions about not only where the threat of intense rainfall and flooding would occur but also about the experimental tools they were utilizing.

2. Science and Operation Goals

The focus of the 2019 FFaIR experiment was to evaluate the usefulness of operational and experimental products from high resolution deterministic and ensemble models to increase forecast skill during the meteorological Day 1 timeframe. The majority of the guidance analyzed was experimental with some operational guidance mixed in; this is further discussed below in Section 3. In addition to high-resolution guidance, the participants also used experimental machine learning forecast products and experimental satellite products.

A secondary component of the experiment included the analysis of the performance of convective allowing models (CAMs) that use the finite volume cubed-sphere dynamic core, referred to as the FV3 core. This core was introduced operationally in June 2019 when GFSv15 was released; also referred to as FV3-GFS or GFS-FV3. The experimental version of the GFSv15 was evaluated in the 2018 FFaIR. A goal of the NWS is to have all meteorological models within the NWS use this core, which led to an increase in the number of high-resolution CAM models utilizing the new FV3 core when compared to last year's experiment.

Another objective of the experiment included evaluating forecaster understanding and their use of probabilistic and ensemble tools. Additionally the HMT team wanted the participants to see what it is like to work at a NWS national center, specifically focused on the Day 1 QPF and MetWatch Desks at WPC. This was done not only by having them issue experimental products similar to those issued by WPC forecasters, but also by including WPC forecasters in the experiment, and having them "drive" the forecast process; see Fig 1. The WPC forecaster also answered any questions the participants had about working at WPC and the methods they used to determine whether or not an event warranted an ERO or MPD.

In summary, the specific experiment goals for 2019 FFaIR were to:

- Evaluate the usefulness of operational and experimental products from high resolution convective-allowing deterministic and ensemble models for forecasting near-term flash flood events.
- Assess the forecasters' understanding of ensemble tools such as probability matched mean (PMM) and local probability matched mean (LPMM) and identify their usefulness in the forecast process.
- Identify ways to incorporate hydrological model guidance into the decision making process for flash flooding guidance issuance.
- Evaluate, subjectively and objectively, the ability of the CSU-MLP "First Guess Field" to predict the Marginal, Slight, Moderate, and High Risks for the Day 1 ERO.
- Identify ways to incorporate advanced remote-sensing and difference fields into the flash flood forecasting process.
- Evaluate, subjectively and objectively, the utility of the FV3-Stand-Alone Regional (SAR) in comparison to the nested version of the FV3-GFS (FV3-Nest), specifically at later forecast times, to determine if the FV3-SAR is a viable alternative to the FV3-Nest.

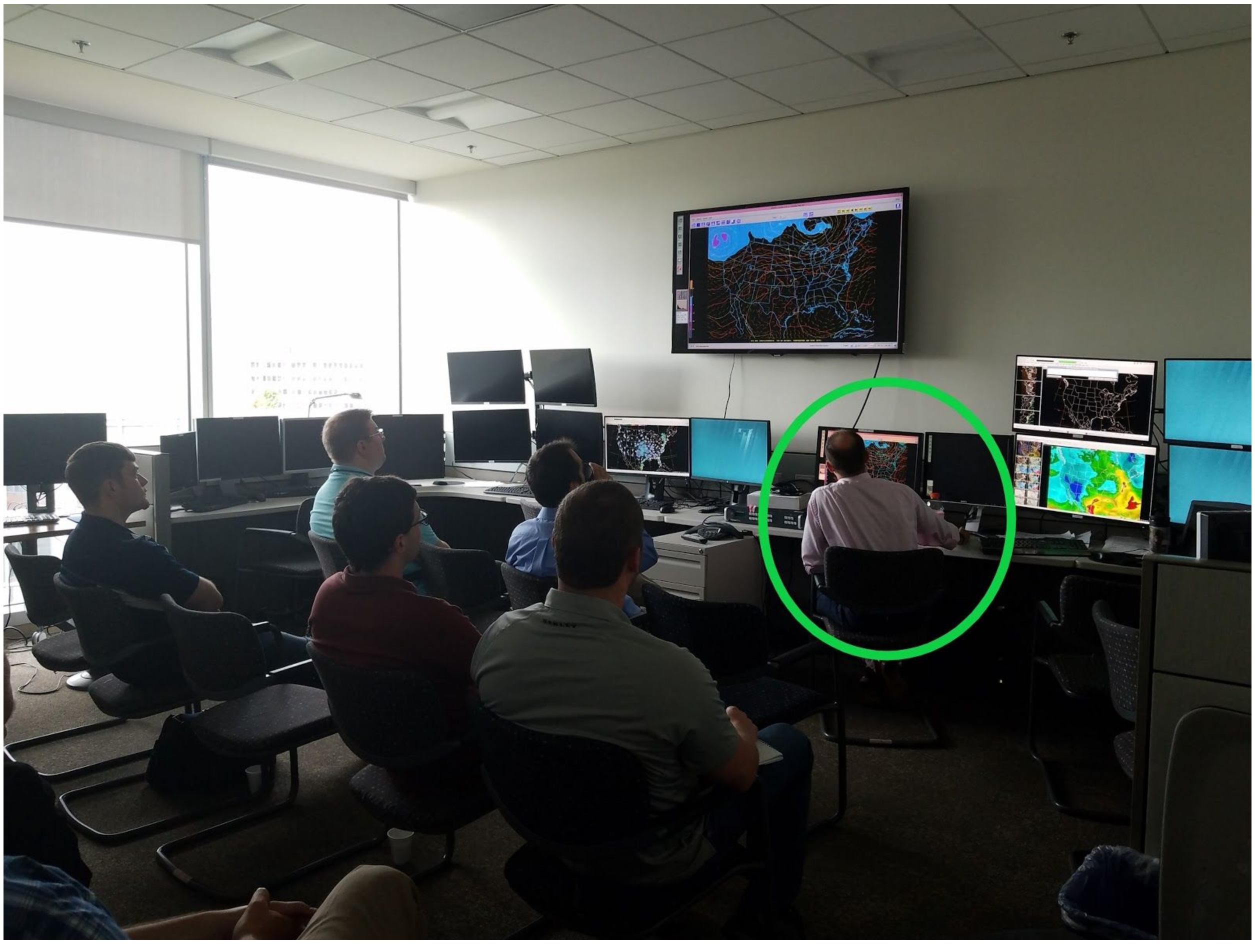


Figure 1: *Example of FFaIR participants working with WPC forecaster Josh Weiss (circled in green) to analyze the day's areas of interest for flash flooding and intense rainfall. Taken on 25 June 2019.*

3. Experiment Methodology and Data

The FFaIR experiment was held for four weeks beginning June 17, 2019 in the WPC-OPC Collaboration Room at the NOAA Center for Weather and Climate Prediction (NCWCP). The four weeks of the experiment were:

Week 1 : June 17-21, 2019
Week 2 : June 24-28, 2019
Week 3 : July 8-12, 2019
Week 4 : July 15-19, 2019

Each week featured a new group of participants and a new WPC forecaster. The list of participants and WPC forecasters for each week can be found in Appendix A in Table A1. The number of participants varied week to week. The products and tools evaluated in FFaIR are listed below in Table 1.

Table 1: *Summary of the operational/experimental model and ensemble guidance, experimental products, and experimental forecasts issued that were subjectively evaluated by the science questions posed by the testbed staff, along with the number of collected scores provided by the participants in parentheses.*

FFaIR Products evaluated and the number of scores for each Model/Cycle/Parameter evaluated in parenthesis						
PRODUCT EVALUATED	**Day 1 QPF**	**Ensemble 6h PMM QPF**	**Ensemble 6h LPMM QPF**	**CIRA/ CSU TPW**	**CSU Day 1 First Guess ERO**	**FFaIR Forecasts**
	FV3-SAR (131)	**HRRRE 00Z** (159)	**HREFv3** (194)	**BTPW** (82)	**GEFS** (178)	**Day 1 ERO** (167)
	FV3-Nest (126)	**HRRRE 12Z** (147)	**HRRRE** (159)	**Merged TPW version 1** (82)	**NSSL** (155)	**PFF1** (168)
	HRRRv4 (129)	**HREFv3** (194)	**NCAR** (80)*			**PFF2** (168)
	HRRRv3 (189)	**NCAR 00Z** (78)*	**SSEFX** (170)			**PFF3** (141)
	NAM-Nest (178)	**NCAR 12Z** (48)*				
	FV3-GSD-SAR (93)*	**SSEFX** (170)				
	SSEF-NSSL (191)					
	SSEF-Morr (191)					
	SSEF-Thomp (138)					
**Counts significantly lower due to data outages.*						

Forecast Activities

The participants were required to issue three experimental forecasts, with an option to issue a fourth if they felt the situation was warranted. The experimental products were all probabilistic products assessing the potential for flash flooding within 40 km of a point, over several different time periods. Table 2 lists the various forecast products issued along with information about them. In the morning, an ERO was issued for the CONUS and was valid from 1500 UTC to 1200 UTC, with probabilistic contours of 5% (Marginal), 10% (Slight), 20% (Moderate), and 50% (High). An example of an experimental ERO is shown in Fig. 2A. In the afternoon, the shorter term Probability of Flash Flooding (PFF) forecasts were issued. The PFF1 and PFF2 were valid for six hours from 1800 UTC to 0000 UTC and 0000 UTC to 0600 UTC, respectively. The PFF3 served as either an update to the PFF1 using new, experimental information that became available later in the afternoon, or could be issued over a new location. The PFF3 was valid for three hours from 2100 UTC to 0000 UTC. Each PFF had probabilistic contours of 10% (Slight), 20% (Moderate), and 50% (High) and were meant to be similar to the WPC MPD by using similar guidance over a limited domain. An example of an experimental PFF can be seen in Fig. 2B.

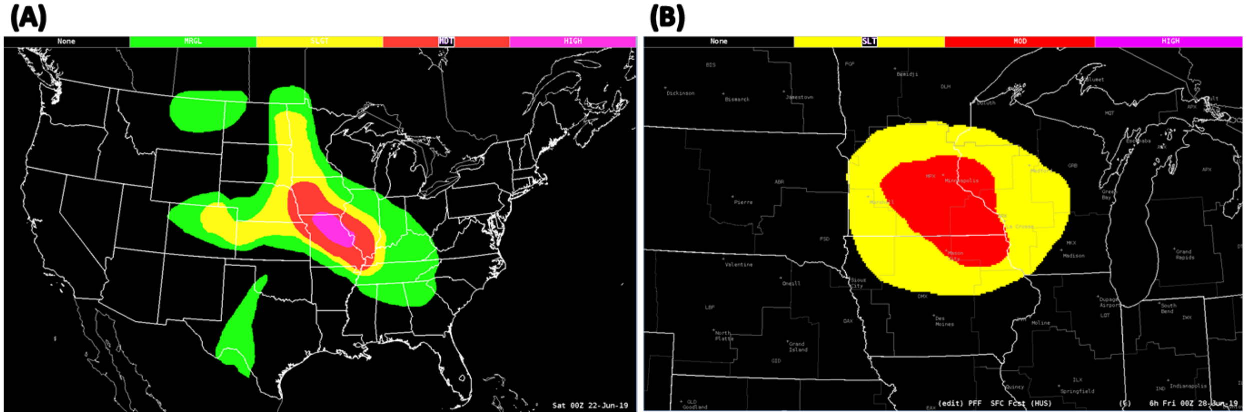


Figure 2: *(A) Example of an experimental ERO issued by Week 1 FFaIR participants valid 1500 UTC 21 June 2019 to 1200 UTC 22 June 2019. (B) Example of an experimental PFF2 issued by Week 2 FFaIR participants valid 0000 UTC to 0600 UTC 28 June 2019.*

Table 2: *Summary of the experimental products that the participants issued daily throughout the 2019 FFaIR experiment.*

Product Name	Issued By	Valid Times	Risk Category	Probability of Flash Flooding within 40 km of a Point
Day 1 ERO	1430 UTC	15 UTC – 12 UTC	Marginal Slight Moderate High	5% 10% 20% 50%
PFF1	1730 UTC	18 UTC – 00 UTC	Slight Moderate High	10% 20% 50%
PFF2	1930 UTC	00 UTC – 06 UTC	Slight Moderate High	10% 20% 50%
PFF3	2030 UTC	21 UTC – 00 UTC	Slight Moderate High	10% 20% 50%

Each morning would begin with the WPC forecaster reviewing what occurred over the past 24 h. This included looking at the 24 h radar loop and local storm reports (LSRs). This was nearly always accompanied with a group discussion assessing the experimental forecasts from the previous day. After the discussion, the WPC forecaster would go over the current meteorological conditions with the participants. This included examining current radar and satellite, as well as upper air soundings and surface observations. The participants, along with the WPC forecaster, would then begin the basic heavy rainfall forecasting process such as: looking at the synoptic setup, low/mid-level moisture, fields, moisture transport and convergence, precipitable water (PWAT) values and anomalies, and various instability fields. After analyzing this model information, the group would have a basic idea of where the "problem" areas for the day would be across the CONUS.

The group would then look at the experimental quantitative precipitation forecast (QPF) guidance from the deterministic and ensemble models. This guidance included hourly to 24 hour deterministic QPF, simulated radar, probability of QPF exceeding various thresholds over varying time periods (i.e. exceeding 2 inches in 3 h), and QPF exceeding flash flood guidance (FFG). Along with model guidance they would also use the experimental satellite products provided by the Cooperative Institute for Research in the Atmosphere (CIRA) at Colorado State University (CSU), mostly using their Merged Total Precipitable Water version 1 product, as well as their Advected Layer Precipitable Water (ALPW) product. The participants would also examine the experimental Machine Learning CSU Day 1 ERO “First Guess” products to help them get an idea of where the risk for flash flooding across the CONUS might be.   Once the participants felt they had a general idea of what the Day 1 ERO should look like, each drew their own ERO on a blank map. Once collected, the WPC forecaster would attempt to combine them and create an experimental ERO. The final result was discussed and a general consensus was reached before the forecast was issued by 1430 UTC.

A similar process was repeated for each of the PFFs, with the addition of determining which area of interest had the greatest risk of flooding in the six or three hour time period. Determining where in the smaller time frame the greatest threat for flash flooding would occur across the CONUS often proved to be a greater challenge than issuing the ERO. There were many instances, especially during the first half of the experiment, when there were multiple regions of interest during the time period a PFF would be valid (see Table 2 for the valid times for each PFF). However, the group would work together to determine where, based on meteorological and antecedent conditions, flash flooding was most likely to be reported. Once the location of interest for the PFF was determined the participants would again use the experimental guidance to draw out the risk areas or decide not to issue the PFF. Note, marginal risks were not included as part of the PFF, so a slight risk was the lowest risk category they could issue.

Daily Weather Briefing

Each day of the experiment consisted of a remote weather briefing centered around the FFaIR experimental ERO and PFF1. The experimental partners of FFaIR and forecasters from NWS Weather Forecast Offices (WFOs) and River Forecast Centers (RFCs) were encouraged to call in. The goals of the briefings were to feature the experimental guidance and tools that were being utilized in the 2019 FFaIR experiment as well as to highlight the regions of interest for flash flooding for the day. The briefing would be given by a volunteer FFaIR participant[1] and would last 20-30 minutes; a list of the briefing presenters and those who called in can be found in Appendix A in Table A2. The briefing would conclude by opening the floor up for discussion and questions for the people who called in. Discussion would range from questions about the experimental products shown, weather that was ongoing, and how active the year has been for flooding along the Mississippi River and its basins. Lastly, since the HMT-Hydro experiment was running concurrently three of the four weeks of FFaIR, there was often collaboration during the open floor portion of the briefing with the HMT-Hydro experiment on where they should focus their efforts for the day. The Final Report for the HMT-Hydro experiment can be found here.

[1] Each Monday of the experiment the WPC Forecaster of the week would be in charge of doing the weather briefing.

Verification

In addition to using the experimental guidance and tools provided by FFaIR partners to issue forecasts, the products were also verified subjectively and objectively throughout the course of the experiment. There were a variety of verification methods used, ranging from general verbal feedback to metrics such as the critical success index (CSI). The various methods will be discussed in the following sections.

Subjective Verification Methodology

The subjective verification done in the 2019 FFaIR experiment centered around fourteen science questions that were developed by the testbed staff in collaboration with this year's science partners. The majority of these questions asked the participants to evaluate the utility, quality, or performance of the guidance, tools, or forecasts on a scale from 0.5 (very poor) to 10 (very good). Participants were also asked to provide verbal feedback about what they liked/disliked about the guidance, tool, or forecast so the testbed team could provide additional information about the usefulness of the product to the developers. If the questions involved evaluating a model or ensemble product, the model/ensemble name was removed from the question to mitigate any biases the participants might have.

A few of the science questions were comparison questions, asking the participants to compare the utility of two products or tools against each other. Using a numerical ranking system, the participants were asked if they preferred one product/tool over the other or if they felt the utility of each was about the same. If they thought both were about the same they were instructed to score 5 for the question, but if they felt one was better they were instructed to adjust their score up or down from 5 towards either 1 if they preferred product/tool A or towards 10 if they preferred product/tool B. A score of 1 or 10 meant the participant completely preferred one product over the other, while anything in between those two scores and a 5 meant there was utility in both products/tools but one of the two was preferred; refer to Fig. 3 to see how the scale worked. For example, if both models had a good footprint of the QPF when compared to observations but Model A QPF had amounts closer to reality a score might be 3.5.

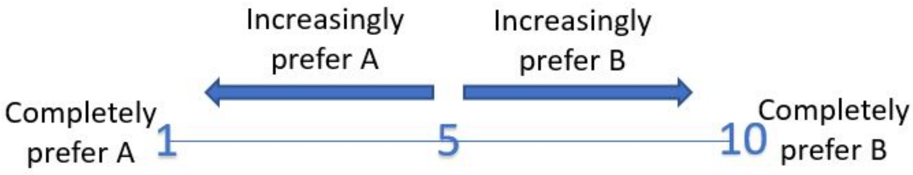


Figure 3: *Diagram depicting the rating scale for the science questions that asked participants to compare two products/tools against one another.*

Table 1 lists the guidance, tools, and forecasts that were evaluated throughout the 2019 FFaIR experiment by these fourteen science questions. It also includes the number of scores each received throughout the course of the experiment. The number of scores varied based on model availability and the number of daily participants providing scores. As can be seen from Table 1 there was a large variance in the number of scores between some of the model products. Specifically when looking at the number of scores for various ensemble products, the National Center for Atmospheric Research (NCAR) Ensemble had significantly fewer scores than the rest of the ensembles. This was the result of maintenance on the NCAR Supercomputer, which precluded running of the experimental guidance for much of the experiment. Therefore, although it is included in Table 1 and there is a description of the ensemble in the 2019 FFaIR Operations Plan, the testbed team feels there was not enough data to provide a valid analysis of the ensemble and thus one will not be provided in this report.

Model and ensemble QPF were verified against Multi-Radar, Multi-Sensor Gauge Corrected (MRMS-GC) Quantitative Precipitation Estimates (QPE) over the forecast period that was being evaluated. In general, model QPF was verified over a 24 h period from 1200 UTC to 1200 UTC across the CONUS. Ensemble products were verified over six hour periods from 1800 UTC to 0000 UTC. The ensemble verification was also regional instead of CONUS wide and was usually centered around the location of the PFF1 for the day being verified. Figure 4 provides an example of what the participants would be shown to verify model and ensemble QPF each day.

Verification for the CSU First Guess fields, the experimental FFaIR ERO, and PFFs issued by the participants was done using a method developed by the HMT testbed team in 2017 called the Unified Flood Verification (UFV) system. The system uses a combination of 1/3/6 hour QPE exceeding flash flood guidance (FFG), 1/6/24 hour QPE exceeding the five year average recurrence interval (ARI), and flash flood/flood NWS Local Storm Reports (LSRs) and U.S. Geological Survey (USGS) gauge reports. A 40 km radius is then applied to each LSR and USGS gauge report and where QPE exceeds either FFG or the 5 year ARI a 25 km radius was applied; these are considered hits. The hits are then combined and plotted on one map along with the probability contours of the product being verified. An example of this method can be seen in Fig. 5A.

In addition to UFV and MRMS-GC QPE, a tool referred to as the “practically perfect technique” was used to help the participants subjectively evaluate the experimental products. Practically perfect is an analysis technique used to measure the “goodness” of a probabilistic forecast. It is configured using Stage IV precipitation data exceeding ARI and FFG along with LSRs and USGS observations. When any of the criteria is met during the specified time period, it is considered a hit. A neighborhood probabilistic forecast is then created from the hits, and a varying radius of influence is applied. A radius of influence of 40 km is applied to LSRs and USGS observation hits while a 25 km radius of influence is applied to instances of QPE data exceeding

FFG or ARI. The binomial field (1 = hit, 0 = no event) is then smoothed using a 105 km Gaussian smoother to create a proxy "probabilistic observation" known as practically perfect. Numerous sensitivity runs of the practically perfect technique were conducted to ensure that the resulting "probabilistic observation" compared well to the operational ERO.

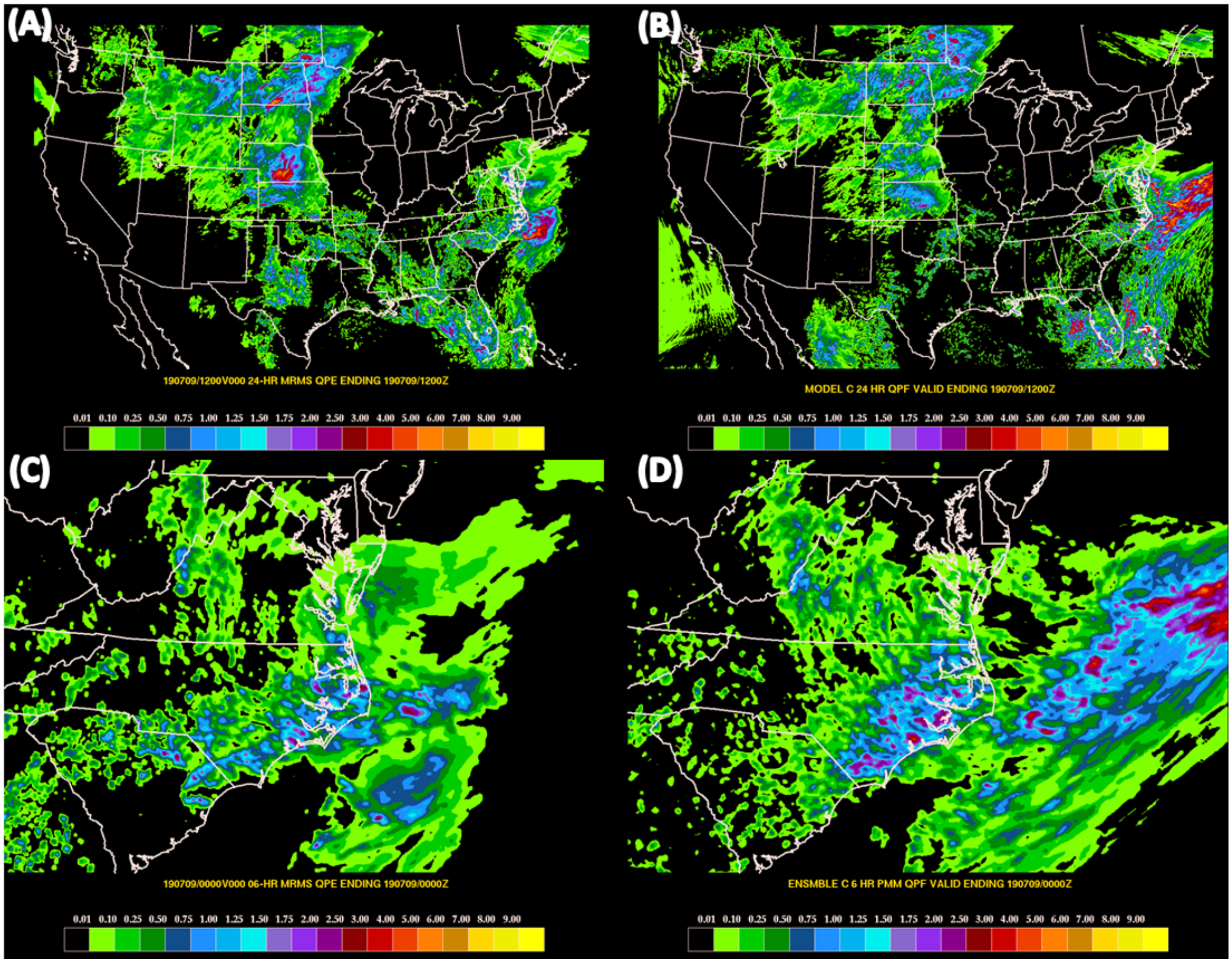


Figure 4: *Across the top is an example of the verification image shown to the participants for Model QPF valid from 1200 UTC 08 July 2019 to 1200 UTC 09 July 2019; where (A) is the 24 h MRMS-GC QPF and the (B) is the 24 h QPF forecast from Model C (HRRRv4). Across the bottom is an example of the verification image shown to the participants for Ensemble Probability Matched Mean QPF valid from 1800 UTC 08 July 2019 to 0000 UTC 09 July 2019; where (C) is the 6 h MRMS-GC QPE and the (D) is the 6 h forecasted Probability Matched Mean QPF from Ensemble C (HREFv3).*

Figure 5C provides an example of a 21 h practically perfect forecast[2]. The product is designed to mimic an ERO and thus the contours are the same as the probabilities contoured for EROs, though the color table differs. For example, referring to Fig. 5 valid from 1500 UTC 21 June to 1200 UTC 22 June 2019, it can be seen that the experimental FFaIR ERO included a high risk of flooding across northeastern MO and a moderate risk from northwestern IA to extreme northwestern TN. However, the practically perfect analysis suggests that the flooding risk was greatly over-forecasted, especially over IA and MO. Based on observations, only a marginal risk should have been issued where the high risk was issued. Additionally, practically perfect indicates that the highest risk over portions of the midwest should have been a slight.

Figure 5 is also an example of how verification information for the EROs evaluated in the experiment were presented to the participants. Using the information from the three different forms of verification, the participants then scored the CSU First Guess fields, the experimental FFaIR forecasts and the PFFs. The image always had the following setup: top left had the experimental forecast along with the UFV plot, top right had the QPE, bottom left had the practically perfect analysis, and the bottom right was empty.

Finally, as previously stated, during the scoring portion of the verification the testbed team encouraged the participants to provide feedback about the guidance and tools. The feedback was in the form of an open discussion while numerical ratings were being given on the product or tool. There were also a couple of science questions that did not require any numerical scoring and only asked for comments and discussion. This type of "verification" was included in the experiment for a variety of reasons. In one instance, direct feedback from forecasters about a product was requested by the developer. It has also been found by developers that knowing what specifically (i.e. color, contour values, etc) the end users like or dislike about a product is just as useful as knowing if the participants felt the product performed well. This is because, in general, it has been found (2018 FFaIR Final Report) that even if a product performs well, if the forecasters don't like the way it looks or can't easily access it they likely will not use the product.

[2] When using the practically perfect for verification of the CSU EROs and the operational ERO from WPC, the verification window spans 24 hours while the products span 21 hours.

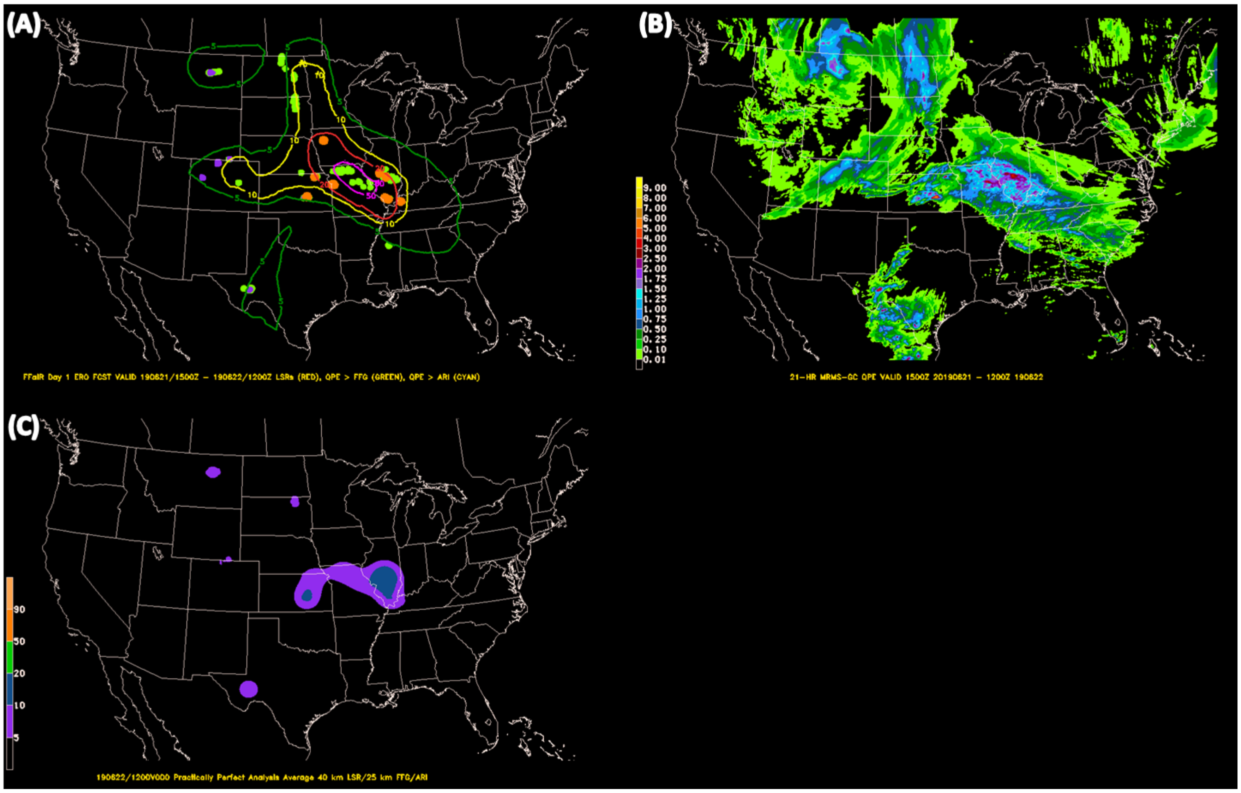

(A)
(B)
(C)
FFaIR Day 1 ERO FCST VALID 190621/1500Z – 190622/1200Z LSRs (RED), QPE > FFG (GREEN), QPE > ARI (CYAN)
21-HR MRMS-GC QPE VALID 1500Z 20190621 – 1200Z 190622
9.00
8.00
7.00
6.00
5.00
4.00
3.00
2.50
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.10
0.01
90
50
20
10
5
190622/1200V000 Practically Perfect Analysis Average 40 km LSR/25 km FTG/ARI

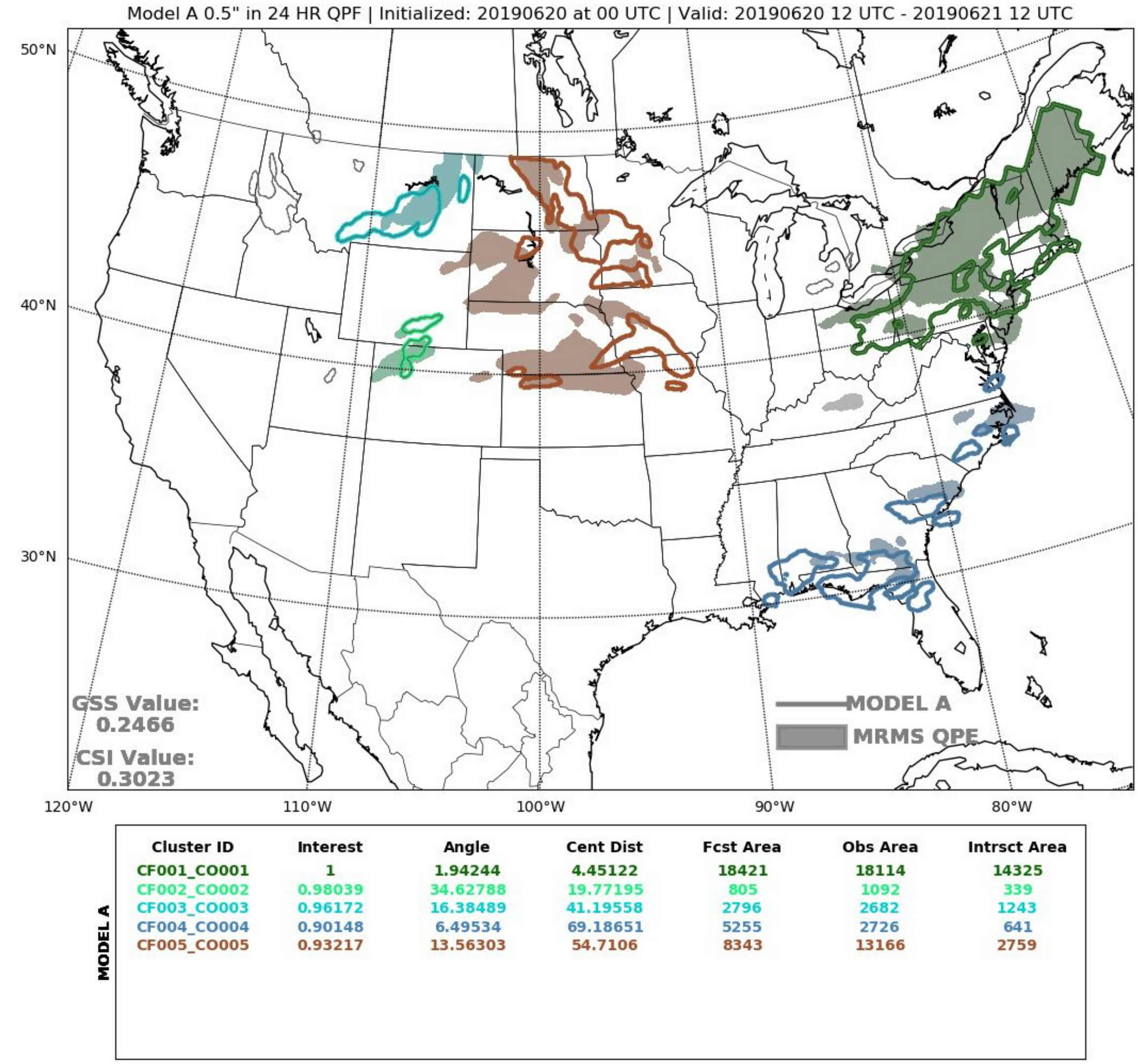
Figure 5: *Example of the image setup for the subjective verification for the CSU First Guess fields and the experimental EROs and PFFs issued by the FFaIR participants. This example is the Day 1 ERO valid 1500 UTC 21 June 2019 to 1200 UTC 22 June 2019*[3]*. (A) The Unified Flood Verification system where green dots are QPE > FFG, purple dots are QPE > ARI, and flash flood LSRs, flood LSRs, and USGS gauge reports and plotted in red. Overlaid are the probability ERO contours. (B) 21 hour MRMS-GC QPE. (C) practically perfect analysis valid for the same time period.*

Objective Verification Methodology

In addition to subjective verification, the Method for Object-Based Diagnostic Evaluation (MODE), part of the Model Evaluation Tools (MET) package, was used to compare various forecasted QPF thresholds from several models to MRMS-GC QPE (see Appendix D 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. An example of the MODE verification used in the 2019 FFaIR experiment can be seen in Fig. 6. MODE works by identifying objects and grouping them into events which are usually geographically separated. The spatial extent of the forecasted event can then be compared to the spatial extent of the observations and plotted as shown in Fig. 6. This example evaluates the 36 h forecast from “Model A” (FV3-SAR) of 24 h QPF at the 0.5 inch threshold valid at 1200 UTC on June 21, 2019. Referring to Fig. 6 it can be seen that there were five events identified;

[3] This is the same date as the example ERO shown in Fig. 2A.

the contour lines are the regions were “Model A” forecasted 24 h QPF to exceed 0.5 inches while the shaded regions are where QPE actually exceeded this threshold. In this instance, MODE suggests that “Model A” did well in forecasting for the Cluster ID: CF001_C0001 (dark green) with 18421 points forecasted within the event area and 18114 points observed within the event area. However, Cluster ID: CF005_C0005 (brown) the model did not do as well. For this event, “Model A” forecasted 8343 exceedance points but 13166 were observed.


Model A 0.5" in 24 HR QPF | Initialized: 20190620 at 00 UTC | Valid: 20190620 12 UTC - 20190621 12 UTC
50°N
40°N
30°N
120°W
110°W
100°W
90°W
80°W
GSS Value:
0.2466
CSI Value:
0.3023
MODEL A
MRMS QPE
MODEL A

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	1	1.94244	4.45122	18421	18114	14325
CF002_CO002	0.98039	34.62788	19.77195	805	1092	339
CF003_CO003	0.96172	16.38489	41.19558	2796	2682	1243
CF004_CO004	0.90148	6.49534	69.18651	5255	2726	641
CF005_CO005	0.93217	13.56303	54.7106	8343	13166	2759

Figure 6: *MODE analysis for the 36 hour “Model A” (FV3-SAR) forecast for 24 hour QPF at the 0.5 inch threshold valid from 12 UTC June 20 to 12 UTC June 21, 2019. Each color represents a different object (or event) while the contour lines represent “Model A” QPF exceeding 0.5 inches and the shaded areas represent the observed QPE exceeding 0.5 inches.*

MODE also computes the Gilbert Skill Score (GSS) and CSI, commonly referred to as Equitable Threat Score and Threat score respectively, over the CONUS domain for several models. Calculations were done by re-gridding all model QPF, along with the QPE, to a common 5 km grid with a CONUS mask applied. Finally, 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 Fig. 7. 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. Less biased forecasts will lie closer to the diagonal line from the bottom left corner to the top right corner. Forecasts with higher threat scores will fall toward the upper right of the graph. Therefore, a perfect forecast, i.e. the most accurate and the least bias, would be located in the top right corner of the diagram.

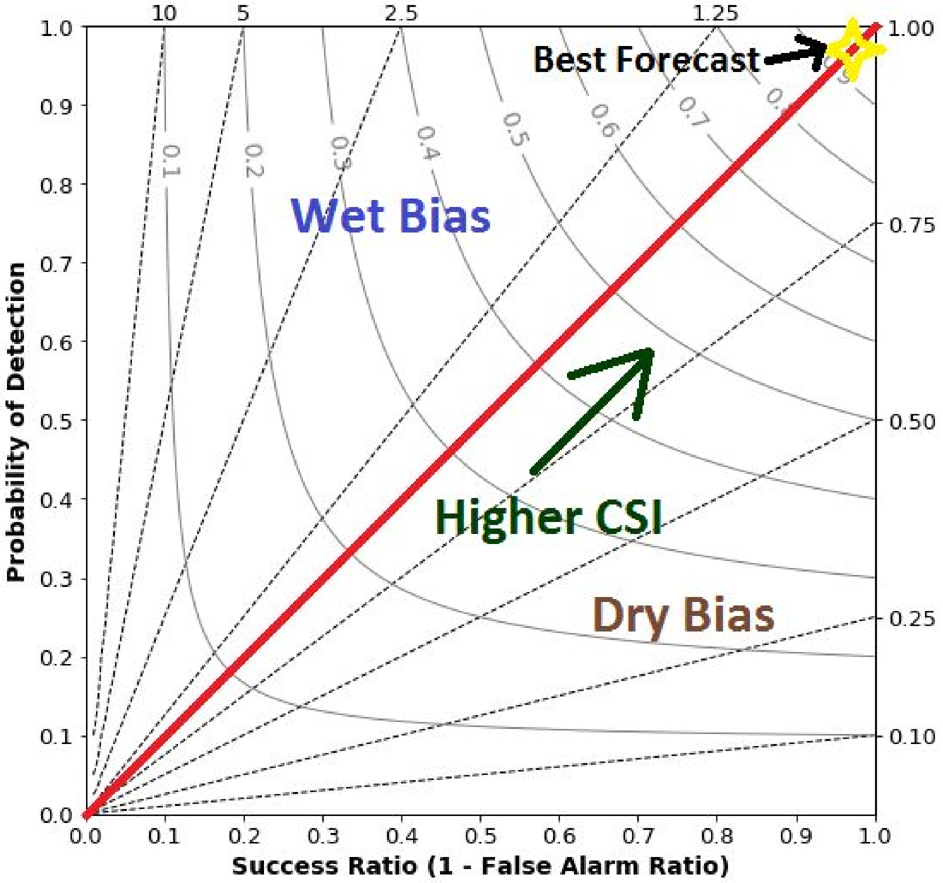


Figure 7. *Example of a Roebber Performance Diagram. Y-axis is the probability of detection, x-axis shows the success ratio (1 - false alarm ratio), dashed diagonal lines represent the bias, and curved solid lines represent CSI.*

Featured Guidance and Tools

There was a multitude of data that was evaluated in this year's FFaIR experiment, and although the majority were from experimental models and ensembles, there were some operational models and ensembles that were also evaluated. For instance, in this year's experiment, two versions of the deterministic high resolution rapid refresh model (HRRR) were evaluated, HRRRv3 (which is operational) and HRRRv4 (which is experimental). Additionally, this year's FFaIR experiment included the evaluation of numerous experimental deterministic and ensemble model guidance that utilized the FV3 core. For instance, the Environmental Modeling Center (EMC) provided deterministic runs of their CAMs that use the FV3 core; the FV3-Nest and the FV3 stand-alone regional model. Likewise, the University of Oklahoma's Center for Analysis and Prediction of Storms (CAPS) research team provided their CAM ensemble, the storm-scale ensemble forecast (SSEF), which is comprised of 14 members, all with the FV3 core. To further assess the performance of the FV3 core in CAMs three members of the SSEF were treated as deterministic runs and analyzed individually as well. A detailed list of the models, ensembles, tools, and guidance that were highlighted during the 2019 FFaIR experiment can be seen in Table 3.

Table 3: *The deterministic and ensemble model guidance that will be evaluated in the 2019 FFaIR experiment (the experimental guidance is in the darker shade).*

Provider	*Model*	*Resolution*	*Forecast Hours*	*Notes*
EMC	*NAM (Nest)*	*3 km*	*Hourly out to 60 h at 00, 06, 12, and 18 UTC .*	*Features an hourly forecast and assimilation cycle for its 3 km CONUS nest. Uses hybrid 3D-EnVar and incorporates radar reflectivity into its assimilation system via a complex cloud analysis approach.*
EMC	*Short Range Ensemble Forecast (SREF)*	*16 km*	*Output every 3 h out to 87 h.*	*Ensemble forecast with 21 members plus the time-lagged WRF-NAM every 6 h.*

EMC	HREFv2	~3 km	36 h forecast run every 6 h.	Operational version of HREF with 8 members, which each member providing a real-time and time-lagged run.
RFCs	Flash Flood Guidance (FFG)	~4 km	01, 03, 06, 12 and 24 h values .	A CONUS mosaic grid created by compiling individual RFC-domain grids. Provides an estimate of how much rain over a specific time-period would cause flooding on small streams.
NSSL/HDSC/ NERFC/CSU	Precipitation Recurrence Interval (RI) using Atlas 14	3 km	1, 3, 6, and 24 h products with RIs at 2, 5, 10, 25, 50, and 100 years.	Precipitation frequency estimates based on historical observations. The Northwest is not covered by Atlas 14 instead Atlas 2 is used.
OWP	National Water Model (NWM) version 1.2	250 m 1 km	Short Range: Forecast 18 h Mid-Rage: Forecast 10 days Long Range: 30 days	Analysis and forecast system that provides streamflow for 2.7 million river reaches and other hydrologic information. Hourly forecasts in the short range.
NSSL	Warn on Forecast System (WoFS) (previously called NEWS-e)	3km Domain 900x900 km	Initialization time varied. Forecasts were out 6 h, with output every 5 min.	HRRRE analysis/boundary conditions. Regional domain of 1000x1000 km. Observations assimilated every 15 min using GSI-EnKF.

ESRL/GSD	*HRRRv3*	*3 km*	*Hourly: 36 h for 00, 06, 12, and 18 UTC run times. All other run times have a 18 h forecast.*	*High resolution, hourly updated, convection allowing nest of the Rapid Refresh (RAP) model.*
ESRL/GSD	*HRRRv4*	*3 km*	*Hourly: 36 h for 00, 06, 12, and 18 UTC run times. All other run times have a 18 h forecast.*	*High resolution, hourly updated, convection allowing nest of the Rapid Refresh (RAP) model.*
ESRL/GSD	*HRRR Ensemble (HRRRE)*	*3 km*	*Hourly out 36 h at 00 and 12 UTC. As needed runs at 15 and 18 UTC.*	*Now covers the entire CONUS.* *9 member ensemble.*
EMC	*HREFv3*	*3 km*	*36 h forecast run daily at 00 UTC.*	*Experimental version of HREF with 9 members, which each member except for FV3-SAR providing a real-time and time-lagged run.*
GFDL/EMC	*FV3-Nest*	*3 km*	*Hourly out to 60 h initiated once daily at 00 UTC.*	*Run simultaneously with its parent domain, FV3.*

GFDL/EMC	FV3-SAR (Stand-alone Regional)	3 km	Hourly out to 60 h initiated once daily at 00 UTC.	The SAR is the stand-alone regional version of FV3 and does not have a global parent domain. **This differs from the FV3-Nest**, which is a forecast over the same domain as the SAR, but is not run simultaneously with it parent model, FV3.
GSD/EMC	FV3-SAR-GSD	3 km	Hourly out to 36h, initiated once daily at 00 UTC.	Uses the FV3-SAR to facilitate CAM-scale development. Has the same initial and boundary conditions as HRRRv3.
OU/CAPS	SSEF	3 km	60 h forecasts, though some of the membership only provide forecasts out 36 h. Initiated at 00 UTC.	14-members run on the FV3 model. ICs are from the NAM (7), a combination of the NAM with a SREF perturbation (6), or the GFS (1).
OU/CAPS	Individual SSEF members with FV3 core: SSEF-Thomp SSEF-Morr SSEF-NSSL	3 km	60 h forecasts, initiated at 00 UTC.	core_cntl has Thompson microphysics core_mp1 has NSSL microphysics core_mp2 has Morrison mircosphics

4. Meteorological Highlights from Throughout the Experiment

There were several notable heavy rain and flash flooding events that occurred during the 2019 FFaIR experiment. This included Hurricane Barry, which is discussed in greater detail below. The first two weeks of the experiment, (refer to Fig. 8), were dominated by 850 mb troughing over the western CONUS and ridging over the eastern portion, with zonal flow aloft. During the second half of the experiment, the majority of the CONUS was under a strong ridge, with a relatively fast moving northern stream along the US/Canada border (see Fig 9).

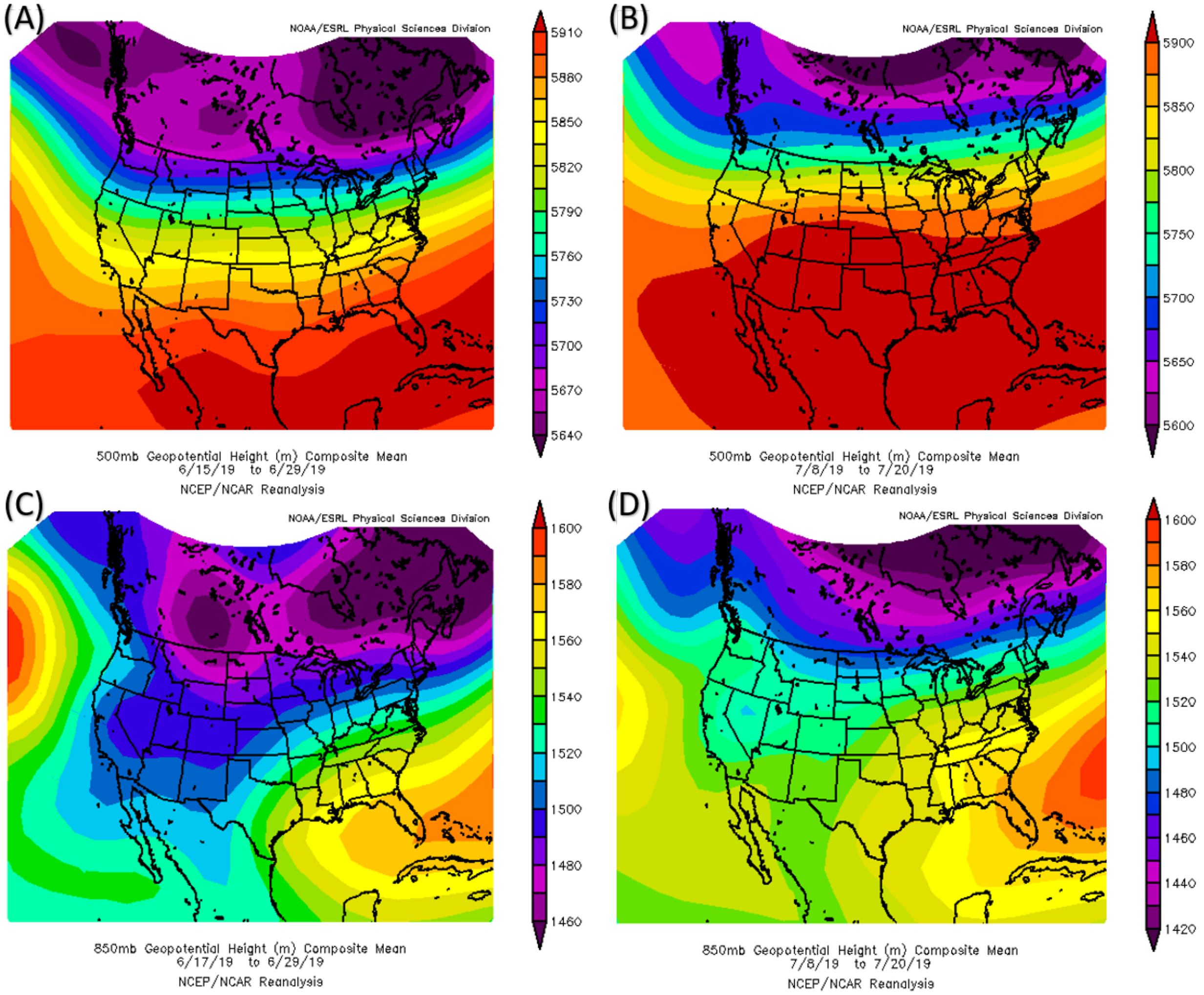

(A)
NOAA/ESRL Physical Sciences Division
5910
5880
5850
5820
5790
5760
5730
5700
5670
5640
500mb Geopotential Height (m) Composite Mean
6/15/19 to 6/29/19
NCEP/NCAR Reanalysis
(B)
NOAA/ESRL Physical Sciences Division
5900
5850
5800
5750
5700
5650
5600
500mb Geopotential Height (m) Composite Mean
7/8/19 to 7/20/19
NCEP/NCAR Reanalysis
(C)
NOAA/ESRL Physical Sciences Division
1600
1580
1560
1540
1520
1500
1480
1460
850mb Geopotential Height (m) Composite Mean
6/17/19 to 6/29/19
NCEP/NCAR Reanalysis
(D)
NOAA/ESRL Physical Sciences Division
1600
1580
1560
1540
1520
1500
1480
1460
1440
1420
850mb Geopotential Height (m) Composite Mean
7/8/19 to 7/20/19
NCEP/NCAR Reanalysis

***Figure 8:** (A)-(B) 500 mb mean geopotential height composites and (C)-(D) 850 mb mean geopotential height composites. (A) and (C) are the mean composite over the first half of the 2019 FFaIR experiment, from 17 June to 29 June 2019. (B) and (D) are the mean composite over the second half of the 2019 FFaIR experiment, from 8 July to 20 June 2019. Composite images were generated from the NCEP/NCAR Reanalysis provided by NOAA/ESRL/Physical Sciences Division (https://www.esrl.noaa.gov/psd/data/composites/day/).*

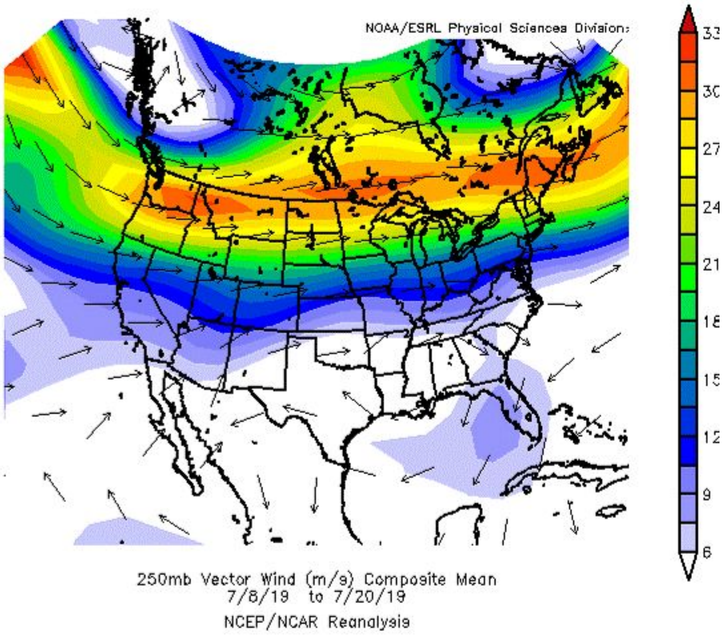

NOAA/ESRL Physical Sciences Division
33
30
27
24
21
18
15
12
9
6
250mb Vector Wind (m/s) Composite Mean
7/8/19 to 7/20/19
NCEP/NCAR Reanalysis

Figure 9: *250 mb vector wind (m/s) composite mean for the second half of the 2019 FFaiR experiment, from 8 July to 20 June 2019. Composite image was generated from the NCEP/NCAR Reanalysis provided by NOAA/ESRL/Physical Sciences Division (https://www.esrl.noaa.gov/psd/data/composites/day/).*

As stated, the dominating 850 mb synoptic pattern for the first two weeks of FFaIR was a trough in the west and a ridge in the east. This pattern resulted in the same regions of the CONUS, especially over the Central Plains, often being at risk for heavy rainfall and flooding for several days in a row, with multiple mesoscale convective systems (MCSs) moving over the same locations. Moreover, there were numerous instances in which a decaying MCS in the morning would leave a mesoscale convective vortex (MCV) that convection would reinitiate along later that evening/night. This scenario was so prevalent during the first two weeks, particularly during Week 2, that one of the participants noted that they did not think that they had ever used the term “MCV” as much as during that week of FFaIR. Although the overall 850 mb pattern between Week 1 and Week 2 was similar, the ridge over the southeastern United States had built farther north and east by the start of Week 2. This resulted in an overall shift of the trough/ridge pattern northward, thus altering the track of the daily MCSs farther north.

Since the orientation of the trough/ridge pattern focused heavy rainfall across the same region for multiple days, antecedent conditions were a major concern of the participants during the first half of FFaIR. In fact, it was often a larger concern than the meteorological setup itself. During Week 1, a slow moving cold front progressed southeastward out of the Northern Plains into the Midwest, while a stationary boundary was present from southern MO through the central Ohio River Valley into NJ (refer to Fig. 10A). These persistent boundaries resulted in the same regions under the threat of heavy rainfall and flooding throughout the week. As can be seen in Fig 11 nearly every day of Week 1 there was a chance for excessive rainfall over the

Central Plains and from the Saint Louis region to the Mid-Atlantic into New England. The pattern finally broke late Thursday night when the two boundaries merged across MO (Fig. 10B). This resulted in a high risk threat for flooding on Friday June 21.

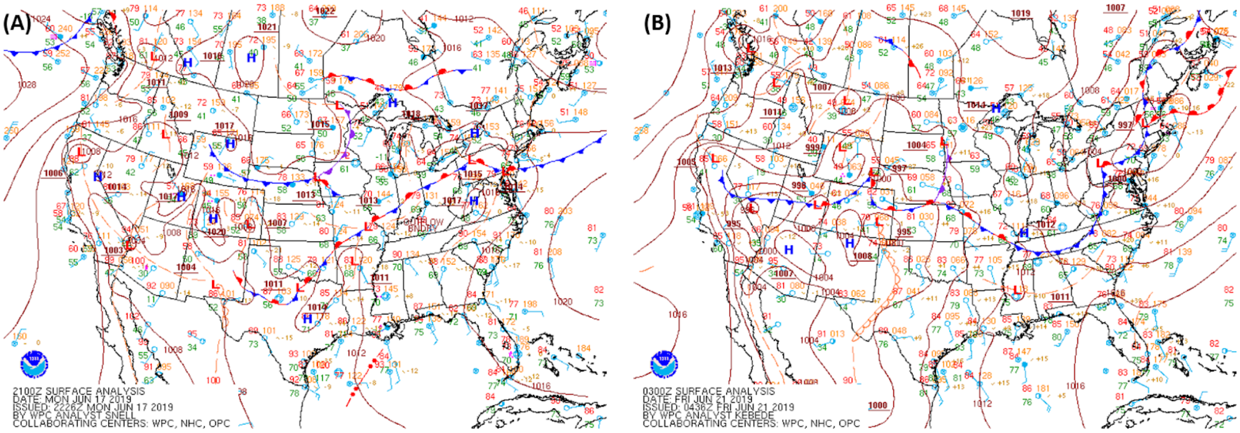


Figure 10: *WPC surface analysis for (A) 17 June 2019 valid at 2100 UTC and (B) 21 June 2019 valid at 0300 UTC. Downloaded from the WPC Surface Analysis Archive webpage* [*https://www.wpc.ncep.noaa.gov/archives/web_pages/sfc/sfc_archive.php*](https://www.wpc.ncep.noaa.gov/archives/web_pages/sfc/sfc_archive.php)*.*

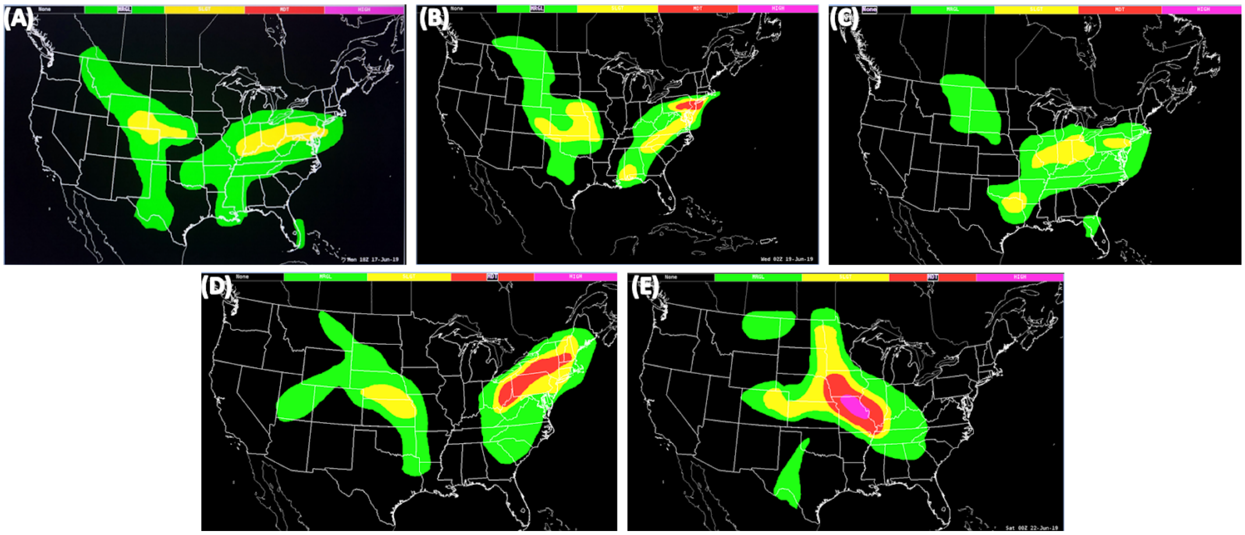


Figure 11: *Experimental EROs issued by the Week 1 participants of 2019 FFaIR valid: (A) 1500 UTC 17 June 2019 to 1200 UTC 18 June 2019, (B) 1500 UTC 18 June 2019 to 1200 UTC 19 June 2019, (C) 1500 UTC 19 June 2019 to 1200 UTC 20 June 2019, (D) 1500 UTC 20 June 2019 to 1200 UTC 21 June 2019, and (E) 1500 UTC 21 June 2019 to 1200 UTC 22 June 2019.*

By Week 2 of the experiment, the ridge over the southeastern U.S. had strengthened and broadened, extending from TX to the Great Lakes. This resulted in a “ring of fire” around the periphery of the ridge. Convective initiation was especially active from AR to MN, with multiple MCSs moving through these regions. In addition to the “ring of fire,” there was a boundary that developed across the Northern Plains into the Northern Rockies for the second half of the week, driving heavy rainfall across the region. The experimental EROs issued by the Week 2 FFaIR participants can be seen in Fig. 12.

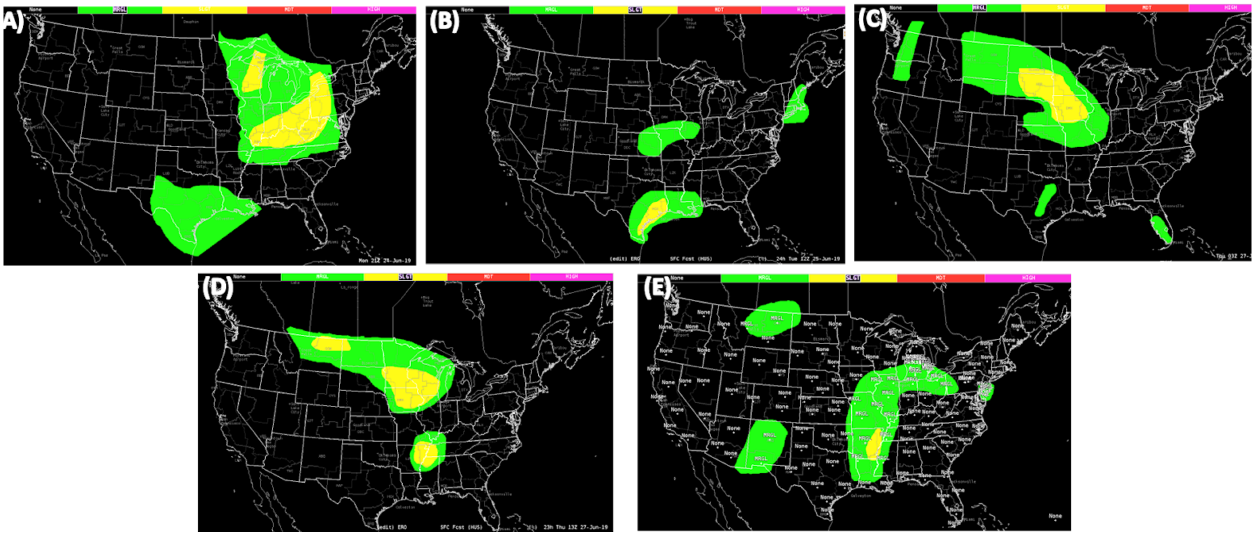
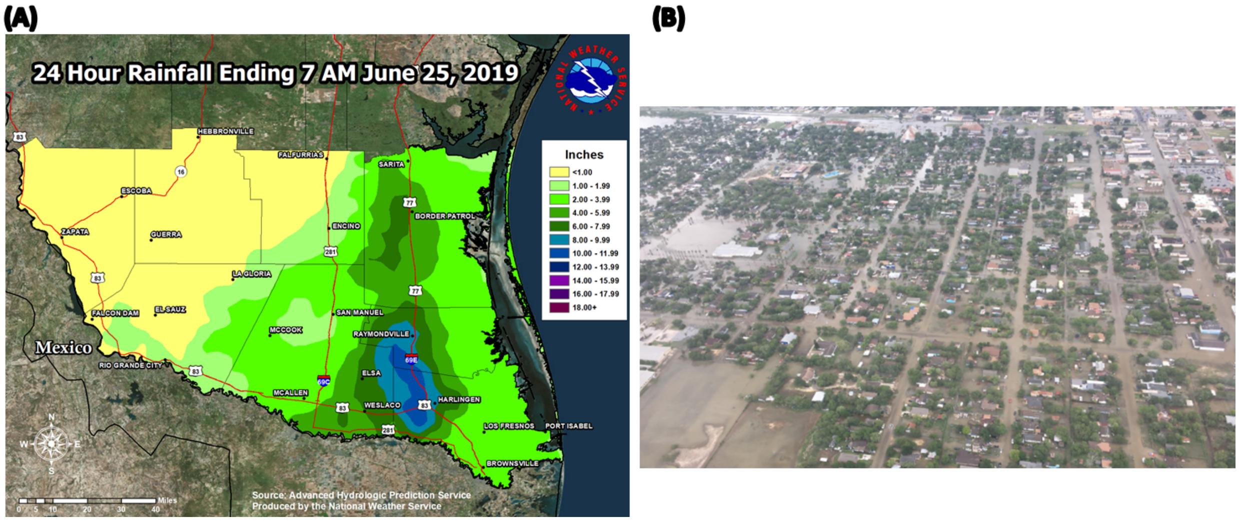


Figure 12: *Experimental EROs issued by the Week 2 participants of 2019 FFaIR valid: (A) 1500 UTC 24 June 2019 to 1200 UTC 25 June 2019, (B) 1500 UTC 25 June 2019 to 1200 UTC 26 June 2019, (C) 1500 UTC 26 June 2019 to 1200 UTC 27 June 2019, (D) 1500 UTC 27 June 2019 to 1200 UTC 28 June 2019, and (E) 1500 UTC 28 June 2019 to 1200 UTC 29 June 2019.*

There were a few significant events that occurred during Week 2 of FFaIR but only one will be highlighted here; for a complete list of notable events that occurred throughout the experiment refer to Table B.2 in Appendix B. Late in the evening Monday June 24, an extremely isolated event occurred in southeastern Texas. According to the NWS WFO Brownsville/Rio Grande Valley (BRO), over a foot of rain fell in six hours across the region (see Fig. 13A), with more than 15 inches falling near Santa Rosa, TX (NWS BRO, 2019). Additionally, multiple cities set new daily rainfall records, such as Raymondville which saw 9.7 inches between 7am 24 June and 7am 25 June CDT. An aerial image of the city can be seen in Fig. 13B (NWS BRO, 2019).

Although this was an extremely impactful event that occurred during the valid time for Monday’s PFF2 (0000-0600 UTC 25 June 2019), the FFaIR participants did not issue a PFF for this region. Figure 14 shows the deterministic model data available for the participants to use for their forecast. As can be seen, there was not a strong signal among the models suggesting such an event would occur (note that model data from the HRRRv4 was unavailable for this day). Some models had no precipitation occurring during the entire 24 h time period. Others indicated the possibility of an event but there was not agreement among these models about the location or intensity of the event. Furthermore, although not shown, the ensemble guidance showed similar discrepancies about the event. However, despite the wide range of forecast outcomes in regards to southern TX, the group did discuss the potential for an isolated heavy shower but this was based solely on the radar and satellite analysis along with the evaluation of current conditions. Ultimately, since there was no model support, the participants focused on the PA region instead and thus "missed" the event. One likely reason for the models' inability to forecast this event was because the majority of the experimental model runs used in FFaIR are 00z runs. This means that since the event began after 00z the next day, it occurred anywhere from 24 to 36 hours into the model forecast. Correctly forecasting such a small scale event that far out is notoriously difficult for mesoscale models and therefore it is not surprising that they did not correctly capture the event.


24 Hour Rainfall Ending 7 AM June 25, 2019

Source: Advanced Hydrologic Prediction Service
Produced by the National Weather Service

Figure 13: *(A) 24 h rainfall totals across southern Texas, from 1200 UTC 24 June 2019 to 1200 UTC 25 June 2019 and (B) an aerial view of the flooding in the city of Raymondville, TX. Both images curiously of the [NWS WFO Brownsville/Rio Grande Valley, TX](#).*

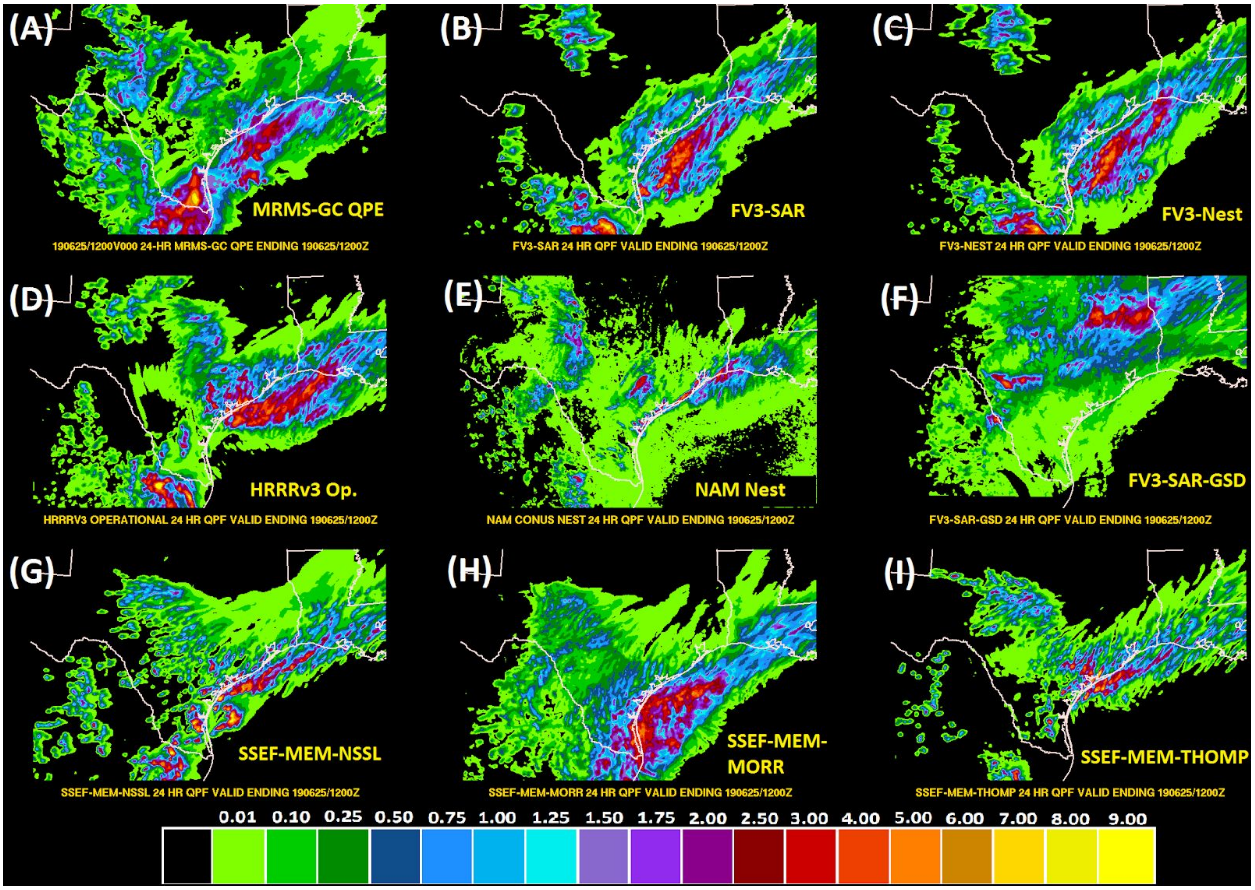
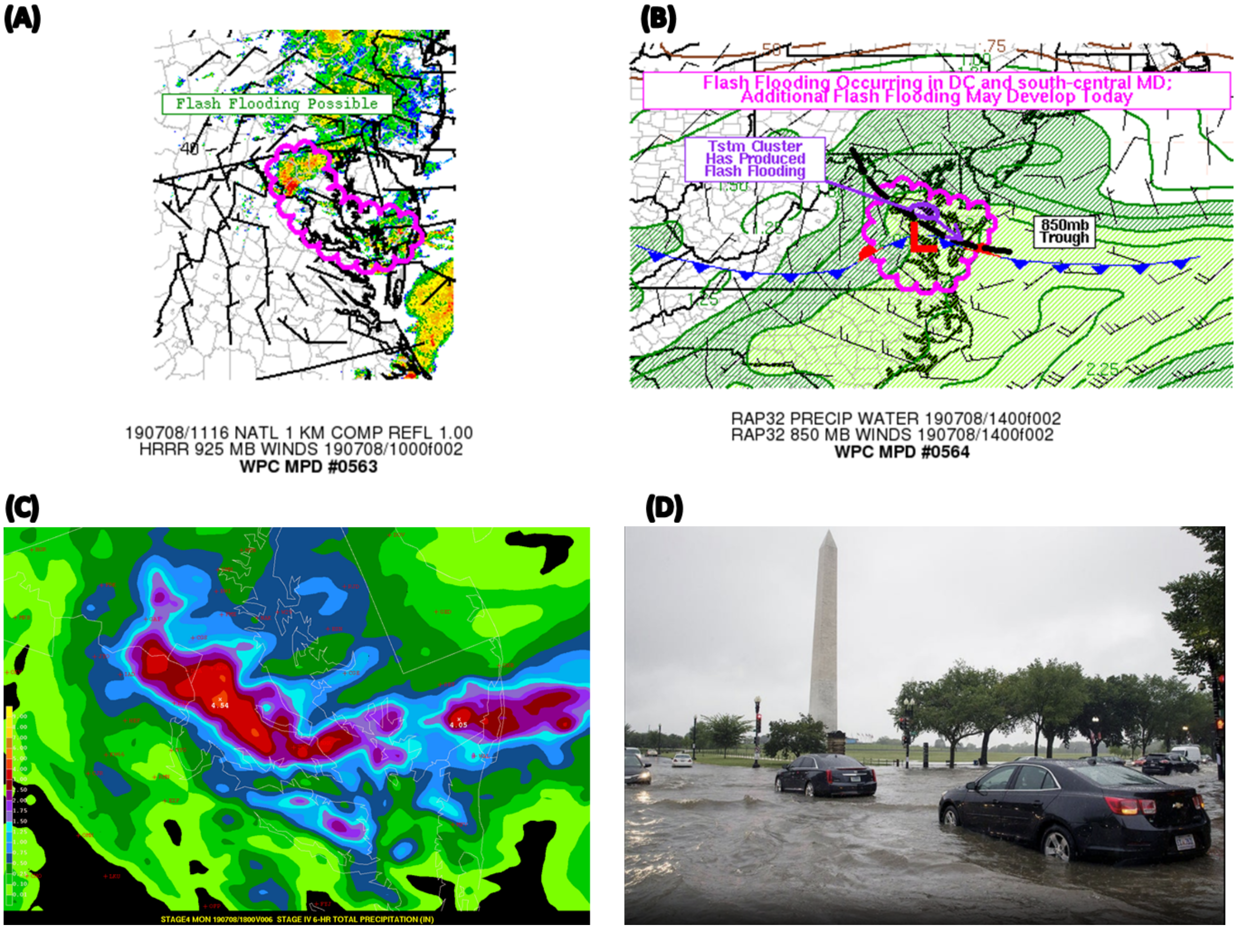

(A)
MRMS-GC QPE
190625/1200V000 24-HR MRMS-GC QPE ENDING 190625/1200Z
(B)
FV3-SAR
FV3-SAR 24 HR QPF VALID ENDING 190625/1200Z
(C)
FV3-Nest
FV3-NEST 24 HR QPF VALID ENDING 190625/1200Z
(D)
HRRRv3 Op.
HRRRV3 OPERATIONAL 24 HR QPF VALID ENDING 190625/1200Z
(E)
NAM Nest
NAM CONUS NEST 24 HR QPF VALID ENDING 190625/1200Z
(F)
FV3-SAR-GSD
FV3-SAR-GSD 24 HR QPF VALID ENDING 190625/1200Z
(G)
SSEF-MEM-NSSL
SSEF-MEM-NSSL 24 HR QPF VALID ENDING 190625/1200Z
(H)
SSEF-MEM-MORR
SSEF-MEM-MORR 24 HR QPF VALID ENDING 190625/1200Z
(I)
SSEF-MEM-THOMP
SSEF-MEM-THOMP 24 HR QPF VALID ENDING 190625/1200Z
0.01 0.10 0.25 0.50 0.75 1.00 1.25 1.50 1.75 2.00 2.50 3.00 4.00 5.00 6.00 7.00 8.00 9.00

Figure 14: *(A) MRMS-GC 24 h QPE valid 1200 UTC 25 June 2019. (B)-(I) Model 24 h QPF valid at the same time, for the following deterministic models: (B) FV3-SAR, (C) FV3-Nest, (D) HRRRv3, (E) NAM Nest, (F) FV3-SAR-GSD, (G) SSEF-NSSL, (H) SSEF-Morr, and (I) SSEF-Thomp.*

Week 3 started off quite eventful for the Washington D.C. area, with a Flash Flood Emergency being issued for the D.C. Metro on July 8, 2019. Across the region, rain rates exceeding 2 inches per hour fell, with WPC issuing two MPDs for the event that are shown in Figs. 15A-B. The first was prior to the system impacting the area (1118 UTC), while the second was after the threat had moved southward (1550 UTC). The latter MPD discussed how much rain had already fallen in the D.C. Metro as well as the additional impacts the slow moving system would cause. One reason this was such a high impact event was because so much water fell in a very short time, during morning rush hour, and in a highly urbanized area. STAGE IV 6 h QPE valid at 1800 UTC on July 8 is plotted in Fig. 15C. The QPE shows that the swath of the heaviest rain fell along the southern portion of the D.C. Metro. Nearly all of the rain seen during those six hours fell in less than an hour. For instance, Reagan International Airport (DCA) measured 3.30 inches between 1255 UTC and 1352 UTC. This resulted in quickly rising flood waters, stranding commuters and bringing the Nation's Capital to a halt. Fig. 15D shows flood waters near the Washington Monument in Washington, D.C.


(A)
Flash Flooding Possible
40
(B)
Flash Flooding Occurring in DC and south-central MD; Additional Flash Flooding May Develop Today
Tstm Cluster Has Produced Flash Flooding
850mb Trough
(C)
STAGE4 MON 190708/1800V006 STAGE IV 6-HR TOTAL PRECIPITATION (IN)
(D)

190708/1116 NATL 1 KM COMP REFL 1.00
HRRR 925 MB WINDS 190708/1000f002
WPC MPD #0563

RAP32 PRECIP WATER 190708/1400f002
RAP32 850 MB WINDS 190708/1400f002
WPC MPD #0564

Figure 15: *(A) WPC MPD #0563 valid 1118 UTC to 1618 UTC 08 July 2019. (B) WPC MPD #0564 valid 1550 UTC to 2000 UTC 08 July 2019. (C) STAGE IV 6 h QPE valid at 1800 UTC 08 July 2019, focused over the D.C., Maryland, and Virgina region. (D) Image of flooding near the Washington Monument courtesy of The Weather Channel (TWC, 2019).*

Unfortunately, because the D.C. event was ongoing when week 3 began the participants were unable to utilize the experimental data to forecast for the event. They were however able to evaluate model performance after the event. A zoomed in forecast of the FV3-SAR and HRRRv4 24 h QPF for the event can be seen in Fig. 16. In this area, participants felt that most of the deterministic models hinted at an isolated heavy rainfall event, though they felt the models had trouble narrowing the exact location. However, as can be seen in Fig. 17, there were two other flooding events that day were not well predicted by multiple models. But, due to the high impact and the participants' proximity to the event over the DC area, most (if not all) of the evaluation focused on this rainfall event, seeming to inflate the subjective scores when a model "got the D.C. event right." This points to some pitfalls in subjective scoring for an entire CONUS.

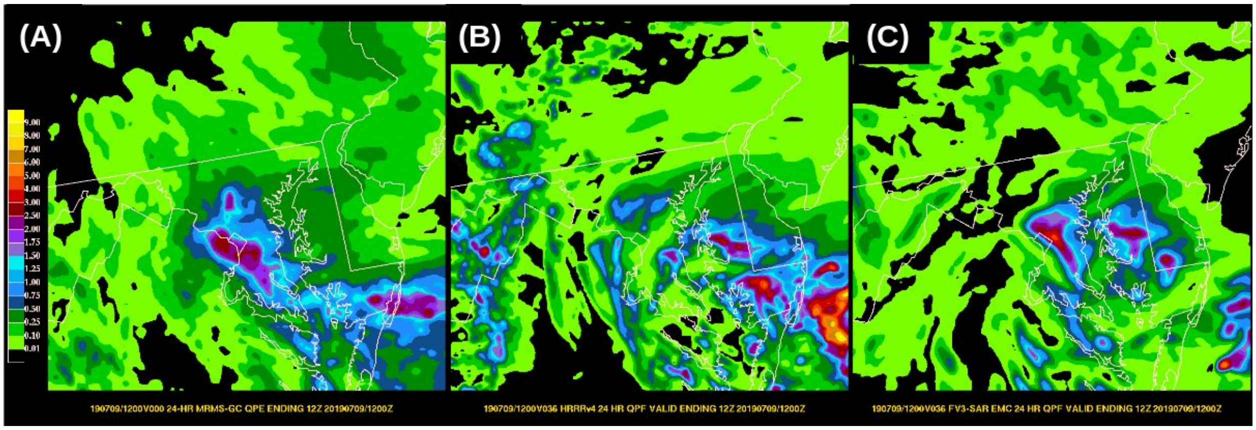

(A)
(B)
(C)
9.00
8.00
7.00
6.00
5.00
4.00
3.00
2.50
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.10
0.01
190709/1200V000 24-HR MRMS-GC QPE ENDING 12Z 20190709/1200Z
190709/1200V036 HRRRv4 24 HR QPF VALID ENDING 12Z 20190709/1200Z
190709/1200V036 FV3-SAR EMC 24 HR QPF VALID ENDING 12Z 20190709/1200Z

Figure 16: *(A) MRMS-GS 24 h QPE, (B) FV3-SAR 24h QPF, and (C) HRRRv4 QPF all valid 1200 UTC 08 July 2019 to 1200 UTC 09 July 2019, zoomed in over the Washington D.C. region.*

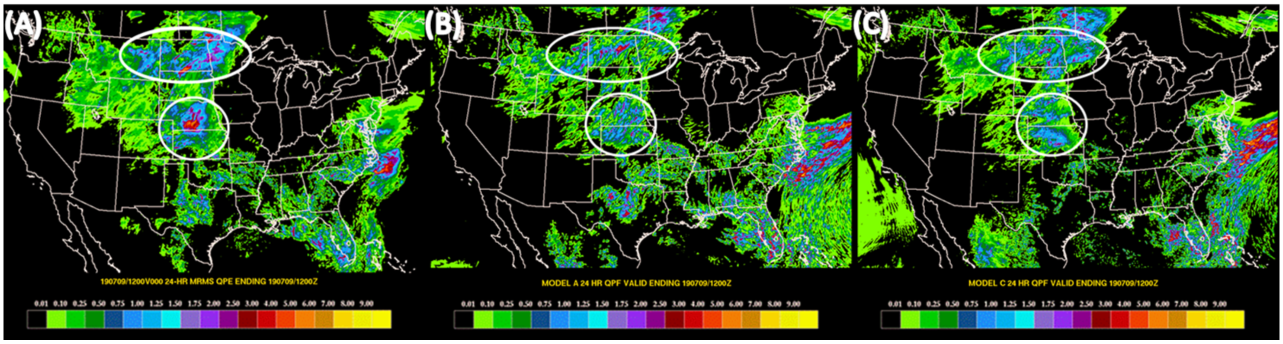

(A)
(B)
(C)
190709/1200V000 24-HR MRMS QPE ENDING 190709/1200Z
MODEL A 24 HR QPF VALID ENDING 190709/1200Z
MODEL C 24 HR QPF VALID ENDING 190709/1200Z
0.01 0.10 0.25 0.50 0.75 1.00 1.25 1.50 1.75 2.00 2.50 3.00 4.00 5.00 6.00 7.00 8.00 9.00

Figure 17: *(A) MRMS-GS 24 h QPE, (B) FV3-SAR 24h QPF, and (C) HRRRv4 QPF all valid 1200 UTC 08 July 2019 to 1200 UTC 09 July 2019. The white circles highlight the two locations in the Central US where major flooding events in addition to the Washington D.C. Flash Flood Emergency occurred.*

The rest of Week 3 continued to be active. Aside from the event that occurred in the D.C. Metro region, the focus for heavy rainfall and flooding during the first couple days of Week 3 was across the Northern Plains. This was driven by a low pressure system that tracked across the U.S./Canada border around the northern extent of a mid-upper ridge. On Wednesday, July 10, precipitation associated with the tropical disturbance that would eventually become Hurricane Barry dropped over 6 inches of rainfall in 6 hours (see Fig. 18) over the New Orleans region. Within some of the convective bands, rainfall rates exceeded 3 inches an hour. Then on July 11, the state of Pennsylvania was hit with multiple rounds of heavy rainfall, resulting in a Flash Flood Emergency being issued for eastern PA and widespread flooding across southwestern PA that unfortunately resulted in two deaths. All the experimental EROs that were issued by the Week 3 participants can be seen in Fig. 19.

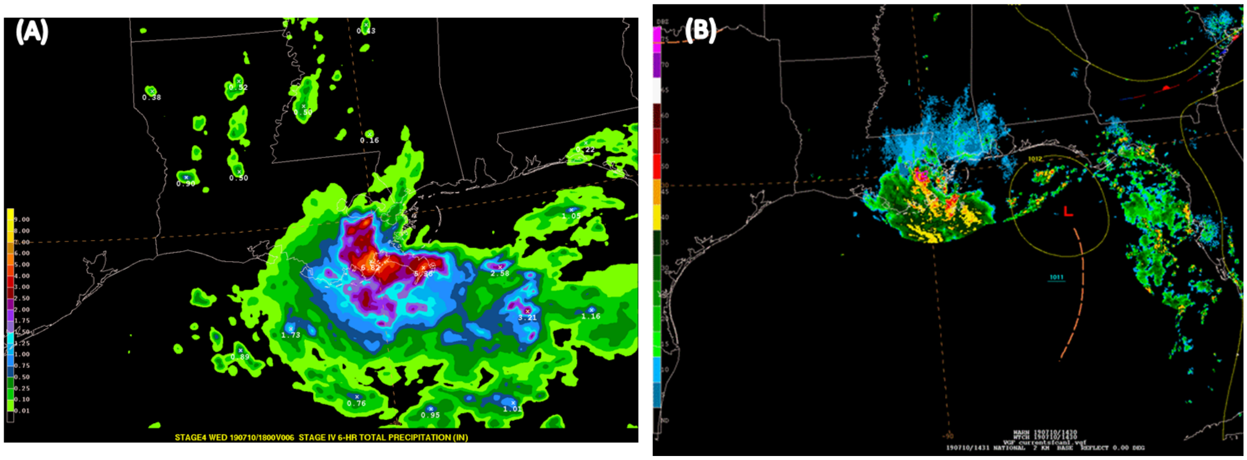

(A)
(B)
STAGE4 WED 190710/1800V006 STAGE IV 6-HR TOTAL PRECIPITATION (IN)
L
1012
1011

Figure 18: *(A) STAGE IV 6 h QPE valid at 1800 UTC and (B) composite reflectivity at 1431 UTC with WPC 1200 UTC surface analysis overlaid for 10 July 2019.*

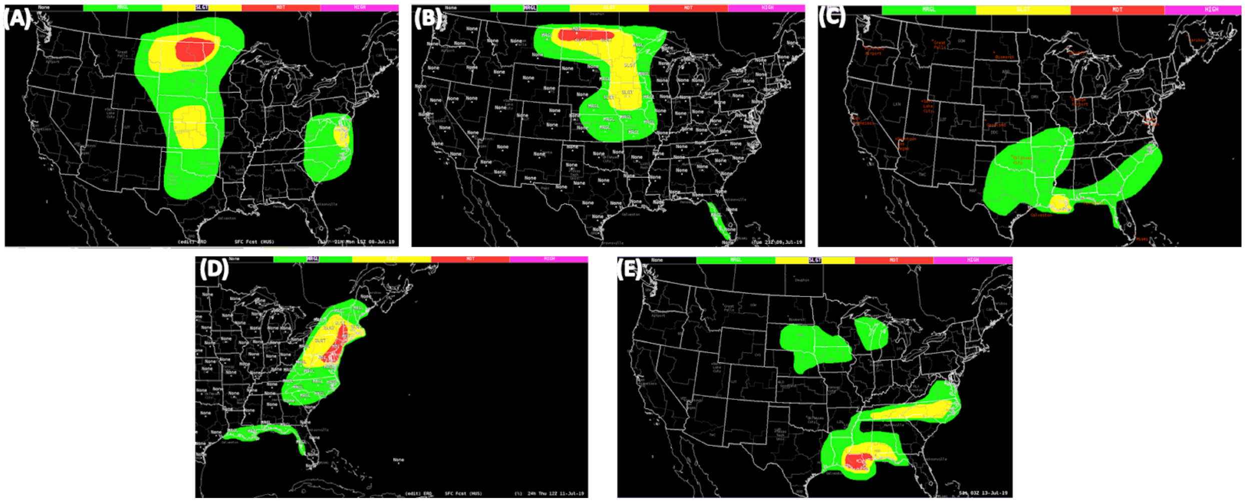


Figure 19: *Experimental EROs issued by the Week 3 participants of 2019 FFaIR valid: (A) 1500 UTC 08 July 2019 to 1200 UTC 09 July 2019, (B) 1500 UTC 09 July 2019 to 1200 UTC 10 July 2019, (C) 1500 UTC 10 July 2019 to 1200 UTC 11 July 2019, (D) 1500 UTC 11 July 2019 to 1200 UTC 12 July 2019, and (E) 1500 UTC 12 July 2019 to 1200 UTC 13 July 2019.*

Like the end of Week 3, the greatest chances for heavy rainfall and flash flooding during the first few days of Week 4 were associated with Hurricane Barry; this will be discussed in further detail below. In addition to the threats driven by Barry, there were multiple rounds of MCS development along the northern edge of the mid-upper level ridge extending from the Dakotas to the Upper Great Lakes. The experimental EROs issued by the Week 4 participants are shown in Fig. 20. An example of a MCS moving across the periphery of the ridge occurred over southeastern MN early in the morning of July 19th. Figure 21A depicts the MPD that was issued by WPC for the event, which was driven by thunderstorms that developed along a warm front located across southern MN and WI. These evolved into a progressive MCS with hourly rain rates of 1-2 inches. However, along the outflow boundary, backbuilding occurred, triggering continual rainfall over a narrow area from southeastern MN into southwestern WI for roughly six hours. This resulted in 4-6 inches of rainfall accumulating the region, see Fig. 21B, with the town of Westby, WI recording 5.74 inches of rainfall by 1200 UTC 19 July 2019 (NWS ARX). This event is discussed further in the results section; the PFF2 for this event can be seen in Fig. 68A.

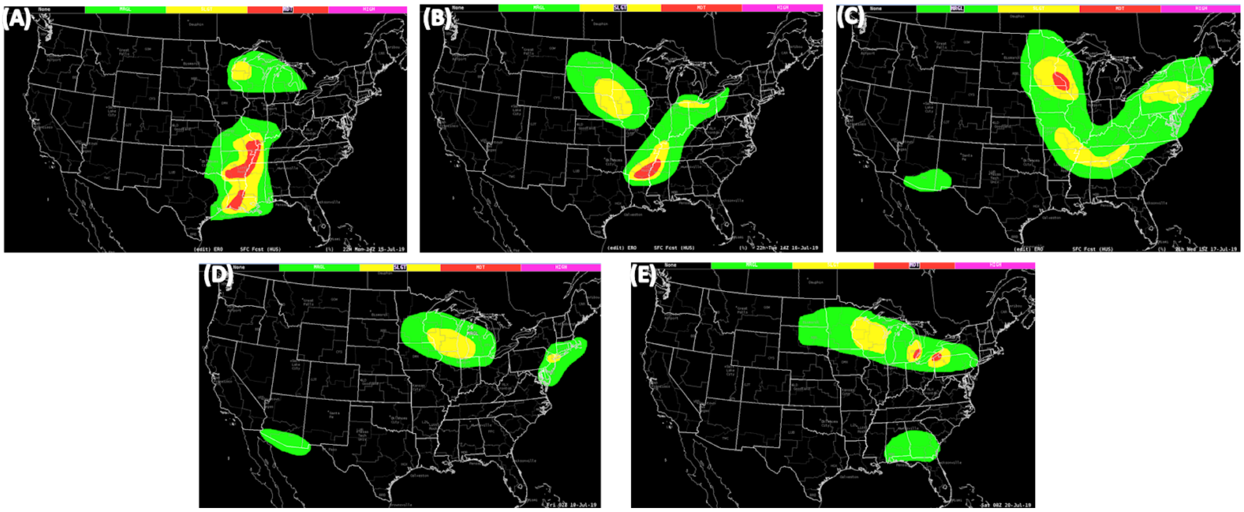


Figure 20: *Experimental EROs issued by the Week 4 participants of 2019 FFaIR valid: (A) 1500 UTC 15 July 2019 to 1200 UTC 16 July 2019, (B) 1500 UTC 16 July 2019 to 1200 UTC 17 July 2019, (C) 1500 UTC 17 July 2019 to 1200 UTC 18 July 2019, (D) 1500 UTC 18 July 2019 to 1200 UTC 19 July 2019, and (E) 1500 UTC 19 July 2019 to 1200 UTC 20 July 2019.*

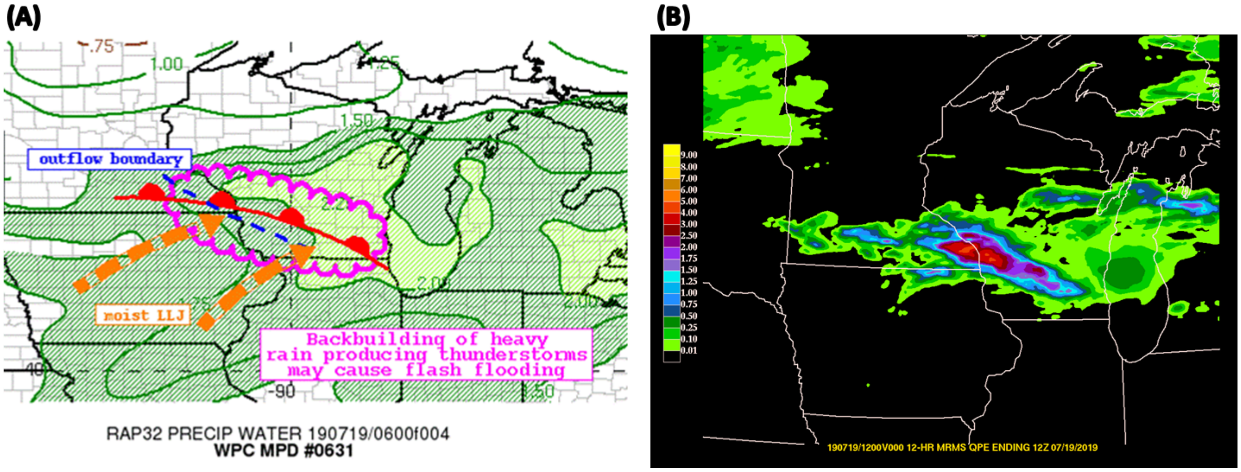

(A)
(B)
.75
1.00
1.50
2.25
2.00
1.75
outflow boundary
moist LLJ
Backbuilding of heavy rain producing thunderstorms may cause flash flooding
40
-90
RAP32 PRECIP WATER 190719/0600f004
WPC MPD #0631
9.00
8.00
7.00
6.00
5.00
4.00
3.00
2.50
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.10
0.01
190719/1200V000 12-HR MRMS QPE ENDING 12Z 07/19/2019

Figure 21: *(A) WPC MPD #0631 valid 0800 UTC to 1315 UTC 19 July 2019. (B) MRMS-Gauge Corrected 12 h QPE valid from 0000 UTC to 1200 UTC 19 July 2019.*

Hurricane Barry

The weekend between Week 3 and Week 4 Tropical Storm Barry briefly transitioned into a category 1 hurricane right before making landfall Saturday morning, 13 July 2019, along the central coast of LA, near Intracoastal City. Hurricane Barry quickly weakened and was downgraded to a tropical storm just four hours after being upgraded to a hurricane. By Sunday afternoon Barry had weakened further, becoming a tropical depression and was located near the western border of LA and AR. As can be seen in Fig. 22 it then gradually tracked northward through AR where it slowly transitioned into a post-tropical system. It officially was deemed a post-tropical system on Monday afternoon, July 15. The remnants of the system then moved eastward across the northern periphery of the ridge with the rest of the synoptic flow. The last weather advisory issued for the remnants of Barry was at 2100 UTC on Wednesday, July 15; at this point WPC noted that the center of the remnants were located over OH.

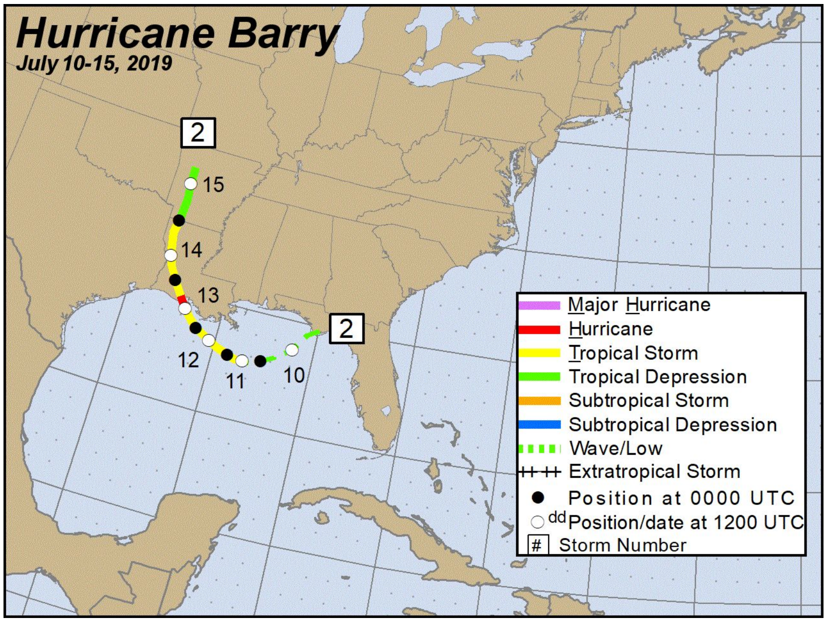


Figure 22: *The smoothed track of Hurricane Barry July 10 to July 15 2019 released by the National Hurricane Center; courtesy of https://www.weather.gov/lch/2019Barry.*

Hurricane Barry offered a unique opportunity for the participants of FFaIR to observe how a national center, specifically WPC, operates during a high impact event. This included witnessing collaboration calls with other national centers such as the National Hurricane Center and the Storm Prediction Center, as well as the extra staff, including managers, needed to ensure everything from normal operations to media interviews are covered. Most importantly, they were able to listen to the discussion that occurred in the days leading up to the event and how the centers worked together to provide clear and consistent information to NWS partners and the public.

As already noted, even though Hurricane Barry did not become a tropical system until the end of Week 3 of FFaIR, the system was largely discussed throughout the week. During this time, the low pressure system was sitting over the northern portion of the Gulf of Mexico. Since the low was close to the shoreline, the main concern was if convection from the system would make its way onto the Gulf coast, from the Florida Panhandle to eastern TX. As already discussed above, this occurred on the morning of July 10, wehn convective bands made their way on shore, resulting in a Flash Flood Emergency in New Orleans. Although the convection weakened, there remained a threat of additional heavy rainfall across LA and into eastern TX throughout the day. Ultimately, the event that occurred over New Orleans on July 10th was the most significant rainfall associated with Barry during Week 3. However, due to how unorganized Tropical Storm Barry was during Week 3, operational and experimental models, both global and CAM alike, had trouble resolving the path of the storm, how much rain would occur, and if/when convective bands would come ashore. For example, Fig. 23A shows that on July 12 the 24 h QPF from the FV3-SAR suggested that all the heavy precipitation associated with Barry would stay offshore. However the HRRRv3 forecast (Fig. 23B) had convective bands coming onshore from MS to the FL Panhandle, dropping over 3 inches of rain. Meanwhile, the SSEF-NSSL (Fig. 23C) was forecasting over 5 inches of rain across southeastern LA. This unpredictability was a staple throughout the week and kept the participants on their toes.

Hurricane Barry had already been downgraded by the NHC to a tropical depression when Week 4 of FFaIR began on July 15. Despite this, heavy rainfall from Barry was still a major concern for the southern and middle Mississippi Valley. Figure 24 shows that a significant amount of rain had fallen during the 24 hours ending at 1200 UTC on July 15, with the rainfall ongoing and portions of central LA already seeing over a foot and a half of rain. As the day continued, the convective bands across central LA strengthened, with convection also initiating along a boundary to the southwest in southeastern TX (Fig. 25B). As can be seen in Fig. 25A, this resulted in an additional 3-6 inches of rain by 1800 UTC. Furthermore, the models were suggesting the rain would continue. Therefore the Week 4 participants decided to issue their PFF1 over the region, feeling the situation warranted a moderate risk where the greatest amount of rain had already fallen. The PFF1 that was issued can be seen in Fig. 25C.

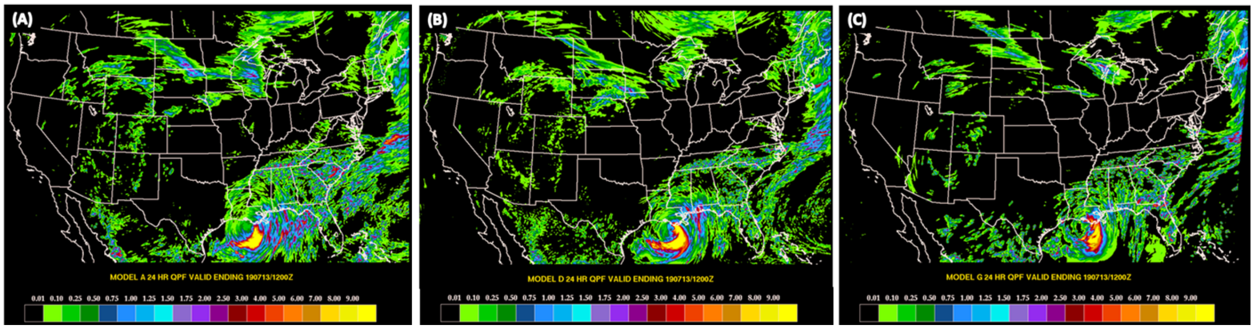


Figure 23: *Model 24 h QPF valid from 1200 UTC July 12 to 1200 UTC July 13, 2019 from (A) the FV3-SAR, (B) the HRRRv3, and (C) the SSEF-NSSL.*

The rainfall that occurred across southwestern AR turned out to be the last big event associated with the decaying tropical system. The remnants of Barry were tracked by the WPC through Wednesday afternoon, with the official post-tropical tag being dropped at 0000 UTC on July 18. The moisture left from Barry then entrained into the westerly flow along the northern periphery of the ridge. In total, the extent of precipitation from Barry extended from the Gulf Coast to southern MI. The highest rainfall total was in Ragley, LA which recorded 23.58 inches of rain throughout the duration of the event (WPC Storm Summary). Other notable amounts were Pass Christian, MS with 13.30 inches, Montrose, AL with 9.33 inches and Baton Rouge, LA with 7.39 inches. For additional rainfall totals refer to WPC's Storm Summary for Barry (WPC Storm Summary).

A similar situation was present over southwestern AR the following morning, July 16. Although Barry had been officially classified as a post-tropical system by this point, the heavy rain threat had not diminished. As can be seen in Fig. 26A, from 1200 UTC 16 July to 1200 UTC 17 July regions of AR had received over a foot of rain. Model guidance indicated the threat would slowly diminish and propagate eastward and northward throughout the day. Therefore the participants shifted the axis of highest probabilities into eastern AR and western MS for the experimental Day 1 ERO shifted (Fig. 26C). However, as can be seen in Fig. 26B, heavy rain continued over the region, particularly western AR, for another six hours or so before it finally moved out of the area. By then, an additional half foot had fallen between 1200 UTC and 1800 UTC. At the conclusion of this event, the rainfall seen across southwestern AR was record setting, with a new state record for the most precipitation from a tropical system being set with 16.59 inches recorded in Dierks, AR (NOAA Climo July 2019).

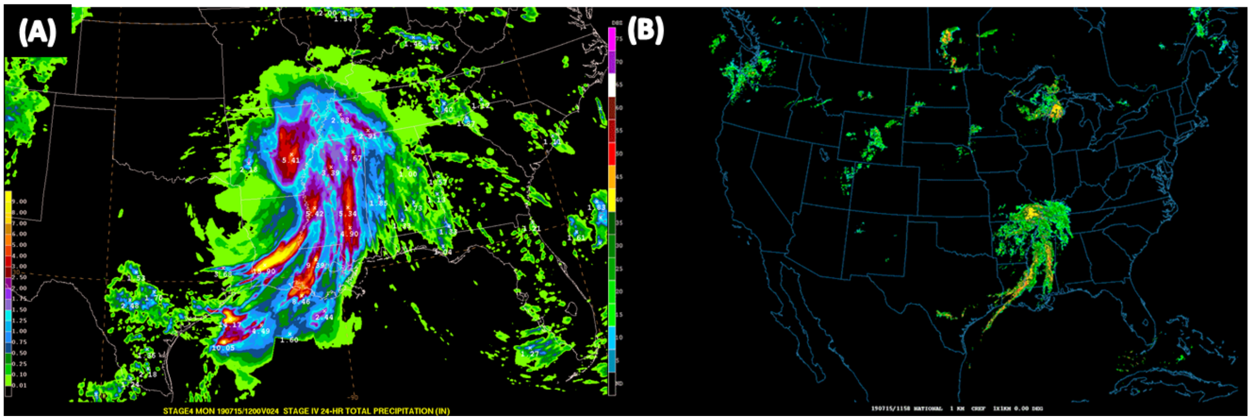

(A)
(B)
STAGE4 MON 190715/1200V024 STAGE IV 24-HR TOTAL PRECIPITATION (IN)
190715/1158 NATIONAL 1 KM CREF 1X1KM 0.00 DEG

Figure 24: *(A) STAGE IV 24 h QPE valid 1200 UTC and (B) composite reflectivity at 1158 UTC 15 July 2019.*

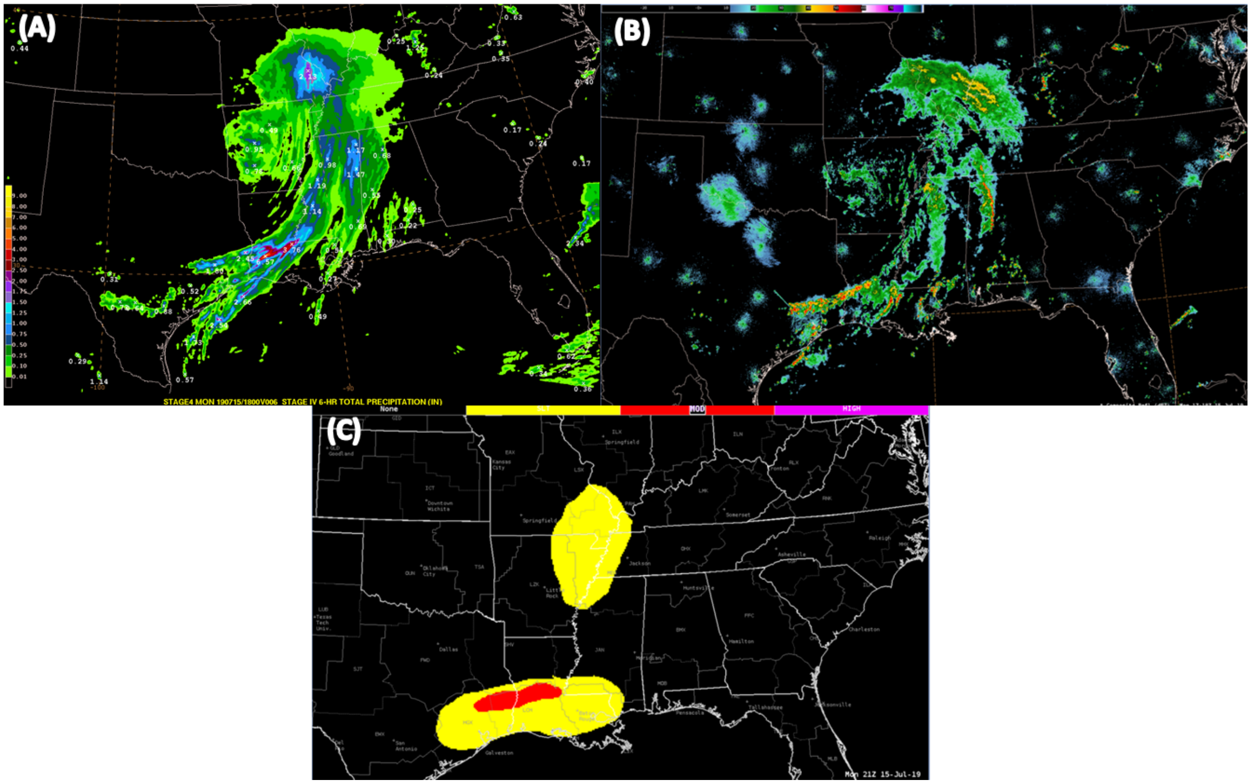

(A)
(B)
(C)
STAGE4 MON 190715/1800V006 STAGE IV 6-HR TOTAL PRECIPITATION (IN)
MOD
HIGH
Mon 21Z 15-Jul-19

Figure 25: *(A) STAGE IV 6 h QPE valid at 1800 UTC and (B) composite reflectivity at 1718 UTC for 15 July 2019. (C) The experimental PFF1 valid from 1800 UTC 15 July 2019 to 0000 UTC 16 July 2019.*

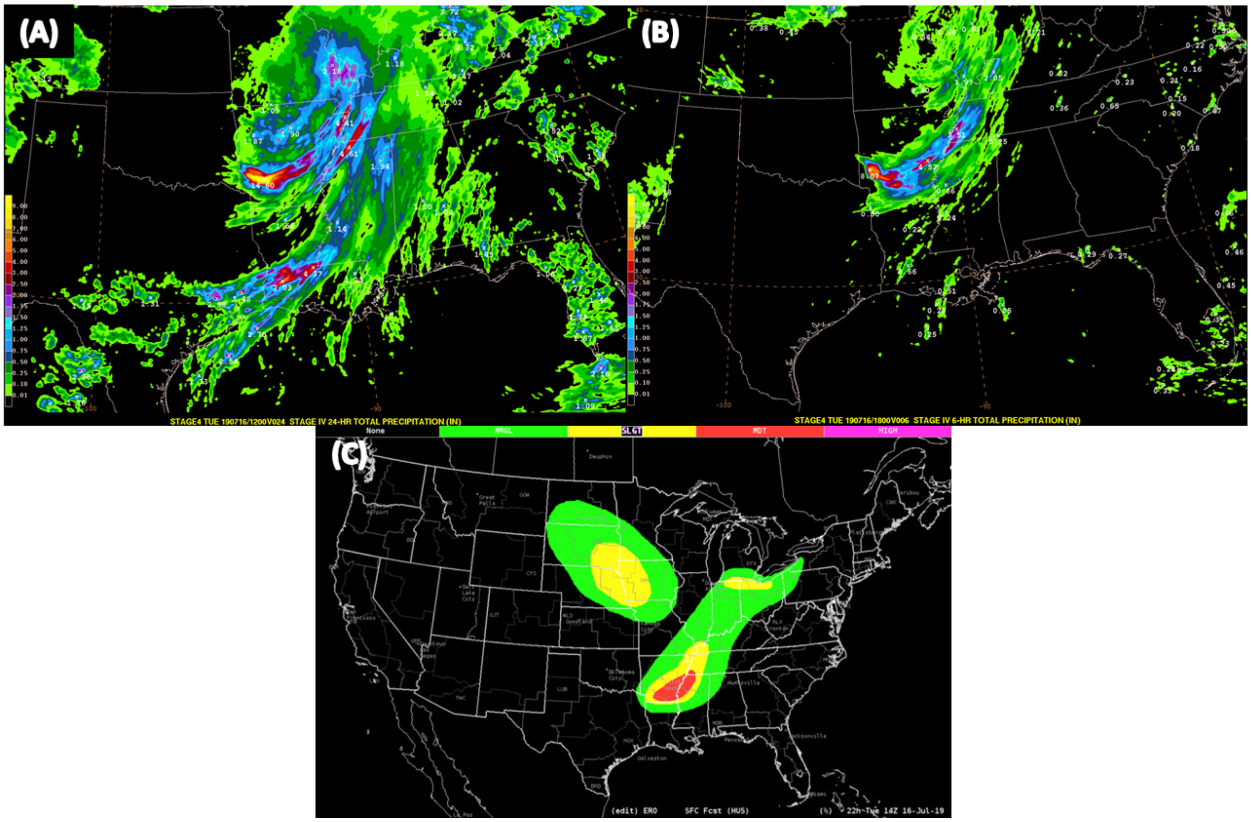

(A)
(B)
(C)
STAGE4 TUE 190716/1200V024 STAGE IV 24-HR TOTAL PRECIPITATION (IN)
STAGE4 TUE 190716/1800V006 STAGE IV 6-HR TOTAL PRECIPITATION (IN)
None
MRGL
SLGT
MDT
HIGH
(edit) ERO
SFC Fcst (HUS)
(%)
22h Tue 14Z 16-Jul-19

Figure 26: *(A) STAGE IV 24 h QPE valid at 1200 UTC and (B) STAGE IV 6 h QPE valid at 1800 UTC for 16 July 2019. (C) Experimentation Day 1 ERO valid from 1500 UTC 16 July 2019 to 1200 UTC 17 July 2019.*

5. Results

The following results section will first discuss the general findings of the experiment before evaluating the various experimental guidance and products individually. As noted previously, since the majority of the model guidance evaluated in FFaIR was experimental, not every product was available every day of the experiment. Because of this, the number of subjective scores collected varied for each tool. Additionally, not every science question was asked every day. Table 1, shown earlier, indicates how many subjective scores were gathered for each model.

Summary of FFaIR Participants' Subjective Verification Scoring

Analysis of the subjective scores for all experimental guidance and forecasts tools found that Week 3 had, on a scale of 0.5-10 for each product, the highest average weekly total score of 6.46. Meanwhile Week 2 participants provided the lowest total average weekly score, 4.72. Week 1 and Week 4 had total average weekly scores of 5.63 and 5.14 respectively[4]. Examining the comments made by each week's participants and evaluating the predictability of the events the participants were scoring, it seems as if the spread between the highest and lowest weekly averages was due to a combination of factors. During Week 2, the models had trouble predicting convective initiation along the western edge of the ridge, explaining why lower scores may have been given by the participants. However, the comments made by the participants suggest that they were also more critical of the models than other participants, even when they verified well based off of the UFV. Therefore, it is likely that the lower weekly average was the result of a combination of model performance during a less predictable pattern which led to a more critical evaluation by the participants themselves.

Opposing this, the higher total weekly average during Week 3 seemed to correspond to the personalities of the participants themselves, though the predictability of the pattern did play a role in the higher values. The higher predictability of the pattern was determined from reviewing the notes taken during Week 3 about the forecasting process, which suggested that there was generally less uncertainty in Week 3 forecasts compared to Week 2. This would suggest that the verification scores for forecasts made during Week 3 would be generally higher than those made during Week 2. However, even when a forecast seemed to be a "miss", the participants of Week 3 still provided high scores for the products. For example, Fig. 27 shows that both the PFF1 valid from 1800 UTC on June 27 to 0000 UTC June 28 (issued by the Week 2 participants but evaluated by Week 3 participants) and the PFF1 valid from 1800 UTC July 8 to 0000 UTC on July 9 (both issued and evaluated by Week 3 participants) did not have any reports within the forecasted risk region. In fact, in the June case there wasn't even rainfall recorded within the southern portion of the forecasted risk region. Even so, on both days the average score from the participants were 6.06 and 6.5, respectively. Furthermore, there was an individual score of an eight for the June event. This suggests that the participants this week were slightly more generous than week two when giving their scores.

[4] The weekly average scores included scores from deterministic and ensemble model guidance, the various experimental Day 1 ERO products, and the experimental PFFs.

The generous scores given by the participants of Week 3 were not just for forecasts issued by other participants or themselves[5]; the group was also less critical of the model forecasts. For instance, for the same two days as discussed above, Fig. 28 shows a comparison of 24 hour QPF and the observed QPE for a couple of cases. For the June case, Fig. 28A-B, "Model A" (FV3-SAR) is shown and, as can be seen, it forecasted a wide area of three or more inches of rainfall across the Midwest, when only a small region saw this amount. However, the average score for this forecast from "Model A" was 6.83. Their reasoning for rating the forecast like this was that although the overall QPF was too wet across the CONUS, placement of the precipitation was good.

Similarly for the July case, see Fig. 28C-D, "Model G" (SSEF-NSSL) incorrectly forecasted most of the convective initiation and propagation across the CONUS, especially across the northern portion of the country. Yet it appeared more accurate with the event that occurred in southern NE, which the group felt was a high impact event. The groups' scores were highly driven by how they felt the model did forecasting this one high impact event, and thus rated the forecast with an average score of 6.47. In general, when comparing how other weeks weighed the importance of getting one major event correct against everything that the model "missed" the group from Week 3 seemed to give more value to the "hit" rather than punishing the model for "misses".

Finally, further support that the personalities of the participants of Week 3 influenced the scores is the fact that the two highest weekly average individual participant scores, were from Week 3, each giving an average score of 7[6]. Opposing this, the two lowest average individual scores came from two different weeks. The lowest was recorded in Week 2 at 4.29. The second lowest individual average score given throughout the week was from Week 4, with an average of 4.30.

[5] During the verification portion of the experiment, the last two forecasts issued (Thursday and Friday) by that week's participants were evaluated by the following week's participants during their first two days (Monday and Tuesday) of verification. The last three days of the week, the participants would evaluate their own forecasts (i.e. on Wednesday they would evaluate their forecast from Monday, then Thursday would evaluate Tuesday's and finally Friday would evaluate Wednesday's forecast. This was done so the data that the forecasts were verified against had time for quality control.

[6] One of these participants only scored 3 of the 5 days. However, the third highest average was also seen in Week 3, 6.87.

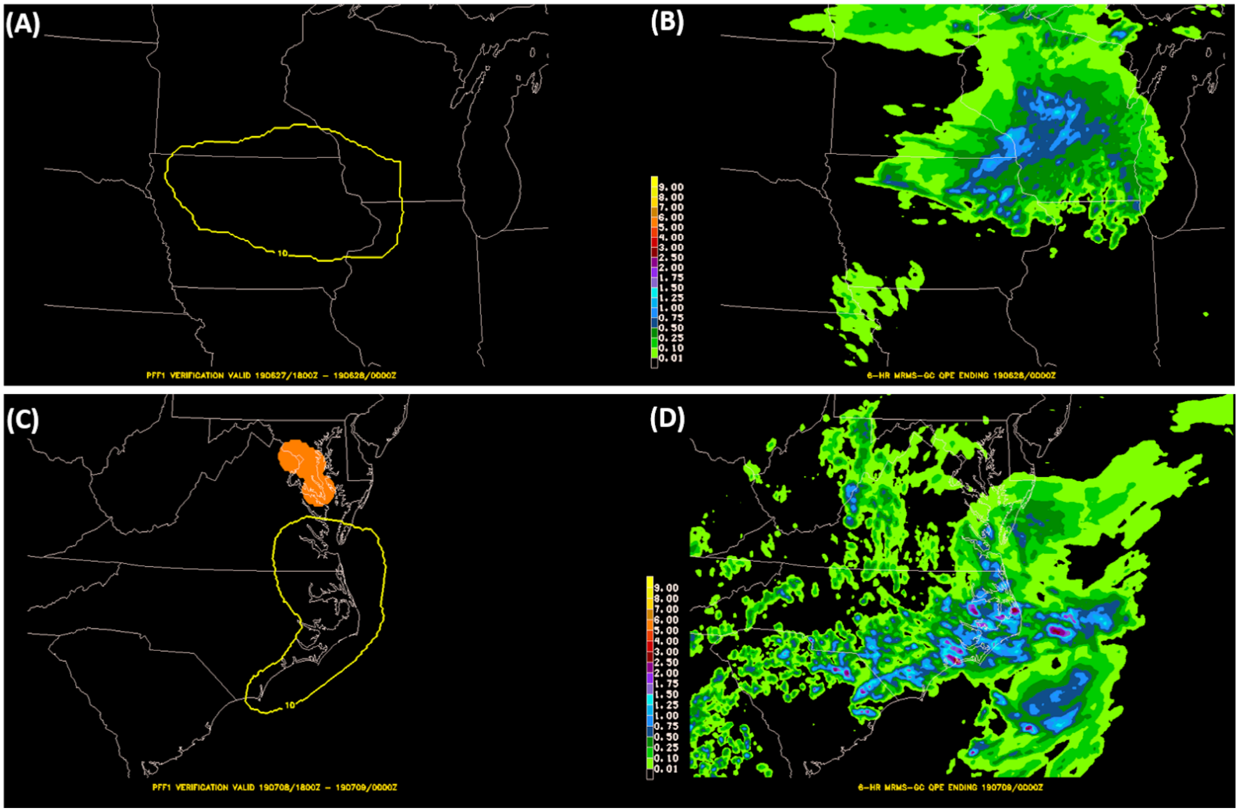

(A)
10
PFF1 VERIFICATION VALID 190627/1800Z - 190628/0000Z
(B)
9.00
8.00
7.00
6.00
5.00
4.00
3.00
2.50
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.10
0.01
6-HR MRMS-GC QPE ENDING 190628/0000Z
(C)
10
PFF1 VERIFICATION VALID 190708/1800Z - 190709/0000Z
(D)
9.00
8.00
7.00
6.00
5.00
4.00
3.00
2.50
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.10
0.01
6-HR MRMS-GC QPE ENDING 190709/0000Z

Figure 27: *Left Column: Experimental PFF1 and right column: 6 h MRMS-GC QPE valid (A)-(B) 1800 UTC 27 June to 0000 UTC 28 June, 2019 and (C)-(D) 1800 UTC 8 July to 0000 UTC 9 July, 2019.*

Deterministic Model Guidance

Analysis of the Subjective Verification

Table 4 and Figs. 29-30 summarize how the individual deterministic models rank amongst each other, based on subjective verification of 24 h QPF guidance. Overall, the HRRRv4 ranked the highest based on the subjective verification scores from the FFaIR participants while the FV3-SAR-GSD had the lowest ranking. The HRRRv4 had a total average during the experiment of 6.56, while the FV3-Nest had the second highest total average score with 6.42. The FV3-SAR-GSD's total average was 4.06, with the SSEF-Thomp having the second lowest average of 4.51. Table 4 shows the total average score each deterministic model received during the experiment. Additionally, Table 4 shows the number of times each model received a rating of 8.5 or better throughout the experiment[7]. Both the experimental and operational versions of the HRRR received a score of 8.5 or better 11 times. Opposing this, the SSEF-Morr, the SSEF-Thomp, and the FV3-SAR-GSD never received a score of 8.5 or higher.

[7] This score range was chosen because it represents the upper 2% of the data.

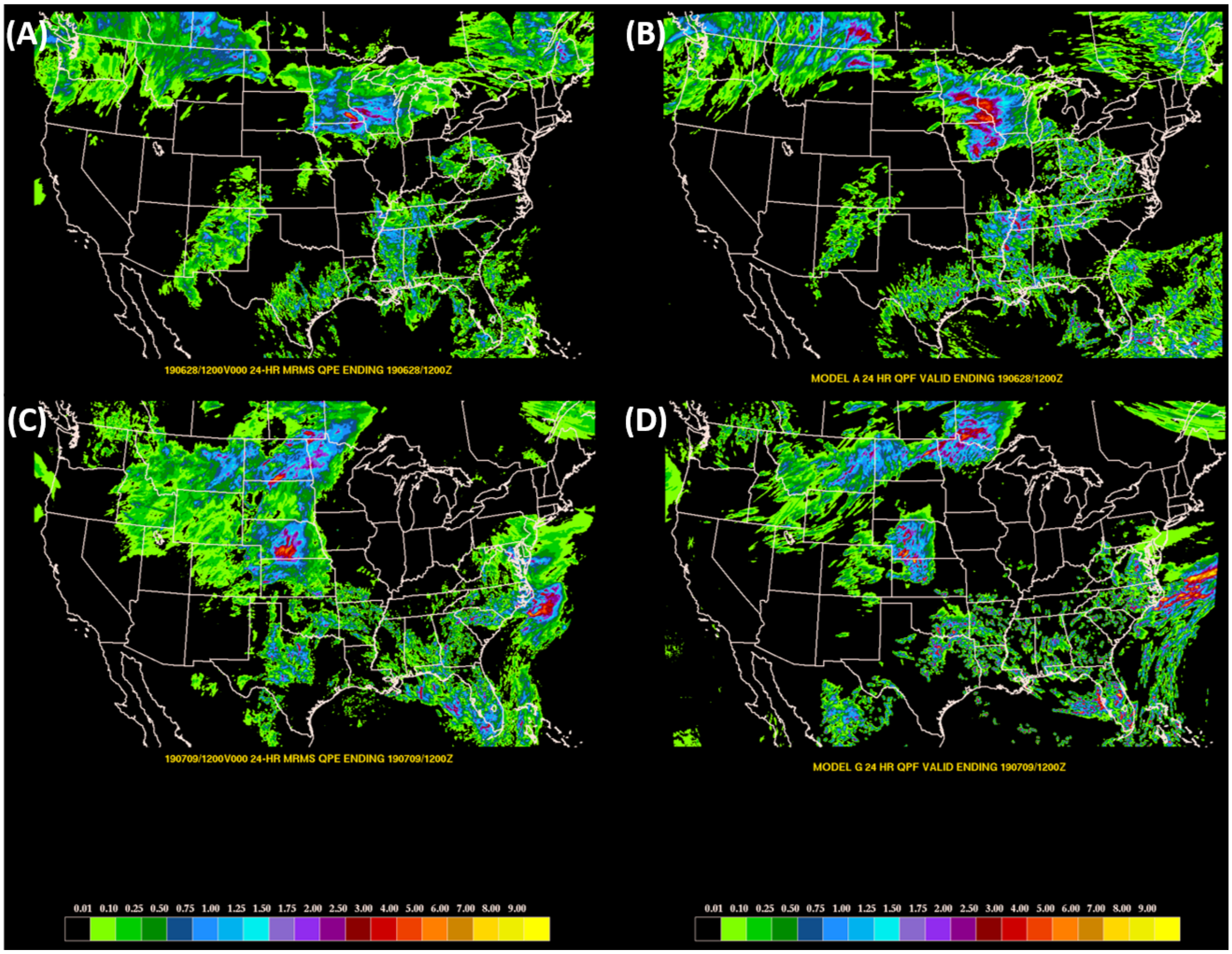


Figure 28: *Left column: 24 hour MRMS-GC QPE and right column model forecasted 24 hour QPF valid (A)-(B) 1200 UTC 27 June to 1200 UTC 28 June 2019 and (C)-(D) 1200 UTC 08 July to 1200 UTC 09 July 2019. (B) is the forecast from "Model A" (FV3-SAR) and (D) is the forecast from "Model G" (SSEF-NSSL).*

Table 4: *The experimental average from the subjective verification portion of the experiment for each deterministic model along with the number of times each model received a score of 8.5 or higher during the course of the experiment.*

Deterministic Model	Experiment Average	Number (percentage) of times model received a score ≥ 8.5
HRRRv4	6.56	11 (9%)
FV3-Nest	6.42	8 (6%)
HRRRv3	6.18	11 (6%)
FV3-SAR	6.17	5 (4%)
NAM-Nest	5.61	5 (3%)
SSEF-NSSL	4.86	3 (2%)
SSEF-Morr	4.67	0
SSEF-Thomp	4.51	0
FV3-SAR-GSD	4.06	0

The spread in the subjective scores over the course of the experiment varied greatly among the models. This can be seen in both the box and whisker plots (Fig 29) and in Fig. 30, which depicts how many times each piece of guidance received a score within a given range. The models with the two highest averages, the HRRRv4 and FV3-Nest, had the smallest spread, with nearly all of their scores falling between 4.5 and 8.5 and a median score of 6.5. The FV3-SAR also had a median score of 6.5 but had a larger spread than the other two models, from 3 to 9. Figure 29 also shows that there were two instances of outliers (depicted by the cross symbol) within the dataset where the scores fell outside the low end of the spread. Both the HRRRv4 and the FV3-Nest had outliers as well, each with four instances, though each of these outliers had a score greater than or equal 3 rather than less than 3, which was seen with the FV3-SAR. Altogether this suggests that although the participants generally felt the FV3-SAR performed as well as the HRRRv4 and FV3-Nest, there were days when its forecast was notably worse than the other two models, thus suggesting it might be less reliable.

Another takeaway from the subjective scoring was the performance of the HRRRv3, and its performance in comparison to the other three top performing models. Table 4 shows that the HRRRv3 had the same number of occurrences of scores at or exceeding 8.5 as HRRRv4, though it also was available to evaluate more often, 189 scores to 129 scores respectively. However, when evaluating the criteria via percentage of scores meeting this threshold, only 6% of the HRRRv3 scores were 8.5 or greater, while HRRRv4 scores meeting this criteria accounted for 9% of the total scores. When compared to the FV3-SAR, the two models had nearly the same total experimental average, 6.18 vs. 6.17 respectively. However, examination of Fig. 30

shows that the score range with the highest number of occurrences throughout the experiment for the HRRRv4, FV3-Nest, and FV3-SAR models was the range from 6.5 and 7.45, while the HRRRv3 saw more scores within the 5.5 to 6.49 range. This resulted in the HRRRv3 having a lower median score than the other three models, 6 vs. 6.5, even though it had a higher count of scores at or exceeding 8.5 than the two FV3 models. This suggests that although there were times when the HRRRv3 was on par with the other three models, even at times outperforming them, routinely its guidance was found to be less useful than the aforementioned models.

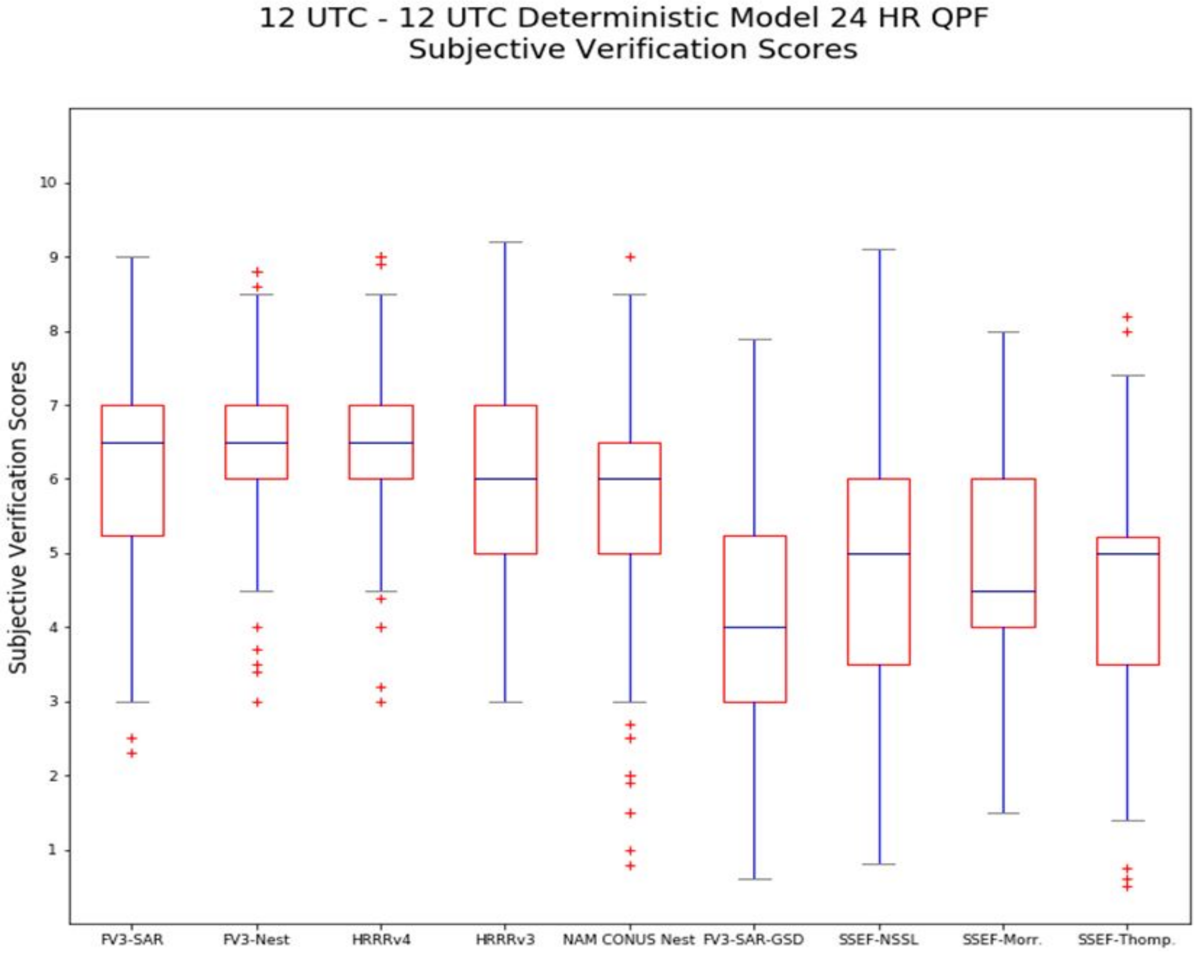


Figure 29: *Box and whisker plot of all the subjective scores given for the deterministic models evaluated in the 2019 FFaIR experiment. Subjective scores were evaluating the 24 h QPF model guidance across the CONUS. All models were initialized at 0000 UTC and the 24 h forecast was from 1200 UTC to 1200 UTC.*

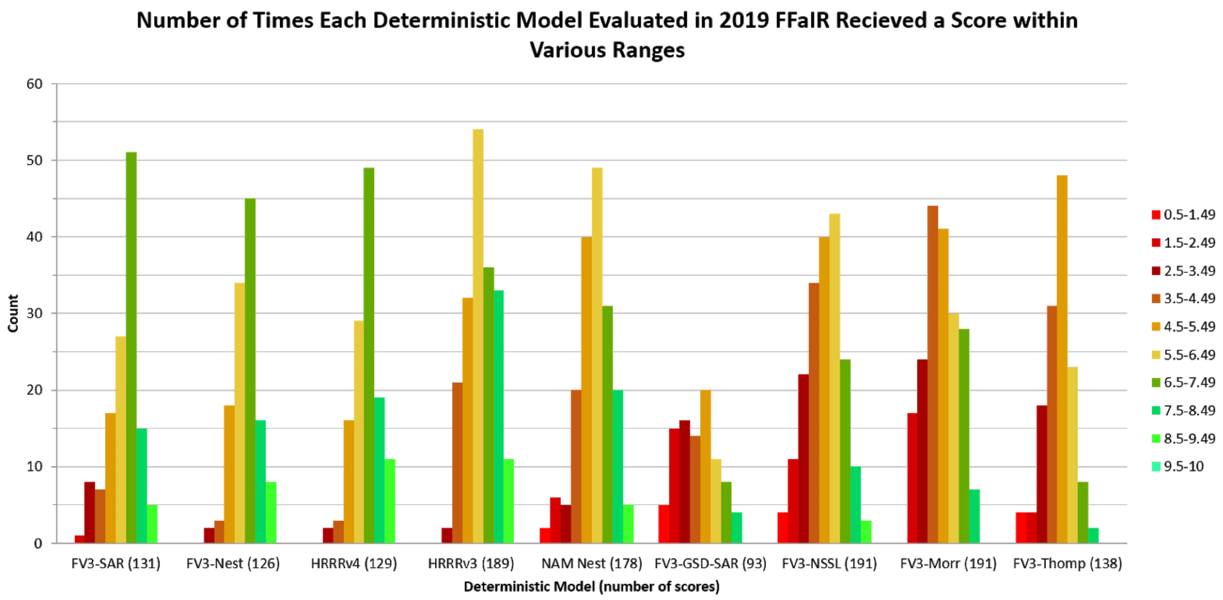


Figure 30: *The number of times each deterministic model received a score that fell within a range of values throughout the 2019 FFaIR experiment. Along the bottom of the graph is the name of the model. In parenthesis is the total number of times the guidance was scored during the experiment.*

The general consensus of the participants during the experiment was that the individual members of the SSEF, which only differed in the microphysics scheme used[8], routinely performed fairly poorly. All three models had a total average subjective score below five, with the SSEF-NSSL having the highest average (4.86) and SSEF-Thomp[9] receiving the lowest average (4.51). However, despite having the lowest average of the three members evaluated, Fig. 29 shows that SSEF-Thomp had the same median as the SSEF-NSSL member (5) while the SSEF-Morr member had a lower median than both of these, 4.5. Interestingly, throughout the experiment it was common for the participants to note that they found the SSEF-Morr member's QPF unrealistic and they would often discard it during the forecasting process. However, as Fig. 30 shows, during the subjective verification it was the only SSEF member to not receive a rating less than 1.5. Conversely, it also did not receive a rating of 8.5 or higher. This suggests that although it was constantly considered a poorer forecast, it was never rated extremely bad (or extremely good) and perhaps provided a more consistent forecast throughout the experiment when compared to the other two members.

[8] The SSEF-Thomp used the Thompson microphysics scheme, the SSEF-Morr used the Morrison Microphysics Scheme and the SSEF-NSSL used the NSSL microphysics scheme. For the full model configuration refer either to Appendix C Table C.3 or Table 5 in the 2019 FFaIR Operations Plan.

[9] The SSEF-Thomp member had 50 fewer scores than the other two models.

Finally, as noted the FV3-SAR-GSD performed the worst of all the deterministic models during the experiment according to the subjective scores from the FFaIR participants. Not only did it have the lowest average, it also had the lowest median, 4, and Fig. 30 shows that it had the lowest bound to its spread. However, it is important to note that this model was available for the participants to evaluate significantly less often than any of the other models. As was shown in Table 1 the FV3-SAR-GSD was only scored 93 times, which is 33 fewer ratings than the next lowest model and 98 fewer than the SSEF-NSSL and SSEF-Morr models. Furthermore, there were many times in which the model data was available for the subjective verification portion of the experiment, but not available for the participants to utilize during the forecasting portion of the experiment. Therefore it is difficult to determine how the participants felt the model did in a real time setting. Even so, it was clear that throughout the experiment the participants felt that the majority of the time the FV3-SAR-GSD failed to provide an accurate 24 hour precipitation forecast; this is further supported through objective verification which is discussed next.

Examples of how the FV3-SAR-GSD performed can be seen in Fig. 31 and Fig. 32. The dates chosen are examples of one of the lower performing days, June 27 (Fig. 31) with an average subjective score of 1.54, and a higher performance day July 12 (Fig. 32) with an average subjective score of 5.88. The 24 h QPF forecast from the FV3-SAR-GSD is compared against the model the participants felt performed the best for each day (FV3-SAR and HRRRv3 respectively) as well as the forecast from the FV3-Morr, which is included as it was also considered a poor performing model. For the June 27 case, the FV3-SAR had an average score of 4.32 (in general all the models struggled with this forecast) while the SSEF-Morr had an average of 3.4. In the July 12 case, the best performing model, the HRRRv3, had an average score of 6.83 while the SSEF-Morr had an average of 4.44. As can be seen from these two cases, the FV3-SAR-GSD often seemed to struggle with low-end precipitation thresholds, predicting widespread light precipitation across large regions. Participants mentioned this over-prediction of light precipitation was a large contributor to the often low scores seen for the model.

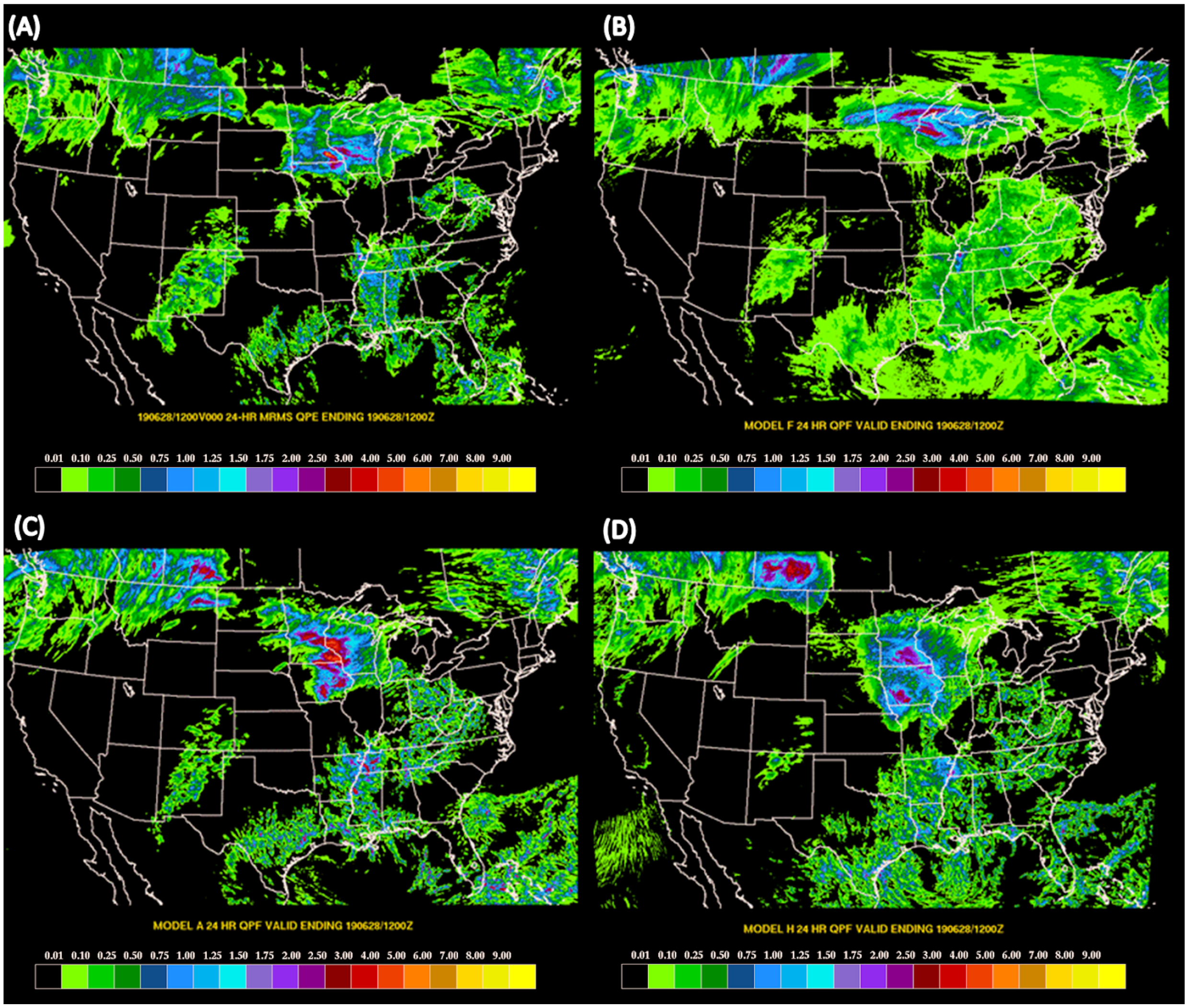


Figure 31: *(A) 24 hour MRMS-GC QPE and (B)-(D) model forecasted 24 h QPF valid 1200 UTC 27 June 2019 to 1200 UTC 28 June 2019 from: (B) FV3-SAR-GSD, (C) FV3-SAR, and (D) SSEF-Morr.*

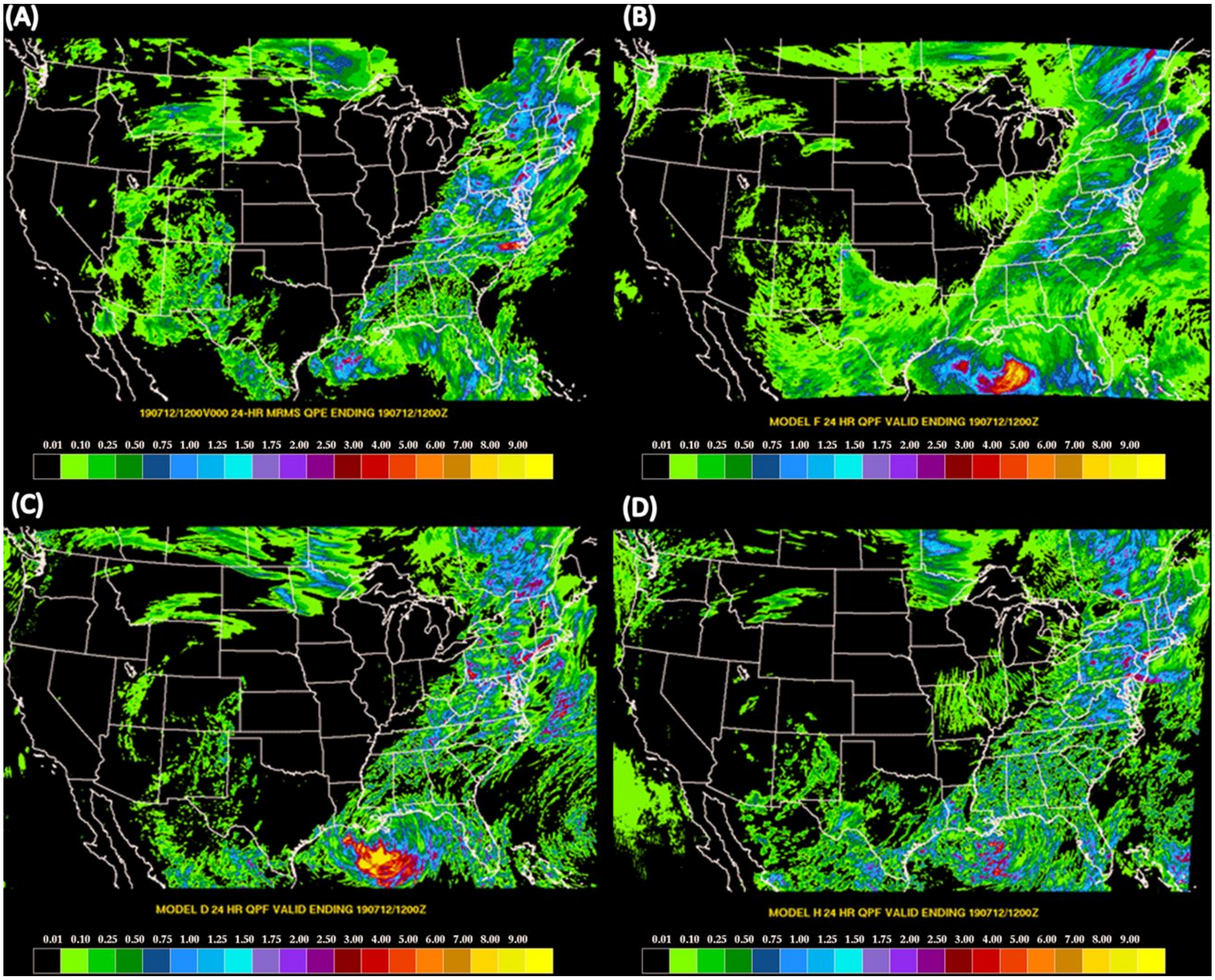


Figure 32: *(A) 24 hour MRMS-GC QPE and (B)-(D) model forecasted 24 hr QPF valid 1200 UTC 11 July 2019 to 1200 UTC 12 July 2019 from: (B) FV3-SAR-GSD, (C) HRRRv3, and (D) SSEF-Morr.*

Analysis of the Objective Verification

Analysis of the 24 h QPF performance diagrams for the deterministic models evaluated in the 2019 FFaIR experiment at thresholds of two inches and lower suggest that the HRRRv4 provided the "best" forecast throughout the experiment. This was in agreement with the results from the subjective evaluation of the models. Figure 33 shows that at the 0.5, 1, and 2 inch thresholds, the HRRRv4 (hrrr_2d in the figure) had the highest CSI and consistently had little bias comparatively, which can be seen by its small distance from the central diagonal line. At higher thresholds, like 4 inches (Fig. 33D), the skillfulness of the HRRRv4 drops off and other models like the NAM-Nest and the FV3-Morr performed better. However, at higher thresholds, model performance decreased across the board and the observational dataset is smaller, so no further analysis will be done. Therefore, unless otherwise stated, all evaluations are referring to 0.5, 1, and 2 inch thresholds.

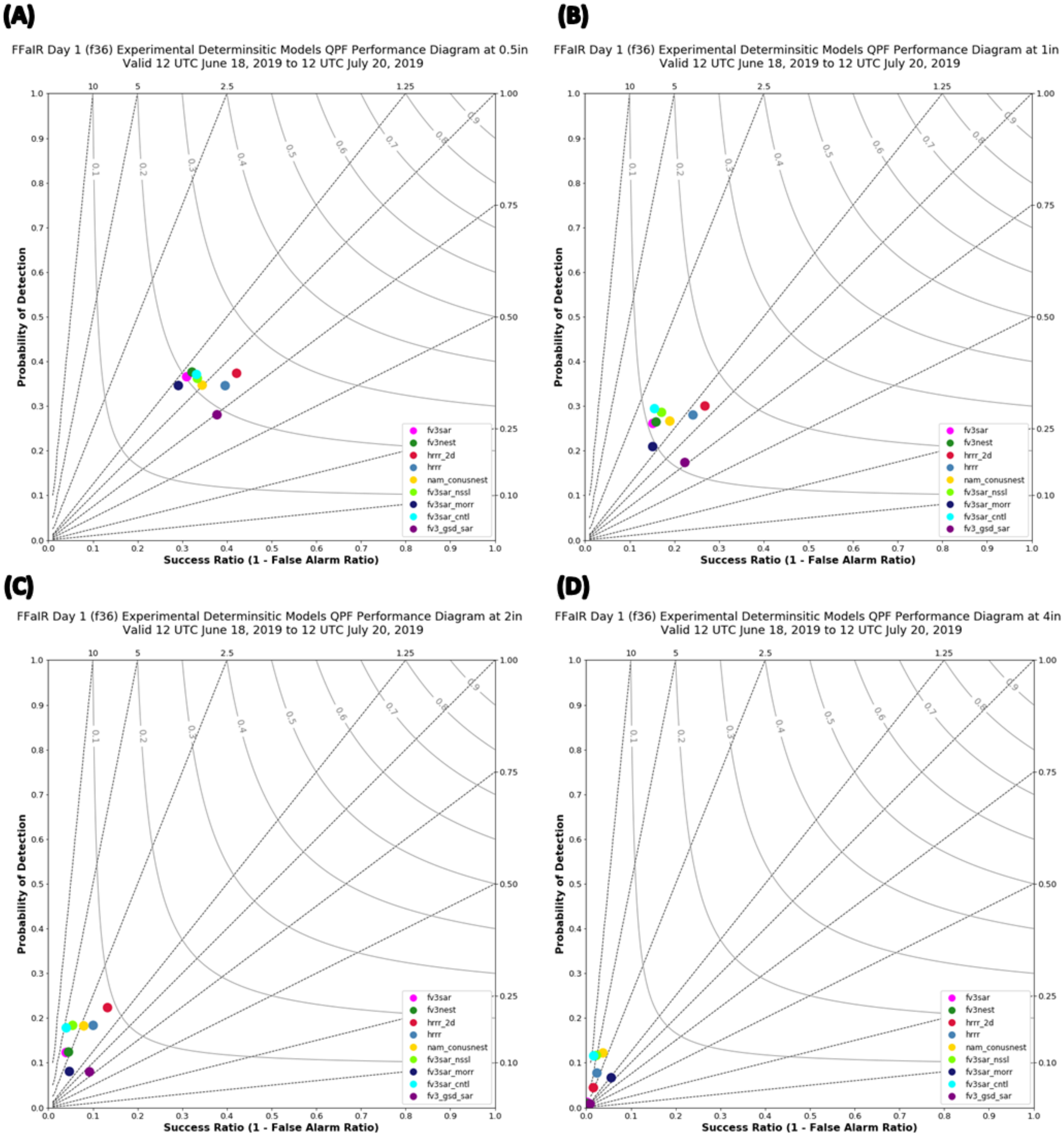


Figure 33: *Performance diagrams for the Day 1, 24 h QPF forecasts valid over the four weeks of the 2019 FFaIR experiment (June 17 to July 20, 2019) for the FV3-SAR (pink), FV3-Nest (dark green), HRRRv4 (red; referenced as hrrr_2d in legend), HRRRv3 (teal; referenced as hrrr in legend), NAM-Nest (yellow), SSEF-NSSL (light green; referenced as fv3sar_nssl in legend), SSEF-Morr (dark blue; referenced as fv3sar_morr in legend), SSEF-Thomp (light blue; referenced as fv3sar_cntl in legend), and the FV3-SAR-GSD (purple). The precipitation thresholds are for: (A) 0.5 inches, (B) 1 inch, (C) 2 inches, and (D) 4 inches.*

Differing from the subjective evaluation, the performance diagram suggests that every model except for the FV3-SAR-GSD and the SSEF-Morr provided a 24 h precipitation forecast that was as good or better than the forecasts by the FV3-Nest and FV3-SAR. For instance, Fig. 33B shows that at the 1 inch threshold, the NAM-Nest had a higher CSI and a smaller wet bias than the FV3-SAR and FV3-Nest. This was extremely noteworthy because the participants regularly commented on the NAM-Nest having a reputation for a high wet bias and the NAM-Nest is being used by EMC as a baseline comparison for the FV3-Nest. This wet bias and low CSI is seen at all the thresholds evaluated in the experiment for the FV3-SAR/Nest.

Another takeaway from the performance diagrams was that both versions of the HRRR outperformed every FV3 CAM that was evaluated, which, as stated, was not the case for the subjective verification. For the lowest threshold, a half inch, the HRRR models had a dry bias but as the threshold increased they began to show a wet bias. However, as can be seen in Fig. 33, nearly every model examined showed a wet bias at higher thresholds. Additionally, aside from the half inch threshold where both HRRR models exhibited a slight dry bias, the FV3-SAR-GSD was the only model to show a dry bias during the experiment. However, differing from the results from subjective verification, the FV3-SAR-GSD did not overwhelmingly underperform compared to other models. At the 1 and 2 inch thresholds the SSEF-Morr, FV3-SAR, and FV3-Nest all had CSI values close to or less than the FV3-SAR-GSD.

Evaluation of MODE along with analysis of the comments made by the participants while using the experimental products in the forecasting process as well as during subjective verification helps shed light on why traditional statistical evaluation differs from forecaster and researcher perspective. Figures 34 to 37 provide an array of examples of how the deterministic models verified using MODE. In general it seems as if the participants focused more on the overall “story the model was telling” rather than a point by point forecast, which is what traditional statistics focus on. As a result, the ranking of model performance was at times considerably different between objective and subjective verification methods.

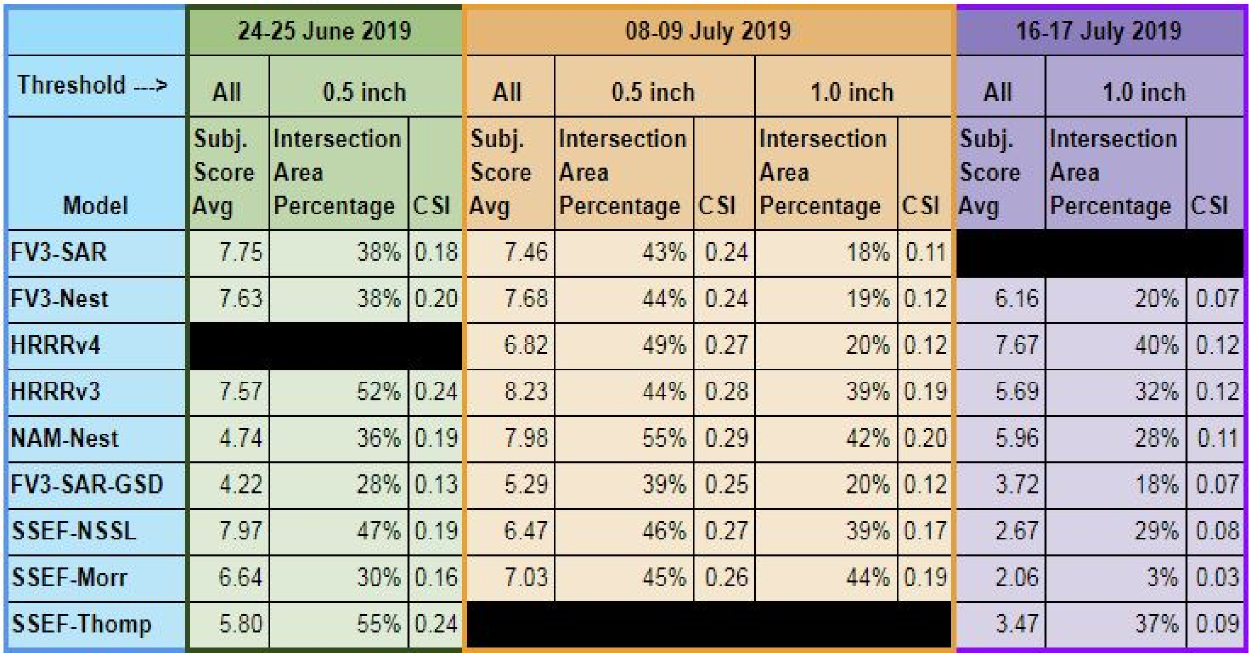
Examination of the MODE product for the half inch threshold ending at 1200 UTC on 25 June 2019, Fig. 34, is a great example of how dissimilar subjective and objective scores can be and why it difficult to mimic forecaster methods’ of model evaluation (which is what MODE tries to do). As can be seen in Fig. 34, MODE defines roughly three regions of rainfall for the event at a half inch threshold. Generally all the models evaluated for this event (HRRRv4 was missing) forecasted a similar rainfall footprint from OH to the mid-Atlantic, with greater differences seen in the forecast of the precipitation over the Northern Midwest and the western coast of the Gulf of Mexico. Table 5 lists each models’ subjective average score, CSI and intersection area percentage (the total number intersection points divided by the total observed points for precipitation objects). Comparing the CSI of the NAM-Nest (Fig. 34D) and SSEF-NSSL (Fig. 34F) it would seem that the two performed about the same, while examination of their intersection area, 36% to 47%, suggests the SSEF-NSSL's QPF coverage was slightly better than the NAM-Nest. However, the difference between their subjective scores was large, 4.74 to 7.97 respectively. This would suggest that MODE isn't telling the whole story and there is some other factor that is heavily weighing the participants' analysis. In this case, based on notes, it seems to be that they felt the NAM-Nest was too wet across the Appalachian Mountains and in GA. However, looking at MODE for higher thresholds there is not a large difference between the two in CSI or QPF hits.

Another example of this difficulty can be seen by comparing the HRRRv3 (Fig. 34C) and SSEF-Thomp (Fig. 34H). As can be seen, their MODE analysis at the half inch threshold is very similar; this is also the case at the inch threshold (not shown). Additionally, Table 5 shows the two models tied for the highest CSI at this threshold, 0.24. Furthermore, comparing their intersection areas, the FV3-Thomp outperformed the HRRRv3, 55% to 52%. Nonetheless, despite seemingly to perform well when objectively analyzed, the FV3-Thomp received the third lowest daily total subjective score (5.8) while the HRRRv3 received a daily average score of 7.57. This supports the conclusion that participants strongly considered things such as timing, QPF footprint, and what they perceive as a wet or dry forecast, as well as their own biases of "wrongness" when verifying a model forecast.

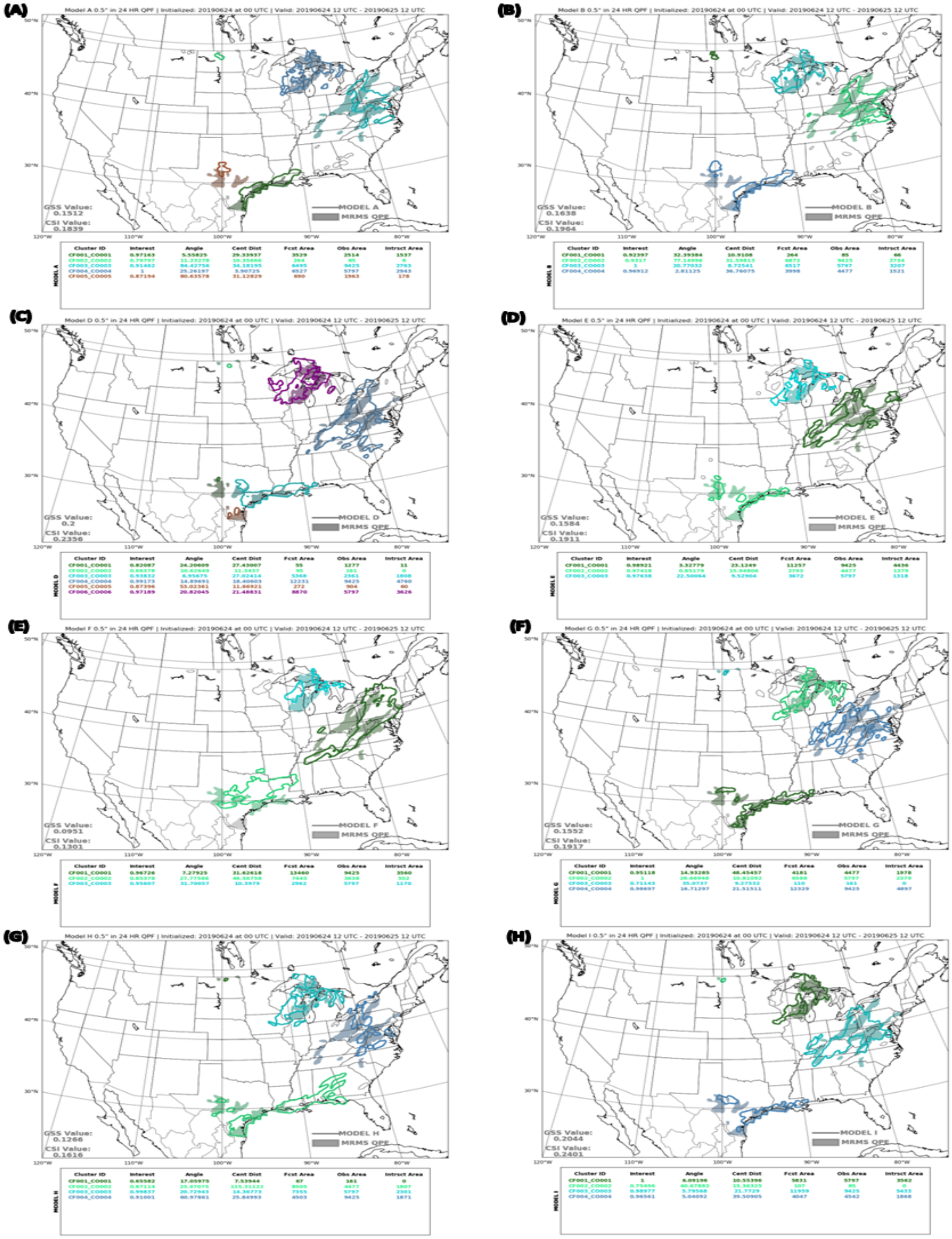
The 24 hour precipitation forecast for July 08, ending 1200 UTC July 09, is another example of how the participants weigh different aspects of the forecast with more importance, which is not always conveyed through objective analysis.   For instance, looking at Table 5, subjectively the FV3-SAR and FV3-Nest outperformed the SSEF-NSSL by approximately a point on this day; 7.46, 7.68, and 6.47 respectively. However, the MODE statistics for the half inch and inch thresholds are higher for the SSEF-NSSL member than either of the aforementioned FV3 CAMs, with the intersection area at the one inch threshold outscoring the FV3-SAR and FV3-Nest 39% to 18% and 19% respectively. Comparison of the three models in Fig. 35 and Fig. 36 shows that for the precipitation across the Northern Plains both FV3 models had a similar footprint to observations, but it was shifted westward, suggesting the models were lagging. Opposing this, the SSEF-NSSL (Fig. 35G and Fig. 36G) had a more zonal orientation to the precipitation pattern. This "miss" in the orientation of the precipitation was noted by the participants and influenced their rating of the forecast. Therefore this suggests that they would prefer a slower model that had the correct orientation of the precipitation over a model that might "correctly" forecast rainfall over more points but have the progression of the system incorrect.

Finally, Fig. 37 and Table 5 depict the various results for the 24 hour precipitation forecast for the AR event that occurred as a result of the remnants of Hurricane Barry, ending 1200 UTC July 17. This is a great example of how a high impact event can greatly influence how the "goodness" of a model can differ between subjective and objective scoring. For instance, the FV3-Nest and the SSEF-Thomp have similar CSI (0.07 and 0.09), with the SSEF-Thomp having a greater intersection area percentage (37% vs 20%) than the FV3-Nest at the one inch threshold. However, subjectively the FV3-Nest outscored the SSEF-Thomp 6.16 to 3.47. This was because although both “missed” the precipitation in the Midwest and through IN and OH, the participants noted that the FV3-Nest did better with the footprint of the high impact event; which is shown by the intersection area percentage for the object 41% to 28%. Moreover, though not shown, it nearly forecasted the exact location of the heaviest rainfall, further suggesting that the participants weighed the correct forecast of this high impact event above all else. This can also be seen when comparing HRRRv3 and HRRRv4 to one another. Again, though they had similar statistics, the participants preferred the HRRRv3 over HRRRv4 because the maximum forecasted by the HRRRv3 was better located with the observed rainfall maximum.

Table 5: *Chart displaying three examples for daily subjective scores, percentage of intersection points to observed points (referred to as intersection area percentage), and CSI for the 24 h QPF guidance from the deterministic models evaluated during the 2019 FFaIR Experiment. Each date is color coded by column and valid from 1200 to 1200 UTC. The “Subj. Score Avg.” column gives the average score from the FFaIR participants after looking at the entire output from each model during that valid period. The “Intersection Area Percentage” column gives the intersection area percentage across the CONUS for the given threshold, which is labeled in the second row of the table. Finally, the “CSI” column gives the CSI score calculated at each threshold over the whole CONUS for the valid time period. The black shading means that the model was not available for evaluation on that day. Each date/threshold corresponds with the MODE images displayed above in Figs. 34 to 37.*



	24-25 June 2019			08-09 July 2019					16-17 July 2019		
Threshold --->	All	0.5 inch		All	0.5 inch		1.0 inch		All	1.0 inch	
Model	Subj. Score Avg	Intersection Area Percentage	CSI	Subj. Score Avg	Intersection Area Percentage	CSI	Intersection Area Percentage	CSI	Subj. Score Avg	Intersection Area Percentage	CSI
FV3-SAR	7.75	38%	0.18	7.46	43%	0.24	18%	0.11			
FV3-Nest	7.63	38%	0.20	7.68	44%	0.24	19%	0.12	6.16	20%	0.07
HRRRv4				6.82	49%	0.27	20%	0.12	7.67	40%	0.12
HRRRv3	7.57	52%	0.24	8.23	44%	0.28	39%	0.19	5.69	32%	0.12
NAM-Nest	4.74	36%	0.19	7.98	55%	0.29	42%	0.20	5.96	28%	0.11
FV3-SAR-GSD	4.22	28%	0.13	5.29	39%	0.25	20%	0.12	3.72	18%	0.07
SSEF-NSSL	7.97	47%	0.19	6.47	46%	0.27	39%	0.17	2.67	29%	0.08
SSEF-Morr	6.64	30%	0.16	7.03	45%	0.26	44%	0.19	2.06	3%	0.03
SSEF-Thomp	5.80	55%	0.24						3.47	37%	0.09


(A)
Model A 0.5" in 24 HR QPF | Initialized: 20190624 at 00 UTC | Valid: 20190624 12 UTC - 20190625 12 UTC
GSS Value: 0.1512
CSI Value: 0.1839
MODEL A
MRMS QPE
(B)
Model B 0.5" in 24 HR QPF | Initialized: 20190624 at 00 UTC | Valid: 20190624 12 UTC - 20190625 12 UTC
GSS Value: 0.1638
CSI Value: 0.1964
MODEL B
MRMS QPE
(C)
Model D 0.5" in 24 HR QPF | Initialized: 20190624 at 00 UTC | Valid: 20190624 12 UTC - 20190625 12 UTC
GSS Value: 0.2
CSI Value: 0.2356
MODEL D
MRMS QPE
(D)
Model E 0.5" in 24 HR QPF | Initialized: 20190624 at 00 UTC | Valid: 20190624 12 UTC - 20190625 12 UTC
GSS Value: 0.1584
CSI Value: 0.1911
MODEL E
MRMS QPE
(E)
Model F 0.5" in 24 HR QPF | Initialized: 20190624 at 00 UTC | Valid: 20190624 12 UTC - 20190625 12 UTC
GSS Value: 0.0951
CSI Value: 0.1301
MODEL F
MRMS QPE
(F)
Model G 0.5" in 24 HR QPF | Initialized: 20190624 at 00 UTC | Valid: 20190624 12 UTC - 20190625 12 UTC
GSS Value: 0.1552
CSI Value: 0.1917
MODEL G
MRMS QPE
(G)
Model H 0.5" in 24 HR QPF | Initialized: 20190624 at 00 UTC | Valid: 20190624 12 UTC - 20190625 12 UTC
GSS Value: 0.1266
CSI Value: 0.1616
MODEL H
MRMS QPE
(H)
Model I 0.5" in 24 HR QPF | Initialized: 20190624 at 00 UTC | Valid: 20190624 12 UTC - 20190625 12 UTC
GSS Value: 0.2044
CSI Value: 0.2401
MODEL I
MRMS QPE
50°N
40°N
30°N
120°W
110°W
100°W
90°W
80°W

MODEL A

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.97163	5.55825	29.33937	3529	2514	1537
CF002_CO002	0.79797	11.23278	10.35666	264	85	8
CF003_CO003	0.91462	84.42756	34.18195	6495	9425	2763
CF004_CO004	1	25.26197	3.90725	6527	5797	2943
CF005_CO005	0.87194	80.63578	31.12829	690	1963	178

MODEL B

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.92397	32.39384	10.9108	264	85	66
CF002_CO002	0.9317	77.14996	31.59813	6872	9425	2734
CF003_CO003	1	20.77032	8.72541	6517	5797	3207
CF004_CO004	0.96912	2.81125	36.76075	3998	4477	1521

MODEL D

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.82087	24.20609	27.43007	55	1277	11
CF002_CO002	0.66378	10.62649	11.3437	95	161	0
CF003_CO003	0.93832	6.95675	27.02414	5368	2361	1808
CF004_CO004	0.99173	14.89691	18.40603	12231	9425	4760
CF005_CO005	0.87396	53.02361	11.66921	272	904	60
CF006_CO006	0.97189	20.82045	21.48831	8870	5797	3626

MODEL E

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.98921	3.32779	23.1249	11257	9425	4436
CF002_CO002	0.97418	0.85179	15.94806	2793	4477	1379
CF003_CO003	0.97638	22.50064	9.52904	3672	5797	1318

MODEL F

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.96726	7.27925	31.62618	13460	9425	3560
CF002_CO002	0.85378	27.77586	46.56758	7445	3638	552
CF003_CO003	0.95607	31.70057	10.3979	2962	5797	1170

MODEL G

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.95118	14.93285	48.45457	4181	4477	1978
CF002_CO002	1	26.66948	10.81092	6588	5797	2379
CF003_CO003	0.71143	35.0737	9.27532	110	161	0
CF004_CO004	0.98697	16.71297	21.51511	12329	9425	4897

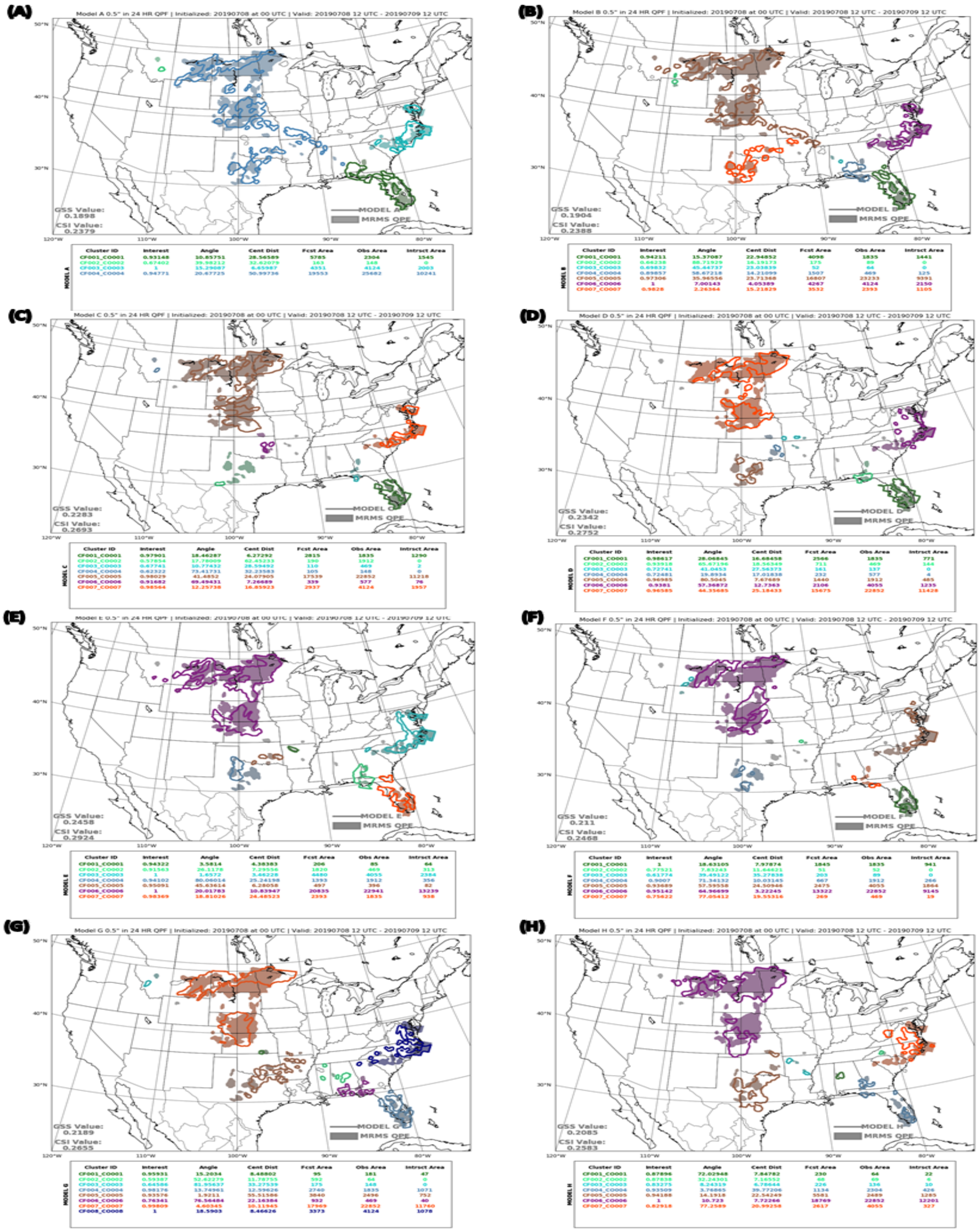
MODEL H

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.65582	17.05975	7.53944	67	161	0
CF002_CO002	0.87114	23.47075	115.31122	8505	4477	1807
CF003_CO003	0.99837	20.72943	14.36773	7355	5797	2301
CF004_CO004	0.91001	60.97861	25.84993	4503	9425	1871

MODEL I

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	1	6.09196	10.55396	5831	5797	3562
CF002_CO002	0.75496	40.67882	15.36325	107	85	0
CF003_CO003	0.98977	5.79568	21.7729	11959	9425	5433
CF004_CO004	0.96561	5.04092	39.50905	4047	4542	1868

Figure 34: *MODE precipitation results for the 0.5 inch threshold over 24 h valid from 1200 UTC 24 June 2019 to 1200 UTC 25 June 2019, showing the (A)-(H) MRMS-GC QPE (shaded) compared to model QPF (contoured) from the following deterministic models: (A) FV3-SAR, (B) FV3-Nest, (C) HRRRv4, (D) NAM-Nest, (E) FV3-SAR-GSD, (F) SSEF-NSSL, (G) SSEF-Morr, and (H) SSEF-Thomp. Below each image are the MODE verification metrics for the corresponding model.*


(A)
Model A 0.5" in 24 HR QPF | Initialized: 20190708 at 00 UTC | Valid: 20190708 12 UTC - 20190709 12 UTC
GSS Value: 0.1898
CSI Value: 0.2379
MODEL A
MRMS QPE
(B)
Model B 0.5" in 24 HR QPF | Initialized: 20190708 at 00 UTC | Valid: 20190708 12 UTC - 20190709 12 UTC
GSS Value: 0.1904
CSI Value: 0.2388
MODEL B
MRMS QPE
(C)
Model C 0.5" in 24 HR QPF | Initialized: 20190708 at 00 UTC | Valid: 20190708 12 UTC - 20190709 12 UTC
GSS Value: 0.2283
CSI Value: 0.2693
MODEL C
MRMS QPE
(D)
Model D 0.5" in 24 HR QPF | Initialized: 20190708 at 00 UTC | Valid: 20190708 12 UTC - 20190709 12 UTC
GSS Value: 0.2342
CSI Value: 0.2752
MODEL D
MRMS QPE
(E)
Model E 0.5" in 24 HR QPF | Initialized: 20190708 at 00 UTC | Valid: 20190708 12 UTC - 20190709 12 UTC
GSS Value: 0.2458
CSI Value: 0.2924
MODEL E
MRMS QPE
(F)
Model F 0.5" in 24 HR QPF | Initialized: 20190708 at 00 UTC | Valid: 20190708 12 UTC - 20190709 12 UTC
GSS Value: 0.211
CSI Value: 0.2468
MODEL F
MRMS QPE
(G)
Model G 0.5" in 24 HR QPF | Initialized: 20190708 at 00 UTC | Valid: 20190708 12 UTC - 20190709 12 UTC
GSS Value: 0.2189
CSI Value: 0.2655
MODEL G
MRMS QPE
(H)
Model H 0.5" in 24 HR QPF | Initialized: 20190708 at 00 UTC | Valid: 20190708 12 UTC - 20190709 12 UTC
GSS Value: 0.2085
CSI Value: 0.2583
MODEL H
MRMS QPE
50°N
40°N
30°N
120°W
110°W
100°W
90°W
80°W

MODEL A

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.93148	10.85751	28.56589	5785	2304	1545
CF002_CO002	0.67402	39.98212	32.62079	163	148	0
CF003_CO003	1	15.29087	6.65987	4351	4124	2003
CF004_CO004	0.94771	20.67725	50.99736	19553	25682	10241

MODEL B

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.94211	15.37087	22.94852	4098	1835	1441
CF002_CO002	0.66238	88.71929	16.19173	175	89	0
CF003_CO003	0.69832	45.44737	23.03839	52	64	0
CF004_CO004	0.89857	58.67218	14.21099	1507	469	125
CF005_CO005	0.97306	35.96556	23.71368	16807	23233	9391
CF006_CO006	1	7.00143	4.05389	4267	4124	2150
CF007_CO007	0.9828	2.26364	15.21829	3532	2393	1105

MODEL C

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.97901	18.46287	6.27292	2815	1835	1290
CF002_CO002	0.57854	17.78009	62.45233	190	1912	0
CF003_CO003	0.67741	10.77432	28.59492	110	469	2
CF004_CO004	0.62322	73.41731	32.23583	105	148	0
CF005_CO005	0.98029	41.4852	24.07905	17539	22852	11218
CF006_CO006	0.91682	69.49431	7.26689	339	577	76
CF007_CO007	0.98564	12.25738	16.85923	2937	4124	1957

MODEL D

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.98617	28.06845	16.68458	2566	1835	771
CF002_CO002	0.93918	65.67196	18.56349	711	469	144
CF003_CO003	0.72741	41.0453	27.56373	161	137	0
CF004_CO004	0.72481	19.8934	17.01838	232	577	4
CF005_CO005	0.96985	80.5045	7.67689	1440	1912	485
CF006_CO006	0.9381	57.36872	12.7363	2106	4055	1235
CF007_CO007	0.96585	44.35685	25.18433	15675	22852	11428

MODEL E

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.94322	3.5814	4.38383	206	85	64
CF002_CO002	0.91563	26.1178	7.29556	1820	469	313
CF003_CO003	1	1.6572	3.46228	4480	4055	2384
CF004_CO004	0.94102	80.06014	25.24198	1393	1912	356
CF005_CO005	0.95091	45.63614	6.28058	497	396	82
CF006_CO006	1	20.01783	10.83947	20835	22941	13239
CF007_CO007	0.98369	18.81026	24.48523	2393	1835	938

MODEL F

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	1	18.63105	7.97874	1845	1835	941
CF002_CO002	0.77521	7.83243	11.64621	51	52	0
CF003_CO003	0.61774	39.49122	35.27838	203	89	0
CF004_CO004	0.9007	71.34132	10.03145	667	1912	266
CF005_CO005	0.93689	57.59558	24.50946	2475	4055	1864
CF006_CO006	0.95142	64.96699	3.22245	13322	22852	9145
CF007_CO007	0.75622	77.05412	19.55316	269	469	19

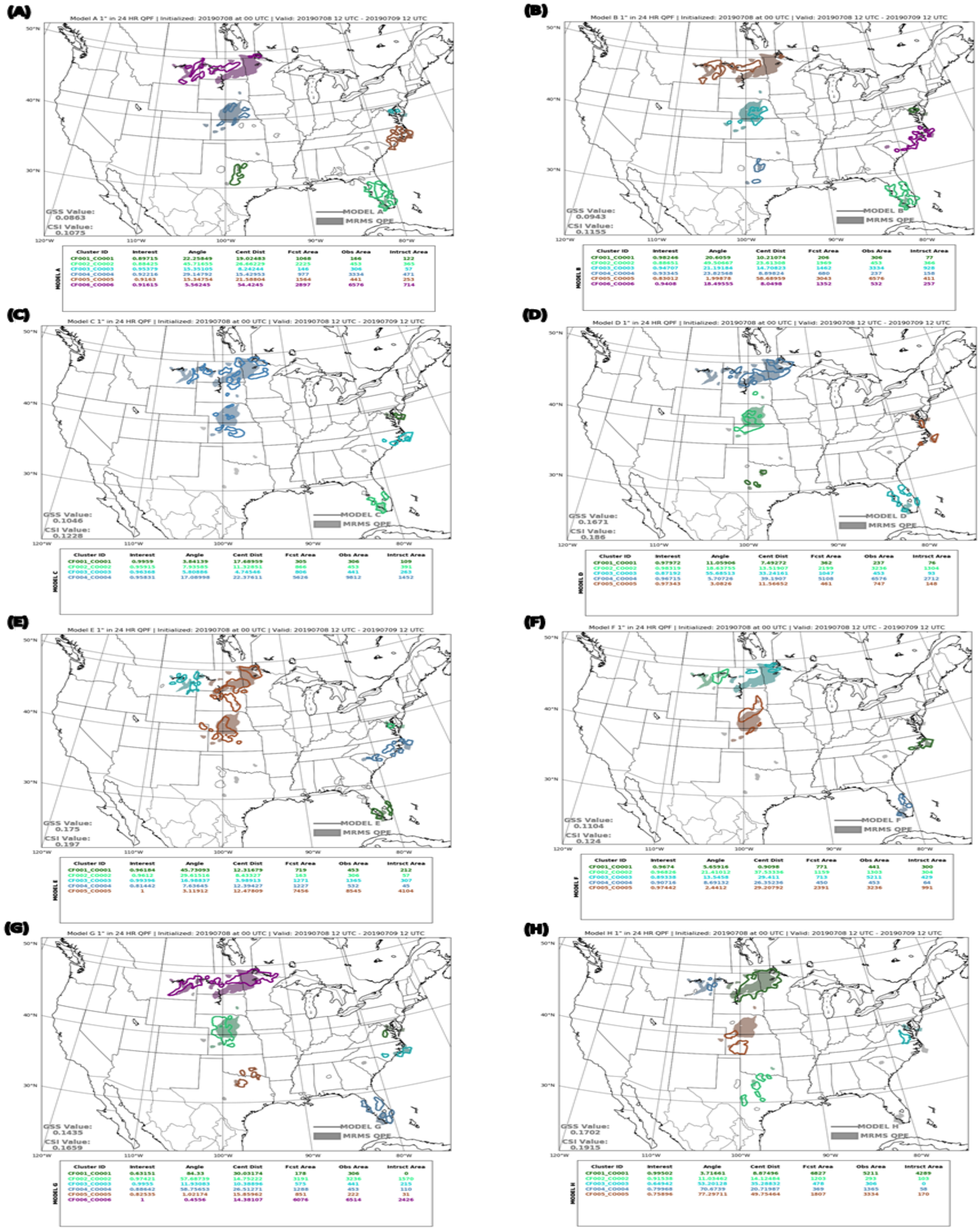
MODEL G

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.95931	15.2034	8.48802	95	181	47
CF002_CO002	0.59387	52.62279	11.78755	592	64	0
CF003_CO003	0.64586	81.95637	33.27539	175	148	0
CF004_CO004	0.98176	13.74961	12.59626	2740	1835	1071
CF005_CO005	0.93576	1.9211	55.51586	3840	2496	752
CF006_CO006	0.76341	76.56484	22.16384	932	469	40
CF007_CO007	0.99809	4.60345	10.11945	17969	22852	11760
CF008_CO008	1	18.5903	8.46626	3373	4124	1078

MODEL H

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.87896	72.02948	7.84782	230	64	22
CF002_CO002	0.87838	32.24301	7.16552	68	69	6
CF003_CO003	0.83275	8.24319	6.78644	226	136	10
CF004_CO004	0.93509	3.76865	39.77206	1134	2304	426
CF005_CO005	0.94188	14.1918	22.54249	5581	2489	1285
CF006_CO006	1	10.723	7.72266	18769	22852	12201
CF007_CO007	0.82918	77.2589	20.99258	2617	4055	327

Figure 35: *MODE precipitation results for the 0.5 inch threshold over 24 h valid from 1200 UTC 08 July 2019 to 1200 UTC 09 July 2019, showing the (A)-(H) MRMS-GC QPE (shaded) compared to model QPF (contoured) from the following deterministic models: (A) FV3-SAR, (B) FV3-Nest, (C) HRRRv3, (D) HRRRv4, (E) NAM-Nest, (F) FV3-SAR-GSD, (G) SSEF-NSSL, and (H) SSEF-Morr. Below each image are the MODE verification metrics for the corresponding model.*


(A)
Model A 1" in 24 HR QPF | Initialized: 20190708 at 00 UTC | Valid: 20190708 12 UTC - 20190709 12 UTC
GSS Value: 0.0863
CSI Value: 0.1075
MODEL A
MRMS QPE
(B)
Model B 1" in 24 HR QPF | Initialized: 20190708 at 00 UTC | Valid: 20190708 12 UTC - 20190709 12 UTC
GSS Value: 0.0943
CSI Value: 0.1155
MODEL B
MRMS QPE
(C)
Model C 1" in 24 HR QPF | Initialized: 20190708 at 00 UTC | Valid: 20190708 12 UTC - 20190709 12 UTC
GSS Value: 0.1046
CSI Value: 0.1228
MODEL C
MRMS QPE
(D)
Model D 1" in 24 HR QPF | Initialized: 20190708 at 00 UTC | Valid: 20190708 12 UTC - 20190709 12 UTC
GSS Value: 0.1671
CSI Value: 0.186
MODEL D
MRMS QPE
(E)
Model E 1" in 24 HR QPF | Initialized: 20190708 at 00 UTC | Valid: 20190708 12 UTC - 20190709 12 UTC
GSS Value: 0.175
CSI Value: 0.197
MODEL E
MRMS QPE
(F)
Model F 1" in 24 HR QPF | Initialized: 20190708 at 00 UTC | Valid: 20190708 12 UTC - 20190709 12 UTC
GSS Value: 0.1104
CSI Value: 0.124
MODEL F
MRMS QPE
(G)
Model G 1" in 24 HR QPF | Initialized: 20190708 at 00 UTC | Valid: 20190708 12 UTC - 20190709 12 UTC
GSS Value: 0.1435
CSI Value: 0.1659
MODEL G
MRMS QPE
(H)
Model H 1" in 24 HR QPF | Initialized: 20190708 at 00 UTC | Valid: 20190708 12 UTC - 20190709 12 UTC
GSS Value: 0.1702
CSI Value: 0.1915
MODEL H
MRMS QPE
50°N
40°N
30°N
120°W
110°W
100°W
90°W
80°W

MODEL A

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.89715	22.25849	19.02483	1068	166	122
CF002_CO002	0.88425	45.71655	26.46229	2225	453	365
CF003_CO003	0.95379	15.35105	8.24244	146	306	57
CF004_CO004	0.92216	29.14792	15.42953	977	3334	471
CF005_CO005	0.9163	15.34754	21.58804	1564	441	253
CF006_CO006	0.91615	5.56245	54.4245	2897	6576	714

MODEL B

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.98246	20.6059	10.21074	206	306	77
CF002_CO002	0.88651	49.50667	23.61308	1969	453	366
CF003_CO003	0.94707	21.19184	14.70823	1462	3334	928
CF004_CO004	0.93345	23.82568	8.89824	680	237	158
CF005_CO005	0.83012	1.99878	58.68959	3043	6576	411
CF006_CO006	0.9408	18.49555	8.0498	1352	532	257

MODEL C

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.9959	3.84139	17.68959	305	306	109
CF002_CO002	0.95915	7.93585	11.32851	866	453	391
CF003_CO003	0.96368	5.80886	4.74546	806	441	263
CF004_CO004	0.95831	17.08998	22.37611	5626	9812	1452

MODEL D

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.97972	11.05906	7.49272	362	237	76
CF002_CO002	0.98319	18.63755	13.51907	2199	3236	1304
CF003_CO003	0.87192	55.68513	33.24161	1047	453	93
CF004_CO004	0.96715	5.70726	39.1907	5108	6576	2712
CF005_CO005	0.97343	3.0826	11.56652	461	747	148

MODEL E

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.96184	45.73093	12.31679	719	453	212
CF002_CO002	0.9612	29.61516	8.43327	163	306	57
CF003_CO003	0.99396	16.98837	3.98913	1271	1365	307
CF004_CO004	0.81442	7.63645	12.39427	1227	532	45
CF005_CO005	1	3.11912	12.47809	7456	8545	4104

MODEL F

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.9674	5.65916	0.9098	771	441	300
CF002_CO002	0.96826	21.41012	37.53336	1159	1303	304
CF003_CO003	0.89338	13.5458	29.411	713	5211	429
CF004_CO004	0.90716	8.69132	26.35236	450	453	64
CF005_CO005	0.97442	2.4412	29.20792	2391	3236	991

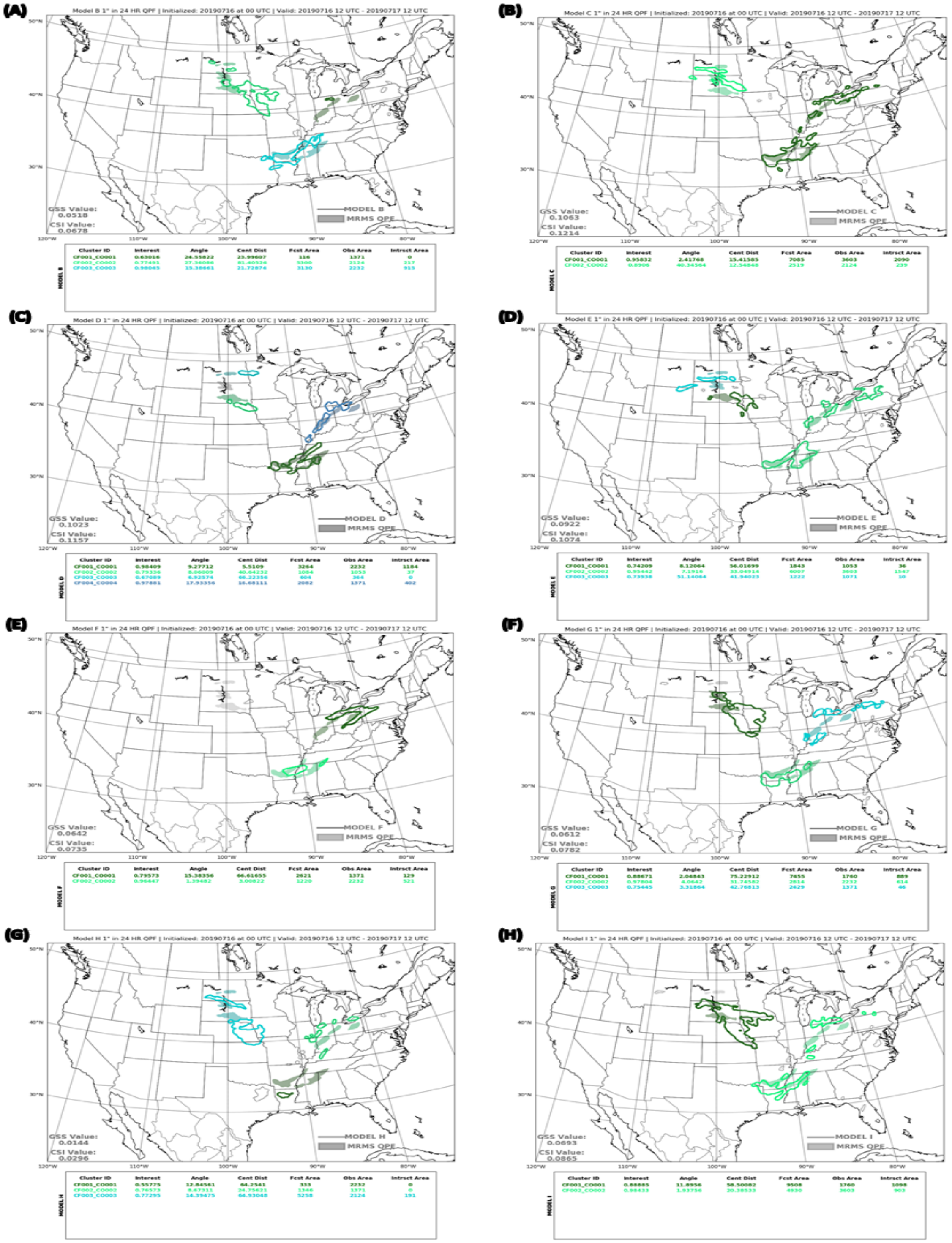
MODEL G

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.63151	84.33	30.03174	178	306	0
CF002_CO002	0.97421	57.68739	14.75222	3191	3236	1570
CF003_CO003	0.9955	11.93083	10.38896	575	441	215
CF004_CO004	0.88642	58.75653	26.51271	1288	453	110
CF005_CO005	0.82535	1.02174	15.85962	851	222	31
CF006_CO006	1	0.4556	14.38107	6076	6514	2426

MODEL H

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.99502	3.71661	8.87496	6827	5211	4289
CF002_CO002	0.91538	11.03462	14.12484	1203	293	103
CF003_CO003	0.64942	53.20128	35.28832	478	306	0
CF004_CO004	0.79968	70.6739	20.71987	369	1365	58
CF005_CO005	0.75896	77.29711	49.75464	1807	3334	170

Figure 36: *MODE precipitation results for the 1 inch threshold over 24 h valid from 1200 UTC 08 July 2019 to 1200 UTC 09 July 2019, showing the (A)-(H) MRMS-GC QPE (shaded) compared to model QPF (contoured) from the following deterministic models: (A) FV3-SAR, (B) FV3-Nest, (C) HRRRv3, (D) HRRRv4, (E) NAM-Nest, (F) FV3-SAR-GSD, (G) SSEF-NSSL, and (H) SSEF-Morr. Below each image are the MODE verification metrics for the corresponding model.*



(A)
Model B 1" in 24 HR QPF | Initialized: 20190716 at 00 UTC | Valid: 20190716 12 UTC - 20190717 12 UTC
GSS Value: 0.0518
CSI Value: 0.0678
MODEL B
MRMS QPE
50°N
40°N
30°N
120°W
110°W
100°W
90°W
80°W

MODEL B

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.63016	24.55822	23.99607	116	1371	0
CF002_CO002	0.77491	27.36086	81.40526	5300	2124	217
CF003_CO003	0.98045	15.38661	21.72874	3130	2232	915

(B)
Model C 1" in 24 HR QPF | Initialized: 20190716 at 00 UTC | Valid: 20190716 12 UTC - 20190717 12 UTC
GSS Value: 0.1063
CSI Value: 0.1214
MODEL C
MRMS QPE

MODEL C

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.95832	2.41768	15.41585	7085	3603	2090
CF002_CO002	0.8906	40.34564	12.54848	2519	2124	239

(C)
Model D 1" in 24 HR QPF | Initialized: 20190716 at 00 UTC | Valid: 20190716 12 UTC - 20190717 12 UTC
GSS Value: 0.1023
CSI Value: 0.1157
MODEL D
MRMS QPE

MODEL D

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.98409	9.27712	5.5109	3264	2232	1184
CF002_CO002	0.79336	8.06009	40.64232	1084	1053	37
CF003_CO003	0.67089	6.92574	66.22356	604	364	0
CF004_CO004	0.97881	17.93356	16.68111	2082	1371	402

(D)
Model E 1" in 24 HR QPF | Initialized: 20190716 at 00 UTC | Valid: 20190716 12 UTC - 20190717 12 UTC
GSS Value: 0.0922
CSI Value: 0.1074
MODEL E
MRMS QPE

MODEL E

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.74209	8.12064	56.01699	1843	1053	36
CF002_CO002	0.95442	7.1916	33.04914	6007	3603	1547
CF003_CO003	0.73938	51.14064	41.94023	1222	1071	10

(E)
Model F 1" in 24 HR QPF | Initialized: 20190716 at 00 UTC | Valid: 20190716 12 UTC - 20190717 12 UTC
GSS Value: 0.0642
CSI Value: 0.0735
MODEL F
MRMS QPE

MODEL F

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.79573	15.38356	66.61655	2621	1371	129
CF002_CO002	0.96447	1.39482	3.00822	1220	2232	521

(F)
Model G 1" in 24 HR QPF | Initialized: 20190716 at 00 UTC | Valid: 20190716 12 UTC - 20190717 12 UTC
GSS Value: 0.0612
CSI Value: 0.0782
MODEL G
MRMS QPE

MODEL G

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.88671	2.04843	75.22912	7455	1760	889
CF002_CO002	0.97804	4.0642	31.74582	2814	2232	614
CF003_CO003	0.75445	3.31864	42.76813	2429	1371	46

(G)
Model H 1" in 24 HR QPF | Initialized: 20190716 at 00 UTC | Valid: 20190716 12 UTC - 20190717 12 UTC
GSS Value: 0.0144
CSI Value: 0.0296
MODEL H
MRMS QPE

MODEL H

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.55775	12.84561	64.2541	333	2232	0
CF002_CO002	0.76573	8.67311	24.75621	1346	1371	0
CF003_CO003	0.77295	14.39475	64.93048	5258	2124	191

(H)
Model I 1" in 24 HR QPF | Initialized: 20190716 at 00 UTC | Valid: 20190716 12 UTC - 20190717 12 UTC
GSS Value: 0.0693
CSI Value: 0.0865
MODEL I
MRMS QPE

MODEL I

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.88885	11.8956	58.50082	9508	1760	1098
CF002_CO002	0.98433	1.93756	20.38533	4930	3603	903

Figure 37: *MODE precipitation results for the 1 inch threshold over 24 h valid from 1200 UTC 16 July 2019 to 1200 UTC 17 July 2019, showing the (A)-(H) MRMS-GC QPE (shaded) compared to model QPF (contoured) from the following deterministic models: (A) FV3-Nest, (B) HRRRv3, (C) HRRRv4, (D) NAM-Nest, (E) FV3-SAR-GSD, (F) SSEF-NSSL, (G) SSEF-Morr, and (H) SSEF-Thomp. Below each image are the MODE verification metrics for the corresponding model.*

Comparison of FV3-Nest and FV3-SAR from EMC

An additional goal of the experiment was to compare the FV3-SAR forecasts with the FV3-Nest forecasts to see if the FV3-SAR could be used exclusively to save on computing expenses. Therefore, in addition to the normal subjective verification that was done for all the deterministic models examined, the two models were evaluated using a comparison question as well. As stated in Section 3, the comparison involved the participants evaluating the utility of the forecast against each other rather than scoring the models on the "correctness" of the forecast. In this question, a score of 5 meant a forecaster thought there was no difference in the utility between the two models, while anything greater than a 5 meant the FV3-Nest was more useful and a score below a 5 meant the FV3-SAR was more useful (refer to Fig. 3). EMC was specifically interested in how the models performed against one another at later forecast times and therefore the models were analyzed at forecast hour 36 with 6 h QPF guidance. Additionally, they were analyzed over a limited domain rather than over the CONUS.

As was seen in the analysis of the subjective model verification, the participants felt that the QPF from the FV3-Nest provided more value to the forecast than the FV3-SAR QPF when compared directly to one another. However, as can be seen in Fig. 38, on average this preference of the FV3-Nest over the FV3-SAR was only slight, with a total average of 5.26, when evaluating the 6 h QPF from each model at forecast hour 36. The only week that leaned towards preferring the FV3-SAR was week 1 (4.73), while week 4 had the highest preference of FV3-Nest (5.91). The average scores however do not tell the whole story due to how the scoring works. As can be seen in Fig. 39 when focusing solely on the scores it can be seen that only 45 of the 114 scores (39%) recorded were a 5, while there were nearly twice as many times in which the FV3-Nest was strongly favored[10] over the FV3-SAR, 32 (28%) to 18 (16%) respectively.

Reviewing how the models scored against one another in the deterministic verification showed that 69% of the time the model that had the higher daily average score for the 24 h QPF evaluation also was the FV3 model that was preferred during the direct comparison. For instance, when evaluating the 24 h QPF valid 1200 UTC June 17 to 1200 UTC June 18, 2019 the average score for the FV3-SAR was 6.48 while the average score for the FV3-Nest was 5.85. From this it should be expected that when comparing the two models against each other that the scores would show that the FV3-SAR was preferred (i.e. scores less than 5). This was in fact what occurred, with an average daily score of 4.65 for the evaluation of their 6 h QPF. Both the 24 h QPF CONUS and the 6 h QPF region verifications can be seen in Fig. 40 and Fig. 41 .

[10] Defined as a score greater than or equal to 6.

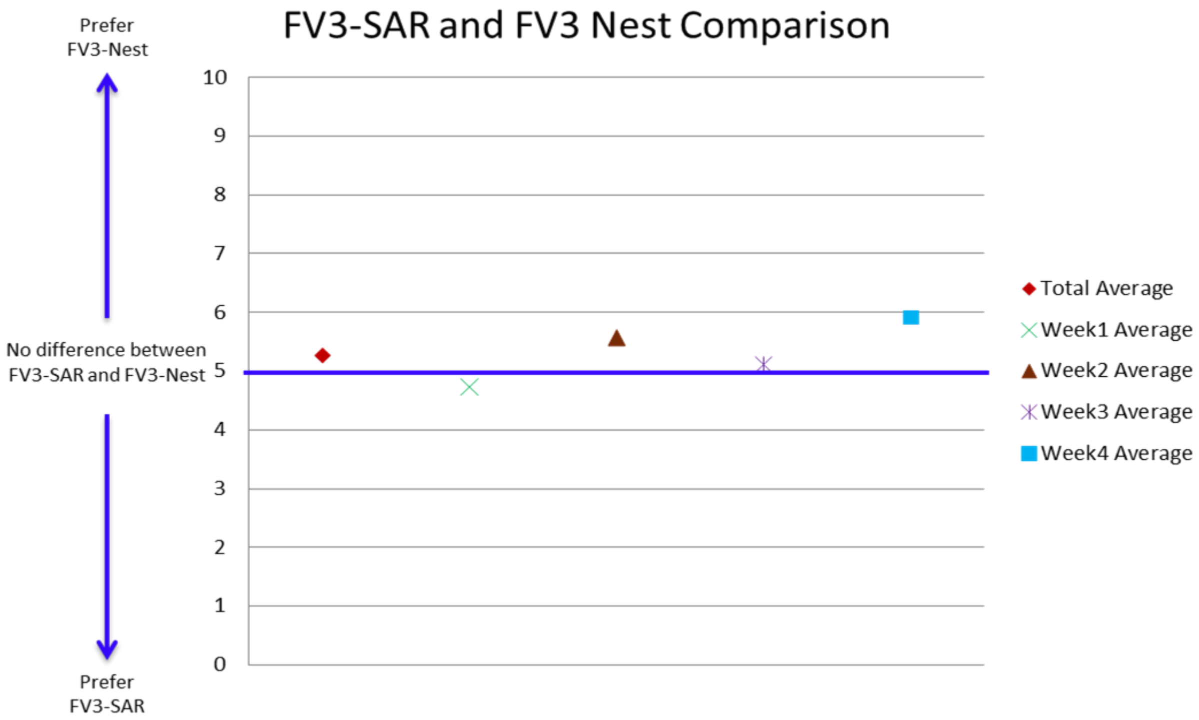


Figure 38: *Chart showing the experiment total and weekly average scores for the FV3-SAR and FV3-Nest comparison of utility for the 6 hour QPF at forecast hour 36. A score of 5 would indicate that there was no discernible difference between the two models, while anything above(below) a 5 indicates that the participants felt the FV3-Nest(FV3-SAR) was more helpful in the forecast.*

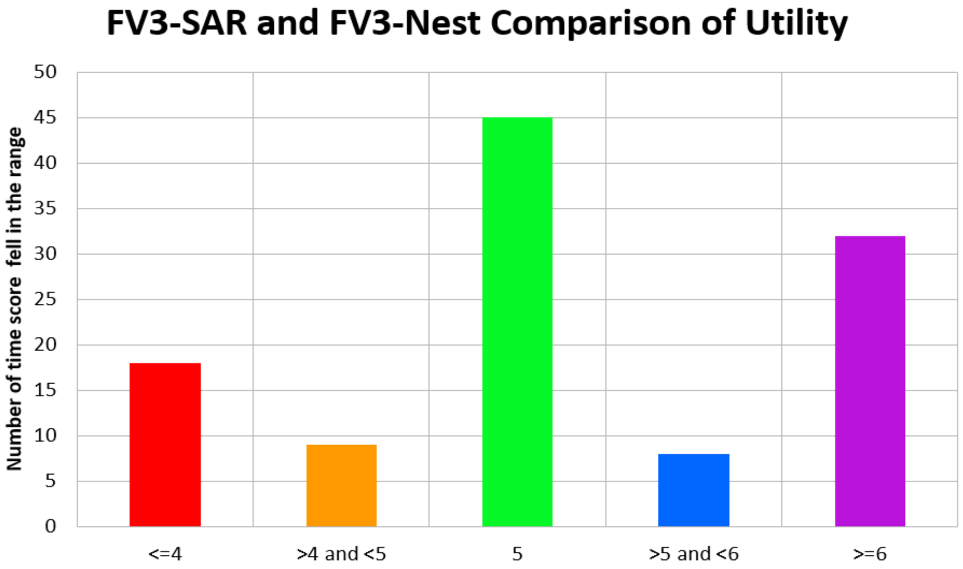


Figure 39: *Chart depicting the number of times the scores for the FV3-SAR and FV3-Nest fell within various ranges.*

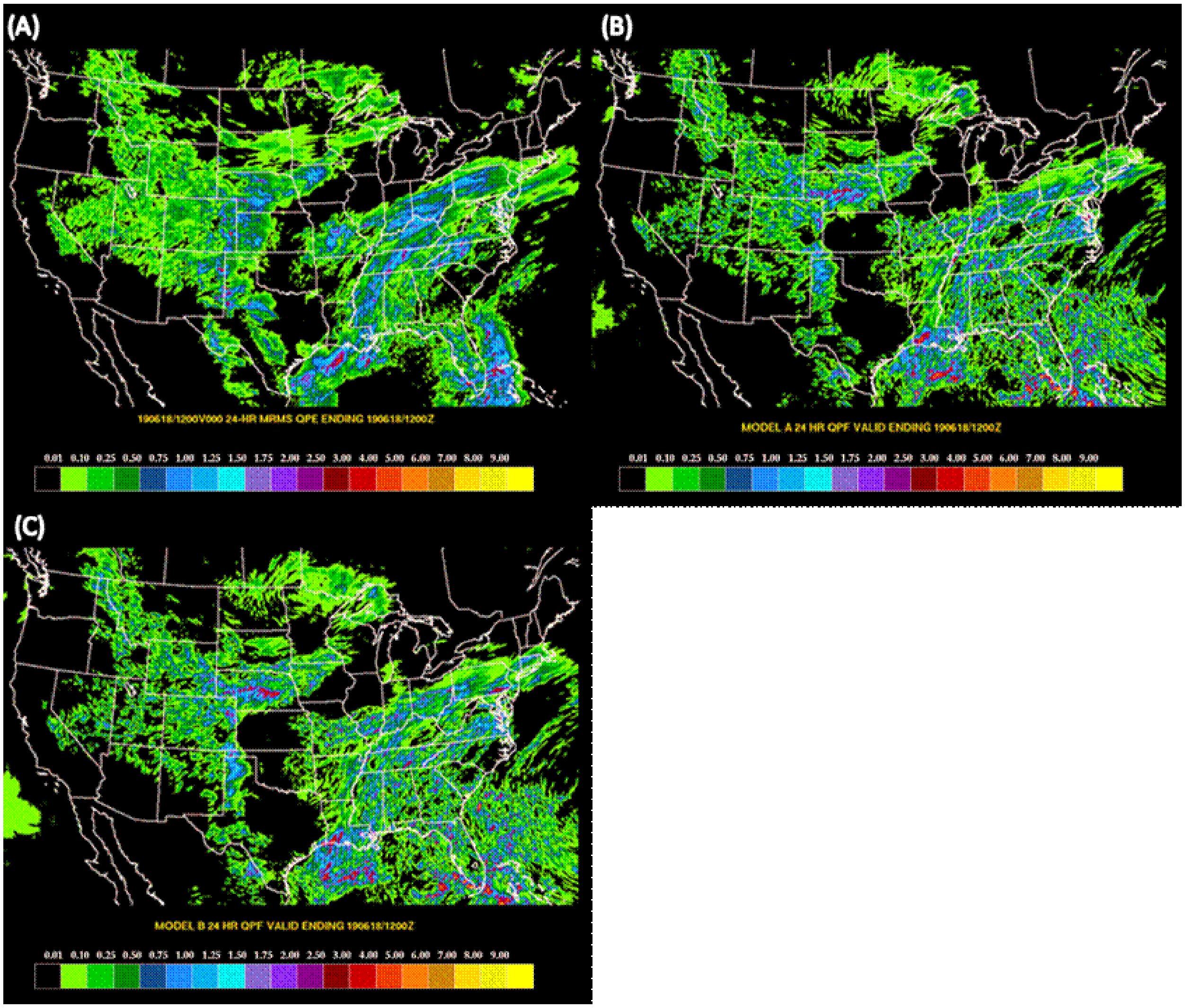


Figure 40: *(A) 24 hour MRMS-GC QPE and 24 hour QPF from (B) FV3-SAR and (C) FV3-NEST all valid 1200 UTC 17 June to 1200 UTC 18 June 2019.*

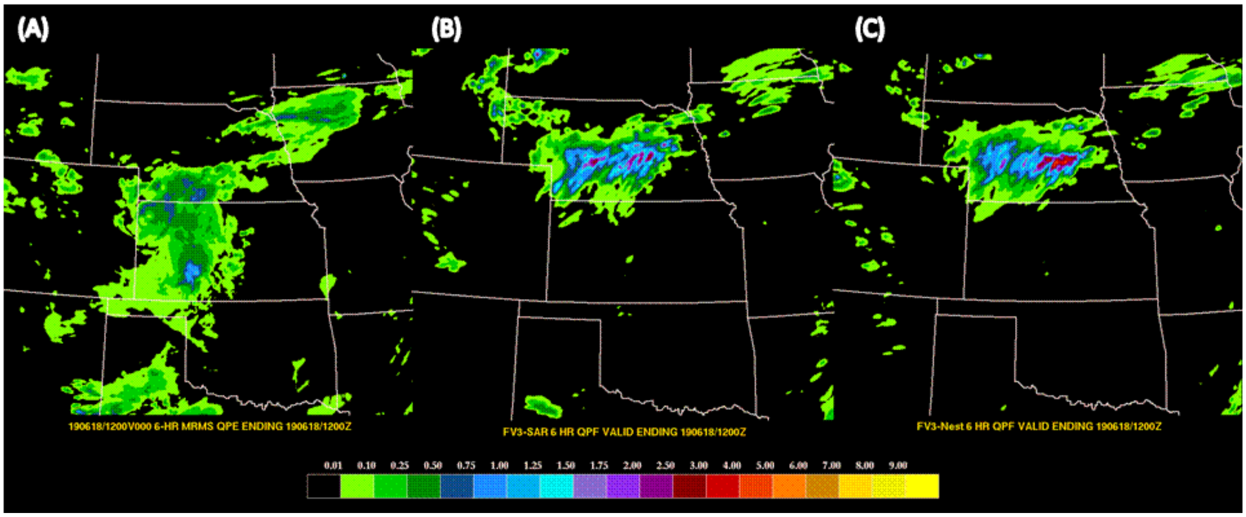

(A)
(B)
(C)
190618/1200V000 6-HR MRMS QPE ENDING 190618/1200Z
FV3-SAR 6 HR QPF VALID ENDING 190618/1200Z
FV3-Nest 6 HR QPF VALID ENDING 190618/1200Z
0.01 0.10 0.25 0.50 0.75 1.00 1.25 1.50 1.75 2.00 2.50 3.00 4.00 5.00 6.00 7.00 8.00 9.00

Figure 41: *(A) 6 hour MRMS-GC QPE and 6 hour QPF from (B) FV3-SAR and (C) FV3-NEST all valid 0600 UTC to 1200 UTC 18 June 2019.*

However, there were some instances in which the results from the two forms of subjective analysis did not agree with one another. One example of this was on June 22, 2019. For the 24 h QPF verification ending 1200 UTC on June 22 the average score for the FV3-SAR was 6.38 while the average score for the FV3-Nest was 5.36. Meanwhile the evaluation of the 6 h QPF valid at the same time suggested that the participants felt the FV3-Nest QPF was marginally (5.08) better. Figures 42 and 43 show the verification for each of these time periods. As can be seen in the 6 h QPF guidance in Fig. 42, although both models did not forecast the precipitation far enough eastward into IL, the FV3-Nest was slower than the FV3-SAR. However, the FV3-SAR significantly over-forecasted precipitation totals, with the highest amounts exceeding 4 inches over a broad area when around 2.5 inches were observed over a relatively small area. The precipitation totals forecasted by the FV3-Nest, on the other hand, were on par with observations.

These differences were also seen when examining the 24 h QPF. Participants noted that these differing misses in the QPF forecast made the comparison between the two models difficult. When looking at the whole CONUS over 24 h (Fig. 43) the participants noted that although both models often failed in the same locations, the degree of failure varied. One example of this is the Bootheel of Missouri, where both models over-forecasted the magnitude of the QPF. Participants noted that although neither forecast was good, the over-forecast was less extreme from the FV3-SAR. Opposing this, along the IA/MO border, the FV3-SAR suggested 24 h totals exceeding 5 inches, when less than 2 inches where seen. The FV3-Nest however did not forecast such high totals and therefore the participants "gave points" to the FV3-Nest for this region. Such vast dissimilarities in the forecasts resulted in the individual scores ranging from a 4 (prefer the FV3-SAR) to a 7 (prefer the FV3-Nest), leading to an average around a 5.

The MODE analysis for this event at the half inch, inch and two inch threshold can be seen in Fig. 44, with intersection area percentages in Table 6. Excluding the system across the Midwest and the Mid-Mississippi Valley, MODE indicates that the FV3-SAR and FV3-NEST performed about the same with an intersect area percentage of 56% and 54% respectively at the half inch threshold. However, if the event is included the FV3-SAR outperforms the FV3-Nest by 7%. The FV3-SAR also outperforms the FV3-Nest at the one inch threshold, especially when focusing on the MODE object for the event, with intersect area percentages of 65% and 42% , respectively. Additionally, it can be seen in Figs. 44C-D that the shape of the one inch QPF contour for the FV3-SAR more closely resembled the observed precipitation footprint than the FV3-Nest's contour shape. However when looking at the two inch threshold the intersect areas suggest that the FV3-Nest slightly outperformed the FV3-SAR (see Table 6), though both significantly over-forecast the extent of QPF equaling two inches. Combined, this suggests that the FV3-SAR provided a better forecast for the lower end amounts and for the location and footprint of precipitation for the event while the FV3-Nest performed better at the higher thresholds, providing less of an over-forecast. This follows the comments from the participants and the results from the subjective verification of the 24 h QPF. It also sheds light on why the results from the two questions about the FV3-SAR and FV3-Nest differed for this event.

Such results show the importance of evaluating the two models across various time scales and thresholds. Additionally it makes it difficult to determine if the FV3-SAR can be used in place of the FV3-Nest. Therefore it has been determined that more testing needs to be done and the FV3-SAR needs to be further refined. Additionally, it is suggested that EMC-FV3-Nest/SAR team further analyze the June 21-22 case. Lastly, as noted above, both models consistently had a high wet bias throughout the experiment which should also be examined further.

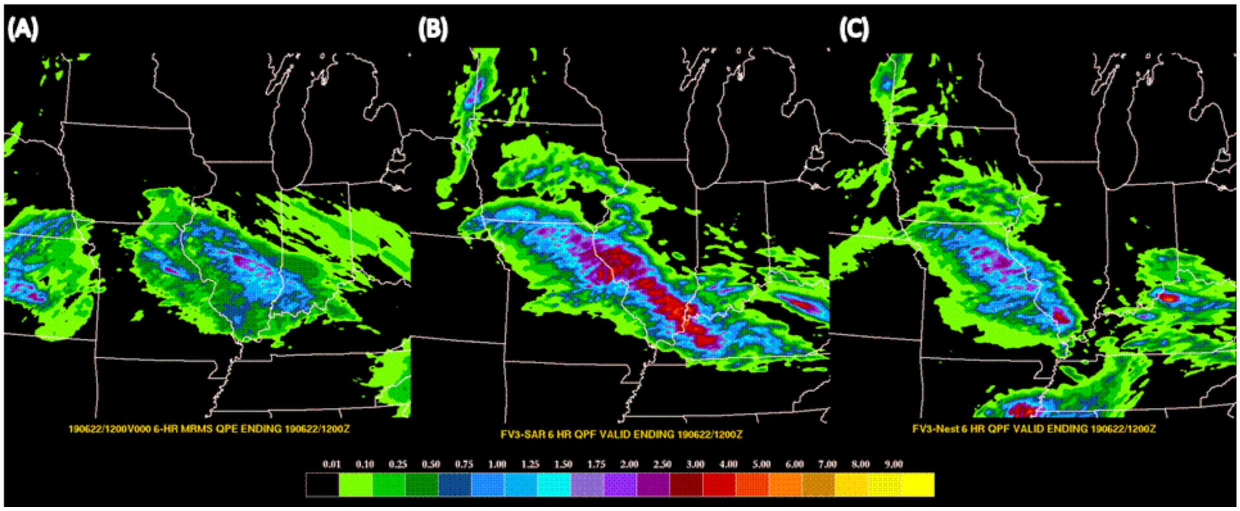

(A)
(B)
(C)
190622/1200V000 6-HR MRMS QPE ENDING 190622/1200Z
FV3-SAR 6 HR QPF VALID ENDING 190622/1200Z
FV3-Nest 6 HR QPF VALID ENDING 190622/1200Z
0.01 0.10 0.25 0.50 0.75 1.00 1.25 1.50 1.75 2.00 2.50 3.00 4.00 5.00 6.00 7.00 8.00 9.00

Figure 42: *Same as Fig. 41 but valid 0600 UTC to 1200 UTC 22 June 2019.*

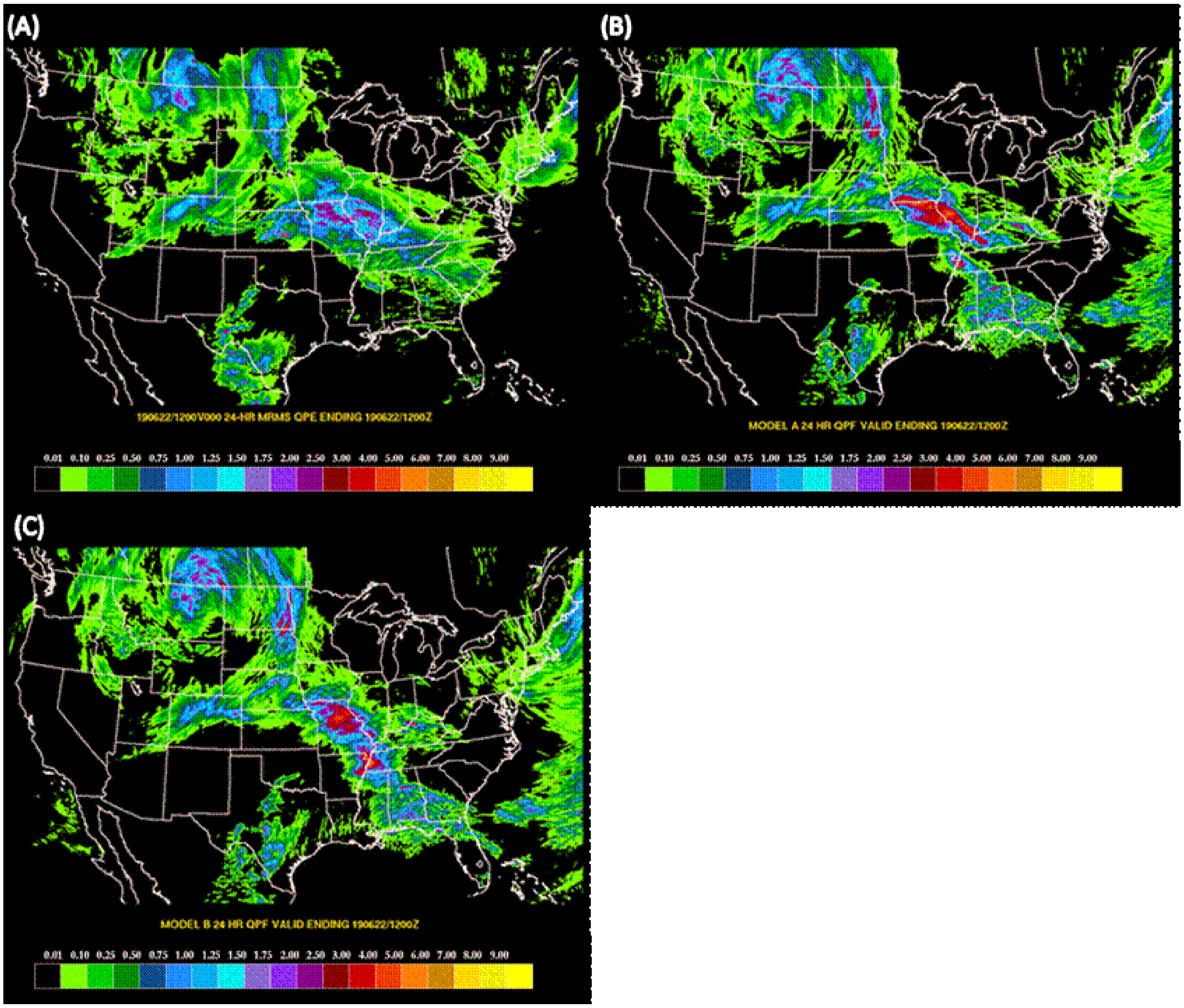
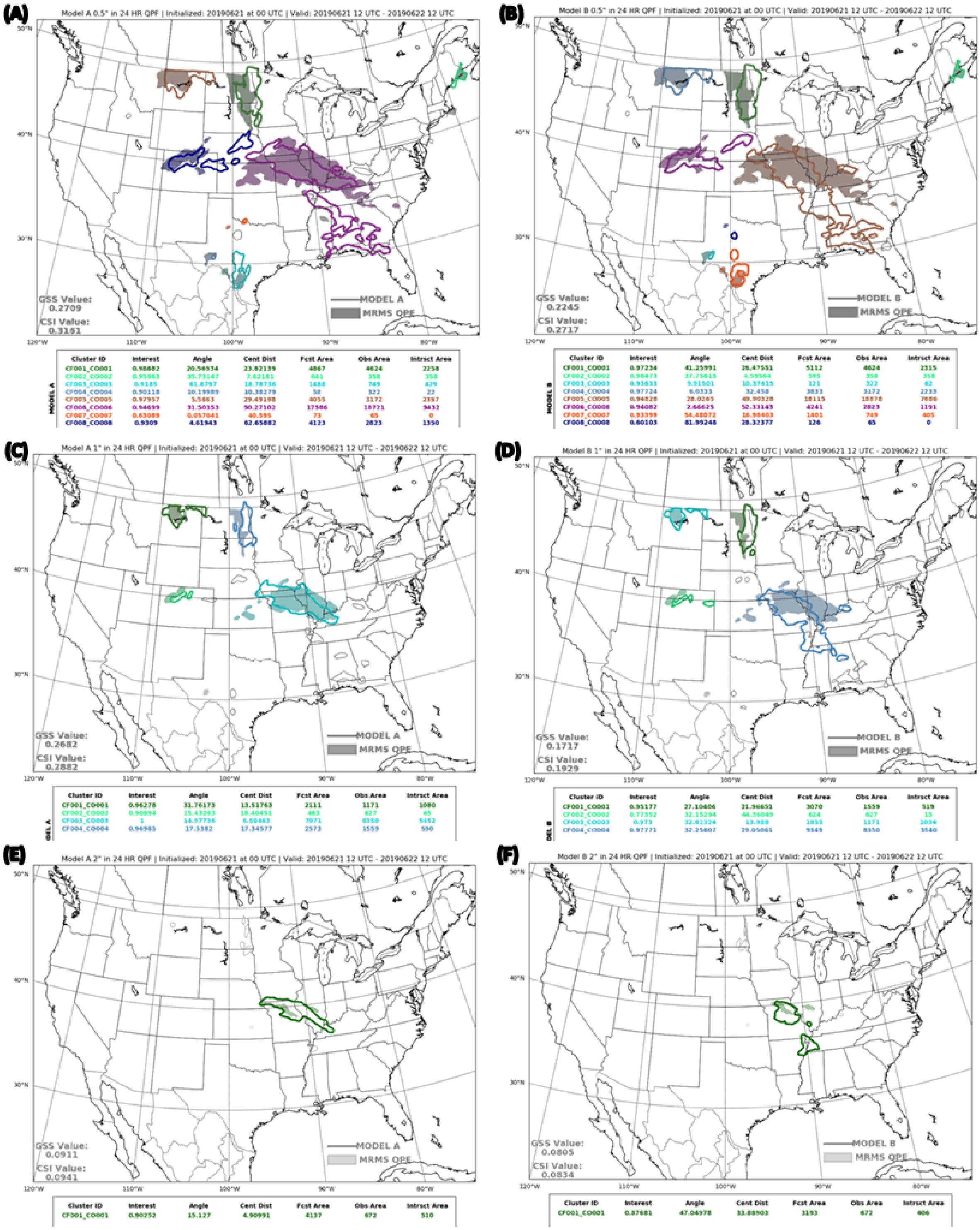


Figure 43: *Same as Fig. 40 but valid 1200 UTC 21 June to 1200 UTC 22 June 2019.*

Table 6: *Chart displaying the MODE intersection area percentage for the FV3-SAR and FV3-Nest at half inch, and inch thresholds valid 1200 TUC 21 June 2019 to 1200 UTC 22 June 2019. The intersection area percentage is the number of intersection points to the number of observed points. Total indicates all the objects identified by MODE were included in the calculation while Midwest indicates that only the points identified as part of the Midwest/Mid-Mississippi Valley system are in the calculation.*

Threshold	0.5”		1”	
	Total	Midwest	Total	Midwest
FV3-SAR	53%	50%	61%	65%
FV3-Nest	46%	41%	44%	42%


(A)
Model A 0.5" in 24 HR QPF | Initialized: 20190621 at 00 UTC | Valid: 20190621 12 UTC - 20190622 12 UTC
GSS Value: 0.2709
CSI Value: 0.3161
MODEL A
MRMS QPE
(B)
Model B 0.5" in 24 HR QPF | Initialized: 20190621 at 00 UTC | Valid: 20190621 12 UTC - 20190622 12 UTC
GSS Value: 0.2245
CSI Value: 0.2717
MODEL B
MRMS QPE
(C)
Model A 1" in 24 HR QPF | Initialized: 20190621 at 00 UTC | Valid: 20190621 12 UTC - 20190622 12 UTC
GSS Value: 0.2682
CSI Value: 0.2882
MODEL A
MRMS QPE
(D)
Model B 1" in 24 HR QPF | Initialized: 20190621 at 00 UTC | Valid: 20190621 12 UTC - 20190622 12 UTC
GSS Value: 0.1717
CSI Value: 0.1929
MODEL B
MRMS QPE
(E)
Model A 2" in 24 HR QPF | Initialized: 20190621 at 00 UTC | Valid: 20190621 12 UTC - 20190622 12 UTC
GSS Value: 0.0911
CSI Value: 0.0941
MODEL A
MRMS QPE
(F)
Model B 2" in 24 HR QPF | Initialized: 20190621 at 00 UTC | Valid: 20190621 12 UTC - 20190622 12 UTC
GSS Value: 0.0805
CSI Value: 0.0834
MODEL B
MRMS QPE
50°N
40°N
30°N
120°W
110°W
100°W
90°W
80°W
MODEL A
MODEL B

(A)

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.98682	20.56934	23.82139	4867	4624	2258
CF002_CO002	0.95963	35.73147	7.62181	641	358	358
CF003_CO003	0.9185	61.8797	18.78736	1488	749	429
CF004_CO004	0.90118	10.19989	10.38279	58	322	22
CF005_CO005	0.97957	5.5663	29.49198	4055	3172	2357
CF006_CO006	0.94699	31.50353	50.27102	17586	18721	9432
CF007_CO007	0.63089	0.057041	40.595	73	65	0
CF008_CO008	0.9309	4.61943	62.65882	4123	2823	1350

(B)

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.97234	41.25991	26.47551	5112	4624	2315
CF002_CO002	0.96473	37.75615	4.59564	595	358	358
CF003_CO003	0.93633	9.91501	10.37415	121	322	62
CF004_CO004	0.97724	6.0333	32.458	3833	3172	2233
CF005_CO005	0.94828	28.0265	49.90328	18115	18878	7606
CF006_CO006	0.94082	2.66625	52.33143	4241	2823	1191
CF007_CO007	0.93399	54.48072	16.98403	1401	749	405
CF008_CO008	0.60103	81.99248	28.32377	126	65	0

(C)

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.96278	31.76173	13.51763	2111	1171	1080
CF002_CO002	0.90894	15.43263	18.40451	463	627	65
CF003_CO003	1	14.97736	6.50463	7071	8350	5452
CF004_CO004	0.96985	17.5382	17.34577	2573	1559	590

(D)

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.95177	27.10606	21.96651	3070	1559	519
CF002_CO002	0.77352	32.15294	44.36049	624	627	15
CF003_CO003	0.973	32.82324	13.988	1855	1171	1034
CF004_CO004	0.97771	32.25607	29.05061	9349	8350	3540

(E)

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.90252	15.127	4.90991	4137	672	510

(F)

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.87681	47.04978	33.88903	3193	672	406

Figure 44: *MODE 24 h precipitation results from the FV3-SAR (left column) and the FV3-Nest (right column) for (A)-(B) the half inch threshold, (C)-(D) the 1 inch threshold, and (E)-(F) 2 inch threshold valid*

from 1200 UTC 21 June 2019 to 1200 UTC 22 June 2019. The MRMS-GC QPE is shaded while the model QPF is contoured . Below each image is the MODE verification metrics for the corresponding model.

Final Recommendations to Operations for the Experimental Deterministic Models

The HRRRv4 is recommended for operations pending additional testing due to a bug that was discovered in the code during the experiment. The bug was in the initial conditions provided by the HRRRDAS. The HRRRv4 initializes off the mean state from the HRRRDAS along with current observations. However the bug was preventing current observations, like radar, from being assimilated into the mean state from the HRRRDAS. The HRRR team discovered this issue and implemented a fix to it on 16 July 2019, at the very end of the FFaIR Experiment. Therefore, although the model performed well and the addition of observed data into the initial conditions should only improve upon the model's performance, it can not be fully recommended for transition until further testing is completed on new cases or retroactively on the cases during FFaIR.

Both the FV3-Nest and FV3-SAR from EMC are recommended for further development and testing. Even though they both performed subjectively well, objective verifications showed that the models had a wet bias throughout the duration of the experiment, which was usually wetter than the NAM-Nest. This wet bias must be addressed before the models can go into operations. Additionally, as already stated, the FV3-SAR is not yet similar enough to the FV3-Nest to be considered as an alternate for the FV3-Nest. Lastly, although the FV3-SAR-GSD was not available for a good portion of the experiment, evaluation of the model on the days it was available suggest that it needs a significant amount of further testing and development before becoming operational.

Ensemble Model Guidance

There were two ensemble products that were evaluated during the 2019 FFaIR Experiment, the PMM and the LPMM. The model products were 6 h forecasts, valid from 1800 UTC to 0000 UTC. For subjective verification, evaluation was focused over a smaller sub-domain that was usually collocated with the domain for same day's PPF1 while the objective verification (MODE statistics) was done over the entire CONUS. The ensemble guidance that was analyzed can be found in Table 1; note that the NCAR Ensemble, though listed, will not be evaluated because it was only available during one week of the experiment. Additionally, as shown in Table 1, all ensemble guidance was initialized at 00z except one of the PMM products. This is because the HRRRE model group provided 00z and 12z runs for evaluation. The 12z run was only analyzed for the PMM product, and hereafter when discussing the PMM results the HRRRE guidance will be referred to by the initialization time: HRRRE 00z and HRRRE 12z.

Analysis of the Subjective Verification

The HREFv3 received the highest total average subjective score for both the 6 h QPF PMM and LPMM guidance during the experiment; with an average of 5.67 and 6.38 respectively. As can be seen in Table 7, the total averages for the experiment for all of the ensemble products evaluated were above a 5. Additionally, each ensemble's LPMM total average exceeded their PMM's total average, suggesting that the participants preferred the LPMM guidance. However, the largest takeaway from the subjective verification was the overwhelming preference for the HREFv3 in general, with the total average for the PMM from the HREFv3 exceeding the total averages not only for the PMM averages of all the other models but also the LPMM total averages of all the other models.

That being said, it is important to note that the HREFv3 was available to evaluate more times than any other ensemble. The HREFv3 had 194 total scores throughout the experiment compared to the HRRRE 00z which was evaluated 159 times. However the influence this had on the results was likely minimal. Looking at Table 7 and Figs. 45-47 it can be seen that HREFv3 score distribution was skewed towards higher end scores while the HRRRE and SSEF forecasts where skewed towards lower end values. For instance, the PMM product had nearly the same percentage of scores exceeding or equal to 7.5 (~16%) for the HREFv3, SSEF and HRRRE 00z ensembles. However when evaluating the lower end scores (<3.5) for these three ensembles only, 8% of the scores from the HREFv3 were less than 3.5 while 22% from the SSEF and 18% from the HRRRE 00z fell within this range. This further supports the conclusion that the HREFv3 performed the best according to the subjective results.

Despite the notable shift in the lower bounds of the subjective scores between each ensembles' PMM and LPMM products that can be seen in Table 7 and Fig. 45, the same trend was not seen when focusing specifically on higher end scores. Both the PMM products for the HRRRE 00z and the SSEF received a greater number of scores at or exceeding 7.5 when compared to their LPMM guidance, though the differences in scores varied; this was also the case when looking at the threshold of 8.5 or greater. The HRRRE 00z had the largest difference between the two products, with 25 instances in which the PMM guidance received a rating of 7.5 or higher but only 17 instances when the LPMM guidance did. The difference seen between the two products in this score range for the SSEF was much smaller, the SSEF PMM and LPMM guidances received a total of 26 and 24 respectively. Contrasting this, the number of scores exceeding or equal to 7.5 for the LPMM product from the HREFv3 doubled when compared to its PMM counterpart, 60 vs 30.

Table 7: *The total subjective average and median for the PMM and LPMM QPF ensemble products, along with the number of times each product received a score equal to or exceeding 7.5 and 8.5 and received a score less than 3.5 and 2.5 throughout the duration of the experiment.*

		HREFv3		**SSEF**		**HRRRE 00z**		**HRRRE 12z**	
PMM Total Average		5.67		5.06		5.34		5.05	
PMM Total Median		6		5		5.5		5	
PMM count (%) greater or equal to:		30 (16%)	7 (4%)	26 (15%)	7 (4%)	25 (16%)	3 (2%)	7 (5%)	2 (1%)
7.5	8.5								
PMM count (%) less than:		15 (8%)	3 (2%)	38 (22%)	11 (6%)	29 (18%)	11 (7%)	24 (16%)	12 (8%)
3.5	2.5								
LPMM Total Average		6.38		5.25		5.54			
LPMM Total Median		6.5		5.4		6			
LPMM count (%) greater or equal to:		60 (31%)	13 (7%)	24 (14%)	7 (4%)	17 (11%)	2 (1%)		
7.5	8.5								
LPMM count (%) less than:		6 (3%)	2 (1%)	29 (17%)	15 (9%)	14 (9%)	5 (3%)		
3.5	2.5								

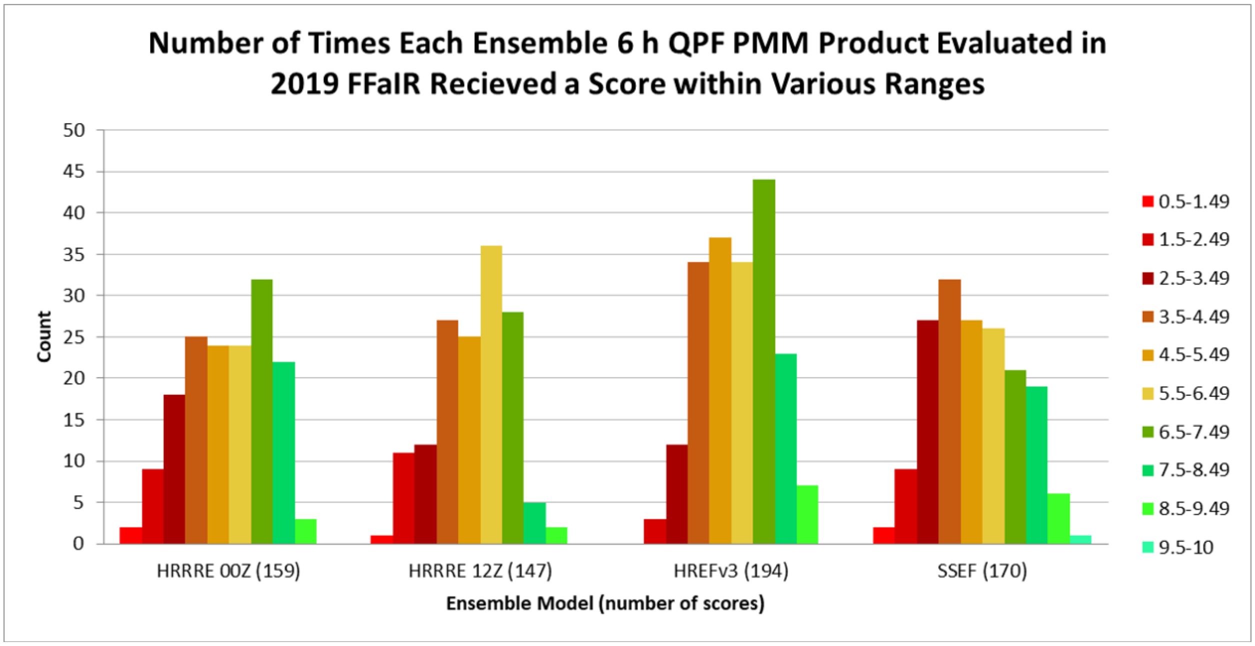


Figure 45: *The total number of times each 6 h QPF PMM ensemble product received a score that fell within a range of values throughout the 2019 FFaIR experiment. Along the bottom of the graph is the name of the ensemble, in parenthesis the total number of times the guidance was evaluated.*

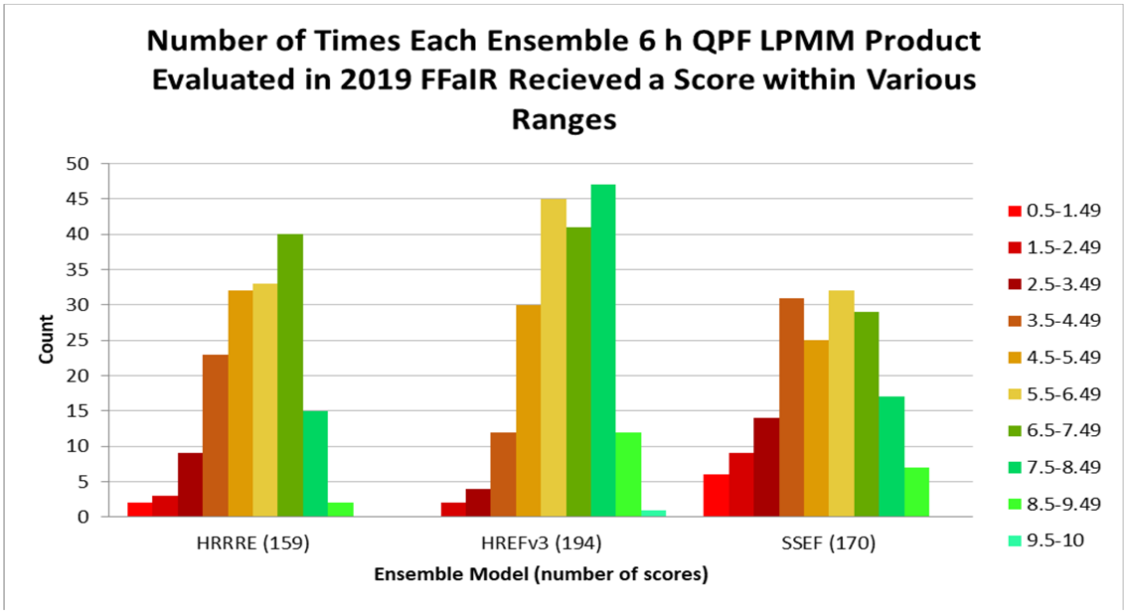


Figure 46: *The total number of times each 6 h QPF LPMM ensemble product received a score that fell within a range of values throughout the 2019 FFaIR experiment. Along the bottom of the graph is the name of the ensemble, in parenthesis the total number of times the guidance was evaluated.*

This puzzling result is most likely due to the order in which the ensemble products were shown. Addressing the latter influence first, throughout the experiment the participants noted that they often adjusted their scoring based off the score they gave to the first product shown. Simply put, they had trouble scoring the models/ensemble products only against observations. Instead they often compared the model/ensemble they were evaluating to other models/ensembles and tried to determine if the current model missed or hit the same events as the previous product. This often led them to say things like "I should not have given such a high score to the last model, this one is way better" or "I am running out of room to score models lower" or "wow, Model # did way better than Model #" or " this LPMM product has the same over-forecasted max seen in the PMM product." Such comments have led to the conclusion that always showing the model/ensemble guidance in the same order leads to some unintended biases in the subjective verification portion of the experiment, especially when evaluating the LPMM QPF products.

During the evaluation sessions of FFaIR, the order in which the PMM products were shown were: HRRRE 00z, HRRRE 12z, HREFv3, and lastly SSEF. The order for the LPMM guidance was: HREFv3, HRRRE 00z, and SSEF. As mentioned previously, analysis of the subjective scores shows that on average when comparing each ensemble LPMM product to its PMM product, the LPMM outperformed its counterpart. Additionally, as noted, the HREFv3 PMM product performed the best out of the ensembles evaluated during subjective verification. Therefore, it can be inferred that likely the best performing LPMM product should be from the HREFv3, which was seen. However, knowing that participants factor in how well the other models/ensembles did, and including the fact that the first LPMM product they were shown was from the HREFv3, it is likely that the high performance from the HREFv3 negatively impacted the scoring for the other two ensembles, especially the HRRRE which was shown immediately after the HREFv3. In other words, it is highly likely that even if the participants thought the ensemble LPMM they were evaluating was good, if they didn't feel it was better than the first one they saw they would reduce its score. Therefore it was "harder" for the SSEF and HRRRE 00z to receive some of the high end scores.

One such instance in which this scenario likely occurred was on June 18 across the Nebraska and Kansas region. As can be seen in Fig. 48, both the HRRRE 00z products incorrectly forecasted the extent of rainfall exceeding 0.75 inches and had the precipitation maximum located along the state border rather than to the north and south of the border; this misplaced maximum was also seen in the HREFv3 or SSEF products. However the extent of the precipitation maximum was smaller in the LPMM guidance than the PMM and the rainfall amounts were lower. Examination of MODE at the half inch threshold (Fig. 49) further supports the idea that the LPMM was likely a better forecast over this region, with the intersection area percentages 46% (PMM) to 63% (LPMM). Combined, this would suggest the LPMM provided a better forecast and therefore would receive a higher average score than its PMM counterpart. However this did not occur, instead the participants gave the PMM guidance a higher average score than the LPMM guidance, 6.14 vs 5.86.

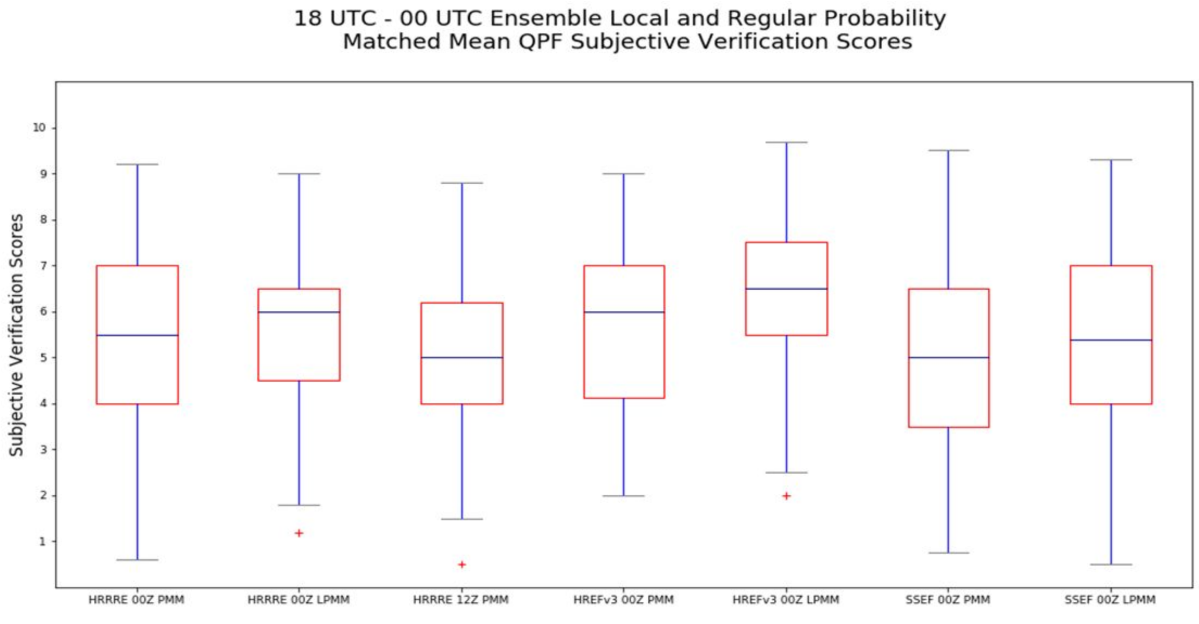


Figure 47: *Box and whisker plot of all the subjective scores given for the ensemble guidance evaluated in the 2019 FFaIR experiment. Subjective scores evaluated the 6 h QPF PMM and 6 h QPF LPMM products over various regions of the United States. All models were initialized at 0000 UTC, except for the HRRRE 12z which was initialized at 1200 UTC. The 6h forecast was from 1800 UTC to 0000 UTC.*

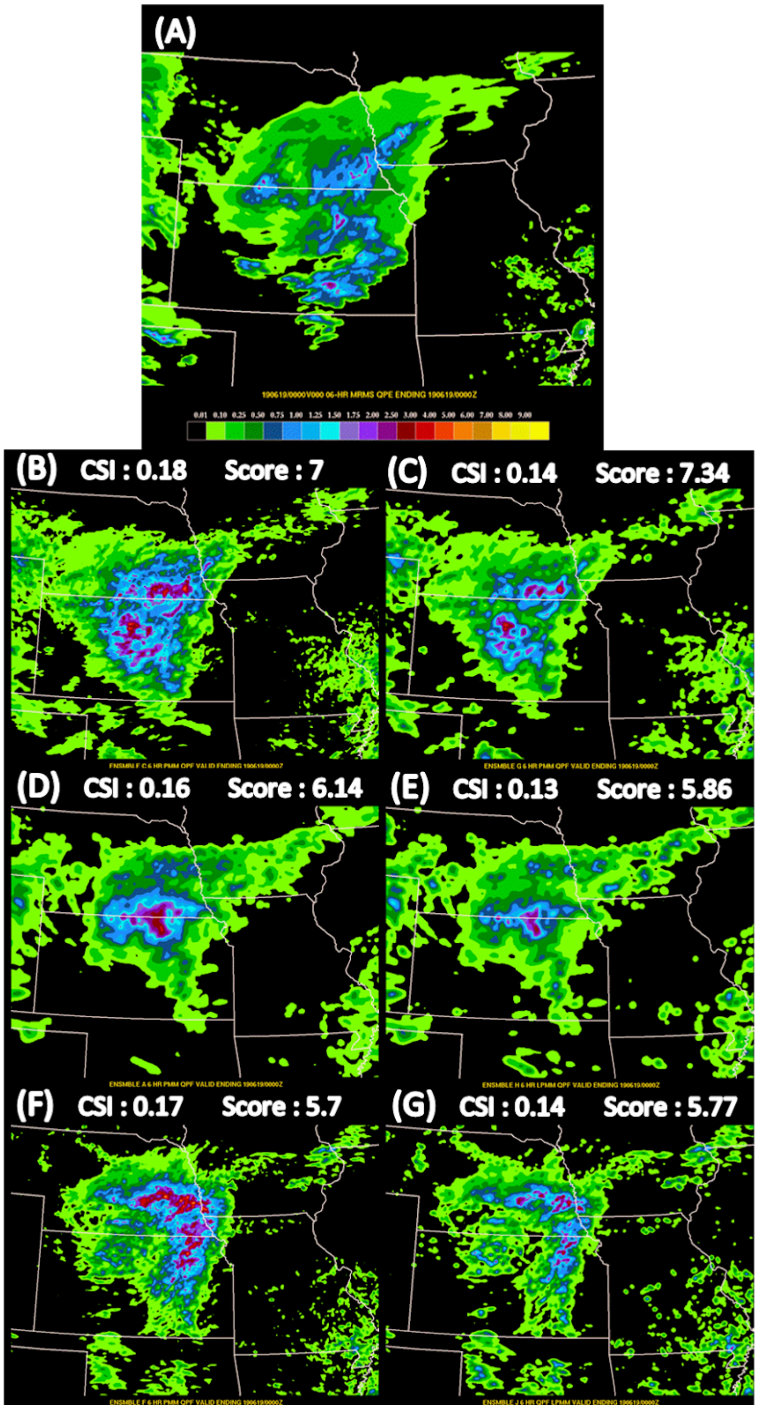
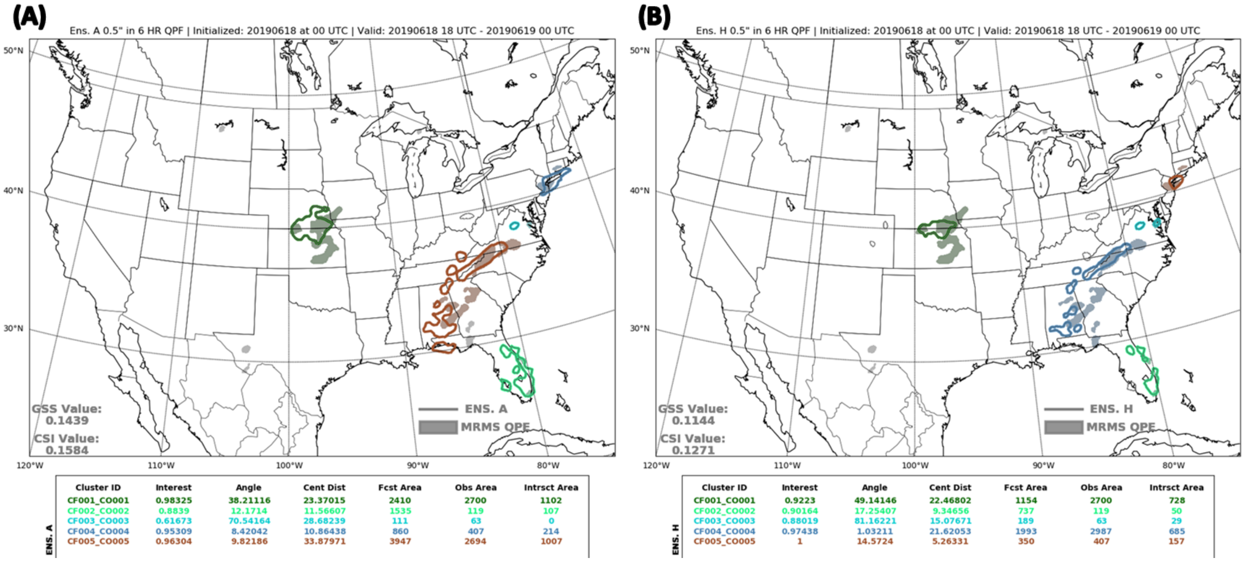

(A)
190619/0000V000 06-HR MRMS QPE ENDING 190619/0000Z
0.01 0.10 0.25 0.50 0.75 1.00 1.25 1.50 1.75 2.00 2.50 3.00 4.00 5.00 6.00 7.00 8.00 9.00
(B) CSI : 0.18 Score : 7
ENSMBLE C 6 HR PMM QPF VALID ENDING 190619/0000Z
(C) CSI : 0.14 Score : 7.34
ENSMBLE G 6 HR PMM QPF VALID ENDING 190619/0000Z
(D) CSI : 0.16 Score : 6.14
ENSMBLE A 6 HR PMM QPF VALID ENDING 190619/0000Z
(E) CSI : 0.13 Score : 5.86
ENSMBLE H 6 HR LPMM QPF VALID ENDING 190619/0000Z
(F) CSI : 0.17 Score : 5.7
ENSMBLE F 6 HR PMM QPF VALID ENDING 190619/0000Z
(G) CSI : 0.14 Score : 5.77
ENSMBLE J 6 HR QPF LPMM VALID ENDING 190619/0000Z

Figure 48: *(A) 6 hour MRMS-GC QPE and 6 hour PMM QPF (left column) and LPMM QPF (right column) from (B)-(C) HREFv3, (D)-(E) HRRRE 00z and (F)-(G) SSEF all valid 1800 UTC 18 June 2019 to 0000 UTC 19 June 2019. Along with each product is the corresponding daily CSI at a threshold of a half inch and the daily average score.*


(A)
Ens. A 0.5" in 6 HR QPF | Initialized: 20190618 at 00 UTC | Valid: 20190618 18 UTC - 20190619 00 UTC
50°N
40°N
30°N
GSS Value:
0.1439
CSI Value:
0.1584
ENS. A
MRMS QPE
120°W
110°W
100°W
90°W
80°W
ENS. A
(B)
Ens. H 0.5" in 6 HR QPF | Initialized: 20190618 at 00 UTC | Valid: 20190618 18 UTC - 20190619 00 UTC
50°N
40°N
30°N
GSS Value:
0.1144
CSI Value:
0.1271
ENS. H
MRMS QPE
120°W
110°W
100°W
90°W
80°W
ENS. H

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.98325	38.21116	23.37015	2410	2700	1102
CF002_CO002	0.8839	12.1714	11.56607	1535	119	107
CF003_CO003	0.61673	70.54164	28.68239	111	63	0
CF004_CO004	0.95309	8.42042	10.86438	860	407	214
CF005_CO005	0.96304	9.82186	33.87971	3947	2694	1007

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.9223	49.14146	22.46802	1154	2700	728
CF002_CO002	0.90164	17.25407	9.34656	737	119	50
CF003_CO003	0.88019	81.16221	15.07671	189	63	29
CF004_CO004	0.97438	1.03211	21.62053	1993	2987	685
CF005_CO005	1	14.5724	5.26331	350	407	157

Figure 49: *MODE precipitation results for the 0.5 inch threshold over 6 h valid from 1800 UTC 18 June 2019 to 0000 UTC 25 June 2019, showing the MRMS-GC QPE (shaded) compared to model QPF (contoured) from: (A) HRRRE 00z PMM and (B) HRRRE 00z LPMM. Below each image are the MODE verification metrics for the corresponding ensemble product.*

Another instance likely occurred during the verification for July 08 (refer to Fig. 50). On this day the participants specifically noted that they felt all three models provided a good forecast. Furthermore they noted that the HRRRE 00z PMM had a good QPF footprint but was too light on precipitation amounts. Therefore it would stand to reason that since the HRRRE 00z LPMM product had a similar QPF footprint but had higher precipitation totals (which were closer to observed amounts) that the average score for the LPMM product would be higher. However, as can be seen at the top of Fig. 50, this was not the case with the PMM average being 7.96 and the LPMM average of 7.08. It is likely that since the HRRRE 00z LPMM product was shown after the HREFv3 LPMM (which received an average of 8.46) that the participants scored the HRRRE LPMM against the HREFv3's performance rather than solely against observations. This hypothesis is further supported by the comments made by the participants that included: "Still good, just not quite as good as the previous ensemble."

It is also possible that the methodologies used by the ensembles to calculate the LPMM product influenced the subjective results. Despite having the same general concept for determining the LPMM, the method used by the HRRRE 00z differed from the method used by the HREFv3 and the SSEF, which used a nearly identical method to create their LPMM products. However it is difficult to determine whether or not the differences in LPMM generation impacted the results. Therefore, further evaluation on best practices for LPMM generation of ensemble QPF must be done.

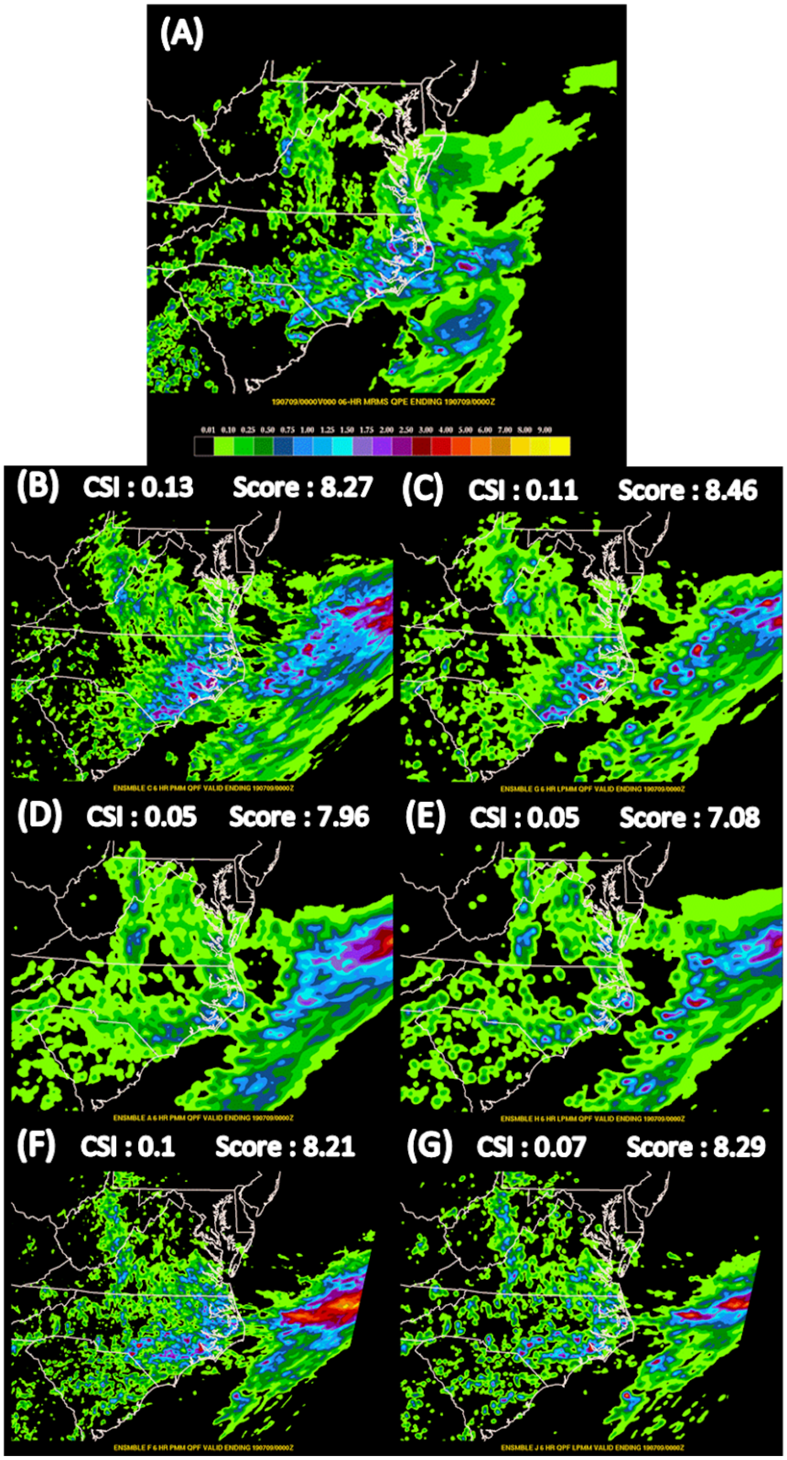

(A)
190709/0000V000 06-HR MRMS QPE ENDING 190709/0000Z
0.01 0.10 0.25 0.50 0.75 1.00 1.25 1.50 1.75 2.00 2.50 3.00 4.00 5.00 6.00 7.00 8.00 9.00
(B) CSI : 0.13 Score : 8.27
ENSMBLE C 6 HR PMM QPF VALID ENDING 190709/0000Z
(C) CSI : 0.11 Score : 8.46
ENSMBLE G 6 HR LPMM QPF VALID ENDING 190709/0000Z
(D) CSI : 0.05 Score : 7.96
ENSMBLE A 6 HR PMM QPF VALID ENDING 190709/0000Z
(E) CSI : 0.05 Score : 7.08
ENSMBLE H 6 HR LPMM QPF VALID ENDING 190709/0000Z
(F) CSI : 0.1 Score : 8.21
ENSMBLE F 6 HR PMM QPF VALID ENDING 190709/0000Z
(G) CSI : 0.07 Score : 8.29
ENSMBLE J 6 HR QPF LPMM VALID ENDING 190709/0000Z

Figure 50: *Same as Fig. 48 but valid 1800 UTC 08 June 2019 to 0000 UTC 09 June 2019.*

Analysis of the Objective Verification

Similar to the results from the subjective verification for the ensemble 6 h QPF PMM and LPMM guidance, the objective verification results are complicated. For each of the thresholds for which the two products were evaluated, the PMM guidance for each ensemble had a higher CSI than its counterpart LPMM; this can be seen in Fig. 51. For instance, at a threshold of 0.5 inches, the CSI for the HREFv3 PMM was 0.138 but the CSI for its LPMM was 0.112. while the LPMM had no bias and the PMM had a wet bias of 1.25. Following this, the SSEF PMM had a CSI of 0.120 but a wet bias around 1.61, while the SSEF LPMM had a CSI of 0.088 but only a slight wet bias of 1.14. Similar differences between the PMM and LPMM products' CSI and biases were seen at the one inch threshold.

Meanwhile the PMM and LPMM QPF guidance from the HRRRE 00z and HRRRE 12z all had a dry bias at the half inch and inch thresholds. Figure 51 shows that at both thresholds, the HRRRE 12z PMM product had the highest CSI, 0.109 and 0.047 respectively, while the HRRRE 00z LPMM product had the lowest CSI, 0.082 and 0.024 respectively. However the lowest frequency bias of the three products was seen in the HRRRE 00z PMM guidance at the one inch threshold, having nearly no frequency bias at the one inch threshold (1.039). Seeing such a stark difference in the bias performance of the HRRRE products compared to the SSEF and HREFv3 products helps support the previous hypothesis that the method for how the LPMM is calculated likely influenced the results seen in the analysis of the subjective verification.

Comparison of SREF and HREFv2 from EMC

Similar to what was done to evaluate the utility of the FV3-SAR against the FV3-Nest, the SREF was compared against the HREFv2. The participants were asked the same type of comparison question, only instead anything less than a 5 meant the SREF was more useful and anything greater than a 5 the HREFv2 was more useful. The two ensembles were evaluated using their 24 h QPF forecasts over the CONUS.

As can be seen in Fig. 52 the HREFv2 was overwhelmingly preferred over the SREF throughout the duration of the experiment. In total 93.9% of the scores were greater than a score 5, with 35.7% equal to or exceeding a score of 8.5. Furthermore, there were 37 times (18.9%) in which the participants completely preferred the HREFv2 over the SREF; i.e. giving a score of 10. Opposing this, only 1.5% of the scores were less than a 5, with no scores lower than 3.5. Examples of a day in which the HREFv2 scored at least one 10 and a day in which there was at least one score favoring the SREF can be seen in Fig. 52.

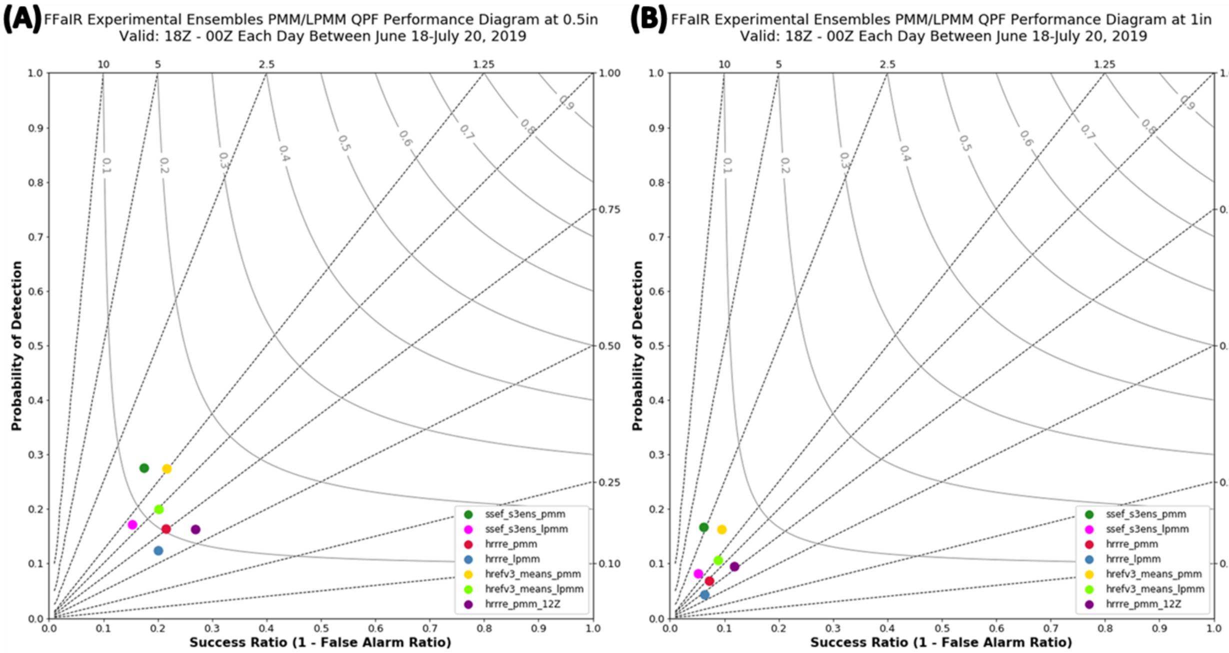


Figure 51: *Performance diagrams for the Day 1, 6 h PMM and LPMM QPF forecasts (1800 UTC to 0000 UTC) valid over the four weeks of the 2019 FFaIR experiment (June 17 to July 20, 2019) for the SSEF-PMM (dark green, SSEF-LPMM (pink), HRRRE 00z-PMM (red), HRRRE 00z-LPMM (teal), HREFv3-PMM (yellow), HREFv3-LPMM (light green), and HRRRE 12z-PMM (purple). The precipitation thresholds are for: (A) 0.5 inches and (B) 1 inch.*

During verification, the participants often noted that the values from the SREF were too low and that it often did not provide a good representation of what had occurred. The SREF often had widespread light precipitation and failed to highlight regions were heavy rainfall had fallen. It also seemed “washed out” though they did recognize that the SREF has a lower resolution than the HREFv2 so they were not surprised by this. Additionally, many of the participants stated that they do not often use the SREF during the warm season but they do use it during the cold season, saying that they have noticed that it seems to perform better in the cold season. Lastly, they noted that although they are aware of the failings of the SREF they feel they do not have a better model to use in the 2 to 3 day range that provides the resolution they would like to use for forecasting in this time period.

Overall, although the results from the subjective verification strongly suggests that participants almost always preferred the HREFv2 to the SREF when it comes to QPF, it seems as if the SREF still has utility. However, the utility appears to be more related to what forecasters are lacking in the forecast process rather than performance of the SREF itself. It is therefore recommended that until there is a replacement for the SREF in the 2 to 3 day range and there is an ensemble that performs well in the cold season that the SREF continue to be available to the forecasters.

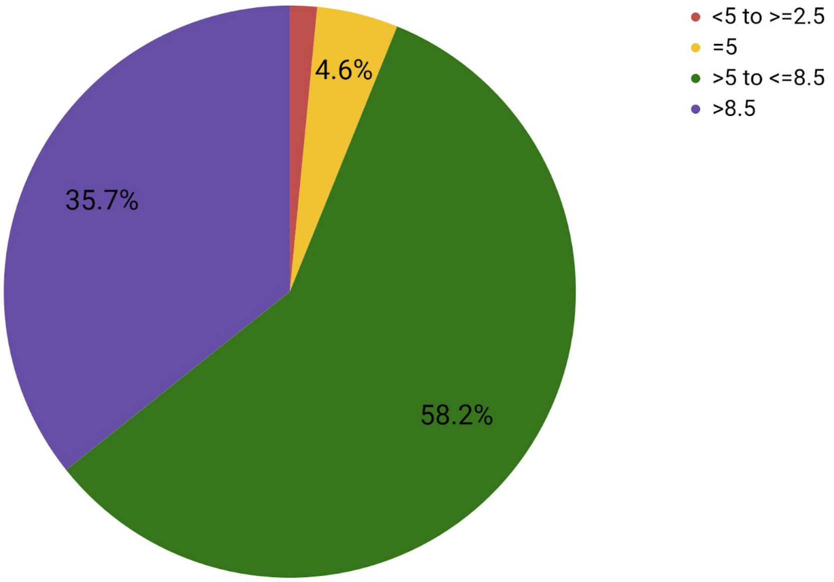


Figure 52: *Pie chart depicting the results from the verification question comparing the SREF v HREFv2, showing the percentage of times in which a score was recorded within various ranges. Anything below a 5 represents the participant finding more utility in the SREF than the HREFv2, while above the 5 means the HREFv2 was more useful. A score of 5 suggests they provided the same amount of utility. The "red slice" representing the <5 to >=2.5 range is 1.5%.*

Final Recommendations to Operations for the Experimental Ensemble Products

All three of the ensembles evaluated during FFaIR are recommended for further testing and development. The HREFv3 would have been recommended for transition into operations but the models that go into the ensembles are not yet frozen. Therefore, it is possible changes to these models will impact the performance of the HREFv3.

In regards to the LPMM, FFaIR advises that the product be readily available for NWS forecasters to use for all ensembles. However, before this can be done, a consensus on how to calculate the LPMM must be done. Based on objective and subjective results from the study, the method the HREFv3 and SSEF used to calculate the LPMM performed better than the method used by HRRRE. Therefore, this calculation should be the one the NWS uses moving forward.

Day 1 CSU-MLP ERO and experimental ERO Results

The two versions of the CSU-MLP EROs (referred to as the GEFS ERO and the NSSL ERO) and the experimental ERO were all evaluated a variety of ways, including being compared against one another. However it is important to remember that a direct comparison between the machine learned EROs and the WPC operational ERO to the experimental ERO must be viewed cautiously because of the differences in the valid times of the products, with the experimental ERO being valid 1500 UTC to 1200 UTC rather than 1200 UTC to 1200 UTC like the other products. In addition, the GEFS ERO and the NSSL ERO are used to create the experimental ERO, and can not be considered completely independent.

The probability of being within a certain ERO category was plotted to show the distribution of the various products across the CONUS throughout the experiment. Figure 53 shows the probability of being in a slight risk contour for all the products except for PFF3 while Fig. 54 shows the same probability but for a moderate risk. Comparing the chance of being under a slight or moderate risk contour between the CSU-MLP EROs and the experiment ERO, there is a notable difference between the experimental ERO and the CSU-MLP EROs. The experimental EROs probabilities were lowest and covered the smallest spatial extent while the GEFS ERO was greatest in both aspects. As can be seen in Fig. 55, this was also the case for the marginal risk contour. This suggests that the machine learned products might have a general tendency of overforecasting the flooding risk. This appears to be especially true across the northern Plains and northern Rockies in the NSSL ERO. This overpredicting of the flooding likelihood in these regions by the NSSL ERO was noted repeatedly by the participants throughout the experiment.

An example of the overforecasting across the northern Rockies and the northern Plains was seen on June 19, 2019. Fig. 56. shows that the NSSL ERO forecasted a marginal risk across a broad region along the northern states west of the Mississippi River when the practically perfect verification (Fig. 56D) suggests the only region that required a marginal risk was extreme northeastern MT. Neither the GEFS ERO nor the experimental ERO had such a large marginal risk across this region. Additionally, Fig. 53 and Fig. 55 show that there also appears to be an over forecast of the flooding risk across the Ohio River Valley and the Mid-Atlantic into the Northeast in the GEFS ERO. However the participants made little comment regarding this region being overpredicted by the GEFS ERO.

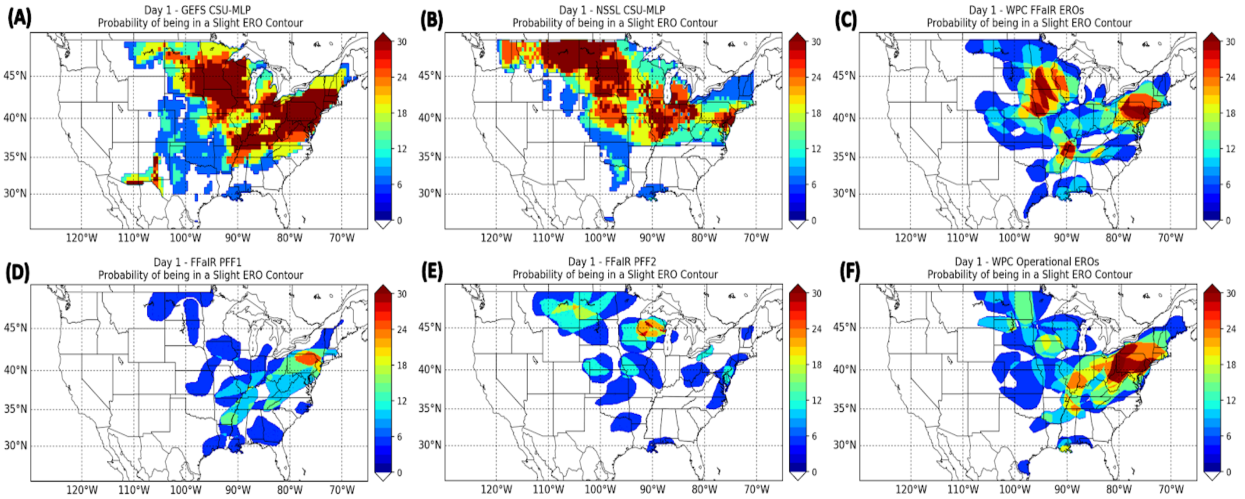

(A)
Day 1 - GEFS CSU-MLP
Probability of being in a Slight ERO Contour
(B)
Day 1 - NSSL CSU-MLP
Probability of being in a Slight ERO Contour
(C)
Day 1 - WPC FFaIR EROs
Probability of being in a Slight ERO Contour
(D)
Day 1 - FFaIR PFF1
Probability of being in a Slight ERO Contour
(E)
Day 1 - FFaIR PFF2
Probability of being in a Slight ERO Contour
(F)
Day 1 - WPC Operational EROs
Probability of being in a Slight ERO Contour
45°N
40°N
35°N
30°N
120°W
110°W
100°W
90°W
80°W
70°W
30
24
18
12
6
0

***Figure 53:** The probability of being under a Slight Risk contour throughout the duration of the 2019 FFaIR experiment for the (A) Day 1 GEFS ERO, (B) Day 1 NSSL ERO, (C) FFaIR Day 1 Experimental ERO, (D) PFF, (E) PFF2 and (F) the WPC Operational ERO.*

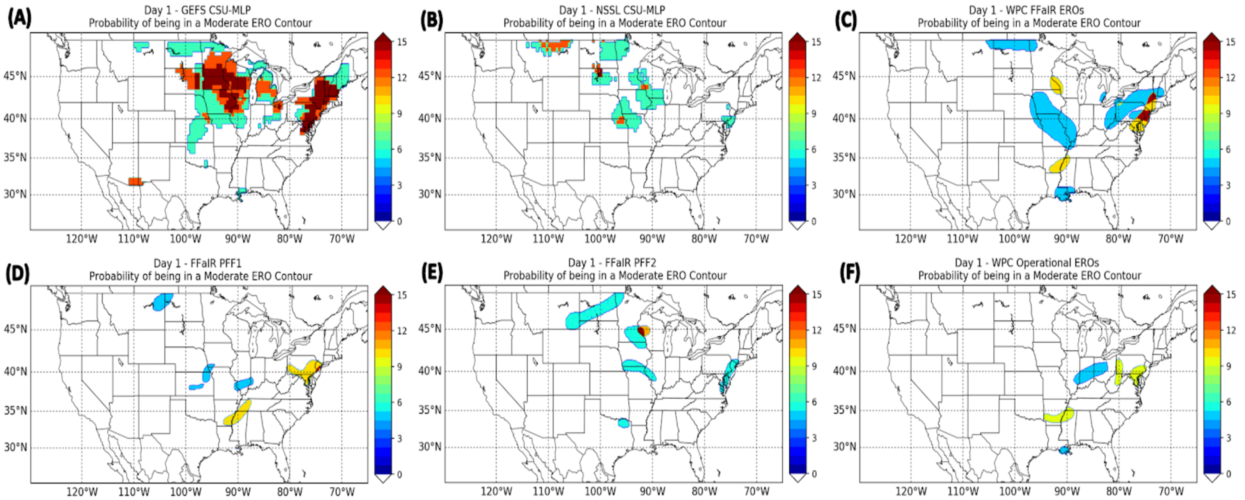

(A)
Day 1 - GEFS CSU-MLP
Probability of being in a Moderate ERO Contour
(B)
Day 1 - NSSL CSU-MLP
Probability of being in a Moderate ERO Contour
(C)
Day 1 - WPC FFaIR EROs
Probability of being in a Moderate ERO Contour
(D)
Day 1 - FFaIR PFF1
Probability of being in a Moderate ERO Contour
(E)
Day 1 - FFaIR PFF2
Probability of being in a Moderate ERO Contour
(F)
Day 1 - WPC Operational EROs
Probability of being in a Moderate ERO Contour
45°N
40°N
35°N
30°N
120°W
110°W
100°W
90°W
80°W
70°W
15
12
9
6
3
0

***Figure 54:** Same as Fig. 53 but for the probability of being under a Moderate Risk.*

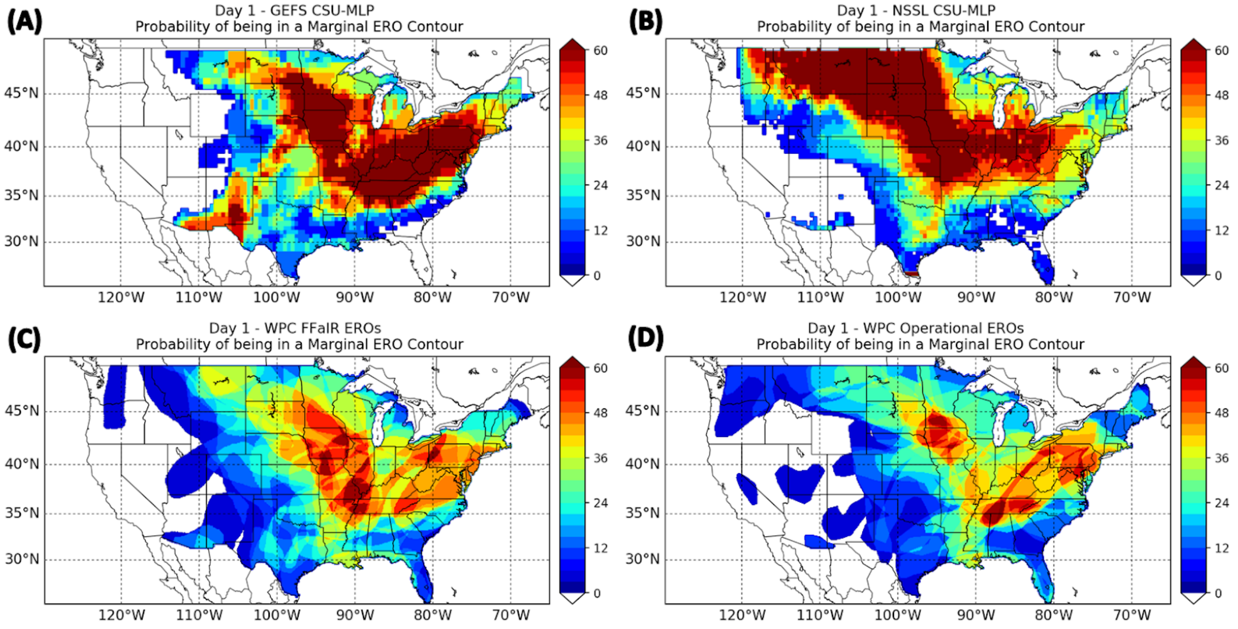

(A)
Day 1 - GEFS CSU-MLP
Probability of being in a Marginal ERO Contour
(B)
Day 1 - NSSL CSU-MLP
Probability of being in a Marginal ERO Contour
(C)
Day 1 - WPC FFaIR EROs
Probability of being in a Marginal ERO Contour
(D)
Day 1 - WPC Operational EROs
Probability of being in a Marginal ERO Contour
45°N
40°N
35°N
30°N
120°W
110°W
100°W
90°W
80°W
70°W
60
48
36
24
12
0

Figure 55: *The probability of being under a Marginal Risk contour throughout the duration of the 2019 FFaIR experiment for the (A) Day 1 GEFS ERO, (B) Day 1 NSSL ERO, (C) FFaIR Day 1 Experimental ERO and (D) WPC Operational ERO.*

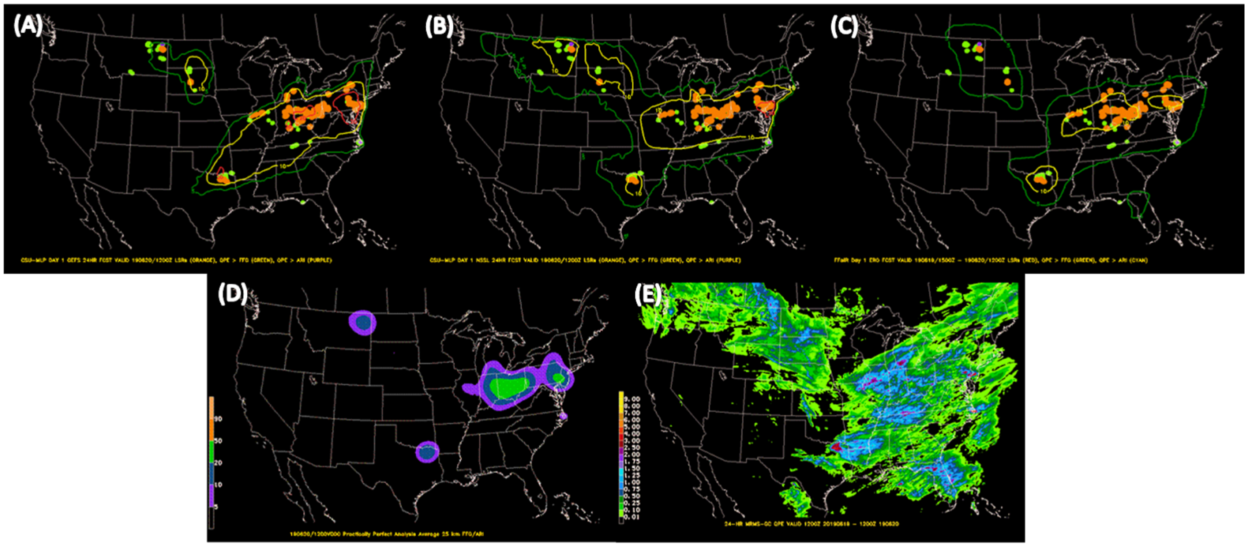

(A)
(B)
(C)
(D)
(E)
CSU-MLP DAY 1 GEFS 24HR FCST VALID 190620/1200Z LSRs (ORANGE), QPE > FFG (GREEN), QPE > ARI (PURPLE)
CSU-MLP DAY 1 NSSL 24HR FCST VALID 190620/1200Z LSRs (ORANGE), QPE > FFG (GREEN), QPE > ARI (PURPLE)
FFaIR Day 1 ERO FCST VALID 190619/1500Z - 190620/1200Z LSRs (RED), QPE > FFG (GREEN), QPE > ARI (CYAN)
10
5
90
50
20
10
5

Figure 56: *(A) GEFS ERO, (B) NSSL ERO, (C) FFaIR Experimental ERO, (D) Practically Perfect analysis and (E) 24 h MRMS-QPE. (A), (B), (D), and (E) are all valid from 1200 UTC 19 June 2019 to 1200 UTC 20 June 2019 and (C) is valid 1500 UTC 19 June 2019 to 1200 UTC 20 June 2019. (A)-(C) the contour coloring follows Table 2. The dots plotted on top of the EROs represent where QPE > FFG (green), QPE > ARI (purple), and flooding/flash flooding storm reports (orange). (D) the purple shading corresponds to a marginal, blue to a slight, green to a moderate, and orange to a high risk*[11]*.*

[11] Additional information about the practically perfect analysis can be found in Section 3.

Analysis of the Subjective Verification

Figure 57 shows how each of the three experimental EROs subjectively performed over the course of the experiment. The ERO produced daily by the FFaIR participants had the highest average total score (6.7) followed by the GEFS ERO (6.07) and then the NSSL ERO (5.76). The experimental ERO also had the highest weekly average each week. This suggests that the participants felt that their forecasts usually improved upon the machine learned products. An example of this can be seen in Fig. 58, where the average daily score for the June 17th forecast was 6.33 for the GEFS ERO, 4.54 for the NSSL ERO, and 7.46 for the experimental ERO.

Despite the FFaIR ERO generally outperforming the machine learning EROs there were a few times in which the participants felt the machine learned forecasts were on par or better than their ERO forecast. An example in which the participants felt all three EROs were good forecasts was the June 17-18 forecast seen in Fig. 58, though they felt the NSSL ERO was the best, receiving an average subjective score of 7.6. The FFaIR ERO had an average score of 7.24 while the GEFS ERO had a 7.09 for this event. Figure 59, on the other hand, shows a time in which the participants felt the FFaIR ERO performed poorly. For this forecast, July 10-11, the participants decided not to include the Ohio River Valley in a flooding risk despite both of the machine learning ERO highlighting at least a marginal threat, with the NSSL ERO suggesting a slight risk from MO to IN. Based on the practically perfect analysis, the NSSL's forecast appears to have been superior because the practically perfect suggests a slight risk was warranted for the area. The participants generally agreed, giving the NSSL ERO an average score of 7.43. The GEFS ERO was considered the next best, receiving an average of 5.71 while the FFaIR ERO was considered the worst, getting an average score of 4.64.

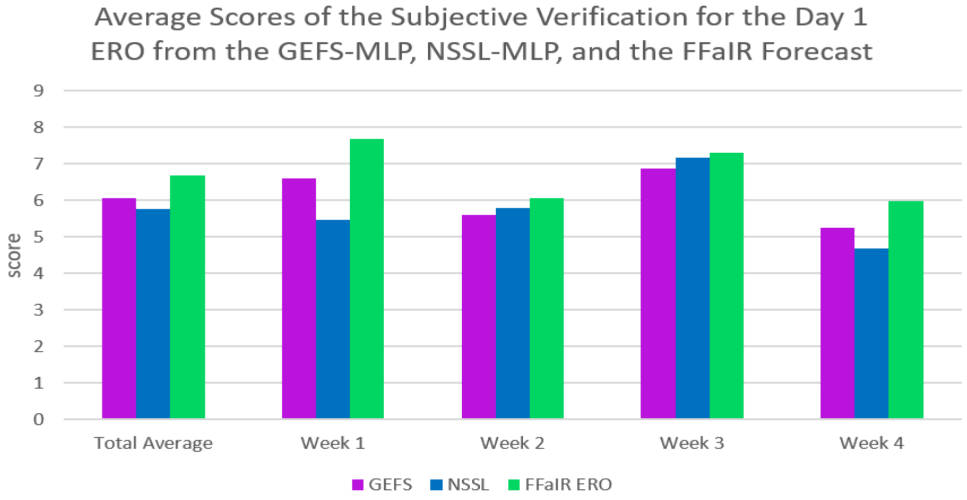


Figure 57: *The total average score and the weekly average scores for the Day 1 ERO from the GEFS-MLP (purple), NSSL-MLP (blue), and the experimental FFaIR forecast (green).*

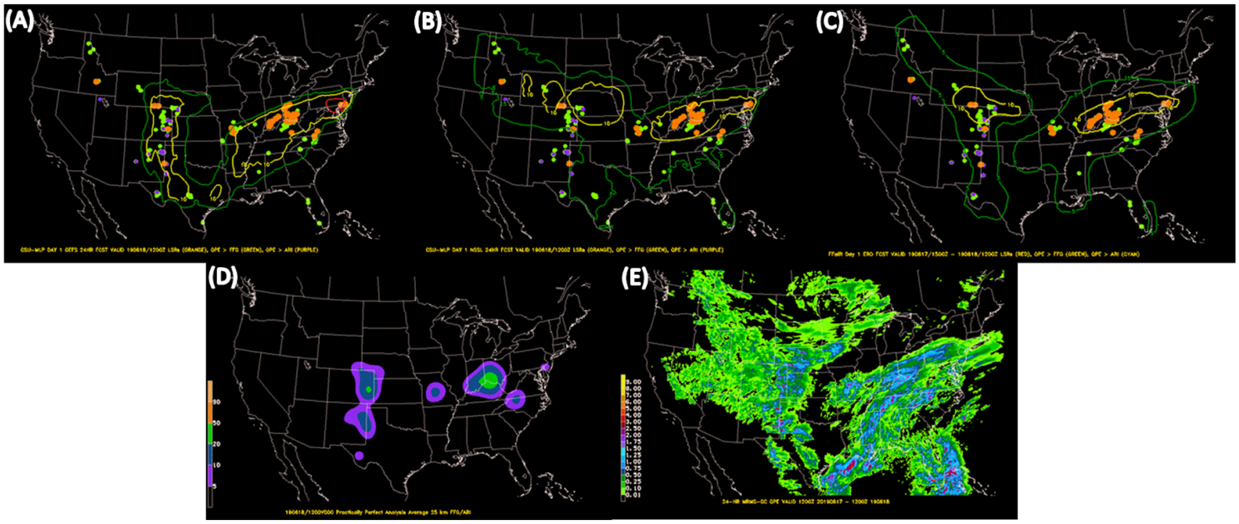


Figure 58*: Same as Fig. 56 but valid 17 June 2019 to 18 June 2019.*

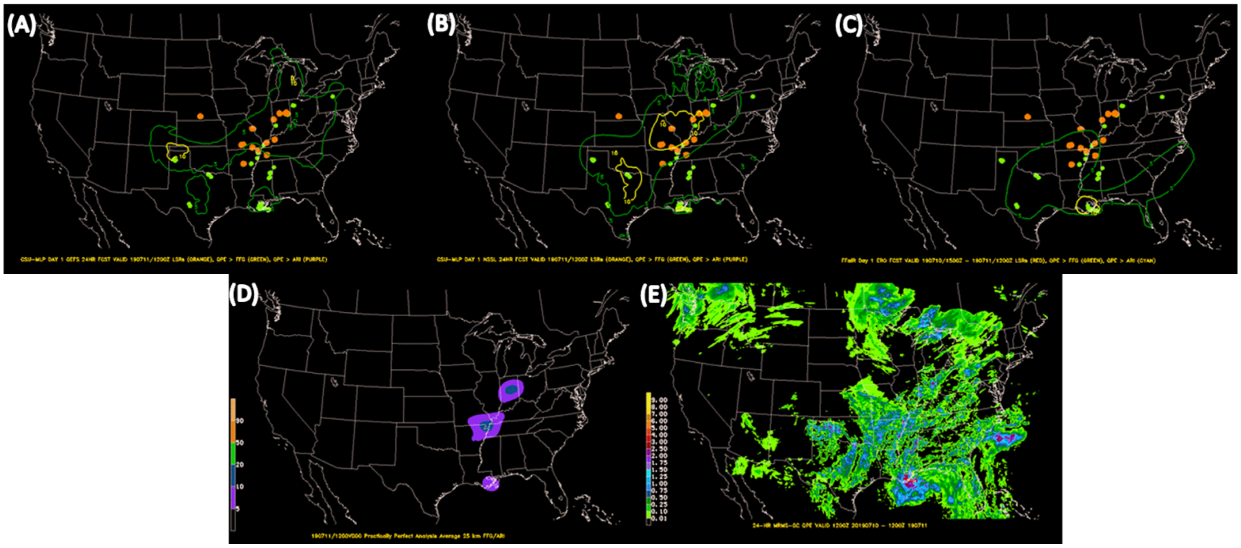


Figure 59: *Same as Fig. 56 but valid 10 July 2019 to 11 July 2019.*

Consistent with the average subjective ratings, participants felt that the areal extent of the risk areas from the machine learned forecasts was too large, particularly for the marginal risk. They felt this led to a high false alarm rate. There were many times participants stated they thought the marginal risk captured the overall rainfall pattern but not necessarily the flood risk. Because of this, their process for creating the ERO often began by looking at what parts of the marginal contours in the machine learned ERO they wanted to "get rid of" or what cohesive risk regions they wanted to separate. Both Fig. and Fig. (June cases) are examples of how the FFaIR participants separated the large marginal risk from the machine learning EROs (especially the one from the NSSL) into two separate regions.

<u>Analysis of the Objective Verification</u>

The experimental products were examined objectively a variety of ways, including comparing the experimental EROs against the WPC operational ERO. Additionally two sets of verification criteria were used, FFG only (QPE > FFG) and the UFV (QPE > FFG, QPE > ARI, and flooding/flash flooding storm reports; referred to as "All" in chart legends. The following discussion will focus on the results using the UFV criteria unless otherwise specified.

Evaluation of the bulk Brier Score and the bulk AuROC[12] suggests that the various EROs performed about the same throughout the experiment. Figure 60 shows how small the differences in these scores were. The bulk Brier Score (lower values indicate less error) for both the operational and experimental ERO was a 0.015, while the GEFS ERO had a score of 0.0152 and the NSSL ERO had the "best" score at 0.0148. Meanwhile the AuROC scores suggest the "best" EROs (higher values are better) were the experimental ERO and the GEFS ERO, each with a value of 0.81, followed by the operational ERO (0.82) then the NSSL ERO (0.84). Note that AuROC measures the ability of the forecast to discriminate between events and non-events, while Brier Score measures the magnitude of the probability forecast errors (i.e. they measure two different things). Overall these results suggest that all the EROs performed well during the course of the FFaIR experiment, with no one ERO significantly performing better than any other.

Using the Brier Skill Score (BSS) as a verification metric however indicates that the WPC operational ERO outperformed all other EROs. Figure 61 shows the experimental FFaIR ERO BSS, the GEFS ERO BSS, and the NSSL ERO BSS all referenced against the operational WPC ERO, where a negative value means the operational ERO performed better and a positive value means it performed worse. As can be seen in Fig. 61A , for the majority of the days of the experiment, the operational ERO performed better than the experimental FFaIR ERO. There were six days, however, in which the experimental FFaIR ERO outperformed the operational ERO, though this was never by a large amount. Figure 61B-C shows how the GEFS ERO and NSSL ERO compared against the operational ERO. As can be seen, the GEFS ERO only performed better than the operational ERO three times throughout the experiment, with the NSSL ERO only outperforming the operational ERO twice. Additionally, there was one day, July 18, in which the machined learned products noticeably underperformed. Each ERO for this day can be seen in Fig. 62 along with the 24 h QPE and the practically perfect analysis. As can be seen, on this day both the GEFS and NSSL EROs forecasted a moderate risk across the Midwest and highlighted a flooding risk in regions along the Mexico border. The NSSL ERO also included a

[12] Au: Area under the curve; ROC: Receiver Operating Characteristic. ROC measures the ability of the forecast to discriminate between events and non-events. AuROC integrates the area under the curve to produce a single value.

marginal risk in the Northern Rockies. Verification shows that in all these instances the machine learned products over forecast the flooding threat.

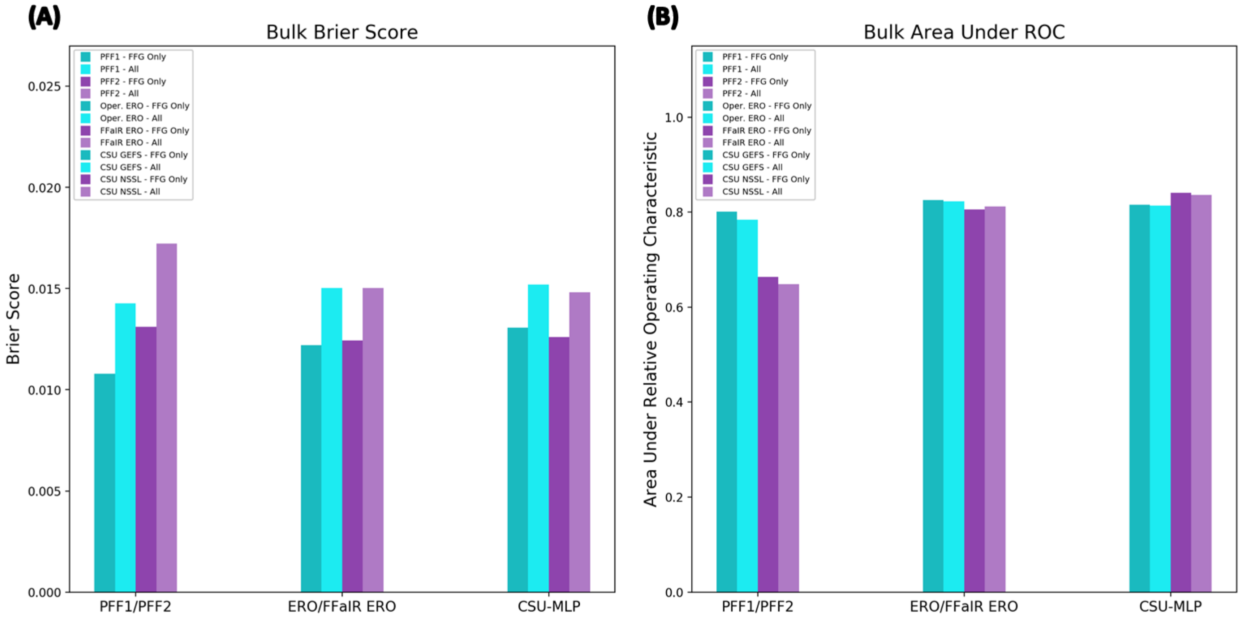


Figure 60: *(A) the bulk Brier Scores and (B) bulk AuROC for the WPC operational ERO, the FFaIR experimental ERO, and the CSU-MLP EROs from the GEFS and the NSSL as well as the PFF1 and PFF2. Each product was verified against FFG only and using the UFV system.*

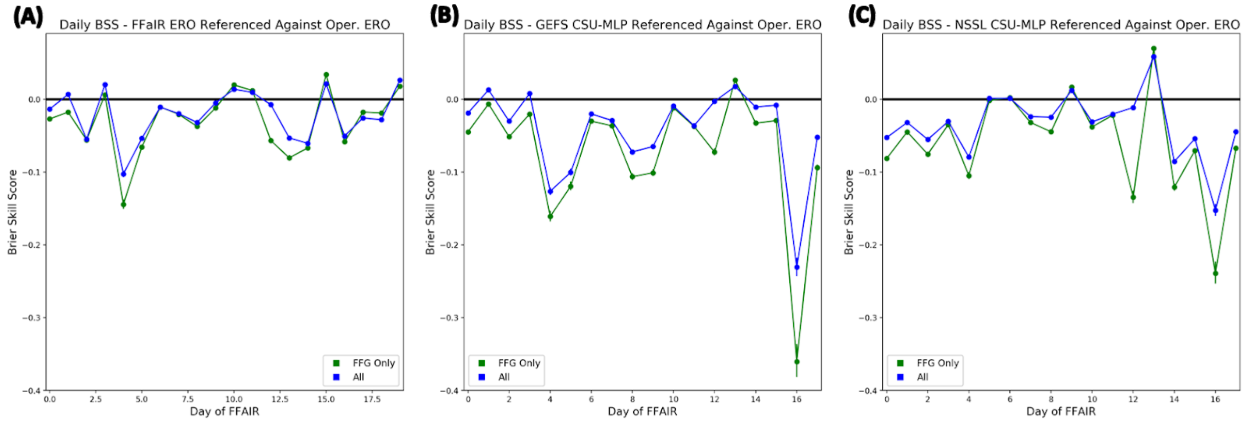


Figure 61*: The daily Brier Skill Score throughout the FFaIR Experiment referenced against the operational ERO for the (A) FFaIR experimental ERO, (B) GEFS ERO, and (C) NSSL ERO. The green line is using only FFG for verification and the blue line is using the UFV system for verification. The negative values mean the operational ERO had better skill.*

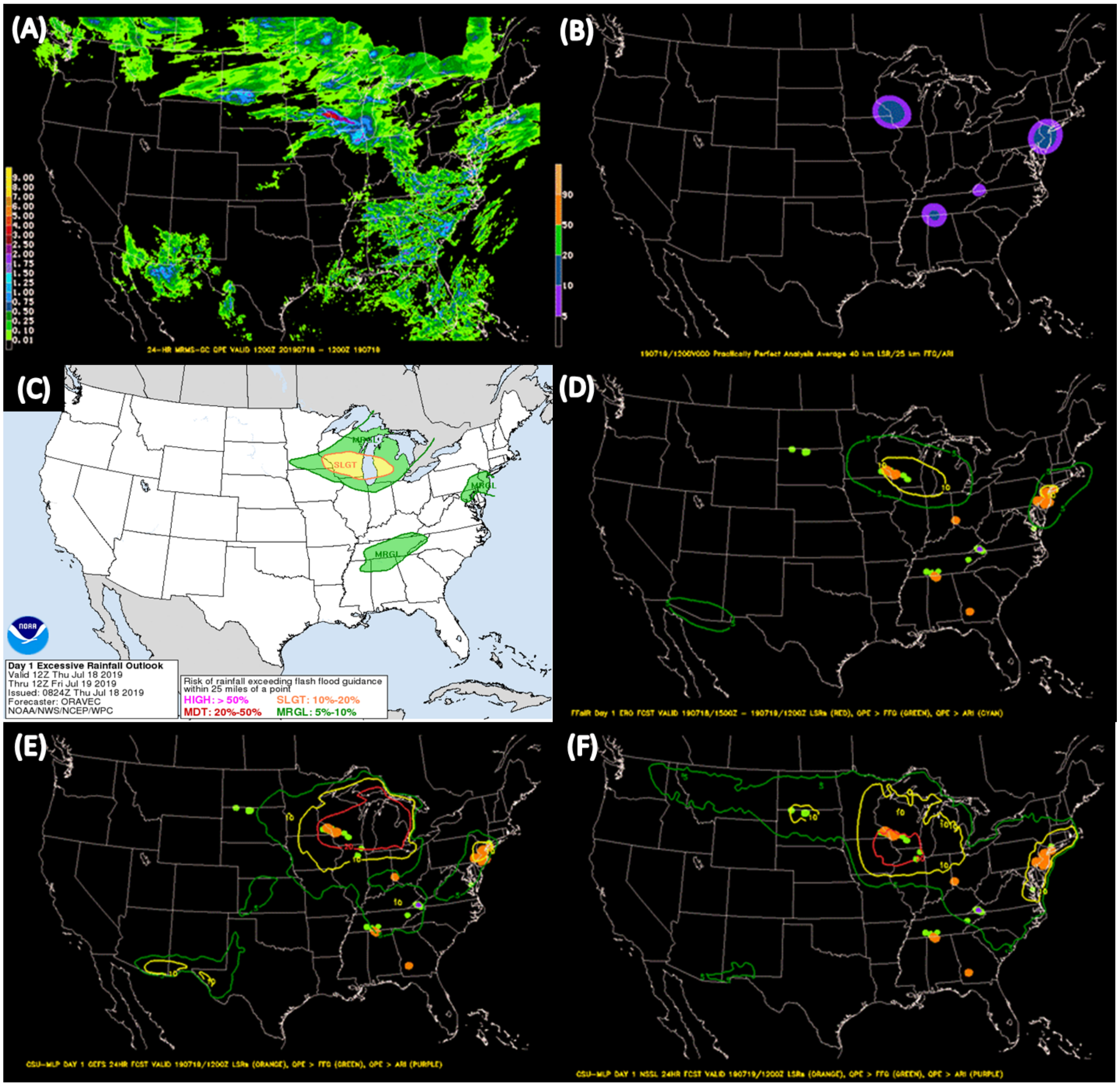

(A)
9.00
8.00
7.00
6.00
5.00
4.00
3.00
2.50
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.10
0.01
24-HR MRMS-GC QPE VALID 1200Z 20190718 - 1200Z 190718
(B)
90
50
20
10
5
190719/1200V000 Practically Perfect Analysis Average 40 km LSR/25 km FFG/ARI
(C)
MRGL
SLGT
MRGL
MRGL
Day 1 Excessive Rainfall Outlook
Valid 12Z Thu Jul 18 2019
Thru 12Z Fri Jul 19 2019
Issued: 0824Z Thu Jul 18 2019
Forecaster: ORAVEC
NOAA/NWS/NCEP/WPC
Risk of rainfall exceeding flash flood guidance within 25 miles of a point
HIGH: > 50%
SLGT: 10%-20%
MDT: 20%-50%
MRGL: 5%-10%
(D)
FFaIR Day 1 ERO FCST VALID 190718/1500Z - 190719/1200Z LSRs (RED), QPE > FFG (GREEN), QPE > ARI (CYAN)
(E)
CSU-MLP DAY 1 GEFS 24HR FCST VALID 190719/1200Z LSRs (ORANGE), QPE > FFG (GREEN), QPE > ARI (PURPLE)
(F)
CSU-MLP DAY 1 NSSL 24HR FCST VALID 190719/1200Z LSRs (ORANGE), QPE > FFG (GREEN), QPE > ARI (PURPLE)

Figure 62: *(A) 24 h MRMS-QPE, (B) Practically Perfect, (C) WPC operational ERO, (D) FFaIR experimental ERO, (E) GEFS ERO, and (F) NSSL ERO all but (D) valid 1200 UTC 9 July to 1200 UTC 10 July. (D) valid 1500 UTC 9 July to 1200 UTC 10 July. (C)-(F) the contour coloring follows Table . The dots plotted on top of the EROs represent where QPE > FFG (green), QPE > ARI (purple), and flooding/flash flooding storm reports (orange). (B) the purple shading corresponds to a marginal, blue to a slight, green to a moderate, and orange to a high risk.*

The Day 1 fractional coverage for each ERO can be seen in Fig. 63. Fractional coverage is used to assess the calibration of the forecasts and when binned by ERO category is analogous to a reliability plot. In general, the results from the fractional coverage suggests that the machine learning products are well calibrated when it comes to identifying slight risks but fall on the low end or just below the probabilistic definition for marginal and moderate risks. It is likely that the CSU methods are producing too large or displaced areas for these risk categories, especially in regards to the NSSL marginal risk. This helps support the comments made by the participants stating that they noticed that the NSSL seemed to over forecast the risk of flooding in the Northern Rockies and Northern Plains.

Opposing this, both the operational and experimental FFaIR EROs appear to be well calibrated at the marginal risk threshold. The operational and experimental FFaIR EROs produce fractional coverage values above their probabilistic definition for slight, suggesting that forecasters did not produce enough slight areas or drew slight areas that were too small. Even so, the experimental FFaIR ERO moderate risk seems to be well calibrated, performing better than the operational moderate risk. Finally, the operational ERO is calibrated for slight with 100% fractional coverage, but not with the experimental FFaIR ERO, although the sample size is limited at this threshold. Neither of the machine learned ERO issued a high risk, while a high risk was only issued once for the experimental ERO; issued on June 21 and can be seen in Fig. 2A.

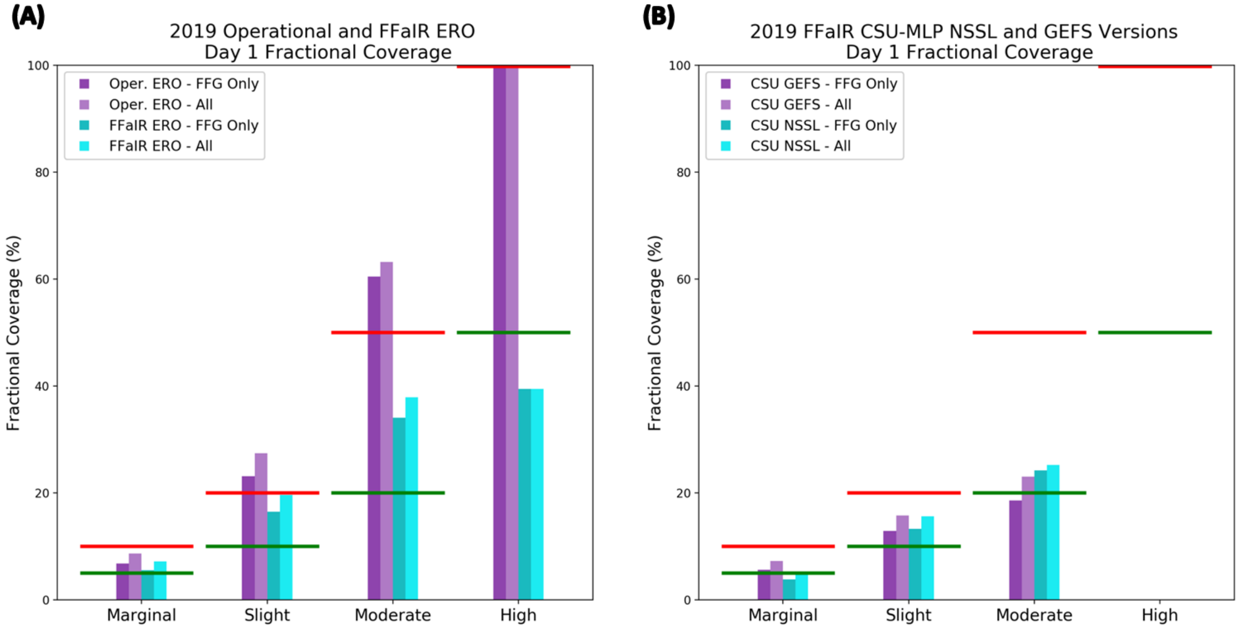


Figure 63: *The fractional coverage of the 2019 Day 1 (A) WPC operational ERO (purple) and FFaIR experimental ERO (blue) and (B) GEFS ERO (purple) and NSSL ERO (blue) for each probabilistic category. Green horizontal lines represent the lower defined bound for each threshold and red horizontal lines represent the higher defined bound. A calibrated forecast should lie between the two bounds.*

Final Recommendations for CSU First Guess Fields for Day 1 ERO

Both of the CSU-MLP first guess fields are recommended for further testing and development. Although both the GEFS and NSSL Day 1 EROs performed well and the participants responded favorably to them, there are a few issues that need to be addressed. The first is the over forecasting seen in the NSSL ERO across the Northern Rockies and Northern Plains. Not only is this impacting the reliability of the product at lower risk thresholds, it likely impacting it at higher risk thresholds as well. Furthermore, the over forecasting is influencing how the participants respond to the product when it highlights a slight and/or moderate risk across these regions. Often forecasters would reduce the CSU product by one probabilistic category (e.g. reduce a moderate to a slight). The GEFS ERO generally also over forecasts, though is somewhat more reliable than the NSSL ERO.

Analysis of FFaIR PFFXs

Although the PFFX products are for research purposes only and are not planned to become operational, their forecast performance throughout the experiment will be briefly discussed. During the course of the analysis it is important to remember that a direct comparison between three PFFs can not truly be done because of the differences in the areal coverage of the products and the amount of time the products are valid for (6 h vs 3h). Additionally, keep in mind the valid times differ as well, with the PFF1 valid 1800-0000 UTC, the PFF2 valid 0000-0600 UTC and the PFF3 is only valid 2100-0000 UTC.

Focusing on the PFF products, Fig. 64 shows that over the course of the experiment PFF3 received the highest average score (6.35) followed by PFF1 (5.85) and then PFF2 (5.7). However, when looking at the weekly averages, it can be seen that the PFF3 forecast was not always the highest performing PFF. For instance, during Week 2 the PFF1 received the highest average score, 5.56, with the PFF3 having the second highest (4.26) and lastly the PFF2 (3.66). Another takeaway from Fig. 64, is that lowest weekly score also varied among the forecasts.

Additionally, there was a large spread in the scores seen for each of the PPFs throughout the experiment. The box and whisker plots in Fig. 65 summarizes this. As can be seen, all three PFFs received scores higher than 8.5 and lower than 2.5 during the course of the experiment, though the PFF3 saw the least amount of these low-end numbers. The large spread in the products' scores, along with the variance in weekly average scores, was likely driven by event predictability. For instance, the pattern on July 9 featured a 1006mb low over the Northern Plains with precipitation wrapping around the low. Guidance was consistent on this feature and the potential for isolated heavy rainfall. Therefore, as can be seen in Fig. 66, all three PFFs issued that day were associated with this feature; the experimental ERO is also included in Fig. 67 to illustrate the 24 h total rainfall from the event. During this event, all the PFF forecasts verified well subjectively with average scores of 8.56 (PFF1), 8.56 (PFF2), and 9.51 (PFF3).

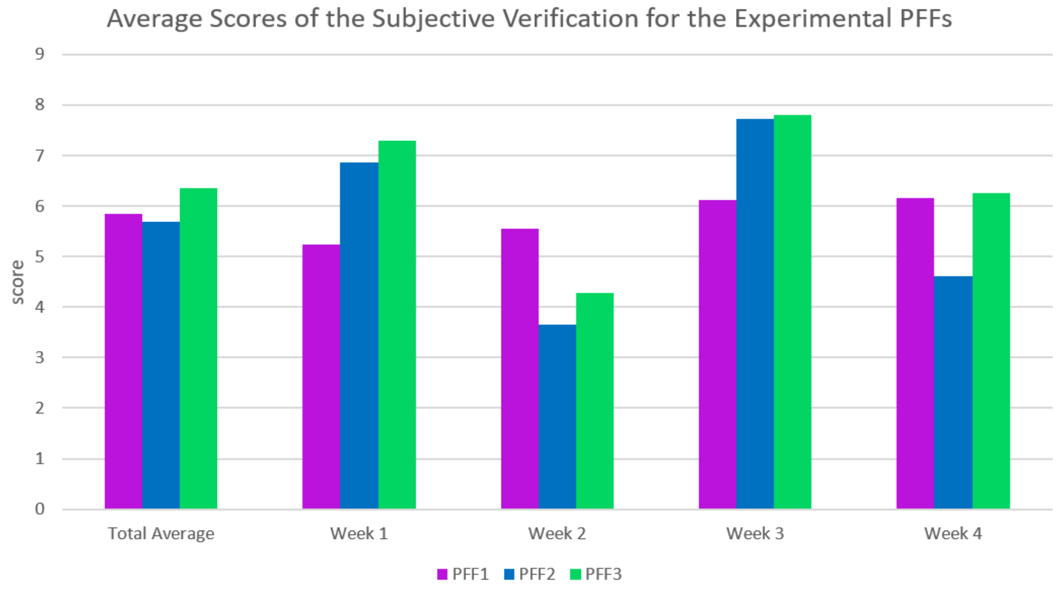


Figure 64: *The total average score and the weekly average scores for PPF1 (pink), PFF2 (blue), and PFF3 (green) during the course of the 2019 FFaIR experiment.*

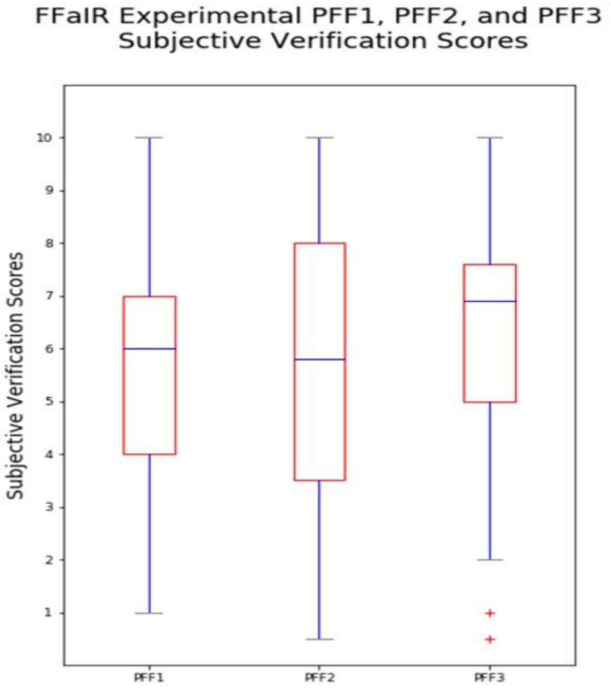


Figure 65: *Box and whisker plot of the subjective scores given during the duration of the 2019 FFaIR experiment for the PFF1, PFF2, and PFF3.*

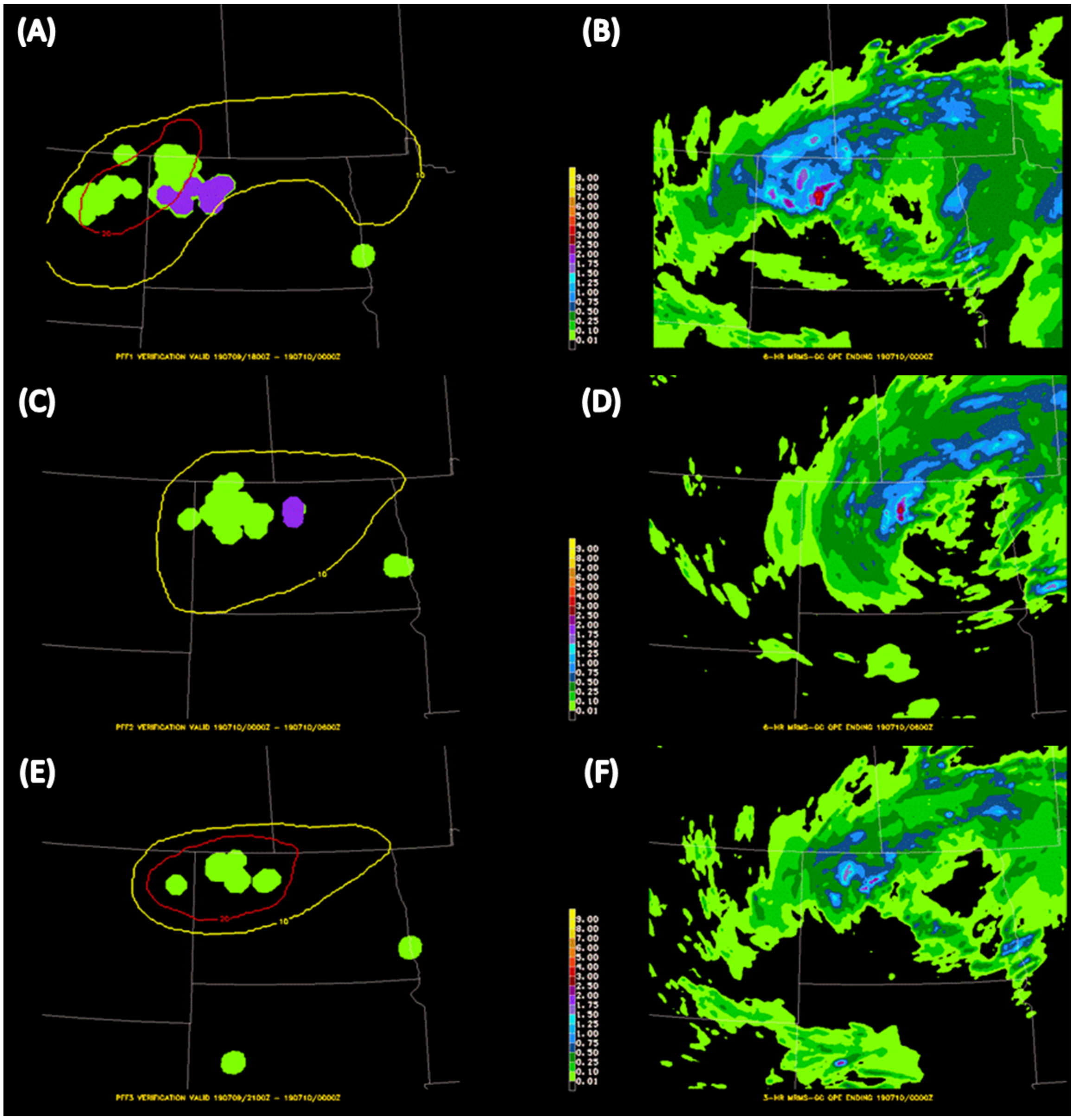

(A)
20
10
PFF1 VERIFICATION VALID 190709/1800Z – 190710/0000Z
(B)
9.00
8.00
7.00
6.00
5.00
4.00
3.00
2.50
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.10
0.01
6-HR MRMS-GC QPE ENDING 190710/0000Z
(C)
10
PFF2 VERIFICATION VALID 190710/0000Z – 190710/0600Z
(D)
9.00
8.00
7.00
6.00
5.00
4.00
3.00
2.50
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.10
0.01
6-HR MRMS-GC QPE ENDING 190710/0600Z
(E)
20
10
PFF3 VERIFICATION VALID 190709/2100Z – 190710/0000Z
(F)
9.00
8.00
7.00
6.00
5.00
4.00
3.00
2.50
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.10
0.01
3-HR MRMS-GC QPE ENDING 190710/0000Z

Figure 66: *Left column is all three of the PFFs issued on July 9 by the participants where (A) is PFF1, (C) is PFF2 and (E) is PFF3. The dots plotted on top of the PFFs represent where QPE > FFG (green), QPE > ARI (purple), and flooding/flash flooding storm reports (orange). Right column is the QPE valid (B) 1800 UTC 9 July 2019 to 0000 UTC 10 July 2019, (D) 0000 UTC to 0600 UTC 10 July 2019 , and (F) 2100 UTC 9 July 2019 to 0000 UTC 10 July 2019.*

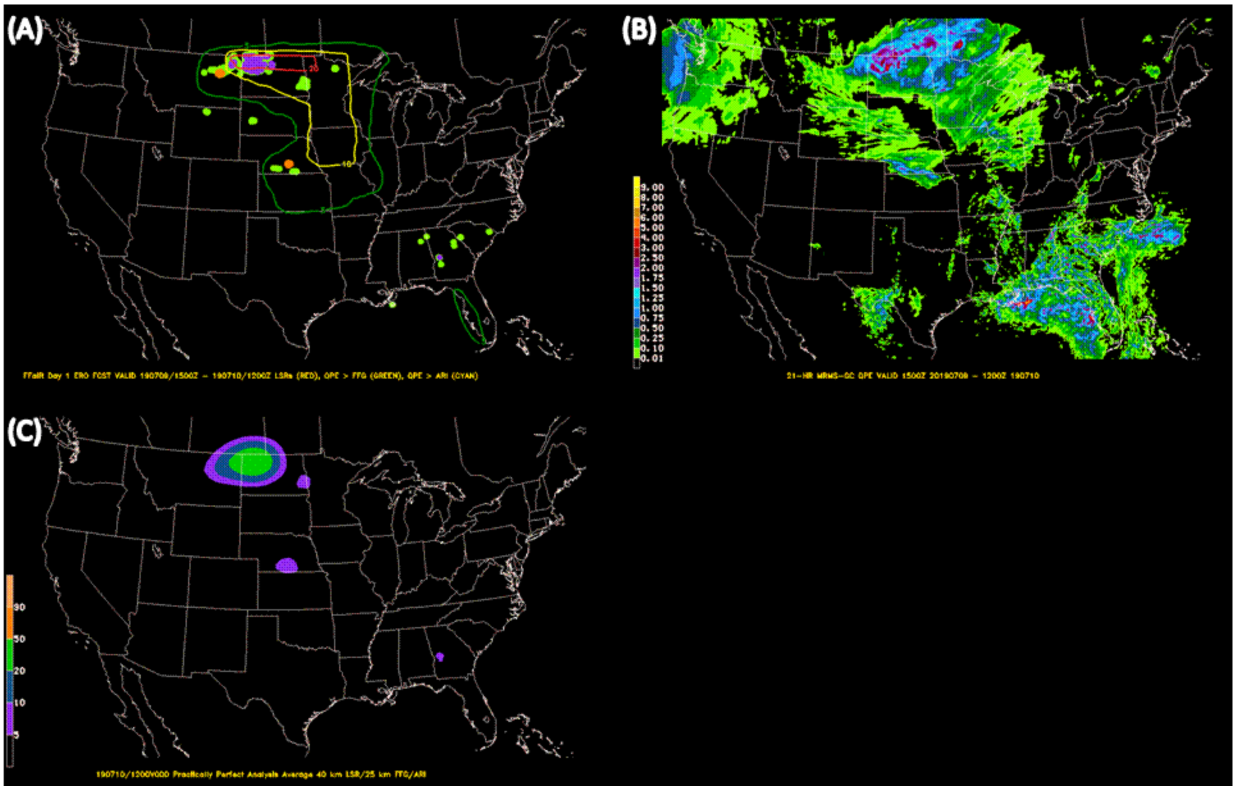

(A)
(B)
(C)
FFaIR Day 1 ERO FCST VALID 190709/1500Z – 190710/1200Z LSRs (RED), QPE > FFG (GREEN), QPE > ARI (CYAN)
9.00
8.00
7.00
6.00
5.00
4.00
3.00
2.50
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.10
0.01
21-HR MRMS-GC QPE VALID 1500Z 20190709 – 1200Z 190710
90
50
20
10
5
190710/1200V000 Practically Perfect Analysis Average 40 km LSR/25 km FFG/ARI

Figure 67: *(A) FFaIR experimental ERO, (B) 24 h MRMS-QPE and (C) Practically Perfect analysis all valid 1500 UTC 9 July to 1200 UTC 10 July. (A) the contour coloring follows Table 2. The dots plotted on top of the EROs same as Fig. 66. (C) Practically perfect analysis.*

Another factor driving the spread in the scores, especially in regards to the PFF2, was the forecast hour of the models/ensembles that was used as guidance for the product. During the course of the experiment, the majority of the guidance used were 00z runs. Therefore the participants were attempting to issue a forecast using CAM data at forecast hours 24 to 30; This issue was briefly mentioned above in Section 4 and, as was noted before, it is challenging for CAMs to produce a reliable forecast that far out in time. Because of this, the participants were often reluctant to issue anything more than slight risk for the PPF2, only issuing a moderate risk for 6 out of the 20 days of the experiment and a high risk only once. This also resulted in difficulty properly forecasting the timing of events. For example, the PFF2 issued on July 18, valid 0000 UTC to 0600 UTC July 19 can be seen in Fig. 68 attempted to capture the heavy rain event that occurred over extreme southeastern MN. CAM guidance was suggesting the event would occur during the PFF2 timeframe, thus the participants issued a PFF2 from southeastern MN, across central WI and into northeastern MI. However the forecast did not verify because the models were too progressive with the event. As can be seen in Fig. 68B the heaviest rainfall and flooding occurred over the following 6 h time period, from 06z to 12z. Participants the next day noted that they likely would not have missed the event if they had access to more recent model runs. This was a common complaint throughout the experiment.

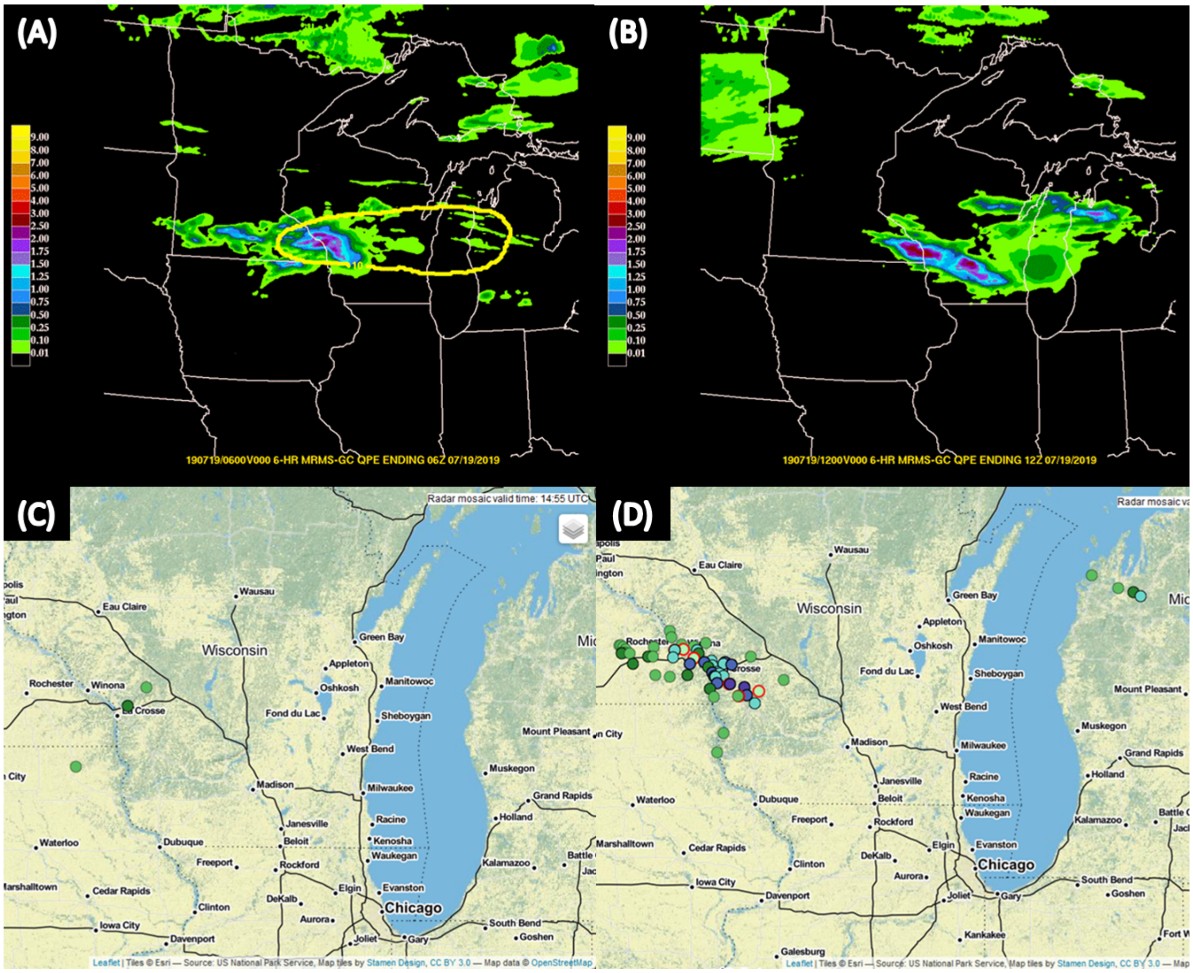


Figure 68: *(A)-(B) 6 h QPE with the PFF2 overlayed on (A). (C)-(D) Local storm flooding, flash flooding, and heavy rainfall reports. Valid 0000 UTC to 0600 UTC (left) and0600 UTC to 1200 UTC (right) 19 July 2019.*

Lastly, objective verification was only done for PFF1 and PFF2. Referring back to Fig. 60, it can clearly be seen that the PFF1 outperformed in the PFF2 in regards to both the Bulk Brier Score and the AuROC. This is not surprising due to the already discussed issues with using CAM data at forecast hours 24 to 30 to predict regional flooding events. However, the Brier Score time series plots of PFF1 and PFF2 (Fig. 60A) , show that there were some days where the PFF2 outperformed the PFF1. One example of this was the forecast from June 19, which can be seen in Fig. 69. As can be seen on this day, the participants missed all the flooding that occurred in Ohio for PFF1 but had high confidence of a heavy rainfall event in the evening/overnight hours in northeastern Texas for PFF2.

Comparing PFF Brier Score is not entirely valid because often the PFF1 and PFF2 were not located over the same region, had the same size, or were never valid at the same time. This can be seen in Fig. 69, where PFF1 spans eastern Great Lakes region while PFF2 focuses on a small region along the eastern border of TX and OK. Instead, a slightly better comparison might be to examine their calibration through fractional coverage of the verifying UFV, which can be seen in Fig. 70. This shows that both products are not well calibrated, but the fractional coverage does exhibit some resolution between each threshold (i.e. there is a difference in fractional coverage between slight and moderate). That being said, it is possible that the forecast themselves aren't necessarily the issue but rather the percentage thresholds selected for slight, moderate, and high. Using the same probabilistic thresholds for the chance of flooding in a six hour time frame compared to a 24 hour time frame might be skewing the results and the FFaIR team should look into these impacts to determine if the PFF products need to redefined.

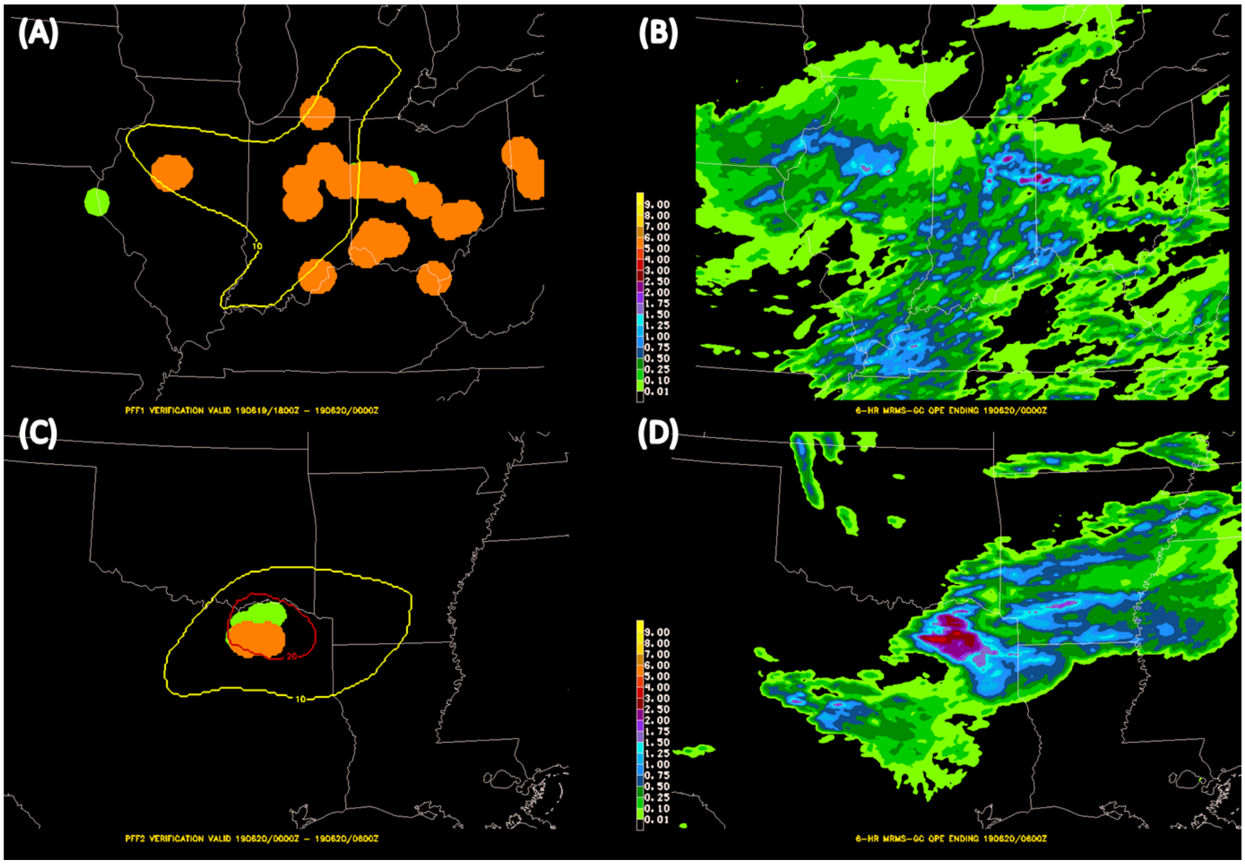

(A)
10
PFF1 VERIFICATION VALID 190619/1800Z – 190620/0000Z
(B)
9.00
8.00
7.00
6.00
5.00
4.00
3.00
2.50
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.10
0.01
6-HR MRMS-GC QPE ENDING 190620/0000Z
(C)
20
10
PFF2 VERIFICATION VALID 190620/0000Z – 190620/0600Z
(D)
9.00
8.00
7.00
6.00
5.00
4.00
3.00
2.50
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.10
0.01
6-HR MRMS-GC QPE ENDING 190620/0600Z

Figure 69: *Left column: the two PFFs issued on June 19 by the participants where (A) is PFF1 and (C) is PFF2. The dots plotted on top of the PFFs represent where QPE > FFG (green), QPE > ARI (purple), and flooding/flash flooding storm reports (orange). Right column: 6 h QPE. Top row: valid 1800 UTC 19 June 2019 to 0000 UTC 20 June 2019. Bottom row: valid 0000 UTC to 0600 UTC 20 July 2019.*

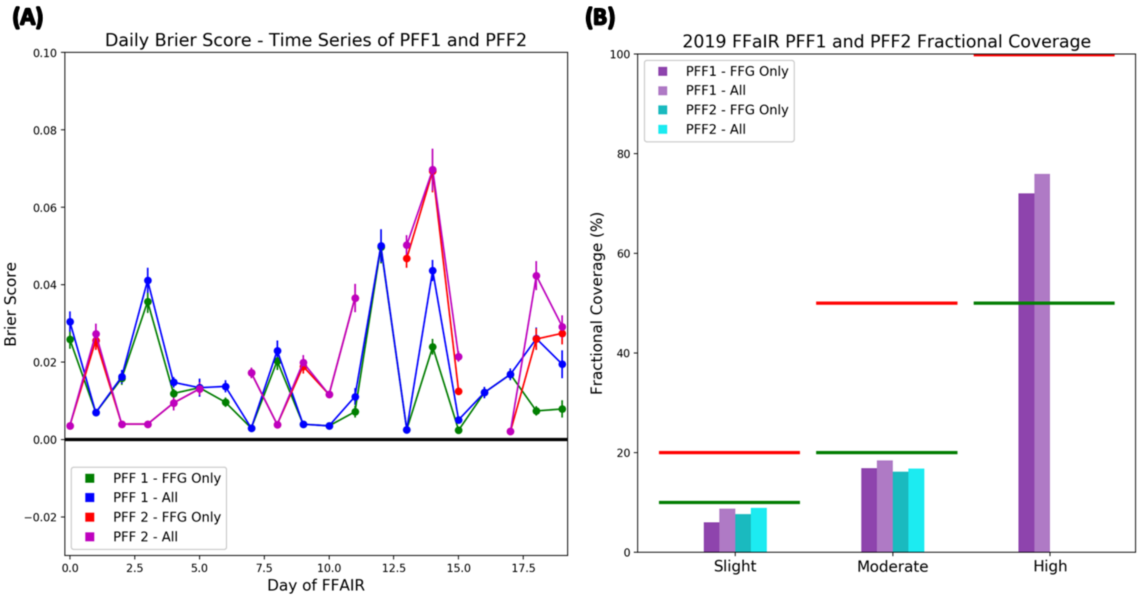


Figure 70: *(A) the daily Brier Score of the PFF 1 and PFF2 throughout the FFaIR experiment. Any missing points means that the participants decided not to issue a PFF. (B) the fractional coverage of the PFF1 (purple) and PFF2 (blue) for each probabilistic category. Green horizontal lines represent the lower defined bound for each threshold and red horizontal lines represent the higher defined bound. A good forecast should lie between the two bounds.*

CIRA CSU Total Precipitable Water and Advected Layer Precipitable Water Products

Throughout the experiment the participants were also able to utilize satellite and model derived precipitable water products provided by CSU-CIRA to aid in the forecasting process. Some of these products, like the Blended Total Precipitable Water (BTPW) and the Advected Layer Precipitable Water (ALPW) products are semi-operational while others, like the Merged Total Precipitable Water v1.0 (Merged TPW) and the various model derived layer precipitable water and difference products, are experimental; more information about these products can be found in Appendix C. Analysis of these products was purely subjective.

Merged vs Blended Total Precipitable Water

To determine if the new method of deriving TPW from satellites provides more detail and is more accurate than the current, semi-operational BTPW method, the Merged TPW product was compared against the BTPW product. This was done using the comparison question, where less than 5 means the BTPW is more useful, greater than 5 means the Merged TPW is more useful, and 5 means they provided the same information. An example of the two products can be seen in Fig. 71. Note that BTPW product is in inches while the Merged TPW product is in millimeters.

Figure 72 shows that overall the participants preferred the Merged TPW product over the BTPW product, with only 4% of the total scores being less than 5. It was also rare (only 14% of scores) that the participants felt that the two products were equally useful. The Merged TPW was generally preferred over the BTPW because the participants liked that it seemed to provide finer detail in the TPW gradient and had less discontinuities within the TPW field. They also noted that the Merged TPW seemed to be more realistic. However, despite the seemingly overwhelming preference to the new product, the participants noted that they had trouble with this question because the two different products were solely compared to one another but not against observations. They often stated that "they did not know which one was right" because there were no observations to verify the products with.

The discontinuities the participants noted were likely the result of the time differences from each pass of the polar orbiting microwave retrievals used to create the BTPW product. Such discontinuities are not present in the Merged TPW product because it uses a mix of GOES-16 TPW and the polar orbiting microwave retrievals which are advected using GFS winds. An example of this can be seen in Fig. 71, with the three regions highlighted are where noticeable discontinuities can be seen in the BTPW (Fig. 71A) but no such discontinuities were seen in the new Merged TPW product (Fig. 71B). The discontinuities are especially apparent along the northern CONUS into Canada, highlighted by the pink rectangle.

In addition to comparing the BTPW and Merged TPW products, the participants were also shown an alternate way of viewing the Merged TPW, which can be seen in Fig. 73. The difference between the two plots was in the color table that was used. Rather than a bright red, yellow, green, blue color scale that is seen in Fig. 71, the alternative product consisted of more colors, that were shaded darker with smoother transitions between values, allowing for more distinct identification of sharp moisture gradients. Although no official question was asked about the participants' thoughts on the alternative way to plot the Merged TPW product, the general feedback when using the product was that they preferred the alternative color scale.

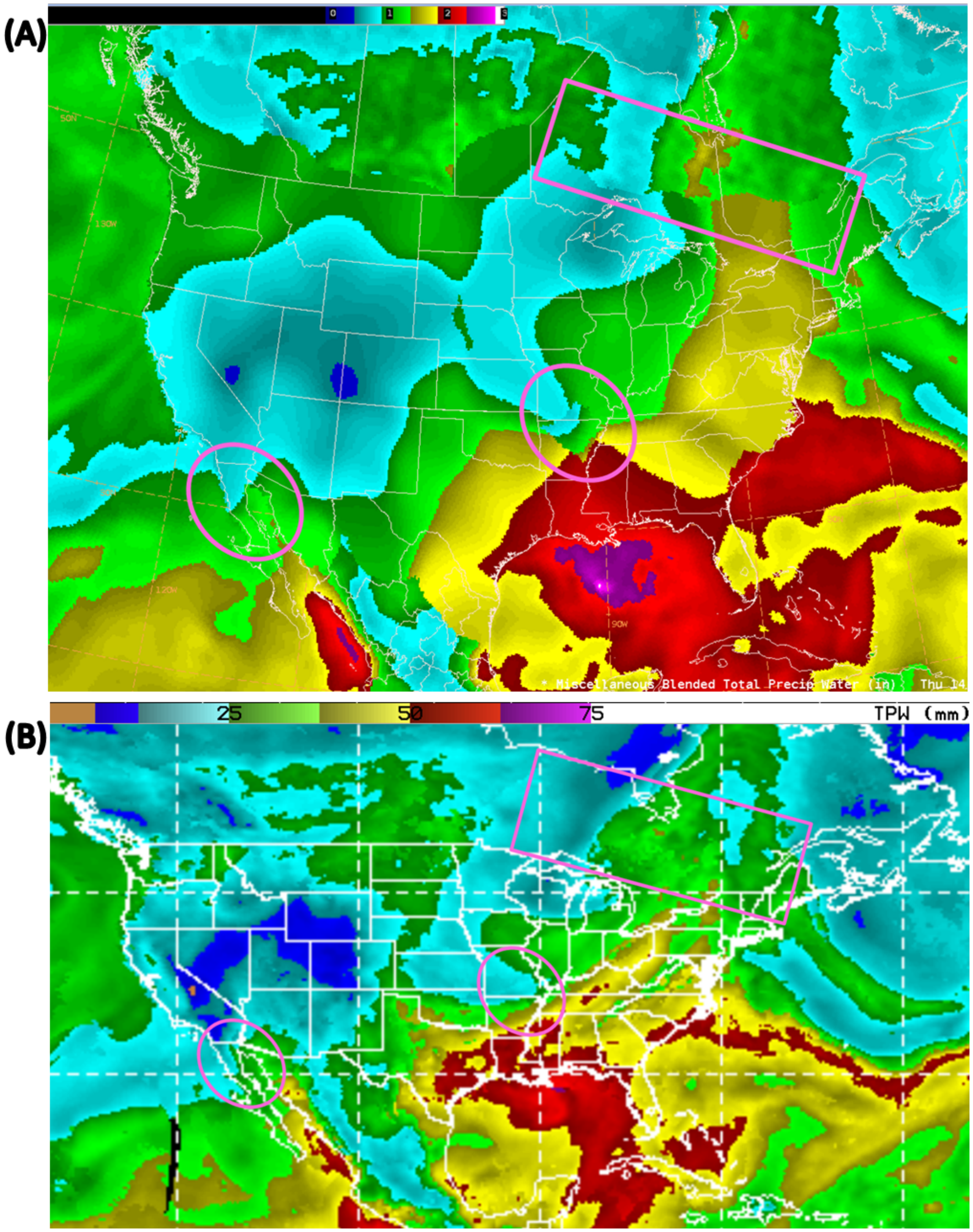


Figure 71: *An example showing the discontinuities noted by the participants between the (A) BTPW product and the (B) Merged TPW product. The pink circles and rectangle highlight areas were differences between the two products were seen. (A) BTPW units are in inches while (B) Merged TPW is in mm. The images are valid at 15 UTC on July 11, 2019.*

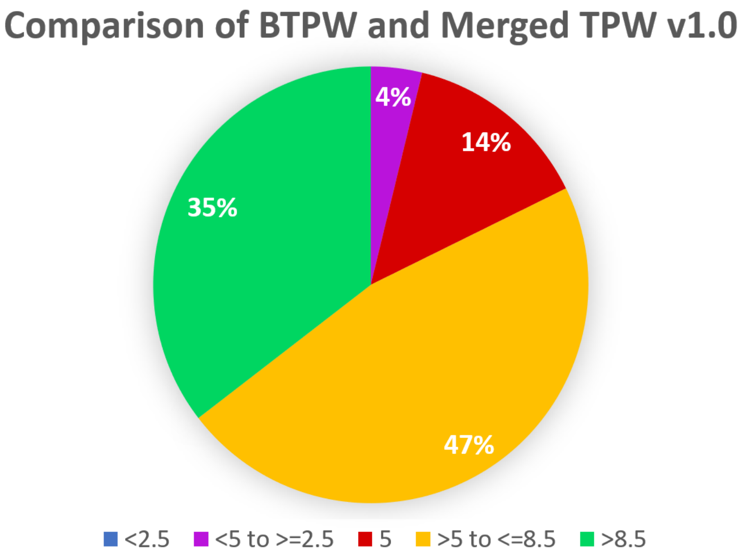


Figure 72: *Pie chart depicting the results from the verification question comparing the BTPW and Merged TPW v1.0 products, showing the percentage of times in which a score was recorded within various ranges. Anything below a 5 represents the participant finding more utility in the BTPW than the Merged TPW, while above the 5 means the Merged TPW product was more useful. A score of 5 suggests they provided the same amount of utility.*

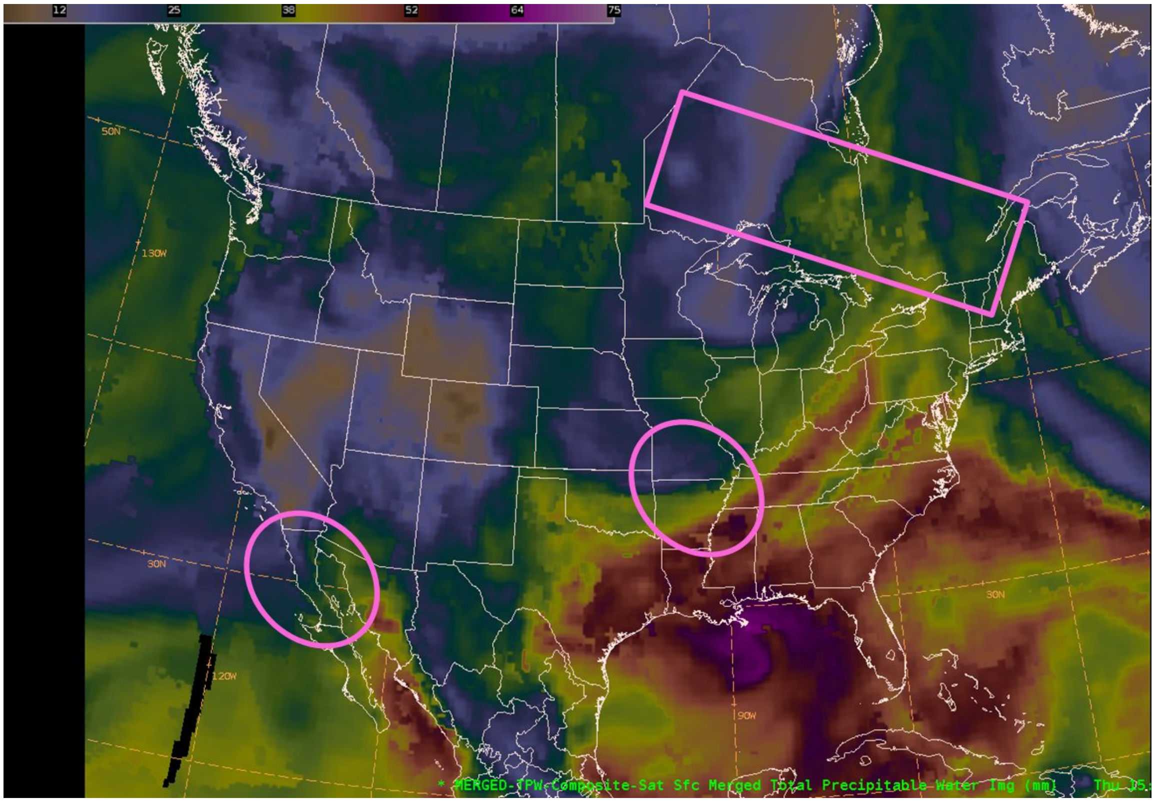

12
25
38
52
64
75
50N
130W
30N
120W
90W
30N
* MERGED-TPW-Composite-Sat Sfc Merged Total Precipitable Water Img (mm)
Thu 15:

Figure 73: *Example of the alternative color scale for the Merged TPW. Valid at the same time as Fig 71.*

HRRR Difference Fields for Advected Layer Precipitable Water

Participants were also asked their opinion of the HRRR[13] difference field product, which is designed to convey the differences between the 3 h HRRR forecast ALPW and observed ALPW. For this product, four layers are plotted: surface to 850 mb, 850 mb to 700 mb, 700 mb to 500 mb and 500 mb to 300 mb. During evaluation of the product, the participants were shown all 4 layers but asked to focus on the 700 mb to 500 mb and 500 mb to 300 mb layers. Additionally, the participants were not tasked with scoring the product but rather they were asked for general feedback about the utility of the product and whether or not they can infer information about the HRRR QPF from the product (i.e. if the HRRR ALPW was greater than observed in a location does that mean an over forecast in QPF in that region). The evaluation was done by showing the participants the 3 h QPE and the HRRR 3 h QPF valid at the same time (1500 UTC). This was done so the participants could identify where the HRRR forecast varied from observations; an example can be seen in Fig. 74. They were then shown the HRRR ALPW difference field (Fig. 75) valid at the same time and asked to discuss the previously stated topics.

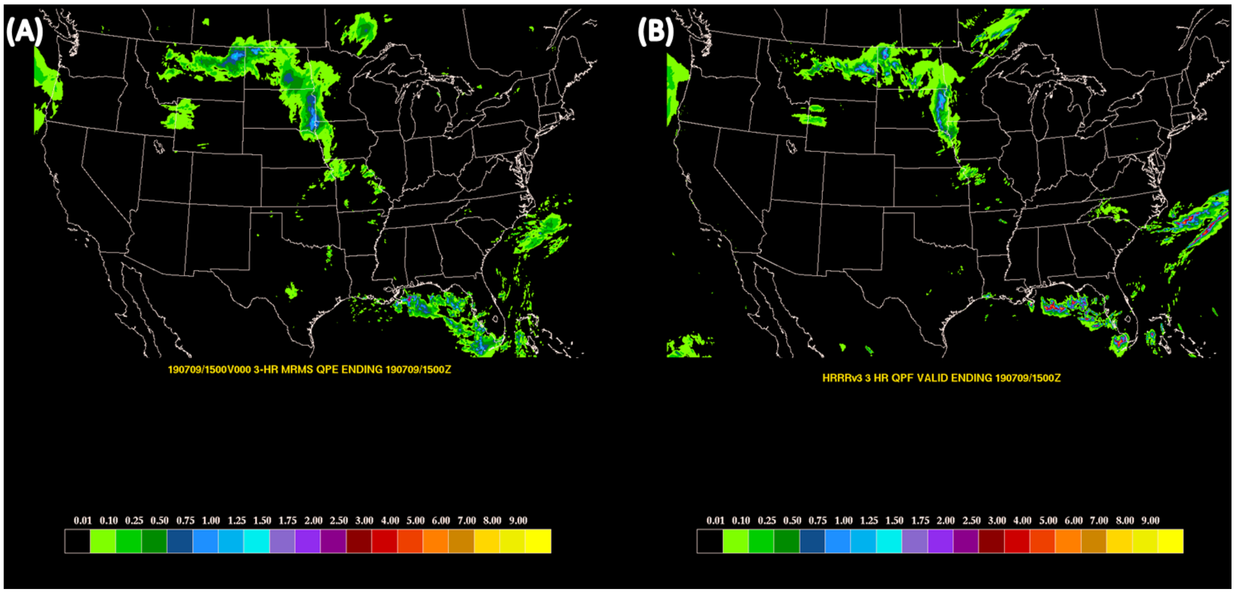


Figure 74: *(A) 6 hour MRMS-GC QPE and (B) HRRRv3 forecasted 6 hour QPF valid 1200 UTC to 1500 UTC 09 July 2019.*

[13] The operational HRRR (HRRRv3) is used for this product.

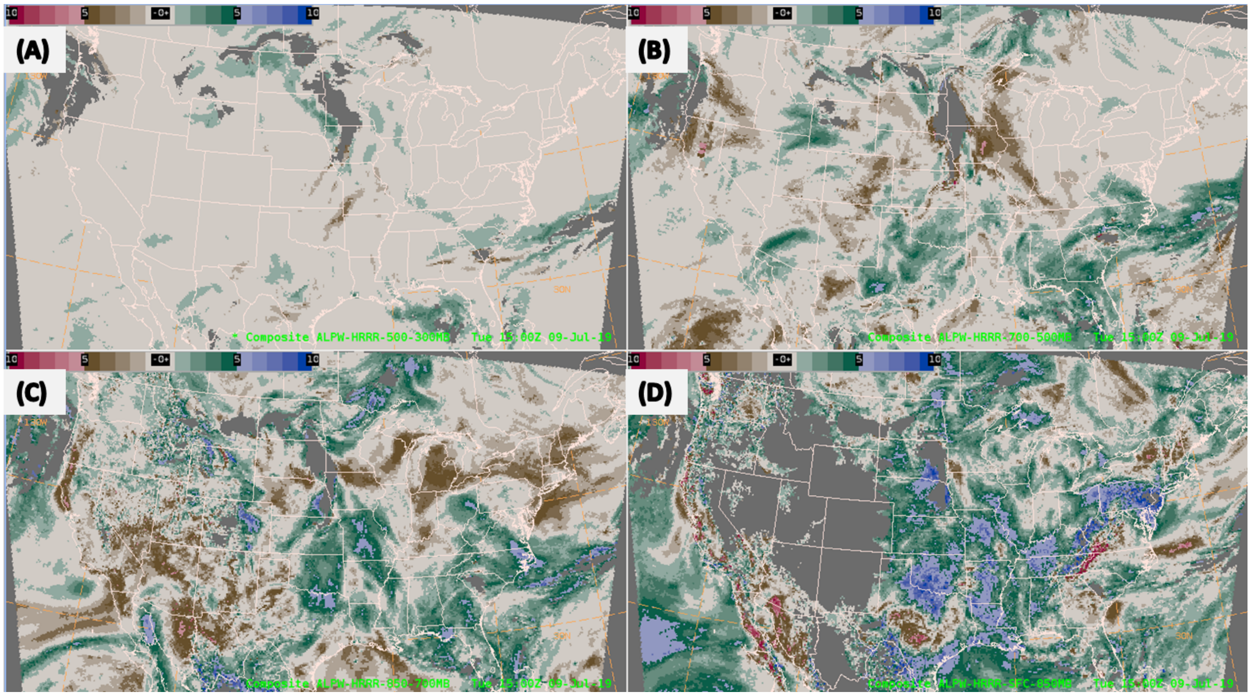

(A)
(B)
(C)
(D)
10 5 -0+ 5 10
* Composite ALPW-HRRR-500-300MB Tue 15:00Z 09-Jul-19
Composite ALPW-HRRR-700-500MB Tue 15:00Z 09-Jul-19
Composite ALPW-HRRR-850-700MB Tue 15:00Z 09-Jul-19
Composite ALPW-HRRR-SFC-850MB Tue 15:00Z 09-Jul-19
30N

Figure 75: *HRRRv3 3 h forecast ALPW minus the CIRA ALPW (mm) valid at 1500 UTC 09 July 2019. (A) 500-300 mb, (B) 700-500 mb, (C) 850-700 mb, and (D) surface to 850 mb.*

Feedback about the product overwhelmingly suggests that the participants did not find the product useful in the forecasting process for heavy rainfall. In general, participants stated that they could see the utility of the product for researchers and developers but that as forecasters they would not use the product. Some noted that it would take extensive training to learn to use the product effectively and felt they could get an idea if the model was running wet or dry by doing a visual comparison of the QPF to radar or MRMS. They also stated that models usually have a wet bias in precipitable water, so they were unsure how seeing this bias depicted would help in the forecast process or what to infer about how it will impact the QPF. Furthermore, they noted that there are other factors that drive QPF, and not just ALPW, so looking at just one product does not provide enough information to make any assumptions about the model over/under forecasting precipitation. Participants however did state that the product helps them quickly identify if the model is too fast/slow with features like a front or a MCS. Though again they stated they could get the same information by comparing model data with radar data.

Elaborating on the difference fields utility in the research realm, many felt that the product would help the developers better understand the model climatology and tendencies. This information then could be used to help improve upon the mid- and upper layers of the atmosphere simulated within the model, which is generally hard to verify. There was also discussion of using the product for case studies, though exactly how the product would be used in the study is beyond the scope of FFaIR.

Final Recommendations for CIRA CSU Satellite Products

The new Merged TPW v1.0 product showed great promise, was well received by the participants and was nearly always preferred over the operational BTPW. However, before it can be made operational, additional verification must be done to ensure the satellite derived product is correctly representing the atmosphere, which could be done using radiosonde data. As for the HRRR APLW difference fields, it is recommended that the product continue to be developed but with the goal of use in research rather than operations.

6. Summary

The seventh annual FFaIR experiment heavily focused on CAM models and ensembles with the newly implemented FV3 core. This included the new version of the HREF (version 3) which switched out the poorest performing NMMB member for the FV3-SAR produced at EMC. The newest versions of the HRRR and HRRRE, which do not have the FV3 core, were also evaluated. In addition to model and ensemble data, the 2019 FFaIR experiment also examined two CSU Machine Learning First Guess Fields for the Day 1 ERO product, one trained on the NSSL-WRF model and the other on the GEFS Reforecast. The experiment also assessed the new TPW product from CSU-CIRA as well as their HRRR ALPW difference product.

Although most of the guidance and tools analyzed during FFaIR received positive feedback from the participants and performed relatively well when evaluated using objective metrics, nearly every product/tool is being recommended for further testing. As can be seen in Table 8, only the **HRRRv4** is being recommended for transitions to operation, albeit conditionally. Before it can be implemented, retrospective runs and testing of the model must be done to ensure the bug fix in mid-July did not negatively impact model performance. Various items that need to be addressed before any of the other products and tools can be transitioned to operations are highlighted in the following bullet points.

- Both the **FV3-Nest** and the **FV3-SAR** from EMC performed well based on the subjective scores given by the participants. The FV3-Nest narrowly outperformed the FV3-SAR in both the 24 h and 6 h QPF subjective verification. Despite the high scores from the subjective portion of the experiment, when evaluated using objective metrics, the FV3-Nest and FV3-SAR performance was not as impressive. Most notable was the high wet bias seen in both models, which at the 1 inch threshold was higher than the wet bias seen in the NAM-Nest. Therefore, until the wet bias is addressed the models should not be transitioned into operations.

- The **FV3-SAR-GSD** from ESRL/GSD differed from the EMC FV3-SAR in the physics suite that was used; see Appendix C. Although the model was only available at a limited capacity during the FFaIR experiment, objective and subjective results indicate that FV3-SAR-GSD underperforms in comparison to the FV3-SAR from EMC.

- The **HREFv3** was the best performing ensemble both objectively and subjectively and would have been recommended for transition into operations. However, the models that the HREF are comprised of are not all frozen (e.g. the model dynamics are still being tested and altered). The still fluid nature of the models that form the HREFv3 could impact model performance in the future. Therefore, until the models are frozen, the HREFv3 must continue to be evaluated.

- Per recommendation from the 2018 FFaIR Experiment, partners providing ensemble data to be evaluated in 2019 FFaIR were asked to produce a **LPMM** product. All three of the ensembles assessed, the **HREFv3**, **HRRRE**, and the **SSEF,** supplied this product. Like last year, overall the product was preferred by the participants over the PMM. However objective results suggest that LPMM performance is somewhat dependent on how the LPMM is calculated. Analysis of ensemble biases showed that the LPMM products from the SSEF and HREFv3 had a lower bias their respective PMM products while the HRRRE LPMM had a higher bias compared to its PMM. Therefore the FFaIR team suggests that the method used by the SSEF and HREF be the method used for LPMM calculations and be the product be produced by all operational ensembles.

- The **CSU-MLP First Guess Day 1 ERO** products were well received by the participants and participants felt that the guidance provided a great “starting spot” for creating the experimental FFaIR ERO. However, there are a few regions across the CONUS that the products do not appear to be well-calibrated for, such as the Northern Rockies and Northern Plains in the NSSL ERO product. Further refinement of the how the flooding risk is determined should be done in these regions. These refinement will likely help with the calibration of the marginal risk, which was generally too large spatially for both the NSSL and GEFS products.

- The **CSU-CIRA Merged TPW v1.0** appears to have improved upon the operational BTPW but needs to be further verified against observations to determine if the new method of deriving the TPW from multiple satellites and combining them is accurate.

- The utility of **CSU-CIRA HRRR ALPW Difference** product was not fully understood by the FFaIR participants and they felt the product would not aid them in their forecasting process. They did however feel that the product would be highly beneficial to model developers.

Table 8: *Research to Operations Transition Metrics for the 2019 FFaIR Experiment.*

Major Models/ Products Evaluated	Recommended for transition to operations	Recommended for further development and testing	Rejected for further testing	Funding Source
FV3-Nest		x		EMC
FV3-SAR		x		EMC
FV3-GSD-SAR		x		ESRL/GSD
HRRRv4	x*			ESRL/GSD
HREFv3		x		EMC
HRRRE		x		ESRL/GSD (OWAQ)
SSEF		x		OU/CAPS (OWAQ)
CSU-Machine Learned Day 1 ERO First-Guess Fields		x		CSU (JTTI)
CIRA-CSU Merged TPW v1.0		x		CSU/CIRA (OWAQ)
CIRA-CSU HRRR ALPW Difference Fields		x		CSU/CIRA (OWAQ)
Total	**1**	**9**	**0**	

Acknowledgments

An enormous thanks to all the staff of HMT and WPC, who worked countless hours to get the experiment ready, especially the AWIPS team of **Chris Bailey** and **Diana Stovern**. Also thank you to our all our partners that provided the products and tools evaluated in FFaIR this year. Thank you to all those who participated in the experiment this year and a special thanks to the WPC forecasters, **Patrick Burke, Josh Weiss, Zack Taylor, Lara Pagano,** and **Frank Pereira** for all the hard work and long hours you put into the experiment. Thank you **Michael Erickson** for your help with the verification and comparison metrics for the operational and experiment EROs and PFFs as well as editing the Final Report. A special thanks to both the WPC Science

Operations Officer, **Mark Klein**, and the DTB Branch Chief, **James Nelson**, for technical and management support. Finally, thank you to the WPC Student Interns **Alyssa Cannistraci, Amanda Wagner,** and **Emily Blumenauer** for their hard work taking notes during the subjective verification section of the experiment.

References

Alexander, C., et al., 2017: WRF-ARW Research to Operations Update: The Rapid Refresh (RAP) version 4, High-Resolution Rapid Refresh (HRRR) version 3 and Convection-Allowing Ensemble Prediction. *18th Annual WRF User's Workshop.* https://ruc.noaa.gov/ruc/ppt_pres/Alexander_WRFworkshop_2017_Final.pdf

Clark, A.J., 2017: Generation of Ensemble Mean Precipitation Forecasts from Convention-Allowing Ensembles. *Wea. Forecasting*, **32**, 1569-1582.

Dowell et al., 2018a: Development and Testing of a High-Resolution Rapid Refresh Ensemble (HRRRE). *The 8th EnFK Data Assimilation Workshop*. https://web.meteo.mcgill.ca/enkf/abstracts_html/c1733a901c.php

Dowell et al., 2018b: HRRR Ensemble (HRRRE) Guidance: 2018 HWT Spring Experiment. https://rapidrefresh.noaa.gov/internal/pdfs/2018_Spring_Experiment_HRRRE_Documentation.pdf

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.

Gourley, J. J., et al., 2017: The FLASH Project: Improving the Tools for Flash Flood Monitoring and Prediction Across the United States. *BAMS*, 361-372.

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, 2016: Extreme Precipitation in Models: An Evaluation. *Wea. Forecasting*, **31**, 1853-1879, https://doi.org/10.1175/WAF-D-16-0093.1.

Herman, G. R., and R. S. Schumacher, 2018: Money Doesn't Grow on Trees, but Forecasts Do: Forecasting Extreme Precipitation with Random Forests. *Mon. Wea. Rev.,* **146**, 1571-1600, https://doi.org/10.1175/MWR-D-17-0250.1.

Jones, T. A., and C. Nixon, 2017: Short-term forecasts of left-moving supercells from an experimental Warn-on-Forecast system. *J. Operational Meteor.*, **5 (13),** 161-170, https://doi.org/10.15191/nwajom.2017.0513.

LZK, NWS Little Rock, AR - Heavy Rain/Flooding with Barry on July 14-17, 2019. Accessed 15 October 2019, https://www.weather.gov/lzk/rain0719.htm

Manikin, G., 2018: Transitioning Model Upgrades into Operations. *13 November 2018 NCEP/EMC Update to WPC*. https://docs.google.com/presentation/d/1engL6G2aPklksDMwtG5a7gAs1JyCr0T6QT3Lqu1gbd8/edit?usp=sharing

March 2019 update: RAP/HRRR National Center Coordination Meeting. https://drive.google.com/drive/folders/0B5TTqI5ra0ZDZ24wbjhfNFdWNjg

McQueen et al., 2004: Overview of the NOAA/NWS/NCEP Short Range Ensemble Forecast (SREF) System. https://www.emc.ncep.noaa.gov/mmb/SREF-Docs/SREF-TPB2004.pdf

Nakanishi, M., Niino, H., 2004. An improved Mellor-Yamada level-3 model with condensation physics: Its design and verification. *Bound. Layer Meteorol.* **112**, 1–31.

NOAA National Centers for Environmental Information, State of the Climate: National Climate Report for Annual 2018, published online January 2019, retrieved on April 12, 2019 from https://www.ncdc.noaa.gov/sotc/national/201813.

Snook, N., F. Kong, K. Brewster, M. Xue, B. Albright, S. Perfater, K. Thomas, and T. Supinie, 2019: Evaluation of Convection-Permitting Precipitation Forecast Products using WRF, NMMB, and FV3 for the 2016-2017 NOAA Hydrometeorology Testbed Flash Flood and Intense Rainfall Experiments. *Wea. and Forecasting*, conditionally accepted.

UFS STI: 2018 Strategic Implementation Plan for Evolution of NGGPS to a National Unified Modeling System. Accessed 9 April 2019, https://www.weather.gov/media/sti/nggps/UFS%20SIP%20FY19-21_20181129.pdf

Appendix A

FFaIR 2019 Participants

Table A.1: *List of the participants for each week of the 2019 FFaIR Experiment. Note the experiment did not run during the week of the Fourth of July.*

Week	WPC Forecaster	WFO/RFC/Other	Research/Academia	EMC
Week 1: June 17-21	Patrick Burke	Brian Lasorsa - WFO LWX Lisa Darby - ESRL/PSD Chris Schaffer - SWRFC Kate Abshire - NWC Alex Korner - AWC/CIRA	Sheldon Kusselson - CSU/CIRA Glen Romine - NCAR Jason English-GSD Taylor Prislovisky - MDL	Logan Dawson* Alicia Bentley* Tomer Burg***
Week 2: June 24-28	Josh Weiss Zack Taylor-June 28	Micheal Brown - WFO PBZ James Telken - WFO ABR Steven VanHorn – WFO OTX	Erik Nielsen - CSU Kent Knopfmeier - NSSL/WoFS Dave Rudack** - MDL	Jacob Carley* Geoff Manikin*
Week 3: July 8-12	Lara Pagano	Caleb Steele WFO VEF Nick Fillo - WFO RNK Erick Boehmler - NERFC	Aaron Hill - CSU Russ Schumacher** - CSU Kim Hoogewind - NSSL/WoFS David Dowell - GSD Erik Engle** - MDL Eric James - CSU/GSD	Ben Blake* Tracey Dorian*
Week 4: July 15-19	Frank Pereira	Charles "Danny" Gant - WFO MRX Andrea Ray - ESRL/PSD Stephen Hrebenach - WFO ILN Jane Marie Wix - WFO JKL Andrew Peck*** - CPC	Keith Brewster - OU/CAPS Jeffrey Duda - GSD Jeff Craven** - MDL	Matt Pyle

*split the week

**only there part of the week

***visiting student/intern

Go-To Meeting Forecasters

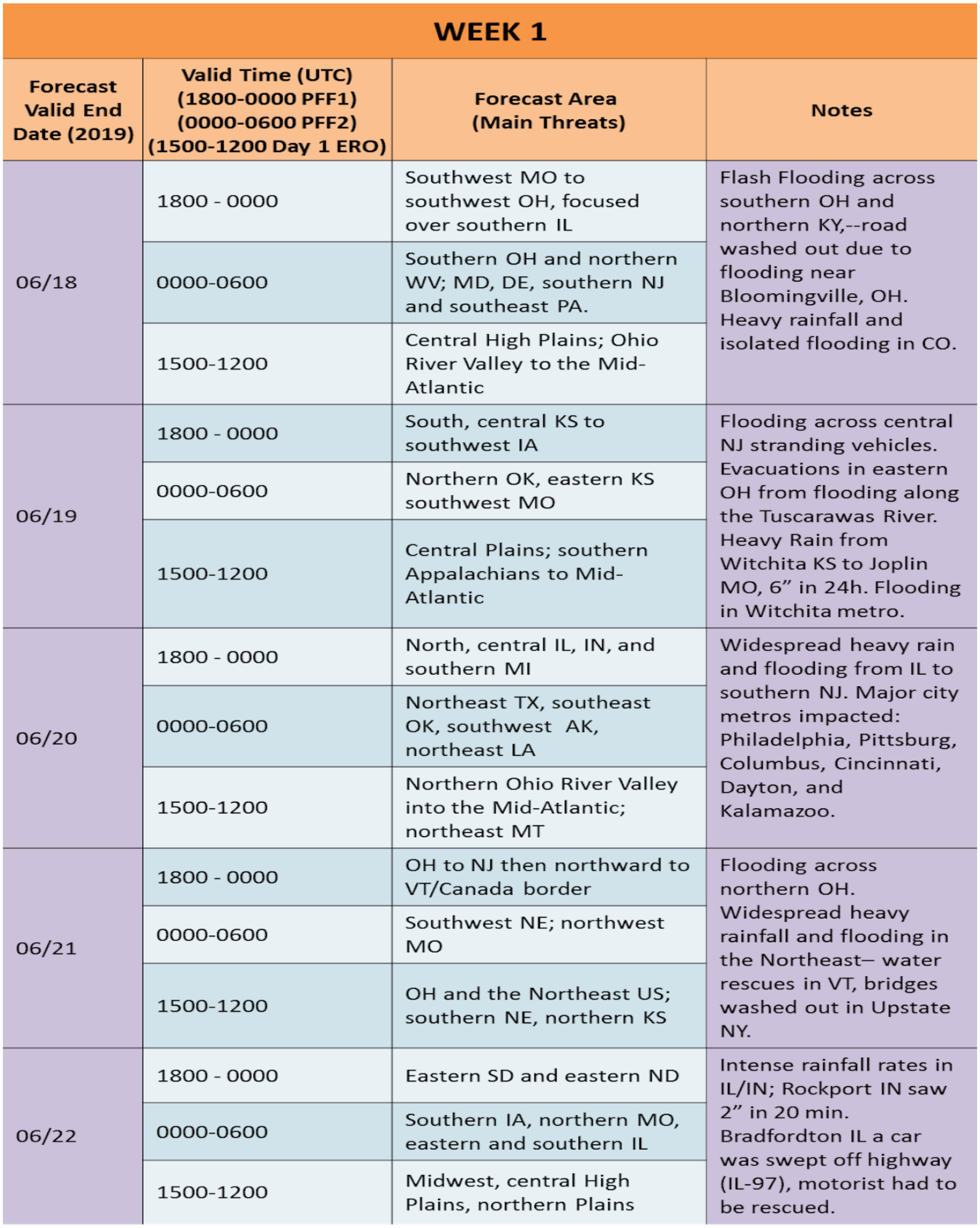
Table A.2: *List of the participants that volunteered to do the daily Go-To meeting highlighting the forecast challenges of the day, the experimental guidance and tools being evaluated in FFaIR, and the experimental FFaIR ERO and PFF1.*

Day	Monday	Tuesday	Wednesday	Thursday	Friday
Week 1: June 17-21	Patrick Burke	Brian Lasorsa	Brian Lasorsa	Chris Schaffer Kate Abshire	Brian Lasorsa Sheldon Kusselson
Week 2: June 24-28	Josh Weiss	James Telken	Steven VanHorn	Erik Nielsen	Michael Brown
Week 3: July 8-12	Lara Pagano	Nick Fillo	Nick Fillo Caleb Steele	Eric James	Erick Boehmler
Week 4: July 15-19	Frank Pereira	Danny Gant	Jane Marie Wix	Stephen Hrebenach	Keith Brewster

Appendix B

FFaIR 2019 Forecast Areas and Significant Events

Table B.1: *Forecast areas and events for the first week of FFaIR taking place June 17-21, 2019.*



WEEK 1			
Forecast Valid End Date (2019)	**Valid Time (UTC) (1800-0000 PFF1) (0000-0600 PFF2) (1500-1200 Day 1 ERO)**	**Forecast Area (Main Threats)**	**Notes**
06/18	1800 - 0000	Southwest MO to southwest OH, focused over southern IL	Flash Flooding across southern OH and northern KY,--road washed out due to flooding near Bloomingville, OH. Heavy rainfall and isolated flooding in CO.
	0000-0600	Southern OH and northern WV; MD, DE, southern NJ and southeast PA.	
	1500-1200	Central High Plains; Ohio River Valley to the Mid-Atlantic	
06/19	1800 - 0000	South, central KS to southwest IA	Flooding across central NJ stranding vehicles. Evacuations in eastern OH from flooding along the Tuscarawas River. Heavy Rain from Witchita KS to Joplin MO, 6" in 24h. Flooding in Witchita metro.
	0000-0600	Northern OK, eastern KS southwest MO	
	1500-1200	Central Plains; southern Appalachians to Mid-Atlantic	
06/20	1800 - 0000	North, central IL, IN, and southern MI	Widespread heavy rain and flooding from IL to southern NJ. Major city metros impacted: Philadelphia, Pittsburg, Columbus, Cincinnati, Dayton, and Kalamazoo.
	0000-0600	Northeast TX, southeast OK, southwest AK, northeast LA	
	1500-1200	Northern Ohio River Valley into the Mid-Atlantic; northeast MT	
06/21	1800 - 0000	OH to NJ then northward to VT/Canada border	Flooding across northern OH. Widespread heavy rainfall and flooding in the Northeast– water rescues in VT, bridges washed out in Upstate NY.
	0000-0600	Southwest NE; northwest MO	
	1500-1200	OH and the Northeast US; southern NE, northern KS	
06/22	1800 - 0000	Eastern SD and eastern ND	Intense rainfall rates in IL/IN; Rockport IN saw 2" in 20 min. Bradfordton IL a car was swept off highway (IL-97), motorist had to be rescued.
	0000-0600	Southern IA, northern MO, eastern and southern IL	
	1500-1200	Midwest, central High Plains, northern Plains	

Table B.2: *Forecast areas and events for the second week of FFaIR taking place June 24-28, 2019.*

WEEK 2			
Forecast Valid End Date (2019)	**Valid Time (UTC) (1800-0000 PFF1) (0000-0600 PFF2) (1500-1200 Day 1 ERO)**	**Forecast Area (Main Threats)**	**Notes**
06/25	1800 - 0000	Eastern TN, eastern KY, southern/eastern OH, northern/western WV, western PA	Raymondville/ Harlingen TX Flooding Event – over a foot of rain in 6 h with ~15" max near Santa Rosa. Hundreds of roads closed. 1188 homes destroyed or had major damage.
	0000-0600	WV to eastern NY	
	1500-1200	Ohio River Valley northward into western NY	
06/26	1800 - 0000	Southeast TX, southwest LA	Isolated flooding in IA/NE/KS/TX. In Guide Rock, NE overflow from a lake damage a short section of railroad track.
	0000-0600	N/A	
	1500-1200	Coast of TX; central Plains	
06/27	1800 - 0000	MO/AK border to southern IL	The majority of the flooding occurred outside of the FFaIR ERO marginal risk across southern MO to central KY. Swift water rescue Gallatin, TN.
	0000-0600	Central MT to central SD	
	1500-1200	Midwest to northern Plains	
06/28	1800 - 0000	IA into southwest WI	Southern MN – Rochester Intl. Airport closed for several hours because of water on run ways. 5" in 6 h. Flash flooding on Zumbro River swept 40-60 cattle off one farm.
	0000-0600	Southern MN, northern IA, and western WI	
	1500-1200	Northern Midwest; northeast MT, Mid-Mississippi Valley	
06/29	1800 - 0000	Southeast AK, eastern MS, northern LA	Isolated flooding from eastern OH to Philadelphia.
	0000-0600	North, northeastern MT	
	1500-1200	Eastern AK; northeast MT; Midwest to western PA	

Table B.3: *Forecast areas and events for the third week of FFaIR taking place July 08-12, 2019.*

WEEK 3			
Forecast Valid End Date (2019)	**Valid Time (UTC) (1800-0000 PFF1) (0000-0600 PFF2) (1500-1200 Day 1 ERO)**	**Forecast Area (Main Threats)**	**Notes**
07/09	1800 - 0000	Coast of NC and southern VA	Loomis NE ~9"of rain in 24h. Kearney NE ~300 people had to be evacuated. Yates ND portion of a highway washed away leading to 2 fatalities. Flash Flood Emergency in Washington DC Metro.
	0000-0600	Eastern MT across ND into northwest MN; southern NE and northwest KS	
	1500-1200	northern Plains to central TX, with large focus on ND	
07/10	1800 - 0000	Eastern MT and northern ND	Heavy Rainfall across northwestern ND – 4.78" in 24h.
	0000-0600	ND	
	1500-1200	Along Canadian border from eastern MT to MN; northern Midwest	
07/11	1800 - 0000	southeast LA	Flash Flood Emergency in New Orleans. Isolated flooding from southern MO to central IN. ~2.5" in under 1.5 h Kempton, IN. Water rescues Muncie, IN.
	0000-0600	N/A	
	1500-1200	Mid- to southern-Mississippi Valley	
07/12	1800 - 0000	TN/NC border, eastern KY north/northeastward to southern NY east to the coast	07/11-morning hours Pittsburg metro widespread flash flooding. Evening into 07/12 morning-flooding from eastern PA to NYC- Lazy-K Campground tents and trailers swept away, campground evacuated. 2 fatalities in Douglass Township, PA from floodwaters.
	0000-0600	Atlantic coast from central NC to New England; VT/NH	
	1500-1200	Mid-Atlantic to the Northeast/ Gulf Coast from LA to Tampa Bay	
07/13	1800 - 0000	central TN to southwest NC; central coast of NC	Isolated events.
	0000-0600	Gulf Coast from eastern LA to western Panhandle of FL	
	1500-1200	LA, southern MS, southern AL; TN to eastern NC	

Table B.4: *Forecast areas and events for the fourth week of FFaIR taking place July 15-19, 2019.*

WEEK 4			
Forecast Valid End Date (2019)	**Valid Time (UTC) (1800-0000 PFF1) (0000-0600 PFF2) (1500-1200 Day 1 ERO)**	**Forecast Area (Main Threats)**	**Notes**
07/16	1800 - 0000	Southeast TX to southwest MS; Mid-Mississippi Valley	TS Barry – heavy rainfall and extreme flooding in southwestern AR. Rain rates exceeded 3” an hour at some points. Murfreesboro, AR 13.5” in 24h. **New state record for** rainfall from tropical cyclone: Dierks, AR -16.59”, with 16.17” of it in 24 h. (lzk)
	0000-0600	northern WI	
	1500-1200	Mid- to Southern-Mississippi Valley; western Great Lakes region	
07/17	1800 - 0000	Southern AK, along the Mississippi River to central TN and western KY.	TS Barry- AR to southeast MI. In Ann Arbor MI 4” in 1.5 h. Flooding continuing in southwest AR. Portions of US-61 and US-51 closed due to flooding.
	0000-0600	N/A	
	1500-1200	AK to southern MI into southwest NY; Dakotas into western IA	
07/18	1800 - 0000	northern PA , southern NY and MA	Heavy Rainfall west of Saint Louis, MO – storm total 5”. Flooding across eastern PA and NYC – creeks overflowing banks, cars stranded by high water, road/highway closures.
	0000-0600	southern MN, northwest WI, northern IA	
	1500-1200	northern Midwest; Mid-Mississippi and Tennessee Valleys; northern PA , southern NY into CT.	
07/19	1800 - 0000	NJ and New York City	NYC and NJ – car rescues due to flooding. Southeastern MN area– 4.5-5” of rainfall overnight leading to rock and mudslides across the area.
	0000-0600	southeast MN, central WI, northwest MI.	
	1500-1200	MN to MI; coastline from eastern MD to central MA	
07/20	1800 - 0000	AL, GA, and northern FL	Southeast MI – heavy rain ~4” in 24h. Northwestern PA – widespread flooding across Oil City, PA. Cars stranded in Orangeville, OH.
	0000-0600	north, northeast WI; southeast coastline of Lake Erie	
	1500-1200	WI; southeast MI; Cleveland/Pittsburg region; southern AL and southern GA	

Appendix C

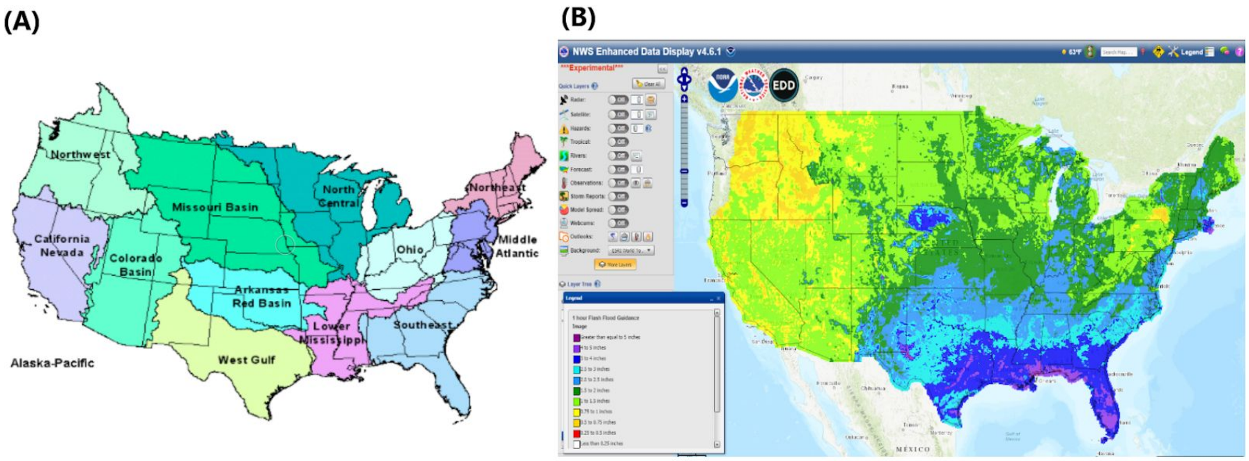
Operational and Experimental Guidance used in FFaIR 2019

C.1 Operational and Experimental Flood Guidance

The 2019 FFaIR Experiment made use of products that are routinely used by the WPC in operations, such as: the River Forecast Center's (RFC) Flash Flood Guidance (FFG) and precipitation Average Recurrence Intervals (ARI). An in-depth discussion of the guidance that was used can be found in the following subsections.

Flash Flood Guidance (FFG)

FFG is a guidance product that is issued by each individual RFC (see Fig. C.1a for the RFC domains) that estimates how much rain over various time-scales would need to fall in order for small streams to overflow their banks. For example, if the 3 h FFG is 2 inches then it is expected that flooding on small streams will begin to occur if 3 h precipitation totals exceed 2 inches. FFG products are generally produced every 6 hours for 1 h, 3 h, and 6 h timescales but can be updated as the RFC sees fit. The FFG system itself is designed to be based off soil moisture conditions and independent of any rainfall-runoff model. Each RFC uses one of four methods to produce the FFG for their domain: Lumped Flash Flood Guidance (LFFG), Gridded Flash Flood Guidance (GFFG), Distributed Flash Flood Guidance (DFFG), and the Flash Flood Potential Index (FFPI). This results in inconsistencies between each RFC's domain. To combat these inconsistencies, the WPC applies a stitching method to create a 5 km CONUS FFG grid; this grid was used during FFaIR. An example of a CONUS-wide FFG product can be seen in Fig. C.1b.



Legend

1 hour Flash Flood Guidance

Image

🟪 Greater than equal to 5 inches
🟪 4 to 5 inches
🟦 3 to 4 inches
🟦 2.5 to 3 inches
🟦 2.0 to 2.5 inches
🟩 1.5 to 2 inches
🟩 1 to 1.5 inches
🟨 0.75 to 1 inches
🟧 0.5 to 0.75 inches
🟥 0.25 to 0.5 inches
□ Less than 0.25 inches

Figure C.1: *(A) The domain for each NWS RFC (NOAA/NWS/https://water.weather.gov/precip/rfc.php) and (B) an example of a 1 h FFG product on the WPC 5 km mosaic grid, valid on 3 April 2019 at 1800 UTC (https://preview.weather.gov/edd/).*

Precipitation Average Recurrence Intervals (ARI)

The WPC uses precipitation ARIs to determine flooding threats across the CONUS. The ARIs are generated statically from rain gauge climatology. Across most of the CONUS the climatology data is from the NOAA Atlas 14 Climatology. This includes the Northeast, which received an update of the climatology data to NOAA Atlas 14 in the fall of 2015 and TX, which was updated in 2018. The only region not covered by the NOAA Atlas 14 is the Northwest, which is defined as WY, MT, ID, OR, and WA. Since Atlas 14 can not be used in this region, the thresholds for the ARIs are derived from NOAA Atlas 2 (Herman and Schumacher 2016); visit either of the following links for more information: PFDS_1 or PFDS_2.

Precipitation ARIs are defined as the "average time between cases of a particular precipitation magnitude for a specified duration and a given location; the term is associated with the analysis of partial duration series" (NOAA Glossary). In other words, ARIs depict how frequently (i.e. every 5 years, 10 years, 100 years, ect.) rainfall accumulation amounts statically happen during a specified time period (i.e. 5 min, 1 hour, 24 hours, ect). This method of depicting rainfall climatology helps to identify how rare any given rainfall event is, aiding in the situational awareness of the forecaster and alerting them to whether or not an event could potentially be extreme. The standard intervals for ARIs are: 1, 2, 5, 10, 25, 100, 500, and 1000 years. Amounts for the ARI product can be given in years or inches and are usually given over time-periods of 1 h, 3 h, 6 h, and 24 h. An example for the ARI for 24 hours during Hurricane Harvey can be seen in Fig. 2. The 2019 FFaIR experiment provided fully-stitched grids for the CONUS, following the methodology discussed in Herman and Schumacher (2016).

Flooded Locations and Simulated Hydrographs (FLASH)

FLASH is a flash flooding forecast tool that utilizes MRMS data to produce high resolution (1 km/5 min) flash flood forecasts (Gourley et al. 2017 and FLASH Website). "The primary goal of the FLASH project is to improve the accuracy, timing, and specificity of flash flood warnings in the US" (FLASH Website). It's run at NOAA's NSSL by a team comprised of both researchers and students in partnership with OU. The FLASH system compares the MRMS data (radar-estimated only) to the NOAA Atlas 14 climatology data to create ARI forecasts for time periods as short as 5 min. However, the actual output product for the estimated rainfall ARIs are for 30 min and 1, 3, 6, 12, and 24 h (Gourley et al. 2017). The FLASH model was developed with an ensemble framework in mind, allowing it to ingest forcings from a wide range of data. Ensemble in this instance refers to multiple water balance concepts. For example, the CREST hydrological model is used for the simulation of surface water fluxes (FLASH Website). Some output products from the FLASH model include, soil saturation (%) and discharge (m^3 s^{-1}).

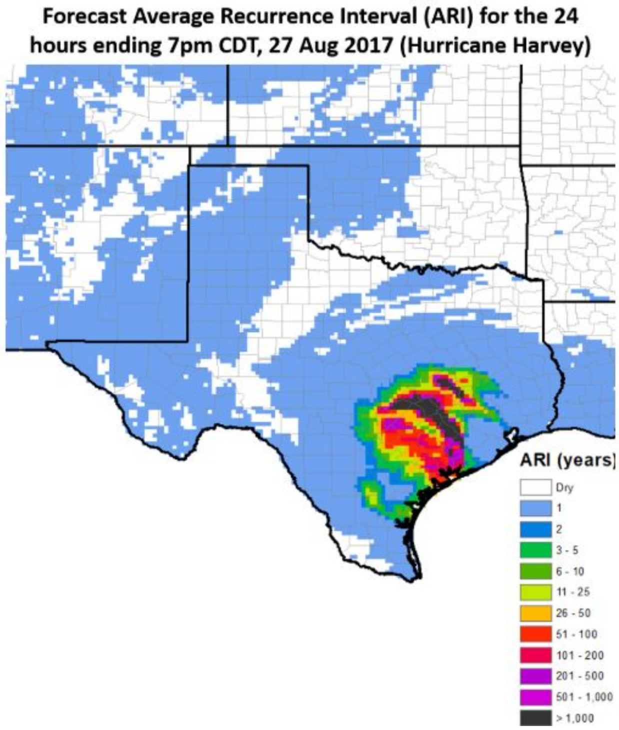

Forecast Average Recurrence Interval (ARI) for the 24 hours ending 7pm CDT, 27 Aug 2017 (Hurricane Harvey)
ARI (years)
Dry
1
2
3 - 5
6 - 10
11 - 25
26 - 50
51 - 100
101 - 200
201 - 500
501 - 1,000
> 1,000

Figure C.2. *An example of a 24 hour ARI for Hurricane Harvey, valid from 00 UTC 27 Aug 2019 to 00 UTC 28 Aug 2019. The pinks/purples are ARIs over 200 years while the darkest color (black) is an ARI greater than 1000 years. Image courtesy of https://metstat.com/real-time/ari/.*

C.2 Operational and Experimental Models: Deterministic and Ensemble

GFDL/EMC GFS-FV3, FV3-SAR, FV3-Nest

GFS-FV3 (GFSv15.1)

As part of the NWS's effort to create a Unified Forecast System (UFS), various NOAA partners were tasked with the development of a new NOAA Global Model core; this will eventually replace the current operational GFS (hereafter GFS). In response the GFSv15, more commonly referred to as the GFS-FV3 (hereafter FV3), was developed. The main difference between the GFS and FV3 is the dynamical core (GFS-FV3 EMC Documentation as of 20190405). The GFS uses a spectral dynamical core while the FV3 is run using a finite volume cubed-sphere dynamic core; for more information on this core click here or here. The finite volume cubed-sphere dynamic core allows the model to be run using Stretched Grid Resolution thus allowing for the resolution of the model to be greater over regions of interest while also enabling the model to zoom in on smaller and smaller storm systems. Conversely, the resolution is then made coarser over the part of the global that NOAA is not concerned with. An example of the FV3 Stretched Grid can be seen in Fig. C.3.

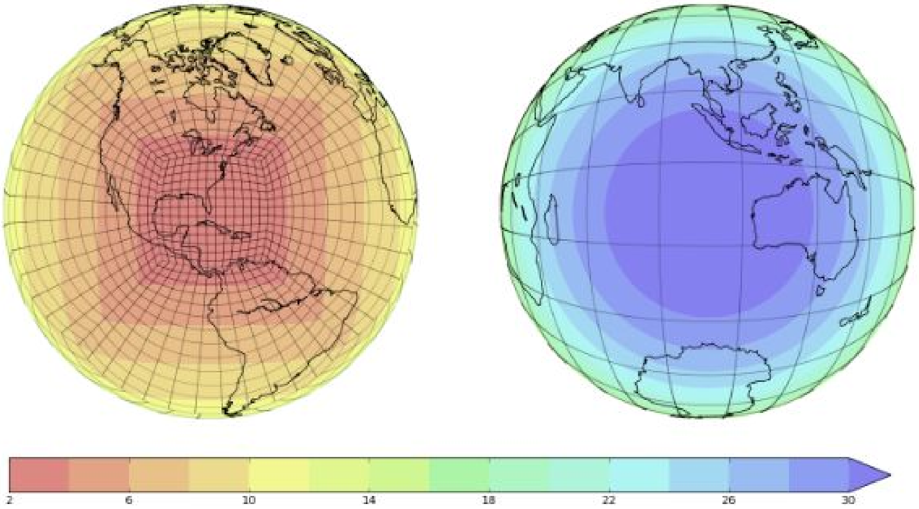


Figure C.3: *An example of the zooming capabilities of the FV3 using the Stretched Grid methodology and the difference in resolution over different parts of the global. The shading represents the horizontal resolution[km]. Image courtesy of NOAA Media Release.*

The FV3, like the GFS, has a horizontal resolution of 13 km and 64 vertical levels. Its output is hourly up to 120 hours and 3 hourly after that up to 384 hours out. It includes the same packages as the GFS except for the following: Zhao-Carr microphysics was replaced with GFDL microphysics, an updated parameterization of ozone photochemistry with additional production and loss terms, a new parameterization of middle atmospheric water vapor photochemistry, a revised bare soil evaporation scheme, a modified convection scheme that reduces the excessive cloud top cooling that is seen in the GFS, and an updated stochastic physics package (GFS-FV3 EMC Documentation). Additionally there were modifications to the data assimilation process as well as changes in the output, post-processed fields and downstream products. Lastly, there were alterations to the vertical velocity, which is now outputted in m/s, and accumulated precipitation variables; see GFS-FV3 EMC Documentation for more detail of all the changes.

FV3-SAR and FV3-Nest (experimental)

"Over the past two years NOAA has developed a limited area modeling capability for the FV3 dynamic core that allows the model to run without the need for a simultaneously integrated global parent domain, thus opening new possibilities for convective-scale modeling with FV3. This limited area capability is informally known as the Stand-Alone Regional (SAR)" - Jacob Carley, EMC. The FV3-SAR therefore differs from the traditional "nested" model, as a "nested" model is run along with the parent global model (in this case FV3) and has a high computational cost. However, the FV3-SAR does not need to be run with its parent model. Instead it uses initial conditions from the 00 UTC FV3 run; this run also provides the lateral boundary conditions (LBC) for the FV3-SAR, which are specified at a 3 hour interval. Both the FV3-SAR and FV3-Nest are currently being run at EMC over the same domain, as can be seen in Fig. C.4. Additionally, a comparison of the NAM-Nest and FV3-SAR can be seen in Fig. C.5.

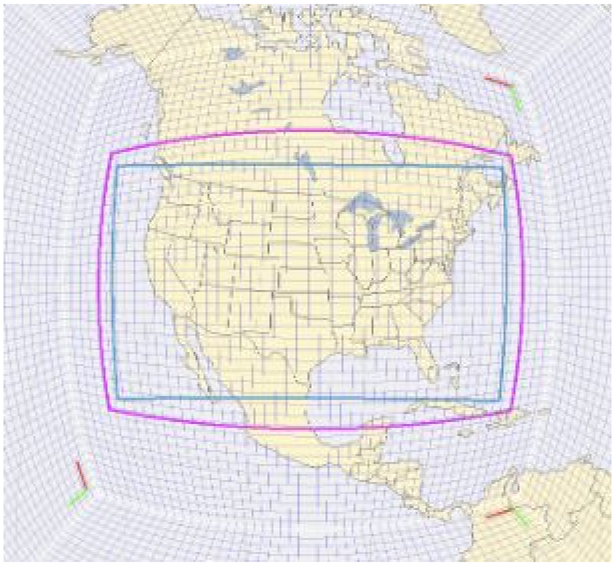


Figure C.4: *The computational grid is in magenta while the output grid is in blue for both the FV3-SAR and the FV3-Nest. This is identical to what is used by the NAM-Nest and HRRR. Image courtesy of Jacob Carley at EMC.*

The FV3-SAR is being developed as part of the UFS goal for a high-resolution regional model. Eventually this regional model suite will replacement the NAM/RAP, the RAP/HRRR, and the SREF. The replacements include all their deterministic and ensemble forecasts, as well as their data assimilation capabilities (UFS 2018). An additional goal of the FV3-SAR is to replace poor performing members of the HREF with FV3-SAR members. This portion of the UFS plan is already underway, with EMC already running both the FV3-SAR and FV3-Nest models and comparing the runs on a regular basis. To help with the evaluation of the FV3-SAR and FV3-Nest, the experiment analyzed the forecasts from both models as well as compared them to one another and to the NAM-Nest. The FV3-SAR and Fv3-Nest domains are the 3 km horizontal grid across the CONUS (Nov2018 update). The models were initialized at 00 UTC with forecasts out 60 h with output hourly.

FV3-SAR-GSD (experimental)

Preliminary testing is now underway for use of the FV3 core at convection allowing scales (referred to as convective allowing models; CAM). The development and testing of FV3-based CAM-scale is currently being worked at EMC, GSD, NSSL, AOML and GFDL, as well as with other partners. Part of this testing involves use of the FV3-SAR to facilitate CAM-scale development, including rapid-cycling data assimilation, that doesn't require use of a global domain and thus reducing the computational cost of the model. GSD is producing a real-time experimental forecast, referred to as FV3-SAR-GSD, at 00 UTC out to 36 hrs. The model uses identical boundary and initial conditions to that of GSD's HRRRv4 experimental forecasts (see HRRRv4 section), but is integrated on the native FV3 gnomonic grid that covers only the CONUS domain of the HRRR with a nominal 3 km grid spacing. The output is mapped to the HRRR lambert-conformal grid for post-processing. Additionally, the FV3-SAR-GSD configuration uses the HRRRv4 physics suite through the Common Community Physics Package (CCPP) interface. No data assimilation is being applied within this forecast.

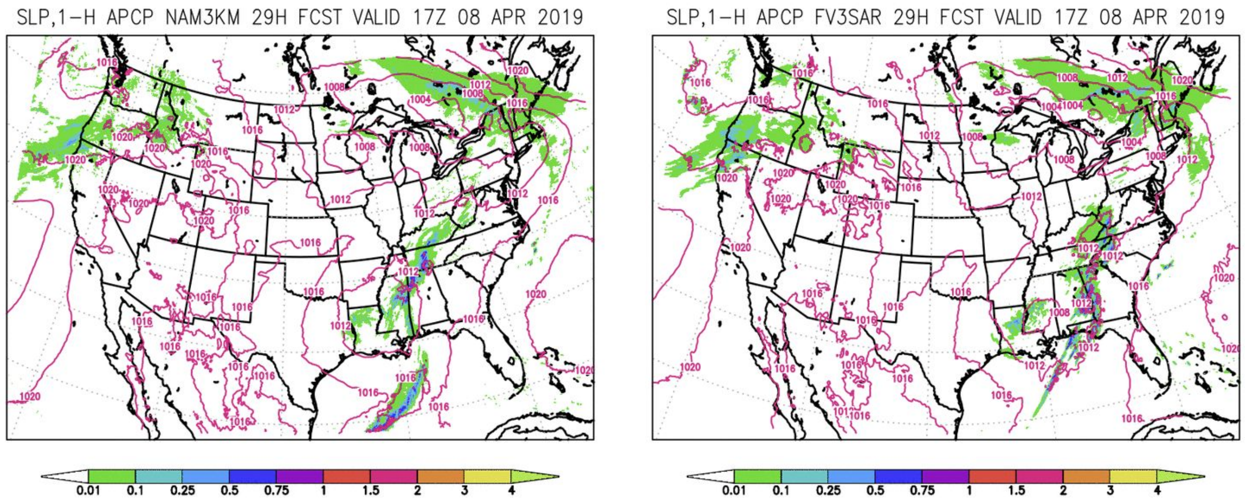

SLP,1-H APCP NAM3KM 29H FCST VALID 17Z 08 APR 2019
SLP,1-H APCP FV3SAR 29H FCST VALID 17Z 08 APR 2019
0.01
0.1
0.25
0.5
0.75
1
1.5
2
3
4

Figure C.5: *An example of a 29 h forecast of SLP and 1 h QPF from the NAM-Nest (left) and the FV3-SAR (right). The images are valid on 08 April 2019 at 1700 UTC. Courtesy of Ben Blake's webpage (EMC).*

ESRL High Resolution Rapid Refresh (HRRR)

The 2019 FFaIR experiment evaluated two related components of the experimental HRRR: the deterministic component of the system, HRRRv4, and the ensemble component of the system, HRRRE (TB-Documentation). The HRRR is updated hourly, and the 3-km grid spacing allows for explicit convention. Whereas the operational HRRRv3 gets its initial conditions from the 13-km Rapid Refresh model (RAP), the experimental HRRRv4 get its initial conditions from the ensemble mean of the HRRRE analysis system (HRRRDAS). Both HRRRv3 and HRRRv4 get lateral boundary conditions from the RAP. Additionally, the initialization includes assimilation of 3 km radar data, with the radar data also being assimilated into the model every 15 min (ESRL-HRRR).

HRRRv4 (deterministic)

The operational version of the HRRR model (HRRRv3) was implemented in late spring 2018. As can be seen in Table C.1, the 00 UTC, 06 UTC, 12 UTC, and 18 UTC cycles of the HRRRv3 have a forecast length of 36 h (TB-Documentation and Manikin 2018), with all other hourly runs out to 18 h. The experimental version (HRRRv4) has increased forecast length to 48 h for the 00 UTC, 06 UTC, 12 UTC, and 18 UTC cycles; refer to Table C.1. For all other forecast cycles, the HRRRv4 runs out to 18 h like HRRRv3. Figure C.6 depicts the HRRRv4 forecast suite. Like HRRRv3, HRRRv4 has a 3 km grid that covers the entire CONUS and the vertical resolution remains unchanged as well. The HRRRv4 also features the WRF-ARW as its core like HRRRv3, however the version of the WRF core has been updated to WRF-ARWv3.9.1. Other updates to the experimental version of the HRRR model include changes to the physics such as an improved land-surface/snow model, a lake model for small lakes and a MYNN PBL (Nakanishi and Niino PBL's surface layer scheme; Nakanishi and Niino 2006) update for better sub-grid clouds, and improved EDMF (eddy diffusivity mass flux) mixing.

There were also changes in the data assimilation (DA) for the model, with the HRRRv4 DA including lightning data from GOES (GLM), aircraft and RAOB moisture observations above 300 mb, and tropical cyclone central pressure estimates from TCvitals for improved position and structure of tropical systems. Furthermore there is the potential for some additional radiance data from GOES-16 (TB-Documentation). A storm-scale ensemble data assimilation system (HRRRDAS) is also being tested in the CONUS HRRRv4 that uses 36 hourly-cycled CONUS HRRR members with assimilation of conventional, radar and satellite observations through GSI-EnKF. This system is designed to improve use of observations during the DA process with a better representation of meso-to-storm scale covariances when compared with the comparatively coarse GDAS used in HRRRv3. The intended benefits of HRRRDAS are: more accurate retention and evolution of meso-to-storm scale features, particularly in the early forecast hours. The HRRRDAS system also forms the basis for HRRR ensemble (HRRRE), which is described below.

Changes were also made in the post-processing; for a full breakdown of common modifications for the HRRRv4 see Fig. C.7. This included addressing the overall warm bias in the 2 m temperatures that is seen in the HRRRv3. Preliminary testing shows that although this bias is not completely corrected in version 4, the bias has been reduced (March 2019 update). A full description of the HRRRv4 model can be found at: https://rapidrefresh.noaa.gov/hrrr/. Finally, because HRRRv4 is considered experimental, part of the evaluation of the model included comparisons between the HRRRv4 and the HRRRv3.

Table C.1: *The forecast length for each forecast cycle for the operational and experimental RAP and HRRR models; courtesy of TB-Documentation.*

Model	Cycle	Forecast Length
RAPv4/v5	04-08, 10-14, 16-20, 22-02 UTC	21 hrs
RAPv4 (operational)	03, 09, 15, 21 UTC	39 hrs
RAPv5 (experimental)	03, 09, 15, 21 UTC	51 hrs
HRRRv3/v4	01-05, 07-11, 13-17, 19-23 UTC	18 hrs
HRRRv3 (operational)	00, 06, 12, 18 UTC	36 hrs
HRRRv4 (experimental)	00, 06, 12, 18 UTC	48 hrs

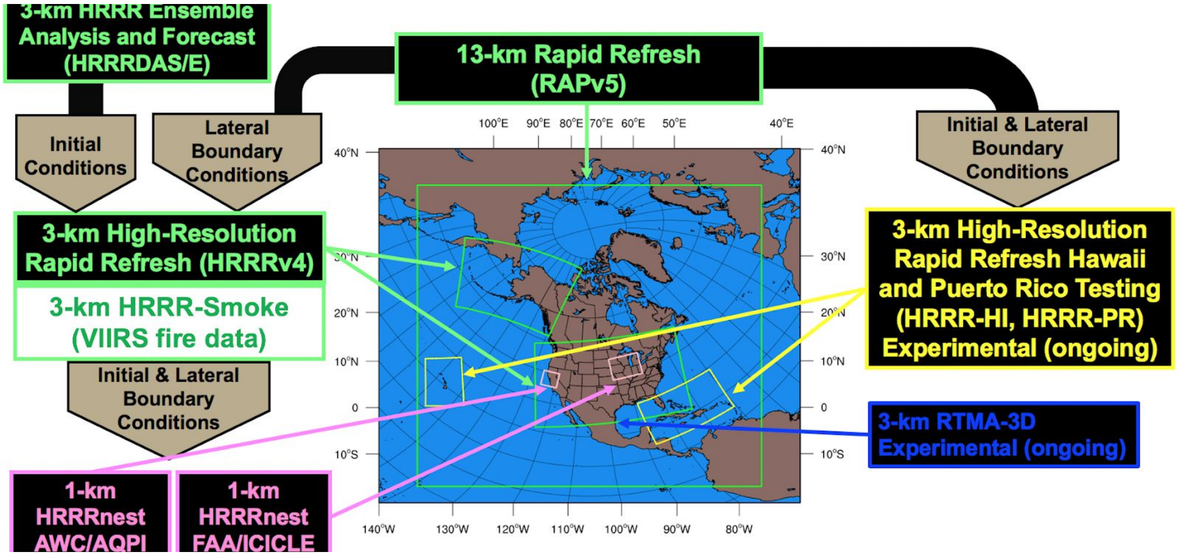
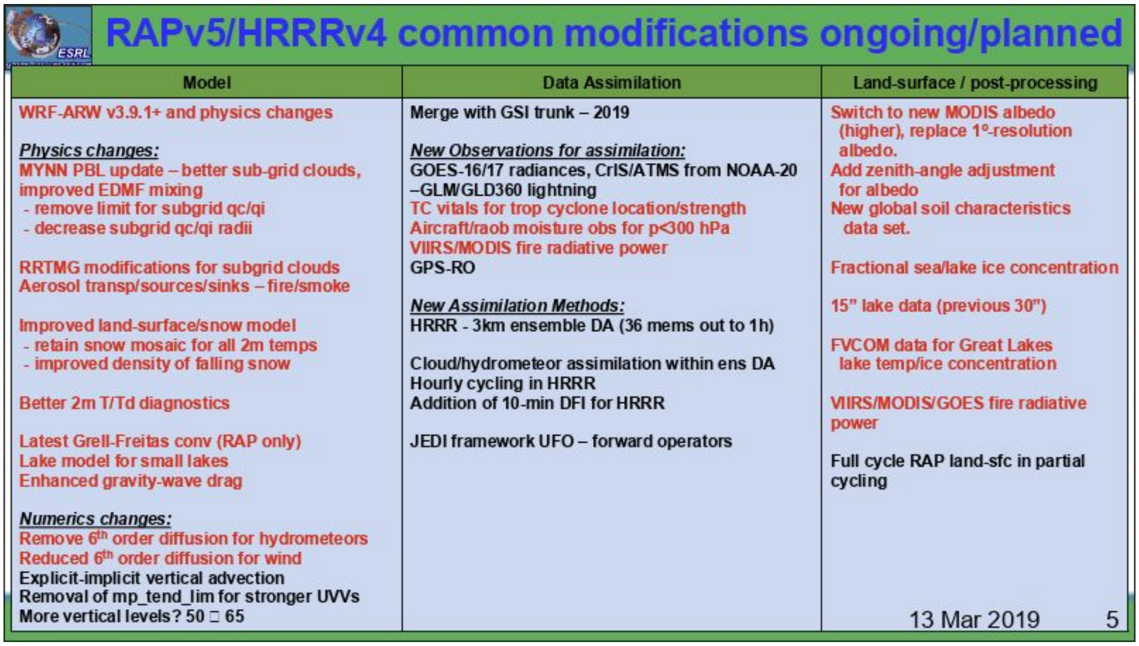

3-km HRRR Ensemble Analysis and Forecast (HRRRDAS/E)
13-km Rapid Refresh (RAPv5)
Initial Conditions
Lateral Boundary Conditions
Initial & Lateral Boundary Conditions
3-km High-Resolution Rapid Refresh (HRRRv4)
3-km HRRR-Smoke (VIIRS fire data)
Initial & Lateral Boundary Conditions
3-km High-Resolution Rapid Refresh Hawaii and Puerto Rico Testing (HRRR-HI, HRRR-PR) Experimental (ongoing)
3-km RTMA-3D Experimental (ongoing)
1-km HRRRnest AWC/AQPI
1-km HRRRnest FAA/ICICLE
100°E
90°E
80°E
70°E
60°E
50°E
40°E
40°N
30°N
20°N
10°N
0
10°S
140°W
130°W
120°W
110°W
100°W
90°W
80°W

Figure C.6: *The experimental HRRRv4 forecast suite provided by David Dowell.*



RAPv5/HRRRv4 common modifications ongoing/planned

Model	Data Assimilation	Land-surface / post-processing
WRF-ARW v3.9.1+ and physics changes *Physics changes:* MYNN PBL update – better sub-grid clouds, improved EDMF mixing - remove limit for subgrid qc/qi - decrease subgrid qc/qi radii RRTMG modifications for subgrid clouds Aerosol transp/sources/sinks – fire/smoke Improved land-surface/snow model - retain snow mosaic for all 2m temps - improved density of falling snow Better 2m T/Td diagnostics Latest Grell-Freitas conv (RAP only) Lake model for small lakes Enhanced gravity-wave drag *Numerics changes:* Remove 6th order diffusion for hydrometeors Reduced 6th order diffusion for wind Explicit-implicit vertical advection Removal of mp_tend_lim for stronger UVVs More vertical levels? 50 □ 65	Merge with GSI trunk – 2019 *New Observations for assimilation:* GOES-16/17 radiances, CrIS/ATMS from NOAA-20 –GLM/GLD360 lightning TC vitals for trop cyclone location/strength Aircraft/raob moisture obs for p<300 hPa VIIRS/MODIS fire radiative power GPS-RO *New Assimilation Methods:* HRRR - 3km ensemble DA (36 mems out to 1h) Cloud/hydrometeor assimilation within ens DA Hourly cycling in HRRR Addition of 10-min DFI for HRRR JEDI framework UFO – forward operators	Switch to new MODIS albedo (higher), replace 1°-resolution albedo. Add zenith-angle adjustment for albedo New global soil characteristics data set. Fractional sea/lake ice concentration 15" lake data (previous 30") FVCOM data for Great Lakes lake temp/ice concentration VIIRS/MODIS/GOES fire radiative power Full cycle RAP land-sfc in partial cycling

13 Mar 2019 5

Figure C.7: *Chart of some of the common modification for HRRRv4; provided during the March 2019 update.*

HRRR Ensemble (HRRRE; experimental)

As stated, in addition to the deterministic HRRRv4, the 2019 FFaIR Experiment also featured the experimental HRRR ensemble model (HRRRE), which saw an increase in its domain since the 2018 FFaIR experiment from half CONUS to full CONUS. Figures C6 and Fig. C.8a show the model domain of HRRRE examined in 2018. This domain did not cover the western portion of the CONUS, however, the HRRRE domain now encompasses the entire CONUS; see Fig. C.8b. The HRRRE domain has 3 km horizontal grid spacing. In total, the model produces 36 HRRRE analyses and 9 HRRRE forecasts. Forecasts are provided three times a day, once at 00 UTC with a forecast length of 36 h, then again at 12 UTC with a forecast length of 24 h and finally at 18 UTC with a forecast length of 18 h (TB-Documentation).

The HRRRE used in the 2019 FFaIR Experiment was initialized daily from the GFS ensemble mean, the GFS Data-Assimilation Ensemble (GDAS) and the RAP/HRRR. The GDAS provides the atmospheric perturbations for the HRRRE while the RAP/HRRR models provide the land surface data. There are 36 HRRRE members, all of which are cycled hourly with GSI-EnFK to assimilate convectional and radar-reflectivity observations (Dowell et al. 2018a, Dowell et al. 2018b and TB-Documentation). The sources of spread are the result of: hourly DA (posterior inflation), lower boundary perturbations (soil moisture), lateral boundary perturbations, and stochastic parameter perturbations across most/all of the RAP/HRRR physics suite.

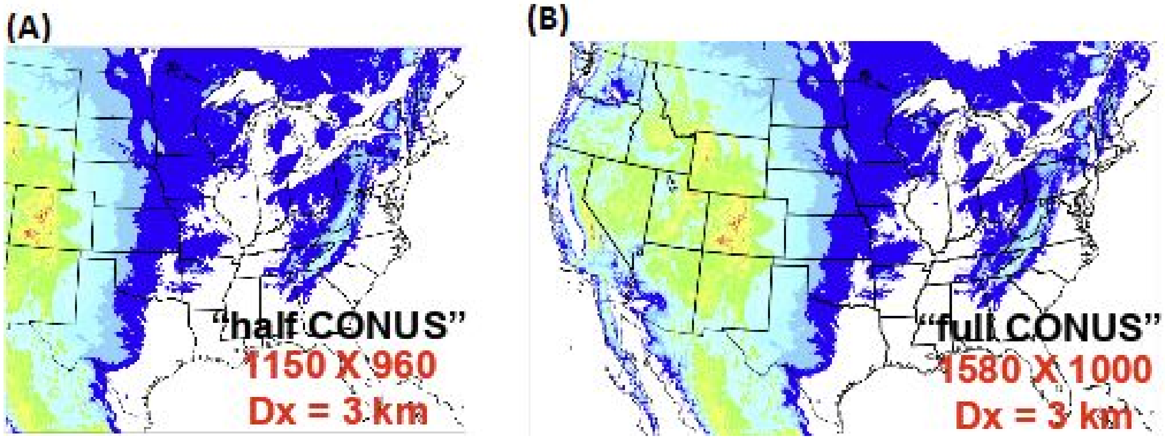


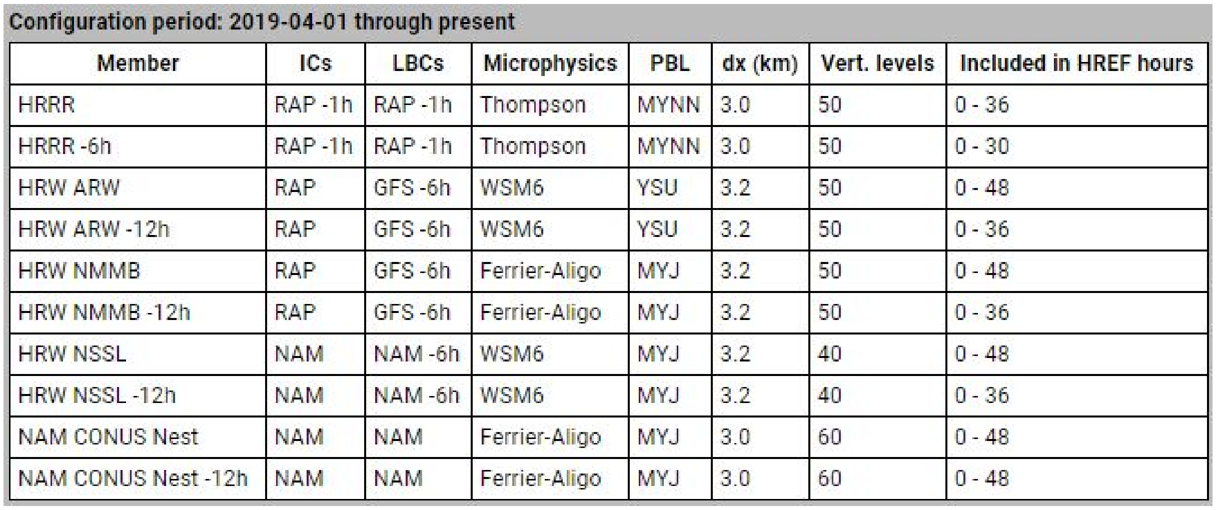
Figure C.8: *The HRRRE model domain for (A) 2018 HRRRE and (B) 2019 HRRRE; courtesy of the March 2019 update.*

EMC Experimental High Resolution Ensemble Forecast (HREFv3)

The HREFv3 is an experimental ensemble product run by EMC that is generated using multiple runs of operational and experimental CAMs with horizontal resolution of ~3 km. The HREFv3 membership differs from the operational HREF membership (which can be seen in Table C.2) consisting of 9 ensemble members instead of 10. Additionally, the experimental model included membership from the HRRR rather than the Hires W-NMMB model. The membership for HREFv3 can be seen in Fig. C.9 and consists of “the two most recent runs of the NAM CONUS nest, the HRRR, the two different HiresW ARW members, and just the 00Z cycle for the regional FV3 run (FV3-SAR). 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 for computation of mean fields” (Documentation from Matt Pyle). The HREFv3 was run once daily at 00 UTC and forecast out 36 h.

Various guidance and products were provided by EMC for the 2019 FFaIR Experiment from HREFv3. This includes probabilistic guidance such as neighborhood probabilities (Harless et al. 2010) of precipitation exceeding FFG, ARI, and various thresholds. The QPF exceedance probabilities were neighborhood probabilities computed over a 40 km radius for an array of thresholds and duration periods (i.e. probability of exceeding 4 inches in 6 h). Furthermore, the following forms of the ensemble mean precipitation were examined: the conventional mean, a full-domain probability matched (PMM; Ebert 2001) mean, a localized PM mean (LPMM; Clark 2017 and Snook et al. 2019). An example of the 3 h QPF output using PMM and LPMM means can be seen in Fig. C.10.

Table C.2: *The operational HREF (version 2.1) ensemble configuration as of 20190401 from https://www.spc.noaa.gov/exper/href/#. Abbreviations are as follows: IC- initial conditions, LBC-lateral boundary conditions, PBL-Planetary Boundary Layer, and dx-horizontal grid size.*



Configuration period: 2019-04-01 through present

Member	ICs	LBCs	Microphysics	PBL	dx (km)	Vert. levels	Included in HREF hours
HRRR	RAP -1h	RAP -1h	Thompson	MYNN	3.0	50	0 - 36
HRRR -6h	RAP -1h	RAP -1h	Thompson	MYNN	3.0	50	0 - 30
HRW ARW	RAP	GFS -6h	WSM6	YSU	3.2	50	0 - 48
HRW ARW -12h	RAP	GFS -6h	WSM6	YSU	3.2	50	0 - 36
HRW NMMB	RAP	GFS -6h	Ferrier-Aligo	MYJ	3.2	50	0 - 48
HRW NMMB -12h	RAP	GFS -6h	Ferrier-Aligo	MYJ	3.2	50	0 - 36
HRW NSSL	NAM	NAM -6h	WSM6	MYJ	3.2	40	0 - 48
HRW NSSL -12h	NAM	NAM -6h	WSM6	MYJ	3.2	40	0 - 36
NAM CONUS Nest	NAM	NAM	Ferrier-Aligo	MYJ	3.0	60	0 - 48
NAM CONUS Nest -12h	NAM	NAM	Ferrier-Aligo	MYJ	3.0	60	0 - 48

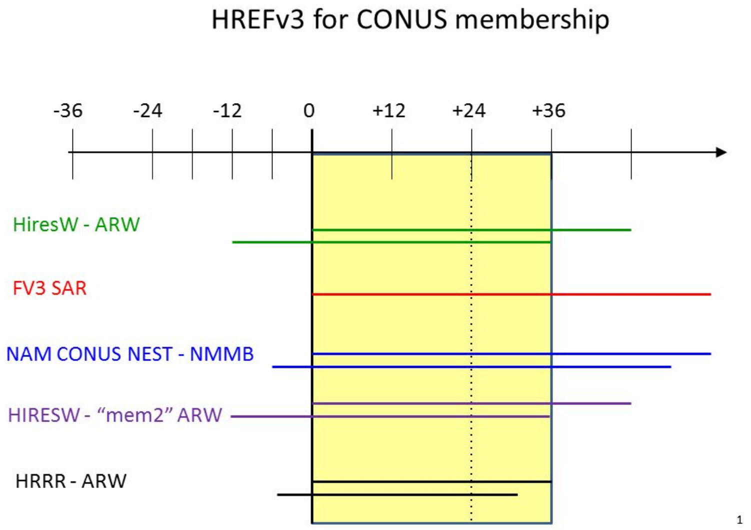


Figure C.9: *The HREFv3 ensemble membership which includes both a real-time and time-lagged member from: HiresW-ARW, NAM CONUS NEST-NMMB, HiresW-"mem2" ARW, and the HRRR-ARW; FV3-SAR only provides a real-time member. Image provided by Matt Pyle at EMC.*

Lastly, EMC tasked FFaIR to evaluate the capabilities of the SREF model in comparison with HREFv3 and other ensemble models such as the HRRRE and the GEFS. The SREF is a Short Range Ensemble Forecast Model that has been run operationally since 2001 (McQueen et al. 2004) and was developed to provide a short-range (0-3 days), multi-regional ensemble model. The model has been frozen, i.e. there will be no more updates to the model, and EMC is working towards turning off the model. However, until EMC knows that other models, such as the HREF and the GEFS, can provide the same utility to the forecast from the 0-3 day time-period it will remain operational. To help with the process, FFaIR evaluated the utility of other ensemble members in comparison to the SREF. During the experiment participants were asked to evaluate various ensemble models alongside the SREF for Day 1. Meanwhile, the FFaIR team performed an evaluation over Day 2 and Day 3.

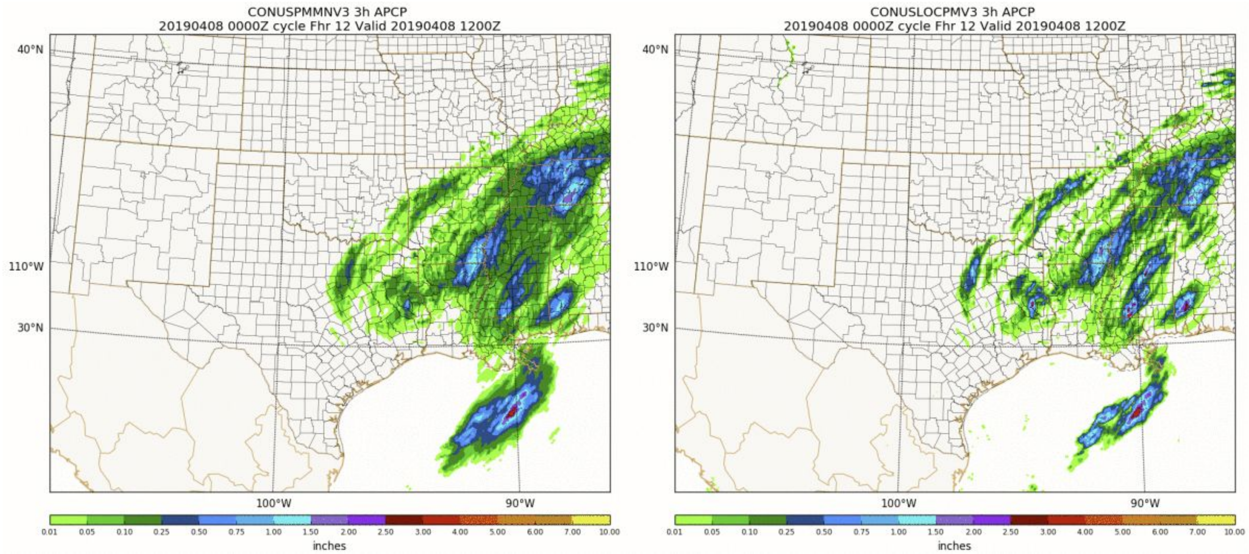


Figure C.10: *Left: the 3 h QPF forecast calculated using the PM mean from the HREFv3, valid at 08 April 2019 at 1200 UTC. Right: same as the left but the mean is calculated using the LPM method. Images courtesy of Matt Pyle.*

OU/CAPS Storm-Scale Ensemble Forecast (SSEF)

The SSEF is an storm-scale ensemble model that is produced by the Center for Analysis and Prediction of Storms (CAPS) and has been tested at both the HWT and HMT for several years. The WRF-ARWv3.9.1.1 is used for the model membership that are 3DVAR-based and the EnKF, coupled with ARPSv5.4, bases. Meanwhile the 3.9_GSD WRF version is used for the members that are single-physics and stochastic-physics based (2018 CAPS). SSEF is run on a 3 km grid across the CONUS and the initial and lateral boundary conditions vary depending on the member; see Table C.3. Radial velocity and reflectivity data are assimilated into the SSEF as well as other available observations using the single-time (non-cycled) ARPS 3DVAR with complex cloud analysis system (http://forecast.caps.ou.edu/). The SSEF produces 60 h forecasts, though some of the membership only provide forecasts out 36 h, and is run at 00 UTC. For the 2019 FFaIR Experiment the SSEF consists of the 14 members listed in Table C.3. Additionally, the FFaIR experiment evaluated three individual members of the SSEF model, the member with Thompson physics (core_cnt1) the member with the Morrison physics (MP-2) and the member with the NSSL physics (MP-1) .

***Table C.3**: OU/CAPS SSEFX configurations for 2019 HMT FFaIR.*

Members	IC	LBC	Microphysics	LSM	Model
core_cntl	NAM	NAM	Thompson	NOAH	FV3
core_lsm1	NAM	NAM	Thompson	RUC	FV3
core_mp1	NAM	NAM	NSSL	NOAH	FV3
core_mp2	NAM	NAM	Morrison	NOAH	FV3
core_pbl1	NAM	NAM	Thompson	NOAH	FV3
core_pbl2	NAM	NAM	Thompson	NOAH	FV3
core_sargfs	GFS	GFS	Thompson	NOAH	FV3
core_sfc1	NAM	NAM	Thompson	RUC	FV3
pert_lsm1	NAM+SREF arwp3	SREF arwp3	Thompson	RUC	FV3
pert_mp1	NAM+SREF arwp1	SREF arwp1	NSSL	NOAH	FV3
pert_mp2	NAM+SREF arwn2	SREF arwn2	Morrison	NOAH	FV3
pert_pbl	NAM+SREF arwn1	SREF arwn1	Thompson	NOAH	FV3
pert_pbl2	NAM+SREF arwp2	SREF arwp2	Thompson	NOAH	FV3
pert_sfc1	NAM+SREF arwn3	SREF arwn3	Thompson	RUC	FV3

C.3 Other Experimental Forecast Tools

In addition to evaluating the previously listed operational and experimental flood and model guidance, the 2019 FFaIR experiment examined the utility of a variety of other data spanning from machine learning to remote sensing products.

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

The ERO "First-Guess" Field that was used and evaluated in the 2019 FFaIR experiment was developed by Greg Herman and Russ Schumacher at Colorado State University (CSU). It is a prediction system that utilizes a random forest (RF) machine learning algorithm (for more information in the algorithm refer to Appendix A in Herman and Schumacher 2018). A RF can be thought of as an ensemble of decision trees, where each individual tree makes a deterministic prediction about a predictand.   These are then put together as an ensemble and the relative frequencies of the predictand outcome are used to create a probabilistic forecast. The goal of the RF method is to have a large set of decision trees that are uncorrelated which can be used to train the model. “One of the most powerful aspects of machine learning algorithms—and RFs in particular—is finding patterns and nonlinear interactions in the supplied training data” (Herman and Schumacher 2018). A large and diverse dataset allows the RF algorithm to diagnose and correct model bias automatically by incorporating information about not just QPF but how other atmospheric ingredients influence the precipitation forecast.

The CSU ERO “First-Guess” model uses RF to train the model to probabilistically predict excessive rainfall (defined here as either a 2 year NOAA 14-based ARI exceedance or a flash flood report) for Days 1-3 across the CONUS. These probabilities are based on a record of comparisons between historical model forecasts and precipitation observations. A general overview of how the process works can be seen in Fig. C.11. Since ARIs vary significantly depending on were in the CONUS a gridpoint is located, the CONUS was divided into eight regions, with the model trained separately for each of these regions.

Two different versions of the of the machine learning model were evaluated during the 2019 FFaIR experiment. The first model, which was referred to as NSSL-WRF based, is trained using the NSSL-WRF; this is also used for its real-time input. This version of the “First-Guess” Field produces a Day 1 forecast for a 24 h period from 12-12 UTC, a 6 h period from 18-00 UTC, and a 6 h period from 00-06 UTC; an example of a Day 1 ERO “First Guess” forecast can be seen in Fig. C.12. The other version of the model, GEFS/O, is trained using the GEFS/R (Global Ensemble Forecast System Reforecast) and uses the operational GEFS for its real-time input. The GEFS/O based ERO fields produce a forecast for Days 1-3, however, only the Day 1 field was evaluated during FFaiR this year. The Day 2 and Day 3 forecasts were evaluated at last year’s FFaIR and deemed ready for operations and are currently being used in operations at the WPC.

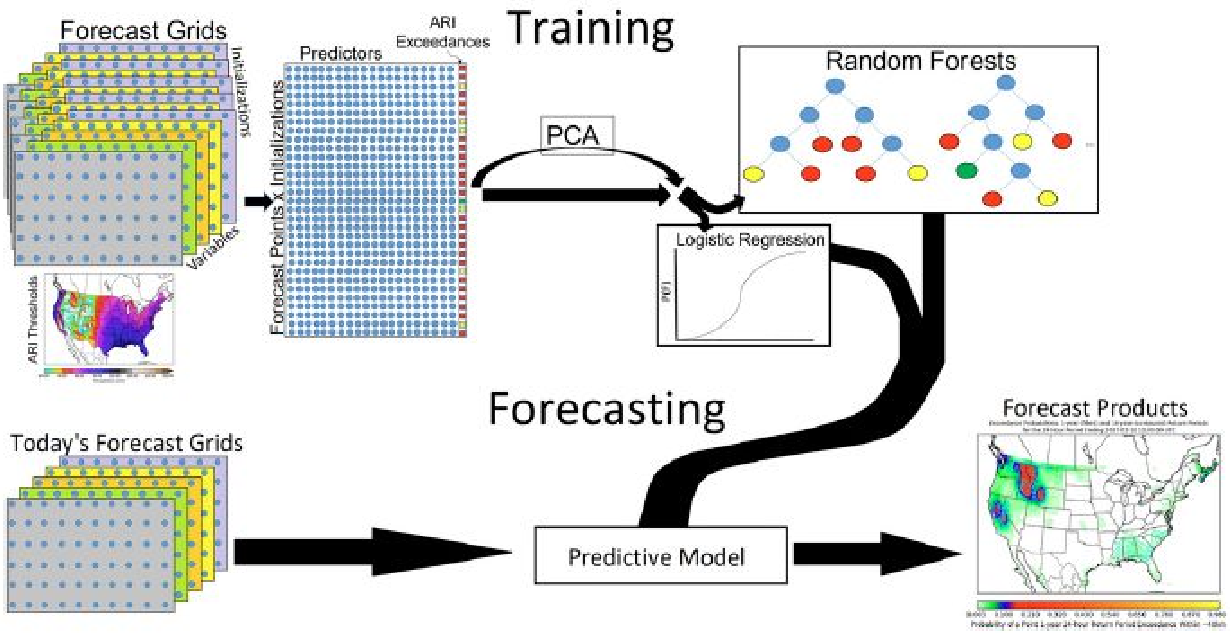


Figure C.11: *A simple schematic of the forecast process for the ERO "First-Guess" model. Forests from various models, such as the GEFS/R, are taken, assembled across fields, space and time to form a training matrix. Past observations are used to associate a label with each forecast initialization, i.e. the forecast day. The training matrix then is preprocessed using Principal Component Analysis (PCA) and then the matrix becomes the input for the ML algorithms. From here, the ARI exceedance probabilities can be generated; image courtesy of Herman and Schumacher (2018).*

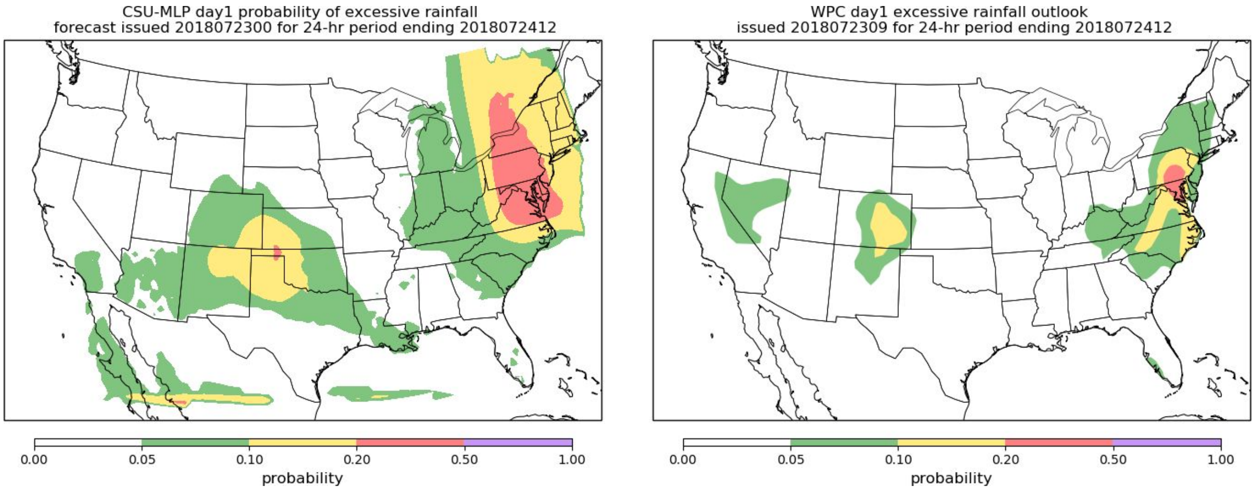

CSU-MLP day1 probability of excessive rainfall
forecast issued 2018072300 for 24-hr period ending 2018072412
WPC day1 excessive rainfall outlook
issued 2018072309 for 24-hr period ending 2018072412
0.00
0.05
0.10
0.20
0.50
1.00
probability
0.00
0.05
0.10
0.20
0.50
1.00
probability

Figure C.12: *An example of the Day 1 ERO forecast from: (left) the NSSL-WRF based CSU "First Guess" model and (right) the Weather Prediction Center (WPC). Both are valid for the 24 h period ending on 24 July 2018 at 1200 UTC. Image provided by Russ Schumacher at CSU.*

CIRA Layer Precipitable Water and Model Difference Products

The Cooperative Institute for Research in the Atmosphere (CIRA) at CSU recently developed a satellite-derived product to examine column and layer precipitable water (PW). The Advected Layer Precipitable Water (ALPW) is derived over for four layers: sfc-850 mb, 850-700 mb, 700-500 mb, and 500-300. This product is from a combination of passive microwave water vapor profile soundings from seven spacecraft (Suomi-NPP, NOAA-19/20, Metop-A/B, and DMSP F-17/18) and the GFS model winds. These products help forecasters determine the depth of water vapor, especially in the mid- to upper-levels, to determine if moisture is available to enhance precipitation. CIRA also created a total column product, TPW, that is currently operational throughout the NWS. Real-time animations of the ALPW and TPW can be found here.

Both the ALPW and TPW products have been part of the FFaIR experiment since 2017. However, CIRA is currently developing a new version of the TPW product. The operational TPW product, referred to as the Blended Total Precipitable Water (BTPW), provides hourly fields of TPW derived from polar orbiter microwave retrievals and surface-based GPS measurements. It does not currently use GOES-16 data. On the other hand, the new product that was tested, referred to the Merged TPW, uses advected polar orbiting microwave retrievals in combination with GOES-16 TPW, which is overlaid in clear regions. A comparison of the BTPW and the Merger TPW products can be found in Fig. C.13. The Merged TPW version 1.0 product is updated every hour at approximately 50min past the hour. The experimentalMerged TPW v1.0 was evaluated during this year's FFaIR experiment and was also compared to the operational BTPW to help determine its utility and forecaster preference.

Lastly, like in previous FFaIR experiments model water vapor difference products was also used/evaluated. Model Layer Precipitable Water (LPW) products are derived from model water vapor profiles over the same layers as the observed products. These profiles are then compared against the observed ALPW products and a difference field is created. This allows forecasters to identify regions where the model is over/underestimating the water vapor. During the experiment only HRRR model derived products were evaluated. CIRA provided the difference fields of the 1 h HRRR forecast from the ALPW for the four layers listed previously. An example of the difference fields can be seen in Fig. C.14 and real-time can be found here.

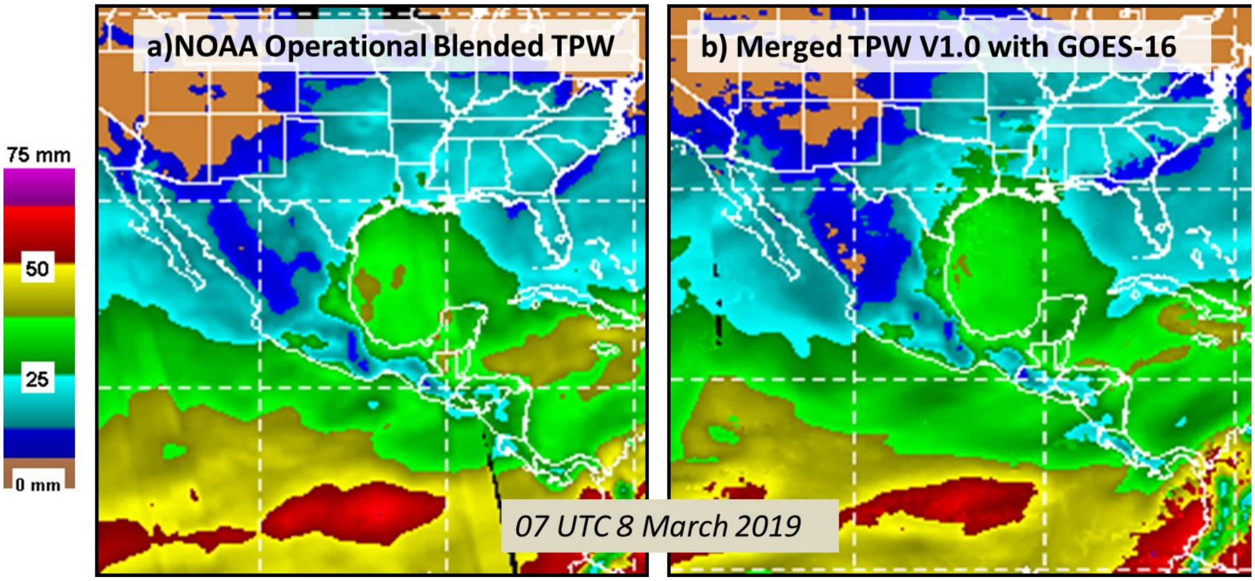

75 mm
50
25
0 mm
a)NOAA Operational Blended TPW
b) Merged TPW V1.0 with GOES-16
07 UTC 8 March 2019

Figure C.13: *An example of the two CIRA TPW products that will be compared during the 2019 FFaIR experiment valid at 0700 UTC on 8 March 2019: (a) the NOAA operational BTPW product and (b) the experimental Merged TPW v1.0; image courtesy of John Forsythe.*

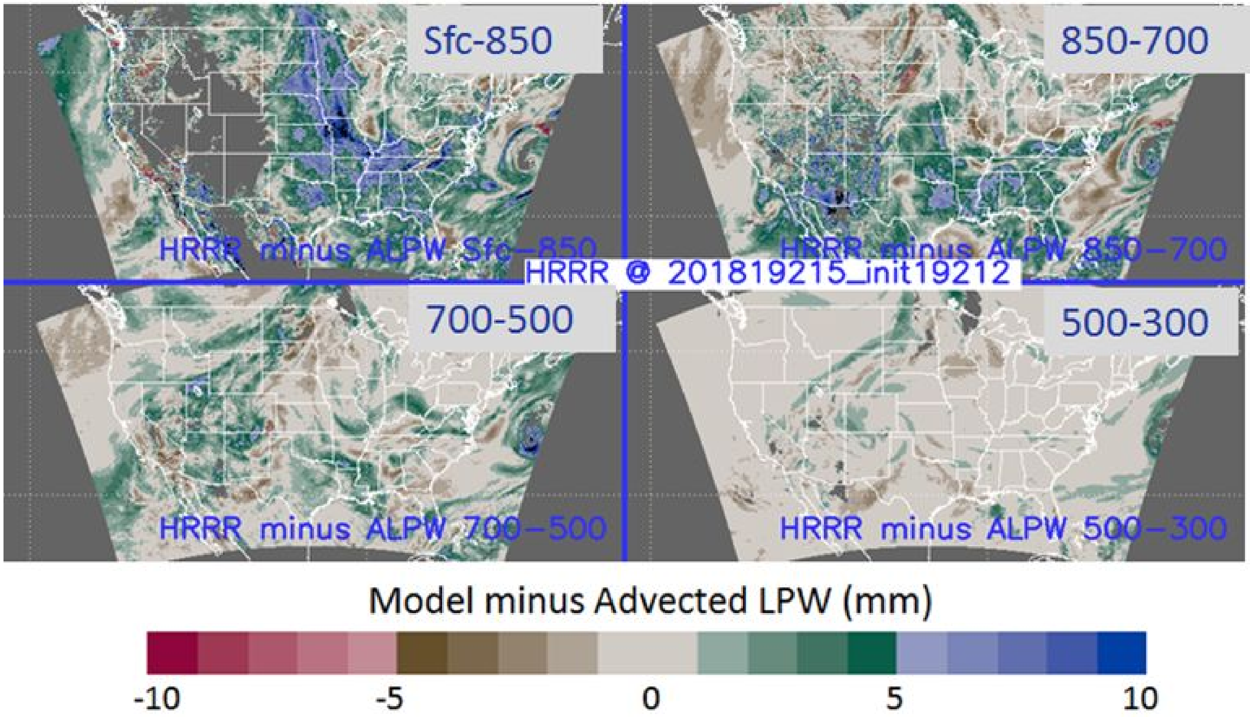


Figure C.14: *Operational HRRR 3 h forecast layer precipitable water minus CIRE LPW (mm) valid at 1500 UTC on 11 July 2018; image courtesy of John Forsyth .*

Appendix D

WPC MODE Settings for the Objective Verification

- 36 HR & 24 HR QPF verified against MRMS-GC QPE
    - 00/12 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 and Configuration File Settings
    - Grid stats harvested from MODE CTS
    - Circular convolution radius of 5 grid squares used
    - Double thresholding technique applied
    - Area threshold of 50 grid squares to keep an object
    - Total interest threshold for determining matches = 0.6