2024 Flash Flood and Intense Rainfall (FFaIR) Final Report

June 10 - August 2, 2024
Weather Prediction Center (WPC)
Hydrometeorology Testbed (HMT)

Sarah Trojniak[1,2] and **James Correia Jr.**[1,2]

[1]CIRES CU Boulder

[2] NOAA/NWS/WPC/HMT

Contents

1 Introduction

The Hydrometeorology Testbed (HMT), which is part of the Weather Prediction Center (WPC) once again hosted the the Flash Flood and Intense Rainfall (FFaIR) Experiment this summer. FFaIR has been run annually since the summer of 2012 and has become on integral part of the research to operations to research (R2O2R) process as it pertains to extreme precipitation and its impacts. This year's experiment focused heavily on the evaluation of the Rapid Refresh Forecast System (RRFS) and it's Quantitative Precipitation Forecasts (QPF). Additionally the experiment evaluated a couple of machine learning (ML) and cluster analysis tools.

This summer, FFaIR ran 5 week-long sessions that were a combination of virtual and hybrid sessions. The two hybrid weeks brought some of the participants for the week to the National Center for Weather and Climate Prediction (NCWCP) in College Park, MD. The dates FFaIR was in session were as follows:

Week 1: June 10 - 14 (virtual)
Week 2: June 24 - 28 (virtual)
Week 3: July 8 - 12 (hybrid)
Week 4: July 22 - 26 (virtual)
Week 5: July 29 - Aug 2 (hybrid)

2 Overview of the Operations and Science

This section briefly highlights the science goals and operations of the 2024 FFaIR Experiment. A full description of FFaIR activities, evaluated guidance, and goals can be found in the 2024 FFaIR Operations Plan (Trojniak and Correia, Jr., 2024).

The overarching science goal in the FFaIR Experiment this year was to evaluate how precipitation forecasts might change when the RRFS suite replaces the current operational suite of models. To do this, a data-denial experiment was created to compare the RRFS ensemble system, referred to as the Rapid Refresh Ensemble

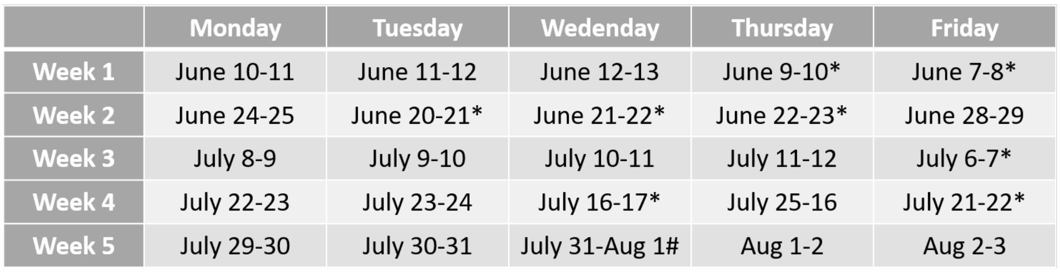
Forecast System (hereafter REFS) and the High Resolution Ensemble Forecast (HREF) and its membership. The data-denial aspect was baked into all of the forecasting activities, with participants being randomly assigned to either the REFS or HREF group.

In addition to the comparison and evaluation of the RRFS and REFS, other experimental datasets included: an ensemble system and a Machine Learning Product (MLP) from the Center for Analysis and Prediction of Storms (CAPS) at the University of Oklahoma (OU), a bias-corrected HREF for Mesoscale Convective Systems (MCSs) ML forecast from Iowa State University (ISU), a National Severe Storms Laboratory Model for Prediction Across Scales (NSSL-MPAS) modeling system, and Cluster Analysis forecasts derived from the REFS. Also incorporated were the exploration of additional risk contours for the Excessive Rainfall Outlook (ERO) and writing of scientific key messages.

The 2024 FFaIR was designed to run in a pseudo-operational setting, forecasting for the Day 1 (12 to 12 UTC) time frame and using real-time data. In the past, if experimental data was missing the experiment would usually still run as normal, just minus the missing data. However, because the experiment this year involved data-denial, if the REFS data was missing the experiment could not run in real-time. When this occurred, the experiment was run in retrospective (retro) mode; performing the same forecast activities but for a past case. When picking a retro case, the team strove to have at least three complete cycles of the REFS suite. However, in some cases the deterministic RRFS (referred to as RRFSp1 in FFaIR) had three cycles available but only 2 cycles of the REFS products. When running in retro, participants were instructed which cycles of model data they could use for each forecasting activity. They were also allowed to use archived radar and satellite data, but only for the current time. This means that if it was currently 15 UTC they could only look at the archived radar and satellite data previous to 15 UTC of the retro day.

Unfortunately, numerous times throughout the experiment the complete REFS suite was not available for use in real-time. Often this was due to one or more members of the ensemble failing. When this occurred, it took two cycles for the ensemble to be available again since it uses time-lagged members. Table 1 shows the breakdown of when the experiment ran in real-time vs retro. In total, 9 of the 25 days had to be run in retro mode, with each week of the experiment having at least one day that included a retroactive case.

Table 1: A list of the dates (12-12 UTC) that were used for the forecasting activities during FFaIR 2024. The dates with an $*$ are dates that were retrospective cases. The # identifies a case in which the experiment did a Day 2 forecast rather than a Day 1 forecast. This is further discussed in Section 2.0.1.



	Monday	Tuesday	Wedenday	Thursday	Friday
Week 1	June 10-11	June 11-12	June 12-13	June 9-10*	June 7-8*
Week 2	June 24-25	June 20-21*	June 21-22*	June 22-23*	June 28-29
Week 3	July 8-9	July 9-10	July 10-11	July 11-12	July 6-7*
Week 4	July 22-23	July 23-24	July 16-17*	July 25-16	July 21-22*
Week 5	July 29-30	July 30-31	July 31-Aug 1#	Aug 1-2	Aug 2-3

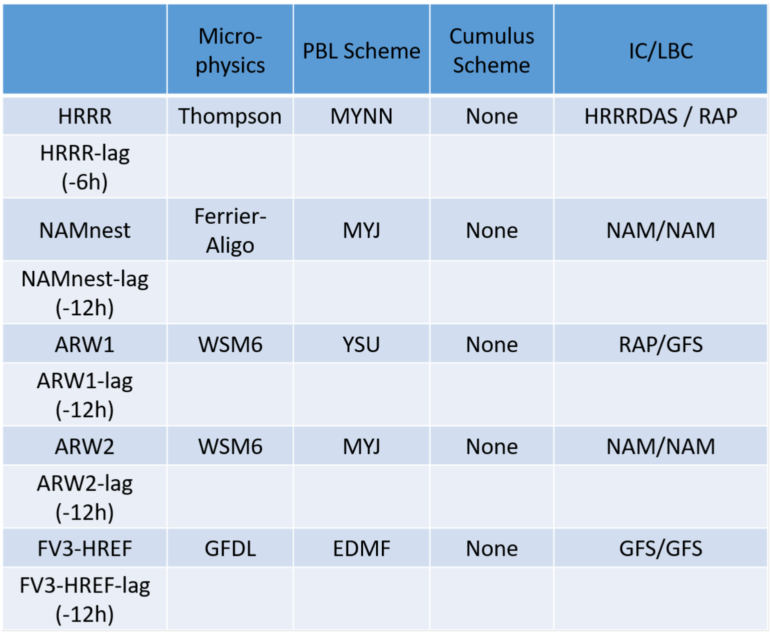
2.0.1 The Unique Circumstance of July 31st - Aug 1st 2024

By the end of the experiment, retrospective cases with an adequate amount of data archived by the team had run out. This resulted in an unique situation in which the experiment created Day 2 forecasts rather than Day 1. As can be seen in Table 1, this was the 24-h period from 12 UTC July 31st to 12 UTC Aug 1st. For the Day 2 forecast in real-time, the participants used model data initialized on July 30th; they could use the 00 UTC and 06 UTC cycles for the morning forecast and the 12 UTC and 18 UTC cycles for the afternoon forecast. Making this day even more distinctive was the fact that the forecast activity itself was *performed* on July 31st, meaning that the start of the Day 2 forecast was actually the current time for them. To help keep the integrity of the Day 2 forecast, the participants were instructed to not look at **ANY** current radar or satellite data during the day, since if it were truly a Day 2 forecast, they would have no knowledge of the ongoing weather on July 31st. If this sounds confusing and challenging, it was! When subjective evaluation was done for the forecasts valid for July 31st - Aug 1st, the REFS data had been back filled by EMC so the REFS could be verified as normal.

2.1 Guidance and Tools Evaluated

For a full description of the guidance evaluated in the 2024 FFaIR Experiment please see Section 4 of the 2024 FFaIR Operations Plan (Trojniak and Correia, Jr., 2024). Tables 2 and 3 show the membership information for the HREF and the REFS. Also evaluated was a convective allowing model (CAM) that uses the Model for Prediction Across Scales (MPAS) provided by NSSL, referred to as the NSSL-MPAS during the experiment. The version provided used Thompson microphysics and was initialized from the HRRR[1].

Table 2: The HREF membership configurations. The HREF is an ensemble of opportunity and consists of 5 convective allowing models and includes time-lagging to create a 10 member ensemble.



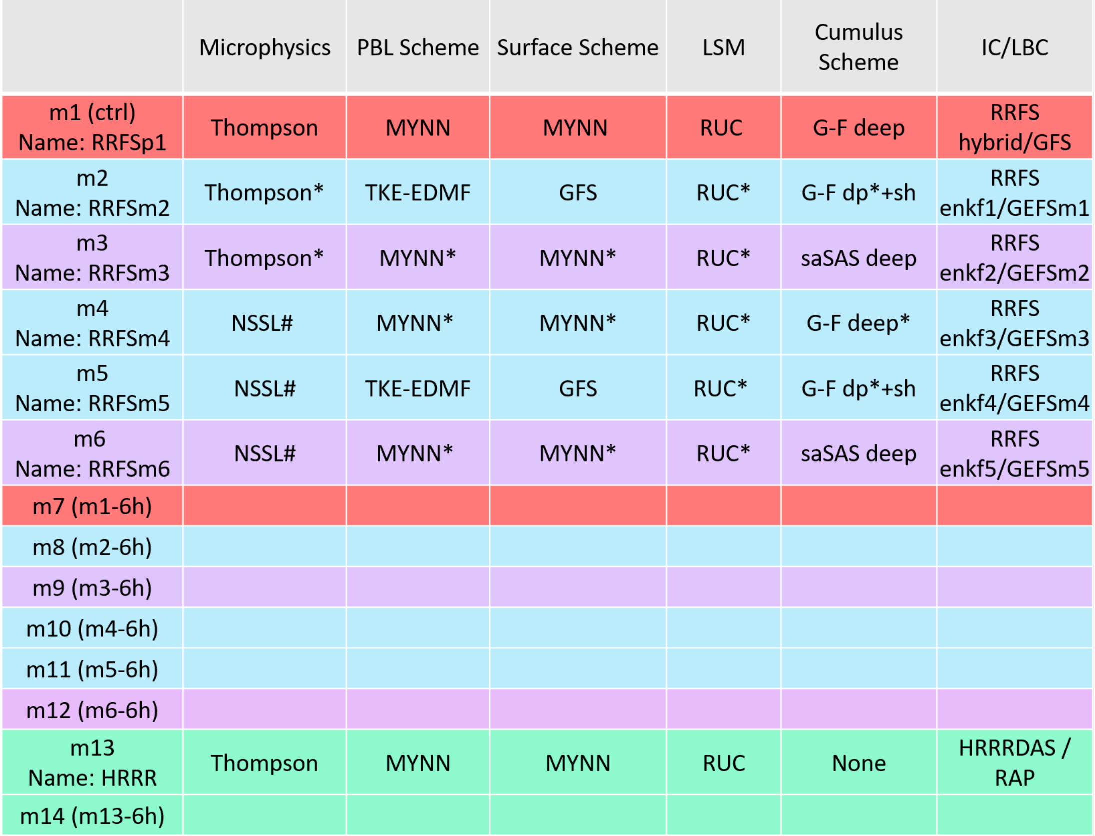
	Micro-physics	PBL Scheme	Cumulus Scheme	IC/LBC
HRRR	Thompson	MYNN	None	HRRRDAS / RAP
HRRR-lag (-6h)				
NAMnest	Ferrier-Aligo	MYJ	None	NAM/NAM
NAMnest-lag (-12h)				
ARW1	WSM6	YSU	None	RAP/GFS
ARW1-lag (-12h)				
ARW2	WSM6	MYJ	None	NAM/NAM
ARW2-lag (-12h)				
FV3-HREF	GFDL	EDMF	None	GFS/GFS
FV3-HREF-lag (-12h)				

For the REFS, aside from traditional ensemble products, 6-h QPF Cluster Analysis products were evaluated. These were supplied by Austin Coleman from CIRES@WPC[2]. The membership for the clustering included time-lagging for 24-h at 6-h intervals, creating a 28 member ensemble (7 real time members for the past 4 cycles). The dominant patterns in the ensemble forecast are derived by calculating the first and second Empirical Orthogonal Functions (EOFs), creating

[1] In the Spring Forecasting Experiment 2024 (Clark et al., 2024), this MPAS version was called the NSSL MPAS HT.

[2] Cooperative Institute for Research in Environmental Sciences (CIRES), CU Boulder

Table 3: The REFS membership configurations. The RRFSp1 (highlighted red) is the control member for the REFS. The HRRR is in green. The members with the G-F deep cumulus scheme are blue and with the saSAS deep cumulus scheme are in purple. Sources of spread in the ensemble are: EnKF ICs, GEFS LBCs, time-lagging, multi-physics, stochastic parameter perturbations(∗), and fixed parameter perturbations (#). The names the members will be referred to as are listed after Name; note what we call as members 2-6 EMC refers to as members 1-5.

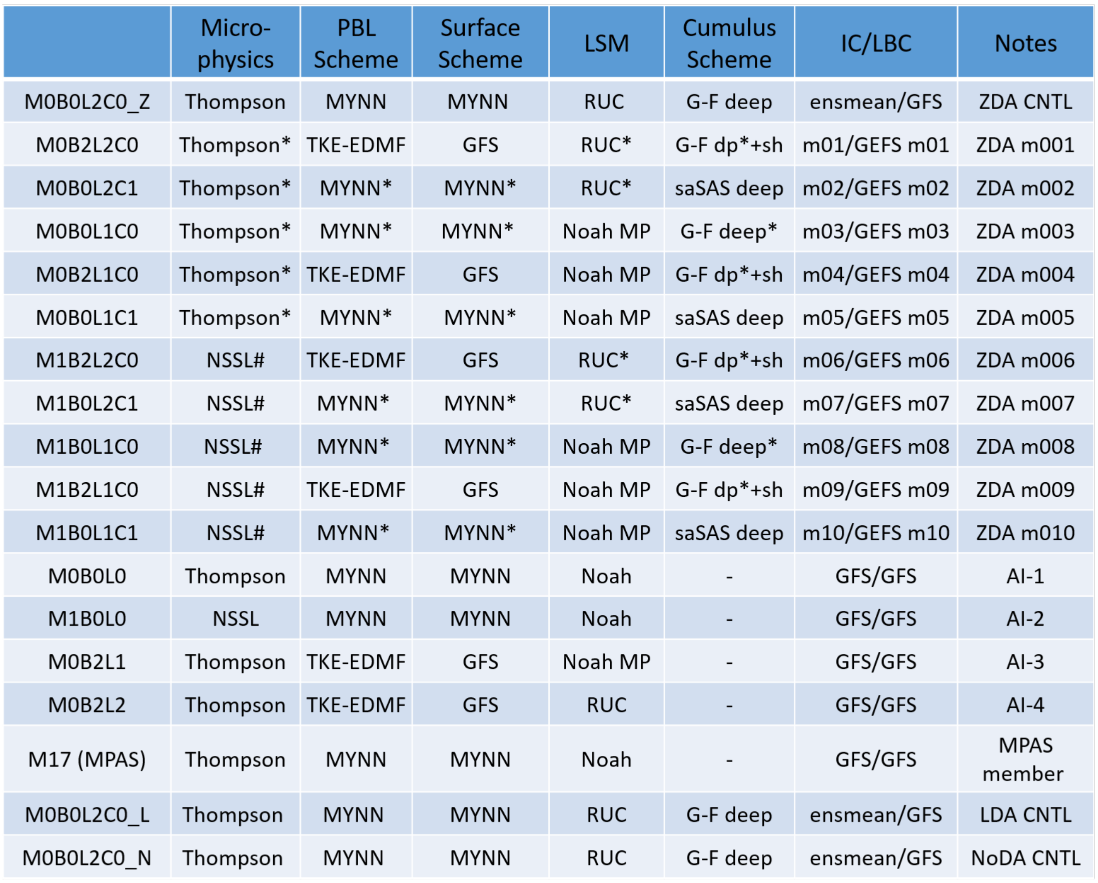


	Microphysics	PBL Scheme	Surface Scheme	LSM	Cumulus Scheme	IC/LBC
m1 (ctrl) Name: RRFSp1	Thompson	MYNN	MYNN	RUC	G-F deep	RRFS hybrid/GFS
m2 Name: RRFSm2	Thompson*	TKE-EDMF	GFS	RUC*	G-F dp*+sh	RRFS enkf1/GEFSm1
m3 Name: RRFSm3	Thompson*	MYNN*	MYNN*	RUC*	saSAS deep	RRFS enkf2/GEFSm2
m4 Name: RRFSm4	NSSL#	MYNN*	MYNN*	RUC*	G-F deep*	RRFS enkf3/GEFSm3
m5 Name: RRFSm5	NSSL#	TKE-EDMF	GFS	RUC*	G-F dp*+sh	RRFS enkf4/GEFSm4
m6 Name: RRFSm6	NSSL#	MYNN*	MYNN*	RUC*	saSAS deep	RRFS enkf5/GEFSm5
m7 (m1-6h)						
m8 (m2-6h)						
m9 (m3-6h)						
m10 (m4-6h)						
m11 (m5-6h)						
m12 (m6-6h)						
m13 Name: HRRR	Thompson	MYNN	MYNN	RUC	None	HRRRDAS / RAP
m14 (m13-6h)						

4 clusters. For each cluster, graphics for the 10th, 25th, 50th, 75th, 90th, and 95th QPF percentiles were created. Along the top of each Cluster graphic are the number of members each cycle has in the cluster group; an example can be found in Section 3. These were available for the afternoon activities only. Although cluster analysis products were not generated for the HREF, a Bias Corrected HREF product was generated from the ISU team. This is a Machine Learning Product (MLP) that identifies the possibility of a Mesoscale Convective System (MCS) from the HREF membership and then attempts to bias-correct location errors. The products provided and evaluated were 24-h QPF mean and probability products.

In addition to the HREF and the REFS, an additional ensemble from CAPS (hereafter CAPSe) was evaluated; Table 4 shows CAPSe membership. A focus of evaluation for this ensemble was the Spatially-Aligned Mean (SAM) QPF and SAM-LPM. The SAM-LPM blends the SAM methodology with that of the local probability match mean (LPM). The CAPS team also provided a MLP for the probability of exceeding 0.5, 1, and 2 inches in 6-h. This uses 4 members from the CAPSe (identified in the Notes section of Table 4) and all members but the FV3-HREF of the HREF for training and forecasts. For further explanation of both SAMs and MLPs, please refer to the 2024 FFaIR Operations Plan (Trojniak and Correia, Jr., 2024).

Table 4: The CAPS ensemble (hereafter CAPSe) membership configurations. In the Notes column it is indicated if the member has DA. If there is an AI-# then it indicates that it is a member used in the CAPS MLP.



	Micro-physics	PBL Scheme	Surface Scheme	LSM	Cumulus Scheme	IC/LBC	Notes
M0B0L2C0_Z	Thompson	MYNN	MYNN	RUC	G-F deep	ensmean/GFS	ZDA CNTL
M0B2L2C0	Thompson*	TKE-EDMF	GFS	RUC*	G-F dp*+sh	m01/GEFS m01	ZDA m001
M0B0L2C1	Thompson*	MYNN*	MYNN*	RUC*	saSAS deep	m02/GEFS m02	ZDA m002
M0B0L1C0	Thompson*	MYNN*	MYNN*	Noah MP	G-F deep*	m03/GEFS m03	ZDA m003
M0B2L1C0	Thompson*	TKE-EDMF	GFS	Noah MP	G-F dp*+sh	m04/GEFS m04	ZDA m004
M0B0L1C1	Thompson*	MYNN*	MYNN*	Noah MP	saSAS deep	m05/GEFS m05	ZDA m005
M1B2L2C0	NSSL#	TKE-EDMF	GFS	RUC*	G-F dp*+sh	m06/GEFS m06	ZDA m006
M1B0L2C1	NSSL#	MYNN*	MYNN*	RUC*	saSAS deep	m07/GEFS m07	ZDA m007
M1B0L1C0	NSSL#	MYNN*	MYNN*	Noah MP	G-F deep*	m08/GEFS m08	ZDA m008
M1B2L1C0	NSSL#	TKE-EDMF	GFS	Noah MP	G-F dp*+sh	m09/GEFS m09	ZDA m009
M1B0L1C1	NSSL#	MYNN*	MYNN*	Noah MP	saSAS deep	m10/GEFS m10	ZDA m010
M0B0L0	Thompson	MYNN	MYNN	Noah	-	GFS/GFS	AI-1
M1B0L0	NSSL	MYNN	MYNN	Noah	-	GFS/GFS	AI-2
M0B2L1	Thompson	TKE-EDMF	GFS	Noah MP	-	GFS/GFS	AI-3
M0B2L2	Thompson	TKE-EDMF	GFS	RUC	-	GFS/GFS	AI-4
M17 (MPAS)	Thompson	MYNN	MYNN	Noah	-	GFS/GFS	MPAS member
M0B0L2C0_L	Thompson	MYNN	MYNN	RUC	G-F deep	ensmean/GFS	LDA CNTL
M0B0L2C0_N	Thompson	MYNN	MYNN	RUC	G-F deep	ensmean/GFS	NoDA CNTL

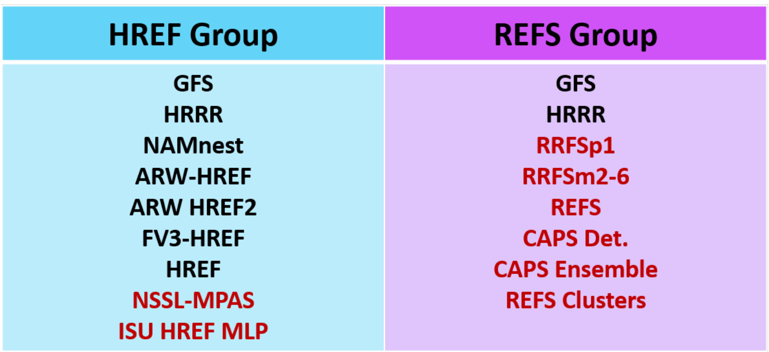
Finally, also evaluated was the ISU Bias Corrected HREF (hereafter ISU MLP or MLP for the HREF). The product attempts to improve the location errors of HREF forecasts of mesoscale convective systems (MCS). It consists of two ML models trained on 24-h QPF from MCSs identified from 2018-2023, along with Storm Prediction Center (SPC) Mesoanalysis data. This is run only for the 12z cycle, so it was not available for the morning forecasting activity. Since it was a 24-h product and the afternoon activity focused on 6-h time windows, aside from days that were run in retro mode, the feedback on the products generated was only from the subjective verification sessions. The ISU team generated domain specific graphics for the 24-h (12-12 UTC) period using the adjusted data of each member plotted with a 5 km grid spacing. The provided graphics were probability matched mean (PMM) and probabilities of exceedance for 1 inch, 2 inches, 3 inches, 5 inches, and 8 inches using a 10 km radius of influence (ROI) and smoother. For a full description of the ISU MLP please refer to 2024 FFaIR Operations Plan (Trojniak and Correia, Jr., 2024).

2.2 Brief Review of Activities

For a full description of the forecasting activities please see Section 3 of the 2024 FFaIR Operations Plan (Trojniak and Correia, Jr., 2024). Like in the past few FFaIR Experiments, the day was broken into three activities: a 20-h Day 1 ERO issuance, subjective verification of the previous day's guidance and tools, and the 6-h Maximum Rainfall and Timing Product (MRTP) exercise. The subjective verification activity will be discussed in Section 3.

As noted previously, this year the experiment was centered around comparison of the HREF and REFS suites and thus the forecasting activities had a data-denial aspect to them. The guidance that each group was able to use for their MRTP and ERO forecasts can be seen in Table 5. Both the HRRR and the GFS were utilized in both groups since the HRRR is a member in both the HREF and REFS and the GFS is a global model. The NSSL-MPAS was included in the HREF group since it has a different model core than the REFS, while the CAPSe was included in the REFS since they have the same model core.

Table 5: The models, ensembles and tools that each Group had access to for the 2024 FFaIR Experiment. The Group names are the names of the operational and experimental ensembles; HREF and REFS respectively. The current operational models/ensembles are in black, while experimental models/ensembles/tools are in red. Note that the HRRR and GFS were used in both Groups.



HREF Group	REFS Group
GFS	GFS
HRRR	HRRR
NAMnest	RRFSp1
ARW-HREF	RRFSm2-6
ARW HREF2	REFS
FV3-HREF	CAPS Det.
HREF	CAPS Ensemble
NSSL-MPAS	REFS Clusters
ISU HREF MLP	

For the ERO forecasting activity, participants were tasked with creating an individual ERO, 1-3 scientific key messages (e.g. short forecast discussions for specific areas and aspects of the ERO), completing a short survey centered around their model data, and creating a collaborative ERO. The ERO mimicked the operational ERO issued by WPC at 16 UTC, which is defined as the risk of exceeding Flash Flood Guidance (FFG) within 25 miles of a point. The WPC ERO currently has 4 risk categories while the FFaIR ERO had 5: Marginal (5%-15%), Slight (15%-25%), Enhanced (25%-40%), Moderate (40%-70%), and High (>70%). The Enhanced Risk was added to encompass the top probabilities included in the operational ERO's definition for a Slight Risk. Also included in the FFaIR ERO was an Intensity contour. This was defined as an area where multiple hours of exceeding the 10-y 1-h ARI would occur within 25 miles of a point. This contour could be drawn independently of the risk contours. An example of the collaborative HREF and REFS EROs along with some individual EROs can be seen in Fig. 1. Results from the testing of the Enhanced and Intensity Contours will not be discussed in the Final Report but will be communicated internally.

The MRTP activity was a 6-h forecast over a domain and time window determined by the facilitators. The facilitators determined the domain and time this year due to the data denial aspect of the experiment. The valid end time of the 6-h period could be anytime between 03 UTC and 12 UTC. For the MRTP,

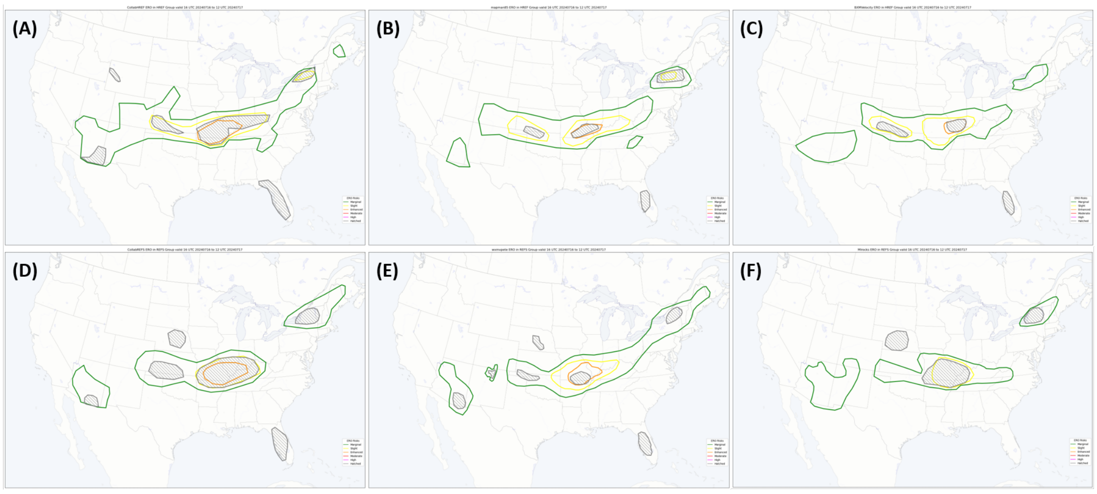
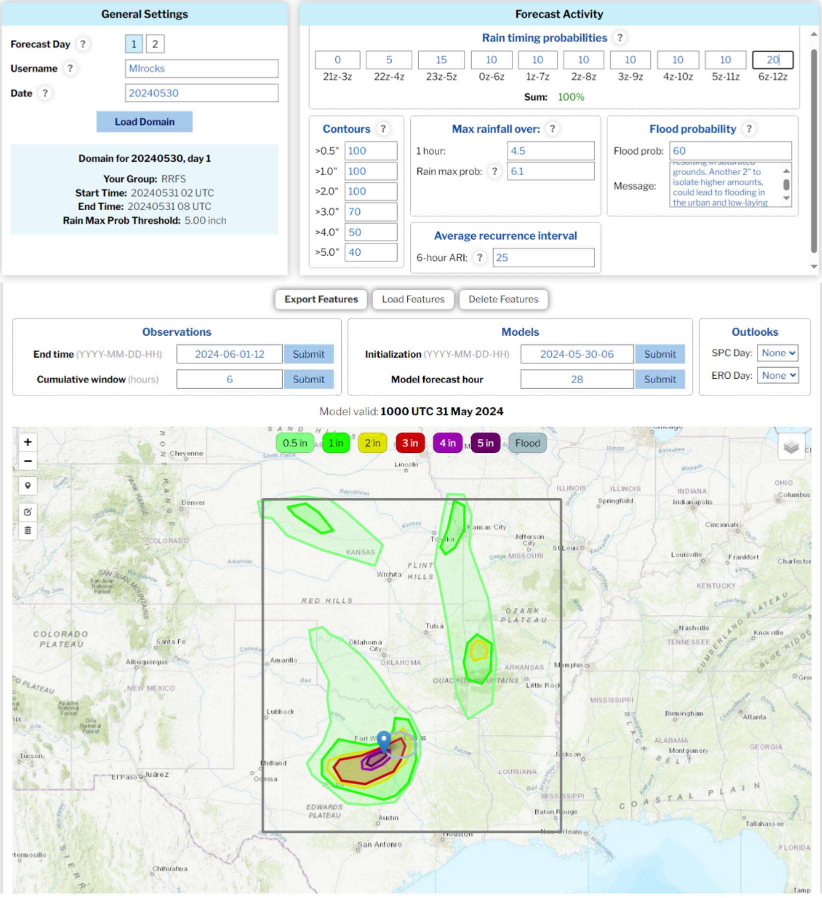


Figure 1: An example of the Day 1 Collaborative ERO for the (A) HREF and (D) REFS Group and individual EROs from members of the (B)-(C) HREF Group and (E)-(F) REFS Group. All valid 16 UTC 16 July to 12 UTC 17 July 2024.

participants were tasked with drawing QPF contours (0.5, 1, 2, 3, 4, and 5 inches) and a flood contour as needed. They also had to provide the location and value of the maximum 6-h rainfall. In addition to this location, they could provide up to two more points where a local maximum might occur. This capability was included in the MRTP this year for cases in which there were multiple areas, separated by a relatively large distance, that could have similar maximum rainfall totals. An example of a drawn MRTP can be seen in bottom part of Fig. 2.

Participants also completed a survey about their model data and thought process, along with a scientific message about the flooding risk. Lastly, they were required to complete the probabilistic component of the MRTP via the prompts along the top of the drawing tool webpage as shown in Fig. 2. This included questions to gather information about the probability that the maximum 6-h rainfall would occur within each of the 10 possible 6-h time windows the MRTP could be valid for (i.e. 21-03 UTC, 22-04 UTC, ... , 05-11 UTC, and 06-12 UTC) and the probability of exceeding a 6-h accumulation threshold decided by the facilitators. They were also asked to assign the probability of exceedance for each of the thresholds that they had the ability to draw contours for. Even if they did not draw a particular threshold, they still could provide a probability greater than zero. In the survey they were asked to provide insight into why they gave a


General Settings

Forecast Day ? 1 2

Username ? MIrocks

Date ? 20240530

Load Domain

Domain for 20240530, day 1

Your Group: RRFS
Start Time: 20240531 02 UTC
End Time: 20240531 08 UTC
Rain Max Prob Threshold: 5.00 inch

Forecast Activity

Rain timing probabilities ?

0	5	15	10	10	10	10	10	10	20
21z-3z	22z-4z	23z-5z	0z-6z	1z-7z	2z-8z	3z-9z	4z-10z	5z-11z	6z-12z

Sum: 100%

Contours ?

>0.5"	100
>1.0"	100
>2.0"	100
>3.0"	70
>4.0"	50
>5.0"	40

Max rainfall over: ?

1 hour: 4.5

Rain max prob: ? 6.1

Flood probability ?

Flood prob: 60

Message: grounds. Another 2" to isolate higher amounts, could lead to flooding in the urban and low-laying

Average recurrence interval

6-hour ARI: ? 25

Export Features | Load Features | Delete Features

Observations

End time (YYYY-MM-DD-HH) 2024-06-01-12 Submit

Cumulative window (hours) 6 Submit

Models

Initialization (YYYY-MM-DD-HH) 2024-05-30-06 Submit

Model forecast hour 28 Submit

Outlooks

SPC Day: None

ERO Day: None

Model valid: 1000 UTC 31 May 2024

Figure 2: An example of a completed MRTP forecast on the MRTP Drawing Website. The domain is highlighted by the gray box. The forecast is valid from 04 UTC to 10 UTC 31 May 2024. This includes the contours for the the 6-h rainfall accumulations: 0.5" (green), 1" (yellow), 2" (red), 3" (dark red), 4" (purple), and 5" (pink) and an area of flood concern (gray). Along the top is the probabilistic information, information about maximum 6-h and hourly accumulation and the Key Message about flooding.

probability greater than zero even though they didn't draw a contour. For instance, they might not have drawn a contour for 5" but put a 5% chance of exceeding 5" and noted in the survey that they didn't draw the contour because the aerial coverage would be too small to contour with our tool.

Lastly, the facilitators determined the MRTP domain and time rather than using the group consensus approach we have used in past FFaIRs. This was done due to the concern that there might be times in which one Model Group's model signaled a heavy rain event but the other group's models did not. Another concern was that even if both groups had a signal in the same region, the timing of the maximum 6-h QPF might differ by several hours. If either of these cases arose, it would be difficult to have to an open discussion to pick the MRTP domain and time without impacting the integrity of the data denial aspect of the experiment. As a result, the facilitators attempted to determine the MRTP domain and time by each picking a domain and time based on the model data from the Model Group they were leading. In instances in which the domain and time could not be agreed upon easily using this method, the facilitators together would look at the all model data to determine the domain/time. When this occurred, the facilitators strove to not favor one Model Group's consensus over the other and the facilitators' MRTP forecasts were not included in the bulk group statistics for that day as to not impact data-denial results.

2.3 Science Questions and Goals

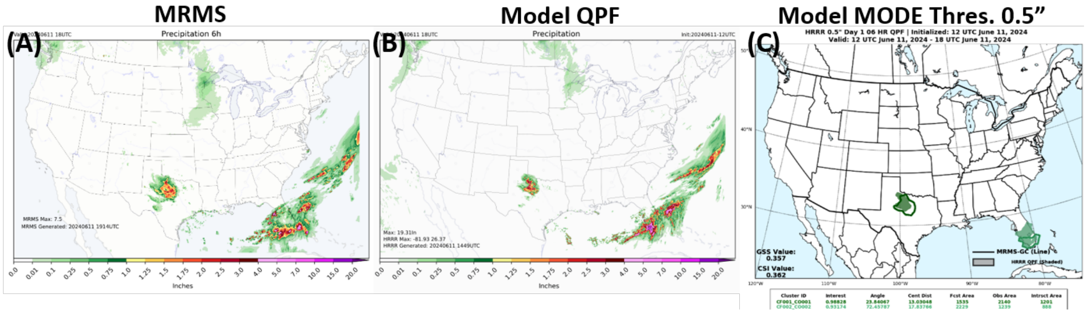
- Evaluate the performance, focusing on the Quantitative Precipitation Forecast (QPF) of the RRFS deterministic (RRFSp1) compared to the HRRR, NAMnest and NSSL-MPAS.
- Although not the main focus, other aspects of the RRFSp1 compared to the aforementioned models will also be noted, such as reflectivity, with performance differences conveyed to developers.
- Evaluate the performance of the REFS through means and probability of exceedances.
- Perform data denial experiments during forecasting activities to simulate what it will be like to forecast only having access to the REFS.
- In addition to evaluating the traditional ensemble probabilities, FFaIR will be evaluating a MLP for QPF probability of exceedance from the CAPS group.
- Continued evaluation of the CAPS Spatially-Aligned Mean (SAM) and a SAM with local probability matched mean (LPM) applied with the SAM methodology, called the SAM-LPM.

- Evaluation of a bias-corrected HREF MLP from ISU over domains of intense precipitation.
- Evaluate cluster analysis based on QPF from the REFS using the previous four cycles of the membership.
- Continue to explore the addition of an ERO risk category between a Slight and Moderate risk, called the Enhanced Risk.
- Explore including an intensity contour on the ERO, defined by exceeding the 10-y 1-h ARI threshold over multiple hours.
- Evaluate the performance (CSI, max QPF) of models and participants for specific 6-h precipitation extreme events via the MRTP.
- Explore participants' perception of probabilities through the MRTP.

3 Verification Methods

Subjective verification and feedback from the participants is at the core of all testbeds. This type of feedback also helps better understand what the statistical results are telling us. For the majority of the subjective verification, "goodness" questions were asked, comparing the forecast to observations and asking the participants to rank them on a scale of 1 (very poor) to 5 (very good). Since the main focus of the FFaIR experiment was on comparing the RRFSp1 and the REFS to operational models, the HRRR, NAMnest, and RRFSp1 were used for deterministic analysis and the HREF and REFS for ensemble performance. The NSSL-MPAS was also included in the deterministic subjective verification so the FFaIR team could start very preliminary evaluation of an MPAS model since that is the planned replacement for the FV3-based REFS system for version 2 of the RRFS.

The Multi-Radar Multi-Sensor Gauge Corrected (hereafter MRMS) Quantitative Precipitation Estimate (QPE) was used as the rainfall observation. This was remapped to the HRRR grid, using the cKDTree package in Python, retaining the maximum value of the 9 grid point neighborhood. Object verification graphics were also created for the CONUS using the Developmental Testbed Center's Model Evaluation Tools (MET) Method for Object-Based Diagnostic Evaluation (MODE). These graphics allowed participants to focus on precipitation thresholds like one-half inch, one, and two inches when evaluating the models. This is done by creating objects based on the given threshold and matching the observed objects to forecast objects (see right column in Fig. 3 for example of what MODE graphics look like). The configuration used for MODE is the same as the previous two years and can be found in Appendix C. MET/MODE was also utilized for some of the statistical analysis discussed later.



Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.98828	23.84067	13.03048	1535	2140	1201
CF002_CO002	0.93174	72.45787	17.83766	2229	1239	888

Figure 3: Example of the CONUS 6-h verification for the 2024 FFaIR Experiment. The participants would see the graphics in a row for the MRMS QPE (left), Model QPF (center) and MODE (right). Each row was a different model: HRRR, NAMnest, RRFSp1 and NSSL-MPAS; HRRR is only shown here. On the verification page, the participants could pick the 6-h synoptic window being evaluated as well as the cycle. (A) 6-h MRMS QPE (B) 12z HRRR and (C) MODE graphic for 0.5” threshold valid 18 UTC 11 June 2024. The MODE graphics have the MRMS object contoured and the model object filled for the given threshold. Objects are matched by color. At the bottom of the graphics is information about the relationship of the QPE/QPF objects to one another.

On the QPE/QPF graphics, there is a grey box highlighting where the maximum value occurred on the domain evaluated. In the bottom left of the graphic is the value. Unfortunately, there were some erroneous MRMS maximum QPE values found during FFaIR. Most of these were related to new wind farms whose locations had not been input into the MRMS QC. Therefore, it was not always helpful to compare the MRMS max with the model max, unless it could be confirmed the max isn't driven by a false reading. This issue was most prevalent in TX and NE. The FFaIR team was in communication with the MRMS team, who have been updating their internal QC to remove the wind farm's false reflectivity but can not update the official QC due to the technical procedures involved to make a change to an official product.

Table 6: Table showing the verification date for the subjective verification sessions during FFaIR. Column 1 is the FFaIR Session date while Column 2 is the Day 1 (12 UTC-12 UTC) period evaluated for the given session's day. The following Columns provide the analysis times for the 6-h time periods evaluated for the CONUS deterministic and ensemble performance, the MRTP and MRTP∗ valid end times, and the CAPSe means and MLP valid end time. For Columns 6-9, if the cell is green that indicates that time period was evaluated over the CONUS for both deterministic and ensemble forecasts.

FFaIR Date	Verification Case Date	MRTP end time (UTC)	MRTP* - $2^{nd}/3^{rd}$ domain end time (UTC)	CAPS product's end time (UTC)	6h time period: 12-18 UTC	6h time period: 18-00 UTC	6h time period: 00-06 UTC	6h time period: 06-12 UTC
June 10	June 9-10	09	06		🟩			🟩
June 11	June 10-11	06	12			🟩	🟩	
June 12	June 11-12	03	00/21		🟩		🟩	
June 13	June 12-13	06	00		🟩	🟩		
June 14	June 9-10	06	09			🟩	🟩	
June 24	June 7-8	09	03				🟩	🟩
June 25	June 19-20	04	03		🟩	🟩	🟩	🟩
June 26	June 20-21	08	11			🟩	🟩	
June 27	June 21-22	07	00				🟩	🟩
June 28	June 24-25	09		12				
June 28	June 22-23	05			🟩	🟩		
July 8	June 28-29	11			🟩		🟩	
July 9	July 8-9	08	03	06	🟩			🟩
July 10	July 9-10	04			🟩			🟩
July 11	July 10-11	03	09	00		🟩	🟩	
July 12	July 11-12	11	12	12	🟩		🟩	
July 22	July 6-7	12				🟩	🟩	
July 23	July 22-23	04	10	00		🟩		🟩
July 24	July 23-24	05		06	🟩	🟩		
July 25	July 16-17	12	05		🟩	🟩		
July 26	July 25-26	12	20	00		🟩	🟩	
July 29	July 21-22	06				🟩		🟩
July 30	July 29-30	06	09	06			🟩	🟩
July 31	July 30-31	12		06	🟩		🟩	
August 1	July 31-Aug 1#	09		06		🟩	🟩	
August 2	Aug 1-2	04	06			🟩		🟩
	Aug 2-3	03				🟩	🟩	

Subjective verification was focused on 6-h QPF rather than 24-h QPF. This was done two different ways, first was to evaluate the CONUS QPF and various ensemble probabilities at the four 6-h synoptic time windows: 12-18 UTC, 18-00 UTC, 00-06 UTC and 06-12 UTC. The other was to evaluate the domain and 6-h time period for which the MRTP was valid. Additional 6-h domains were sometimes evaluated; these will be referred to as MRTP∗ and unlike the MRTP domains/times, were chosen after an event occurred. For both, the 18z (longest lead time), 00z, 06z, and 12z (shortest lead time) cycles of a given model/ensemble were evaluated. Due to survey fatigue and time constraints, typically only 2 of the 4 synoptic 6-h windows (chosen by the FFaIR facilitators) were evaluated on most days. Time periods with interesting weather or difficult forecasts were usually chosen and often the facilitators attempted to not have too much of the MRTP time period overlap with the synoptic windows chosen. Table 6 lists the valid times for verification each day of FFaIR. Figures 3-5 show what the participants would see for the verification of the CONUS 6-h QFE, 6-h domain specific QPF (hereafter referred to as MRTP and/or MRTP∗), and 6-h ensemble QPF probabilities.

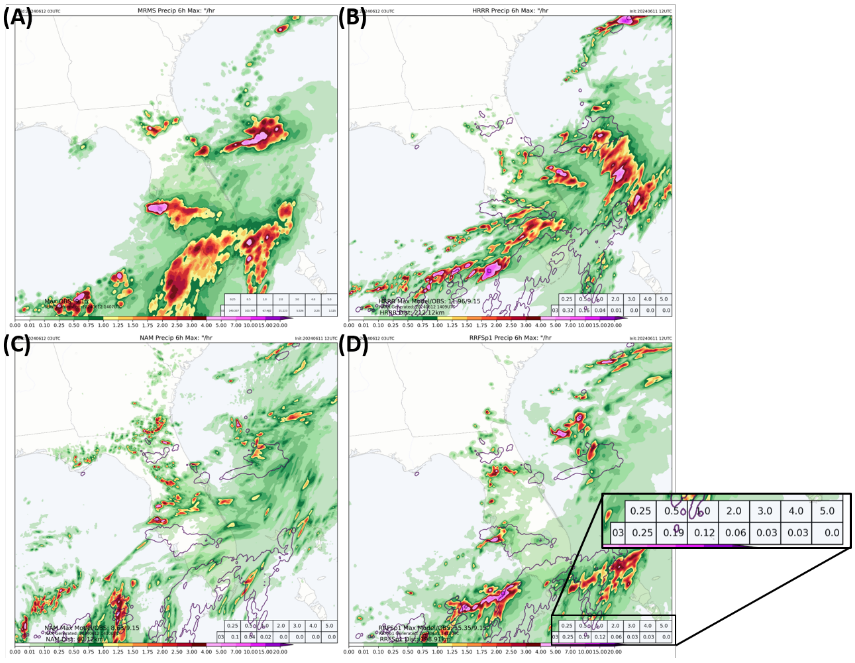


Figure 4: Example of the MRTP and MRTP∗ verification graphics. 6-h (A) QPE and 12z (B) HRRR, (C) NAMnest, and (D) RRFSp1 QPF valid 03 UTC 11 June 2024. The MRMS 1" object is outlines in purple on (B)-(D). In the bottom right of the graphics is a table showing the CSI at various thresholds; this is zoomed in for (D). Text on the graphic includes the MRMS and model max as well is the distance the observed max was from the model max.

For the evaluation of the REFS Clusters, the 75^{th} percentile was chosen empirically. This was used due to the size of the ensemble being evaluated; looking at any percentiles higher would be driven by only 1 or 2 members, while reviewing lower percentiles would not capture heavier QPF amounts, which was the main

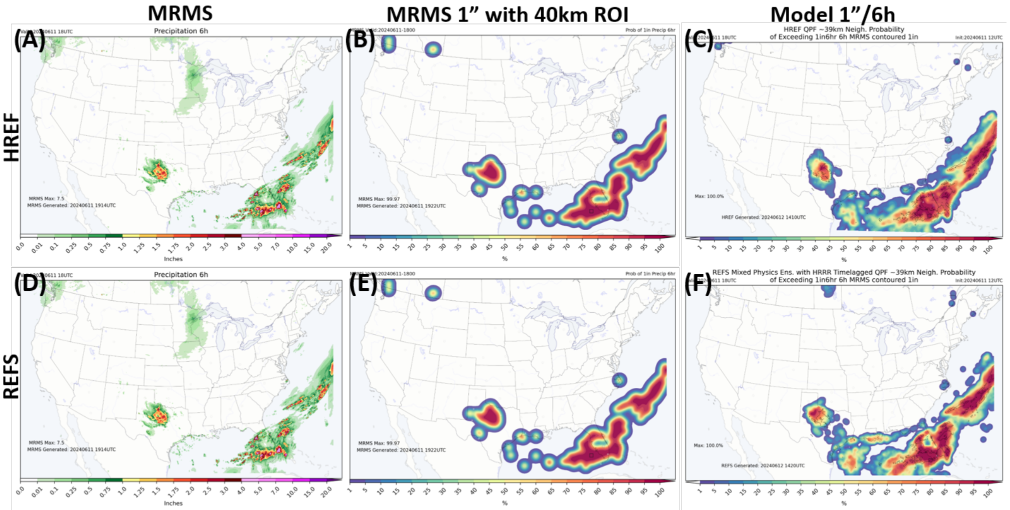

MRMS
MRMS 1" with 40km ROI
Model 1"/6h
HREF
REFS
(A)
Precipitation 6h
MRMS Max: 7.5
MRMS Generated: 20240611 1914UTC
Inches
(B)
MRMS Valid:20240611-1800
Prob of 1in Precip 6hr
MRMS Max: 99.97
MRMS Generated: 20240611 1922UTC
%
(C)
HREF QPF ~39km Neigh. Probability of Exceeding 1in6hr 6h MRMS contoured 1in
Init:20240611 12UTC
Max: 100.0%
HREF Generated: 20240612 1410UTC
%
(D)
Precipitation 6h
MRMS Max: 7.5
MRMS Generated: 20240611 1914UTC
Inches
(E)
MRMS Valid:20240611-1800
Prob of 1in Precip 6hr
MRMS Max: 99.97
MRMS Generated: 20240611 1922UTC
%
(F)
REFS Mixed Physics Ens. with HRRR Timelagged QPF ~39km Neigh. Probability of Exceeding 1in6hr 6h MRMS contoured 1in
Init:20240611 12UTC
Max: 100.0%
REFS Generated: 20240612 1420UTC
%

Figure 5: Valid the same time as Fig. 3 but showing the verification graphics/setup for evaluating the HREF and REFS. The MRMS 6-h QPE is on the left. In the center [(B) and (E)] is the 1" 6-h QPE with a 39-km radius of influence applied to create a probabilistic QPE. (C) and (F) are the 12z HREF and REFS, respectively, 1" 6-h^{-1} probabilities.

goal of the evaluation. The 75th allows for larger membership contribution while still highlighting higher values. Participants were shown the 4 Clusters and the full ensemble 6-h QPF 75th percentile, along with the 6-h MRMS and were asked which, if any, of the Clusters felt performed better than the full ensemble 6-h QPF 75th. An example of this can be seen in Fig. 6.

The ISU MLP was evaluated by comparing it to the standard EMC HREF products for 24-h. The probability matched mean (PMM) from the ISU MLP was compared against the local probability matched mean (LPMM) from the HREF. Ensemble probabilities compared were the probability of exceeding 1", 2", 3", 5" and 8" in 24-h. An example of the verification setup/graphics can be seen in Fig. 7. The ISU MLP for the HREF is domain specific based on where the machine learning determines a possible MCS. Because it is small domain, the PMM from the MLP is more comparable to the HREF CONUS LPMM[3], thus the reason that the HREF PMM was not compared to the ISU MLP PMM. As for the probabilities, the HREF probabilities use a ROI and smoother of roughly 39-km but the ISU MLP for the HREF used 10 km. This was chosen since the domain for the MLP

[3] PMM ranks and matches member magnitude across the CONUS while the LPMM does this over small patches (in similar size to the MLP domain) and then stitches these together.

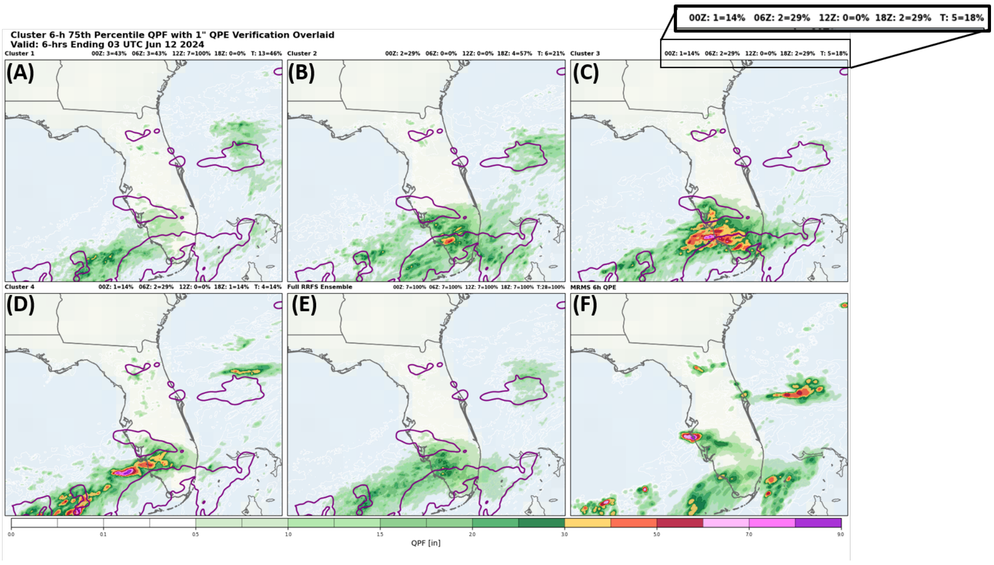

00Z: 1=14%   06Z: 2=29%   12Z: 0=0%   18Z: 2=29%   T: 5=18%
Cluster 6-h 75th Percentile QPF with 1" QPE Verification Overlaid
Valid: 6-hrs Ending 03 UTC Jun 12 2024
Cluster 1
00Z: 3=43%   06Z: 3=43%   12Z: 7=100%   18Z: 0=0%   T: 13=46%
(A)
Cluster 2
00Z: 2=29%   06Z: 0=0%   12Z: 0=0%   18Z: 4=57%   T: 6=21%
(B)
Cluster 3
00Z: 1=14%   06Z: 2=29%   12Z: 0=0%   18Z: 2=29%   T: 5=18%
(C)
Cluster 4
00Z: 1=14%   06Z: 2=29%   12Z: 0=0%   18Z: 1=14%   T: 4=14%
(D)
Full RRFS Ensemble
00Z: 7=100%   06Z: 7=100%   12Z: 7=100%   18Z: 7=100%   T:28=100%
(E)
MRMS 6h QPE
(F)
0.0
0.1
0.5
1.0
1.5
2.0
3.0
5.0
7.0
9.0
QPF [in]

Figure 6: Valid the same time as Fig. 4 but showing the verification graphics/setup REFS Clusters for the 6-h QPF 75th percentile for (A) Cluster 1, (B) Cluster 2, (C) Cluster 3, (D) Cluster 4 and (E) the full ensemble. (F) is the MRMS 6-h QPE. On (A)-(E) the purple contours are the 1" MRMS QPE. Along the top of each Cluster graphic are the number of members each cycle has in the cluster group, along with the total number of members. Note: the QPF/QPE scale plotted on these graphics differs from the other scales used for QPF/QPE graphics in FFaIR.

was small and in the hope that a smaller ROI and smoother would allow for more characteristics of the forecast to show through. The two questions asked for the subjective verification were:

- "How would you rate the ISU MLP guidance performance for PMM compared to the standard HREF guidance?"
- "How would you rate the ISU MLP guidance performance for Exceedance Probabilities compared to the standard HREF guidance?"

The scale was from 1 to 5, with 1 being ISU MLP much worse than HREF and 5 being ISU MLP much better than HREF.

For evaluation of the CAPS products, the 6-h QPF mean, the Spatially-Aligned Mean (SAM) and the SAM-LPM were shown along with the CAPSe and REFS LPM. Participants were asked questions about the size and shape of each mean's footprint compared to observed, as well as the magnitude of the maximum. The possible choices were:

- Footprint is larger than observed
- Footprint looks similar to observed
- Footprint is smaller than observed
- Shape of footprint is similar to observed
- Shape depicts what you expected to happen based on observation
- Maximum is higher than observed
- Maximum is about the same as observed
- Maximum is lower than observed

Participants were able to chose any of the descriptions they thought applied.

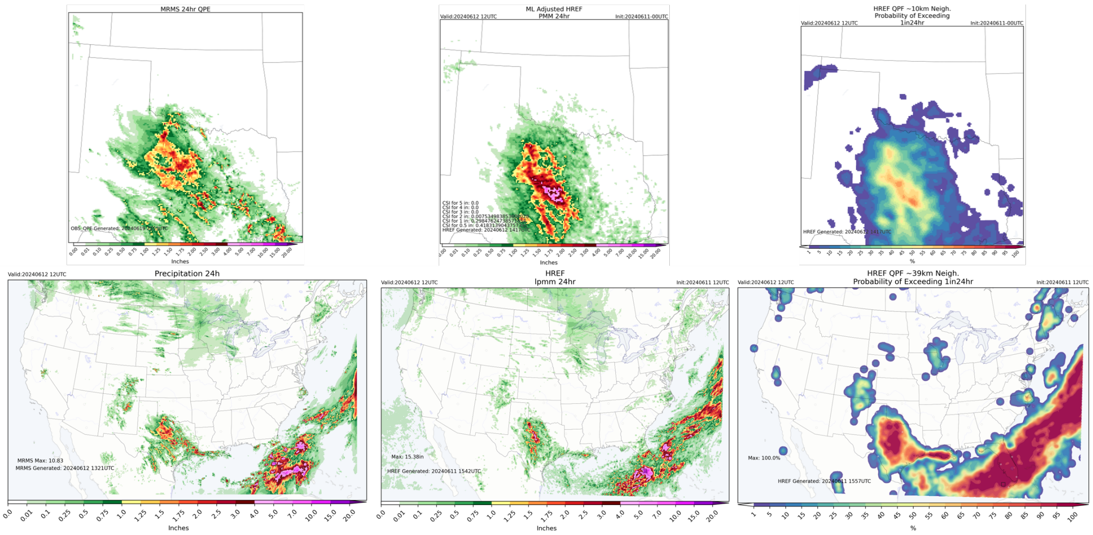


Figure 7: Example of the verification graphics for evaluating the standard HREF products and the ISU MLP for the HREF, valid 12 UTC 11 June 2024. Along the top are the graphics generated by the ISU team for their ISU MLP and along the bottom are the HREF CONUS graphics. The left column is the 24-h MRMS QPE. The middle column is the 24-h ISU MLP PMM (top) and HREF LPMM (bottom). The right column is the ISU MLP and HREF probability of 1"/24-h.

For the evaluation of the CAPS MLP for probability of exceedance, the 1"/6-h threshold was chosen and compared to the HREF and CAPSe. Participants were asked on a scale of 1 to 5 to "Rate how the 1"/6-h probabilities from the CAPS

MLP performed in comparison to the HREF/CAPSe", with 1 being CAPS MLP was much worse and 5 being the CAPS MLP was much better.

Bulk objective verification was done using a variety of methods and over different time periods and domains. MRMS QPE was the observation dataset used for every verification method done but there was some variability in how the MRMS was remapped to the coarser resolution of the models. For any analysis done using MET/MODE, the grid mapping of MRMS to the model or ensemble grid was done using the configuration described in Appendix C. For all other verification analysis methods, the aforementioned method of remapping the MRMS to the HRRR grid was done. Verification done using MET/MODE encompassed the period from 12 UTC 01 June to 12 UTC 04 August 2024. MET/MODE verification is CONUS wide and was done for 24-h and 6-h rainfall. Additional deterministic CONUS wide verification was done across what will be referred to as the Testbed Season: valid 12 UTC 02 May to 12 UTC 14 August 2024. CONUS wide ensemble verification was valid 12 UTC 08 June to 12 UTC 06 August 2024; this was shorter due to missing REFS runs in early June and mid-August.

6-h verification was done for the MRTP domains and additional domains chosen throughout the Testbed Season by the FFaIR facilitators. Each domain covered roughly a 10°x10° latitude/longitude box. How MRTP domains were chosen is discussed in Section 2.2 and were picked prior to an event occurring. The rest of the domains were chosen after the event occurred and based on if the event resulted in a high impact flooding event, had 6-h rainfall $\geq$3", were challenging to forecast, and/or had high areal coverage of one inch or greater of 6-h rainfall. Evaluation on only the domains that were done as a part of the MRTP activity will be referred to as the MRTP verification. Analysis done for all the domains during the Testbed Season, including the MRTP domains, will be referred to as MRTP+, which has 129 6-h domains in total.

4 Brief Summary of the Weather During FFaIR

The 2024 FFaIR verification season, May 1st to August 12th included two hurricanes (Beryl and Debby), two dam breaks, and nearly weekly burn scar flooding near the town of Ruidoso, NM. Figure 8 summarizes the season in terms of ERO risk, Flash Flood Warnings (FFW), and Local Storm Reports (LSRs) with some of the more notable events highlighted in Fig. 8B. Figure 9 shows the total rainfall for the same period for this year's FFaIR and the 2023, 2022, and 2021 FFaIRs. After two years of relatively dry summers across the Midwest and the Great Lakes, due to a lull in Mesoscale Convective Systems (MCSs), this year saw an active MCS season. However, there was a decrease in rainfall for this time period over the Northern Plains, while from southern Ohio to western Maryland/Virginia saw drought conditions.

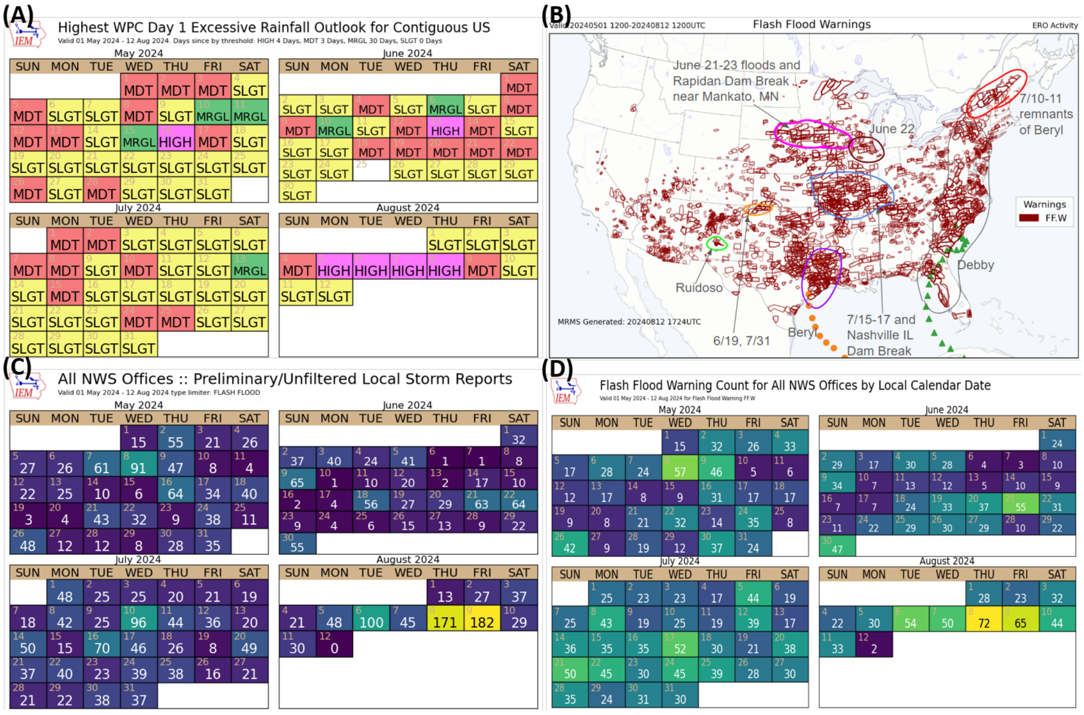


Figure 8: (A) Highest Day 1 ERO Risk, (B) issued FFW along with notations of significant events during the time period, (C) all NWS Offices' LSRs, and (D) the number of FFWs issued across all NWS offices from 01 May to 12 August 2024.

During the first week of FFaIR, Florida was impacted by a weak tropical disturbance, resulting in near to record breaking precipitable water (PWAT) across southern FL (NWS-Miami, 2024). From June 11-13 portions of southern FL saw 10-15" of rainfall, with areas of Miami-Dade County seeing over 21", which resulted in a Flash Flood Emergency being issued for Miami. On the other side of the

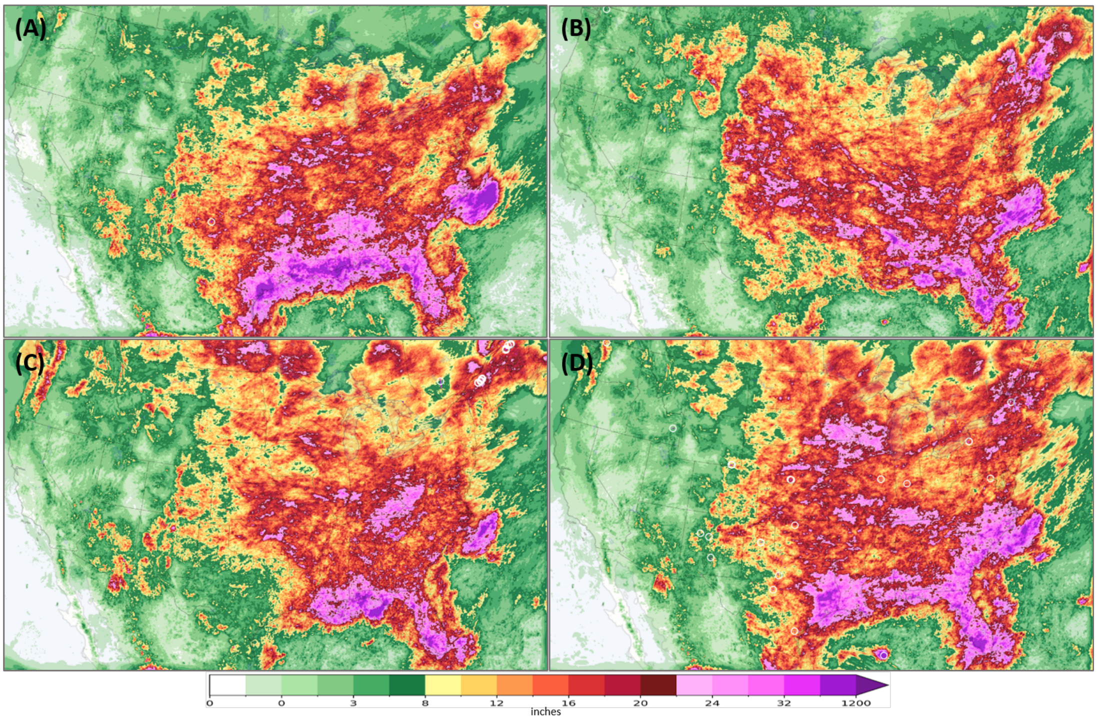


Figure 9: Accumulation of MRMS QPE from 01 May to 12 August (A) 2024, (B) 2023, (C) 2022, and (D) 2021.

state, Sarasota saw nearly four inches in an hour and over 8" in three hours in the barrier islands to its west. Hurricane Beryl made landfall in eastern Texas in early July, during week 3, with its remnants bringing heavy rainfall and flooding to northern VT on July 10^{th} and 11^{th}. Hurricane Debby formed at the end of FFaIR and thus was not actively evaluated by participants, however it is included in the bulk evaluation done for the Testbed Season: 02 May to 14 August 2025.

Unfortunately, two dam breaks occurred during FFaIR, the Rapidan Dam near Mankato, MN and the Nashville, IL dam. The dam break in MN was the result of several days of heavy rainfall over the region from June 21-23, with some parts of the area seeing 15 inches of rainfall in 48-h (NBC-News, 2024). The performance of the models and ensembles during these events will be discussed in great detail in the results section. The Nashville, IL reservoir dam collapsed on July 16 due to excessive rainfall that occurred (KSDK-News, 2024), with the area seeing over 6 inches in 12-h. FFaIR was not in active session for this event, nor was this event run as one of the retrospective cases. That said, Fig. 10 shows the 00z and 12z

6-h QPF from the HRRR, NAMnest, RRFSp1, and NSSL-MPAS for part of this event. Note that the 00z RRFSp1 has a significantly different event evolution than the MRMS and the other models while the 12z forecast is more similar to MRMS and the other models. This stark difference in performance between 12z RRFSp1 forecasts and those prior to the 12z cycle is a constant theme noted by the participants and will be discussed further in the Results section.

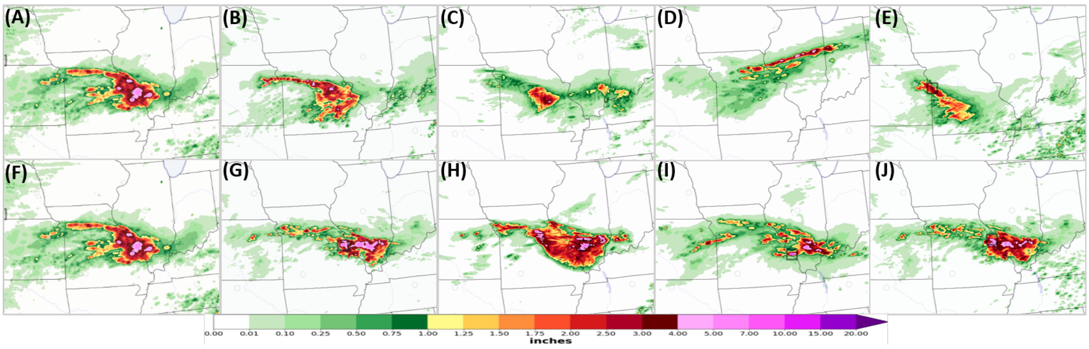


Figure 10: 6-h (A) and (F) MRMS QPE and (B) 00z HRRR, (C) 00z NAMnest, (D) 00z RRFSp1, (E) 00z NSSL-MPAS, (G) 12z HRRR, (H) 12z NAMnest, (I) 12z RRFSp1, and (J) 12 NSSL-MPAS QPF valid 18 UTC 16 July 2024.

5 Results

This section will discuss the results from both the subjective and objective verification done for the products evaluated in the 2024 FFaIR Experiment. Not all of the products evaluated will include a quantitative analysis. The majority of this section will focus on the evaluation of the RRFSp1[4] and the rest of the REFS FV3 membership, hereafter referred to as the REFS members or membership, compared to the HRRR and the NAMnest as it relates to QPF. As noted in Section 3, QPF was compared against the MRMS-GC QPE. Additionally, the 18z cycle of the models/ensembles during subjective verification is the oldest cycle (i.e. has the longest lead time) of the available model runs.

[4] The deterministic member of the REFS, which is often referred to as RRFS_a or just RRFS.

5.1 Analysis of the Deterministic QPF

Analysis of the deterministic forecasts, for both subjective and objective verification, was done two different ways, one being CONUS wide and the other being domain specific. Subjective evaluation occurred for the days that FFaIR was in session; Table 6 shows the dates verified. Section 3 explains in detail the subjective methodology of evaluation, but as a reminder, for the domain specific subjective verification results, there are two different ways the analysis was done: MRTP-only results (27 domains) and the results from all the domains that were subjectively evaluated, referred to as the MRTP$*$, which has 45 domains.

For subjective evaluation, participants scored the available 18z, 00z, 06z, and 12z cycles for the HRRR, NAMnest, RRFSp1, and NSSL-MPAS. Table 6 shows the 6 h time periods that were evaluated over the CONUS (cells highlighted green) as well as the valid end time of the MRTP and the MRTP+ domains, while Figure 15A shows the location of all of the MRTP domains examined during FFaIR. Scores were given on a scale of 1 to 5, with 1 being a very poor forecast and 5 being very good. The NSSL-MPAS was missing about one-third of the time during the experiment and therefore had fewer times scored than the other models. Because of the amount of missing data and since the NSSL-MPAS was not the main focus of the FFaIR, the general discussion of the deterministic subjective performance will exclude the NSSL-MPAS, though it will still be represented in the graphs/charts shown. A brief discussion of the NSSL-MPAS subjective feedback/performance will be provided in Section 5.1.5.

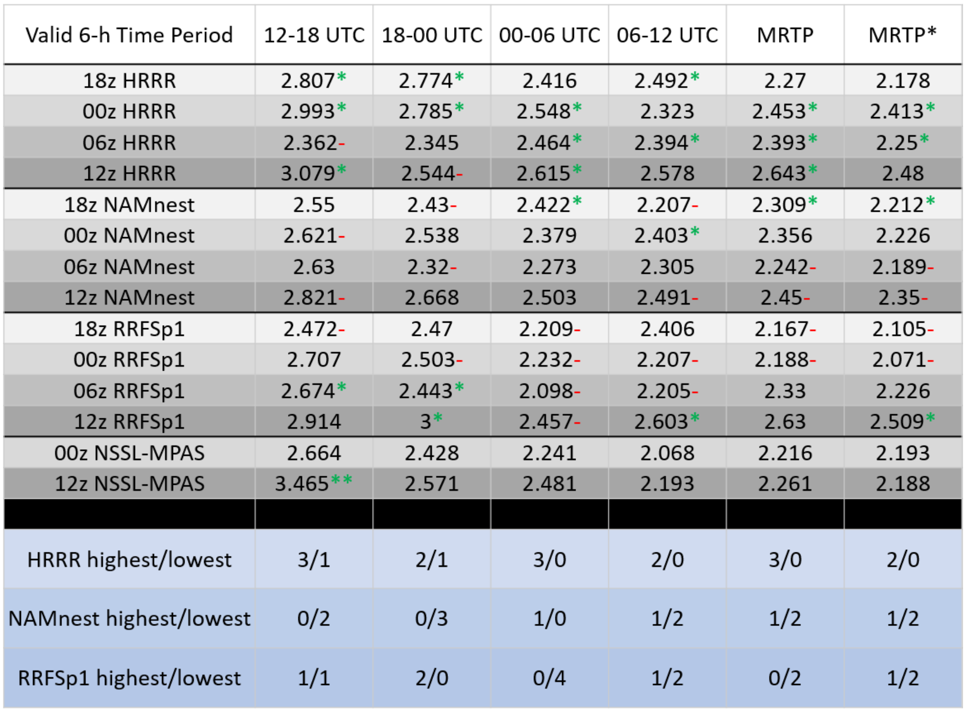
5.1.1 Subjective: CONUS

For the CONUS 6-h analysis, there were 16 different time period/cycle combinations (4 cycles multiplied by 4 6-h synoptic windows) that could be evaluated for a given case. The total mean score for each combination can be seen in Table 7. Among the HRRR, NAMnest, and RRFSp1, the HRRR was the most likely to have the highest total score, 10 times out of 16. The RRFSp1 had the highest total mean 4 times and the NAMnest had the highest mean twice. The RRFSp1 never had the highest mean for its 18z or 00z cycles, with its 00z cycle having the lowest total average compared to the HRRR and NAMnest for each period except the 12-18 UTC time frame. The 00-06 UTC time period (after the diurnal max) is when the RRFSp1 performed the poorest, with every model cycle having the lowest mean score of the three models. Furthermore, the lowest total average score for any model/cycle/time period was the 06z RRFSp1 (2.098) valid 00-06 UTC. The poor performance of the RRFSp1 after the diurnal maximum (after 00z for Day 1), was a consistent pattern seen during FFaIR and was often noted by participants as well as seen in the objective analysis. In total, although the RRFSp1 had the highest total mean score more often than the NAMnest (4 vs. 2), it also had the lowest mean score more often than the NAMnest (8 vs. 6), suggesting that there are significant jumps in forecast quality. Finally, the 12z RRFSp1 always had the highest mean score across the cycles for each CONUS time period when compared against its other cycles. The improvement of the 12z cycle for the RRFSp1 compared to its other cycles was often noted by the participants and was not always seen in the HRRR and NAMnest.

Figure 11A-D[5] show the number of times each model/cycle combination had the highest average score while Fig. 12 shows the percent of times each model received a score of 1 (very poor) to 5 (very good) for the CONUS 6-h analysis; refer to Appendix B for the numeric values for Fig. 12. For the first 6-h CONUS time window (12-18 UTC), despite the lack of an advanced data assimilation (DA) system, the NAMnest was competitive with the HRRR and RRFSp1, even for the 12z cycle, which had the shortest lead time (forecast hour 06). That said, Fig. 12D shows that the distribution of scores for the NAMnest were shifted more to the left (poor) than the HRRR and RRFSp1. Interestingly, even though the 06z RRFSp1 had the highest total mean score for this 6-h period, it had two fewer days with the highest daily average score than the NAMnest, again hinting at forecast quality inconsistencies. Finally, this time period saw the highest percentage of 5s (best score) for the 4 synoptic 6-h periods.

[5] Removal of the NSSL-MPAS from the dataset does not change the overall results of the model winner spread across the cycles. Generally, the HRRR was the model most likely to see an increase in the number of times with the highest daily mean per cycle when the NSSL-MPAS was not included. This is likely in part due to the fact that if there was a tie for the highest average, both models were rewarded.

Table 7: Table showing average subjective score across the FFaIR dates for a given model/cycle for the 6-h time periods evaluated for the CONUS deterministic and the MRTP and MRTP∗ domains. The highest average score among the HRRR, NAMnest, and RRFSp1 for a given cycle for the time period/domain has a green ∗ while the lowest has a red -; the differences between the lowest/highest scores are not necessarily statistically significant. There was an instance in which the NSSL-MPAS had the highest average; this is indicated by a green ∗∗. Also included is a summary of the number of times the HRRR, NAMnest, and RRFSp1 had the highest/lowest mean for each period/domain summed across the four cycles.



Valid 6-h Time Period	12-18 UTC	18-00 UTC	00-06 UTC	06-12 UTC	MRTP	MRTP*
18z HRRR	2.807*	2.774*	2.416	2.492*	2.27	2.178
00z HRRR	2.993*	2.785*	2.548*	2.323	2.453*	2.413*
06z HRRR	2.362-	2.345	2.464*	2.394*	2.393*	2.25*
12z HRRR	3.079*	2.544-	2.615*	2.578	2.643*	2.48
18z NAMnest	2.55	2.43-	2.422*	2.207-	2.309*	2.212*
00z NAMnest	2.621-	2.538	2.379	2.403*	2.356	2.226
06z NAMnest	2.63	2.32-	2.273	2.305	2.242-	2.189-
12z NAMnest	2.821-	2.668	2.503	2.491-	2.45-	2.35-
18z RRFSp1	2.472-	2.47	2.209-	2.406	2.167-	2.105-
00z RRFSp1	2.707	2.503-	2.232-	2.207-	2.188-	2.071-
06z RRFSp1	2.674*	2.443*	2.098-	2.205-	2.33	2.226
12z RRFSp1	2.914	3*	2.457-	2.603*	2.63	2.509*
00z NSSL-MPAS	2.664	2.428	2.241	2.068	2.216	2.193
12z NSSL-MPAS	3.465**	2.571	2.481	2.193	2.261	2.188
HRRR highest/lowest	3/1	2/1	3/0	2/0	3/0	2/0
NAMnest highest/lowest	0/2	0/3	1/0	1/2	1/2	1/2
RRFSp1 highest/lowest	1/1	2/0	0/4	1/2	0/2	1/2

The 18-00 UTC and 00-06 UTC times will be discussed together since they span the diurnal maximum for precipitation and convective initiation (CI) and dissipation; the latter 6-h period is also usually when Mesoscale Convective Systems (MCS) begin to form. Across these two time periods, there is a noticeable shift in RRFSp1's performance. As stated previously, the RRFSp1 always had the lowest mean score for 00-06 UTC. Although this in itself speaks to the struggles that the RRFSp1 had in forecasting precipitation moving into the overnight hours, the changes in mean scores and the distribution of scores help tell how different this shift in performance is between the two periods compared to the operational models. For instance, for the two cycles with the shortest lead time, 06z and 12z, the difference in total average score for the NAMnest and HRRR varies by about

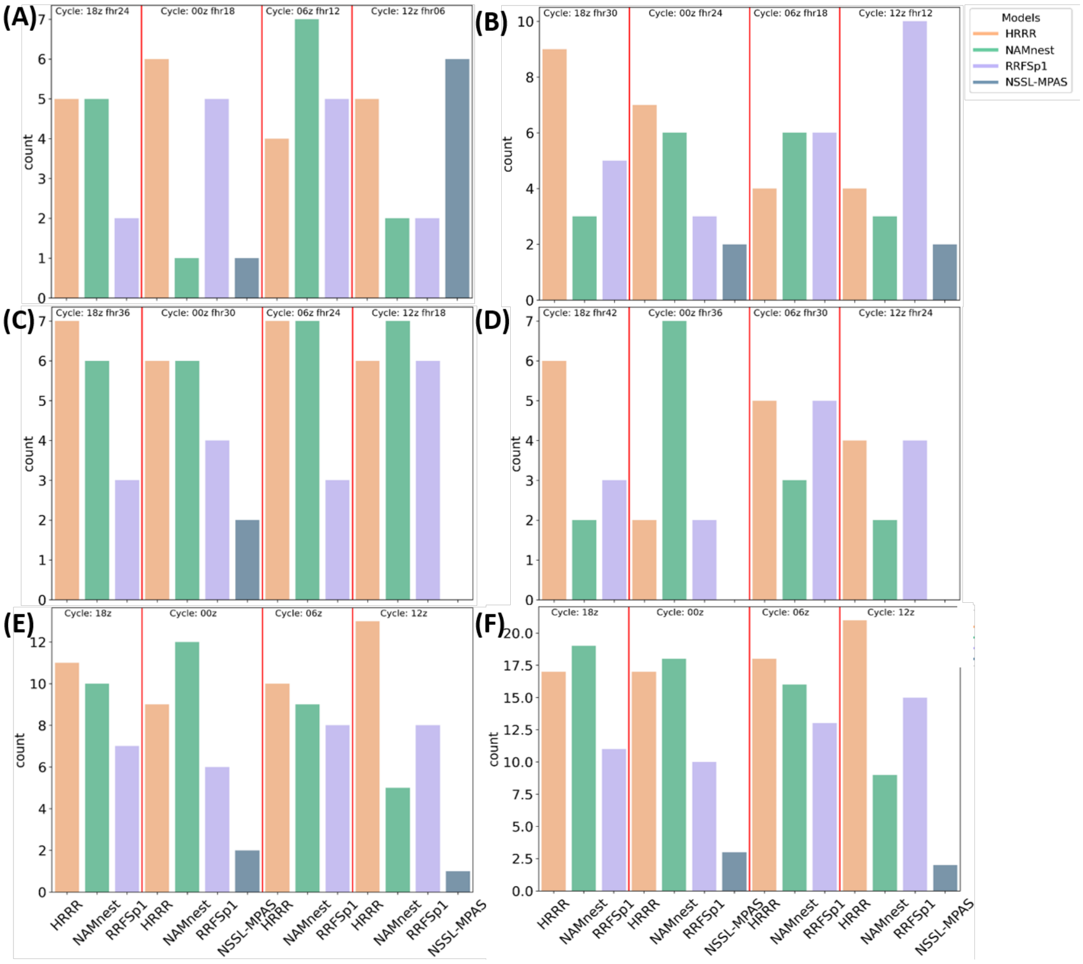


Figure 11: The number of times each model had the highest average subjective score for the day for the given 6 h period evaluated for the 18z, 00z, 06z and 12z cycles. The HRRR (orange), NAMnest (green), RRFSp1 (light purple) and NSSL-MPAS (grey-blue). (A)-(D) are for the CONUS subjective verification for the four synoptic windows for the Day 1 forecast: (A) 12-18 UTC, (B) 18-00, (C) 00-06 UTC and (D) 06-12 UTC. (E) and (F) show the MRTP and MRTP∗ results. The red vertical line indicates the separation between model cycles. Along the top of each cycle grouping is the cycle information and the valid forecast hour. Note: If there was a tie, then each model was counted as a winner. Also, not all 6 h time windows had the same number of times (days) evaluated.

0.15 while the RRFSp1 sees a decrease in score by 0.34 or greater. The change for the 12z RRFSp1 is the most drastic, going from a total average score of 3 to 2.457, while the HRRR/NAMnest went from 2.544/2.668 to 2.615/2.503.

Looking at the number of times each model/cycle had the highest daily mean score (Fig. 11B-C), two things stand out. First is the decrease in the HRRR's

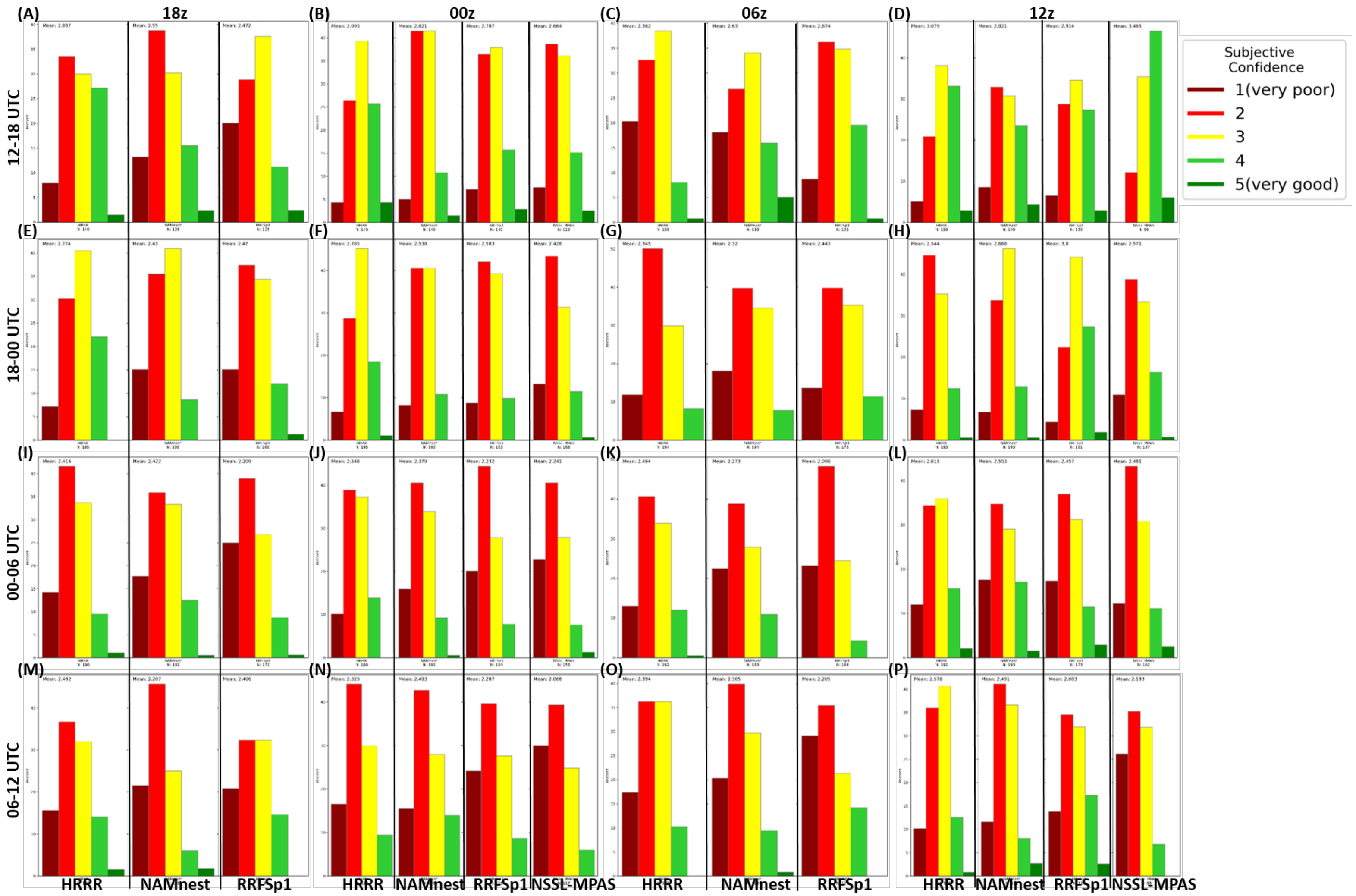


Figure 12: Percent of times each model/cycle combination received a score of 1 (very poor; dark red), 2 (red), 3 (yellow), 4 (green), and 5 (very good; dark green) for each 6 h synoptic time period for Day 1 during subjective verification of model performance of QPF. From top to bottom the synoptic time periods are: 12-18 UTC, 18-00 UTC, 00-06 UTC, and 06-12 UTC. The cycle evaluated from left to right is 18z (oldest cycle), 00z, 06z, and 12z (newest). Within the charts the models are listed HRRR, NAMnest, RRFSp1 and NSSL-MPAS (00z and 12z only). Refer to Appendix B for the numeric values.

number of highest daily mean scores as the lead time decreases while the scoring trend for the RRFSp1 increases for the 18-00 UTC period. Second is that for the 00-06 UTC period, the NAMnest and HRRR performed well and had comparable scores, while the RRFSp1 lags behind the two operational models except for its 12z forecast. The first point is likely driven in part by the HRRR's known bias in delayed development of CI (ex. James et al., 2022). The second point is likely due to two things; the RRFSp1's tendency quickly "kill off" convection after the diurnal maximum and difficulty initiating convection outside of the diurnal maximum. These tendencies can be seen in the objective analysis and will be discussed further in sections below.

Not surprisingly, the difference in distribution of the subjective scores for all models also differs across the two 6-h time windows; Fig. 12E-H for 18-00 UTC and M-I for 00-06 UTC. In general, the participants felt the forecasts for 18-00 UTC were most likely to be considered average to poor while for 00-06 UTC the forecasts were most likely to be considered poor. Furthermore, all model/cycle combinations saw an increase in the percentage of scores of a 1 (very poor) for 00-06 UTC forecasts when comparing these two time periods. This shows that, as a whole, when looking across the CONUS, models were felt to perform worse after the diurnal max. That said, the change specifically in the perceived performance of the RRFSp1 between these two periods is of concern given that for each cycle the RRFSp1 had the lowest total mean and its increase in the percentage of 1s (very poor) and 2s (poor) across the diurnal peak was always greater than the change seen for either of the two operational models. For instance, for the 12z cycles the percentage of 1s and 2s the RRFSp1 change from 26.7% to 54.3%, while the HRRR (NAMnest) ranged from 51.8% (40.4%) to 46.4% (52.3%). Additionally, for the good (4) and very good (5) scores, both operational models saw an increase in their percentage, 13% to 17.7% and 13.5% to 18.7% respectability, while the RRFSp1 saw a drop from 26.7% to 14.6%.

For the 06-12 UTC period, aside from the 18z cycle, the RRFSp1 had the highest daily mean score for the respective cycles the same number of times as the HRRR. The 00z NAMnest had the highest daily mean 7 times while the HRRR and RRFSp1 each had the highest mean twice. Even more interesting is that the

NAMnest had the highest total mean score for this cycle (2.403); HRRR was 2.323 and RRFSp1 was 2.207. For all but the 18z cycle, where it tied with the NAMnest, the RRFSp1 had the highest percentage of poor scores (1s). This was even true for the 12z cycle, despite having the highest total mean across all cycle/model combinations (2.603), albeit the difference in the poor score percentages among the models is small; HRRR 10%, NAMnest 11%, and RRFSp1 14%. In contrast, for all but the 00z cycle, the RRFSp1 had the greatest percentage of a score of 4 (good). This once more suggests inconsistencies in forecast quality.

Additionally, there are some general findings across the time periods that need to be mentioned. The 00z cycle for the RRFSp1 had the highest daily score the fewest number of times for all but the 12-18 UTC period. The poor performance of the 00z cycle for the RRFSp1 can also be seen in the total means, with the RRFSp1 having the lowest mean of the 3 models for all but the 12-18 UTC cycle. Differing from this was the performance of the 12z cycle. Here participants often noted that the RRFSp1 seemed to struggle forecasting events at other cycles, then at the 12z cycle it would suddenly depict the event. This can be most easily seen in the total means in Table 7, where the RRFSp1 is the only model where the highest total cycle mean is found at 12z for every period. In theory, this should be expected, the shorter the lead time, the better the forecast should be (less time for errors to accumulate). However previous FFaIRs and general discussion with forecasters have found that this is not always case. Thus, seeing this result would appear to be encouraging, but instead participants generally did not feel that way. They noted that since the model often had no signal of an event until the 12z cycle, they not only had lower confidence in the forecast being suddenly correct but they also would struggle to message for the event. For the operational models, even if the 12z cycle wasn't the best forecast, participants noted that at least they depicted an event across all the cycles, rather than it suddenly appearing, like what seemed to happen the RRFSp1. This feedback carried to the MRTP evaluation as well, and will be discussed below.

Another characteristic of the RRFSp1 that was mentioned, regardless of the 6-h time window evaluated, was the abundance of light precipitation. This was particularly notable across the southeastern US for air mass thunderstorms. The

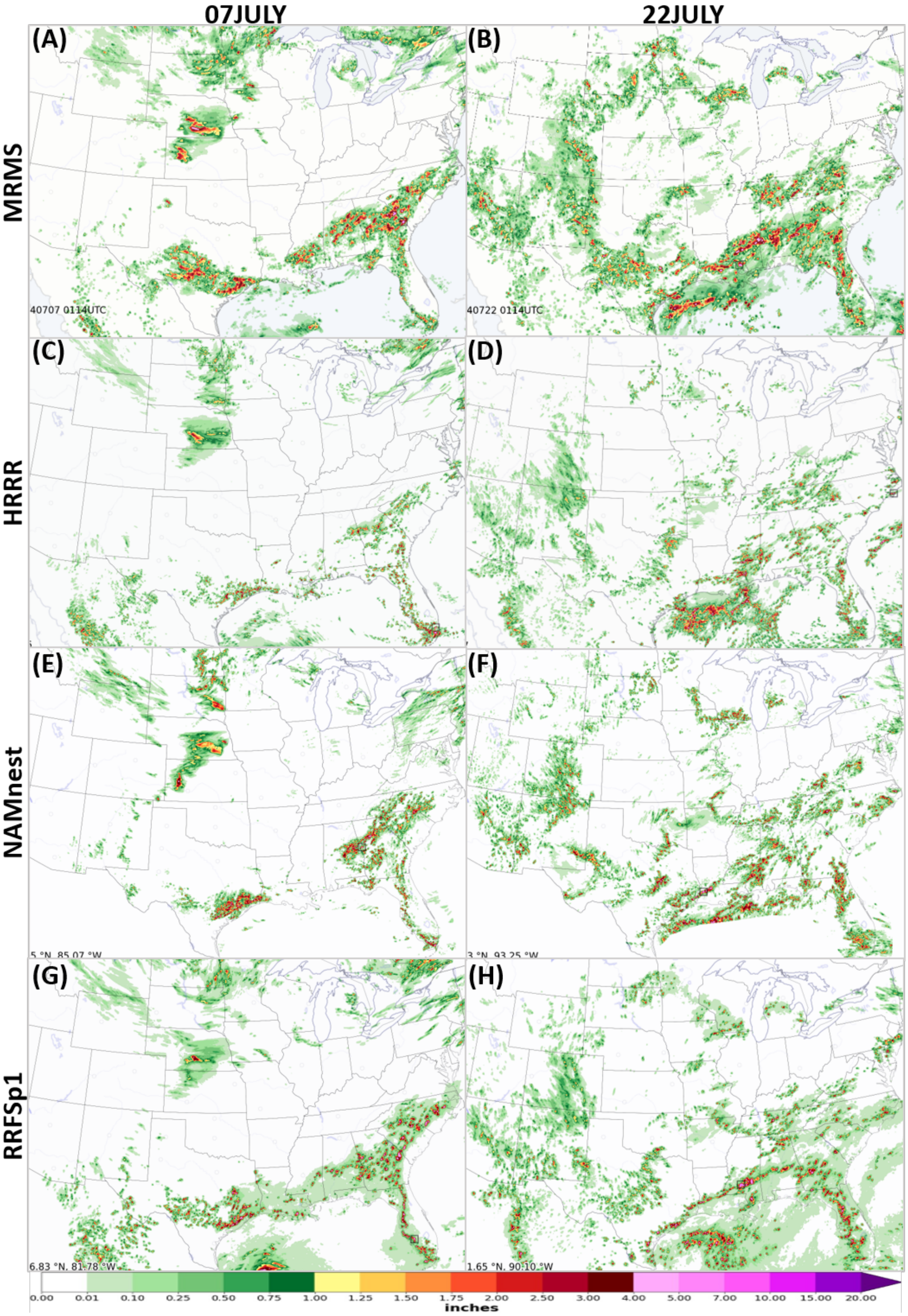


Figure 13: LEFT: 6-h MRMS QPE and model QPF valid 00 UTC 07 July 2024. RIGHT: 6-h MRMS QPE and model QPF valid 00 UTC 22 July 2024. (A) and (B) are MRMS. Models are initialized at 12z (left) 06 July and (right) 21 July. (C)-(D): HRRR, (E)-(F): NAMnest. (G)-(H) RRFSp1.

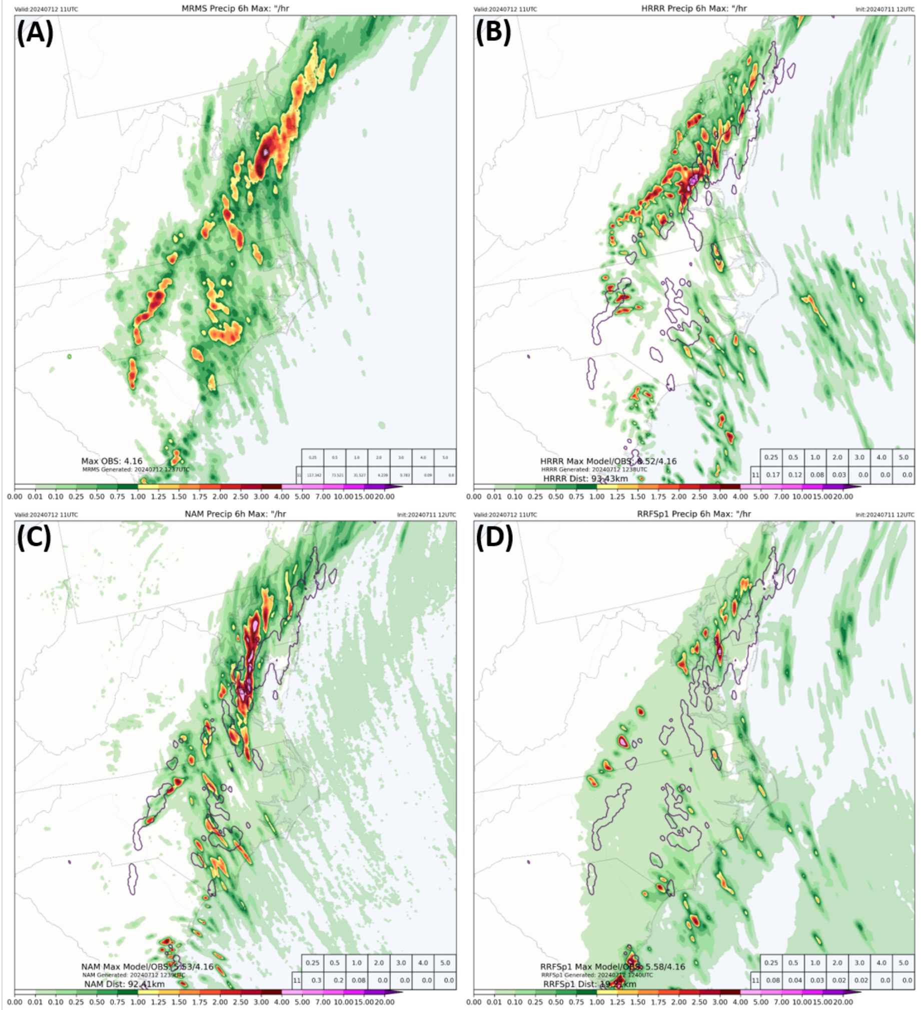


Figure 14: Verification graphics for the MRTP: 6-h (A) MRMS QPE and QPF from the 12z initialization on July 11 from the (B) HRRR, (C) NAMnest, and (D) RRFSp1, valid 11 UTC 12 July 2024. The MRMS 1" object is outlines in purple on (B)-(D).

RRFSp1 would forecast light precipitation anywhere there was instability driven by diurnal heating. Then contained within the light precipitation would be speckles of heavy rainfall. Sometimes these speckles or popcorn storms were so abundant it resulted in participants noting a high bias in magnitude. An example of both these instances can be seen in Fig. 13E-H, with light precipitation and random 3+ inches embedded across the Great Lakes region while across the southeast there is a sharp gradient between a trace (≤)0.01" and 3+ inches. In Fig. 13A-D (hereafter 07JULY), erroneous continuous light precipitation extended along the coast from

TX to VA. Farther north in KS/NE the RRFSp1 has light precipitation with a single nearly circular blob of 3+ inches; the HRRR has a similar forecast but its 1+ inch region shows a progressive structure to the footprint rather than a single, high value that appears unrealistic. This type of characteristic was also seen in more organized convection, sometimes even occurring when convection developed along a front, like is seen in Fig. 14; hereafter 12JULY.

5.1.2 Subjective: MRTP and MRTP∗

Discussion will now shift from the CONUS subjective analysis to the MRTP and MRTP∗ domains, which generally focus on the 6-h time period and domain in which the heaviest precipitation fell. The domains for the MRTP can be seen in Fig. 15A while Fig. 15B shows the MRTP∗ domains, including the domains analyzed outside of the subjective analysis. Table 6 provides the information for the valid end time of the MRTP and MRTP∗ forecast periods while the last two columns in Table 7 show the experimental means for the respective two domain classifications.

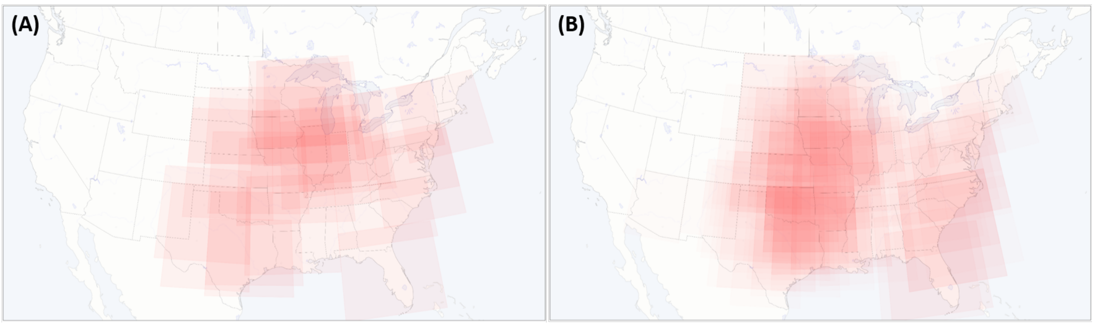


Figure 15: The domains for the (A) MRTP activity and the (B) MRTP∗ objective analysis for the 2024 FFaIR Experiment.

For the MRTP total average mean, the RRFSp1 never had the highest cycle mean but it had the lowest mean twice, for the 18z and 00z cycles. The HRRR had the highest total cycle mean for all but the 18z cycle, where the NAMnest had the highest. When the additional domains are added to the analysis (aka the MRTP∗), the NAMnest and RRFSp1 each had the highest cycle average once and the lowest cycle average twice. The 12z cycle of the RRFSp1 has the highest cycle mean for the MRTP∗. When shifting from MRTP to MRTP∗, the HRRR mean for the 12z cycle changes from 2.643 to 2.48 while the RRFSp1 saw a smaller change, 2.63 vs 2.509, thus resulting in the RRFSp1 overtaking the HRRR for the highest cycle mean. For all but the 12z cycle, for both the MRTP and MRTP∗ analyses, the RRFSp1 always had the fewest times in which it had the highest daily mean (Fig. 11E-F). For the 12z cycle, the NAMnest is the least likely to have the highest mean. It is this cycle in which the RRFSp1 "wins" the most times compared to its other cycles. Interestingly, even though the NAMnest has the lowest total mean for the 06z, it had the highest daily mean for that cycle (regardless of MRTP or MRTP∗) more often than the RRFSp1.

Fig. 16 show the distribution of the scores for the MRTP and MRTP∗ across the cycles. Perhaps not surprising, given the results from the total means, the RRFSp1 is the model most likely to receive a score of 1 (very poor) for all but the MRTP∗ 12z cycle. However, for a score of 4 (good) the RRFSp1 either tied or had a higher percentage than the NAMnest. When focusing on just the MRTP∗ results (which includes results from all domains subjectively evaluated), the NAMnest for all but the 00z cycle had the lowest percentage of 1s. However, it was also the model least likely to receive a score of 5 (very good). This is particularly interesting since the NAMnest is generally thought of as a poor performing model. However, these results and those from past FFaIRs suggest that the NAMnest has utility in identifying the extremes in precipitation for the late evening and overnight hours when MCSs are likely to occur (when the MRTPs are valid). The results also suggest that the RRFSp1 struggles in this aspect, which would coincide with the CONUS results.

Based on the subjective evaluation, the RRFSp1 seems to struggle in domains/6-h windows with heavy precipitation, though not always in the same way. Unlike in the past FFaIRs where participants overwhelming noted the wet bias from the RRFSp1, this year the model swung between too wet and too dry depending on the scenario. When the event is driven by strong convective dynamics, like the development of an MCS or explosive convection along an outflow boundary, the participants often mentioned that the RRFSp1 was too dry. This was especially true when the convection driving the event develops between 21 and 22 UTC

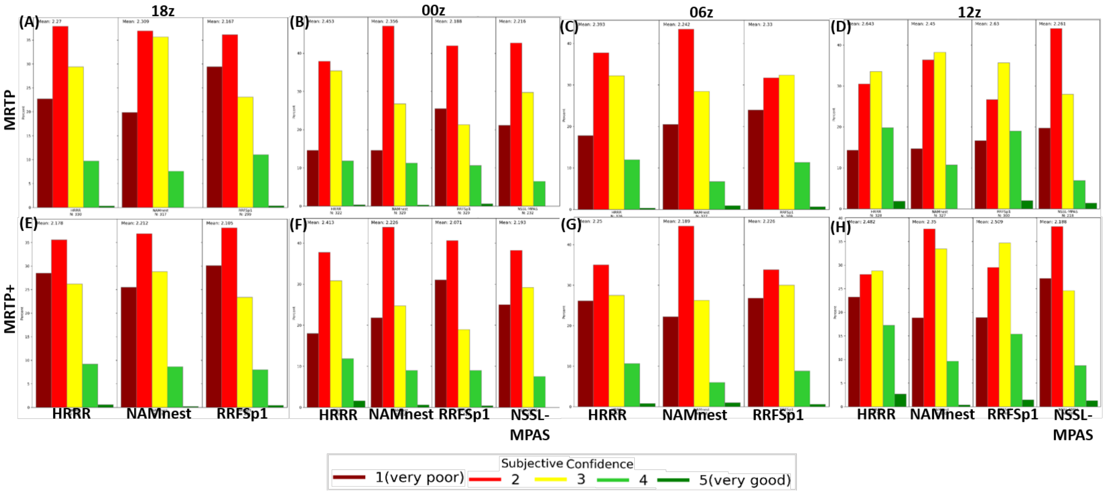


Figure 16: Like Fig. 12 but for the MRTP (top) and MRTP (bottom) subjective analysis, for the (A)/(D) 18z, (B)/(E) 00z, (C)/(G) 06z, and (D)/(H) 12z cycle scores.*

(after the local diurnal heating maximum) through roughly 06 UTC. This is most prevalent when organized convection was not already established in the area by the model prior to roughly 22 UTC. Analysis of the simulated reflectivity suggests that the CI is being suppressed and its energy dispersed or the simulated environment is too stable. This shows itself as ripples or waves emanating outward from where convection is expected to develop.

An example of the RRFSp1's inability to develop deep convection in situations like that described above can be seen in Figs. 17 and 18 for an MCS event from 24-25 June 2024 (hereafter 25JUNE). Although all models struggled with the evolution of this event, the RRFSp1 had no areas of one inch or greater QPF, meaning it had little signal of a heavy rainfall event[6]. Looking at the simulated reflectivity, it appears as though the RRFSp1 is attempting to develop convection (≥50 dbz) but instead, what look like gravity waves or undulations are simulated, similar to ripples from a pebble dropped in water. This results in widespread light precipitation with no real discernible features in the QPF pattern.

This feature of the RRFSp1 is particularly interesting when convection is ongoing prior to 21/22z. The model tends to simulate storm evolution correctly

[6] Although the 12z cycle is being used for this example, the same thing was seen in the other cycles.

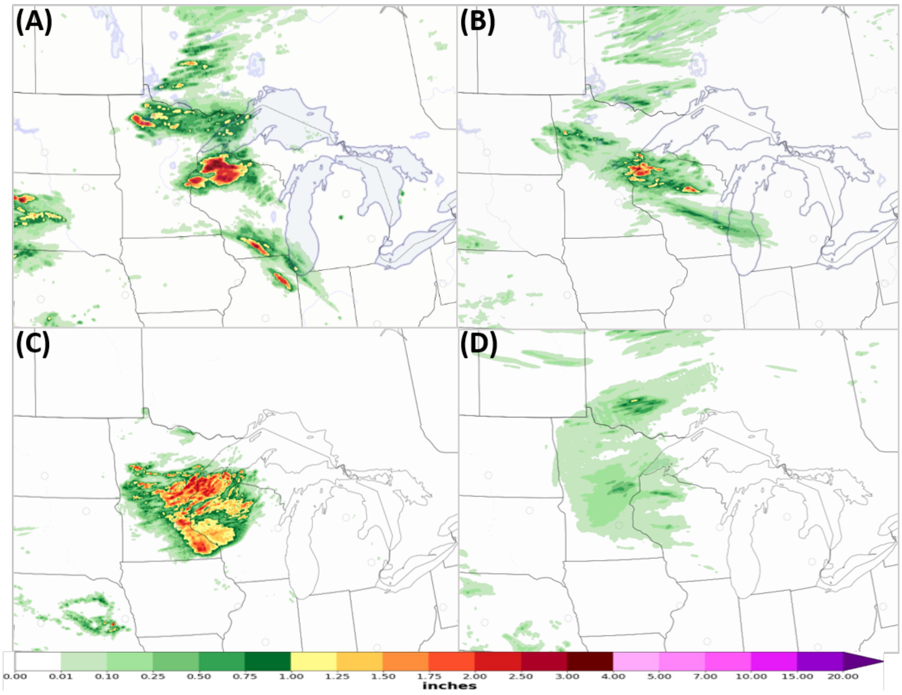


Figure 17: 6-h (A) MRMS QPE and (B)-(D) QPF from the 12z initialization on June 24 from (B) HRRR, (C) NAMnest and (D) RRFSp1 valid 06 UTC 25 June 2024.

until roughly 21/22 UTC. After this, any outflow or convergent boundary created from the ongoing simulated convection is muted to non-existent. An example of this can be seen from 20-21 June 2024 (hereafter 21JUNE) in Figs. 19 and 20. In Fig. 19, the magenta circle indicates where the RRFSp1 appears to attempt CI along the outflow from the northern convection that will eventually interact with the southern convection. But rather than organized convection, light simulated reflectivity is seen that slowly propagates south (looks like gravity bores). This does not interact with the southern system like is seen in the observations and the operational models, and the propagating area of simulated low reflectively creates erroneous widespread light precipitation along an east/west axis to the southeast ahead of the region of concern (Fig. 20).

Interestingly, the same apparent convective suppression does not occur along the outflow from the convection off the High Plains in WY/NE/CO. Additionally, this southern area of convection does not weaken after the diurnal max but rather develops into a MCS. Consequently, the event is not completely missed like the

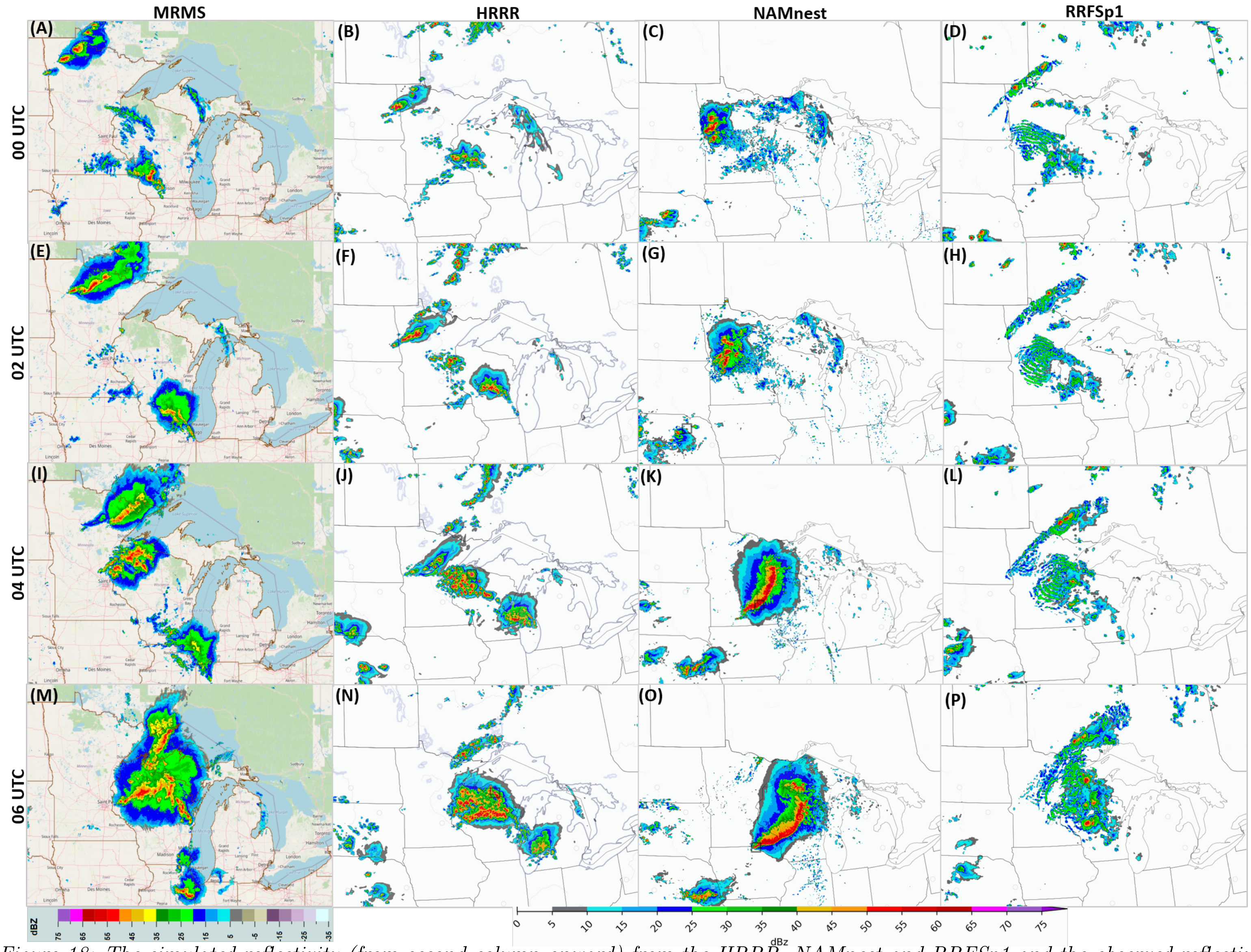


Figure 18: The simulated reflectivity (from second column onward) from the HRRR, NAMnest and RRFSp1 and the observed reflectivity (first column) for the event in Fig 17. From top to bottom, valid at: 00, 02, 04, and 06 UTC on 25 June 2024. Note the ripples in the simulated reflectivity for the RRFSp1 (last column).

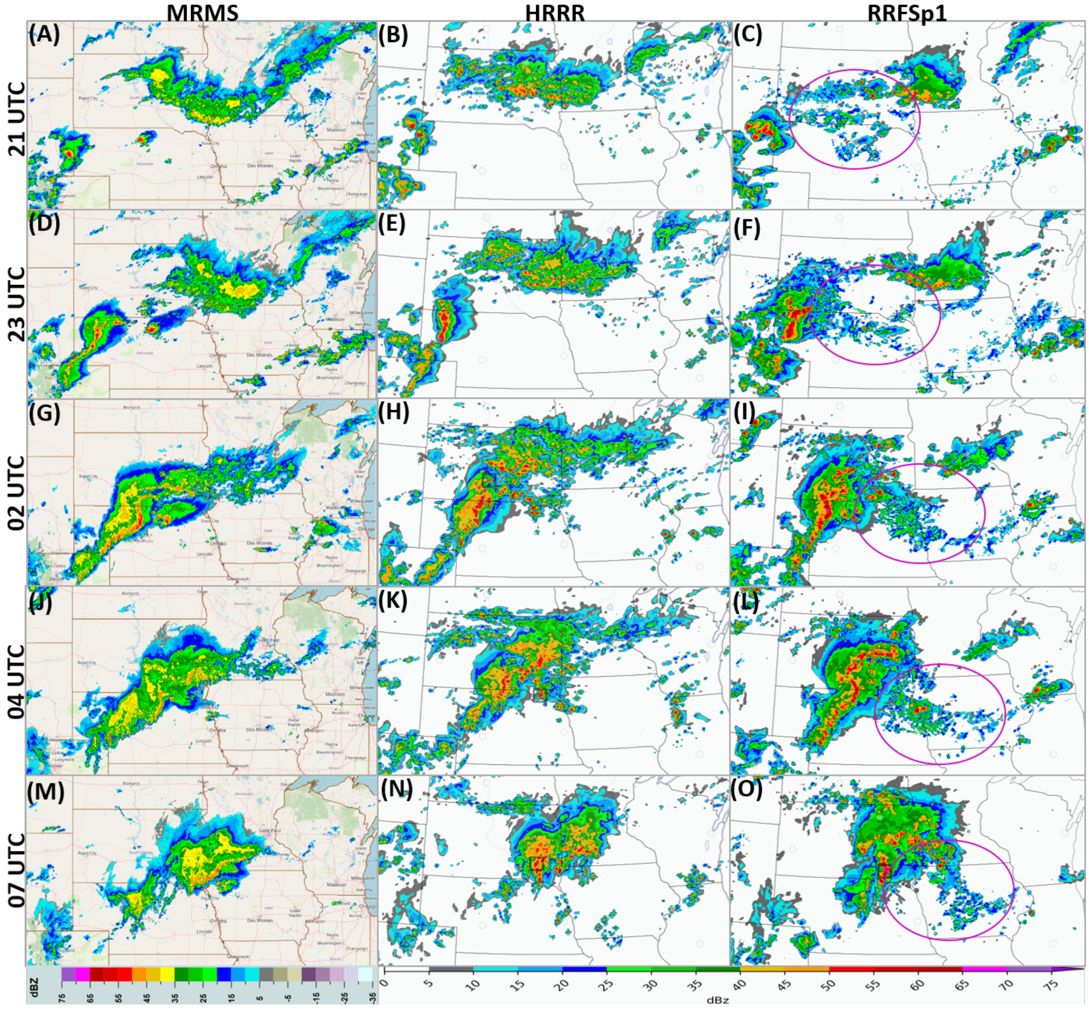


Figure 19: The observed reflectivity (left) and the simulated reflectivity from the HRRR (middle) and RRFSp1 (right) for the 21JUNE case. Reflectivity is valid (from top to bottom) at 21 UTC and 23 UTC 20 June and 02 UTC, 04 UTC, and 07 UTC 21 June 2021.

25JUNE event. However, the axis of the precipitation is incorrect; it is not east/west-oriented as seen in HRRR, NAMnest (not shown) and MRMS in the northern portion of the footprint. This results in the RRFSp1 having the MCS progress northeastward rather than east-northeastward, thus allowing significant precipitation to extend farther northward than the HRRR and NAMnest (not shown) by 11 UTC; see Fig. 20G-I. This resulted in a shift of mostly positive comments about the RRFSp1 for the 08 UTC verification to mostly negative for the 11 UTC verification. For example, a participant evaluating the 08 UTC time frame

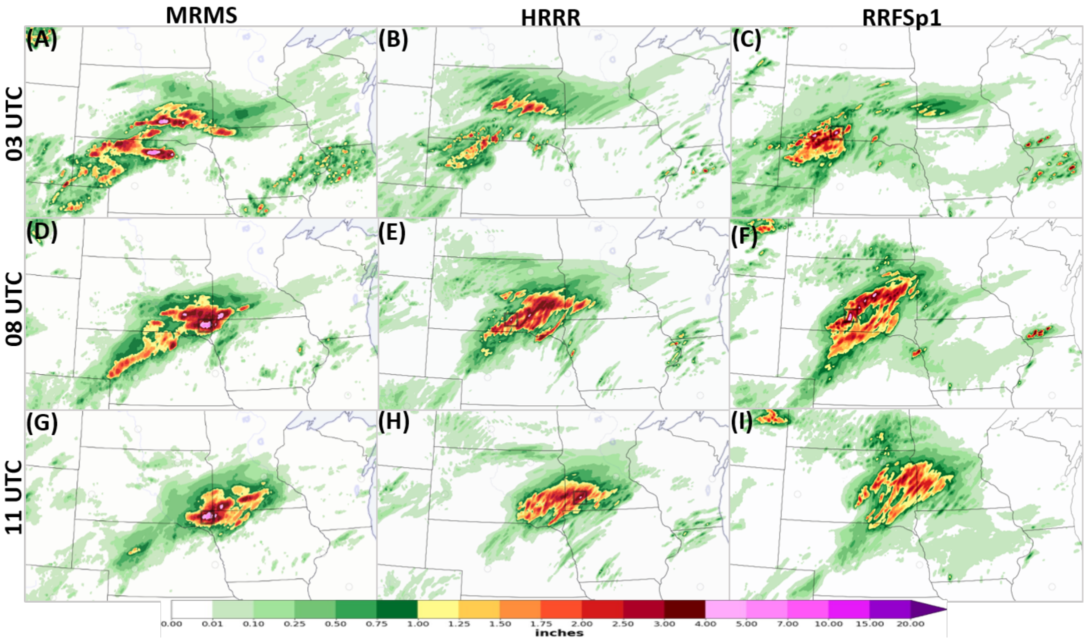


Figure 20: Similar to Fig. 19 but 6-h (left) MRMS QPE and QPF from the 12z initialization on June 20 from (middle) HRRR and (right) RRFSp1 valid t 03 UTC (top), 08 UTC (middle) and 11 UTC (bottom) 21 June 2024.

wrote: "The 12Z Cycle for all models created a forecast that, while displaced north, would absolutely have been helpful to understanding storm mode, QPF footprint, and maximum amounts. The HRRR/RRFS seemed a bit better with the HRRR performing the best at 12Z, while the NAMnest had the highest amounts and was displaced too far north and missed the SW flank extension that is pretty common in MCS like these." Then for the 11 UTC verification wrote "Once again 12Z suite was much better and would lead to an improved confidence forecast, but overall the CAMs really missed this event. That being said, HRRR/NAMNest at least produced some semblance of the footprint and amounts, RRFS really struggled with this being too far north."

When it comes to participant comments about the wet bias in the RRFSp1, the character and occurrence has changed from the previous FFaIRs. As noted above, often the forecasts were on the dry side due to what appears to be convective suppression by the model going into the evening hours. However, that does not mean that the RRFSp1 was not able to produce high QPF totals during the MRTP∗ events. If events were driven by large scale influences (like the remnants of Hurricane Beryl) or organized convection that is established prior to the diurnal maximum, the RRFSp1 had a tendency to be aggressive with totals. Generally, it was the magnitude of the QPF maximum that the participants commented on rather than the coverage itself. They mentioned that the areal coverage of high totals was small and isolated within light precipitation and looked unrealistic. That is not to say that there weren't instances of the over forecasting of the extent of higher totals, just that it was significantly less prevalent than in previous years. This is likely due to the suppression of explosive convection.

In regards to RRFSp1 performance, like was noted for the CONUS evaluation, the best-performing cycle of the RRFSp1 was at 12z. This was also true when compared to the NAMnest and HRRR, with the 12z RRFSp1 average MRTP$*$ score just barely edging out the HRRR's, despite the HRRR having more times in which it had the highest daily score. However, the RRFSp1's 18z and 00z forecasts had the lowest average score (refer to Table 7) of the three models. This supports the feedback that the RRFSp1 often seemed to completely miss heavy rainfall events until the 12z cycle, after which it seemingly "got it right". An example of this can be seen in Fig. 21; hereafter 22JUNE. The comments from the participants for this event are below; note the mention of the sudden improvement for the 12z cycle. Although the HRRR had a similar precipitation pattern as the RRFSp1 for the 00z and 06z cycles, it forecasted a larger footprint of $\geq$2".

- "The NAMnest did have some aspects of the double maxima rainfall early on but placement was well off while the RRFSp1 outlined very little rainfall and towards the last two cycles increased rainfall significantly along with developing the double maxima solution. The HRRR never really latched onto the double maxima solution and placement was slightly off."
- "The RRFS really had a great forecast once it figured it out, but it took a while to get there."
- "Was excited about location on 12Z RRFSp1, then saw the 2+ amount and wasn't quite as excited."
- "Once again there is a significant improvement in the 12z suite vs the prior cycles. The RRFS at 12Z had the best overall coverage and footprint of the

heavy rain, but all 3 models the last 2 cycles would have helped produce a forecast and message.”

- “Overall the Namnest seemed to do the best across all initializations. The 12z p1 did great with location. ”
- “Models did a pretty good job in this domain. HRRR struggled a bit with the convection west of Omaha, but main area of impact in Minn/Iowa forecasted fairly well.”

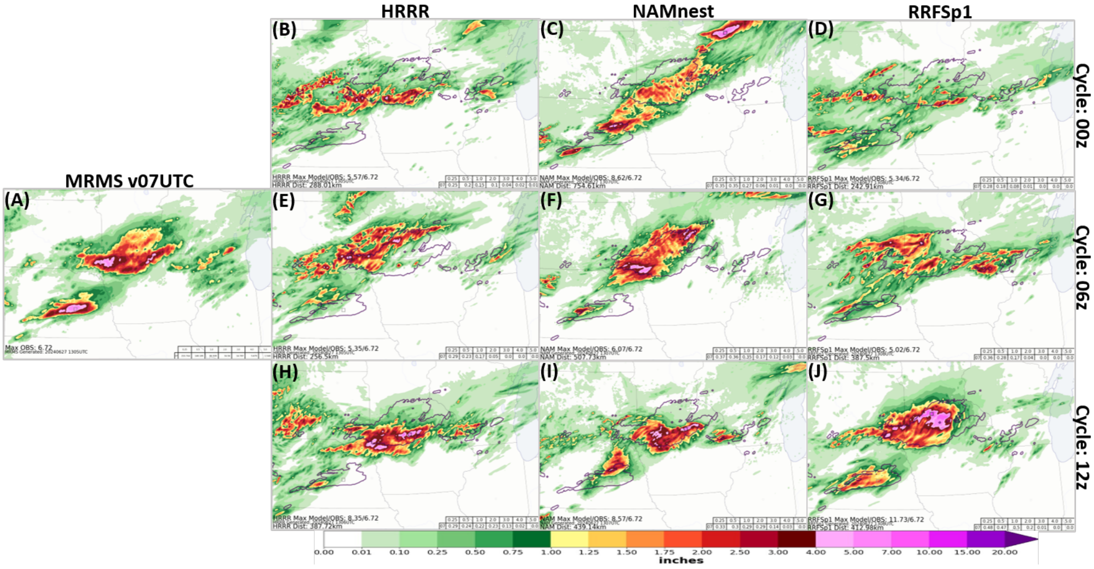


Figure 21: Verification graphics for the MRTP for 22JUNE event. 6-h (A) MRMS QPE and (B)-(J) QPF valid 07 UTC 22 June 2024. The MRMS 1” object is outlines in purple on (B)-(J). For (B)-(J) columns left to right are HRRR, NAMnest, and RRFSp1 while from top to bottom are the model’s initialization cycle: 00z, 06z, 12z.

5.1.3 Subjective: REFS Members

Although the REFS members 2-6 were not formally evaluated by participants during the daily verification sessions, participants in the REFS group had the members available to use in their forecasting process. Unfortunately, the 12z cycle of the REFS was the most likely to fail, so often the comparison of the REFS members to the control (RRFSp1) was done for the 06z cycle (since it had the next shortest forecast lead time for events). From the smaller dataset, similar characteristics found in the deterministic RRFSp1 can be seen in the REFS members. Specifically the over forecasting of widespread light precipitation with isolated high precipitation “popcorn” storms embedded within. An example of this can be seen for the 07JULY case in Fig. 22. Additionally, like the RRFSp1 (Fig. 13D), each member struggled to forecast the event in KS/NE regardless of cumulus parametrization scheme (hereafter CU scheme). In fact, RRFSm3, whose configuration differs only in the CU scheme (saSAS vs GF; see Table 3), was similar to RRFSp1 with the small maximum of 3”.

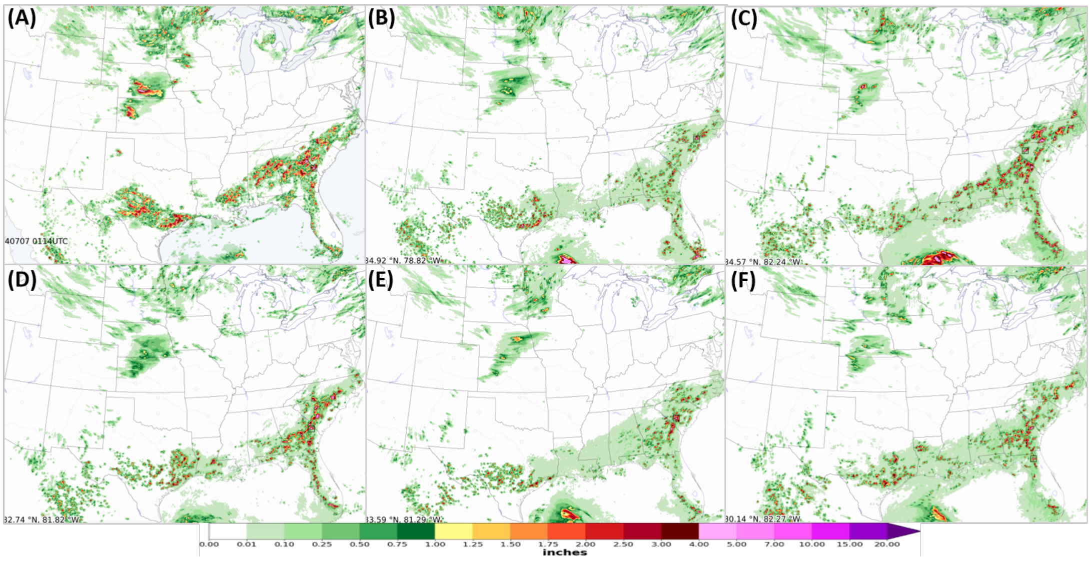


Figure 22: 6-h (A) MRMS QPE and (B)-(F) QPF initialized at 12z valid 00 UTC 07 July 2024 for the REFS membership: (B) RRFSm2, (C) RRFSm3, (D) RRFSm4, (E) RRFSm5, and (F) RRFSm6. Refer to Fig. 13B-D for the operational models and RRFSp1 forecast.

The impacts of the CU scheme on explosive convection were seen regardless of what scheme was used for the member configuration, though subjectively it seemed less egregious in the other members than seen in RRFSp1. For example, for the 25JUNE case (Fig. 23) the QPF for 3 of the 5 members shows at least a signal for two areas of ≥1”/6-h, each with a shift in location and size. Two of these members, RRFSm3 (Fig. 23C) and RRFSm6 (Fig. 23F) have saSAS deep as their CU scheme, while RRFSm4 (Fig. 23D) has G-F deep. The suppression of convection in the memberships’ simulated reflectivities can be seen in Fig. 24 and 25. For members 3 and 6, light simulated reflectivity (5dBz) can be seen during CI while a more ripple structure is seen in the other members. Note that by the end of the time period

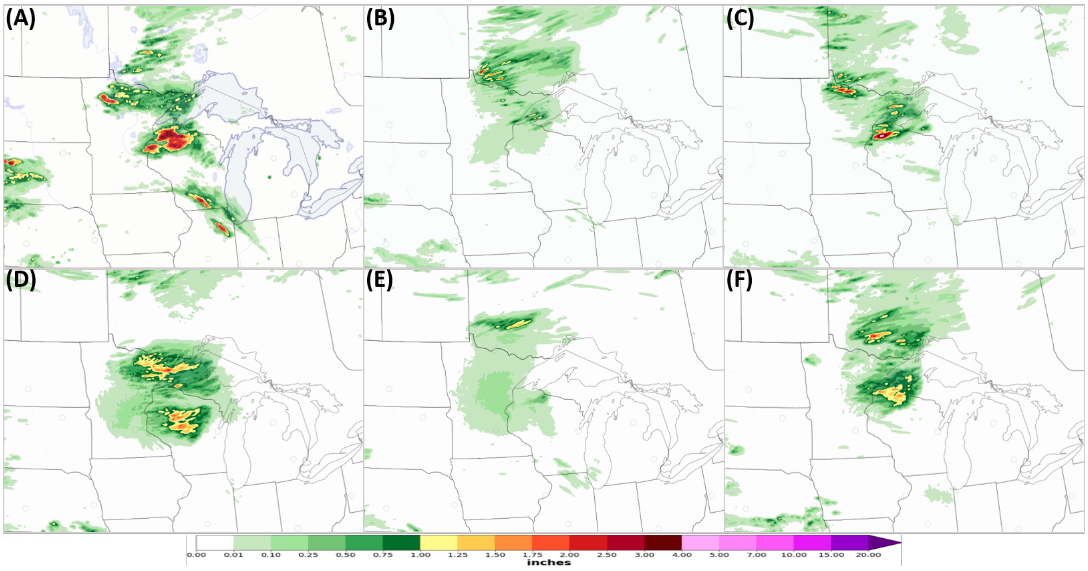


Figure 23: Like Fig. 17 but for the REFS members for the 25JUNE case initialized 06z 24 June 2024. (A) MRMS, (B) RRFSm2, (C) RRFSm3, (D) RRFSm4, (E) RRFSm5, and (F) RRFSm6.

(06 UTC), only two members have some signal of MCS development (RRFSm4 and RRFSm6). Most interesting is RRFSm3's forecast, although it is the only member to forecast over 1.5" in the region of concern, and in the correct location, the coverage of the ≥1.5" was much smaller than see in observations. Furthermore, the QPF was generated over a 2-h period by a large simulated cell rather than organized multi-cellular convection; see Fig 26. This shows that even with the CU scheme, the RRFS system is still able to produce the erroneous, large, high precipitation generating convective cells as seen in previous years.

As discussed above for the 25JUNE case, the characteristic of the CI suppression for RRFSm3 and RRFSm6 differs from the other REFS members, with low simulated reflectivity values (5-10dBz) that look as though they are a shadow of the higher values of simulated reflectivity (convection). Often the area of low simulated reflectivity is of similar shape and size to the closest simulated convection. The signature of this suggests the models are attempting to suppress strong CI that should be developing along the outflow boundaries of the already ongoing convection. The low simulated reflectivity values (5-10dBz) does not seem to be associated with light precipitation in the same location; see the example of 21JUNE, Fig. 27 across IA. Meanwhile, the ripple-like features seen in the other members are

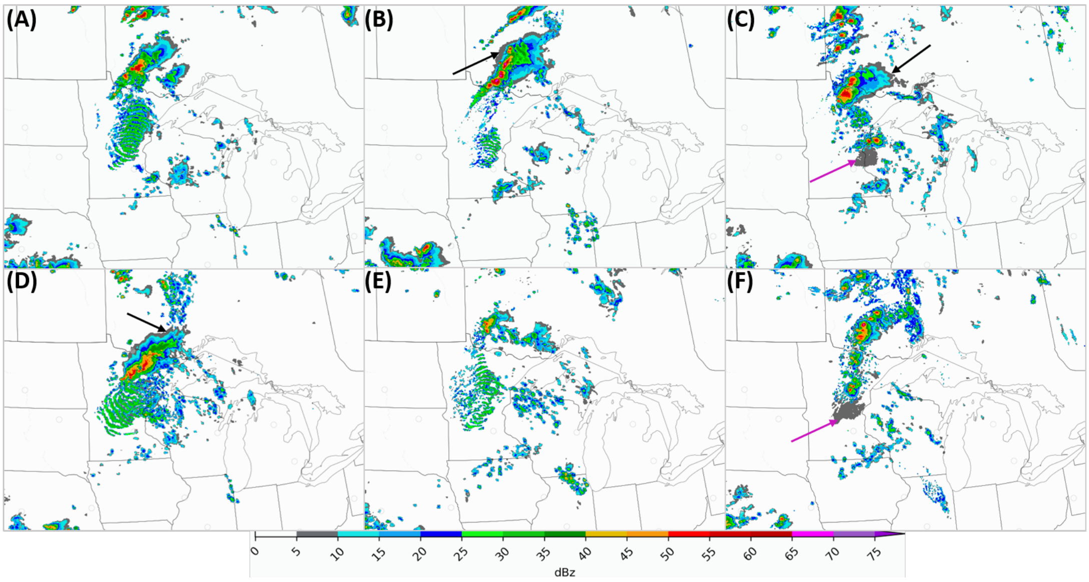


Figure 24: Simulated reflectivity (25JUNE case) valid at 02 UTC (A) RRFSp1, (B) RRFSm2, (C) RRFSm3, (D) RRFSm4,(E) RRFSm5, and (F) RRFSm6 on 25 June 2024, initialized 06z 24 June 2024. The pink arrows highlight the light simulated reflectivity (5-10dBz) signatures that appear to do due to suppression of CI while the black arrows show examples of how the low sim. ref. values typically are seen. Refer to Fig. 18E for the observed MRMS Reflectivity valid for this time.

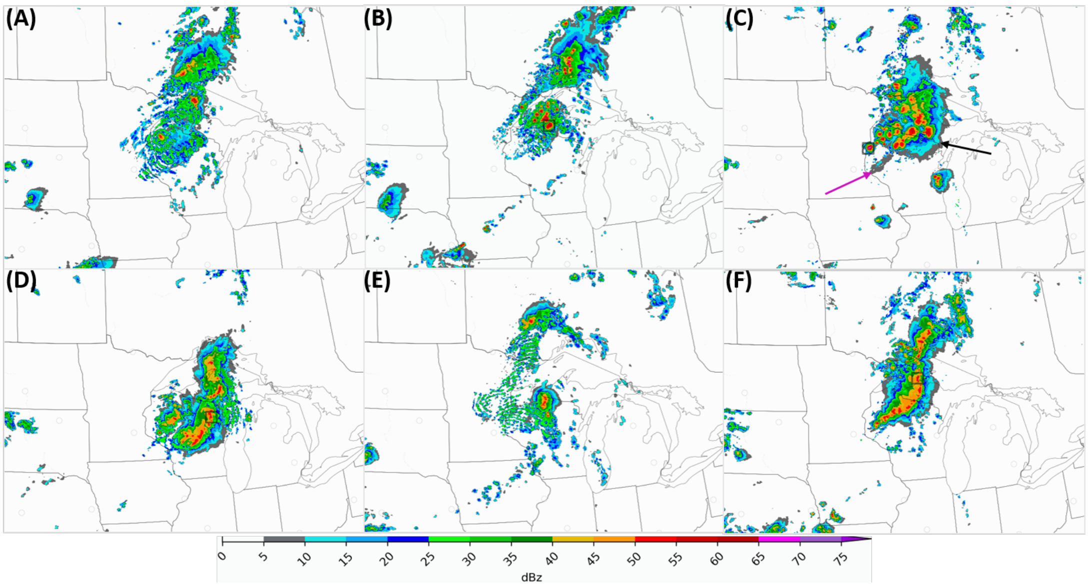


Figure 25: Like Fig. 24 but valid 06 UTC on 25 June 2024. Refer to first column in Fig. 18 for observed reflectivity. Refer to Fig. 18M for the observed MRMS reflectivity valid for this time.

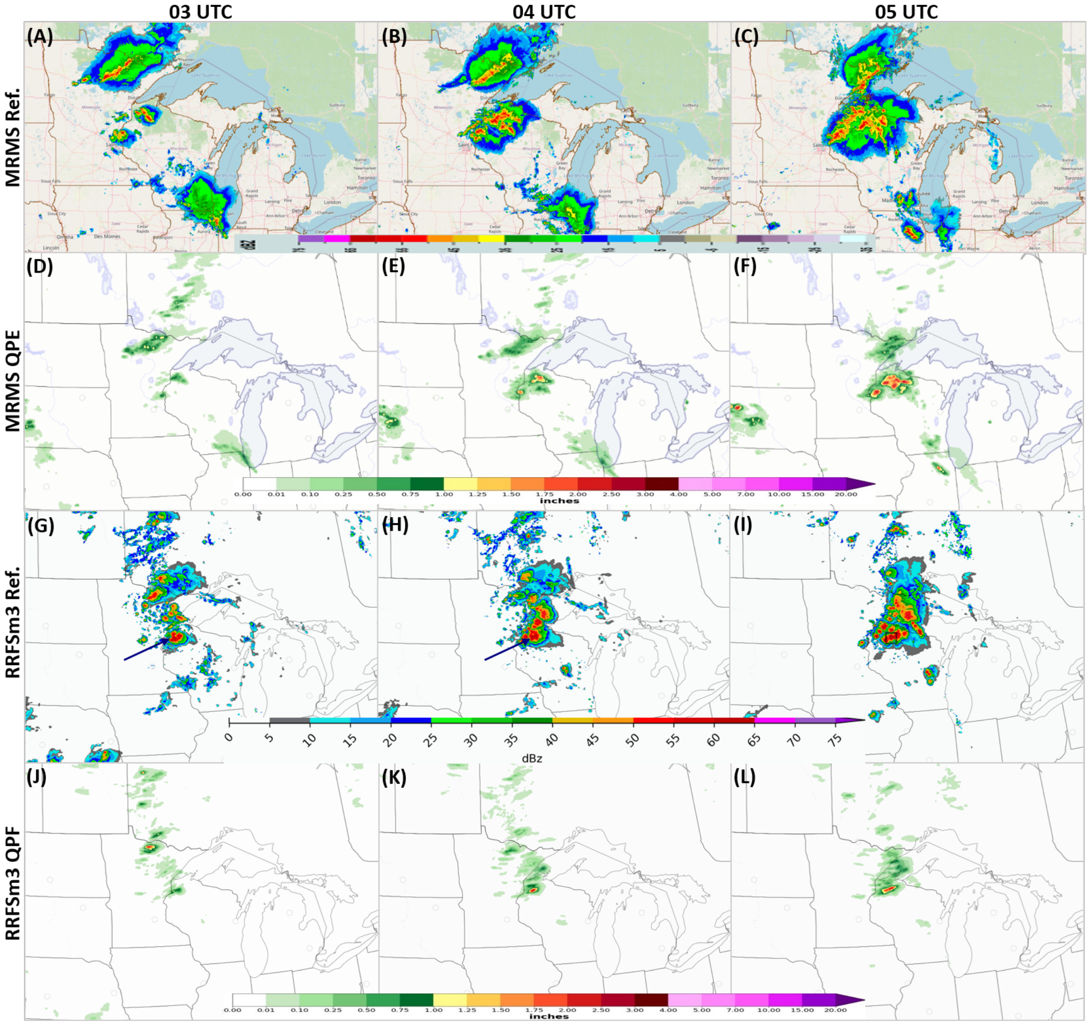


Figure 26: Hourly QPF/QPE and reflectivity for the 25JUNE case for the RRFSm3 compared to observations. (A)-(C) MRMS reflectivity and (G)-(I) RRFSm3 simulated reflectivity. 1-h (D)-(F) MRMS QPE and (J)-(L) RRFSm3 QPF. Dark blue arrow indicates the large convective cell discussed in the text.

associated with trace amounts of QPF. It is worth noting that the light simulated reflectivity (5dBz) is not absent in the other members, nor in operational models; its coverage is less prevalent. For all members, as well as for the operational models, light reflectivity is generally simulated in conjunction with well developed convection patterns, along the edges of the edges of the area of convection, not disconnected from the swath; the pink and black arrows in Figs. 24 and 25 show an example of the worrisome features and the typical way the low dBz is seen.

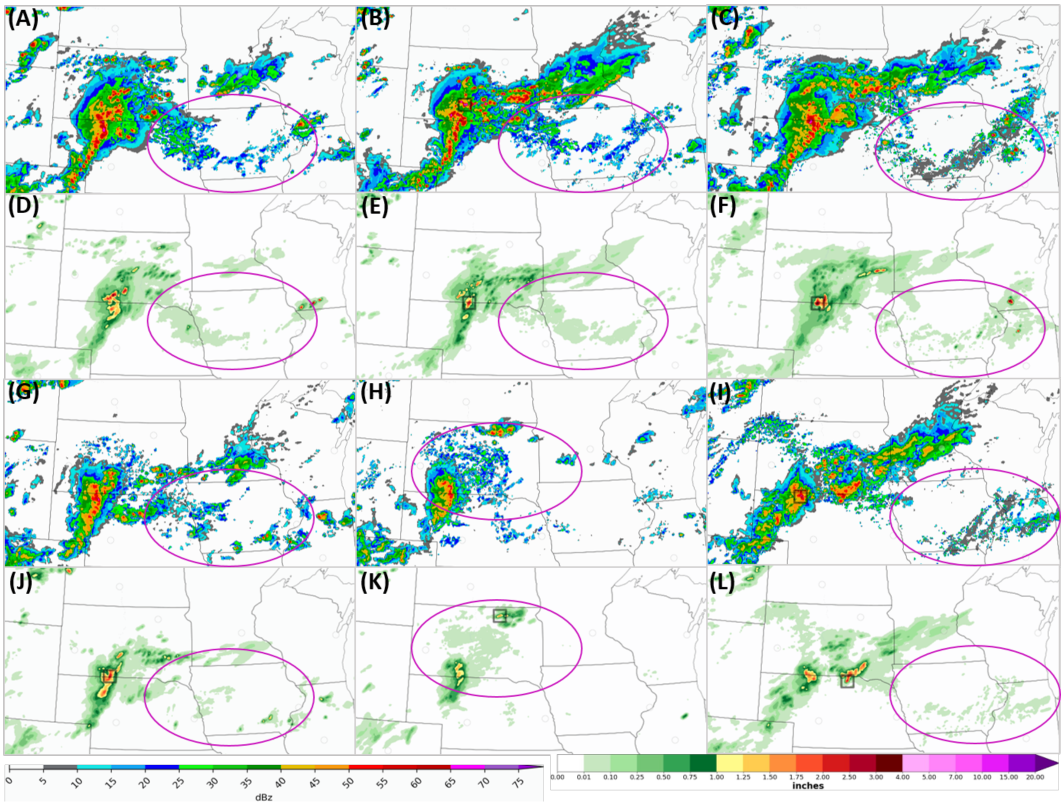


Figure 27: 21JUNE case: Simulated reflectivity/1-h QPF valid at 02 UTC/03 UTC for (A)/(D) RRFSp1, (B)/(E) RRFSm2, (C)/(F) RRFSm3, (G)/(J) RRFSm4, (H)/(K) RRFSm5, and (I)/(L) RRFSm6 on 21 June 2024 for the 12z 20 June 2024 initialization.

5.1.4 Subjective: End of the Week Survey

At the end of each week, participants were asked general questions about the performance of the RRFSp1. Figures 28 and 29 show the results of the some of the questions. Not all the participants completed the end of the week survey. Only 1 participant throughout the length of the experiment felt that the RRFSp1 was generally better than the HRRR's performance and 4 felt that it performed better than the NAMnest. The RRFSp1 was perceived to perform worse than the HRRR by more participants (11) than when the RRFSp1 was compared to the NAMnest (7). In terms of the RRFSp1's QPF coverage and the magnitude of the maximum, participants' results were skewed towards the RRFSp1's coverage and maximum QPF being too low. 42.8% felt the QPF coverage was not expansive enough compared to MRMS. Meanwhile, 52% of the respondents felt the maxima were low to too low. Roughly a third of the participants felt the coverage and maximum QPF values matched well with observations. No one indicated that they thought the that the QPF coverage was too expansive or that the maxima were too high, though 20.8% felt it was slightly too expansive and 12% slightly too high.

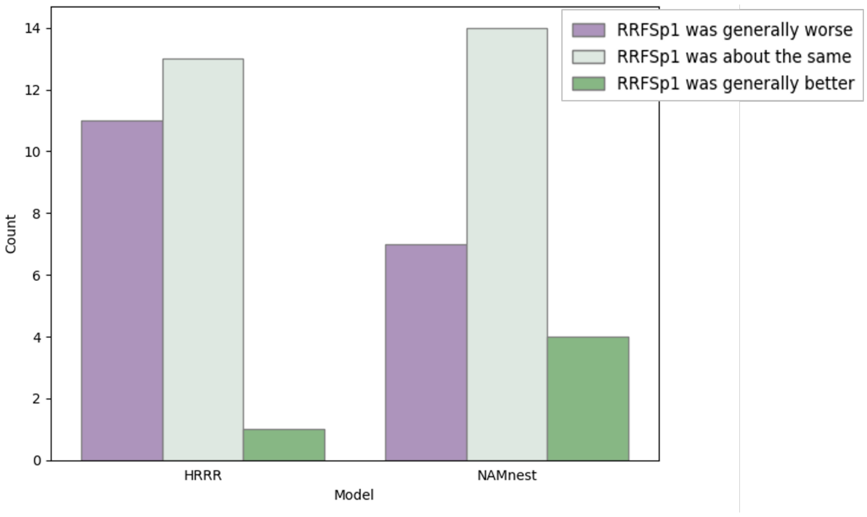


Figure 28: Results from the end of week survey for the question: "During your week, how do you feel the RRFSp1 performed compared to the following models?" with the HRRR and NAMnest as the choices.

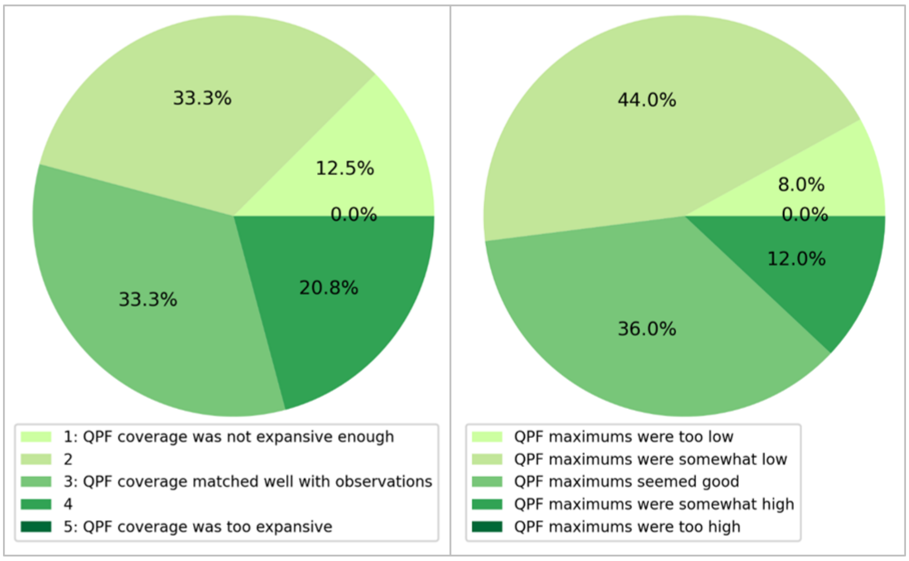


Figure 29: Results from the end of week survey for the questions: [LEFT] "During your week, what was your general feelings on the RRFSp1's QPF coverage?" and [RIGHT] "During your week, what was your general feelings on the RRFSp1's maximum QPF magnitudes?"

Participants were asked to elaborate on the questions discussed above. Participants noted the RRFSp1 seemed to have some bad misses, struggled with nocturnal convection, and even if it had a good forecast for the 12z cycle, it struggled with the forecast in the lead up to the event. They did note that general evolution and coverage seemed good. Some comments such as these were common:

- "We all know HRRR has an early dry bias and can have great longer-length forecasts and poor shorter-length forecasts, meaning there is plenty of room for improvement from using a different model, but sadly I found many examples in which the RRFSp1 had really bad misses (in the contingency table sense). It rarely had bad false alarms. In general, I was a bit dismayed at how poorly RRFSp1 seemed to perform."
- "The RRFS seemed to have repeatable issues, particularly underestimating nocturnal convection"
- "Often times the RRFSp1 was underdone in the magnitude of the precipitation, but the general evolution of the precipitation and the coverage was good."
- "The RRFSp1 did have a couple fairly good forecasts (notably in the 12Z cycles) but seemingly struggled in the lead up to the event."
- "RRFSp1 seemed to struggle to make realistic MCSs compared to the HRRR and NAMnest."
- "I did notice the RRFSp1 rippling a couple of times and that it tended to be on the quicker end of convective initiation."
- "The differences were quite subtle in many cases, but I think the HRRR still has a slight edge on the RRFSp1, and I struggled to notice much difference between it (RRFSp1) and the NAMNest many times. Occasionally the NAMNest was the best of the bunch, especially when there was heavy precip...although it often got the magnitude and not the location. The RRFSp1's bullseyes were often a little too small and it had unrealistically large areas of light forecast precip."

- “Coverage in trace amounts were too high, but coverage in higher amounts was slightly too low with the MCS-type events (with pulse storms the coverage in higher amounts seemed more realistic).”
- “I think it did tend to struggle more with nocturnal convection than the HRRR and NAMnest.”

5.1.5 Subjective: Notes on NSSL-MPAS

As mentioned, due to computing resource issues, the NSSL-MPAS was not available for at least one third of the FFaIR experiment. For example, the 00z forecasts valid 12-18 UTC for the HRRR, NAMnest, and RRFSp1 were scored 140 times, while the NSSL-MPAS was scored only 119 times. The 00z NSSL-MPAS for the MRTP$*$ was evaluated 110 times vs the 185 for the other models. Because of this, the performance in comparison to the other guidance of the NSSL-MPAS cannot be fully examined; however, some general feedback will be discussed.

Table 7 and Figs. 11 and 12 include the subjective results when the NSSL-MPAS was available (the model initializations were only at 00z and 12z). The 12z NSSL-MPAS actually had the highest total average score for the 12-18 UTC valid period in the CONUS evaluation, with an average of 3.465 (the double green $*$ in Table 7); the next highest was 3.079 from the HRRR. However, that was the only instance of the NSSL-MPAS having a higher total mean score. For the 06-12 UTC valid period, both cycles of the NSSL-MAPS had a lower total mean score than the lowest mean from the other models; this was also true for the MRTP$*$.

Participants noted that the NSSL-MPAS had a lot of erroneous light precipitation like the RRFSp1 but its characteristic was much different than the RRFSp1. Rather than a continuous area of light precipitation ($\leq$0.1”) the QPF field was covered by light precipitation speckles, with very few instances of 6 h precipitation greater than one inch. Figure 30 shows the NSSL-MPAS forecasts valid at the same time as the two forecasts shown in Fig. 13. Note that compared to the RRFSp1, across the southeast where the RRFSp1 has widespread coverage of $\leq$0.1” with embedded grid points that jump to 3+ inches of QPF, the NSSL-MPAS has widespread grid point speckles that vary between ≤0.1” and ≤1”. There are few instances of >1”.

This odd looking QPF characteristic seemed to be less prominent outside of the diurnal maximum time period (18-00 UTC), unlike the RRFSp1’s light precipitation feature that was seen across all time periods. For example, Fig. 31 shows the RRFSp1 and NSSL-MPAS QPF valid 00-06 UTC 07 and 22 July 2024; the RRFSp1’s QPF still has widespread light precipitation as seen in the prior six hour period, but the NSSL-MAPS QPF no longer has a speckled look to it and has a bit more structure to the pattern. Another example of this can be seen in the 12JULY case, at 00z on July 12 (Fig. 32) the speckled structure is present across the CONUS, including in the convection associated with the front along the coast. However, by 11 UTC, Fig. 33 (matching (Fig. 14), the NSSL-MPAS has a more organized structure to the QPF while the RRFSp1 still has the light precipitation as was discussed previously.

Reflectivity analysis (not shown) suggests that during the convective maximum, the NSSL-MPAS creates small popcorn cells across the CONUS, with little to no organization or evolution to the cells. However, differing from past discussions of popcorn cells in FFaIR for the RRFSp1, these are small and have low simulated reflectivity values (≤40 dBz). Additionally, differing from this year’s RRFSp1, the NSSL-MPAS does not appear to struggle to develop organized convection (e.g. MCSs) after the diurnal maximum.

5.1.6 Subjective: Main Takeaways

The HRRR subjectively outperformed the NAMnest and RRFSp1 throughout FFaIR, with the exception of its 06z and 12z cycles valid for the 18-00 UTC time period. This is likely due to the known struggles the HRRR has with delayed CI. For the 00-06z valid period the RRFSp1 had the lowest total mean regardless of the model cycle. Furthermore, the 00z cycle of the RRFSp1, excluding the period valid 12-18 UTC, always had the lowest total mean for the experiment; this includes both the MRTP and MRTP$*$ analyses. Additionally, the RRFSp1, especially when evaluating the MRTP$*$ domains, was noted by the participants to struggle with the

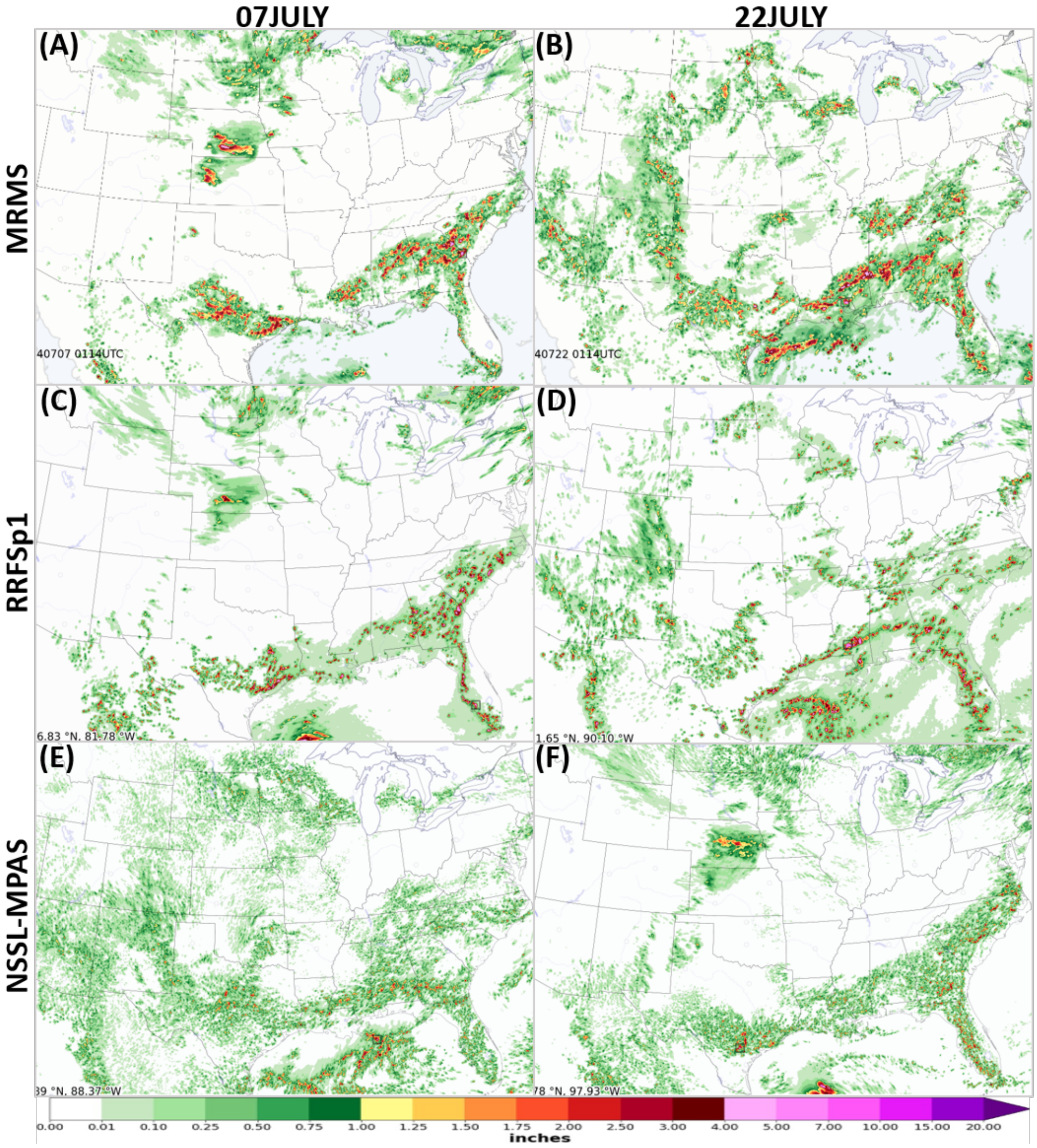


Figure 30: LEFT: 6-h MRMS QPE and model QPF valid 00 UTC 07 July 2024. RIGHT: 6-h MRMS QPE and model QPF valid 00 UTC 22 July 2024. (A) and (B) are MRMS. Models are initialized at 12z (left) 06 July and (right) 21 July. (C)-(D): RRFSP1 and (E)-(F): NSSL-MPAS. (A)-(D) are the same as Fig. 13A, B, G, and H.

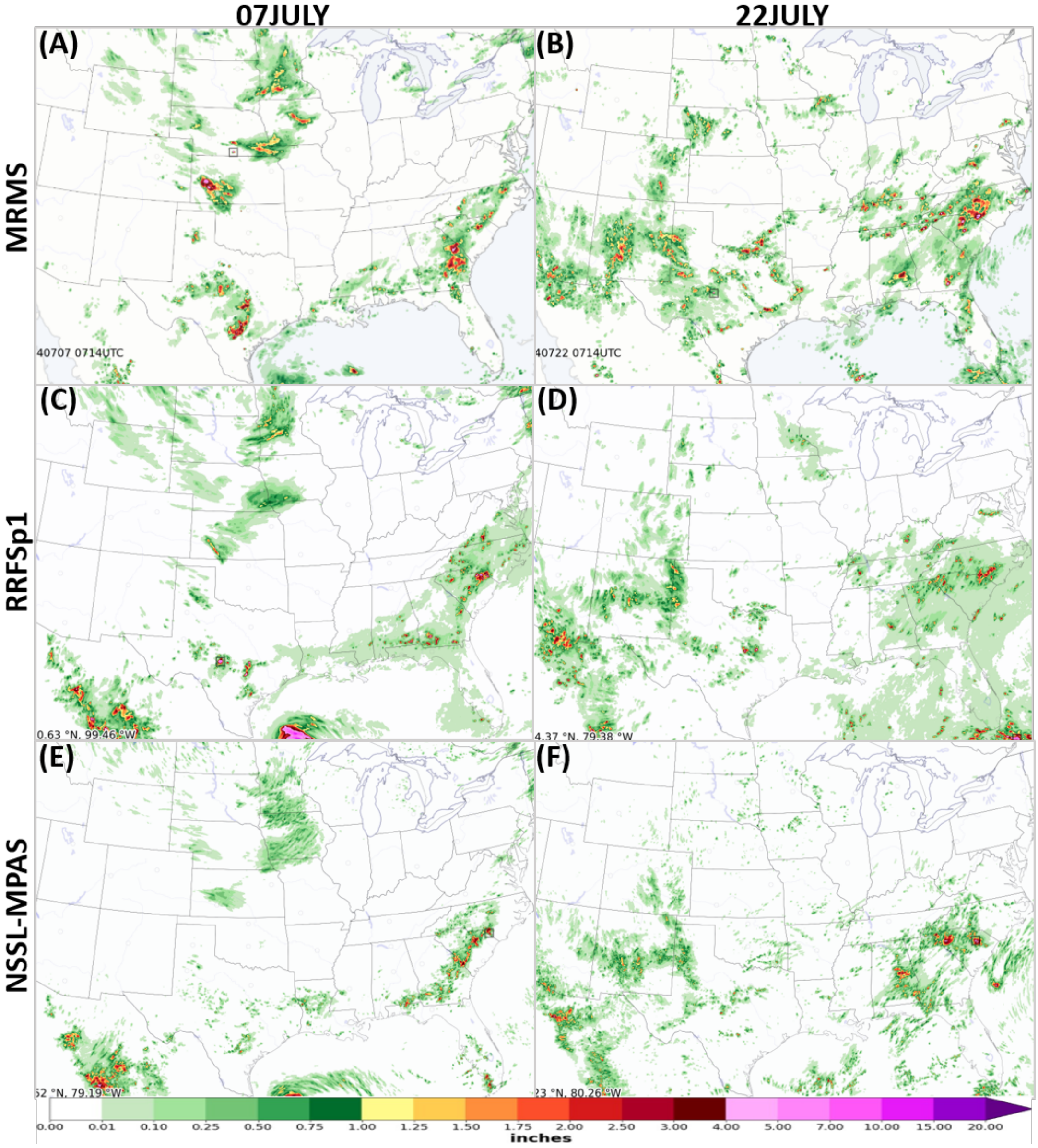


Figure 31: Like Fig. 30 but valid 06 UTC 07 July (left) and 22 July (right) 2024.

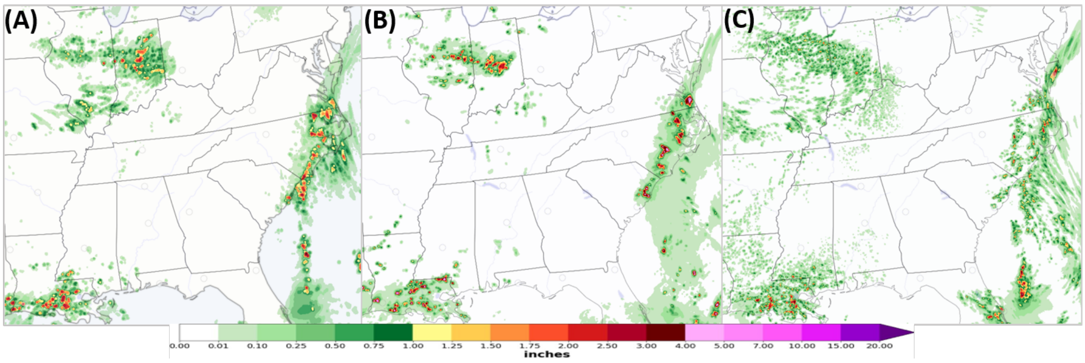


Figure 32: 6-h (A) MRMS QPE and QPF from the 12z initialization on July 11 for the (B) RRFSp1 and (C) NSSL-MPAS, valid 00 UTC 12 July 2024.

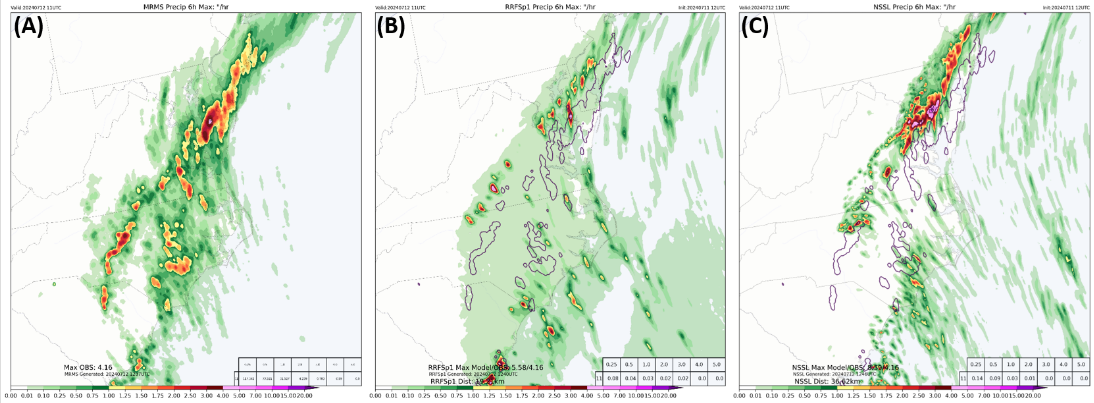


Figure 33: Verification graphics for the MRTP: 6-h (A) MRMS QPE and QPF from the 12z initialization on July 11 for the (B) RRFSp1 and (C) NSSL-MPAS, valid 11 UTC 12 July 2024. The MRMS 1" object is outlines in purple on (B)-(D). (A) and (B) are the same as Fig. 14A and D.

forecast until its 12z cycle, when it seemed to abruptly "get it right" (see 22JUNE discussion).

The RRFSp1 seems to either significantly over- or under-predict coverage of heavy rainfall. To some extent, this appears to be partially driven by the time of the occurrence of the organized convection. If the convection is not well established in the model prior to 21 UTC/the diurnal maximum (see discussion of 21JUNE and 25JUNE), under forecasting occurs. Often after the diurnal maximum, CI suppression was seen in the simulated reflectivity, preventing the development of organized convection such as MCSs. However, in synoptically-driven events or within tropical systems, the CI suppression was not present (the model was able to forecast organized convection), and the model tended to over forecast QPF magnitudes, like has been noted in past FFaIRs. The high bias as it relates to popcorn storms still exists but how it presents itself has changed from past FFaIRs. Rather than multiple popcorn storms anywhere there is instability, that appear to "rain out" all the available moisture (ex. Fig. 28 in 2023 FFaIR Final Report (Trojniak et al., 2023)), this year the bias depicted as isolated high QPF totals embedded in a large coverage of trace amounts of precipitation (ex. Fig. 13 and Fig. 14). This was a feature seen in both large scale precipitation events and during general thunderstorm days.

Overall participants felt the RRFSp1 was similar to the NAMnest and worse than HRRR but for different reasons than last year. This year they didn't discuss the high magnitudes in their comparison to the NAMnest (aka having large magnitudes like the NAMnest), but noted that the RRFSp1 can miss events like the NAMnest. The missed events, however, are different between the two models; the RRFSp1 struggles with big events like MCSs, while the NAMnest has trouble with convective evolution but rarely misses MCS develop, even if the location is off. This aspect of the RRFSp1 is particularly concerning since the struggle to create nocturnal convection and MCSs was also seen in the REFS membership. This resulted in the ensemble forecast also failing to highlight the risk (ex. 25JUNE); this aspect will be discussed in Section 5.2.

5.1.7 Objective: CONUS

Model performance across the CONUS using MET-MODE (see Section 3) was done for 24-h and 6-h QPF, from 01 June to 04 August 2024. Figure 34 shows the 24-h performance diagrams (hereafter PD) for the 00z and 12z cycles for the HREF and REFS membership at one-half, one, and two inch thresholds. Figure 35 is the PDs for the four 6-h synoptic time periods for the HRRR, NAMnest, RRFSp1 and RRFSm2-6, for the same thresholds as the 24-h PDs.

Focusing on the 24-h PDs, the main takeaway is that there is more spread in the performance of the HREF membership than the REFS membership. It isn't until the two inch threshold that the REFS membership shows some spread in performance, and the greatest is seen for the 12z cycle. At one-half and one inch, the

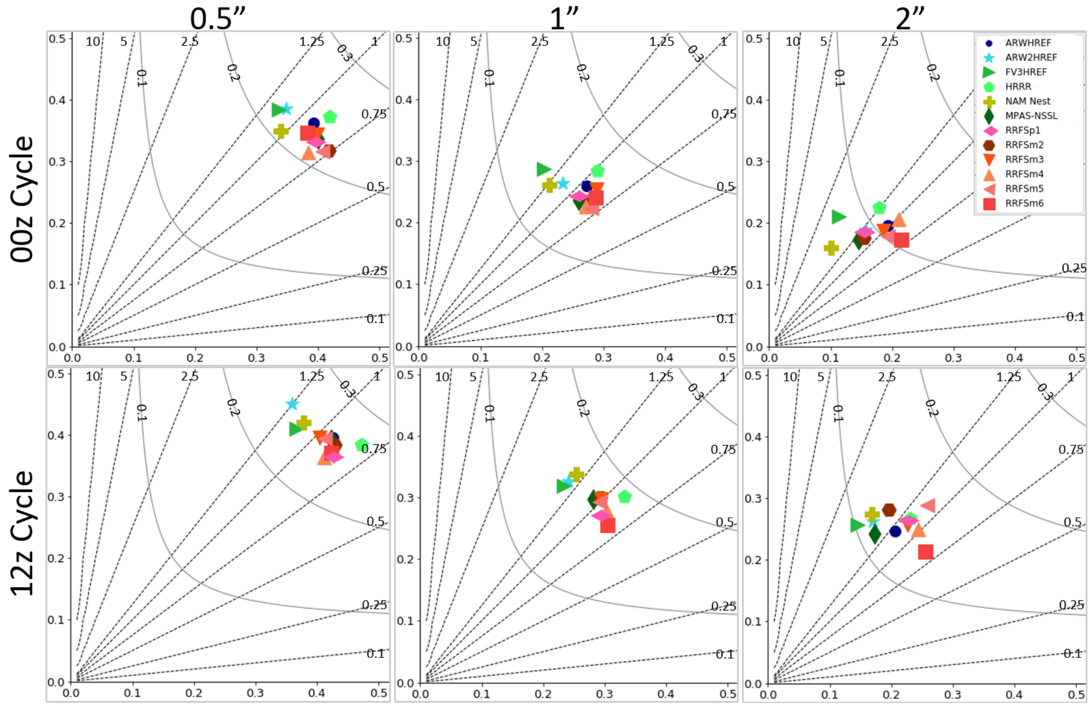


Figure 34: Performance diagrams for CONUS 24-h QPF from 01 June to 04 August 2024 for the HREF and REFS membership. Top is the 00z initialization and bottom is the 12z, for the thresholds of 0.5 (left), 1 (middle), and 2 (right) inches.

REFS membership (including the HRRR) generally has a dry bias; this is true for the 18z and 06z cycles (not shown) as well. Additionally, for those two thresholds, the HREF membership tends to have a higher probability of detection (POD) while the REFS members have a higher success ratio (SR). For all but the 2” threshold for the 12z cycle, the HRRR has the best or is tied for the best critical success index (CSI). This includes the 12z cycle, despite the HRRR having the largest dry bias. The RRFSp1 never has the highest CSI of the REFS membership. For the 12z cycles, the NAMnest ties or has a better CSI than all of the REFS members. Finally, for the three inch threshold (not shown), for all cycles the NAMnest is the worst performer based on the CSI. The HRRR was always outperformed by one or more of the REFS members but which member(s) that outperform the HRRR varied depending on the cycle. That said, the RRFSm5 always had or was tied for the highest CSI.

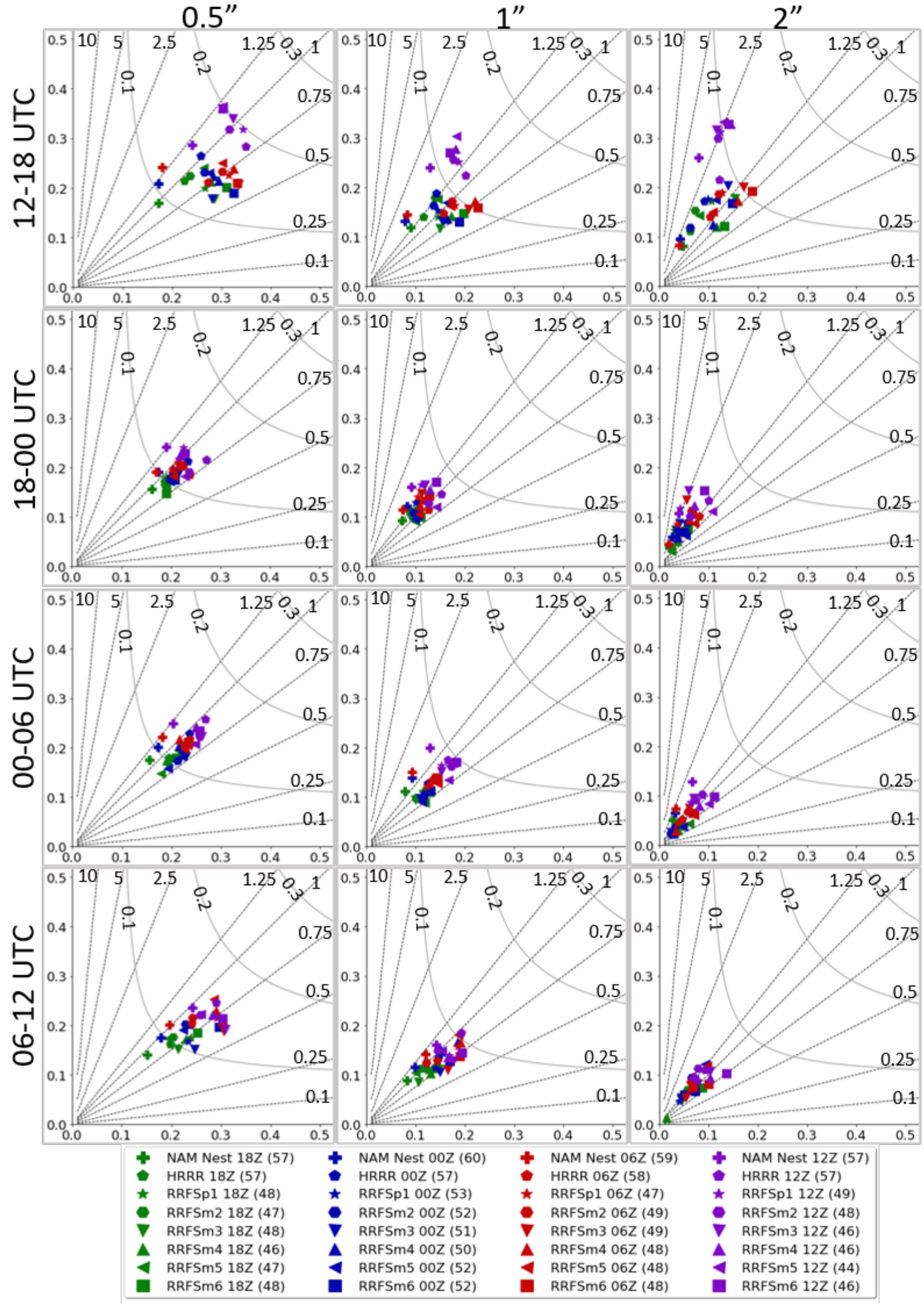


Figure 35: Performance diagrams for the CONUS 6-h QPF from 01 June to 04 August 2024 for the HRRR (pentagon), NAMnest (cross), RRFSp1 (star) and RRFSm2-6 for their forecasts initialized (ordered from longest to shortest lead time) at 18z (green), 00z (blue), 06z (red) and 12z (purple). From top to bottom, the columns are valid for the following 6-h periods: 12-18 UTC, 18-00 UTC, 00-06 UTC and 06-12 UTC. The thresholds shown are 0.5 (left), 1 (middle), and 2 (right) inches.

For the 6-h PDs, Fig. 35, the cycles are color coded: green for 18z, blue for 00z, red for 06z and purple for 12z. In general, the quality of the forecast is stratified by initialization time. The operational models (HRRR and NAMnest) generally tend to follow the same bias tendencies across cycle initializations and 6-h time periods; NAMnest mostly has a high bias; HRRR generally shows a low bias. The REFS members do not show this pattern; across cycles there can be large jumps in bias. For example, for the 12-18 UTC valid period for 1" the membership goes from a dry bias for the 06z cycle to a wet bias (approaching 2) for the 12z cycle. In the most extreme case, the RRFSm4 and m6 00z, 06z, and 18z cycles had a bias around 0.75 then jumped to a bias around 2 for their 12z cycle. Another general characteristic of all models was the increase in bias as threshold increased, however the operational models tended to have a less dramatic change in bias across thresholds than the REFS members. The one exception was for the 00-06 UTC time period, which saw the NAMnest have the greatest increase while the HRRR and REFS members hovered between a low bias and a bias of 1. However, for the one and two inch thresholds, the NAMnest had a noticeably greater POD for the 12z cycle than the other models. The RRFSp1 rarely was the best performer of the REFS members. RRFSm6 was the most likely of the REFS members to have the highest CSI.

Additional performance results based on the 6-h time periods are list below:

- **12-18 UTC:** All models' 12z cycles performed the best in terms of CSI. For all thresholds, there was a large jump in bias/POD seen in the REFS members at the 12z cycle. For this valid period, for nearly every cycle/threshold, the REFS members had a better CSI than either of the operational models.
- **18-00 UTC:** For the half inch threshold, the REFS membership performance generally falls between the HRRR's and NAMnest's. As thresholds increase, the REFS members' CSI tends to outperform NAMNest and be on par or better than HRRR (depends on the member).
- **00-06 UTC:** For one-half inch, all models but the NAMnest have a dry bias. This time period is the only one that doesn't have a strong shift from dry to wet bias in the REFS membership across the thresholds. In fact, some

members actually increase their dry bias as thresholds increase for all cycles but the 12z. For example, the 00z RRFSm5 (blue left pointing triangle) goes from a bias around 0.8 for one-half inch to roughly 0.65 for the 2” threshold. For the 1” and 2” thresholds, the 12z NAMnest had a noticeably higher POD than all other models. As threshold increased the NAMnest, RRFSm2, and RRFSm6’s CSI became better or equal to the HRRR.

- **06-12 UTC:** For one-half and one inch, all models but the NAMnest have a dry bias, with the RRFSm3 usually having the greatest dry bias for the two thresholds. For the 2” threshold, the RRFSm6 has a consistent bias of 0.75 across its cycles and the highest CSI of all models for its 12z cycle. Unlike the other time periods, the highest CSI across cycles is not always 12z.

Analysis of hourly precipitation totals can be seen in Figures 36-38 across the Testbed Season[7] for both the 00z and 12z cycles of the HRRR, NAMnest, REFS members, and the NSSL-MPAS. In terms of the number of occurrences of hourly accumulation (Fig. 36), the characteristics of REFS membership performance was separated by type of microphysics scheme used. The RRFSp1, RRFSm2, and RRFSm3, which use a version of the Thompson scheme, begin to diverge from the members that use the NSSL scheme (RRFSm4-6) around 0.75” (green vertical line in Fig. 36), resulting in a shallower slope to their survival curve than RRFSm4-6. RRFSp1 and RRFSm2/3 began to over forecast instances of hourly totals starting at ∼1.75”; indicated by the blue vertical line in Fig. 36. These members also had a shallower slope at 00z than the NAMnest, suggesting a more significant high precipitation bias at higher thresholds than the NAMnest. For the 12z cycle, these members’ slopes were more similar to the NAMnest.

Opposing this, RRFSm4-6 overall under forecasted nearly all instances of hourly accumulations. Two exceptions were the 00z RRFSm5, which followed the MRMS curve from 4-5.5” and the 12z RRFSm4-6 from roughly 3.5” to 4.75”. All of the REFS members, aside from the RRFSm5, and the NSSL-MPAS, saw a decrease in their max hourly accumulation from the 00z cycle to 12z cycle, though the shift for RRFSm4 and RRFSm6 was much smaller than for the other members.

[7] Discussed in Section 3 and defined as valid from 01 May to 14 August 2025.

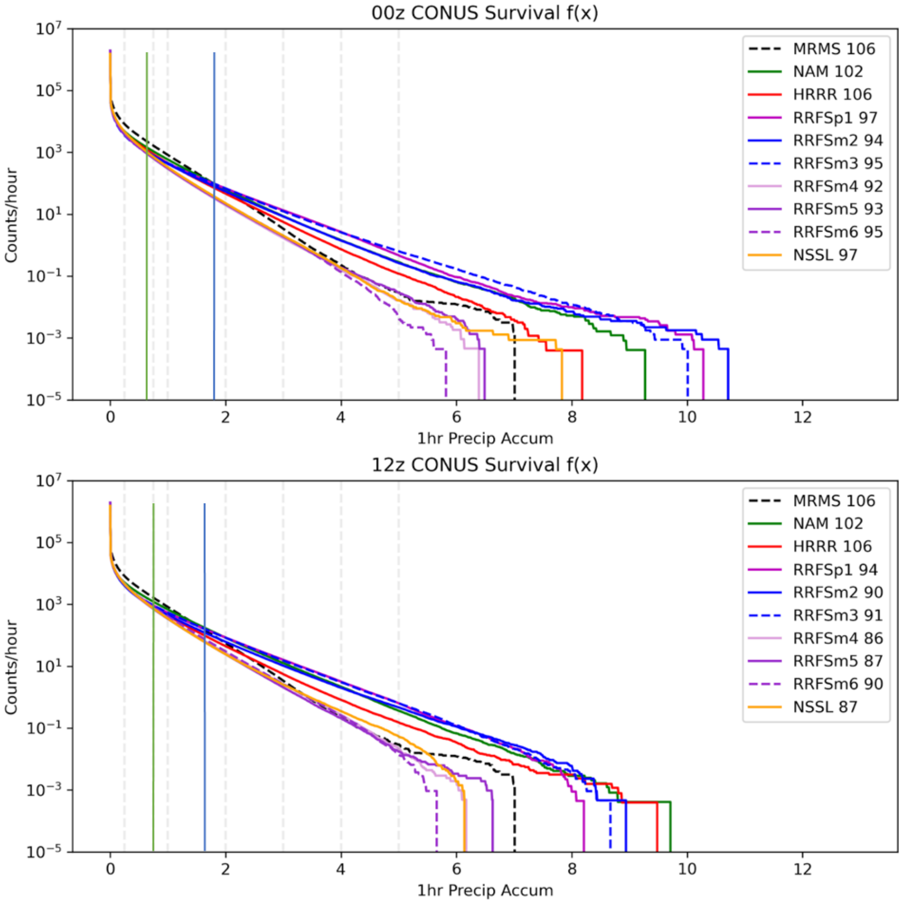


Figure 36: Hourly QPE/QPF survival functions for the 00z (top) and 12z (bottom) model cycles valid from 01 May to 12 August 2024. The QPE is in dashed black. For the REFS members, the solid colored lines represent members with the G-F CU scheme and colored dashed ones represent members with saSAS CU scheme. The green vertical lines indicates roughly where the REFS members that under forecast diverge from those that over forecast. The vertical blue line on each chart indicates when the majority of members cross the MRMS curve.

For 00z, the max hourly accumulation for the RRFSp1 (10.25"), RRFSm2 (10.75"), and RRFSm3 (10") was greater than both the HRRR (8.25") and NAMnest (9.25"); for 12z these members decreased to between 8 and 9 inches, while the operational models increased to 9.5 and 9.75 inches respectively.

Figure 37 shows the average hourly accumulation across the diurnal cycle for Day 1 (12-12 UTC), with zeros included (right side of the figure) and not included (left). Excluding the zeros can be thought of model forecast precipitation intensity, while including the zeros representation the coverage of rainfall. When focusing on

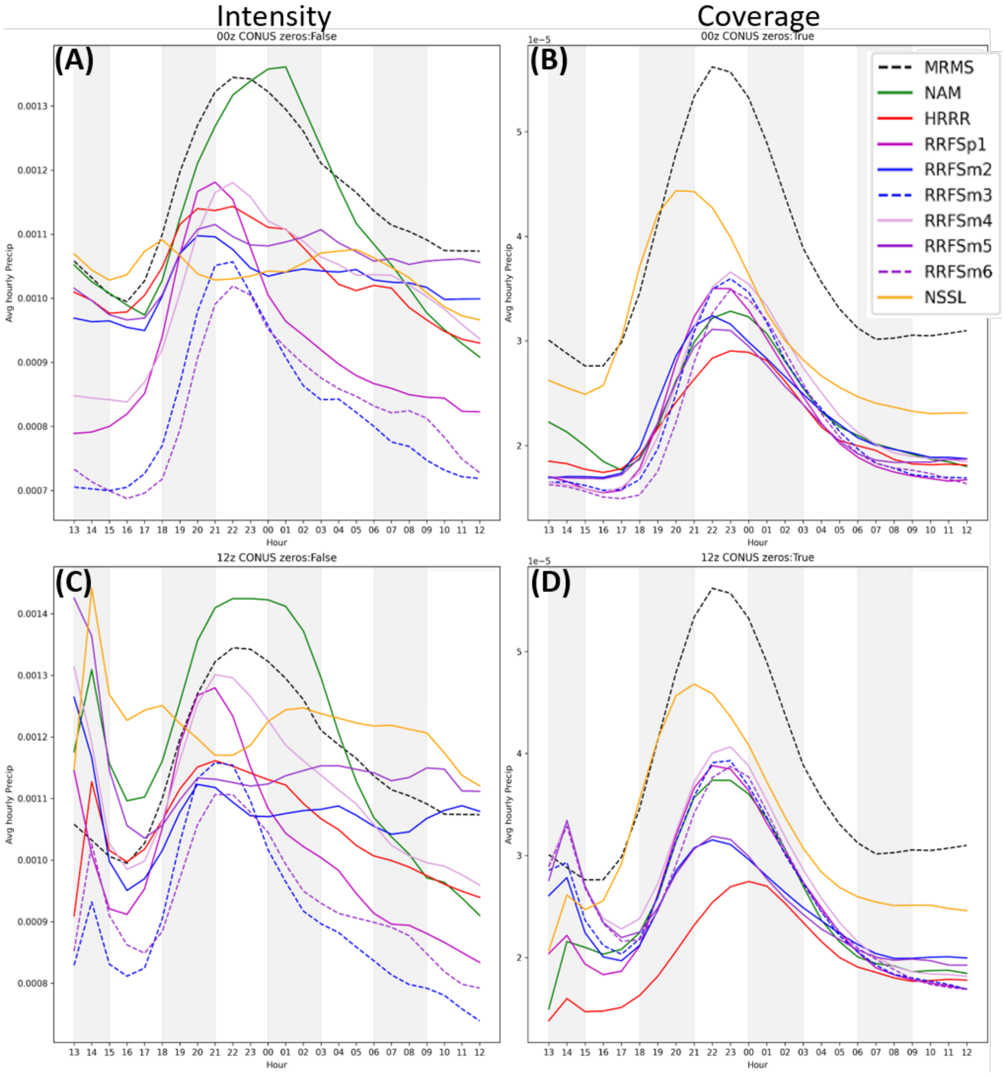


Figure 37: Diurnal analysis of the average hourly QPE/QPF for the 00z (top) and 12z (bottom) model cycles valid from 01 May to 12 August 2024. [LEFT] zeros are NOT included in average (can be thought of as Intensity) and [RIGHT] zeros are included in average (an be thought of as Coverage). See caption in Fig. 36 for information about dashed vs solid colored lines for the REFS members.

intensity (Fig. 37A and C), the peak in the diurnal cycle generally occurs earlier than observed for the REFS membership. The HRRR's peak is relatively flat while the NAMnest shifts from a late peak for the 00z to a broader peak centered across the observed diurnal max. All models except the HRRR see an overall increase in forecasted intensity across the diurnal cycle from the 00z to 12z cycles. The

RRFSp1 (G-F CU Scheme) and the two members with the saSAS CU Scheme (m3 and m6) have curves of similar shapes, though the magnitude across the diurnal cycle for the RRFSm3/m6 was always less than RRFSp1, as well as less than all other models evaluated. The RRFSp1 and RRFSm3/m6 also have a sharp drop in the average hourly precipitation after their forecasted diurnal peaks (∼21 UTC) that is not observed in the other members or the operational model. The 12z RRFSp1 follows the observed line from roughly 19-20 UTC before peaking at it's diurnal max at 21 UTC, an hour before the observed maximum. RRFSm4 is the least like the other REFS member. At the start of the diurnal cycle it is similar to the RRFSp1 but at the peak in the observed diurnal max, it is most similar to the observed timing of the peak. After the diurnal maximum it has a slope similar to the HRRR's, meaning its drop after the maximum is less steep than RRFSp1/m3/m6. Finally, RRFSm2 and m5, which are the only members that use the TKE-EDMF PBL Scheme, have an average hourly total that remains relatively constant after the diurnal maximum. In contrast to all the other models, the NSSL-MPAS has a minimum in the intensity during the observed maximum and characteristics that differ from the rest of the models evaluated.

Figures 37B and D show the average hourly rainfall across the CONUS when zeros are included, which, as stated, can be thought as highlighting the coverage of rainfall. Excluding the NSSL-MPAS, the shape of the REFS members and operational models are similar to one another, with the HRRR perhaps being the least similar of the group. Going from 00z to the 12z cycle, there is an increase in the variability of coverage among the models. Every model sees an increase in the hourly average to the diurnal max aside from the HRRR and RRFSm2, which see a slight decrease, and RRFSm5, which remains basically the same between the two cycles. The RRFSm2 and m5 are the least like the other REFS members in terms of magnitude and timing of the diurnal maximum, being lower and slightly early. The greatest difference between them and the rest of the membership is for the 12z cycle from 20-04 UTC. Despite the lower hourly average after the diurnal peak (after roughly 06 UTC), they have greater coverage than the RRFSp1 and RRFSm3/4/6. For the 12z cycle the HRRR has the lowest coverage across all the entire diurnal cycle.

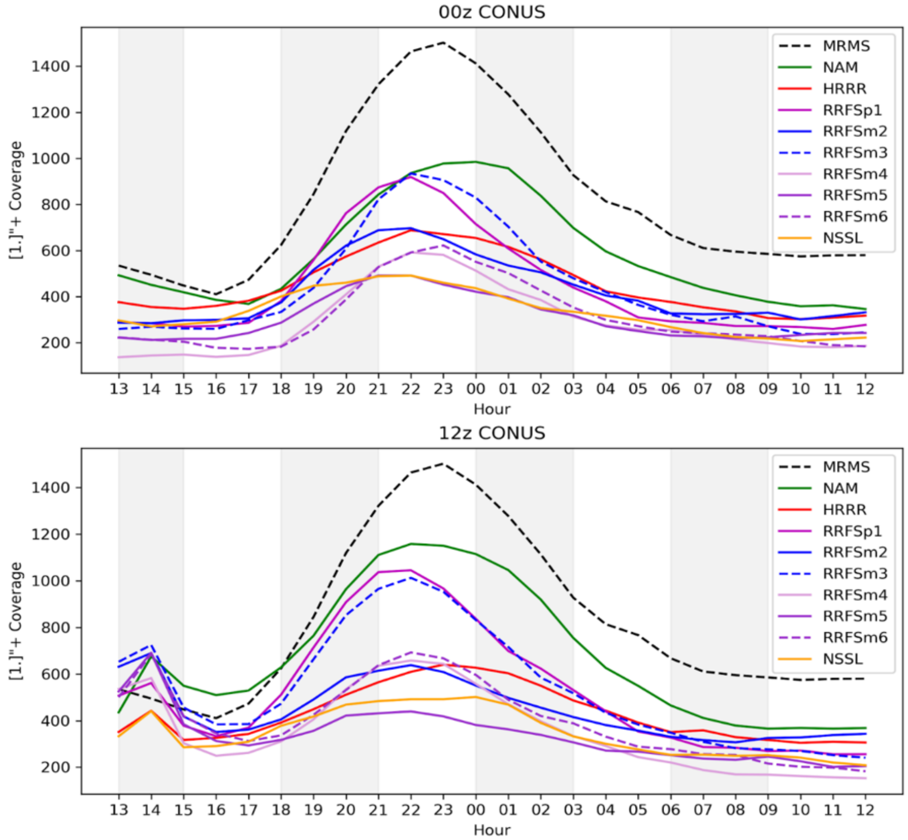


Figure 38: Diurnal analysis of the coverage of the hourly QPE/QPF of 1" the 00z (top) and 12z (bottom) model cycles valid from 01 May to 12 August 2024. See caption in Fig. 36 for information about dashed vs solid colored lines for the REFS members.

When comparing the left side of Fig. 37 to the right side, possible tendencies can be inferred. The RRFSm3 and m6 go from having the lowest hourly average at the diurnal maximum to the highest when zeros are included in the average. This suggests that these two members are missing fewer events in the lead up to and at the diurnal max. Differing from this, the RRFSp1 and RRFSm4 go from having the highest average of the REFS members when zeros aren't included to being on par with RRFSm3 and m6, suggesting they are missing events. However, when they do forecast rainfall, it is more intense (higher totals) than the latter two; this follows a similar pattern to the NAMnest. After the diurnal max, the RRFSp1, RRFSm3 and m6 are likely both missing events and under forecasting the intensity of the events. The RRFSm4, which has a different intensity curve than the other 3 after the max but looks similar to them in terms of coverage, is likely missing approximately the same number of events after the diurnal max, but when it forecasts rainfall, it is much more intense than the other members. Finally, the RRFSm2 and m5 are most like the HRRR, and appear to forecast both fewer and weaker events leading up to the diurnal max.

The coverage of 1"/h across the diurnal cycle can be seen in Fig. 38. For both the 00z and 12z, the NAMnest most closely matches the diurnal cycle compared to the other models, with the exception of the first 6-h of its 12z cycle. The RRFSp1 and RRFSm3, which only differ in CU Scheme, perform similarly, with the exception of the timing of their respective diurnal maxima; the RRFSp1 having a slightly earlier and shorter max compared to the RRFSm3, though both maxima are occurring before the observed max. These two models are similar to the NAMnest in the lead up to the diurnal max but have a rapid decrease in coverage of $\geq$1" after their respective maxes that is not seen in the NAMnest or in observations. RRFSm4 and m6 perform relatively similar to one another, with the RRFSm4 generally having less coverage of 1" than the RRFSm6 but similar slopes across the Day 1 time period. The RRFSm2 and m5 (the only ones without MYNN PBL Scheme) differ the most from the other members, with a longer maximum and smaller change in the coverage of 1" across the time period than the other members. That said, the coverage for 1" from RRFSm2 is always more than RRFSm5.

Combining these analyses of hourly QPF across the CONUS some overall findings are:

- Under each method of analysis, the REFS members that perform similarly to one another aren't always grouped by the CU Scheme. When looking at the number of times an hourly QPF value was forecasted, the REFS members were grouped by Microphysics Scheme. For the average hourly QPF across the diurnal cycle, there were groupings by both the CU Scheme and the PBL Scheme. Interestingly, the RRFSp1 performed more like the RRFSm3 and m6, which have a different CU Scheme, than it did to the members with the same scheme. Finally, in terms of the coverage of 1" across the diurnal cycle, RRFSp1 and RRFSm3, which differ only in the CU Scheme used, had nearly identical curves.

- REFS members that use the Thompson Microphysics Scheme over forecast the occurrence of hourly totals beginning between 1.5" and 2", depending on the cycle; they under forecast occurrences below 1.5". Members using NSSL Microphysics nearly always under forecast all hourly totals.
- RRFSm2 and m5, which are the only members that use the TKE-EDMF PBL Scheme, have different characteristics in the shape of their curve after the diurnal max when not including zeros (intensity) in the hourly average compared to the other members; i.e. they don't show a decrease in the average after the max. These two models also do not see an increase in their hourly average QPF when going from the 00z to 12z cycle at the diurnal max when zeros are included; this was seen in the other REFS members.
- In terms of intensity across the diurnal cycle, of the REFS members RRFSm4 has a slope after the diurnal max that is most similar to observations, though the decrease in average hourly QPF happens faster than observed QPE. RRFSp1 has the steepest drop off after its forecasted diurnal max in terms of both intensity and coverage, supporting the subjective feedback about how it is quick to kill off convection after the max.
- Although RRFSp1 over forecasts instances of hourly totals beginning between 1.5 and 2", when looking at the hourly average QPF and the hourly coverage of 1" across the diurnal cycle, its average is under forecasted. This is most notable after its forecast diurnal maximum. This suggests that the RRFSp1, especially after 21 UTC, is missing events. It also implies that for the events it does forecast, it is over forecasting magnitude. The over forecasting of the magnitude most often occurs leading up to and around 21 UTC, given the sharp drop off in the 1" coverage and hourly average after this time.

5.1.8 Objective: MRTP/MRTP+

Figure 15B shows the location of all of the MRTP+ domains for the objective verification. In total there were 129 6-h domains that made up the MRTP+ dataset, with valid dates ranging from May 2^{nd} to August 14^{th} 2024. The MRTP$*$ domains are included in this dataset. Nearly all the cases, aside from the MRTP domains, were selected after the event occurred, as discussed in Section 3. Of these 129 domains 63% of the them had a valid end time of 04 to 12 UTC, with 85% of them valid with a valid end time of 21 to 12 UTC.

Focusing first on the MRTP domains and the performance of the participants, the PDs for all Day 1[8] MRTP forecasts can be seen in Figs. 39 and 40. As has been the case in previous FFaIRs, the participants tended to have a higher probability of detection (POD) than the models and thus a higher bias. This is likely due to the tendency of the participants to draw large contours for their MRTP. Generally this is due to two factors: the capabilities of the drawing tool used and participants' implying a degree of probabilistic thinking to their contour despite the forecast definition being deterministic. This year, in an attempt to better understand the implicit probabilistic aspect of their methodology, the team collected feedback on when and if they applied a probability to each contour they did or did not draw. The team plans to use this information to help with future FFaIR planning but it will not be addressed in this report.

In general, both the participants and models/ensembles, not surprisingly, perform better on days when the one inch area had large areal coverage; these days are identified by red or purple text for the 1" areal coverage on each PD in Fig. 39 and 40. Participants often had the same or better CSI than the models and ensembles they had access to during the forecast process[9]. In terms of those participants assigned the HREF versus REFS, most days there was not a clustering of results based on the model. The HREF and REFS groups each had a visually noticeable difference in their daily aggregated CSI 16 of the 24 times, with each one having the highest daily aggregated CSI 8 times; HREF/REFS having the highest outlined in blue/red in Fig. 39 and 40. Even so, the separation between groups is relatively low most of the time and was more pronounced in POD. The HREF group had 15 days with higher POD than REFS but only 7 days with higher

[8] Remember we did one MRTP that was a Day 2 forecast.

[9] In real time the HREF and REFS Group had access to the Day 1 18z (run after the start of Day 1) HRRR. Aside from the HRRR, they 12z cycles were available for all models included in the HREF Group. The REFS Group was a bit more complicated. Often they only had access to the ensemble data for the 06z cycle but did have the 12z RRFSp1 and some REFS members, but not all. In retro, the two groups were allowed to look at 18z Day 1 data.

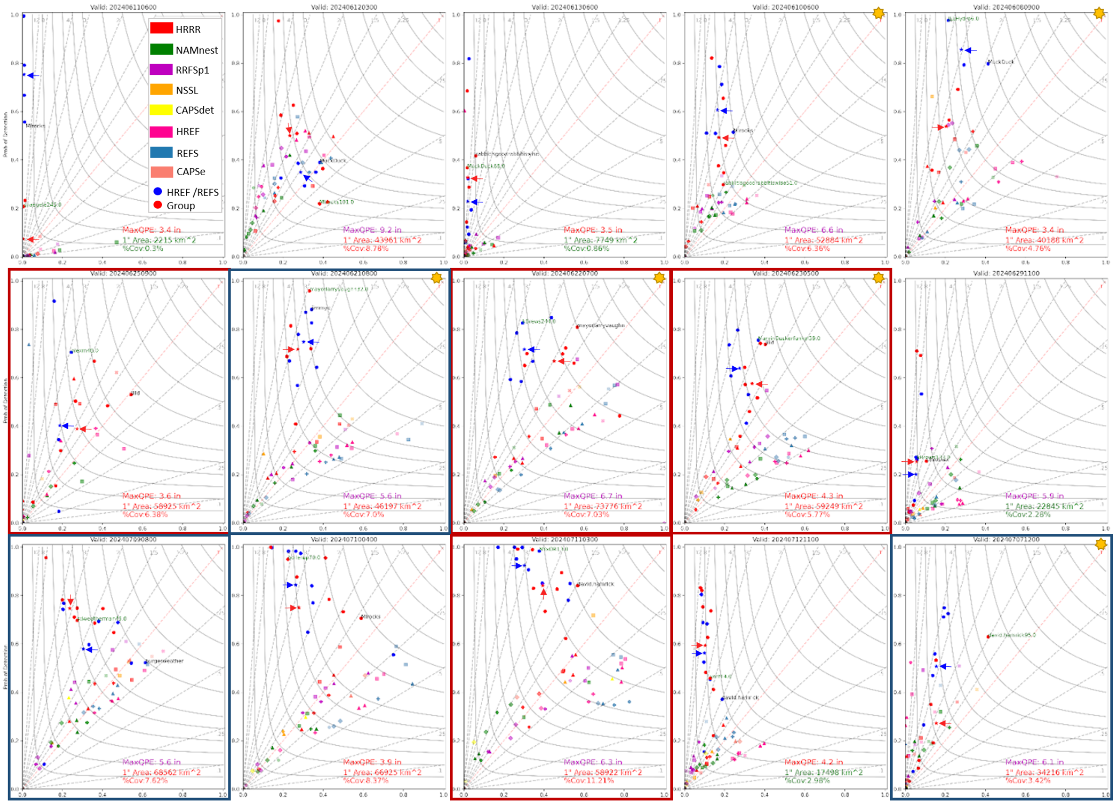


Figure 39: Weeks 1-3 Day 1 MRTP forecast days' performance diagram for a 1". Blue/Red circles are participants in the HREF/REFS Group. Blue/Red star (a blue/red arrow points to the star) is the aggregate average of the HREF/REFS Groups. Other colors represent a model or ensemble (HRRR: red, NAMnest: dark green and RRFSp1: purple), with symbols representing a time grouping. Lighter colors represent cycles that were not available to forecasters when doing the MRTP. PMM is plotted for the ensembles. Each day the maximum rainfall, maximum 1" areal coverage in km^2 and percent coverage of the MRTP domain are shown in the bottom part of their respective diagrams and color coded into 3 groups (low: green, medium: red, large: purple). If the HREF or REFS Group's aggregate average daily CSI was noticeably higher than the other, that day's diagram is outline in blue or red respectively. MRTPs that were done in retro mode have a yellow $*$.

SR. This means that the HREF Group overall caught more events while having a higher bias at 1".

Fig. 41 shows the aggregated performance of the HREF and REFS groups along with HRRR, NAMnest, RRFSp1, and NSSL-MPAS for the 24 MRTP forecasts. The REFS group had a slightly higher CSI (0.149 vs 0.137) while the HREF had a higher POD (0.645 vs 0.604). As was seen in the daily PDs, both groups had significantly higher POD than any of the models. They also had higher CSI than

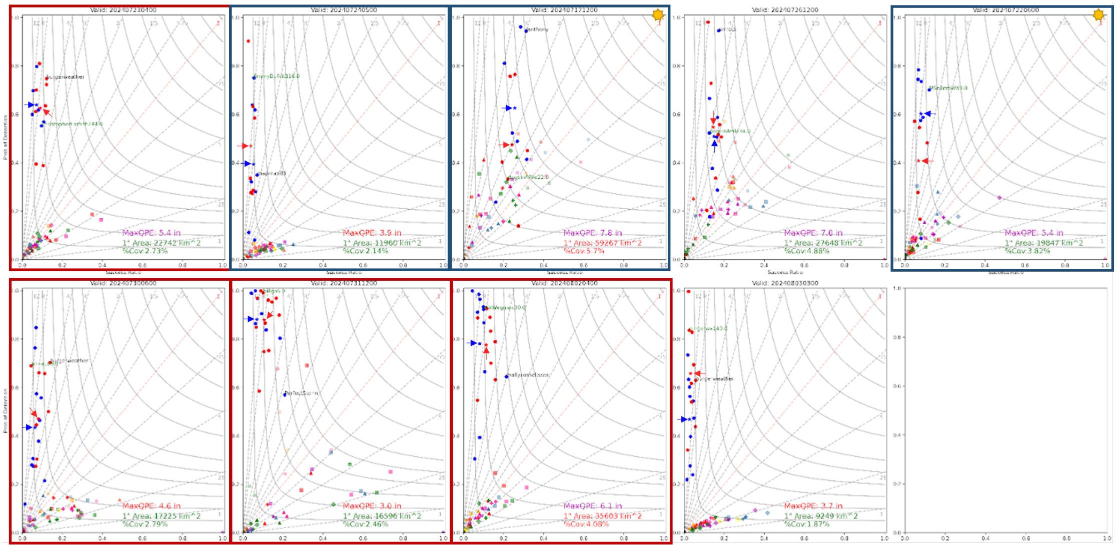


Figure 40: Same as Fig. 39 but for Weeks 4 and 5.

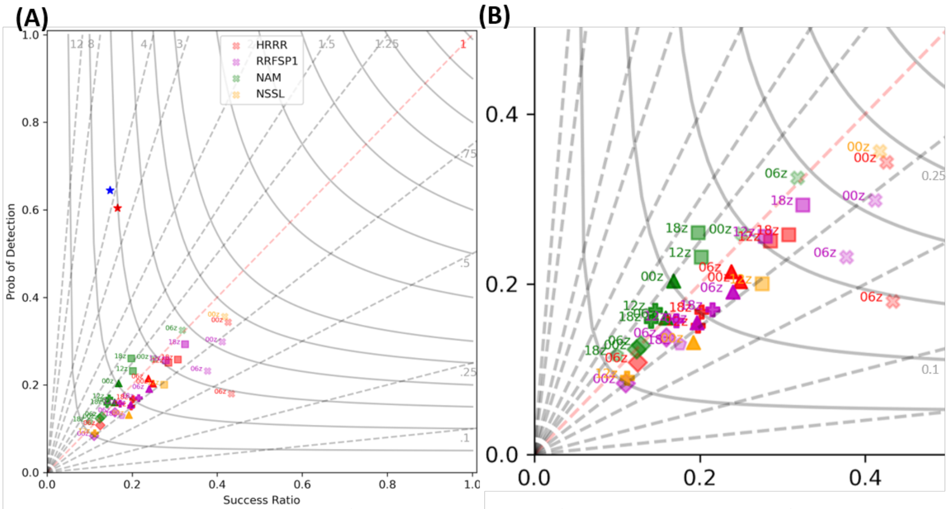


Figure 41: The aggregate MRTP forecast days' performance diagram for a threshold of 1". Blue/Red star is the aggregate average of the HREF/REFS group. HRRR: red, NAMnest: dark green, RRFSp1: purple and NSSL-MPAS: yellow. Each symbol a different time grouping: X - Day 0 06z and 00z cycles, Square - Day 1 18z and 12z cycles, Triangle - Day 1 06z and 00z cycles, Pentagon - Day 2 18z and 12z cycles, Diamond: Day 2 06z and 00z (notional 60-h forecast) cycles, and Pentagon: Day 3 18z and 12z cycles. Darker shading for Day 1 and 2 and lighter for cycles participants likely didn't have access to during the forecasting activity. (A) performance diagram including the HREF/REFS Group and (B) zoomed to only show the model forecasts.

the majority of the model/cycle initializations. The model/cycles that had higher CSI than the two groups were almost always cycles that were not available to participants (X: Day 0 06z and 00z cycles and Square: Day 1 18z), the exception being the Day 1 12z (square), which the participants sometimes had access too. That said, the Day 0[10] analysis has a smaller dataset, since only MRTPs valid later than 06 UTC for the Day 0 00z cycle or valid at 12 UTC for the Day 0 06z can be included.

In terms of the models themselves, the NAMnest consistently had a bias between 1 and 1.25, while the HRRR and RRFSp1 always had a bias below 1, generally centered around 0.75. The NAMnest rarely had a higher SR than the other two models. It nearly always had a higher POD, the exception being the Day 1 12z and 18z cycles. The RRFSp1 and HRRR had a higher CSI than the NAMnest for every cycle but the Day 0 06z cycle. In other words, The RRFSp1 1" is on par with HRRR, yet the NAMnest leads in event prediction at longer lead times (i.e. prior to the Day 1 12z cycle). Although across the MRTP domains/times the NAMnest falls short in the CSI comparison, the importance of the NAMnest's ability to identify the possibility of an event at longer lead times can not but understated, nor the fact that it is the only one of the 3 models to not have a dry bias when focusing on domain specific 6-h heavy rainfall events. Finally, analysis of the CDF of the each models' normalized 1" CSI at various initializations (not shown) for the MRTPs showed that for CSI greater than roughly 0.3 the RRFSp1 outperforms the other models. However for the lowest CSIs (0-0.05), the RRFSp1 also had the highest probability of occurring for all cycles but the 12z cycle. The NAMnest typically had the lowest probability for the low CSIs. Combined, this suggests that the RRFSp1's performance for any given event could be very good or very bad. It also suggests the the NAMnest, like seen in the PDs, is less likely to miss an event compared to the other two models.

Figure 42 shows the aggregated performance of the PMM[11] for all deterministic and probabilistic guidance, including the REFS and CAPSe for the full MRTP+. The analysis is aggregated across the model cycles available for each given domain/valid time, providing an aggregate skill at forecasting 6-h heavy rainfall

[10] Day 0 refers to cycles that initialized starting at 00z of the 12-12 UTC day definition. For example, if the forecast for Day 1 was 12 UTC June 10 to 12-UTC June 11, then any cycle initialized on June 10 is referred to as a Day 1 forecast. If the cycle was initialized at 00z or 06z on June 11 then it is referred to as a Day 0 forecast.

[11] HREF PMM was not included because it is on a different grid.

events. All deterministic forecasts have a negative bias, including the NAMnest. The REFS and CAPS PMM shift from a negative to a positive bias as the threshold increases, though the performance of the PMM should be interpreted carefully since the PMM method is calculated over the whole CONUS. The HRRR and the REFS members are clustered together, especially for thresholds ≤1 inch with a similar bias between each other and more spread in SR than in POD.

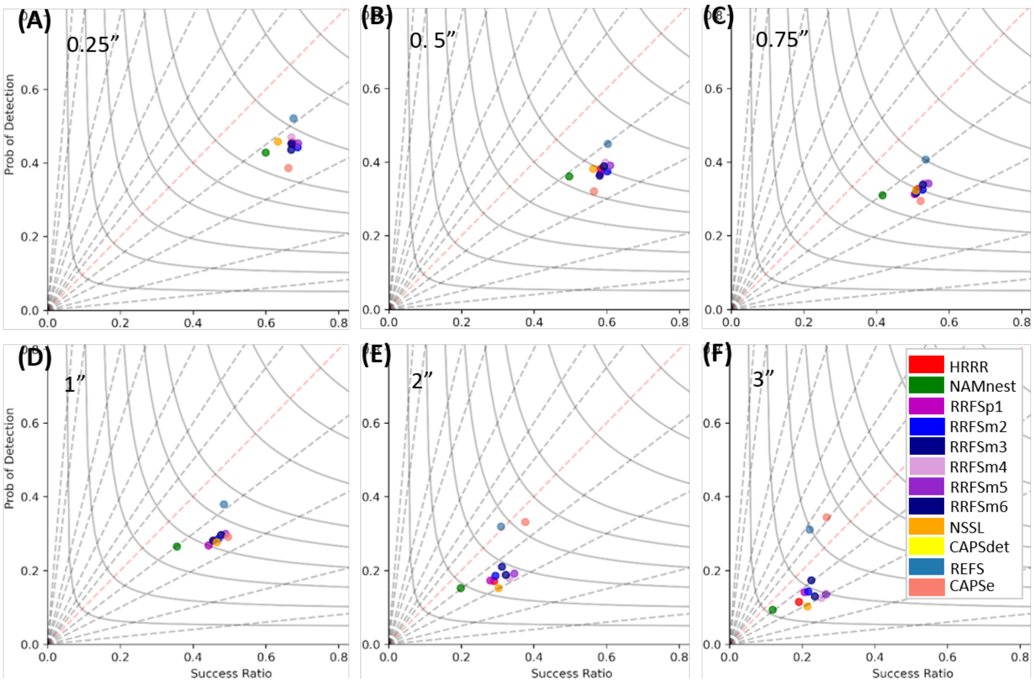


Figure 42: MRTP+ performance diagrams, aggregated across the model cycles and the 129 6-h domains that make up the MRTP+ dataset for the following thresholds: (A) 0.25", (B) 0.5", (C) 0.75", (D) 1", (E) 2", and (F) 3".

Figure 43 shows the domain max of the MRMS/models along with the 1" coverage for the MRTP+ at 48 and 24 hour forecasts. The HRRR, NAMnest, and RRFSp1 all have too much density in areal coverage of 1" bins <10,000 km^2 (hereafter "k" is refer to thousand, so 10,000 is 10k) except 24-h NAMnest, which still over forecasts but to a lesser extent. This is especially true for the first bin (<10k km^2 with maximum rainfall of <1"), where NAMnest has no instances (like observed) but the HRRR and RRFSp1, regardless of lead time, has density in this bin. For the RRFSp1, the density is significantly more than the HRRR. This suggests that the RRFSp1 is either missing events (0" max and 0 coverage of 1" is in that bin) or that it is more likely to under-forecast the maximum rainfall and coverage of 1" for the extreme rainfall events of the MRTP+; ex. the 25JUNE case (see Fig. 17). The white box in Fig. 43 highlights where the majority of the MRMS QPE falls[12], comparing each model's density distribution within the box it can be seen that the density is skewed more left in the RRFSp1 than the HRRR or NAMnest; i.e. the RRFSp1 forecasts are more likely to fall in the lower bins. Although all models under forecast the right most bins in this box, the NAMnest's density is the least different from MRMS. That said, all, models also have too much density in the bottom left of the boxed area compared to observations, with the 24-h NAMnest having the highest density.

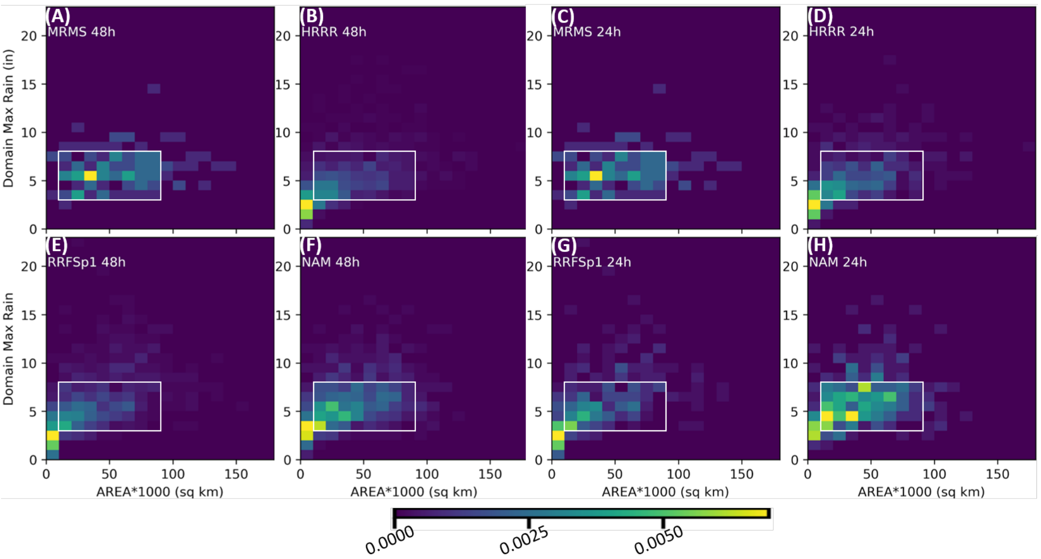


Figure 43: Heat maps comparing the domain 6-h rainfall maximum (y-axis) to the 1" coverage area in thousands of square kilometers, starting at 10,000, or 10k, km^2 (x-axis) for the MRTP+ dataset. (A)/(C) MRMS, (B) 48-h HRRR, (D) 24-h HRRR, (E) 48-h RRFSp1, (F) 48-h NAMnest, (G) 24-h RRFSp1 and (H) 24-h NAMnest. Here 24-h and 48-h refers to the forecast hour.

Specifically comparing the domain observed and model maximum magnitude[13], also gives us an insight on the model ability to forecast the maximum value. Figure. 44 shows this for the 24-h forecast for the REFS members, HRRR and NAMnest. Overall, the density for all models is shifted left, meaning an under

[12] Max domain rainfall magnitude outside of the box is mostly driven by wind farm errors discussed in Section 3.

[13] The distance between this two is not considered.

forecast of the maximum rainfall amount. Although not shown, as lead time decreases, the left skewness decreases, meaning with time the degree in which the magnitude of the rainfall maximum is under forecasted generally decreases. The 24-h NAMnest clustering is the most centered around the 1:1 line while the RRFSm5 was the most/least likely to under/over forecast the maximum. Finally, RRFSp1, RRFSm2 and RRFSm3, which have the Thompson Microphysics Scheme, have greater instances of over forecasting than RRFSm4-6, which have the NSSL Microphysics Scheme. This follows what was seen in the CONUS analysis (ex Fig. 36), where the characteristics of the hourly QPF were grouped by the Microphysics Schemes. Of the first three REFS members, the RRFSm3 (which has the saSAS CU scheme) has the greatest density of over forecasting the maximum, especially when looking at model maximum magnitudes greater than 9 inches.

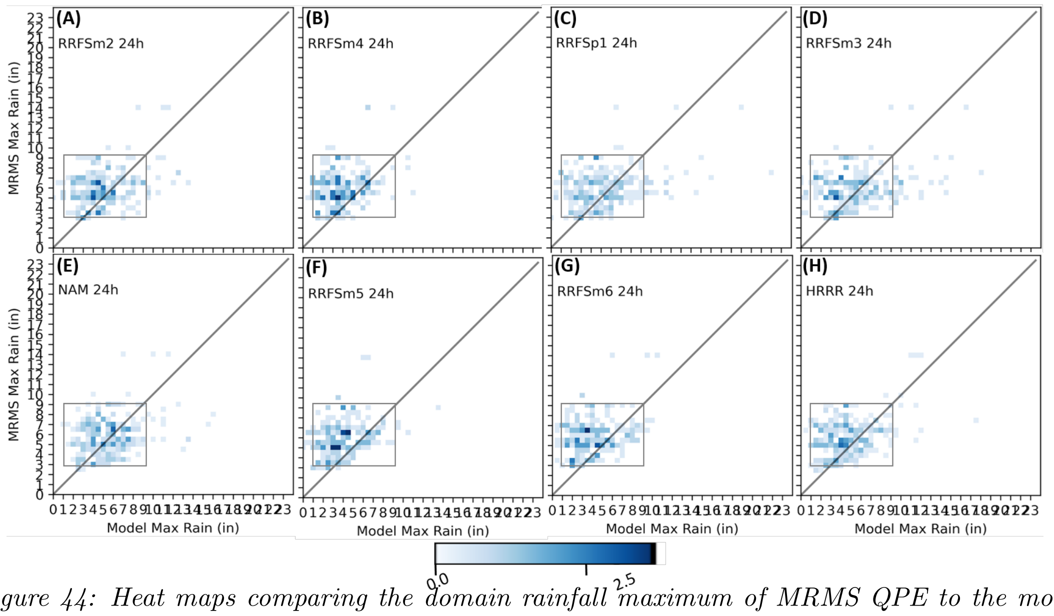


Figure 44: Heat maps comparing the domain rainfall maximum of MRMS QPE to the model QPF for the 24-h foreacst from: (A)RRFSm2, (B) RRFSm4, (C) RRFSp1, (D) RRFSm3, (E) NAMnest, (F) RRFSm5, (G) RRFSm6 and (H) HRRR for the 24-h the forecast hour.

5.2 Ensemble Analysis of the QPF

5.2.1 Subjective Analysis

This year for the subjective analysis of ensembles, participants were asked to evaluate the HREF and REFS 6-h QPF probabilities for the 1" and 5" thresholds. These are the probabilities generated by the post-processing methods at EMC. Like with the deterministic 6-h evaluation across the CONUS, the synoptic time periods for Day 1 were evaluated: 12-18 UTC, 18-00 UTC, 00-06 UTC and 06-12 UTC. Table 6 shows which of these time periods were evaluated each day while Fig. 45 shows the normalized results for the participants' impression of the probabilities compared to MRMS QPE. The choice "No Area" was meant for instances when both MRMS and ensembles had no area exceeding a given threshold and the participants felt that was the correct forecast, however, it is possible that participants misinterpreted this and put "No Area" when one of the two had no area. Figures 46-48 are examples of the 12z HREF/REFS forecasts for the 1" and 5" thresholds along with the number of times one of the three subjective scores were picked to help provide some context for the evaluation. Due to differences in how the CAPSe calculates their probabilities, subjective evaluation of their neighborhood probabilities was not done. However, the mean products were subjectively evaluated and will be discussed in this section, along with an objective overview of the mean products compared to the HREF and REFS PMM and LPMM.

For the ensembles' 1"/6-h probabilities, across all 4 time periods and cycles the participants rarely felt that either the HREF or REFS were too high. In general, both the HREF and REFS saw an increase in the number of times participants felt the probabilities were good as lead time decreased. For the 12-18 UTC and 18-00 UTC time periods, the 06z/12z REFS was more likely to score than than the same HREF cycles. For all but the 18-00 UTC time, the 18z/00z HREF subjectively was better than the REFS, meaning that at longer lead times the participants tended to feel the HREF probabilities better represented the risk of exceedance than the REFS. For 00-06 UTC (after diurnal max) period, for all cycles the HREF probabilities mostly scored "good", while for all but the 12z

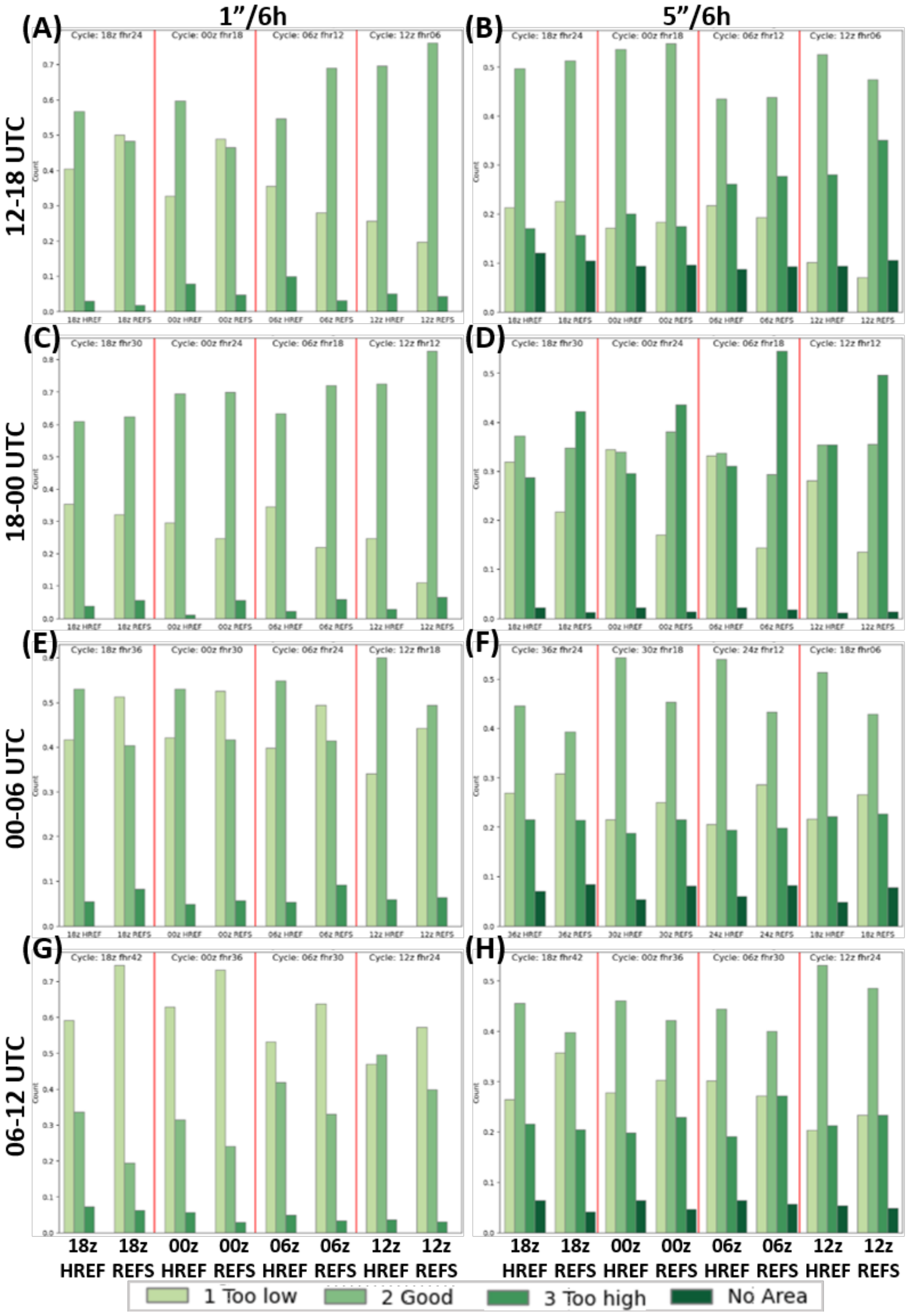


Figure 45: The normalized subjective results for the HREF and REFS the magnitudes of the 1” (left) and 5” (right) in 6-h neighborhood probabilities for the CONUS valid (A)-(B) 12-18 UTC, (C)-(D) 18-00 UTC, (E)-(F) 00-06 UTC and (G)-(H) 06-12 UTC, for the 18z (longest lead time), 00z, 06z, and 12z (shortest lead time) for the HREF and REFS. On each chart the model cycles are divided by a red vertical line.

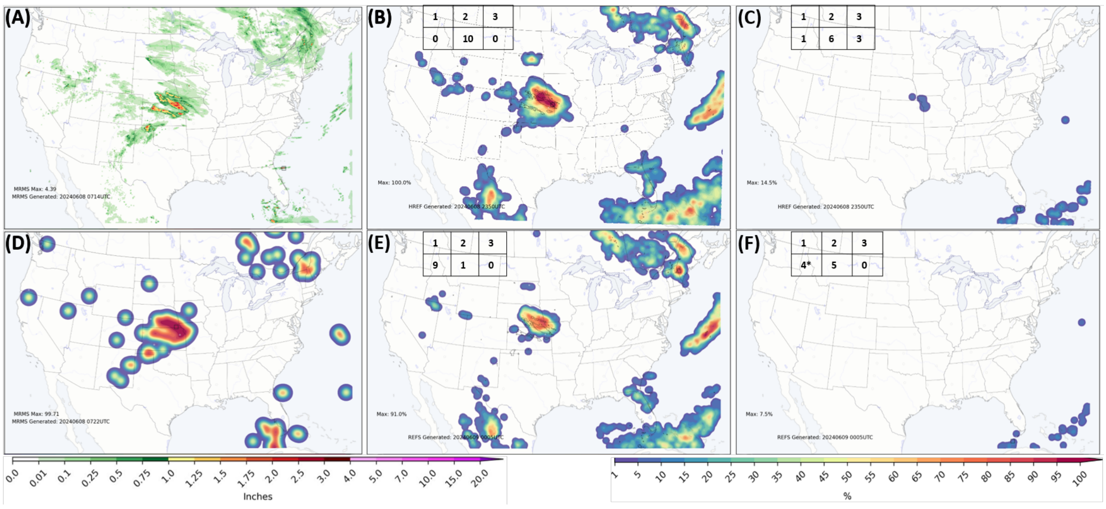


Figure 46: The 6-h (A) MRMS QPE and (D) MRMS 1" QPE with a 40 km ROI applied, valid 06 UTC 08 JUNE 2024. The 1" and 5" in 6-h probabilities for the 12z (B)/(E) HREF and (C)/(F) REFS valid for the same time. On (B)-(C) and (E)-(F) are the number of times the magnitude of the probabilities received a subjective score of (1) too low, (2) good and (3) too high.

cycle the REFS magnitudes were most likely to be considered "too low". For the 06-12 UTC time, for all cycles of the REFS and all but the 12z cycle for the HREF the most likely choice for 1"/6-h probability magnitudes was "too low". That said, for each cycle, the HREF always had a higher frequency of "good" being chosen than the REFS.

The high percentage of "too low" responses for the REFS for the latter two time periods follows what was noted in the deterministic analysis of the REFS members, that they miss events after the diurnal maximum. Figure 46 shows an instance when all participants felt the 1"/6-h magnitudes for the HREF forecast were "good" while all but one participant felt the REFS's magnitudes were "too low". For this event one participant wrote:

> "For the area of interest in the Plains, the HREF outperformed the REFS at each time step, both for 1" and 5" due to its higher probs for 1" and at least some coverage of 5". Although MRMS had no coverage of 5" here, the REFS showing 0% did not adequately reflect the potential. However, in New England, i felt the REFS did a bit better than the HREF."

The characteristics of the participants' subjective evaluations for 5"/6-h differs from 1"/6-h; most notable is the increase in the number of times that "too high"

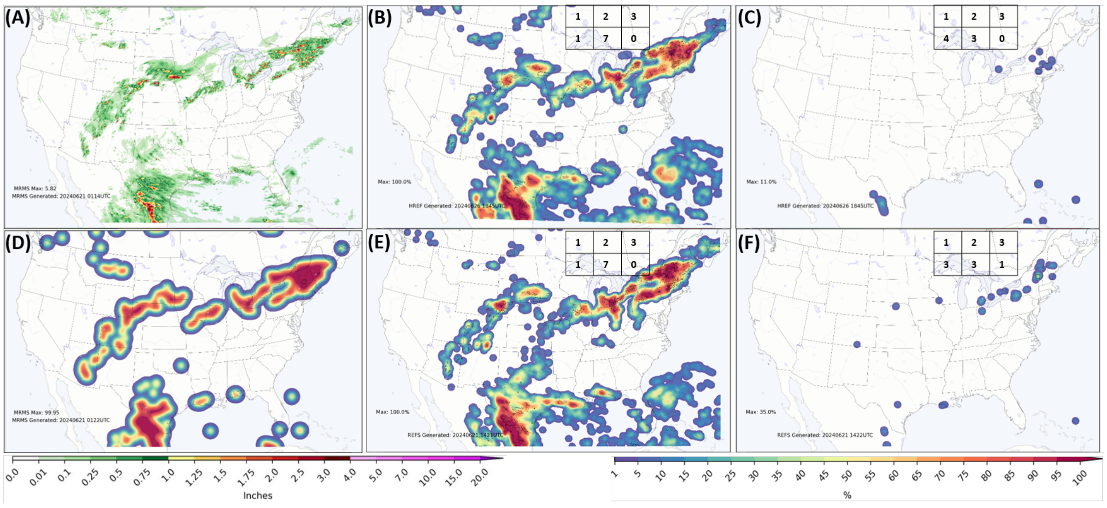


Figure 47: Like Fig. 46 but valid 00 UTC 21 JUNE 2024 (21JUNE case). The 1" and 5" in 6-h probabilities for the 12z (B)/(E) HREF and (C)/(F) REFS valid for the same time. On (B)-(C) and (E)-(F) are the number of times the magnitude of the probabilities received a subjective score of (1) too low, (2) good and (3) too high.

was chosen for the probability magnitudes. This is specifically true for the REFS at all cycles for the 18-00 UTC period, with "too high" always being the most likely choice; in fact, for the 06z cycle, over 50% of the time participants ranked the probabilities as being "too high". Differing from this, the HREF generally had similar percentages across the three scoring options ("too low", "good", "too high"). For the first two 6-h windows, the higher percentage of a "good" scores for the 5"/6-h probability varied by the cycle between the two ensemble systems. For the last two time periods, the HREF always had a higher chance of receiving a score of "good" than the REFS. For all but one cycle/period (06z cycle for 06-12 UTC) the REFS had a higher chance that the probability magnitudes were scored "too low"; aside from the 06z cycle, the chance for scores of"too high" were generally similar between the two ensembles. This further supports the finding that the REFS as a whole struggles with forecasting events after the diurnal maximum compared to the HREF. An example of the shift in ensemble performance across the diurnal maximum can be seen in Figs. 47 and 48, the 21JUNE case discussed in the deterministic section. For the forecasts valid at 00 UTC June 21, the results from the participants was nearly the same for the HREF and REFS, except for the time period valid at 06 UTC; the REFS goes from every participant feeling the

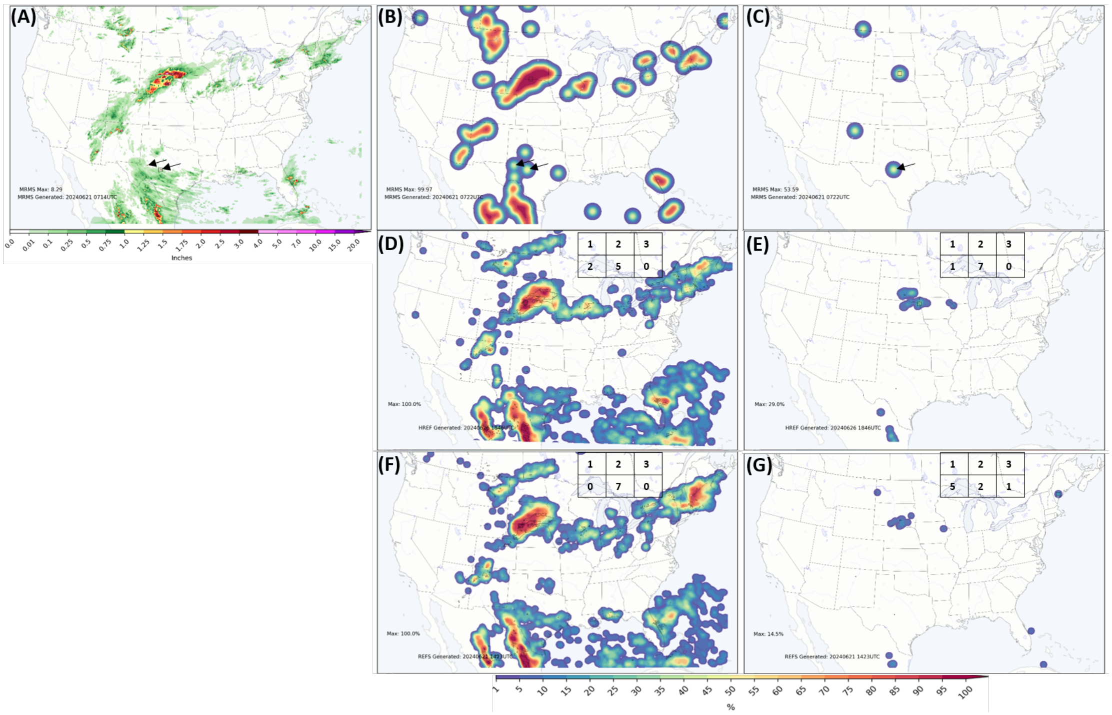


Figure 48: Valid same day as Fig. 47 but at 06 UTC (21JUNE case). 6-h (A) MRMS QPE and (B) MRMS 1" and (C) 5" QPE with a 40 km ROI applied. The 1" and 5" in 6-h probabilities for the 12z (D)/(F) HREF and (E)/(G) REFS. The arrows on (A)-(C) indicate erroneous MRMS data. and (D)-(G) are the number of times the magnitude of the probabilities received a subjective score of (1) too low, (2) good and (3) too high.

1"/6-h magnitudes are good to 5 of the 8 participants feeling the 5"/6-h magnitudes are too low. This is opposite of the HREF which had 5 of the 7 participants feel the 1"/6-h magnitudes were good; this increased to 7 of the 8 or the 5"/6-h[14].

In addition to the subjective verification of the CONUS probabilities, analysis of the participants' perception of the HREF and REFS performance was done through the MRTP. Figure 49 shows the results for the various ways this evaluation was done. Figure 49 First, during the forecasting process for the MRTP, the participants were ask to identify whether the HREF and REFS ensemble probabilities and LPMM/PMM were used (included in forecast), useful (helped with forecast), considered (looked at or thought about as possible outcome) and/or not considered (not looked at or thought about) in their forecast. Then during

[14] One participant only provided an answer for the 5"/6-h, thus there being one score for the 5"/6-h than for the 1"/6-h subjective evaluation on this day.

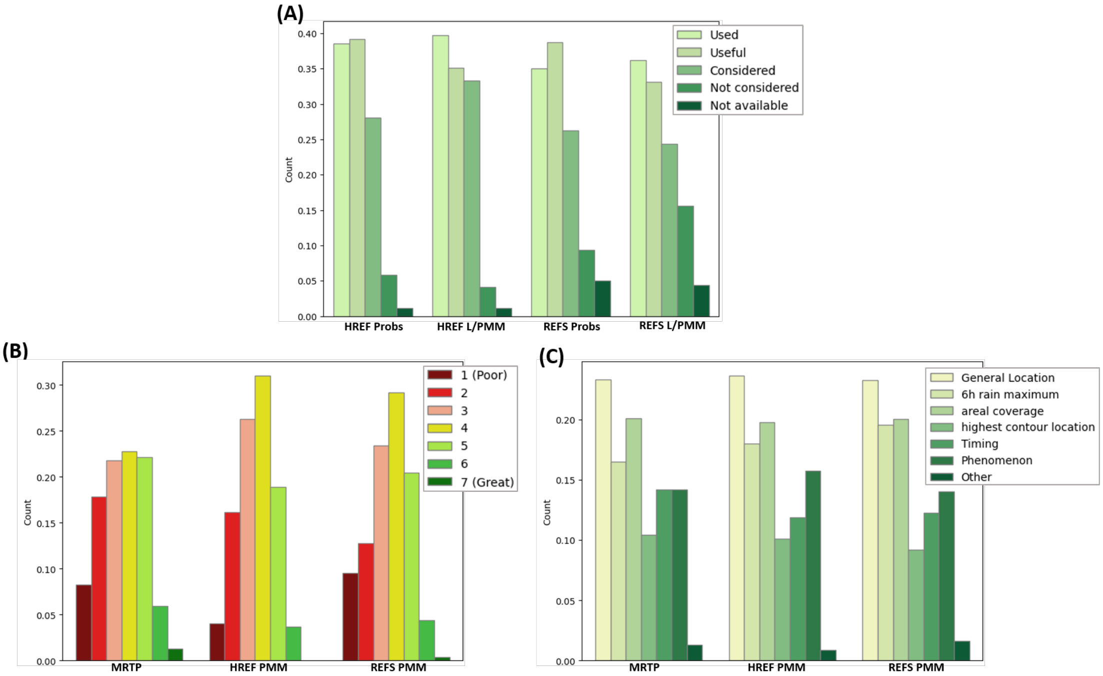


Figure 49: (A) Results from the MRTP activity survey for the question "How did you use the suite of Model and Ensemble Guidance available to you?" for the HREF and REFS's probabilities and L/PMM. (B) Results for subjective verification of the HREF and REFS's PMM along with the participants' MRTP forecast on a scale of 1 (poor) to 7 (great). (C) Same as (B) but the feedback from the question "What were the good aspects of those forecasts?", participants could chose any option that applied. All results were normalized.

subjective verification participants also evaluated the PMM for each ensemble for the MRTP domain/time by providing a "goodness" score[15] and indicating what qualities of the forecast they thought were good.

Overall, the REFS products were more likely to be "not considered" than the HREF's products. This was especially true for the REFS LPMM products, which were "not considered" 15% of the time compared to the HREF's 4%. The REFS probabilities were useful roughly the same amount of the time as the HREF probabilities but identified as "used" less than the HREF (35% to 40%). The scores for the participants' MRTPs and the HREF PMM had an average of 3.56, while the REFS PMM had a marginally lower average of 3.53. In general, the shape of the distribution for the goodness scores for the two ensembles were similar, with

[15] Same 1 (very poor) to 5 (very good) scale used in the deterministic verification.

the exception of the score of 1 (poor). The REFS saw more scores of 1 (15% to 5%) than the HREF. The HREF had slightly higher percentages of scores of 2 and 3 than the REFS. Thus, is it likely that, similar to what was seen with the RRFSp1, when the REFS didn't perform well, it often performed quite poorly, while a less-than-good performance from the HREF was likely to still provide some utility (i.e., rarely deserving a 1). In terms of what aspect of the forecasts were good, the distributions were relatively similar. That said, the participants did feel that the forecast timing of the MRTP was more likely to be good compared to the ensembles. The HREF was the most likely to have "Phenomenon" (i.e. type of rainfall event) chosen as a good aspect of the forecast while the REFS was slightly more likely to have the 6-h rain maximum be chosen.

Figures 50 and 51 show the HREF and REFS PMM and probabilities of exceeding 1"/6-h for the 25JUNE MRTP case, while Figs. 52 and 53 show similar plots for the 21JUNE and 22JUNE MRTP cases respectively. The 25JUNE case was discussed in length in Section 5.1.2, as an example when the REFS members struggled with CI after the diurnal maximum. Evaluation of the PMM and probabilities for this case (Figs. 50 and 51) shows, not surprisingly, that both the PMM and probabilistic forecasts failed to highlight the risk of northern WI seeing at least 1" in 6-h. While the HREF forecasts the location of the event incorrectly, southern WI vs northern WI, the REFS misses the event completely. Forecasters noted the REFS "severely underperformed" for the event. When events, like the 21JUNE and 22JUNE cases, were driven by convection that was ongoing prior to the diurnal maximum, the REFS performed as good or better than the HREF. For instance, the for each of these cases, the average "goodness" scores for the REFS PMM was slightly higher than the HREF PMM; 4.4 vs 4.64 and 4.4 vs 4.5 for each respective case.

5.2.2 Objective Analysis

Figure 54 shows the performance of the HREF and REFS 24-h PMM and LPMM over the CONUS; also included is the CAPSe's PMM, LPMM, SAM and SAM-LPM. For both the HREF and REFS, the PMMs always have a higher CSI than the corresponding LPMMs, but as the threshold increases the difference in

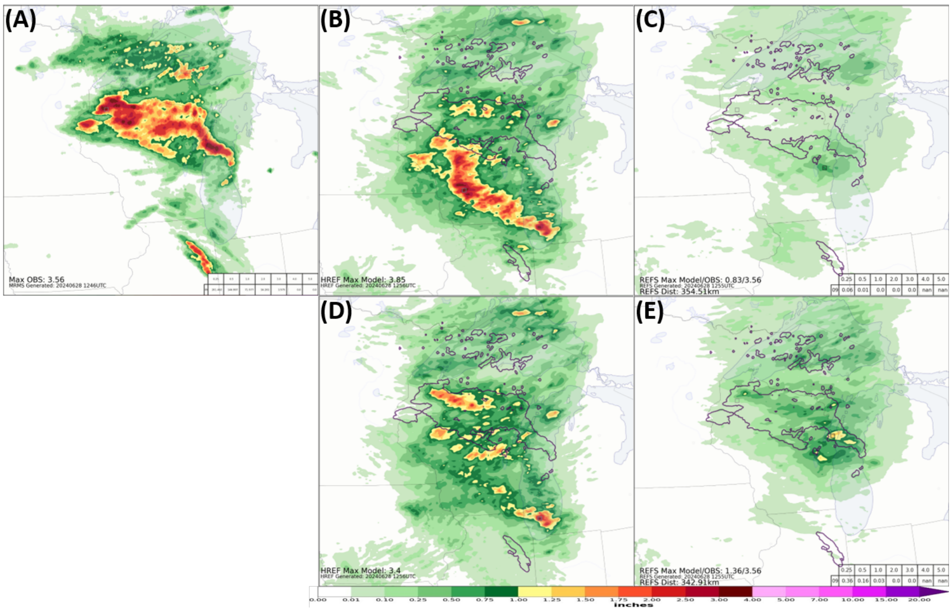

(A)
Max OBS: 3.56
MRMS Generated: 20240628 1246UTC
(B)
HREF Max Model: 3.85
HREF Generated: 20240628 1256UTC
(C)
REFS Max Model/OBS: 0.83/3.56
REFS Generated: 20240628 1255UTC
REFS Dist: 354.51km
0.25 0.5 1.0 2.0 3.0 4.0 5.0
09 0.06 0.01 0.0 0.0 0.0 nan nan
(D)
HREF Max Model: 3.4
HREF Generated: 20240628 1256UTC
(E)
REFS Max Model/OBS: 1.36/3.56
REFS Generated: 20240628 1255UTC
REFS Dist: 342.91km
0.25 0.5 1.0 2.0 3.0 4.0 5.0
09 0.36 0.16 0.03 0.0 0.0 nan nan
0.00 0.01 0.10 0.25 0.50 0.75 1.00 1.25 1.50 1.75 2.00 2.50 3.00 4.00 5.00 7.00 10.00 15.00 20.00
inches

Figure 50: 6-h (A) MRMS QPE and (B) 06z HREF, (C) 06z REFS, (D) 12z HREF, and (E) 12 REFS PMM valid 09 UTC 25 June 2024 (25JUNE MRTP case). The purple contour on (B)-(E) is the 1" area from MRMS.

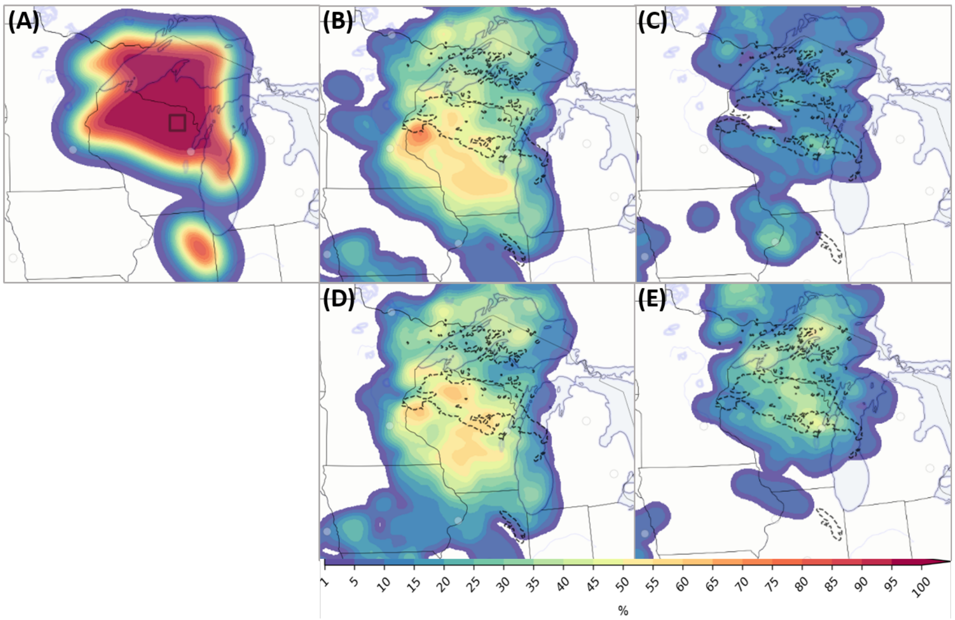


Figure 51: 6-h (A) MRMS QPE 1" with a 40 km ROI applied and (B) 06z HREF, (C) 06z REFS, (D) 12z HREF, and (E) 12 REFS probabilities of 1"/6-h valid 09 UTC 25 June 2024 (25JUNE MRTP case). Refer to Fig. 50 for QPE and ensemble PMM. The dashed black contour on (B)-(E) is the 1" area from MRMS.

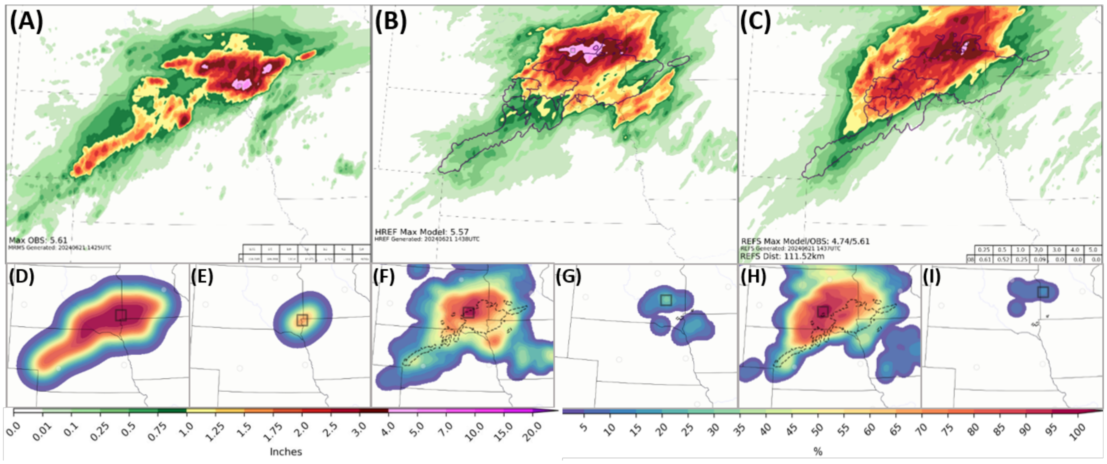


Figure 52: 6-h (A) MRMS QPE and 6-h PMM with the 1" area from MRMS contoured in purple from the (B) 12z HREF and (C) 12z REFS valid 07 UTC 21 June 2024. Along the bottom are 6-h probabilities for the (D)-(E) MRMS QPE with a 40 km ROI applied to the given threshold, (F)-(G) HREF and (H)-(I) REFS valid at 06 UTC the same day; 6-h probabilities are not generated for 07 UTC, when this MRTP was valid. (D), (F), and (H) are for 1"/6-h and (E), (G), and (I) are for 5"/6-h. On (F)-(I) the MRMS QPE is contoured in dashed black for the respective threshold.

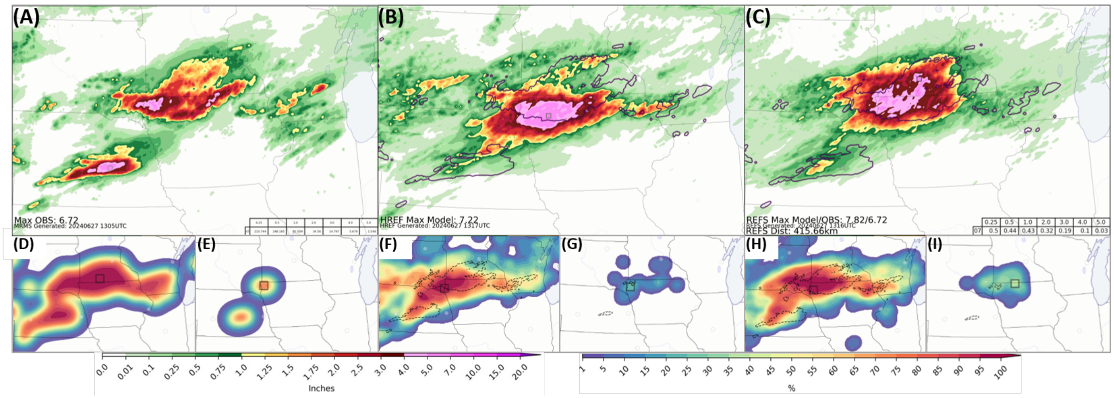


Figure 53: Like Fig. 52 but all images valid 06 UTC 22 June 2024.

CSI between the two means decreases. The PMM for both ensembles also always has a higher POD than the matching LPMM. Compared to one another, for all but the 2" threshold for the 12z cycle, the HREF means have a higher POD than the corresponding mean from the REFS. On the other hand, the REFS means always had a higher SR than their HREF counterparts. This results in a similar CSI between the two ensembles and their respective means. In terms of bias, the HREF PMM and LPMM always have a higher bias than the respective REFS means, with the REFS means having a bias below 1 for the one-half and one inch thresholds

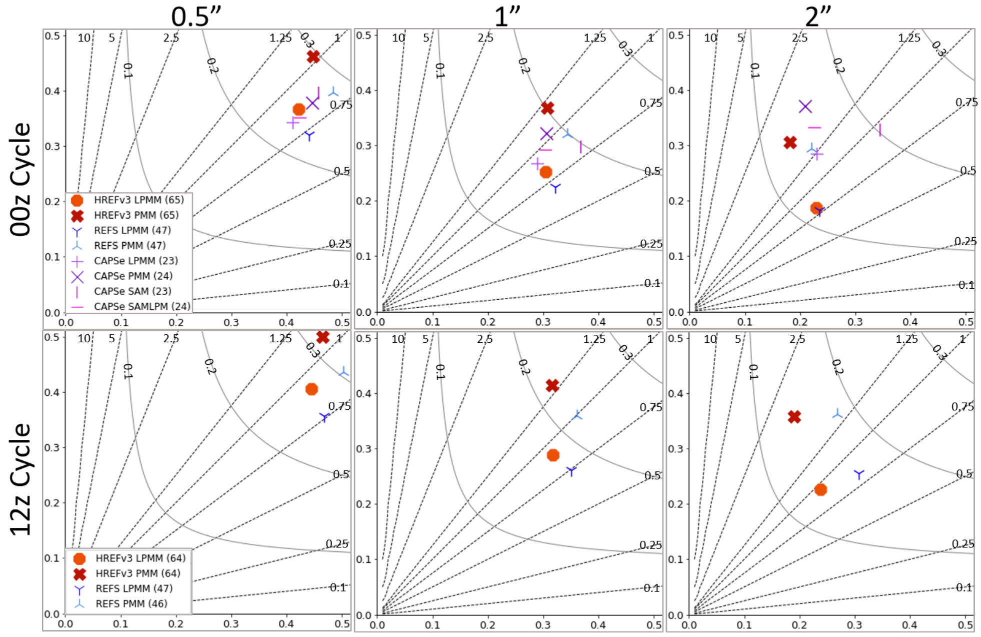


Figure 54: Performance diagrams for CONUS 24-h QPF from 01 June to 04 August 2024 for the 00z and 12z HREF and REFS PMM and LPMM. The CAPSe (valid only for 00z cycle) is also shown for the PMM, LPMM, SAM, and SAM-LPMM. Top is the 00z cycle initialization and bottom is the 12z, for the thresholds of 0.5 (left), 1 (middle), and 2 (right) inches.

(exception is the 12z PMM for 1”, which has a bias of 1). The LPMM bias for both ensembles is always dry, but the REFS LPMM has a stronger dry bias for nearly every threshold/cycle. For the 00z cycle at the 2” threshold, the CSI of each system’s means are similar to one other, while for 12z, the REFS outperforms the HREF.

As mentioned above, forecasters use the PMM to get a sense of the possible maximum rainfall, so they want and expect the PMM to over forecast totals. The dry bias or lack of high bias, though encouraging in terms of statistics, is worrisome in terms of forecaster utility. It is also important to note that this evaluation was done on 24-h totals, not 6-h, which is what the subjective results focused it. It is likely that the increased duration, covering the whole Day 1 time period, helped to wash out some of the characteristics of the means at shorter durations. This is supported by the the differences discussed in Section 5.1.7 seen in the individual members of the REFS when the 24-h and 6-h performance characteristics are compared; i.e. Figs. 35 and 34. However, due to compute and personnel constraints, 6-h mean performance PDs were not generated.

Another way to evaluate performance of the REFS across the Day 1 time period is to look at the distribution of the probabilities across the time period. Since this year's FFaIR focused on a 6-h forecasts, the normalized distribution of 6-h MRMS QPE and QPF probabilities was plotted every 3 hours for both the REFS and HREF; see Figs. 55 and 56. For this analysis, zeros were not included so the first bin is from 0.01% to <5%. After that, bins are every 5%, with the low end value included in the binning process and the high value excluded; meaning for the 5% bin, $5\% \leq x < 10\%$ and for the 10% bin, $10\% \leq x < 15\%$ and so on. Before comparing the two ensembles, it is important to note that although the results have been normalized, the HREF and REFS do not share the same spatial resolution (REFS is 3 km while HREF is 5 km) so probabilities themselves are not apples to apples, more like a golden apple to a granny smith apple.

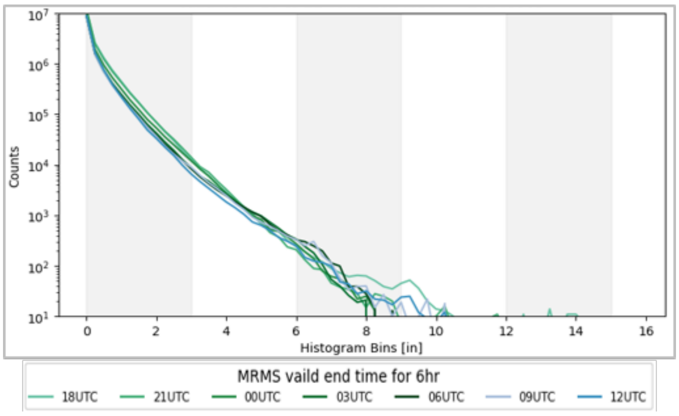


Figure 55: Normalized histogram of the occurrence of MRMS QPE 6-h accumulations from 07 May to 05 August 2025 for forecasts from 12z-12z for Day 1. Color shading indicates the valid end time within the Day 1 time period every 3-h.

The distribution of MRMS QPE 6-h precipitation is similar among the valid hours until between 5 and 6 inches. After that, there is a bit more spread in the instances of 6-h accumulation $\geq$6”; some of the higher totals are the result of the erroneous MRMS data discussed in Section 3. Comparing the HREF and REFS to MRMS, both ensembles have a lower chance of a given probability being forecast by their respective systems for those valid across the diurnal min, v18 (top row in Fig. 56), v09 (bottom row), and v12 (not shown) UTC[16]. This was not seen in the MRMS distribution and could suggest that both systems struggle with sustaining overnight convection, like MCSs, into the morning, early afternoon hours or that there is a delay in convective initiation in the early afternoon in the ensemble systems. Interestingly, the 12z REFS (dashed purple) v18 had a significantly higher distribution for the 1" 6-h^{-1} for probabilities ≥30% compared the its other cycle and both HREF cycles. This is likely a reflection of the wet bias discussed previously in the first 6-h of the 12z forecast.

In terms of the probability distribution between the HREF (green Fig. 56) and the REFS (purple), regardless of the threshold, (1", 3" or 5" in 6-h), the REFS had the same or higher likelihood of low end (<5%) probabilities forecasted than the HREF. Across the other probability bins, the REFS nearly always had a higher chance of the probability occurring for all thresholds and times. The greatest difference in ensembles probability distribution between the HREF and REFS occurred for the forecasts valid 21 and 00 UTC. The characteristic of higher likelihoods might suggest that the REFS has lower spread than the HREF, though this could also be a function of the higher resolution in the REFS.

For the 1" and 3" probabilities, the HREF's slope is relatively constant compared to the REFS's slope. This is particularly true for from roughly the 5% to 15% forecast probability bins, where the REFS's slope, regardless of cycle or valid time, has a sharp decrease; this is also present for 5" 6-h^{-1}. This is most notable for the forecasts valid 06, 09 and 12 (not shown) UTC. This steep slope in the REFS's distribution, results in the REFS and HREF having nearly the same chance of probabilities across the 10-20% probability space for the forecasts valid 18, 06, 09 and 12 UTC. For both 09 and 12 UTC the 1" 6-h^{-1}, from roughly 10% to 20%, the 12z REFS has less occurrences of those probabilities than the HREF. Once the forecast probability is greater than 40% the REFS begins to diverge from the HREF, resulting in higher occurrences of the higher probabilities; 3" 6-h^{-1} is similar. This disparity between the HREF and REFS could suggest that for marginal events or events that are harder to predict (e.g. MCSs) that happen

[16] Reminder for the Day 1 time period we are referring to 12 UTC to 12 UTC so v18 UTC is the first 6-h time period of the day and v09 and v12 UTC are the last two 6-h time periods of Day 1.

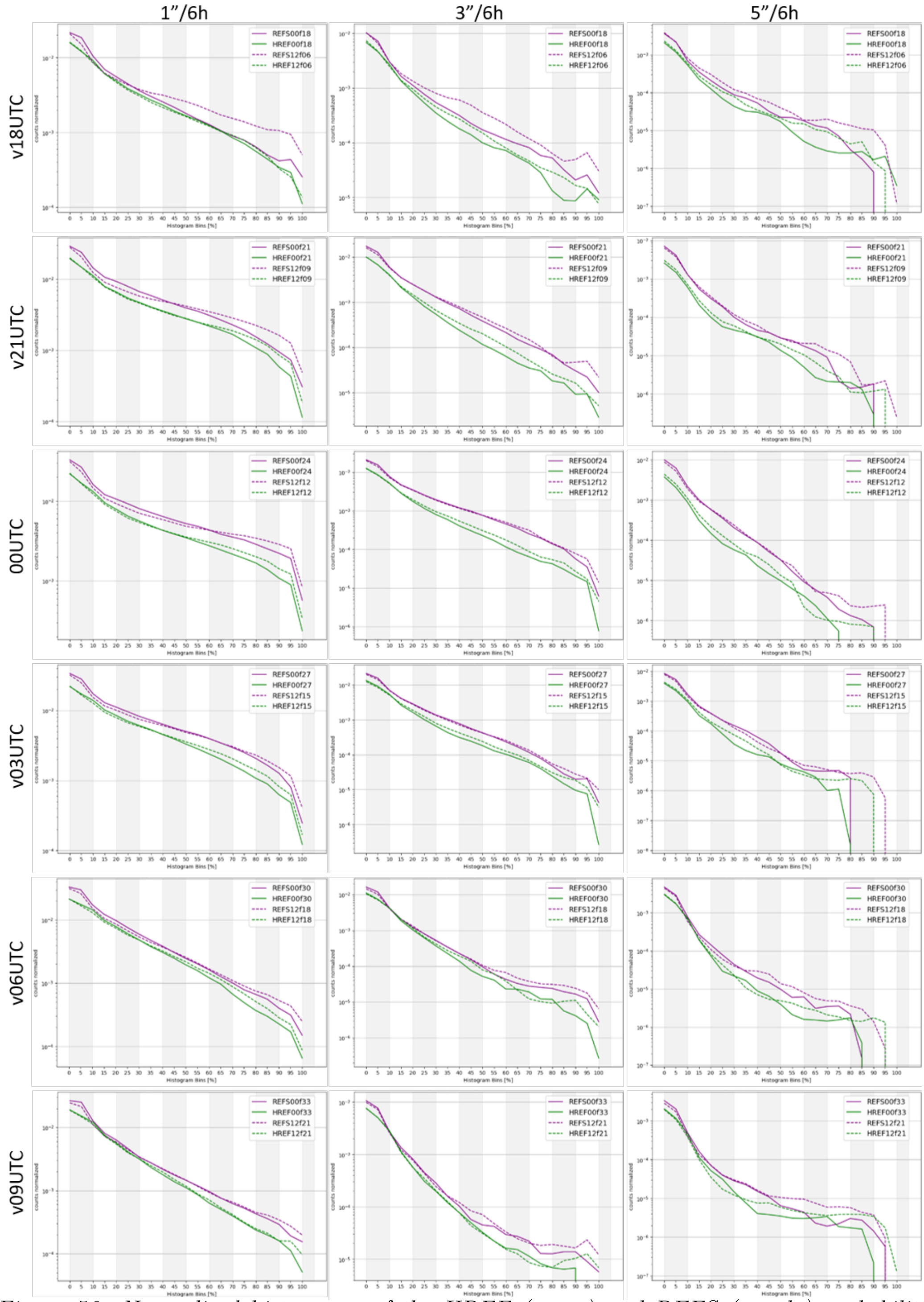


Figure 56: Normalized histograms of the HREF (green) and REFS (purple) probabilities for thresholds of 1”/6-h (left), 3”/6-h (middle) and 5”/6-h (right) over same time as Fig. 55. Solid lines are for the 00z cycle and dashed are 12z cycle. Forecasts from both ensembles are every 3-h, valid top to bottom at 18, 21, 00, 03, 06, and 09 UTC. The histogram for the forecasts valid 12 UTC is not included.

after the diurnal max and across the diurnal minimum, the REFS under forecasts the risk of 1"/6-h, though it is possible that the HREF over forecasts the risk. More analysis will need to be done to confirm one or the other, but based on the previously discussed analysis done on the membership, the former seems more likely than the later.

Both ensemble system had instances in which they forecasted a 85% chance or greater of 5" 6-h^{-1}. The forecasts valid at 00 UTC saw the greatest difference in probability distributions between the two, while the 09 and 12 UTC times saw both ensembles have a nearly consistent chance of forecasting probabilities ≥40%. Like with the other two thresholds, for the probabilities between 10% and 20% saw nearly identical occurrences for the two ensembles.

5.2.3 End of the Week Survey

Like with the deterministic models, participants were asked general questions about their overall perceptions of the REFS in the end of the week survey, to gauge their impression of its performance across the week. Figure 57 shows the results from the questions asking about their impressions on the maximum values for the PMM and the magnitudes and spatial coverage of the neighborhood probabilities from the REFS. Interestingly, there was large agreement (75%) among the participants that completed the survey that the PMM provided reasonable maximum values for QPF. This seems to be in slight contrast to the daily survey results shown in Fig. 49A and B, with the participants noting they often did not use the PMM or LPMM in their forecast, and the REFS PMM was most likely to receive a score of 1 (poor) in terms of its performance for the MRTP. That said, participants were more likely to highlight the 6-h PMM maximum from the REFS as a good aspect of the forecast than they did for their own MRTP or the HREF PMM. Unfortunately, in the written comments on the three aforementioned questions, the focus was more on the neighborhood probabilities, so it difficult to determine why there seems to be a slight disconnect between the daily and weekly surveys.

As for the neighborhood probabilities, no one indicated that the magnitudes of the probabilities were often too high or too low. That said, 45.8% of the respondents noted that they felt that the magnitudes were occasionally too low

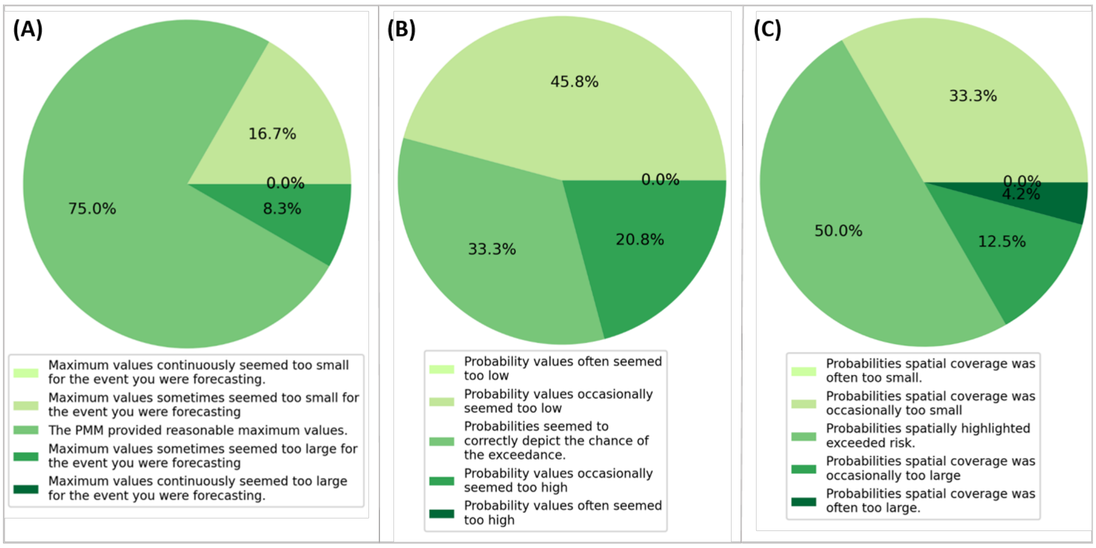


Figure 57: Results from the end of week survey for the questions: (A) ”During your week, what was your general feelings on the REFS’s PMM?” (B) “During your week, what was your general feelings on the REFS’s neighborhood probabilities’ magnitudes?” and (C) ”During your week, what was your general feelings on the REFS’s neighborhood probabilities’ spatial coverage?”

verse 20.8% feeling that they were occasionally too high. In terms of spatial coverage of the probabilities, half of the participants felt the coverage of the probabilities highlighted the exceedance risk. 33.3% felt that the coverage was slightly too small compared to 12.5% that felt it was slightly too large. Below are some of the written comments that summarize the generally feelings from the participants on the probabilities and the REFS as a whole.

- “As mentioned, there tended to be an underdispersive tilt compared to the HREF. The probabilities were much tighter focused with sharper gradients. This probably verifies well in higher predictability events, but there were several examples where the REFS missed some areas of more localized/random heavy rainfall while the HREF had more widespread low probabilities in the given area.”
- “REFS seemed to be slightly underdispersive, as indicated by probabilities that were sometimes higher than HREF and some missed events.”
- “Very generally, the RRFS was on the dry side of the guidance. Worked well in hindsight for those events that underperformed, but was usually too dry.”

- “Overall, I really liked the neighborhood REFS probabilities.”
- “The probabilities seemed generally decent, but I believe in most instances the HREF did a bit better.”
- “I thought they did good and were useful.”
- “They were still a little under-dispersive, but definitely helped to highlight some of the uncertainty.”

5.2.4 CAPSe QPF Means

During the course of the five FFaIR weeks, there were unplanned outages of the CAPSe[17] and therefore it was only subjectively evaluated 10 times; see Table 6. Thankfully, data was able to be collected retroactively and was included in the 24-h performance diagrams of ensemble means. Figure 58 shows the subjective results for the CAPS 6-h QPF SAM and SAM-LPM, which included evaluating the CAPSe and REFS 6-h QPF LPMMs. The REFS LPMM was more likely to be chosen to have a footprint that was smaller than observed compared to the CAPSe LPMM. In terms of maximum magnitude, participants were more likely to choose the ‘maximum is lower than observed” for the REFS LPMM than the CAPSe LPM.

Focusing specifically on the three CAPS means, hereafter SAM, LPM, and SAM-LPM, participants felt that the SAM maximum QPF had a strong likelihood of being lower than observed while the SAM-LPM was the most likely of the means to have participants note that they felt the maximum value was too high compared to observations. All three means had roughly the same chance of participants noting that the footprint was too small compared to observations. The SAM-LPM fell between the SAM and the LPM in terms of chance of footprint being similar to observations (LPM had the greater chance, SAM the lower) and between the two for likelihood of the footprint being smaller (LPM chance was lower, SAM’s was higher). Figure 59 shows an example of the differences among the three means. Note that the SAM and SAM-LPM have a similar looking footprint to one another, though the SAM has widespread coverage of light ($\leq$0.1”) precipitation that is

[17] As a reminder CAPSe is the CAPS Ensemble.

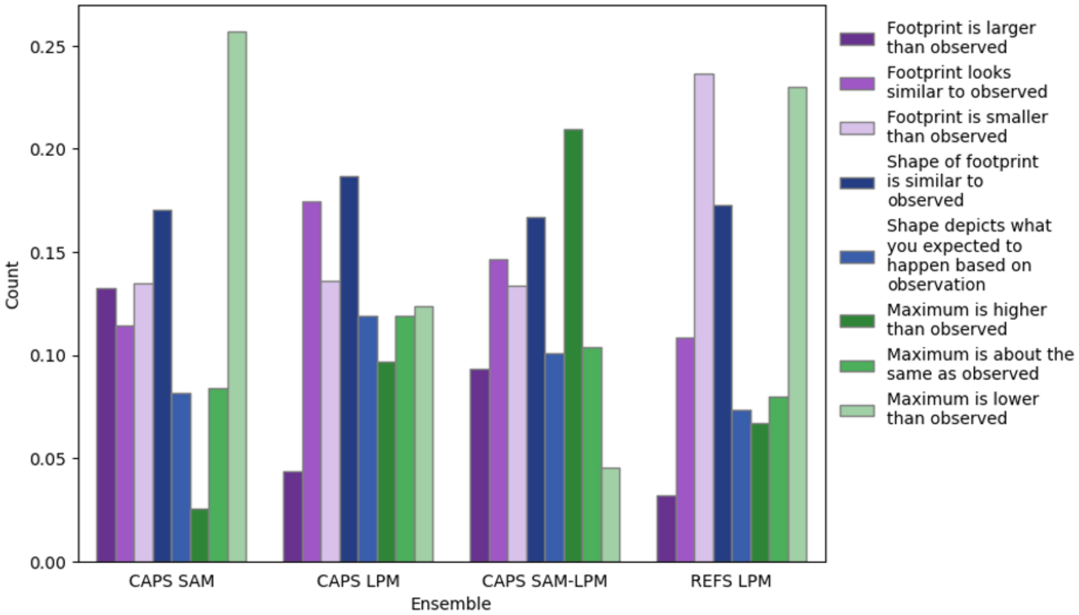


Figure 58: The normalized count of the number of times each type of ensemble mean fit the given description for the 2024 FFaIR dates that the CAPSe was available.

not seen in the SAM-LPM. Meanwhile, the SAM-LPM increases maximum values with the heaviest axis of precipitation. Finally, despite the methodology of the SAM being designed to retain the character of the precipitation footprints from the individual members, the SAM and SAM-LPM had a lower chance of participants choosing "Shape of footprint is similar to observed" than the LPM did.

Some general features were noted by participants. First, it was really case dependent on which of the three CAPS means captured the shape of the footprint best, as can be seen by the small differences in the dark blue bar in Fig. 58. Secondly, the SAM and SAM-LPM looked similar, aside from magnitude, which is to be expected. Often the increase in magnitude seen in the SAM-LPM was not preferred over the SAM magnitudes. This perception of not preferring the higher values seen in the SAM-LPM is supported by Fig. 54, which shows that for the 24-h QPF for the one-half, one, and two inch thresholds, the SAM always outperforms the SAM-LPM. Often they have the same POD but the SAM has a noticeably higher SR. Interestingly, despite the subjective feedback suggesting that the LPM performed better overall than the SAM and SAM-LPM, based on the 24-h QPF performance diagrams, the CAPS LPM always performed worse

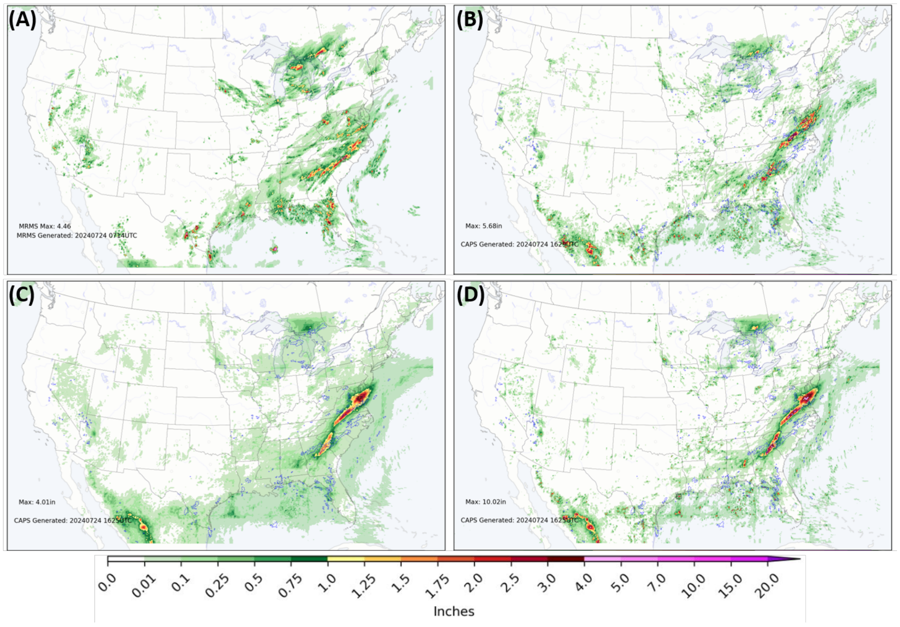


Figure 59: 6-h (A) MRMS QPE and CAPS (B) LPM, (C) SAM, and (D) SAM-LPM from the 00z initialization on July 23, valid 06 UTC 24 June 2024.

compared to the SAM and SAM-LPM. That said, as noted earlier, when looking at the 24-h performance of the REFS and HREF, subjective evaluation was done on 6-h QPF, so tendencies seen in the 6-h might be "washed out" in the 24-h analysis.

Participants in the end of the week survey had positive comments about the two products. For instance, one participant wrote:

> "Yes. The SAM definitely ""focuses"" the precipitation field better, leading to higher precipitation maxima without ""blowing out"" the field with extreme precipitation everywhere." Another participant wrote: "I've been evaluating the SAM product from CAPS for years. It has clearly shown some benefit (although not dramatically). I'd have to finish by commenting that I have grown a bit tired of verifying this technique every year in the verification exercise. I suggest denoting the ""SAM"" as a success, using it in the future, and replacing a question verifying it with something else in future FFaIRs."

The facilitators of FFaIR tend to agree with the last comment and would like to see the SAM methods transitioned into the operational output from the NWS ensembles.

5.3 CAPS MLP 6-h QPF Probabilities

The CAPS MLPs provided were trained across 4 members from the CAPSe and 8 members from the HREF. CAPS provided both ensemble-based probabilities of rainfall exceeding 0.5 and 1 inch in 6-h and ensemble mean ML 6-h precipitation. Only the MLP for 1"/6-h (hereafter MLP Probs.) were evaluated during the subjective verification session, though the MLP means were available on the HMT website for participants to use during their forecast activities. Like was mentioned in Section 5.2.4, this product was only evaluated 10 times during FFaIR.

Figure 60 shows the results of the CAPS MLP 1"/6-h compared to the CAPSe and HREF 1"/6-h probabilities. As has been the trend in FFaIR and in the HMT Winter Weather Experiment (WWE), the CAPS MLP probabilities are considered poor in most instances when compared to both the CAPSe and the HREF. During the cases in which the MLP was evaluated in FFaIR this year, the MLP Probs. were never considered much better than either of the ensemble probabilities they were compared to. However, there were a few instances in which the participants felt the MLP did slightly better than the CAPSe or HREF probabilities. An example of one of these instances is shown in Fig. 61A-D; on this day 2 of the 14 participants indicated they felt the MLP Probs. were slightly better than the CAPSe's; 3 of 14 said they were about the same while 2 of 14 said it was slightly worse and half said it was much worse. One of the participants that picked slightly better stated "CAPS MLP seems to reduce probabilities across all areas, not sure if that is desirable" while the other made no comment. This comment does not provide much insight into what they liked about the CAPS MLP over the CAPSe, though it is possible that the "look" of the CAPSe probabilities, which only include an ROI but not a smoother, could be the reason for the MLP Probs. being preferred (i.e. they liked the "look" of the product compared to the CAPSe's). For comparison, this participant indicated that the MLP Probs. were slightly worse than the HREF for this day while the other participant indicated that MLP Probs. were much worse than the HREF.

Reading though the written comments that were made during the evaluation, it appears that there were instances in which the participants misunderstood which

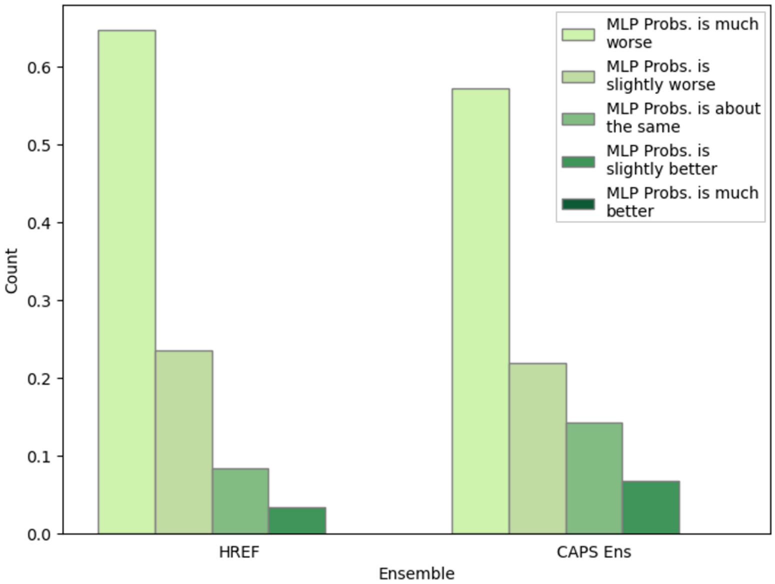


Figure 60: The normalized subjective results for the comparison of the CAPS MLP probabilities to the CAPSe and HREF for 1"/6-h probabilities. Participants scores ranged from MLP Probs. are much worse than the CAPSe/HREF to MLP Probs. are much better than the CAPSe/HREF.

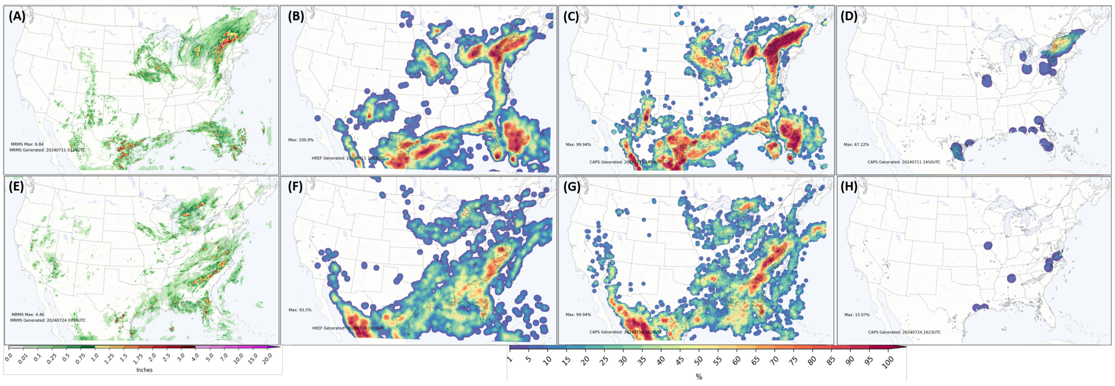


Figure 61: The 6-h (A) and (E) MRMS QPE and probabilities of 1"/6-h from the (B) and (F) HREF, (C) and (G) CAPSe and (D) and (H) the CAPS MLP. (A)-(D) are valid 00 UTC 11 July 2025. (E)-(H) are valid 06 UTC 24 July 2025. Dashed black on (B)-(D) and (F)-(H) are the 1" MRMS QPE objects.

probabilities were being compared. For instance, for the example in Fig. 62, when comparing the MLP Probs. to the HREF a participant chose that the MLP Probs. were slightly better than HREF, and for the CAPSe comparison probabilities, indicated that the MLP Probs. were much worse but in the comments wrote "The CAPS MLP was very underdone in comparison to MRMS, CAPS, and HREF both in the the areal extent of 1" probabilities and the magnitude/actual probabilities themselves. For this particular forecast, CAPS did extremely well.” This suggests that they compared the CAPSe to the HREF and the MLP Probs. rather than the MLP Probs. to the HREF and CAPSe. Other examples of this were found when reviewing the scores with the comments. This confusion was likely due to the naming convention used and the way the question was asked and will be noted for next year’s FFaIR. Despite some of this confusion, the written comments and discussions during the experiment made it abundantly clear that participants felt that the CAPS MLP Probs. were continuously too low. In one case, the differences were so extreme; despite both the HREF and CAPSe having probabilities of exceeding 1” ≥80% in the same location, the CAPS MLP 1”/6-h had 0% chance of exceedance; verification showed widespread areas of ≥1” (ex. Fig. 61E-H in the Carolinas). Finally, in the events shown in Figs. 61 and 62, the MLP Probs. had 0% chance of 1”/6-h rainfall in the west (Front Range westward) while both parent ensembles had probabilities greater than 0; verification showed that there were corresponding observations of 1”/6-h. Based on these results and those from previous FFaIRs, this product is not recommended for transition.

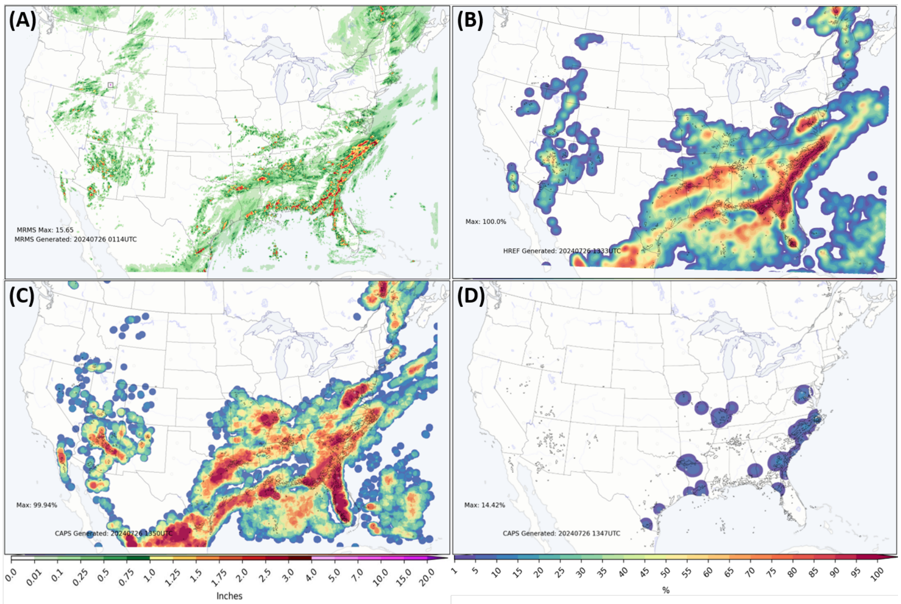


Figure 62: The 6-h (A) MRMS QPE and 1”/6-h from the (B) HREF, (C) CAPSe and (D) CAPS MLP, valid 00 UTC 26 July 2025. Dashed black on (B)-(D) are the 1” MRMS QPE objects.

5.4 ISU Bias Corrected HREF MLP

The ISU Bias Corrected HREF MLPs, hereafter the HREF MLPs, attempt to identify an MCS-like feature from the HREF membership and bias-correct for the location errors over the area in which the ML model identified the feature. The ISU group provided domain specific graphics for 24-h PMM and 10 km neighborhood probabilities of exceeding 1, 2, 3, 5, and 8 inches in 24-h. HREF MLP were only provided for the 12z cycle, so these products were not used in real-time forecasting activities. For those participants who compared the operational HREF output to the HREF MLPs, Fig. 63 shows the results from the subjective evaluation. For the PMM comparison, the HREF LPMM was compared to the HREF MLP PMM since the LPMM methodology would result in a more accurate comparison to the HREF MLP PMM given that the operational HREF PMM covers the full CONUS, while the LPMM is for more limited-area domains across the CONUS. Thus, the LPMM is less likely to be impacted by QPF outside of the HREF MLP domain.

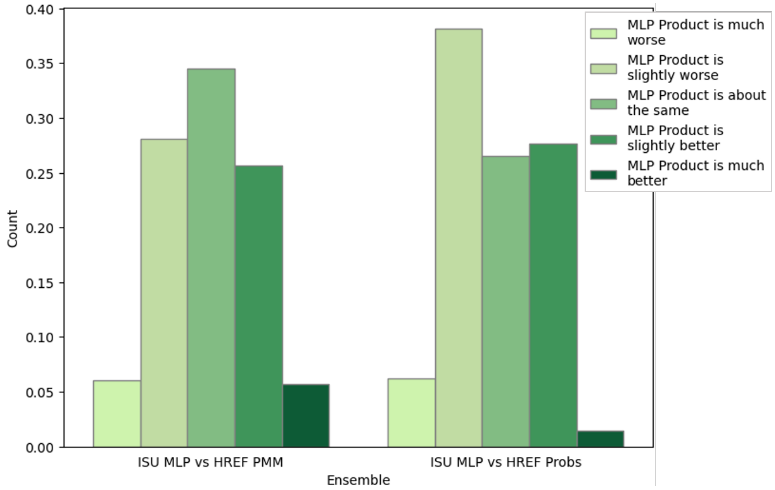


Figure 63: The normalized subjective results for the comparison of the ISU MLP to the HREF. Participants scores ranged from ISU MLP is much worse than the HREF to ISU MLP is much better than the HREF. On the left are the results comparing the ISU MLP PMM to the HREF LPMM and on the right are the results comparing the ISU MLP probabilities to the HREF probabilities.

As was shown in Section 3 Fig. 7, the comparison graphics between the HREF and the HREF MLP did not cover the same area, with the HREF products shown over the CONUS and the HREF MLP products shown only for the domain over which the ML models identified. This resulted in some difficulty for the participants during evaluation, since the “details” of the footprint of the area of concern (i.e. the ML identified domain) were easier to see on the domain-specific graphics compared to the CONUS graphics. The FFaIR team hopes to address this in FFaIR 2025 for the next round of HREF MLP evaluations by making domain specific graphics or changing the evaluation from a comparison to a question about the utility of the products. Additionally, at the FFaIR team’s recommendation, the ISU team created 10 km neighborhood probabilities rather than using 40 km, like the operational HREF output does. The thought behind this was that since the calculation was only over a small domain, rather then the CONUS, the detail from a small ROI/smoother would be beneficial in the evaluation. However, feedback from the participants showed that it was difficult to compare the HREF MLP probabilities to the HREF probabilities. Therefore, this will also be changed for next year’s evaluation in FFaIR.

In terms of the results themselves, the distribution of scores for the PMM/LPMM comparison was gaussian. Interestingly, whether the participants felt the the MLP improved upon the HREF forecast or degraded the forecast was not always dependent on if the MLP correctly identified an MCS. For example, for an event across the Great Lakes regions which had multiple rounds of convection and linear systems move through, (Fig. 64A-C) of the 17 participants, 10 if them said the HREF MLP PMM was slightly better than the HREF LPMM and 4 said the MLP was much better than the HREF product. The rest said the two products were about the same. Unfortunately, the verification survey did not include a space for written feedback, so it is unclear what aspect of the HREF MLP PMM they preferred over the HREF LPMM. Differing from this was the 25JUNE case; Fig. 64D-F. In this event, the HREF MLP incorrectly removed the signal for two lines of precipitation, removed the well-forecast line of precipitation across northern WI that was seen in the HREF LPMM, and shifted the bulk of the heavy rainfall erroneously too far southwest. For this case, 7 out of the 11 participants felt that the HREF MLP PMM was significantly worse than than the HREF LPMM; the remaining participants “chose about the same”.

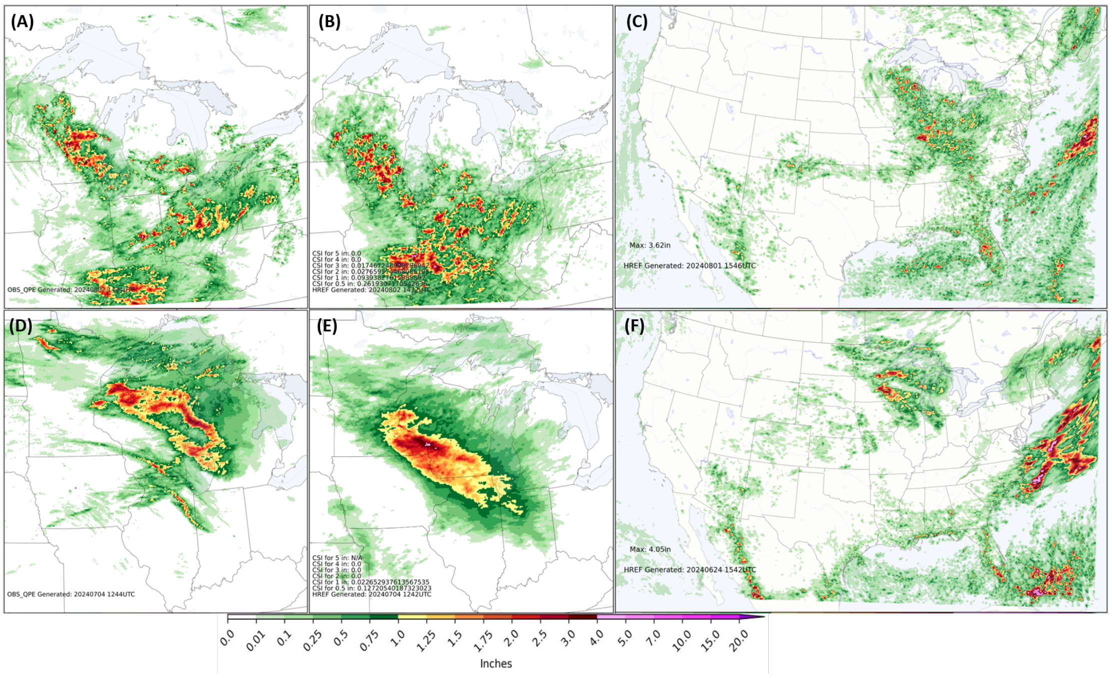

(A)
OBS_QPE Generated: 20240802 1434UTC
(B)
CSI for 5 in: 0.0
CSI for 4 in: 0.0
CSI for 3 in: 0.017467248908296942
CSI for 2 in: 0.027659574468085105
CSI for 1 in: 0.09393827615885897
CSI for 0.5 in: 0.26193071170542636
HREF Generated: 20240802 1432UTC
(C)
Max: 3.62in
HREF Generated: 20240801 1546UTC
(D)
OBS_QPE Generated: 20240704 1244UTC
(E)
CSI for 5 in: N/A
CSI for 4 in: 0.0
CSI for 3 in: 0.0
CSI for 2 in: 0.0
CSI for 1 in: 0.022652937613567535
CSI for 0.5 in: 0.12720540187323023
HREF Generated: 20240704 1242UTC
(F)
Max: 4.05in
HREF Generated: 20240624 1542UTC
0.0
0.01
0.1
0.25
0.5
0.75
1.0
1.25
1.5
1.75
2.0
2.5
3.0
4.0
5.0
7.0
10.0
15.0
20.0
Inches

Figure 64: The 6-h (A) and (D) MRMS QPE, (B) and (E) HREF MLP PMM and (C) and (F) HREF LPMM valid 12 UTC (A)-(C) 02 Aug and (D)-(F) 25 June 2025.

The distribution of scores for comparison of the probabilities (Fig. 63) was not gaussian and nearly 40% of the time participants felt that the HREF MLP probabilities were slightly worse than the HREF probabilities; ratings of ‘about the same” and “HREF MLP slightly better” had roughly the same chance of occurring. This suggests that participants felt the HREF MLP PMM was more beneficial over the HREF LPMM than the MLP probabilities were over the HREF probabilities. Figures 65 and 66 show the 1, 2, and 3 inch 24-h probabilities for the events discussed above. For the first event, despite the high praise for the MLP PMM, the results for the probabilities were mixed. 5 of the participants indicated they felt the HREF MLP probabilities were slightly better than the HREF’s but 2 felt the HREF MLP probabilities were significantly worst and 5 felt they were slightly worse than the HREF’s. For the latter case, only two participants felt that the MLP and HREF probabilities were of similar quality; the rest felt the MLP probabilities were slightly or significantly worse.

The end of the week comments on the HREF MLP products noted the aforementioned difficulties of comparing the HREF MLP to the HREF given the

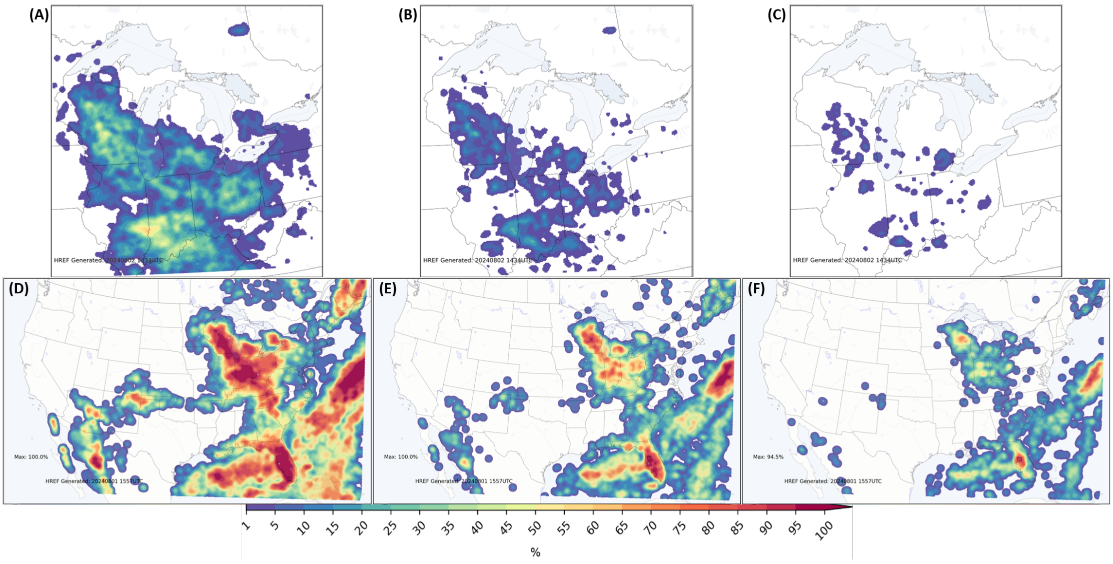


Figure 65: (A)-(C) the MLP 10 km neighborhood probabilities and (D)-(F) the HREF 40 km neighborhood probabilities, for 1" (left), 2" (middle) and 3" (right) in 24-h valid 12 UTC 02 Aug 2025. 24-h MRMS QPE can be found in Fig. 64A.

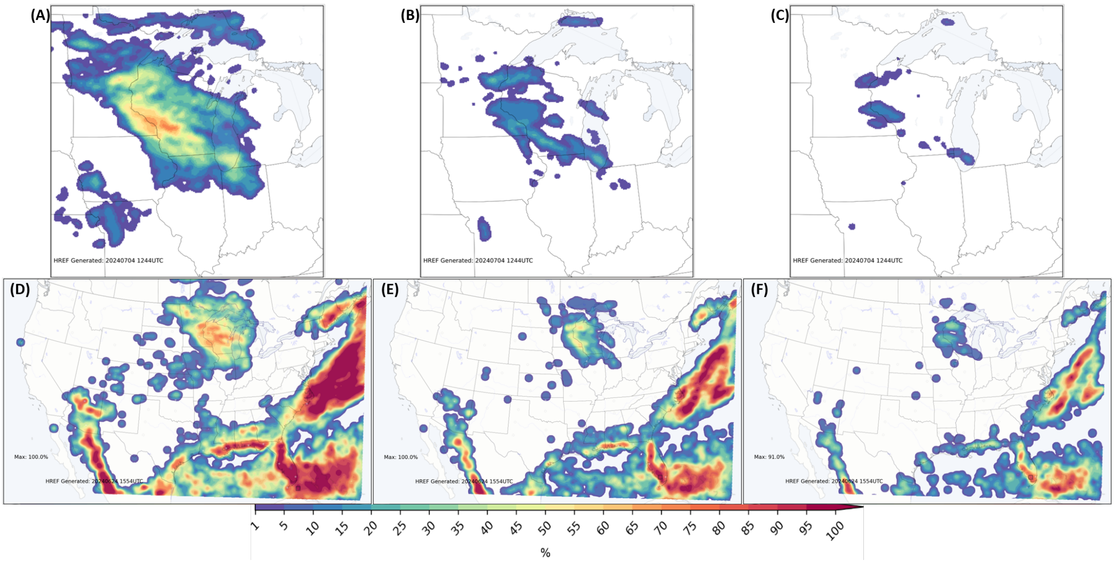


Figure 66: Like Fig. 65 but valid 12 25 June 2025 and the 24-h MRMS QPE can be found in Fig. 64D.

difference in how the graphics were generated; i.e. domain vs CONUS. Again, in the 2025 FFaIR experiment the FFaIR facilitators plan to address this. Apart from that, results varied on how participants felt the MLP HREF products did.

Regardless, participants appear to be in favor of a MLP tool like this (see comments below) and it is recommended that development continues.

> "It was another tool in the toolbox. It wasn't perfect, but something I took into consideration. I did rate it better than HREF during portions of the week."

> "It seemed hit or miss depending on the day. Neither seemed to be better overall in the short time period we looked at them. It is nice to see another data set though from an operational standpoint."

5.5 REFS Clusters

Ensemble clustering is a powerful post-processing method that condenses ensemble forecasts down to their prevalent scenarios, acting as a helpful tool to visualize and message forecast uncertainty. While clustering is applied extensively to global ensemble forecast data at WPC and through the Dynamic Ensemble-based Scenarios for IDSS platform (DESI), development of clustering applications for convection-allowing ensemble systems at WPC is beginning to be explored. For the 2024 FFaIR Experiment, ensemble clustering was applied to 6-h QPF from a 28-member time-lagged REFS to begin exploring clustering on convection-allowing ensemble systems.

Science goals for this experiment included:

- Understanding if any REFS QPF cluster scenarios were able to outperform the full ensemble forecasts according to subjective evaluation.
- Determining if any particular REFS QPF cluster scenario outperformed the full ensemble QPF fields most often according to subjective evaluation.
- Ensuring that REFS QPF cluster scenarios did not simply stratify the full ensemble membership by initialization time.

Analysis of cumulative subjective results showed that at least one cluster for 6-h 75th percentile QPF fields outperformed the full ensemble in 83.5% of participant evaluations, highlighting the method's utility for forecasting extreme rainfall (left pie chart in Fig. 67). Further examination of whether any particular cluster scenario outperformed the full ensemble system most often (right pie chat in Fig. 67) reveals that participants found Cluster 4 to provide the best representation of the forecast most often. This result was surprising given that the clusters are sorted by decreasing membership, meaning that Cluster 4 always had the smallest percentage of ensemble membership.

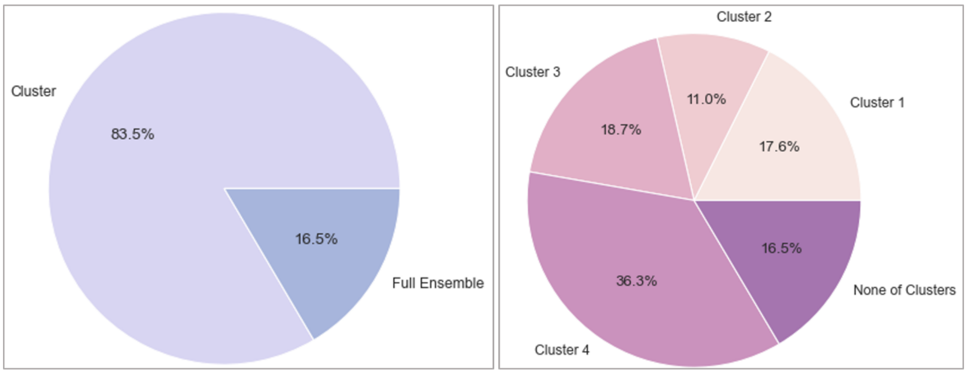


Figure 67: [LEFT] Analysis of cumulative participant responses to the question: "Which cluster performed better than the full membership for the 75th percentile?". Responses are aggregated over all clusters to understand how often any cluster subjectively outperformed the full ensemble forecast. [RIGHT] Analysis of cumulative participant responses to the question: "Which cluster performed better than the full membership for the 75th percentile?".

Two potential reasons for this result exist. Additional findings from 2024 FFaIR and the 2024 HWT SFE experiments suggest that the REFS exhibits signs of under dispersion within its forecasts. Assuming the 28-member REFS configuration used here is similarly under dispersive and the true solution lies outside of the ensemble envelope, and thus, the small membership clusters on the outskirts of the ensemble envelope may have a better opportunity to capture an outcome closer to observed. Alternatively, many participant comments suggested that they prefer the visualization of Cluster 4 to the other scenarios due to its more deterministic appearance, given that the cluster percentile calculations take place over far fewer members than with the other clusters. For example, one participant observed the following: "I think the finer resolution detail of cluster 4 bumps it ahead of the full ensemble, in my opinion. I want to see that higher resolution detail from a deterministic member and if it can be captured in a cluster of the ensemble, even better!" While the cluster forecast contains members of the same horizontal resolution as the rest of the ensemble system, the percentile calculations over such a small sample size create the illusion of higher resolution. An example of this phenomenon can be seen from 6-h QPF Clusters for the MRTP domain ending 0300 UTC July 11, 2024 (Fig. 68), where participants unanimously selected the 75th percentile QPF from Cluster 4 as the most similar to observed precipitation for that 6-h window. To definitively determine which hypothesis led to the subjective preference of Cluster 4 will need to be further explored in future work with the HMT.

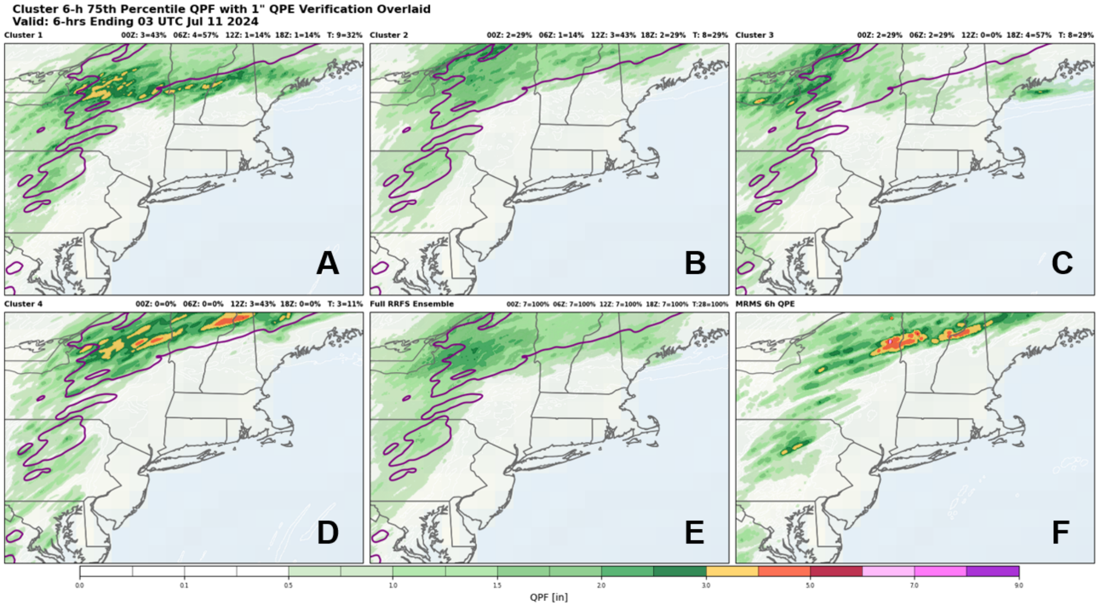

Cluster 6-h 75th Percentile QPF with 1" QPE Verification Overlaid
Valid: 6-hrs Ending 03 UTC Jul 11 2024
Cluster 1
00Z: 3=43% 06Z: 4=57% 12Z: 1=14% 18Z: 1=14% T: 9=32%
A
Cluster 2
00Z: 2=29% 06Z: 1=14% 12Z: 3=43% 18Z: 2=29% T: 8=29%
B
Cluster 3
00Z: 2=29% 06Z: 2=29% 12Z: 0=0% 18Z: 4=57% T: 8=29%
C
Cluster 4
00Z: 0=0% 06Z: 0=0% 12Z: 3=43% 18Z: 0=0% T: 3=11%
D
Full RRFS Ensemble
00Z: 7=100% 06Z: 7=100% 12Z: 7=100% 18Z: 7=100% T:28=100%
E
MRMS 6h QPE
F
0.0
0.1
0.5
1.0
1.5
2.0
3.0
5.0
7.0
9.0
QPF [in]

Figure 68: 6-h 75th Percentile QPF for each cluster (panels A - D), the full REFS (E), and MRMS 6-h QPE (F) valid between 2100 UTC July 10 and 0300 UTC July 11, 2024. The 1 inch MRMS 6-h QPE contour is overlaid on each cluster and full ensemble forecast for convenience. Note that the colorbar increments are slightly different from other figures herein which use the same colormap.

Finally, participants were asked the following about the intercluster spread: “Do you feel there was sufficient spread with respect to initialization time amongst the clusters or did the clustering approach simply group solutions by cycle?” Participants generally either commented that the clusters tended to have sufficient distribution of different initialization time membership among the clusters, or they stated that there were not enough cluster runs available to determine whether there was sufficient spread by cycle among the cluster scenarios. Further investigation would need to be conducted to conclude that the clustering algorithm is able to meaningfully group members by physical characteristics rather than simply by initialization time, but this result is encouraging for the validity of ensemble clustering within the REFS configuration, and for more broadly generating useful spread from a time-lagged REFS configuration.

6 Final Thoughts and Recommendations

6.1 RRFSp1/REFS Specific Findings

It is concerning that despite the advanced DA and cycling of the RRFSp1, all cycles but the 12z struggled in terms of performance compared to the operational models. Feedback across all times/domains often noted that the RRFSp1 did not perform well leading up to the event and then for the 12z cycle the forecast would dramatically improve. This is interesting given that 12z is when there is generally a lull in convection/precipitation and thus less impacts from the cycling in of radar observations. Participants also noted that this was concerning and would result in a decrease in their lead time in messaging events. Two quotes, one from daily verification and one from the weekly feedback, that summarize this are:

> "RRFS guidance in the first two cycles was awful; just awful. It rebounded quickly for the 06/12Z cycles. But that gives me some concern as an operational system. The HRRR had some errors at the beginning cycles but still looked reasonable given what occurred"

> "The RRFSp1 did have a couple fairly good forecasts (notably in the 12Z cycles) but seemingly struggled in the lead up to the event."

The RRFSp1 and the REFS members (RRFSm2-m6) tend to suppress convection, especially if convection initiation happens after the diurnal maximum (between 21 and 22 UTC). There were multiple instances (ex. 25JUNE) in which the operational models forecasted the development of an MCS or organized convection leading to heavy rainfall overnight but the RRFSp1 and the REFS members did not. These types of events already have low predictability but at least the operational models, specifically NAMnest, are good at highlighting the potential risk, even if they don't get the exact location correct. This is preferred by participants over the lack of signal/dry bias seen in the REFS. Differing from this, if a heavy rainfall event in the evening or overnight hours is driven by convection ongoing during the diurnal max, then the RRFSp1 tends to do well with event location but is likely to have a wet bias. The sharp drop off in precipitation after 21/22 UTC was most pronounced in RRFSp1 and RRFSm3 and m6; see Fig.37. Additionally, the suppression of convection from the CU Schemes is driving an abundance of erroneous light precipitation (<0.1"). Often embedded within the widespread light precipitation are small, isolated high precipitation popcorn storms, resulting in comments like these: "RRFS seems too isolated with heavier rain" and "The RRFSp1's bullseyes were often a little too small and it had unrealistically large areas of light forecast precip." This characteristic was seen in nearly all types of weather patterns, including general thunderstorm days in the southeast, frontal convection, and MCSs.

Generally, the RRFSp1 and RRFSm2-m6 have a dry bias at lower thresholds (missing events) but a wet bias at higher thresholds (when it forecasts it over forecasts magnitude) for 24-h QPF across the CONUS. The REFS membership was also more clustered among itself than the HREF membership was. When separated into a 6-h QPF analysis, there is a noticeable jump in the REFS membership's 12z cycle's performance, especially for the 12-18 UTC time period, where the membership jumps from a slight dry bias to a large wet bias (Fig. 35). When focusing on the heavy rainfall events themselves (i.e. the MRTP/MRTP+ dataset), the clustering of the REFS members and HRRR was pronounced and continually had a dry bias. The NAMnest typically had a higher POD than the other models, especially at longer lead times. REFS members can perform at a high level for bigger/predictable events (top 10-20%) but are inconsistent on a regular basis when using CSI as the metric (see Figs. 39 and 40).

Despite the aforementioned REFS members' dry bias and struggles to develop convection after the diurnal max, some members can still generate hourly QPF extremes, especially for the 00z cycle. The members that generate extreme hourly maxima vs those that do not (Fig. 36) are tied to the type of microphysics scheme used. RRFSp1, RRFSm2 and m3 all use the Thompson Scheme and over forecast the occurrence of hourly totals beginning around 1.5", with instances of hourly

QPF $\geq$10" from all of them. For the RRFSp1 specifically, although it over forecasts instances of hourly totals, its hourly average QPF and the hourly coverage of 1" across the diurnal cycle is under forecasted, most notably after its forecasted diurnal maximum. This suggests that the RRFSp1 is missing events (especially after 21 UTC) and implies that for the events it does forecast, it is over forecasting magnitude. This over forecasting of the magnitude most often occurs leading up to and around 21 UTC, given the sharp drop off in the 1" coverage and hourly average after this time. Of all the members, the RRFSm3, which differs from RRFSp1 only in the CU parm used, performed the most similar to RRFSp1.

In regards to the REFS, similar to what was seen with the evaluation of the individual members, a decrease in perceived performance was noted across the diurnal maximum. After the diurnal max (00-06 UTC and 06-12 UTC), for both the 1"/6-h and 5"/6-h neighborhood probabilities, the HREF was more likely to have participants identify the magnitudes as "good" compared to the REFS, while the REFS was more likely than the HREF to have participants feel the magnitudes were "too low". Prior to the diurnal max (12-18 UTC and 18-00 UTC), the REFS was more likely to rate more highly than the HREF. For the 12z cycle valid 12-18 UTC, for 5"/6-h the REFS was less likely than the HREF to be considered "good", with the REFS having a greater chance of the magnitudes being "too high" than the HREF; this perception is supported by the high bias seen in the membership for this time period and cycle. Additionally for the 5"/6-h subjective evaluation during the 18-00 UTC valid period, the most likely choice to be picked by participants, regardless of cycle, for the REFS was "too high"; this was not the case for HREF, which had relatively equal chances of any given choice being selected by participants. Combined, this further supports the analysis that the REFS as a whole over forecasts QPF magnitudes in the lead up to the diurnal max and under forecasts after the max. For the MRTP domains, the REFS PMM saw more spread in its goodness scores than the HREF, receiving more "poor" scores (1) than the HREF but also more "somewhat good" (5) and "good" (7) scores than the HREF. This perhaps alludes to an inconsistency in forecast goodness, where a forecast could be either really good or really bad; example Figs. 51 and 50 compared to Fig. 53.

Reviewing 24-h performance diagrams, for all but the 2” threshold for the 12z cycle, the HREF means have a higher(lower) POD(SR) than those from the REFS. This results in similar CSIs between each ensemble’s corresponding mean. Again, the exception is the 2” threshold for the 12z cycle, where the REFS means have the same POD as the HREF means but their SR is roughly greater by 0.1. Statistically this seems promising but the PMM, which is designed in a way that typically results in a high bias, has a low bias until the 2” threshold where the bias is 1.25. For comparison the HREF PMM bias is between 2 and 2.5. This could suggest that REFS membership struggles to either product high 24-h totals or that it’s underdispersive. Also, given the method that forecasters employ the PMM for, a change from an expected high bias product could be worrisome, given that they use the PMM to determine a “reasonable” worst case scenario.

Finally, the end of the week feedback collected from the participants to gauge their feelings on the performance of the RRFSp1 and the REFS as a whole (Figs. 28 and 29 and Fig. 57) suggest that the participants still prefer the HRRR and HREF over the RRFSp1 and REFS. The NAMnest results are more mixed, but there is concern that low POD in the HRRR and RRFSp1 that is not seen in the NAMnest will result in missing events if the NAMnest was retired. To fully understand how the participants felt between the two systems, the participants were asked: “How do you feel forecast quality will change if the HREF suite (HRRR, NAMnest, ARW, ARW2, and FV3-HREF) is retired and replaced with the REFS suite (RRFSp1, RRFSm2-5, and HRRR)?”. The results can be seen in Fig. 69. 58.3% of the participants indicated that they felt forecast value would be slightly degraded, with 4.2% indicating they felt it would be severely degraded. No one indicated they felt the forecast value would be significantly increased, thought 8% did note they felt the value would slightly increase.

Participants provided written feedback to this question as well. All of the written responses can be found in Appendix B but the following two quotes seem to summarize the results from all the subjective feedback:

> “The REFS suite noticeably struggled at certain points of the week, particularly in some of the intense (retro) cases that were ran throughout the week. It was of note that the current model suite handled these events slightly

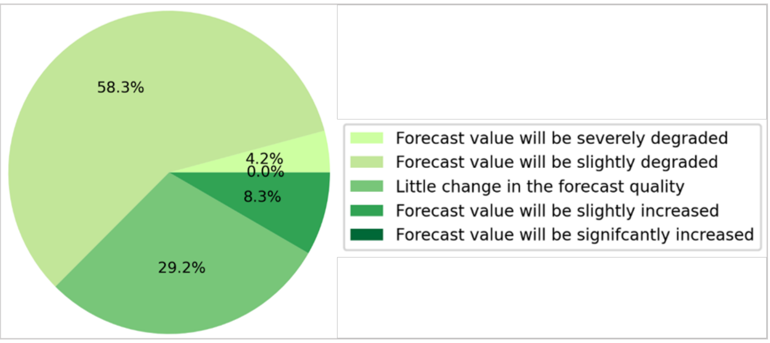


Figure 69: Results from the end of week survey for the questions: "How do you feel forecast quality will change if the HREF suite (HRRR, NAMnest, ARW, ARW2, and FV3-HREF) is retired and replaced with the REFS suite (RRFSp1, RRFSm2-5, and HRRR)?"

> better even with some uncertainty in regards to the forecast. The REFS suite was not completely the worst throughout the week and did have some events that it forecasted better with the 12Z runs consistently having slightly better skill before potentially tapering back rainfall amounts/moving location or footprint orientation in later model cycles."

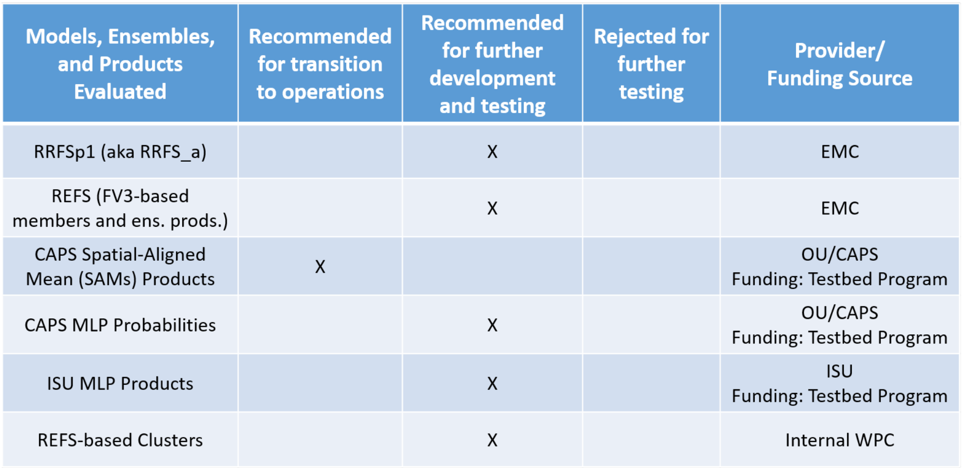
> "This is hard to know in advance. I do think the REFS suite will be useful in many situations. However, the HREF does seem to have a larger spread in outcomes (my subjective opinion) across more marginal areas of extreme precipitation, which I think would be taken away by replacing the HREF."

6.2 Recommendations

Table 8 shows the transition recommendations for the funded and unfunded products, models and tools that were evaluated in the 2024 FFaIR Experiment. Only one of these is recommended for transition to operations, the CAPS spatially-aligned mean (aka the SAM and SAM-LPM) method. This was also tested in the 2023 FFaIR; the SAM-LPM had more positive feedback than the SAM but this year the SAM was preferred over the SAM-LPM. Different types of events and the makeup of participants likely resulted in the difference of which of the two methods were preferred but either way, participants from both years felt the SAM method added value from just the simple mean and the LPMM and they would like to see the SAMs operational.

The RRFSp1, REFS (membership and ensemble products), REFS Clusters, ISU HREF MLP and CAPS MLP are all recommended for further testing. The

Table 8: Table of recommendations for transition to operations for the evaluated funded and unfunded products, models and tools during the 2024 FFaIR Experiment.



Models, Ensembles, and Products Evaluated	Recommended for transition to operations	Recommended for further development and testing	Rejected for further testing	Provider/ Funding Source
RRFSp1 (aka RRFS_a)		X		EMC
REFS (FV3-based members and ens. prods.)		X		EMC
CAPS Spatial-Aligned Mean (SAMs) Products	X			OU/CAPS Funding: Testbed Program
CAPS MLP Probabilities		X		OU/CAPS Funding: Testbed Program
ISU MLP Products		X		ISU Funding: Testbed Program
REFS-based Clusters		X		Internal WPC

above subsection summarized the shortcomings in the RRFSp1 and REFS and why they are not recommended for transition. Because of this, the REFS Clusters can not be recommended, though the method itself showed utility and participants generally preferred one of the Cluster's $75th$ percentile forecast to the full ensemble's. The ISU HREF MLPs for PMM and probabilities showed promise but were hard to properly compare to the HREF due to the setup of the verification questions. This will be addressed for the 2025 FFaIR Experiment. That said, the ISU MLP, which attempts to bias-correct location errors for MCS and MCS-like events, did not always add value over the HREF EMC generated products for such events (ex. 25JUNE event). However, there were instances in which it added value to events that were not classic MCS events (ex. forecast valid 12 UTC 02 August 2025). Finally for the CAPS MLP, the FFaIR team recognizes that the evaluation period was hindered due to compute issues but when it was available to evaluate, it severely under performed compared to the two ensembles that the MLP membership is take from, the CAPSe and the HREF. There were instances in which both parent ensembles had 1"/6-h probabilities exceeding 80% and the MLP probabilities were 0. It is the FFaIR team's advice that the methodology for this MLP be changed.

References

Clark, C. J., and Coauthors, 2024: Spring forecasting experiment 2024 conducted by the experimental forecast program of the noaa hazardous weather testbed preliminary findings and results. Tech. rep., NCEP SPC-HWT. URL https://hwt.nssl.noaa.gov/sfe/2024/docs/SFE2024_tech_memo.pdf.

James, E. P., and Coauthors, 2022: The high-resolution rapid refresh (hrrr): An hourly updating convection-allowing forecast model. part ii: Forecast performance. *Wea. Forecasting*, **37**, 1397–1417, https://doi.org/https://doi.org/10.1175/WAF-D-21-0130.1.

KSDK-News, 2024: 'this was like a raging river': Homes in nashville, illinois, evacuated after dam failure. URL https://www.ksdk.com/article/news/local/homes-in-nashville-illinois-being-evacuated-as-dam-failure-deemed-imminent/63-2100b38c-2c59-447b-a00a-6a2c430bd7a0.

NBC-News, 2024: Record flooding inundates northwest iowa, prompts evacuations, isolates one city. URL https://www.nbcnews.com/news/weather/record-flooding-inundates-northwest-iowa-prompts-evacuations-isolates-rcna158513.

NWS-Miami, 2024: June 11-13th, 2024: Rainfall & urban flooding event. URL https://storymaps.arcgis.com/stories/9a696025541145b3a06662a65574c301, this is an ArcGIS Story Map.

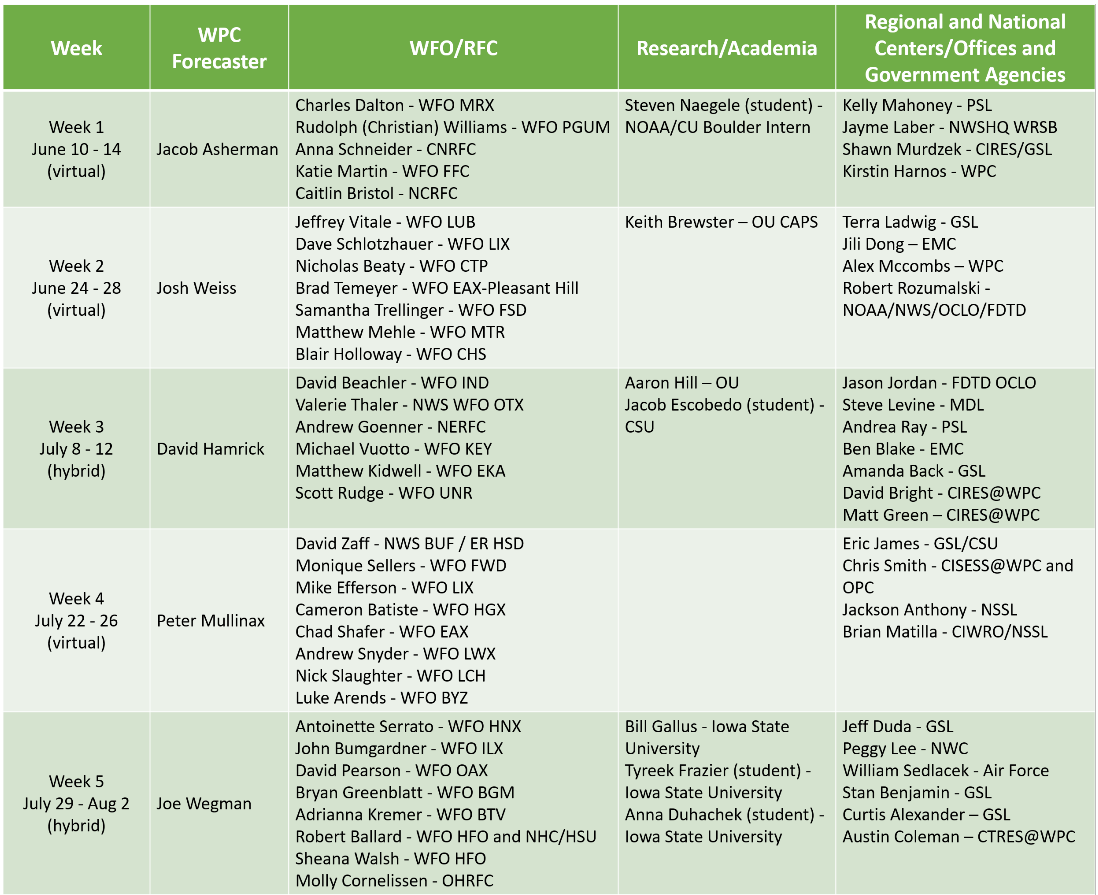
Trojniak, S., and J. Correia, Jr., 2024: 2024 ffair operations plan, published online at https://www.wpc.ncep.noaa.gov/hmt/Reports/FFaIR/2024FFaIR_OpsPlan.pdf. If missing please contact WPC.

Trojniak, S., J. Correia, Jr., and W. M. Bartolini, 2023: 2023 flash flood and intense rainfall (ffair) final report: Part 1 - rrfs related results and findings. Tech. rep. Published online at https://www.wpc.ncep.noaa.gov/hmt/Reports/FFaIR/2023_FFaIR_Final_Report_Part1.pdf. If missing please contact WPC.

Appendices

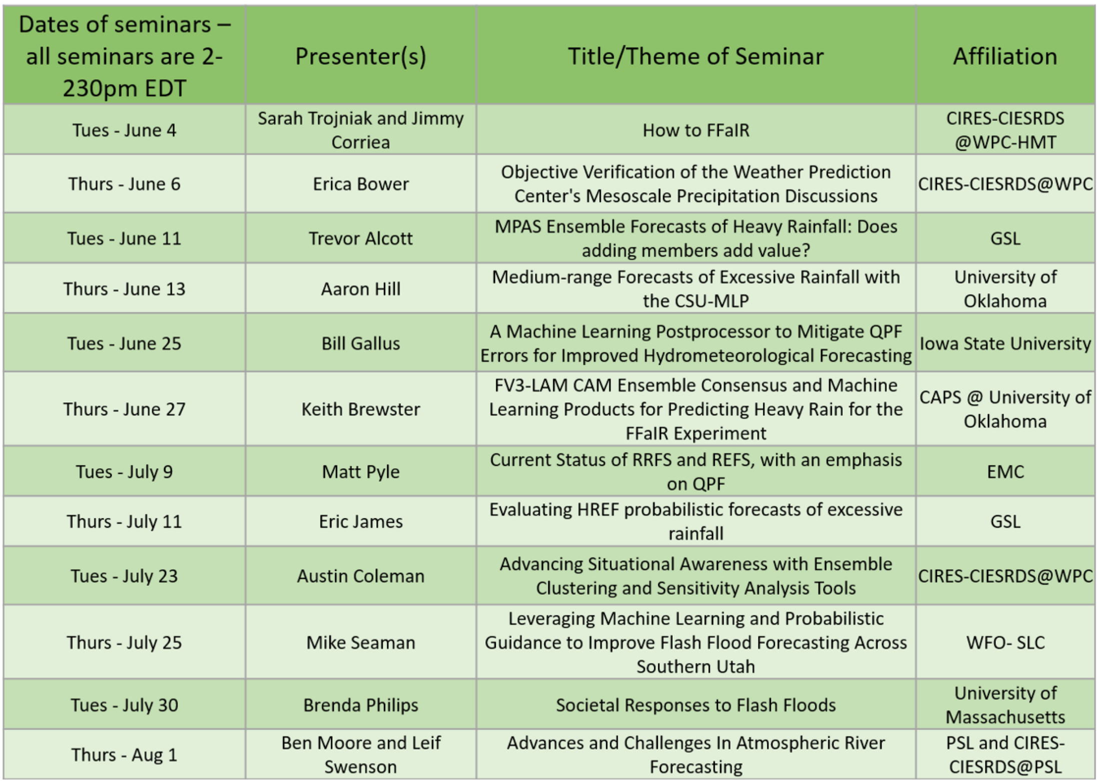
A List of Participants and Seminars

Table 9: List of the participants for each week of the 2024 FFaIR Experiment.



Week	WPC Forecaster	WFO/RFC	Research/Academia	Regional and National Centers/Offices and Government Agencies
Week 1 June 10 - 14 (virtual)	Jacob Asherman	Charles Dalton - WFO MRX Rudolph (Christian) Williams - WFO PGUM Anna Schneider - CNRFC Katie Martin - WFO FFC Caitlin Bristol - NCRFC	Steven Naegele (student) - NOAA/CU Boulder Intern	Kelly Mahoney - PSL Jayme Laber - NWSHQ WRSB Shawn Murdzek - CIRES/GSL Kirstin Harnos - WPC
Week 2 June 24 - 28 (virtual)	Josh Weiss	Jeffrey Vitale - WFO LUB Dave Schlotzhauer - WFO LIX Nicholas Beaty - WFO CTP Brad Temeyer - WFO EAX-Pleasant Hill Samantha Trellinger - WFO FSD Matthew Mehle - WFO MTR Blair Holloway - WFO CHS	Keith Brewster – OU CAPS	Terra Ladwig - GSL Jili Dong – EMC Alex Mccombs – WPC Robert Rozumalski - NOAA/NWS/OCLO/FDTD
Week 3 July 8 - 12 (hybrid)	David Hamrick	David Beachler - WFO IND Valerie Thaler - NWS WFO OTX Andrew Goenner - NERFC Michael Vuotto - WFO KEY Matthew Kidwell - WFO EKA Scott Rudge - WFO UNR	Aaron Hill – OU Jacob Escobedo (student) - CSU	Jason Jordan - FDTD OCLO Steve Levine - MDL Andrea Ray - PSL Ben Blake - EMC Amanda Back - GSL David Bright - CIRES@WPC Matt Green – CIRES@WPC
Week 4 July 22 - 26 (virtual)	Peter Mullinax	David Zaff - NWS BUF / ER HSD Monique Sellers - WFO FWD Mike Efferson - WFO LIX Cameron Batiste - WFO HGX Chad Shafer - WFO EAX Andrew Snyder - WFO LWX Nick Slaughter - WFO LCH Luke Arends - WFO BYZ		Eric James - GSL/CSU Chris Smith - CISESS@WPC and OPC Jackson Anthony - NSSL Brian Matilla - CIWRO/NSSL
Week 5 July 29 - Aug 2 (hybrid)	Joe Wegman	Antoinette Serrato - WFO HNX John Bumgardner - WFO ILX David Pearson - WFO OAX Bryan Greenblatt - WFO BGM Adrianna Kremer - WFO BTV Robert Ballard - WFO HFO and NHC/HSU Sheana Walsh - WFO HFO Molly Cornelissen - OHRFC	Bill Gallus - Iowa State University Tyreek Frazier (student) - Iowa State University Anna Duhachek (student) - Iowa State University	Jeff Duda - GSL Peggy Lee - NWC William Sedlacek - Air Force Stan Benjamin - GSL Curtis Alexander – GSL Austin Coleman – CTRES@WPC

Table 10: List of the 2024 FFaIR Science Seminars. The slides for the seminars can be found here.



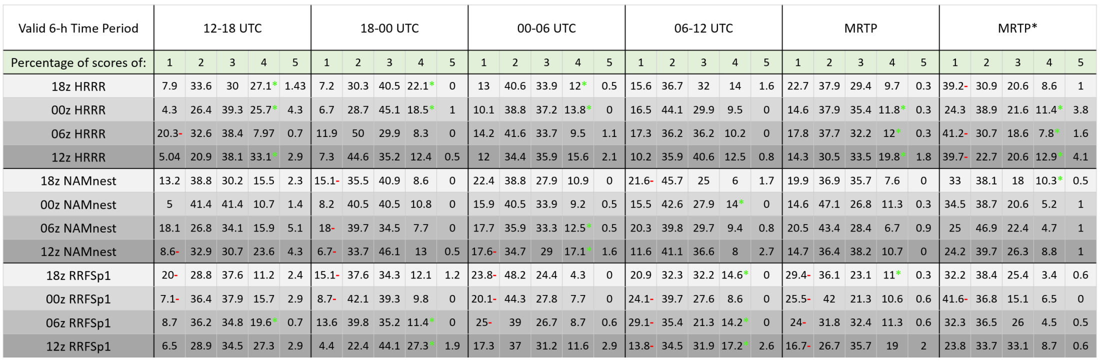
Dates of seminars – all seminars are 2-230pm EDT	Presenter(s)	Title/Theme of Seminar	Affiliation
Tues - June 4	Sarah Trojniak and Jimmy Corriea	How to FFaIR	CIRES-CIESRDS @WPC-HMT
Thurs - June 6	Erica Bower	Objective Verification of the Weather Prediction Center's Mesoscale Precipitation Discussions	CIRES-CIESRDS@WPC
Tues - June 11	Trevor Alcott	MPAS Ensemble Forecasts of Heavy Rainfall: Does adding members add value?	GSL
Thurs - June 13	Aaron Hill	Medium-range Forecasts of Excessive Rainfall with the CSU-MLP	University of Oklahoma
Tues - June 25	Bill Gallus	A Machine Learning Postprocessor to Mitigate QPF Errors for Improved Hydrometeorological Forecasting	Iowa State University
Thurs - June 27	Keith Brewster	FV3-LAM CAM Ensemble Consensus and Machine Learning Products for Predicting Heavy Rain for the FFaIR Experiment	CAPS @ University of Oklahoma
Tues - July 9	Matt Pyle	Current Status of RRFS and REFS, with an emphasis on QPF	EMC
Thurs - July 11	Eric James	Evaluating HREF probabilistic forecasts of excessive rainfall	GSL
Tues - July 23	Austin Coleman	Advancing Situational Awareness with Ensemble Clustering and Sensitivity Analysis Tools	CIRES-CIESRDS@WPC
Thurs - July 25	Mike Seaman	Leveraging Machine Learning and Probabilistic Guidance to Improve Flash Flood Forecasting Across Southern Utah	WFO- SLC
Tues - July 30	Brenda Philips	Societal Responses to Flash Floods	University of Massachusetts
Thurs - Aug 1	Ben Moore and Leif Swenson	Advances and Challenges In Atmospheric River Forecasting	PSL and CIRES-CIESRDS@PSL

B Supplement Information

In this appendix you will find additional charts, graphics and written feedback from the participants that supplement the Final Report and provide additional information when needed.

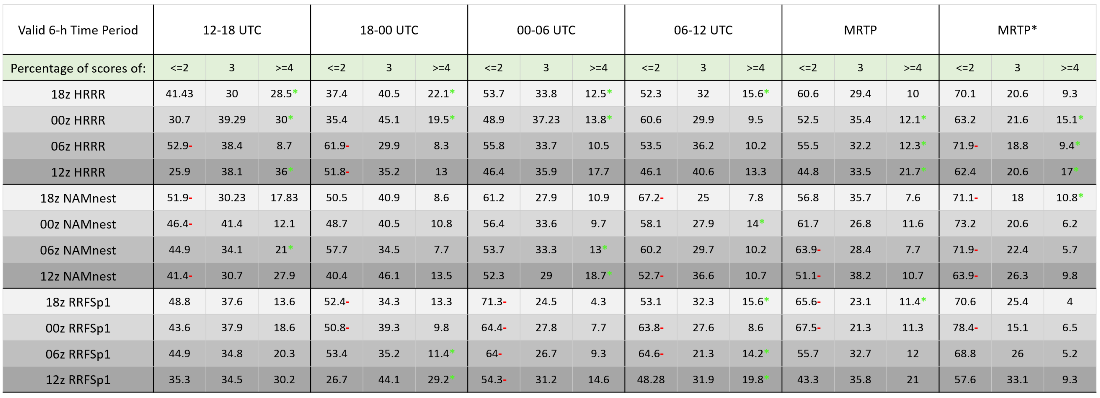
B.1 Charts and Graphs

Table 11: Table showing percentage of times that the HRRR, NAMnest, and RRFSp1 model initializations received a score of 1 (very poor) to 5 (very good) during the subjective evaluation for the 6-h time periods evaluated for the CONUS deterministic and the MRTP and MRTP∗ domains. The highest percentage of 4s (a 5 was rare so 4s are highlighted) for a given cycle for the time period/domain has a green ∗ while highest percentage of 1s has a red -; the differences between the lowest/highest scores are not necessarily statically significant.



Valid 6-h Time Period	12-18 UTC					18-00 UTC					00-06 UTC					06-12 UTC					MRTP					MRTP*				
Percentage of scores of:	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
18z HRRR	7.9	33.6	30	27.1*	1.43	7.2	30.3	40.5	22.1*	0	13	40.6	33.9	12*	0.5	15.6	36.7	32	14	1.6	22.7	37.9	29.4	9.7	0.3	39.2-	30.9	20.6	8.6	1
00z HRRR	4.3	26.4	39.3	25.7*	4.3	6.7	28.7	45.1	18.5*	1	10.1	38.8	37.2	13.8*	0	16.5	44.1	29.9	9.5	0	14.6	37.9	35.4	11.8*	0.3	24.3	38.9	21.6	11.4*	3.8
06z HRRR	20.3-	32.6	38.4	7.97	0.7	11.9	50	29.9	8.3	0	14.2	41.6	33.7	9.5	1.1	17.3	36.2	36.2	10.2	0	17.8	37.7	32.2	12*	0.3	41.2-	30.7	18.6	7.8*	1.6
12z HRRR	5.04	20.9	38.1	33.1*	2.9	7.3	44.6	35.2	12.4	0.5	12	34.4	35.9	15.6	2.1	10.2	35.9	40.6	12.5	0.8	14.3	30.5	33.5	19.8*	1.8	39.7-	22.7	20.6	12.9*	4.1
18z NAMnest	13.2	38.8	30.2	15.5	2.3	15.1-	35.5	40.9	8.6	0	22.4	38.8	27.9	10.9	0	21.6-	45.7	25	6	1.7	19.9	36.9	35.7	7.6	0	33	38.1	18	10.3*	0.5
00z NAMnest	5	41.4	41.4	10.7	1.4	8.2	40.5	40.5	10.8	0	15.9	40.5	33.9	9.2	0.5	15.5	42.6	27.9	14*	0	14.6	47.1	26.8	11.3	0.3	34.5	38.7	20.6	5.2	1
06z NAMnest	18.1	26.8	34.1	15.9	5.1	18-	39.7	34.5	7.7	0	17.7	35.9	33.3	12.5*	0.5	20.3	39.8	29.7	9.4	0.8	20.5	43.4	28.4	6.7	0.9	25	46.9	22.4	4.7	1
12z NAMnest	8.6-	32.9	30.7	23.6	4.3	6.7-	33.7	46.1	13	0.5	17.6-	34.7	29	17.1*	1.6	11.6	41.1	36.6	8	2.7	14.7	36.4	38.2	10.7	0	24.2	39.7	26.3	8.8	1
18z RRFSp1	20-	28.8	37.6	11.2	2.4	15.1-	37.6	34.3	12.1	1.2	23.8-	48.2	24.4	4.3	0	20.9	32.3	32.2	14.6*	0	29.4-	36.1	23.1	11*	0.3	32.2	38.4	25.4	3.4	0.6
00z RRFSp1	7.1-	36.4	37.9	15.7	2.9	8.7-	42.1	39.3	9.8	0	20.1-	44.3	27.8	7.7	0	24.1-	39.7	27.6	8.6	0	25.5-	42	21.3	10.6	0.6	41.6-	36.8	15.1	6.5	0
06z RRFSp1	8.7	36.2	34.8	19.6*	0.7	13.6	39.8	35.2	11.4*	0	25-	39	26.7	8.7	0.6	29.1-	35.4	21.3	14.2*	0	24-	31.8	32.4	11.3	0.6	32.3	36.5	26	4.5	0.5
12z RRFSp1	6.5	28.9	34.5	27.3	2.9	4.4	22.4	44.1	27.3*	1.9	17.3	37	31.2	11.6	2.9	13.8-	34.5	31.9	17.2*	2.6	16.7-	26.7	35.7	19	2	23.8	33.7	33.1	8.7	0.6

Table 12: Like Table 11 but the percentage of 1s (very poor) and 2s (poo) are combined together and 4s (good) and 5s (very good) combined.



Valid 6-h Time Period	12-18 UTC			18-00 UTC			00-06 UTC			06-12 UTC			MRTP			MRTP*		
Percentage of scores of:	<=2	3	>=4	<=2	3	>=4	<=2	3	>=4	<=2	3	>=4	<=2	3	>=4	<=2	3	>=4
18z HRRR	41.43	30	28.5*	37.4	40.5	22.1*	53.7	33.8	12.5*	52.3	32	15.6*	60.6	29.4	10	70.1	20.6	9.3
00z HRRR	30.7	39.29	30*	35.4	45.1	19.5*	48.9	37.23	13.8*	60.6	29.9	9.5	52.5	35.4	12.1*	63.2	21.6	15.1*
06z HRRR	52.9-	38.4	8.7	61.9-	29.9	8.3	55.8	33.7	10.5	53.5	36.2	10.2	55.5	32.2	12.3*	71.9-	18.8	9.4*
12z HRRR	25.9	38.1	36*	51.8-	35.2	13	46.4	35.9	17.7	46.1	40.6	13.3	44.8	33.5	21.7*	62.4	20.6	17*
18z NAMnest	51.9-	30.23	17.83	50.5	40.9	8.6	61.2	27.9	10.9	67.2-	25	7.8	56.8	35.7	7.6	71.1-	18	10.8*
00z NAMnest	46.4-	41.4	12.1	48.7	40.5	10.8	56.4	33.6	9.7	58.1	27.9	14*	61.7	26.8	11.6	73.2	20.6	6.2
06z NAMnest	44.9	34.1	21*	57.7	34.5	7.7	53.7	33.3	13*	60.2	29.7	10.2	63.9-	28.4	7.7	71.9-	22.4	5.7
12z NAMnest	41.4-	30.7	27.9	40.4	46.1	13.5	52.3	29	18.7*	52.7-	36.6	10.7	51.1-	38.2	10.7	63.9-	26.3	9.8
18z RRFSp1	48.8	37.6	13.6	52.4-	34.3	13.3	71.3-	24.5	4.3	53.1	32.3	15.6*	65.6-	23.1	11.4*	70.6	25.4	4
00z RRFSp1	43.6	37.9	18.6	50.8-	39.3	9.8	64.4-	27.8	7.7	63.8-	27.6	8.6	67.5-	21.3	11.3	78.4-	15.1	6.5
06z RRFSp1	44.9	34.8	20.3	53.4	35.2	11.4*	64-	26.7	9.3	64.6-	21.3	14.2*	55.7	32.7	12	68.8	26	5.2
12z RRFSp1	35.3	34.5	30.2	26.7	44.1	29.2*	54.3-	31.2	14.6	48.28	31.9	19.8*	43.3	35.8	21	57.6	33.1	9.3

B.2 Written Feedback

Below are all of the written responses for the question "How do you feel forecast quality will change if the HREF suite (HRRR, NAMnest, ARW, ARW2, and FV3-HREF) is retired and replaced with the REFS suite (RRFSp1, RRFSm2-5, and HRRR)?"

- "We don't use the HREF much when doing QPF in CA, but from these evaluations not sure the REFS is ready to replace the HREF."
- "It's only one week so it's difficult to gage if the models were in a rut or not, but the performance around the board was generally lacking outside of a few runs here and there."
- "The REFS suite noticeably struggled at certain points of the week, particularly in some of the intense (retro) cases that were ran throughout the week. It was of note that the current model suite handled these events slightly better even with some uncertainty in regards to the forecast. The REFS suite was not completely the worst throughout the week and did have some events that it forecasted better with the 12Z runs consistently having slightly better skill before potentially tapering back rainfall amounts/moving location or footprint orientation in later model cycles."
- "I don't feel the RRFS is ready for real time yet, as once it becomes operational, the code will be locked. It does seem like the over abundance of low level moisture problem has been fixed, but now fails to develop convection at times when convection develops.
- "The NAMnest contributes the least, but occasionally does get lucky."
- "This is hard to know in advance. I do think the REFS suite will be useful in many situations. However, the HREF does seem to have a larger spread in outcomes (my subjective opinion) across more marginal areas of extreme precipitation, which I think would be taken away by replacing the HREF."
- "It seems like the NAM nest captures some patterns better than most of the other models, but in general it doesn't seem like it will degrade it too much."

- "Losing the NAM nest would be tough based on the model performance I observed this week. Further improvements should be made to the RRFS system before it is implemented."

- "I think overall, the forecast quality will remain unchanged or slightly changed to the forecast value being increased. I think there was only the one event where the RRFS missed an event whereas the HREF had multiple hiccups throughout the week. From what I remember I think the REFS did well most of the week except for the one day."

- "I found the NAMnest to be especially helpful, so it would be a loss to lose it."

- "Right - we come to the key issue - can RRFSp1 replace NAMnest or not? It (RRFSp1) may be slightly weaker from my very limited exposure this week. HREF would be somewhat stronger if FV3-HREF was eliminated."

- "Seems like a "grand" ensemble (HREF and REFS combined) might be a better idea. Hybrid forecasting systems always seem to perform better than individual forecast systems."

- "It's tough to answer this question after just one week, but I think it'll just take some getting used to. I'm accustomed to the HREF ensemble, while I'm less familiar with the REFS at this point."

- "I thought on the whole the REFS performed slightly worse than the HREF for our cases, but in actuality there were plenty examples where all the modeling systems were awful. I think in day to day usage, without the HREF available, one wouldn't notice a difference. However, I don't think it will be a noticeable improvement either."

- "I think the REFS suite still needs a little more work before the HREF suite is retired. While the HRRR is still part of that group, I do think the NAMnest did outperform the RRFSp1 on occasion so you would be losing some of that."

- "I think it varies considerably. As mentioned in a previous question, it did well on one of the days, but was similarly variable and wrong as the HREF group...therefore really difficult to say. It certainly wasn't any better than the HREF suite on the whole."

C MET-MODE Settings for Objective Verification

Below is the MODE Configuration used.

- Both QPF and MRMS-GC QPE re-gridded to a common 5km lat/lon grid
- 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