2020 Flash Flood and Intense Rainfall Experiment: *Findings and Results*

June 15 - July 17, 2020
Weather Prediction Center
Hydrometeorology Testbed

Sarah Trojniak - Systems Research Group, NOAA/NWS/WPC/HMT
James Correia Jr. - CIRES CU Boulder, NOAA/NWS/WPC/HMT
Benjamin Albright - Systems Research Group, NOAA/NWS/WPC/HMT

Table of Contents

1. Introduction

The Flash Flood and Intense Rainfall (FFaIR) Experiment has been held annually since 2012 at the Weather Prediction Center (WPC) and is part of the Hydrometeorology Testbed (HMT). The experiment spans four weeks between June and July and brings together people from across the meteorological community to focus on increasing the skill of forecasting heavy rainfall and flash flooding. Usually the experiment takes place in person at the National Oceanic and Atmospheric Administration (NOAA) Center for Weather and Climate Prediction. However, because of the national pandemic, this year's experiment was held virtually for the first time, leading to its own set of challenges and lessons learned.

An overarching goal of FFaIR each year is to help with the National Weather Service's (NWS) research to operations (R2O) process by evaluating new heavy rainfall and flash flooding guidance in a semi-operational setting. This year focused on evaluating the utility of the FV3[1] based Convective Allowing Models (CAMs) for assessing the excessive rainfall and flash flooding threat in the Day 1 timeframe. New tools such as a heavy rainfall object tracker were also evaluated. A full list of what was evaluated can be seen in Appendix A.

2. Operations and Science

As stated, the FFaIR Experiment brings together meteorologists from a variety of backgrounds, ranging from academia and students, to model and product developers to forecasters and hydrologists; see a full list of the participants in Appendix B. This year the experiment participant list even included a forecaster from the German Meteorological Service (Deutscher Wetterdienst or DWD). The diverse group allows for unique conversations and communication between the various fields, leading to scientific growth, new collaboration, and better understanding among the disciplines of the challenges experienced in other parts of the weather enterprise. This interaction, though not necessarily a formal goal, is a vital element of not only FFaIR but for all NWS Testbeds.

The FFaIR Experiment uses a pseudo-operational environment to expose the participants to the new models and tools. The pseudo-operational setting is achieved by mimicking operations at WPC's Day 1 Quantitative Precipitation Forecast (QPF) and MetWatch desk, requiring the participants to issue experimental forecasts in real time, with product issuance deadlines; this is discussed in further detail below. In addition to using them in a real time setting, the participants also spend time subjectively evaluating various aspects of the guidance.

[1] FV3 stands for finite volume cubed-sphere. It is the dynamical core that is planned to be the backbone of all NWS models going forward.

Combined, the operational setting and subjective evaluation help to address the scientific objectives of the FFaIR Experiment. The main focus in regards to model evaluation is the analysis of different FV3-CAM configurations. During FFaIR, eight different configurations were analyzed along with an FV3-CAM ensemble. Below is the full list of science objectives the 2020 FFaIR Experiment focused on:

- Evaluate the usefulness of operational and experimental products from high resolution convective-allowing deterministic models and ensembles for forecasting near-term flash flood events.
- Evaluate various configurations for the FV3-SAR[2] (SAR: stand along regional) from both the Environmental Modeling Center (EMC) and Global Systems Laboratory (GSL), focusing on things like timing, amount, and location of QPF.
- Evaluate the utility of the WPC's experimental Heavy Precipitation Object Tracker (HPOT) and collect feedback about possible improvements to the product and how it is displayed.
- Evaluate, subjectively and objectively, the Colorado State University Machine Learning Products (CSU-MLP) "First Guess Field" Marginal, Slight, Moderate, and High Risks for the Day 1 ERO.
- Evaluate the spatial aligned mean (SAM) and the SAM-LPM (a blend of SAM and the local probability matched mean) from CAPS-OU SSEF[3].
- Use a new method of subjective verification to see if it provides more useful information to developers. The method will focus on understanding what participants felt was useful and what information they receive from the models rather than simply assigning a "goodness" value to the product.
- Unique aspects of the forecast such as timing, duration, location, and maximum accumulation were forecasted interactively by the participants. Implementation of the new method of interactive forecasting included the participants creating their own forecast of QPF exceedances and was used to provide a perspective similar to a Mesoscale Precipitation Discussion (MPD).

[2] As of September 2, 2020 EMC changed the name of the FV3-SAR to the FV3-LAM (limited area model).
[3] CAPS-OU: Center for Analysis and Prediction of Storms at the University of Oklahoma.
SSEF: Storm Scale Ensemble Forecast.

2.1 Virtual Operations

Like in previous years, FFaIR took place between the months of June and July, with a week off for the Fourth of July. The experiment was run daily, with new participants each new week. However, unlike previous years, the remote nature of the experiment allowed for larger attendance each week (see Appendix B), from around 10 participants each week to around 20. Additionally, because the participants were spread out across the country, the daily start time of FFaIR was later than usual to accommodate those in the more western time zones. Below is the list of the four weeks of FFaIR and the daily schedule (Table 1), which was altered after the first week and therefore is different from the schedule seen in the Operations Plan, due to the overwhelming success of our new forecast product, the Maximum Rainfall and Timing Product (MRTP).

Week 1 : June 15-19, 2020
Week 2 : June 22-26, 2020
Week 3 : July 6-10, 2020
Week 4 : July 13-17, 2020

The experiment was hosted via Google Meet and utilized most of the Google Suite platforms available to NOAA Employees. This included creating a shared folder that all the participants had access to with information such as a document listing the important links the participants would need daily (such as the FFaIR Website), attendance sheets, and topography maps. Communication between the participants was encouraged, not only through the virtual meeting, but also within the Meet Chat and a living "blog" document in the shared folder.

Due to connectivity issues and large lag times, the previous methods used to show the experimental data for FFaIR, such as AWIPS2 and NMAP, were not possible. To overcome this obstacle, a website was developed to display experimental model and ensemble data and other experimental products. Operational products and guidance were viewed via various NWS and non-NWS websites, with the FFaIR Website[4] providing links to these websites for quick access during the forecasting exercises. Another barrier was the inability to use WPC drawing tools to create our forecast products. To combat this issue, the FFaIR team worked with members of WPC's Development Testbed Branch to create a drawing tool via the web[5]. This utilized geolocation information so that verification could still be done on the forecast products.

[4]FFaIR Website: https://origin.wpc.ncep.noaa.gov/hmt/ffair2020/FFaIR_Webpage.php
[5] Drawing Tool Website: https://origin.wpc.ncep.noaa.gov/hmt/ffair2020/ERO/ero.html#

Table 1: *The weekly schedule for 2020 FFaIR Experiment.*

Time	Monday	Tuesday	Wednesday	Thursday	Friday
13z-14z Situational Awareness		On your own	On your own	On your own	On your own
1330z-1430z Greetings and Orientation	Call-in				
14z-1445z Forecast Exercise Day 1 ERO		Call-in	Call-in	Call-in	Call-in
1445z-16z Forecast Exercise Day 1 ERO	Call-in				
16z – 1730z Subjective Verification	Call-in	Call-in	Call-in	Call-in	Call-in
1730-1830z Lunch					
1830z-20z Forecast Exercise MRTP	Call-in		Call-in		Call-in
1830z-1930z Experiment Presentations		Call-in 2 presentations		Call-in 1 presentation	
1930z-2030z Optional MRTP		Call-in		Call-in	

2.2 Forecasting Activities

This year, the forecasting activities for FFaIR consisted of two products, a Day 1 Excessive Rainfall Outlook (ERO) and a new experimental product developed by the FFaIR team called the Maximum Rainfall and Timing Product (MRTP). The Day 1 ERO was the morning forecasting exercise and was valid from 16 UTC the current day to 12 UTC the following day. This valid time period matches the 16 UTC update for the Day 1 operational ERO issued by WPC. Similar to the operational WPC Day 1 ERO, the FFaIR ERO was issued for the continental US (CONUS) and identifies where there is a potential for flash flooding within 40 km of a point with probabilistic risks: 5%-10% (Marginal), 10%-20% (Slight), 20%-50% (Moderate), and >50% (High). As in previous years, the forecast discussion leading up to the collaborative issuance of the Day 1 ERO was led by a WPC Forecaster and focused on looking at

the experimental guidance. Throughout the forecast process, open discussion between participants was encouraged and more often than not participants were more than eager to share their thoughts and what they noticed from the guidance.

Normally, as part of the process for creating the collaborative ERO, the participants would each draw their own ERO on a provided printed map (see the front page of the 2019 FFaIR Final Report for an example of this). These would be used to guide the overall discussion for what the Experimental Day 1 ERO should look like. Because of the virtual nature of the experiment this method of drawing individual EROs was not feasible and thus the use of Google Slides was adapted so participants, in real-time, could "draw" an ERO and the WPC forecaster and the rest of the participants could view it. Although the drawing tool on Google Slides presented some challenges, the system worked well overall and an example of some of the EROs drawn for July 10, 2020, along with the official experiment ERO, can be seen in Fig. 1.

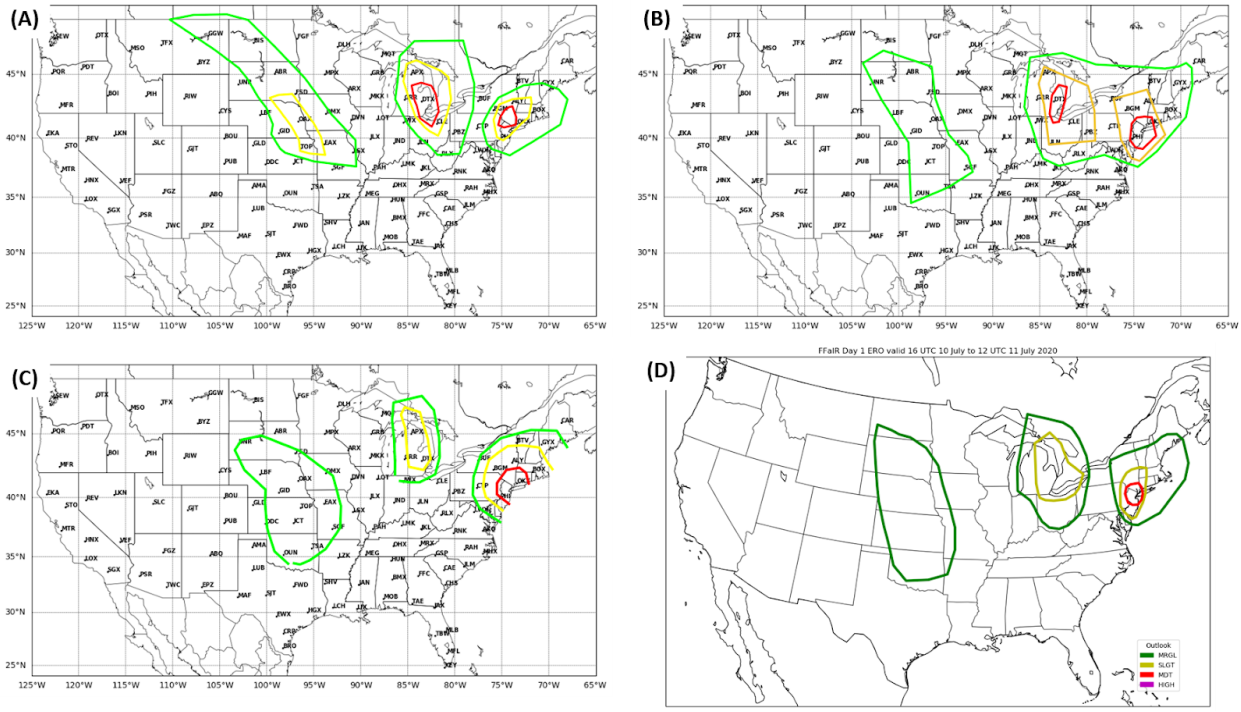

Figure 1: *(A)-(C) Example of the Day 1 EROs drawn by the participants using Google Slides. (D) The final, collaborative FFaIR ERO created and drawn by the WPC Forecaster. All valid 16 UTC 10 July to 12 UTC 11 July, 2020.*

The second forecast product, the MRTP, was issued in the afternoon. It was modeled to be a product that conveys information similar to the information provided from WPC's Mesoscale Precipitation Discussion (MPD). The goal of the MRTP was to help assess the ability of models and ensembles to identify the threat for, including the ingredients of, heavy rainfall and flooding. Following the general idea of the MPD, the MRTP was focused on a sub-regional to regional scale and was valid for no longer than six hours. In addition to this, the forecasters

were required to list what models/ensembles/guidance they used for the forecast. The forecasting exercise would begin with a general overview of the current conditions and the synoptic setup by the WPC forecaster, followed by a discussion of what the various experimental guidance was predicting. During this time, like with the ERO forecast exercise, open discussion was encouraged. The area of focus and the valid time for the product would then be agreed upon. The valid start time of the MRTP could be as early as 21 UTC and as late as 06 UTC. The participants would then break off on their own to create their individual MRTPs and complete their MRTP Survey. Despite working individually to draw these, everyone remained on the virtual call and often continued to discuss what they were thinking among each other. An example of what the MRTP drawing tool looked like can be seen in Fig. 2. The participants were required to contour where they thought the total observed precipitation for the time period would be one inch or greater. They were also required to identify the location where they thought the greatest total rainfall would occur. If the participants wanted to, they could also draw 2, 3, and 4 inch precipitation contours, but this was not required nor verified. Despite this being optional, many participants chose to do this extra work of forecasting extreme precipitation.

Depending on the day, the participants were assigned a model or ensemble to focus their MRTP discussion and forecast on. This was done with the idea that by assigning guidance to all of the participants, the FFaIR team could gather feedback about all the models and ensembles analyzed in the 2020 FFaIR. Participants, however, did not need to stick with their assigned guidance and a survey was used each day to record this information and other feedback about the forecasts. The questions on the survey can be found in Section 4.5 and the locations of the MRTPs can be found in Appendix C.

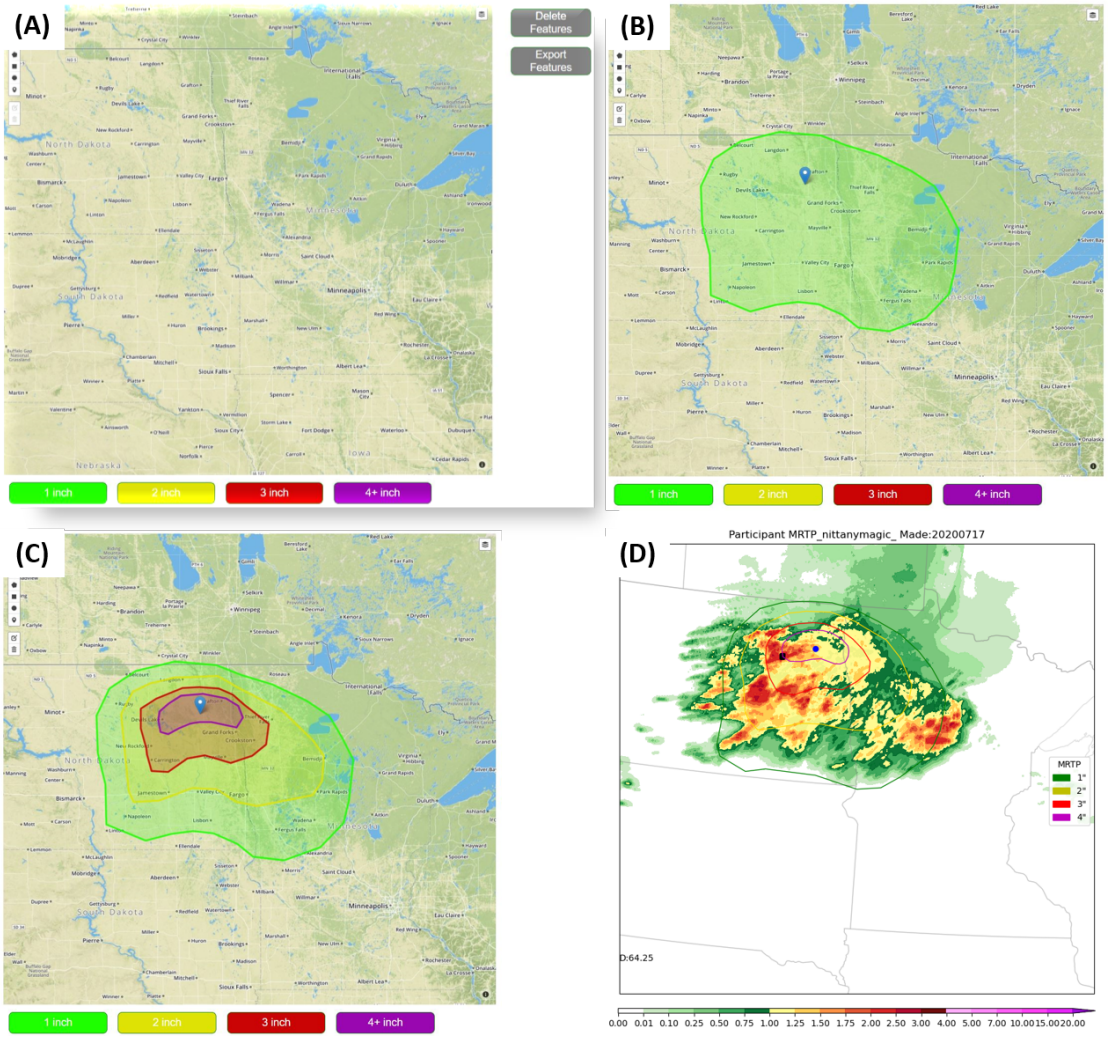

Figure 2: *(A)-(C) Example of the process for drawing a MRTP via the web drawing tool. (D) The verification image for the MRTP drawn, verified using MRMS-GC QPE. The green, yellow, red, and purple contours/fill represent 1, 2, 3, and 4 inches of QPF. (C) and (D) the inverted tear drop is the forecasted location of maximum rainfall. (D) the blue dot is the forecast max, the black dot is the observed max and the observed rainfall is underlaid. All valid 21 UTC 17 July to 03 UTC 18 July, 2020.*

2.3 Guidance and Verification

A diverse collection of guidance was evaluated in FFaIR this year, ranging from global models to machine learning products. Overall, the FFaIR experiment focused on the analysis of eight different configurations of the FV3-SAR model, three ensembles, three first guess Day 1 ERO products, a heavy precipitation object tracker (HPOT) and the National Water Model (NWM) Visualization Services version 1.2. A detailed description of the guidance can be found in the 2020 FFaIR Operations Plan.

The participants used these products and tools to guide them in their experimental forecasts and openly discussed their utility both formally and informally. A variety of analysis methods were used to assess the tools, including subjective verification by the participants. A list of what was subjectively verified can be seen in Table 2. In general, the same guidance was also objectively verified, with the exception of the HPOT and the NWM Visualization Services version 1.2.

Table 2: *Summary of the experimental model and ensemble guidance, experimental products, and experimental forecasts that were subjectively evaluated. The number of scores recorded for each guidance evaluated is in parenthesis. For the Day 1 QPF models there are two numbers, one for the evaluation of the 00z run and the other for the 12z run. Some guidance was only subjectively evaluated at the end of the week and did not receive numerical scores. *Number of days available to be evaluated rather than number of scores.*

PRODUCT EVALUATED	Day 1 QPF	6h QPF SARs only	Ensemble 6h QPF	CSU First Guess Day 1 ERO	FFaIR Forecasts	Not Scored
	GFSv16 (130)/(96)	**EMC FV3-SAR** (19)*	**HREFv3 LPM** (307)	**GEFS** (327)	**Day 1 ERO** (325)	**HPOT**
	HRRRv4 (269)/(247)	**EMC FV3-SARX** (19)*	**HRRRE LPM** (307)	**NSSL1** (327)	**MRTPs** (325)	**NWM Viewer v1.2**
	EMC FV3-SAR (288)/(232)	**EMC FV3-SARDA** (16)*	**SSEF LPM** (222)	**NSSL1** (321)		**WPC PQPF**
	EMC FV3-SARX (293)/(189)	**GSL FV3-SAR1** (11)*	**SSEF SAM** (129)			**NBMv4 PQPF**
	EMC FV3-SARDA (232)/(147)	**GSL FV3-SAR2** (8)*	**SSEF SAM-LPM** (130)			
	FV3-SAR OU (272)	**GSL FV3-SAR3** (7)*	**NBMv4 DMO-PM** (252)			
		GSL FV3-SAR4 (7)*	**NBMv4 QMD-PM** (288)			
		SSEF Member (20)*				

2.3.1 Subjective Verification

It is common practice for NWS Testbeds to use subjective verification methods to help assess the utility of experimental products and tools from the eyes of the user. Although this is useful, the results are strongly controlled by biases such as the order in which the products are shown, asking too general of a question, forecaster/modeler prejudice (i.e. always uses one model over another) or object of concern (i.e. cares more about one area of the country than another), and general unwillingness of evaluators to say anything is "perfect". Some of these biases were noted during the analysis for 2019 FFaIR and facilitators attempted to mitigate these issues for this year's subjective evaluation.

Various strategies were employed to potentially lessen biases like order effects or favorite model effects. For instance, when evaluating model performance for 24 h QPF, the names of the models were hidden, the participants were told to only score the model's utility for a given region (though they could write comments about the whole CONUS) and each day the order in which the models were shown differed. The daily region in which the subjective verification for the 24 h QPF was focused on can be found in Appendix C. In other instances, direct comparison questions were asked, such as "Do you feel guidance A did better, worse, or about the same as guidance B?" For additional value, each question allowed for the participants to provide written feedback about the guidance and the facilitators encouraged open discussion among the participants about the verification while taking notes on what was discussed. The hope is that steps like these will help provide more useful feedback to our partners about their products.

Table 2 brings to light another obstacle in the evaluation process, the availability of experimental data. As can be seen above, the number of scores or times available to be subjectively analyzed differed among the guidance. This was particularly noticeable when looking at the number of times that the eight different configurations of the FV3-SAR were available to be compared to one another during the subjective verification comparison question for 6 h QPF; explained further in Section 4.1.2.

2.3.2 Objective Verification

Along with the subjective evaluation, daily and bulk (beyond the scheduled experiment days) objective verification was done via a variety of methods. The number of days included in the bulk verification varied depending on the model or ensemble and the initialization time. The number of days each model/ensemble was available for bulk verification can be found in Table 3. It is important to note that these numbers differ from the number of times a model/ensemble was available for subjective evaluation.

Table 3: *Summary of the number of days included in the bulk verification for each model or ensemble during the 2020 FFaIR Experiment. Days were limited to only the days in which FFaIR was in session, from June 16 to July 18, 2020.*

Model/Ensemble	6 h QPF Init. 00z	24 h QPF Init. 00z	24 h QPF Init. 12z
EMC FV3-SAR	19	19	17
EMC FV3-SARX	19	19	16
EMC FV3-SARDA	15	15	14
GSL FV3-SAR1	17	16	16
GSL FV3-SAR2	13	12	13
GSL FV3-SAR3	13	12	18
GSL FV3-SAR4	13	12	19
SSEF FV3-SAR-OU		20	
HRRRv4		18	20
GFSv16		14	
HREFv3 (LPM mean)	18 (19)	19 (18)	15 (15)
HRRRE	19	18	10
SSEF	19	20	
SSEF - SAM (SAM-LPM)	13 (13)	13 (12)	
NBM - PM means		18	10

The majority of the objective evaluation was done using the the Method for Object-Based Diagnostic Evaluation (MODE), which is part of the Model Evaluation Tools (MET) package[6] (Bullock 2016); see Appendix D for the MODE configuration. MODE was used to evaluate model and ensemble QPF at various thresholds (i.e. 1 inch, 2 inches, etc.) by comparing the forecast rainfall object to the observed rainfall object. An example MODE can be seen in Fig. 3 for the 1 inch threshold. In Fig. 3C the comparison between the forecast rainfall object and the observed rainfall object can be seen, where the Multi-Radar Multi-Sensor Gauge Corrected (MRMS-GC) data is contoured and the HRRRv4 QPF is shaded. The different colors correspond to the Cluster ID, which is determined by the spatial proximity of rainfall objects. For instance, Cluster 3 (blue) extends from central KS to central MS. In addition to identifying and clustering precipitation objects visually, MODE provides statistics comparing the forecast objects to the observed objects including centroid distance, angle, and intersection area. These can be used to determine information like the area ratio (forecast area/observed area) and intersect ratio (intersect area/observed area).

[6] Information on MET can be found at the Developmental Testbed Center website: https://dtcenter.org/community-code/model-evaluation-tools-met.

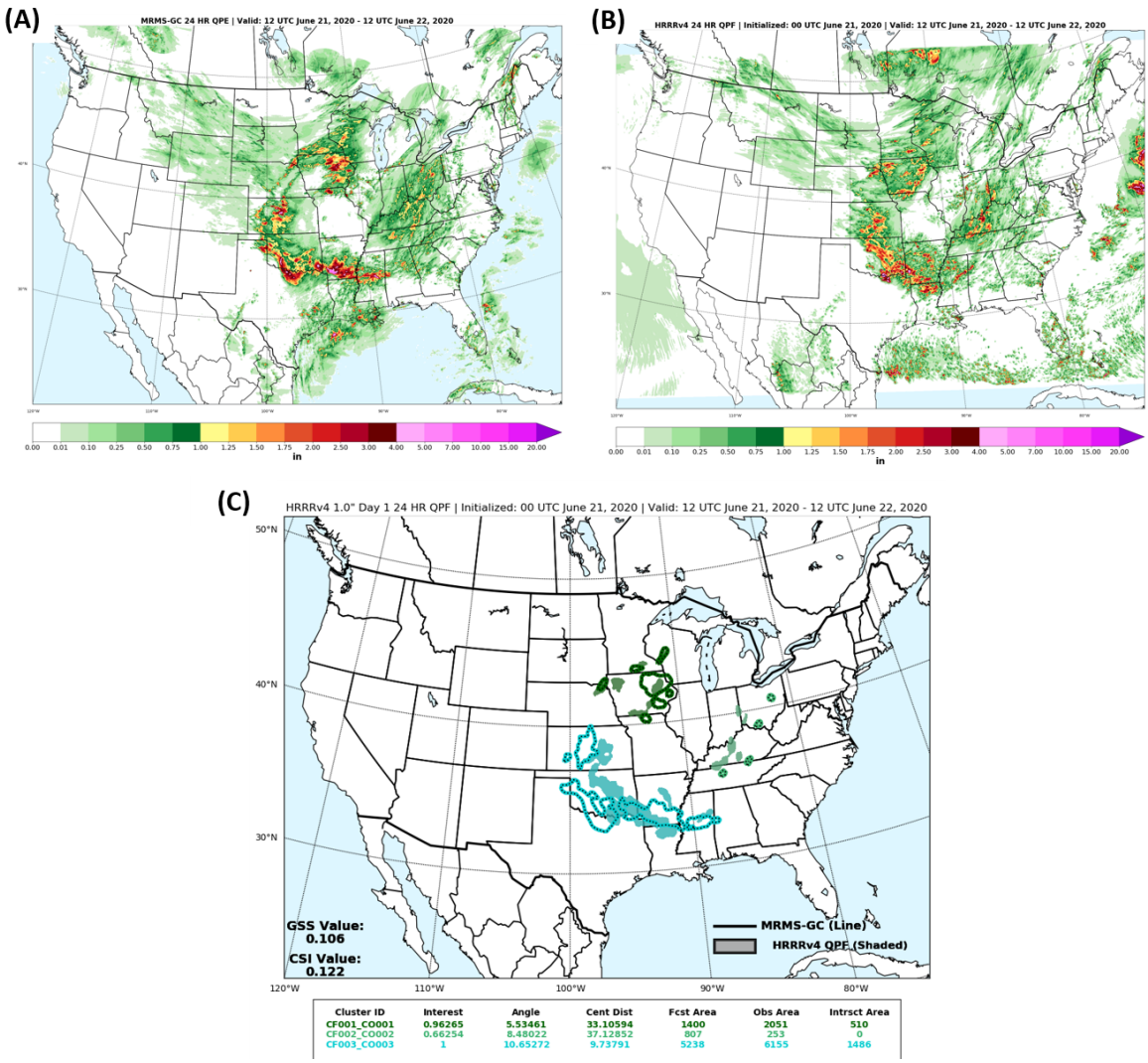

Figure 3: *(A) the 24 h MRMS-GC QPE, (B) HRRRv4 24 h QPF, and (C) MODE analysis for 1 inch 24h QPF valid 12 UTC 21 June to 12 UTC 22 June, 2020.*

Additionally, grid-to-grid verification was performed using the output from statistics calculated using the MODE package. The MODE methodology is similar to the Grid-Stat tool in the MET package (MET 2018), which uses a filter scale to remove objects less than a specified size. Daily contingency tables were created from this output. The statistics from each daily contingency table were then accumulated and used to create performance diagrams for the entire duration of FFaIR (see Fig. 4).

Another verification tool used was the Practically Perfect analysis, which was developed to evaluate probabilistic forecasts, such as the ERO. It is derived from local storm reports, exceedances of average recurrence intervals, flash flood guidance, and USGS gauge data. A radius of influence is then applied to each data point and smoothed to create what the probabilistic forecast would be given perfect knowledge of the verification. Figure 5 provides an

example of the Practically Perfect analysis compared to the FFaIR Day 1 ERO forecast for July 10, 2020. In this case, the Practically Perfect is suggesting that a moderate risk (20%-50% chance of flash flooding within 40km of a point) was warranted from from the DE/MD/VA peninsula to southeastern NY, approaching the need for a high risk. Meanwhile over MI, only a marginal risk (5%-10%) should have been forecasted. This methodology also allows for evaluation of more traditional statistics such as factional coverage and bulk AuROC[7].

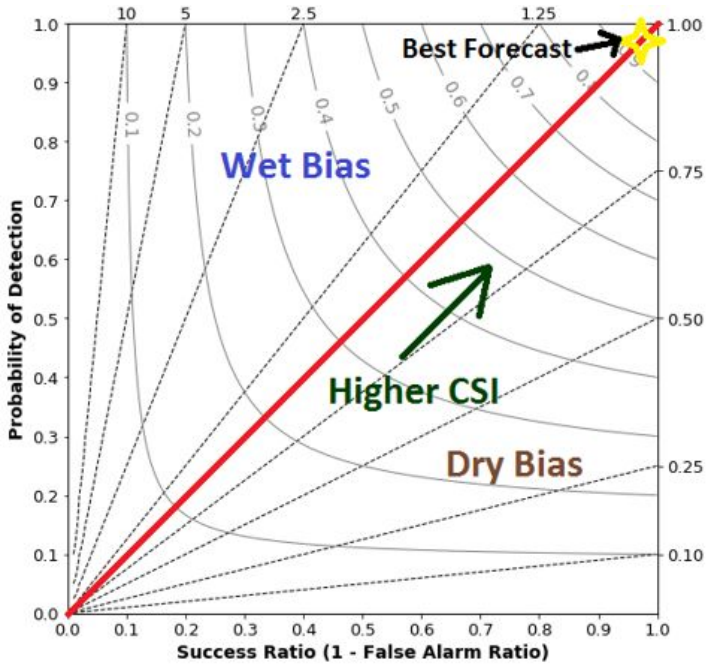

Figure 4: *Example of a Roebber Performance Diagram. Y-axis is the probability of detection, x-axis shows the success ratio (1 - false alarm ratio), dashed diagonal lines represent the bias, and curved solid lines represent CSI.*

[7] 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.

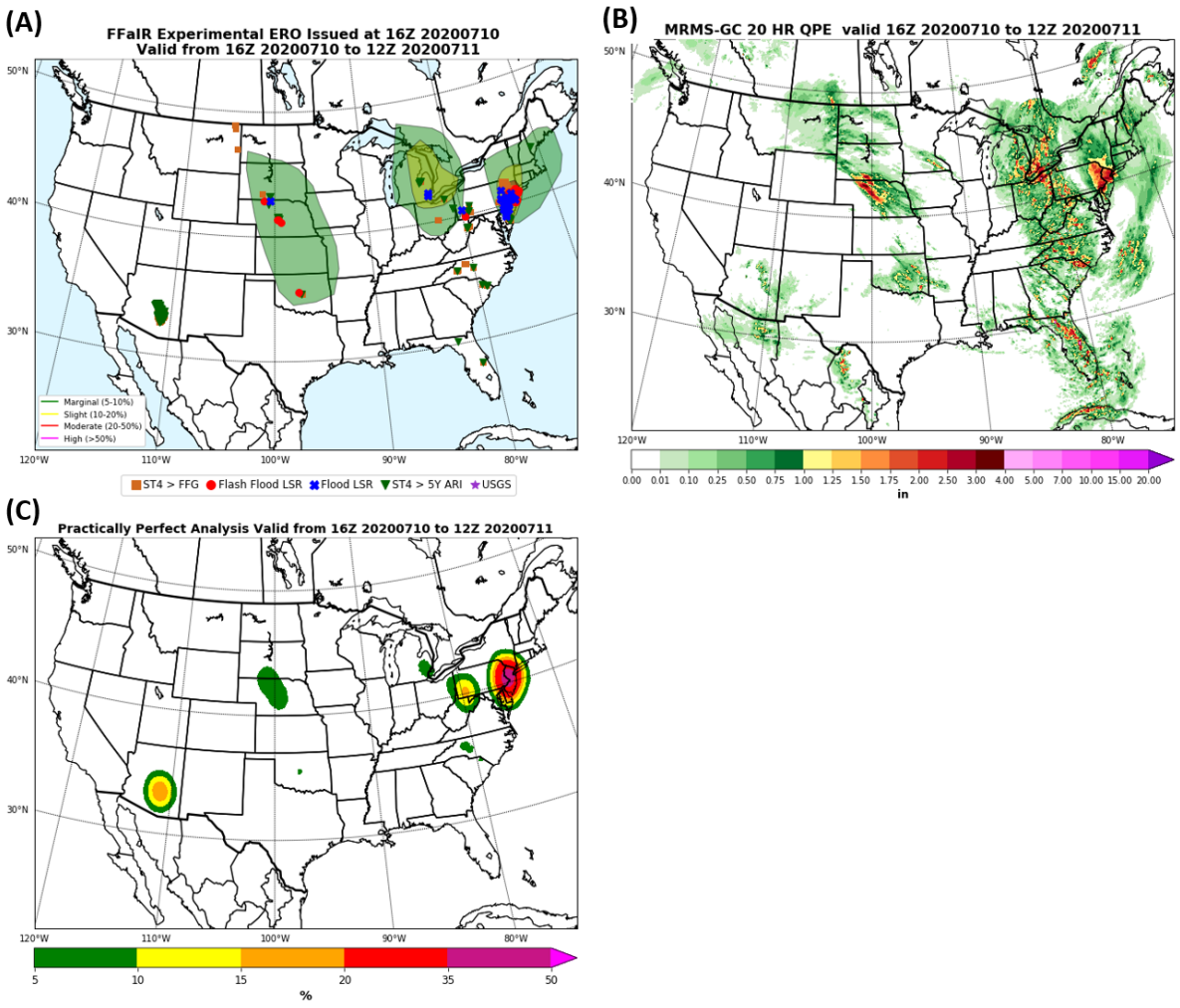

Figure 5: *Example of a subjective verification image for the CSU MLP First Guess forecasts and FFaIR Day 1 ERO. (A) FFaIR Day 1 ERO overlaid with: QPE > FFG (orange box), QPE > ARI (green triangle), flash flood LSRs (red dots), flood LSRs (blue x), and USGS gauge reports (purple star). (B) 20 h MRMS-GC QPE and (C)Practically Perfect analysis. Valid 16 UTC 10 July to 12 UTC 11 July, 2020.*

3. Meteorological Highlights During the Experiment

Although a full description of notable events and areas of interest during the forecasting exercises can be found in Appendix C, a brief summary of the overall meteorological conditions throughout the experiment and some higher impact events will be discussed here. FFaIR began this year with a synoptic setup atypical for the summer, a deep trough in the Northwest and a closed low over the Mid-Atlantic. As can be seen in Fig. 6, this pattern persisted the whole first week, resulting in a nearly continuous risk for excessive rainfall from the Carolina's to West Virginia. In fact, on June 16, 2020 portions of North Carolina saw over 5 inches of rain (Fig. 6D). The Day 1 ERO's forecasted for Week 1 can be seen in Fig. 7.

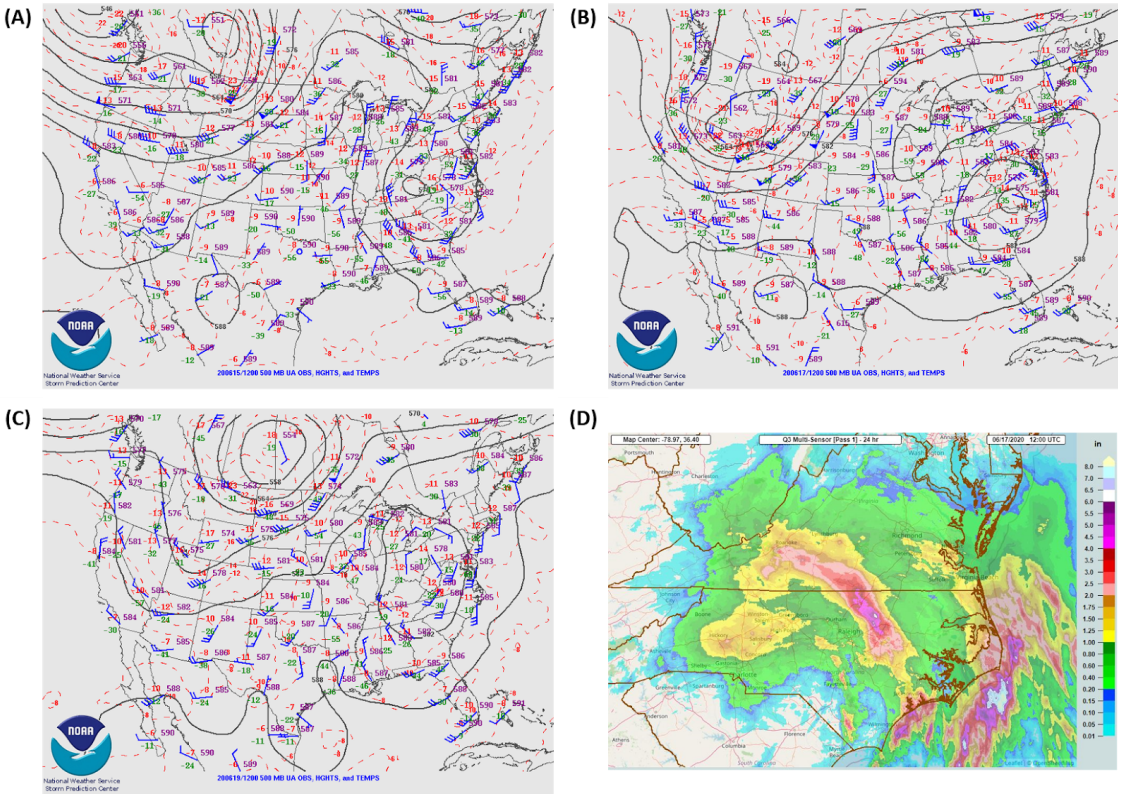

Figure 6: *SPC 12z 500mb analysis for (A) 15 June, (B) 17 June, and (C) 19 June, 2020. (D) 24 h MRMS-GC QPE valid 17 June 2020.*

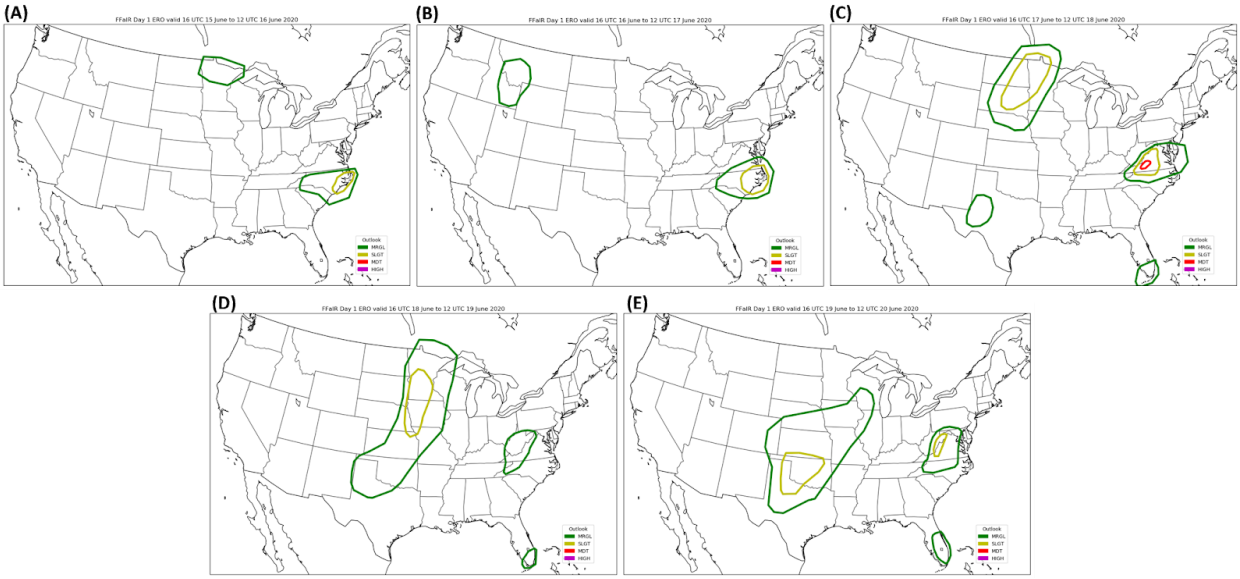

Figure 7: *Experimental EROs issued by the Week 1 participants, valid: (A) 16 UTC 15 June to 12 UTC 16 June, 2020, (B) 16 UTC 16 June to 12 UTC 17 June, 2020, (C) 16 UTC 17 June to 12 UTC 18 June, 2020, (D) 16 UTC 18 June to 12 UTC 19 June, 2020, and (E) 16 UTC 19 June to 12 UTC 20 June, 2020.*

After the first week, ridging began to move in across the southern US and remained in place for the rest of the experiment. A composite of the 500mb heights for each week of FFaIR can be seen in Fig. 8. The focus of Week 2 was across the western Gulf Coast, due to the nearly continuous presence of a slow moving Mesoscale Convective Vortex (MCV). Meanwhile, Week 3 included a tropical system impacting the Northeast and during Week 4 a new daily rainfall record occurred in Illinois. Figures 9-11 show the FFaIR Experimental Day 1 ERO forecast for weeks 2-4.

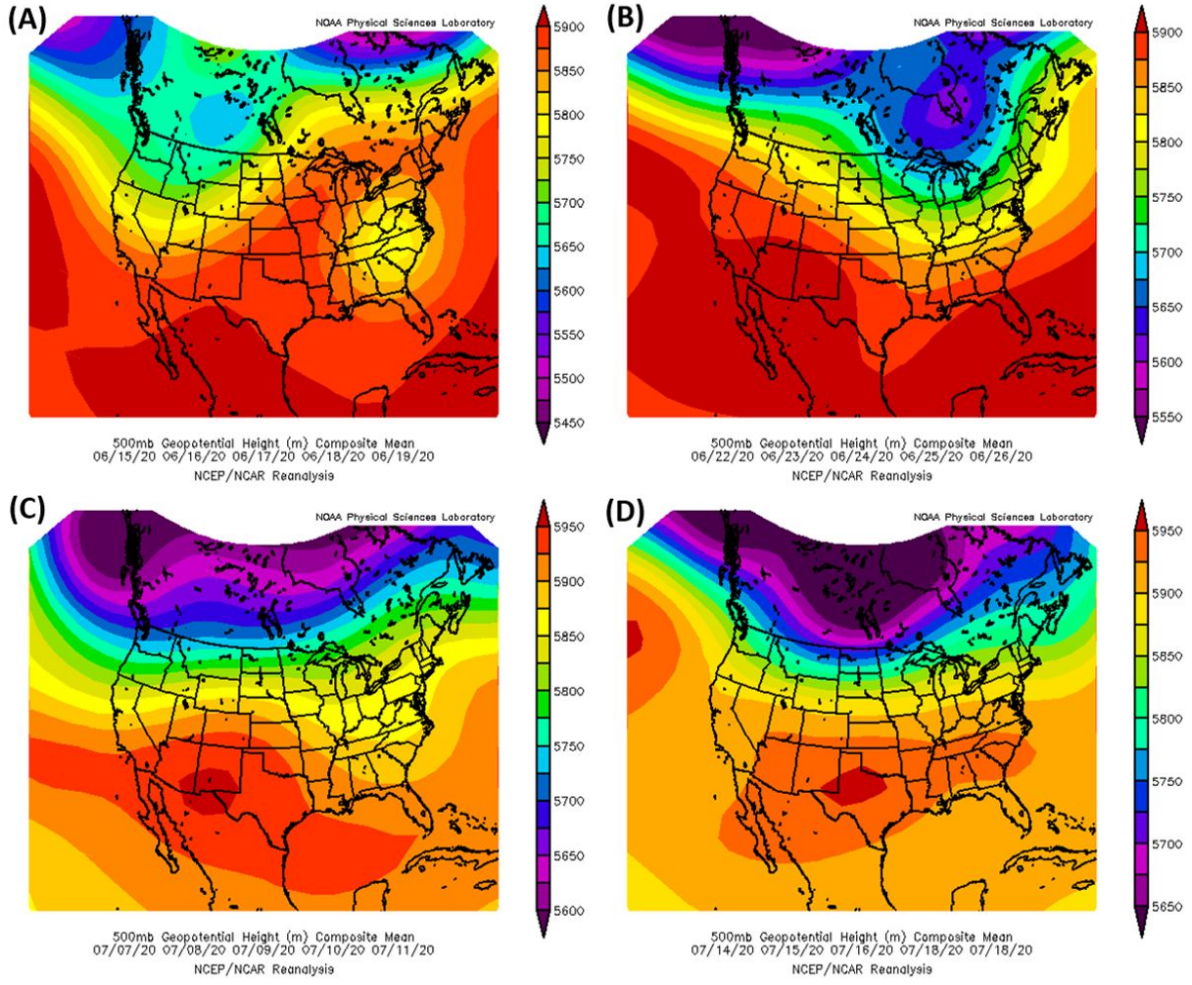

Figure 8: *500 mb mean geopotential height composites for (A) Week 1, (B) Week 2, (C) Week 3, and (D) Week 4 of the 2020 FFaIR Experiment. Composite images were generated from the NCEP/NCAR Reanalysis provided by NOAA/ESRL/Physical Sciences Division[8] (PSL 2014).*

[8] PSL, 2014; https://www.esrl.noaa.gov/psd/data/composites/day/

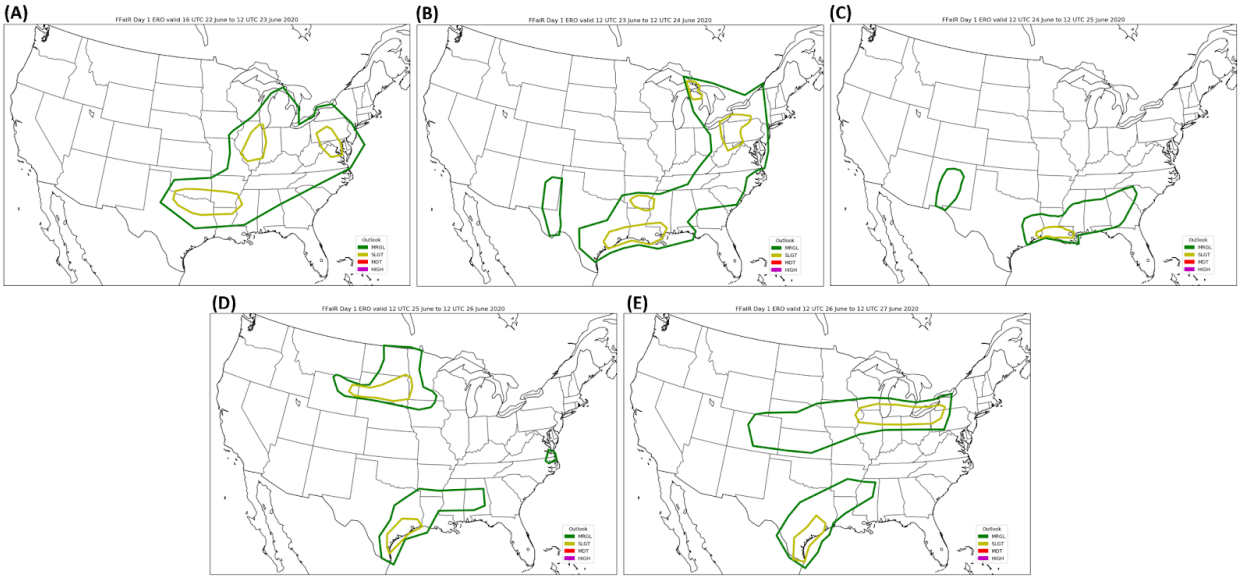

Figure 9: *Experimental EROs issued by the Week 2 participants, valid: (A) 16 UTC 22 June to 12 UTC 23 June, 2020, (B) 16 UTC 23 June to 12 UTC 24 June, 2020, (C) 16 UTC 24 June to 12 UTC 25 June, 2020, (D) 16 UTC 25 June to 12 UTC 26 June, 2020, and (E) 16 UTC 26 June to 12 UTC 27 June, 2020.*

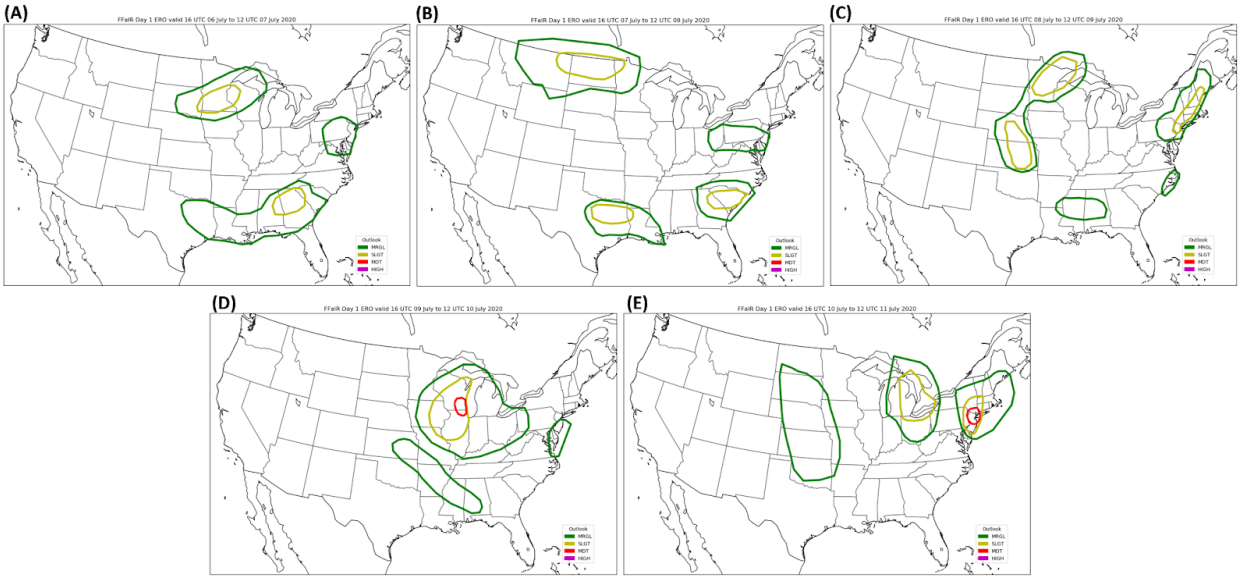

Figure 10: *Experimental EROs issued by the Week 3 participants, valid: (A) 16 UTC 06 July to 12 UTC 07 July, 2020, (B) 16 UTC 07 July to 12 UTC 08 July, 2020, (C) 16 UTC 08 July to 12 UTC 09 July, 2020, (D) 16 UTC 09 July to 12 UTC 10 July, 2020, and (E) 16 UTC 10 July to 12 UTC 11 July, 2020.*

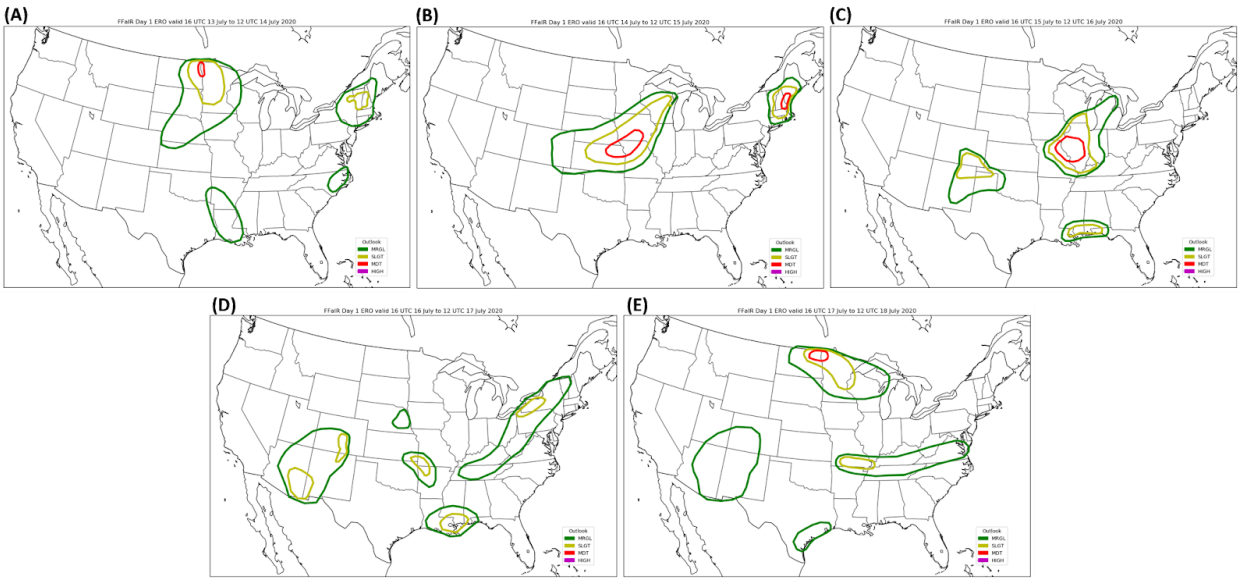

Figure 11: *Experimental EROs issued by the Week 4 participants, valid: (A) 16 UTC 13 July to 12 UTC 14 July, 2020, (B) 16 UTC 14 July to 12 UTC 15 July, 2020, (C) 16 UTC 15 July to 12 UTC 16 July, 2020, (D) 16 UTC 16 July to 12 UTC 17 July, 2020, and (E) 16 UTC 17 July to 12 UTC 18 July, 2020.*

As stated, the heavy rainfall risk for Week 2, was generally associated with a Mesoscale Convective Vortex (MCV), which developed along the OK/TX border early in the week and slowly moved southward and became nearly stationary near the Texas Coast. This drove many flash flooding events from New Orleans to Houston, including a highly impactful event in Katy, TX, a suburb of Houston, on June 24-25, 2020. As can be seen in Fig. 12, more than 6 inches of rain fell across the region in six hours, leading to widespread flooding, numerous stranded vehicles, and streams and creeks overflowing their banks and closing roads.

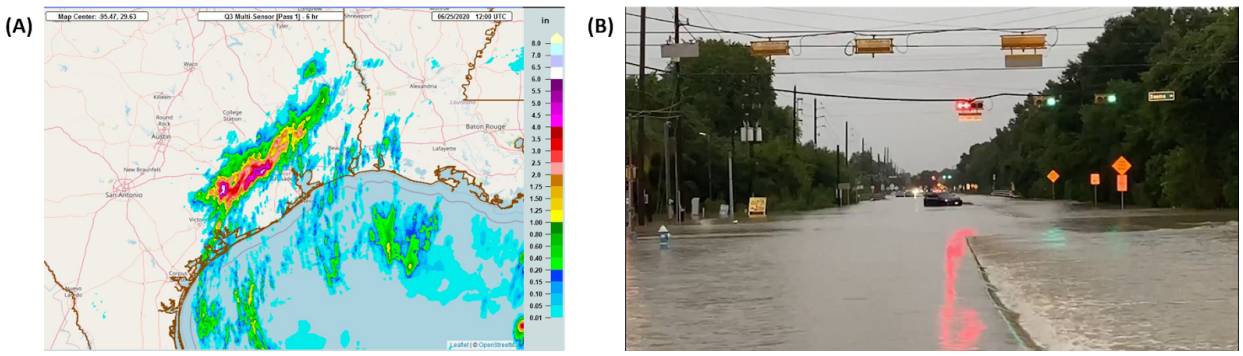

Figure 12: *(A) 6 h MRMS-GC QPE valid 12 UTC 25 June 2020. (B) Picture of flooding in Katy, TX courtesy of KHOU 11 News*[9] *(2020).*

[9] KHOU 11 article can be found at: https://www.khou.com/article/news/local/several-people-stranded-in-cars-due-to-flash-flooding-near-katy/285-e2160aad-32c0-4b35-b329-8b608c11eebe

By the end of Week 2, the threat from the MCV had begun to diminish and the focus was shifted to a shortwave making its way across the Upper Midwest. Even though the initial threat of the storms associated with the shortwave resulted in severe weather, the main threat shifted to heavy rainfall by the evening of June 26. The change in weather threat was due to the flow weakening, high PWAT values (>1.75 inch) and convection shifting from progressive to backbuilding. The event resulted in a couple of MPDs issued by WPC and is recapped by a storm summary written by the Chicago's Weather Forecast Office (WFO) about the event; see Fig. 13.

Week 3 of the experiment brought multiple rounds of heavy rainfall across Wisconsin, resulting in widespread flooding across the state on July 9-10, 2020. The event was driven by a strong shortwave over IA and weak southerly flow that continuously filtered moist air into the region (see Fig. 14A). When all was said and done most of southern Wisconsin had received more than 2 inches of rain and parts of the region saw closed roads and highways due to flood waters; refer to Fig. 14.

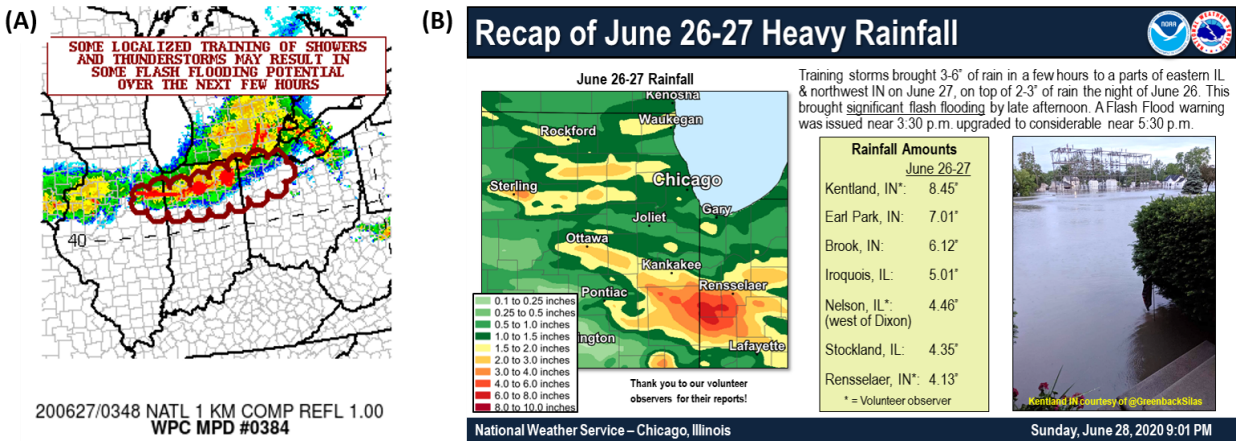

Figure 13: *(A) WPC MPD #0384 issued 0348 UTC 27 June 2020 and (B) the information graphic issued by Chicago's WFO[10] for the heavy rainfall that occurred June 26-27, 2020.*

[10] WFO LOT, 2020; *https://www.weather.gov/lot/26june2020*

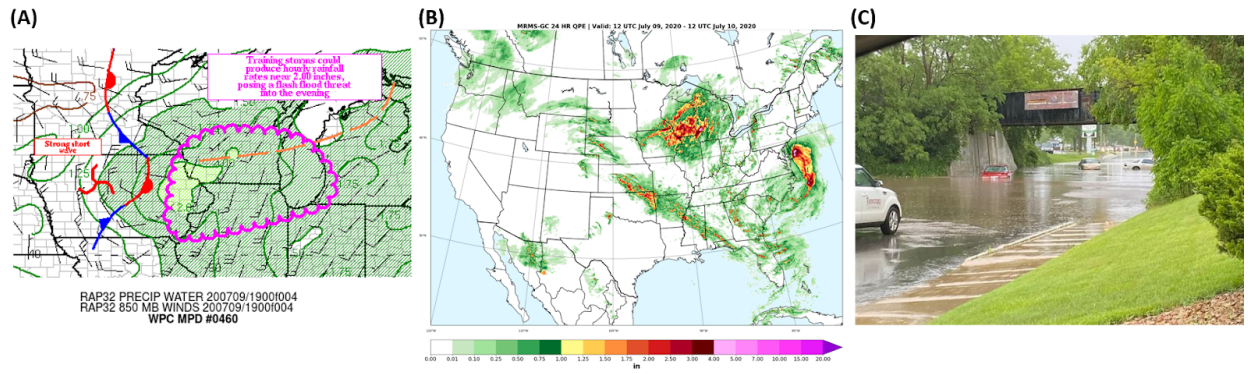

Figure 14: *(A) WPC MPD #0460 issued 2130 UTC 09 July 2020. (B) 24 h MRMS-GC QPE valid 12 UTC 09 July to 12 UTC 10 July, 2020. (C) Image of flooding on Interstate 41 in Fond du Lac, WI on 10 July 2020; courtesy of the FDL Reporter[11] (2020).*

Another center of focus during Week 3 extended from northern VA to the Northeast, first with a high impact event from Washington DC to NYC which was then followed by Tropical Storm Fay (hereafter TS Fay). The first event occurred from July 6 to July 7, 2020, with some of the greatest effects of the event felt in Philadelphia. Although the likelihood of excessive rainfall and flooding was appreciated by the FFaIR participants, the severity of the event came as somewhat of a shock. As can be seen in Fig. 10A and Fig. 15, the FFaIR Day 1 ERO had a Marginal Risk for the southern portion of the Northeast Corridor, but practically perfect analysis suggests a Moderate Risk was needed. In fact, the flooding was so prolific in the Philadelphia Metro area that a Flash Flood Emergency was issued (Fig. 16A). A highly moist and unstable environment was coupled with weak steering flow, resulting in rain rates exceeding 3 inches an hour in some areas. According to the Mount Holly WFO event review (https://www.weather.gov/phi/EventReview20200706), upwards of 6 inches of rain accumulated in less than 2 hours. Farther south, around Baltimore MD, storm totals exceeded 7 inches. All in all the event led to widespread water rescues and road closures (Fig. 16).

[11] Taima, K 2020; Article found at: https://www.fdlreporter.com/story/news/2020/06/10/fond-du-lac-interstate-41-closed-between-u-s-151-and-military-road/5338575002/

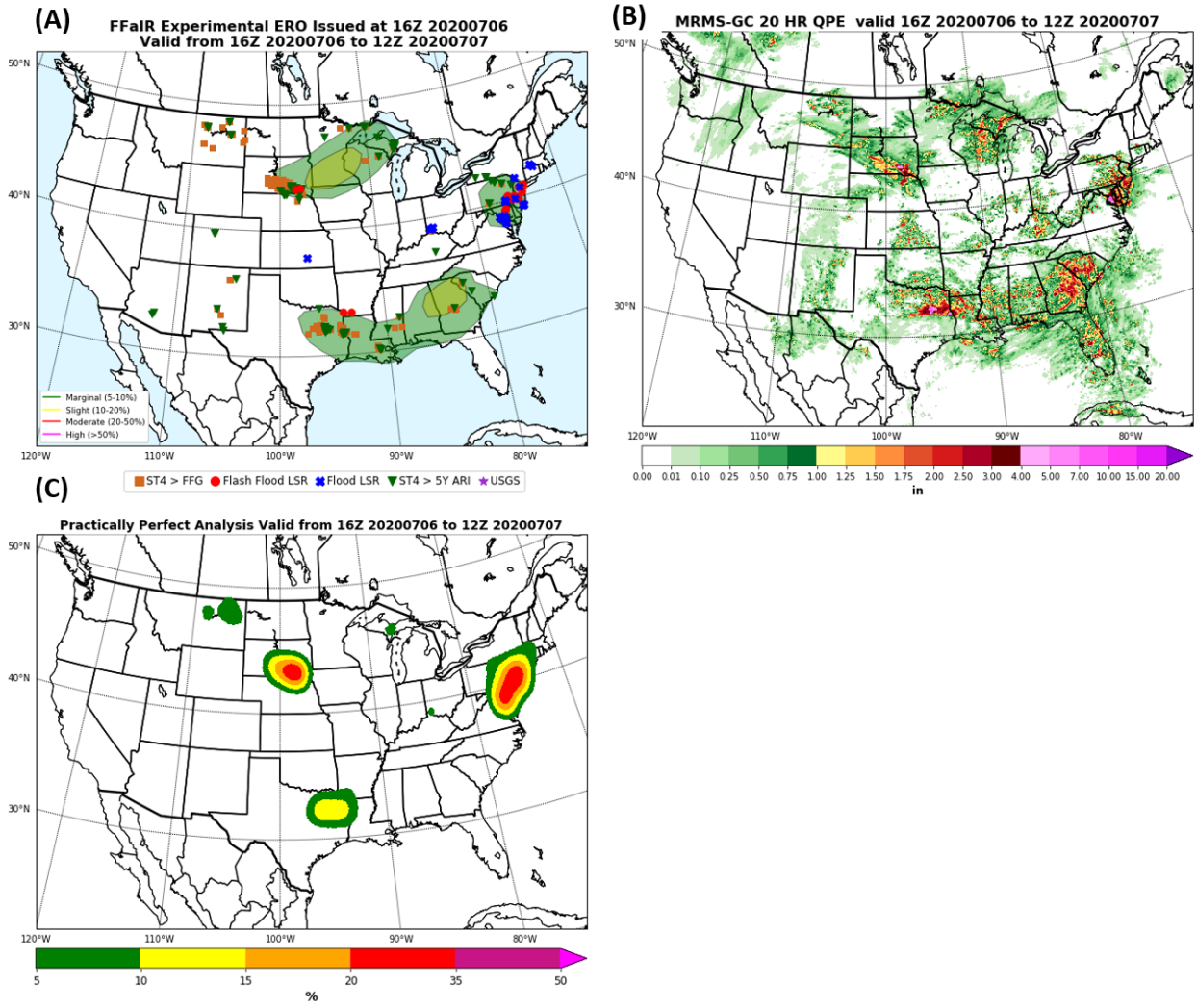

Figure 15: *Valid 16 UTC 06 July to 12 UTC 07 July, 2020: (A) FFaIR Day 1 ERO overlaid with: QPE > FFG (orange box), QPE > ARI (green triangle), and flash flood LSRs (red dot), flood LSRs (blue x), and USGS gauge reports (purple star), (B) 21 hour MRMS-GC QPE and (C) Practically Perfect Analysis.*

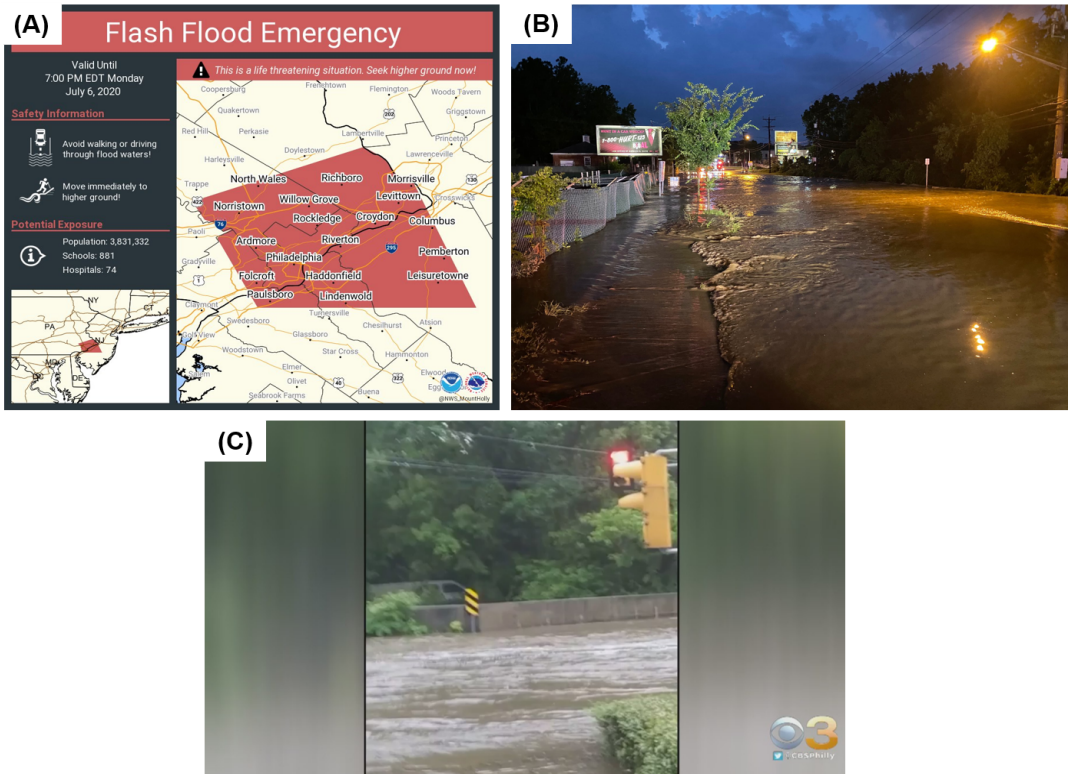

Figure 16: *(A) The Flash Flood Emergency issued for the Philadelphia region, valid 6 July, 2020. (B) and (C) images of flooding across the region on July 6th; courtesy of CBS Local News Philadelphia (2020) and the Twitter account of the Prince George's County Fire/EMS Department*[12].

Then, after an already wet week in the northeast, TS Fay set its sights for the region. As can be seen in Fig. 17A, TS Fay developed from a MCV over GA that moved into the Atlantic Ocean and took on tropical characteristics, moving northward towards New Jersey. The system was mainly a heavy rainfall threat, with WPC forecasting widespread 2+ inches of rainfall from NJ/eastern PA, northward (Fig. 17B) and isolated totals exceeding 7 inches. Figure 18 shows the 24 hour QPF forecasts, valid from 12 UTC July 10 to 12 UTC July 11, from some of the experimental deterministic guidance evaluated in FFaIR. Interestingly, despite pending landfall (around 00 UTC July 11), the June 10, the models initialized at 00z had some difficulty predicting the axis of heavy rainfall. For instance, along the northeastern side of the storm some models struggled with the impact TS Fay would have across Long Island. They also had difficulty with the outer bands of the system and how far north they would progress into NH and ME, with nearly all the models predicting multiple bands of an inch or more of rainfall across the area. That said, aside from GFSv16 and the EMC FV3-SAR, there was agreement among the models that the focus of heavy rainfall would be central NJ and overall had a handle of the northern extent of the main swath of rainfall from TS Fay.

[12] Prince George's County Fire/EMS Department Twitter (2020): https://twitter.com/PGFDNews/status/1280341872783888384

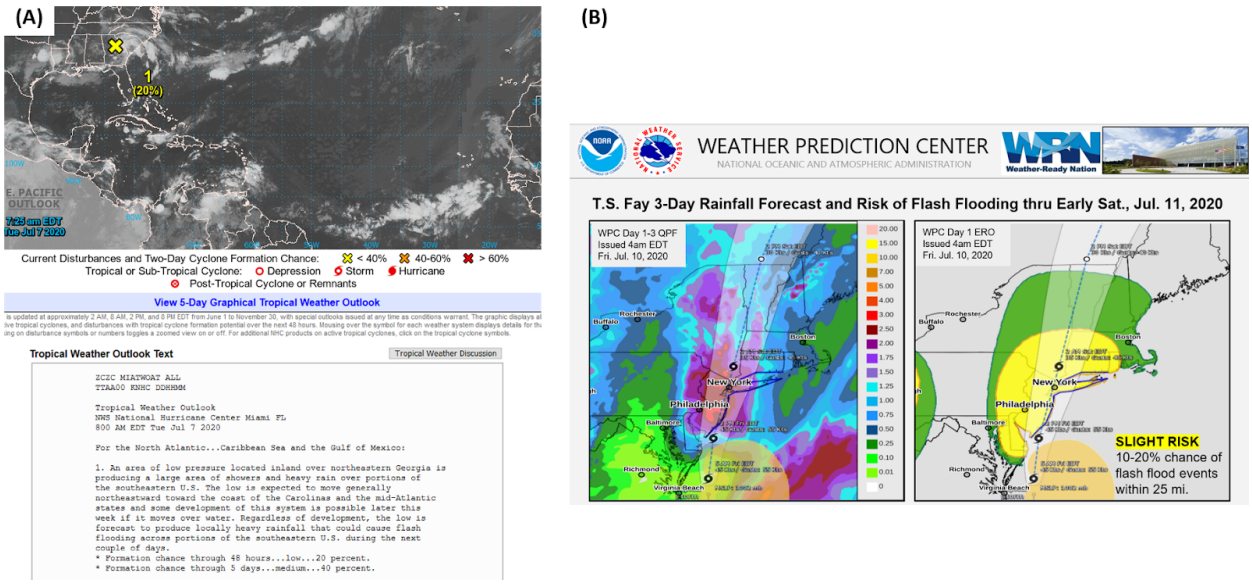

Figure 17: *(A) NHC Tropical Weather Outlook issued at 12 UTC 7 July 2020 for what would become Tropical Storm Fay. (B) The WPC graphic for the rainfall and flash flooding threat associated with Tropical Storm Fay issued 08 UTC 10 July 2020.*

The Northeast remained under a flooding risk going into Week 4 of FFaIR. An upper level vorticity maximum and climatologically high precipitable water values led to the development of heavy showers in the morning hours of July 14, 2020 across Vermont and New Hampshire. By mid-morning the rainfall had resulted in flooding across the states, with water inundating a hospital's ER and some of the operating rooms (WMUR9 Article 2020), leading to the temporary closure of the hospital. Heavy rainfall and flooding continued for the remainder of the day in NH and ME.

The second major event occurred the following day across central IL, where the record for 24 hour rainfall for the month of July in Peoria IL was broken (National Climatic Data Center 2020). The July 15, 2020 event was driven by a weakening MCV combined with a frontal system draped across the region; see Fig. 19A for WPC's MPD for the area. This setup resulted in central IL receiving rainfall first from the storms that initiated along the warm front followed by those associated with the cold front; see Fig. 19B-C. This "double whammy" gave rise to a prolonged period of rainfall rates up to 1.5 inches an hour. At Peoria International Airport 5.19 inches fell, beating out the old July daily rainfall record of 4.09 inches that was set in 1895. In addition to becoming the new daily record for July, the event was the second highest 24 hour rainfall total for any month by less than a half inch; 5.52 inches still holds the record, set on May 18, 1927. Figure 20 shows the 24 hour MRMS-GC rainfall for the region along with some photos of the flooding that occurred.

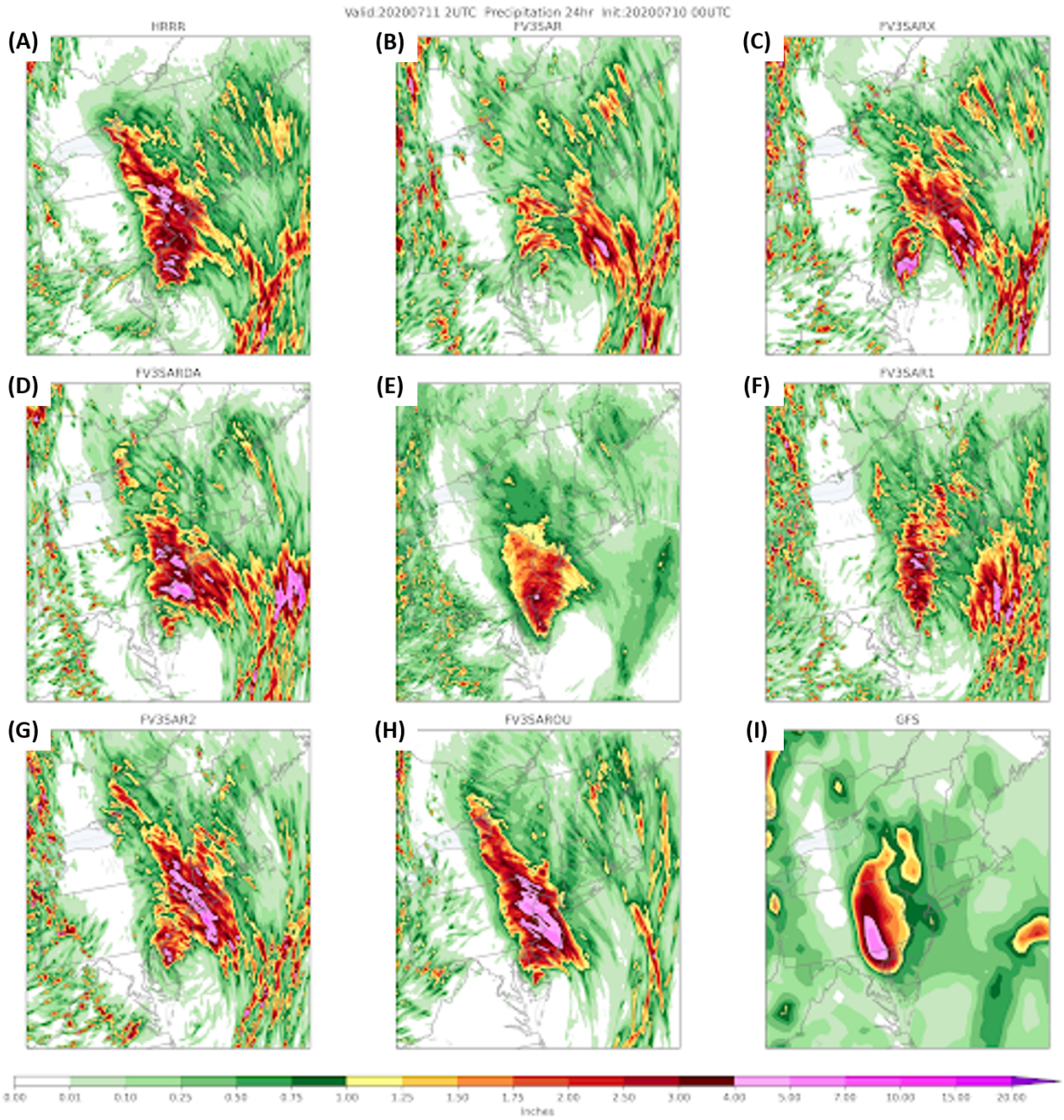

Figure 18: *The 24 h QPF forecasts from the (A) HRRRv4, (B) EMC FV3-SAR, (C) EMC FV3-SARX, (D) EMC FV3-SARDA, (F) GSL FV3-SAR1, (G) GSL FV3-SAR2, (H) SSEF contorl member (FV3-SAROU), and (G) GFSv16 and (E) the MRMS-GC QPE valid from 12 UTC July 10 to 12 UTC July 11.*

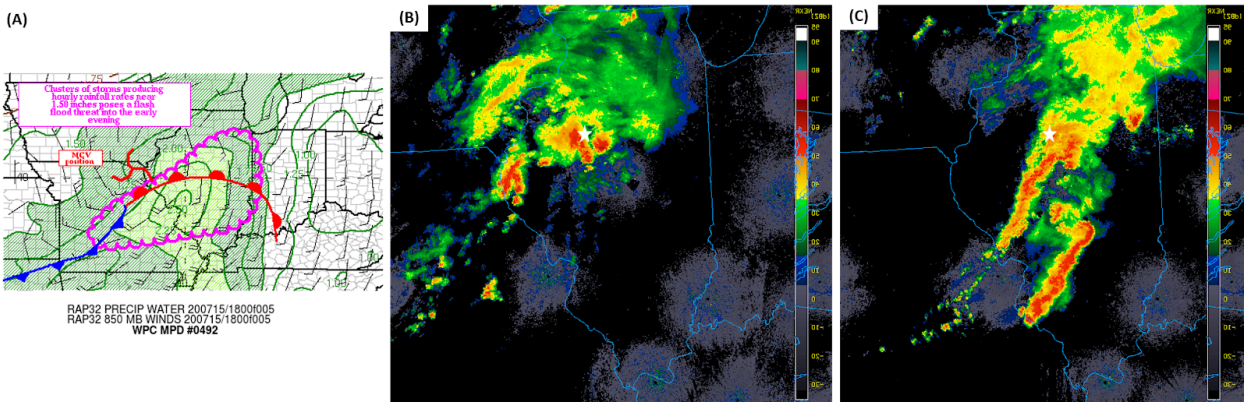

Figure 19: *(A) WPC MPD #0492 issued 2015 UTC 15 July 2020. Composite radar analysis valid at (B) 19 UTC and (C) 23 UTC on 15 July 2020; courtesy of UCAR Image Archive (2020).*

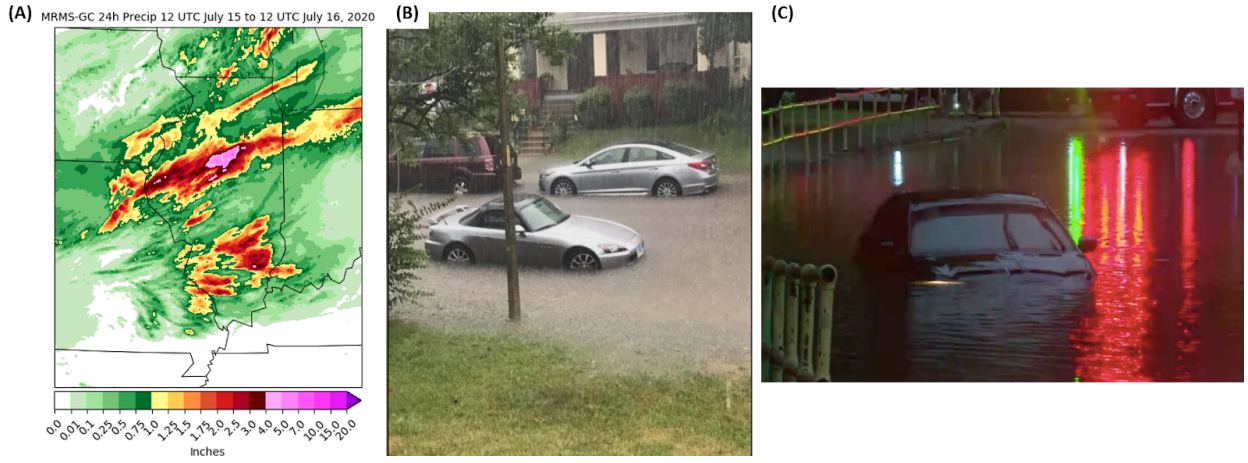

Figure 20: *(A) 24 h MRMS-GC QPE valid 12 UTC 15 July to 12 UTC 16 July 2020. (B) and (C) pictures of flooding across the Peoria region due to the record 24 h rainfall. Images from NBC 25 News[13].*

Because of the rather obvious impact and uniqueness of the event across central IL, the FFaIR participants decided that they wanted to do a MRTP-like forecast for the region, in the form of a Nowcast[14]. The Nowcast was valid from 21 UTC July 15 to 00 UTC July 16 and like the MRTP the participants were only required to contour where they thought 1 inch of rain would fall in that time period and place a point where they thought the maximum rainfall would occur. However, as can be inferred in Fig. 21, most of the participants contoured for other thresholds as well (2, 3, and 4 inches). Additionally, a survey was quickly created, gathering information about the process the participants used creating their forecast. The survey included asking if they relied more on current radar and satellite data or model and ensemble guidance.

[13] Links to the articles where the images were taken: Swathwood (2020): https://week.com/2020/07/15/flash-floods-strand-many-peoria-drivers-prompting-rescues/
Guerrero (2020): https://week.com/2020/07/15/flooding-hits-several-cities-across-central-illinois/
[14] The NWS specifies a Nowcast as zero to three hours, though up to six hours may be used by some.

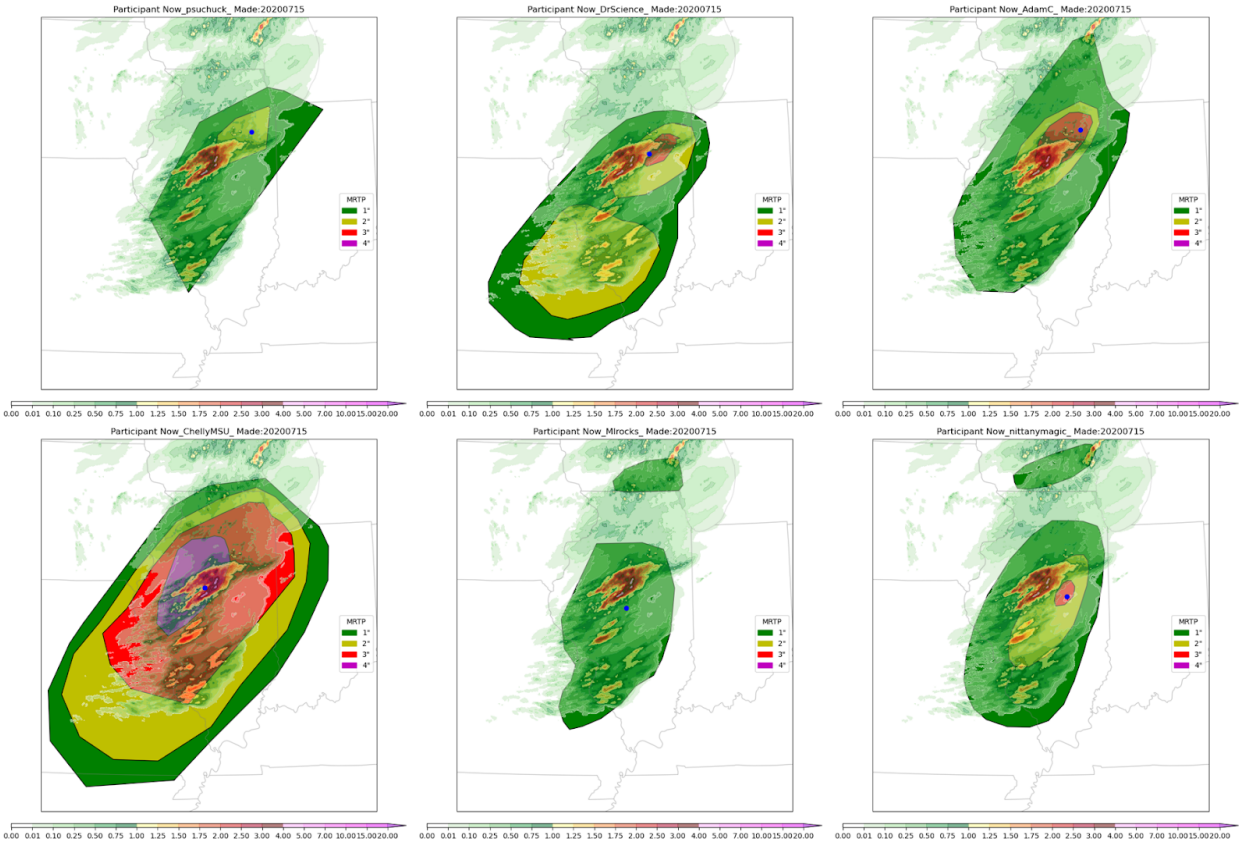

Figure 21: *Six examples of Nowcast forecasts issued by Week 4 participants overlaid with the observed MRMS-GC QPE valid from 21 UTC 15 July to 00 UTC 16 July 2020. The Nowcast contours are for 1 (green), 2 (yellow), 3 (red), and 4 (purple) inches. The blue dot indicates where the participant forecasted the highest rainfall totals to occur.*

4. Results

The following sections will discuss both subjective and objective verification for the experimental models, ensembles and products evaluated in the 2020 FFaIR experiment. Along with a summary of the results, overall recommendations about the experimental guidances' readiness for operations will be given. The MRTP forecasts, methodology and general feedback about the product will also be examined.

4.1 Deterministic Guidance

As already mentioned, a large focus of FFaIR this year was the evaluation of multiple configurations for the FV3-SAR model; the configurations for the FV3-SARs provided by EMC and GSL can be seen in Table 4. The GSL configurations varied in both initial conditions (IC) and Lateral Boundary Conditions (LBC) and in the "Horizontal Order" or Hord used in the dynamics suite. In past years of the experiment, the participants have been asked to assess the 24 h QPF from all modeling being evaluated. However the subjective verification portion of the

experiment can get overwhelming and therefore, as not to overload the participants, the number of FV3-SAR configurations individually assessed during the subjective evaluation of the 24 h QPF was limited. Instead a comparison question of 6 h QPF was asked to help examine how well the various FV3-SAR configurations performed compared to one another; the setup of this question is discussed below.

(A)

Model	ICs	LBCs	Microphysics	Land Surface Scheme	PBL	Radiation	Referred to as:
FV3-SAR	GFSv15	GFSv15	GFDL	NOAH	EDMF	RRTMG	SAR
FV3-SARX	GFSv15	GFSv15	Thompson	NOAH	MYNN	RRTMG	SARX
FV3-SARDA	GFSv15	GFSv15	Thompson	NOAH	MYNN	RRTMG	SARDA

(B)

Model	ICs	LBCs	Physics Suite	Dynamics Option Suite	Referred to as:
FV3-SAR1	GFS	GFS	RAPv5+/HRRRv4+	Hord 5	SAR1
FV3-SAR2	GFS	GFS	RAPv5+/HRRRv4+	Hord 6	SAR2
FV3-SAR3	HRRRv4	RAPv5	RAPv5+/HRRRv4+	Hord 5	SAR3
FV3-SAR4	HRRRv4	RAPv5	RAPv5+/HRRRv4+	Hord 6	SAR4

Table 4: *The FV3-SAR configurations for the FV3-SARs evaluated in the 2020 FFaIR Experiment from (A) EMC and (B) GSL.*

4.1.1 Evaluation of 24 h QPF

Evaluation of model 24 h QPF was done by showing the participants the model forecast next to the 24 h MRMS-GC QPE for the CONUS. The name of the model was unknown and each day the order in which the models were shown differed. Additionally, each day the region over which they were to evaluate the model changed. Table C.1 in Appendix C lists the region that was assessed each day, this region generally coincided with the location of the previous day's MRTP. As shown in Table 2 , the HRRRv4, GFSv16, EMC FV3-SAR (hereafter SAR), EMC FV3-SARX (hereafter SARX), EMC FV3-SARDA (hereafter SARDA), and FV3-SAR OU (hereafter SAROU) were all evaluated subjectively for their 24 h precipitation forecast. The 24 h QPF from both the 00z and 12z runs were scored when available; the SAROU did not have a 12z run. Importantly, after the completion of FFaIR, the CAPS-OU team found a bug in the SAROU related to the coupling of the MYNN and the radiation schemes. Therefore, discussion of the results from the SAROU will be limited.

Figure 22 and 23 show the percentage of the time that each model received a score ranging from 1 to 10, along with the average score for the experiment. The HRRRv4 and SARX, for both the 00z and 12z initialization cycles, were identified subjectively as the best performing models. For the 00z runs the SARX had a marginally higher average score than the HRRRv4;

6.14 vs 6.06, while the HRRRv4 slightly outperformed the SARX during the evaluation of the 12z runs; 6.4 to 6.31. Interestingly, the 12z cycle of the GFSv16 had an average score of 6.3, which was on par with the SARX. This average was a large jump from the 00z average of 4.39 and seemingly indicates that the later cycle was overwhelmingly preferred to the earlier run. However, it is important to note that the 12z run of the GFSv16 only was scored 96 times, which is 34 scores less than the total for the 00z cycle and around 50 scores less than the next lowest score amount for the other 12z runs; refer to Table 2. Lastly, the SAROU was generally considered the worst performer, even compared to the GFSv16, which is a global model rather than a CAM.

The SAROU continuously received lower scores due to an apparent systematic inability to develop convection. An example of this can be seen in Fig. 24; for this day, July 6-7, 2020, the focus of the verification was across the Mid-Atlantic. Note that all the models, except the SAROU, had a north/south orientation of the precipitation over the area of interest, with most of the models indicating a heavy rainfall event from Washington DC northward. However the SAROU predicted two bands of convection across central NJ rather than a widespread threat. Furthermore, expanding to the CONUS, it can be seen that precipitation was observed over most of the CONUS but the SAROU failed to develop convection across the CONUS, most notably over the Northern Plains and the central US. Throughout the experiment "misses" like this were seen for nearly every 24 hour period, resulting in the SAROU being drastically different from the other models. This low bias in precipitation was constantly noted by the participants both verbally and in the written comments. For example, a comment from one of the participants for this case was: "Except for eastern Texas and Minnesota-Canadian border, it was a complete miss. Had most of the precipitation over the mid-Atlantic in the wrong location (off the coast) and a dry bias." The drastic difference between the SAROU configuration and the other FV3-CAMs (e.g. especially the HRRRv4 and the GFSv16) could also be seen in the 24 h QPF performance diagrams at the half inch and inch thresholds. The discovery of the bug in this model helps to explain why this configuration of the FV3-CAM differed so dramatically from the other configurations.

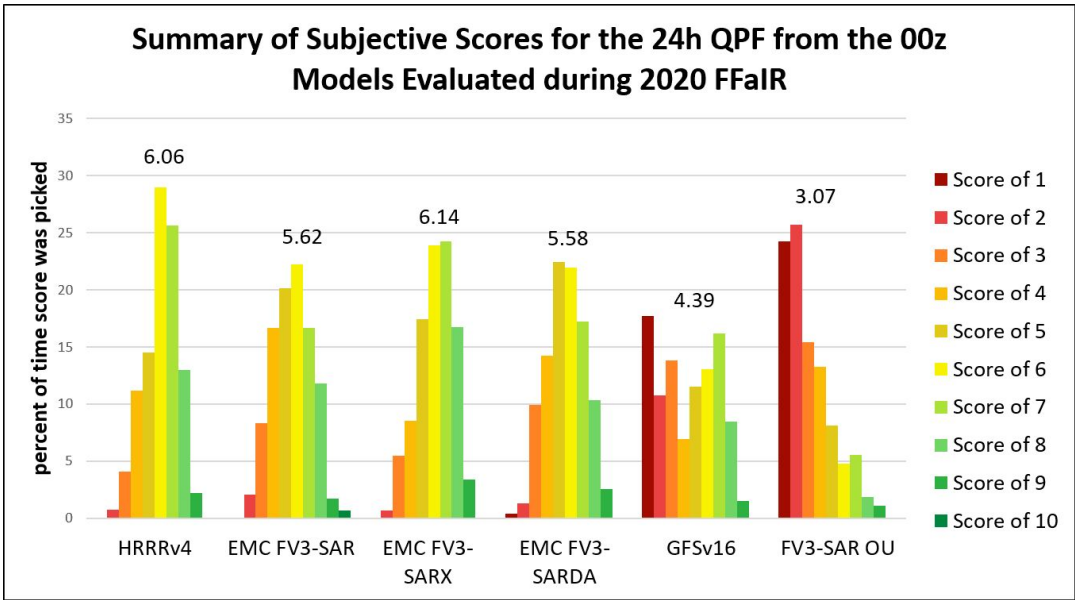

Figure 22: *The percentage of time each deterministic model received a score of 1 through 10 during the course of the experiment during the subjective verification for 24 h QPF for 00z model initialization along with the experimental average plotted above the percentage analysis.*

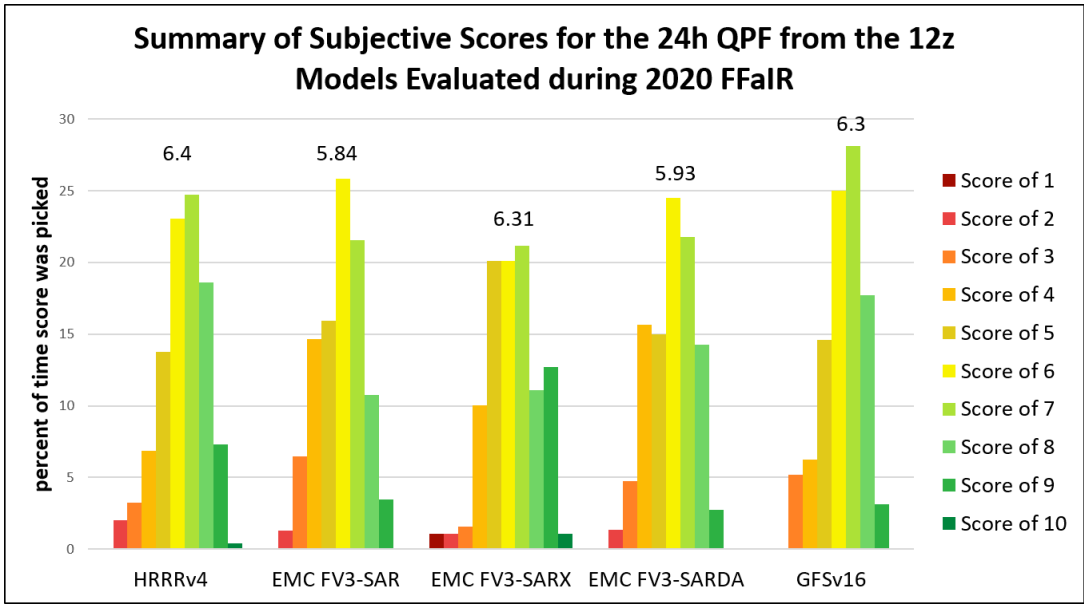

Figure 23: *The percentage of time each deterministic model received a score of 1 through 10 during the course of the experiment during the subjective verification for the 24 h QPF for the 12z model initialization along with the experimental average plotted above the percentage analysis.*

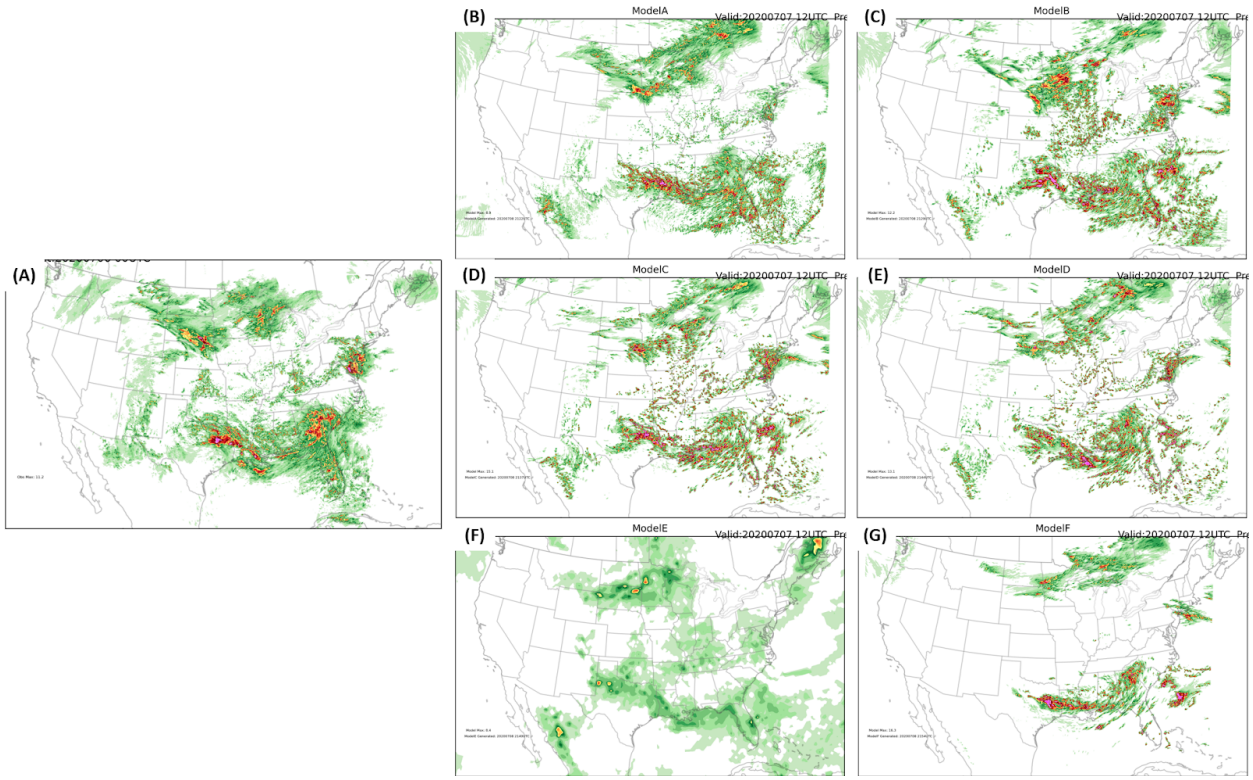

Figure 24: *(A)24 h MRMS-GC QPE and 24 h QPF from (B) HRRRv4, (C) EMC FV3-SAR, (D) EMC FV3-SARX, (E) EMC FV3-SARDA, (F) GFSv16, and (G) SSEF control member (SAROU) valid 12 UTC 06 July to 12 UTC 07 July 2020.*

4.1.2 FV3-CAM Analysis for 6 h QPF

To evaluate the performance of the various FV3-SAR model configurations against each other, the participants were shown the 6 h QPF for each model valid from 18 UTC to 00 UTC, along with the observed precipitation all at once; see Fig. 25. Looking across the CONUS, the participants were instructed to identify which, if any, of the models captured the day's 6 h rainfall totals. The participants were able to choose multiple configurations if they felt two or more performed about the same. There was also an option for the participants to select that none of the configurations performed well and all of the configurations performed about the same. The participants were encouraged to explain why they felt one model did better or worse than another. Unfortunately, due to data flow issues there were only five days in which all 8 of the model configurations were available for analysis. The models that were most likely not to be present for verification were the four configurations of GSL FV3-SAR (hereafter SAR1, SAR2, SAR3, and SAR4).

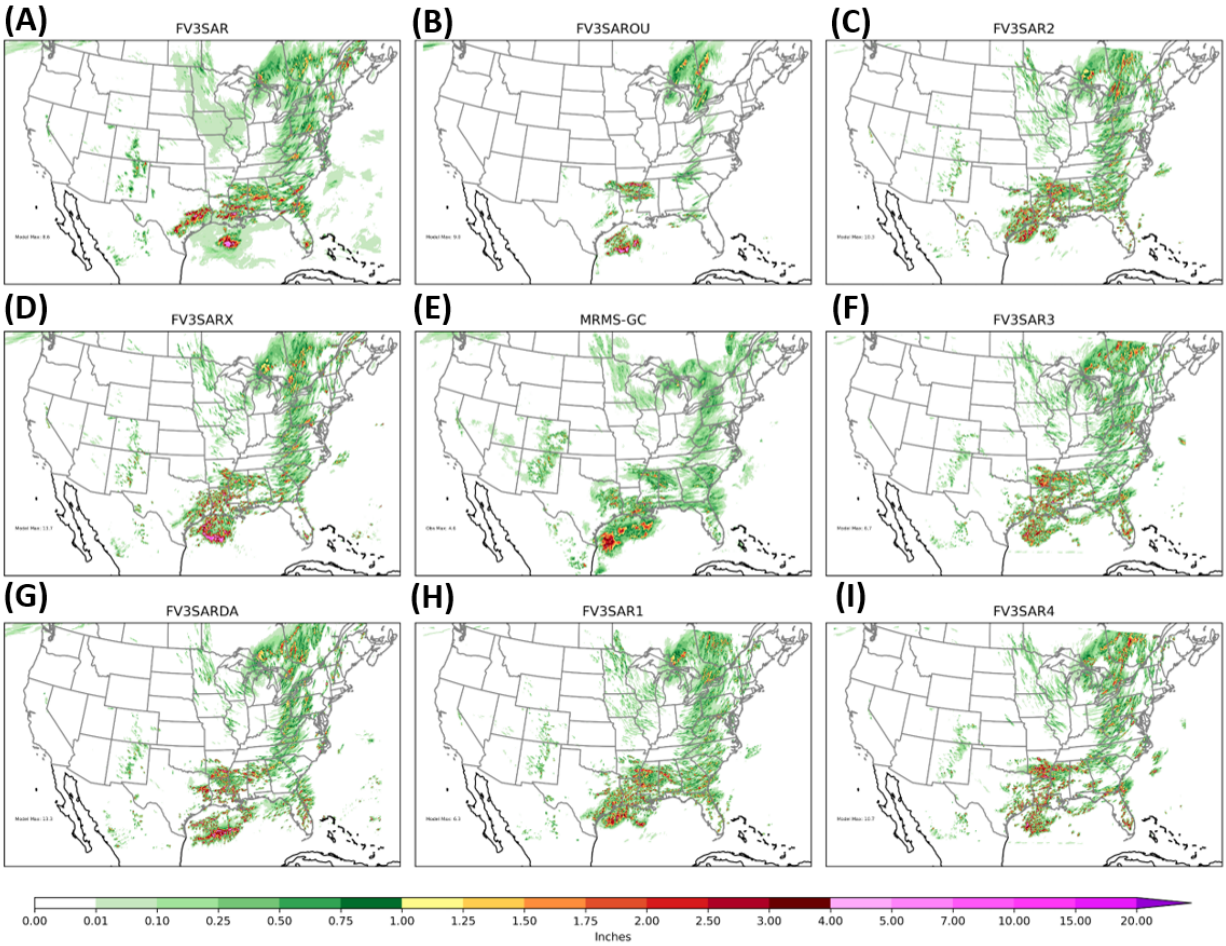

Figure 25: *Example of the FV3-SAR configuration comparison image shown to participants for subjective verification. (E) 6 h MRMS-GC QPE. 6 h QPF from (A) EMC FV3-SAR, (B) SSEF control member (SAROU), (C) GSL FV3-SAR2, (D) EMC FV3-SARX, (F) GSL FV3-SAR3, (G) EMC FV3-SARDA, (H) GSL FV3-SAR1, and (I) GSL FV3-SAR4. Valid 18 UTC 23 June to 00 UTC 24 June 2020.*

The days in which all FV3-CAM configurations were available for comparison were, June 24, June 26, June 27, July 7, and July 8, 2020[15]. During these five days, the SAR3, SAR1 and the SARDA were the most likely to be picked as the "best" forecast while the SAROU was the least likely to be chosen. Figure 26 shows the number of times each FV3-CAM configuration was picked as one of the best CONUS precipitation forecasts along with the daily Critical Success Index (CSI) for the 1 inch QPF threshold; SAROU was not included. As can be seen, the "model of day" chosen by the participants did not always correspond with the highest CSI. For instance, on June 26 the model with the highest CSI was the SARDA but it was only picked once by a participant as the best forecast for the day. Instead the SAR1 was overwhelmingly

[15] Date is the valid end date for the forecast. So June 24 would be valid 18 UTC June 23 to 00 UTC June 24 2020.

favored by the participants. Reviewing the comments from the participants there was no clear reason why this was the case.

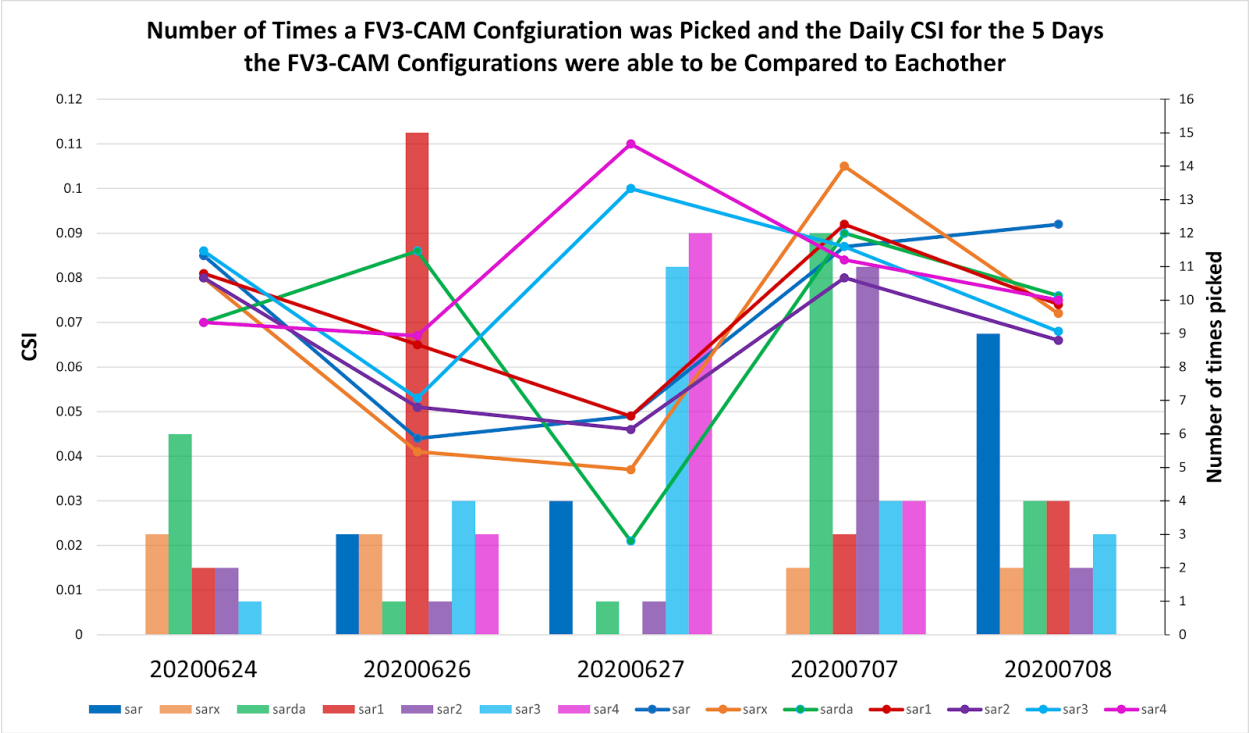

Figure 26: *Summary of the 6 h QPF subjective verification and daily CSI at the 1 inch QPF threshold for the five days in which all FV3-SAR configurations were available for the subjective evaluation portion of the experiment. Left axis is CSI, right axis is the number of times a configuration was chosen (participants could choose more than one configuration) as a good forecast, lines are the CSI, and bar plots are the subjective evaluation picks.*

During the course of the experiment, the most preferred FV3-CAM configuration, when available, was the SAR3 followed by the SARX and SAR4. Figure 27 shows that SAR3 was chosen 35% of the time when it was available, while the SARX and SAR4 were chosen approximately 31% of the time. The subjective preference differs from the objective results, both for 24 h and 6 h QPF, and will be discussed further below.

Repeatedly discussed by the participants, both verbally and written, was the pronounced wet bias in the QPF from all the FV3-SAR configurations (excluding SAROU). This wet bias was noted in both synoptically forced and mesoscale events but what was most concerning to the participants was the precipitation forecast for single cell convection, often referred to as "popcorn" convection. For instance one of the participants wrote: "The 'popcorn' storms in all of the models have too much rain compared to observations - they all have 2-3" of rain in each little blob. That makes it hard to distinguish between everyday storms and more organized threats."

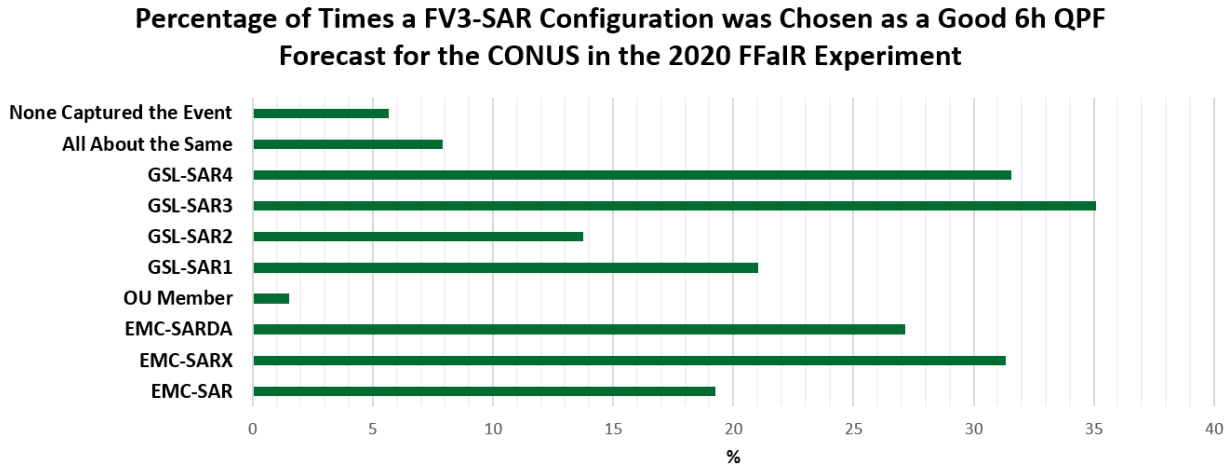

Figure 27: *The percentage of time each FV3-SAR configuration was chosen as a good forecast across the CONUS for the 6 h QPF during subjective verification. Participants were allowed to pick multiple models/choices and not all models were available the same number of times; refer to Table 2.*

When this type of convection was being simulated by the FV3-SARs, nearly every convective cell was forecasted to be a heavy rain event. Furthermore, the structure of the "popcorn" convection looked gridded in nature, almost as if the whole convective cycle was occurring within one grid cell of the model. An example of this can be seen in Fig. 28 in the 24 h QPF across Ohio and Tennessee Valleys. Observations show that scattered convection occurred across the region, with only some areas of embedded precipitation totals exceeding 2 inches. However, the SAR, SARX, and SAR3 all had widespread convection producing rainfall totals exceeding 3 inches; this was not unique to these three models. Additionally, the size of the convective cells were larger than observed. Most troubling was that these 24 h totals generally occurred on the hourly timescale. Comparing Fig. 28 to Fig. 29 and Fig. 30, it can be seen that the same gridlike looking, high precipitation single cell convection is also seen in the 1 h QPF forecast. This suggests that the wet bias in the popcorn convection seen in the 24 h QPF is actually hourly QPF totals for the individual cells, not precipitation that was accumulated on the gridscale over 24 hours. Looking at Fig. 29F and Fig. 30F, it can be seen that the HRRRv4 does not have this same issue and is therefore likely a problem confined to the FV3 core. Since this over-forecasting of both convective development and the associated QPF was constantly seen throughout the duration of the experiment and across every FV3-SAR configuration, it is imperative that developers of the FV3-SAR work to identify the cause of wet bias.

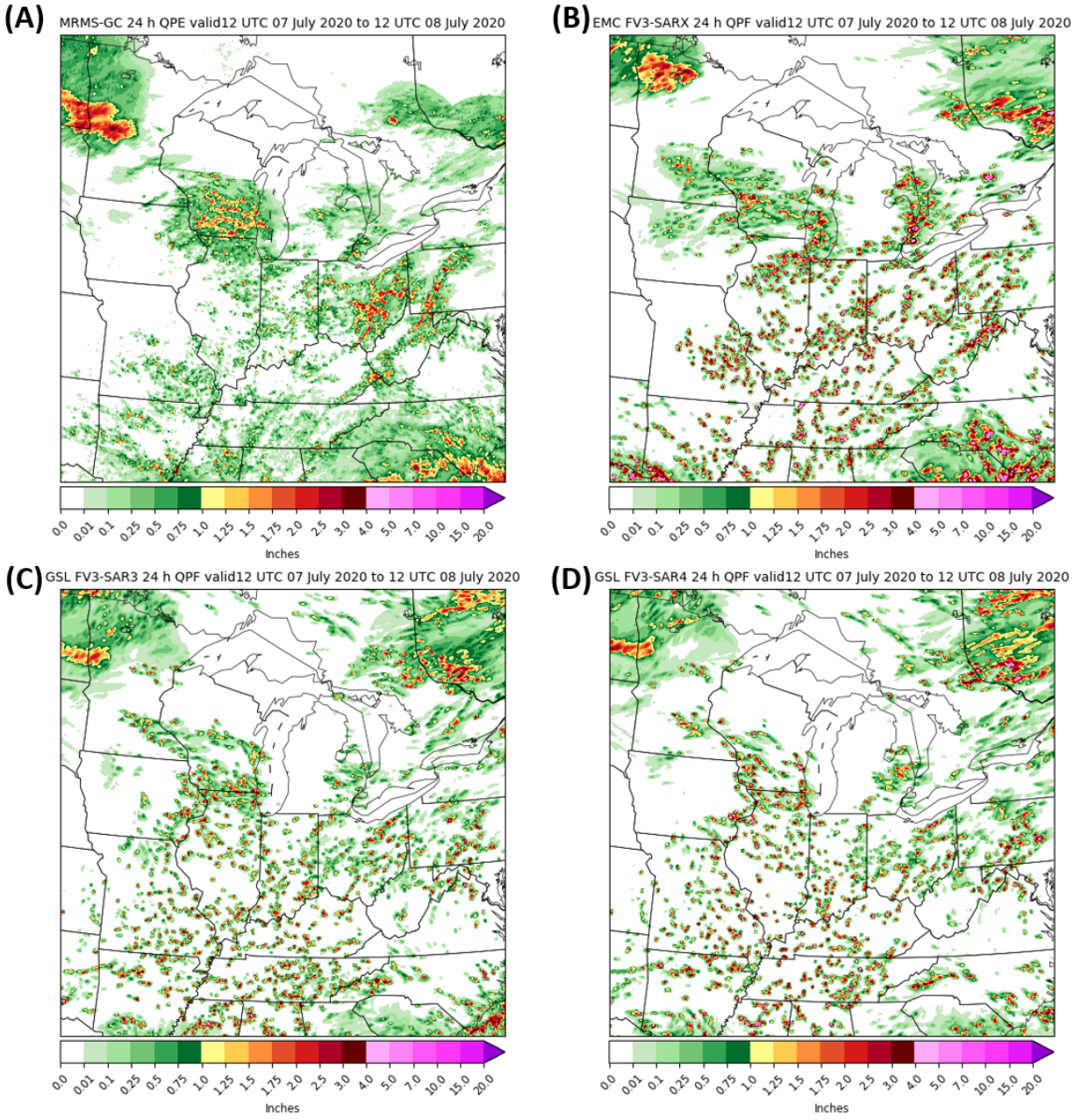

Figure 28: *(A)24 h MRMS-GC QPE and 24 h QPF from (B) EMC FV3-SARX, (C) GSL FV3-SAR3, and (D) GSL FV3-SAR4 valid 12 UTC 07 July to 12 UTC 08 July 2020.*

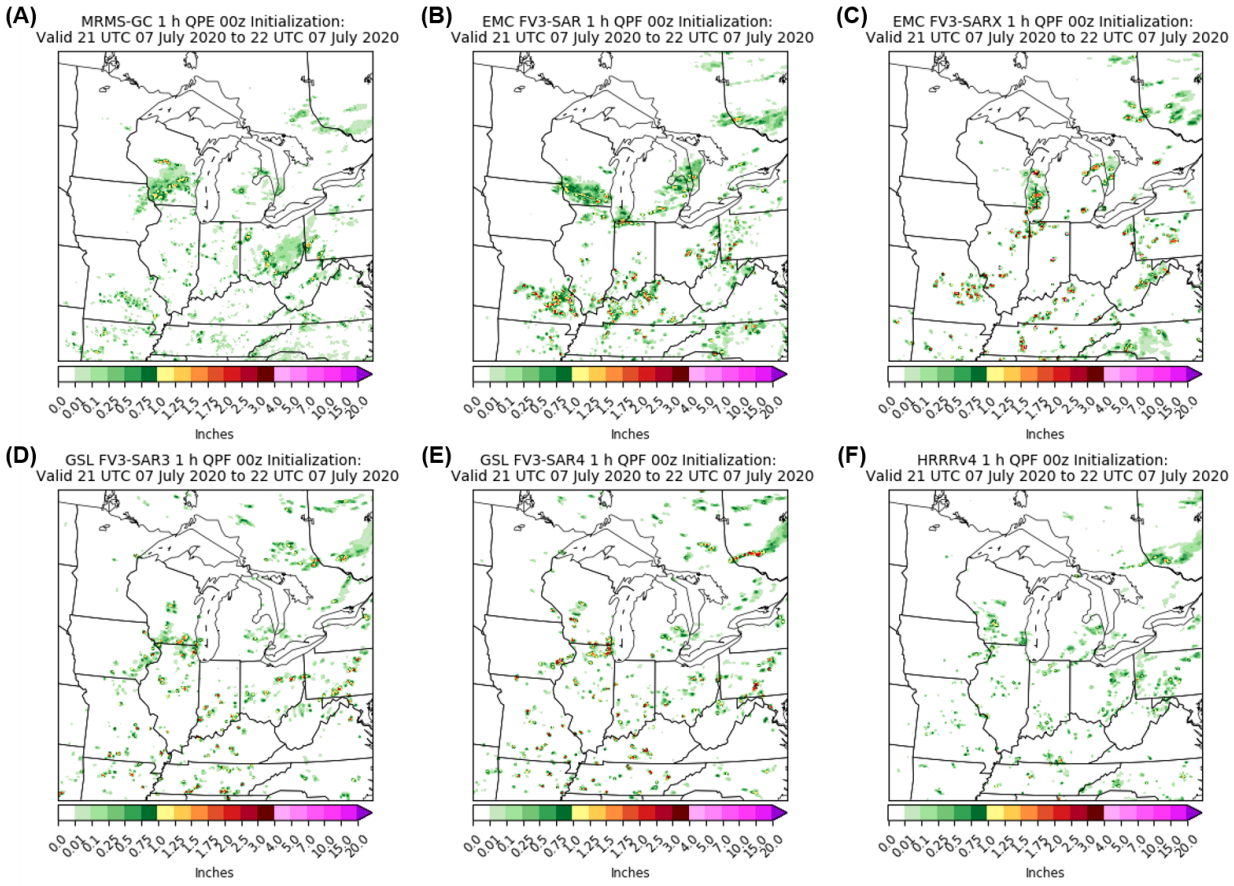

Figure 29: *(A)1 h MRMS-GC QPE and 1 h QPF from (B) EMC FV3-SAR, (C) EMC FV3-SARX, (D) GSL FV3-SAR3, (E) GSL FV3-SAR4 and (F) HRRRv4 valid 21 UTC to 22 UTC 07 July 2020.*

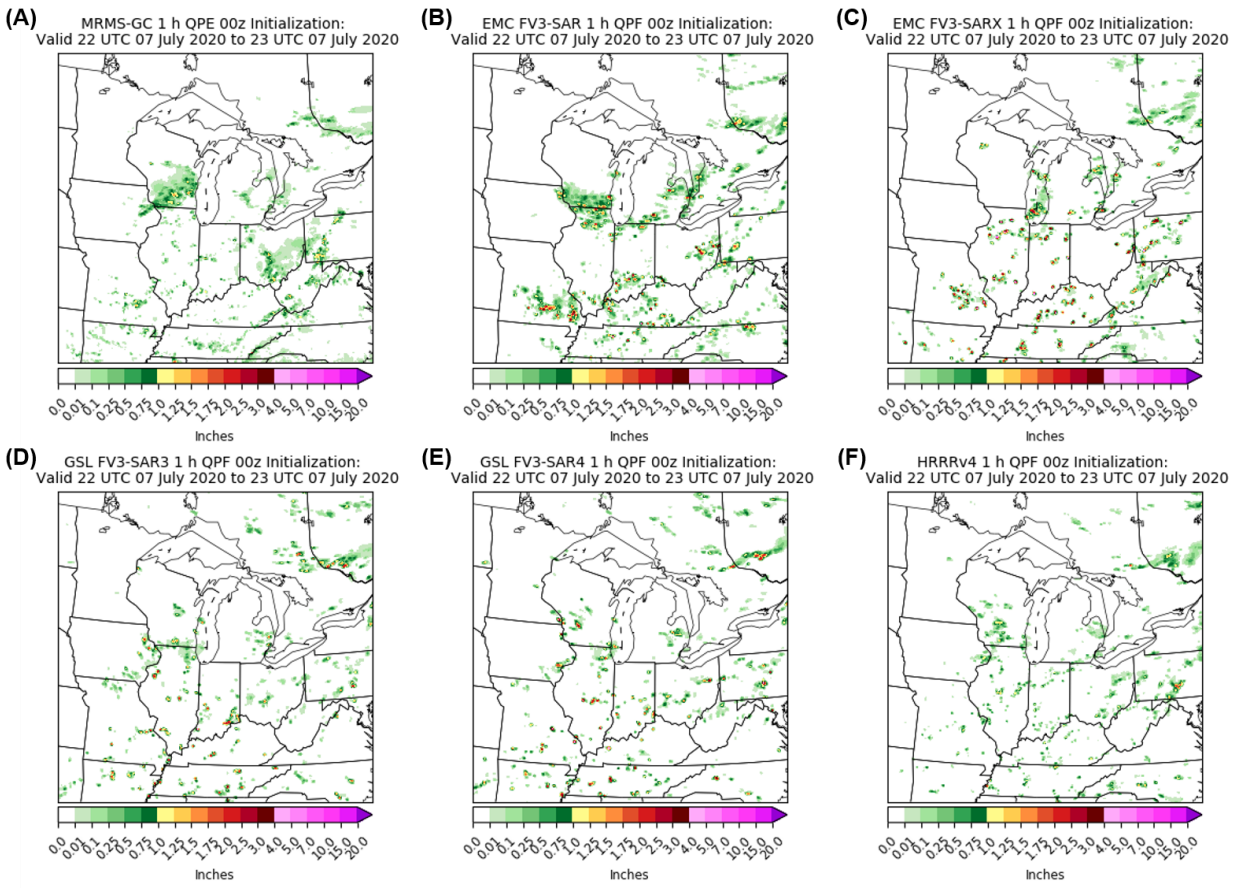

Figure 30: *Same as Fig. 29 but valid 22 UTC to 23 UTC 07 July 2020.*

4.1.3 Objective Evaluation

As stated in Section 2.3.2, objective verification was done using MODE and grid-stat in the MET package to produce the performance diagrams. This method may not account for single cell or popcorn storms when they fall below the filter scale and are not in proximity to other filter scale precipitation objects. Therefore, the extent of the wet bias in the FV3-SAR configurations are likely not fully recognized. The FFaIR team plans to perform grid point verification at a later date to identify the effect that different verification methods have on model performance and to better document the performance associated with the high QPF in nearly all popcorn convection in the FV3-SAR models.

The 24 h grid-based QPF performance diagrams valid 12 UTC to 12 UTC for 0.5, 1, and 2 inch thresholds for the 00z and 12z model runs can be seen in Fig. 31; refer to Table 3 to see the number of days included in each model's bulk verification. Over the course of the experiment the overall performance of the 12z run of the deterministic models was better than the 00z. This is to be expected and a good thing to see. An exception to this was in the bias, with nearly every model seeing a slightly worse bias in the 12z run compared to the 00z. For instance, the SAR and

SARX both had an increase in their wet bias at the half inch and inch threshold in their 12z runs. However, the differences seen in the model biases between the two runs is minimal and does not deter from the increase in performance seen for the other indices. Additionally, the HRRRv4 had the highest CSI throughout the experiment, outperforming all FV3-SAR configurations.

Across nearly all thresholds examined and for both model run times, the SAR had the highest wet bias; this was especially notable at the half inch threshold. This wet bias might have helped lead to the higher probability of detection (POD) seen at the half inch and inch thresholds compared to the other models, though it also constantly had one of the lowest Success Ratios (SR). The HRRRv4 had the highest Critical Success Index (CSI) across all thresholds examined, followed by the SARX and then the SAR. This result was similar to the subjective results, though subjectively there was a greater separation between the HRRRv4 and SARX to the SAR. At the 2 inch threshold for the 00z runs, the SARX and HRRRv4 stood out as the better forecasts but for the 12z runs the models, the performance was roughly the same across the models evaluated.

Lastly, all the GSL FV3-SAR configurations were generally clustered together on the 24 h QPF grid-based performance diagrams, so it is difficult to determine if one was constantly better than the other. Overall the configurations with the same ICs and LBCs (SAR1 and SAR2 vs SAR3 and SAR4) were more similar to one another than those with the same dynamics suite (SAR1 and SAR3 vs SAR2 and SAR4). This suggests that the ICs/LBCs are more influential in the model performance than the parameters forcing the dynamics, at least when comparing Hord 5 to Hord 6[16]. The SAR3 and SAR4 nearly always had a higher CSI than SAR1 and SAR2, while SAR1 and SAR3 generally had a lower wet bias than their counterparts (SAR2 and SAR4 respectively).

[16] Reminder Hord is an abbreviation for “Horizontal Order”.

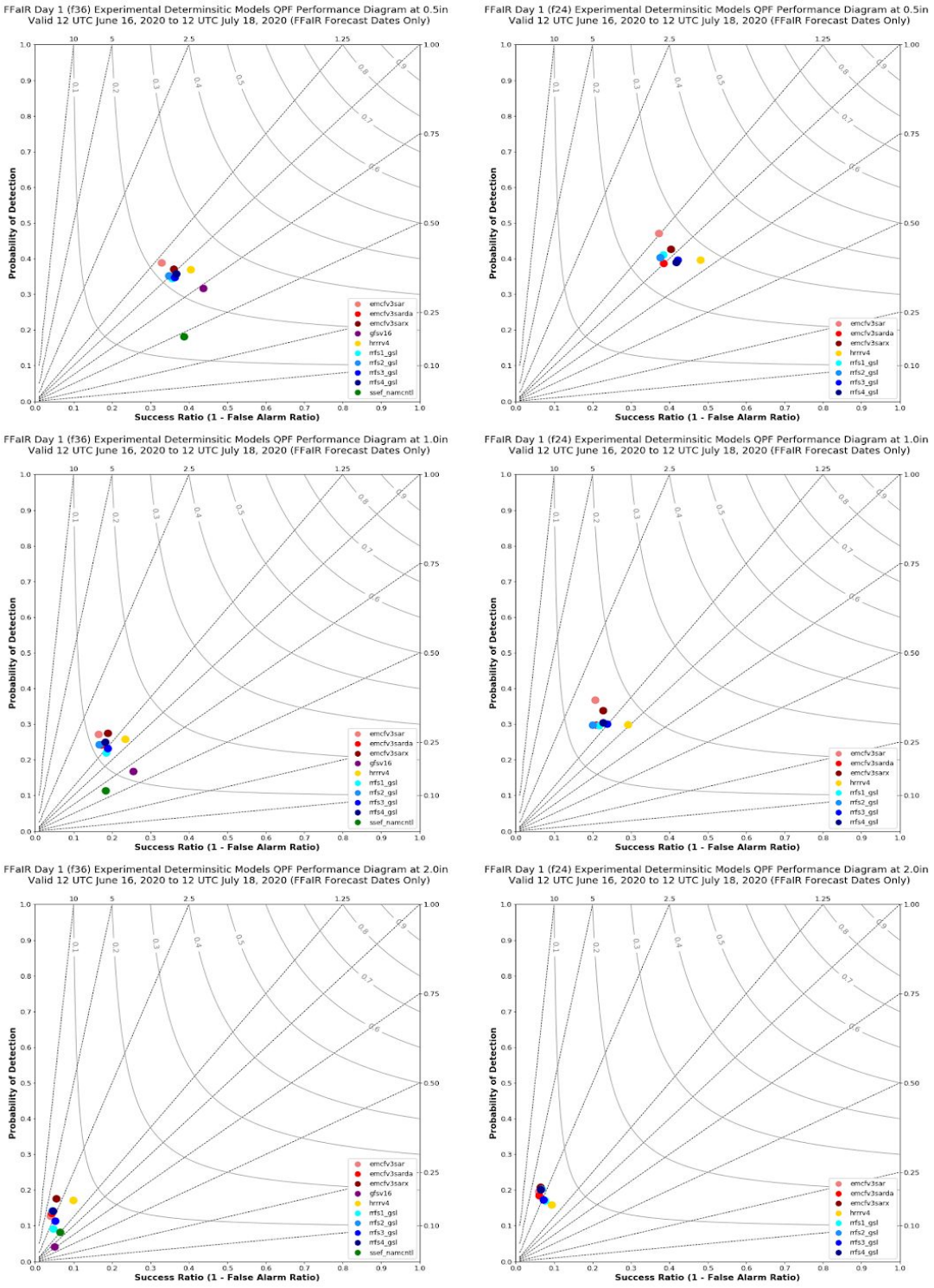

Figure 31: *Performance diagrams for the Day 1, 24 h QPF forecasts valid for only the days in which FFaIR was in session, from June 16 to July 18, 2020 for the deterministic models evaluated during FFaIR. Left side: model initialization at 00z. Right side: model initialization at 12z. Precipitation thresholds are for: 0.5 inches (top), 1 inch (middle) and 2 inches (bottom).*

The 6 h QPF grid-based performance diagrams, valid 18 UTC to 00 UTC, can be seen in Fig. 32. The higher wet bias for the SAR at the half inch threshold was also seen in the 6 h grid-based QPF analysis, while all the other FV3-CAMs (excluding SAROU) were clustered together and the SAR2 and SAR3 had the lowest CSI. At the inch and 2 inch threshold, the EMC FV3-SARs outperform the GSL FV3-SARS, though it is important to note there was a larger sample size from EMC than from GSL. Of the GSL models, as was seen in the 24 h QPF evaluation, the SAR4 performed the best across all precipitation thresholds examined. The statistical results for the 18 UTC to 00 UTC time period differ from the subjective results. As noted above, when all FV3-CAMs were compared against one another the SAR3 was chosen 35% of the time as a useful forecast while the SAR4 and SARX were chosen 31% of the time. Again, availability of the product comes into play, with the SARX being available over 50% more of the time than the SAR3 or SAR4, which were available the same number of times.

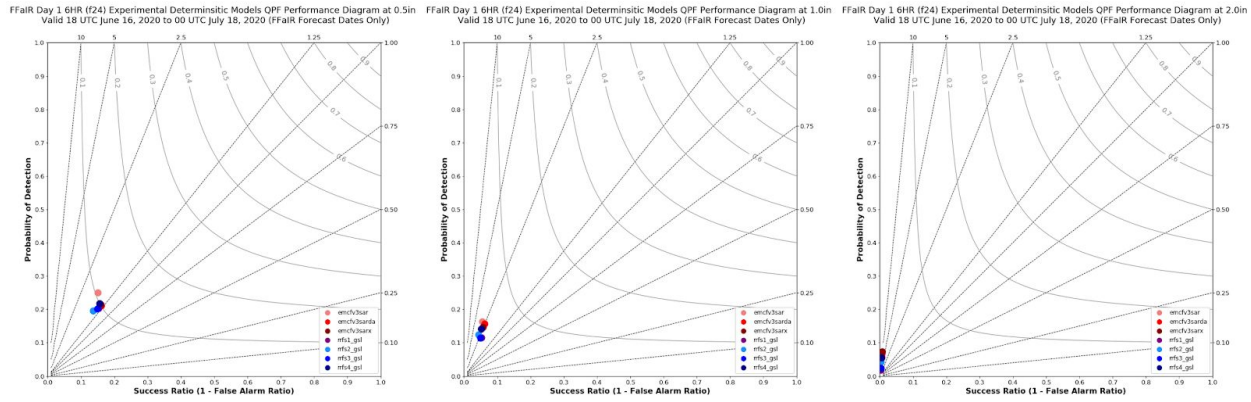

Figure 32: *Performance diagrams for the 6 h QPF forecasts valid 18 UTC to 00 UTC valid for only the days in which FFaIR was in session, from June 16 to July 18, 2020 for the EMC FV3-SAR and GSL FV3-SAR configurations evaluated during FFaIR. Precipitation thresholds are for: 0.5 inches (left), 1 inch (middle) and 2 inches (right).*

The reason for the difference between the objective and subjective results is unknown, though a hypothesis is that the differences were due to the aforementioned over forecasting of precipitation in popcorn type convection. Due to the size of these apparent popcorn convective cells, it is possible that they were not always identified by the verification process. Therefore this type of wet bias would not fully show through in grid-based verification even though it can easily be seen with the naked eye; refer to page 63 in the MODE Technical Report, Section 7.4 (Bullock 2016). An example of how the small scale aspect of the popcorn convection is not always identified as an object, even at the half inch threshold, can be seen in Fig. 33; refer to Figs. 28 to see an example regional 24 h QPF forecast with popcorn convection. Focusing on Fig. 33A, it can be seen that some of the popcorn convection is identified across the Ohio and Tennessee River Valleys as a forecast object (identified by the gray fill) but no observed object was matched. However, comparing the MODE analysis (Fig. 34A) from the SARX to the

zoomed in regional SARX 24 h QPF forecast, Fig. 28B, QPF > 0.5 inches can be seen speckled across the region that were not identified by MODE.

During the times in which the SARX, SAR3, and SAR4 were all available for comparison against one another, the SAR3 appeared to be the least likely of the three to forecast 4+ inches of QPF in 6 h within the popcorn storms. Since the participants often discussed the inaccuracy of the precipitation intensity forecasts for what was general daytime scattered convection, they would be more prone to chose the FV3-CAM model that had the lowest QPF within the cells, when all else was similar between the models (i.e. pattern, storm mode).

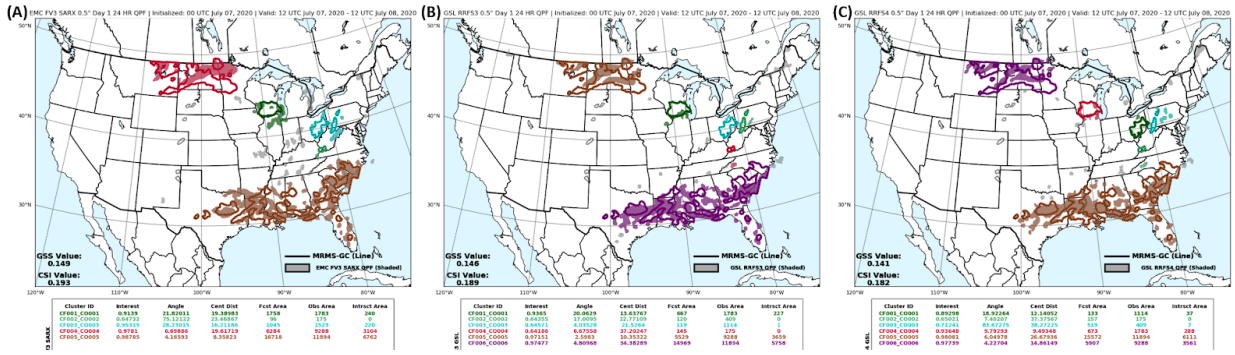

Figure 33: *MODE precipitation results for the 0.5 inch threshold over 24 h valid from 12 UTC 07 July to 08 UTC 25 July 2020, showing the (A)-(C) MRMS-GC QPE (shaded) compared to model QPF (contoured) from the following deterministic models: (A) EMC FV3-SARX, (B) GSL FV3-SAR3, (C) GSL FV3-SAR4. Below each image are the MODE verification metrics for the corresponding model.*

4.2 Ensemble Guidance

The CAM ensembles evaluated during the Day 1 time frame in FFaIR this year were the HREFv3 from EMC, the HRRRE from GSL, and the SSEF from OU-CAPS. The HREFv3 is a multi-core ensemble, while the HRRRE and SSEF are single core, with all members in the SSEF from the FV3 core. The HREFv3 was supposed to include the HRRRv4 in its membership, but prior to the start of FFaIR, instabilities in the HRRRv4 system resulted in a change in the ensemble membership from the HRRRv4 to the operational HRRR.

4.2.1 Subjective Evaluation

Because the HREF is often considered as the ensemble forecast to beat, the FFaIR team this year was interested in a direct comparison of the SSEF[17] and HRRRE to the HREFv3. This was accomplished by showing the 6 h Local Probability Matched (LPM) Mean forecast valid 18 UTC to 00 UTC of the HREFv3 next to one of the other ensembles' LPM mean (either HRRRE

[17] The results for the SSEF are not a complete representation of the ensemble because one of the members of the ensemble was the SAROU, which, as previously stated, was found to have a bug related to the coupling of the MYNN and the radiation schemes.

or SSEF) to compare against each other and the observed 6 h precipitation. An example of what the participants were shown can be seen in Fig. 34 (example uses HRRRE, HREFv3 comparison). The participants were then asked if the ensemble's forecast was better, slightly better, about the same, slightly worse or worse than the HREFv3 forecast. For this day, 7 out of the 13 participants felt the HRRRE performed better than the HREFv3, while 4 felt they performed about the same and the remaining 2 felt the HRRRE performed slightly worse than the HREFv3.

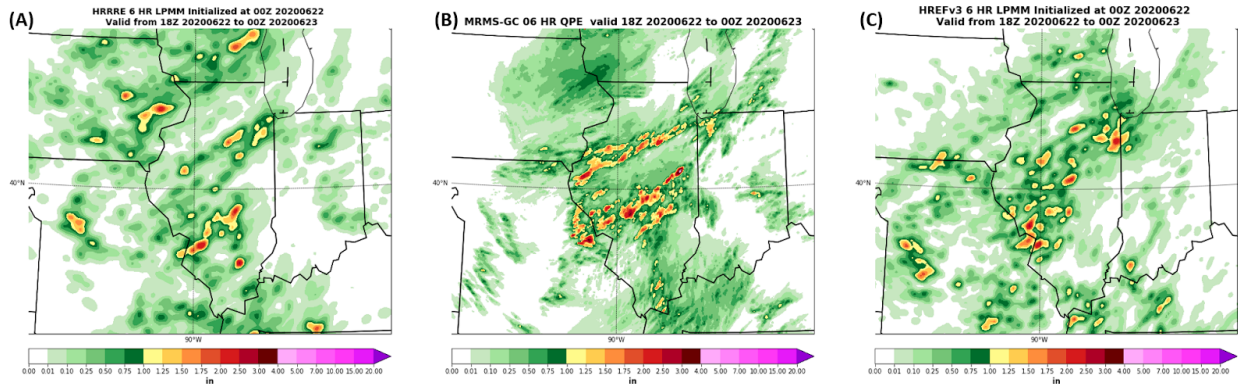

Figure 34: *Example of the subjective verification image shown for the ensemble comparison question. 6 h LPM mean QPF for (A) HRRRE and (C) HREFv3 and (B) 6 h MRMS-GC QPE valid 18 UTC 22 June to 00 UTC 23 June 2020.*

Figure 35 shows that over the course of the experiment, no matter if compared to the HRRRE or the SSEF, the HREFv3 was more often identified by the participants as the better or slightly better forecast. However, the HRRRE was more likely to be chosen over the SSEF as being comparable or slightly better than the HREFv3. Additionally, participants regularly commented that they liked aspects from each of the ensembles and that looking at various ensemble outputs helped them get a better idea of where the uncertainty is. They also often noted that the HRRRE was on par with the HREFv3 when it came to forecast problems such as location or amounts.

Another takeaway from the subjective evaluation was how forecasters interpret ensemble mean products during verification. For example, in the subjective verification discussions, forecasters would often comment on whether or not the LPM mean (which was used for subjective evaluation of the ensembles) resembled the observed storm mode (i.e. banded, popcorn, etc.) or convective evolution. However, such a comparison is not wholly reasonable, since all mean products, including the LPM mean, are a combination of the ensemble members and therefore the structure/evolution of the convective system cannot be construed if there is large spread among the members. An example of why this should not be common practice during

subjective verification was provided to the FFaIR team by a GSL participant who is part of the HRRRE team and can be seen in Figs. 36 - 37 .

Figure 36 shows the LPM mean comparison image the participants were shown for the HRRRE and HREF from 18 UTC June 12 to 00 UTC June 13, 2020. Upon seeing this, participants made statements like: "The 'mode' of precipitation looks much poorer in the HRRRE (isolated convection?)" and "HRRRE provides idea of extremes better, where HREF more closely resembled discreteness/coverage of QPF." However, as can be seen in Fig. 37A-B, if other types of means are examined, the same convective mode inferred from the LPM mean would likely not be deduced from these. Furthermore, there is a wide variety of storm modes and evolutions that can be implied when looking at the individual members' QPF (Fig. 37C-F). Therefore, it is important to remind participants that interpreting storm mode from a single image mean product during verification should not be common practice.

The feedback from the participants of how they interpreted the LPM mean and other mean products, though secondary to the objective of the subjective evaluation question, shows the need for continuous training on how to interpret various ensemble products. With the widespread use of ensembles, especially in the day 1 time period, it is important to ensure that forecasters are not unknowingly applying deterministic evaluation methods to ensemble products. Additionally, these results led to a discussion among the FFaIR team on how to best compare the performance of one ensemble to another in future testbed experiments.

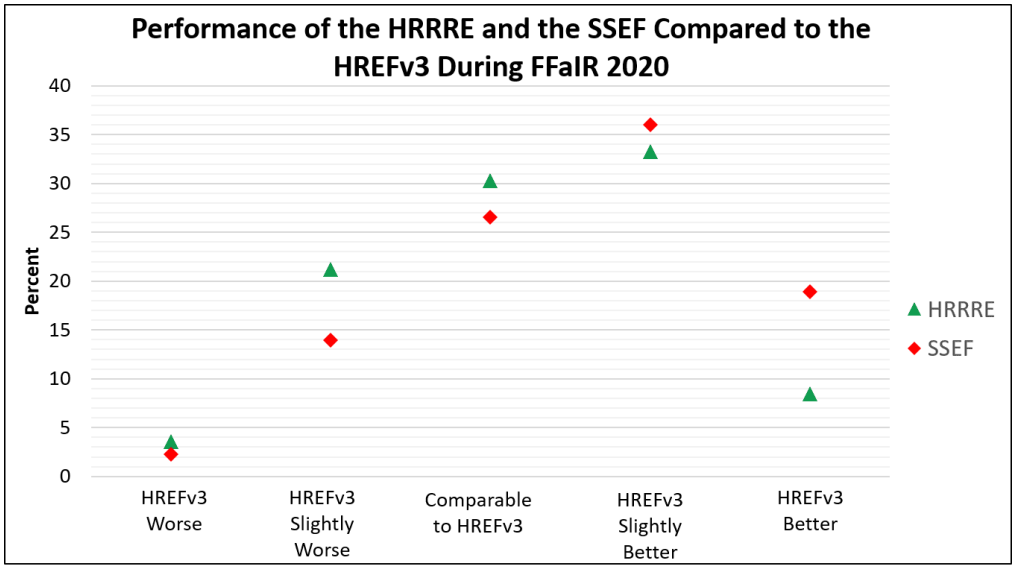

Figure 35: *The percentage of time that the HRRRE and SSEF was chosen to be worse, slightly worse, comparable to, slightly better or better than the HREFv3 by the FFaIR participants during the subjective evaluation portion of the experiment.*

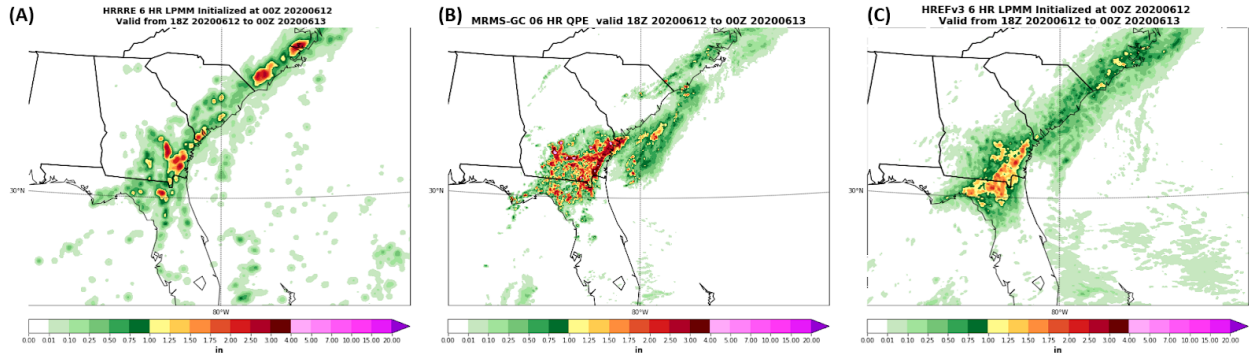

Figure 36: *6 h LPM mean QPF for (A) HRRRE and (C) HREFv3 and (B) 6 h MRMS-GC QPE valid 18 UTC 12 June to 00 UTC 13 June 2020.*

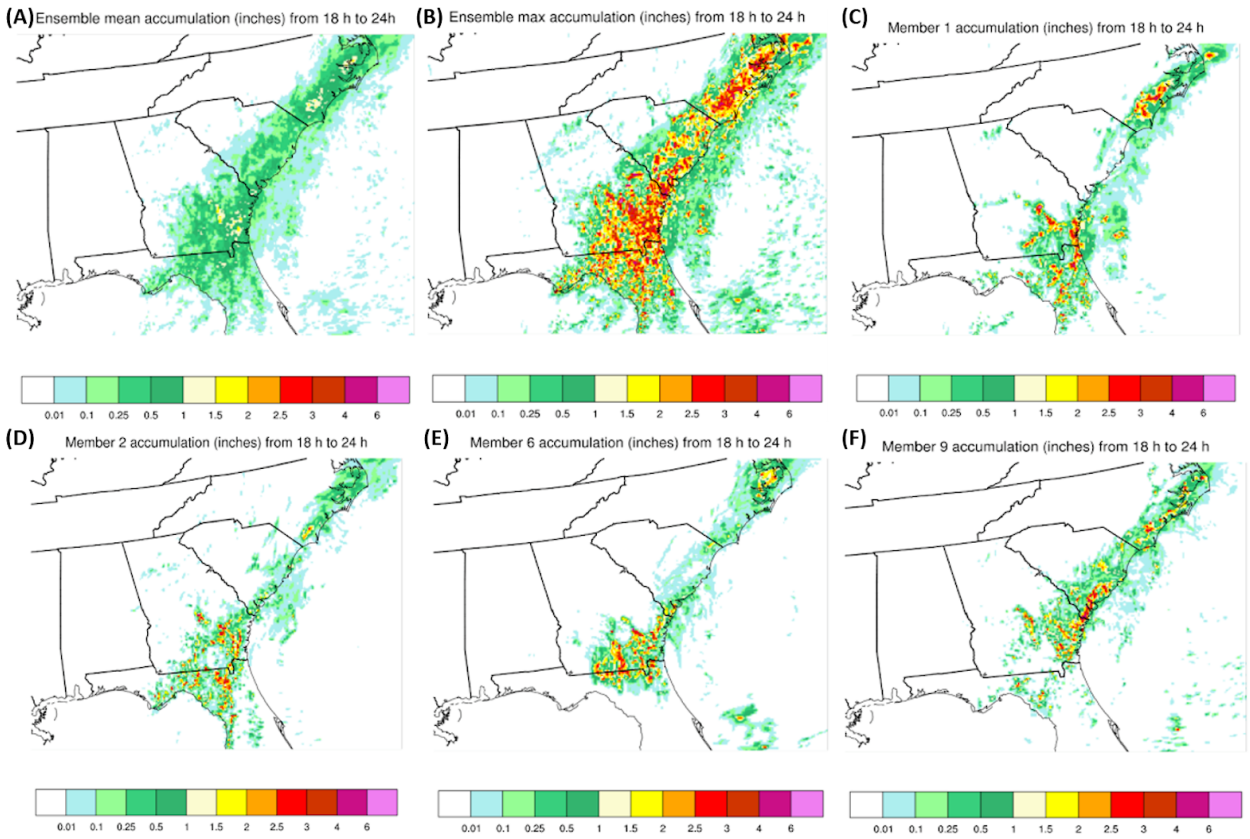

Figure 37: *HRRRE 6 h (A) mean QPF, (B) ensemble max accumulation, (C) member 1 QPF, (D) member 2 QPF, (E) member 6 QPF, and (F) member 9 QPF valid 18 UTC 12 June to 00 UTC 13 June 2020.*

4.2.1.1 SSEF Spatially Aligned Mean Products

This year the OU-CAPS team provided two new mean products from the SSEF to be evaluated, the Spatially Aligned Mean or SAM and a blend of the SAM and LPM mean or SAM-LPM[18]. Participants were shown ensemble mean (LPM mean) alongside the SAM (SAM-LPM) and asked if they felt the SAM (SAM-LPM) added value to the forecast, degraded the forecast, or neither. It can be seen in Fig. 38 that over 40% of the time the participants felt that the SAM products did not add value to the forecast process. This was also apparent when reading through their comments and listening to their discussions about the products. Often they would note that it was difficult to see a difference between the SAM product and their mean counterpart; an example of how subtle differences could be sometimes can be seen in Fig. 39.

[18] Refer to the FFaIR 2020 Operations Plan for more information on these products.

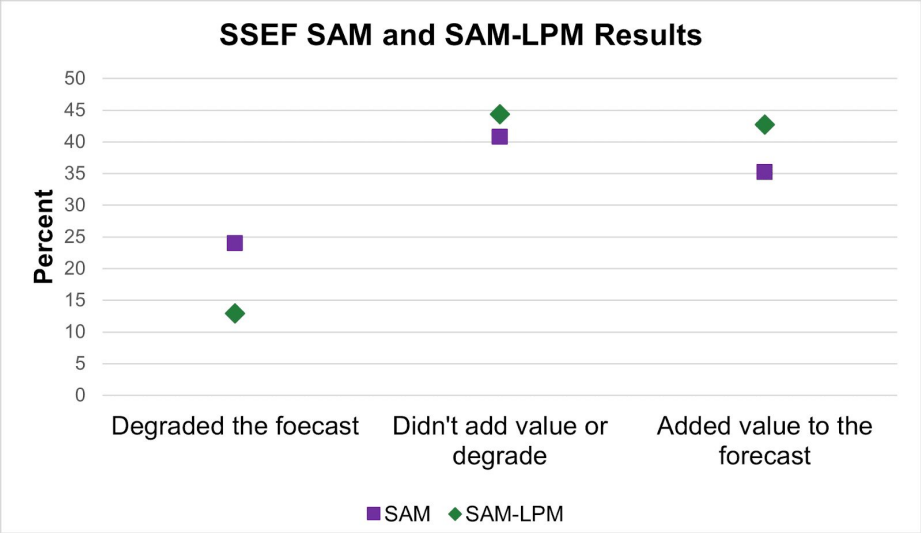

Figure 38: *The percentage of time that the SSEF SAM and SAM-LPM were chosen to have degraded, added value or neither compared to the SSEF mean and LPM mean by the FFaIR participants during the subjective evaluation portion of the experiment.*

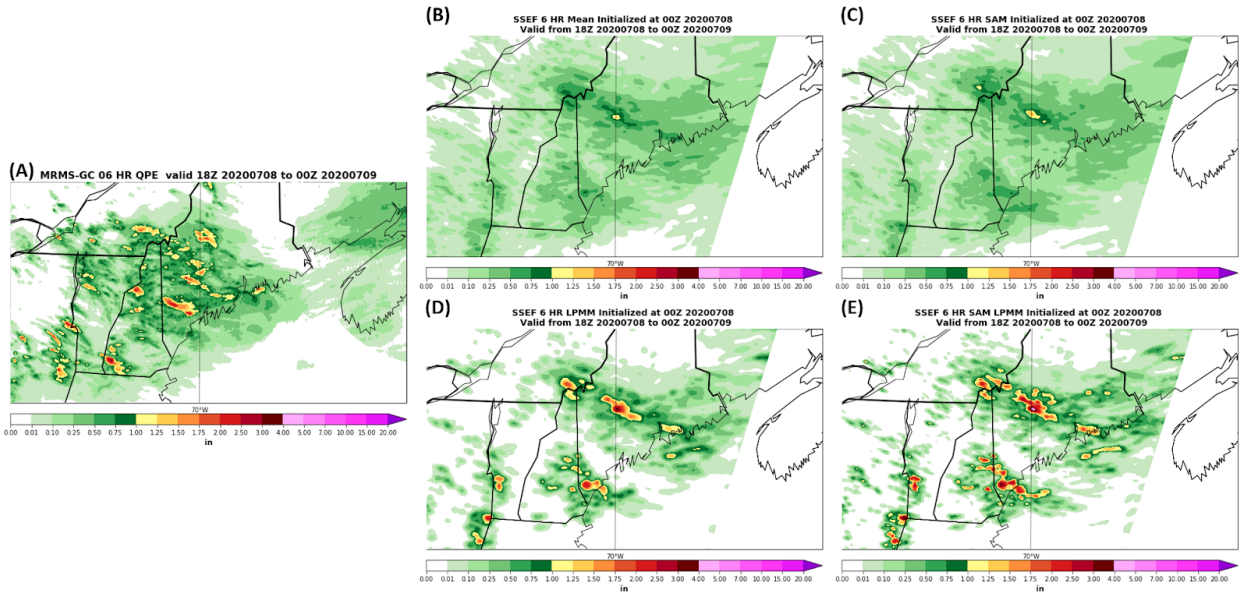

Figure 39: *(A) 6h MRMS-GC QPE and 6 h SSEF (B) mean QPF, (C) SAM QPF, (D) LPM mean QPF, and (E) SAM-LPM mean QPF valid 18 UTC 08 July to 09 July 2020.*

Focusing solely on the SAM-LPM, Fig. 38 shows that the participants felt that the product added value to the forecast 43% of the time and 45% of the time felt it neither added value or degraded the forecast. Participants stated that when the LPM mean had the general idea of the forecast correct, applying the spatial alignment algorithm to the LPM mean appeared to enhance the LPM product. For instance one participant noted: “The most noticeable impact of applying the SAM technique seems to be to increase the intensity of the LPMM values rather

than displacing or changing the shape of features. So if the regular/non-SAM mean is too low, then applying SAM will improve the forecast quality and vice versa."

The OU-CAPs team performs their own statistical analysis of their ensemble (the SSEF) and their products, therefore one was not done by the FFaIR team. Additionally it should be noted that the SAM products were only available for the second half of FFaIR and therefore, due to the small sample size, a comprehensive conclusion should not be made. Instead additional analysis should be done, including further evaluation in FFaIR.

4.2.1.2 NBM Probability Matched Mean Products

The National Blend of Models (NBM) is looking into developing a PM mean and this year provided FFaIR with two versions of the product, one that does not include the QMD (Quantile Mapping and Dressing) system and one that does. The non-QMD products are referred to as DMO PM mean[19] and the QMD product is referred to as QMD PM mean. Over the course of the experiment the two products generally were not well received by the participants. Figure 40 shows the results from the subjective evaluation of the products; the median score for each of the products was a 3.

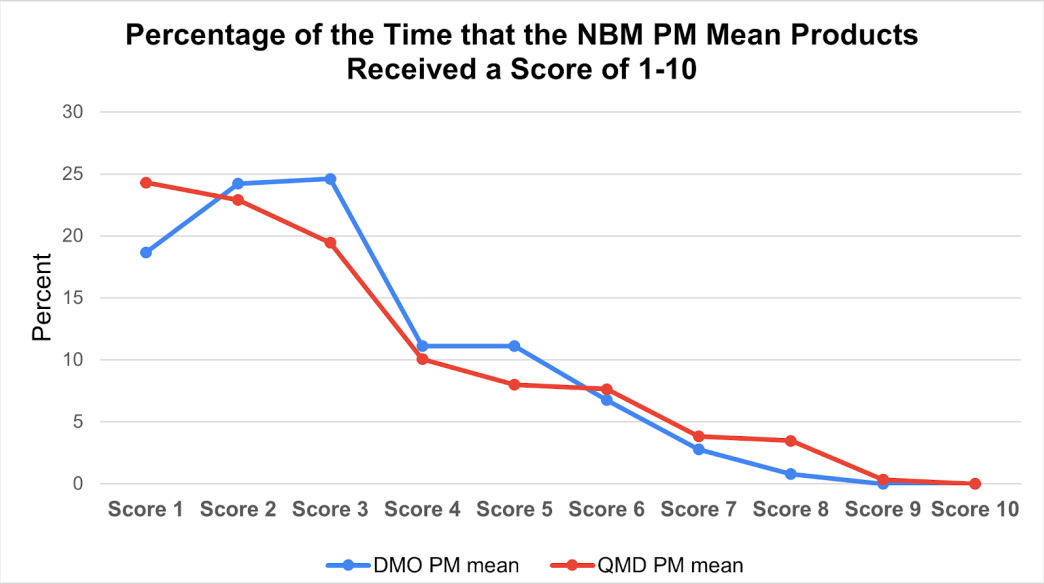

Figure 40: *The percentage of time that the NBM DMO PM mean and QMD PM mean received a score by the FFaIR participants during the subjective evaluation portion of the experiment*

A major concern among the participants was under forecasting of precipitation amounts along with the inability to highlight important features of the event (i.e. things like banded heavy precipitation). An example of these challenges can be seen in Fig. 41A-F; for both days shown, the two products did well predicting the footprint of the lower end rainfall amounts (< 1 inch) but

[19] DMO stands for Direct Model Output.

missed the heavy amounts. Additionally the PM mean products did not help highlight where or what the "forecast problem of the day was." That said, there were instances, for example see Fig. 41G-I, in which both NBM PM means were able to identify the heavy rainfall threat. One theory for the NBM PM mean products having difficulty isolating the heavy rainfall threat is that rather than using a weighted blend, consisting mostly of CAM scale models like is done for the Day 1 QPF product, every member that goes into the NBM was used, including global models. This resulted in around 170 forecasts being used to calculate the PM mean. Such a wide dispersion of forecasts and model resolution likely washed out heavy rainfall signals except for in strongly forced cases, such as July 16-17 (Fig. 41G-I). It is the FFaIR team's suggestion that a weighted blend similar to that used for Day 1 QPF should be considered to be used to calculate the PM mean rather than all the members available for blending.

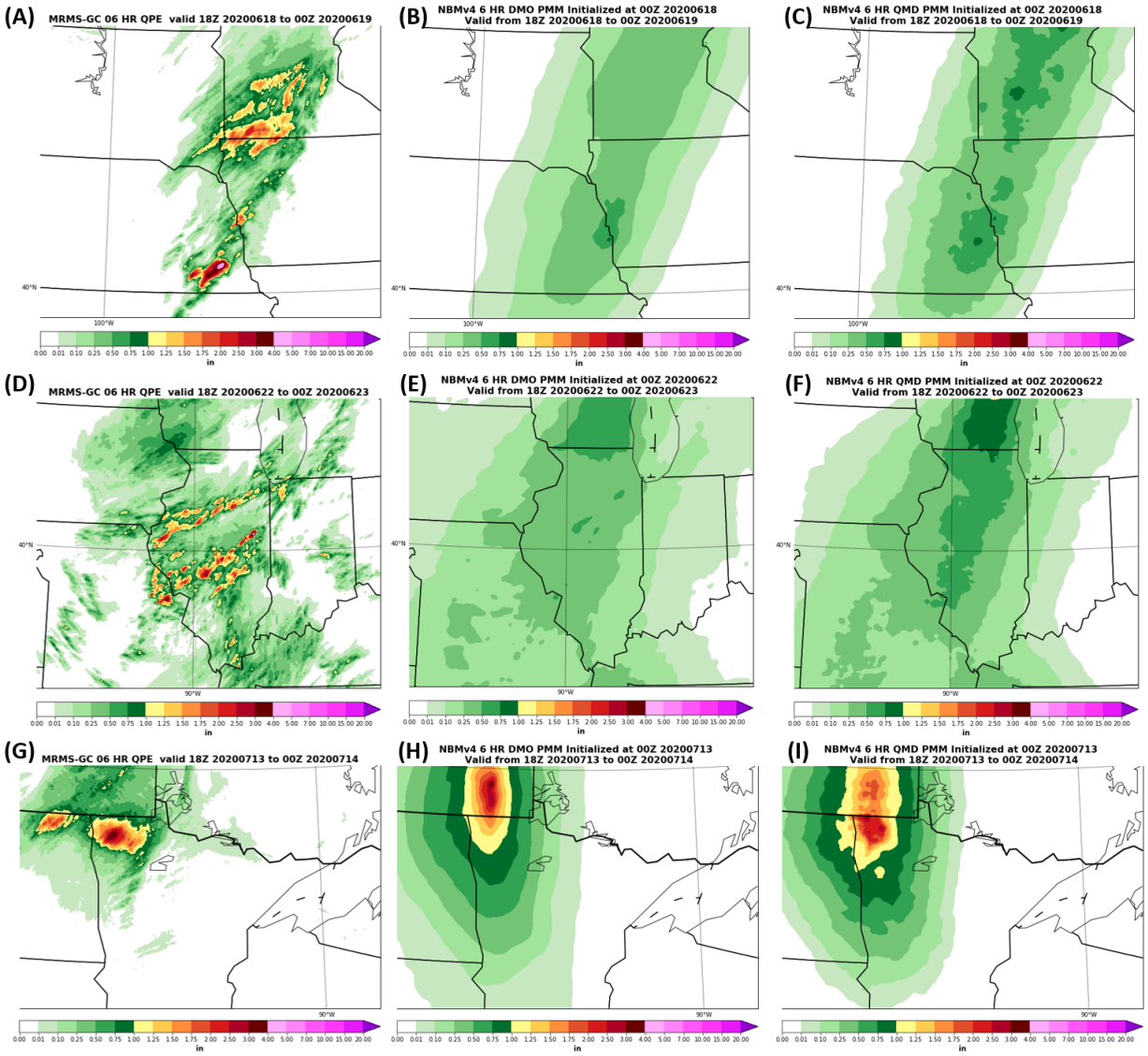

Figure 41: *Left: 6 h MRMS-GC QPE. Middle: 6 h NBM DMO-PM mean QPF. Right: 6 h NBMQMD-PM mean QPF. (A)-(C) valid 18 UTC 18 June to 00 UTC 19 June 2020, (D)-(F) valid 18 UTC 22 June to 00 UTC 23 June 2020, and (G)-(I) valid 18 UTC 13 July to 00 UTC 14 July 2020.*

4.2.2 Objective Evaluation

The performance diagrams for the ensemble products evaluated in FFaIR can be found in Fig. 42 and Fig. 43. The HREFv3 mean products repeatedly had the highest CSI of all the ensemble products evaluated. This included both run initialization times and forecast length (24 h and 6 h). Of the HREFv3 products, the probability matched mean (PMM) outperformed the general mean and the LPM mean. Furthermore, at the half inch precipitation threshold, the three highest CSI values were the HREFv3 PMM, LPM, and general mean. The HREFv3 products also generally had less of a bias for the 24 h precipitation forecasts compared to the other ensembles. The HRRRE consistently had a dry bias across all its mean products, though at the 2

inch threshold, the HRRRE PMM had a slight wet bias. This was in noticeable contrast to the HRRRE mean at the same threshold, which had bias approaching 0.10 for the 24 h QPF initized at 12z; the bias for the HRRRE PMM was near 1. There was also a large difference in bias between the 12z HREFv3 mean and 12z HREFv3 PMM for the 2 inch precipitation threshold, 0.25 compared to almost 2.5. The bias approaching 2.5 seen for the 12z HREFv3 PMM at the 2 inch threshold was the greatest bias seen for any of the products for all thresholds and run initializations for 24 QPF. The higher bias of the PMM compared to the general mean is not surprising since the PMM product is designed to amplify the mean values using individual members' QPF values. The fact that the HREFv3 repeatedly had the highest CSI and POD but also the highest bias shows the importance of not using only one verification metric to determine the "value" of a product.

Focusing on the performance diagrams for the 6 h QPF, Fig. 43, valid from 18 UTC to 00 UTC, the HREFv3 LPM mean had the highest CSI compared to the HRRRE and SSEF LPM mean products. This aligns with the subjective results, where the HREFv3 was considered the best of the three LPM mean products by the participants. The HRRRE LPM outperformed the SSEF LPM during this timeframe, which also agrees with the subjective results. However, the difference in performance between the HRRRE and SSEF LPM was less than the difference of either to the HREFv3 LPM. Additionally the LPM mean for the HRRRE and the SSEF had a dry bias for the half inch, inch, and two inch precipitation thresholds while the HREFv3 had a wet bias. The wet bias might be one of the reasons the participants generally preferred the HREFv3 over the other two ensembles, since they stated they use LPM mean guidance to get an idea of the reasonable worst case scenario. Therefore they expect the product to have a slight wet bias and use that knowledge in their forecast.

Comparing the SSEF mean to the SSEF SAM, for both the 6 h and 24 h forecast, for all thresholds except for the half inch, 24 hr QPF, the general mean performed better than the SAM. Interestingly, for the 24 h forecast, the general mean had a higher CSI, though the SAM had a slightly better SR but lower POD. However, for the 6 h QPF their CSI was roughly the same but the general mean had a higher SR. The opposite was true when comparing the SSEF LPM to the SSEF SAM-LPM, with the SAM-LPM generally having a higher CSI than the LPM. The SAM-LPM also consistently had a higher POD while the LPM had a higher SR. Overall, the results are a mixed bag, it seems as if the SAM itself does not add much increased value compared to the general mean, but the SAM-LPM does add some value to the QPF forecast.

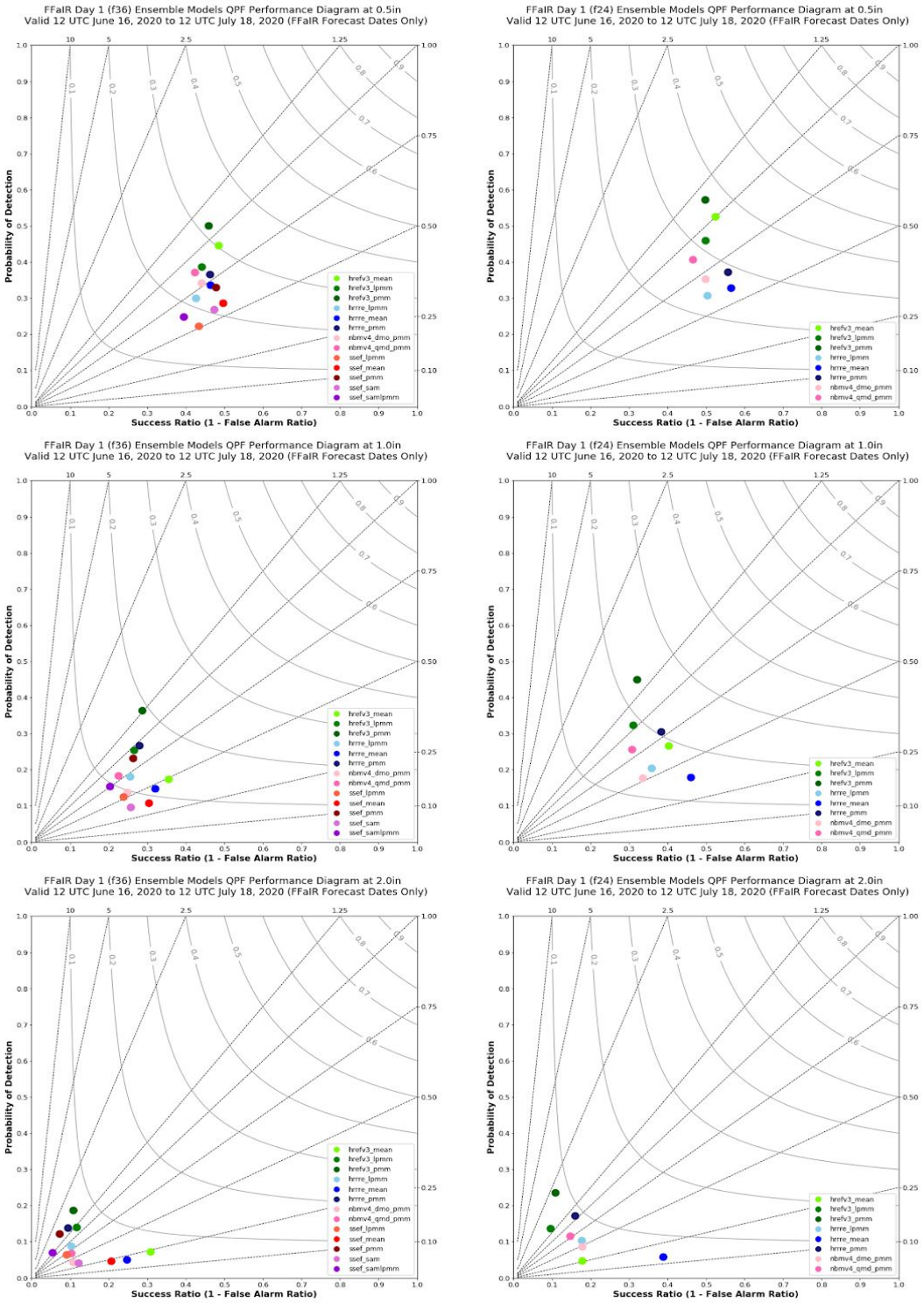

Figure 42: *Performance diagrams for the Day 1, 24 h mean, PM mean, and LPM mean forecasts valid over the four weeks of the 2020 FFaIR experiment (June 16 to July 18, 2020) for the ensembles evaluated during FFaIR. Left side: ensemble initialization at 00z. Right side: ensemble initialization at 12z. Precipitation thresholds are for: 0.5 inches (top), 1 inch (middle) and 2 inches (bottom).*

Lastly, looking at the NBM PM mean products, compared to the PM mean from the HREFv3, HRRRE, and SSEF, the performance of the products were notably worse for the inch and two inch precipitation thresholds, for both initialization time periods. The most notable difference was seen between the HREFv3 PM mean and the NBM DMO PM mean for the 12z run at the inch threshold. Although both had approximately the same SR, the HREFv3 guidance had a CSI near 0.24 and a POD of 0.45, while the DMO-PM had a CSI around 0.13 and a POD of 0.18. Comparing the DMO-PM to the QMD-PM, the QMD-PM product always had a higher CSI and POD, as well as a bias closer to one. The increased performance of the NBM PM product when using the QMD method is not surprising as it is a bias correction method.

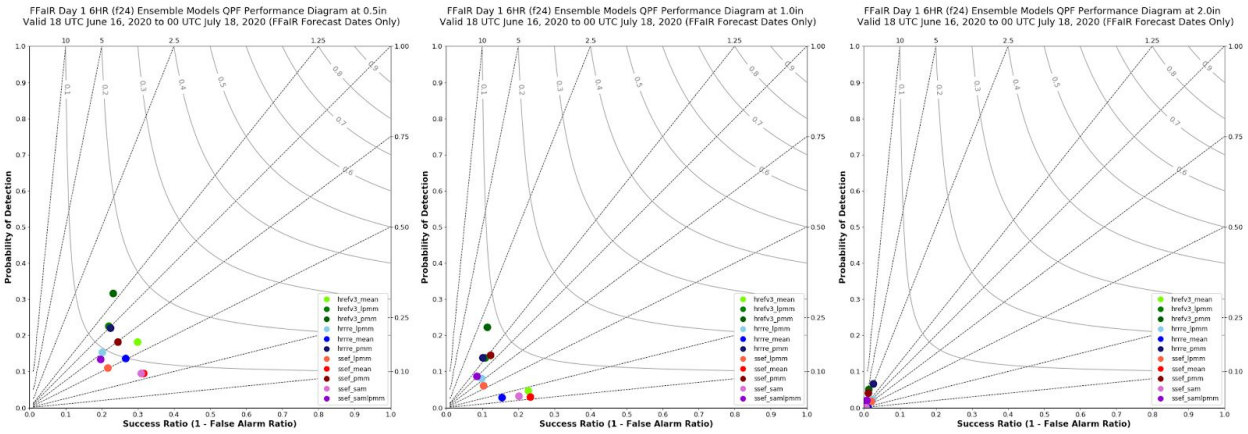

Figure 43: *Performance diagrams for the 6 h QPF forecasts valid 18 UTC to 00 UTC over the four weeks of the 2020 FFaIR experiment (June 16 to July 18, 2020) for the ensemble means evaluated during FFaIR. Precipitation thresholds are for: 0.5 inches (left), 1 inch (middle) and 2 inches (right).*

4.3 CSU Machine Learning "First-Guess" ERO and FFaIR ERO

Three versions of a machine learning first guess product for the Day 1 ERO were evaluated during FFaIR. One trained on GEFSv11 (hereafter GEFS ERO) climatology and excessive rainfall identifiers, the other two are trained on the NSSL model climatology but with two different sets of excessive rainfall identifiers; refer to Appendix D. One of the versions, hereafter NSSL1 ERO, had the same training as the NSSL ERO evaluated in the FFaIR 2019 Experiment, while the new version, hereafter NSSL2 ERO, was trained on the new set of identifiers. Each of the products were valid from 12 UTC to 12 UTC; this differed from Experimental ERO issued by the FFaIR participants, which was valid from 16 UTC to 12 UTC.

Both in the subjective and objective analysis of the three first-guess Day 1 ERO products, the GEFS ERO performed the best while the NSSL1 ERO performed the worst. Figure 44 provides a summary of the subjective evaluation, and it can clearly be seen the participants overwhelmingly favored the GEFS ERO compared to the two NSSL versions. A comment that was repeatedly made by the participants about the GEFS ERO was that although it appeared to

over forecast the excessive rainfall risk, it was a quick way to identify the areas of concern and where to focus their efforts during the forecast process. An example of this can be seen in Fig. 45, for the ERO forecast valid 12 UTC July 08 to 12 UTC July 09, 2020. Here it can be seen that the GEFS Day 1 ERO highlighted the region from the Central Plains to Lake Superior for an enhanced slight risk for excessive rainfall, while both NSSL EROs had a marginal risk for the area. During the discussion for creating the Experimental ERO, the participants discussed adding a Moderate Risk across MN due to the First-Guess from the GEFS, which lead them to focus their time on more deeply investigating heavy rainfall ingredients. Ultimately, as can be seen in Fig 45F, the participants chose to stay with a Slight Risk, with the practically perfect verification suggestions the event was a high end Sligh Risk.

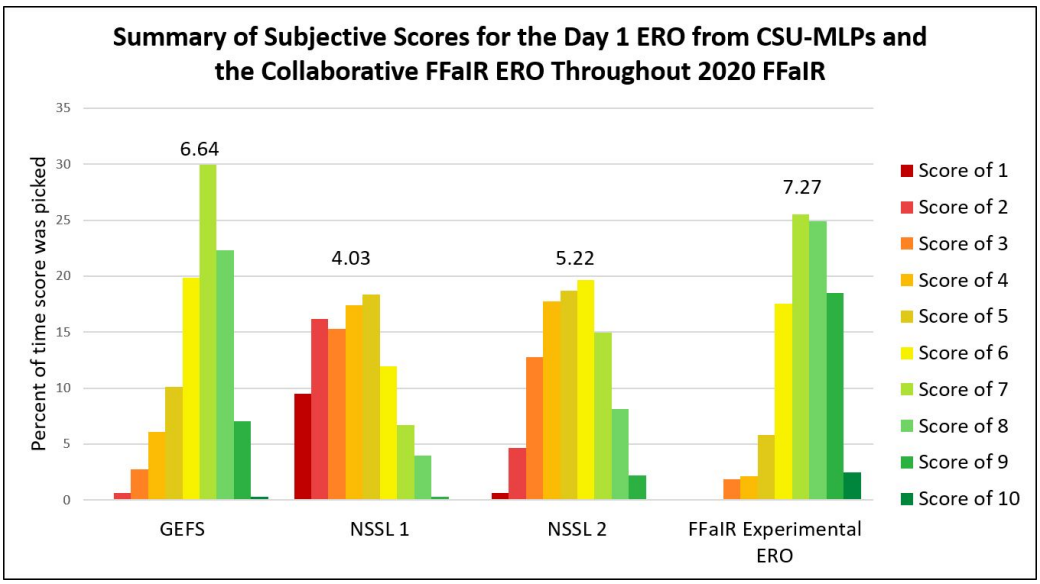

Figure 44: *The percentage of time the CSU First-Guess EROs and the FFaIR ERO received a score during the course of the experiment during the subjective verification along with the experimental average plotted above the percentage analysis.*

Focusing on the two versions of the NSSL EROs, NSSL2 was preferred over NSSL1, with an average score of 5.22 to 4.03. During subjective verification the participants noted that it seemed as if the NSSL EROs had difficulty identifying areas of the higher magnitude excessive rainfall risks. Often they appeared to have a broad brush of Marginal Risk which was not always helpful in the forecasting process at highlighting the "problem of the day". The NSSLs' lack of higher magnitude risk forecasts when compared to the GEFS ERO, the FFaIR ERO and the WPC ERO can be seen when comparing the probability of being under a Slight Risk for each of the various EROs in Figs. 46-47. The NSSL EROs (Figs. 46C and 46E)cover a similar spatial area as the other EROs for Marginal Risk but this spatial coverage drastically decreases for a Slight Risk forecast, along with the probability of being under a slight risk (Figs. 46D and 46F). Additionally, the NSSL EROs have a speckled look to them, meaning there are many areas of

small contoured risk areas. An example of this can be seen in Fig. 46E across IL and IN. The developers of the products believe that this unique aspect of the NSSL EROs is likely a smoothing issue.

As stated, focusing on the GEFS ERO, a quick glance at Fig. 46 it can clearly be seen that the GEFS ERO was more prone to "issue" excessive rainfall risks than either of the NSSL EROs. However, comparing the GEFS probabilities to the FFaIR and WPC ERO probabilities of being in a Marginal or Slight Risk (Fig. 47), it can also be seen that the GEFS appears to generally over forecast the excessive rainfall risk. This coincides with the comments made by the participants about the product. This is especially notable across the Central Plains for both risk categories and along the East Coast from northern Virginia to New York. However, the apparent over forecasting of the risk across the latter of the regions is not surprising because that area is densely populated and therefore it is more likely that a flash flood will be seen and reported than in other portions of the CONUS. Since part of the training in the first-guess field includes Local Storm Reports (LSRs), an area that often gets numerous LSRs is seen by the model as a location that floods easier than other locations and thus the region has a greater chance to be flagged for a risk of flooding.

Comparing the NSSL EROs to one another it can be seen that the change in training between NSSL1 and NSSL2 led to an increase in both the probability of being in Marginal and a Slight Risk. This increase seen in the NSSL2 compared to the NSSL1 is particularly noticeable across the Upper Mississippi Valley and the Carolinas for the Marginal Risk. The NSSL2 product was also able to identify the risk of excessive rainfall associated with the Southwest Monsoon. One region however where the NSSL2 might have been over zealous compared to the NSSL1 (using the WPC ERO as baseline) was across the Carolinas, especially for the Marginal Risk. In Fig. 46E it can be seen that the probability of being in a Marginal Risk across northern SC and central NC is around 60% in the NSSL2 but around 30% in the NSSL1 and the WPC forecast. This suggests that although overall the NSSL2 was an improvement over NSSL1, there are still some biases in the NSSL2 that need to be addressed. Additionally, the speckled appearance to the risk contours that was seen in the ERO forecasts were prevalent enough to show through in the probability of being in a Marginal Risk image; refer to Fig. 46.

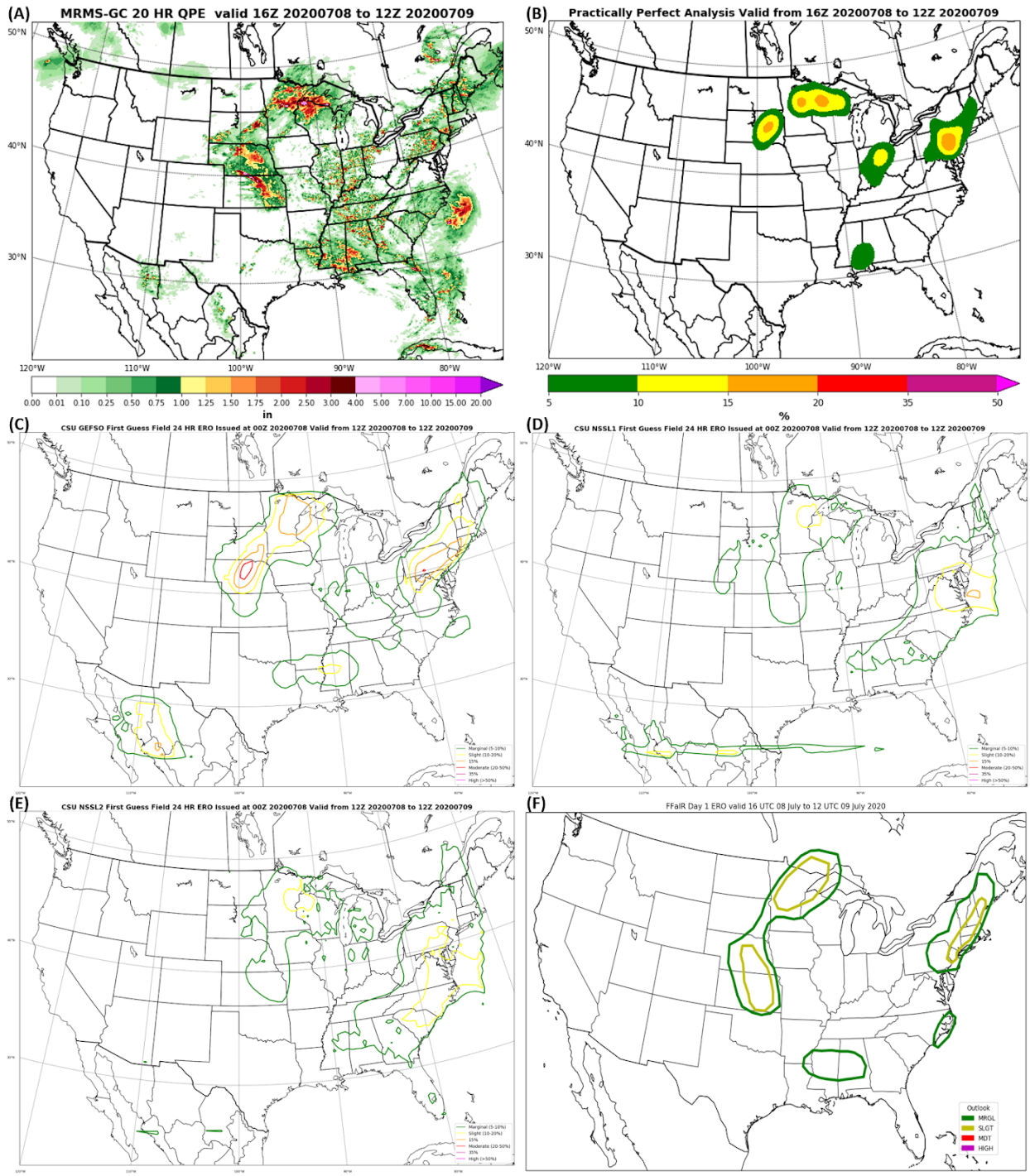

Figure 45: *(A) 24 h MRMS-GC QPE, (B) Practically Perfect analysis, (C) GEFS Day 1 ERO, (D) NSSL1 Day 1 ERO, and (E) NSSL2 Day 1 ERO valid 12 UTC 08 July to 12 UTC 09 July 2020. (F) FFaIR Day 1 ERO valid 16 UTC 08 July to 12 UTC 09 July 2020.*

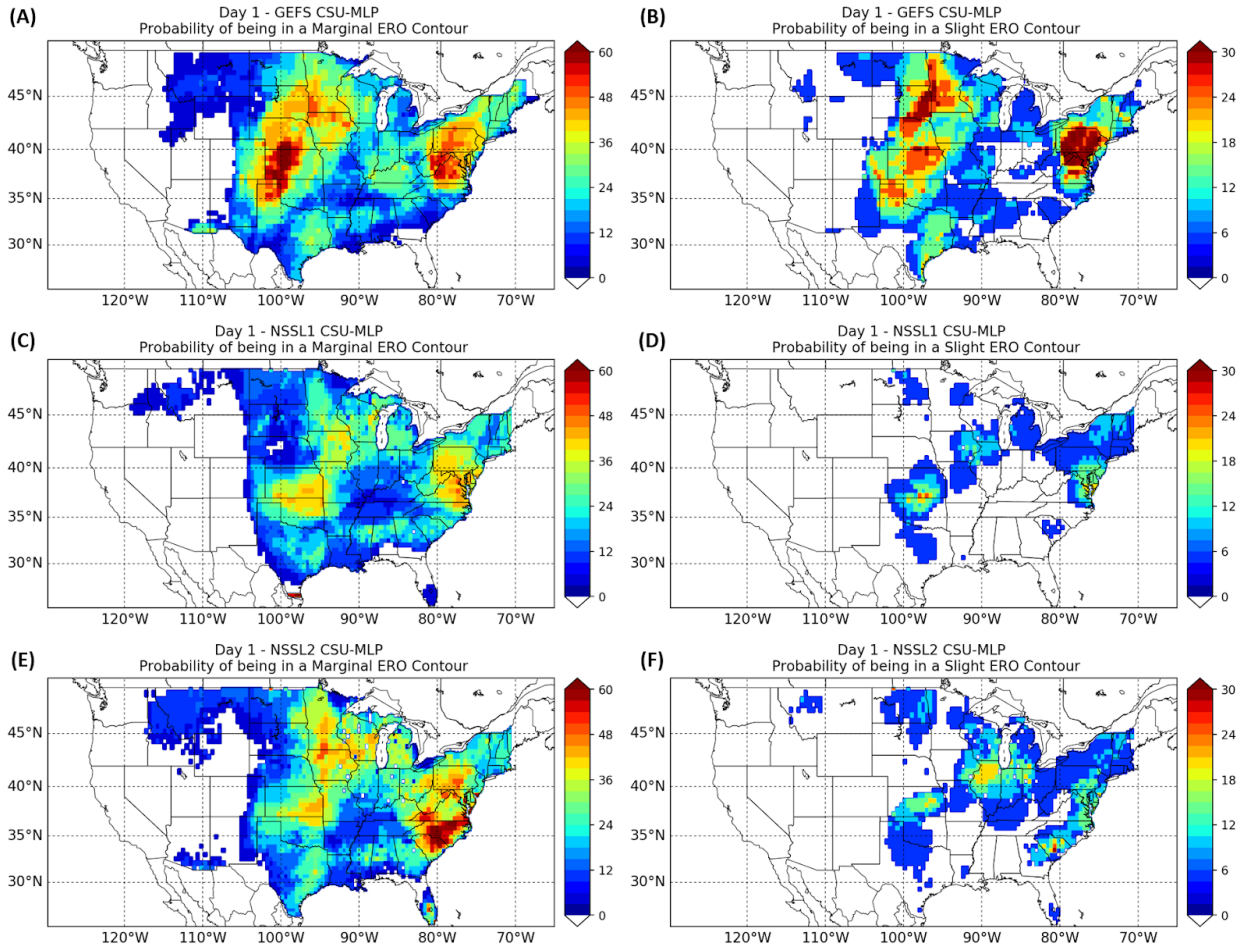

Figure 46: *Probability of being in a Day 1 ERO Marginal (left) or Slight (right) risk during the 2020 FFaIR Experiment in the CSU First-Guess Day 1 ERO (A)-(B) GEFS , (C)-(D) NSSL1, and (E)-(F) NSSL2.*

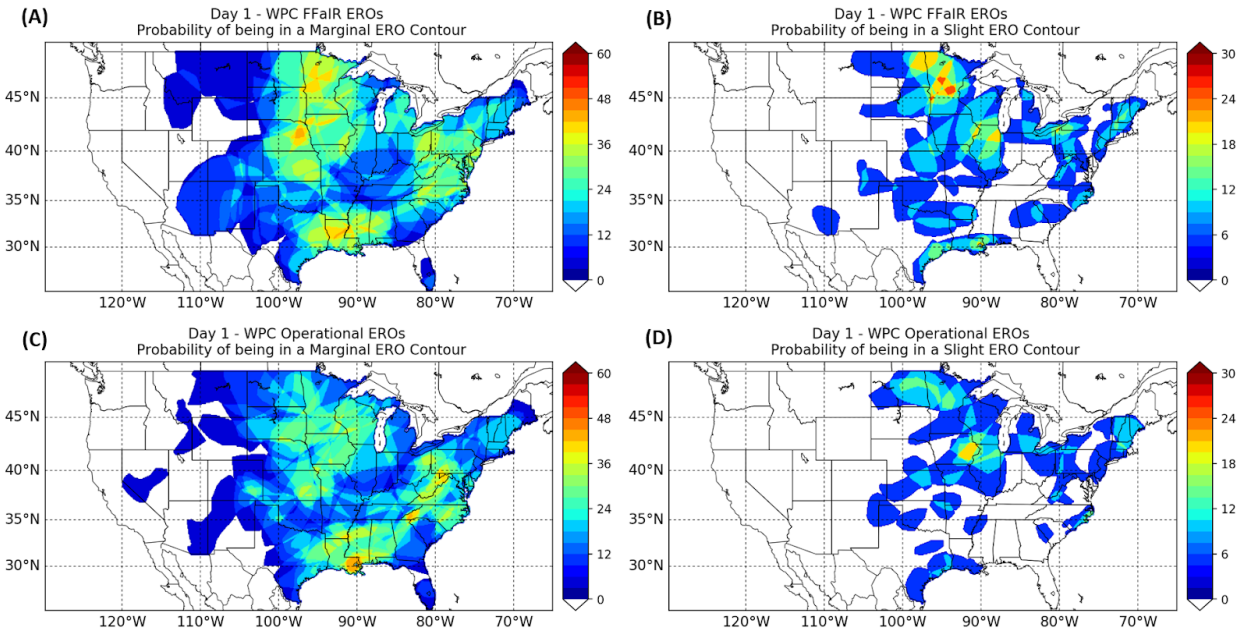

Figure 47: *Probability of being in a Day 1 ERO Marginal (left) or Slight (right) risk during the 2020 FFaIR Experiment for (A)-(B) FFaIR ERO and, (C)-(D) WPC Operational ERO.*

Figure 48 shows that although the NSSL2 ERO has a slightly lower Brier Score (BS) than the GEFS ERO, the GEFS ERO AuROC[20] is notebable higher than both of the NSSL EROs and comparable with both the FFaIR Experimental and the WPC Operational EROs. This means that although the NSSL2 ERO was slightly more accurate than the GEFS ERO during the experiment, the GEFS ERO did a better job distinguishing events from non-events. Additionally the daily Brier Skill Score (BSS), which used the WPC ERO as the reference forecast, showed that the GEFS ERO was more likely to "beat" the skill of the WPC ERO than either of the NSSL EROs were. The NSSL2 ERO had a better daily BSS during the duration of the experiment than the NSSL1 ERO. Lastly, the daily BSS trends indicated that the Experimental ERO had roughly the same skill as the Operational ERO, with just under half the days being more skilled than the reference forecast.

The changes made to the GEFS ERO from last year's GEFS Day 1 ERO have led to an improvement in the product and it should be under consideration to transition into operations. However, with the implementation of the new version of the ensemble, GEFSv12, re-training of the machine learning algorithm will likely need to be done and tested. Therefore it is difficult to recommend for operations until this concern is addressed. As for the changes made to the NSSL

[20] 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.

ERO from last year, the new version, NSSL2, performed better and should be the new configuration the CSU team uses as they continue to refine the CAM-scale first guess ERO.

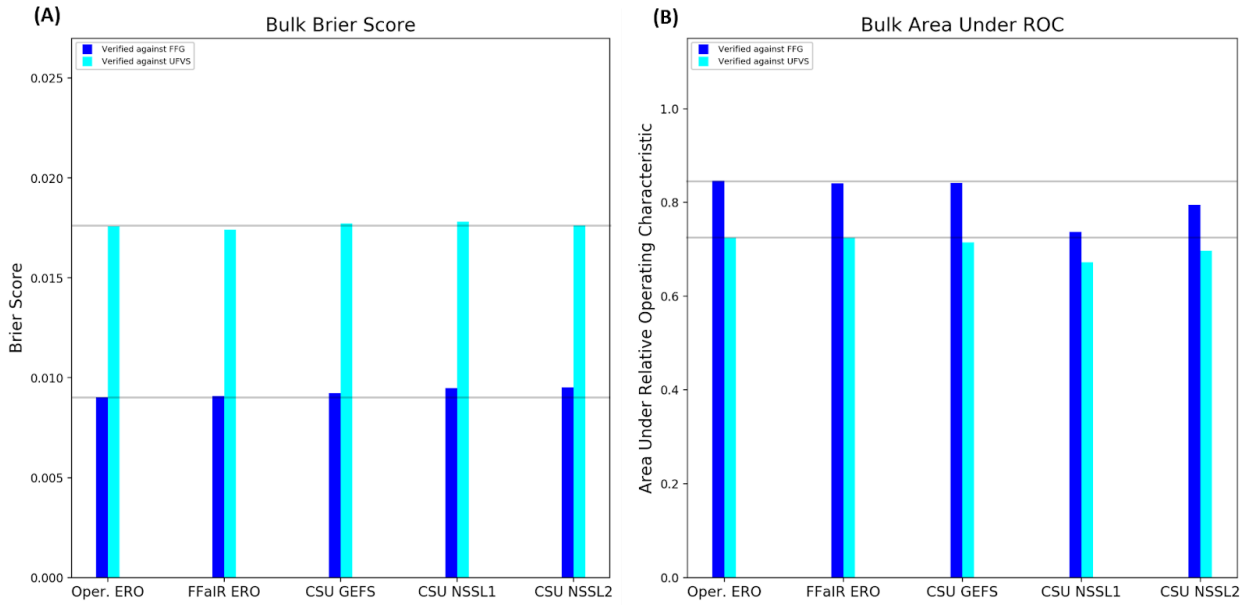

Figure 48: *(A) the bulk Brier Scores and (B) bulk AuROC for the WPC Operational ERO, the FFaIR experimental ERO, and the CSU-MLP EROs from the GEFS, NSSL1 and NSSL2. Each product was verified against FFG only and using the UFV system (QPE > FFG, QPE > ARI , and flash flood LSRs, flood LSRs, and USGS gauge reports).*

4.3.1 Intermediate ERO Risks

During FFaIR, the CSU First Guess ERO fields were plotted with additional probabilities of excessive rainfall contours, 15% and 35%; see Fig. 45C-E. These two probabilities are in the middle of the probabilities of exceeding FFG guidance for the Slight (10-20%) and Moderate (20-50%) risks. The 15% and 35% probability for excessive rainfall were plotted as an indicator to when a region is approaching the higher end of the probabilities of either the Slight or Moderate risk and can be thought of as intermediate risk identifiers. Feedback about the additional risk contours was positive and participants would like the products to continue to be plotted with the intermediate risk contours. However, it was noted that the intermediate risks should not be extended to the actual ERO issued by WPC. They felt that it would confuse the public, especially since the coloring is similar to the risks for the SPC Severe Weather Outlook.

4.4 End of the Week Survey Results

Some of the experimental guidance that is highlighted during FFaIR is not evaluated daily; instead the participants answer questions about the products at the end of the week after utilizing them in their forecasting exercises. The following subsections will highlight the takeaways from the participants' responses.

4.4.1 NBMv4 PQPF

Participants were asked if it was valuable to have access to multiple, high end PQPF percentiles from the NBMv4: the 90th, 95th, and 99th for 6h and 24h time intervals. The general consensus from the participants was that although they liked the PQPF products, three high percentile products was a bit too much. Many noted that they already have a plethora of information to look at when completing their forecasts and that they can get the information they need about the "worst case scenario" from the 90th percentile. Furthermore, they felt there was often little difference between the 90th and 95th percentile. Many also noted that although they found the product useful they did not find it particularly helpful in forecasting the flash flood risk, stating that often the signals were washed out in the product, lacking the small scale details. For instance, referring to Fig. 49 it can be seen that all three of the 24 h PQPF products have a large area of >1 inch of precipitation extending from NE to MN, suggesting a widespread threat for rainfall. However observations show that there were two separate regions of rainfall, with light to no rain between the maximums across the NE/KS region and northern WI.

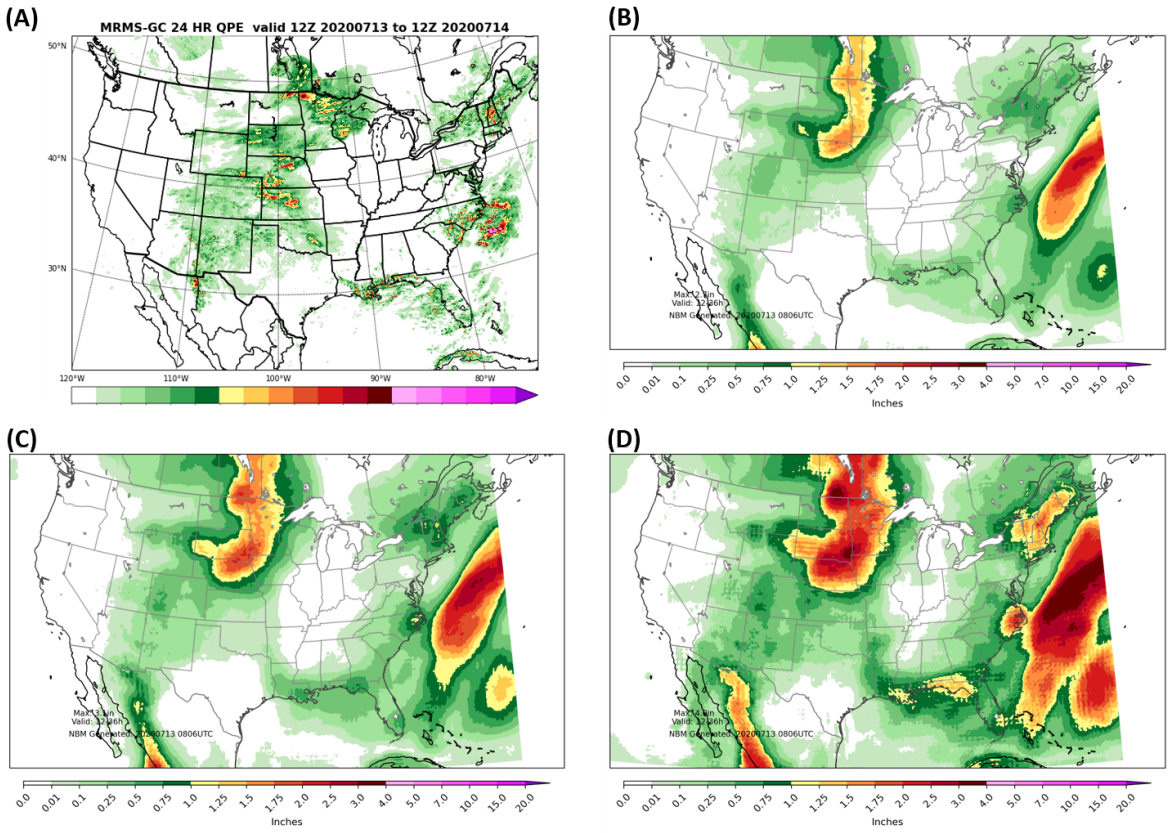

Figure 49: *(A) 24 h MRMS-GC QPE. The NBMv4 (B) 90th percentile, (C) 95th percentile, and (D) 95th percentile for 24 h precipitation totals. Valid 12 UTC 13 July to 12 UTC 14 July 2020.*

4.4.2 Experimental Heavy Rainfall and Object Tracker (HPOT)

HPOT[21] was developed to help convey the variability in the tracks of heavy rainfall regions among the members of an ensemble. Using the Model Evaluation Tool (MET) Method for Object-Based Diagnostic Evaluation (MODE) time-domain tool, heavy rainfall objects are tracked in space and time for each member, then the ensemble probability of being in a heavy rainfall object is calculated. This method was also applied to time-lagged deterministic ensembles (TLE).

Over the course of the experiment HPOT guidance was produced for the operational HREF, HRRRE, HRRRv3 TLE, and HRRRv4 TLE. HPOT was utilized during the collaborative drawing of the ERO and most participants mentioned using the product to help draw their MRTP. Additionally it was utilized during the drawing of the one Nowcast forecast exercise the Week 4 participants did for the Peoria, IL record rainfall, and this case will be used as an example for the utility of HPOT. The radar analysis and the HPOT guidance from the HREF,

[21] HPOT Website: https://origin.wpc.ncep.noaa.gov/verification/mtd/view.php

HRRRv3 TLE and HRRRv4 TLE, from 19 UTC 15 July to 00 UTC 16 July 2020, can be seen in Figs. 50-53 for the NowCast (valid 21-00UTC) forecasting activity.

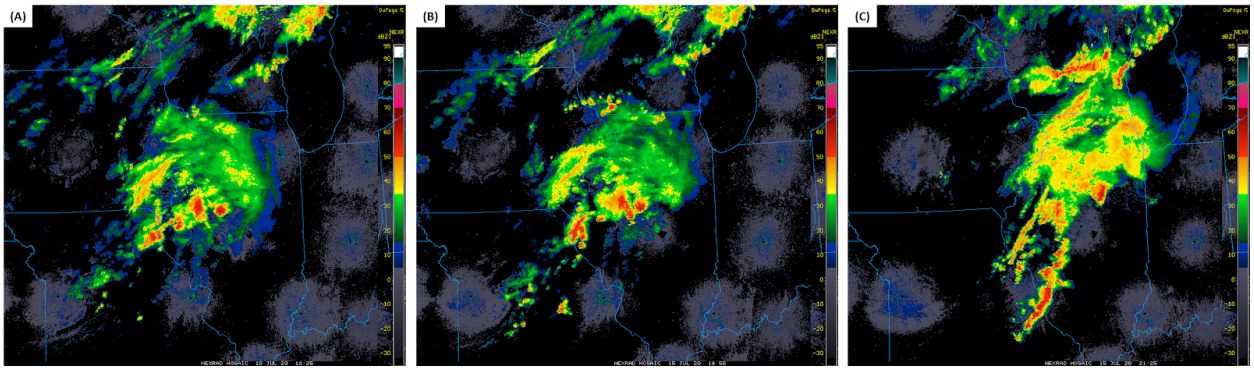

Figure 50: *Composite radar analysis valid at (A) 1830 UTC, (B) 19 UTC and (C) 23 UTC on 15 July 2020; courtesy of UCAR Image Archive.*

The observed reflectivity shows that by 19 UTC, the regions of heavier precipitation had moved out of northern MI and into west-central IL. However, the HPOT guidance valid 19 UTC for both the HREFv3 (initialized at 12z) and the HRRRv4-TLE (initialized at 15z) had the highest probabilities of being under a heavy rainfall object across northern MO and eastern IA. On the other hand the HRRRv3-TLE (initialized at 15z) had the highest probabilities over the MO/IA/IL border and into northwestern IL, closer to reality. As time progressed the HRRREv3-TLE kept the highest probabilities of heavy rainfall across central IL and south-southwestward into southeastern MO. Looking at the individual heavy precipitation objects identified (the symbols in the product), the HRRRv3-TLE had the heavy rainfall objects progressing southeastward across IL rather than having an eastward progression across north-central IL with a second line developing to the south; though again, the probabilities did suggest such an evolution.

Although the HREFv3 and HRRRv4-TLE HPOT guidance had a similar areal extent and location for the highest probabilities for heavy rainfall at 19 UTC, Fig. 51 and Fig. 53 show they diverged from one another as time progressed. The HREFv3 probabilities became more widespread and extended toward the east and northeast while the HRRRv4-TLE kept the highest probabilities farther south, over the St. Louis region and eastward. The heavy rainfall objects symbols from both the HREFv3 and the HRRRv4-TLE guidance showed that the northern objects were forecast to progress northeastward while the more southern objects would continue on a more eastern path. However, this split progression was less apparent in the HRRRv4-TLE, due to smaller membership. The intensity of the heavy rainfall objects, identified by the color of the symbol with the color indicates the 90th percentile of object intensity[22], however, differed between the two. The HRRRv4-TLE object intensity for most of the objects and timesteps was

[22] For the intensity scale of the heavy rainfall object refer to the legend on the left side of the product.

generally less than a half inch an hour while the HREFv3 object intensity was generally greater than a half inch an hour, with most of the southern objects approaching or exceeding one inch an hour.

Overall, the HPOT guidance for the HRRR TLEs kept the greatest threat for heavy rainfall farther south than observed, though the HRRRv4-TLE (Fig. 53) had a better forecast for the progression of the heavy rainfall objects than the HRRRv3-TLE (Fig. 52). Additionally, although both had the threat too far south, the HRRRv4-TLE had a higher probability for heavy rainfall across the Peoria region than the HRRRv3-TLE. Lastly, even though the HREFv3 guidance extended probabilities for heavy rainfall further north than the HRRR forecasts, it generally had probabilities less than 50% over the Peoria area. The lower probabilities were likely due to the greater spread in the members.

Feedback from the participants about the HPOT product was overwhelmingly positive. Many of the participants discussed that they liked that the tool not only conveys the heavy rainfall threat but also that it helped them easily see the spread in the members and the differences in rainfall intensity in each member. For instance, one participant noted: “Best thing about this is that it puts probabilities along with centroids. The existence of "objects" in a proposed outlook area seemed to add to my and the groups confidence during most of the ERO forecast exercises.” This merger of information, however, was found to be difficult to decipher for some of the participants, though some that noted this said they felt they would like the product more with additional training and exposure to it.

Another comment was about the thresholds used to identify heavy rainfall objects. Currently the guidance uses 0.05” or 0.1” hourly rates within a precipitation object to identify heavy rainfall objects (the user chooses the threshold). Many participants noted this seemed like a low threshold and that they would like to see higher thresholds as identifiers. Although this threshold seems small, developers performed sensitivity studies by varying the convolution radius and rainfall thresholds. In general, due to the smoothing of heavy precipitation objects and a minimum object size requirement, the maximum intensity of objects found with a threshold of 0.1 inches per hour usually exceeds 1 inch per hour at some point in the object's evolution. In other words, users should not focus on the threshold used to identify the objects, but instead the 90th percentile of object intensity. A solution to the suggestion by the participant would instead be to add a threshold to 0.2” per hour, which would miss marginal heavy precipitation objects, but still identify the more intense heavy rain events.

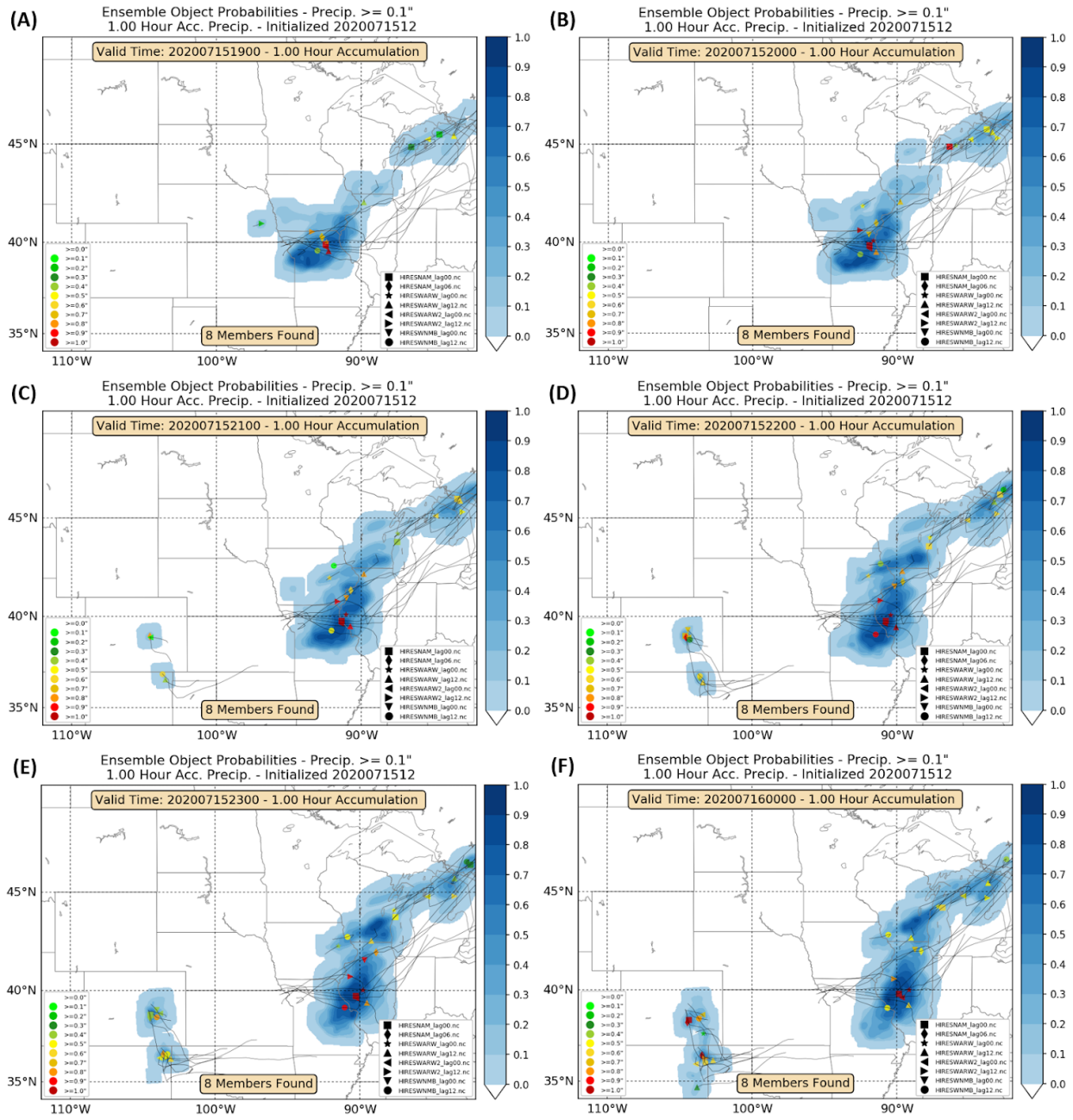

Figure 51: *HPOT for the operational 12z HREF valid (A) 19 UTC, (B) 20 UTC, (C) 21 UTC, (D) 22 UTC, and (E) 23 UTC 15 July 2020. Right side of each image is the probability of being in a heavy rainfall object, bottom right corner is the legend for the symbol associated with each member, and left bottom corner is the legend for the 90th percentile of object intensity.*

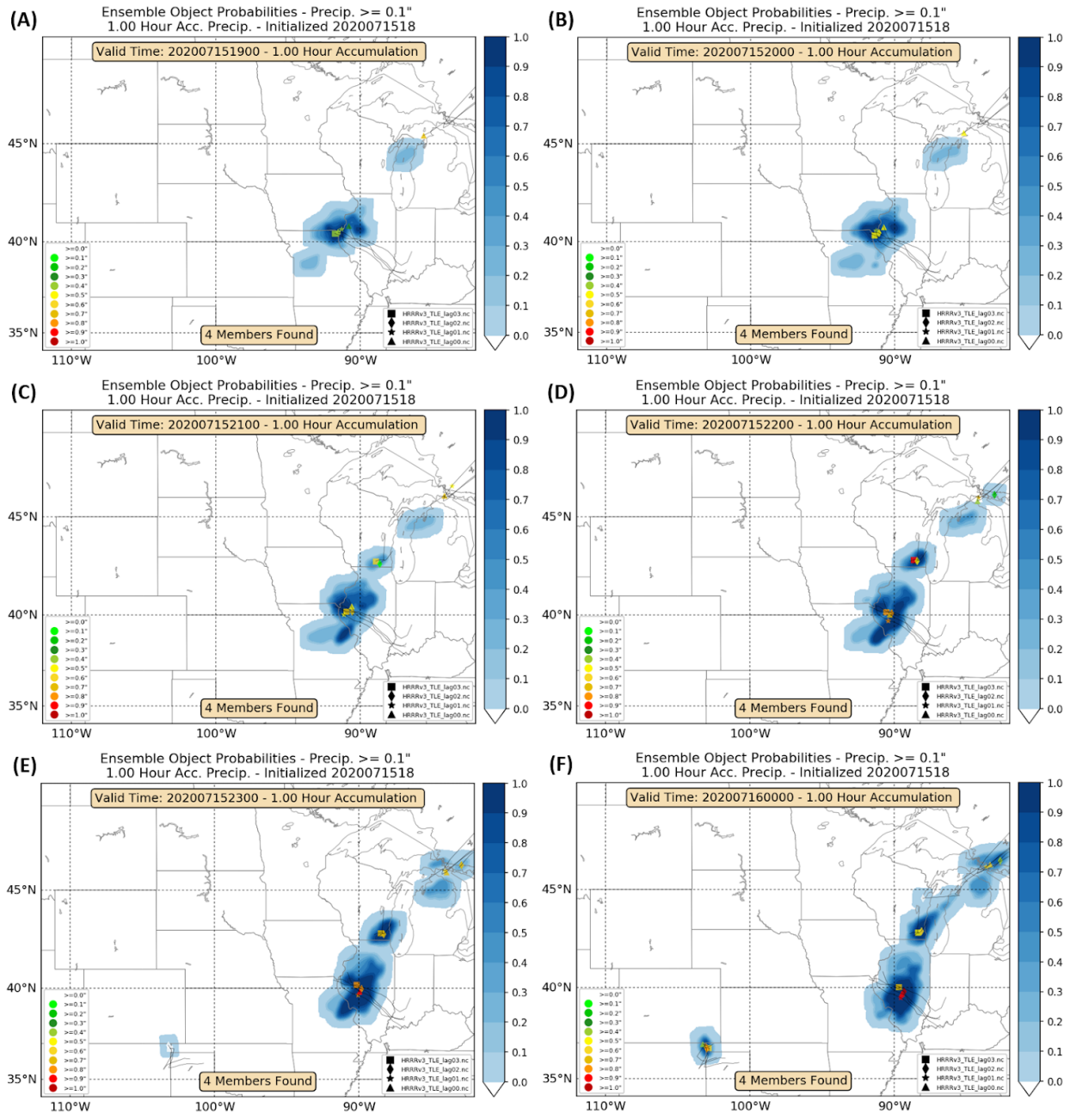

Figure 52: *Same as Fig. 51 but for HRRRv3-TLE.*

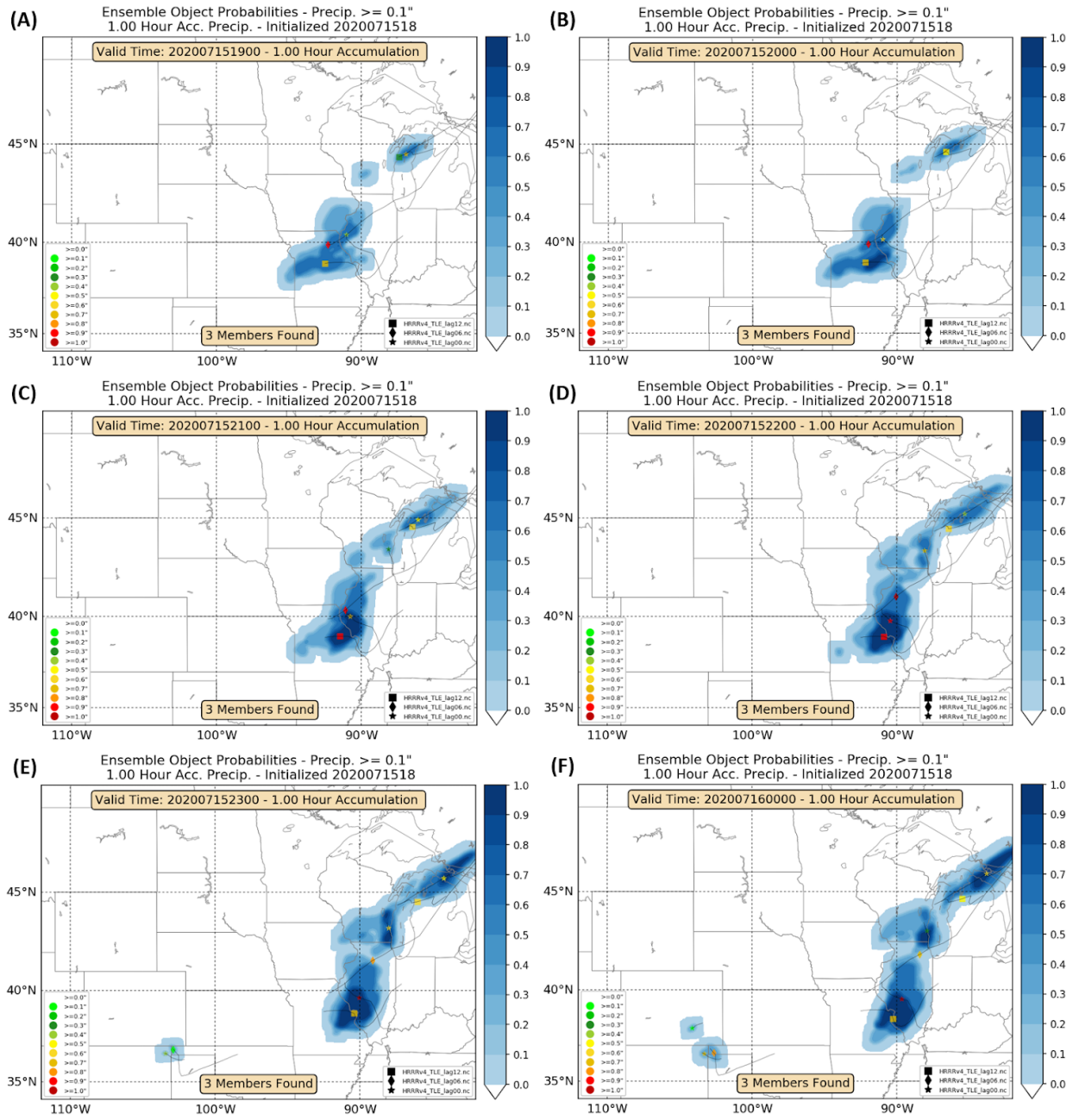

Figure 53: *Same as Fig. 51 but for HRRRv4-TLE.*

4.4.3 National Water Model Visualization Services v1.2

The National Water Center (NWC) has developed a GIS-based tool that is hosted on the web to view the output from the National Water Model (NWM). The NWC Visualization Services version 1.2 (hereafter WaterViewer v1.2) was introduced to the participants as a hydrologic tool to use to help identify the hydrological response to excessive rainfall, specifically to identify the risk for flash flooding. However, because the WaterViewer v1.2 is still in active development and only available to a small portion of the NWS, the participants themselves were

not able to individually utilize the tool. Instead the FFaIR facilitator led the group through the hydrological analysis from the products available in the WaterViewer v1.2. This included products such as the Bankfull[23] Arrival Time and the Streamflow Rate of Change.

Overall feedback about the WaterViewer v1.2 and its products were mixed. Although participants commented they felt there could be utility, the usability isn't there yet. For instance, one participant said, "I think it was helpful, but the ease of using the NWM site I think is still challenging and requires a fair amount of dedicated time to not only navigate the site, but to analyze all of the information within it." Many participants noted the slowness of the WaterViewer v1.2 to load as an issue. Utility was also brought up as an issue due to the fact only one model input, HRRR, was used for the short term forecast. They stated that if the HRRR had the location of the precipitation wrong, the hydrologic response would also be incorrect. A few commented that they felt the utility lied more on a regional or national level, rather than on a WFO level. Most participants did however like the ability to see stream response in small rivers and streams. Finally, product wise, the Streamflow Rate of Change forecast was well liked by the participants since it helped to identify the "flashiness" of a river or basin. Participants said it was helpful to see the streamflow was changing rather than if it was just high or not because the streamflow might be high for reasons unrelated to QPF. Therefore just seeing that the streamflow was high was not enough information to identify the flash flooding threat but how the rate changed provided much more insight.

4.4.4 New Color Curve for QPF

As the NWS continues to work towards a more uniform communication of their graphics and products, one area of focus has been identifying color tables that are colorblind friendly. In March of 2019 the NWS released the proposed new standard color curves[24] and these guidelines were the basis for the QPF color curve used for the 2020 FFaIR Experiment. Difficulty in creating the color table arose due to the number of intervals WPC uses for their QPF compared to the number of intervals for the QPF color scale guidelines. To adjust for this, the suggested colors for the probability of ice were used for the higher QPF values. The color table used by WPC and FFaIR can be seen in Fig. 54. At the end of the week, the participants were asked for general comments about the new scale.

[23] "Bankfull" conditions are approximated by the 67% annual exceedance probability (AEP). Bankfull flows and AEPs were derived using a 25-year retrospective analysis of the NWM v2.0 (OWP NWM 2020).

[24] The Standard Color Curve Summary slides can be found at: https://docs.google.com/presentation/d/1UEGGXWYa7c7awcOFdczRAWBByq3Oi5LL2wYWBfimBlg/edit#slide=id.p24

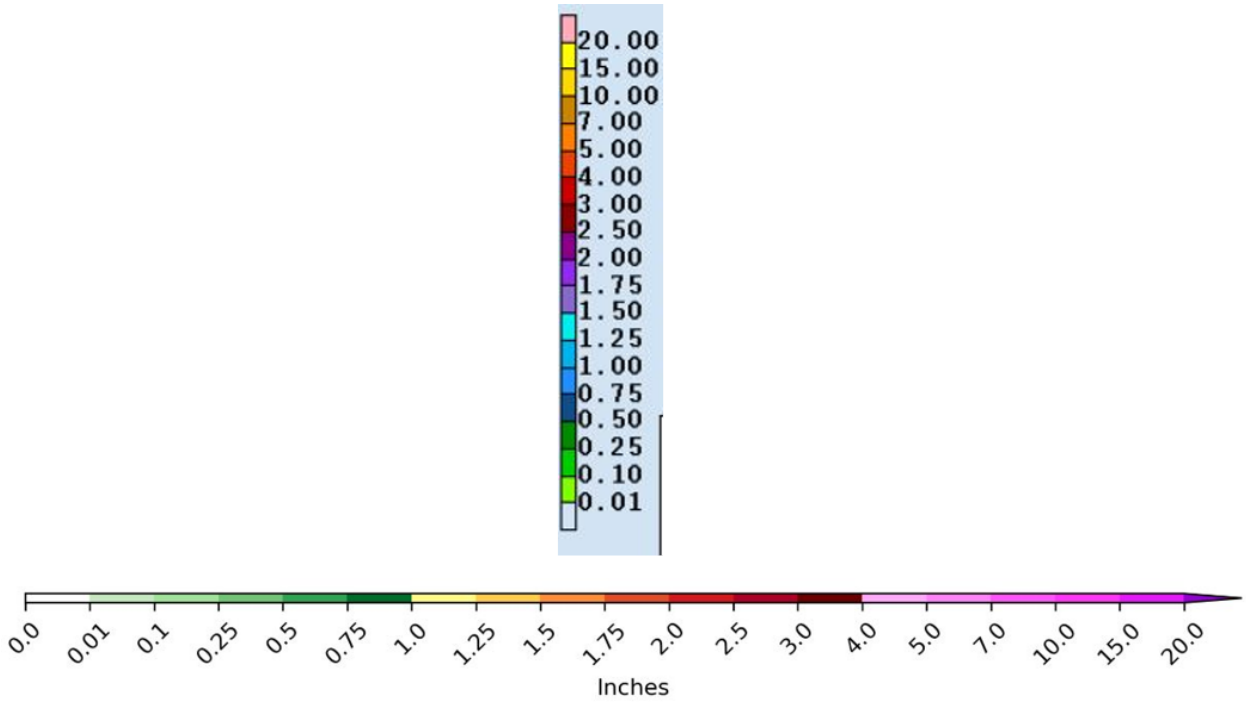

Figure 54: *Top: the operational WPC QPF legend. Bottom: the experimental color blind friendly QPF legend used in the 2020 FFaIR Experiment.*

Overall the feedback was positive, though most of the participants noted they were not colorblind so they could not say whether or not the change was beneficial to those who are. A constant praise of the color scale was the distinct color change that made it easy to quickly discern areas of higher precipitation. The change from green to yellow for 1 inch was the most discussed as a change that helped guide the eye to areas of concern. Some participants did note however that some of the pink colors were hard to determine from one another. Given that it was within the 4+ inch intervals that the colors were altered from the guidelines this comment was not surprising. Additionally come participants noted that the color scale reminded them of the colors used for reflectivity, though they did state that fact did not mean they did not like the scale. Lastly, there was overwhelming support of a universal NWS QPF scale and the experimental color scale is a step in the right direction.

4.5 Maximum Rainfall and Timing Product (MRTP)

The goal of the MRTP exercise was to use the experimental guidance in the forecasting of extreme precipitation and its ingredients. As stated in Section 2.2, MRTPs were issued daily with each participant creating their own product. The location and valid time period of the MRTP however was determined as a group; the time period and location of each MRTP and NowCast can be found in Appendix C Table C.2. In addition to the interactive activity of drawing precipitation contours of 1 inch or greater, participants completed a survey during the forecast, answering questions as to the value of the maximum rain in the forecast domain, the

maximum duration of the rain in that time period, the models or ensembles they were assigned, and which models/ensemble influenced their forecast decisions; the survey can be seen in Fig. 55 . The forecast data was combined with the survey data and a number of analyses were produced to assess forecasters behavior and performance during the 20 forecast days of FFaIR. Experimental models were also evaluated post-experiment to compare with forecasters. In general, because the model/ensemble QPF data was shown as a static image the forecasters did not have the ability to click on the images to determine the specific accumulation value at a point nor duration information. In this way, the model images were generally helpful but not specifically designed for forecasters to directly use any of the calculations in their own forecast.

4.5.1 MRTP Case Study 17-18 July 2020 MCS

The MRTP forecast on the last day of the experiment was valid from 21 UTC July 17 to 00 UTC July 18, 2020. As can be seen in Fig. 56, the forecast concern involved a threat for a rapidly developing squall line transitioning into a MCS. One of the key forecast challenges was predicting this MCS and deriving information about its initiation, spatial coverage, propagation speed, and the potential for convective storm training on the southwest side. Many participants questioned the ability of the MCS to produce a coherent 1 inch rainfall swath given its forward speed, which was predicted to be near 50 mph.

Focusing on the forecast from the HRRRv4, not shown, the HRRRv4 initialized at 00z suggested that the 6 h QPF greater than 1 inch would be focused in North Dakota and far western Minnesota. Meanwhile, the 12z HRRRv4 forecast trended slightly further south and east. Comparing these two model forecasts leading up to and at 21 UTC to observations, the 00z guidance was considerably better in depicting the initiation, extent, and orientation of the MCS than the 12z guidance. This was due to the 12z guidance both initiating convection too late and developing the MCS slower. However, despite the slower initiation process seen in the 12z model, both HRRRv4 forecasts correctly matched the western border of the intense rainfall area. Even though the 00z run better depicted the development of the system, by 06 UTC (after the MRTP was no longer valid) it is clear that both HRRR forecasts are about 3 or 4 hours behind, consistent with the lack of organization at 21 UTC.

Examination of the hourly QPF from the two HRRRv4 runs (not shown) revealed that the largest hourly rates were in the convective leading line. The large accumulation amounts forecasted by the models was the result of convective cell training in the southwest and western portions of the domain combined with the overlap of convective and stratiform precipitation to the central and northern portions of the MCS. Training along the outflow at and near the apex of the bow later in the lifecycle of the MCS also helped to drive the QPF amounts seen. Figures 57 and 58 show that when modeled and observed precipitation rates were evaluated at thresholds of 0.1, 0.25, and 1 inch for various durations (1-6 hours), the modeled precipitation rates do not

compare favorably to observations at longer durations. For example, the HRRRv4 forecasts are missing longer duration occurrences but are capturing some aspects of the rates at higher thresholds and the lowest rainfall durations.

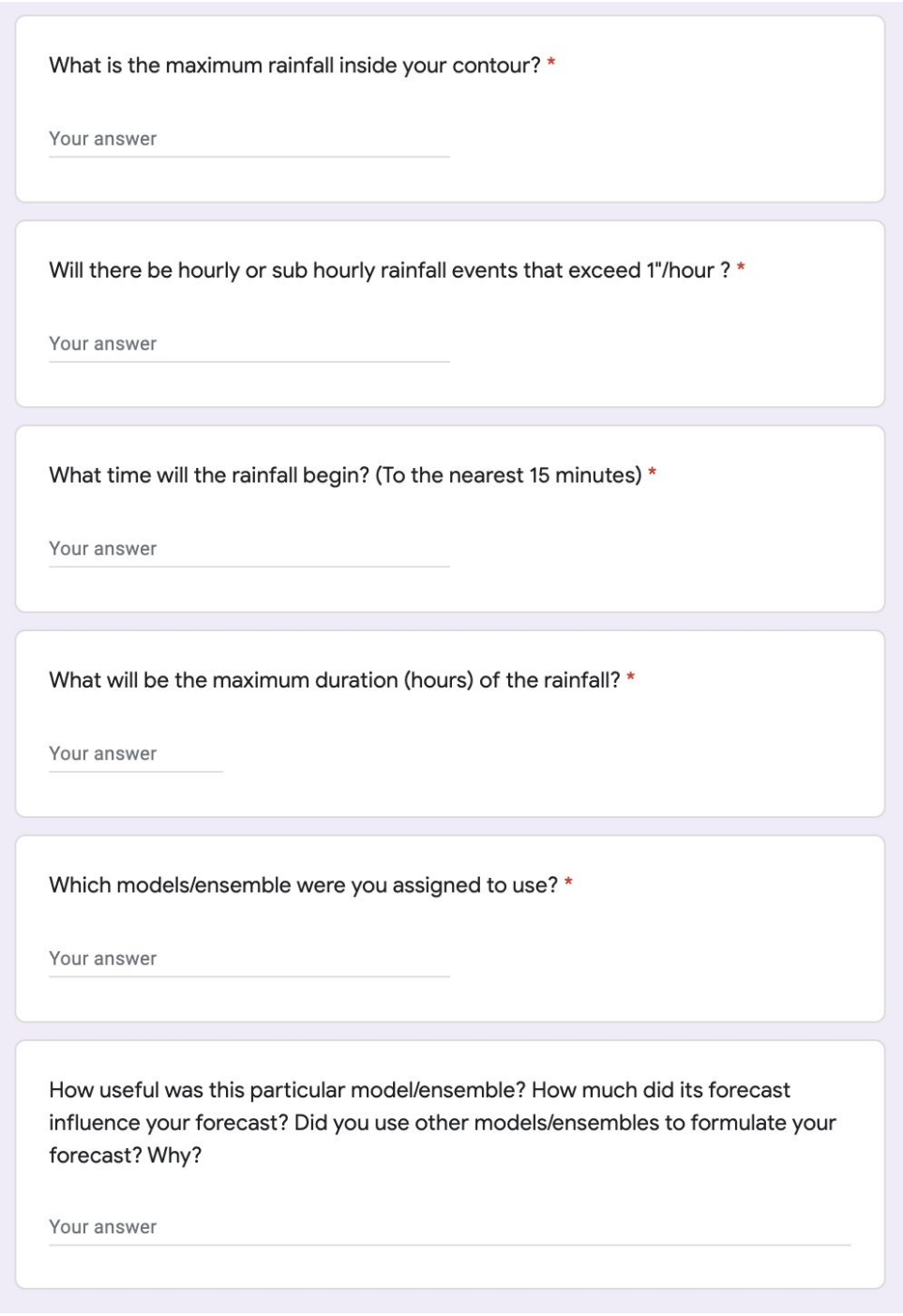

Figure 55: *The questions asked in the MRTP survey the participants filled out as they did the MRTP forecast exercise each day.*

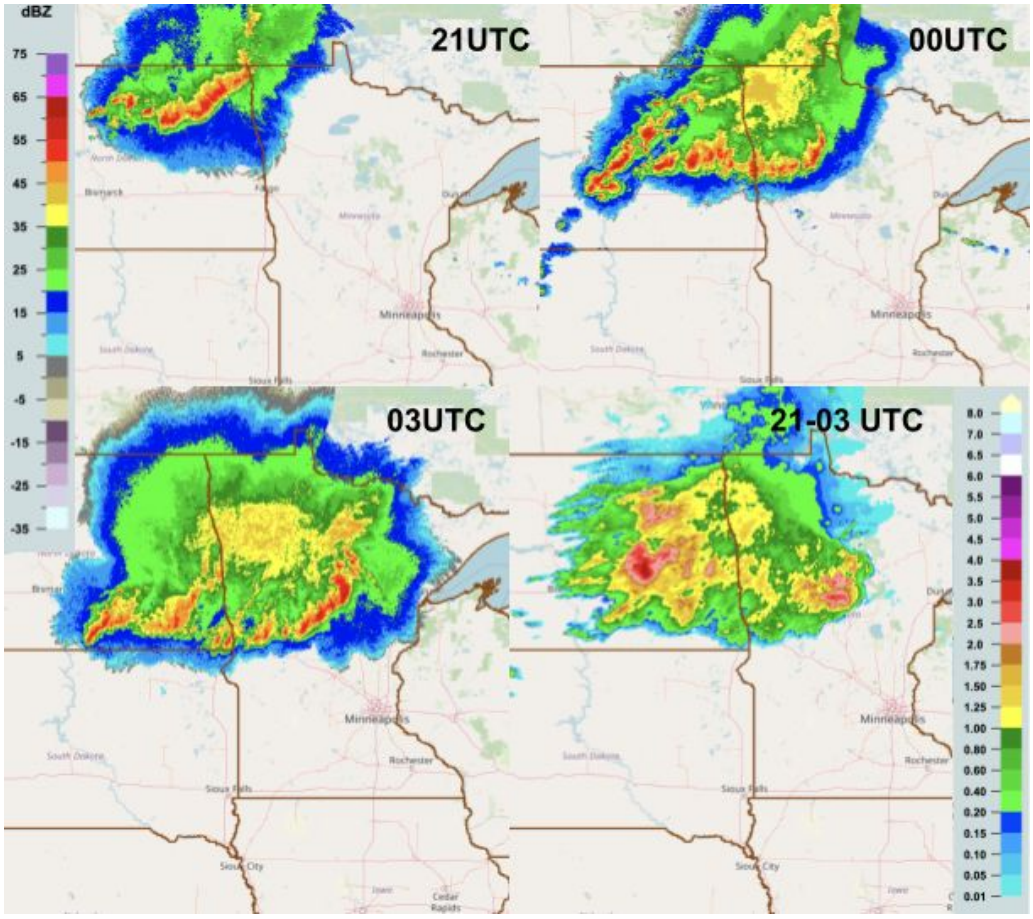

Figure 56: *MRMSv12 reflectivity for the 17-18 July 2020 MCS case at 21, 00, and 03 UTC along with the accumulated precipitation over the 6 hour period from 21 UTC July 17 to 00 UTC July 18, 2020.*

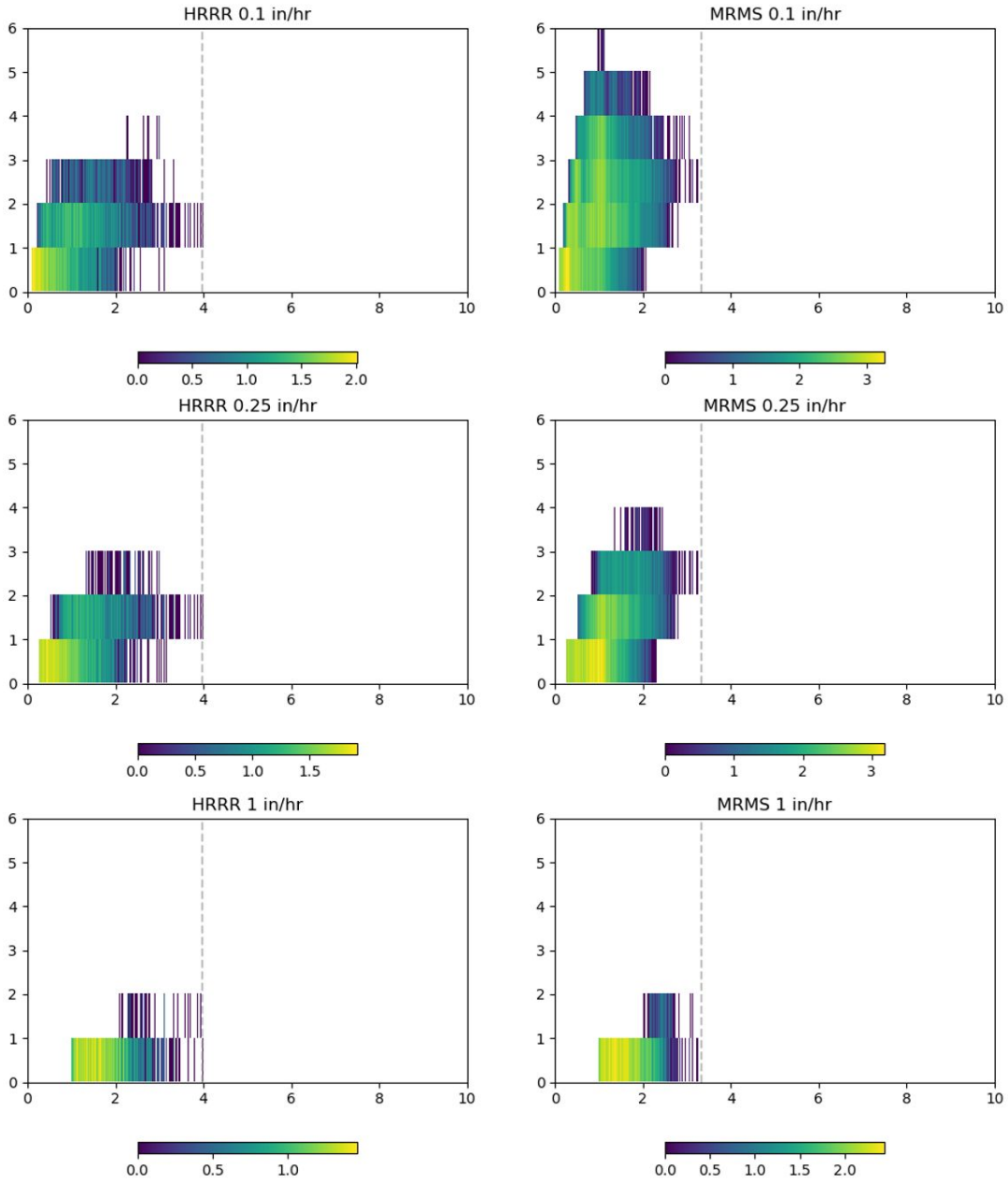

Figure 57: *Two dimensional histograms comparing HRRRv4 initialized at 00z and MRMS for 6 h accumulations on x-axis and rainfall rate durations (0.1 inch/h top, 0.25 inch/h middle, 1 inch/h bottom) on the y-axis valid 21 UTC July 17 to 03 UTC July 18, 2020.. Maximum rainfall for the 6 h period is shown by the dashed vertical line.*

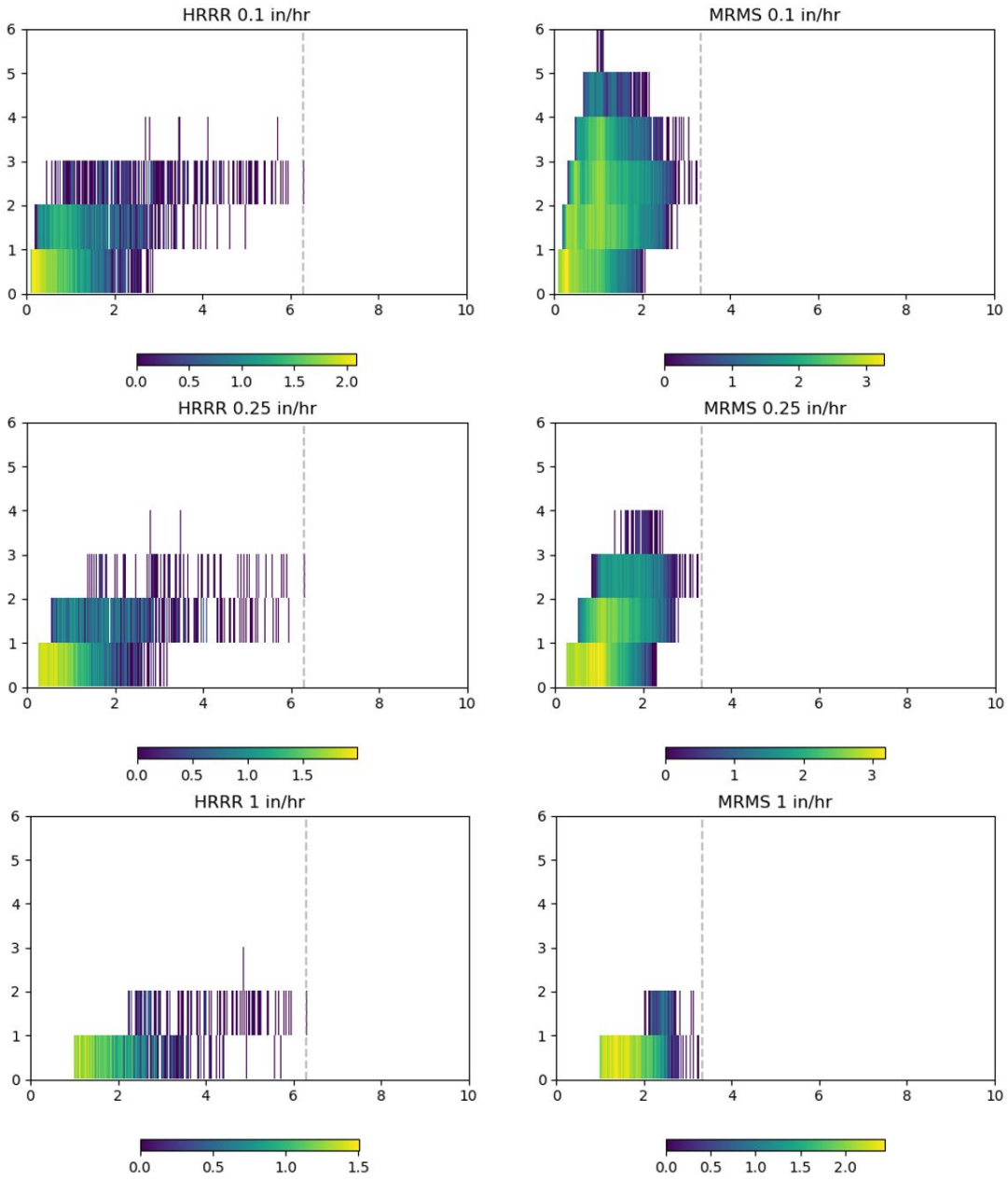

Figure 58: *Same as Fig. 57 but comparing HRRRv4 initialized at 12z to MRMS.*

There were numerous instances of observed convective cell clusters that initiated west-southwest of the MCS and led to training convective cells following the outflow southward. This continuous generation along the southwestern flank of the MCS resulted in the heaviest precipitation being observed along the southwestern portion of the MCS, see Fig. 56. As the MCS increased in size and moved eastward a concentrated area of precipitation formed in central Minnesota, associated with a developing mesoscale convective vortex. Around this area, cell motions were east-northeast and resulted in training along and rearward of the convective line. The HRRRv4 did not forecast this evolution, instead it kept the largest rainfall accumulations to

the east-southeast of the MCS rather than to the east-northeast after 03 UTC. Interestingly, later in the observed MCS the rainfall pattern mimics the HRRRv4 but is 3-6 hours later than forecasted (not shown). The realism depicted in the HRRRv4 forecasts was hard to deny but the timing of features was a different story, ranging from too slow to develop to too quick to evolve. The struggle to simulate specific aspects of MCS evolution could be due to any number of reasons from resolution issues to physics parameterization challenges.

Focusing on the MRTPs issued by the forecasters (Fig. 59), they largely agreed on the forecast areas, with little variation in the southward extent in North Dakota but did vary on their placement of the eastern edge of the 1inch isohyet. Since the MCS organized earlier than models forecast, forecasters improved upon model predictions and expanded their 1 inch isohyets farther east than most models forecasted. While most forecasters placed their maximum rainfall locations to the north, where both precipitation from the leading line and trailing stratiform precipitation would occur. However the observed maximum actually occurred to the west where early convective cell training occurred (not shown). In Minnesota, another local maximum emerged from convective cell training near the leading line, due to the formation of an MCV. About six forecasters correctly extended their 1 inch area to cover this region; see Fig. 60. In summary, there are many ways to accumulate precipitation and this event showcased all of them in a fast moving MCS. Thus proving the old adage, “the heaviest precipitation occurs where the rainfall rate is the highest for the longest time” (Doswell et al. 1996).

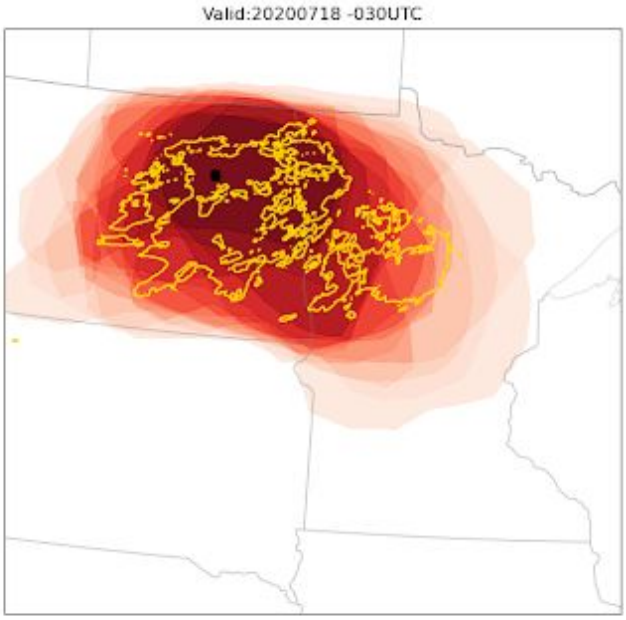

Figure 59: *Forecast domain for 17-18 July 2020 MCS case study, depicting the ensemble of human forecasts (shaded red) with the MRMS 1 inch rainfall contour (yellow).*

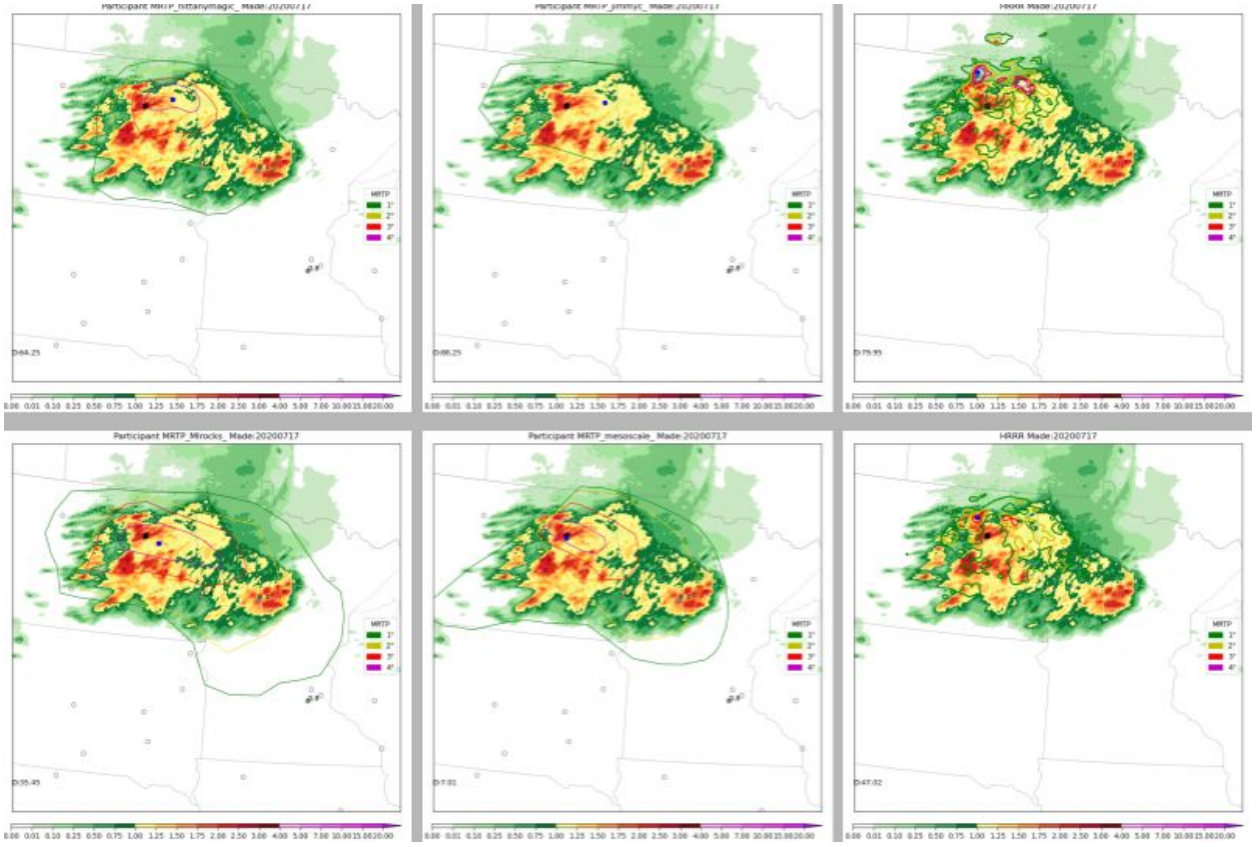

Figure 60: *MRTP forecasts (contours according to legend) from WPC (upper left), Jimmy (upper middle), HRRRv4 00z (upper right), Sarah (lower left), participant "mesoscale" (lower middle; shortest distance to maximum), and HRRRv4 12z (lower right) with MRMS precipitation shaded according to lower colorbar.*

4.5.2 Daily and Aggregate Statistics

Statistical analysis of the MRTPs included performance diagrams for each day of the experiment were constructed by mapping each forecaster's 1 inch contour to the MRMS 1 km grid and deriving contingency table elements. Additionally the distance from each forecaster's point maximum to the MRMS maximum rainfall was derived off this grid. The rainfall duration was computed using 0.10, 0.25, 1, and 2 inch per hour precipitation thresholds as an hourly exceedance of rainfall. These were then summed over a 3 or 6 hour period using hourly MRMS-GC QPE for the area that received 1 inch or greater total precipitation. The following discussion compares the duration for the 0.25 inch per hour exceedance, which generally agrees with the 0.10 inch per hour threshold to within one hour.

To begin the forecasters' ensemble of forecasts compared to the 1 inch MRMS contours for each of the 20 days were examined. As can be seen in Fig. 61 the experiment covered a variety of geographical areas and phenomena, ranging from tropical storms and MCSs to individual bouts of severe convection, mesoscale convective vortices, and squall lines. Most

MRTP were valid for a 6 h period and generally ended at or prior to 6 UTC. Some of the most accurate days were when the participants were tasked with forecasting for overnight MCSs, likely due to the nature of MCSs having larger areal coverage of precipitation than other phenomena.

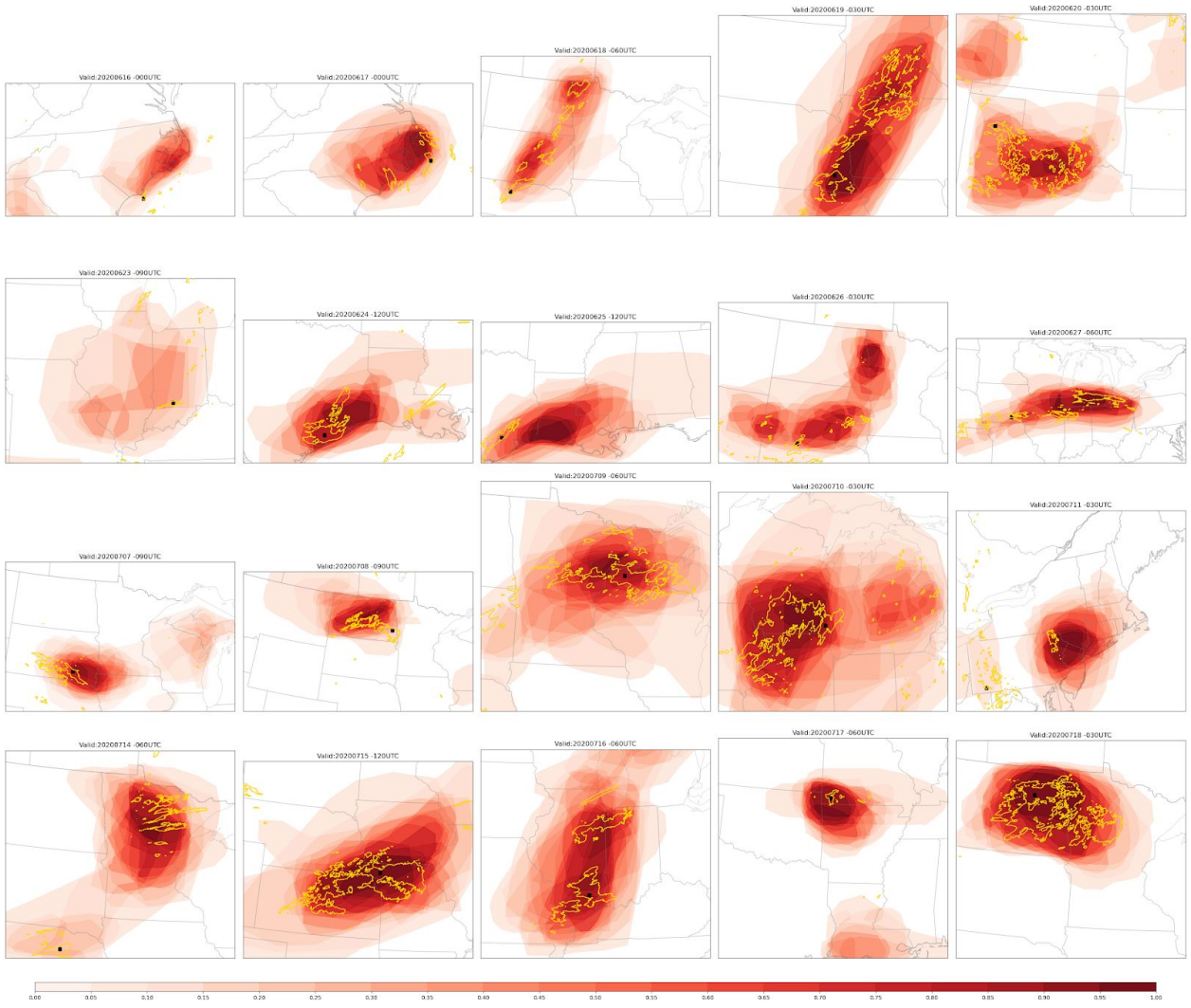

Figure 61: *MRTP human ensemble of forecast areas (red shading, normalized by number of forecasters per day), with MRMS 1 inch contour (yellow).*

Examples of the eight best daily forecasts, based on minimum distance to the observed maximum rainfall location, are shown in Fig. 62. Participants utilized a strategy of drawing a large generally smooth 1 inch contour, attempting to find the right scale to draw the forecast that encompassed the most events rather than drawing many objects to capture individual storms. The placement of maximum rainfall forecast points were consistent with the forecasts, but varied widely based on the guidance being used to inform these forecasts. Only in a few events, most of them uncertain, did the scale of the forecasts not relate to the scale of the observed rainfall as

depicted in the forecast ensemble above. Thus the challenge for participants was to recognize and account for the scale of the event, and determine where the maximum in rainfall would emerge.

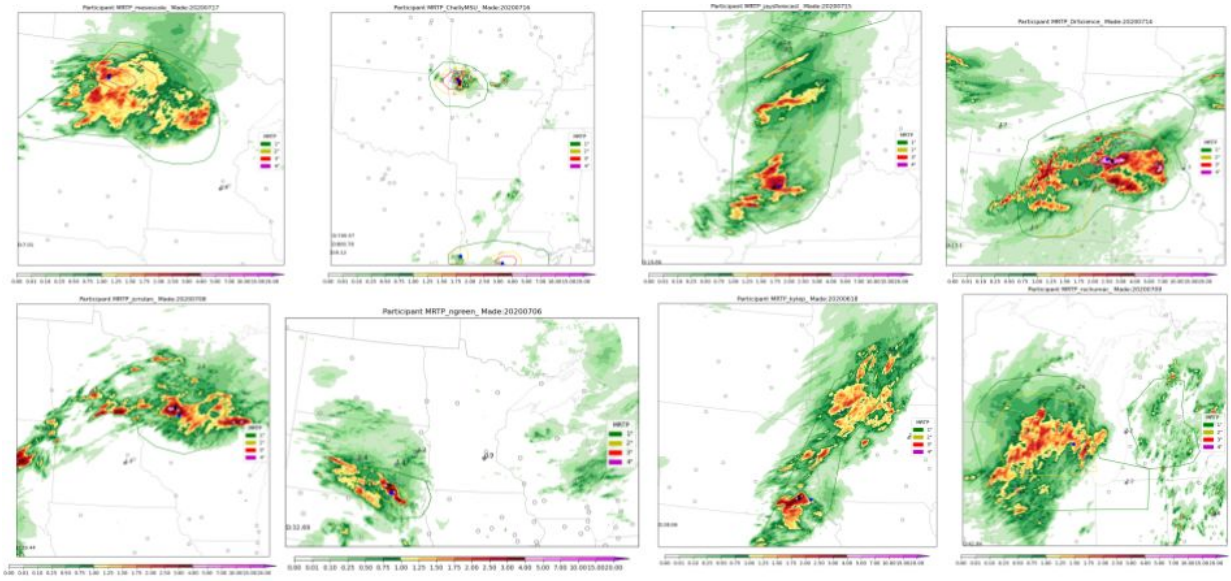

Figure 62: *Eight of the best MRTP forecasts with respect to minimum distance from their forecasted maximum rainfall location to the observed maximum. These eight cases ranged from 7 to 42km away. Each of the eight best MRTP maximum rainfall location forecasts shown is for a different event during the experiment.*

Performance diagrams for each day a MRTP was issued are shown in Fig. 63, with forecasters in blue dots and models specified according to the legend, broken up by model cycle (square 00z, triangle 12z). Forecasters drew much larger contours of 1 inch QPF than models forecasted, thus resulting in substantial wet bias in the MRTP forecasts. However it also resulted in relatively high probability of detection (POD), though a reduced success ratio (SR). Daily, models rarely breached PODs of 0.5, while critical success index (CSI) values are generally consistent with the bulk of forecasters. Note that forecasters seldom had complete access to 12z guidance by the deadline for product submission. Despite this, the top forecasters of the day generally had CSI approaching or better than the top models.

In general, the performance diagrams show that when the models do well, the forecasters do well. This can be seen in the Fig. 63 in the daily variation SR, where for the most part, when SR was higher for models it was also higher for the forecasters. The positive movement (from low SR to high SR) is probably for a variety of reasons, most of which center on the increased predictability for events that are synoptically driven. Trying to control for maximum rain or area, it was evident that areal extent was more important for predictability to the participants. Our collective forecast experience and discussions in these events was that there would be an important forecastable event in these time periods. The fact that forecasters tended to have high

POD and low SR in the smallest events speaks to that approach. That performance is a high bias and low SR leads us to conclude that these events require precision that is not currently achievable with this suite of guidance.

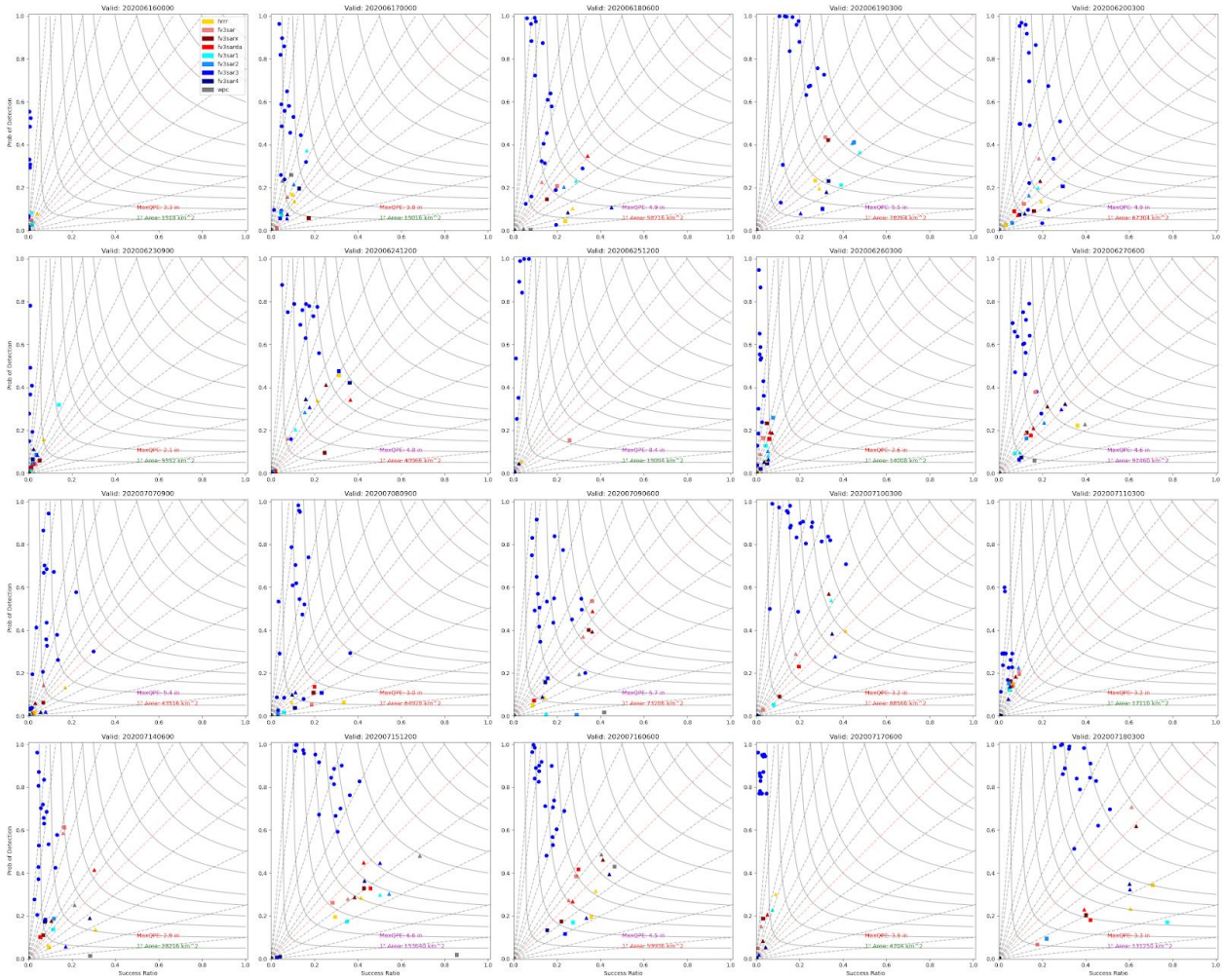

Figure 63: *Daily performance diagrams from MRTP with blue dots representing each forecaster, models colored according to the legend with 00z (squares) and 12z (triangles) represented separately. The Areal coverage in thousands of square km, and the maximum accumulation from MRMSv11 is shown. The quantities are both color coded: Area 0-30k (green), 30-90k (red), 90k+ (magenta) and rainfall 1-2 inches (green), 2-4 inches (red), 4+ inches (magenta).*

Like was seen with the MRTP forecasts, model performance varied based on event type (MCS, tropical cyclone, etc), especially in relation to the size of the precipitation area. On the largest coverage days, model performance tends towards a bias around 1 (near equal POD and SR) but rarely did models exceed a POD of 0.5, while for the smallest coverage days model bias is much higher and thus SR much lower. The largest areal coverage days may be some of the more predictable events while some of the most challenging days typically occur when the precipitation accumulations are relatively large but the areal coverage is smallest. This latter

point might mean that these small events are either not predicted in the models, significantly displaced (in time or space), or precipitation is not as extreme as forecasted.

In aggregate statistics, model forecasts across the 20 days were compared with three forecasters who each forecasted for 19+ days of FFAIR, refer to Fig. 64. The bulk statistics for models at 00z (squares) show lower skill than the 12z guidance with somewhat similar bias. Forecasters generally had similar CSI to most of the 12z guidance with much higher POD. In general, it appears that the individual model guidance was comparable in area to observations since their bias is generally around 1. Additionally, most models saw an increase in their bias when comparing the 00z model initialization to the 12z, which improved their accuracy as well.

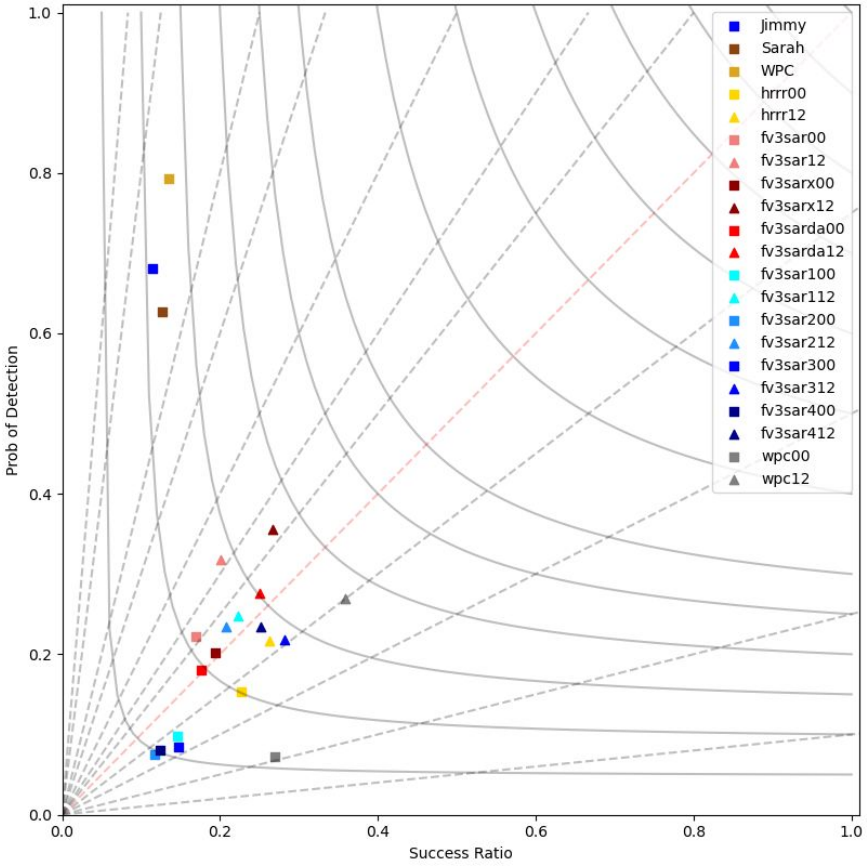

Figure 64: *Similar to Fig. 63, except Aggregate statistics during the MRTP experiment. Three long term forecasters are square, 00z models square and 12z models triangles.*

The survey data collected helped assess how well forecasters did when forecasting the location and amount of maximum rainfall as well as the maximum duration within the 1 inch precipitation contour. Histograms of the duration difference, maximum rainfall difference, and distance from forecast to observed rainfall maximum location can be seen in Fig. 65. Forecasters distance errors, consistent with their higher PODs, were relatively small compared to model guidance (here split up by cycle). Forecasters were within 200 km more than 50% of the time,

while the models' histograms were relatively flat out to distances less than 500 km and 50% of the time models were within 250 km and 300 km for the 12z and 00z cycles respectively.

For precipitation duration, forecasters were biased slightly low, between 1-2 hours, with a mean difference around -1 hour. Meanwhile models were generally the polar opposite, forecasting longer durations. Maximum rainfall differences for forecasters were generally under forecast 90% of the time, with a mean difference near -1 inch. Models tended toward over forecasting, with a mean difference of +1 inch; the 12z cycle being better dispersed around 0 than the 00z cycle. Examining two-dimensional histograms of duration difference versus maximum rainfall differences, when max rainfall is over forecast, forecasters under forecast duration and vice versa.. The relationship is much weaker or absent in models, and therefore it seemingly appears duration errors have little relationship to the max rainfall errors.

In general, these results point to the fact that during a variety of cases, CAMs provide decent guidance on duration, location, and magnitude of the maximum rainfall within these domains. Likewise, forecasters were competitive with the model guidance that was available to them. Had this guidance been explicitly quantitative it is plausible they could have used this information more directly in the decision making process. In the future, we hope to document how such guidance is used to see if it improves forecaster decision making.

While the results shown here are promising, we would caution that on many days, the location and amount of rain can be highly suspect from particular models. However, it is encouraging that the ensemble of models evaluated in FFaIR have error distributions small enough to compare favorably with forecasters' error distributions. A more comprehensive and rigorous analysis of the models' data is planned to see how these error distributions compare for the whole CONUS domain over a 35 day period.

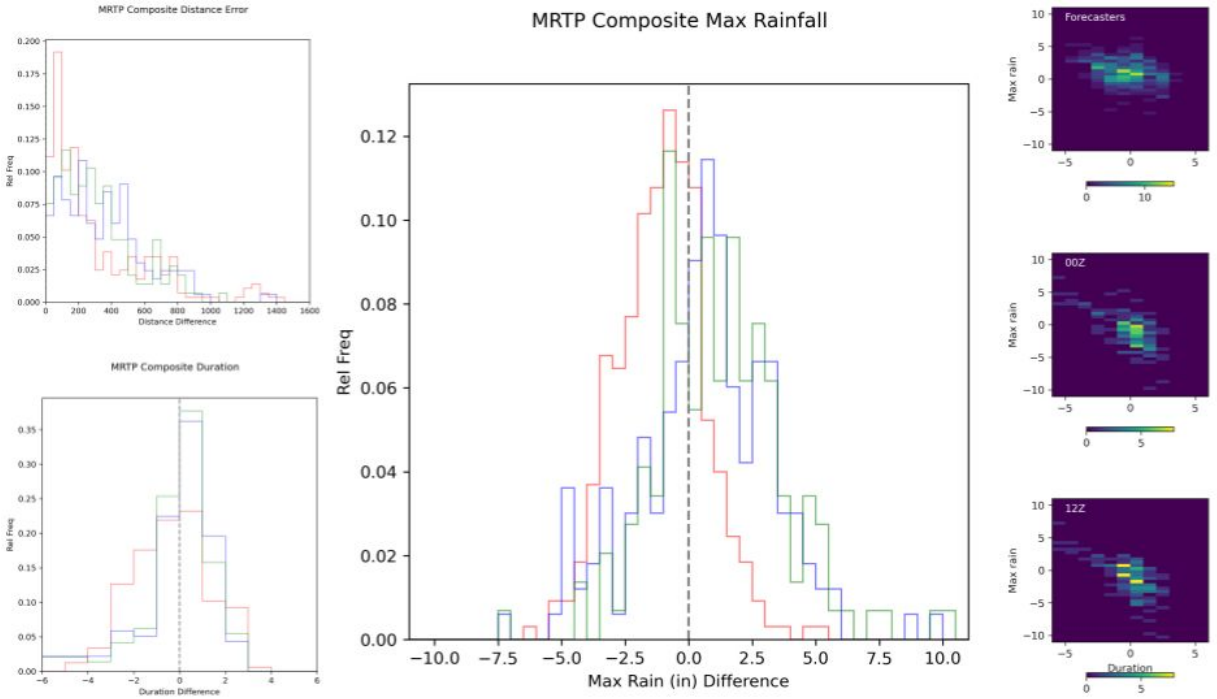

Figure 65: *Various errors gathered from forecasters (red) surveys compared to model information at 00z (blue) and 12z (green) for Distance error (upper left), duration error (lower left), Maximum rainfall (middle), and 2d histograms of max rain vs duration (right) for forecasters (upper right), 00z models (middle right), and 12z models (lower right).*

5. Summary and Conclusions

The 2020 Flash Flood and Intense Rainfall Experiment was heavily centered around the evaluation of the various FV3-SAR configurations, with eight different configurations evaluated as well as an FV3-SAR based ensemble. Understanding how the performance of the various configurations differ among one another will help developers converge on the optimal dynamics and physics suites for the FV3-SAR prior to its planned implementation in 2023. Additionally it will aid in determining any systematic deficiencies in the FV3 core that were not present in other model cores that might need to be addressed.

This year, FFaIR evaluated spatially aligned means, SAM and SAM-LPM, developed by the CAPS-OU. These two means were only evaluated using the SSEF and had mixed results on their performance when compared to the general mean and the LPM mean. However, this might have been in part due to the bug found in one of the member's code. FFaIR also evaluated a product that helps identify the heavy rainfall risk referred to as HPOT. It tracks heavy rainfall objects forecasted by each member of an ensemble and then provides the probability of being in a heavy rainfall object. This was not statistically evaluated but rather feedback from the participants was collected and it was extremely well liked by the majority of the participants.

Aside from evaluating experimental models and ensembles, new/experimental forecasting tools were also assessed. This included an updated version of the CSU First Guess Day 1 ERO trained on the GEFS as well as a new version of the NSSL ERO First Guess Field to compare to the version from last year. The updated GEFS ERO tool was praised by the participants and performed well compared to both the operational and FFaIR Day 1 ERO. Such results, along with the CSU Day 2 and Day 3 GEFS First Guess EROs having already been successfully transitioned into operations, helps to show the utility of machine learning tools in the forecasting process.

The main findings and recommendations from the 2020 FFaIR Experiment can be found in the following bullets points. Table 5 indicates which models/ensembles/products are recommended to be transitioned to operations and which are recommended for further development.

- EMC provided three different configurations of the FV3-SAR, referred to as the **SAR, SARX,** and **SARDA.** Of these three configurations the **SARX** performed the best both subjectively and objectively compared to the other two configurations. The SAR had the highest wet bias of not only these three configurations, but of all the deterministic models evaluated, for both the 24 h and 6 h QPF. Subjectively, the 24 h QPF forecasts from the SAR and SARDA were comparable to one another but for the 6 h QPF forecast, the SARDA was preferred over the SAR. **Therefore, it is recommended that EMC**

continues with FV3-SAR development using the SARX and SARDA configurations while stopping development on the SAR configuration.

- GSL provided four different configurations of the FV3-SAR, referred to as the **SAR1, SAR2, SAR3,** and **SAR4.** Subjectively, the 6 h QPF from the **SAR3** and **SAR4** were preferred over the forecasts from the SAR1 and SAR2. Objective verification was less clear cut. The grid-based 24 h QPF performance diagrams showed that the SAR3(SAR1) and SAR4(SAR3) had comparable CSI to one another with SAR3 and SAR4 always having greater CSI than SAR1 and SAR2. Meanwhile, SAR3(SAR1) nearly always had a bias closer to 0 than SAR4(SAR2). Both the SAR2 and SAR4 use the Hord 6 dynamics suite but differ in initial and lateral boundary conditions, GFS vs HRRRv4/RAPv5 respectively, while SAR1 and SAR3 use Hord 5 and use the GFS vs HRRRv4/RAPv5 respectively, as their initial and lateral boundary conditions. This suggests that the dynamics suite used by SAR1 and SAR3 verified slightly better than the one used by SAR2 and SAR4. **Therefore, it is recommended that EMC continues with FV3-SAR development using the Hord 5 dynamics suite.**

- All FV3-SAR configurations had a notable wet bias that must be addressed. The most troubling part of this wet bias is in the QPF associated with single cell or popcorn convection. Each configuration appears to be creating gridscale convection that "rains out" all the moisture available in the grid cell column. The FV3-SARs are forecasting hourly rates exceeding 3 inches for nearly every single cell storm in the model. **It is the FFaIR's team request that this issue becomes a top priority as the FV3-SAR development continues.**

- The HREFv3, even with the substitution of the operational HRRR for the HRRRv4, outperformed the HRRRE and SSEF. The HRRRE performed better than the SSEF but the control member of the SSEF had a bug in the code that was discovered after the experiment so it is difficult to quantify the true performance of the SSEF. Although the HRRRE was considered to be less useful than the HREFv3, they still liked to look at the ensemble and were happy to have additional CAM ensemble information aside from the HREFv3. Furthermore, discussion among the participants suggest that additional ways to convey ensemble output should be explored. **The FFaIR team supports the implementation of the HREFv3. Additionally, development work on the HRRRE and SSEF should continue to help increase same core ensemble performance.**

- The SSEF Spatially Aligned Means (SAMs) were met with mixed results. The participants noted that it was often difficult to see much of a difference between the SAM/SAM-LPM and the general/LPM mean. Statistically, the general mean performed better than then SAM while the SAM-LPM performed better than the LPM mean.

- The NBM PM mean products were found to have little utility in the day one time frame. This is likely due to the products including all 170+ NBM members available, rather than just high resolution models. **It is recommended that MDL looks into using an ensemble suite similar to the one used for the NBM day 1 QPF as the base for the PM mean products rather than including all the members available.**

- Objectively and subjectively the CSU ML First Guess Day 1 GEFS ERO performed well and was often comparable to both the operational and FFaIR ERO. The two NSSL Day 1 EROs were less favorable, though the changes made to the training of the model (NSSL1 vs NSSL2) did appear to enhance the forecast. **The GEFS Day 1 ERO should be transitioned to operations, however with the upgrade to the GEFS system it must be demonstrated that the retrained ML product on the new GEFS climatology is comparable to the one evaluated in FFaIR.**

- The Heavy Rainfall Object Tracker (HPOT) was well received by participants. They felt it helped convey the heavy rainfall threat in a concise way. It was often used by the participants when working on their MRTP forecasts. They did note some alterations to the images and website that need to be addressed but overall the results for the product were positive.

- The MRTP exercise was a tremendous success. The participants enjoyed the activity and it helped create a more engaging experiment. The survey responses and the products themselves have helped the FFaIR team dive deeper into model characteristics and hopefully the finding can help enhance model development in the future. The FFaIR team plans to continue using the product as a catalyst for participant engagement and model/ensemble evaluation in future experiments. In fact, the product has been adapted for use in the WPC-HMT Winter Weather Experiment.

- There were many pros and cons to having to hold the FFaIR Experiment virtually. A major advantage of having a virtual experiment was the ability to have more people participate in the experiment and allowed for a larger diversity of participants. It should be noted however, that day-long virtual sessions are much more exhausting than in-house sessions. Since it is likely that the FFaIR Experiment will still hold some aspects of the virtual experiment in the future, it is important that the team learn more about best practices for virtual meetings so burn out does not occur.

Table 5: *Research to Operations Transition Metrics for the 2020 FFaIR Experiment. * indicates conditional recommendations which were discussed in the above bullet points.*

Models, Ensembles and Products Evaluated	Recommended for transition to operations	Recommended for further development and testing	Rejected for further testing	Provider/Funding Source
EMC FV3-SARs		X		EMC
GSL FV3-SARs		X		GSL
HRRRv4	X			EMC/GSL
GFSv16		X		EMC
HREFv3	X*			EMC
HRRRE		X		ESRL/GSL
SSEF		X		OU/CAPS
NBM PM means		X		MDL
SSEF Spatially Aligned Means		X		OU-CAPS
CSU-ML Day 1 ERO GEFS	X*			CSU/JTTI
CSU-ML Day 1 ERO NSSL1/NSSL2		X		CSU/JTTI
HPOT		X		WPC/CIRES

Acknowledgments

The FFaIR team would like to extend a tremendous thank you to the HMT and WPC staff that helped us prepare for the experiment and all the support they provided during the experiment. With a special thanks to **Ben Albright** and **Mike Erickson** for helping with the verification of the guidance evaluated. We can not thank our WPC Forecasters (**Rich Otto, Bryan Jackson, Josh Weiss,** and **Andrew Orrison**) enough for all their hard work and help during the forecasting portions of the experiment. Your insight and guidance were astronomical and helped enhance the participants' experience. And thank you to all the forecasters at WPC who covered their shifts while they helped us with the experiment. We would like to recognize and thank all the developers who provided us the data to evaluate during the experiment, especially those for GSL, EMC, CAPS, and CSU. We enjoy working with you and look forward to our continued partnership. **Sara Sienkiewwicz** thank you for sharing your web code, without it the FFaIR website likely would have not happened. Lastly, the team would also like to acknowledge and thank our intern **Katie Bachli** for all her hard work leading up to and during the experiment. We are very proud of you, your hard work, and the research you helped with.

References

Bullock, R. G., Brown, B. G., & Fowler, T. L. (2016). Method for Object-Based Diagnostic Evaluation (No. NCAR/TN-532+STR). doi:10.5065/D61V5CBS.

CBS Local News Philadelphia, 2020: Accessed 7 July 2020, https://philadelphia.cbslocal.com/wp-content/uploads/sites/15116066/2020/07/cheltenham-creek-flooding.jpg.

Doswell, C. A., H. E. Brooks, and R. A. Maddox, 1996: Flash Flood Forecasting: An Ingredients-Based Methodology. Wea. Forecasting, 11, 560–581, https://doi.org/10.1175/1520-0434(1996)011<0560:FFFAIB>2.0.CO;2.

KHOU 11 News, 2020: Several people stranded in stalled cars due to flash flooding near Katy. Accessed 25 June 2020, https://www.khou.com/article/news/local/several-people-stranded-in-cars-due-to-flash-flooding-near-katy/285-e2160aad-32c0-4b35-b329-8b608c11eebe.

Guerrero, G., 2020: Flooding hits several cities across Central Illinois. NBC 25 News, Accessed 16 July 2020, https://week.com/2020/07/15/flooding-hits-several-cities-across-central-illinois/.

Hou, D., M. Charles, Y. Lou, Z. Toth, Y. Zuh, R. Krzusztofowicz, Y. Lin, P. Xie, D.-J. Seo, and M. P. B Cui, 2014: Climatology-Calibrated Precipitation Analysis at Fine Scales: Statistical Adjustment of Stage IV toward CPC Gauge-Based Analysis. *J. Hydrometeor.*, **15** (6), 2542–2557, https://doi.org/10.1175/JHM-D-11-0140.1.

MET 2018: Grid-Stat Tool. *MET Tutorial Presentations: Winter 2018*, Boulder, CO, http://www.dtcenter.org/sites/default/files/community-code/met/docs/presentations/met-tutorial-20180131/12_Grid_Stat_Jan18.pdf.

National Climatic Data Center, 2020: Record Event Report (RER) Peoria, IL. Accessed September 2020, https://w2.weather.gov/climate/index.php?wfo=ilx.

NHC, 2020: Accessed July 2020, https://www.nhc.noaa.gov/.

OWP NWM, 2020: Handbook: NWC Visualization Services Version 1.4. https://docs.google.com/document/d/1RbPTtV4VbIBFcG0WId1J9VZwZFFIK8lp-mfGoyHzJvc/edit.

Prince George's County Fire/EMS Department Twitter, 2020: Accessed 7 July 2020, https://twitter.com/PGFDNews/status/1280341872783888384.

PSL, 2014: Daily Mean Composites. Accessed August 2020, https://psl.noaa.gov/data/composites/day/.

Swathwood, S., 2020: Flash floods strand many Peoria drivers, prompting rescues. NBC 25 News, Accessed 16 July 2020, https://week.com/2020/07/15/flash-floods-strand-many-peoria-drivers-prompting-rescues

Standard Color Curve Summary 2019: Proposed National Weather Service Standard Color Curves. NWS Color Curve Working Group, Accessed Spring 2020, https://docs.google.com/presentation/d/1UEGGXWYa7c7awcOFdczRAWBByq3Oi5LL2wYWBfimBlg/edit#slide=id.p24.

Taima, K, 2020: Flooded Interstate 41 in Fond du Lac reopens after more than 5 hours closed. FDL Reporter, Accessed 11 July 2020, https://www.fdlreporter.com/story/news/2020/06/10/fond-du-lac-interstate-41-closed-between-u-s-151-and-military-road/5338575002/.

Trojniak, S.M., and B. Albright, 2019: 2019 Flash Flood and Intense Rainfall Experiment: Findings and Results. 122 pp, https://www.wpc.ncep.noaa.gov/hmt/Final_Report_2019_FFaIR.pdf.

Trojniak, S.M., J. Correia Jr, B. Albright, 2020: 2020 Flash Flood and Instance Rainfall (FFaIR) Experiment: Program Overview and Operations Plan. 35 pp, https://docs.google.com/document/d/1fUGXVZd6MMMY7Ek-5HLp95fFZo9jcqx3X3SSTd1FMuQ/edit#heading=h.gohd31isqrpk.

WFO LOT, 2020: June 26 & 27, 2020: Severe Storms Bring Numerous Areas of Wind Damage & Flash Flooding. Accessed September 2020. https://www.weather.gov/lot/26june2020.

WMUR9 Article, 2020: Grafton County flooding forces hospital to cancel surgeries.Written by Amy Coveno, Accessed 14 July 2020, https://www.wmur.com/article/line-of-storms-featuring-heavy-downpours-trigger-flash-flood-warnings-in-new-hampshire/33308677#.

WPC, 2020: https://www.wpc.ncep.noaa.gov/index.php#page=ovw.

WPC Met Watch Archives: Mesoscale Precipitation Discussions, https://www.wpc.ncep.noaa.gov/metwatch/metwatch_mpd.php.

UCAR Image Archive, 2020: https://www2.mmm.ucar.edu/imagearchive/.

Appendix A

A.1 Guidance and Products Evaluated

Table A.1 *The deterministic and ensemble model guidance and products that will be evaluated in the 2020 FFaIR experiment.*

Provider	Model	Resolution	Forecast Hours	Notes
OWP	***National Water Model (NWM) version 2.0***	*250 m 1 km*	*Analysis and Assimilation: 0h Short Range: Forecast 18 h Mid-Range: 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.*
OWP	***NWM Visualization Services version 1.2***			*The NWM visualization services leverages GIS technology and is used to display output from the NWM. The NWM visualization service is currently in demonstration mode and is only accessible to a limited number of forecasters and researchers throughout the NWS.*
ESRL/GSL	***HRRRv4* GSL version not parallel because it was halted***	*3 km*	*Hourly forecasts. Forecast length: 00, 06, 12, and 18 UTC runs are 48 h. All other run times are 18 h.*	*High resolution, hourly updated, convection allowing nest of the Rapid Refresh (RAP) model.*

EMC	***HREFv3* altered membership after HRRRv4 halt***	*~3 km*	*48 h forecast run daily at 00 and 12 UTC.*	*Consists of 10 members, each member provides a real-time and time-lagged run.*
GFDL/EMC	***GFSv16***	*13 km*	*3-hourly output up to 384 hours.*	*Finite-volume cubed-sphere dynamical core. Increased vertical resolution to 64 layers.*
ESRL/GSL	*HRRR Ensemble (**HRRRE**)*	*3 km*	*Hourly out to 48 h at 00 and 12 UTC. Hourly out to 24hr at 06 and 18 UTC.*	*Now covers the entire CONUS. Initialized from the first 9 members of the HRRRDA analysis.*
GFDL/EMC	***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. Has the same physics suite as the GFSv15.*

GFDL/EMC	***EMC-FV3-SARX***	*~3 km*	*Hourly out to 60 h initiated once daily at 00 UTC.*	*Experimental Stand-alone Regional. Configuration uses the Thompson microphysics and MYNN planetary boundary layer schemes.*
GFDL/EMC	***EMC-FV3-SARDA***	*~3 km*	*Hourly out to 60 h initiated once daily at 00 UTC.*	*Experimental Stand-alone Regional with hourly data assimilation, same physics suite as FV3-SARX.*
ESRL/GSL	*Multiple (4) versions of the* ***GSL-FV3-SAR***	*~3 km*	*Hourly out to 36h, initiated once daily at 00 UTC.*	*GSI will be providing various versions of their FV3-SAR, changing the IC/LBC and dynamics suites.* ***4 configurations evaluated from GSL:*** ***FV3-SAR1*** ***FV3-SAR2*** ***FV3-SAR3*** ***FV3-SAR4***
OU/CAPS	***SSEF***	*3 km*	*60 h forecasts, though some of the members 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 GEFS perturbation (6), or the GFS (1).*

MDL	***NBMv4***	*2.5 km*	*Hourly out 36 h 3-hrly out 66 h 12-hrly out 264 h*	*A calibrated blend of forecast guidance, with an "expert weight" given to each system.*
WPC/CIRES	***Heavy Rainfall Object Tracker (HPOT) Product***		*Varies based on model*	***Product available for:*** ***Operational HRRR*** ***HRRRv4*** ***HREF*** ***HRRRE***
CSU	***Machine Learning Day 1 ERO First Guess Product***		*36 h forecast*	***Product available for:*** ***GEFS*** ***NSSSL (two versions of product)***
WPC/HMT	***Color Blind Friendly WPC QPF colorbar***			*Combination of the WPC QPF intervals and the NWS suggested QPF curved for colorblindness*

Appendix B

B.1 Participant and Presenter Information

Table B.1: *List of the participants for each week of the 2020 FFaIR Experiment. Note the experiment did not run during the week of the Fourth of July.*

Week	WPC Forecaster	WPC/RFC/Other	Research/Academia/ Student	EMC	GSL	MDL
Week 1	Rich Otto	Glen Merrill - WFO SLC Caleb Steele - WFO VEF Ian Lee - WFO DTX Andrew Mangham - NCRFC HAS Sarah Jamison - WFO CLE Robert Munroe - WFO GSP Kyle Pallozzi - WFO LWX Josh Whisnant - WFO EKA Amanda Schroeder - WGRFC Jeremy Buckles - WFO MRX Kelly Mahoney - PSL Keith White - WFO EWX	Erik Nielsen - CSU/Texas A & M Aaron Hill - CSU	Shannon Shields Geoff Manikin	David Dowell	Steve Levine
Week 2	Bryan Jackson	Robin Fox -WFO OTX Paul Iñiguez - WFO PSR Kate Abshire - NWC Nick Carletta - WFO FGF Margaret Curtis - WFO GYX Morgan Simms - WFO MHX Ross Giarratana - WFO RLX Mike Johnson - WFO MEG Jessica Smith - WFO MLB Sid King - WFO FFC	Allie Mazurek - CSU Student	Matt Pyle Cory Martin	Terra Ladwig	David Rudack
Week 3	Josh Weiss	Brad Carlberg - WFO PUB Treste Huse - WFO BOU Barrett Smith - WFO RAH John Cristantello - WFO OKX Nick Greenawalt - WFO CLE Jordan Dale - OAR/WPO Patrick Blood - WFO HFO Allan Diegan - WFO Melissa Beat - WFO AMA Tim Brice - WFO EPZ Linda Gilbert - WFO MQT Tyler Castillo - WFO CRP Eric Marello - WFO FWD Kent Knopfmeier - NSSL/WoFS	Russ Schumacher - CSU Ian Russell - PSU Student	Matt Morris Chris Macintosh	Eric James	Eric Engle
Week 4	Andrew Orrison	Paul Fajman - WFO OAX Timothy Lynch - WFO FGF Mike Montefusco - WFO AKQ Charles Ross - WFO CTP Jay Engle - WFO OKX Adam Clark- NSSL/HWT HelgeTuschy - DWD Michelle Amin - WFO HUN Patricia Sanchez - WFO FWD Aviva Braun - WFO CYS	Keith Brewster - OU/CAPS Jacob Escobedo - CSU Student	Ben Blake	Jeff Duda	Mark Antolik

Table B.2: *List of the presenters and the title of their presentations during the 2020 FFaIR Experiment. Two presentations were given on Tuesdays and one on Thursdays. The presentations can be found in this shared folder.*

Week	Presenter	Title of Presentation
Week 1	Aaron Hill - CSU	"First Guess" Excessive Rainfall Outlooks from Machine Learning Models
	Isidora Jankov - GSL	Stochastic Approaches in High Resolution Rapid Refresh Ensemble (HRRRE)
	Kelly Mahoney - PSL	Disentangling Uncertainties in Streamflow Predictions: 2018 Ellicott City MD Flash Flood Case Study
Week 2	Kate Abshire - NWC	National Water Model Visualization Services: Background and Overview
	Matthew Pyle - EMC	High-Resolution Ensemble Forecast (HREF) Upgrades and Plans
	John Forsythe - CIRA/CSU	Towards the Next Generation of Blended Total Precipitable Water (TPW) for Operations
Week 3	Mike Erickson - WPC/CIRES	Development and Usage of the Heavy Rainfall Precipitation Object Tracker (HPOT)
	Eric Engle - MDL	NMB v4.0: Gamma Quantile Mapping and PMM Generation
	Russ Schumacher - CSU	From random forests to flood forecasts: Evaluation and implementation of the CSU-Machine Learning Probabilities at WPC
Week 4	Ben Blake - EMC/IMSG	FV3 Stand Alone Regional (SAR): Configurations at EMC
	Keith Brewster - OU/CAPS	SAR-FV3 Storm-Scale Ensemble Forecasts and Ensemble Consensus Products for the 2019-20 HMT FFaIR Experiments
	JJ Gourley - NSSL	Use of MRMS and FLASH products in operations and future directions

Appendix C

C.1 Subjective Verification Locations and Noteable Events

Table C.1: *List of the locations over which the 24 h QPF subjective scores were focused on during the 2020 FFaIR Experiment. The valid date of verification is from 12 UTC the date prior to 12 UTC of the day listed. The Notable Events column refers to interesting events that happened throughout each week.*

Week of Evaluation	Valid Date of Verification	Region Evaluated for 24 h QPF Subj. Analysis	Notable Events
Week 1	June 12	East Coast	Slow moving closed low present most of the week over Mid-Atlantic resulted in widespread flash flooding reports from northwest of Raleigh to southeastern WV. June 18-19 heavy rainfall event south of Lincoln NE with some areas of 7+ inches in 24 h.
	June 13	N/A	
	June 16	VA to SC	
	June 17	VA to SC	
	June 18	Northern Plains	
Week 2	June 19	TX to MN	Nearly continuous threat for heavy rainfall and flooding along the western Gulf Coast throughout the week leading flooding in the greater Houston Metro. By the end of the week a surface low developed over NE and brough heavy rainfall from the Midwest to the Great Lakes.
	June 20	Southern Plains	
	June 23	OK/TX/LA region	
	June 24	Eastern TX to GA	
	June 25	Gulf Coast to Carolinas	
Week 3	June 26	Central to Northern Plains	All MRTPs but one were focused somewhere between the Dakotas to MI. Flash Flood Emergency issued for the Philadelphia region July 6, with flooding across the DC Metro into southern NJ. TS Fay made landfall in NJ in evening of July 10. Widespread flooding in northeastern NJ.
	June 27	Great Lakes	
	July 7	Mid-Atlantic	
	July 8	Northern Plains	
	July 9	MN into WI	
Week 4	July 10	Midwest to Great Lakes	Scattered heavy rainfall across New England in the beginning of the week, leading to an ER flooding on July 14. July 15 saw widespread flooding across Central IL. This led to a new record 24 h total rainfall in Peoria, IL and parts of Illinois State University's campus flooding.
	July 11	New England	
	July14	Minnesota	
	July 15	Central Plains	
	July 16	Peoria IL event	
Additional Verification Sessions	July 17	WV to NY	
	July 18	Dakotas to WI	

C.2 MRTP Information

Table C.2: *List of the valid times for the MRTPs and their region of interest during the 2020 FFaIR Experiment.*

Week	MRTP Valid Time and Date (2020)	MRTP Forecast Region
Week 1	21 UTC June 15 - 00 UTC June 16	North Carolina and southern Virginia
	21 UTC June 16 - 00 UTC June 17	North Carolina and southern Virginia
	00 UTC - 06 UTC June 18	Dakotas to Minnesota
	21 UTC June 18 - 03 UTC June 19	Eastern Central Plains to the Upper Mississippi Valley
	21 UTC June 19 - 03 UTC June 20	Texas, Oklahoma Border
Week 2	03 UTC - 09 UTC June 23	Mid-Mississippi Valley
	06 UTC - 12 UTC June 24	Southeastern Texas to Louisiana
	06 UTC - 12 UTC June 25	Southeastern Texas to Louisiana
	21 UTC June 25 - 03 UTC June 26	Northern Plains
	00 UTC - 06 UTC June 27	Mid-Mississippi Valley through the Great Lakes Region
Week 3	03 UTC - 09 UTC July 07	South Dakota into Northern Mississippi Valley
	03 UTC - 09 UTC July 08	North Dakota
	00 UTC - 06 UTC July 09	Minnesota/Wisconsin
	21 UTC July 09 - 03 UTC July 10	Western Great Lakes
	21 UTC July 12 - 03 UTC July 11	Philadelphia/NYC Corridor
Week 4	00 UTC - 06 UTC July 14	Minnesota
	06 UTC - 12 UTC July 15	Central Plains to Mid-Mississippi Valley
	00 UTC - 06 UTC July 16	Mid-Mississippi Valley to Central Great Lakes
	00 UTC - 06 UTC July 17	Southwestern Missouri/Northwest Arkansas
	21 UTC July 17 - 03 UTC July 18	North Dakota/Minnesota
NowCast	21 UTC July 15 - 00 UTC July 16	Illinois

Appendix D

D.1 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

D.2 NSSL1 and NSSL2 Configuration Differences

The main difference between the NSSL1 and NSSL2 is how many predictors are at each grid point and how the predictors are evaluated at a point. The predictors for NSSL1 are constructed in a forecast-point relative sense; i.e. there is an 11x11 grid-point box that surrounds the forecast point at a particular time. The inputs for the training of the model are grabbed from each grid point. Opposing this, the NSSL2 takes each of those inputs (for a particular variable like CAPE) and just spatially averages the predictors. This results in the NSSL1 having ~15,000 predictors and NSSL2 having 142 when making a forecast at one point.

Another difference is in how the NSSL1 and NSSL2 are regionally trained. The definition of excessive rainfall differs for each region since they are essentially “optimized” to get the best forecast; the GEFS-based ML models are optimized regionally too. The details of the regional training can be seen in Table D.1 and the location of the regions in Fig. D.1.The acronyms used in the table defined as:

- FC1 - 1-yr ARI exceedance from Climatology-Calibrated Precipitation Analysis (CCPA; Hou 2014)[25] dataset and Flash Flood Reports (FFRs)
- FC2 - 2-yr ARI exceedance from CCPA dataset and FFRs
- FCS1 - 1-yr ARI from CCPA, Stage-IV, and FFR
- FCS2 - 2-yr ARI from CCPA, Stage-IV, and FFR

Table D.1: *The observations used for training each region of the NSSL1 and NSSL2 First-guess Machine Learning Day 1 ERO fields.*

Region	NSSL1	NSSL2
Pacific Coast	FC2	FC2
Rockies	FC2	FC2
Southwest	FC2	FC2
Northern Great Plains	FC1	FC1
Southern Great Plains	FCS2	FC1
Midwest	FC1	FC1
Southeast	FCS1	FCS1
Northeast	FCS2	FC2

[25] A rain-gauge calibrated product based on Stage-IV

Figure D.1: *The breakdown of the regions used in the training for the CSU Day 1 ERO First-guess Fields.*