2023 Flash Flood and Intense Rainfall (FFaIR) Final Report: *Part 1 - RRFS Related Results and Findings*

June 5 - August 11, 2023
Weather Prediction Center (WPC)
Hydrometeorology Testbed (HMT)

Sarah Trojniak[1], **James Correia Jr.**[1], and **W. Massey Bartolini**[1]

[1]CIRES-CIESDRS CU Boulder, NOAA/NWS/WPC/HMT

Contents

1 Introduction

The Flash Flood and Intense Rainfall (FFaIR) Experiment has been held annually since 2012 and is part of the Hydrometeorology Testbed (HMT) at the Weather Prediction Center (WPC). FFaIR sits at the intersection of research and operations (R2O), bringing together developers, forecasters, and researchers in a pseudo-operational setting to use and evaluate new products and tools as they relate to flash flood and intense rainfall forecasting.

The FFaIR Experiment's 11th season was full of firsts. Understanding the need to capture a plethora of heavy rainfall events across varying spatial and temporal timescales and meteorological forcing, the team decided to extend the length of the experiment. Rather than the traditional 4 weeks, with 2 weeks on either side of the week of July 4th, FFaIR was in session for 6 weeks, spanning from 5 June to 11 August 2023. This allowed for more subjective analysis of the tools and products as well as a longer time period to objectively evaluate the data. It was also the first time since the summer of 2019 that FFaIR had participants in-person at WPC. However, since it was not feasible to have 6 weeks of in-person sessions, a mix of completely virtual and hybrid (another first) sessions were held; with 2 weeks of hybrid and 4 weeks of virtual.

FFaIR was in session for the following weeks:

Week 1: June 5 - 9 (virtual)
Week 2: June 12 - 16 (virtual)
Week 3: June 26 - 30 (hybrid)
Week 4: July 10 - 14 (virtual)
Week 5: July 31 - Aug 4 (hybrid)
Week 6: Aug 7 - 11 (virtual)

FFaIR had approximately 90 participants this year. Participants were mostly from across the National Weather Service, coming from Weather Forecast Offices (WFOs) and National Centers and Labs, with participation also from teams at Colorado State University (CSU). Figure 1 shows the WFOs that participated in FFaIR this year while a full list of participants and their affiliations can be found in Appendix A. A seminar series was also hosted by FFaIR. Speakers ranged from forecasters to academia to model developers and were generally held on the Tuesdays and Thursdays that FFaIR was in session. The seminars were open to all of the NWS and our partners, roughly averaging 40 people per seminar. A list of the presenters and their topics can be found in Appendix A as well.

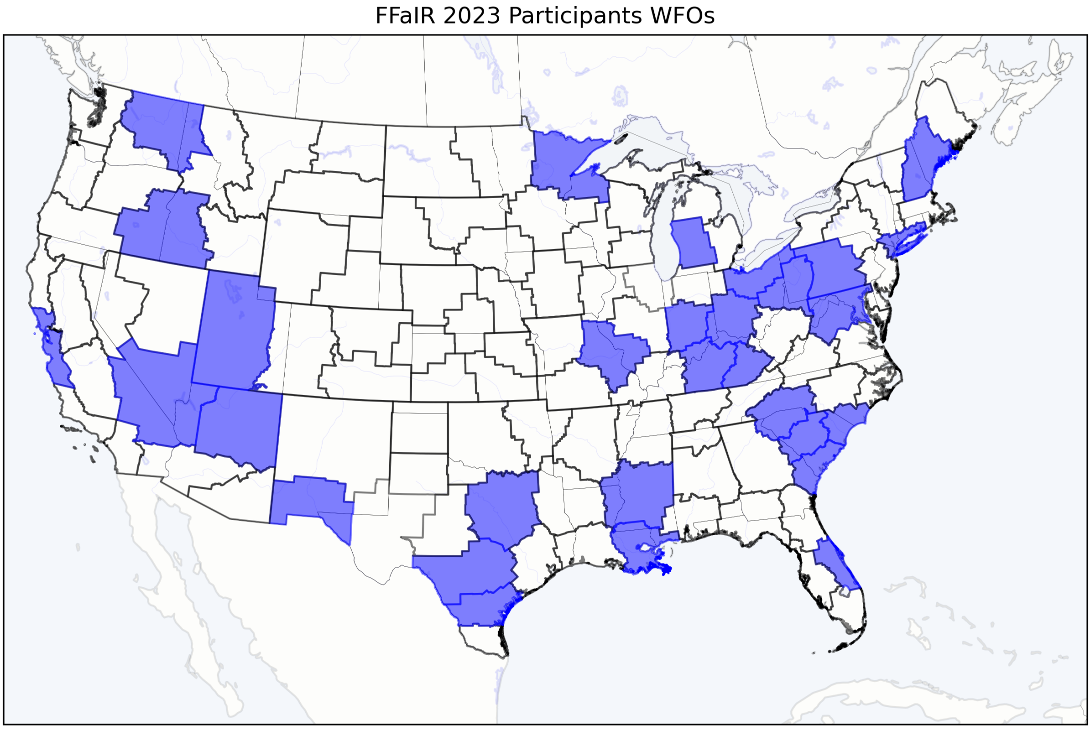


Figure 1: Graphic of WFOs that participated in the 2023 FFaIR Experiment, indicated by the blue fill. Not included is WFO Honolulu and NWS National Centers, which also had participates.

Finally, for the first time the FFaIR team is releasing the Final Report in two parts. Part 1 (aka this document) consists of results centered around the Rapid Refresh Forecast System (RRFS). This includes models or ensembles that employ the same model core, the Finite Volume Cubed-Sphere (FV3) core. Part 1 also includes a comprehensive explanation of the experimental operations and discussion of the weather during FFaIR. The description and discussion of the results of the forecasting activities will be spilt between Part 1 and Part 2. The Maximum Rainfall and Timing Product (MRTP) is the forecasting activity that is included in Part 1 since it included analysis of model 6-h Quantitative Precipitation Forecast (QPF). Part 2 of the 2023 Final Report will focus on the Day 1 FFaIR forecasting activities, the Excessive Rainfall Outlook (ERO) and an Average Recurrence Interval (ARI) based ERO called the AERO. It will also include analysis of Colorado State University's (CSU) First-Guess EROs and satellite products from a team at CSU's Cooperative Institute for Research in the Atmosphere (CIRA).

2 Science and Operations

While an in-depth summary of the daily operations of FFaIR and the data and products that were evaluated can be found in the 2023 FFaIR Operations Plan (Trojniak and Correia, Jr., 2023), a brief summary will be given in this section. This section will also cover the questions that were asked during the verification portion of FFaIR, as it relates to Part 1.

2.1 Daily Schedule

Although there was some variability in the day-to-day operations of FFaIR, in general the morning consisted of activities focusing on the Day 1 (16-12 UTC) time frame while the afternoon focused on verification and short-term (6 hr) forecasts. There were two Day 1 activities, the creation of an Excessive Rainfall Outlook (ERO) or an Average Recurrence Interval (ARI) based ERO, hereafter the AERO. After a weather briefing given by a WPC forecaster, the participants were broken into two groups, one worked to issue a Day 1 ERO, the other a Day 1 AERO. This part was similar to last year's FFaIR. However, unlike last year, this year participants were tasked with first creating an individual ERO or AERO before working to create a collaborative product. Volunteers shared their screen to discuss their individual ERO or AERO and products/tools they looked at to create their Day 1 product. Then, working as a team, each breakout group created a collaborative ERO/AERO. The two breakout groups would then come back together and discuss each group's collaborative Day 1 product and their thought process behind the forecast. Examples of this can be found in Part 2 of the report.

The verification session occurred after the Day 1 forecasting activities and will be discussed in Section 2.6. Once done with verification, the participants were again briefed by a WPC forecaster, with a focus on the current conditions, trends, and the threat for heavy rainfall. This was in preparation for the Maximum Rainfall and Timing Product (MRTP) forecasting activity. Using the information from the briefing, their own analysis of data, and open discussion, participants voted on a region and a 6-h window in which they expected the heaviest rainfall to occur. They also voted on a likely on a threshold for which they forecast the probability of exceedance; see the second to last bullet point in the next section listing MRTP required forecast elements. After voting, they then worked on their individual MRTP.

2.2 Description of Forecasting Activities - MRTP

Since the forecasting activities closely followed those done in the 2022 FFaIR Experiment and are explained in great detail in the 2023 FFaIR Operations Plan (Trojniak and Correia, Jr., 2023), the MRTP will only be briefly explained. The MRTP was comprised of multiple parts: a collaborative process of deciding over what region to forecast and for what 6-h time window between 21 UTC and 12 UTC that the forecast was valid for, evaluation of a randomly assigned model or ensemble, completion of a survey, drawing a 6-h precipitation forecast, and forecasting values for given thresholds or aspects of the forecast. The collaborative process was previously discussed above in Section 2.1. Participants were randomly assigned a model/ensemble to evaluate so that all the experimental data "had eyes on it". However, they were not required to use it in their forecast. In the survey, the participants answered questions about their assigned model/ensemble, indicated what data they found useful in the forecast process, and answered questions related to things like timing and flooding likelihood.

The latter two parts of the MRTP were the bulk of the MRTP process. Like in previous years, the participants had the option to draw 6-h QPF contours for: 0.5, 1, 2, 3, 4, and 5 inches. They also could draw a contour for where they thought flooding would occur. Additionally, they had to place a point where they thought the maximum 6-h total would be. The other aspects of the forecast required were:

- The 6-h maximum rainfall (corresponds with the point they placed).
- The maximum 6-h ARI to be exceeded.
- The 1-h maximum rainfall in the domain.
- The probability of flash flooding.
- The probability of the flash flooding leading to damage[1].
- The probability that the 6-h maximum rainfall would exceed the value that was voted on by the participants for the event; i.e. the participant's confidence an extreme event of the chosen threshold would occur.
- Considering all possible 6-h periods that comprised a valid MRTP time frame[2], what was the probability that the maximum 6-h rainfall would occur in each of those 6-h time periods.

The last two bullet points are explained in great detail in the 2023 FFaIR Operations Plan (Trojniak and Correia, Jr., 2023) and in the MRTP tutorial created for the participants[3].

2.3 Overview of RRFS and RRFS-like Data and Products

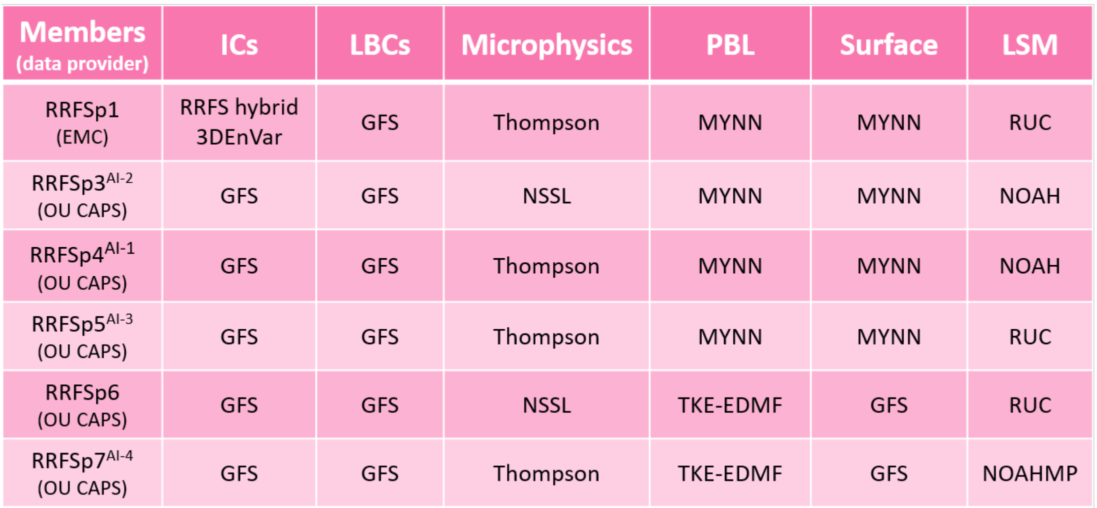
This section will serve as a brief summary of the data and products evaluated in FFaIR that will be discussed in Part 1 of the 2023 Final Report. Unfortunately, in some instances a full evaluation of the data or product could not be completed due to limited availability, changes to model configurations, or errors found in product code. This will be elaborated on further in this section, as well as in the Experiment Goals and Results sections.

[1] It was up to the participant to determine what they considered damage. There was no strict definition to this.

[2] There were 10 possible time periods, the first was from 21-03 UTC while the last was 06-12 UTC.

[3] Tutorial location: https://www.wpc.ncep.noaa.gov/hmt/hmt_webpages/drawingtools/tutorial.pdf

Table 1: The deterministic model configurations that were evaluated in 2023 FFaIR. For the models provided by the OU CAPS team, if the model is part of their machine learning product, the member number is super-scripted as AI-#. This is the same Table as Table 3 in the 2023 FFaIR Operations Plan (Trojniak and Correia, Jr., 2023).



Members (data provider)	ICs	LBCs	Microphysics	PBL	Surface	LSM
RRFSp1 (EMC)	RRFS hybrid 3DEnVar	GFS	Thompson	MYNN	MYNN	RUC
RRFSp3^{AI-2} (OU CAPS)	GFS	GFS	NSSL	MYNN	MYNN	NOAH
RRFSp4^{AI-1} (OU CAPS)	GFS	GFS	Thompson	MYNN	MYNN	NOAH
RRFSp5^{AI-3} (OU CAPS)	GFS	GFS	Thompson	MYNN	MYNN	RUC
RRFSp6 (OU CAPS)	GFS	GFS	NSSL	TKE-EDMF	GFS	RUC
RRFSp7^{AI-4} (OU CAPS)	GFS	GFS	Thompson	TKE-EDMF	GFS	NOAHMP

2.3.1 Rapid Refresh Forecast System

The Rapid Refresh Forecast System (RRFS[4]) has been in development since roughly 2018 and is slated to replace most of the operational convective allowing modeling (CAM) systems by Fall of 2025. Therefore, strenuous testing of the modeling system continues to be done across the NWS Testbeds and Proving Grounds.

This year FFaIR planned to evaluate both the deterministic RRFS and four RRFS ensemble configurations across the entirety of the experiment. Following last year's naming convention, the deterministic RRFS being developed for operational implementation, will hereafter be referred to as the RRFSp1[5]. The configuration for the RRFSp1 can be found in Table 1.

[4] This year one of the participants suggested verbally referring to the RRFS as the R2FS rather than Rufus, which has been the naming convention for a few years. Another suggestion was to call it aRRFS. As a Star Wars and dog lover, the facilitator fully supports referring to the RRFS as either R2FS or aRRFS.

[5] The Environmental Modeling Center (EMC) and the Global Systems Laboratory (GSL) refer to this as the RRFS_a or the RRFSa.

The RRFSp1 is also the control member for the RRFS ensembles provided for evaluation. The four ensemble configurations were designed to evaluate the impact of single and mixed physics, as well as time-lagging membership. The ensembles each consisted of 10 members; their configurations can be found in Tables 4 and 5 of the 2023 FFaIR Operations Plan (Trojniak and Correia, Jr., 2023). The single physics, non time-lagged ensemble was referred to as the RRFSe1, while the time-lagged version was called the RRFSe1tl. The multi-physics with no time-lagging was referred to as RRFSe2, while the time-lagged version called the RRFSe2tl. Please refer to Section 3.1 of the Operations Plan for a full description of the configurations for the ensembles.

Since the RRFS is in active development, the system was not "frozen" during the FFaIR Experiment. For the most part, changes made to the system were considered non-science changes that should not have had an impact on the model forecasts; a full list of the changes can be found at: https://www.emc.ncep.noaa.gov/users/emc.campara/rrfs/log.html[6]. That said, there was one change that occurred during FFaIR that resulted in significant down time and model availability. From June 30 to July 17 the system was turned off to change from the CONUS domain to the North American (NA) domain; see Fig. 2. Although the RRFSp1 was available beginning July 17, the RRFS ensemble was not up and running until July 24. After this change, the flow of data became less stable and there were cycles/days where no RRFS data was available. The down time resulted in week 4 of FFaIR having no RRFS data to evaluate, and weeks 3, 5, and 6 having at least one day in which the data was not available for multiple cycles. Additionally, since the computational expense increased tremendously with the domain increase, only one ensemble configuration was provided after the switch, the RRFSe1tl.

2.3.2 Data and Products Provided by the Center for Analysis and Prediction of Storms

The Center for Analysis and Prediction of Storms (CAPS) at the University of Oklahoma (OU) once again provided a variety of model data and products to be evaluated during FFaIR. Five RRFS-like model configurations were provided,

[6] The log is provided and updated by Ben Blake.

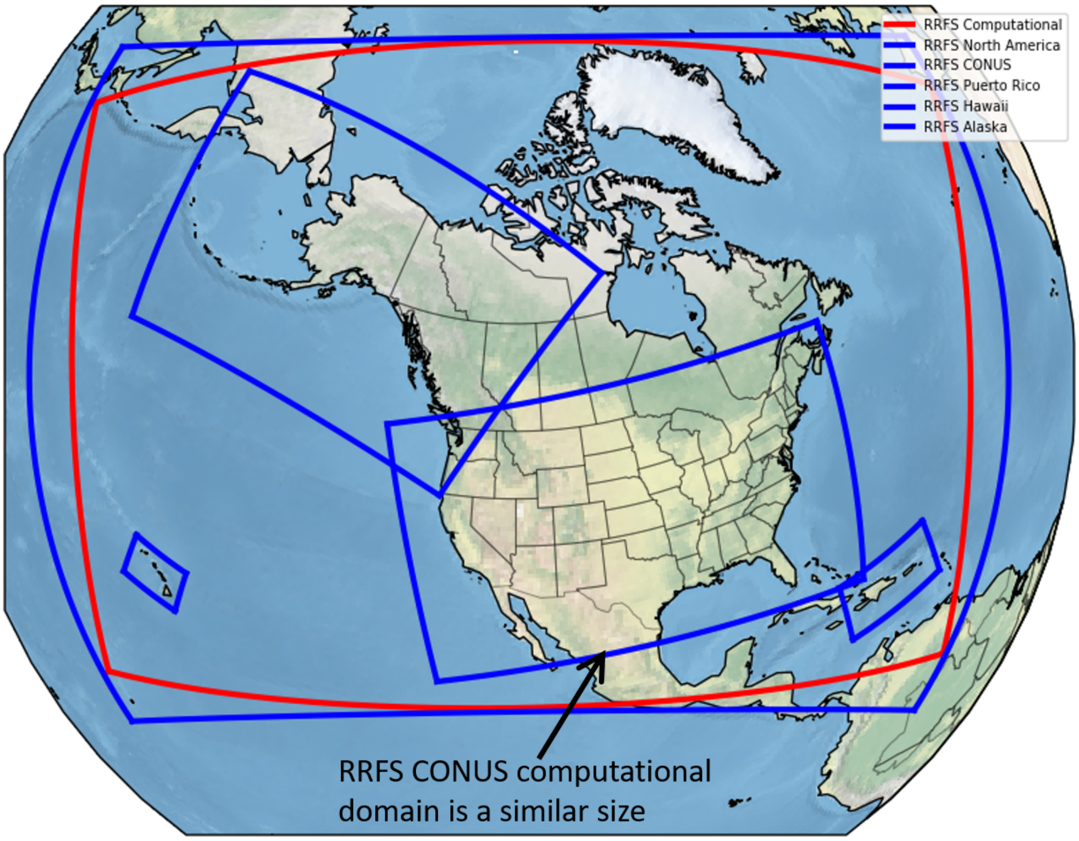


Figure 2: The outline of the various RRFS domains run by EMC in blue. The red outline is the computational area for the North American domain. The CONUS output domain is roughly the same size as its the computational domain (see black arrow). Image provided by Ben Blake at EMC.

ranging from changes in the microphysics to changes in the land-surface scheme. These configurations were referred to as RRFSp3, RRFSp4, RRFSp5, RRFSp6, and RRFSp7 and their makeup can be found in Table 1. These were only available to evaluate for the 00 z cycle and did not include any data assimilation outside of the Global Forecast System (GFS), which is what the models were initiated from. CAPS also provided an ensemble, referred to as the CAPS_RRFSe, which was a 10 member mixed physics ensemble (see Table 6 in the Operations Plan (Trojniak and Correia, Jr., 2023)) and used Global Ensemble Forecasting System (GEFS) members for the initial and lateral boundary conditions. Like the deterministic models, the CAPS_RRFSe was only available for the 00 z cycle.

As part of their ensemble product suite, in addition to the arithmetic mean (hereafter mean), the probability matched mean (PM or PMM), and the local probability matched mean (LPM or LPMM), the CAPS team calculated two spatial-aligned means (SAM). One was the normal SAM, which used the methods described in Brewster (2003) for the spatial alignment of the background forecast to observations. The other combines the SAM methodology with the LPMM method to create what is referred to as the SAM-LPM. For this mean, the SAM method was applied first to the ensemble members, followed by the LPMM method.

A machine learning (ML) product (MLP), called the HREF+, was also provided. This MLP used U-net ML methodology to create probabilistic forecasts for 6-h QPF exceeding one-half, one, and two inches. It is referred to as the HREF+ because its training used membership from the HREF combined with some of the CAPS deterministic model configurations. The CAPS contribution to the membership are noted by a super-scripted AI followed by a number in Table 1 while the HREF members used were the HRRR, NAMnest, hiresw_arw, and hiresw_nssl, along with their time-lagged forecasts. The Operations Plan (Trojniak and Correia, Jr., 2023) provides an in-depth explanation of this MLP. Unfortunately, near the end of the experiment, a bug was found in the calculation of the probabilities. Therefore, although participants evaluated the product, results from MLP forecasts will not be discussed.

2.4 Other Models

The operational versions of the HRRR, NAMnest, and HREF were used as comparison to the experimental models and ensembles. Additionally, the FV3 member of the HREF, hereafter FV3-HREF, was loosely evaluated during the experiment as a baseline for the RRFSp1 since the FV3-HREF is the “primitive” version of the RRFSp1. Lastly, to gather performance information for the Hazardous Weather Testbed (HWT) NSSL team, one of the three Model for Prediction Across Scales (MPAS) CAMs evaluated in the Spring Forecasting Experiment (SFE) was included in some of the verification and forecasting activities. Tables 1 and 2 in the HWT SFE Preliminary Findings (Clark et al., 2023) describe the configuration of the MPAS-NSSLs, the version used in FFaIR was the one initiated with HRRR

IC/LBCs and included Thompson microphysics. During FFaIR this version was referred to as the MPAS-NSSL.

2.5 Science Questions and Goals for Part 1

As mentioned previously, a large portion of FFaIR was planned around the evaluation of the RRFS, both its deterministic and ensemble output. With the system moving closer to its planned code freeze and the NWS science evaluation, it was top priority for the HMT team to scrutinize its warm season precipitation forecasting performance. This included not only rainfall accumulations, but storm mode/progression, location, areal coverage and rates. This was compared against the operational CAMs, the HRRR and the NAMnest, as well as the aforementioned CAPS provided configurations. Probabilistic forecasts of 6-h QPF were the focus for evaluating the RRFS ensemble(s).

The full list of science questions and goals that will be addressed in Part-1 of the Final Report are listed below. Note that some of these were only able to be partly addressed do to missing data or errors in code that were found; these questions will be italicized.

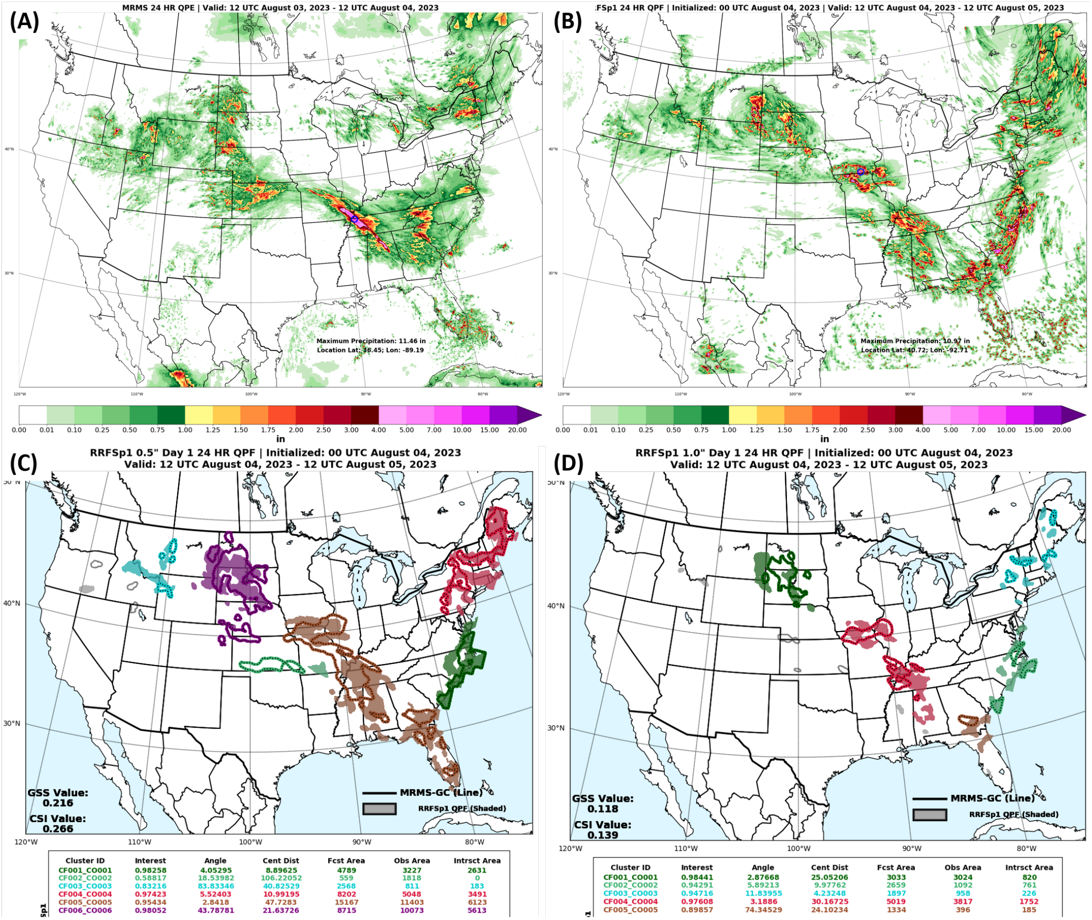
- Evaluate the performance, focusing on Quantitative Precipitation Forecast (QPF) and precipitation rate, of the RRFS_a (referred to as the RRFSp1 in the FFaIR experiment) compared to the HRRR and NAMnest.
- Evaluate the performance of different configurations of the RRFS other than the planned operational version of the RRFS deterministic.
- *Analyze the impact of the RRFS DA and compare it to the DA done in the HRRR.*
- *Identify the pros and cons of a multi-physics ensemble compared to a single physics ensemble with stochastic perturbations. Compare their performance to the HREF.*
- *Evaluate the performance of ensembles with time-lagged members.*

- *Evaluation of a machine learning product for the probability of QPF exceedance, called the HREF+, from the CAPS group.*
- Evaluation of a OU-CAPS Spatially-Aligned Mean (SAM) and a SAM with local probability matched mean (LPM) applied with the SAM methodology, called the SAM-LPM.
- Evaluate the performance (CSI, max QPF) of the various models for specific 6-h precipitation extreme events via the Maximum Rainfall and Timing Product (MRTP).

2.6 Verification Methods

As is the case with most testbeds, participant feedback and subjective evaluation of products and tools are an integral part of the testbed environment. Feedback happens both naturally during the forecasting activities, as the participants are actively utilizing the tools, and during the scheduled verification sessions. With the exception of Mondays, verification is of the previous day's guidance and forecasts used during the previous day's forecasting activities. Generally, the guidance is compared against observations and participants rank the guidance on a scale of goodness from 1 (poor) to 10 (great). However, this year some questions extended beyond the general goodness question, including scoring for confidence added to a forecast or how forecasted maxima relate to observed maxima. Additionally, most questions included asking the participants to provide written feedback on their ratings and their overall thoughts on the performance of the product/model being evaluated. Finally, differing from previous years, this year participants were divided into two groups to expand verification without increasing the burden on participants. Group 1 evaluated the earlier forecasts/cycles and Group 2 the later ones. Not all questions were broken into two groups.

Precipitation verification was done using Multi-Radar Multi-Sensor Gauge Corrected (hereafter MRMS) Quantitative Precipitation Estimate (QPE) as truth. The MRMS data was remapped to the HRRR grid, using the cKDTree package in Python and retained the maximum value of the 9 grid point neighborhood. The MRMS/model comparison was shown both as a side by side comparison and via


(A)
MRMS 24 HR QPE | Valid: 12 UTC August 03, 2023 - 12 UTC August 04, 2023
Maximum Precipitation: 11.46 in
Location Lat: 38.45; Lon: -89.19
(B)
FSp1 24 HR QPF | Initialized: 00 UTC August 04, 2023 | Valid: 12 UTC August 04, 2023 - 12 UTC August 05, 2023
Maximum Precipitation: 10.97 in
Location Lat: 40.72; Lon: -92.71
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
in
(C)
RRFSp1 0.5" Day 1 24 HR QPF | Initialized: 00 UTC August 04, 2023
Valid: 12 UTC August 04, 2023 - 12 UTC August 05, 2023
GSS Value: 0.216
CSI Value: 0.266
MRMS-GC (Line)
RRFSp1 QPF (Shaded)
40°N
30°N
120°W
110°W
100°W
90°W
80°W
(D)
RRFSp1 1.0" Day 1 24 HR QPF | Initialized: 00 UTC August 04, 2023
Valid: 12 UTC August 04, 2023 - 12 UTC August 05, 2023
GSS Value: 0.118
CSI Value: 0.139
MRMS-GC (Line)
RRFSp1 QPF (Shaded)

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.98258	4.05295	8.89625	4789	3227	2631
CF002_CO002	0.58817	18.53982	106.22052	559	1818	0
CF003_CO003	0.83216	83.83346	40.82529	2568	811	183
CF004_CO004	0.97423	5.52403	10.99195	8202	5048	3491
CF005_CO005	0.95434	2.8418	47.7283	15167	11403	6123
CF006_CO006	0.98052	43.78781	21.63726	8715	10073	5613

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.98441	2.87668	25.05206	3033	3024	820
CF002_CO002	0.94291	5.89213	9.97762	2659	1092	761
CF003_CO003	0.94716	11.83955	4.23248	1897	958	226
CF004_CO004	0.97608	3.1886	30.16725	5019	3817	1752
CF005_CO005	0.89857	74.34529	24.10234	1334	396	185

Figure 3: 24-h (A) MRMS QPE and (B) RRFSp1 QPF valid at 12 UTC 05 Aug. 2023. MODE verification for (C) 0.5 in and (D) 1 in with the MRMS QPE contoured and the model QPF filled. Matching MRMS and model clusters have the same color. Included in the MODE plot is statistical information for the forecast threshold as a whole and for the individual clusters identified by MODE.

images showing object verification. The object verification graphics were created using the Developmental Testbed Center's Model Evaluation Tools (MET) Method for Object-Based Diagnostic Evaluation (MODE). The MODE verification allowed participants to focus on precipitation thresholds while evaluating the models. An example of the RRFSp1 MODE verification for one half inch and one inch can be seen in Fig. 3. The configuration used for MODE is the same as the previous two years and can be found in Appendix D of the 2020 FFaIR Final Report (Trojniak et al., 2020). MET/MODE was also utilized for some of the statistical analysis discussed later.

Table 2: The number of times the MRTP was done for a given 6-h time window. This also corresponds to the number of times a 6-h time window was used for the 6-h QPF subjective evaluation during the daily verification sessions. The * indicates that the given 6-h was used one additional time for the 6-h verification question compared to the MRTP.

21-03 UTC: 4	22-04 UTC: 2*	23-05 UTC: 0	00-06 UTC: 4	01-07 UTC: 0
02-08 UTC: 0	03-09 UTC: 6	04-10 UTC: 2	05-11 UTC: 0	06-12 UTC: 12

Precipitation verification was done for 24-h and 6-h QPFs. For 24-h QPF verification, only the HRRR, NAMnest, and RRFSp1 were verified. The 24-h period verified was defined as 12-12 UTC for the 18z, 00z, 06z, and 12z cycles. The 18z cycle was the oldest cycle, with a valid forecast hour (referred to as f or fhr) of f42. The 00z cycle was valid at f36, 06z was valid at f30 and 12z was valid at f24. Participants were asked to evaluate each model's goodness and to pick the "best" model of the cycle. This was to see what would happen if the participants were forced to pick a winner even if they gave multiple models the same score. For the 6-h verification, forecasts were verified for the 6 hours that corresponded with when the previous day's MRTP was valid. For example, on Aug 1 the MRTP was forecast for the 06-12 UTC period, therefore the 6-h QPF verification question was valid 06-12 UTC Aug 2. Verification was done for the 00z, 06z, and 12z cycles, with all deterministic models being verified for the 00z cycle, while only the HRRR, NAMnest, and RRFSp1 were verified for the other two cycles. Table 2 lists the number of times a valid 6 hour time period for the MRTP was done, which corresponded to the number of 6-h QPF verification periods in question.

Model maximum precipitation rate (pmax) for the NAMnest, RRFSp1, RRFSp3 and RRFSp4 for the same 6-h time window as the 6-h QPF verification was also done[7]. The pmax was plotted across the time period, using the maximum for a grid point over the 6 hours. MRMS maximum prate from 2-min data was used for truth. Rather than asking the typical goodness question, participants instead were asked two comparison questions. (1) Compare the MRMS and the 00z suite maximum

[7] The HRRR was not used because it does not provide pmax only instantaneous precipitation rate.

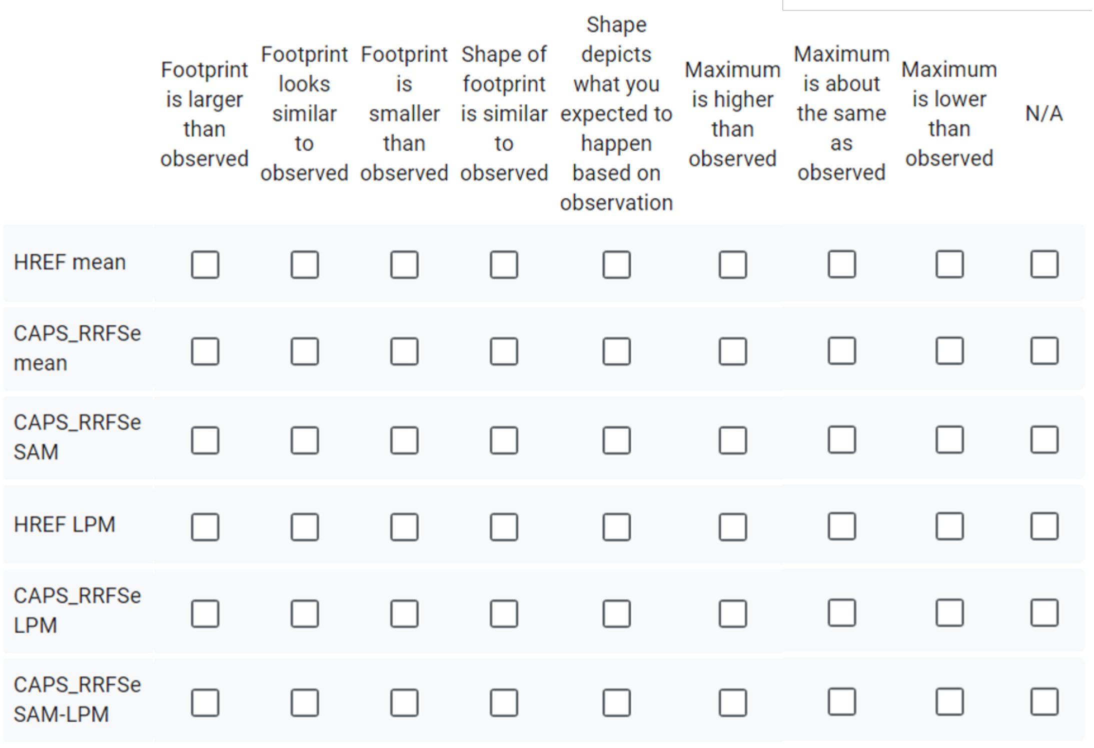
precipitation rates for magnitude. They could pick as many of the following choices as they felt applied for each model: too high, reasonable, too low, and/or max value is too high. (2) Compare the MRMS and the 00z maximum precipitation rates for areal coverage and position. Here too they could pick as many of the following choices as they felt applied for each model: too large, just right, too small, similar location, and/or large displacement.

For ensemble verification, the 1-in 6h and 5-in 6h probabilities were evaluated. The valid time for the verification was either 03, 06, 09, or 12 UTC, whichever was closest to the MRTP valid end time[8]. For evaluation, the 1 or 5 in 6-h MRMS QPE was contoured on the ensemble probability graphic. Additionally, using a 39 km gaussian smoother, the MRMS QPE was shown probabilistically for each of the thresholds. The goal was to evaluate the 00z and 12z cycles of the ensembles but due to the inconsistent availability of data, the two cycles (18z, 00z, 06z, or 12z) with the most data available were evaluated each day. For this question, participants were asked "On a day like today would the probabilities you see give you more or less confidence in the forecast?" They could give a score of 1 (less confidence) to 5 (more confidence).

Since the RRFSe1tl was the ensemble configuration that was provided by EMC after the RRFS downtime to change over to the NA domain, the hope was that it would have been the most consistently available for evaluation pre-change. Unfortunately, the RRFSe1tl was the least available during the first 3 weeks of FFaIR for subjective evaluation, and thus evaluated the least for that time period. Therefore it is extremely difficult to complete an evaluation of the RRFSe1tl or any other ensemble provided by EMC. As for the CAPS_RRFSe, most of the first two weeks of FFaIR the ensemble was missing but it was provided consistently through the rest of the experiment. Therefore, a brief discussion of the CAPS_RRFSe compared to the HREF will be provided.

Evaluation of the two experimental spatially aligned means provided by CAPS_RRFSe, the SAM and SAM-LPM, was done by comparing these means to either the general mean or the LPM mean from the HREF and CAPS_RRFSe. For

[8] So if a MRTP was valid at 04 UTC then the following day was verified for 03 UTC.

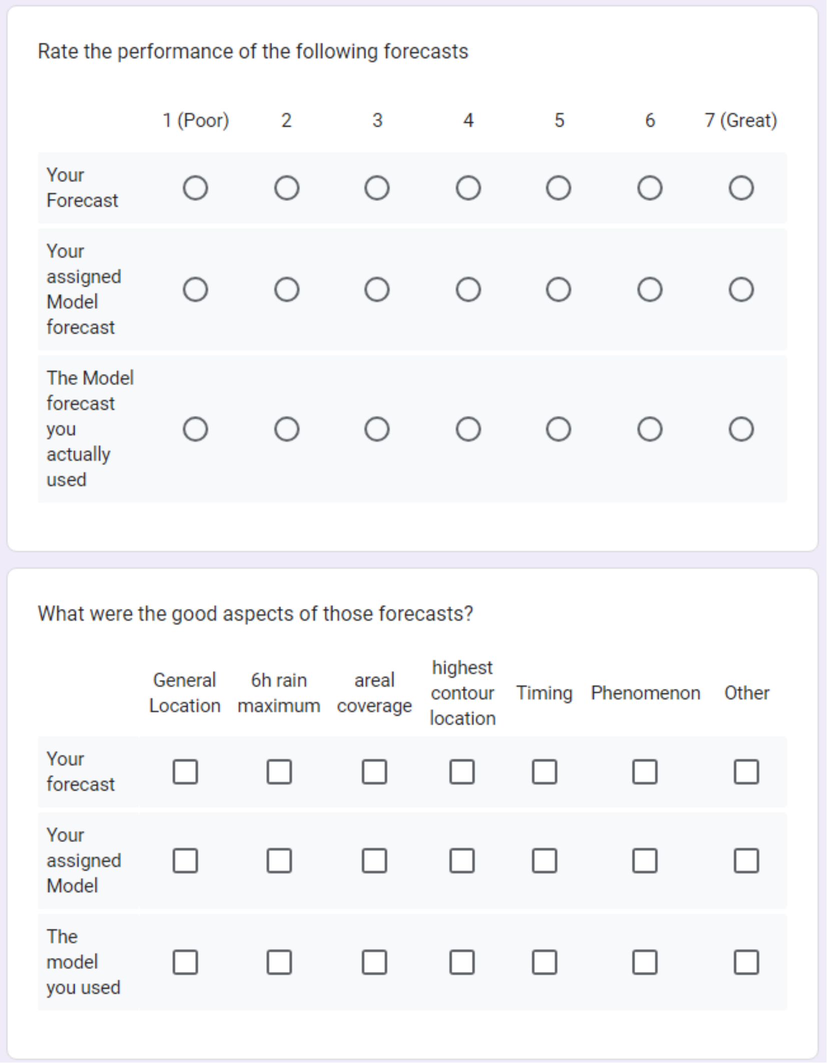


	Footprint is larger than observed	Footprint looks similar to observed	Footprint is smaller than observed	Shape of footprint is similar to observed	Shape depicts what you expected to happen based on observation	Maximum is higher than observed	Maximum is about the same as observed	Maximum is lower than observed	N/A
HREF mean	□	□	□	□	□	□	□	□	□
CAPS_RRFSe mean	□	□	□	□	□	□	□	□	□
CAPS_RRFSe SAM	□	□	□	□	□	□	□	□	□
HREF LPM	□	□	□	□	□	□	□	□	□
CAPS_RRFSe LPM	□	□	□	□	□	□	□	□	□
CAPS_RRFSe SAM-LPM	□	□	□	□	□	□	□	□	□

Figure 4: The question setup for evaluation of the CAPS_RRFSe SAM and SAM-LPM.

this, participants were shown the 6-h means for either 00 UTC or 06 UTC and asked "How well does each mean capture the following? Please check as many descriptions that apply to each ensemble mean." Figure 4 shows the choices the participants had.

Participants also evaluated their MRTPs from the previous day, as well as the model/ensemble they were randomly assigned to evaluate during the activity. Figure 5 shows the cadence of questions for the MRTP verification. The verification graphics for the participant MRTP included their forecast overlaid with the 6-h MRMS, along with their forecasted and the observed location of maximum rainfall. It also included information about what they forecasted the previous day, e.g. what ARI they expected to be exceeded and how far their forecasted rainfall maximum was from the observed maximum. Their CSI at each MRTP threshold they drew was included along with their assigned model's/ensemble's CSI at 1 in for all available cycles. For evaluation of the model QPF, the 1 in contour for QPF was overlaid with 6-h MRMS while for the ensembles, the PMM and LPMM



Rate the performance of the following forecasts

	1 (Poor)	2	3	4	5	6	7 (Great)
Your Forecast	○	○	○	○	○	○	○
Your assigned Model forecast	○	○	○	○	○	○	○
The Model forecast you actually used	○	○	○	○	○	○	○

What were the good aspects of those forecasts?

	General Location	6h rain maximum	areal coverage	highest contour location	Timing	Phenomenon	Other
Your forecast	□	□	□	□	□	□	□
Your assigned Model	□	□	□	□	□	□	□
The model you used	□	□	□	□	□	□	□

Figure 5: The question setup for the MRTP verification.

1 in contour was overlaid. These graphics included similar additional information to what the MRTP graphics included. Figure 6 shows an example of what the verification graphics looked like, along with labels for the information included.

2.6.1 Verification Terminology for Dates

Although there were 30 days that FFaIR was in session, there were actually 31 days used for the subjective evaluation during FFaIR. For the first day of the experiment, participants verified data from the previous Friday, valid ending

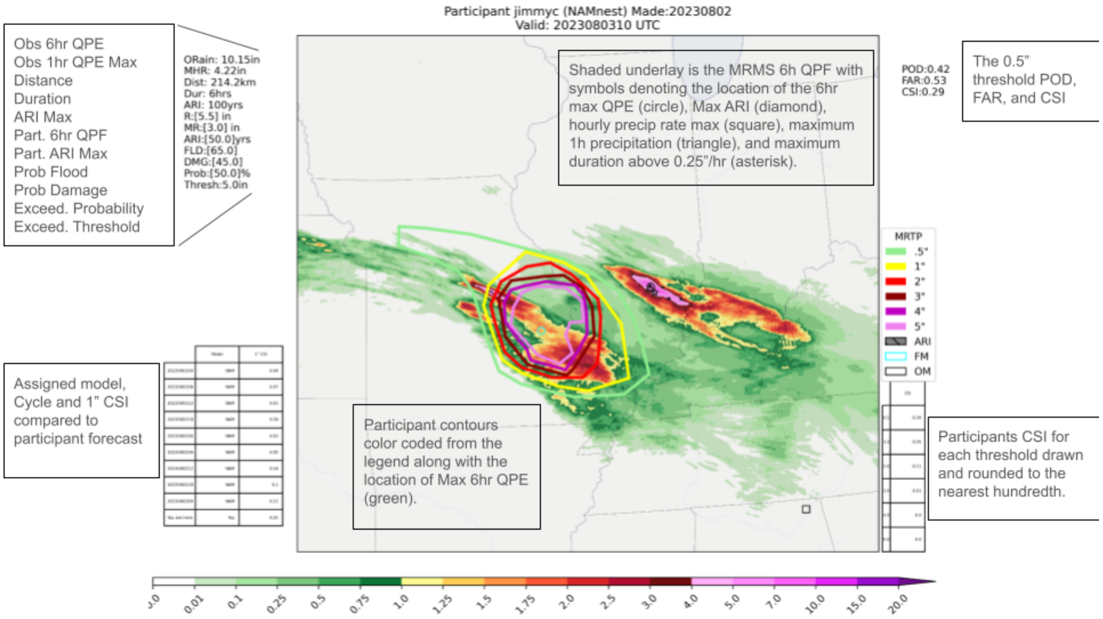

Participant jimmyc (NAMnest) Made:20230802
Valid: 2023080310 UTC
Obs 6hr QPE
Obs 1hr QPE Max
Distance
Duration
ARI Max
Part. 6hr QPF
Part. ARI Max
Prob Flood
Prob Damage
Exceed. Probability
Exceed. Threshold
ORain: 10.15in
MHR: 4.22in
Dist: 214.2km
Dur: 6hrs
ARI: 100yrs
R:[5.5] in
MR:[3.0] in
ARI:[50.0]yrs
FLD:[65.0]
DMG:[45.0]
Prob:[50.0]%
Thresh:5.0in
Shaded underlay is the MRMS 6h QPF with symbols denoting the location of the 6hr max QPE (circle), Max ARI (diamond), hourly precip rate max (square), maximum 1h precipitation (triangle), and maximum duration above 0.25"/hr (asterisk).
POD:0.42
FAR:0.53
CSI:0.29
The 0.5" threshold POD, FAR, and CSI
MRTP
.5"
1"
2"
3"
4"
5"
ARI
FM
OM
Assigned model, Cycle and 1" CSI compared to participant forecast
Participant contours color coded from the legend along with the location of Max 6hr QPE (green).
Participants CSI for each threshold drawn and rounded to the nearest hundredth.
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

Figure 6: Example verification image from the MRTP experiment, labeling the relevant elements that participants could examine for quality.

12 UTC 03 June 2023; an exception to this were the verification questions associated with MRTPs, EROs and AEROs. These 31 days will be referred to as FFaIR Dates from here on out. In addition to the FFaIR Dates, evaluation across the Testbed Season was also done. For this time span, data was collected from May 1 to Aug. 11 2023, regardless of whether the FFaIR was in session or not. Furthermore, due to the emphasis placed on the RRFSp1 evaluation, three additional date groups were analyzed, referred to as RRFS Dates, CONUS Dates, and NA Dates. The RRFS Dates includes only the days that the 00 z RRFSp1 was available for verification during FFaIR. The CONUS Dates are the RRFS Dates in which the CONUS domain was used, while the NA Dates are the RRFS Dates that the NA domain was used.

Finally, for the ensemble evaluation, the dates that the CAPS_RRFSe data was available will be called the CAPS Dates. This is because of data from the RRFS missing during week 3. Additionally, since the CAPS_RRFSe was only provided at 00z all ensemble discussion will be for 00z. Regarding data availability, for the 00z cycle the number of scores each ensemble received were: HREF - 193, RRFSe1 - 24, RRFSe1tl - 42, RRFSe2 - 25, RRFSe2tl - 28, and CAPS_RRFSe - 149.

3 A Brief Discussion on the Weather for FFaIR

Understanding the synoptic pattern, rainfall coverage, and types of events that occurred during this year's experiment will help with the overall analysis of the subjective and objective results. Even though this year's experiment lasted longer than previous years by 2 weeks[9], general comparisons can still be made. In Fig. 11 is MRMS-QPE summed over the FFaIR Dates for 2021, 2022, and 2023. For the 2023 accumulation, the FFaIR dates in which the RRFSp1 was available for evaluation is also shown; meaning most of July was not included in the sum.

Both 2023 accumulations and 2021 look more similar to one another than either looks to 2022. One exception to this is over the western Gulf States, where it was unseasonably dry this year. In fact from May 2023 to Aug 2023, drought conditions developed from the TX coast to the FL Panhandle; see Fig. 8. Furthermore, unlike last year, this year there was not a lull in precipitation across the Midwest or the Plains. In fact, during week 5 of FFaIR, numerous events occurred in the Midwest. This led to every day of that week's MRTP domain being located over Missouri. Figure 9 shows the 6-h MRMS QPE for each of the valid MRTP periods during the first week of August, while Fig. 10 and Fig. 3A show the 24-h total for each day of week 5 across the CONUS. Lastly, Fig. 11 compares the summed total for the week and the four days that RRFSp1 was available.

Since the RRFS was not available for all the days, evaluation of the RRFSp1 and the RRFSe1tl against the operational systems for the entire active period was not possible. That said, the 4 events that the RRFS was running for helped provide some insight to how the RRFS performed for convection with Mesoscale Convective System (MCS) characteristics that developed/strengthened overnight and were driven by Mesoscale Convective Vortexes (MCVs) and the Low Level Jet (LLJ). An in-depth discussion of model performance of one of these events will be done in Section 4.4. This time span also provided 3 of the 10 days in which the MRTP domain saw an event where the 6-h 1 in coverage was >30,000-km^2.

[9] Two weeks refers to the number of extra weeks FFaIR hosted participants, dates for these weeks are called FFaIR dates. However, statistical analysis from May 1 to Aug 12, 2023 was also done.

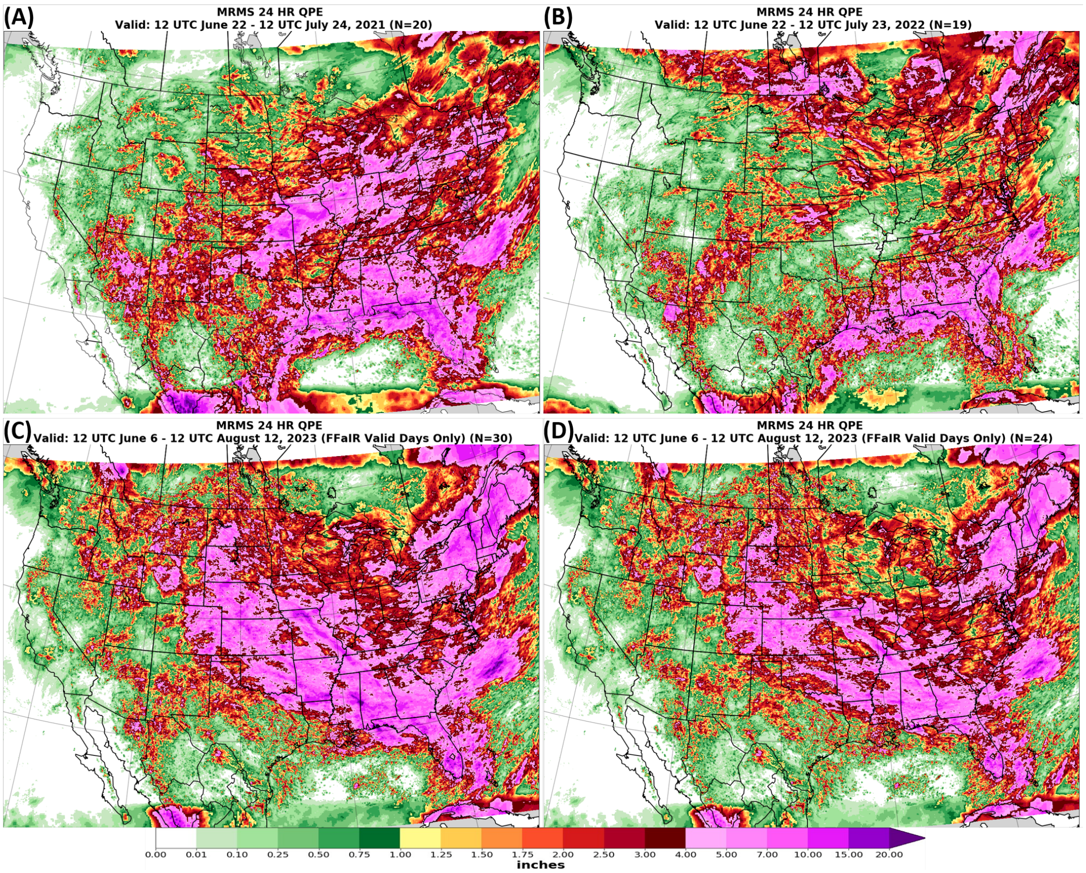

(A)
MRMS 24 HR QPE
Valid: 12 UTC June 22 - 12 UTC July 24, 2021 (N=20)
(B)
MRMS 24 HR QPE
Valid: 12 UTC June 22 - 12 UTC July 23, 2022 (N=19)
(C)
MRMS 24 HR QPE
Valid: 12 UTC June 6 - 12 UTC August 12, 2023 (FFaIR Valid Days Only) (N=30)
(D)
MRMS 24 HR QPE
Valid: 12 UTC June 6 - 12 UTC August 12, 2023 (FFaIR Valid Days Only) (N=24)
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 7: Accumulation of MRMS QPE for the days FFaIR was in session for (A) 2021 [n=20], (B) 2022 [n=19], (C) 2023 [n=30], and (D) RRFS 2023 [n=24], where RRFS 2023 refers to the dates in which the 00z RRFSp1 was available for evaluation using the 2023 FFaIR sessions.

Another difference between the 2023 FFaIR season and the two previous was the storm mode seen across the Southeast. Rather than numerous days of weakly forced convection that was diurnally driven, more organized convection was seen. This was driven by anomalously low heights across the eastern CONUS. Figure 12 shows the difference among the 500-mb and mean sea level pressure for the months that the 2021, 2022, 2023 FFaIR Experiments were in session. This difference in the weather pattern over the Southeast made it difficult to evaluate if updates to RRFSp1 helped to address the excessive development of popcorn convection that has been reported on by the FFaIR team starting in the 2020 FFaIR Final Report

(Trojniak et al., 2020), due to the environment being unfavorable for this type of convection.

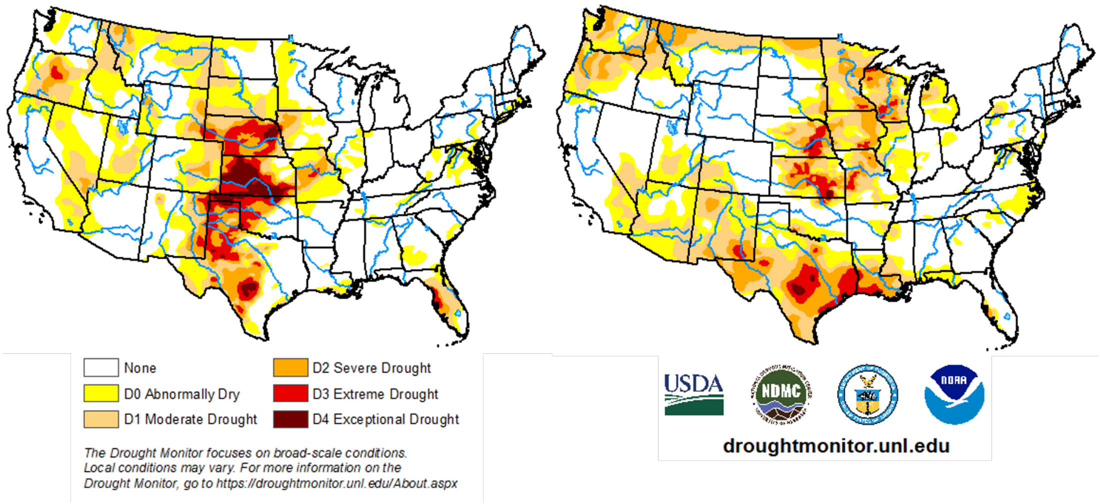

None
D0 Abnormally Dry
D1 Moderate Drought
D2 Severe Drought
D3 Extreme Drought
D4 Exceptional Drought
The Drought Monitor focuses on broad-scale conditions.
Local conditions may vary. For more information on the
Drought Monitor, go to https://droughtmonitor.unl.edu/About.aspx
USDA
NDMC
NOAA
droughtmonitor.unl.edu

Figure 8: The change in the US Drought Monitor from (left) May 9 2023 to (right) Aug 15, 2023. Images taken taken from https://droughtmonitor.unl.edu/.

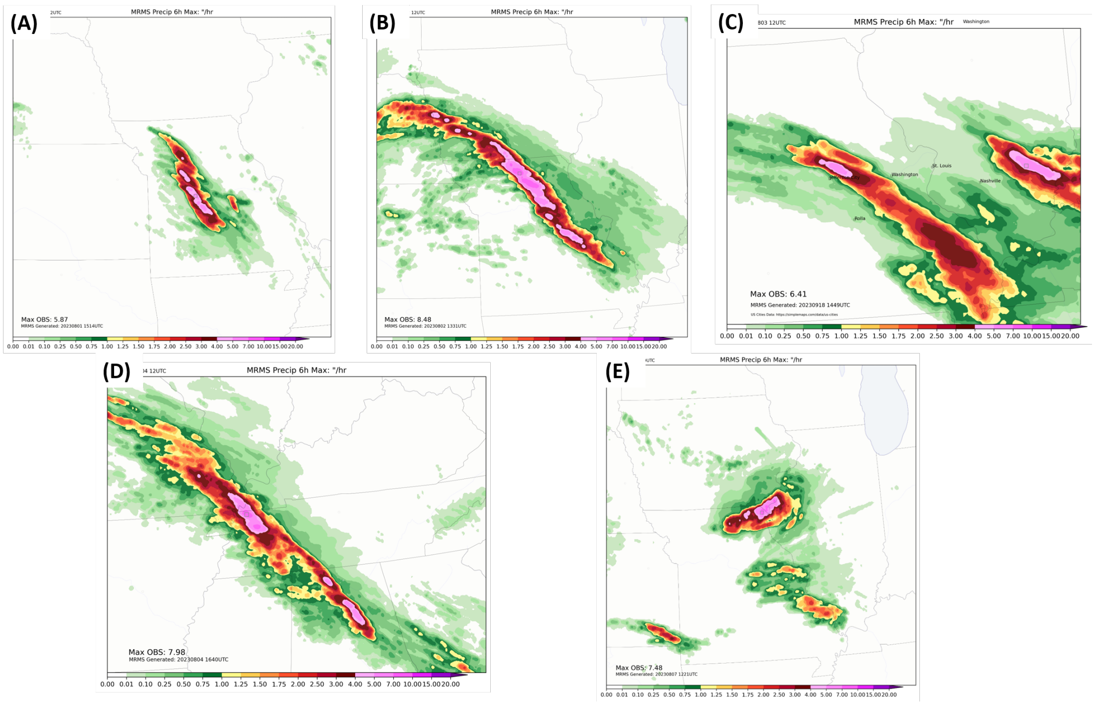


Figure 9: 6-h MRMS QPE for the MRTPs during week 5 of the 2023 FFaIR Experiment valid: (A) 06-12 UTC 01 Aug., (B) 06-12 UTC 02 Aug., (C) 06-12 UTC 03 Aug., (D) 06-12 UTC 04 Aug., and (E) 03-09 UTC 05 Aug. 2023.

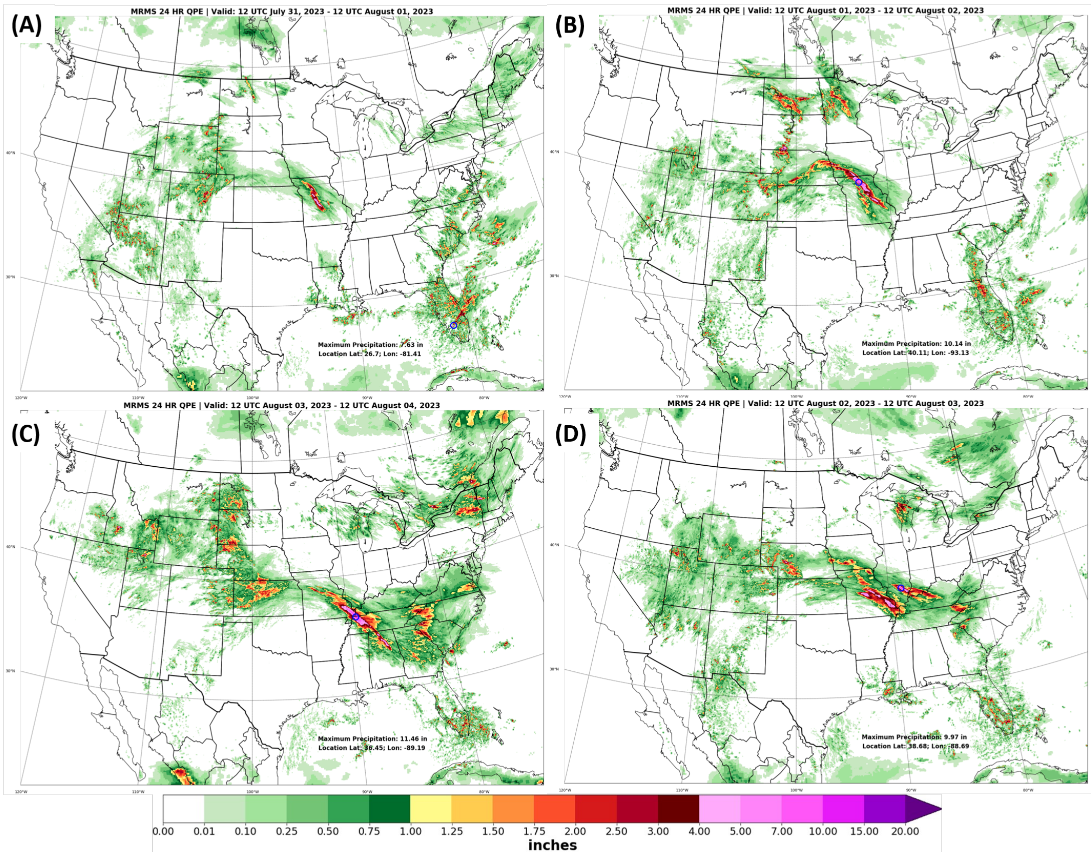

(A)
MRMS 24 HR QPE | Valid: 12 UTC July 31, 2023 - 12 UTC August 01, 2023
Maximum Precipitation: 7.63 in
Location Lat: 26.7; Lon: -81.41
(B)
MRMS 24 HR QPE | Valid: 12 UTC August 01, 2023 - 12 UTC August 02, 2023
Maximum Precipitation: 10.14 in
Location Lat: 40.11; Lon: -93.13
(C)
MRMS 24 HR QPE | Valid: 12 UTC August 03, 2023 - 12 UTC August 04, 2023
Maximum Precipitation: 11.46 in
Location Lat: 36.45; Lon: -89.19
(D)
MRMS 24 HR QPE | Valid: 12 UTC August 02, 2023 - 12 UTC August 03, 2023
Maximum Precipitation: 9.97 in
Location Lat: 38.68; Lon: -88.69
40°N
30°N
120°W
110°W
100°W
90°W
80°W
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 10: The 24-h MRMS QPE valid ending at 12 UTC on (A) 01 Aug., (B) 02 Aug., (C) 03 Aug., and (D) 04 Aug, 2023. These are 4 of the 5 days of week 5, the last day's, 05 Aug. 2023 24-h MRMS QPE can be found in Fig. 3A.

For the 30 days that FFaIR was in session, there were 10 days in which the 16z Day 1 WPC ERO had at least one area with a Moderate Risk across the country; Week 1: 2 days, Week 2: 1 day, Week 3: 0 days, Week 4: 4 days, Week 5: 3 days, and Week 6: 0 days. Furthermore, the first day of Week 4, July 10, had a High Risk over Vermont. This High Risk was driven by the same system that brought deadly flooding across PA, NJ, NY and DE the previous day (ex: Medina et al. (2023) and Espinoza and FOX 29 staff (2023)). Vermont and the surrounding areas saw prolonged rainfall spanning the two days, thus antecedent conditions helped to drive the High Risk issuance on July 10; see Fig. 13 to compare the WPC and FFaIR Day 1 Products. The event resulted in widespread flooding, washouts of bridges and roadways and mud/landslides; a summary of the event by the Burlington WFO can be found on their website (Banacos, 2023). Unfortunately, FFaIR was not in session for the first day of the event (it was a Sunday) and the RRFS was unavailable due to domain transition.

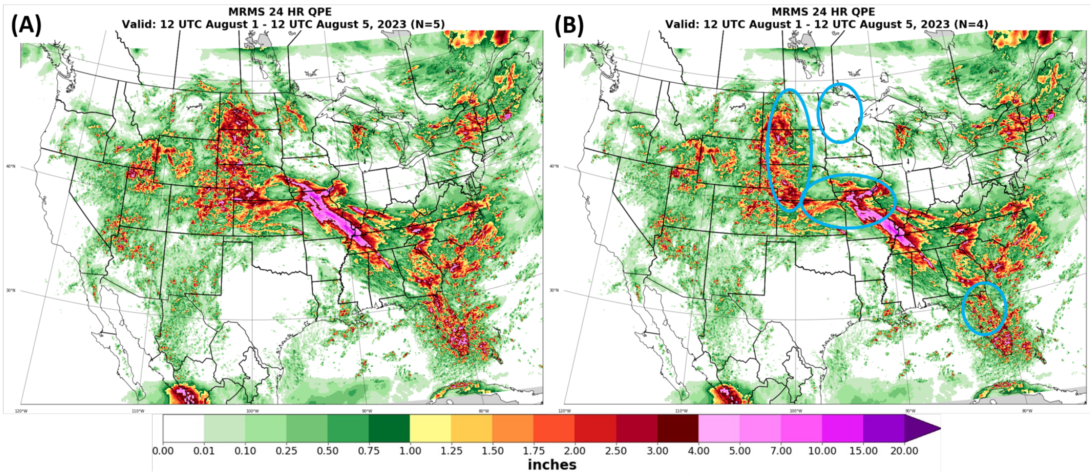


Figure 11: Accumulation of MRMS QPE for week 5 (July 31 - Aug 5 2023) of FFaIR which (A) includes all five days and (B) includes only the days RRFSp1 was available; i.e. the 24-h period ending at 12 UTC on Aug. 2 is not included. The blue circles in (B) indicate where there a noticeable differences between the accumulations.

4 Results

The results section will encompass both subjective and objective results from the 2023 FFaIR experiment. Much of the section will focus on the performance of the RRFS deterministic model (RRFSp1) compared to the HRRR and NAMnest. This will be followed by a brief summary of ensemble performance and the evaluation of the mean products supplied by CAPS. Lastly evaluations related to the MRTP activity will be discussed.

4.1 Deterministic

As discussed in Section 2.6.1, missing data impacted the ability to compare models across the entirety of FFaIR. Because of this, data was evaluated across various date groups (ex. FFaIR Dates and RRFS Dates). This resulted in a plethora of additional data to analyze. Therefore, all data groups will not always be discussed, especially when the results are comparable across the date groups.

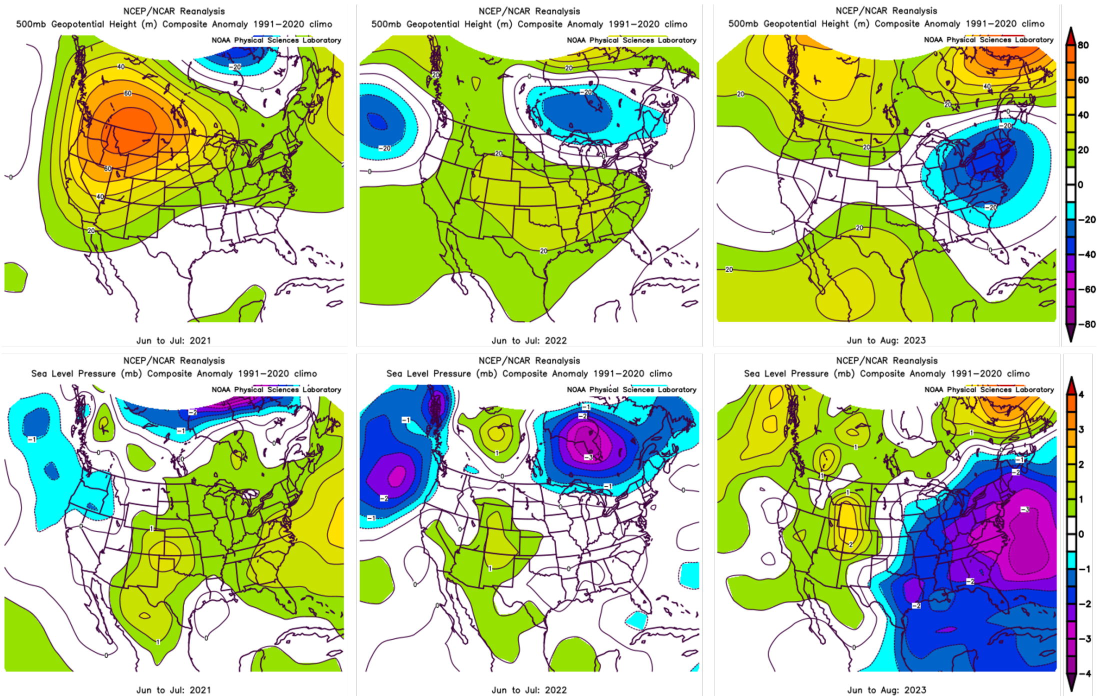

NCEP/NCAR Reanalysis
500mb Geopotential Height (m) Composite Anomaly 1991–2020 climo
NOAA Physical Sciences Laboratory
Jun to Jul: 2021
Jun to Jul: 2022
Jun to Aug: 2023
Sea Level Pressure (mb) Composite Anomaly 1991–2020 climo
80
60
40
20
0
-20
-40
-60
-80
4
3
2
1
0
-1
-2
-3
-4

Figure 12: [TOP] 500-mb geopotential height (m) and [BOTTOM] sea level pressure (mb) composite anomalies for [LEFT] June-July 2021, [MIDDLE] June-July 2022, and [RIGHT] June-August 2023. Composite images created using PSL's Monthly/Seasonal Climate Composites Website (NOAA Physical Sciences Laboratory, 2023).

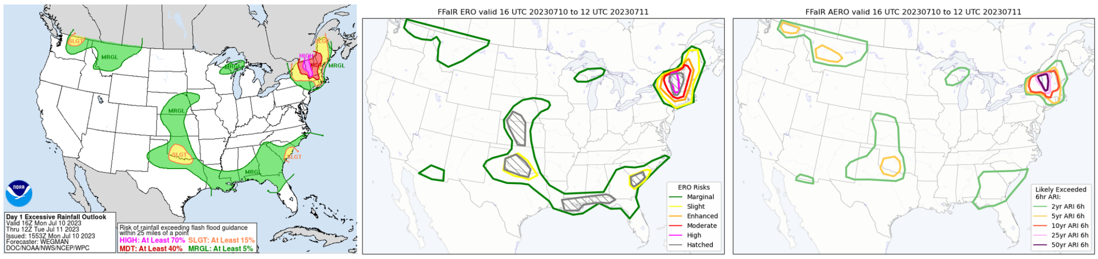
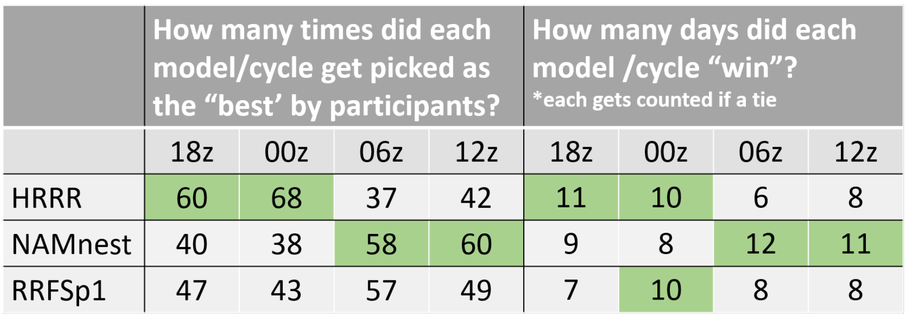


Figure 13: The Day 1 [LEFT] WPC ERO, [MIDDLE] FFaIR ERO and [RIGHT] FFaIR AERO valid 16 UTC 10 July to 12 UTC 11 July 2023.

4.1.1 24-h QPF

Subjective evaluation for 24-h QPF was only done for the HRRR, NAMnest and RRFSp1 valid ending at 12 UTC, for their 18z (fhr42), 00z (fhr36), 06z (fhr30) and 12z (fhr24) cycles, therefore only dates in which the RRFS was available will be discussed. Table 3 shows the results when participants were asked to pick which of the three models had the best forecast for a given cycle while Figs. 14 and 15

Table 3: Results of the 24-h QPF subjective evaluation when participants were asked: “If you had to pick the best {cycle} forecast, which model would it be?”



	How many times did each model/cycle get picked as the “best’ by participants?				How many days did each model /cycle “win”? *each gets counted if a tie			
	18z	00z	06z	12z	18z	00z	06z	12z
HRRR	60	68	37	42	11	10	6	8
NAMnest	40	38	58	60	9	8	12	11
RRFSp1	47	43	57	49	7	10	8	8

summarize the results of the numeric scores given by participants when evaluating the “goodness” of a model/cycle. When choosing a favorite, the HRRR was picked the most for its performance for the 18z and 00z cycle while the NAMnest was picked the most for the 06z and 12z cycles. That said, the 06z cycle the RRFSp1 was chosen only one less time than the NAMnest, 57 vs. 58 respectively. Surprisingly, despite the large disparity between the 00z HRRR (chosen best 68 times) and the RRFSp1 (chosen best 43 times), when breaking it down in terms of the number of days the model/cycle was picked the most by the participants for the day/cycle, the two models tied with 10 days each. For the other cycles, the model chosen most often was also the model that “won the day” the most. Finally, the HRRR and NAMnest never tied for the “best” model. This differs from the RRFSp1, which tied with the NAMnest 5 times and the HRRR 3 times across the evaluated days/cycles. This suggests that overall the participants felt that even when the RRFSp1 performed well, it was less likely to outright be the best model for the day/cycle.

Similar results were found when using the mean of each model/cycle’s subjective score to determine the number of times a model/cycle combination was the daily winner and the number of times a model had highest average daily score for a given cycle[10]; the left side and right side of Fig. 14 respectively. When looking at the “winner” per cycle, like seen for “pick a winner” in the previous table, the HRRR was the “best” most often for the 18z and 00z cycles, beating out the RRFSp1 at

[10] If there was a tie in the average, each model in the tie was counted as a winner of that day/cycle.

00z by 1 (11 vs 10 times being the cycle winner). The NAMnest again had the most "wins" for both the 06z and 12z cycles. However, in the case of the NAMnest winning, the HRRR and RRFSp1 looked more competitive for the two cycles than when the "pick a winner" method was applied. When evaluating the daily winner across all cycles for the RRFS Dates (Fig. 14A), the 12z NAMnest had the most "best" forecasts with 5 times, followed by the 12z RRFSp1 with 4 times.

Despite the NAMnest being the daily winner the most times for the 06z and 12z cycles, when looking at the mean across the entirety of the RRFS dates, the RRFSp1 edged out the NAMnest for these cycles with an average of 6 vs 5.888 and 6.266 v 6.076 respectively; the blue stars in Fig. 15 signify the highest mean for each cycle. The HRRR has the highest mean for the 18z (5.655) and 00z (6.102) cycles. Interestingly, despite not having the highest mean for the 06z and 12z cycles, the highest score most often chosen (7 out of 10) was greater than that for the RRFSp1 and NAMnest (6 out of 10). Additionally, even though the HRRR had a high percentage of scores of 7, it's mean decreased from 6.181 for the 00z cycle to 5.783 and 5.82 for the 06z and 12z cycles respectively. This tracked the comments that participants made during discussion, stating that they often felt the 00z run of the HRRR was the best cycle for the model. They also noted that the HRRR seemed dry, so even though it may have had a good overall rainfall footprint they would lower its rating due to the sometimes significant under forecast of the coverage of higher rainfall totals.

Opposing this, the mean subjective scores increased as the fhr decreases for both the NAMnest (5.444, 5.594, 5.888 and 6.076) and the RRFSp1 (5.594, 5.676, 6, and 6.266). They were also more similar to one another when looking at the percent of good (≥7), average (5 and 6) and bad (≤4) scores. That said, the RRFSp1 always had a higher percentage of 6's for the RRFS Dates than NAMnest and for all cycles the RRFSp1 had an edge on the NAMnest for the percent of times it was scores 7 or higher. For the low end scores, it was cycle dependant on which of these two models had a higher percentage of scores 1-4. This suggests that the performance of the RRFSp1 more closely followed that of the NAMnest than the HRRR.

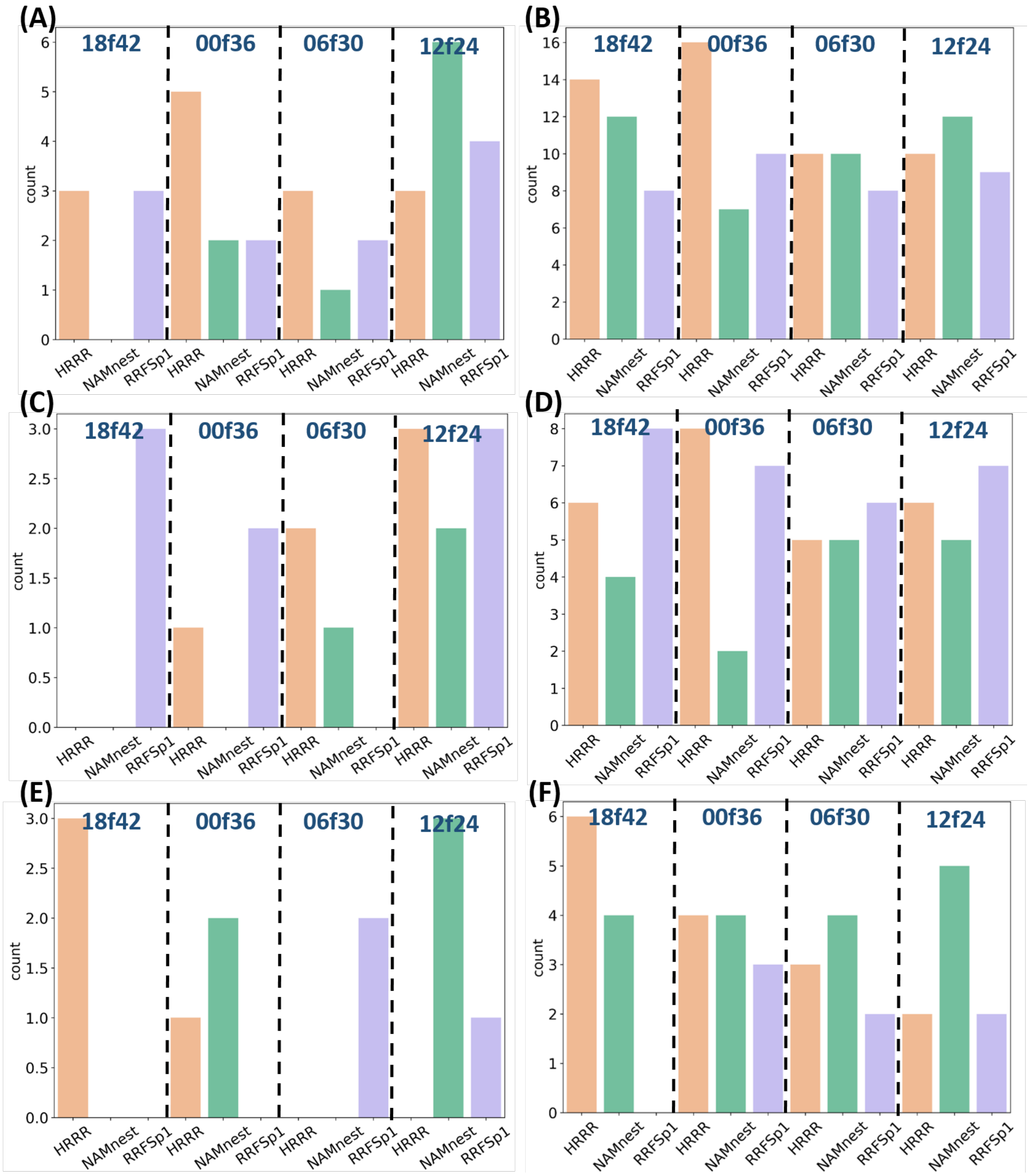


Figure 14: The number of times each model had the highest average subjective score for the day for the HRRR (orange), NAMnest (green) and RRFSp1 (light purple), for the 18z, 00z, 06z and 12z cycles valid at fhr42, fhr36, fhr30, and fhr24 respectively. (A), (C), and (E) are the number of times a model/cycle combination had the highest average score for the verification day regardless of cycle. (B), (D), and (F) are the number of times a model had highest average daily score for a given cycle. Number of dates or dates times cycle along with the dates evaluated are as follows: RRFS Dates: (A) n = 26 and (B) n = 102, CONUS Dates: (C) n = 17 and (D) n = 64, and NA Dates: (E) n = 9 and (F) n = 38. Note: If there was a tie, then each model was counted as a winner, therefore the summed counts across the models/cycles will not always be equal to n.

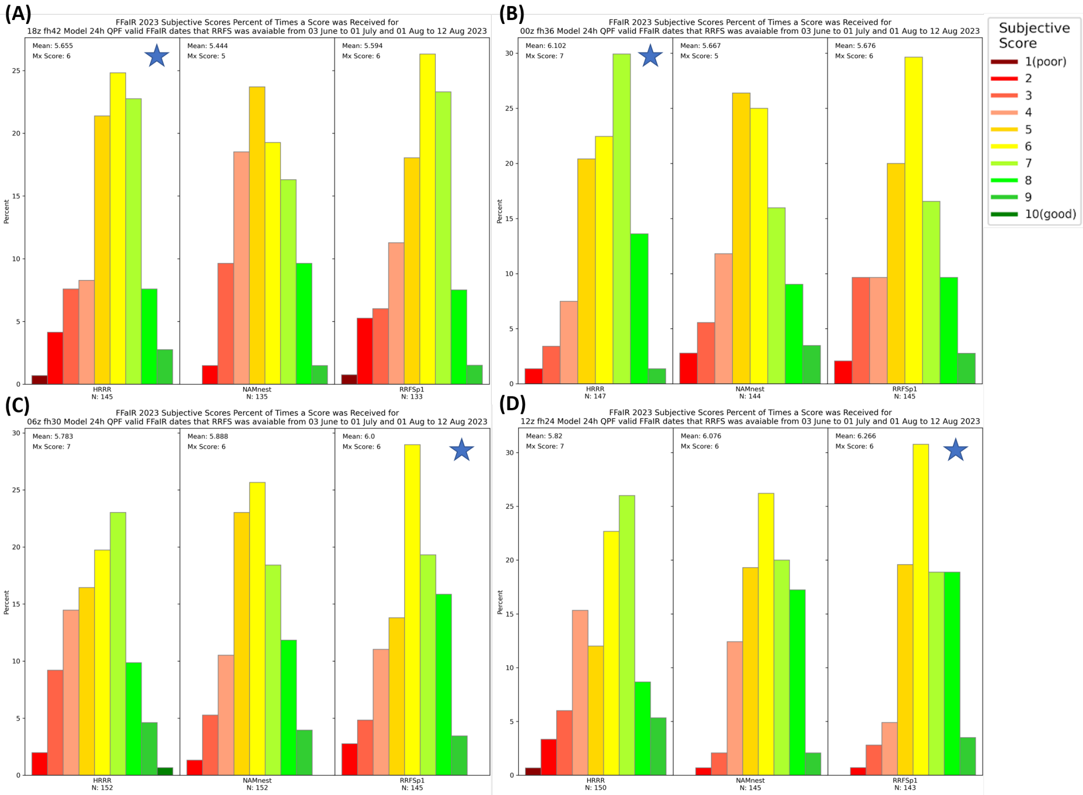


Figure 15: Results from the subjective verification for 24 h QPF showing the percent of the time each model received a score from 1 (dark red) to 10 (dark green) for the RRFS Dates for the (A) 18z, (B) 00z, (C) 06z and (D) 12z cycles valid at fhr42, fhr36, fhr30, and fhr24 respectively. For each panel, the models are as ordered from left to right: HRRR, NAMnest, RRFSp1. The number of scores received (N) is plotted below the model name. The score received the most and the mean score for each model is plotted long the top. The model with the highest mean for each cycle has a blue star.

In terms of pre-/post-domain change, when examining Figs. 14C-F, the RRFSp1 scored highest for the CONUS dates. while the NAMnest was most likely to win for the NA dates. When evaluating based on winning scores for the two date ranges (Fig.16 , there was an increase in the perceived "goodness" of the NAMnest and RRFSp1 from the CONUS dates to the NA dates, while the HRRR's performance worsened between the two time frames. For instance, for the percent of scores (≥7) the 00z/12z cycles of the NAMnest and RRFSp1 increased from 20.5%/40.5% and 33%/23% for the CONUS Dates to 35%/47% and 23%/40.5% respectively. On the other hand, the HRRR's percentage for the same scores and cycles decreased slightly from 46%/44.5% to 39.5%/40.5%.

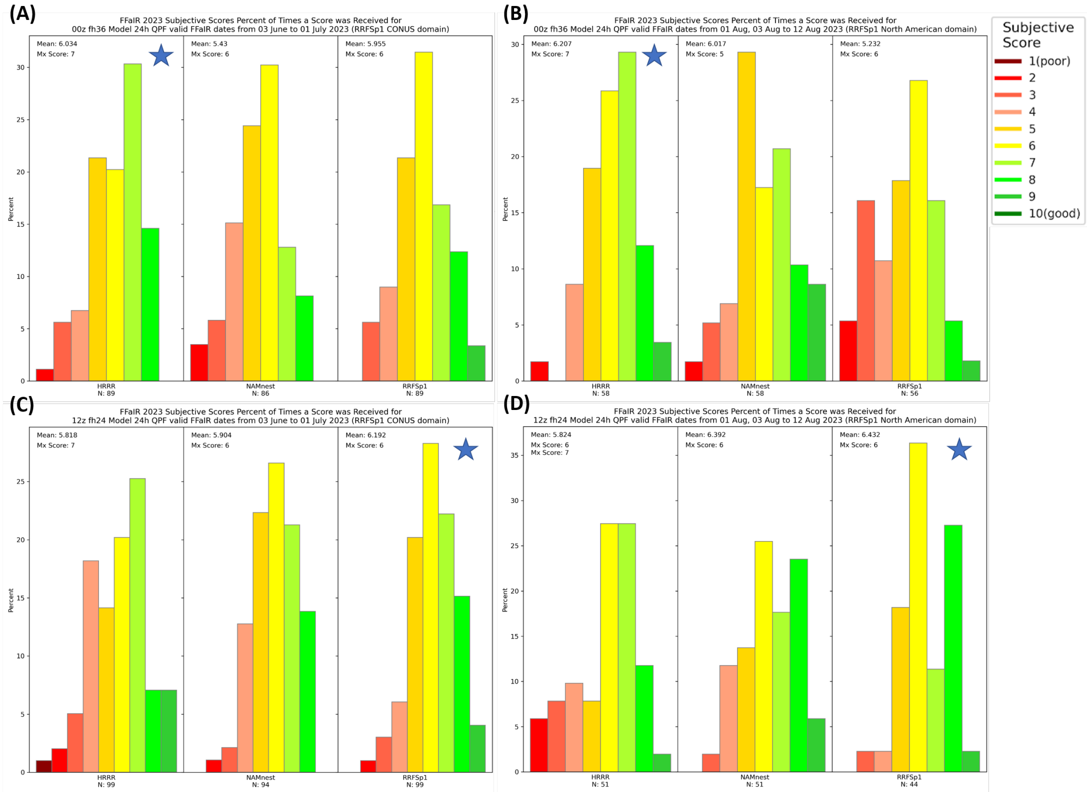


Figure 16: Like Fig. 15 but for (A)-(B) the CONUS and (C)-(D) NA Dates. (A) and (C) are the 00z cycles and (B) and (D) are the 12z cycles.

The goal of looking at the performance pre-/post-domain change was to see if changes made to the RRFSp1 during this time impacted the RRFSp1 performance. However, seeing as how there were also notable differences in the operational models' performances between the two date sets, it is likely that the driving factor seen in the change was due to the differences between types of events seen pre-/post-domain change. Aside from the number of NA Dates (9) being smaller than the CONUS Dates (15), generally the CONUS domain saw smaller events, both in terms of precipitation coverage and magnitude (i.e less extreme) than the NA domain dates. The NA dates saw more precipitation coverage across the country, with MCVs larger in magnitude and coverage (i.e. more extreme). For reference, in terms of the MRTP domain pre-/post-change, the number of days that had 1" coverage >30k km^2 was 2/15 days for the CONUS Dates and 6/9 days for the NA Dates. In general, FFaIR has noted that the HRRR recently has tended to under perform in weather regimes like those seen in the NA Dates, whereas the NAMnest tends to perform better. With a jump seen in the RRFSp1 performance similar to the NAMnest, this suggests that participants felt that the RRFSp1 performs better in regimes that drive MCSs/organized convection, at least when looking at 24-h precipitation totals.

That said, participants noted that the RRFSp1 tended to do a good job at identifying the overall rainfall footprint across the CONUS, though slightly displaced. They also felt that despite a reasonable footprint, they did not trust the magnitudes forecasted, stating they felt the coverage of higher totals was too large. Therefore, another possibility for why the RRFSp1 saw an increase in the subjective scores ≥ 7 could be because since it tends to have a high magnitude bias, it appears to capture more extreme events. This differs slightly from the NAMnest, which also has a known wet bias, in that the NAMnest is not known for excelling at the rain/no-rain footprint but is good at identifying the potential for overnight MCS/MCV/training/backbuilding events. For example, for the first in the series of events that impacted the MO region, the NAMnest was the first to identify the threat of convection developing overnight and diving south across the middle of the state while neither the HRRR or RRFSp1 suggested heavy rainfall would occur[11]; see Fig. 17. Furthermore, even at 24-h out from the valid end time of the event, the HRRR and RRFSp1 still struggled to identify the risk; see Fig. 18.

In summary, when focusing on 24-h QPF subjective verification, the participants felt that the performance of the RRFSp1 fell somewhere between the HRRR and NAMnest. The overall feeling was that the RRFSp1 was useful in terms of highlighting the general rainfall footprint (rain versus no rain). Many participants noted that they felt the RRFSp1 rainfall coverage (both 0.01" and higher thresholds) was more likely to be displaced from the MRMS QPE than the HRRR's. It was also almost always wetter, both in terms of magnitude of maximum amounts and coverage of 1+ inches, which reminded the forecasters of the NAMnest. Additionally, the look of the precipitation pattern in the RRFSp1, from which the evolution of the convection can be inferred, differed from the operational

[11] This was not unique to this cycle, both the RRFSp1 and HRRR struggled to identify any risk of heavy rainfall at or around this time no matter the cycle.

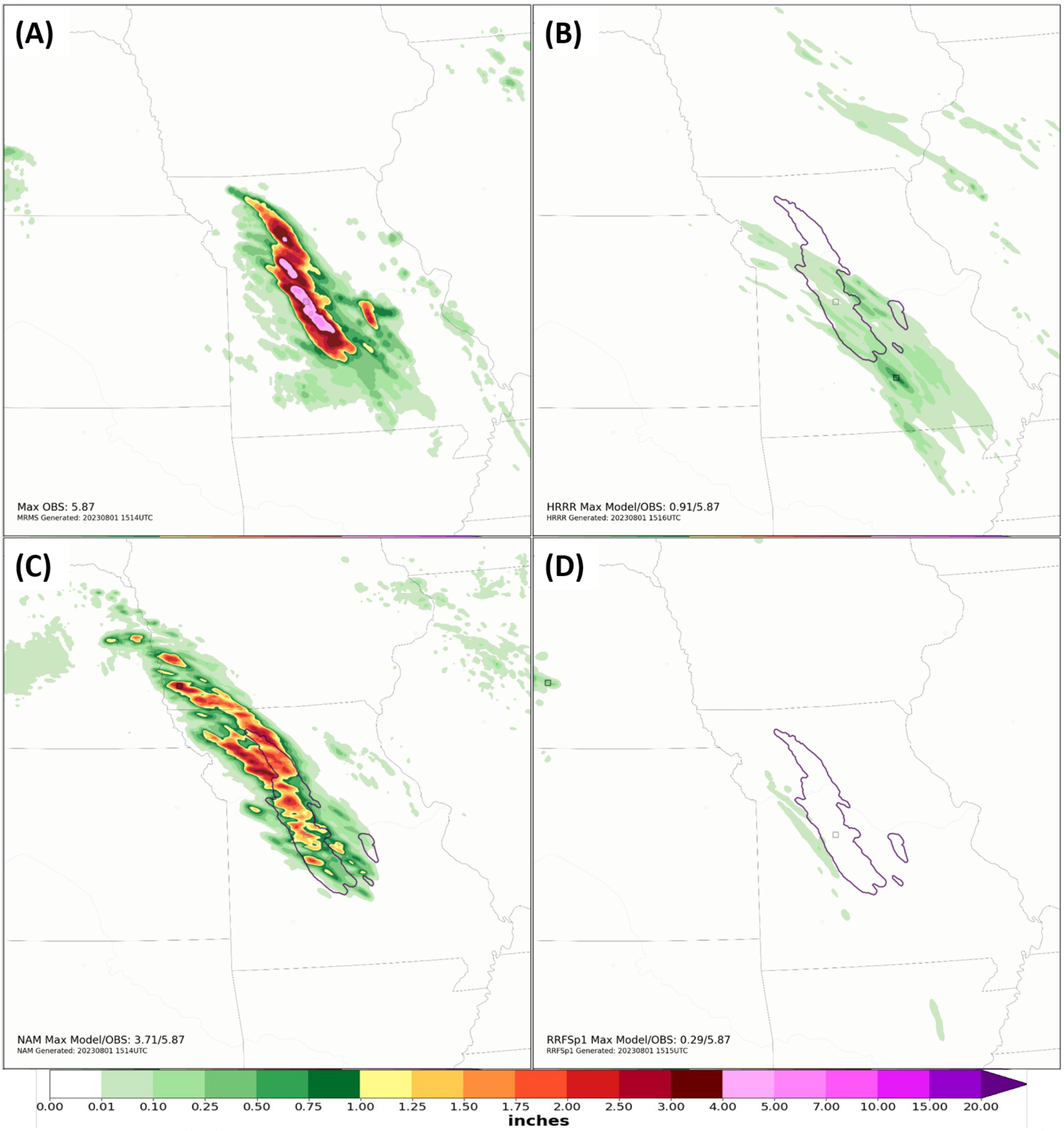

(A)
Max OBS: 5.87
MRMS Generated: 20230801 1514UTC
(B)
HRRR Max Model/OBS: 0.91/5.87
HRRR Generated: 20230801 1516UTC
(C)
NAM Max Model/OBS: 3.71/5.87
NAM Generated: 20230801 1514UTC
(D)
RRFSp1 Max Model/OBS: 0.29/5.87
RRFSp1 Generated: 20230801 1515UTC
0.00
0.01
0.10
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
2.50
3.00
4.00
5.00
7.00
10.00
15.00
20.00
inches

Figure 17: 6-h (A) MRMS QPE and the 18z cycle from 30 July 2023 QPF for the (B) HRRR, (C) NAMnest, and (D) RRFSp1 valid 06-UTC to 12-UTC 01 Aug 2023. The purple contour in (B)-(D): the MRMS QPE 1” contoured.

models and MRMS. The NAMnest seemed more realistic in this aspect, even with its known high bias, according to the participants, with many comments following the theme of this statement: “it (RRFSp1) presents unrealistic structures that don’t align with how I expect severe storms to look in modeled environments.” This included how the precipitation accumulated within the 24-h QPF footprint.

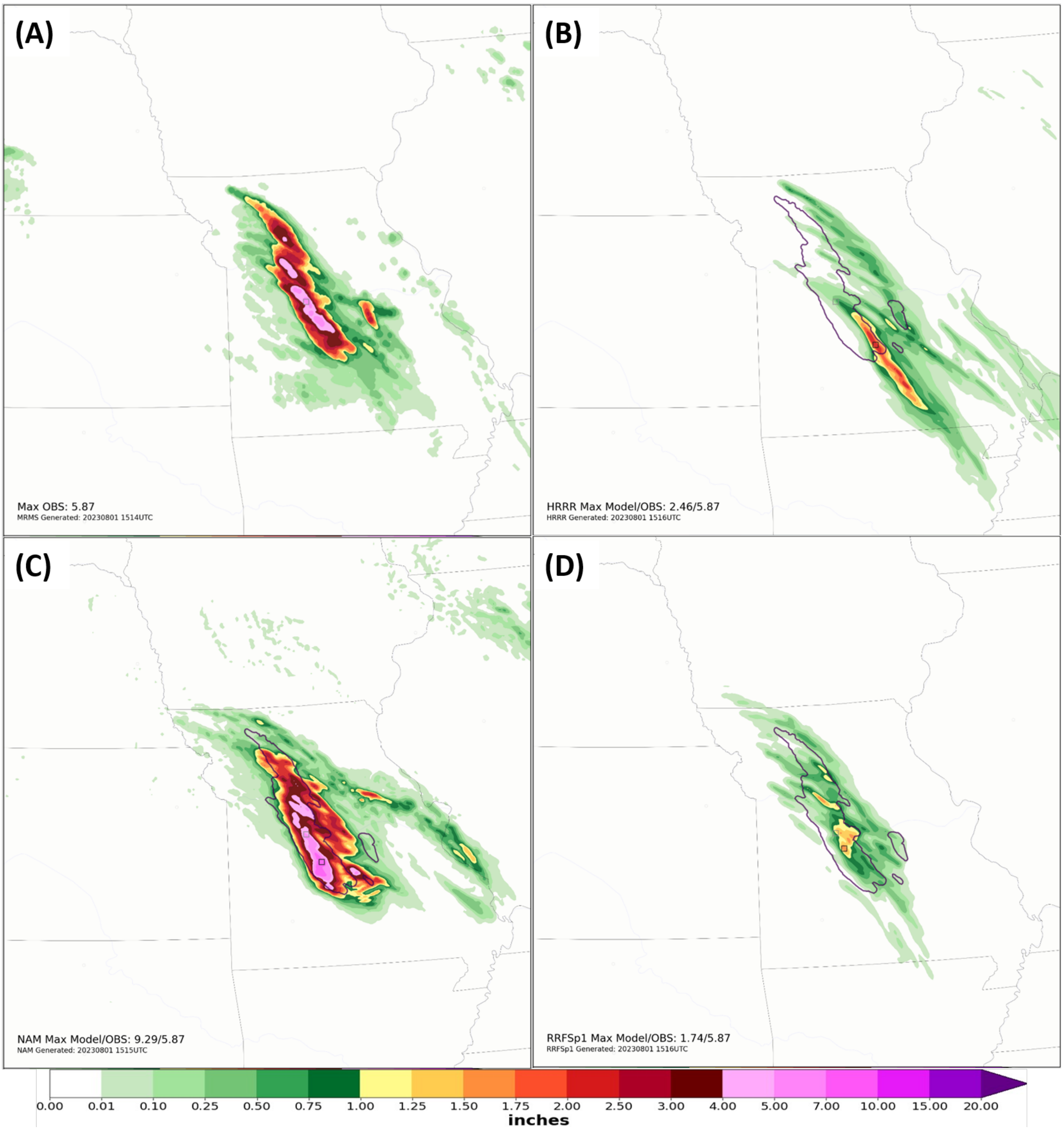

(A)
Max OBS: 5.87
MRMS Generated: 20230801 1514UTC
(B)
HRRR Max Model/OBS: 2.46/5.87
HRRR Generated: 20230801 1516UTC
(C)
NAM Max Model/OBS: 9.29/5.87
NAM Generated: 20230801 1515UTC
(D)
RRFSp1 Max Model/OBS: 1.74/5.87
RRFSp1 Generated: 20230801 1516UTC
0.00
0.01
0.10
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
2.50
3.00
4.00
5.00
7.00
10.00
15.00
20.00
inches

Figure 18: Same as Fig. 17 but for the 12z cycle from 01 Aug 2023.

Switching from subjective to objective verification, the 24-h QPF performance diagrams can be seen in Figs. 19 and 20. For these, even though the FFaIR participants subjectively evaluated only the HRRR, NAMnest, and the RRFSp1's 24-h, the performance diagrams for the 00z cycle include the OU CAPS RRFS-like deterministic members (RRFSp3-RRFSP7), the FV3 member of the HREF (FV3

HREF), and a MPAS-NSSL configuration[12]. When looking across the RRFS Dates, going from the 00z cycle to the 12z cycle, the HRRR's dry bias was worse at the 12Z cycle for the half and one inch thresholds. This agrees with what participants noted, that they felt the HRRR was often drier at 12z than at 00z. On the other hand, the RRFSp1's bias increased from the 00z to 12z cycle at all thresholds. This increase resulted in the RRFSp1's bias being nearly identical to the NAMnest's at 12z; the NAMnest's bias remained roughly the same between the two cycles.

At 00z, the RRFSp1's Critical Success Index (CSI) falls between the HRRR and NAMnest until the 3" threshold, after which its performance compared to the operational models becomes more mixed. At the 3" threshold, the RRFSp1 has a higher CSI than both models, while the NAMnest seemingly has no skill. But by 5" (not shown) the RRFSp1 and NAMnest performance is the same, albeit not good but greater than 0 (bottom left of diagram), which is where the other models lie. However, it is important to note that the sample size for the higher thresholds is small. For 12z, the RRFSp1's bias increased to be similar to the NAMnest, but with a slightly higher CSI; both models still scored slightly worse than the HRRR. Like at 00z, at 3+ inches there is variability in the models' performance against once another, though again, there are very few events at these thresholds. One thing that remains constant is that the RRFSp1 and NAMnest continue to have nearly identical bias for the 12z cycle, with the RRFSp1 minimally wetter than the NAMnest at these thresholds. When separating into the CONUS and NA date groups, Fig. 20, two differences jump out. First, there is an the increase in the overall performance of the NAMnest from Pre- to Post-change, especially for the 12z cycle. This change follows what was seen in the subjective verification. In fact, the NAMnest outperforms the other two models in terms of POD for both cycles at thresholds $\leq$2 in. Second is the RRFSp1's bias as it relates to the operational models. Pre-change, (the left 2 columns in Fig. 20), the RRFSp1's bias was either the same or greater than the NAMnest for both the 00z and 12z and at all thresholds, but with a higher POD/CSI. However, when looking at the Post-change Dates, differences can be seen between the two cycles and thresholds.

[12] Reminder, the FV3-HREF and the MPAS-NSSL were not part of the formal evaluation for FFaIR; see Section 2.4

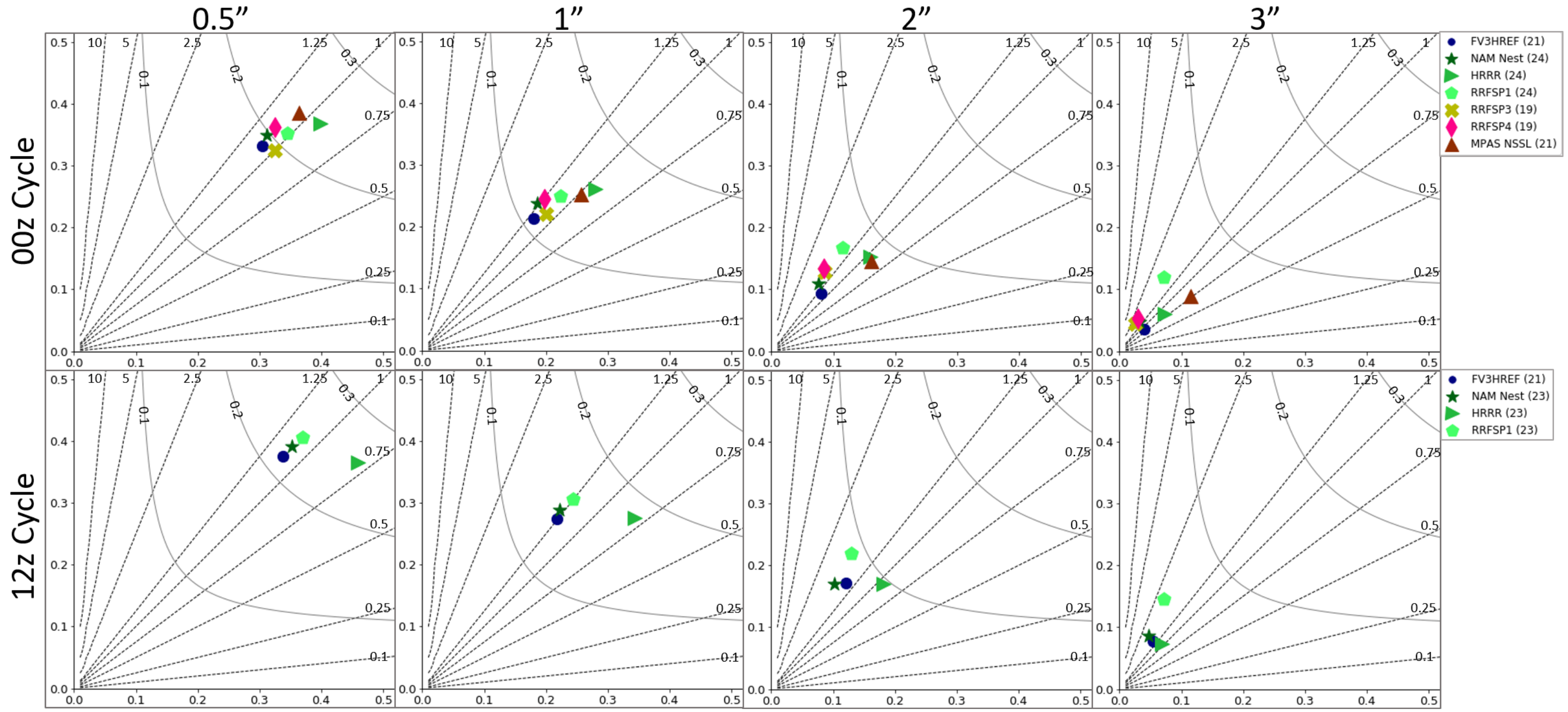


Figure 19: The 24-h performance diagrams for the RRFS Dates (FFaIR in-session only dates) during the 2023 FFaIR Experiment. The 00z cycle is along the top and the 12z cycle is along the bottom. From left to right, the thresholds evaluated are: half inch, one inch, two inches and three inches. For the 00z cycle the models evaluated are: FV3HREF (blue circle), NAMnest (dark green star), HRRR (green triangle to right), RRFSp1 (light green pentagon), RRFSp3 (yellow x), RRFSp4 (pink diamond), and MPAS NSSL (dark red triangle). The models evaluated at 12z are: FV3HREF (blue circle), NAMnest (dark green star), HRRR (green triangle to right), and RRFSp1 (light green pentagon). In the legends to the right of the performance diagrams are the number of forecasts that went into the analysis for each model and cycle in the ()’s next to the model name.

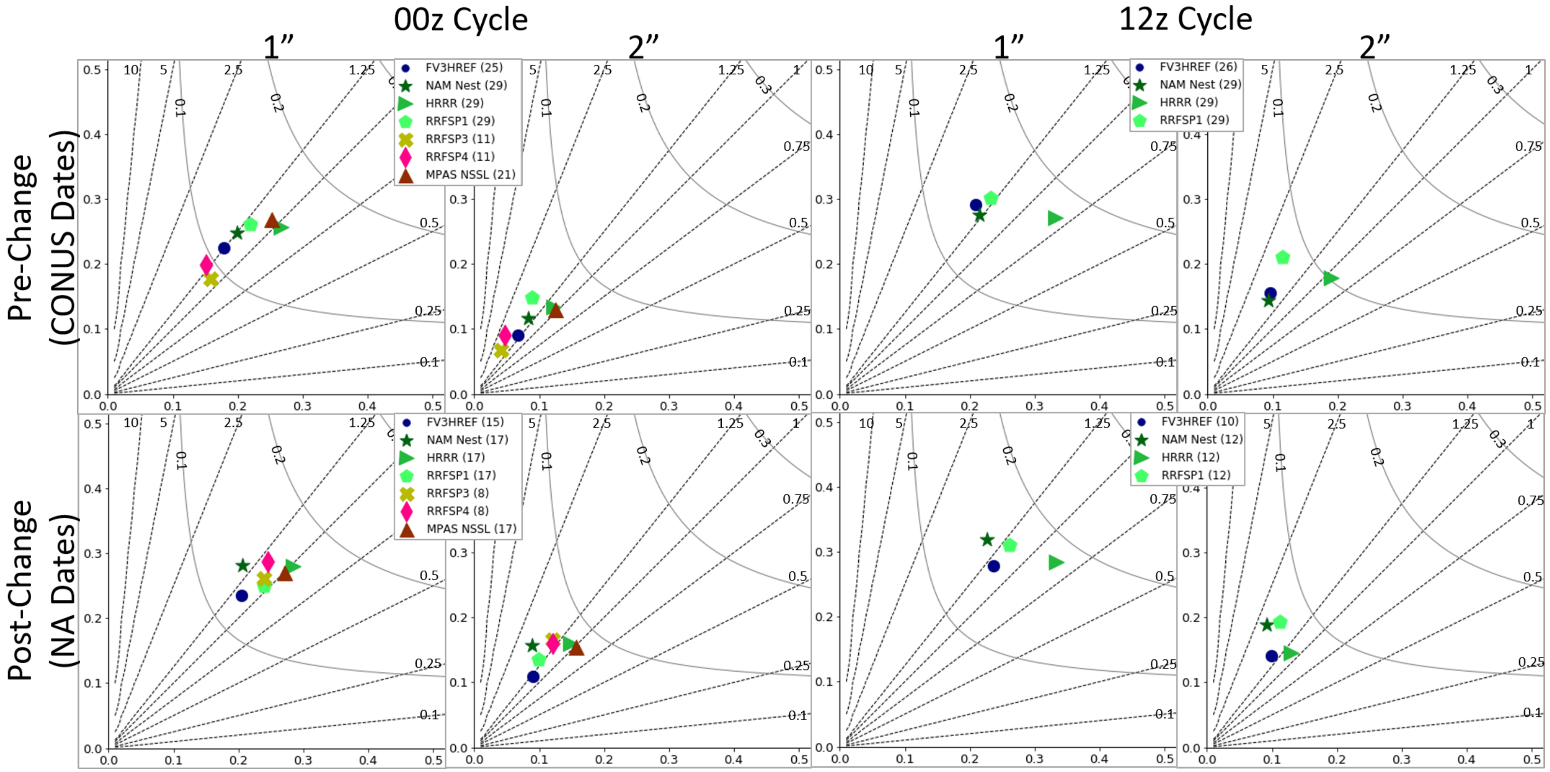


Figure 20: Similar to Fig. 19 but for the Pre-Change (CONUS) Dates and the Post-Change (NA) Dates, though not unique to FFaIR in-session dates. The 4 left panels are the Pre-Change Dates and the 4 right panels are the Post-Change Dates. Columns 1 and 3 are for the one inch threshold and columns 2 and 4 are for the two inch threshold. Icons for the models are the same as listed in Fig. 19

At 00z the RRFSp1's bias shifts to be closer to the HRRR's for the half and one inch thresholds, even having a slight dry bias at half inch (not shown). For the same thresholds but for the 12z cycle, the RRFSp1's bias also shifts to be drier than the NAMnest, but still remains closer in magnitude to the NAMnest's bias than the HRRR's. For both cycles, at thresholds ≥3 in the bias of the RRFSp1 and NAMnest are nearly identical, like that seen in the CONUS and RRFS Dates.

Overall, both the subjective and objective verification, which was done over the full CONUS for 24–h verification, would suggest that the RRFSp1's performance lies somewhere between the NAMnest and the HRRR. When looking at the characteristics of the performance diagrams, there was a shift in how the models performed relative to one another depending on the cycle and over what date group was examined; this was very similar to what was noted in the subjective results. Although the three models saw a general increase in CSI from 00z to 12z, there was a shift in their biases between the two cycles; the HRRR trending overall drier, the RRFSp1's trending overall wetter and the NAMnest's staying relatively similar. One constant in the performance is the NAMnest was nearly always positioned to the lower left of the HRRR and RRFSp1, thus one could infer that the forecasts were generally worse. However, subjectively the participants often picked the NAMnest as the best forecast. This discrepancy most likely is driven by three things: (1) the participants' ability to factor in what they consider to be the important aspects of the forecast. (2) How useful they feel the forecast is. (3) Unlike when using the performance diagrams they are not evaluating the forecast at specific thresholds but rather at all thresholds simultaneously, thus comparing how the higher magnitude QPF pattern fits within the lower values and their conceptual model of what the precipitation pattern should look like given how the event evolved.

Using the 24-h period from 12-UTC 31 July to 12-UTC 01 Aug 2023 as an example to show how (1) and (2) could play out, the 24-h MRMS QPE and the HRRR, NAMnest, and RRFSp1 12z QPF can be seen in Fig. 21. Looking at the MRMS QPE, it is clear there was a heavy rainfall event in MO. Other rainfall was generally scattered with pockets of strong convection. For this event, participants felt the NAMnest performed the best in MO, stating that it nearly nailed the magnitude and location of the event while the other two models seemed to struggle. The only other location in the CONUS they discussed was over CO and this discussion was minimal. This suggests that a large part of what the participants used to evaluate the model's forecast was the MO event. The winner for this cycle was the NAMnest, with 4/5 of the participants that were assigned to review the NAMnest this cycle voting for it; all participants were part of the oral discussion that was aforementioned. However, when looking at the CSI for the day (in the bottom left of each model's QPF image) one can see at the three models were comparable to one another at the quarter, half, and one inch thresholds. This suggests that looking at the performance diagram alone is not enough to understand the true goodness or utility of a model/forecast.

Although this is just one case, the discussions during the verification sessions generally followed suit. Generally, if there was an area that the group (or participant) felt was the most important to get right, the best model was the one that was closest to getting that event right[13]. Objective analysis does not work like this and thus will not always align with the subjective results. Therefore it is important to not only rely on one method or the other to evaluate model performance; instead work to meld what the two methods are showing to understand the complete picture of model performance.

4.1.2 6-h QPF

As stated in Section 2.6, the subjective evaluation encompassed all the operational and experimental models that were available. However, aside from the RRFSp1, the experimental data was only provided by the model developers for the 00z cycle. It also differed from the 24-h QPF verification in that was was not done over the entire CONUS (with a few exceptions) but rather over the MRTP region. Also, the 6-h time frame was not constant, it changed depending on the valid time for the MRTP; see Table 2 for the verification periods.

[13] There are some exceptions of course. For instance if the NAMnest had no rain anywhere else in the domain but over MO, it is highly unlikely that the participants would have picked this model as the winner.

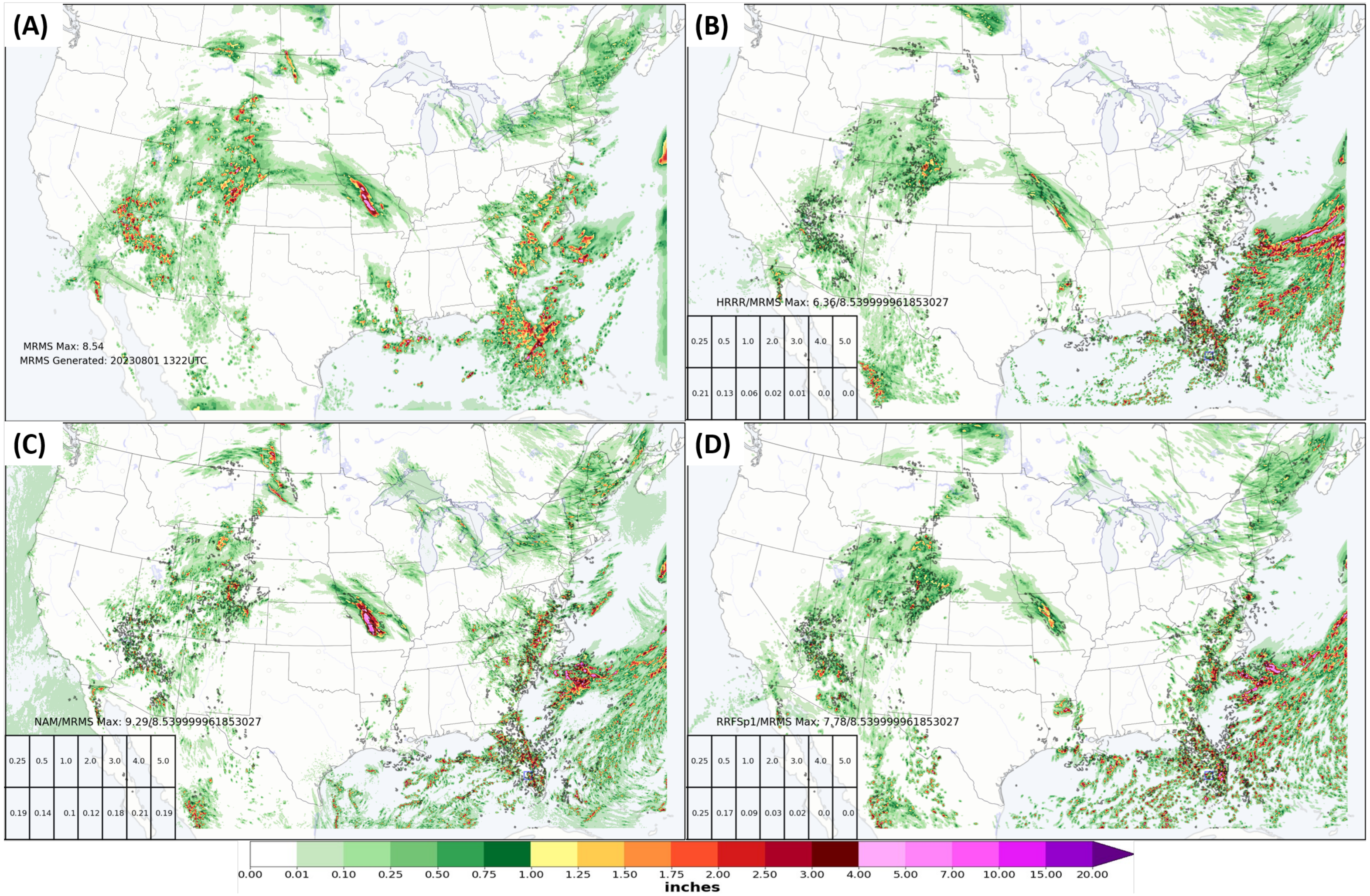


Figure 21: 24-h (A) MRMS QPE and the 12z cycle QPF for the (B) HRRR, (C) NAMnest, and (D) RRFSp1 valid 12-UTC 31 July to 12-UTC 01 Aug 2023. In dashed black in (B)-(D) is the MRMS QPE 1" contoured. In the bottom left of (B)-(D) are the given model's CSI for the 0.25, 0.5, 1, 2, 3, 4, and 5 inch threshold.

Evaluating the models in terms of number of times they had the highest average score for a given cycle/day (Fig. 22) across all 6 weeks of FFaIR, every model but the RRFSp7 was a cycle ”winner” least once. For 00z, the HRRR was the most likely to have the highest average score (10 times), followed by the NAMnest (9) and the RRFSp1 (7). Of the CAPS deterministic members and the NSSL-MPAS, the RRFSp4 had the most “wins” with 3. For the 06z cycle, the RRFSp1 had the highest average score the most times at 12 days, with the HRRR and NAMnest tying with 9 times each. For the 12z cycle, the HRRR lead the way with 12 times, followed by the NAMnest (10) and the RRFSp1 with 6. A similar pattern is seen when only looking at the RRFS Dates, with slight changes to how the HRRR and NAMnest compared to one another at 00z and 06z, with the HRRR and NAMnest tying for most “wins” at 00z and the NAMnest pulling ahead of the HRRR for number of winds at 06z. Lastly, just as was seen with the 24-h QPF evaluation, the NAMnest went from winning the least amount of times at the 12z cycle for the CONUS dates (not shown) to winning the most times at the 12z cycle for the NA dates. Interestingly, for the 06z cycle the RRFSp1 won the most for the NA dates when looking at the 6-h verification but won the least when looking at the 24-h verification.

Fig. 23 shows the distribution of the subjective scores for the 00z cycle, for the FFaIR (top) and RRFS (bottom) Date groups. Along the bottom of each chart is the number of times the model received a score. Note that for the FFaIR Dates, the RRFSp1-RRFSp7 all have around the same number of scores while the NSSL-MPAS has a total that falls between the operational and the other experimental models. Meanwhile for the RRFS Dates, the RRFSp3-RRFSp7 (i.e. the CAPS runs) have half the number of scores that the RRFSp1, HRRR and NAMnest have. Again, the NSSL-MPAS totals falls between these two groups. This means that aside from the HRRR, NAMnest, and RRFSp1 for the RRFS Dates, the events that drive the distributions are not comparable among the models.

With that in mind, for the two date groups in Fig. 23 the HRRR had the highest mean (4.9/4.715) followed by the NAMnest (4.514/4.556). Focusing on the All FFaIR dates, the RRFSp4 had the third highest mean with 4.455 followed by the NSSL-MPAS (4.391) and the RRFSp1 (4.243). Even though the RRFSp4

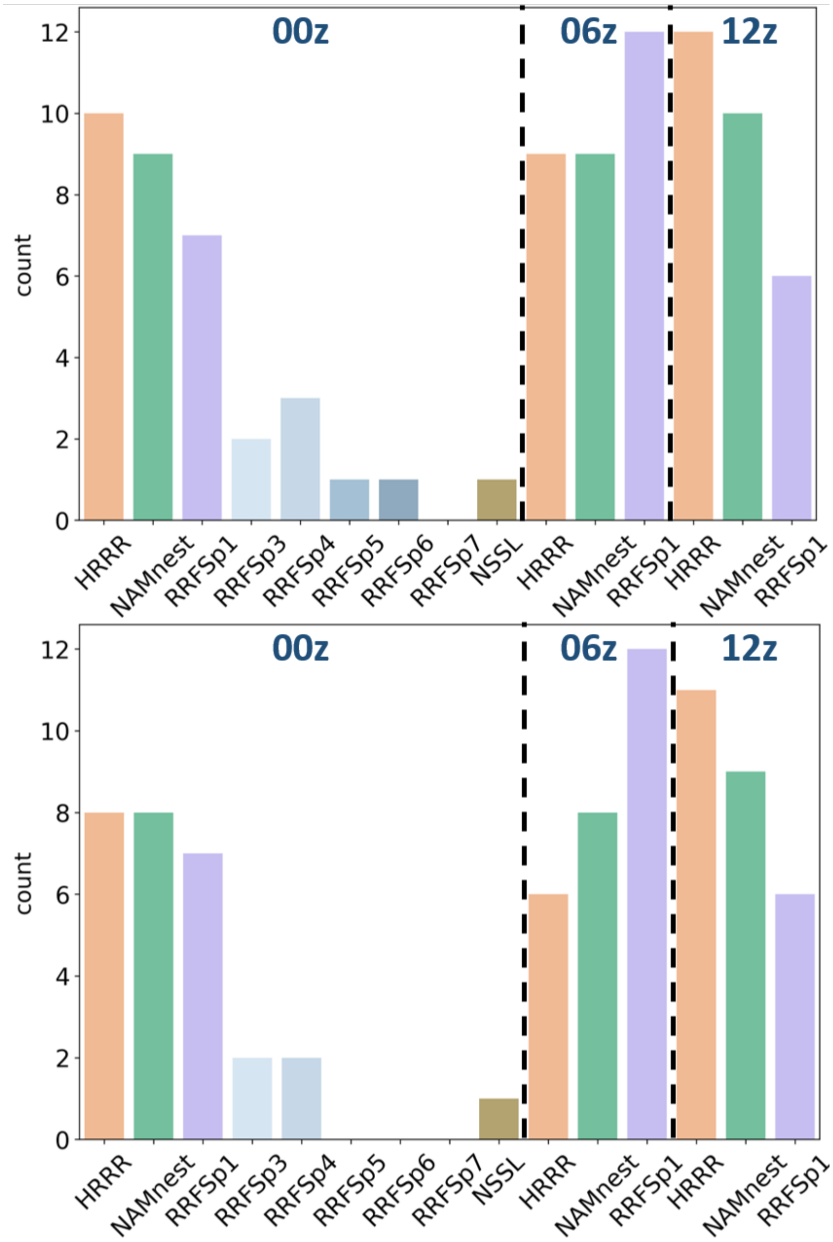


Figure 22: The number of times each model had the highest average subjective score for the day for the 00z, 06z and 12z cycles for their 6-h QPF valid during the MRTP time period for all 6 FFaIR weeks (top) and RRFS Dates (bottom).

has a higher mean than the NSSL-MPAS and the RRFSp1, it was less likely to receive a score of 7 or higher than them, each receiving these scores 15%, 21%, and 17% of the time respectively. On the lower end of the spectrum, the RRFSp4 and NSSL-MPAS scored 4 or less roughly 51% of the time, with the RRFSp1 around 56%. For reference, the HRRR and NAMnest saw scores ≥7/≤4 roughly 22%/43% and 17.5%/56% of the time, respectively. When shifting to the RRFS Dates, the RRFSp4 (3.865) was the best performing of the CAPS models, while the NSSL-MPAS (4.257) and RRFSp1 (4.243) have higher means.

Furthermore, when evaluating the NA Dates (not shown), which is the time period in which the NSSL-MPAS and RRFSp1 have practically the same number of scores, the NSSL-MAPS has a mean of 4.473 while the RRFSp1's mean was 3.491. These were both less than the HRRR (4.517) and the NAMnest (5.034). It is only when CONUS dates (not shown) are evaluated, when there were roughly 30 less scores recorded for the NSSL-MPAS than the operational models and RRFSp1, that the RRFSp1 had a higher average score than the NSSL-MPAS, 4.785 vs. 4.037. The HRRR and NAMnest had an average score of 4.861 and 4.226 respectively for these dates. Interestingly, even though the NSSL-MPAS's mean is higher than the RRFSp1 for all but one date group, it only once had the highest average daily mean for the 00z cycles. It is unclear why the NSSL-MPAS was rarely the "winner" of the day despite it seeming to perform better across the experiment than the RRFSp1, which had a comparable number of "wins" to the operational models. Perhaps it suggests that the NSSL-MPAS is more likely to have a "middle-of-the-road" forecast while the RRFSp1 is more likely to score either quite high or quite low.

For the 06z and 12z cycles, as stated earlier, only the HRRR, NAMnest, and RRFSp1 were evaluated. Their subjective performance for the RRFS Dates can be seen in Fig. 24A-B. For 06z, the average scores for the models were basically within 0.1 of each other. In fact, the HRRR and RRFSp1 were essentially tied with means of 4.645 and 4.65 respectively. For 12z, like what was seen in the 24-h QPF evaluation, the NAMnest (5.014) and RRFSp1 (4.993) pulled ahead of the HRRR (4.786). When going from CONUS to NA Dates, the NAMnest goes from having the lowest mean of the three models to having the highest for both the 06z and 12z cycles. Across all models and both cycles, the subjective average decreased from CONUS to NA Dates, aside from the 12z NAMnest. When looking at Fig. 24C-D, 12z CONUS and NA domains respectively, the HRRR becomes more skewed towards the lower end of the "goodness" scale (to the left). The

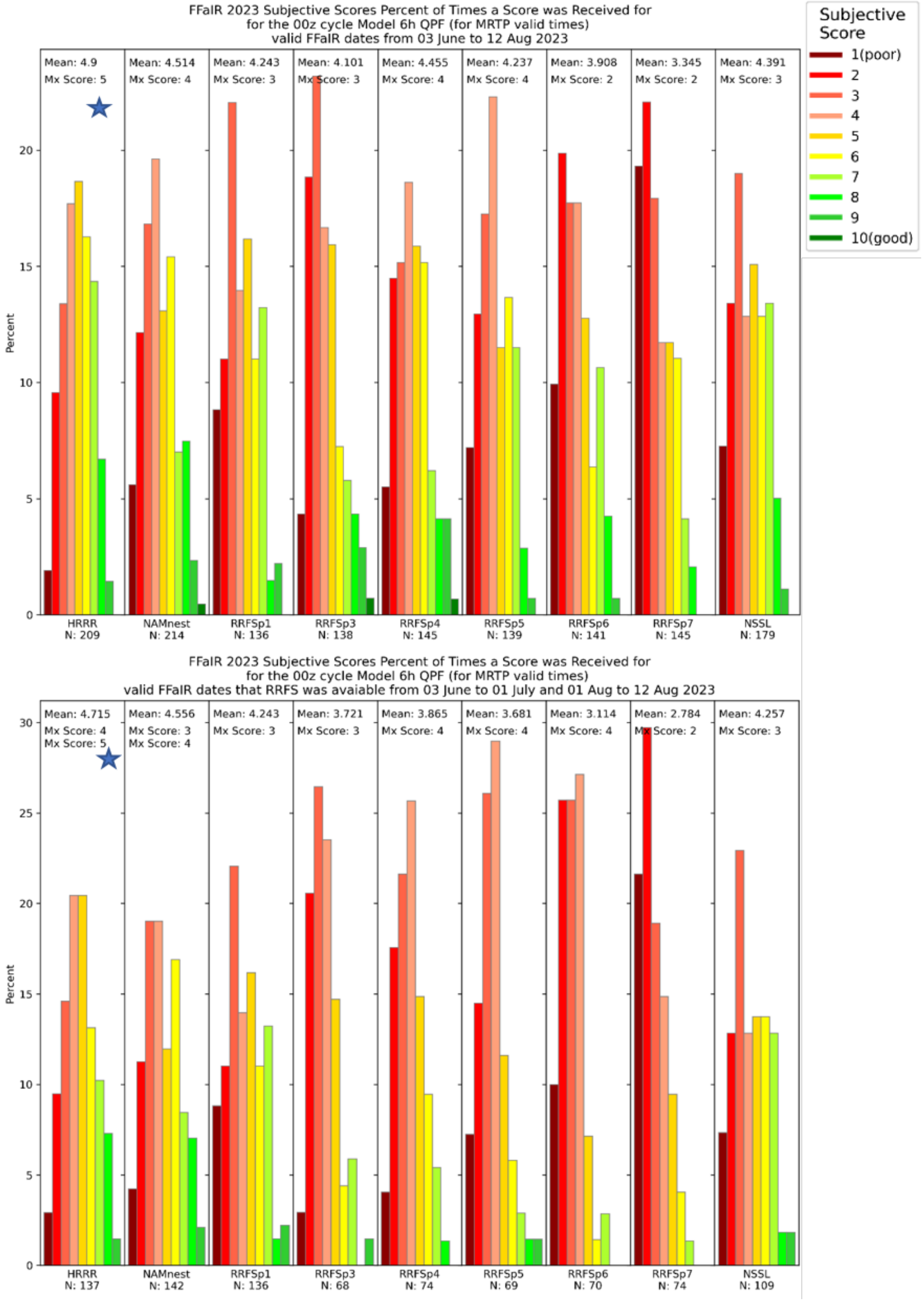


Figure 23: Like Fig. 15 but for 6-h QPF subjective verification scores for [TOP] all 6 FFaIR weeks and [BOTTOM] RRFS Dates.

NAMnest changes to being skewed slightly right, with a jump in the number of times it received a high score of 8 or 9. Meanwhile the RRFSp1 skewness shifted slightly left towards more average and lower end scores, though the change in the distribution of its scores was less exaggerated than seen in the operational models.

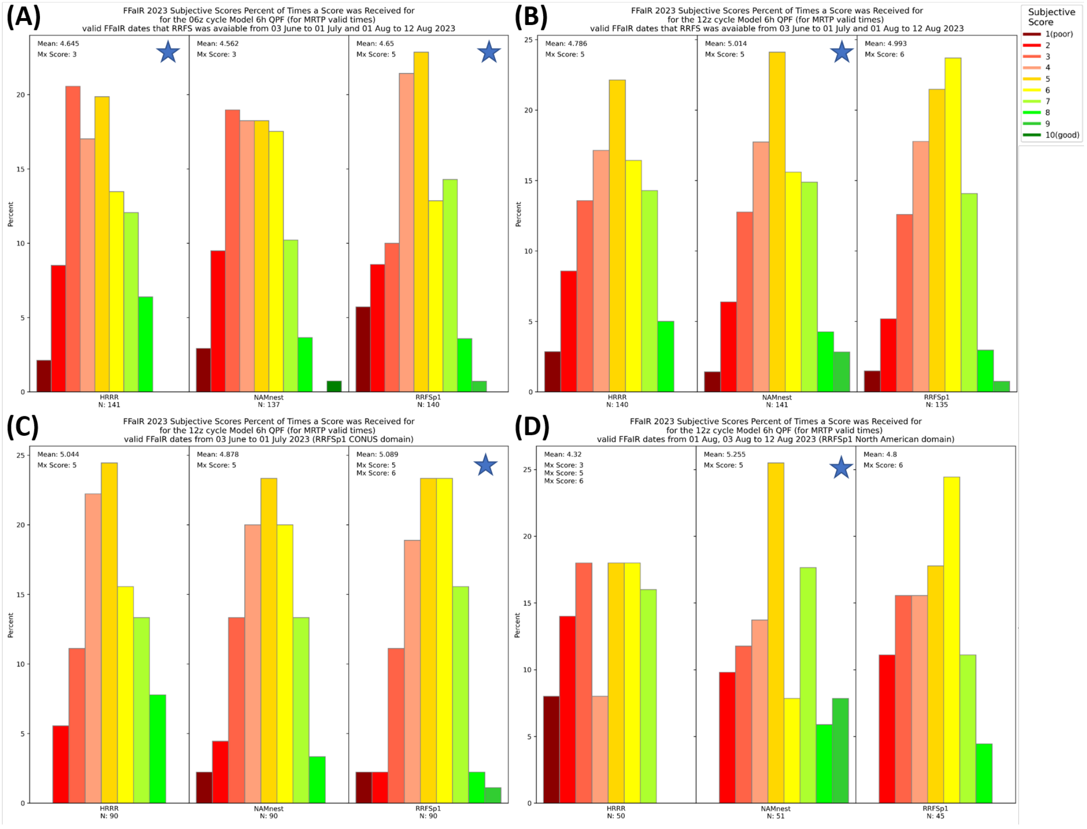


Figure 24: Like Fig. 15 but for 6-h QPF subjective verification scores for the for the RRFS Dates for cycles 06z (A) and 12z (B). (C) and (D) are the 12z cycle score for CONUS and NA Dates respectively. Reminder, for each panel, the models are as ordered from left to right: HRRR, NAMnest, RRFSp1.

As discussed previously, the availability of experimental data was inconsistent, with data missing for a week or more across the FFaIR dates for most of the models, and the CAPS models (RRFSp3-7) and the NSSL-MPAS were only run for the 00z cycle. Additionally, the evaluation of the RRFSp1 was the top priority of FFaIR. Therefore, the 6-h performance diagrams and following discussion will focus on the HRRR, NAMnest, and RRFSp1 for RRFS Dates. The performance diagrams for the four synoptic 6-h windows (12-18, 18-00, 00-06, and 06-12 UTC) can be seen in Fig. 25. Since the performance diagrams only show the synoptic 6-h windows, the 6-h time period which saw the greatest rainfall might span two forecast periods. Additionally, these performance diagrams show scores for the full CONUS rather than the region of interest as was done for the subjective verification. For a summary of how the models/cycles performed for higher end 6-h rainfall events (aka during the MRTP time periods), please referred to Section 4.4.1.2.

Depending on the cycle, 6-h time period, and threshold, the operational model that scored most comparably to the RRFSp1 varies. Therefore it is difficult to make overarching statements about its performance compared to the operational models. However, there are a few things that are worth noting. First is the shift of the RRFSp1 bias from the afternoon (18-00 UTC) to the late night/early morning hours (06-12 UTC). As the forecast shifts across the three 6-h windows, the RRFSp1's bias at all thresholds changes from having a noticeable wet bias (often wetter than the NAMnest) to having a slight dry bias. For example, looking at the one inch threshold (middle row in Fig. 25), the RRFSp1 for all cycles has a bias between 1.5 and 2 for 18-00 UTC but by the last 6-h period (06-12 UTC) the bias is approximately 0.75. Although both the HRRR and NAMnest also tend to shift to a drier bias across the time windows, the shift is less drastic. The NAMnest shifts from left of a bias of 1.25 to right of it, while the HRRR shifts from around 1 to 0.75.

Next is the change in performance across the different cycles. As lead time decreased, which in this case refers to moving from cycle initialization at 18z to 12z, the model performance tended to improve. However, the manner in which the performance got better (aka closer to the upper right corner of the diagram) varied across the models, as well as for the 6-h time periods and thresholds. Overall, for all but the first 6-h window (first column in Fig. 25) the differences between the cycle runs for each model generally were mostly in the SR and POD scores, while bias remained fairly constant. However, for the first 6-h window (12-18 UTC) similar trends are not seen among the models as the cycle initialization changes. The HRRR's forecast performance varies depending on the cycle, with perhaps the only trend being that the 06z forecast is always the worst performing of the 4 forecast cycles across each threshold. Opposing this, the 18z, 00z, and 06z cycles

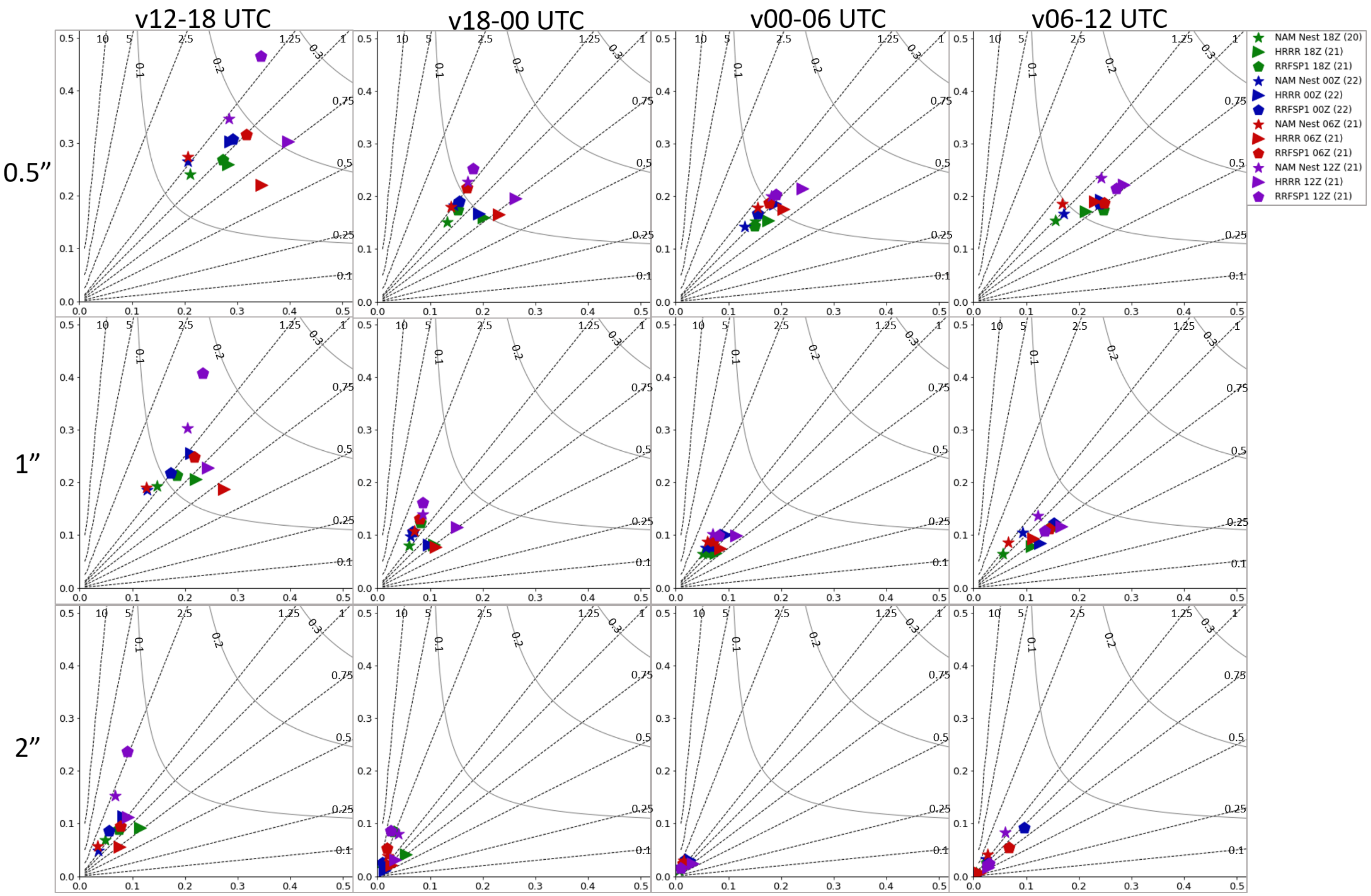


Figure 25: Performance diagrams for 6-h QPF for the HRRR (sideways triangle), NAMnest (star) and the RRFSp1 (pentagon) for their forecasts initialized (ordered from furthest from valid time to closest) at 18z (green), 00z (blue), 06z (red) and 12z (purple). From left to right, the columns are valid for the following 6-h periods: 12-18 UTC, 18-00 UTC, 00-06 UTC and 06-12 UTC. The thresholds shown are 0.5 (top), 1 (middle), and 2 (bottom) inches.

of the NAMnest and RRFSp1 tended to be clustered by model on the performance diagram; though there was more separation among the three cycles for the RRFSp1 than the NAMnest. For their 12Z cycles, each model saw a dramatic increase in CSI.

Continuing the analysis of the model/cycle performance for the 12-18 UTC time period, it is important to note that this time frame is unique in two ways. First, there is generally a lull in precipitation from 12-18 UTC in the warm season, with most overnight convection dying out and daytime heating to initiate convection lacking. Typically, precipitation during this time period is driven by a well developed system, like a MCS or MCV. Models tend to struggle in both the longevity of these features and their propagation speed. This might suggest that the NAMnest and RRFSp1 are relatively constant across the older cycles with evolution of large scale systems as they progress into the overnight hours (based on the clustering), while the HRRR is less consistent. Or that the the HRRR is less likely to maintain MCS/MCV strength.

Secondly, the 12-18 UTC time period represents the first 6 forecast hours of the 12z cycle. The first 6 hours of the forecast are typically when the influences of the DA system are most apparent. Therefore, the impact of the various DA systems and how the models respond to them is likely being seen. As one would expect, each model's CSI is the highest for the 12-18 UTC period, but as noted previously, the jump in CSI was more drastic for the RRFSp1 and NAMnest than the HRRR. Additionally, the RRFSp1 has the highest CSI and POD for the 12z cycle of the three models. At first glance, this jump in CSI would represent a positive outcome, but there are characteristics of this jump that are disconcerting. The increase appears to be driven by an increase in POD with SR remaining relatively static, thus resulting in an increase in the bias relative to the other cycles. The NAMnest, on the other hand, sees an increase in both SR and POD, keeping it's bias similar to the 3 previous cycles. Combined, this suggests that the RRFSp1 does a better job at sustaining ongoing convection that is cycled into the model during its DA process than the operational models. However, this is likely the result of it overdeveloping the convection present at initialization.

This conclusion is supported by the results of the analysis done in the 2023 HWT SFE and in FFaIR on comparing the impacts of DA on the forecast for the HRRR and RRFSp1 (called the RRFS in the 2023 HWT SFE). Though, for FFaIR, as stated in Section 2.6, technical difficulties resulted in the DA evaluation being completed for only the first 3 weeks of FFaIR. The HWT SFE Final Report noted that participants regularly commented that "simulated reflectivity in RRFS was too high, and that RRFS often had spurious storms" (Clark et al., 2023). Similar feedback was given by FFaIR participants, with a large emphasis on the tendency of the RRFSp1 to overdevelop ongoing convection, thus leading to a wet bias. Comments like the ones listed were common during the evaluation:

- "The RRFS seemed to initialize better over the first few hours with location but had a wet bias with high intensity. The HRRR seemed to do better with time and had more realistic intensity, and possibly a slight dry bias."
- "Too many and too intense storm cores in the RRFS compared to the the HRRR, while both still had too much coverage compared to MRMS."
- "The RRFS shows the lower dBZ much better than the HRRR. The RRFS though also runs very hot in convective cells that do form (very high dBZ) whereas the HRRR is a bit more moderated (maybe 55 vs 65 dBZ) for single cells."
- "The RRFS did slightly better than the HRRR at maintaining strength in some of the storms, but I'm not sure if this is a model positive or if it ties into the overall wet bias of the RRFS. Every convective cell appears to max out reflectivity and be much heavier than MRMS obs."

Even though formal evaluations were never performed for the 6-h time window being discussed presently (12-18 UTC), it helps shed light on how DA could be impacting the results seen for the 12z cycle.

The last participant comment listed above emphasizes some of difficulties faced when analyzing the results (both subjectively and objectively) of the RRFSp1. Is a high wet bias and a model that tends to over develop convection acceptable as long as it has a high POD and CSI? What is the bias that would be acceptable? Is the model still providing useful information, even if it has an abundance of erroneous convective strength and precipitation? Is it okay that the model gets something right for the wrong reasons? These questions are not meant to be answered here, but rather to point out considerations that should be discussed when considering if a model should be transitioned into operations.

4.1.3 Additional Analysis of QPF

Like last year, the QPF was examined outside of the traditional contingency table methods. Figure 26 shows the survival functions for 6-h and 1-h QPF for both the 2022 and 2023 Testbed seasons. Comparing the MRMS from 2022 to 2023 reinforces the change in precipitation patterns between the two years, with 2023 seeing more cases of 6-h totals exceeding a foot, but less instances of high hourly totals. For 6-h totals, the count/hour is similar between the two years until roughly 5 in. At this accumulation the number of occurrences decreases faster for 2022 than 2023. For 1-h, counts for the two years remain similar until ∼4.25 in, then the 2023 occurrences decrease faster than 2022.

The HRRR and RRFSp1 for 6-h QPF did not follow what was seen in MRMS, with model maximum QPF decreasing from 2022 to 2023. For the RRFSp1 this resulted in a move closer to the MRMS line from 2022 (MRMS shifted up while RRFSp1 shifted down). These opposing shifts resulted in the difference in the maximum between MRMS and the RRFSp1 decreasing, from ∼7.75 in to ∼4.5 in. The RRFSp1 2023 closely resembled MRMS until roughly 9 in, as opposed to diverging from MRMS at 2.5 in last year. After 9 in the curve becomes flatter than than MRMS, denoting a wet bias. Opposing this, the HRRR shifted from having a slight wet bias beginning around 7 in last year to having a dry bias (lower counts of accumulations than MRMS) until around 15 in, yet its overall slope was similar to MRMS.

For 1-h QPF, both models show a general decrease in hourly totals from 2022 to 2023, which also follows the trend in the MRMS. Additionally, the HRRR and RRFSp1 crossed the MRMS curve at approximately the same location for both years, between 3 and 3.25 inches for the HRRR and between 1.75 and 2 inches for the RRFSp1 (representing a change in character from under to over forecasting).

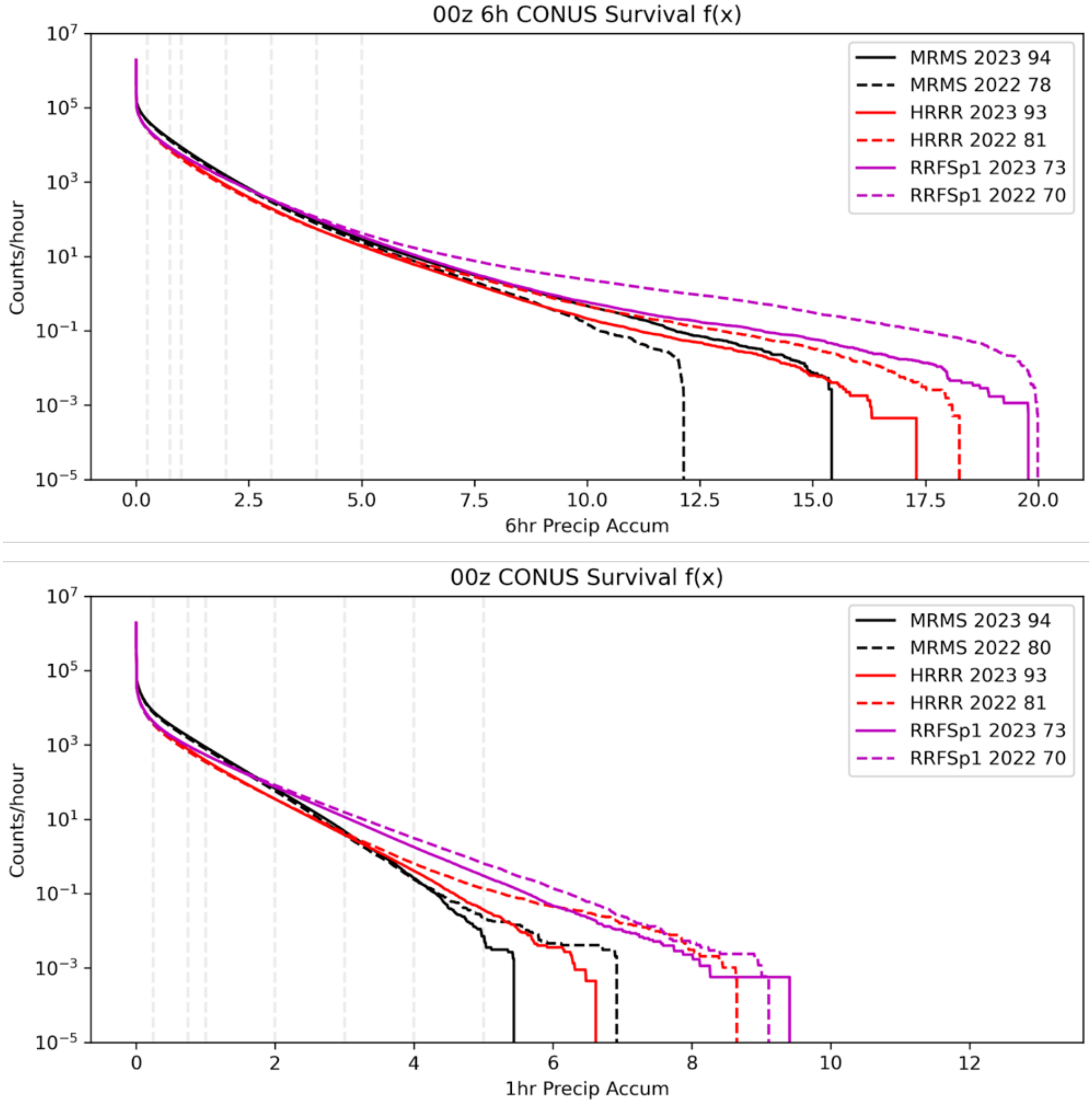


Figure 26: 6 hourly (top) and hourly (bottom) QPE/QPF survival function for the 2022 (dashed) and 2023 (solid) Testbed Seasons over the CONUS comparing MRMS (black), HRRR (red) and RRFSp1 (purple). On the y-axis is the counts per hour and on the x-axis is the 6 or 1-h precipitation accumulation.

However, the characteristics of the decrease in the occurrences from 2022 for the two models was noticeably different. The HRRR's shift between the two years more closely resembles the MRMS change, in terms of slope change and maximum hourly total (∼3.5 in vs ∼4.5 in). For MRMS the maximum hourly total decreased from 7 in to ∼5.5 in while the HRRR decreased from ∼8.5 in to 6.75 in. The

RRFSp1, on the other hand, only had a small change in the slope of the curve from 2022 to 2023. So although the number of higher hourly totals decreased between the two years, it was less of a decrease than seen in the HRRR or MRMS. Furthermore, the maximum hourly total actually increased slightly in the model, rather than decreasing. Combined, with what was seen for the 6-h comparison, this suggests that the 6-h totals are largely driven by short duration rainfall in the RRFSp1 rather than steadier rain over a longer period of time. It also suggests that although the wet bias appears to have decreased in the RRFSp1, the problem of over forecasting high hourly accumulations is still present.

Rainfall over the diurnal cycle can be seen in Fig. 27; coverage means that both rain and no rain (zeroes) were included in the average while intensity means only accumulated precipitation was included in the average (i.e. no zeroes). For MRMS, the average precipitation decreased in both coverage and intensity for 1 and 6-h totals from 2022 to 2023 across the entirety of the diurnal cycle. An exception is hourly coverage (bottom left) from around 18z to 03z, which saw nearly the same average hourly precipitation between the two years. The HRRR followed a pattern similar to this from 2022 to 2023.

On the other hand, the changes from 2022 to 2023 for the RRFSp1 differed across the diurnal cycle. The peak of the average rainfall shifted right (i.e. later) from 2022 to 2023 to be better aligned with the observed diurnal cycle of convective initiation, especially in terms of coverage. However, along with the shift, there was also an increase in average precipitation in terms of both coverage and intensity. During the typical lull in convective initiation, from approximately 06 UTC to 18 UTC, the average decreased from 2022 to 2023. For coverage, this decrease actually brings the RRFSp1's average precipitation nearly to the HRRR's totals, which is not ideal given how notable the dry bias is in the HRRR.

The difference between the two years for the RRFSp1 in terms of intensity was relatively small, though, like with the average precipitation coverage, the average hourly rate was lower during the less convective times of the day in 2023 for 6-h QPF. Differing from the 6-h rainfall in 2022, the intensity of the RRFSp1's QPF shifted from being similar near the peak of the diurnal cycle to being 0.005 in 6h^{-1} greater

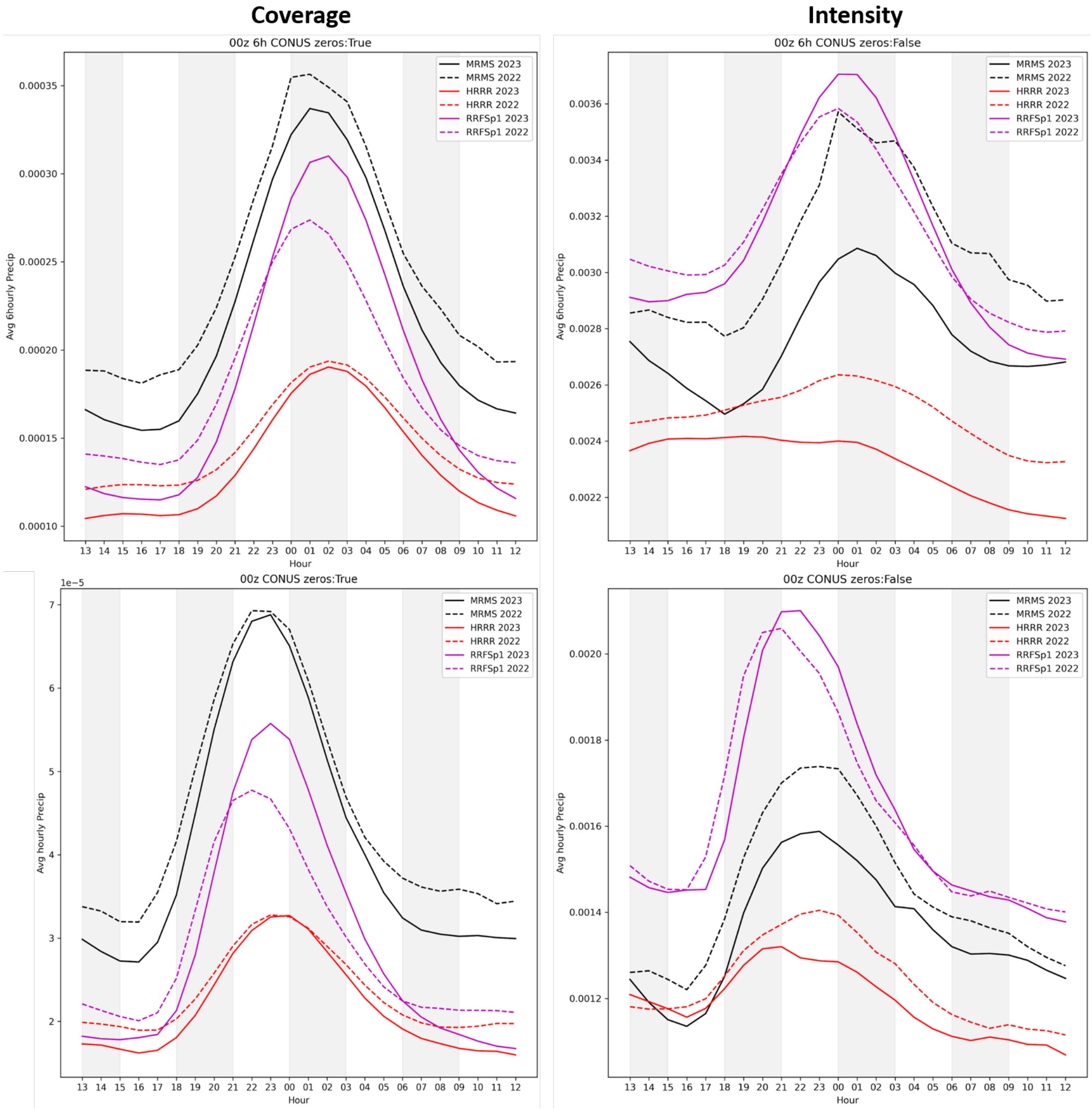


Figure 27: Diurnal analysis of the average 6 hourly (top) and hourly (bottom) QPE/QPF for the 2022 (dashed) and 2023 (solid) Testbed Seasons. [LEFT] coverage (zeros included in average) and [RIGHT] intensity (zeros NOT included in average). MRMS (black), HRRR (red) and RRFSp1 (purple). Note: the average hourly precipitation are scaled differently for each of the images.

than MRMS (note that this is across the whole CONUS, so averaged values are expected to be small). This differs from the HRRR which is 0.007 in 6h^{-1} less than MRMS. For hourly intensity, the difference between MRMS and RRFSp1 at their respective peaks also increased from 2022 (0.00025 in h^{-1}) to 2023 (0.0005 in h^{-1}).

Therefore, like last year, the RRFSp1 still appears to generating convection that is too intense. Since the greatest intensity is seen in the 1-h averages, this further supports the conclusion from the survival diagram analysis that short duration rainfall is likely driving the 6-h totals, rather than more steady precipitation. It also suggests that even though the RRFSp1's curve shifted closer to MRMS from 2022 to 2023, there was not a substantial difference seen in the slope. This is troublesome, given that the conditions that allowed widespread pulse thunderstorms (aka popcorn convection) was not as prevalent as it had been in previous years; see Section 3. That is not to say that popcorn convection did not occur; it just means that unlike last year when nearly every day somewhere in the southeast experienced pulse thunderstorms, those days were much rarer in 2023. The exception to this would be over Florida. An example of on such day can be seen in Fig 28, with scattered storms across the southeast.

The aforementioned conclusion does correspond with trends seen by the FFaIR team and noted by participants: 1) convective cells did not appear as large as in previous years; instead the cells were small but still intense. Because they were smaller, the coverage of these high precipitation cells appeared to be less. Again, referring to Fig 28, more instances of $\geq$2.5 in can be seen in the RRFSp1 forecast than in the operational models or MRMS. This is especially noticeable over western Florida. 2) The "look" of the QPF footprint often did not fit the participants' perceived conceptual model of how rainfall should evolve across an event. For instance, in the end-of-the-week survey, a participant wrote "The storm structures in RRFSp1 did not match MRMS (storms were too big/too many). The location of rainfall tended to be off as well." Another wrote, "RRFSp1 struggled with storm structure, even in the first few hours."

4.1.4 Precipitation Rates

In the past two FFaIRs, unrealistically high instantaneous (prate) and maximum (pmax) precipitation rates were identified in the RRFSp1. It was thought that these high rates were happening over multiple timesteps within the hour, thus helping to drive the high hourly totals and the overall wet bias seen in the RRFSp1. Thus the RRFSp1 pmax was compared against the NAMnest (HRRR

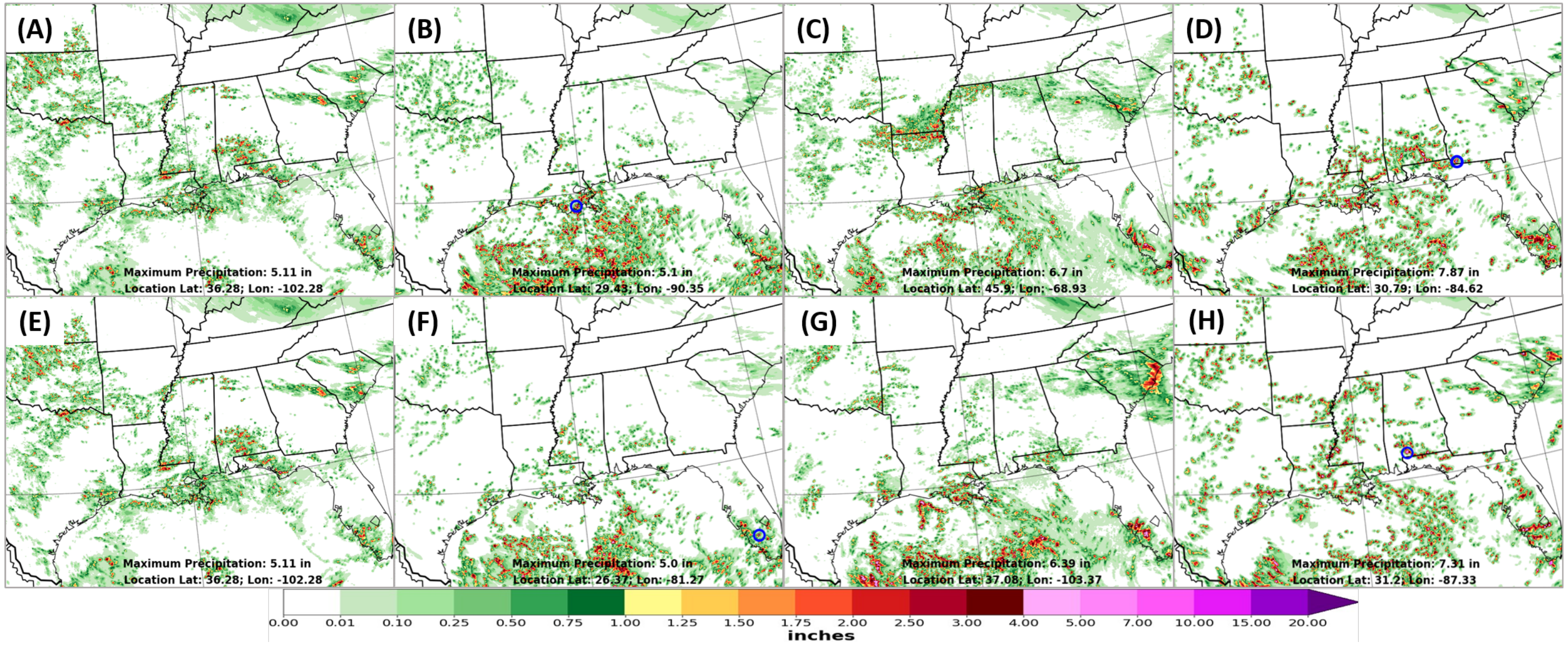

(A)
Maximum Precipitation: 5.11 in
Location Lat: 36.28; Lon: -102.28
(B)
Maximum Precipitation: 5.1 in
Location Lat: 29.43; Lon: -90.35
(C)
Maximum Precipitation: 6.7 in
Location Lat: 45.9; Lon: -68.93
(D)
Maximum Precipitation: 7.87 in
Location Lat: 30.79; Lon: -84.62
(E)
Maximum Precipitation: 5.11 in
Location Lat: 36.28; Lon: -102.28
(F)
Maximum Precipitation: 5.0 in
Location Lat: 26.37; Lon: -81.27
(G)
Maximum Precipitation: 6.39 in
Location Lat: 37.08; Lon: -103.37
(H)
Maximum Precipitation: 7.31 in
Location Lat: 31.2; Lon: -87.33
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 28: 24-h (A) and (E) MRMS QPE and (B)-(D) and (F)-(H) model QPF valid for 12 UTC 07 June 2023. (B) and (F): HRRR, (C) and (G) NAMnest, and (D) and (H): RRFSp1. (B)-(D) are the 00z cycles and (F)-(H) are the 12z cycles.

doesn't output pmax) and the RRFSp3 and RRFSp4; the latter two were used since they have the same model core as RRFSp1.

However, this year the prate/pmax values appear to be more reasonable. Across the various convective patterns that were observed during FFaIR, it was found that the values tended to fall between the maximum values seen in the MRMS and those seen in the NAMnest, generally staying below 20 in h^{-1}. Though the maximums did still tend to be larger than the RRFSp3 and RRFSp4 values. When asked about the footprint of the pmax and the maximum value compared to MRMS, participants nearly always said the overall footprint was smaller than MRMS but the maximum values were still higher. They also noted that rates of 3-4 in h^{-1} seemed to appear less in all models than observed. An example of this can be seen in Fig. 29.

Also noted last year was that the RRFS ensemble members tended to see even higher pmax values than the RRFSp1; there were instances of ensemble members reaching up to 200 in h^{-1}, though rates this high were often confined to the first hour or two after initialization. This year, the highest rate from an ensemble member was 80 in h^{-1}. A more typical high value was around 52 in h^{-1}, like that shown in the pmax daily plot for 02 Aug 2023 in Fig. 30. The changes made to the RRFS over the past year appear to have helped improve the high precipitation rate issue in the model, though the wet bias in terms of QPF still remains, so the factors driving this still need to be identified.

4.2 Short Summary on RRFSp1 QPF Findings

In general, subjectively the RRFSp1 has improved since last year. Participants felt that typically the RRFSp1 did a good job with the overall spatial distribution of precipitation but still tended to have unreasonable maximum precipitation rates, so they found it hard to trust. They also noted that there was still a speckled look to the precipitation, even in regimes that don't support it. In terms of the subjective verification, the "best" model for 24-h QPF at times differed from the 6-h "best" QPF. The HRRR and RRFSp1 had highest averages for 24-h QPF, while the NAMnest generally had higher scores for 6-h, both in terms of distribution and

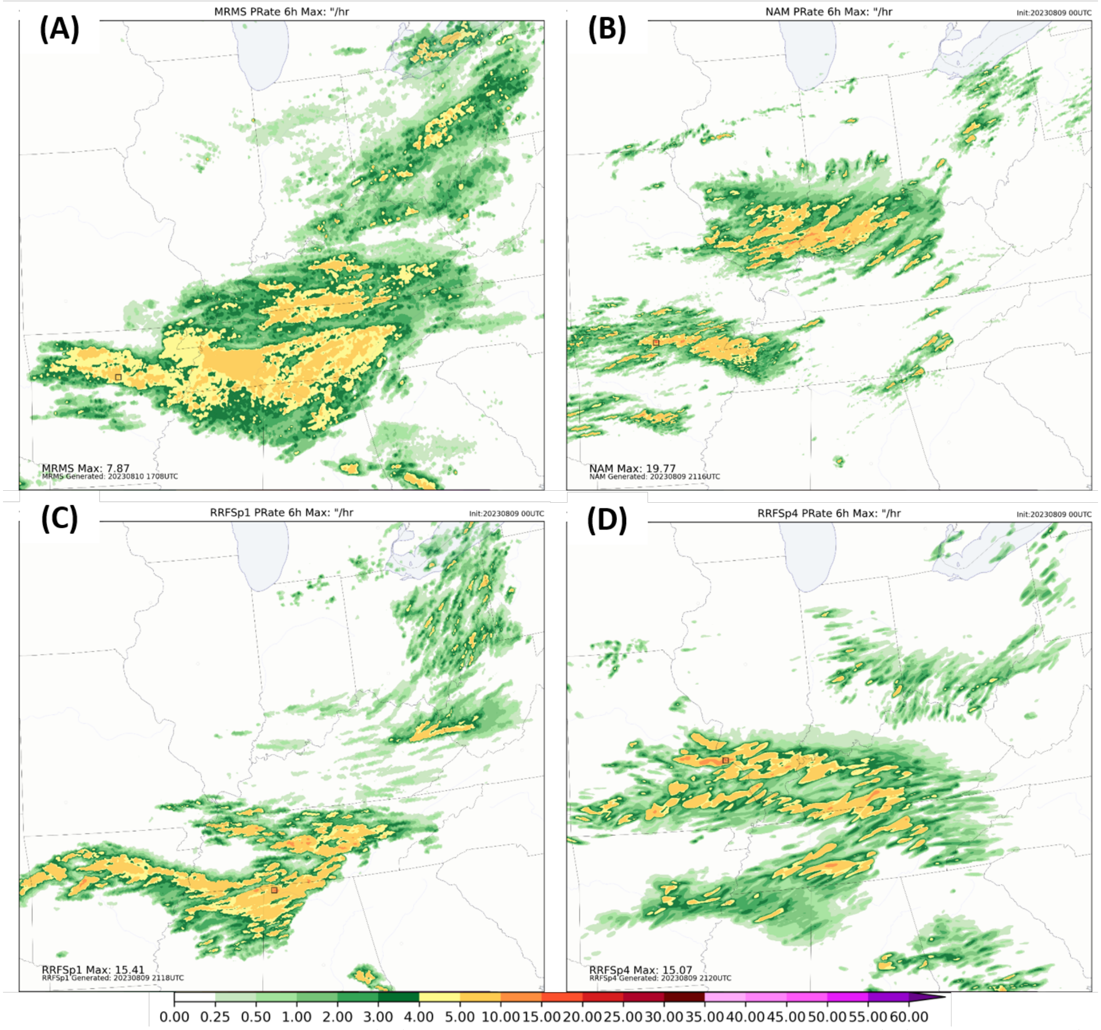


Figure 29: 6-h maximum of (A) MRMS-GC precipitation rate and (B) NAMnest, (C) RRFSp1, and (D) RRFSp4 model pmax valid 0000 UTC 08 Aug 2023.

averages. This difference likely stems from two things: first that 24-h accumulations can often “hide” errors such as precipitation timing, location, size in forecasts of specific events. For example, if there were two events were observed over a similar area in 24-h, the events would look like one event. But when looking at the events individually, it might be found that for the first event the model coverage was too small and shifted south but for the second event it was too large and covered both areas of observed rainfall. This could result in the 24-h totals looking similar to observations even though location and size for each individual event was incorrect.

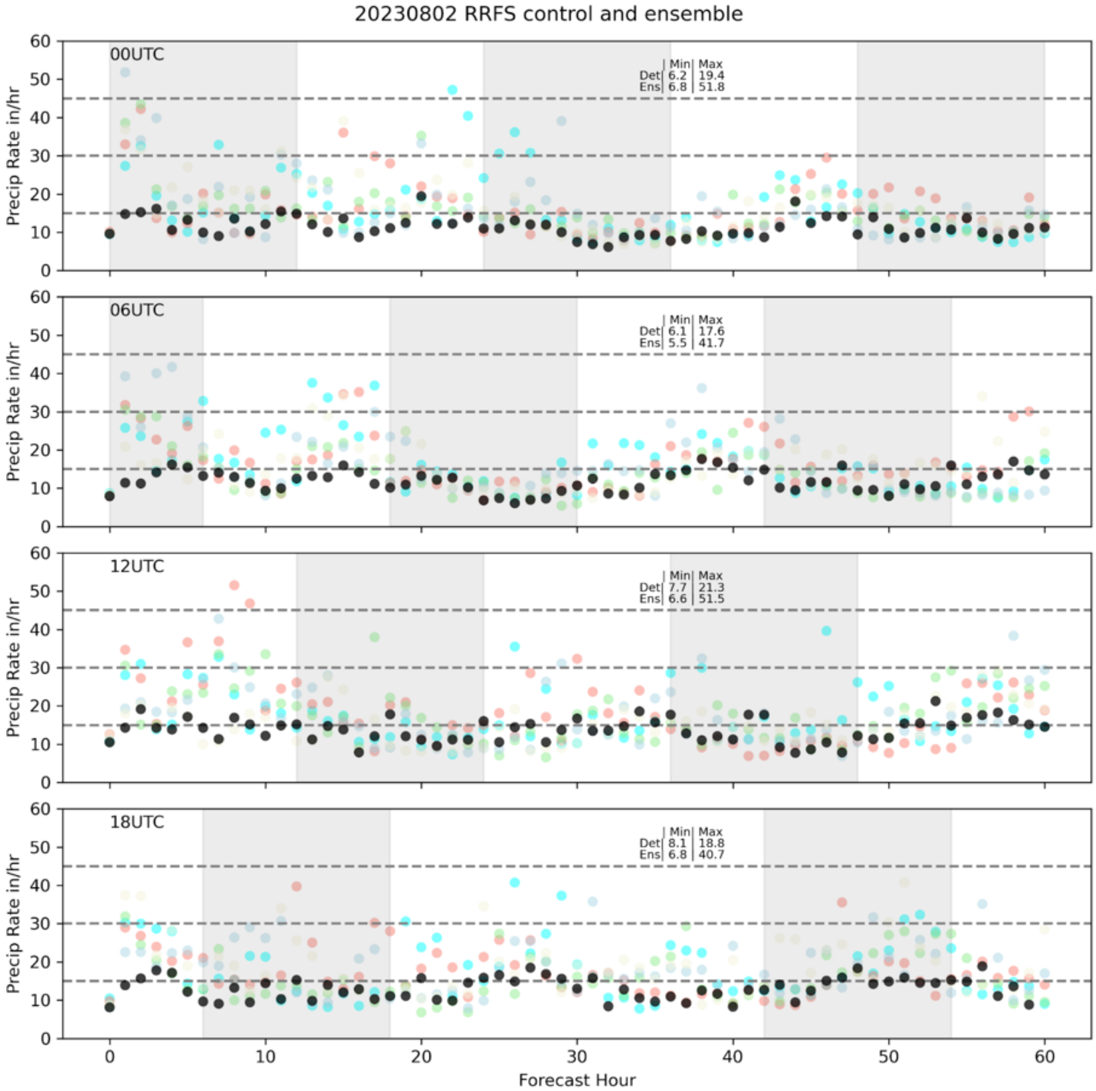


Figure 30: PMAX for the RRFS deterministic (black dots) and ensemble members (color by member) for each cycle on 2 August 2023; cycle order from top to bottom: 00z, 06z, 12z, 18z. On the top of each panel is a summary table that shows the Min and Max pmax from the deterministic and the ensemble members. Dashed lines at 15, 30 and 45"/hr are shown for reference. Gray patches indicate night time for each cycle; here defined from 00 UTC to 12 UTC.

Secondly, the 24-h QPF was evaluated over who CONUS while the 6-h QPF was over the MRTP domain, for a single, usually extreme rainfall event.

Additionally, in both the subjective and objective verification, the performance of RRFSp1 compared to NAMnest shifted when looking at the dates before and after the change from the CONUS domain to the NA domain in RRFSp1. Though it is possible the change in domain had an impact on the RRFSp1's performance, it is more likely that the change in the type of events between the two time periods was the reason. For the CONUS domain dates, spatially smaller, less organized heavy rainfall events were seen while the NA domain had better organized and larger scale heavy rainfall events. For example, over the MRTP domain, 6-h rainfall of 1 in or greater over >30,000 km^2 occurred on 2 of the 15 days for the CONUS domain but 6 of the 9 days for the NA domain. Furthermore, when looking at the 24-h rainfall totals for the NA domain days, there was very little rainfall across the CONUS outside these MRTP regions, so they were large drivers in how the participants scored the models and in the overall contingency table statistics.

Analysis of 6-h and 1-h rainfall survival functions and average rainfall for the CONUS over the diurnal cycle (in terms of coverage and intensity) show that the RRFSp1 continues to have a wet bias driven by short duration, intense rainfall. The survival function in Fig. 26 shows a decrease in observed hourly accumulations >3 in from 2022 to 2023. However, unlike MRMS, the maximum hourly accumulation from 2022 to 2023 in the RRFSp1 increased slightly. This differed from the the 6-h MRMS observed counts, which saw an increase between the two years of accumulations $\geq$5 in. Meanwhile RRFSp1's counts decreased between the two years starting around 5 in, resulting in slope and counts similar to MRMS until 10 in. The results suggest that for the 2023 Testbed season, MRMS 6-h totals at the higher accumulations were mostly driven by prolonged rainfall across the time period, while intense hourly totals tended to drive the 6-h totals seen in the RRFSp1. This was further supported when comparing the coverage of rainfall at both thresholds to the intensity across the diurnal cycle; see Fig. 27. Hourly average QPF from the RRFSp1 increased from 2022 to 2023 despite it decreasing in MRMS. This is particularly troublesome, given that popcorn convection was not as abundant this year due to abnormally low heights over the eastern CONUS, and it has been thought that the RRFSp1's overzealous forecasting of popcorn convection was driving the intense hourly totals seen.

Finally, at the end of the week participates were asked how they felt the RRFSp1 performed compared to the HRRR and NAMnest for their week. Figure 31 shows the results from this question; note that there are no comments from week 4 of FFaIR due to the RRFSp1 being unavailable. Overall participants felt that the RRFSp1 performed similar to the NAMnest but slightly worse than the HRRR. They were also asked to provide comments on the performance of the RRFSp1 and if they noticed any biases in the model. Overall, comments from the CONUS domain dates (more synoptically driven events) were more neutral to positive. For the NA domain dates (more MCS/MCV driven, larger scale events), the comments were more negative for the RRFSp1, while there was praise for the performance of the NAMnest. Below are some comments that summarize the overall feedback about the RRFSp1.

- "RRFS and NAM(nest) seemed to perform similarly, though with different looks to them."
- "The RRFS was comparable to the NAMNest, but when the Nest was properly placed, it was typically better...but when it was not, it was severely worse. The RRFS was typically close in placement overall."
- "It was definitely consistently way too hot in terms of precip totals. And its spatial distribution was a little odd. The storms it produced were unrealistically small and intense with precipitation."
- "The storm structures in RRFSp1 did not match MRMS (storms were too big/too many). The location of rainfall tended to be off as well."
- "RRFS is still a little "hot" in precip areas, but at least didn't miss as many areas as HRRR did. "
- "In general, sometimes the RRFS nailed the forecast given very specific parameters/conditions. But overall looking at the bigger picture, the HRRR and NAMNest especially outperformed the RRFS."
- "There were times when RRFSp1 was the best model, but overall the QPF amounts tended to be high and I did not perceive there to be any advantage in placement vs. the other models."

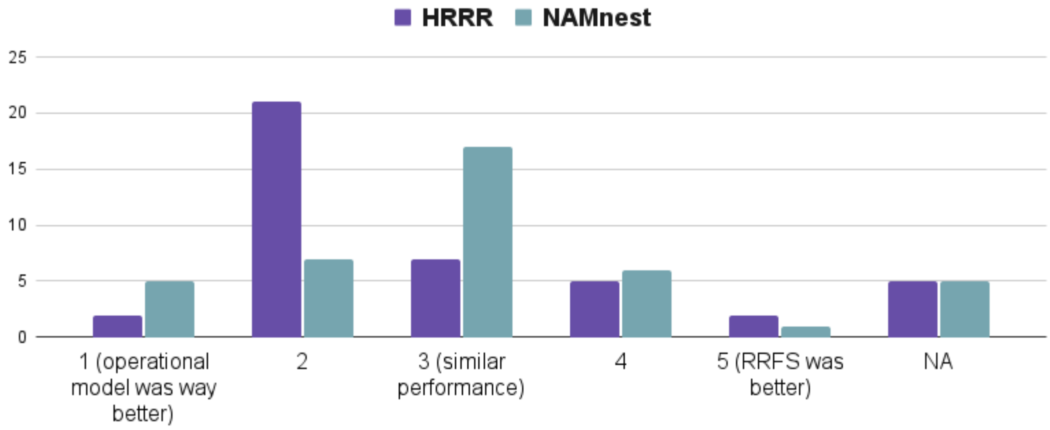


Figure 31: Results from the end of the week survey sent out to the participants, asking them to rate how they felt the RRFSp1 performance compared to the HRRR and NAMnest for their week. Here the RRFSp1 is the baseline, a score of 1 means the RRFS was much worse than either the HRRR or NAMnest while a score of 5 means the RRFSp1 was much better.

4.3 Ensemble

As mentioned in Section 2.3, the experimental ensemble data was available inconsistently, thus it is difficult to do a proper comparison of the ensembles against each other or against the HREF. Additionally, before analysis from FFaIR could be completed, the RRFS team had decided to change from the planned single physics, time-lagged ensemble (RRFSe1tl) to a multi-physics ensemble that differed from both the RRFSe2 and RRFSe2tl seen in June of FFaIR. This RRFS multi-physics ensemble configuration includes membership from the HRRR, both current and time-lagged runs. Thus none of the ensembles evaluated in FFaIR are being considered by the RRFS team to be the ensemble for RRFSv1. As for the CAPS_RRFSe, the probability files provided did not include a smoothed probability field, resulting in a different look to the field compared to the other ensembles; see Fig. 32. Participants noted that trying to compare the unsmoothed probabilities to smoothed probabilities was difficult. Consequently, only a brief analysis of ensemble performance will be provided.

Figure 33 shows the 00z confidence ranking for all of FFaIR (top) and days in which the CAPS_RRFSe was available (bottom) since it was the experimental

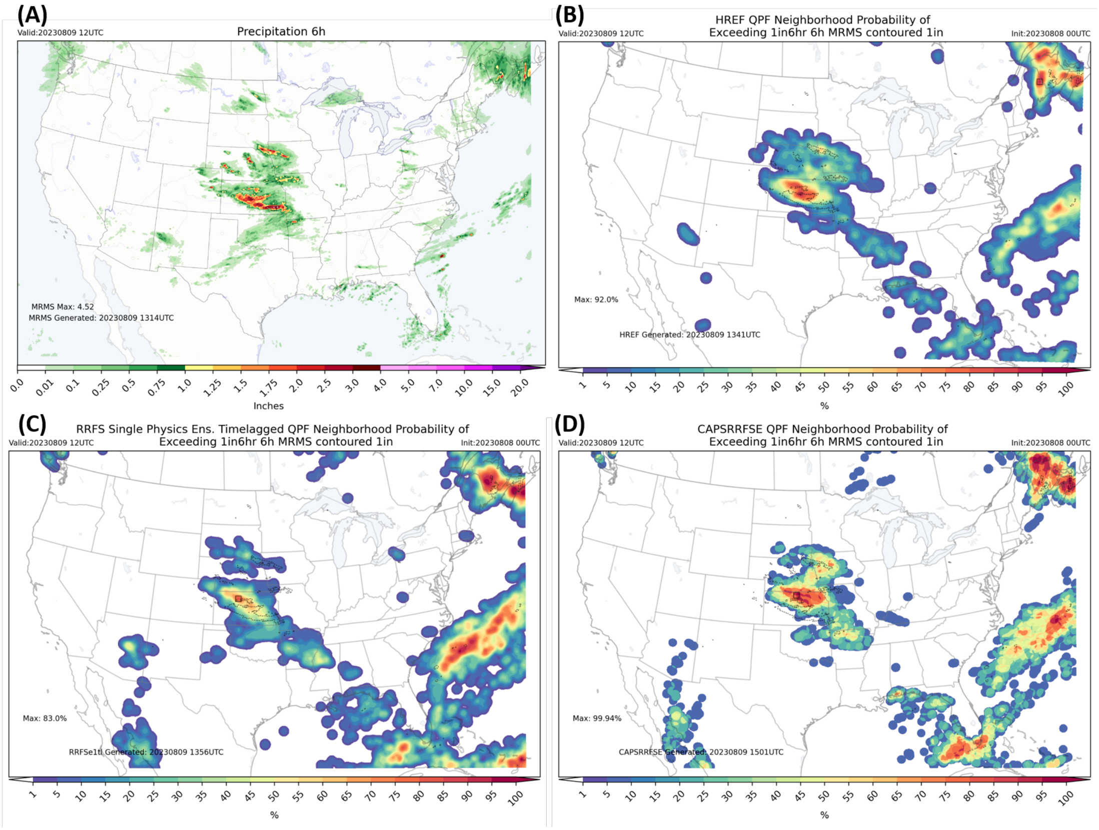


Figure 32: (A) 6-h MRMS and (B) HREF, (C) RRFSe1tl, and (D) CAPS_RRFSe probabilities of 1 in 6 h^{-1} valid 12 UTC 09 Aug 2023. Dashed contour in (B)-(D) is the MRMS 1 in.

ensemble that was available the most consistently. The all-FFaIR results will not be discussed, nor will the RRFS ensembles from EMC, but they are provided for completeness. Over the CAPS dates, the probability of exceeding 1 in 6 h^{-1} which provided confidence in the forecast was about the same for the HREF and the CAPS_RRFSe. However, for the higher probability evaluated, 5 in 6 h^{-1}, the HREF was the most likely of the two ensembles to provide confidence in the forecast for the participants. 55% of the time, participants felt the CAPS_RRFSe provided little confidence (scores of 2 or 1) in the forecast for 5 in 6 h^{-1}. Meanwhile 50% of the time the HREF provided confidence (scores of 4 or 5) in the forecast.

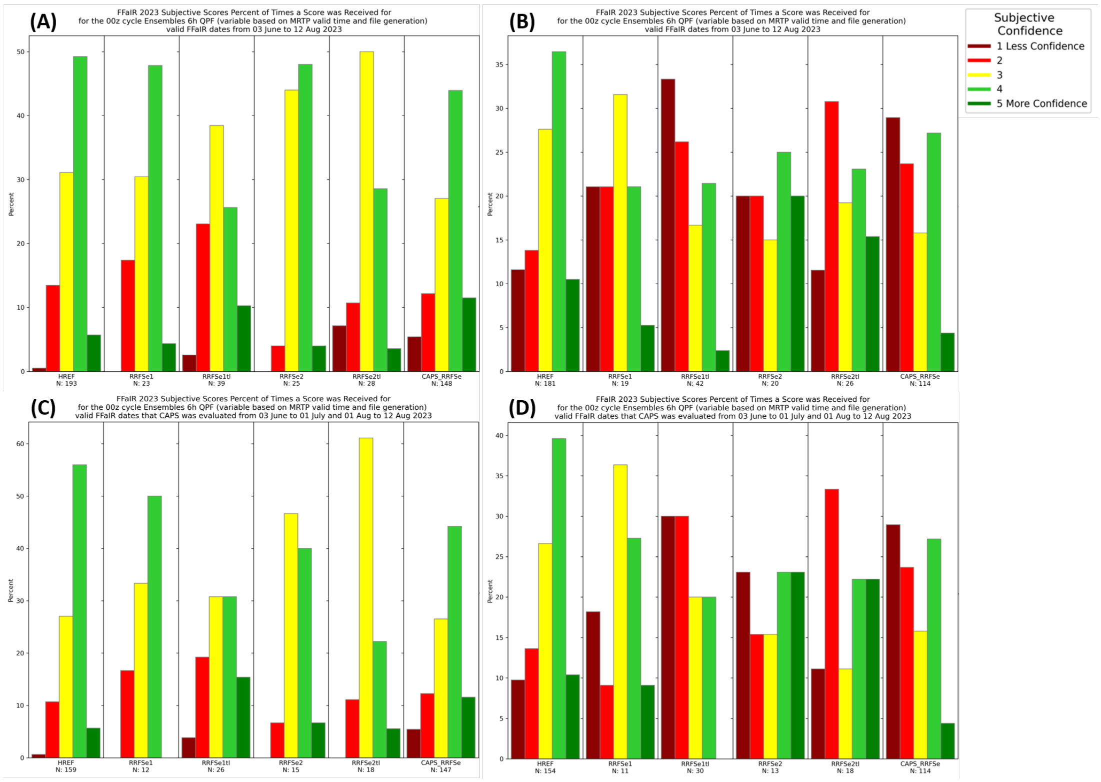


Figure 33: Like Fig. 15 but for the 00z ensemble subjective verification scores. (A)-(B) are for FFaIR Dates and (C)-(D) are specific to the days the CAPS_RRFSe was evaluated. (A) and (C) are for 1 in 6 h^{-1} and (B) and (D) are for 5 in 6 h^{-1}.

Although not shown, performance diagrams of the 24-h PMM and LPM means were done over all FFaIR dates for all the ensembles[14]. For the HREF, the PMM at all thresholds evaluated always had a higher CSI than the LPM, with a slight wet bias seen in the PMM and a slight dry bias seen in the LPM. This is consistent with past FFaIRs. Opposing this, the CAPS_RRFSe LPM always had the higher CSI while both of the means hovered around a bias of 1. Last year, the CAPS_RRFSe's means were not evaluated but were in 2021 (though the configuration differed greatly and was called the SSEF). In the 2021 FFaIR experiment, the performance of the PMM and LPMM means followed that of the HREF. Additionally, although the sample size is smaller, ensemble evaluation this year found that the RRFSe1tl means performed like the HREF, with the PMM having a higher CSI than LPM

[14] The HREF was available for all 30 days. The CAPS_RRFSe had 25 days of availability. The RRFSe1tl had the most of all the RRFS ensembles at 11 days.

and a high bias for the PMM and a low bias for the LPM. Thus, it is unlikely that the differing model cores in the HREF (multiple cores) and CAPS_RRFSe (FV3 core only) drove this difference. It would be useful for the CAPS team to examine why the bias of the CAPS_RRFSe PMM and LPM means did not follow convention.

4.3.1 OU CAPS Mean Products

Using a spatial alignment method, the CAPS team supplied two additional means, referred to as the SAM and SAM-LPM, with their ensemble product data aside from the typical set of means (the arithmetic mean, PMM mean and the LPM mean). The SAM and SAM-LPM were compared against the regular mean and LPM mean from the CAPS_RRFSe and HREF. The results from subjective verification of the ensembles and their means can be seen in Fig. 34. Comparing the HREF and CAPS_RRFSe means to the CAPS_RRFSe SAM, the differences are relatively small. However, comparing the CAPS_RRFSe mean to the SAM, the SAM increased in the number of times participants felt that the "footprint looks similar to observed" while decreasing for "footprint is larger than observed", with little change in how they viewed the shape of the footprint. This suggests that the participants felt the SAM methodology shrunk the footprint while retaining its shape. A similar shift, but greater in magnitude, was also seen when comparing the LPM and SAM-LPM. This corresponds with what participants noted in their comments, often stating that they felt that the SAM methods helped decrease the coverage of the light precipitation. It seemed to be a wash on whether or not this feature of the two SAMs was useful. For instance, one participant noted "The SAM definitely is a little less detailed than the regular CAPS, and the SAM LPM was definitely a little too "washed out". Someone else stated, "the SAMs are pretty good at not having too large of a low value field."

Another difference can be seen in the how the participants felt the magnitude of the maximum accumulation compared to observations for the various LPM means. For both the HREF and CAPS_RRFSe LPMs, the general feeling was that the maximum was about the same or slightly lower than observed. However, for the CAPS_RRFSe SAM-LPM, the choice that was picked the most in terms of

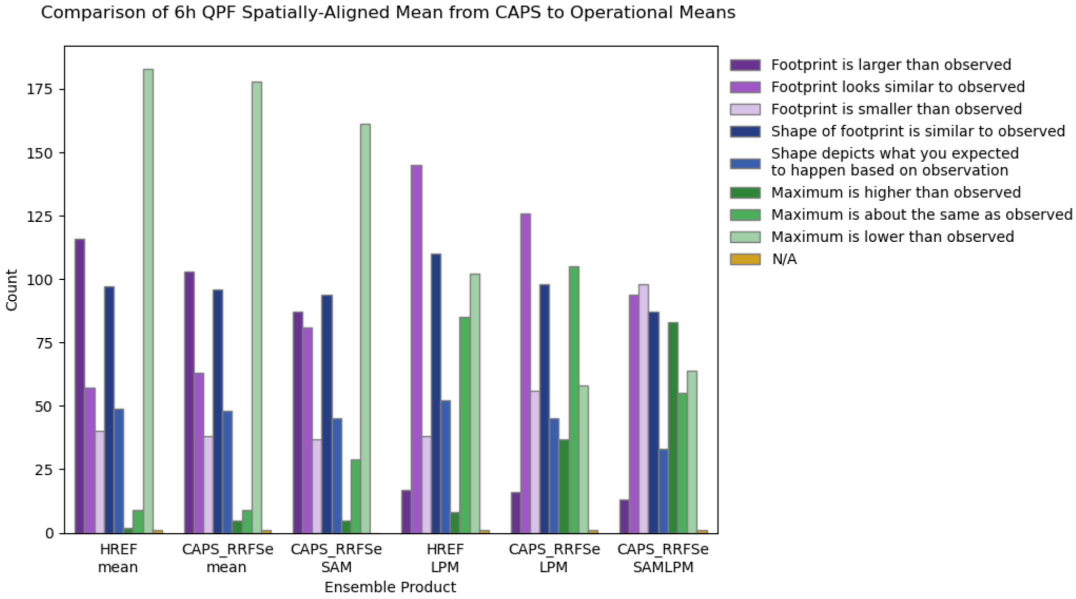


Figure 34: The number of times each type of ensemble mean fit the given description for the 2023 FFaIR dates that the CAPS_RRFSe was available.

maximum magnitude by the participants was "Maximum is higher than observed." Based on discussion during the experiment and written feedback, participants didn't always choose this option because the value of the magnitude was too large but instead picked this option because the aerial extent of the higher values was too large. This also impacted how they scored the footprint size question. Some participants noted they would pick both too large and too small because the rain/no rain footprint was too small but the coverage of the maximum values was too large. That is not to say that there weren't instances of the magnitude itself was larger than observed. For instance, for the forecast shown in Fig. 35, the observed max (6.15 in) and the CAPS_RRFSe LPM max (6.21 in) were similar both in magnitude and location while the CAPS_RRFSe SAM-LPM had a maximum about 3 inches greater than both (9.59 in).

Additionally, it was noted how drastically the footprint can change in this forecast. Focusing across the KS/OK border and into northeastern AR the aerial extent of a hundredth of an inch or greater decreased between the CAPS_RRFSe LPM and SAM-LPM. Opposing this, the higher totals (≥1 in) in the region became more cohesive and resulted in a larger, continuous area of one or more inches of

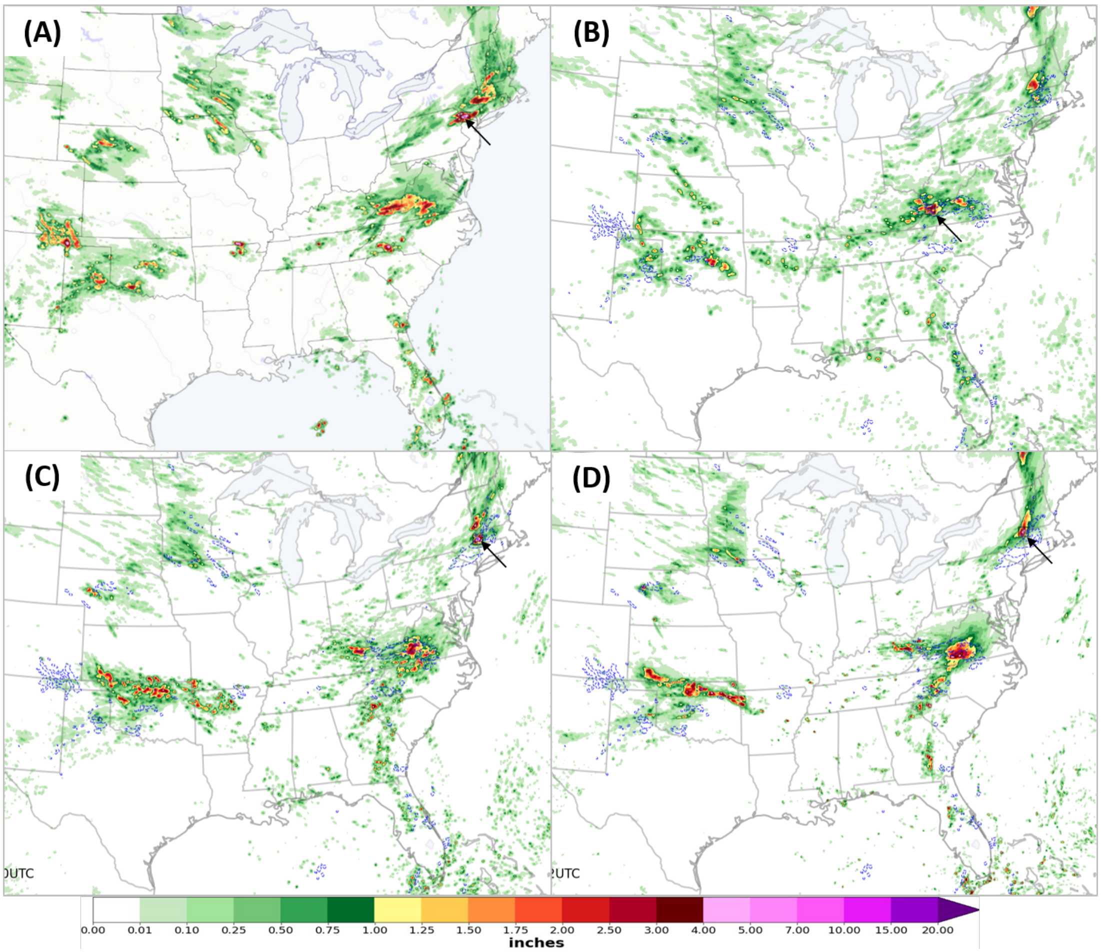


Figure 35: 6-h (A) MRMS QPE and (B) HREF LPM mean, (C) CAPS_RRFSe LPM mean and (D) CAPS_RRFSe SAM-LPM valid 06 UTC 14 July 20230. The dashed blue on (B)-(D) is the 1 in MRMS contour. The black arrow points to the location of the maximum: (A) 6.15", (B) 3.92" (C) 6.21", and (D) 9.59".

precipitation. A similar result occurs over southern VA when focusing on values ≥2.5 in. These sorts of shifts are examples of what the participants meant when they said they felt the aerial extent of the higher totals was too large. Over the course of the experiment, whether or not this difference in the "look" of the SAM-LPM compared to the LPM was liked or disliked seemed varied among the participants and event type/location . For instance, one participant wrote: "(t)he SAM-LPM seems to do a much better job with the higher intensities overall, even though they're a bit misplaced. In fact the CAPSRRFSE samlpm looks a little too hot overall, but in some localized areas, this is a better depiction of threat."

Interestingly, even though participants sometimes didn't like the increase in the coverage of the footprint of higher magnitude accumulations, it was consistently mentioned that they liked the less speckled character of the precipitation in the SAM-LPM versus the LPM. The SAM-LPM helped to smooth out the gradient between thresholds and tended not to have a large number of small bulleyes of high magnitude totals, which is often a characteristic of the LPM mean. For example, one participant stated "Generally like the SAM-LPM best. SAM-LPM seems like a useful technique that may be worth exploring operationally. LPM often is too bullseyed with the maxes without enough moderate amounts...and the SAM-LPM almost looks like a smoothed LPM which seems to fix the bias seen in the LPM alone" while another said "SAM-LPM was the best. It had the best combination of coverage and magnitude. LPMs on their own tend to not have enough "moderate" QPF and are too bullseyed. SAM-LPM corrects for that issue."

Based on how often comments like these were seen and the generally positive feedback given about the product even if the SAM-LPM magnitude was too high (value) or too extensive (coverage), the SAM-LPM added value to the forecasting process over the LPM. Though not discussed in much detail here, the same can not be said for the SAM. For the most part participants noted that they felt it was difficult to find differences between the mean and the SAM. Since it wasn't easy to find differences between the two means, they generally did not note that there was significant value added when looking at the SAM over the conventional mean. Two statements from the end of the week survey summarized this well: (1) "Only the SAM-LPM product appeared to be useful. The SAM product just looked like a straight ensemble mean (i.e., reduced precipitation maxima magnitudes and broader areas)." (2) "The SAM-LPM added more value than just the SAM. It had a better handle on location and threat level compared to the SAM which was over done at lower thresholds and underdone at higher thresholds." Therefore, the FFaIR team feels it would be most beneficial to continue development on the SAM-LPM.

4.3.2 OU CAPS Probability of Exceedance Product the HREF+

As stated in Section 2.3, a bug was found in the HREF+. Unfortunately, this was not discovered until the last week of FFaIR, therefore the feedback collected is not valid. The bug resulted in high rainfall probabilities often being located where model input did not have rainfall. It also removed large areas of rainfall erroneously. An example of this can be seen in Fig. 36. Note that across the Northern Rockies, both the HREF and CAPS_RRFSe have probabilities approaching 80% for 0.5 in $6h^{-1}$ but the HREF+ had low to 0 chance of exceedance over the same area. Opposing this, in southwestern MO, the HREF+ had probabilities of 100% while the other two had low to 0 chance of exceedance. The orientation of the footprints and location of maximum probabilities also differ (ex. the Plains), though some of this might be partially driven by the ML aspect of the HREF+. For instance, in NE the HREF and CAPS_RRFSe have a footprint of 50% or greater centered over northern NE, just south of its border with SD. However the HREF+ has a larger extent of +50% that is centered over southeastern NE. Despite the setback of the bug found in the HREF+, the FFaIR team supports continued development on this tool.

4.4 Maximum Rainfall and Timing Product

As explained in Section 2.2, the MRTP was designed to have all participants draw multiple rainfall contours in a chosen 6 hour period. The MRTP activity started with participants collaborating to choose the domain and time period where either the maximum 6-h rainfall and/or the largest areal coverage of rainfall would occur, with both criteria having some correlation to the occurrence of potential flash flooding. Additionally, participants answered questions about the amount and location of the maximum rainfall, flood probability, damaging flood probability, the hourly maximum rain amount and two new probabilistic questions. The first had the group vote on a maximum rainfall threshold value. This value represented a reasonable extreme for the event. Participants would then enter a probability that the threshold would be exceeded. The second was a general timing question, with ten bins representing each hour from 03 to 12 UTC. Participants were asked to enter probabilities for when they thought the maximum rain amount would occur.

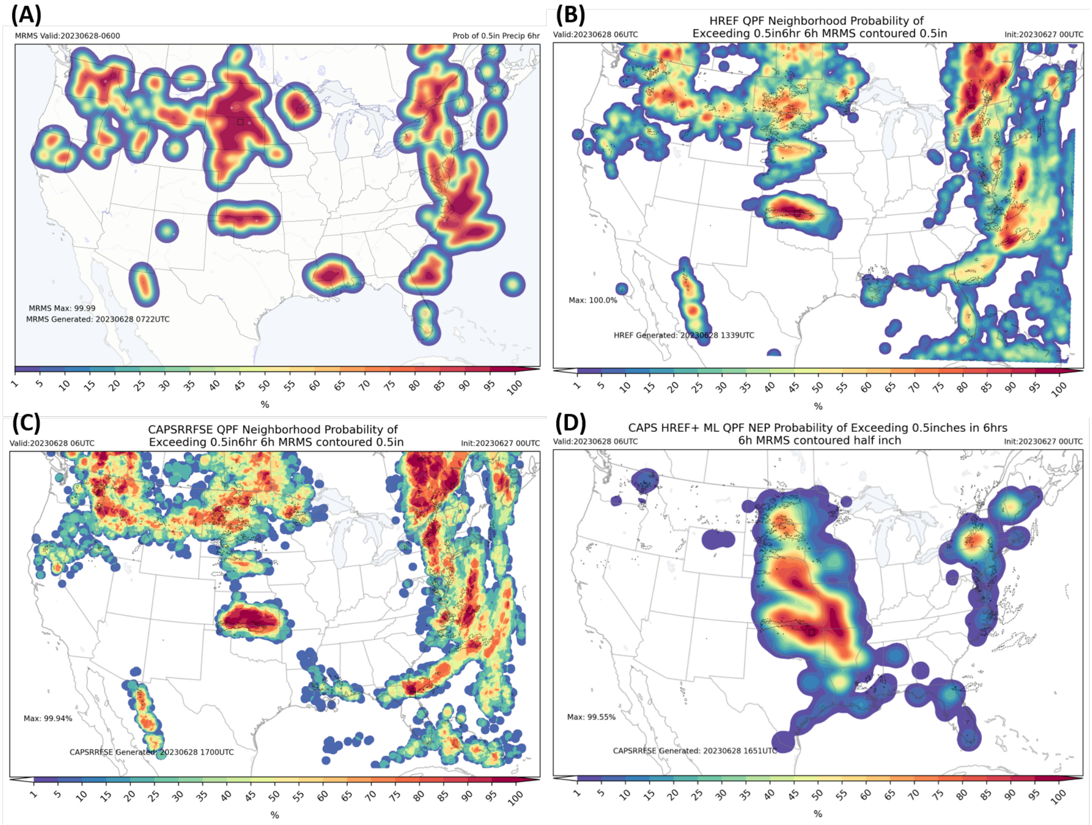


Figure 36: The 0.5 in6h^{-1} probabilities (A) derived from MRMS QPE and from the (B) HREF, (C) CAPS_RRFSe and (D) HREF+ valid 06 UTC 28 June 20230. The dashed black on (B)-(D) is the 1 in MRMS contour.

New this year was the addition of model precipitation forecasts directly in the browser where participants could draw their forecasts. This enabled quite a few participants to test out drawing smaller polygons more consistent with the model forecasts. Participants noted that they liked this feature since it made it easier for them to see the location differences across the models/cycles. In the past they stated that it was difficult to "eyeball" these differences when they were small.

The 2023 FFaIR season was quite different than 2021 and 2022 in terms of location of heavy rain events and general accumulated rainfall[15]. Of the 30 MRTP events, there were 20 that occurred in the early morning hours ending at 09, 10 and 12 UTC (Table 2). Eleven events had rainfall maximum exceeding 5" per 6

[15] See Section 4.2 for other differences.

hours, and twenty events had 1” areal coverage below 30,000 km^2 (Fig. 37). For reference, in the 2021 FFaIR season only 7 events had 1” in 6-h areal coverage below 30,000 km^2, while 6 events had areal coverage above 90,000 km^2. The largest 1 in 6 h^{-1} areal coverage event in for the 2023 FFaIR season was approximately 75,000 km^2.

In total, 55 Mesoscale Precipitation Discussions were issued by WPC during FFaIR. For 27 of the 30 days, an MPD was issued in the MRTP domain and in close temporal proximity to the chosen MRTP time period (Fig. 38). In all cases, MRTPs were submitted prior to 21z, and the first available 6 hour window ended at 0300 UTC. On all but ten days, we were successful in selecting the time of peak rainfall for the day, and on all but 5 days were successful in capturing the peak areal coverage within the MRTP window; represented by the star and open circle in Fig. 38, respectively. The MRTP activity was effective in determining the time and location of finding extreme rainfall and its potential impacts, even in this slower than usual rainy season. That participants could locate and temporally determine where and when MPDs might be issued in advance indicates that model guidance is positively contributing to the extreme rainfall forecast challenge.

4.4.1 Human and Model Performance

4.4.1.1 Model Data Usage

The participants of the MRTP experiment were randomly assigned a model or ensemble data to evaluate[16], though they were not required to use the assigned guidance and could use any available guidance they wished (Fig. 39A) for their forecast. More participants used the HREF than the HRRR, but deterministic models were generally used more than ensembles in their forecast process. However, for those using primarily a deterministic forecast, they still relied upon an ensemble, namely the HREF, (Fig. 39B) for guidance in their forecast. For those that initially used an ensemble, the main deterministic model referenced for additional guidance was the HRRR (Fig. 39C). The popularity of operational models can be frustrating for us as we try to evaluate new models, but this feature can be

[16] This was done to make sure all the guidance was being evaluated and had feedback provided.

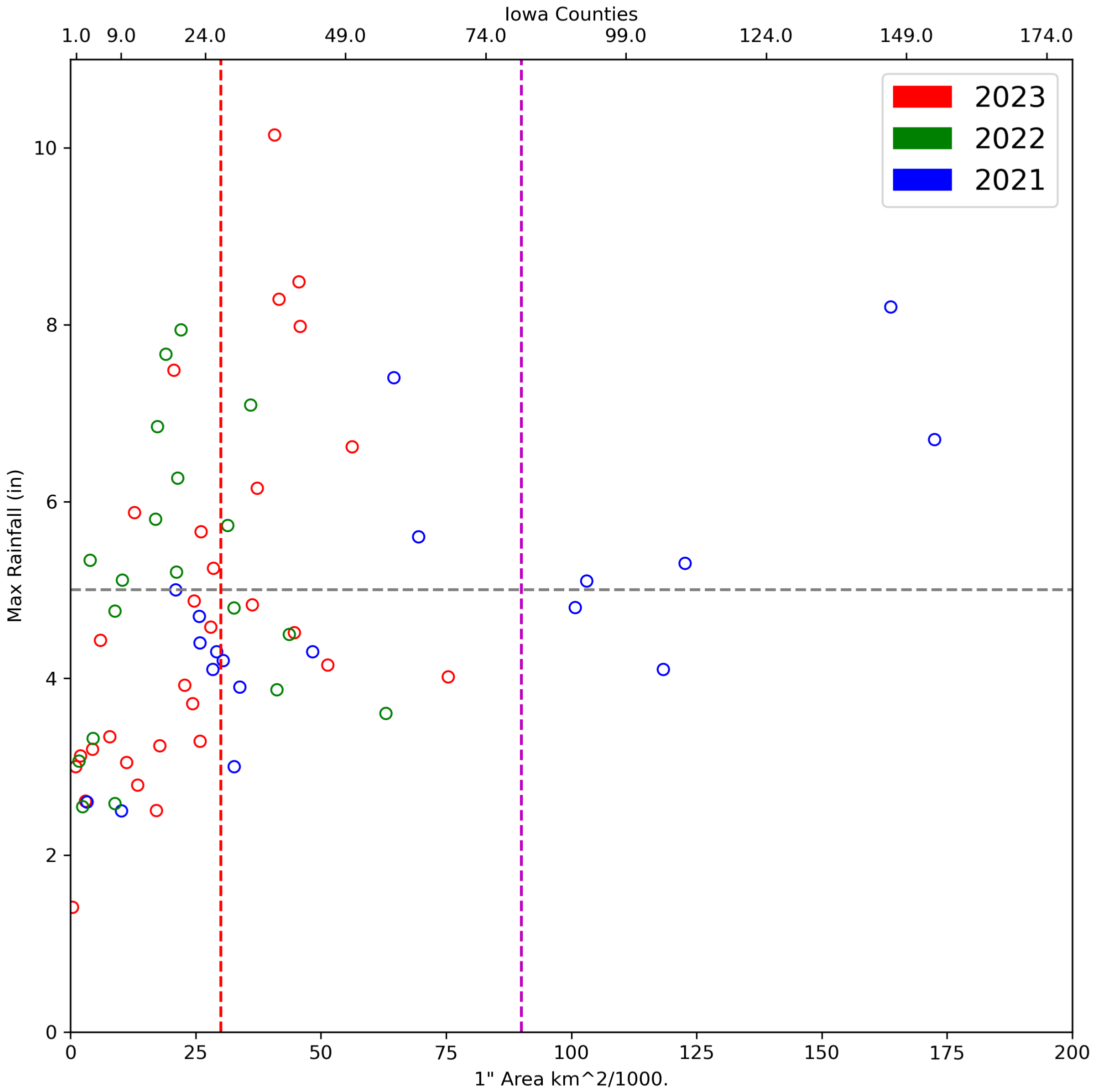


Figure 37: A comparison of the 2021 (20 days), 2022 (19 days) and 2023 (30 days) MRTP events in terms of areal coverage of 1 inch and the maximum rainfall in the domain as determined by the MRMS. The red dashed line denotes 30k km^2 and the purple dashed line 90k km^2. The grey dashed line denotes 5 in.

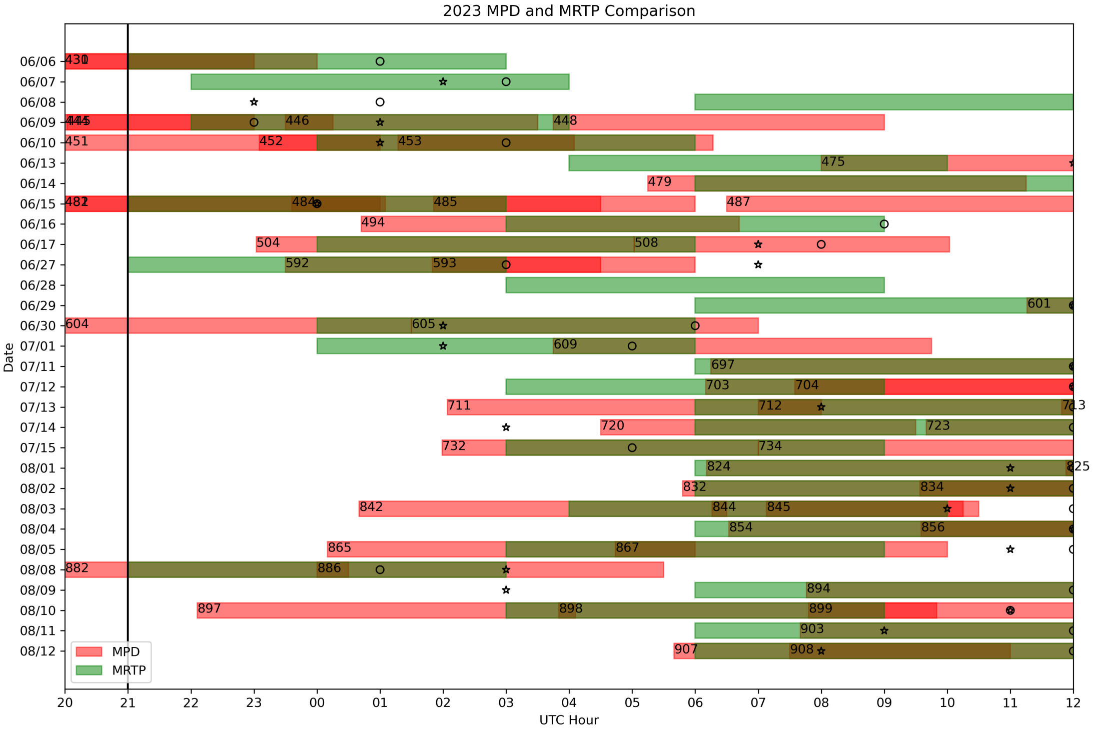


Figure 38: A comparison of the issuance times of the operational WPC Mesoscale Precipitation Discussions (red) and the timing of the MRTP activity time window (green). MPDs which cover the geographical domain are shown no matter the time of day. Also shown are the time of the maximum 6hr rainfall (star) in the domain and the time of the maximum 1" areal coverage (open circle). The number seen at the start of the MPD window is the issuance number.

helpful as experimental models improve and we can longitudinally track progress via popularity. Despite the minimal use of the RRFSp1 in the MRTP experiment, RRFSp1 and its ensemble were competitive in the post event evaluations. It will always be difficult, however, for known operational models to be usurped by an limited use and not-yet-available experimental model.

4.4.1.2 Performance Diagrams for Accumulation

Performance diagrams (Roebber, 2009) were produced for each day of the MRTP activity, showing the participant's performance along with every cycle of model guidance that was available for the MRTP time period[17] (see Fig. 40.)

[17] As many as 18 model cycles were available for the GFS model, running every 6h and verified out to 84h.

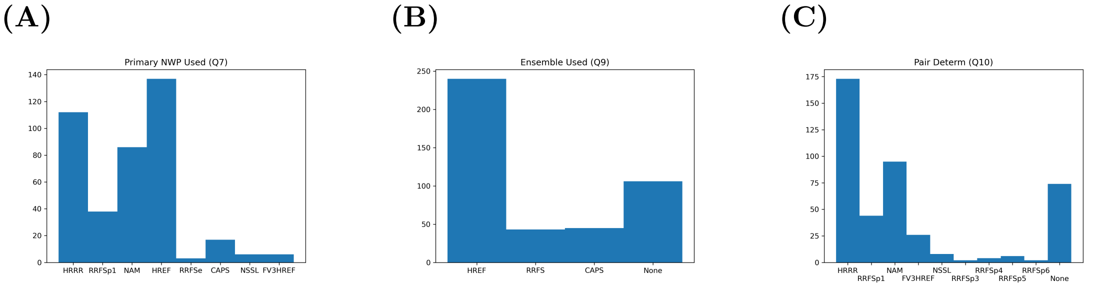


Figure 39: (A) Primary deterministic or ensemble model used to make MRTP forecasts. (B) If deterministic as primary, then used this ensemble model. (C) If ensemble as primary, then used this deterministic model.

Participants, as in previous years, typically had relatively higher Probability of Detection (POD) and lower Success Ratios (SR), indicated by the blue dots up along the left side of the performance diagram. There was considerably more spread amongst this year's participants. Although this could be contributed to having a larger number of veteran participants compared to new ones than in previous years, it is more likely that the spread came from the struggle associated with the type of events that were being forecast for; i.e. small and localized areas of heavy precipitation. Participants tended to draw their forecast areas in a more coarse fashion (i.e. generally larger areas), thus resulting in higher frequency bias during smaller-scale/localized events. To some extent, this is a challenge noted in the web-based drawing tools. Consequently, on larger areal coverage days, participants generally scored higher SRs and thus higher CSI. The SR improvement arises because of how participants draw their forecasts in larger polygons, a feature of our web based drawing tools. This is in contrast to the models, where grid points offer precision in their depiction of rainfall. Typically, as the model guidance improved so do the best participants, as seen in the CSI scores. The best participants compete with the best model guidance, indicating at least anecdotally, that participants take advantage of the most skillful guidance.

Daily model performance indicates that Day 1 guidance frequently improves upon the Day 2 forecasts, though there is considerable variability in best model/cycle within the Day 1 time period. No model performed consistently best, no matter the cycle, on a daily basis. This perspective is incorporated into most participant's forecasting approach where we sift through all the guidance looking for similarity and agreement, knowing that "all models are wrong but some are useful" Box (1979). Given the relatively smaller areas of extreme precipitation seen this season, a good portion of model guidance had frequency biases less than 1 on many days. Only on the days with $\geq$30k km^2 did frequency bias have values near or greater than 1. In previous years, model guidance has routinely been associated with frequency bias near 1, when larger scale extreme precipitations events occurred.

This year we kept the half inch contour as the lowest starting point because a dominant strategy used by all participants has been to draw a large encompassing contour of where rain will fall. The addition of the model data within the drawing activity web page did result in more participants drawing smaller polygons at times but this wasn't a systematic finding. Uncertainty of event location dominates how large the polygons will be, especially if multiple precipitation bands might be in relatively close proximity.

The cumulative distribution function (CDF) of CSI scores were broken down by model cycle for Day 1 and compared against participants (Fig 41). On most days, CSI scores indicate that participants are equally competitive with short term model guidance prior to the event. Generally, participants had access to most of the 18z guidance with the exception of the RRFS. At lead times of 36-30 hours (Day 1 00z and 06z cycles) participants were outperforming models across the CSI spectrum. At 24 hours, only the NAMnest had similar scores in the 85th percentile of the CDF to the participants. By 18 hours, the models and participants have converged and the participants are able to improve upon guidance at the 90th percentile. While the sample size declines for 12 hours, participants have been subsumed by the models with HRRR and RRFSp1 having better distributions through most of the CSI spectrum except where they are tied at the upper percentiles. This confirms our graphical depictions from the performance diagrams, that the best participants take advantage of the guidance to make forecasts that score on par with the most recent model guidance, while consistently improving upon guidance from earlier model runs.

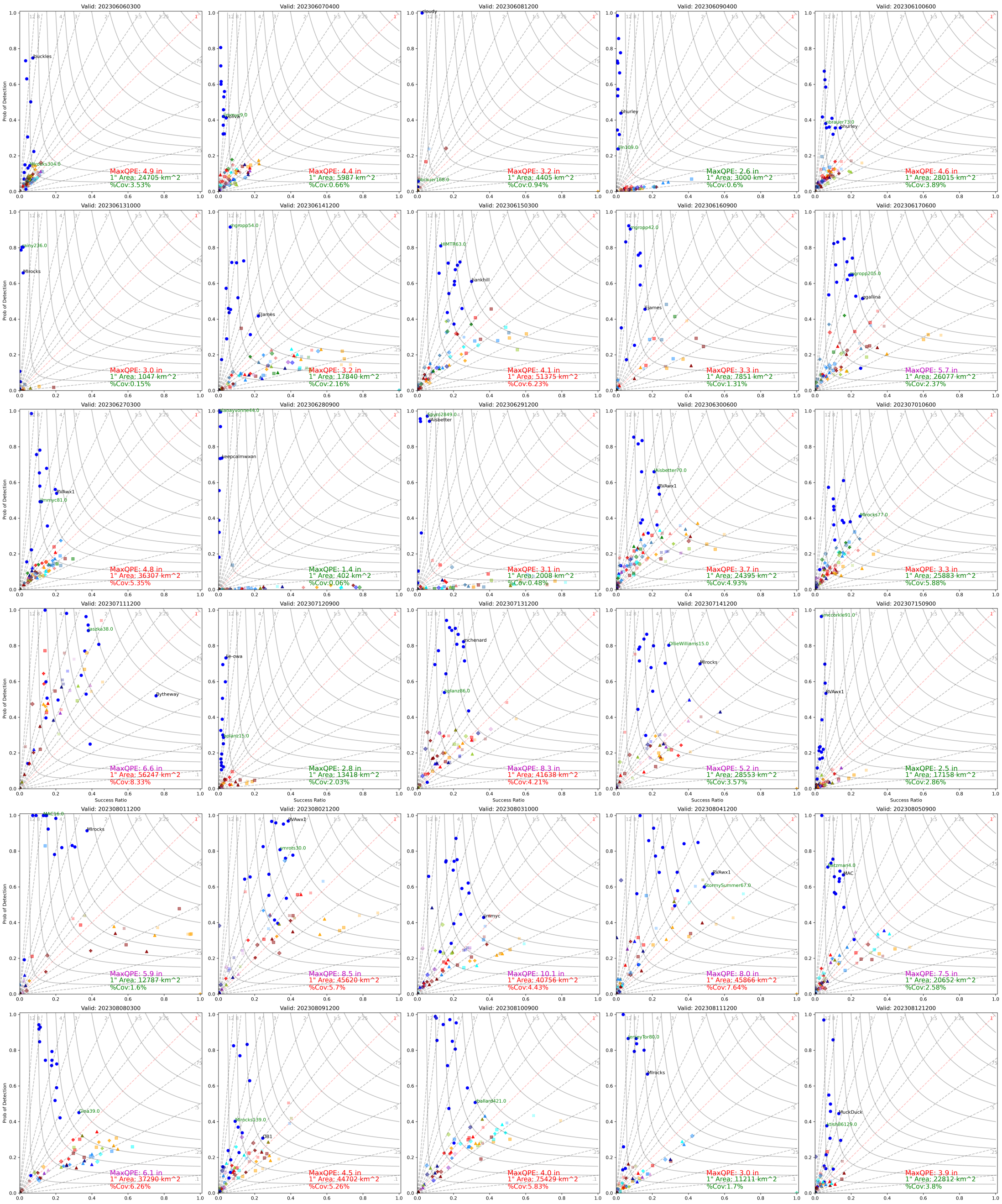


Figure 40: Each MRTP forecast days' performance diagram for a threshold of 1". Blue circles are participants while the other colors represent a model or ensemble, with each symbol a different time grouping. Lighter colors represent models at times that were not available to forecasters. For each day 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).

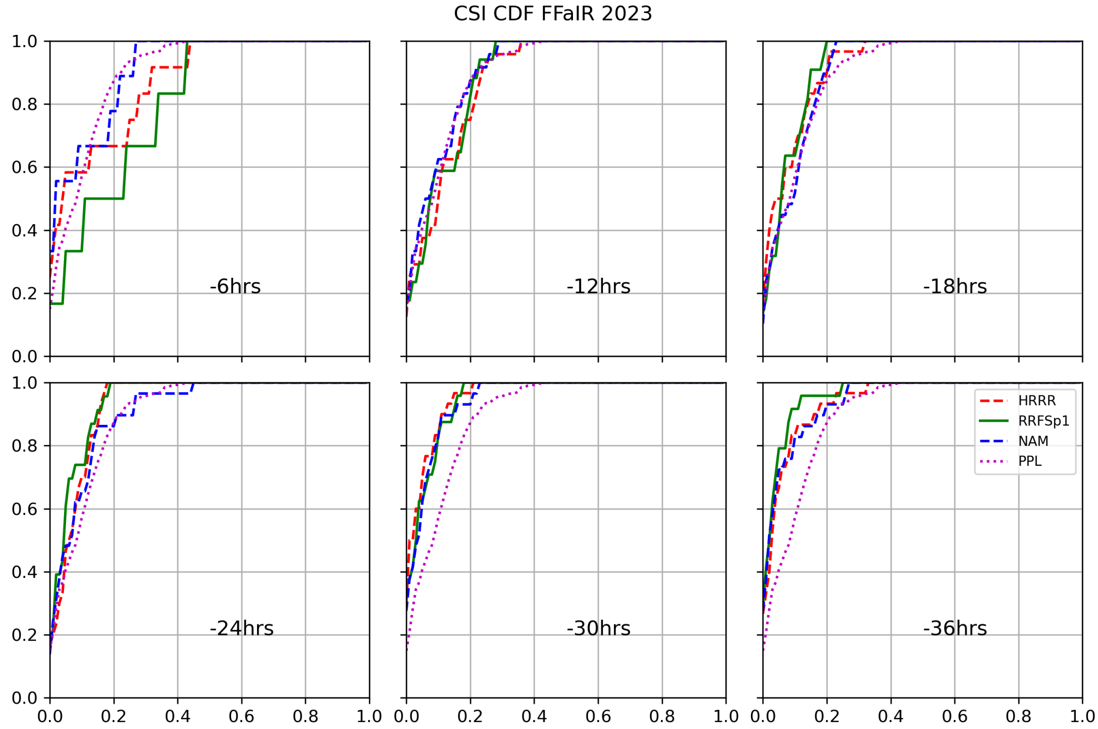


Figure 41: Cumulative distribution function of 1" CSI for the 3 primary models and participant group throughout the MRTP experiment.

Skill over the whole experiment can be seen in Fig 42 for the 1 in threshold. This was done by comparing the 3 facilitators, merging the 6 WPC forecasters into a group called WPC Forecaster and including the 3 primary deterministic models HRRR, NAMnest and RRFSp1 by cycle (max lead time of 60 hours). The 4 aforementioned participants had a minimum number of 27 forecasts, compared to the RRFSp1, which had 24 forecasts. The participants performed well, having higher CSI than most of the guidance with the exception of 00/06z, in aggregate. The HRRR generally performed the best of the 3 models at most cycles, having a dry bias and higher Success Ratio. The NAMnest tended to have a wet bias and higher POD, while the RRFSp1 generally placed in between these two extremes, similar to what was found in the general evaluation of QPF in Section 4.1. Given the relatively small areal coverage of the MRTP events, a cursory analysis was done to understand how the aerial extent of the precipitation impacted the aggregated results. It was found that the top 5 events accounted for 30% of the total aerial coverage of 1", or about twice the expected value. Likewise 50% of the events

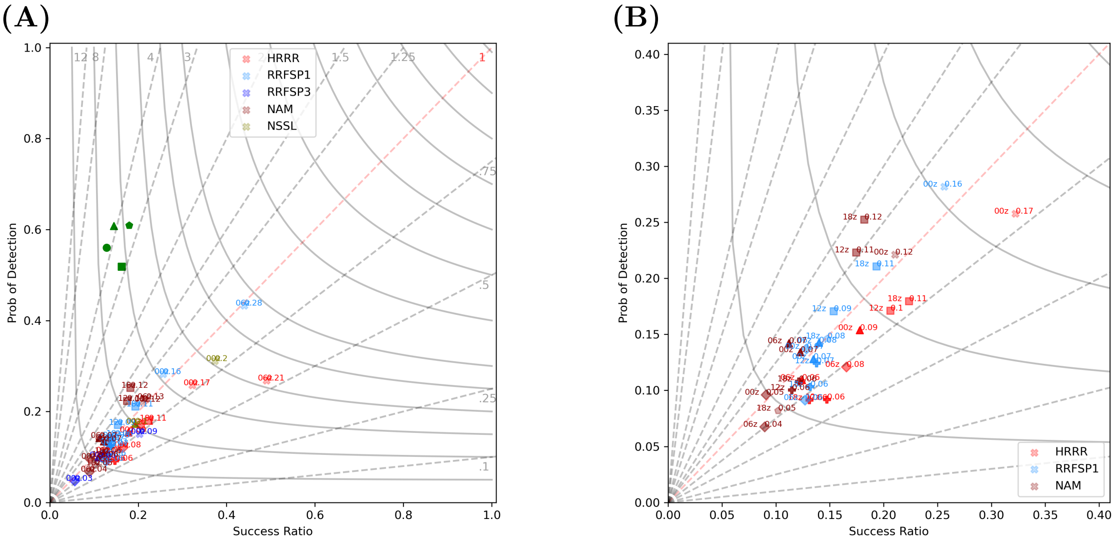


Figure 42: (A) MRTP Performance diagram for the 1” threshold with participants (green symbols) and models and (B) the zoomed performance diagram with just models. The zoomed image does not include model/cycles that had a CSI greater than 0.2: i.e. the 00z NSSL-MPAS, 06z RRFSp1, and the 06z HRRR. Included on the diagrams with the model symbols are the cycle and CSI values.

contribute only 25% of the overall potential skill for 2023, rather than 50% of the skill. This means that the largest events contribute the most to the models skill during FFaIR; aka, not all events have the same impact to the aggregated skill.

4.4.1.3 Distance and Maximum Rain Quality

Given the models’ difficulty with forecasting location precisely, we turn to other evaluation methods to extract utility from the guidance and participants. Throughout the experiment we tracked the MRTP domain maximum rainfall and the distance between the observed and forecast maximum locations. A comparison between the models and participants follows.

The maximum rainfall amounts were compared for each day as averages (model or participant) versus the observed MRMS rainfall maximum for the MRTP domain and time chosen (Fig 43). The symbols get larger as you move further away from perfect calibration line and this yields 8 model days and 9 participant days where the errors were largest. Note that there is a tendency for the forecast rain maximum

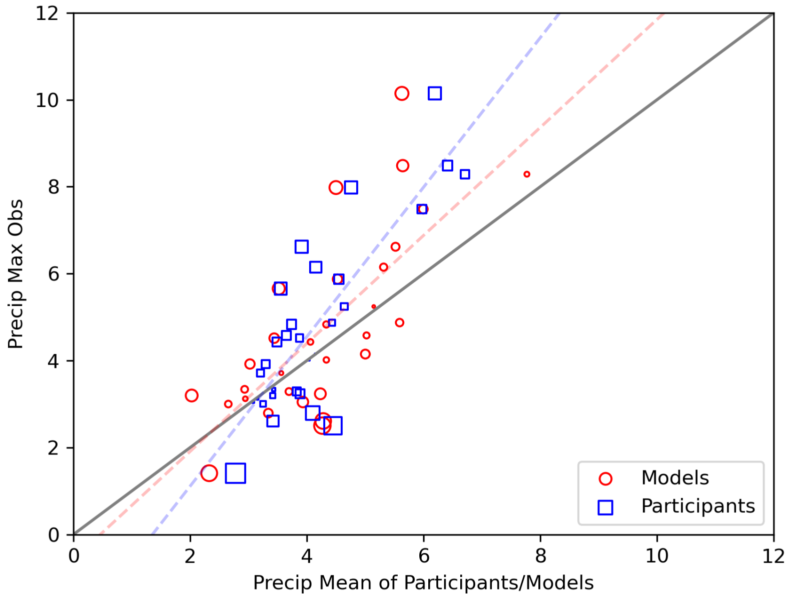


Figure 43: The mean of the model (red/circle) and participant (blue/square) forecasts available for each MRTP event versus the observed maximum rainfall. Best linear fit lines for each group are shown in light dashed lines.

to be overestimated at lower observed rainfall maxima. As the observed rainfall maxima gets larger, the forecasts move toward lower maxima, or an underestimate.

Comparing the maximum rainfall amount via normalized error, and breaking it down by Day 1 models, we can see some trends. Figure 44 shows that the HRRR is the most conservative with 70% of its forecasts below 0%; aka 70% of it's forecasts were below the observed max. The NAMnest and RRFSp1 were more similar to each other, with less than 50% of their forecasts below 0% error; 38% and 43% respectively. The NAMnest has the longest tail of the CDFs indicating its tendency to over predict the rainfall maximum value by more than 80%. The rain error characteristics shown in Fig 45 for modeled rain versus percent error, shows that the percent errors for the NAMnest occur somewhat evenly across low and high maximum rainfall totals for its overforecasts. Note that the participants CDF tracked more closely with the HRRR than the other models, Fig. 44, indicating that they either preferred the HRRR, or maybe that participants favored a more conservative approach to forecasting or both.

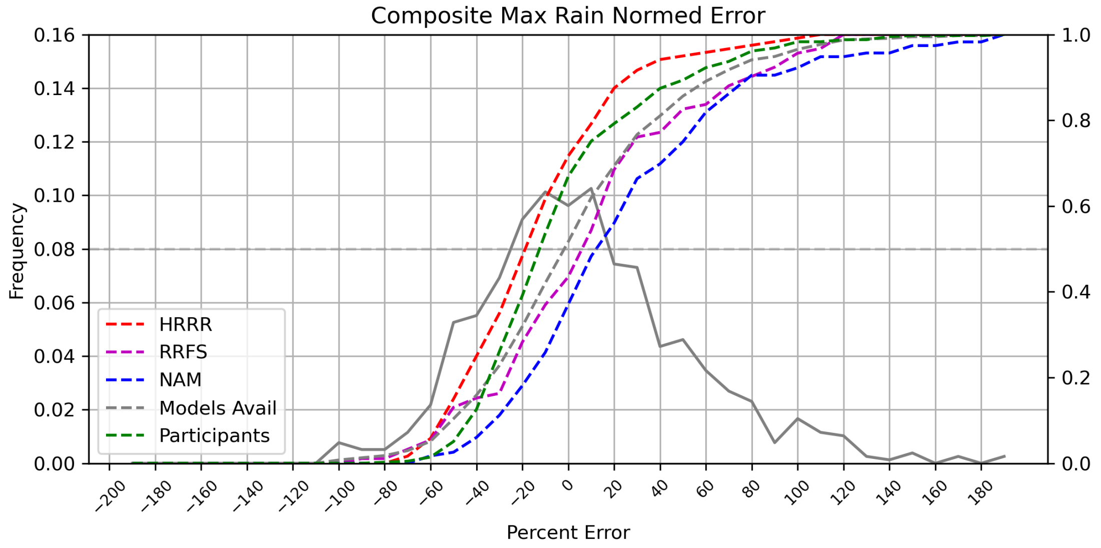


Figure 44: Composite CDF of maximum rainfall amount for normalized error for the valid 6-h MRTP time window for the Day 1 cycles for the HRRR (red), RRFSp1 (purple), NAMnest (blue) and the Participants (green). The dashed grey line represents Models Avail., which is defined as all models and forecast hours available for that valid time and solid grey is the underlying histogram for models available.

Distance errors broken down by Day 1 models were very different from rainfall maxima. The CDFs for models and participants (Fig 46) shows that the HRRR was within 100 km 14% of the time compared to the NAMnest at 7.8% and the RRFSp1 at 5.5%. Participants were more similar to the NAMnest at this range but quickly transitioned to be more consistent with the HRRR and better than all models after 175 km. Thus participants' locations of maximum rainfall were more likely to be closer to the observed maximum rainfall.

4.4.1.4 Probabilistic Components

The latest addition to MRTP was to add probabilistic components, including forecasting the probability of exceeding a threshold for the maximum rainfall and the timing probability. These followed the methods of Lawson (2023) and references therein to evaluate the forecast using Information Theory. The forecasts for exceedance are evaluated using Lawson (2023) Game 1, where we assume there is an equal likelihood that the rain threshold will be exceeded or not. Figure 47 shows a comparison of the Information Gained in bits and evaluates all the forecasts in a

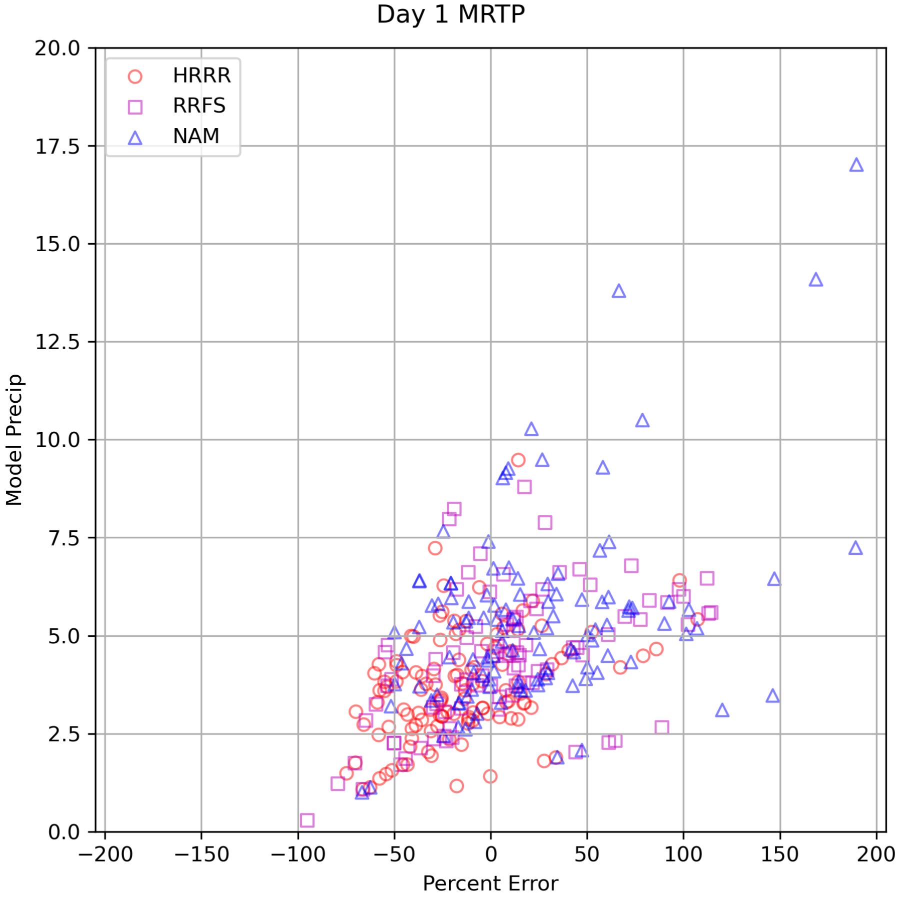


Figure 45: Maximum rain normalized percent error versus predicted 6-h rainfall maximum for the valid 6-h MRTP time window for the Day 1 cycles.

single CDF. The evaluation reveals that roughly 45% of the participants forecasts were associated with positive information gain, or that when the threshold was exceeded the forecasters had indicated that there was greater than 50 percent probability of exceedance. This is lower than expected by chance.

Pairing the probability of rainfall exceedance over the max rain threshold against the participants' rainfall maxima results in the construction of an attributes or reliability diagram Hsu and Murphy (1986). This can be thought of as something akin to consistency of the predictions. The attributes diagram constructed for the MRTP experiment (Fig 48) reveals that participants under forecast for all but the 40 and 100 percent bins. So although some participants added information with the probability values, the resulting maximum rainfall they chose were under

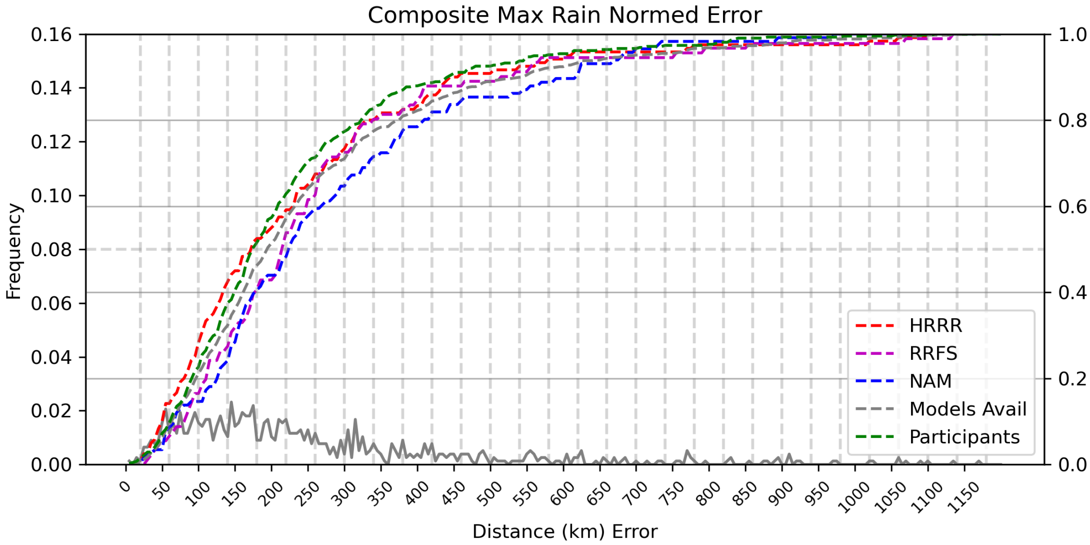


Figure 46: Like Fig. 44 but for the composite CDF of the distance between forecast and observed domain maximum rainfall of normalized error for each model and the participants.

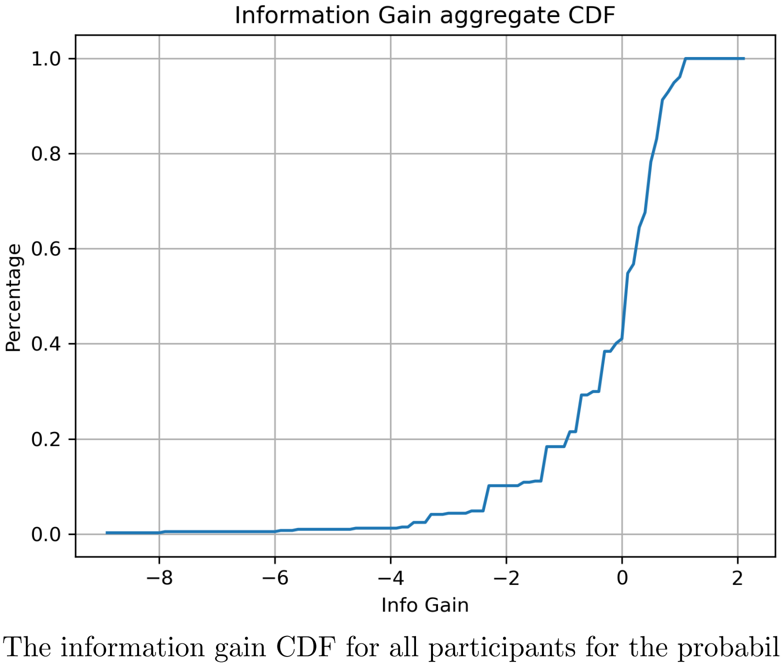


Figure 47: The information gain CDF for all participants for the probability of exceeding a threshold for the maximum rainfall across the MRTP experiment.

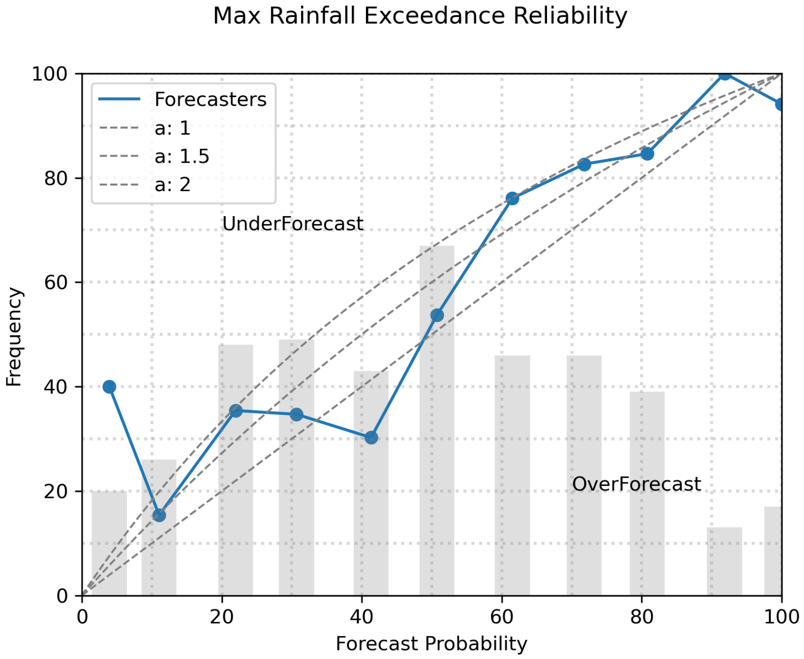


Figure 48: Attributes diagram showing frequency of forecast probability (gray bars) using the corrective strategy of Bröcker and Smith (2007) computing the mean probability in each decile probability bin, along with the reliability of the forecast probabilities (blue line and circle markers).

forecast and thus slightly inconsistent or perhaps uncalibrated with probabilistic judgments.

Participants were further asked when the peak in rainfall would occur as a probabilistic judgement. Ten bins, each representing the valid end time of a 6-h window from 03 to 12 UTC (ex. 03 UTC would be for the 6-h period from 21 UTC to 03 UTC) were used along with a maximum allocation of 100 percentage points. The results from this exercise can be seen in Fig 49. In general participants expressed uncertainty in the timing of maximum 6-h rainfall (more than one bin had a nonzero value), with the number of bins typically used ranging from 3 to 5 bins. Throughout the course of the experiment it was unlikely that participants used only 1 bin (i.e. putting 100% in a single bin), happening about 25 times. This indicated a certain and deterministic forecast. The maximum probability used in the latter case was 100 percent, but the former had a sharp distribution centered around 20 percent with a long tail.

The probability distribution at the time that verified as the peak rainfall was Poisson-like from 0 indicating that participants had the correct time in their envelope of probability but were not correctly forecasting the correct bin with their peak probability values. Thus the Ignorance Score results were shifted to below zero, with little information value added. Figure 50 shows the timing difference between observed peak rainfall and consensus time of the MRTP forecast (red), along with the participants difference derived from the peak probability in the

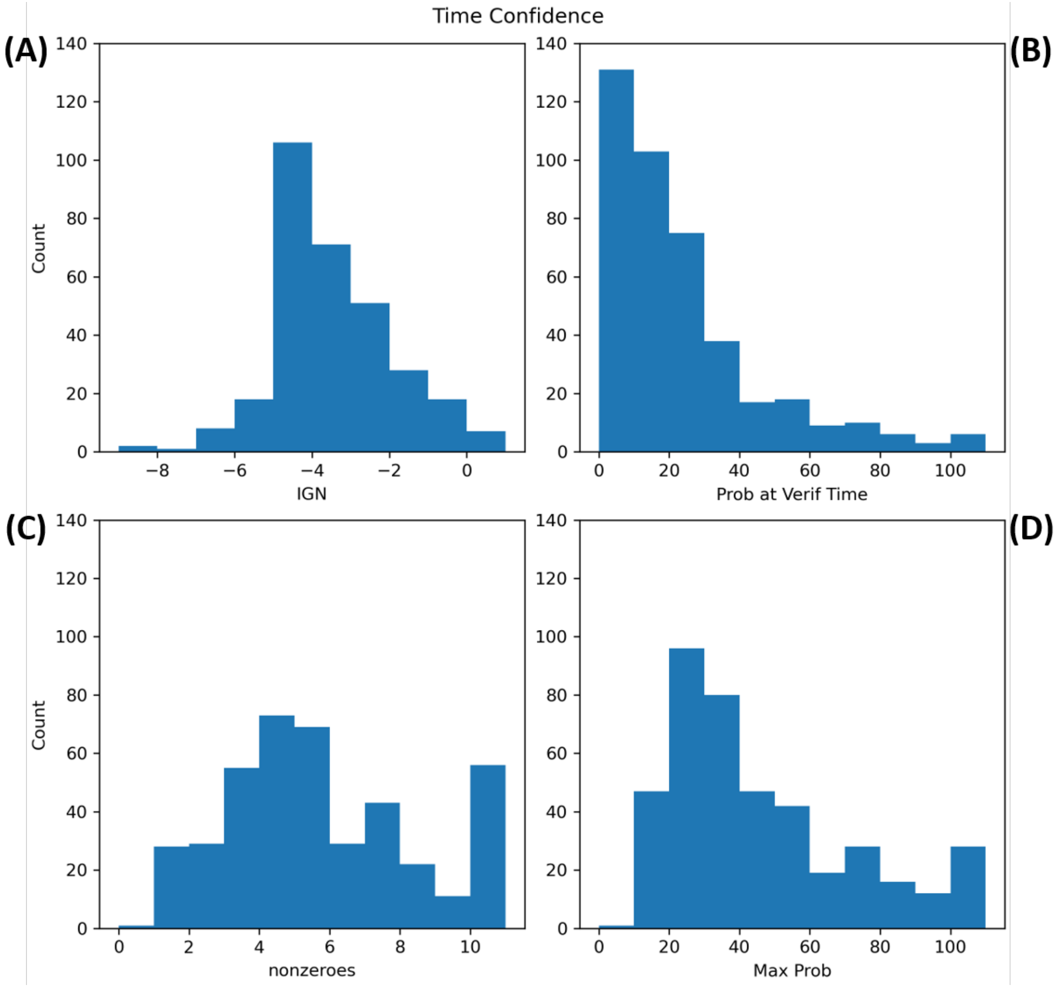


Figure 49: The evaluation of the timing confidence as (A) the Ignorance score and (B) Histogram of ignorance values. (C) the forecast probability histogram valid at the verification time and the histogram of the number of nonzero bins over all forecasts and (D) the histogram of the Maximum Probability for each forecaster.

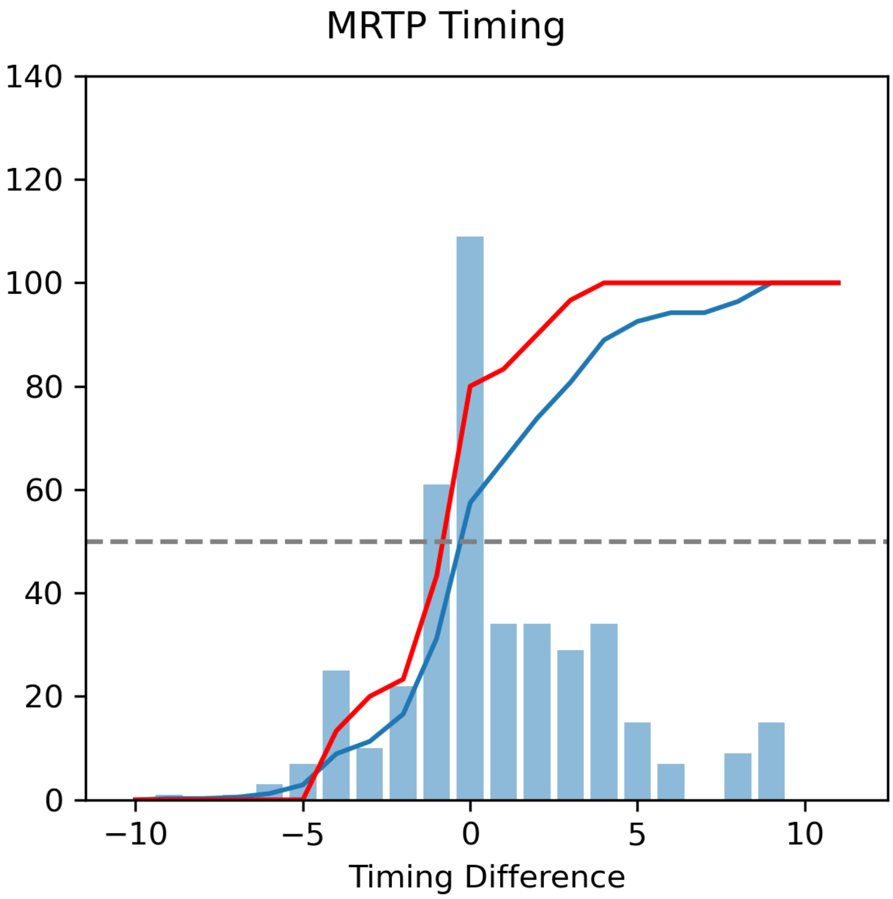


Figure 50: MRTP timing difference of the participants across FFaIR. Histogram of the occurrence of the differences (blue bars). The red and blue lines represent the timing difference between observed peak rainfall and consensus time of the MRTP forecast and the participants difference derived from the peak probability in the timing confidence, respectively.

timing confidence (blue). This shows that the timing differences from individuals are strongly peaked at 0 and -1 hours relative to the observed.

The subjective evaluation of MRTP forecasts involved "goodness" rankings from 1 (worst)-7 (best) for the assigned model, the most used model and the participants own forecasts, as shown in Fig 51. All 3 peaked around a likert score of 4/7. The assigned model generally performed worse than either the participants or the used model. The Used (Assigned) Model becomes worse than participant scores in normalized percentage difference at a score of 5, and is 25 (50) percent lower than participants score at values greater than 5, respectively. Our hypothesis

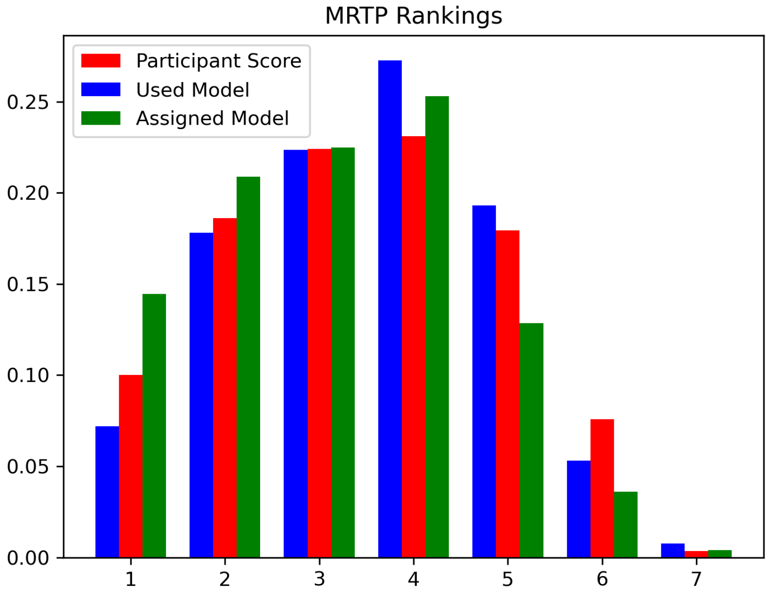


Figure 51: Frequency of ratings for the participants forecast (red) and the corresponding model forecasts that were used as assigned (green) or actually used (blue). The Likert scale ranged from 1 to 7.

is that the precision that models can provide in terms of location, areal coverage, number of features and precipitation maxima all play a role in perceptions of quality or goodness (Murphy, 1993). Participant strategy is in contrast to model precision, mostly being broader in forecast areal coverage. However, many participants attempted to be precise to match model bias. This seldom leads to quality as evaluated quantitatively using CSI. A breakdown of those forecasts employing different strategies and their quantitative outcome via CSI is needed.

5 Summary

Part 1 of the 2023 FFaIR Final Report focused on the evaluation of the RRFS and other similar CAM deterministic and ensemble models and associated products. With the approaching science freeze for the RRFS, the team felt it was important to provide an in-depth analysis of the system during the 2023 Testbed Season[18]. Evaluation was done for QPF at 24-h, 6-h and 1-h accumulations. Additionally, the MRTP actively was used to evaluate model performance at identifying the extreme precipitation risk, both location and timing.

Both subjective and objective verification suggest that the RRFSp1 performance falls between the NAMnest and the HRRR, with a bias similar (sometimes larger) to the NAMnest but a CSI closer to the HRRR[19], especially for 24-h. For 6-h accumulations, the exception to this was during the overnight hours, roughly from 06-12 UTC. Here the RRFSp1 was closer to the HRRR in having a dry bias. This likely suggests that the RRFSp1 struggles to keep convection active or develop new convection once daytime heating diminishes. This is a trend the team has also been seeing in the HRRR over the past two FFaIR seasons, with it struggling to identify the development of MCSs. An example of this was shown in Fig. 17, where the NAMnest was able to identify the MCS risk at forecast hour 42 but neither of the other two models did. Even at forecast hour 24 the HRRR and RRFSp1 struggled to identify the risk for heavy rainfall; see Fig. 18.

This lack of convection, from May thru Aug 2023, is further supported when looking at precipitation across the diurnal cycle (Fig. 27), in terms of average precipitation across the CONUS for both coverage and intensity, the RRFSp1 2023 version shows lower totals than the 2022 version; it is even lower than the HRRR in some instances. It is possible that some of the changes made between 2022 and 2023 to address the over abundance of popcorn convection and strong updrafts had resulted in preventing the system from developing convection when diurnal heating waned.

[18] Reminder, this is May - Aug of 2023 to include when the HWT SFE is running.

[19] The HRRR had a noticeable dry bias this year.

Regarding popcorn convection, as mentioned in the summary of the weather during FFaIR (Section 2.3), the synoptic conditions favoring popcorn convection were missing during the majority of FFaIR and the Testbed season this year. Therefore, aside from over Florida, coverage of popcorn convection was minimal, which perhaps suggests that the RRFSp1 correctly wasn't developing it in a regime that doesn't support it. That said, participants still noted that they saw very small storms (rather than the larger sized ones from last year) with high precipitation totals often. They also noted that they felt the evolution of systems did not look correct, stating: "it (RRFSp1) presents unrealistic structures that don't align with how I expect severe storms to look in modeled environments" and "RRFSp1 struggled with storm structure, even in the first few hours." This, along with the high bias, made it difficult for participants to trust the RRFSp1 forecasts.

As for the performance of the RRFS ensemble, due to data flow challenges, an analysis of the four ensemble configurations provided was not completed. Moreover, even if the data had been consistent, the evaluation would be obsolete given that right after FFaIR it was decided that ensemble membership would consist of the HRRR and its time-lagged forecast for implementation, which was not a configuration during FFaIR. Therefore, due to the high bias in QPF, especially at the hourly scale, the unrealistic looking footprint of the QPF, the lull in convection when diurnal heating is not present, and the fact that the ensemble was not able to be properly evaluated, we do not support the transition of the RRFS to operations but recommend it for further development and testing.

For the CAPS provided products, the spatially aligned means and the ML 6-h QPF product are both recommended for further development and testing. The SAM-LPM was well received by the participants and they liked the ability of the SAM method to remove the speckled look of the standard LPM mean. However, they did note that they felt the method increased the size of the footprint of high-end totals too much while making the footprint of low end (0.01-0.1 in) amounts too small. The ML product, referred to as the HREF+, was found to have a bug in the calculations for the neighborhood probabilities, and therefore was not evaluated in this year's experiment. However, the HREF+ was evaluated in the

2022 Experiment (only for the half inch threshold) and showed promise, so the team is excited to evaluate the product in the upcoming 2024 FFaIR Experiment.

References

Banacos, P., 2023: The great vermont flood of 10-11 july 2023: Preliminary meteorological summary. URL https://www.weather.gov/btv/The-Great-Vermont-Flood-of-10-11-July-2023-Preliminary-Meteorological-Summary, accessed on Nov 27, 2023.

Box, G. E. P., 1979: Robustness in the strategy of scientific model building. *Robustness in Statistics*, G. N. W. R. L. Launer, Ed., Academic Press, 201–236.

Brewster, K. A., 2003: Phase-correcting data assimilation and application to storm-scale numerical weather prediction. part i: Method description and simulation testing. *Mon. Wea. Rev.*, **3**, 480–492, https://doi.org/https://doi.org/10.1175/1520-0493(2003)131\%3C0480:PCDAAA\%3E2.0.CO;2.

Bröcker, J., and L. A. Smith, 2007: Increasing the reliability of reliability diagrams. *Weather and Forecasting*, **22 (3)**, 651 – 661, https://doi.org/https://doi.org/10.1175/WAF993.1, URL https://journals.ametsoc.org/view/journals/wefo/22/3/waf993_1.xml.

Clark, A. J., and Coauthors, 2023: Spring forecasting experiment 2023 conducted by the experimental forecast program of the noaa hazardous weather testbed. Tech. rep., NCEP SPC and NOAA/OAR NSSL. URL https://hwt.nssl.noaa.gov/sfe/2023/docs/HWT_SFE_2023_Prelim_Findings_v1.pdf.

Espinoza, M., and FOX 29 staff, 2023: Delaware valley, lehigh valley overwhelmed as heavy rain causes flash flooding. URL https://www.fox29.com/news/delaware-valley-lehigh-valley-overwhelmed-with-heavy-rain-causing-flash-floods, accessed on July 11, 2023.

Hsu, W., and A. H. Murphy, 1986: The attributes diagram a geometrical framework for assessing the quality of probability forecasts. *International Journal of Forecasting*, **2 (3)**, 285–293, https://doi.org/https://doi.org/10.1016/0169-2070(86)90048-8, URL https://www.sciencedirect.com/science/article/pii/0169207086900488.

Medina, E., L. Albeck-Ripka, and J. McKinley, 2023: At least 1 dead as heavy rains set off flash flooding in new york. URL https://www.nytimes.com/2023/07/

09/nyregion/flooding-west-point-orange-county.html, accessed on July 11, 2023.

Murphy, A. H., 1993: What is a good forecast? an essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, **8 (2)**, 281 – 293, https://doi.org/https://doi.org/10.1175/1520-0434(1993)008⟨0281:WIAGFA⟩2.0.CO;2, URL https://journals.ametsoc.org/view/journals/wefo/8/2/1520-0434_1993_008_0281_wiagfa_2_0_co_2.xml.

NOAA Physical Sciences Laboratory, 2023: Monthly/seasonal climate composites. URL https://psl.noaa.gov/cgi-bin/data/composites/printpage.pl, accessed on Oct. 19, 2023.

Roebber, P. J., 2009: Visualizing multiple measures of forecast quality. *Weather and Forecasting*, **24 (2)**, 601 – 608, https://doi.org/https://doi.org/10.1175/2008WAF2222159.1, URL https://journals.ametsoc.org/view/journals/wefo/24/2/2008waf2222159_1.xml.

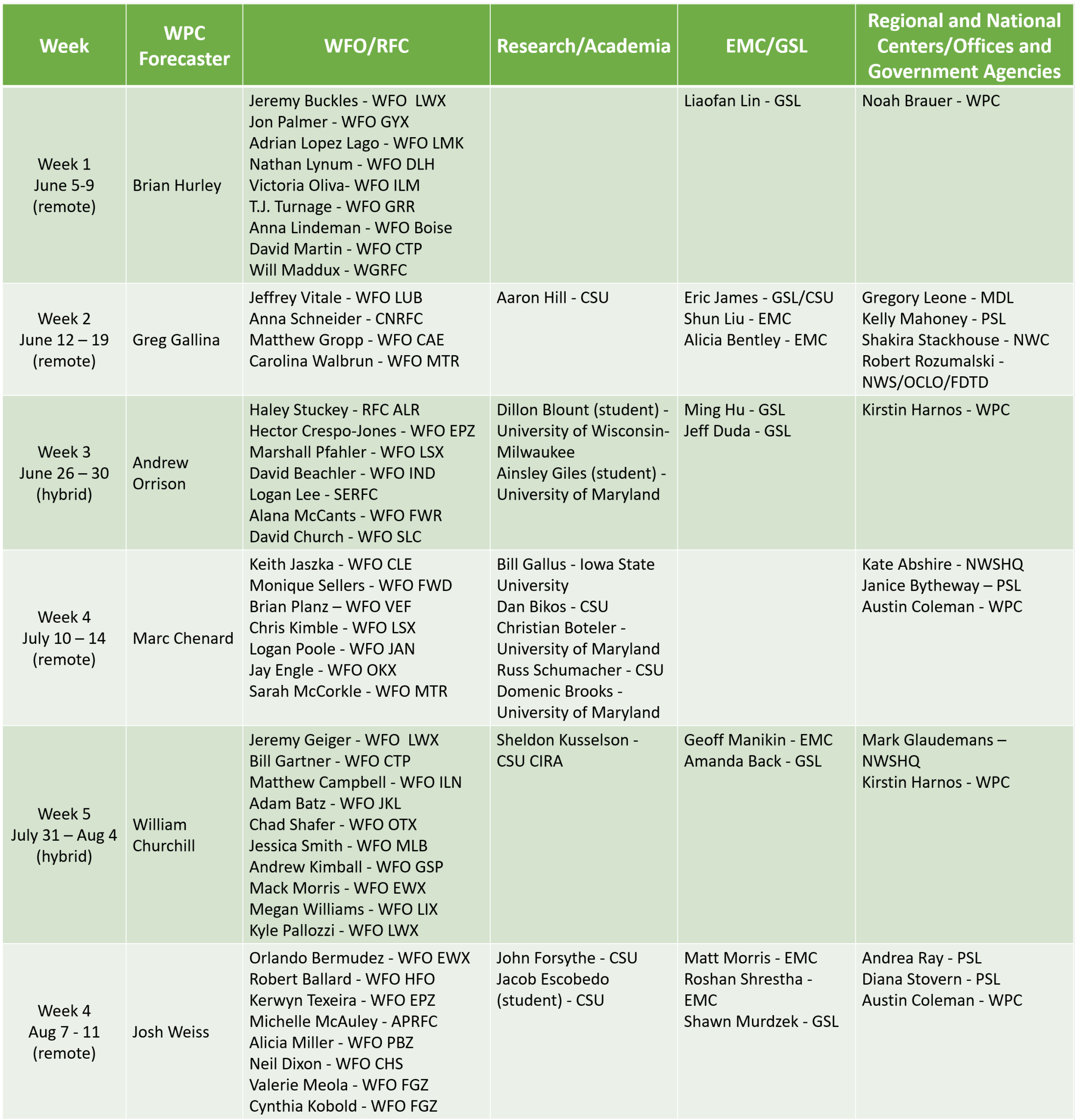
Trojniak, S., and J. Correia, Jr., 2023: 2023 ffair operations plan, published online at https://www.wpc.ncep.noaa.gov/hmt/Reports/FFaIR/2023_FFaIR_Operations_Plan.pdf. If missing please contact WPC.

Trojniak, S., J. Correia, Jr., and B. Albright, 2020: 2020 flash flood and intense rainfall experiment: Findings and results. Tech. rep., NCEP WPC-HMT. URL https://www.wpc.ncep.noaa.gov/hmt/Final_Report_2020_FFaIR_Experiment_Nov13.pdf.

Appendices

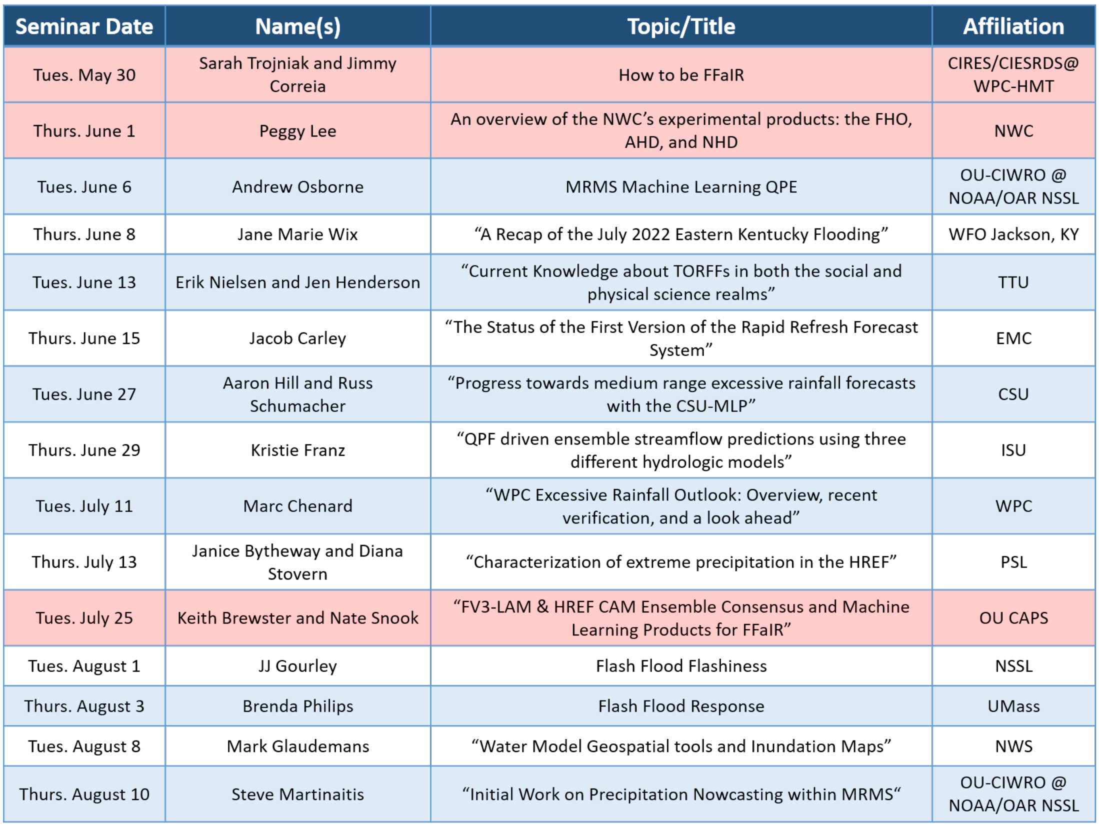
A List of Participants and Seminars

Table 4: List of the participants for each week of the 2023 FFaIR Experiment.



Week	WPC Forecaster	WFO/RFC	Research/Academia	EMC/GSL	Regional and National Centers/Offices and Government Agencies
Week 1 June 5-9 (remote)	Brian Hurley	Jeremy Buckles - WFO LWX Jon Palmer - WFO GYX Adrian Lopez Lago - WFO LMK Nathan Lynum - WFO DLH Victoria Oliva- WFO ILM T.J. Turnage - WFO GRR Anna Lindeman - WFO Boise David Martin - WFO CTP Will Maddux - WGRFC		Liaofan Lin - GSL	Noah Brauer - WPC
Week 2 June 12 – 19 (remote)	Greg Gallina	Jeffrey Vitale - WFO LUB Anna Schneider - CNRFC Matthew Gropp - WFO CAE Carolina Walbrun - WFO MTR	Aaron Hill - CSU	Eric James - GSL/CSU Shun Liu - EMC Alicia Bentley - EMC	Gregory Leone - MDL Kelly Mahoney - PSL Shakira Stackhouse - NWC Robert Rozumalski - NWS/OCLO/FDTD
Week 3 June 26 – 30 (hybrid)	Andrew Orrison	Haley Stuckey - RFC ALR Hector Crespo-Jones - WFO EPZ Marshall Pfahler - WFO LSX David Beachler - WFO IND Logan Lee - SERFC Alana McCants - WFO FWR David Church - WFO SLC	Dillon Blount (student) - University of Wisconsin-Milwaukee Ainsley Giles (student) - University of Maryland	Ming Hu - GSL Jeff Duda - GSL	Kirstin Harnos - WPC
Week 4 July 10 – 14 (remote)	Marc Chenard	Keith Jaszka - WFO CLE Monique Sellers - WFO FWD Brian Planz – WFO VEF Chris Kimble - WFO LSX Logan Poole - WFO JAN Jay Engle - WFO OKX Sarah McCorkle - WFO MTR	Bill Gallus - Iowa State University Dan Bikos - CSU Christian Boteler - University of Maryland Russ Schumacher - CSU Domenic Brooks - University of Maryland		Kate Abshire - NWSHQ Janice Bytheway – PSL Austin Coleman - WPC
Week 5 July 31 – Aug 4 (hybrid)	William Churchill	Jeremy Geiger - WFO LWX Bill Gartner - WFO CTP Matthew Campbell - WFO ILN Adam Batz - WFO JKL Chad Shafer - WFO OTX Jessica Smith - WFO MLB Andrew Kimball - WFO GSP Mack Morris - WFO EWX Megan Williams - WFO LIX Kyle Pallozzi - WFO LWX	Sheldon Kusselson - CSU CIRA	Geoff Manikin - EMC Amanda Back - GSL	Mark Glaudemans – NWSHQ Kirstin Harnos - WPC
Week 4 Aug 7 - 11 (remote)	Josh Weiss	Orlando Bermudez - WFO EWX Robert Ballard - WFO HFO Kerwyn Texeira - WFO EPZ Michelle McAuley - APRFC Alicia Miller - WFO PBZ Neil Dixon - WFO CHS Valerie Meola - WFO FGZ Cynthia Kobold - WFO FGZ	John Forsythe - CSU Jacob Escobedo (student) - CSU	Matt Morris - EMC Roshan Shrestha - EMC Shawn Murdzek - GSL	Andrea Ray - PSL Diana Stovern - PSL Austin Coleman - WPC

Table 5: List of the 2023 FFaIR Science Seminars. The slides for the seminars can be found here. The rows in red indicate seminars given outside of the weeks the FFaIR was in session.



Seminar Date	Name(s)	Topic/Title	Affiliation
Tues. May 30	Sarah Trojniak and Jimmy Correia	How to be FFaIR	CIRES/CIESRDS@ WPC-HMT
Thurs. June 1	Peggy Lee	An overview of the NWC's experimental products: the FHO, AHD, and NHD	NWC
Tues. June 6	Andrew Osborne	MRMS Machine Learning QPE	OU-CIWRO @ NOAA/OAR NSSL
Thurs. June 8	Jane Marie Wix	"A Recap of the July 2022 Eastern Kentucky Flooding"	WFO Jackson, KY
Tues. June 13	Erik Nielsen and Jen Henderson	"Current Knowledge about TORFFs in both the social and physical science realms"	TTU
Thurs. June 15	Jacob Carley	"The Status of the First Version of the Rapid Refresh Forecast System"	EMC
Tues. June 27	Aaron Hill and Russ Schumacher	"Progress towards medium range excessive rainfall forecasts with the CSU-MLP"	CSU
Thurs. June 29	Kristie Franz	"QPF driven ensemble streamflow predictions using three different hydrologic models"	ISU
Tues. July 11	Marc Chenard	"WPC Excessive Rainfall Outlook: Overview, recent verification, and a look ahead"	WPC
Thurs. July 13	Janice Bytheway and Diana Stovern	"Characterization of extreme precipitation in the HREF"	PSL
Tues. July 25	Keith Brewster and Nate Snook	"FV3-LAM & HREF CAM Ensemble Consensus and Machine Learning Products for FFaIR"	OU CAPS
Tues. August 1	JJ Gourley	Flash Flood Flashiness	NSSL
Thurs. August 3	Brenda Philips	Flash Flood Response	UMass
Tues. August 8	Mark Glaudemans	"Water Model Geospatial tools and Inundation Maps"	NWS
Thurs. August 10	Steve Martinaitis	"Initial Work on Precipitation Nowcasting within MRMS"	OU-CIWRO @ NOAA/OAR NSSL