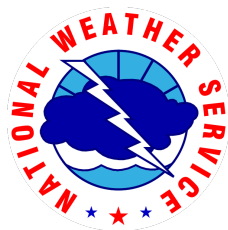


2022 Flash Flood and Intense Rainfall (FFaIR) Final Report: *Results and Findings*

June 21 - July 22, 2022
Weather Prediction Center
Hydrometeorology Testbed

Sarah Trojniak - CIRES CU Boulder, NOAA/NWS/WPC/HMT

James Correia Jr. - CIRES CU Boulder,
NOAA/NWS/WPC/HMT



Contents

1	Introduction	2
2	Science and Operations	5
2.1	Daily Operations	6
2.2	Forecasting Activities	9
2.3	Data and Products Provided	12
2.3.1	Model Information	14
2.3.2	Machine Learning Products	16
2.4	Science Questions and Verification Methods	18
2.4.1	Experiment Goals	18
2.4.2	Verification	20
3	Results	27
3.1	QPF	28
3.1.1	A Short Summary	47
3.2	Note on Ensembles	48
3.3	Precipitation Rate	50
3.4	The EROs and AERO	57
3.4.1	CSU MLP and FFaIR EROs	58
3.4.2	FFaIR AERO	71
3.5	MRTP	85
3.5.1	Human and Model Performance	87
3.5.1.1	Performance Diagrams for Accumulation	87
3.5.1.2	Distance and Maximum Rain Quality	92
4	Summary and Conclusions	95
	Appendices	100
A	List of Participants and Seminars	100
B	MRTP Workflow	101
C	FFaIR Surveys	107
C.1	2022 FFaIR Verification Survey	107
C.2	2022 End of the Week Feedback on FFaIR and its Products Survey	118

1 Introduction

The 10th Annual Flash Flood and Intense Rainfall (FFaIR) Experiment, which is part of the Hydrometeorology Testbed (HMT) at the Weather Prediction Center (WPC), was once again held completely virtual. FFaIR focuses on the evaluation of the utility of new guidance and tools to help forecast heavy rainfall and flash flooding. It also strives to better understand the challenges of forecasting these events. National Weather Service (NWS) Testbeds are unique as they bring together people from across the weather enterprise, ranging from developers to forecasters and researchers, to work collaboratively to advance the science.

The 10 year anniversary of FFaIR included the challenges of fine scale extreme rainfall and an overall reduction of extreme rainfall coverage. Although extreme rainfall and flooding are never wanted, it is difficult to evaluate the performance of model guidance and tools when events do not occur. Figure 1 shows the rainfall totals for the 2021 FFaIR season compared to this year using Multi-Radar Multi-Sensor Gauge Corrected (hereafter MRMS) Quantitative Precipitation Estimate (QPE). The scarcity of rainfall is very apparent from the Ohio River Valley to the Central Plains and into eastern Texas. This resulted in very few large scale, heavy rainfall events and no a Moderate or High risk was issued for the WPC or FFaIR Excessive Rainfall Outlook (ERO)¹. The 24-h QPE for the 19 days of FFaIR can be seen in Figs. 2 - 5.

Additionally, very few Mesoscale Convective Systems (MCSs) were observed during FFaIR. The lack of MCSs resulted in heavy rainfall precipitation events that were generally small in scale and thus harder to forecast. For instance, one of the most impactful events that occurred was located along the southern border of Virginia and West Virginia, where over 6 in of rain fell in six hours (Fig. 6A) overnight from July 12th to the 13th in the slot canyons of the region. This led to flooding that washed away homes and roads, caused mudslides, and initially led to (~40) missing people; thankfully they all ended up being accounted for (Whetstone et al., 2022). Figures 6B-D show some of the destruction caused by

¹A forecasting activity in FFaIR is issuing an ERO, the product and process will be discussed in Section 2.

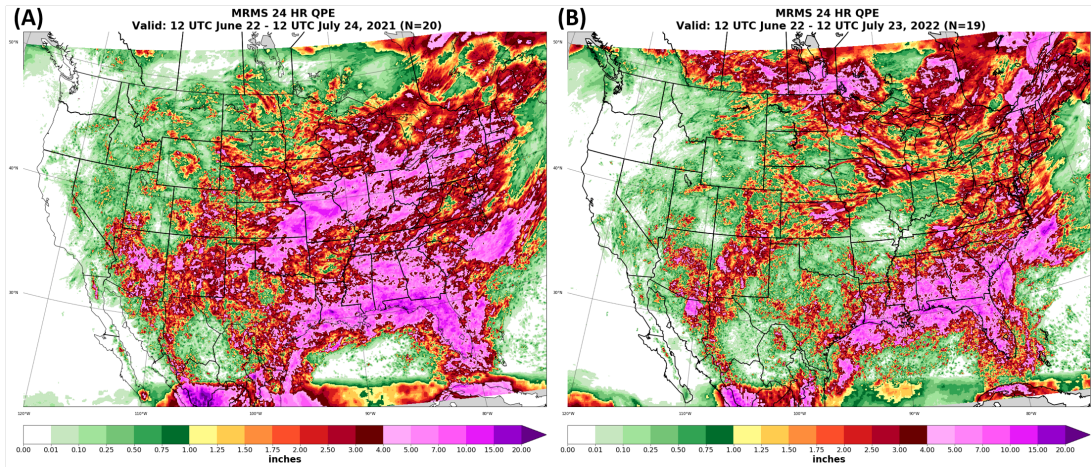


Figure 1: Accumulation of MRMS QPE for the days FFaIR was in session for (A) 2021 and (B) 2022.

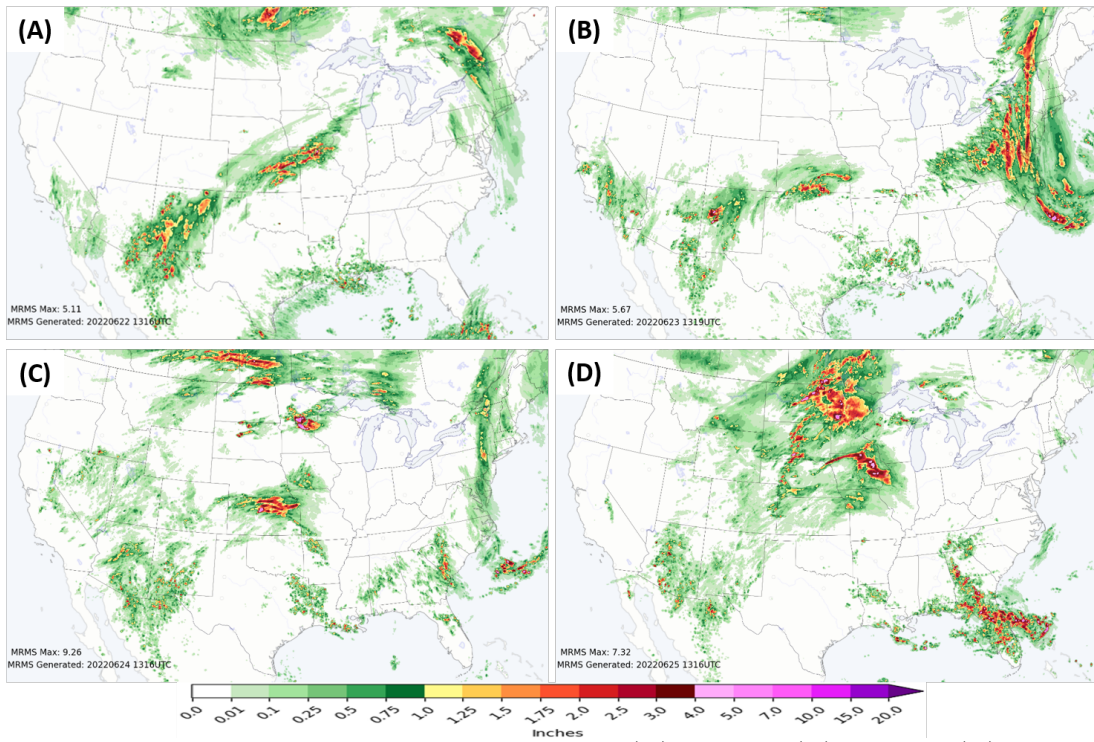


Figure 2: 24-h MRMS QPE valid at 12 UTC on (A) June 22 (B) June 23 (C) June 24 and (D) June 25, 2022.

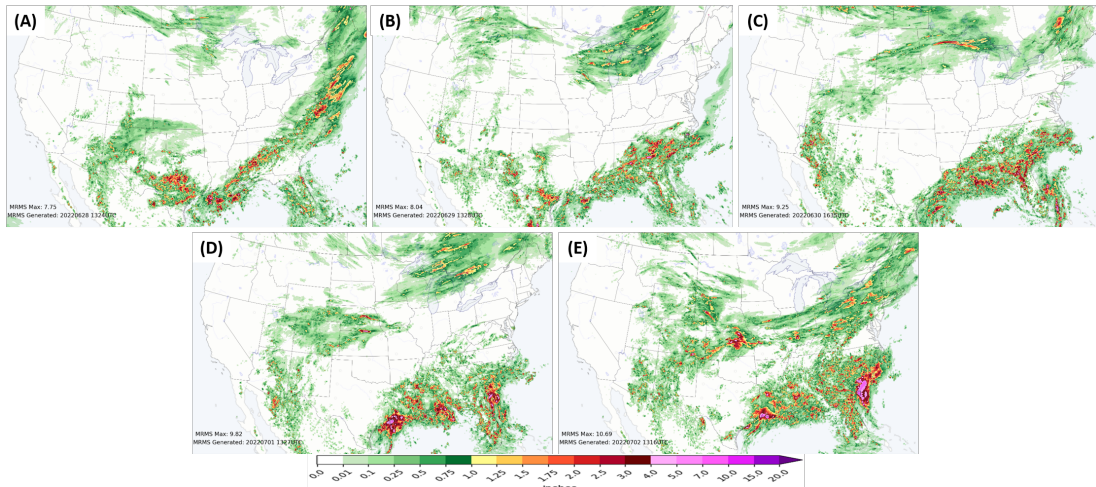


Figure 3: 24-h MRMS QPE valid at 12 UTC on (A) June 28 (B) June 29 (C) June 30 (D) July 01 and (E) July 02, 2022.

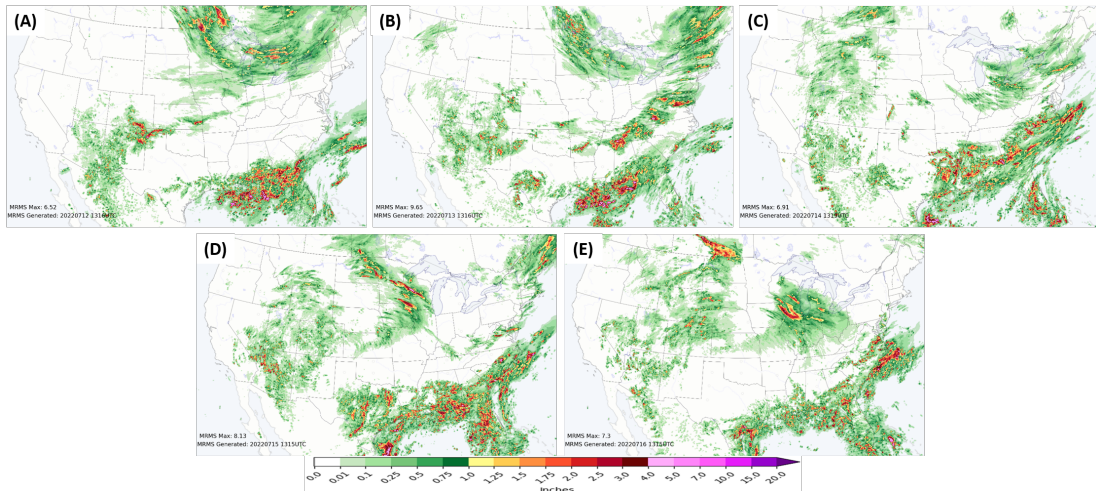


Figure 4: 24-h MRMS QPE valid at 12 UTC on (A) July 12 (B) July 13 (C) July 14 (D) July 15 and (E) July 16, 2022.

the rushing water and mudslides. In comparison, one of the larger events occurred over night from July 20th to the 21st. Across the Knoxville region, a back-building line over the city resulted in 8+ inches of rain and widespread flooding in the area (Fig. 7). Comparison of the 3 inch contours in Figs. 6A and 7A helps to show the difference in scale and continuity between the two events, with isolated spots of heavy precipitation in the July 12-13 case rather than the larger, continuous region of higher precipitation totals seen in the July 20-21 case.

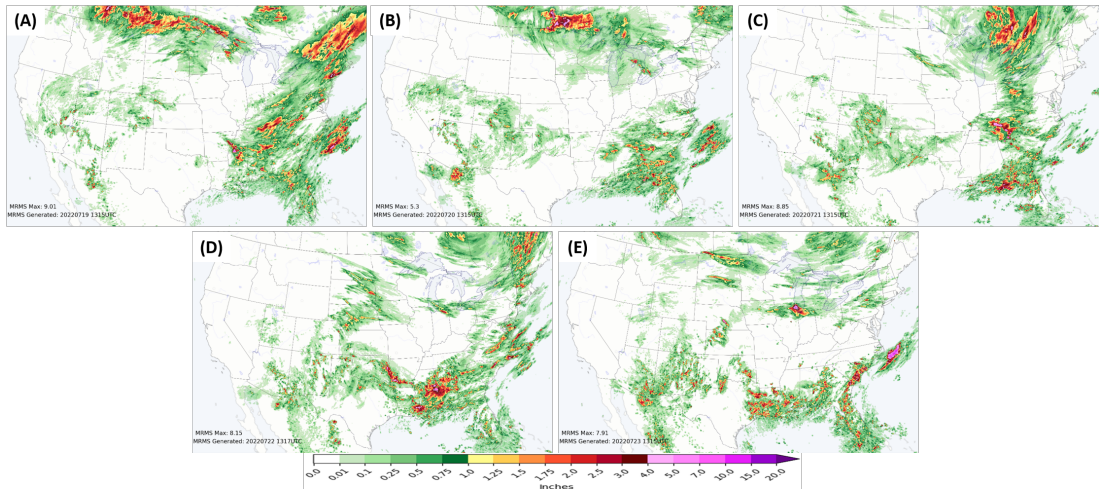


Figure 5: 24-h MRMS QPE valid at 12 UTC on (A) July 19 (B) July 20 (C) July 21 (D) July 22 and (E) July 23, 2022.

Due to the relatively inactive pattern across most of the country, a lot of FFaIR was centered around forecasting for the Southwestern Monsoon and convection across the Southeast. In both instances, the majority of the precipitation across these two regions was disorganized and/or clustered convection. Figure 8 shows the type of six hour precipitation events participants were forecasting for. Events such as these are challenging because either heavy rainfall did not occur or heavy rainfall coverage was low and/or isolated. The former was less of a challenge and more of an inconvenience since FFaIR experiments’ goal is to identify the extreme events. The latter involves the limited predictability of Convective Allowing Models (CAMs) for convective/pulse thunderstorms. Because of the limits in predictability for small scale events, subjectively evaluating the performance and utility of the operational and experimental CAMs was difficult.

2 Science and Operations

FFaIR continues to be a valuable avenue for the evaluation of new products and models. With the quickly approaching scheduled implementation of the Rapid Refresh Forecast System (RRFS), FFaIR helps to get the new model guidance in front of forecasters for realtime forecasting activities. Feedback from participants as they used the various configurations of the RRFS, along with daily verification of

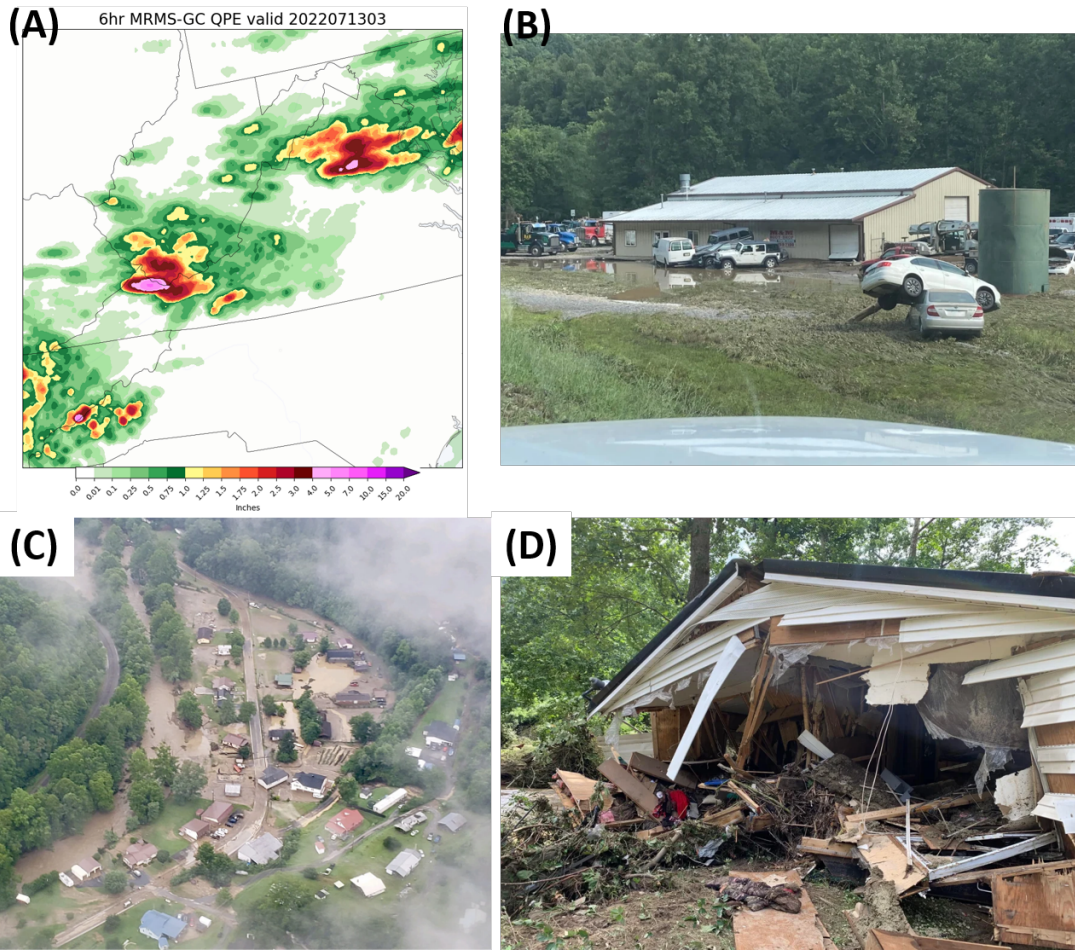


Figure 6: (A) 6-h MRMS QPE valid 21 UTC 12 July to 03 13 July 2022. (B)-(D) Images from the flash flooding that occurred in Buchanan County Virginia. Images taken from Whetstone et al. (2022).

the Quantitative Precipitation Forecast (QPF), helps inform the model developers of possible biases, tendencies, or shortcomings of the model that might not be identified by traditional verification methods. More about the daily activities and the data that were evaluated can be found in the following subsections.

2.1 Daily Operations

This year there were 80 participants across the four weeks of FFaIR. Participants ranged from NWS forecasters to academic and NOAA researchers. Nearly every NWS Region had at least one participant. Weather Forecast Offices (WFOs)

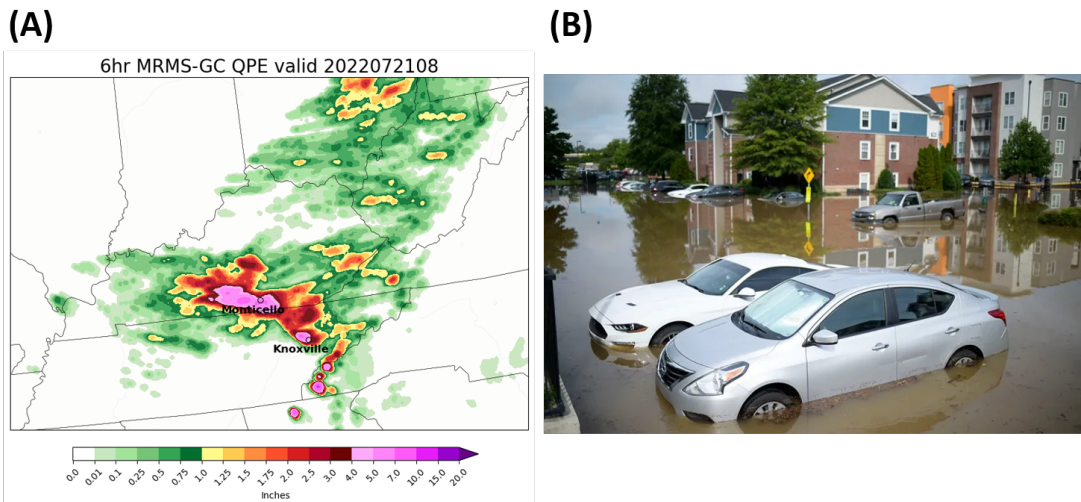


Figure 7: (A) 6-h MRMS QPE valid 02 UTC to 08 21 July 2022. (B) Image from Guerry and Fisher (2022) of flooding in Knoxville, TN. For all images, if there is a grey box present anywhere on the image, it is representing where the maximum is located over the CONUS.

outside the continental United States (CONUS) also participated this year, with participants from Puerto Rico, Guam, and Hawaii. In fact, the Honolulu office had a participant for each week of FFaIR. Additionally, national centers and labs took part, with the Environmental Modeling Center (EMC), the Global Systems Laboratory (GSL), the Physical Science Laboratory (PSL), the National Water Center (NWC), and the Hazardous Weather Testbed (HWT) all in attendance. Groups from the Center for Analysis and Prediction of Storms (CAPS) at the University of Oklahoma (OU) and Colorado State University (CSU) participated, as well as supplied data for evaluation. The full list of participants can be found in Appendix A. The four weeks that FFaIR was in session were:

- Week 1: June 21 - 24**
- Week 2: June 27 - July 1**
- Week 3: July 11 - 15**
- Week 4: July 18 - 22**

The experiment started a half hour earlier this year, at 1330 UTC, to allow more time for completion of all activities. The mornings would start with an open discussion of what happened over the past 24 hrs, followed by a weather briefing

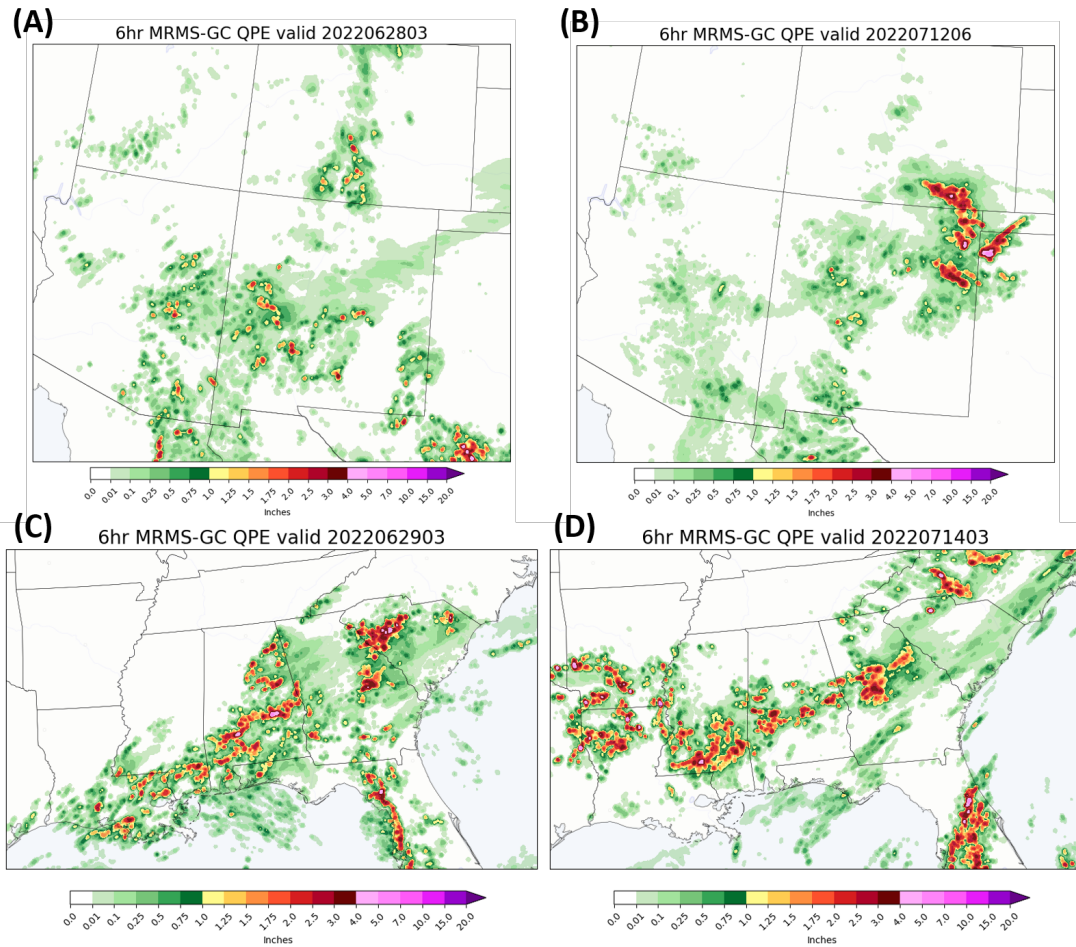


Figure 8: 6-h MRMS QPE valid (A) 21 UTC 27 June to 03 UTC 28 June 2022, (B) 00 UTC to 06 UTC 12 July 2022, (C) 21 UTC 28 June to 03 UTC 29 June 2022, and (D) 21 UTC 13 July to 03 UTC 14 July 2022.

for the day ahead provided by a WPC forecaster. The group then split into two breakout rooms for the first forecasting activity of the day, creating either a Day 1 Excessive Rainfall Outlook (ERO) or an ERO based on Average Recurrence Intervals (referred to as the AERO). Both products were to be “issued” by 16 UTC. After the forecasting activity the verification session started, which was broken up by lunch most days. Once verification was completed, the Maximum Rainfall and Timing Product (MRTP) activity began. A briefing of current observations and forecast was provided and participants began evaluating the near-term heavy rainfall and flash flooding threat, identifying the region/time of the greatest threat

and then forecasting the 6-h rainfall totals for the region. Depending on how active the weather was, participants were also tasked with picking a location to forecast for the Day 2 time frame. On Tuesdays and Thursdays FFaIR also hosted seminars that were open to anyone in the NWS. A list of the seminars can be found in Appendix A.

2.2 Forecasting Activities

As noted above, two of the forecasting activities were the Day 1 ERO and AERO. These were both done in the morning and were valid from 16 UTC to 12 UTC. Each morning, participants would be randomly assigned to either the ERO or AERO group. Participants were encouraged to draw their own outlook and provide rationale for their product. Then the group would discuss the individual products created and what areas of concern they had and work together to create a collaboration Day 1 ERO or AERO.

The FFaIR ERO mimics the operational ERO issued by WPC and is defined as “the probability that rainfall will exceed Flash Flood Guidance (FFG) within 40 kilometers (25 miles) of a point.” It has four risk categories, Marginal (5-15%), Slight (15-40%), Moderate (40-70%), and High (>70%), that can be drawn. An example of the FFaIR ERO compared to the operational ERO can be seen in Fig. 9A-B. The product can be thought of as a way of indicating the coverage of FFG exceedances and flash flood or flood reports. Figure 10 shows the coverage of reports expected for each risk category.

The AERO², on the other hand, attempts to identify the heaviest rainfall that can be expected as it relates to climatology. In other words, the product focuses on intensity rather than coverage of events. Like last year, the exceedance of the six hour ARIs were chosen as the climatological threshold for the product³. However, unlike last year, this year no set probability of exceedance was given. Meaning rather than saying there is a 75% chance of a given ARI being exceeded, participants were told to identify the 6-h ARI most likely to be exceeded, within 25

²The AERO was called the ARI-ERO in the 2021 FFaIR Experiment.

³Refer to Section 2.3 in the 2022 FFaIR Operations Plan (Trojniak and Correia, Jr., 2022) for why the six hour ARI is used.

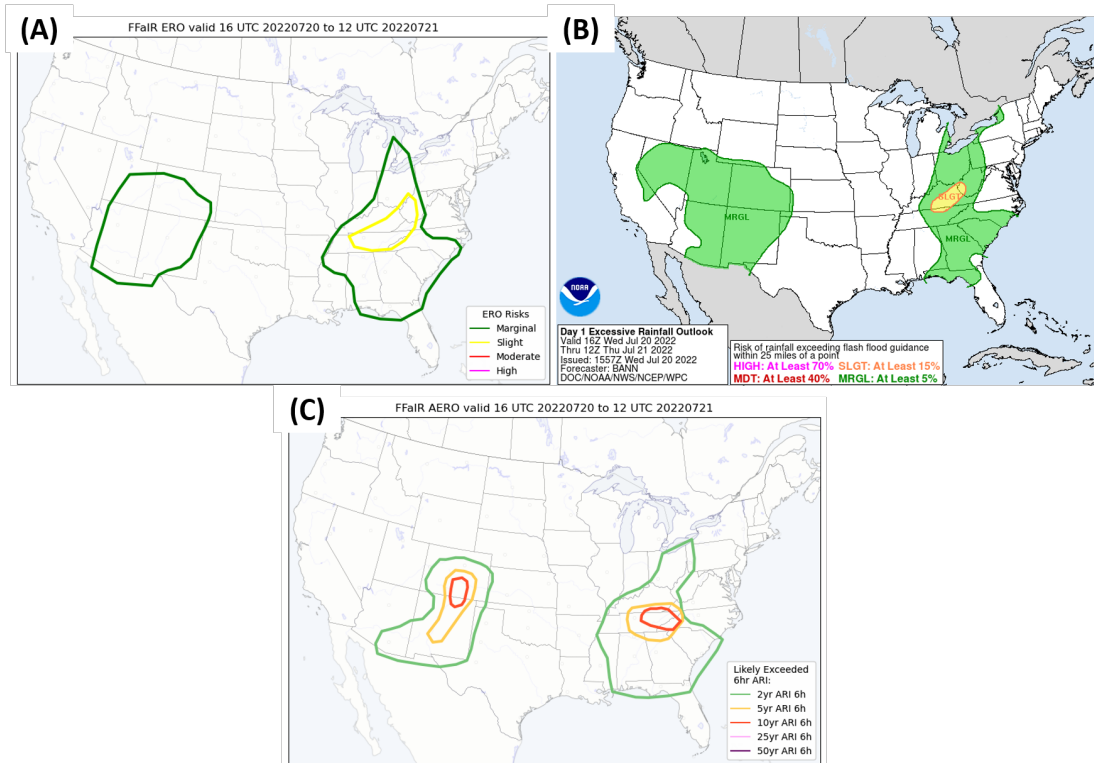


Figure 9: The Day 1 (A) FFaIR ERO, (B) WPC ERO, and (C) FFaIR AERO valid 16 UTC 20 July to 12 UTC 21 July 2022.

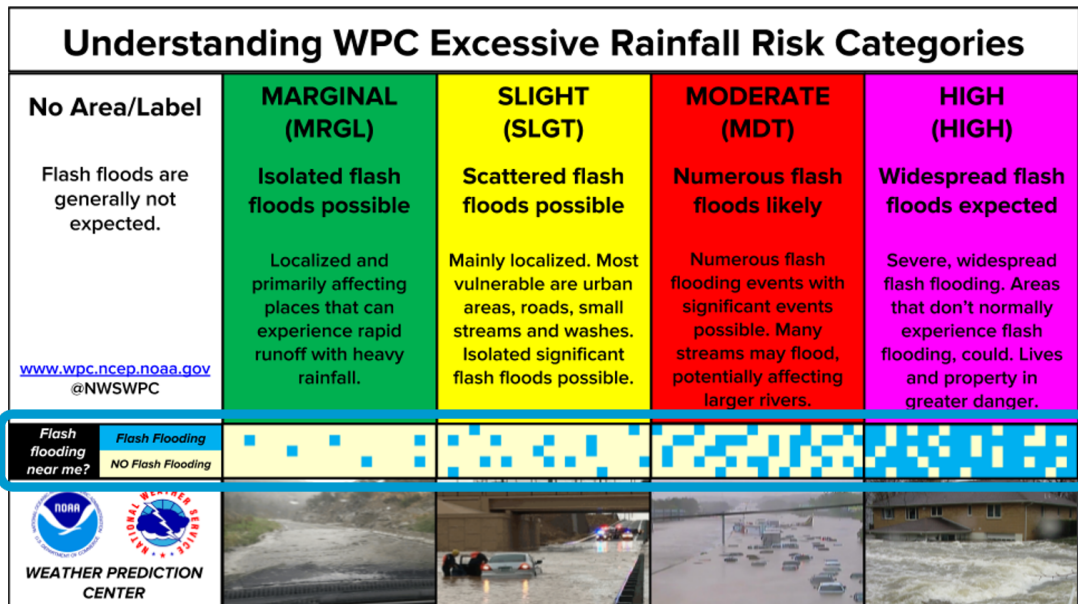


Figure 10: WPC graphic depicting what impacts can be expected for a given ERO category. Circled in blue is the expected coverage of flash flooding with each category.

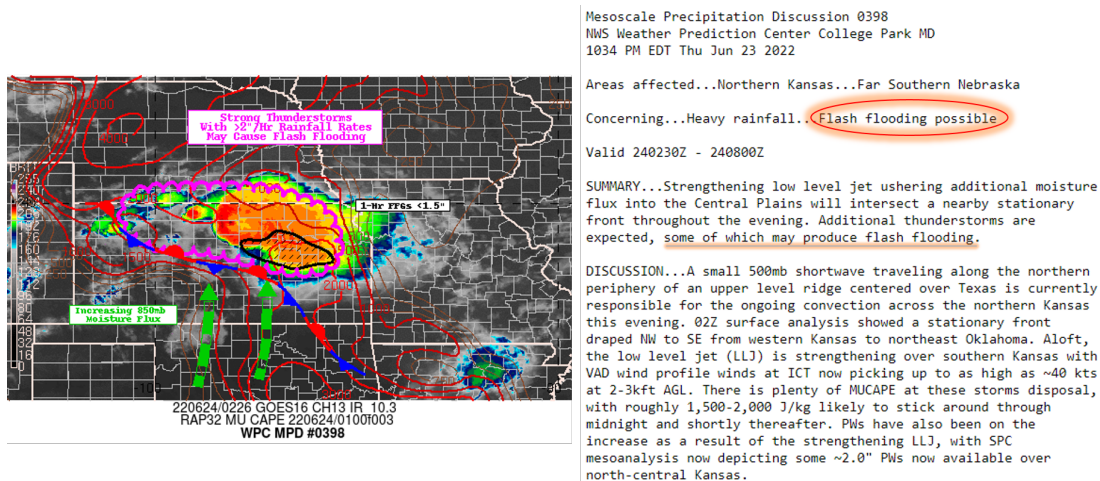


Figure 11: Taken from WPC’s MPD archive at https://www.wpc.ncep.noaa.gov/metwatch/metwatch_mpd_multi.php?md=398&yr=202. Left is the image the forecaster used for the MPD. Right is a screen capture of some of the text of the MPD. Circled in red is the flash flooding possible tag. Underlined shows the verbiage the forecaster used to again convey the threat for flash flooding.

miles of a point, for any six hour time period within the valid time of the product (16 UTC to 12 UTC). Exceedances of the 2, 5, 10, 25, and 50 year ARIs in six hours were used for the product. An example of what the AERO might look like can be seen in Fig. 9C.

The afternoon forecast activity was the Maximum Rainfall and Timing Product (MRTP). This activity was originally introduced in the 2020 FFaIR Experiment and was developed to loosely mimic the Mesoscale Precipitation Discussion (MPD) product issued by WPC. The MPD is a product that identifies where a near-term (0-6-h) risk for heavy rainfall exists. An example of a MPD can be seen in Fig. 11. Like the MPD, the MRTP tries to identify where the greatest risk for heavy rainfall is, in a forecast context, while also providing additional information about the risk. However, unlike the MPD, the MRTP includes drawing isopleths for the 6-h QPF. Participants could draw for 6-h rainfall totals from a half inch to five inches and, new this year, for the highest ARI to be exceeded based on each participants confidence.

The MRTP begins as a collaborative process, with the participants working together to determine where and for what six hour interval the greatest risk for

heavy rainfall is over the CONUS. The earliest the product's start time could be valid for was 21 UTC and the latest the valid end time could be was 12 UTC. A Day 1 MRTP was issued every day of the experiment. On some days, when weather and time allowed, a Day 2 MRTP was also issued. The Day 2 MRTP had the same earliest(21 UTC) and latest (12 UTC) valid time limitations as the Day 1 MRTP but for the following day. If a Day 2 MRTP was issued, the next day, the Day 1 MRTP activity was done for the same region and time period to evaluate how the model and participants' forecasts changed. An example of MRTP verification images for a Day 2 and Day 1 case can be seen in Fig. 12.

In addition to drawing various isopleths, participants were asked to input information about the event such as amount and location of the maximum rainfall, the max 6-h ARI that would be exceeded, the probability of flooding, the probability of the flooding causing damage, and the maximum hourly rainfall total. Additionally, as part of the activity, participants were randomly assigned a model or ensemble. They were not required to use the model or ensemble in their forecast but they were required to evaluate the model and complete a survey about how they felt the model/ensemble was performing in real-time. The survey they were required to complete can be found in Appendix B while a screenshot of the drawing tool used to create the MRTP can be seen in Fig. 13.

2.3 Data and Products Provided

This section is intended to provide a brief overview of the data and products for the experiment. For additional information about the following data/products, please refer to the 2022 FFaIR Operations Plan (Trojaniak and Correia, Jr., 2022). The Operations Plan also includes information about products and tools that were initially planned for evaluation but for one reason or another were not able to be analyzed. Therefore, there will be some discrepancies between this section and the data/tools sections in the Operations Plan.

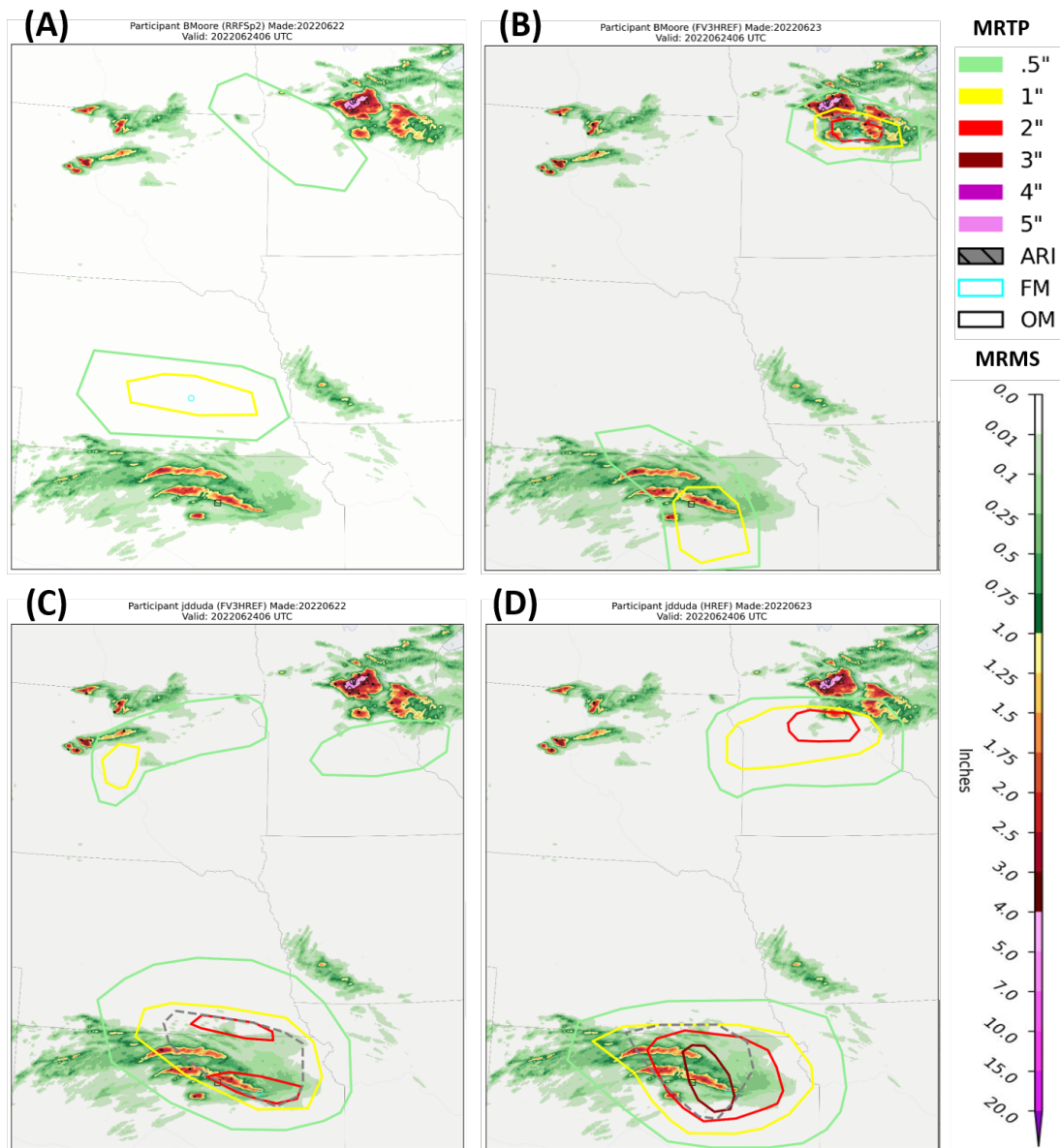


Figure 12: Two randomly chosen participant Day 2 (A and C) and Day 1 (B and D) MRTP forecast verification images all valid 00 UTC to 06 UTC 24 June 2022. For each image, the MRMS QPE is filled while the MRTP forecast is contoured: 0.5" (green), 1" (yellow), 2" (red), 3" (dark red), 4" (purple), and 5" (pink). The dashed gray is the where the participant believes 6-h ARIs will be exceeded. The blue circle is where the forecasted maximum rainfall will be and the black circle is where it was observed.

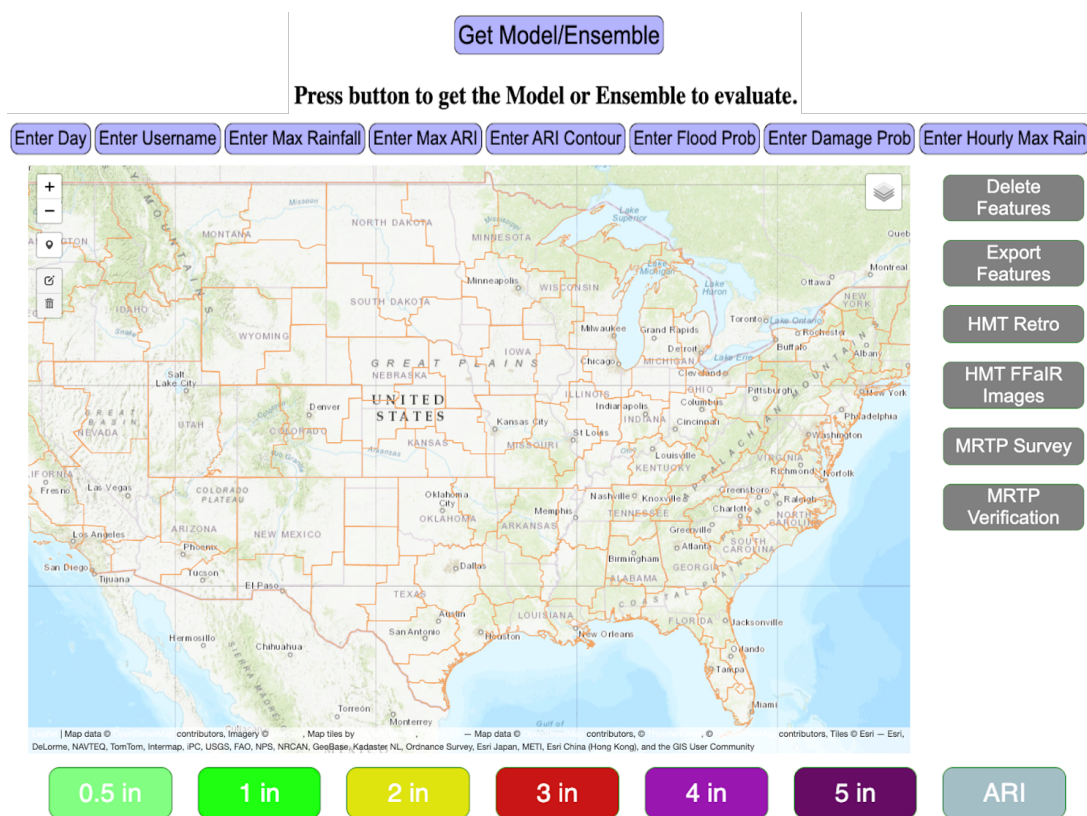


Figure 13: Screen capture of what the MRTP Drawing Tool Website looks like. The “Get Model/Ensemble” button randomly assigns the participant a model or ensemble to evaluate. The line of buttons at the top of the map are what the participants used to input various aspects of the forecast. The bottom row of buttons are the various 6-h thresholds they could forecast for.

2.3.1 Model Information

As noted above, a large portion of the experiment was focused on evaluation of the RRFS. The original plan was to analyze eight different configurations of the deterministic RRFS and two versions of the RRFS ensemble. However, due to numerous unforeseen circumstances, data availability was limited, particularly during the latter half of the experiment. The greatest impact of these outages were for the ensembles, which were not available for over half of FFaIR. Due to missing data, very little meaningful analysis could be done for the ensemble configurations and therefore very little time will be spent discussing the RRFS ensembles.

Table 1: The deterministic model configurations that were evaluated in FFaIR 2022 along with the number of days each model was available for objective verification out of the 19 days of the experiment (red column). *indicates that the model configuration was not constant during the experiment.

Model	Days Available	Cycles/ Fhr	ICs	Microphysics	PBL	LSM	Data Assimilation
RRFSp1	18	00 and 12z Fhr 60	own	Thompson-Eidhammer	MYNN	RUC	GFS cold start
RRFSp2	13	00z Fhr 36	18Z RRFSp1 central state, hourly 3km hybrid 3DnVar	Thompson-Eidhammer	MYNN	RUC	Hybrid 3DnVar
RRFSp3 (MOB0LO_P)	13	00z Fhr 84	RRFSe mean	Thompson	MYNN	NOAH	Inherited from the RRFSe
RRFSp4 (MOB0LO_PG)	19	00z Fhr 84	GFS	Thompson	MYNN	NOAH	GFS cold start
RRFSp5 (M1B0LO_P)	17*	00z Fhr 84	RRFSe mean	NSSL	MYNN	NOAH	Inherited from the RRFSe
RRFSp6 (MOB0L1_P)	17*	00z Fhr 84	RRFSe mean	Thompson	MYNN	NOAH-MP	Inherited from the RRFSe
RRFSp7 (M1B2L2_P)	17*	00z Fhr 84	RRFSe mean	NSSL	TKE-EDMF	RUC	Inherited from the RRFSe
RRFSp8 (MOB2L1_P)	17*	00z Fhr 84	RRFSe mean	Thompson	TKE-EDMF	NOAH-MP	Inherited from the RRFSe

Table 1 lists the RRFSe deterministic (hereafter RRFSe) models that were evaluated during FFaIR. For simplicity, the configurations were numbered 1-8 and were referred to as RRFSe prototypes (RRFSp). In addition to information about each RRFSp’s configuration, Table 1 also includes the number of times the models were available for analysis, either retrospectively or in real-time, during the experiment. RRFSp2 and RRFSp3 were available the least amount of times (13). RRFSp5-8 were available for 17 of the days but for 4 of those days (the last week of FFaIR) they were run with initial conditions from the Global Ensemble Forecast System (GEFS) mean since the RRFSe ensemble provided by GSL (RRFSe) was not available to use. Therefore, results from the RRFSp5-8 should be taken with a grain of salt.

This year, the RRFSe configurations provided by EMC, GSL, and CAPS were designed to be interconnected. RRFSp1 provided the initial conditions for the RRFSp2 at 18 UTC and the baseline for the GEFS perturbations to re-center

around, which were used to start the EnKF “RRFSDAS”⁴ ensemble. Six hours of cycling with the RRFSDAS, which provided flow-dependent information in the EnVar cost function for a hybrid analysis, then followed. The RRFSp2 is the central state of the RRFSDAS and is used to recenter the EnKF ensemble mean each hour, which also serves as the control member for the RRFS ensemble (referred to as RRFSe during the experiment). The first 8 members of the RRFSDAS were perturbed to create the members of the RRFSe, which also included the RRFSp2 as a member. The RRFSe mean was used as the initial conditions for RRFSp3-8, which differ among themselves via their parameterizations (refer to Table 1).

Finally, like the RRFSe, the second ensemble evaluated, referred to as CAPS_RRFSe, used members of the RRFSDAS to initialize its members. RRFSe had 9 members (including RRFSp2) while CAPS_RRFSe had 11 (including RRFSp3) members. An additional difference was that RRFSe had no mixed physics while CAPS_RRFSe had mixed physics. Please refer to Section 2.1.3 in the 2022 FFaIR Operations Plan (Trojaniak and Correia, Jr., 2022) for additional information about the membership.

2.3.2 Machine Learning Products

In addition to the CAPS team providing RRFSp3-8 and CAPS_RRFSe, they also developed a probabilistic rainfall machine learning product (MLP) from four members of the HREF and four members of their ensemble. The product identified the probability of exceeding a half inch of rainfall in six hours and is called the HREF+. However, due to lack of data for evaluation, feedback for this is not included in the report.

Day 1 MLP of EROs were once again provided by CSU, three versions trained on the GEFS, two CAM versions, and a blended version that includes a GEFS version along with the two CAM versions. A list of the CSU ML EROs and how they differ can be found in Table 2. One of the GEFS-based EROs is already running operationally at WPC and will be referred to as the GEFSO ERO. The GEFSO has been evaluated previously in FFaIR and was trained on the previous

⁴Rapid Refresh Forecast System Data Assimilation System

Table 2: CSU MLP of EROs that were evaluated during the 2022 FFaIR experiment. This table is the same as Table 8 in the 2022 FFaIR Operations Plan (Trojniaik and Correia, Jr., 2022).

ERO MLP	Training Model	Forecast Model	Observational training set	Availability
GEFSO	GEFSv11	GEFSv12	Flood/Flash Flood LSRs and regional specific CCPA ARI exceedances	00z
FV3GEFSR	GEFSv12	GEFSv12	Flood/Flash Flood LSRs and regional specific CCPA ARI exceedances	00z and 12z
UFVSGEFSR	GEFSv12	GEFSv12	UFVS	00z and 12z
HRRR	HRRR	HRRR	Flood/Flash Flood LSRs and regional specific CCPA ARI exceedances	00z
NSSL2	NSSL_ARW	NSSL_ARW	Flood/Flash Flood LSRs and regional specific CCPA ARI exceedances	00z
BLEND	n/a	Weighed blend of the FV3GEFS, NSSL2 and HRRR	n/a	00z

version of GEFS (i.e. GEFSv11) while using input from the most recent upgrade of the ensemble (GEFSv12) for its forecast.

Recently, reanalysis data was provided for the GEFSv12⁵. Using the reanalysis data, the CSU created an updated version of the GEFSO that is trained on the GEFSv12 reanalysis. This new GEFS-based ERO MLP will be referred to as FV3GEFSR. A second GEFSv12 trained ML ERO was provided as well. This version, referred to as the UFVSGEFSR, used a different observation set for its training than the GEFSO and FV3GEFS ERO MLPs. The observational dataset used for the UFVSGEFS is called the Unified Flooding Verification System (UFVS; (Erickson et al., 2019)). This dataset is what WPC uses to verify their ERO and has been used by CSU as part of their verification process as well.

Two CAM versions of the CSU MLPs (HRRR and NSSL WRF based) were evaluated in FFaIR. The NSSL version, called NSSL2, was evaluated in both the

⁵This version has the FV3 core.

2020 and 2021 FFaIR experiments and is used as a measure for the performance of CSU CAM based ERO MLPs. The HRRR trained version was updated from last year’s version with training extended to include forecasts from 2021, which had an active monsoon season unlike the 2020 season that the version evaluated last year was trained on. Lastly, the blended version (hereafter the BLEND) combined the HRRR, NSSL2, and FV3GEFSR ERO forecasts. The forecast weight for each of the models in the BLEND was determined from their relative skill over the last 90 days.

2.4 Science Questions and Verification Methods

2.4.1 Experiment Goals

Below is the list of the planned research objectives for the 2022 FFaIR Experiment.

- Evaluate the usefulness of operational and experimental products from high resolution convective-allowing deterministic and ensemble models’ (CAM) QPF. This includes focusing on QPF thresholds exceeding 1 inch, as the previous two FFaIRs have noted that at precipitation thresholds greater than 1 inch, the wet bias from FV3-CAMs increases quickly as the threshold value increases.
- Assess the impact of FV3-CAMs configuration changes that may reduce the prolific precipitation and/or precipitation rates related to grid point storms (also referred to as popcorn storms) that were identified in the 2020 and 2021 FFaIR Experiment. This will include assessing the newly output hourly maximum precipitation rates out of the FV3-CAMs; only the instantaneous precipitation was available last year.
- Evaluate the impact of cycled data assimilation on the first six hours of the different RRFS members.
- Use the MRTP to identify timing errors during MCS events using the 6 hour precipitation verification. In past FFaIRs, it has been noted that models might have the correct idea of how an event might evolve but do not have

the timing correct (ex. initiation, change in progression direction and speed, etc.).

- Analyze the utility of various RRFS ensemble configurations, from multi-physics to stochastic parameter perturbations (SPP), and compare their performances to the HREF. In addition to evaluating “classic” ensemble probabilities, FFaIR will be evaluating a machine learning product for the probability of exceedance from the CAPS group.
- Evaluate the CSU MLP for the Day 1 ERO. This year, three new versions of their operational GEFS-based ERO will be analyzed, as well as the updated version of the HRRR-based ERO from the 2021 FFaIR experiment.
- Evaluate the utility of creating an ERO that is centered around the exceedance of 6-h ARIs. These forecasts will attempt to predict locally heavy rain that the traditional ERO may not consider due to small spatial or temporal scales. It will be referred to as the AERO and created by the participants.
- In the 2020 FFaIR experiment, a product tracking heavy precipitation objects (then called HPOT) was evaluated and had positive feedback from the participants. A new website is in development for the product, now called the Tracking of Heavy Precipitation Objects (THEPrO), and feedback will be collected on both the product and the website.

The previously noted data issues impacted the ability to address some of the experimental goals. Some of the questions were abandoned while others, for instance trying to assess the impact of cycled DA for the RRFS configurations, did not have enough days for an evaluation to provide a useful analysis. In other instances, the lack of organized rainfall (ex. MCSs) made it difficult to identify possible shortcomings in models. For example, questions about model timing were often skipped since the focus of these questions were to identify timing biases in the evolution of organized convection. The lack of organized convection also prevented the analysis of the websites hosting object tracking products. Therefore, the results from this year’s FFaIR experiment will focus on evaluation of the deterministic RRFS configurations, CSU’s ML EROs, and analysis of the AERO and MRTP.

2.4.2 Verification

A large majority of FFaIR centers around participants subjectively evaluating the guidance and providing both verbal and written responses. This happens in real time (i.e. while the participants are using the data to create forecasts) and during the verification session. The verification session usually involves analyzing the performance of the previous day’s guidance. However, on Monday’s the previous Friday’s forecast is evaluated. Evaluation is done using verification websites designed by the FFaIR team and via Google Forms. Generally, the subjective verification requires the participants to compare the forecast/guidance to observations and then put a value to its "goodness". For evaluation of the QPF and the EROs/AERO, participants were asked to rate the guidance on a scale of 1 (poor) to 10 (great) individually. In addition to a ranking, all questions in the survey included at least one voluntary question that required a written response to provide feedback on the guidance. Often these questions just asked for the participants’ overall thoughts on the performance of the product/model being evaluated.

The model QPF evaluation was on the 24-h accumulation, valid at 12z for both 00z and 12z cycles. This was compared against 24-h MRMS QPE that was remapped to the HRRR grid, using the cKDTree package in Python and retain the maximum value of the 9 grid point neighborhood. The MRMS/model comparison was shown both as a side by side comparison and via 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 a half inch and two inches can be seen in Fig. 14. 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.

For evaluation of the EROs, each ERO (including the FFaIR ERO) was shown along with the MRMS and the practically perfect; Fig. 15 provides an example of what the verification looked like. The practically perfect method used is explained in great detail in Erickson et al. (2019) and can be thought of as the ERO that

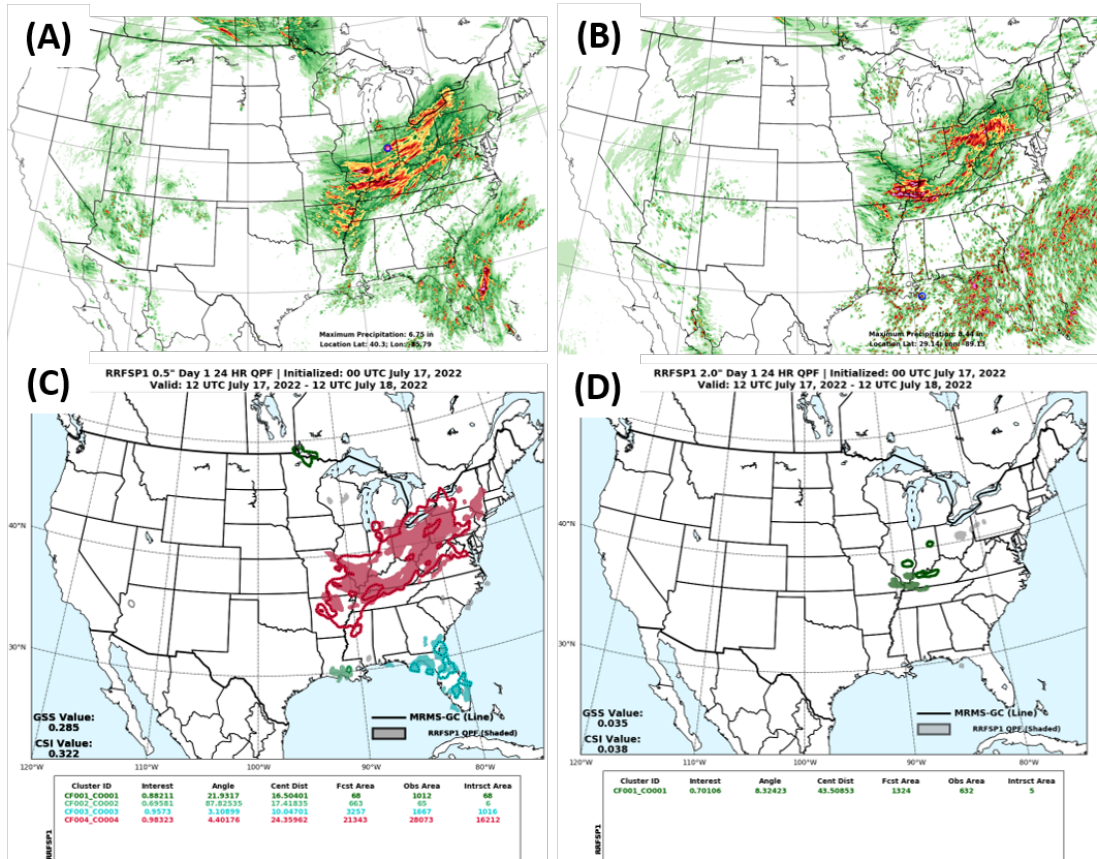


Figure 14: 24-h (A) MRMS QPE and (B) RRFSp1 QPF valid at 12 UTC 18 July 2022. MODE verification for (C) 0.5 in and (D) 2 in where the contour is the MRMS QPE and the fill is the model QPF. Matching MRMS and model clusters have the same color. Included in the MODE plot is statistical information for the forecast threshold as a whole and for the individual clusters identified by MODE.

would have been drawn if we knew what the rainfall impacts would be as they relate to flooding ahead of time. It has the same risk definitions as the ERO: Marginal (5-15%), Slight (15-40%), Moderate (40-70%) and High (>70%). The FV3GEFSR and the UFVSGEFSR MLP EROs were run at both 00z and 12z, all other CSU MLPs were only run for the 00z of their respective parent models. For verification of the CSU ML EROs with two forecast initialization times, the 00z and 12z forecasts were set side-by-side. Examples of this can be seen in Fig. 16. In addition to the MRMS and practically perfect, the observations that make up the UFVS were plotted on the same image as the forecast ERO. Furthermore, in the case of the CSU MLP EROs, an additional contour (2.5%) was shown outside

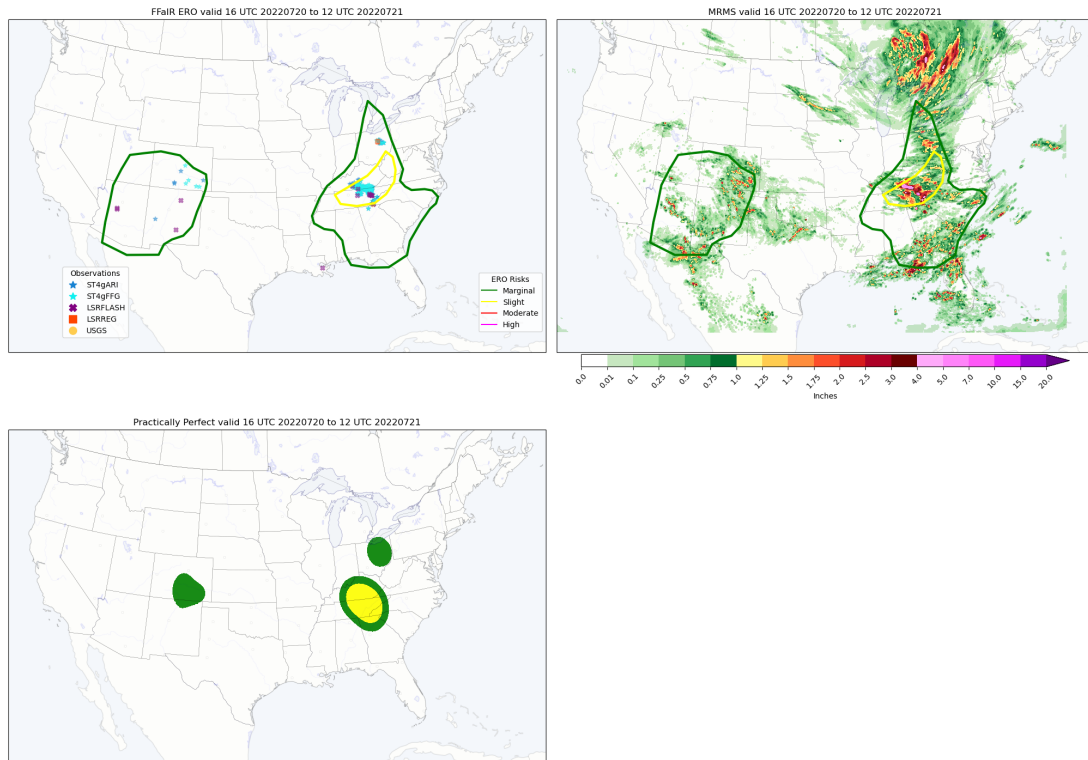


Figure 15: FFaIR ERO verification image valid 16 UTC 20 July to 12 UTC 21 July 2022. Top left: FFaIR ERO with UFVS overlaid. Risk categories - Marginal: 5%-15% (green), Slight: 15%-40% (yellow), Moderate: 40%-70% (red) and High: >70% (purple/pink). Top right is the 20hr QPE and bottom left is the practically perfect verification.

of the values that define the WPC ERO risk categories. The contour was added after the first week of FFaIR at the suggestion of a participant.

Verification of the AERO involved had two ranking questions associated with it. The first followed the “goodness” 1-10 ranking question aforementioned. The second question asked participants “Focusing specifically on the LSRs, how well do you feel ARI exceedances matched up with reports of heavy rainfall?” For this question they were given a ranking from 1 (poor) to 5 (great). The written response question was “Please comment on your thoughts on the AERO. Include things like if you thought it is useful for identifying heavy rainfall, if you thought the thresholds are good, etc.” Since the AERO is a product in development, the FFaIR team wanted as much feedback about the utility of the product, and about

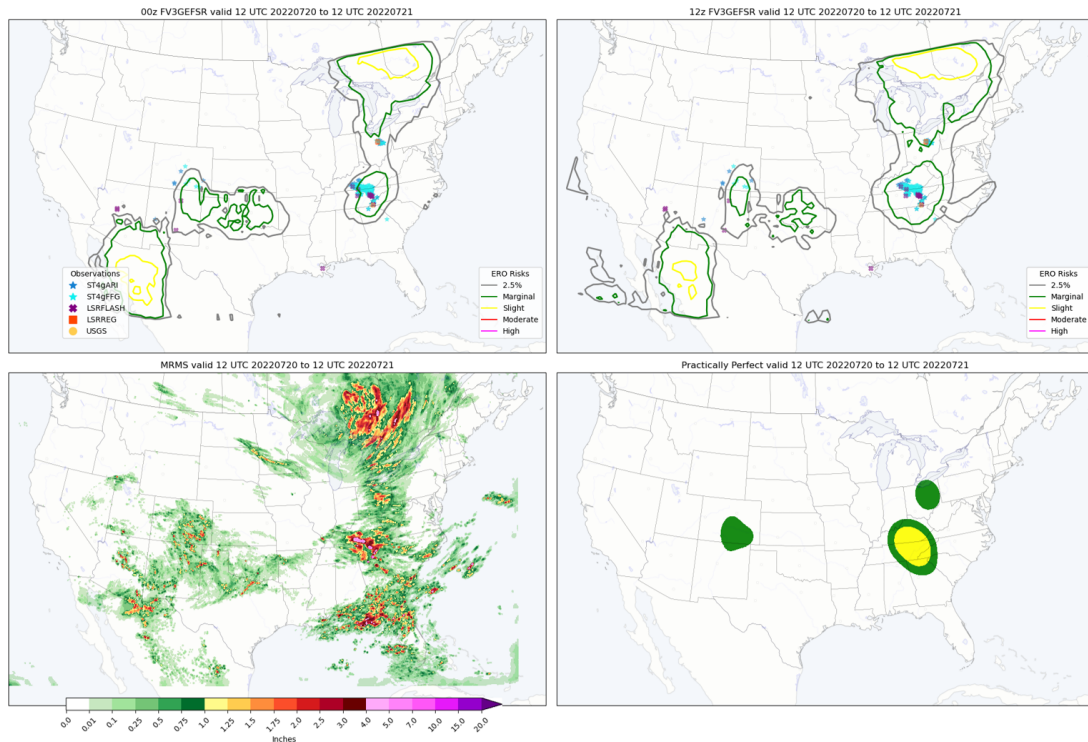


Figure 16: Along the top: the 00z and 12z FV3GEFSR ERO respectively, with the WPC ERO risk probabilities contoured [Marginal: 5% (green), Slight: 15% (yellow), Moderate: 40% (red) and High: 70% (purple/pink)] and the 2.5% probability contoured in gray. Additionally the UFVS data points are overlaid on the ERO images. Bottom left is the 24-h QPE and bottom right is the practically perfect verification. All valid 12 UTC 20 July to 12 UTC 21 July 2022.

the thresholds/probabilities chosen, as possible. Figure 17 shows the setup for the AERO verification. Since no practically perfect had been developed for the AERO yet, participants used the rainfall footprint and ARI exceedance locations for verification. Also provided for reference were the 2-y and 25-y 6-h ARIs for the CONUS.

The MRTPs were objectively and subjectively evaluated. When possible, participants evaluated their own MRTP. On Mondays, when a new set of participants started, MRTPs from the previous week were randomly assigned for evaluation. To help with the analysis of the MRTP, verification graphics included the MRMS QPE overlaid with the MRTP QPF contours, various contingency table statistics, information about the observed values compared to the participant's forecast

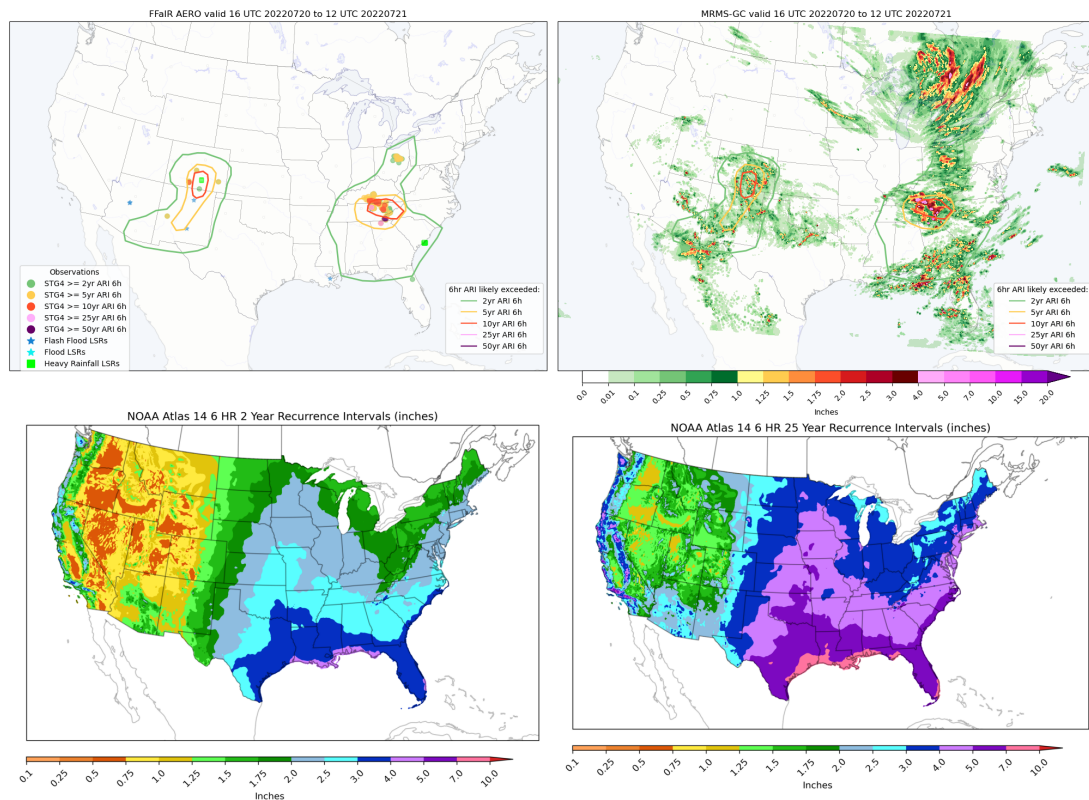


Figure 17: FfaIR AERO verification image valid 16 UTC 20 July to 12 UTC 21 July 2022. Top left: FfaIR AERO with STG4 exceedance of the AERO ARI thresholds, Flash Flood and Flood LSRs, and Heavy Rainfall LSRs. AERO contours and observations colors for the 6-h ARI are: 2-y - green, 5-y - yellow, 10-y - red, 25-y - pink, and 50-y - dark purple. Top right is the 20hr MRMS QPE with the AERO overlaid. Along the bottom are the 2-y and 25-y 6-h ARIs respectively.

values, and information about the CSI of the model they were assigned to evaluate. A similar graphic was created for all models and cycles that had a valid forecast for the MRTP time. Figs. 18 and 19 show an example of the graphics with labels to help decipher the images. The forecast was valid 02 UTC to 08 UTC 21 July 2022, the participant's username was MIrocks and they were assigned the HRRR for evaluation. For comparison, the HRRR 20220720 12z run is shown in Fig. 19 since it is likely the cycle of the model that was evaluated by MIrocks during the MRTP activity the previous day. Comparing the values outlined by the green box in each image, it can be seen that for POD, FAR, and CSI at the 1 in threshold, MIrocks outperformed this cycle of the HRRR. However, comparing the values in

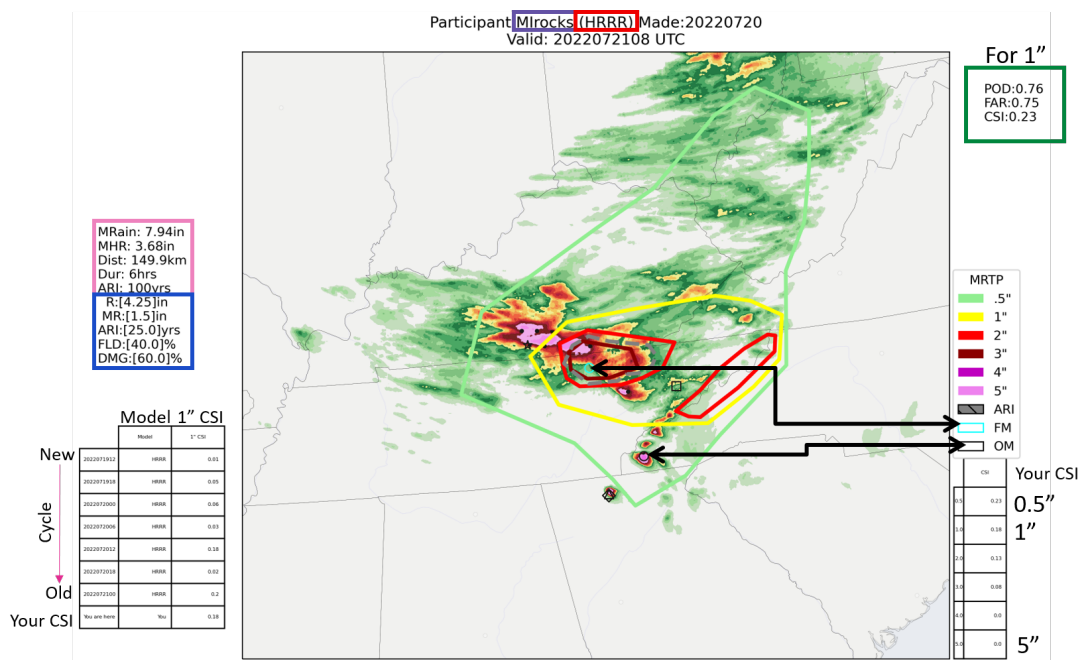


Figure 18: MIrocks's Day 1 MRTP drawn on July 20, valid 06-08 UTC 21 July 2022. The MRTP thresholds are contoured as: 0.5" green, 1" yellow, 2" red, 3" dark red, 4" light purple, and 5" dark purple. The 6-h MRMS QPE is filled. The forecasted (cyan) and observed (black) rainfall maximum are plotted as circles. For the purpose of this figure, black arrows were added to point to the locations of the aforementioned maximums. Within the pink box is information about observations. In the blue box is information about forecasted values. In the grid below the blue box is the CSI for 1" for all model cycles (name boxed in red) that had a forecast valid for the MRTP time period. In the grid on the right side of image is the participant's (username boxed in purple) CSI for each MRTP contour. The green square shows their statistics for 1".

the pink (observed values) and blue boxes (forecasted values), MIrocks forecast the maximum rainfall to be 4.25 in but the max in the domain was 7.94 in. The forecast location of the max rainfall was 149.9km away from the observed max; compare the location of the two black arrows on the map.

Using all this information, the participants evaluated the one inch contour they drew against the model they were assigned and ranked these from very poor to very good. They assessed the accuracy of the area of the contour and its orientation. If a Day 1 and Day 2 forecast were made, they answered the same questions for both forecasts, as well as how they felt their Day 2 MRTP did compared to their Day 1. Lastly, from past experience, the FFaIR team has found that partic-

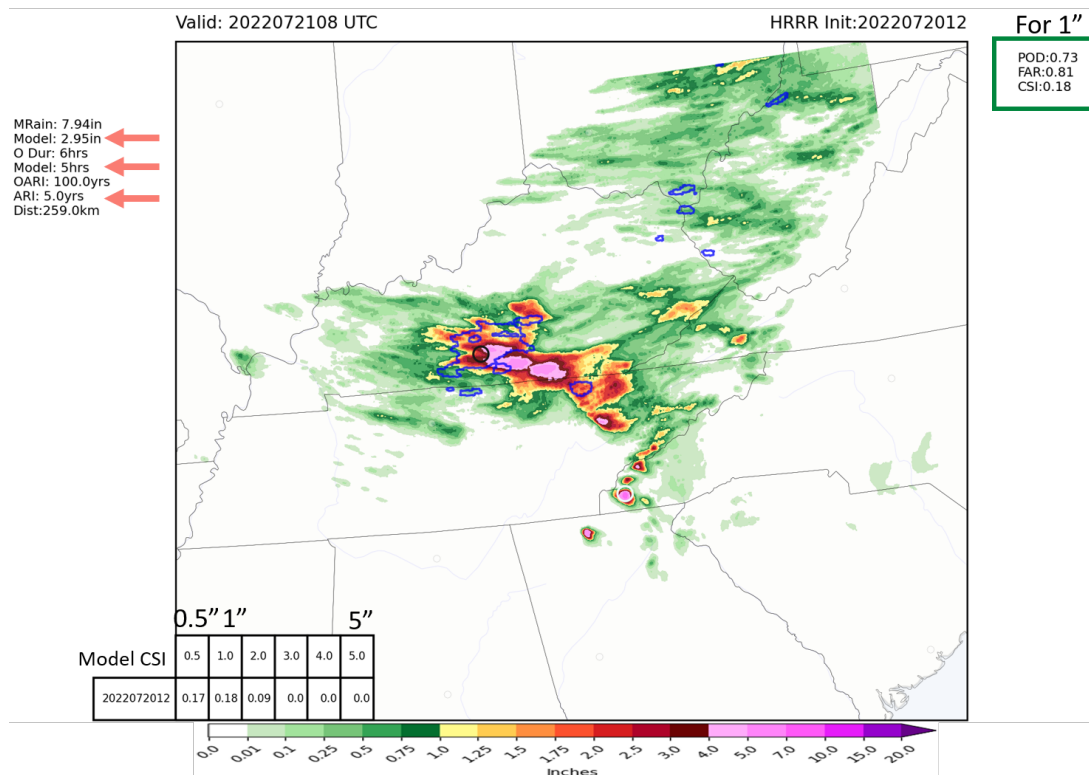


Figure 19: The HRRR's 20220720 12z cycle valid 06-08 UTC 21 July 2022 with where the model forecasted 1" or greater rainfall to occur within the valid 6-h time frame contoured in blue. The 6-h MRMS QPE is filled. Orange arrows identify what the forecasted maximum rainfall, ARI, and duration were with the observed plotted above the forecasted values. In the grid at bottom of graphic is the HRRR's CSI at each MRTP threshold. The green square shows its statistics for 1".

Participants enjoy looking at the MRTP verification. Therefore, extra time was spent allowing participants to talk about their forecast with the group. Additionally, composites (a human ensemble) of every MRTP done for that day were made for each threshold participants were able to draw for; see Fig. 20. This allowed for discussion about consensus among the group.

Finally, there was an end of the week survey that went out to all the participants. This was not required to be completed but was highly encouraged. Of the 80 participants, roughly half completed the end of the week survey. Both surveys can be found in Appendix C.

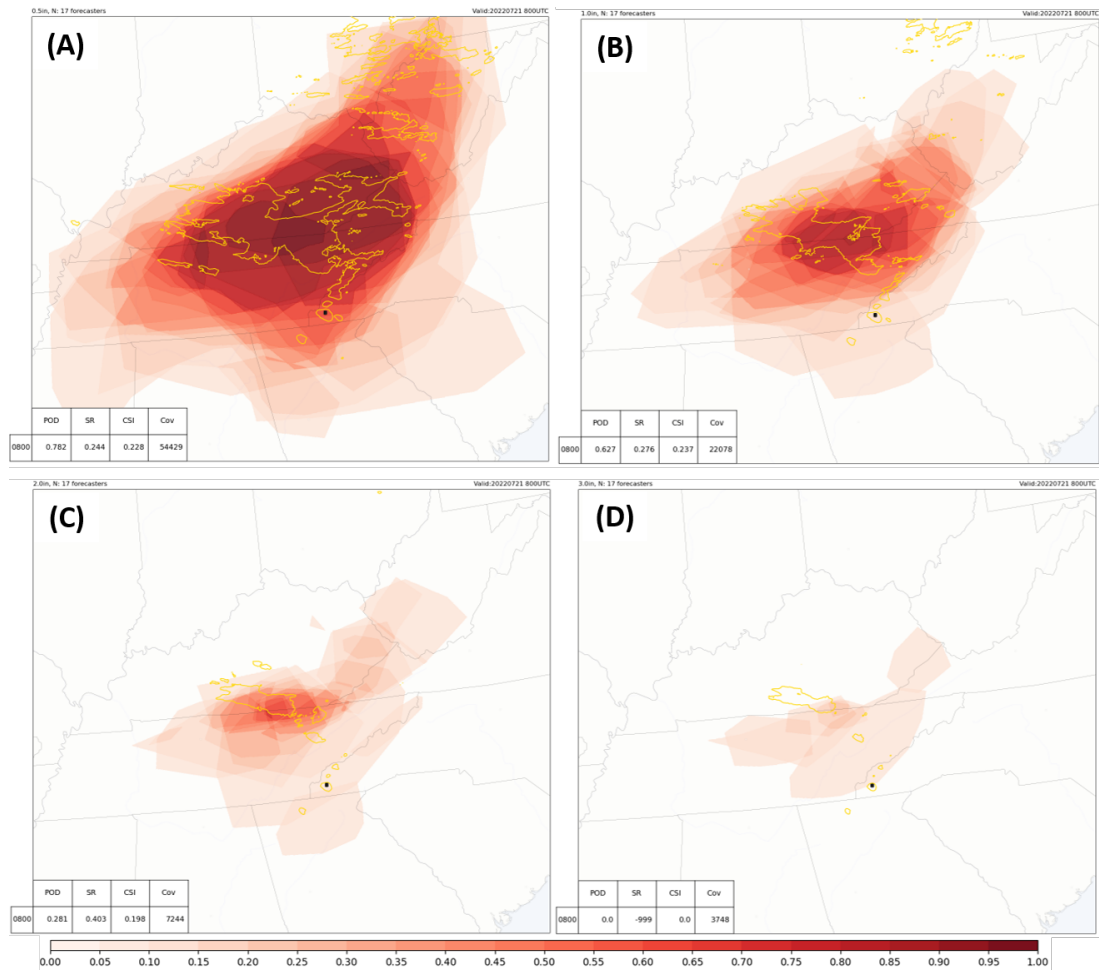


Figure 20: Composite of all the MRTPs drawn on July 20, valid 06-08 UTC 21 July 2022 for: (A) 0.5", (B) 1", (C) 2" and (D) 3". Contoured in yellow is each images' respective MRMS QPE. The grids at the bottom of each image have the POC, SR, CSI, and converge information for their receptive threshold.

3 Results

The following sections will highlight both the subjective and objective (a.k.a. qualitative and quantitative) findings from the 2022 FFaIR Experiment. These results will help determine the readiness of the evaluated model data and products to be transitioned to operations. The RRFS⁶ configurations listed in Table 1 were compared against two operational CAMs, the HRRR and the NAMnest. These two

⁶Reminder, RRFS in this context refers to the RRFS deterministic CAMs.

operational models were chosen since the Unified Forecast System (UFS) science evaluation team identified them as the systems to measure RRFS performance against (Kinter et al., 2020). This section will first focus on QPF verification from deterministic models followed by a brief note on the ensembles, before discussing precipitation rates. Then findings related to the EROs and AERO will be summarized and finally the MRTP activity.

3.1 QPF

As noted in Section 2.4.2, participants were asked to subjectively score 24-h QPF on a scale of 1 (poor) to 10 (great). This was done for both the 00z and 12z forecasts but only the results from the 00z forecast will be discussed since 12z forecasts were only available for the HRRR, NAMnest, and RRFSp1. Figures 21 and 22 summarize the results for the subject scores. The inconsistent flow of data resulted in overall lower sample sizes for the experimental versus operational models, with the operational models receiving just over 300 scores compared to the RRFSp3 which had the fewest scores, 109. Furthermore, there were only **FOUR** days where all the models were available. Because of this, any general comparisons are difficult to make. The four 00z cycles that all models had available data were: June 30, July 1, July 12, and July 13 2022.

As a result of the differing sample sizes, models were grouped based on the number of scores received. The HRRR, NAMnest and RRFSp1 (hereafter Group 1) had a similar number of scores, while the RRFSp2 and RRFSp4 (hereafter Group 2) were similar with 213 and 187 times scored respectively. RRFSp3 and RRFSp5-8 (hereafter Group 3) all had 137 or less scores recorded. To help quickly identify model results based on groups, the columns in charts shown in Figs. 21 and 22 have been outlined for Groups 1 and 2. The results from Group 3 will only briefly be touched on because they were only evaluated for about half the days of FFaIR, and the configuration for RRFSp5-8 was changed during the last week of FFaIR.

During the course of the experiment, only the HRRR and RRFSp4 received a score of 10 (each once) while the HRRR, RRFSp6, and RRFSp8 were the only ones

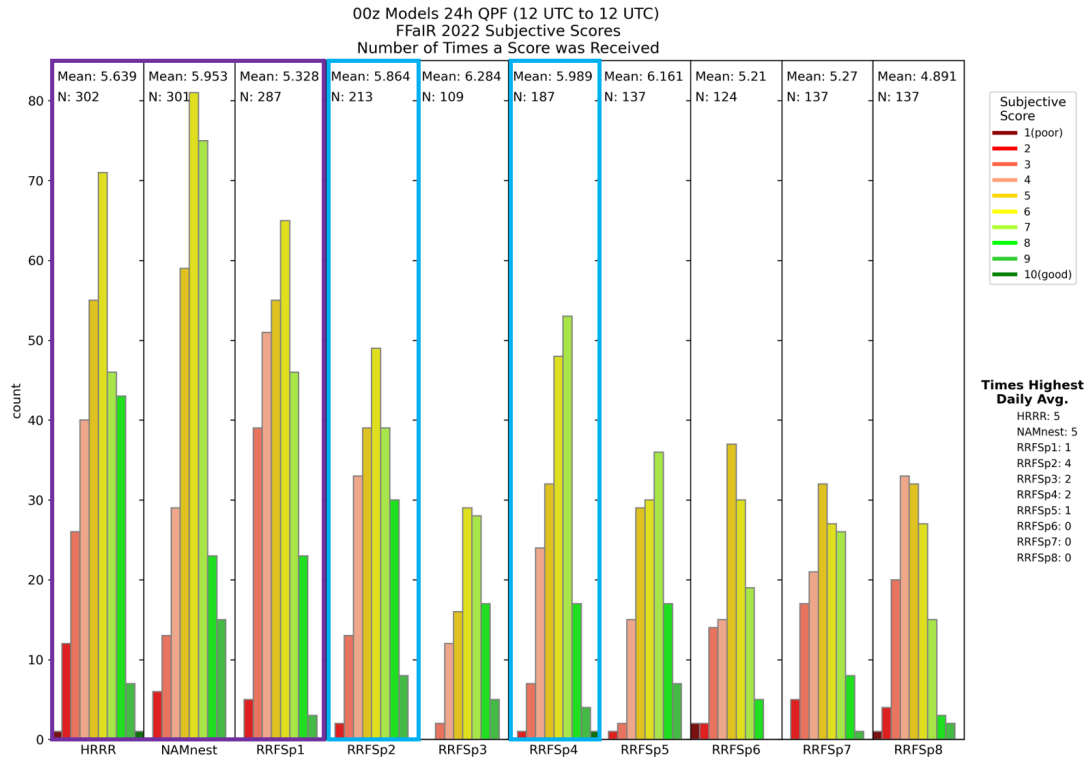


Figure 21: Results from the subjective verification for 24 h QPF for 00z model initialization showing number of times each model received a score from 1 (dark red) to 10 (dark green) during the duration of FFaIR. The number of scores received (N) and the mean score for each model is plotted along the top. Outlined in dark purple are the models in Group 1, outlined in blue is Group 2, and the ones not outlined are in Group 3. On the right, below the legend, is the number of times each model had the highest daily average score.

to receive a score of 1. Among all the models evaluated, RRFSp3 had the highest average score, 6.284, but it was only evaluated 109 times. The lowest mean, evaluated 137 times, was the RRFSp8 (4.891). In Group 1, the NAMnest had the highest mean score, 5.953, followed by the HRRR (5.639), with the RRFSp1 (5.328) having the lowest score. In Group 2 the RRFSp4 (5.989) had a higher mean than the RRFSp2 (5.864). When looking at daily average scores, the HRRR and NAMnest both had 5 days in which they had the highest average score, while RRFSp4 had the highest average 4 of the days. When solely looking at the 4 days that had all of the models, the NAMnest had the highest average score (6.76) followed by RRFSp5 (6.52), RRFSp3 (6.33), RRFSp4 (6.21), and the HRRR (6.05).

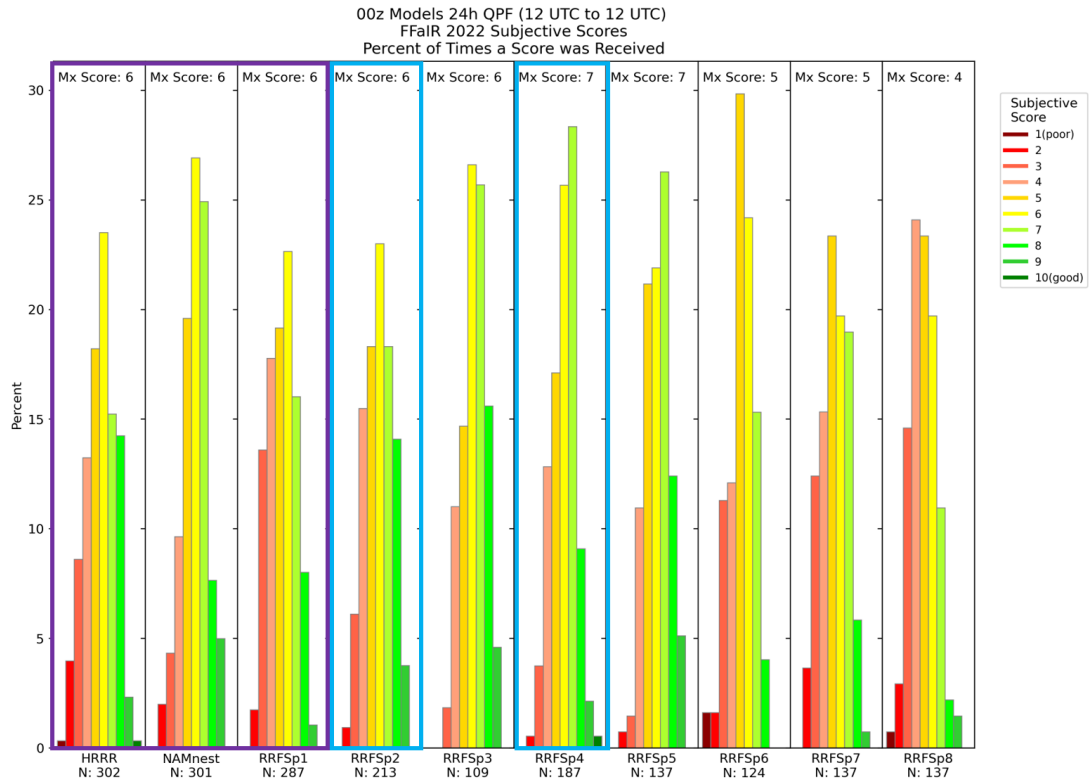


Figure 22: Similar to Fig. 21 but for the percent of times each model received a score. N is now along the bottom and along the top is the score each model received the highest percent of the time.

The rest averaged below a score of 6, with average scores of 5.76 (RRFSp7), 5.64 (RRFSp2), 5.6 (RRFSp6), 5.24 (RRFSp1), and 5.213 (RRFSp8).

An example of the model 24-h QPF compared to MRMS for one of the days that all the data was available for evaluation by the participants can be seen in Fig. 23 along with their daily average scores in the caption. This was one of the more challenging forecasts during FFaIR, with a tropical low off the coast of LA, a low off the Carolina's, and an MCS that moved southward from extreme southeastern NE into northeastern KS. On this day, average scores ranged from 7.143 (RRFSp5) to 4.238 (RRFSp1). The distribution of the scores can be seen in Fig. 24. Overall, participants noted that nearly all the models over forecast rainfall totals in LA, particularly because models brought the system too far inland. Participants were impressed that most of the RRFS configurations developed the

MCS, while the HRRR did not. However, they also noted the over forecast of totals across the southeast. For example two participants wrote:

“A few of the models (NAMNest, p3, and p5) captured the MCS in Kansas quite well. RRFS had a lot of 3-6” totals for pulse storms in the SE CONUS. This overestimation was evident last year as well.”

“Most of the RRFS models showed too high in rainfall amounts and too much coverage with the “typical” afternoon convection across the southeast CONUS. They did fairly well with the other features, such as the front from NE to northern VA, and with the locations of the tropical disturbance near the TX/LA coast and Carolina coast. ”

The distribution of scores in Figs. 21 and 22 show that although the average scores for the models in Group 1 were similar (separated by 0.625), the skewness is more similar between the HRRR and the RRFSp1 than the NAMnest is to either. The NAMnest’s scores are skewed right (higher) while the other two are skewed left. Thus the RRFSp1 was more similar to the HRRR, which was overall perceived as the poorer performing operational model of the two. Common comments about the HRRR’s performance noted by participants were:

“Overall, they definitely outperformed the HRRR. The HRRR really seemed to struggle this week and pretty much everything else did better than it.”

“The HRRR really struggled last week, so it wasn’t hard for the RRFSs to beat it.”

“The HRRR seemingly struggled quite a bit compared to the RRFSs and the NAMnest.”

“The HRRR I think was consistently poor. The RRFS was often down but the couple times I looked at it it seemed similar or just slightly better/worse each time.”

The perception of the participants of HRRR’s performance, specifically its dry bias, during FFaIR is supported by the objective verification and will be discussed further below. Due to the poor performance of the HRRR during FFaIR, along

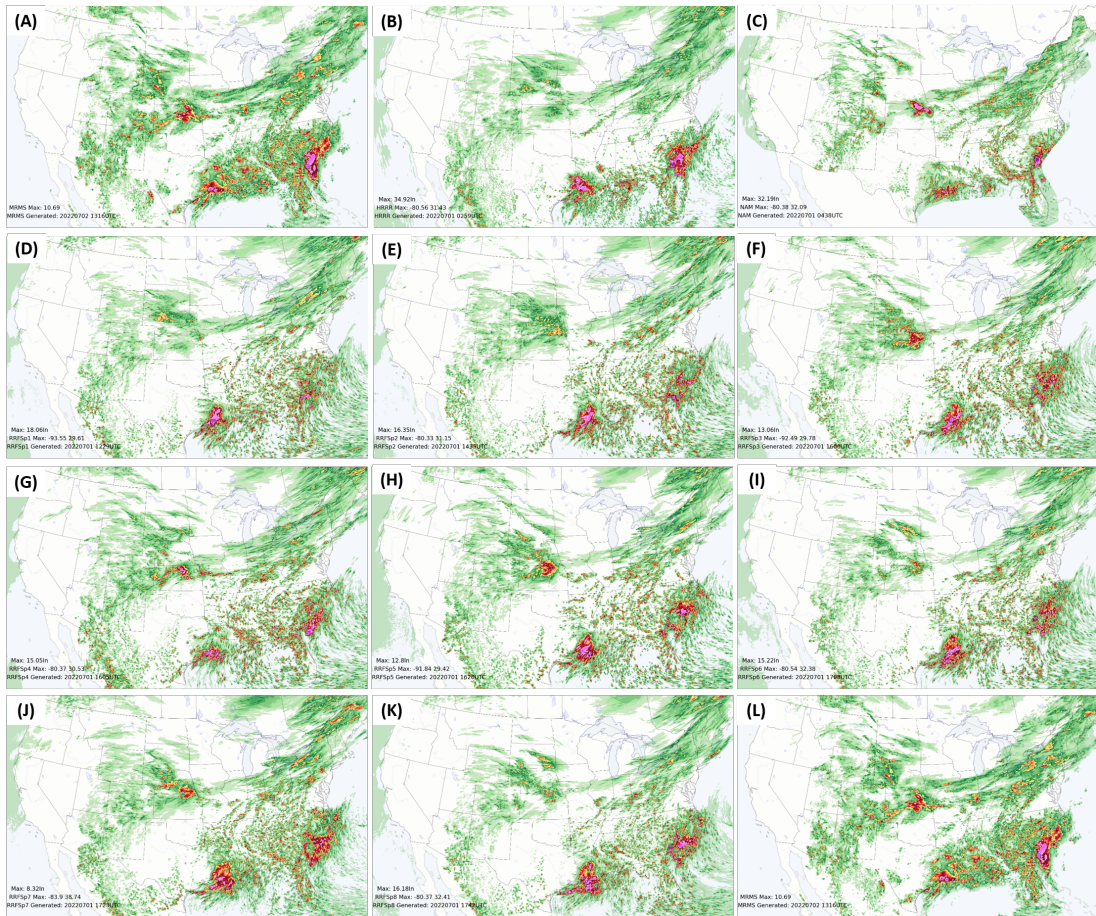


Figure 23: 24-h rainfall forecast for all of the models evaluated during FFaIR compared to MRMS valid 12 UTC 01 July to 12 UTC 02 July 2022. The daily average score received from the subjective verification will follow the model name in []. (A) and (L) MRMS QPE. (B) HRRR [5.238], (C) NAMnest [6.81], (D) RRFSp1 [4.238], (E) RRFSp2 [4.722], (F) RRFSp3 [7.19], (G) RRFSp4 [6.762], (H) RRFSp5 [7.143], (I) RRFSp6 [5.286], (J) RRFSp7 [5.905], and (K) RRFSp8 [4.571] QPF.

with the higher performance of the NAMnest, using the HRRR as a baseline for the performance of the RRFSp models during FFaIR is not advised.

Between Groups 1 and 2, the RRFSp4 was the model most likely to receive a 7, while the other models were most likely to receive a 6. It also had lowest percent of scores between 1 and 3. The subjectively good performance of the RRFSp4, especially compared to RRFSp2, is interesting since RRFSp2 had the RRFSDAS⁷

⁷The RRFSDAS is similar to the HRRRDAS but is in early development so not all the same data assimilation information is included in the RRFSDAS as the HRRRDAS.

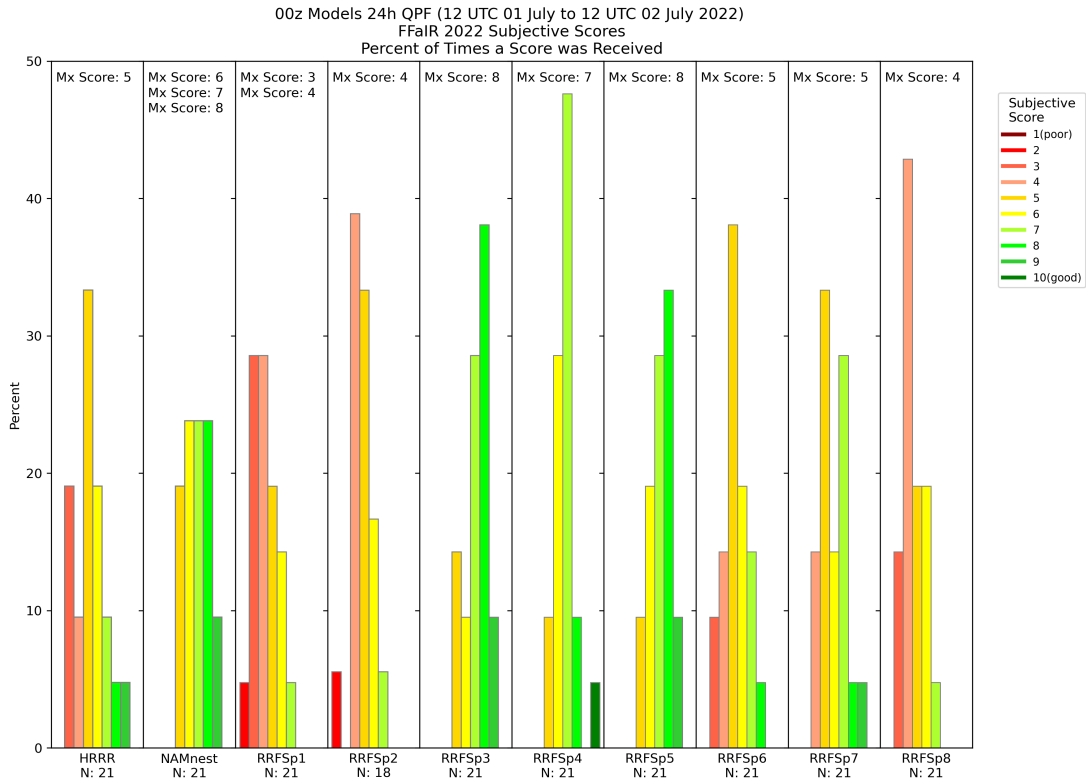


Figure 24: Similar to Fig. 22 but for a single day. Scores are for verification of the 00 UTC cycle 20220701F36.

and RRFSp4 is a cold start from the GFS initial conditions. The RRFSp4 also performed well objectively. In fact, when looking at performance diagrams in Fig. 25A and B, the RRFSp4 has a higher CSI and comparable frequency bias to RRFSp1 and RRFSp2. That said, at the higher thresholds like 2 and 3 inches (Fig. 25C and D), its performance declined in comparison to RRFSp1 and RRFSp2.

Figures 25 and 26 provide a summary of model and ensemble performance during the forecast days of the experiment and across what will be called the Testbed Season (May 12 to July 31, 2022). RRFSp3 and RRFSp5-8 were not included in this analysis since their availability was lacking during the experiment. Additionally, RRFSp4 is not included in the Testbed Season analysis as it was only run during FFAIR. For both time periods, the RRFSp1 and RRFSp2 generally performed more similar to the NAMnest than the HRRR. As already noted, during the experiment the participants felt that the HRRR often under performed,

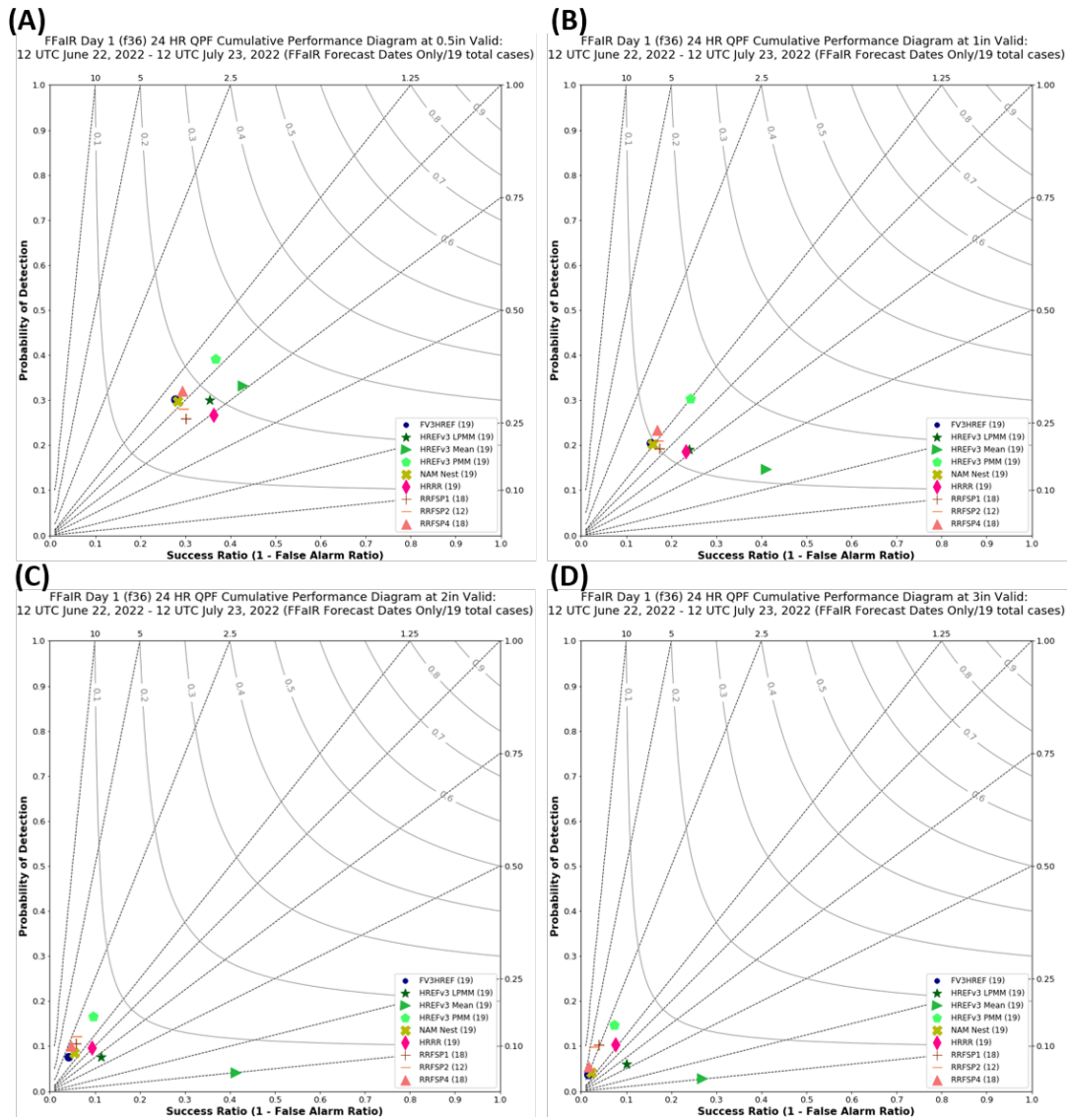


Figure 25: Performance diagrams for the 00z QPF forecasts valid for Day 1 for only the days in which FFaIR was in session, from June 21 to July 22, 2021 for the deterministic models evaluated during FFaIR. Precipitation thresholds are for: (A) 0.5 inches, (B) 1 inch, (C) 2 inches and (D) 3 inches. The symbols/colors for each model are: HRRR is pink \diamond , NAMnest is yellow \times , RRFSp1 is purple $+$, RRFSp2 is pink $-$, RRFSp4 is pink \triangle . Also include but not discussed are the FV3 member of the HREF which is the dark blue \bullet and the HREF mean (\triangleright), probability matched mean (\diamond) and local probability matched mean (\star) all in various shades of green.

regularly noting that the HRRR seemed too dry. This perception is verified when looking at the frequency bias at a half and one inch during FFaIR (Fig. 25A-B). However, when looking across the Testbed Season, the dry bias was less prevalent. Meanwhile, the CSI for all thresholds, except 3 inches, is similar among the Group 1 and 2 models during FFaIR, but over the entire Testbed session the HRRR's CSI was the highest at all thresholds evaluated. Therefore, as already stated, although the RRFSp1/2 seem to be comparable to the HRRR, the performance of the HRRR during the 2022 FFaIR Experiment seems to be uncharacteristically low and thus may not be the best baseline to use to evaluate the performance of the RRFS during this period.

Echoing what has been noted in previous FFaIR Experiments, participants once again commented on the over abundance of weakly forced (aka popcorn or weakly forced) convection, along with a wet bias at higher thresholds. They often noted that the RRFS was wetter than the NAMnest, which is generally considered a “hot” model by the weather community. Furthermore, when it came to the weakly forced convection, they were hesitant to trust the evolution of storms since the RRFS kept these as isolated, strong storms rather than clustering them into more realistic areas of storms. For instance, when asked about model performance throughout the week most comments had a similar sentiment to the following:

“The RRFSs generally performed similar to the NAM. They both had pros and cons. One thing that stuck out for me was that while the NAM has high rain rates, the RRFSs have even higher rain rates.”

“One bias I generally noticed from the RRFSp1 and p2 was the tendency to over forecast coverage and rainfall maxima pockets for the pulse thunderstorm convection over the Texas/Louisiana/Arkansas Gulf Coast area. Otherwise it handled the northeast CONUS convection and southwest convection fairly well.”

“RRFS seemed to really overdo precip accums, especially in moist air masses like with the tropical system or the southeast.”

“Overall, I thought it was too popcorny and wasn't showing enough more organized thunderstorm clusters.”

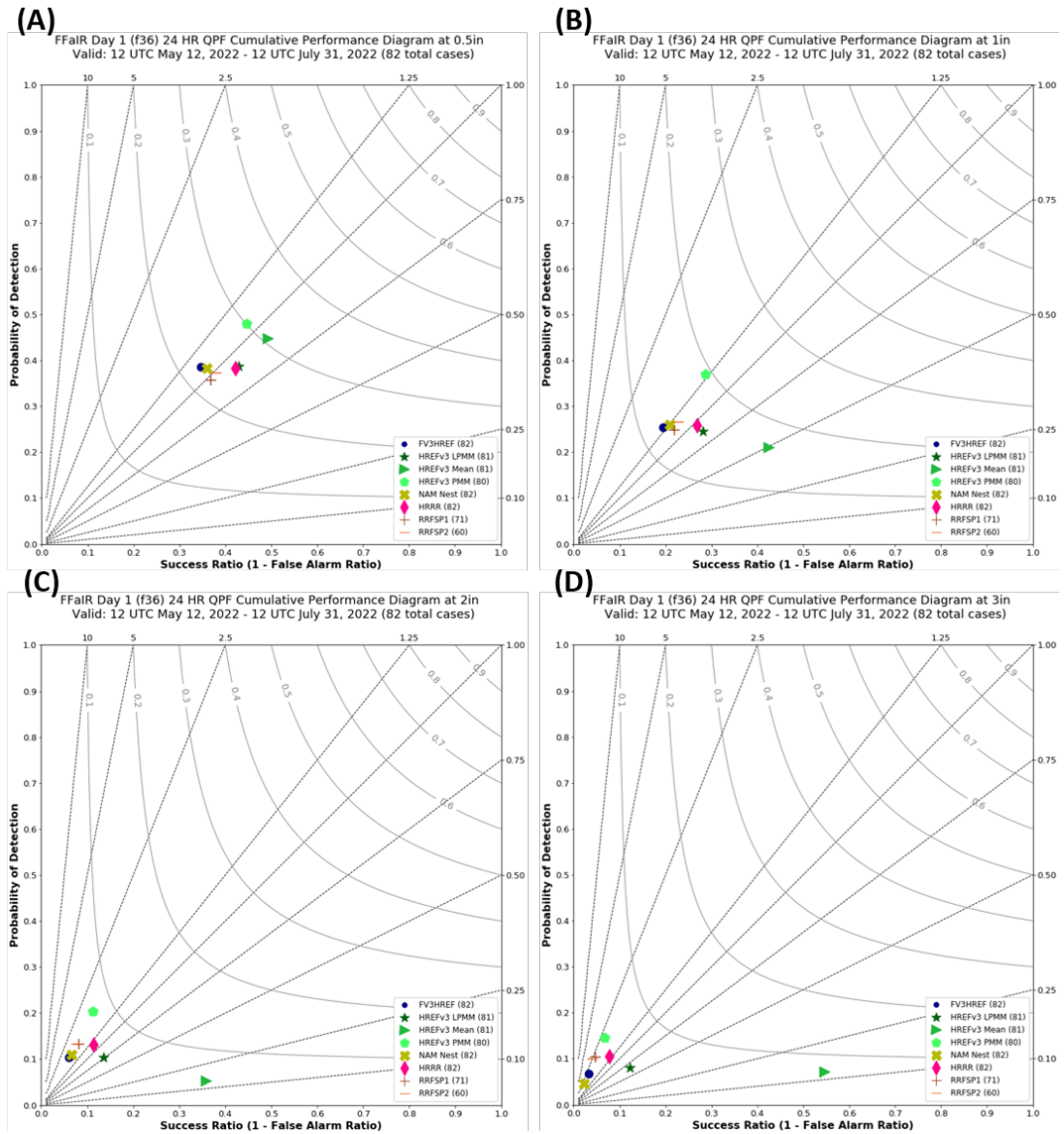


Figure 26: Similar to Fig. 25 but across the Testbed Season, May 12-July 31, 2022.

The comments on the popcorn convection and wet bias for individual days were more specific. For instance, for the day shown in Fig. 23 participants noted:

“A few of the models (NAMNest, p3, and p5) captured the MCS in Kansas quite well. RRFS had a lot of 3-6” totals for pulse storms in the SE CONUS. This overestimation was evident last year as well.

“Most of the RRFS models showed too high in rainfall amounts and too much coverage with the “typical” afternoon convection across the southeast CONUS. They did fairly well with the other features, such as the front from NE to northern VA, and with the locations of the tropical disturbance near the TX/LA coast and Carolina coast.

Figures 27 and 28 provide an hourly look at the forecast popcorn storms across the southeast for July 1st. Although none of the models shown did particularly well with the forecast at forecast hours shown (21 and 22 UTC), there are themes between the RRFS forecasts (D-F in the aforementioned figures) that are not seen in the operational models (B-C) or in the MRMS QPE (A). For instance, across MS/AL/GA/SC there were widespread areas of light precipitation (≤ 1 in.) with embedded pockets of 1-3 inches. Although underdone on both coverage and amount, the NAMnest forecast light rain with embedded heavier amounts. On the other hand, RRFSp1/2/4 all had scattered, single cell convection across the area. While clusters of heavier amounts were observed, single cell storms populate the region, with the majority of the cells seeing at least an inch of rainfall per hour. Although an inch in an hour or greater can occur in diurnally driven storms like these, to have nearly every cell have such totals is unlikely given this day’s environment.

Another example of the difference between RRFS and operational models for hourly precipitation can be seen for the June 29th forecasts in Figs. 29 and 30. Again, the lack of cell clustering is apparent in RRFSp1/2/4 (D-F in the aforementioned figures) when compared to the operational models and MRMS. The storms forecast by the RRFS models seem more isolated in nature and more numerous. Zooming in further (Fig. 31) the RRFSp1/2 rainfall distribution in the cells often resembles thunderstorm cores; the highest total is centered in the cell, with a tight gradient to the maximum precipitation. The maximum hourly QPF,

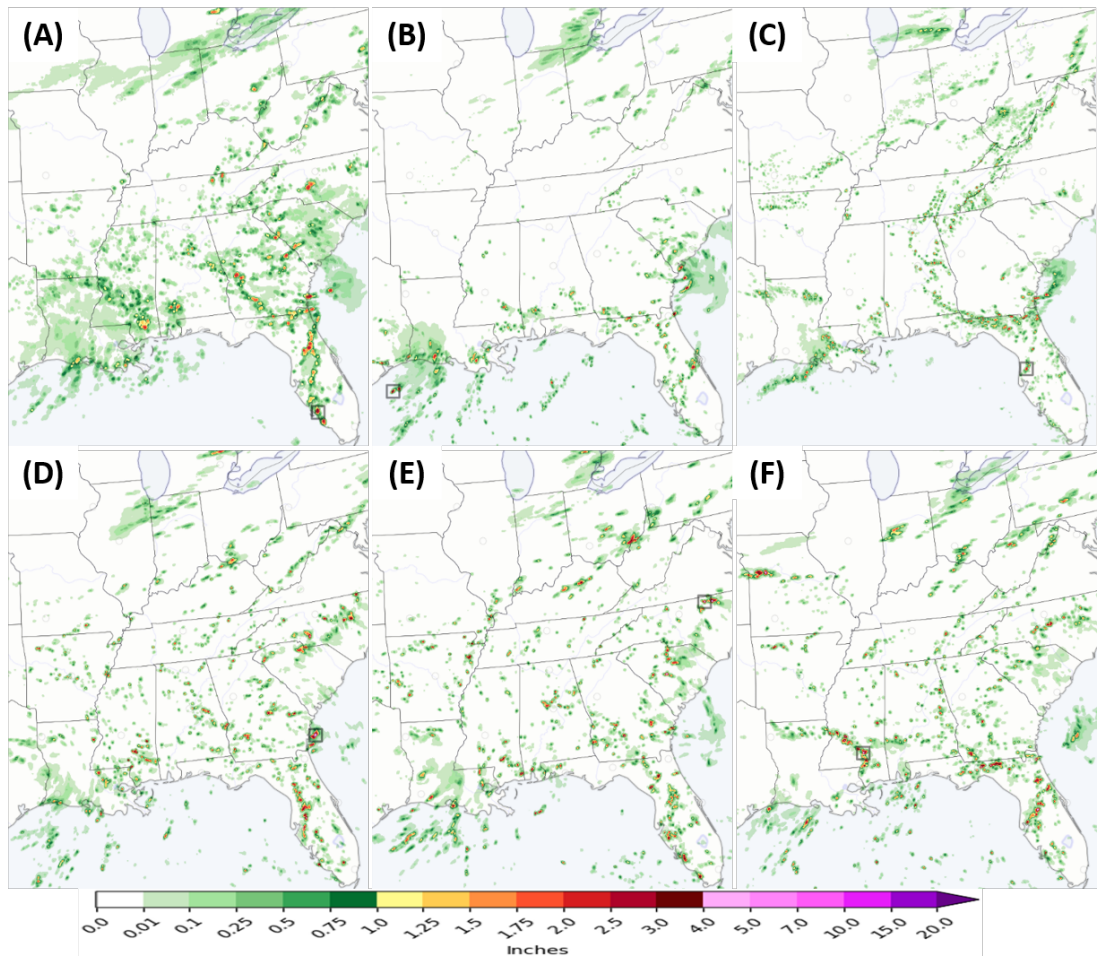


Figure 27: Hourly (A) MRMS QPE and (B) HRRR, (C) NAMnest, (D) RRFSp1, (E) RRFSp2, (F) RRFSp4 QPF valid 21 UTC 01 July 2022.

identified by the grey box, occurred over this zoomed in area in the NAMnest (3.96), RRFSp1 (5.51), and RRFSp2 (6.1). The observed maximum for this time period was in North Carolina (3.54"). Not only was the NAMnest's magnitude similar to MRMS at this time, it was also similar in location, along the Carolina coast albeit the wrong Carolina (SC rather than NC).

The prolific occurrence of strong popcorn convection does not appear under more strongly forced environment. For instance, for the 00z forecast on July 01, 2022 across the Ohio River Valley and into southern MI, the evolution/structure of hourly (Figs. 27 and 28) and 24 hour (Fig. 23) precipitation from the RRFS models look comparable to the operational models. Even across the southwest,

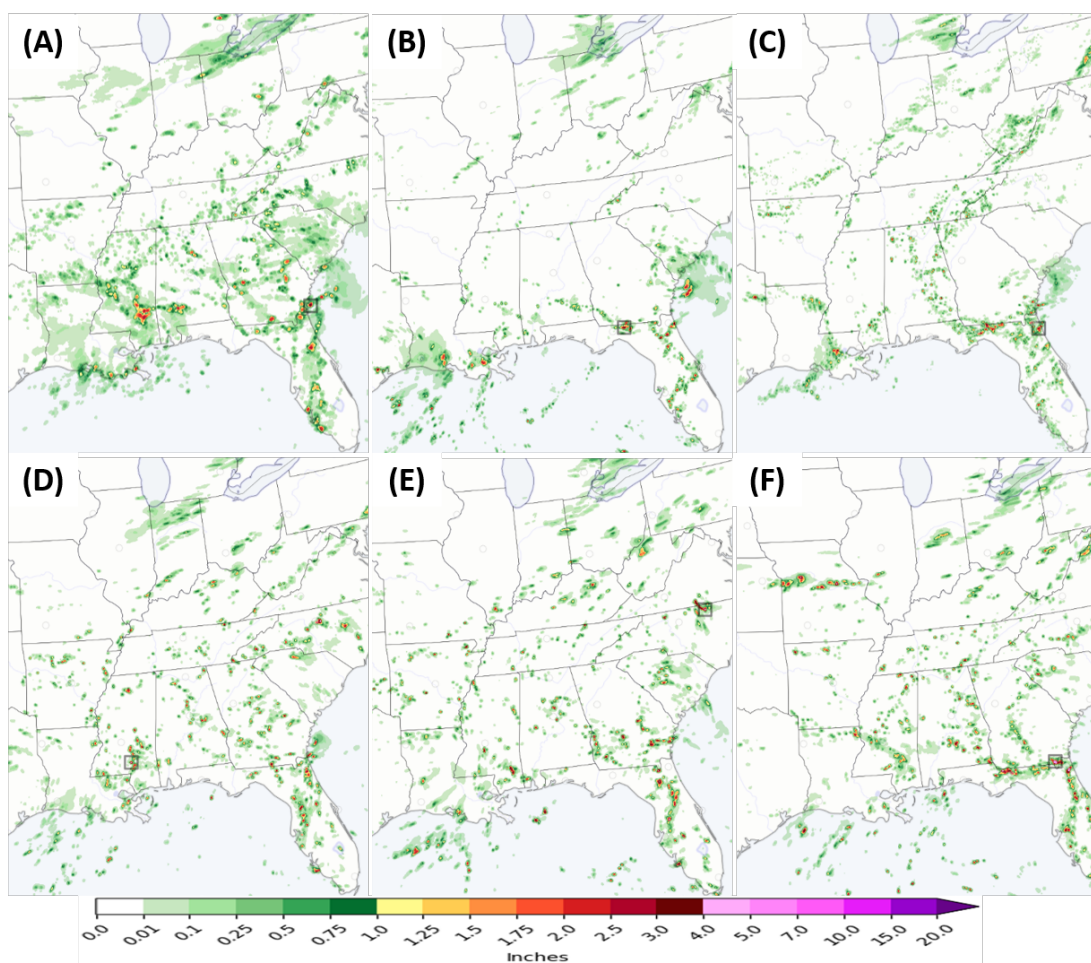


Figure 28: Hourly (A) MRMS QPE and (B) HRRR, (C) NAMnest, (D) RRFSp1, (E) RRFSp2, (F) RRFSp4 QPF valid 22 UTC 01 July 2022.

with an active Monsoon, the forecast precipitation from RRFs models appears more realistic, with rainfall spread out across the region rather than a speckling of isolated cells.

Since the excessive development of popcorn convection during the Testbed Season appeared to be confined in and around the southeastern US, hourly precipitation was compared over 3 domains: the southeast, southwest (due to the active monsoon) and the CONUS (see Fig. 32), with a focus on the HRRR, NAMnest, RRFSp1 and RRFSp2. The survival function, here defined as the sum of all histogram bins at and above each bin threshold, for the hourly precipitation across each of the aforementioned domains can be seen in Fig. 33. This one dimensional

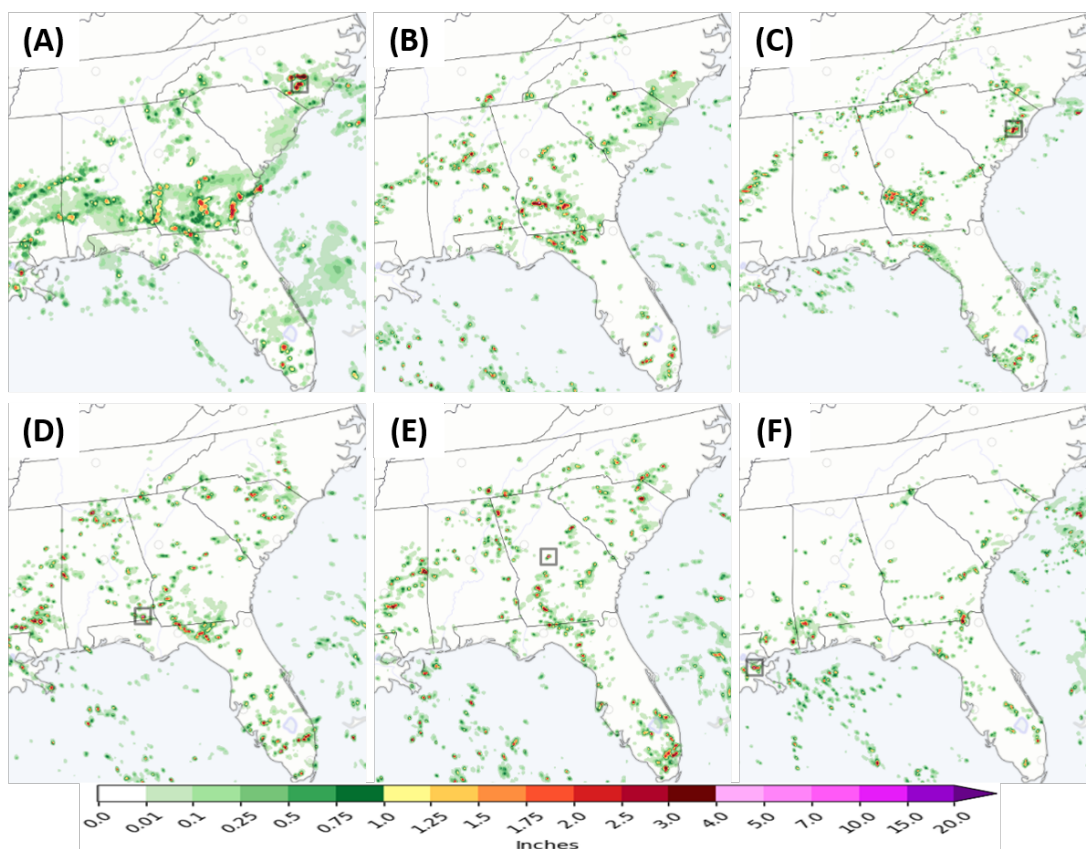


Figure 29: Hourly (A) MRMS QPE and (B) HRRR, (C) NAMnest, (D) RRFSp1, (E) RRFSp2, (F) RRFSp4 QPF valid 21 UTC 29 June 2022.

view is akin to using each bin as the precipitation areal coverage for any precipitation threshold.

Focusing on the CONUS (top image in Fig. 33), all the models have a wet bias at higher thresholds, while slightly under forecasting the coverage of hourly totals ≤ 1.5 in compared to MRMS. After 1.5 in, RRFSp1 and RRFSp2 begin to differ from the operational models and MRMS, having greater coverage of hourly precipitation totals ranging from 1.5 to ~ 6.75 inches. After ~ 6.75 in, MRMS coverage becomes negligible while the NAMnest's coverage of hourly accumulation greater than 6.75 in surpasses the RRFSp1. Meanwhile, the RRFSp2 continues to diverge from the other models and has a longer tail. In fact, although the plot stops at 13 inches, the highest hourly precipitation total output by the RRFSp2 during the Testbed Season was 41.69 in.

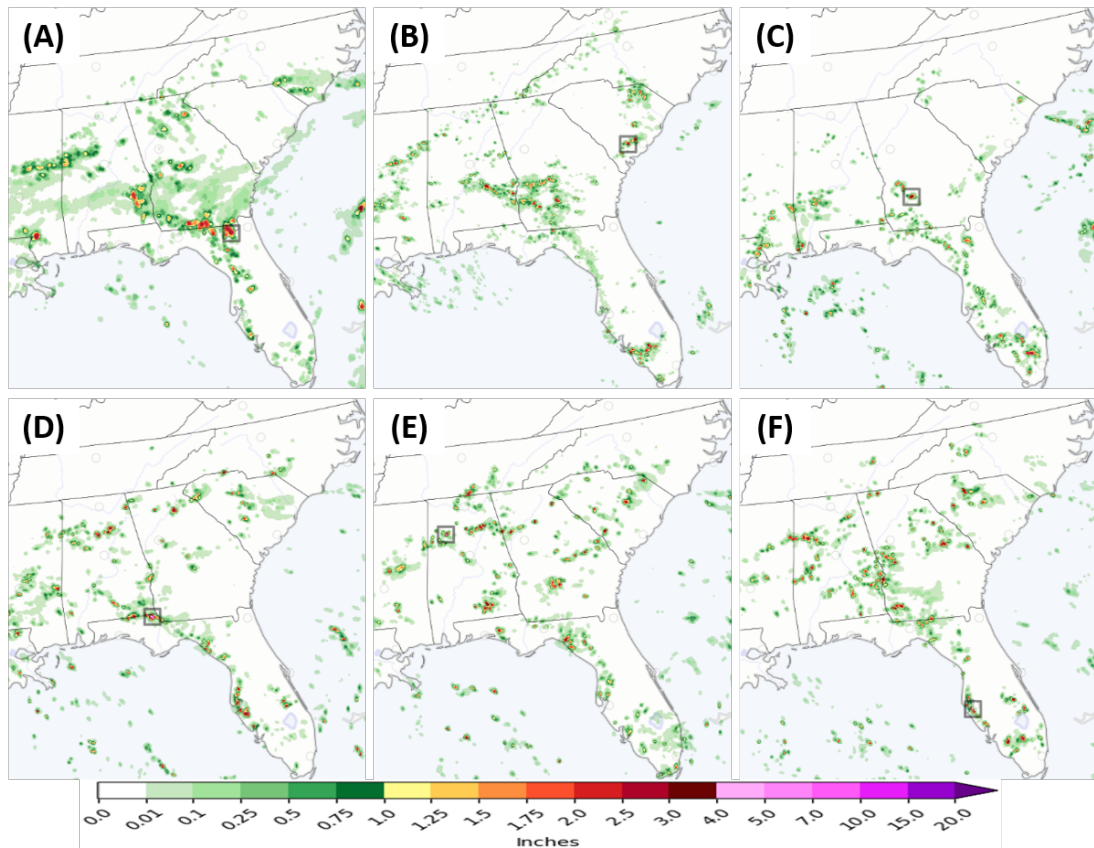


Figure 30: Hourly (A) MRMS QPE and (B) HRRR, (C) NAMnest, (D) RRFSp1, (E) RRFSp2, (F) RRFSp4 QPF valid 23 UTC 29 June 2022.

The differences in the coverage of hourly rainfall between the RRFs prototypes versus the operational models and MRMS are more pronounced when the domain is narrowed to the southeast. Around 1.25 in, the RRFSp1/2 coverage outpaces MRMS without their curve ever resembling the MRMS curve. On the other hand, the HRRR and NAMnest have a survival curve that is more similar to MRMS, though both do begin to have a wet bias around 2.5 in. Furthermore, unlike in the CONUS analysis, neither operational model outpaces the RRFSp1 in terms of coverage. Opposing both the CONUS and southeast, across the southwest (bottom image in Fig. 33) all the models have a dry bias for thresholds near and below 3.25 in, with the NAMnest, RRFSp1, and RRFSp2 all performing similar to one another.

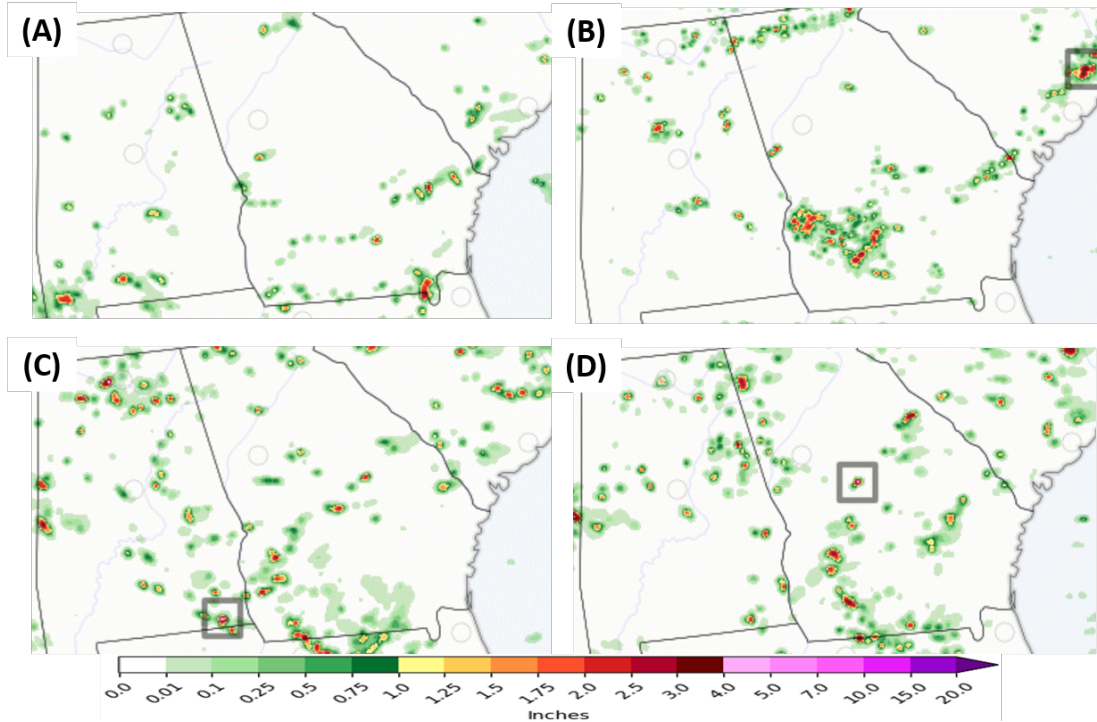


Figure 31: Similar to Fig. 29 but zoomed in on AL/GA for the (A) HRRR, (B) NAMnest, (C) RRFSp1, and (D) RRFSp2 forecasts. If a grey box is present in the zoomed image, it indicates that the CONUS max hourly QPF was observed there.

The characteristics of the hourly rainfall survival functions over the CONUS gives further insight into the 24-h QPF performance diagrams seen in Figs. 25 and 26. Like in the 24-h QPF, the wet bias from the RRFSp1/2 noticeably increases as the threshold increases. The survival function also helps identify both where the models' wet bias begins and to pin-point at what threshold the greatest wet bias is seen. This could suggest that the wet bias seen in the 24-h QPF is driven by short duration rainfall accumulation totals, though further analysis will need to be done to fully understand the relationship. Combined, these results support participant feedback about the over forecasting of precipitation, especially across the southeast and mostly in the form of popcorn thunderstorms.

The diurnal cycle of hourly precipitation for the CONUS and the southeast can be seen in Fig. 34. The left images includes zeros (i.e. grid points where QPF is zero) in its analysis, which can be thought of as showing the coverage of rainfall over the respective region. The right images do not include grid points where QPF

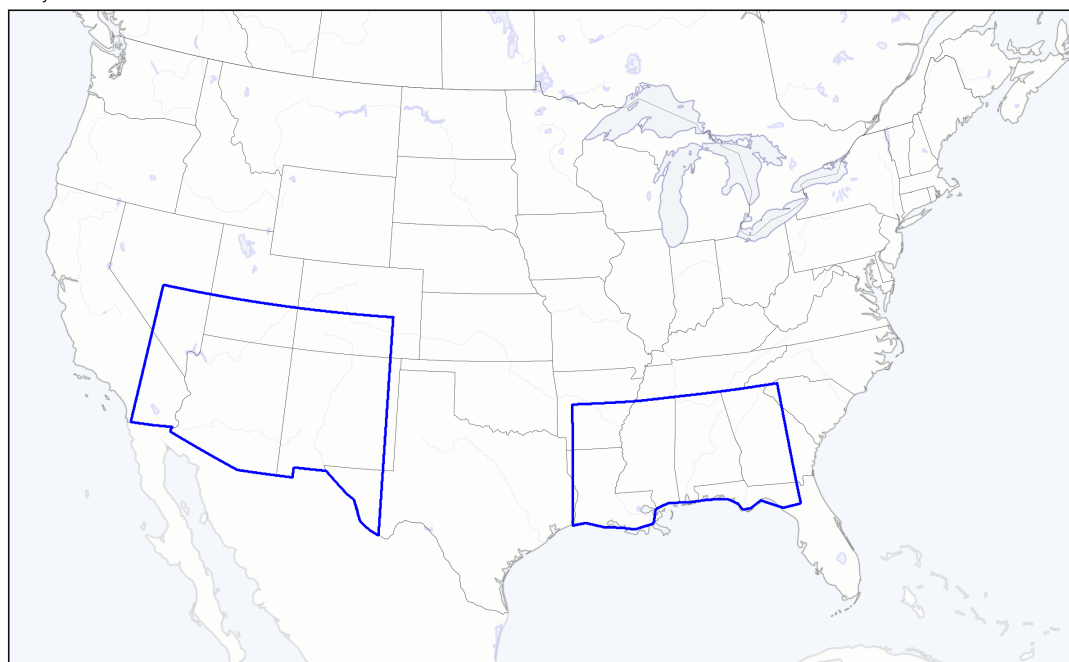


Figure 32: The boundaries for the southeast and southwest domains used in the precipitation analysis.

is zero, meaning only grid points that had rainfall were counted. These can be thought of as showing the average intensity of the rainfall across their respective domains.

In terms of coverage, the HRRR, NAMnest, RRFSp1, and RRFSp2 all have lower average hourly precipitation than MRMS over the diurnal cycle. However, the RRFSp models both more closely resemble the curve than the operational models, in magnitude and timing. When compared to the HRRR, both experimental models' average hourly rainfall is 1.5 to 2 times as large across the CONUS and 5 to 6 times as large over the southeast. Upon first glance, this increase in what can be thought of as precipitation coverage suggests an improvement in QPF forecasting over the operational models. However, when examining just when it is raining (i.e. intensity), the RRFSp models actually exceed MRMS average hourly rainfall at every time period. At their respective peaks, the RRFSp1 and RRFSp2 over forecast the magnitude by approximately 0.00025 to 0.00065 inches across the CONUS and 0.001 to 0.002 inches in the southeast. Combined, this suggests that

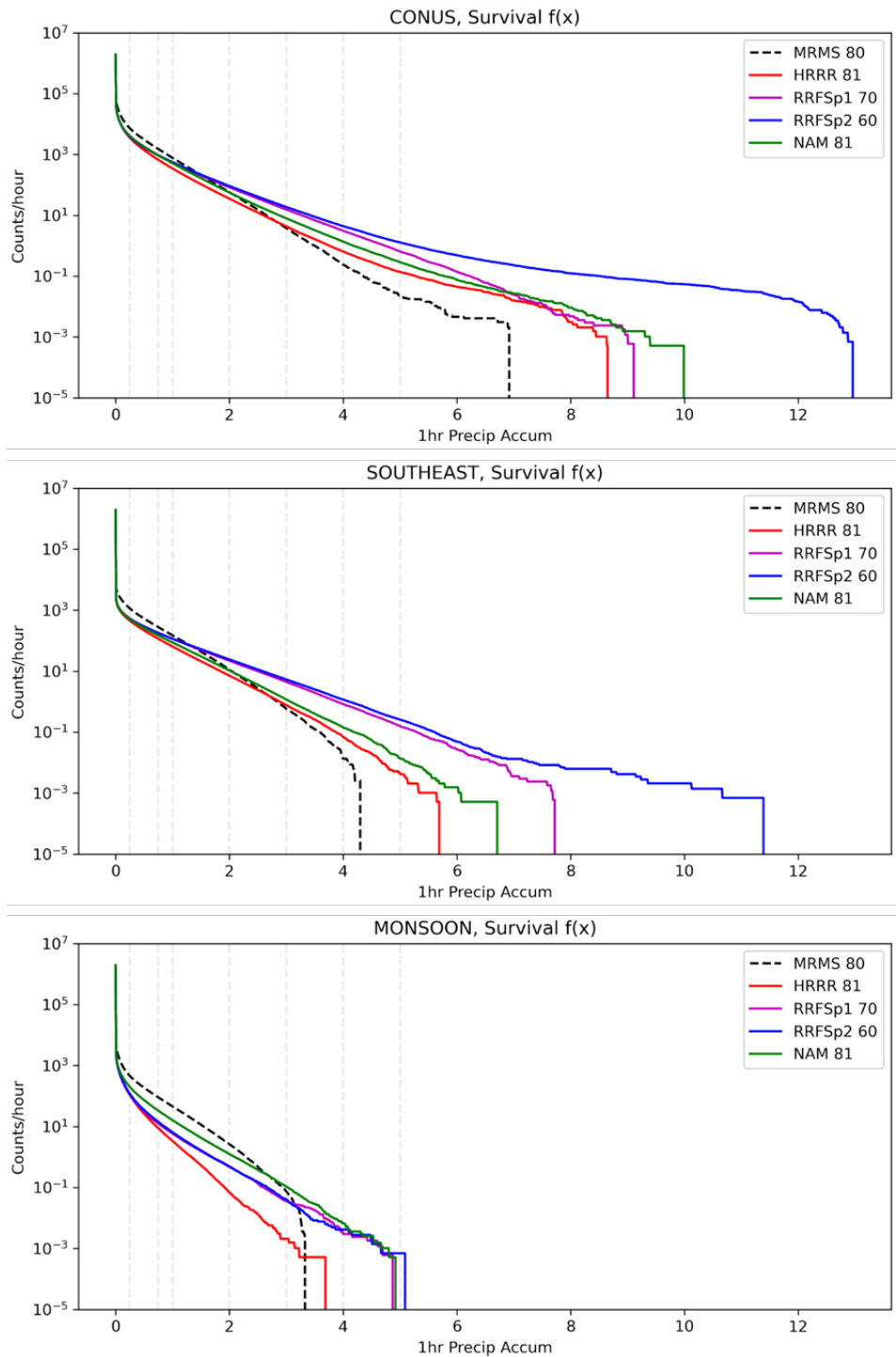


Figure 33: Hourly QPF survival function for the Testbed Season over the CONUS (top), the southeast (middle), and the southwest (bottom), comparing MRMS (dashed), HRRR (red), NAMnest (green), RRFSp1 (purple), and RRFSp2 (blue). On the y-axis is the counts per hour and on the x-axis is the 1 hr precipitation accumulation.

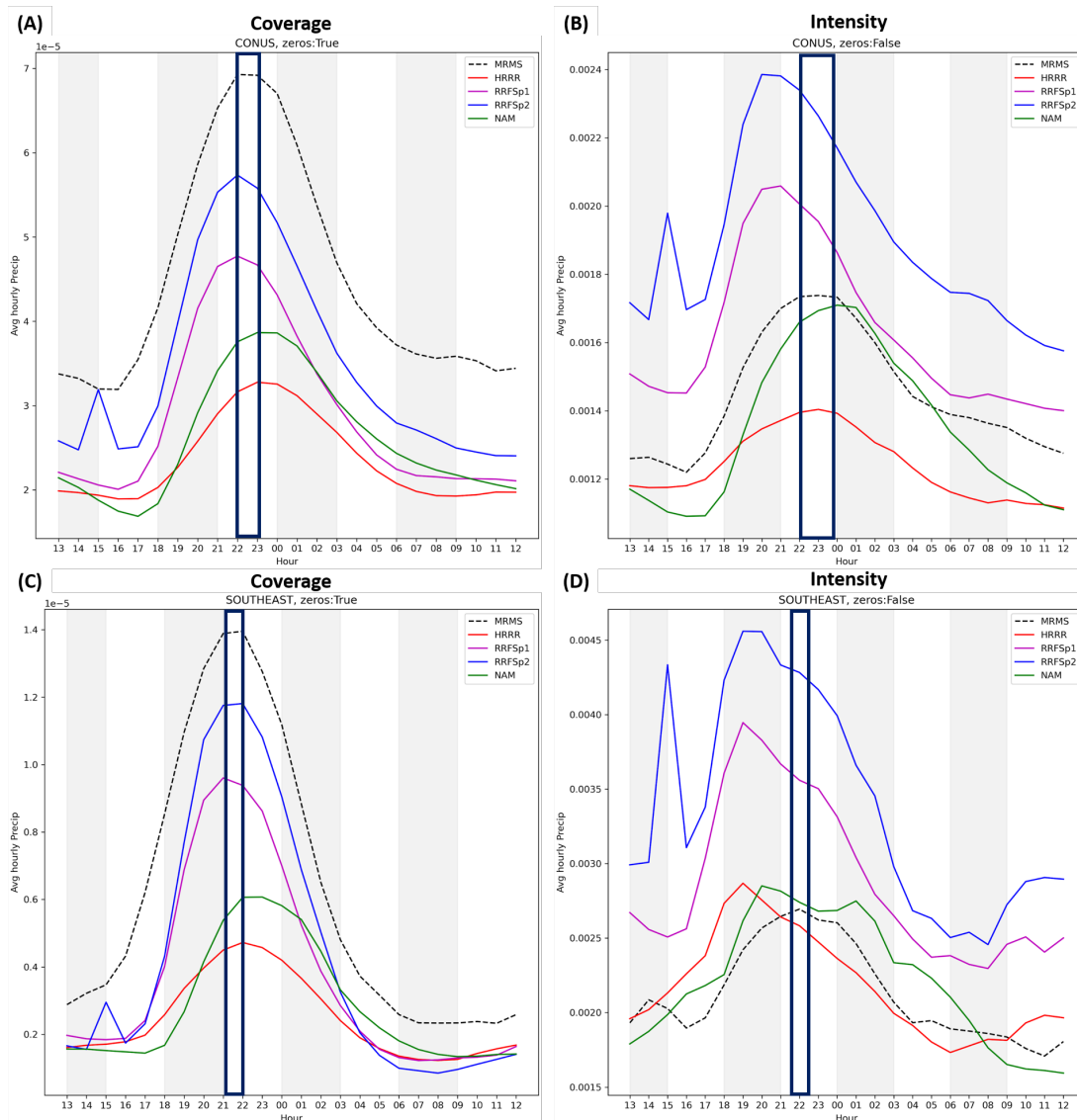


Figure 34: Diurnal analysis of hourly QPF for the Testbed Season over the CONUS (top) and the southeast (bottom). (A) and (C) show coverage (zeros included in average) of hourly precipitation. (B) and (D) show the intensity (zeros NOT included in average). The black rectangles indicate the time in which MRMS is at its maximum; the width of the rectangle varies among the plots based on the duration of the MRMS maximum. Note: the average hourly precipitation are scaled differently for each of the images.

although the RRFS models might better simulate rainfall occurrence (rain vs no rain), when it does rain, it is too intense, which agrees with the survival function results. It also supports what the participants noted about the rainfall footprint from the RRFS models generally being good but the totals being too high.

To help easily identify trends in the timing of convective initiation, each image in Fig. 34 has a black rectangle that approximates the time frame in which the maximum average rainfall is observed in MRMS. RRFSp1/2 have similar timing of maximums. When looking at the timing in terms of coverage, their maximum slightly precedes the observed maximum but is still closer to the observed maximum than the operational models. The HRRR and NAMnest both lag the observed maximum by roughly an hour. This might indicate that the RRFSp1/2 better forecast the onset of afternoon convection than the operational models. However, the RRFSp1/2 diurnal maxima in rainfall intensity over the CONUS and southeast occur earlier compared to MRMS. Over the CONUS, both maxima are roughly 2 hours earlier than observed whereas the HRRR has roughly the same timing in maximum as observed and the NAMnest still slightly lags MRMS. The HRRR has noticeably lower average intensities over the duration of the day. On the other hand, between 19 UTC and 5 UTC, the NAMnest most closely resembles MRMS.

Across the southeast, the time of the peak in average hourly rainfall in the HRRR closely matches the RRFS models, with the max near 19 UTC rather than 22 UTC. Meanwhile, the NAMnest peak more closely resembles MRMS, though rather than a smooth increase/decrease in intensity, it has two peaks in the maximum, roughly at 1930 UTC and 0030 UTC. Despite the differences in the tendencies of their maxima, both operational models have average hourly values more similar to MRMS than either RRFS model. The most pronounced change between the CONUS and southeast is seen in the HRRR. Across the CONUS, the HRRR's change in average hour precipitation over time is subtle in comparison to MRMS and the other models. However, over the southeast the curve of the diurnal increase is more similar to MRMS and the NAMnest, at least in terms of magnitude. This suggests that although the HRRR tends to struggle in simulating the increase in convection associated with the maximum daytime heating, when

there is abundant moisture (which is common over the southeast during the warm season) it can produce QPF similar to observations.

On the other hand, there is little difference in the pattern of precipitation intensity over the course of the day for the RRFSp1/2 between the two domains. The one exception is after the diurnal maximum. Both models have a sharper decrease in magnitude around 00 UTC over the southeast than is seen over the CONUS or in MRMS. This drop off is more intense in RRFSp2. Additionally, each model simulates an increase in average hourly rainfall starting around 08 UTC. This is also seen in the HRRR but it is not as large. This might suggest that when there is abundant moisture available, the RRFS models quickly “use up” all the moisture (thus the quick drop) and once the simulated environment rebounds from the sudden drop in moisture the cycle rapidly starts again, as is suggested by the increase seen around 08 UTC (i.e. forecast hour 32) in Fig. 34D.

3.1.1 A Short Summary

Due to previously mentioned data flow issues, most of the analysis performed was on the RRFSp1 and RRFSp2 over the span of the Testbed Season. The wet bias that has been noted in the previous versions of the RRFS (aka LAMs and SARs) by the FFaIR team is still present in the versions evaluated this year. Subjectively, participants noted the prolific simulation of day time convection that is diurnally driven, what is referred to as popcorn storms. These were often confined to the southeastern US. The most common comments about the popcorn storms on the QPF magnitudes and how isolated the storms were. The participants noted that it appeared that nearly every storm simulated produced 2+ inches of rainfall in an hour. Although it is not unlikely to see such accumulations from any given storm, it is unlikely that every storm that develops will have such high rainfall rates. As to the isolation of the cells, they often mentioned that rather than simulating clustering of storms or a broad area of convectively driven precipitation, the RRFS models had cells that were distinctly individual; noting that it appears as though the cells do not move.

Digging deeper into the comments about the high hourly rainfall totals being simulated in the RRFS models, hourly QPF was analyzed; ex. Fig. 33. Analysis

showed that over the CONUS starting around 1.5 in the RRFS models began to have a wet bias, while the NAMnest's wet bias began around 2 in and the HRRR's around 3 in. The RRFSp2 hourly totals distinctly outpaces all the models and MRMS while the RRFSp1's wet bias outpaces the operational models until about 7 in. At this point the NAMnest's wet bias becomes greater than the RRFSp1's. However, when focusing on the southeast, the RRFSp1/2 start to have a wet bias sooner than over the CONUS and neither operational model ever exceeds them. When looking at rainfall diurnally (Fig. 34), both RRFS models have convective initiation too early (by roughly 2 hours). Additionally, comparing coverage to intensity showed that although the average hourly rainfall coverage is more similar to MRMS for RRFSp1/2 than the operational models, when focusing on intensity both models have higher averages than MRMS for all hours of the day. This suggests that the RRFS models are more likely to simulate rainfall than the operational models but when it does rain, the intensity is too high.

3.2 Note on Ensembles

As noted in section 2.3.1, the ensemble guidance was the least available. For the RRFS_e, issues arose not only with the running of the system but also in the post-processing. The FFaIR team relies on data providers to create the ensemble products (ex. means, probabilities, etc.) rather than running in-house post-processing on the members of the ensemble. Although the ensemble products were created, an error was discovered in the post processing code during the experiment. Similarly, the ensemble provided by OU CAPS, the CAPS_RRFS_e, ran into computational errors when creating the post-processed ensemble guidance. Therefore, although planned, an analysis of the RRFS-based ensembles and how they compare to the HREF could not be performed.

Despite the difficulties with the ensemble products, the individual members of the RRFS_e and CAPS_RRFS_e ran in a semi-reliable cadence. This made it possible for some ensemble analysis to be performed. On a daily basis, a python-based website was used to quickly look at either 1/3/6 hour precip and hourly prate and pmax. An example of the display can be seen in Fig. 35, showing the hourly max precipitation values over the CONUS from the ensemble members compared

to the deterministic RRFSe runs and MRMS for the 00z June 29 2022 cycle. Note that the members from the RRFSe (top section in image) tend to have the highest max hour precipitation of all the configurations on this day. A similar dashboard was created for the deterministic runs, which included looking at the operational models.

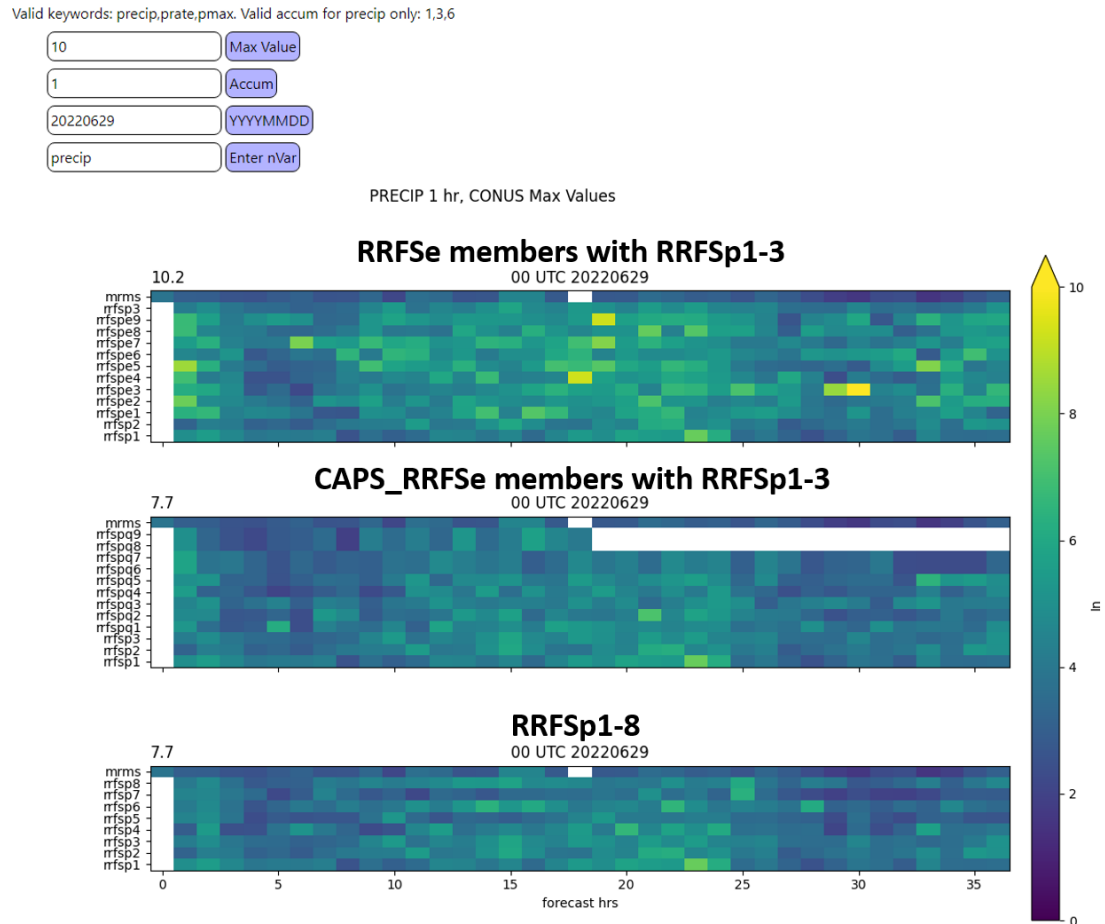


Figure 35: Screen capture of the Ensemble Member Dashboard developed for quickly viewing ensemble member output of max QPF (1, 3, or 6 inches), prate, and pmax. User is prompted to chose the parameter they want to view along with the max value for the colorbar and the initialization date. Top: RRFSe members with RRFSp1-3. Middle: CAPS_RRFSe members with RRFSp1-3. Bottom: RRFSp1-8. MRMS as the top “line” in each section. Valid for the 00z cycle on June 29, 2022.

3.3 Precipitation Rate

In previous FFaIRs, it was hypothesized that high precipitation rates were contributing to wet bias seen in the various RRFs versions evaluated. Last year, the team analyzed the instantaneous precipitation rate (hereafter p-rate) and found instances of extremely high p-rates from the LAMs (exceeding 150 in h^{-1})⁸. Based on these findings and feedback from the FFaIR team, developers worked to identify what could potentially be causing these large rates. Updates were made to the system to address the potential culprits and implemented in this year's RRFs. Additionally, EMC included the hourly maximum precipitation rate (hereafter pmax) in its GRIB2 files, at the request of the FFaIR team.

Although both p-rate and pmax were evaluated, most of this discussion will focus on pmax. One continuous comment about looking at p-rate versus pmax made by participants was that the footprint of pmax more closely resembled MRMS than p-rate. Specifically, even though the coverage of both model precipitation rate variables nearly always appeared to be less than observed, pmax tended to have higher coverage than p-rate, and thus was preferred by the participants. Another consistent comment was that despite the low areal coverage of p-rate/pmax, the highest values for each variable output by the RRFs models were routinely larger than the MRMS or from operational models. For instance, looking at Figs. 36 and 37, coverage of the rates is lower than observed and the “look” of the rates in the models are more cellular than MRMS. This is to be expected due to the differences spatially and temporally between the models and MRMS. When focusing on the magnitude of the highest observed rate over the domain at this time (denoted by the small black box in each image with the magnitude listed in the caption of the figures), on average the RRFs models produce values approximately 13 in h^{-1} to 27 in h^{-1} greater than MRMS.

The hourly rainfall at f19 for this day can be seen in Fig. 38. In this case, the maximum hourly precipitation accumulation does not necessarily correspond with the location of greatest p-rate/pmax. However, for the RRFSp1, the highest

⁸See Fig. 67 in Trojaniak and Correia, Jr. (2021) for a summary of the p-rates observed compared to MRMS.

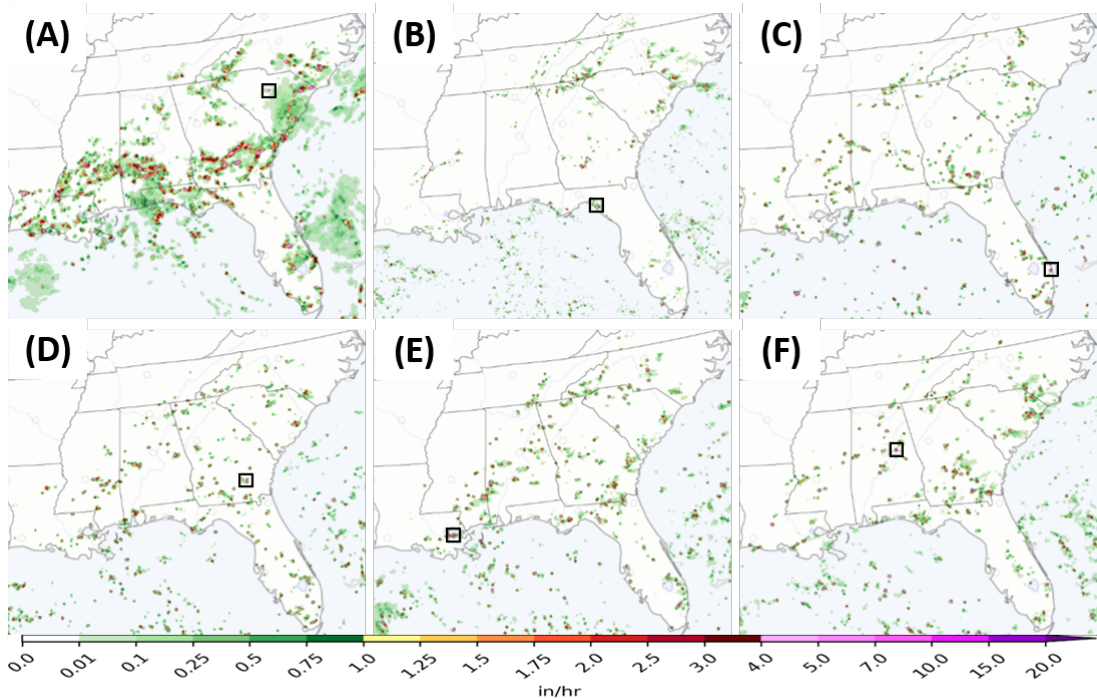


Figure 36: Hourly p-rate and the maximum rate from over the CONUS, which is depicted by a black square on the image and magnitude is shown within the [] of this caption. (A) MRMS QPE [7.87 in h^{-1}] and (B) NAMnest [10.6 in h^{-1}], (C) RRFSp1 [14.17 in h^{-1}], (D) RRFSp2 [10.63 in h^{-1}], (E) RRSFp3 [12.11 in h^{-1}], and (F) RRSFp4 [21.7 in h^{-1}] valid 19 UTC 29 June 2022.

pmax and hourly rainfall were both forecast to occur in the same cell, off the coast of Florida's Panhandle. Figure 39 zooms in on this region to show the pattern of rainfall and rates. The maximum pmax is 20.27 in h^{-1} . The RRFs models have a time step of 60-s which means that the model is producing 0.34 in rainfall, or roughly 7% of the forecasted 4.9 in of QPF in one minute. Although this is large, the RRFSp2 for this forecast time had a pmax of 31.15 in h^{-1} which equates to 0.52 in min^{-1} . In other words, the RRFSp2 is forecasting a half inch of rainfall to occur in one minute. The corresponding cell for the maximum pmax in RRFSp2 is 3-4 inches. That means in the RRFSp2, 15%-20% of the hourly QPF is being forecast to fall in one minute. For comparison, the time step for the HRRR is 20-s and for the NAMnest is 6.25-s. The maximum p-rate for the HRRR⁹ for this forecast hour was 6.25 in h^{-1} or 0.1 in min^{-1} or 0.035 inches in 20 seconds. For the

⁹Reminder, the HRRR does not output pmax.

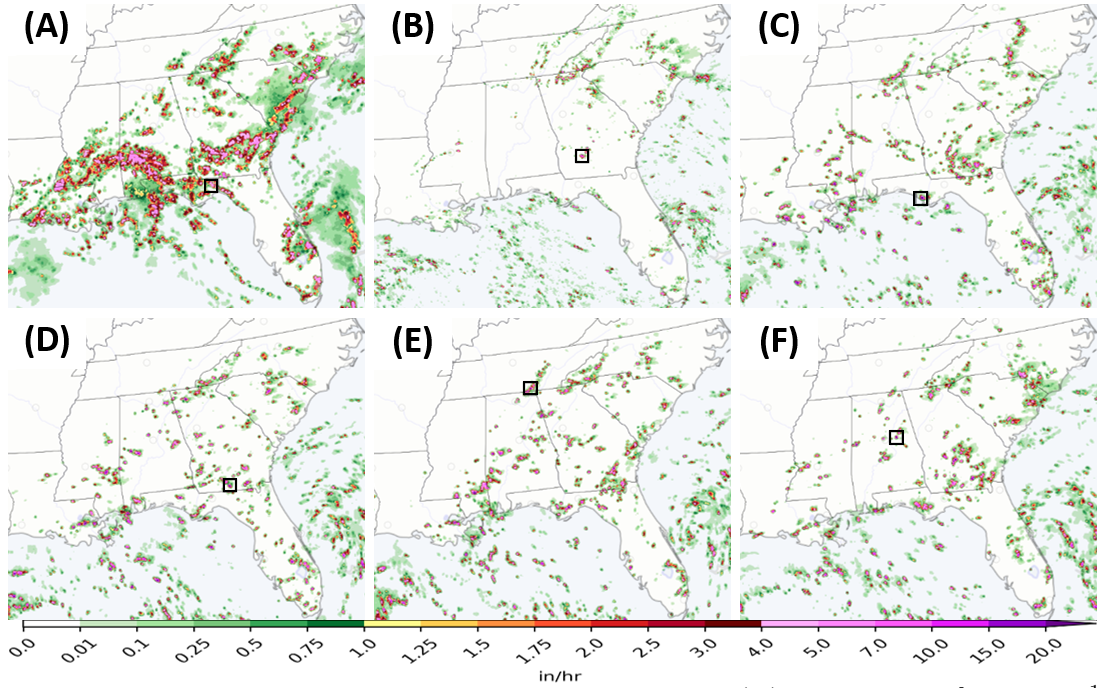


Figure 37: Same as Fig 36 but valid 19 UTC 29 June 2022. (A) MRMS QPE [7.87 in h⁻¹] and (B) NAMnest [12.36 in h⁻¹], (C) RRFSp1 [20.27 in h⁻¹], (D) RRFSp2 [31.15 in h⁻¹], (E) RRFSp3 [32.55 in h⁻¹], and (F) RRFSp4 [34.79 in h⁻¹].

NAMnest, the maximum pmax was 12.36 in h⁻¹, which is roughly 0.21 in min⁻¹ or 0.021 inches in 6.25 seconds.

Figure 40 shows how the maximum values of hourly QPF, p-rate, and pmax differed between MRMS, the operational models and RRFSp1/2 for June 29, 2022 over the CONUS. Comparing these maximum as a time series we can see the RRFSp models consistently have the highest value for each of the parameters. In fact, the 00z cycle of the RRFSp1 at f15 forecast a location to receive 7.7 in in an hour while MRMS maximum rainfall was around 4.5 in. Across the time period shown, the average maximum hourly rainfall for the RRFSp1 and RRFSp2 were 4.23 in and 4.28 in respectively while the average for the HRRR was 3.55 in, the NAMnest was 3.5 in, and MRMS was 2.84 in¹⁰. The maximum average p-rate(pmax) for the forecast were: MRMS - 7.11(7.8) in h⁻¹, HRRR - 7.16(N/A) in h⁻¹, NAMnest -

¹⁰Averaged over all forecast hours, respective of each models varying forecast lengths. See Table 1 for experimental model forecast length.

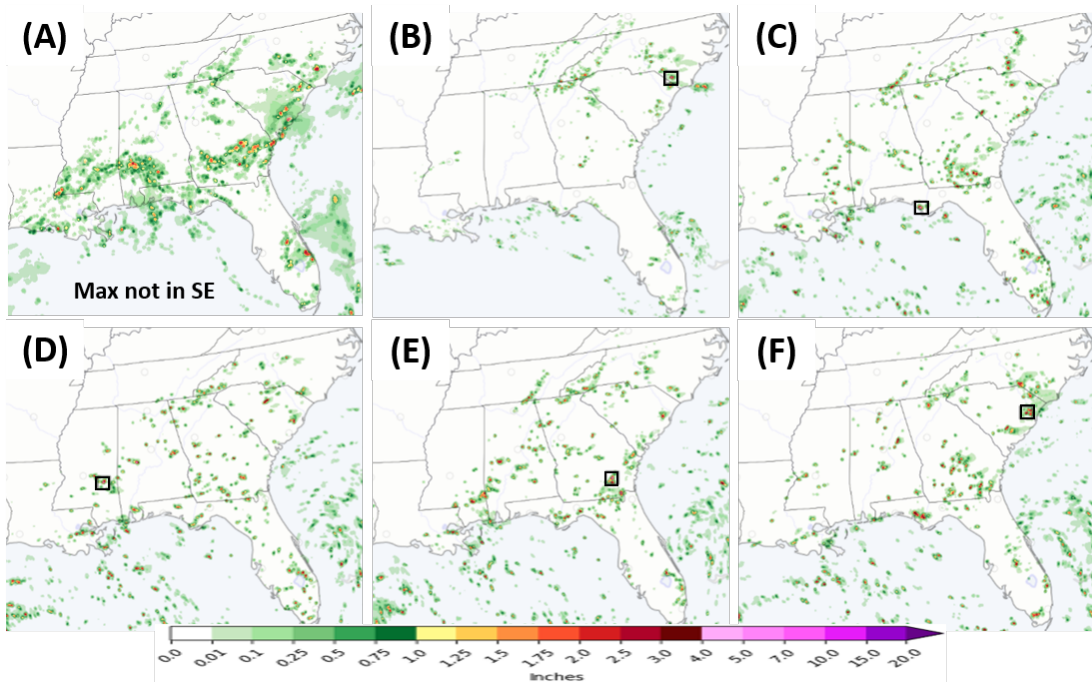


Figure 38: Hourly QPE/QPF and the accumulation from over the CONUS, which is depicted by a black square on the image and magnitude is shown within the [] of this caption. (A) MRMS QPE [2.8 in] and (B) NAMnest [4.82 in], (C) RRSFp1 [4.9 in], (D) RRSFp2 [4.83 in], (E) RRSFp3 [5.02 in], and (F) RRSFp4 [4.08 in] valid 19 UTC 29 June 2022.

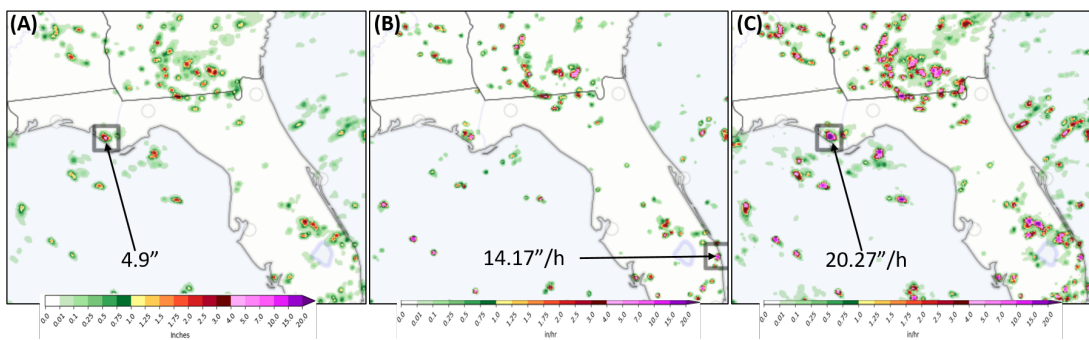


Figure 39: Hourly (A) QPF, (B) p-rate, and (C) pmax zoomed in on the Florida Panhandle, valid 19 UTC 29 June 2022.

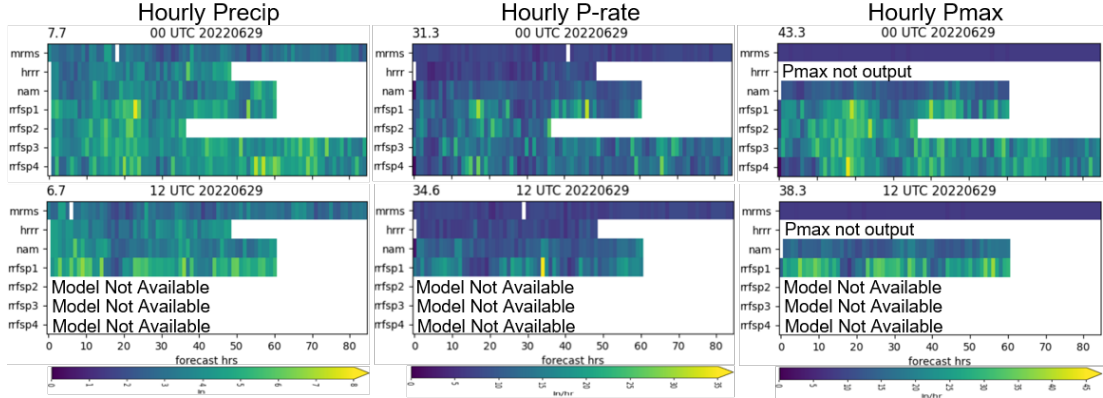


Figure 40: Hourly maximum value of (left) precipitation, (middle) p-rate, and (right) pmax for the 00z (top) and (bottom) 12z cycles on 29 June 2022. The model/observation shown in each image is ordered as follows: MRMS, HRRR, NAMnest, RRFSp1, RRFSp2, RRFSp3, and RRFSp4. The upper bound for colorbar is: (left) 8 in, (middle) 35 in h^{-1} and (right) 45 in h^{-1} . In the top left corner of each image is the maximum value from all models/observations across all hours.

9.52(13.12) in h^{-1} , RRFSp1 - 13.54(22.69) in h^{-1} , RRFSp2 - 13.59(24.92) in h^{-1} , RRFSp3 - 13.11(22.47) in h^{-1} and RRFSp4 - 12.14(21.04) in h^{-1} .

The high precipitation rates were not confined to the deterministic RRFSe models. Figure 41 shows the ensemble members from both the RRFSe (top graph) and CAPS_RRFSe (bottom graph) compared to the RRFSe deterministic members. The average pmax from the RRFSe members are 1.5-2 times larger than those of RRFSp1-RRFSp4 while their maximum pmax are 2-4 times greater. Additionally, the “bump” in the pmax seen at the start of the forecast in the deterministic RRFSe is exacerbated in the RRFSe members, with a few members seeing a maximum near 200 in h^{-1} . The peak differences between the deterministic and ensemble members seems to occur at the aforementioned time and around the diurnal maximum, suggesting the RRFSe is more excitable at these times. It is hypothesized that the higher pmax values output by the ensemble members are driven by the perturbation methods used to create the ensemble. Looking at the top graph in Fig. 42, hourly average and maximum rainfall accumulation are also greater in the RRFSe members than the deterministic RRFSe. However, the difference between the two categories is not as great as what is seen in the pmax, with the ensemble members being roughly 0.5 times greater than the deterministic.

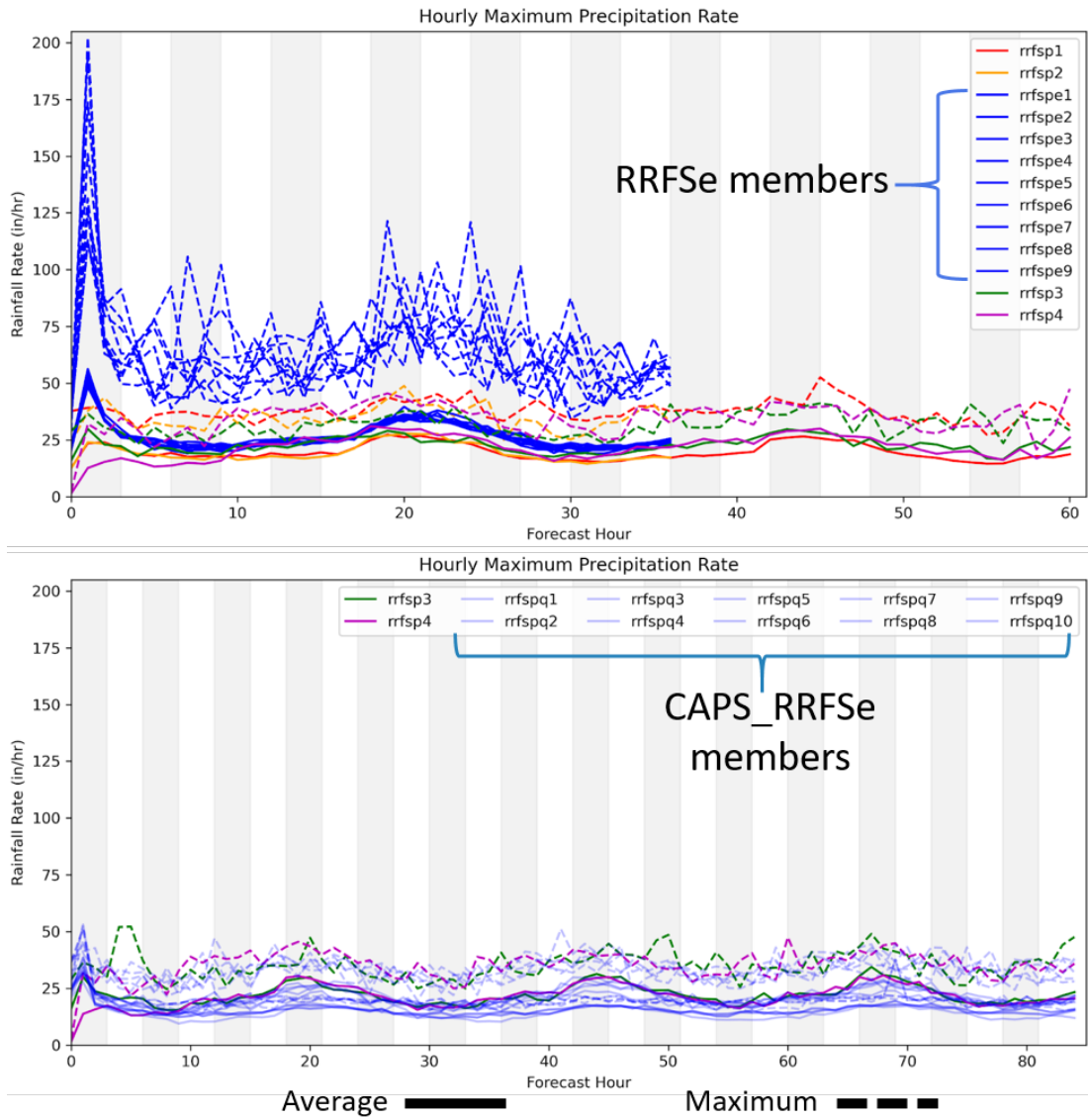


Figure 41: Hourly pmax across forecast hour for the TOP: RRFSp1 (red), RRFSp2 (yellow), RRFSp3 (green), RRFSp4 (purple), and the 9 members of the RRFSe (all in blue) valid over the Testbed Season. BOTTOM: RRFSp3 (green), RRFSp4 (purple), and the 10 members of the CAPS_RRFSe (all in light blue) valid over FFaIR dates. The solid line is the averaged pmax for each model and the dashed line is the maximum pmax seen at each forecast hour.

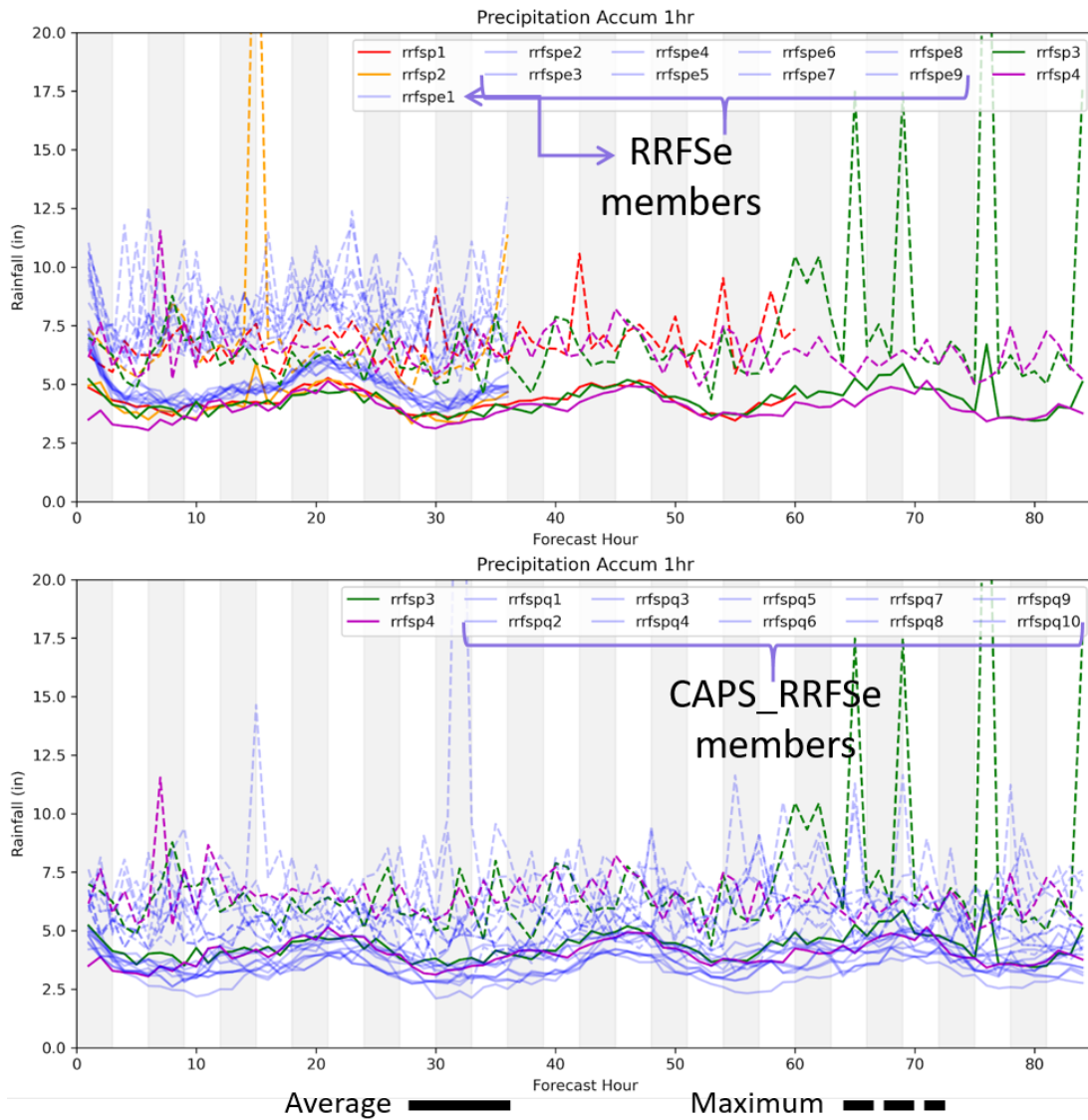


Figure 42: Similar to Figure 41 but for hourly precipitation and the RRFSe members are all light blue rather than blue.

The same differences in pmax are not seen between the RRFSp3 and RRFSp4 (deterministic CAPS models) compared to the CAPS_RRFSe members (Fig. 41 bottom graph). Rather, the CAPS_RRFSe members tend to have similar maximum pmax magnitudes to the RRFSp3 and RRFSp4, while their average pmax tends to be lower. A similar tendency is also seen when comparing the hourly precipitation totals (Fig. 42 bottom graph). The CAPS_RRFSe membership does not use perturbations to create member spread¹¹, rather differences in physics are employed; see Table 7 in Trojniaak and Correia, Jr. (2022). This further supports the hypothesis that the perturbation methods employed by the RRFSe to create ensemble spread could be exacerbating the already high precipitation rates and accumulations seen in the RRFSe members.

3.4 The EROs and AERO

This section will focus on the analysis of the CSU MLP EROs and the FFaIR ERO and AERO. The performance of the EROs will be discussed first, followed by the AERO. Please refer to Table 2 for the differences between the CSU MLP EROs.

The less than abundant number of heavy rainfall events during FFaIR also had an impact on evaluating the ERO and AERO's ability to identify high end events, although the MLP EROs' ability to correctly forecast null events is just as important as finding the extremes. Based on the practically perfect verification used by WPC for the ERO, no Moderate or High risk days occurred during the experiment. Moreover, even Marginal and Slight risk days were scarce this year in comparison to previous years of FFaIR.

The low coverage of the excessive rainfall Marginal and Slight risks for the Operational and FFaIR EROs during FFaIR this year compared to the FFaIR dates from 2021 can be seen in Figs. 43 and 44. These comparisons reiterate the stark difference between the events that occurred in FFaIR 2021 compared to this season. Apart from over the southeast and southwest, the CONUS had a frequency of a Marginal risk being draw over a region generally less than 18% of the time,

¹¹Some impact of perturbations can be found in the initial and lateral boundary conditions.

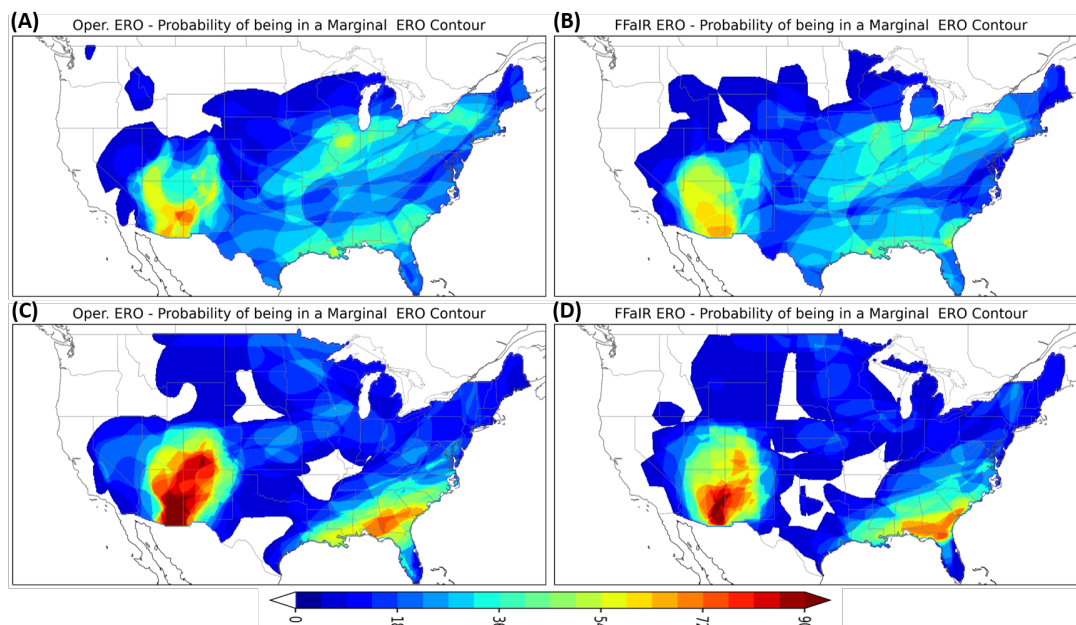


Figure 43: Probability of being in a Day 1 ERO Marginal risk during the (A)-(B) FFaIR 2022 and (C)-(D) FFaIR 2021 experiments. (A) and (C) Operational ERO and (B) and (D) FFaIR probabilities.

with some regions in the central Mississippi Valley never having a Marginal risk forecast during FFaIR this year. In contrast during FFaIR 2021, nearly all of the CONUS, east of the Rockies, had a 36% chance or greater. A larger difference between the two years is seen when comparing the probability of being in a Slight risk. Figure 44 shows that the coverage of the forecasted Slight risk was low this year, especially across the Central US. Additionally, unlike last year, it was rare for an area to have Slight risk forecasted over it multiple times. However, comparing the Operational and FFaIR EROs, one can see that the participants were more likely to draw a Slight risk than the forecasters at WPC for the 2022 FFaIR season.

3.4.1 CSU MLP and FFaIR EROs

As stated in Section 2.4.2, the participants were asked to rate the quality of the various versions of the ERO. Figure 45 shows the distribution of scores during FFaIR. Similar to previous FFaIR Experiments, the FFaIR ERO routinely scored higher than any of the MLP EROs, with an average score of 6.77, receiving a score of 7 or greater roughly 61% of the time. For reference, 38% of the 00z

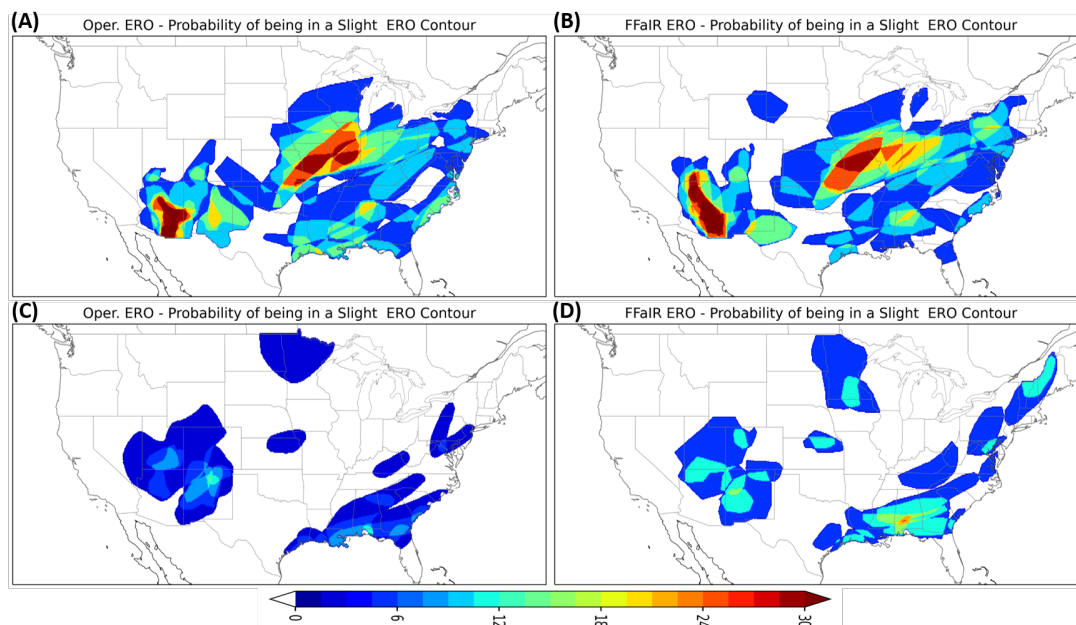


Figure 44: Same as Fig. 43 but for the Slight risk.

UFVSGEFSR, 31% of the GEFSO, and 26% of the 00z FV3GEFSR scores were 7 or higher. Of the CSU MLP EROs, the UFVSGEFSR ERO was the most preferred. The UFVSGEFSR ERO 12z forecast had a slightly higher average subjective score than the 00z forecast, 5.9 vs. 5.83 respectively. However the 12z forecast had a wider distribution of scores, receiving scores of 1 and 10 whereas the 00z UFVSGEFSR ERO forecast did not receive either of those two scores. This could suggest that the “goodness” of the 12z is less consistent than the 00z forecast. A similar pattern was seen when comparing the 00z and 12z FV3GEFSR. The 12z forecast had an average of 5.64 and the 00z forecast’s average was 5.41. However, the low/high end scores of the two different initialization times differ less than was seen for UFVSGEFSR, though the 12z forecast did receive more scores of 1 than the 00z forecasts.

Interestingly, although participants seemed to prefer the new versions (FV3GEFSR and UFVSGEFSR) of the GEFS-based EROs to the operational version (GEFSO), the subjective scores tell a different story. The GEFSO, which was only available at 00z, had an average score of 5.53 and was slightly higher than the average score for the 00z FV3GEFSR (5.41). It is possible that during open discussion participants

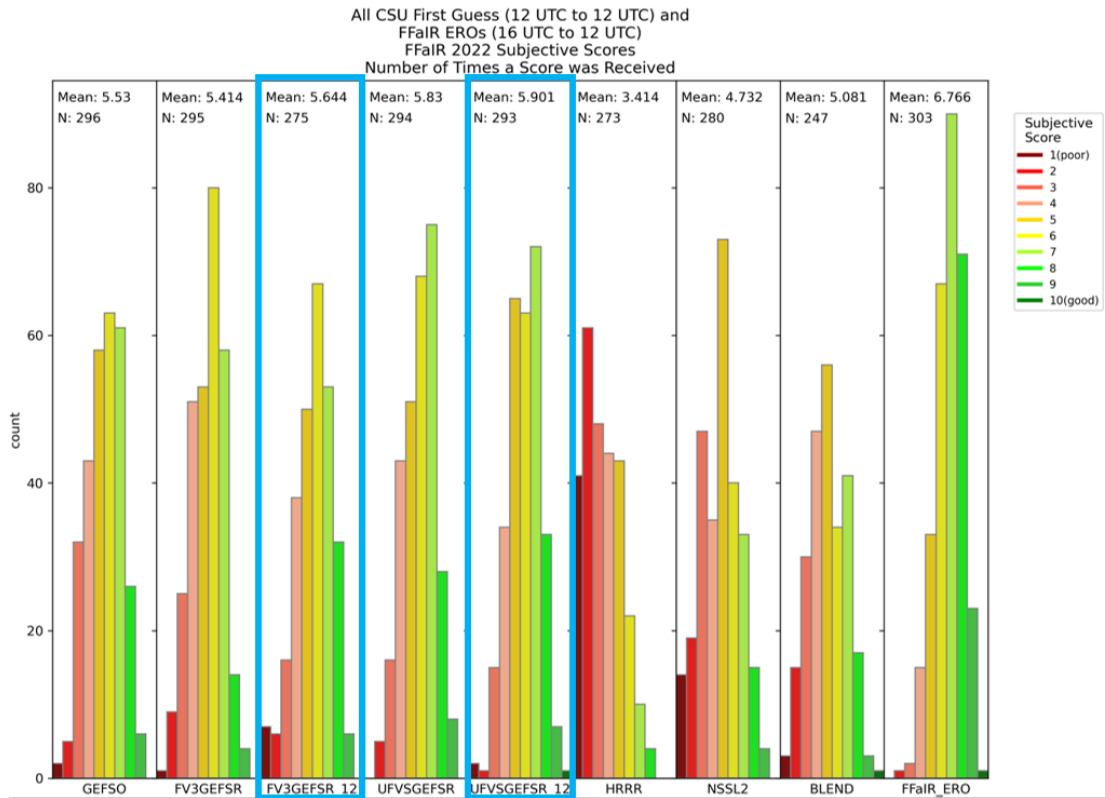


Figure 45: Similar to Fig. 21 but for the CSU MLP (valid 12-12 UTC) and FFaIR EROs (valid 16-12 UTC). Boxed in blue are the CSU MLP EROs that are initialized with 12z model data; all others are initialized from 00z. Along the top is each forecast’s mean and number of scores received over the duration of FFaIR.

were not specifying what cycle of the FV3GEFSR they were comparing to the GEFSO. However, in the written feedback there was a clearer distinction between the 00/12z cycles for the FV3GEFSR and UFVSGEFSR. In the written comments for the question: “Were there noticeable differences between the GEFS EROs that used the old training for heavy rainfall and the GEFS EROs that use UFVS?”, three overarching themes were seen: 1) Due to the lack of extreme rainfall it was hard to judge the ERO performance. 2) Often there were few differences seen between the three versions of the GEFS EROs, especially when comparing GEFSO and FV3GEFSR. 3) The UFVSGEFSR looked smoother, which participants liked, than the other two versions.

To highlight these points, consider the verified Slight risk day shown in Fig. 46 and evaluation scores listed in Table 3. The majority of the comments mentioned

Table 3: The participant “goodness” scores for the CSU MLP GEFS EROs and FFaIR ERO valid at 12 UTC 13 July 2022. Last row is the average score for each ERO on this day. Refer to Fig. 46 to see the valid EROs with verification.

Participant	GEFSO	00z FV3	12z FV3	00z UFVS	12z UFVS	FFaIR
1	8	7	8	7	9	7
2	6	6	7	6	7	8
3	5	4	5	4	5	7
4	3	3	6	3	6	7
5	4	3	5	4	5	6
6	7	6	7	7	6	8
7	3	2	5	4	7	6
8	5	5	6	6	7	6
9	5	4	6	5	6	7
10	4	5	6	4	5	6
11	4	3	4	4	5	6
12	6	4	7	5	8	6
13	4	4	5	5	4	7
14	4	2	5	2	5	7
15	4	5	7	6	7	8
16	3	2	3	3	3	4
17	5	5	7	5	7	7
18	4	4	6	5	4	7
19	7	7	7	7	7	6
Average	4.79	4.26	5.89	4.84	5.95	6.63

that all the CSU MLP GEFS EROs looked similar to one another. Several participants noted that the UFVSGEFSR had a slight advantage over the other CSU MLP EROs. One participant commented, “The old GEFS was much more conservative and ended up under estimating ERO rates.” Another stated “The ones with the old training were too “high-res” with their ERO outlines. The UFVS more broad brush effect seems to cover the heavy rain areas better. However, the 12Z UVFS was way overdone with its areas (too large).”

All of the CAM versions of the CSU MLP EROs were less preferred than the GEFS versions. The BLEND had an average score of 5.08, followed by the NSSL2

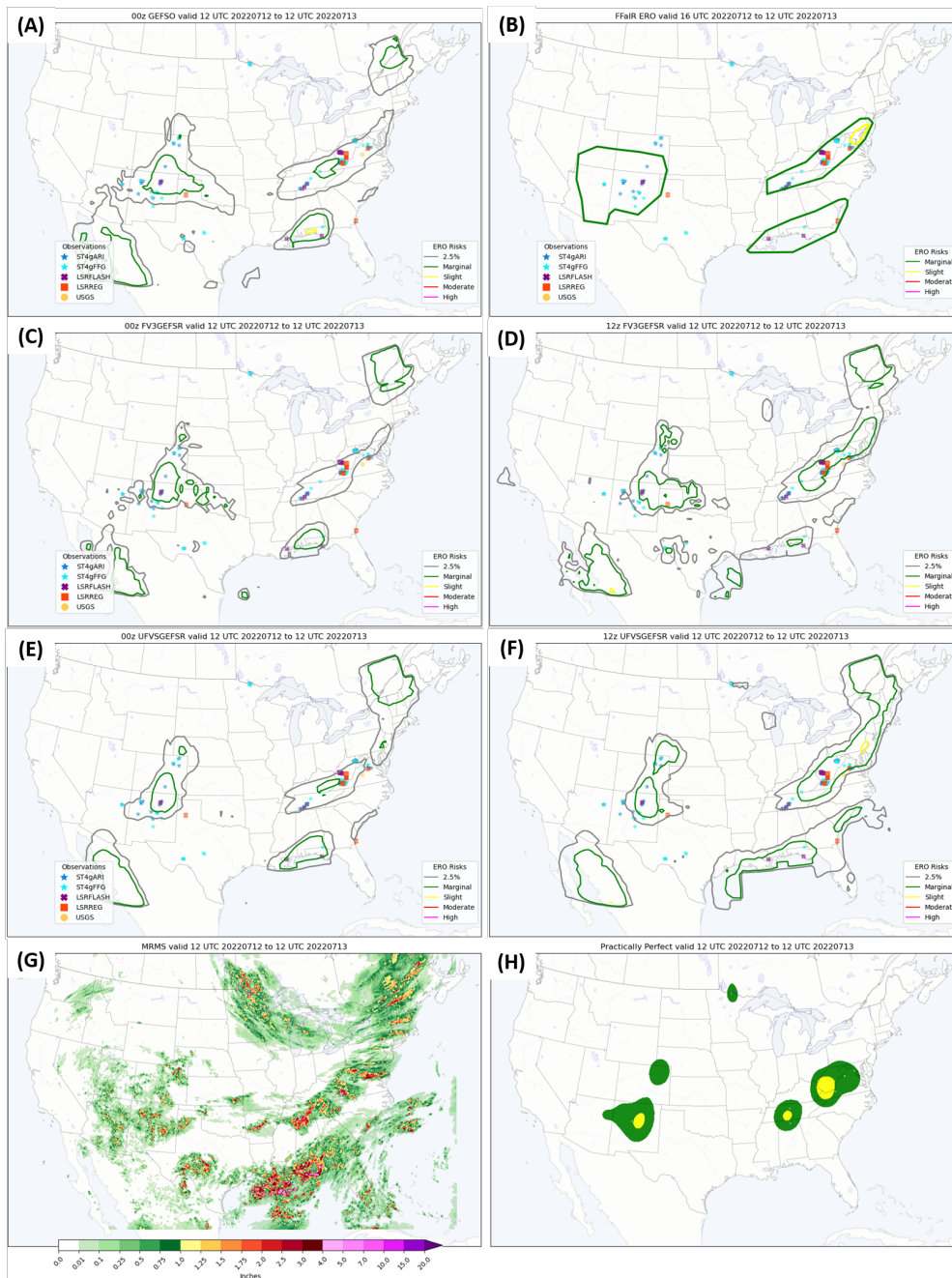


Figure 46: (B) FFaIR ERO valid 16 UTC 12 July to 12 UTC 13 July 2022. (A) GEFSO, (C) 00z FV3GEFSR, (D) 12z FV3GEFSR, (E) 00z UFVSGEFSR, and (F) 12z UFVSGEFS EROs valid 12 UTC 12 July to 12 UTC 13 July 2022. The WPC ERO risk probabilities contoured [Marginal: 5% green, Slight: 15% yellow, Moderate: 40% red and High: 70% purple/pink]. The 2.5% probabilities are contoured gray on the CSU MLP EROs. The UFVS data points are overlaid on ALL the ERO images. (G) 24-h QPE and (H) practically perfect verification, valid 12 UTC 12 July to 12 UTC 13 July 2022.

with an average of 4.73. The HRRR had the lowest average score, 3.41. Looking at the distribution of its scores (Fig. 45), the HRRR ERO rarely received a score of 7 or greater. In fact, 71% of the scores for the HRRR ERO performance were 4 or lower. For comparison, the NSSL2 ERO saw 41% of the score in that same range, while the various GEFs EROs' had scores of 4 or lower between 18% and 29%. Over the course of the experiment, the participants routinely stated that the HRRR ERO was not helpful. They felt it struggled to forecast an excessive rainfall risk, but when it did, the Marginal risk areas were too large and noisy. Additionally, they noted that the HRRR ERO nearly always had a Slight risk over the waters south of the United States, which they felt made it difficult to look at. Since ERO forecasts do not extend over the oceans and therefore is not verified, one way to address this issue would be to mask out the ocean when plotting the HRRR ERO for verification. This non-scientific fix will be discussed with the CSU team. Some comments from the participants that summarize the overall feedback were:

“What is going on with the HRRR? Very overforecasted over a very large area, and underforecasted some areas where there were reports.”

“HRRR seems to always handle this poorly, only highlight areas outside of the CONUS (i.e. Pacific Ocean, Mexico, Gulf of Mexico and western Atlantic).”

“The HRRR was way too broad to the point of being nearly unusable.”

“HRRR: All I can say is yikes. That level of false alarming is just way too much and not helpful.”

“Honestly, all did generally well. HRRR was the one that performed noticeably worse than all of the above.”

Figures 48 and 49 show the frequency of being in Marginal and Slight risks during FFaIR for the 00z CSU ML EROs; see Figs. 43C-D and 44C-D for the Operational and FFaIR ERO probabilities. Like the Operational and FFaIR EROs, the probability of being within a forecasted Marginal or Slight risk from any of the CSU ML EROs was low. The NSSL2 ERO was the most likely to forecast

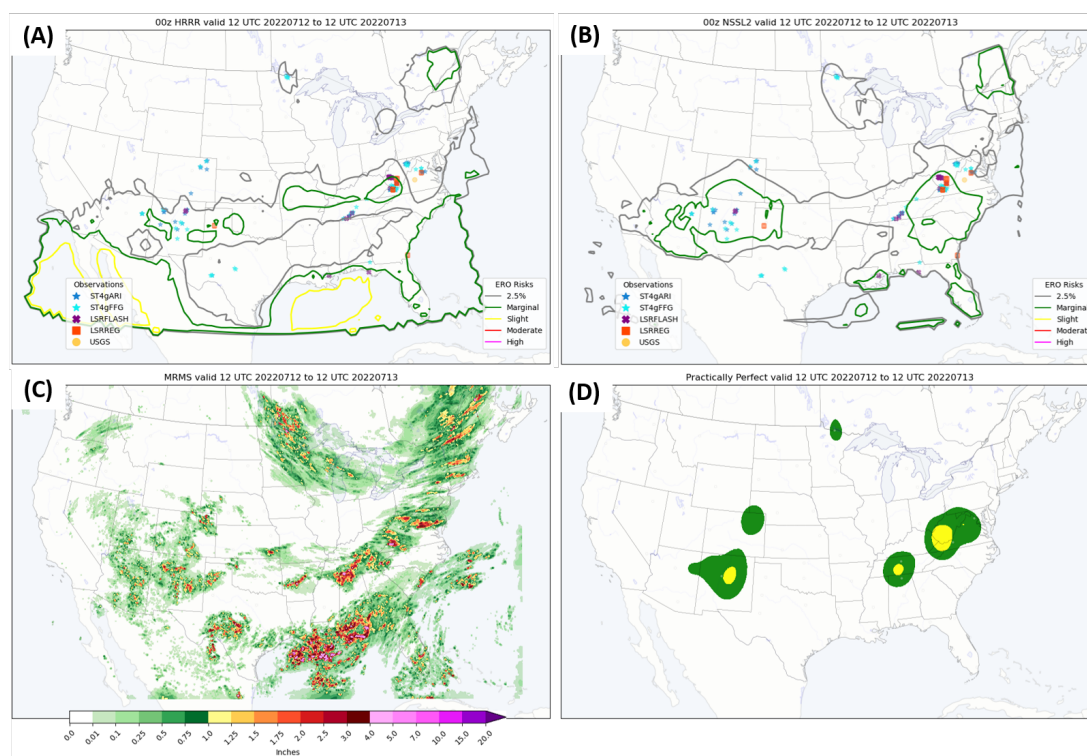


Figure 47: (A) HRRR and (B) NSSL2 EROs. The WPC ERO risk probabilities contoured [Marginal: 5% green, Slight: 15% yellow, Moderate: 40% red and High: 70% purple/pink]. The 2.5% probabilities are contoured gray. The UFVS data points are overlaid on ALL the ERO images. (C) 24-h QPE and (D) practically perfect verification. All valid 12 UTC 20 July to 12 UTC 21 July 2022.

a Marginal risk and more closely matched the Operational ERO. The highest probabilities were over the Carolinas and the Southwest. Similar maxima were seen in the Operation ERO, though the extent was slightly larger and the coverage of the higher probabilities were shifted to the northeast for both regions in the NSSL2 ERO. Despite being comparable to the Operational ERO for the Marginal risk, the NSSL2 ERO almost never forecast a Slight risk during the course of FFaIR.

Even though there was a clear preference to the GEFS-based EROs over the CAM-based EROs, statistically all the CSU ML EROs performed similar to one another during FFaIR. Figure 50¹² shows the Brier Score (BS) along side of the Area Under the Curve (AUC) Receiver Operating Characteristics (ROC) respectively. Only small differences in the BS can be seen among the CSU ML EROs,

¹²The CSU BLEND ERO is not plotted in these two images since multiple days were missing.

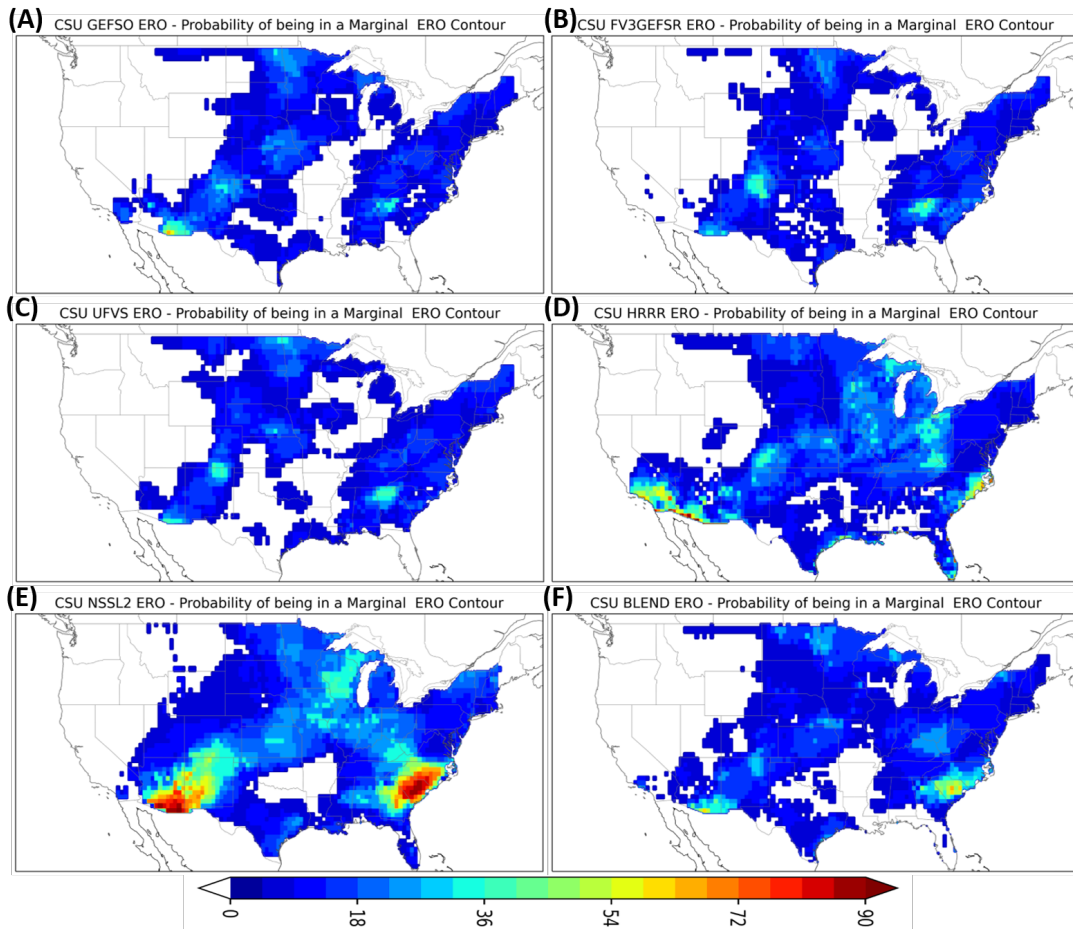


Figure 48: Probability of being in a Day 1 ERO Marginal risk during FFaIR 2022 for the CSU MLP EROs. (A) GEF50, (B) 00z FV3GEFSR, (C) 00z UFVSGEFSR, (D) HRRR, (E) NSSL2, and (F) BLEND.

suggesting that during FFaIR they all had about the same accuracy. How their accuracy varied day-by-day compared against the Operational ERO can be seen in Fig. 51, which further shows the similarity among the GEFS-based and CAM-based EROs. All of the CSU ML EROs seemed to struggle on Day 10 of FFaIR, which was the forecast ending at 12 UTC 02 July 2022. Figure 52 shows the the Operational ERO along with verification and the CSU FV3GEFSR, UFVS-GEFSR, and HRRR based EROs for this day. The 00z CSU ML EROs forecasted nearly no Slight risks across the CONUS but practically perfect identified 3 areas corresponded to a Slight risk. Additionally, unlike the Operational ERO, the CSU

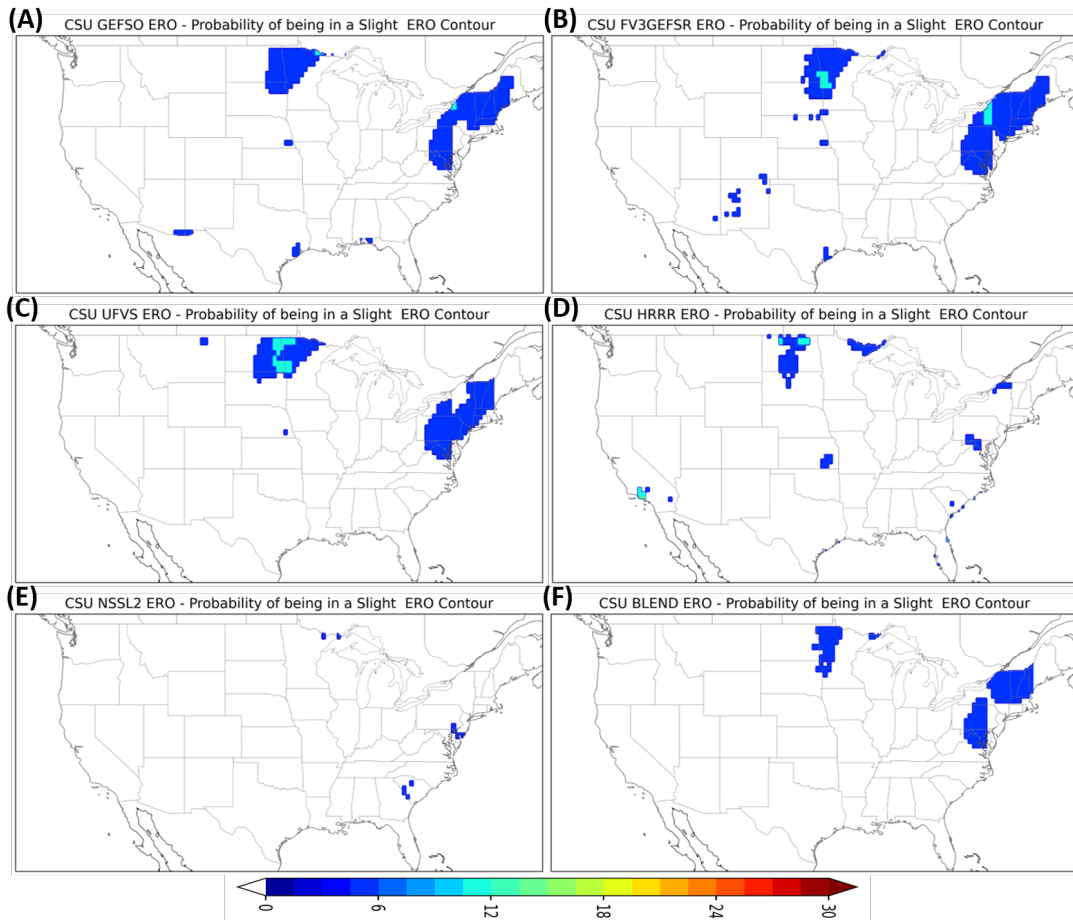


Figure 49: Same as Fig. 48 but for the Slight risk.

EROs failed to identify how broad the Marginal risk was across Colorado and New Mexico.

The AUC-ROC (Fig. 50B) has more variability among the EROs, though keep in mind that the differences are on the scale of a hundredths. Verifying again the UFVS (light blue), of the CSU ML EROs, the HRRR-based one had the highest AUC-ROC, followed closely by the GEFSSO. However, all of the ML EROs hover around the 0.5 mark, meaning overall, during the four weeks of FFaIR, the models had trouble distinguishing events from non-events. The unimpressive AUC-ROC scores were likely driven by the lack of spatially large events and the unusually dry spell that happened during FFaIR this year. This is supported by evaluating the performance of the NSSL2 during the 2020 and 2021 FFaIR

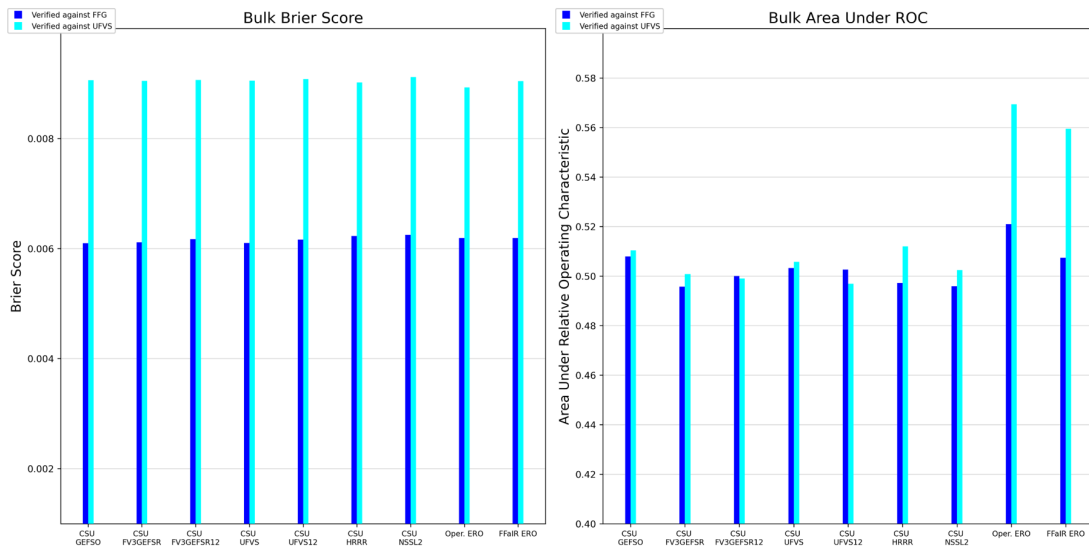


Figure 50: CSU, FFaIR, and the Operational EROs' (left) Brier Score (BS) and (right) Area Under the Curve (AUC) Receiver Operating Characteristics (ROC) verified against FFG exceedances only (dark blue) and the UFVS (light blue) across the 2022 FFaIR experiment days.

experiments. The NSSL2 configuration was first tested in the 2020 FFaIR and has not changed since. Figure 53 shows that in 2020 its AUC-ROC verified against the UFVS, was roughly 0.7 and in 2021 was about 0.8. The impact the abnormal FFaIR session had on ERO performance is also supported by the higher values seen in the performance of both the Operational and FFaIR EROs for both the 2020 and 2021 FFaIR Experiments¹³. Furthermore, comparison of the BS of these EROs from this year and the previous two FFaIRs show a similar trend. Another interesting difference between the three years is that in the previous two FFaIRs, the AUC-ROC scores were higher when verified against FFG than against UFVS for all EROs evaluated. However, for this year, the opposite was seen for almost all the EROs. This difference was seen most for the Operational and FFaIR EROs. Why this might be the case is beyond the scope of this analysis, but it is highlighted to once again emphasize the uniqueness of the 2022 FFaIR session to past ones.

Despite lackluster events during the FFaIR 2022 season, a few conclusions can be made for the CSU MP EROs. First, the new version of the GEFS-based

¹³This also applied to the CSU GEFSO, called CSU GEFS in 2020. However the GEFSO was not formally evaluated in the 2021 FFaIR Experiment.

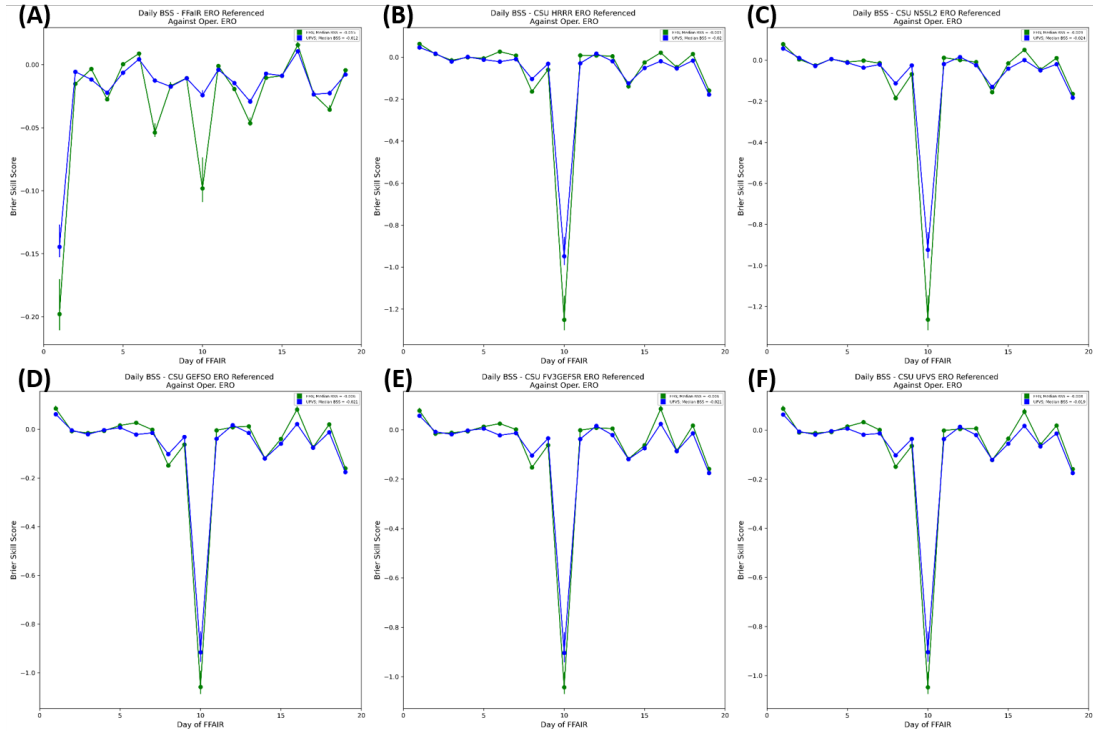


Figure 51: (A)FFaIR, (B) HRRR, (C) NSSL2, (D) GEFSO, (E) FV3GEFSR, and (F) UFVSGEFSR EROs' Brier Skill Score (BSS) referenced against the Operational ERO. Verified against FFG exceedances only (green) and the UFVS (blue) across the FFaIR experiment days.

ERO (FV3GEFSR) trained on GEFSv12 was comparable subjectively and objectively to the operational GEFS-based ERO (GEFSO). Therefore, transitioning the FV3GEFSR ERO to the operational CSU ERO is supported. Additionally, there was an overwhelmingly positive response to the CSU GEFS-based ERO that was trained using the UFVS dataset (UFVSGEFSR). It is recommended that the CSU team continue development of the system. Subjectively, the HRRR-based ERO did not outperform the NSSL2-based ERO and the objective verification was not particularly telling. Comparison of the BS and AUC-ROC to the version last year could suggest that the changes to the MLP hindered the HRRR ML ERO performance. However, as discussed, even the ML EROs that did not have changes between the two years saw a notable drop in performance when compared to last year's (and 2020's) experiment. Therefore it is difficult to say that the decreased

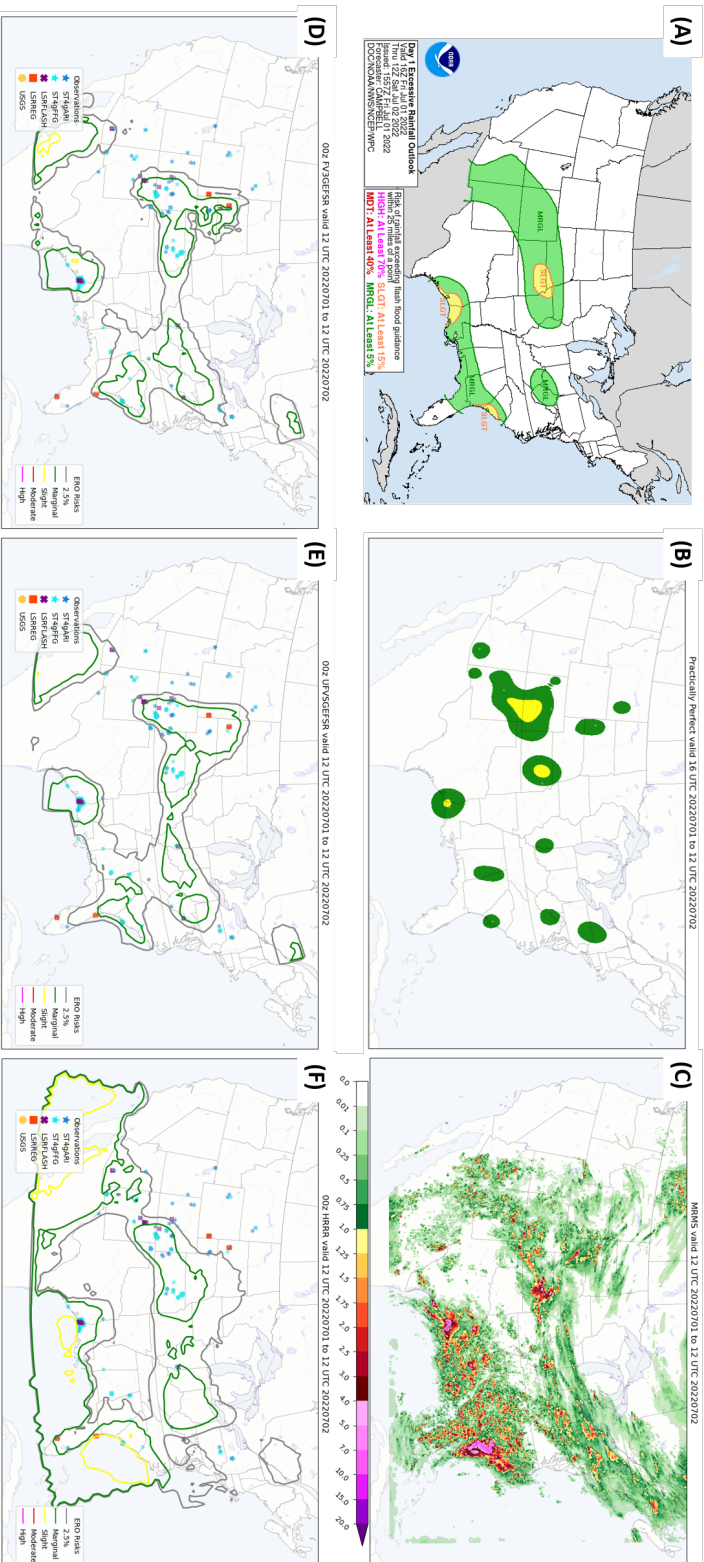
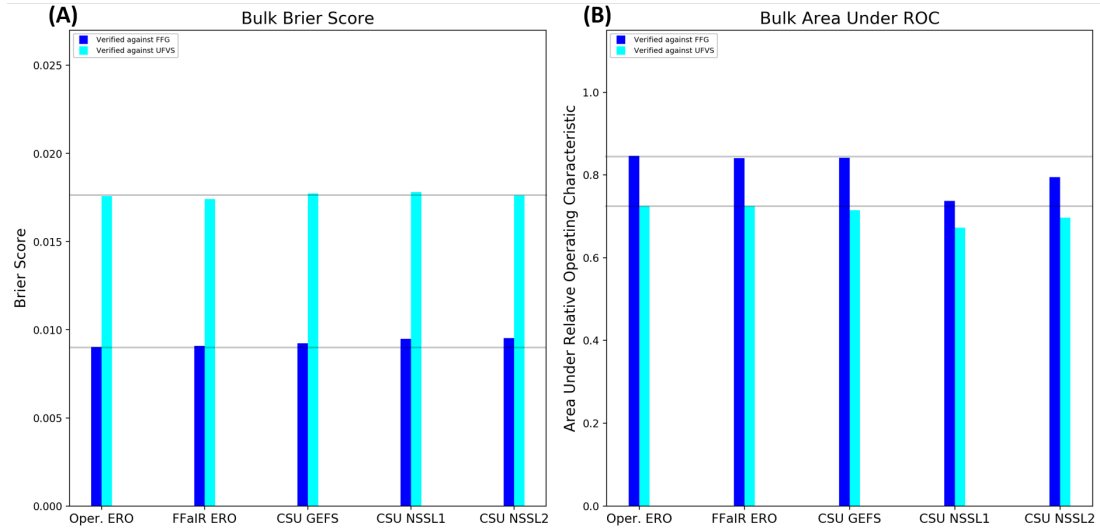


Figure 52: (A) Operational ERO valid 16 UTC 01 July to 12 UTC 02 July 2022. (D) 00z FV3GFSR, (E) 00z UFV3GFSR, and (F) HRRR EROs valid 12 UTC 01 July to 12 UTC 02 July 2022. The WPC ERO risk probabilities are contoured [Marginal: 5% green, Slight: 15% yellow, Moderate: 40% red and High: 70% purple/pink]. The 2.5% probabilities are contoured gray on the CSU MLP EROs. The UFVS data points are overlaid on ALL the ERO images. (B) Practically perfect verification and (C) 24-h QPE valid 12 UTC 01 July to 12 UTC 02 July 2022.

FFaIR 2020 Experiment



FFaIR 2021 Experiment

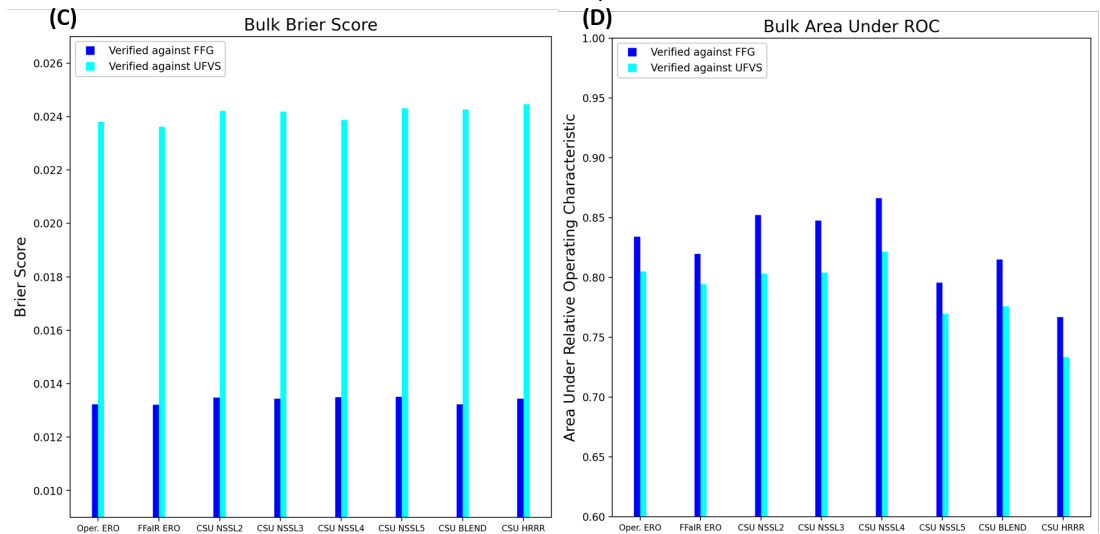


Figure 53: CSU MP, FFaIR, and the Operational EROs' (A) and (C) the Brier Score (BS) and (B) and (D) the Area Under the Curve (AUC) Receiver Operating Characteristics (ROC) verified against FFG exceedances only (dark blue) and the UFVS (light blue) for the (A)-(B) 2020 FFaIR and (C)-(D) 2021 FFaIR experiments. This figure is a combination of Fig. 48B in Trojnia et al. (2020) and Fig. 77B in Trojnia and Correia, Jr. (2021) with a few minor additions to add titles.

performance was a result of the updates to the system. Thus, the FFaIR team encourages further development of a HRRR-based MLP for the ERO.

3.4.2 FFaIR AERO

As discussed in Section 2.2, the AERO attempts to identify rainfall intensity as it relates to climatology rather than highlighting the coverage of the impacts of the heavy/extreme rainfall like the ERO does. Figure 54 shows how the differences in the ERO and AERO's methodologies can lead to differences in highlighted risk areas. For these Day 1 forecasts, valid 16 UTC 20 to 12 UTC 21 July 2022, both products identify similar areas for their lowest threshold; i.e. the Marginal risk and the 2-y 6-h ARI exceedance (both contoured green). The FFG valid for this ERO forecast and the 2-y, 5-y, and 10-y 6-h ARIs can be seen in Fig. 55.

Across the Southwest, the ERO group forecast a Marginal risk that encompassed most of the Four Corner states. Meanwhile, the AERO's 2-y exceedance contour did not cover the whole Four Corners region, but rather highlighted the heavy rainfall risk to CO and NM. Additionally, they chose to draw for higher ARI exceedances across south/central CO, identified by the black arrow in Fig. 54. Using the AERO methodology, participants were able to convey where the most intense rainfall was expected to occur. Various accumulation amounts during the valid forecast time period can be seen in Fig. 56.

Similar differences between the two products can be seen in the eastern US, which includes the Knoxville flooding event discussed in the introduction (see Fig. 7), with the lowest thresholds of each product having a similar look and spatial extent. Over this region, the participants in the ERO breakout group decided to introduce a Slight risk across TN and into the eastern portions of the Ohio River Valley (the hot pink arrow in Fig. 54). The AERO group also chose to highlight the risk that higher ARIs could be exceeded. The two areas with higher thresholds for the respective products overlapped, though the area drawn by the AERO group encompassed a smaller area and was focused across eastern TN.

The point of highlighting the differences between the two forecasts is not to suggest that one method is superior to the other. Rather, the goal is to emphasize

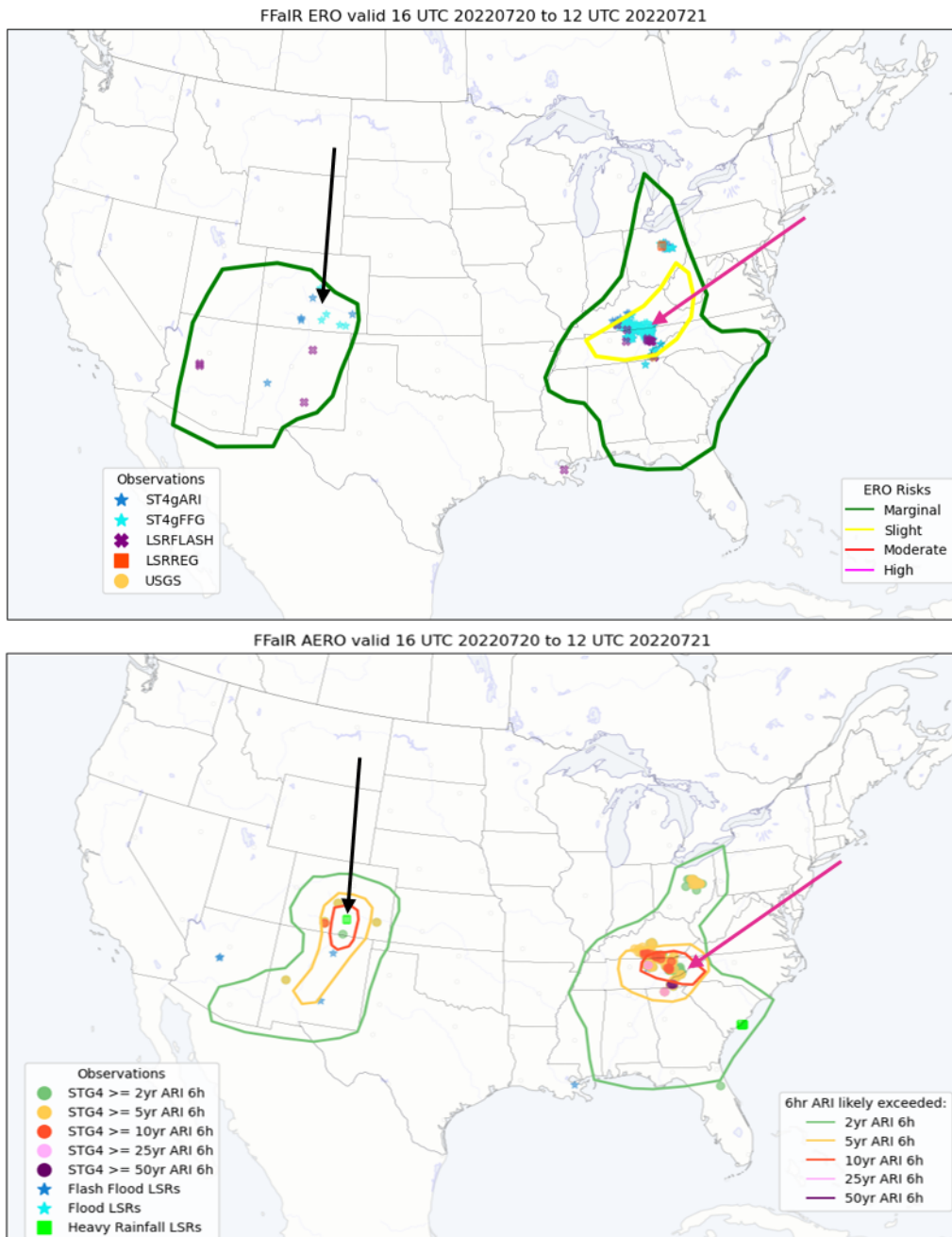


Figure 54: Comparison of the Day 1 TOP: FFaIR ERO and BOTTOM: FFaIR AERO valid 16 UTC 20 July to 12 UTC 21 July 2022. Plotted along with each are the observations used for their respective verification. These are the same forecasts shown in Section 2.4.2 Figs. 15 and 17, refer to their captions for observation descriptions and the observed MRMS for this day. The arrows highlight regions in which the differences in methodology for the two products are emphasized.

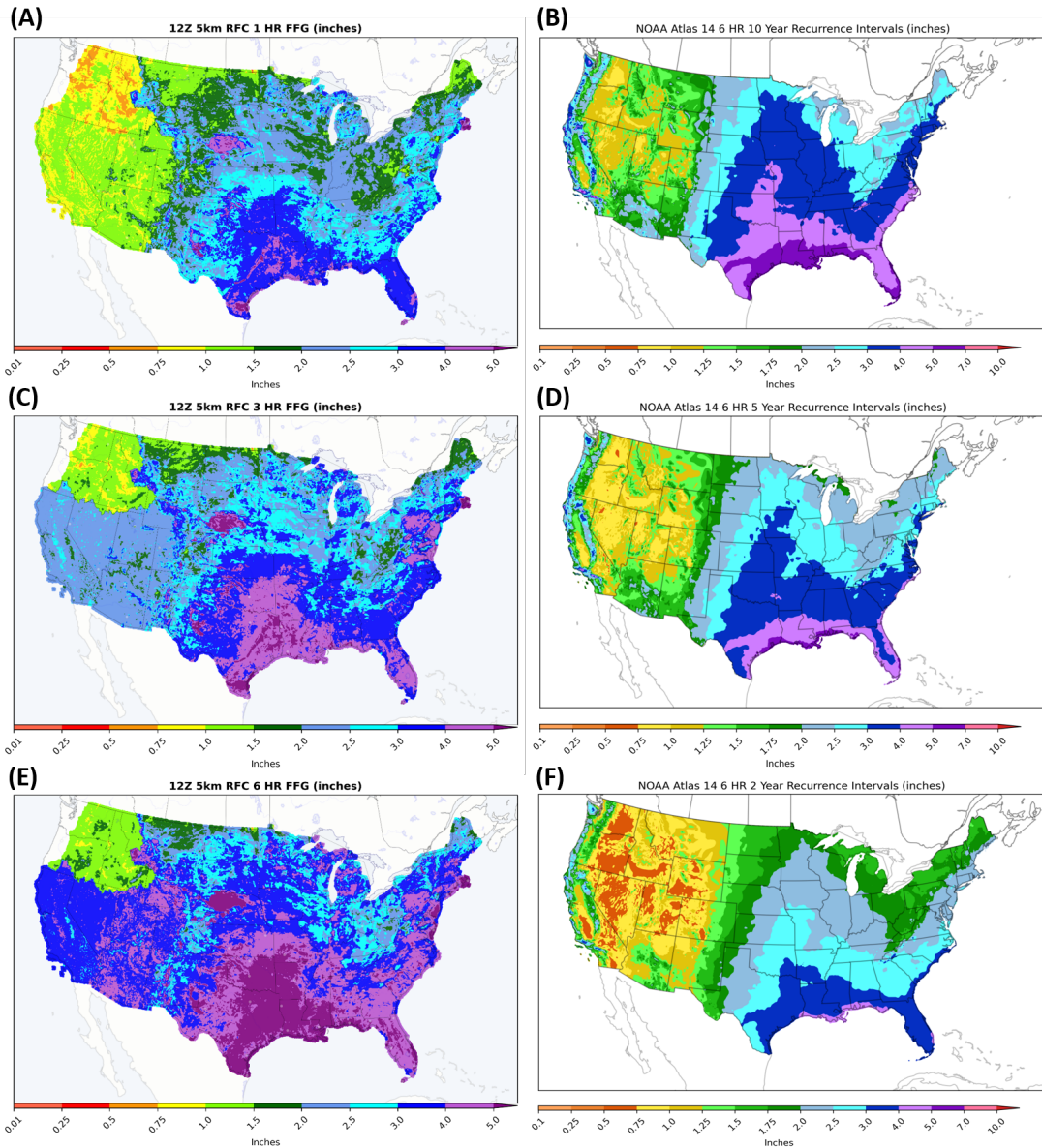


Figure 55: Left side from top to bottom: the 1-h, 3-h, and 6-h FFG issued at 12 UTC 20 July 2022. Right side from top to bottom: the 6-h ARIs for 2-y, 5-y and 10-y.

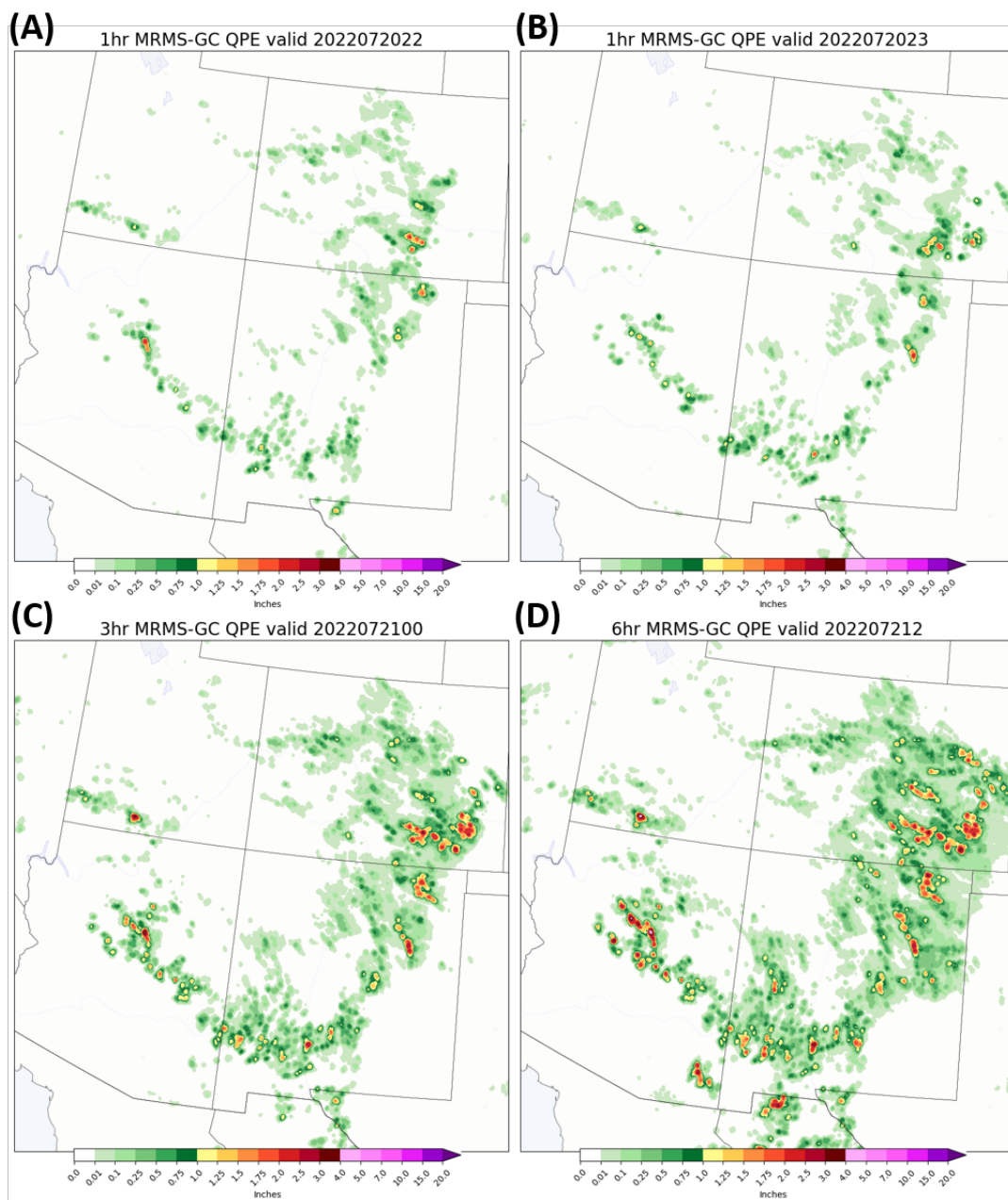


Figure 56: Example of some of the observed precipitation accumulation thresholds over the Four Corners region during the time period that the Day 1 ERO and AERO issued on July 20, 2022 were valid. 1-h MRMS QPE valid (A) 22 UTC and (B) 23 UTC 20 July 2022. (C) 3-h MRMS QPE valid 00 UTC and (D) 6-h MRMS QPE valid 02 UTC 21 July 2022.

the difference in utility and specificity despite both outlooks being used to identify excessive rainfall. Specifically, the ERO is used to identify the coverage of potential impacts from soil moisture, forecast rainfall, topography, etc. For instance, it is likely that the ERO group extended the slight risk into western WV since the region has many slot canyons and if the forecast rainfall fell in one of these areas there could be multiple flood events. On the other hand, the AERO does not take into account anything other than precipitation and its climatology. Nor does it try to identify the impacts exceeding some climatological threshold (in this case a given 6-h ARI) will have. **When creating the AERO, the interest is in highlighting what the maximum potential intensity could be.**

The subjective verification from the participants for the AERO can be seen in Fig. 57. On the left side of the figure is the distribution of the scores evaluating the forecast's utility. The distribution of AERO scores were similar to what was seen for the ERO (Fig. 45), though it has a longer tail on the lower score side. For comparison, the AERO received a score of 7 or better 48% of the time and a score of 4 or less 12% of the time vs the ERO's 61% and 6% respectively. It is possible that some of the lower scores resulted from lack of participant understanding of the goals of the AERO at the start of each week. Even though the forecast activity was explained in the operations plan (Trojniak and Correia, Jr., 2022) and reviewed during each week's orientation, during the first verification session participants that were not in the morning breakout room creating the AERO often stated that they were not sure they were evaluating the product correctly since they had not gone through the process of creating it yet. This sentiment is supported when reading through the written comments provided by the participants during verification. For example, one participant wrote "Would need more experience with the product to make that determination, but after a first try, I'm interested in using it again!" Another commented that they would "(c)an't really comment on this as I hadn't done the AERO forecast yet and don't think I truly wrap my head around it."

Since the AERO is attempting to identify where heavy/excessive rainfall will occur, participants were also asked "Focusing specifically on the LSRs, how well do you feel ARI exceedances matched up with reports of heavy rainfall?" This question

FFaIR AERO (16 UTC to 12 UTC) Overall and LSR-based
FFaIR 2022 Subjective Scores
Percent of Times a Score was Received

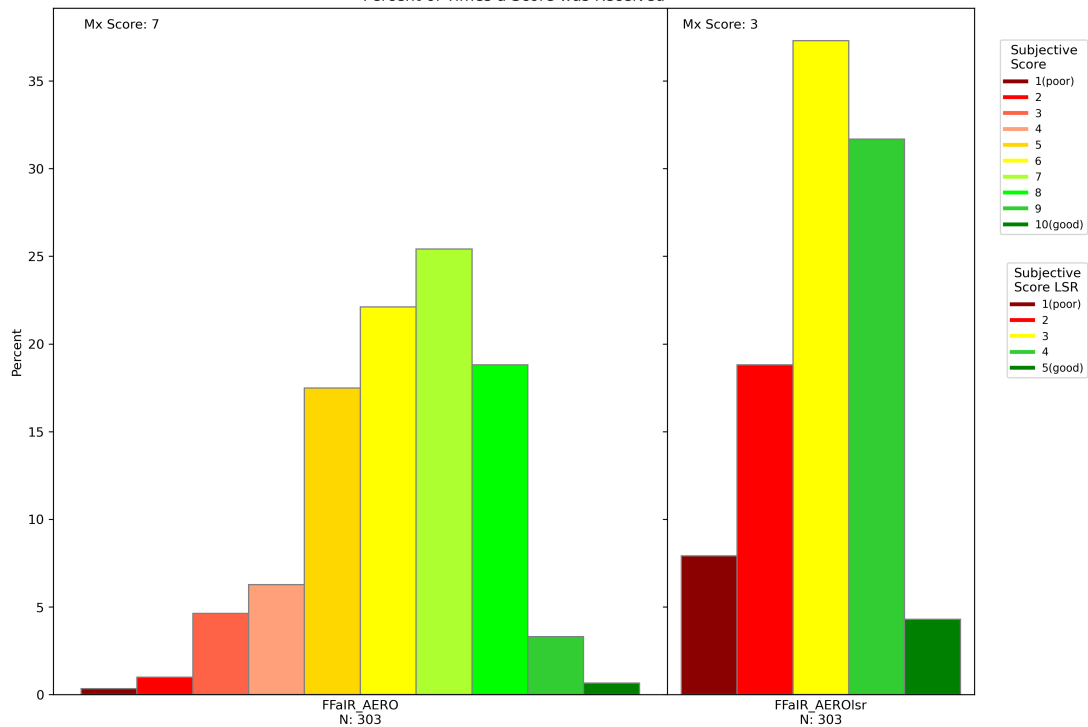


Figure 57: Results from the subjective verification of the AERO during the 2022 FFaIR Experiment showing the percent of the time it received a given score. LEFT: scale is from 1 (dark red) to 10 (dark green) and reflects the perceived utility or “goodness” of the forecast. RIGHT: scale is from 1 (dark red) to 5 (dark green) and reflects perceived correspondence of the AERO to heavy rainfall LSRs.

was on a scale of 1 to 5 and the results can be seen on the right of Fig. 57. To clarify, verification of the AERO did NOT include heavy rainfall LSRs, this question was asked simply out of curiosity of the FFaIR team. Specifically, the team was interested if there would be any correlation between higher ARI exceedances and heavy rainfall reports. The results suggest that some days correspond well and others not at all, which is perhaps to be expected given that heavy rainfall LSRs were scarce during FFaIR. However, participant written comments suggest that there was some misunderstanding of the question and rather than only comparing the heavy rainfall LSRs to the AERO, many participants discussed comparing flood and flash flood LSRs to the forecast product as well.

Like last year, the concept of a product forecasting ARI exceedances received positive feedback. One theme that emerged was that participants liked that the product was rooted strictly in precipitation exceedance and had less subjectivity in the drawing of contours than the ERO has. For instance:

“It’s a fun little product to provide an indication of how predicted rainfall will compare to climatology irrespective of impacts to infrastructure or human life. As opposed to the ERO, which includes the socioeconomic impacts, the AERO product keeps things strictly scientific, which is an important aspect of this experiment.”

“I think identifying extreme rainfall from a purely meteorological perspective is useful, in contrast to the current ERO, which relies on hydrological information from FFG. However, I am not convinced that Atlas 14 values are a good benchmark for this purpose.”

“I like the use of aero because it avoids people reporting things. Those can be missed but a gauge and rain totals for an area are very useful as a flood predictor.”

“Yes, it would be helpful. It is easier to verify, and highlights the relative rarity of events rather than the ultimate societal impact.”

“The AERO does provide a good “unconditional” outlook on heavy rainfall.”

In other words, participants seemed to like that the product was strictly grounded in exceedance of ARIs rather than the ERO, which has some amount of forecaster subjectivity to it given all the factors that can influence how the ERO risks are drawn.

Another theme was that although they liked the product, they felt it was most useful in conjunction with the ERO. For example, comments like these were common during verification discussion and in the written feedback:

“I think it is useful to identify places of heavy rainfall, which helps to be on the lookout for specific places that maybe the ERO does not present. ”

“In combination with considerations of antecedent conditions and land type, the AERO nicely highlight where anomalously heavy rainfall is expected to fall and may lead to impacts.”

“From an impacts perspective, I think it is quite useful to forecast the ARIs. These geographically thresholds of precipitation are far more meaningful and actionable than the static thresholds (1”, 2”, etc.).”

“I like the perspective it offers, though struggle with using it in isolation, i.e., completely replacing the ERO with it.”

“If the goal is to identify intense or anomalously high rainfall, then it can be useful. But I wouldn’t use it to predict flash floods due to lack of info on antecedent conditions.”

Again, turning to the example shown in Fig. 54, one can theorize how using the two products in tandem might be done. Focusing on the Four Corner’s region, where flash flooding is common during the monsoon season because of land type (i.e. desert and mountain desert), it does not necessarily take climatologically high rainfall totals to result in flooding. Because of this, when the monsoon is active, the desert southwest is often under a Marginal risk. Therefore it is difficult to identify if factors other than the typical risk associated with the monsoon were included in the reasoning for the issuance of a Marginal of the area. However, when combined with the AERO, the presence of the 5 and 10-y ARI exceedance would indicate that central New Mexico and south central Colorado should expect to see higher than normal rainfall totals and that the flash flooding associated with the monsoon moisture might be more widespread across this area than for the rest of the region that the Marginal risk encompasses. Furthermore, in relation to flash flooding and debris flow due to burn scars, highlighting what area is likely to see the heaviest rainfall could help them more strategically place resources.

Participants also noted that they would have liked a concrete probability defined for the AERO contours. Meaning they did not like that the definition of the AERO for drawing exceedance contours did not specifically state what the probability of exceedance was but rather just said “6-h ARI most likely to be succeeded.” Feedback about the lack of a probability being associated with the product was expected. As noted in Section 2.2, unlike in FFaIR 2021, this year’s experiment did not include a definitive probability of exceedance value because the FFaIR team wanted to “see” at what confidence level participants were willing to draw for and how the product would look if participants were given free reign to

draw for whatever probability they were envisioning in their head. It also allows the team the freedom to explore what the “look” of the contours should be; i.e. larger more inclusive contours like is typically drawn for ERO Marginal risks or smaller, more precise contours like are typically seen when Moderate or High risks are issued.

To evaluate at what probabilistic threshold participants seemed to be using when drawing the AERO exceedance contours, practically perfect (P-P) methodology was applied to observed exceedances of 6-h MRMS QPE for each ARI threshold. To remain consistent with the WPC ERO verification, analysis was done using a radius of influence (ROI) of 40 km and gaussian smoothing of 105 km¹⁴. Examples of the P-P verification for the Day 1 AERO for each ARI threshold along with the MRMS ARI exceedances plotted with an ROI of 40 km applied can be seen in Figs. 58 and 59 for 24 June and 01 July 2022; the observed MRMS seen in Fig. 2D and Fig. 3E. These two days were chosen because they represent two different scenarios. The June event had two distinct areas of heavy rainfall across the Northern Plains; these were two different systems. The July event had scattered, widespread exceedances from the Four Corners to Montana and across the the eastern US.

The proximity of the two heavy rainfall areas, and a smaller third area over the MT/WY border, for June 24 leads to the question, should the AERO contour for the 2-y 6-h ARI be two separate areas or one continuous area over the region? Depending on what P-P verification probability is used, depends on whether or not the two systems are encompassed by the same contour. An example of how the AERO 2-y contour might look using the probability value used to define the start of the Marginal (5%), Slight (15%), and Moderate (40%) ERO risks contoured in black can be seen in the left side of Fig. 60. If the AERO was defined as the 5% probability that the 2-y 6-h ARI will be exceeded than a large 2-y contour would verify, covering most the north, central US into the the Central Plains. However if 40% is the probability threshold, then the two areas discussed above are not

¹⁴Starting in the spring of 2022, WPC implemented a dynamic P-P, where the ROI and smoothing factor changes for each category. The ROI and smoother values were not changed for the Marginal risk P-P verification.

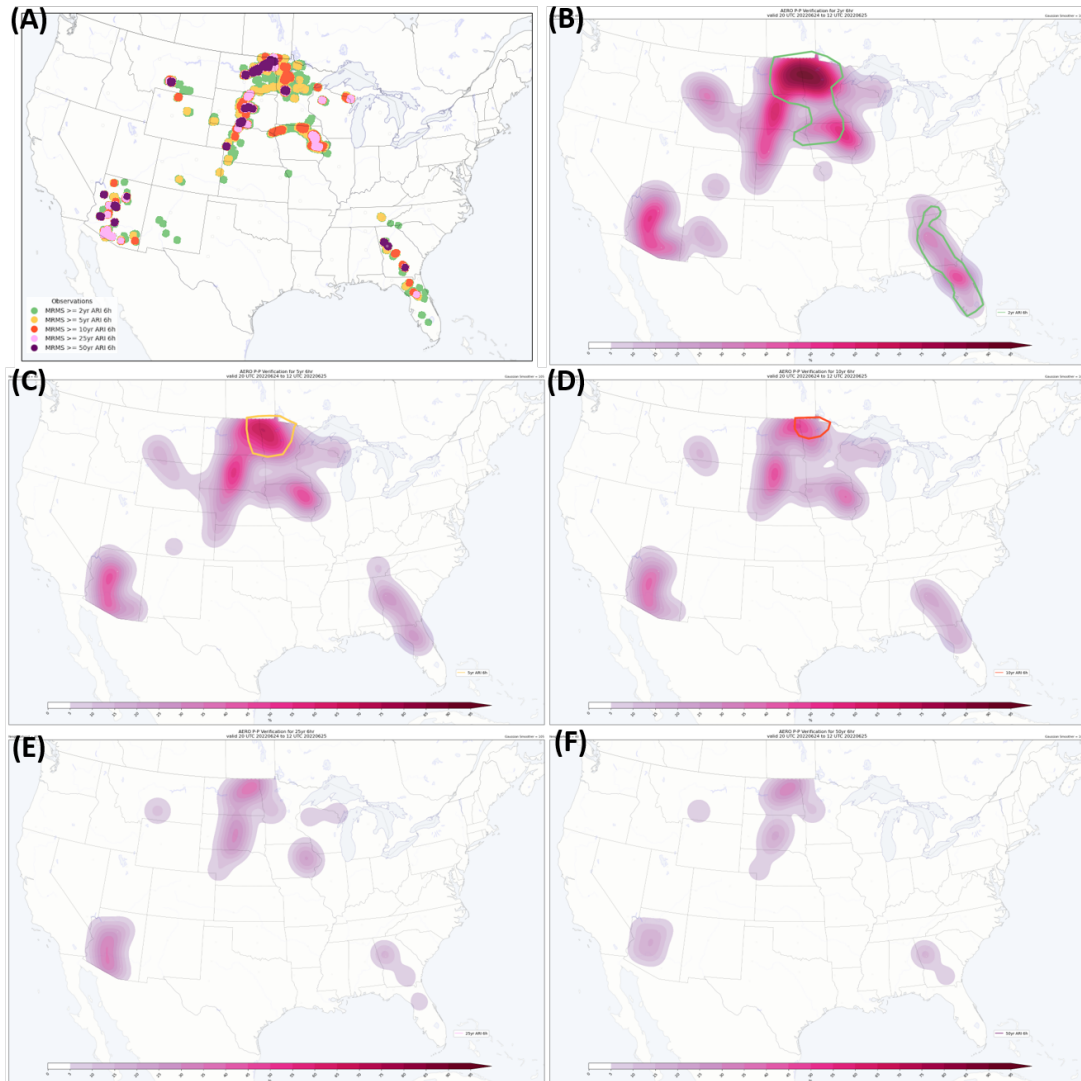


Figure 58: (A) Representation of 2-y (green), 5-y (yellow), 10-y (red), 25-y (pink) and 50-y (purple) 6-h ARI MRMS QPE exceedances when a 40 km neighborhood is applied. P-P using a 40 km neighborhood and 105 km gaussian smoother for the (B) 2-y, (C) 5-y, (D) 10-y, (E) 25-y, and (F) 50-y 6-h ARI. (B)-(F) If a FFaIR AERO contour was drawn for the given threshold, it is overlaid in its respective color. All valid 16 UTC 24 June to 12 UTC 25 June 2022.

encompassed by the same contour and the smaller, third area to the west would not be contoured at all. Furthermore, over the southeast only a small region in FL is encompassed in a contour while an area over GA, where exceedances of up to the 50-y ARI were observed, is excluded.

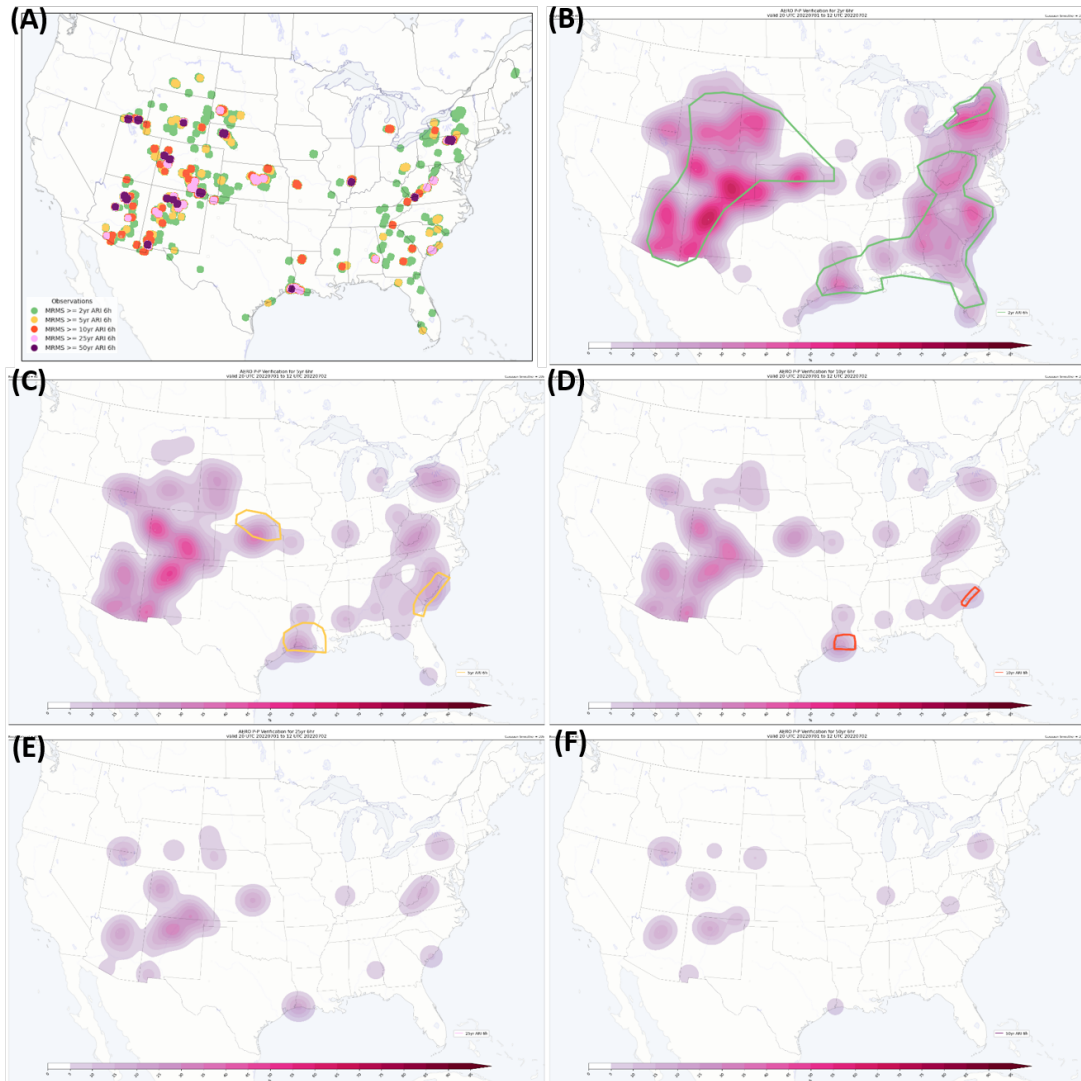


Figure 59: Same as Fig. 58 but for 16 UTC 01 July to 12 UTC 02 July 2022.

Similarly, Fig. 61 shows the P-P verification for the Day 1 AERO issued on 1 July 2022. Focusing on the 2-y P-P verification (left side) across the western portion of the CONUS, using 5% or 15% as the confidence level for the AERO on this day would have lead to a large 2-y contour similar to what was drawn by the participants. However, if a 40% probability of exceedance is used, the 2-y contour would be drawn to separate out the risk of exceedance across AZ, NM, and CO from areas further north and east.

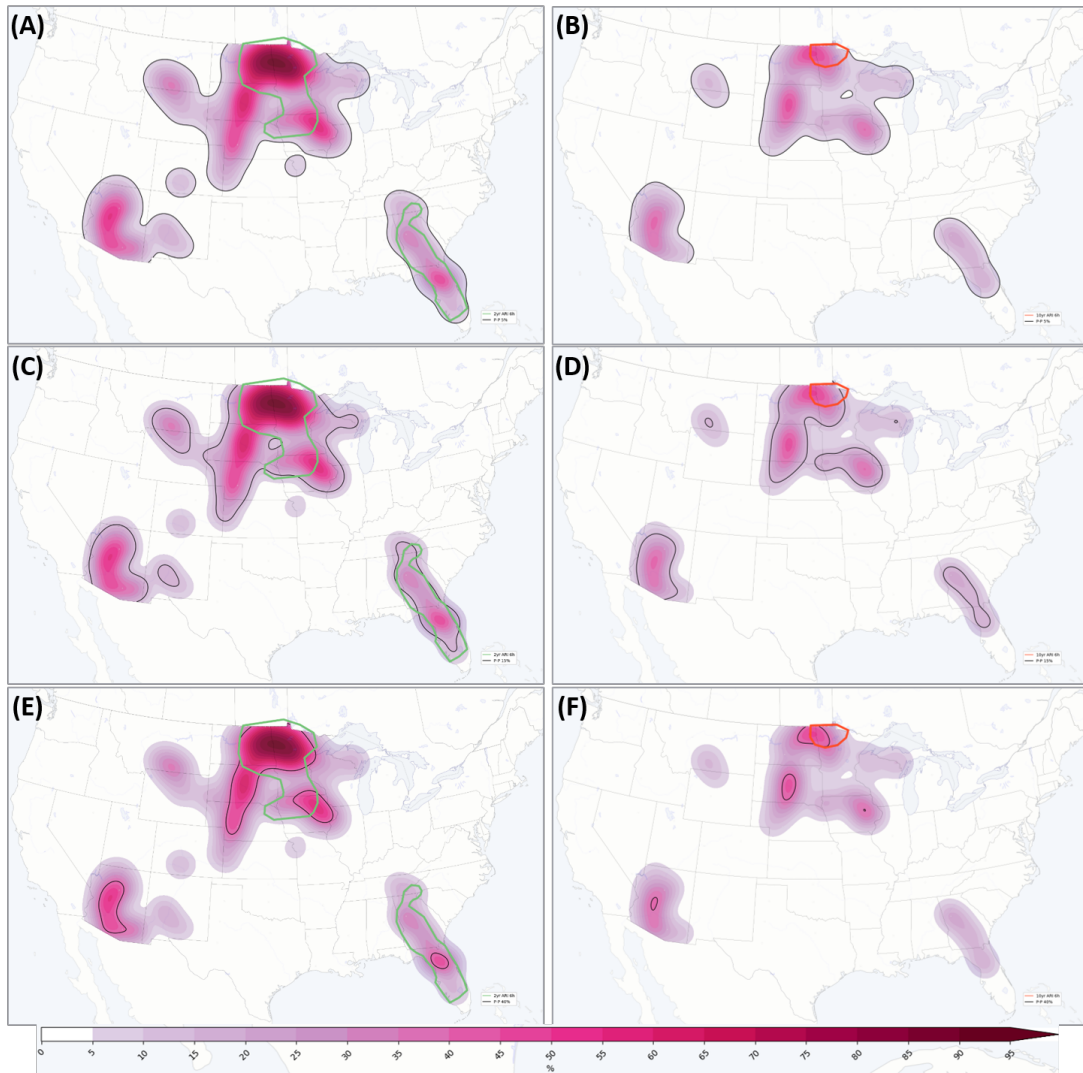


Figure 60: P-P using a 40 km neighborhood and 105 km gaussian smoother for the 2-y (left) and 10-y (right) 6-h ARI, valid 16 UTC 24 June to 12 UTC 25 June 2022. Contoured is black is the probability threshold (A)-(B) 5%, (C)-(D) 15%, and (E)-(F) 40%. If a 2-y (left) or 10-y (right) contour was drawn for the FFaIR AERO it is contoured in green or red respectively.

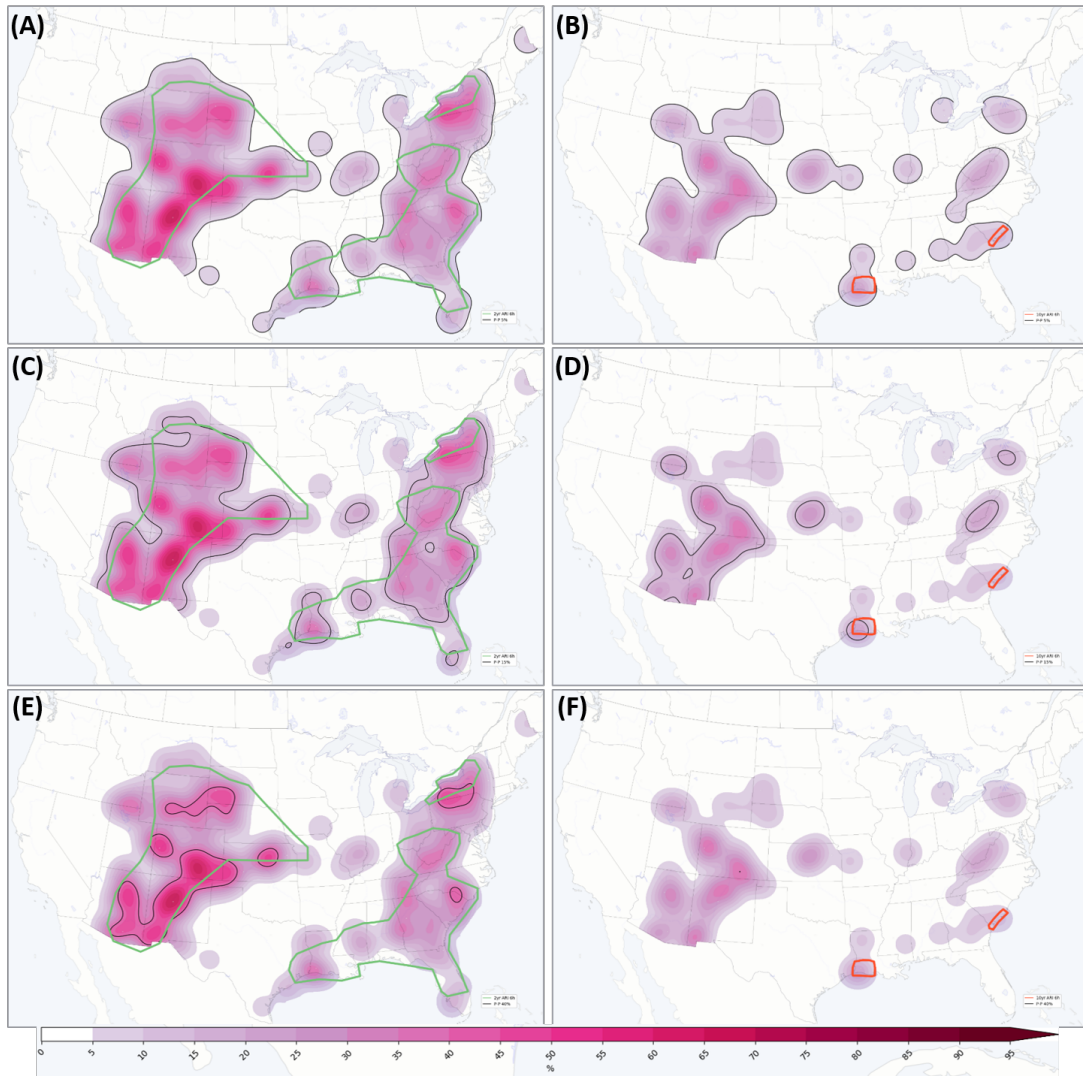


Figure 61: Same as Fig. 60 but for 16 UTC 01 July to 12 UTC 02 July 2022.

Another consideration that factors into choosing a probability for the AERO is related to coverage to some degree, though not necessarily in the way the coverage of flood reports is related to the ERO risk. More in the sense of observation proximity to each other and how extremely isolated occurrences (i.e. a single storm) should be handled. For instance, let's say that for the June 25 event it was decided that the two areas in the Northern Plains should be two separate areas in the AERO, then perhaps 40% should be used for the likelihood confidence works. However, this threshold doesn't convey the risk in south, where the spatial extent

of exceedance coverage over GA/FL is relatively small. Differing from this is the 1 July 2022 case, with widespread but numerous the reports over the western and eastern CONUS. Out west perhaps the 5% is the best choice since there it is difficult to discern breaks in the observances. But then you look across the east and southeast and the 5% P-P verification would suggest that a nearly continuous contour should be present across the eastern CONUS and the southeast extending into eastern TX. Which, some might argue, would create too much false alarm. In fact, on this day and on similar days, participants noted that they felt a large contour across the southeast for summertime storms was over kill. In both these cases, different probabilities might be preferred for the AERO on the same day depending on what type of risk is expected (ex. MCS vs Monsoon).

To further complicate determining what confidence probability would be optimal to encapsulate what the AERO is trying to convey, there is the question of whether or not the probability values should be consistent across all ARI thresholds of the AERO. For instance, perhaps a higher probability is needed for the lower ARI thresholds to reduce the false alarm but lower probabilities are needed for higher ARI exceedances. This however might lead to instances in which the verification of the higher ARI exceedances spatially is larger than the area encompassed by the P-P verification for the lower ARI exceedance. An example of how this could be the case can be seen by comparing the footprint for the 24 June case of the 2-y P-P using 40% as the confidence (Fig. 60E) to the 10-y contour using a 15% confidence (Fig. 60D).

In summary, participants like the AERO because it is grounded in precipitation only i.e. 6-h ARI exceedance rather than the EROs' usage of precipitation, antecedent ground conditions, and flood reports. However, they generally would rather use it in tandem with the ERO. Tandem usage of AERO with the ERO was surprising and the HMT will continue exploring the AERO approach to complement the ERO. Since this is an exploratory product, certain aspects of the product were loosely defined this year, specifically for what confidence probability the AERO contours should be drawn for. This was done to see what the AERO would look like if participants were not constrained by a strict definition, helping the FFaIR team better understand what the participants felt the product should

highlight or not. Although the ROI and smoothing were held constant, future development and verification work for the AERO will explore changing these values to calibrate the AERO to better convey the risk. The AERO activity will continue to be done for next years experiment and the team will likely test different definitions for the product to evaluate how best to utilize the AERO.

3.5 MRTP

The Maximum Rainfall and Timing Product (MRTP) was designed to have all participants draw multiple rainfall contours in a chosen 6 hour period in addition to drawing an area of six hourly maximum ARI. Additionally, participants answered questions about the amount and location of the maximum rainfall, flood probability, damaging flood probability, and the hourly maximum rain amount. 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.

The 2022 FFaIR was quite different than 2021, as shown previously in the FFaIR 24h rainfall difference between 2021 and 2022 in Fig. 1. This resulted in the MRTP 6 hour period forecast activity being comprised of mostly low in areal coverage but high accumulation maxima precipitation events, as shown by MRMS observations (Fig. 62). There were 10 days with maximum rainfall above 5" (compared to 7 in 2021) and all events had a footprint that was <90k km² (compared to 6 events above in 2021). This presented many challenges since most of these events were not synoptically obvious. The precipitation location, extent, duration, and maxima were typically seen as lower predictability on all but 3 days.

Flash flood warnings (Fig. 63) were present on 13 of the 19 days during the 2022 FFaIR Experiment, though only 5 days had broad coverage of warnings rather than isolated warnings. Looking at the seasonal Flash Flood warnings (May-July), the dates that FFaIR covered were low in June of both 2021 and 2022, while July 2022 had about half the number of warnings as July 2021. In total, 29 Mesoscale Precipitation Discussions were issued by WPC. For 17 of the 19 days, the MPD

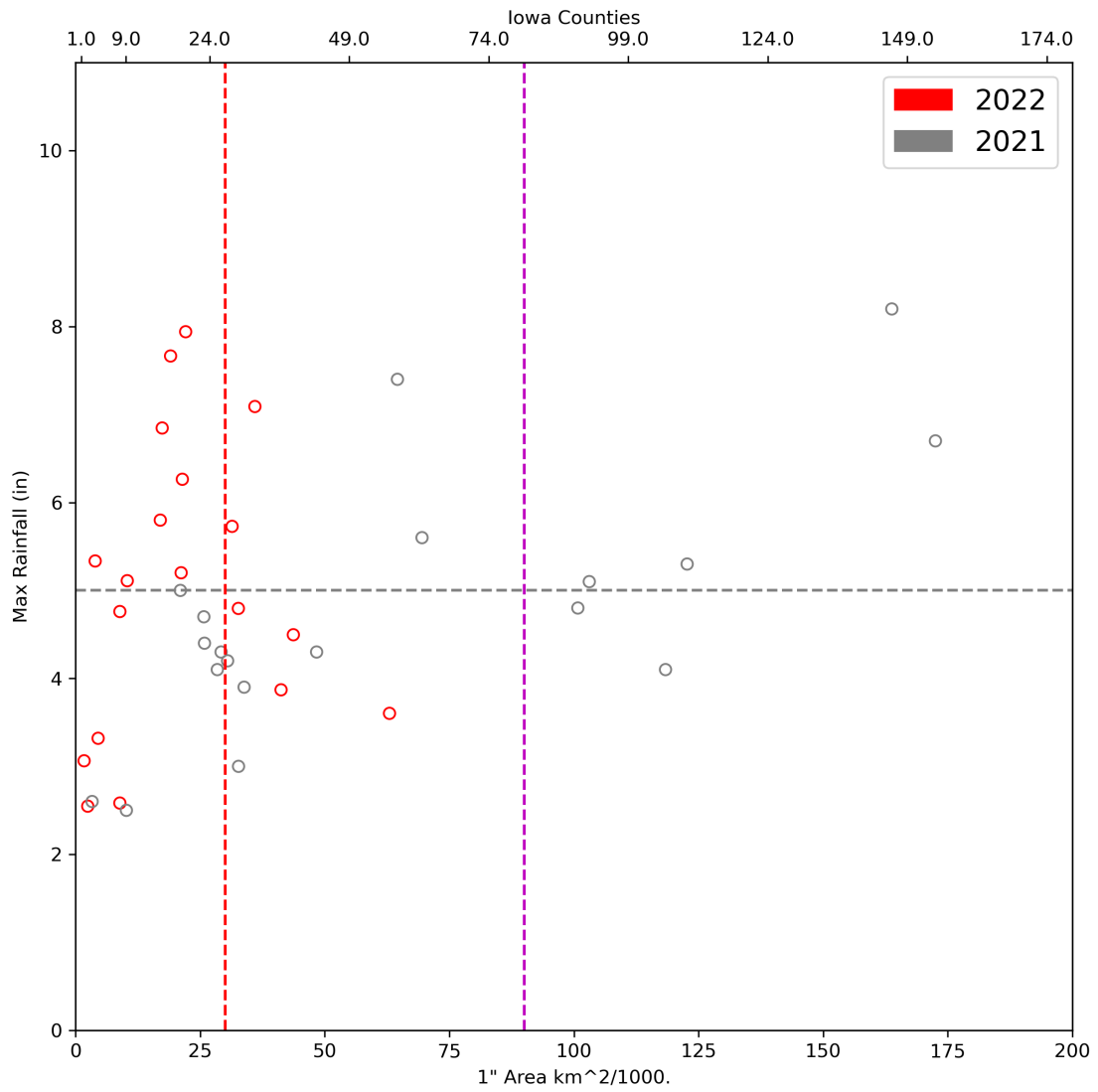


Figure 62: A comparison of the 2021 (20 days) and 2022 (19 days) MRTP events in terms of areal coverage of 1 inch and the maximum rainfall in the domain as determined by the MRMS. The red dashed line denotes 30k km² and the purple dashed line 90k km². The grey dashed line denotes 5 in.

issuance occurred in the MRTP domain and in close temporal proximity to the chosen MRTP time period (Fig. 64). In all cases, MRTPs were submitted prior to 21z, and the first available 6 hour window ended at 0300 UTC.

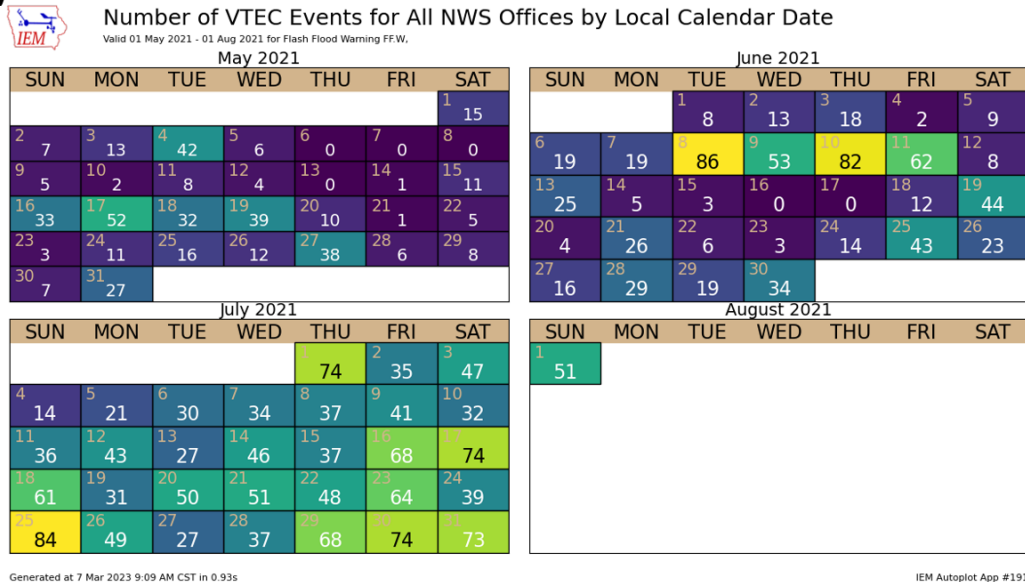
The ability of the participants to accurately pick a MRTP domain that encompassed the maximum 6-h rainfall or the maximum areal coverage of 1" area along with the correct valid time was evaluated. This was done by plotting the accumulated max rainfall and the areal coverage for various precipitation threshold over the MRTP domain and the CONUS. An example of what these plots look like can be seen in Fig. 65, if the red (CONUS) and blue (MRTP) lines intersect that means the maximum 6-h accumulation over the CONUS was located over the MRTP domain. The dashed yellow indicates the maximum 1" areal coverage. Using this method it was found that on 9 of the 19 days participants chose the time period where either the maximum rainfall 6-h accumulation (7) or maximum 1" areal coverage (2) in the MRTP domain occurred. On 5 days, the maximum 6-h rainfall occurred within 1 hour of the valid end time of the MRTP and on 9 days the maximum 1" areal coverage in the MRTP domain was within 1 hour of the valid end time. There were only 2 days in which the participants missed either of these metrics in the chosen time period; example can be seen in Fig. 65A. While there is much variation between maximum rainfall and maximum 1" areal coverage, the MRTP domains contained maximum rainfall equal to the daily CONUS maximum on 9 days and on only 2 days did the rainfall maximum fall below 60% of the daily CONUS maximum. The MRTP activity was effective in finding relevant areas at relevant times in the hunt for 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 problem.

3.5.1 Human and Model Performance

3.5.1.1 Performance Diagrams for Accumulation

Performance diagrams (Roebber, 2009) were produced for each day of the MRTP activity, showing the participants performance along with every cycle of

(A)



(B)

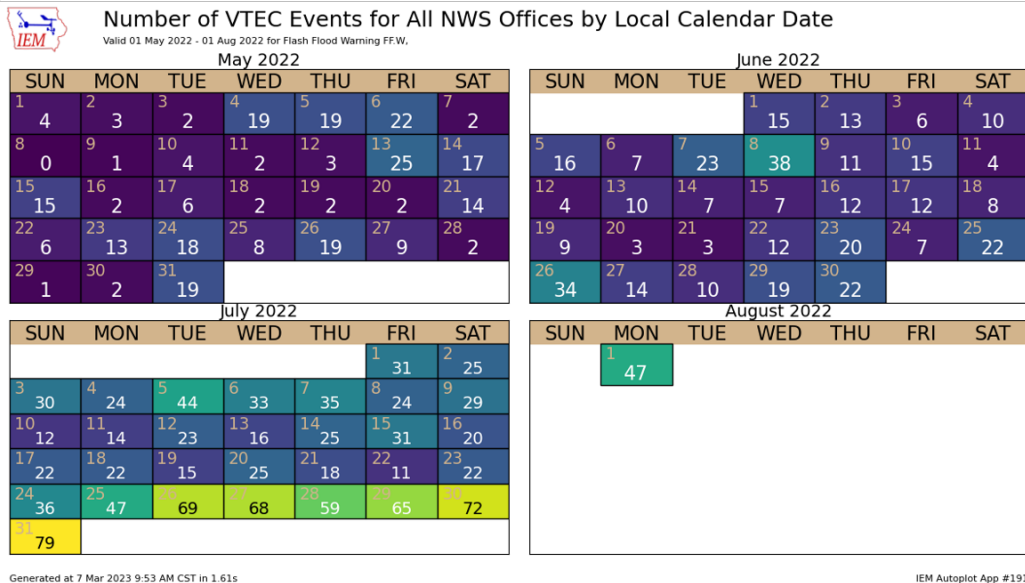


Figure 63: Daily Flash Flood warnings from (A) 2021 and (B) 2022 shown for each calendar day.

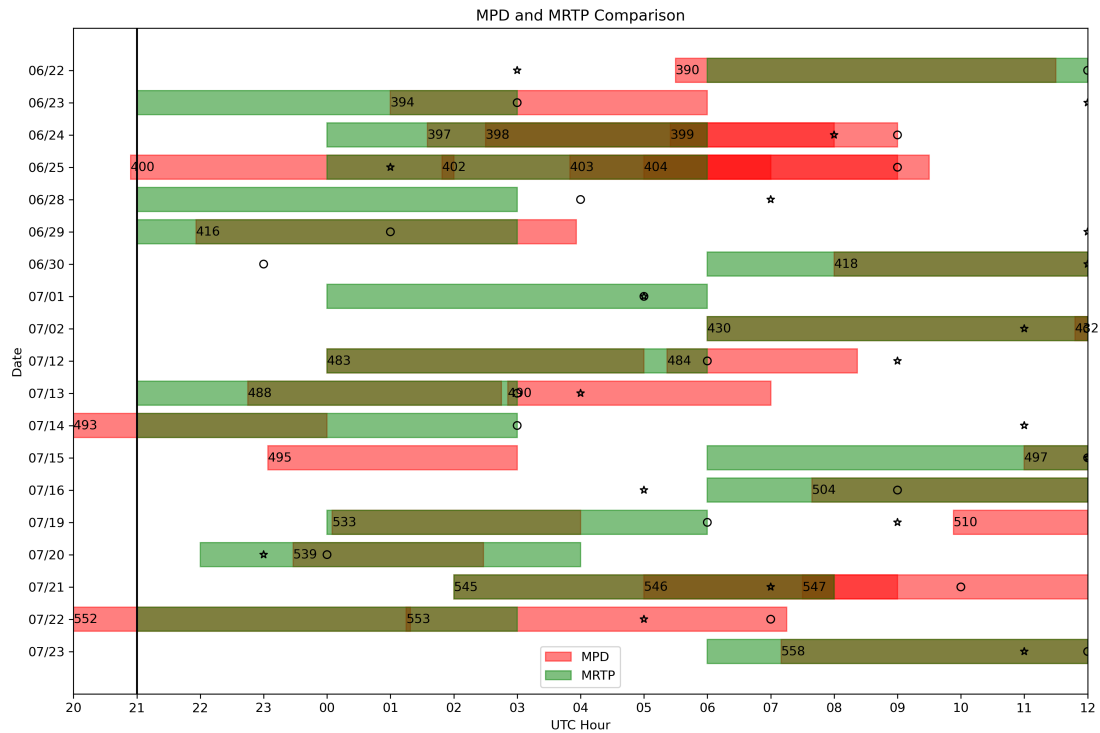


Figure 64: A comparison of the issuance times of the operational WPC Mesoscale Precipitation Discussions (red) and the timing of the MRTP activity time window (green). MPDs which cover the geographical domain are shown no matter the time of day. Also shown are the time of the maximum 6hr rainfall (star) 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.

model guidance that was available for the MRTP time period¹⁵ (Fig. 66). Participants, as in previous years, typically had with relatively higher POD and lower Success Ratios (SR), indicated by the blue dots up and to the left on the performance diagram. There is considerably more spread amongst this years participants, some being new to FFaIR, but most of the spread likely came from the struggle associated with the type of events that were forecast for; i.e. small and localized areas of heavy precipitation. On larger areal coverage days, participants mostly shifted to 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

¹⁵As many as 18 model cycles were available for the GFS model, running every 6h and verified out to 84h.

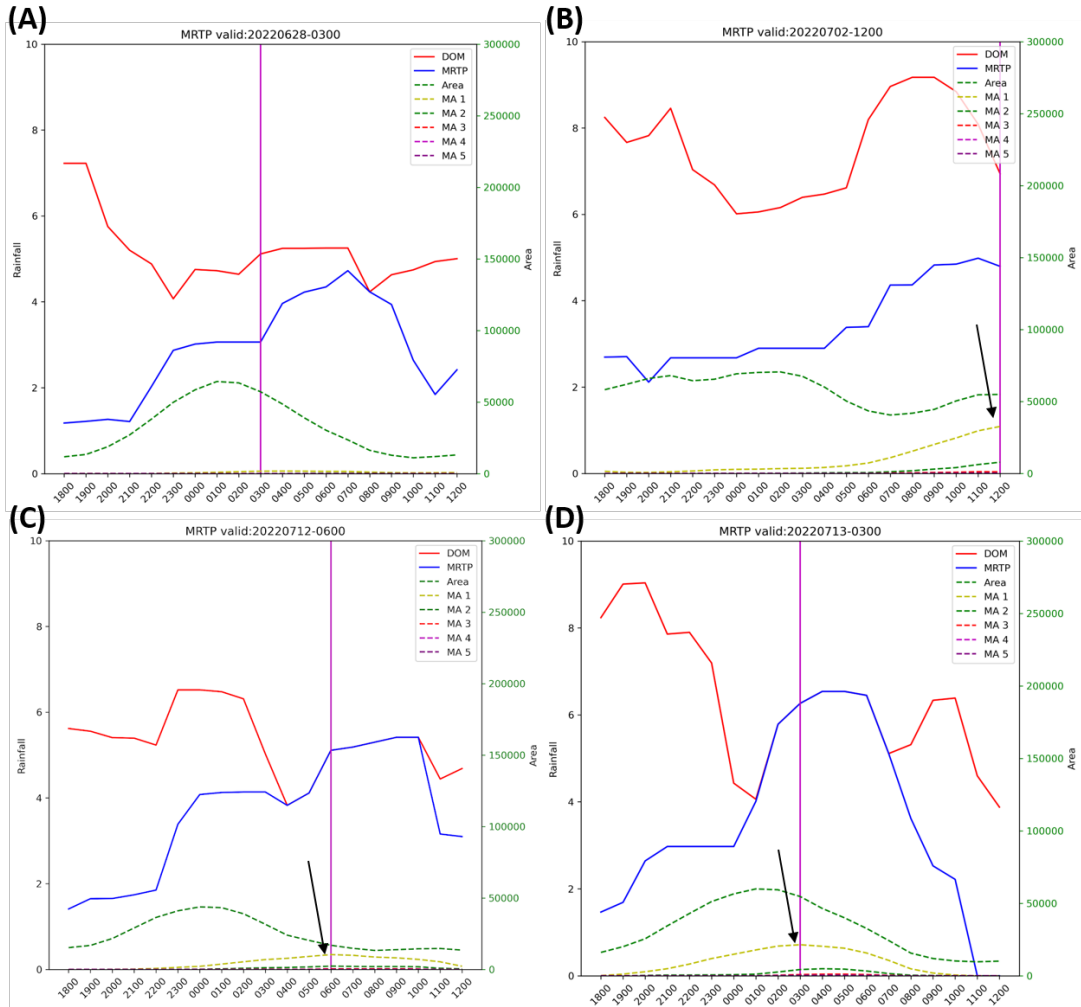


Figure 65: The maximum accumulated 6-h rainfall observed across the CONUS (red line) and over each day's MRTP domain (blue line). The areal coverage, in km², of 1" over the CONUS is dashed green. The other dashed lines are the areal coverage for 1" (yellow), 2" (dark green), 3" (red), 4" (light purple) and 5" (dark purple) within the MRTP domain. If a black arrow is present it is highlighting when the peak 1" areal coverage was within the MRTP domain. The vertical purple line is the valid end time of the MRTP. Valid time for analysis runs from 1800 UTC to 12 UTC. Analysis valid for MRTP issued on (A) 27 June, (B) 1 July, (C) 11 July, and (D) 12 July 2022.

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 guidance. On most days, CSI scores indicate that participants are equally competitive with short term model guidance just prior to the event, which participants were not able to use for their forecast.

Daily model performance indicates that Day 1 guidance is frequently the best guidance of the day, 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 participants forecasting approach where we sift through all the guidance looking for similarity and agreement, knowing full well that “all models are wrong but some are useful” Box (1979). Given the relatively smaller areas of extreme precipitation this season, a good portion of model guidance had frequency biases less than 1 on many days. Only on the days with $\geq 30\text{k km}^2$ 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.

Identifying extreme precipitation is normally a challenge but the spatial scale of this years events made model guidance skill even harder to achieve. While the areal coverage of events was distinctly different between 2021 and 2022, nearly half of the overall quality was driven by 5-6 days or 25% of the events with the other half of the quality driven by the remaining 75% of days. The main difference between the two years was the scale of the medium events, with 2022 having more, and the largest events, 2021 saw more of. Thus the outcome of lower quality in 2022 was related more to the lower predictability of smaller scale events.

Similar results were found by Griffin et al. (2022) when examining the half inch rainfall using the performance diagram. In previous years, the one inch contour was the lowest available for participants to draw. This year we added the half inch contour because a dominant strategy used by all participants was to draw a large encompassing contour. This strategy made interpreting the one inch results difficult since it was a manifestation of probabilistic thinking (encompassing where

it will rain heavily) and deterministic drawing. We will continue using the half inch contour to encourage precision in the one inch contour for better comparisons to models.

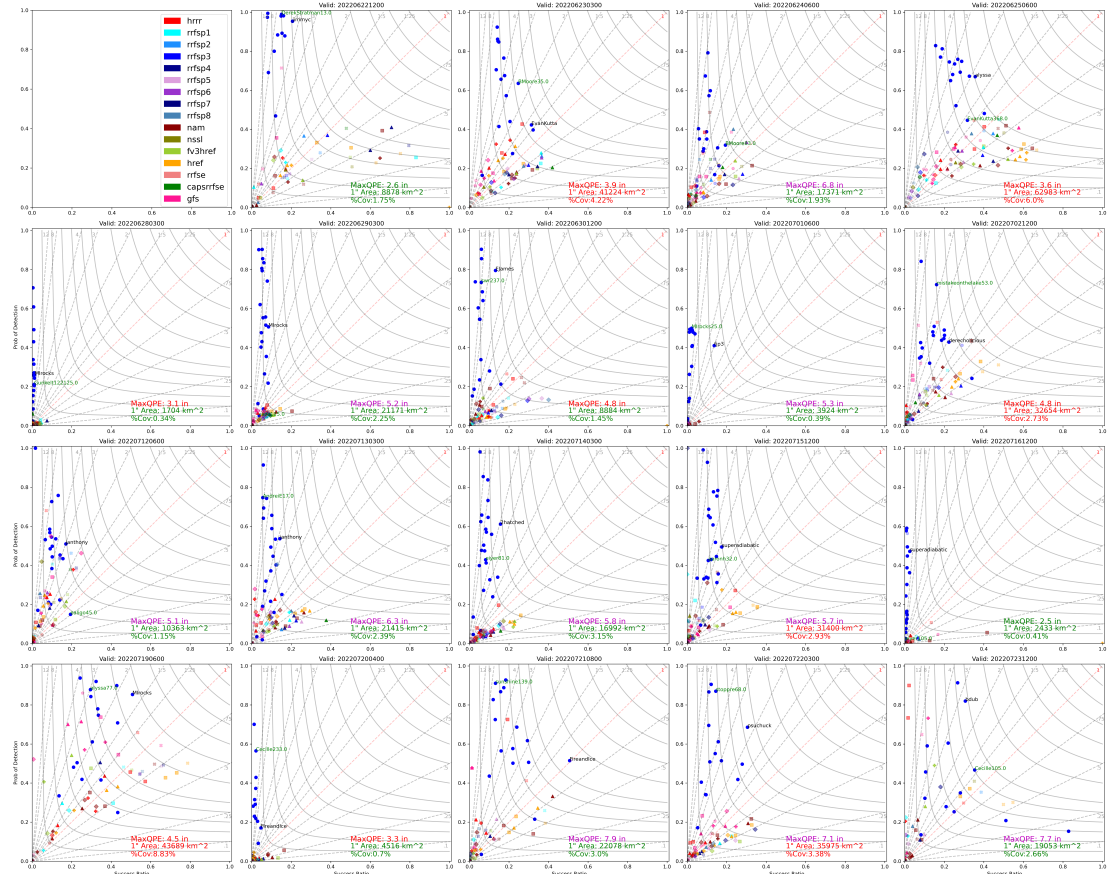


Figure 66: Each MRTP forecast days' performance diagram for a threshold of 1". Blue circles are participants while the other colors represent a model or ensemble, with each symbol a different time grouping. Lighter colors represent models at times that were not available to forecasters. For each day 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).

3.5.1.2 Distance and Maximum Rain Quality

A new diagram of performance was created to show how close, distance wise, participants were to the observed location of maximum rainfall and also the maximum value of rainfall in the MRTP domain (Fig. 67). The best quality of the

forecast are thus the observed maximum and a distance of zero (perfect). With models having 2.5 times more forecasts than participants, due to cadence and lead time, the smallest distance of the day was typically from a model (15 of 19 days). However, looking at the percentage of models that are within 75km, 6 days are near 0%, 11 are $\leq 5\%$, and 17 are below $\leq 10\%$. On 6 days, participants were not competitive with models for the lowest distance.

For maximum rainfall, most days saw predictions clustered around the observed maximum with the exception of 4 days where only a few models or participants were close. Participants fell at or below the maximum on 10 days, and were in close proximity on 7 days, only exceeding on 2 days. On only 3 days were participants modestly different than the clustered Day 1 guidance. Participant means were much lower on 7 days compared to the maximum observed rain. A composite error analysis across all days (Fig. 68) shows a -40% error peak for participants, similar to the model bi-modal peaks at -60 and -25% respectively. The cumulative distribution functions (Fig. ??) for participants has the same value at the participant relative frequency peak (-40%), with nearly 75% of the distribution below zero, while for models it is closer to 65%. For the most part, participants were aligned with modeled maximum rainfall.

In general, the results of the MRTP forecast activity were positive despite the difficult fine scale nature of the extreme events this year. Models and participants were competitive on multiple aspects of quality as shown in performance diagrams, distance to max rain and max rain itself. The MRTP domain frequently contained a WPC operationally issued MPD in close proximity to the time selected by the groups. Participants nearly always selected domains where extreme rainfall occurred and most often where the areal coverage of 1" rainfall was close to the largest for that convective day. The outcome of the MRTP activity is that participants successfully used model information to predict aspects of the extreme rainfall events in advance.

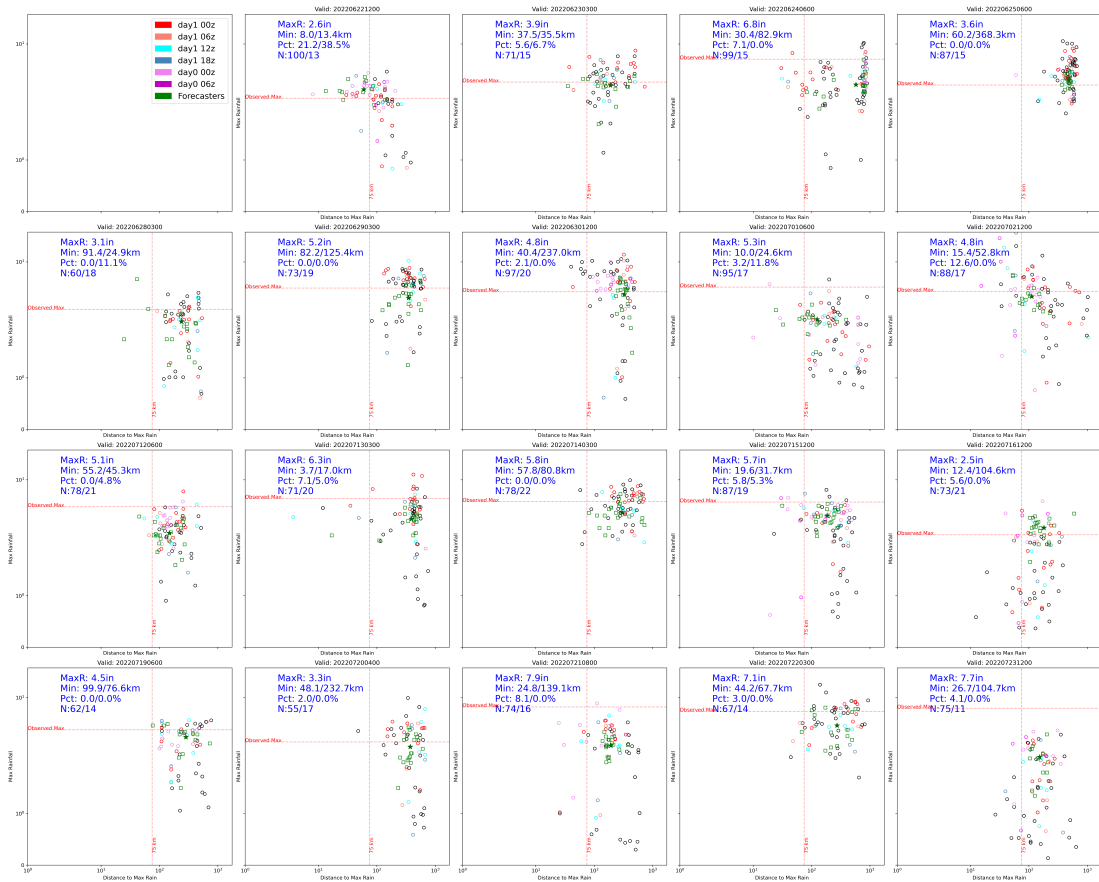


Figure 67: Each MRTP maximum rainfall distance and maximum amount diagram for each day. Participants are represented in the green squares, while models (circles) are represented in time by color, with the day 2 or 3 models in black. Both axes are on the log10 scale. Dashed lines represent a distance of 50km and the maximum observed rainfall targets.

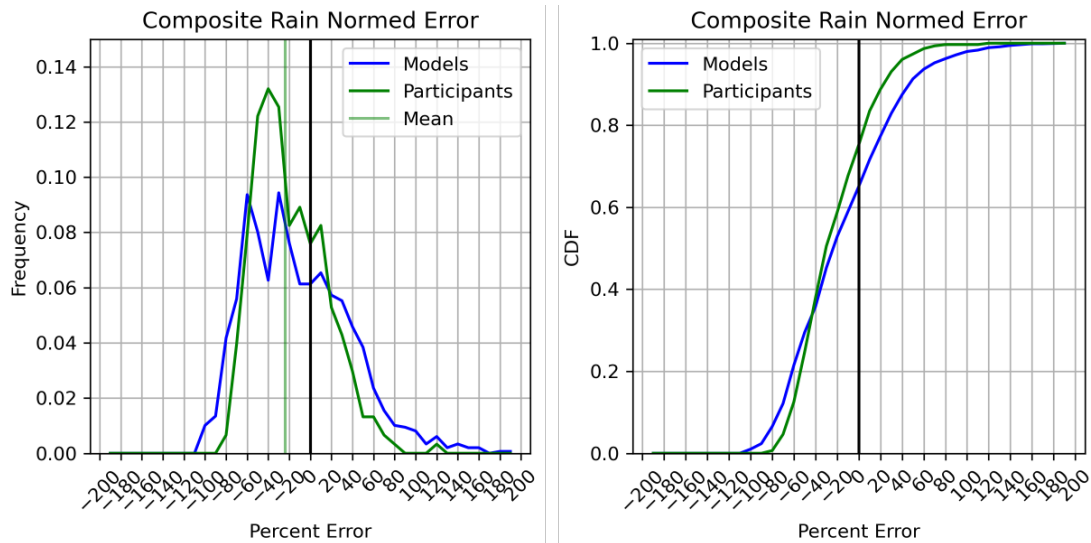


Figure 68: Composite analysis of the maximum rainfall (left) and the cumulative density function (right) across models (blue) and participants (green) expressed as normalized percent error, relative to the observed point maximum rainfall from MRMS. Shown for reference are zero error (vertical black line) and participant mean maximum rainfall (vertical green line) expressed as a percentage of the observed maximum rainfall.

4 Summary and Conclusions

The 2022 FFaIR Experiment was challenged by an abnormal season, with most extreme rainfall events being small in scale. Despite this, participants were still able to identify the general areas where these smaller events would occur, especially when it came to picking the MRTP domain and valid time. Additionally, despite this and the data flow issues that occurred during FFaIR, the team was able to tease out useful verification information about the data and products. A summary of the findings is listed below and the team’s recommendations for transition to operations can be seen in Table 4.

- The wet bias that has been noted in the previous versions of the RRFS is still present in the versions evaluated this year. This bias appears to be most extreme over the southeast United States.
- Participants once more commented on the prolific simulation of popcorn storms in the RRFS models. When analyzing the hourly precipitation, the

RRFSp1 and RRFSp2 outpaced MRMS in terms of coverage at 1.5" across the CONUS and 1.25" over the southeast.

- When compared to the operational models and MRMS precipitation rates, both instantaneous (p-rate) and hourly max (pmax), the RRFS deterministic and ensemble members had high to extreme rates. This included some simulated pmax values over 100 in h^{-1} . Often these high rates were collocated with popcorn storms.
- The HREF+ was only available for a handful of days and therefore could not be properly evaluated. However, the FFaIR team likes the concept and recommends for continued development of the MLP.
- The FV3 GEFS based ML Day 1 ERO provided by CSU performed similarly to the operational version (GEFSO) and is recommended for transition into operations.
- The CSU UFVS-based Day 1 ML ERO is recommended for continued development. Although the most preferred of the three GEFS versions by the participants, this season did not have a wide verity of heavy rainfall events, preventing the newly developed observational training method from being adequately challenged.
- The CSU HRRR-based Day 1 ML ERO is recommended for continued development. It was the least preferred of the CSU ML EROs by participants. It appears to over forecast for the Marginal risk and under forecast for the Slight risk.
- The AERO activity and product was well liked by participants. An interesting finding that came out of feedback from them was that they thought the AERO complimented the ERO and they liked to use the ERO and AERO in tandem.

Table 4: Research to Operations Transition Metrics for the 2022 FFaIR Experiment.

Models, Ensembles and Products Evaluated	Recommended for transition to operations	Recommended for further development and testing	Rejected for further testing	Provider/Funding Source
RRFSp1		X		EMC
RRFSp2	Not Applicable			GSL
CAPS ML HREF+		X		OU/CAPS Funding: Testbed Program
CSU ML FV3GEFSR	X			CSU Funding: JTTI
CSU ML Day 1 ERO UFVSFV3GEFS		X		
CSU ML Day 1 ERO HRRR		X		

References

- 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.
- Erickson, M. J., J. S. Kastman, B. Albright, S. Perfater, J. A. Nelson, R. S. Schumacher, and G. R. Herman, 2019: Verification results from the 2017 hmt–wpc flash flood and intense rainfall experiment. *Journal of Applied Meteorology and Climatology*, **58** (12), 2591 – 2604, <https://doi.org/10.1175/JAMC-D-19-0097.1>, URL <https://journals.ametsoc.org/view/journals/apme/58/12/jamc-d-19-0097.1.xml>.
- Griffin, A., S. Trojaniak, J. Correia, Jr., and J. Nelson, 2022: An analysis of the maximum rainfall and timing product forecast activity during the flash flood and intense rainfall experiment 2022, published online at <https://docs.google.com/presentation/d/1-aD6uFGNmUj2o5LAXGO-8oie8boMH8DA-UZOOh-8LL6w>. If missing please contact WPC.
- Guerry, C., and K. Fisher, 2022: Uncommon type of storm causes flooding across east tennessee, with some rescues in knoxville. Knoxville News Sentinel, accessed 11 Nov 2022, <https://www.knoxnews.com/story/weather/2022/07/21/knoxville-floods-storms-over-east-tennessee-lead-closed-roads/10115526002/>.
- Kinter, J., V. Tallapragada, and J. Whitaker, 2020: Unified forecast system research-to-operations (ufs-r2o) project proposal. Ufs-r2o proposal, COLA/GMU and NOAA/NCEP/EMC and NOAA/ESRL/PSD, 121 pp. <https://www.weather.gov/media/sti/UFS-R2O-Project-Proposal-Public.pdf>.
- Roebber, P., 2009: Visualizing multiple measures of forecast quality. *Weather and Forecasting*, **24**, 601 – 608, <https://doi.org/10.1175/JAMC-D-19-0097.1>.
- Trojaniak, S., and J. Correia, Jr., 2021: 2021 flash flood and intense rainfall experiment: Findings and results. Tech. rep., NCEP WPC-HMT. URL https://www.wpc.ncep.noaa.gov/hmt/Final_Report_2021_FFaIR_Experiment.pdf.
- Trojaniak, S., and J. Correia, Jr., 2022: 2022 ffair operations plan, published online at <https://docs.google.com/document/d/e/2PACX-1vTtE-HC4dhnR1kOjwc36>

CAftnFh0-E4Z8xVv2n9rP7sMNXH7dSVyXCa2ooDhxcX5iY5F0YTXzTk8Fg
O/pub. 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.

Whetstone, T., R. Wilusz, and S. Riley, 2022: Southwest, central virginia crews return home after helping with buchanan co. flooding response. WFXRtv.com, accessed 11 Nov 2022, <https://www.wfxrtv.com/news/local-news/southwest-central-virginia-crews-travel-to-buchanan-co-to-help-with-flood-response/>.

Appendices

A List of Participants and Seminars

Table 5: List of the participants for each week of the 2022 FFaIR Experiment. The experiment did not run during the week of the Fourth of July.

Week	WPC Forecaster	WFO/RFC	Research/Academia	EMC/GSL	Regional and National Centers/Offices and Government Agencies
Week 1 June 21 - 24	Zack Taylor	Rob Megnia - WFO BOX Evan Kutta - WFO MQT Linda Gilbert - WFO MQT Joe Clark - WFO HFO Cory Rothstein - WFO SGF Bryan Greenblatt - WFO BGM		Olivia Ostwald - EMC Ruiyu Sun - EMC Geoff Manikin - EMC Jeff Duda - GSL Curtis Alexander - GSL	David Wright - GLERL Ben Moore - PSL Nachiketa Acharay - PSL Andrea Ray - PSL Derek Stratman - NSSL
Week 2 June 27 - July 1	Frank Pereira	Marshal Pfahler - WFO LSX Emanuel Rodriguez – WFO SJU Odalys Martinez – WFO SJU Scott Rozanski – WFO HFO Douglas Kahn - WFO OAX TJ Gunkel - WFO CLE	Russ Schumacher - CSU Bill Gallus - Iowa State	Matt Morris - EMC Logan Dawson - EMC Terra Ladwig - GSL Eric James - GSL/CSU	Race Clark - NSSL Janice Bytheway - PSL Kelly Mahoney - PSL Kent Knopfmeier - NSSL/HWT Diane Cooper - FEMA
Week 3 July 11 - 15	Marc Chenard	Erin Walter - WFO GJT Dominic Paoletti - RFC AB Andrew Goenner - WFO TAR Andrei Evbuoma - WFO ALY James Danco - WFO RAH Kevin Strongman - WFO TWC Eswar Iyer - WFO AKQ Tina Stall - WFO HFO Anna Schneider - RFC CN Maura Casey - WFO GYX Cody Poche - WFO AKQ Kyle Perez - WFO SGF Gino Izzi - WFO LOT	Keith Brewster - OSU Aaron Hill - CSU	Eric Aligo - EMC Xiaoyan Zhang - EMC Binbin Zhou - EMC Ed Szoke - GSL	Heather Grams - NSSL Jackson Anthony - NSSL Burkley Twiest – SPC/HWT
Week 4 July 18 - 22	Josh Weiss	Matt Czikowsky - RFC LM Brittney Whitehead - WFO OHX Kevin Kacan - WFO DTX Charles Ross - WFO CTP Mike Colbert - WFO CTP Cecille Villanueva - WFO SJU Carlos Anselmi - WFO SJU Nicholas Slaughter - WFO GUM Robert Ballard - WFO HFO Jessie Smith - WFO MLB Jeremy Buckles - WFO MRX Chad Swain - WFO IND Arin Peters - WFO TFX		Marcel Caron - EMC John Brown - GSL	Peggy Lee - NWC Diana Stovern - PSL Rob Cifelli - PSL

Table 6: List of the 2022 FFaIR Science Seminars. The slides for the seminars can be found [here](#).

Dates	Presenter(s)	Topic/Title of Seminar	Affiliation
June 7	Jennifer Shoemake	Integrating WoFS into May 30th, 2021 Flash Flood Operations	WFO - ABQ
June 14	Jacob Carley	An Overview of the Rapid Refresh Forecast System	EMC
June 21	Brenda Phillips	Survey Results on Motorists Decision-making and Urban Flash Floods in the DFW Area	U Mass and CASA
June 23	Pat Spoden	Forecasting and IDSS for Dual Weather Threats	WFO - PAH
June 28	Erik Nielsen and Jen Henderson	Physical and Social Science Aspects of TOR/FF Events	Texas A&M and Texas Tech University
June 30	Russ Schumacher and Aaron Hill	Updates and Improvements to Colorado State University-Machine Learning Probabilities Excessive Rainfall Forecasts	Colorado State University
July 12	Marty Baxter	Spatiotemporal Changes in Michigan Rainfall and Using the NWM's Retrospective Analysis to Provide Context for Real-time Forecasts	Central Michigan University
July 14	Keith Brewster and Tim Supinie	OU CAPS RRFS Ensemble Performance and Discussion of a New Machine Learning Mean Product	OU CAPS
July 19	Andrew Orrison	A Discussion of WPC's METwatch Operations	WPC
July 21	Shakira Stackhouse	Evaluating the Skillfulness of the Hurricane Analysis and Forecast System (HAFS) Forecasts for Tropical Cyclone Precipitation using an Object-Based Methodology	AFSO-FDS Water Resources Services Branch

B MRTP Workflow

Below are the input information that participants were required to fill out on the MRTP Drawing Tool webpage prior to exporting their MRTP. The max ARI was not always the same as the ARI contour drawn. For instance, if a participant thought that a large area would see 6-h ARI exceedances of 5-y but that somewhere within that area might see an exceedance of the 100 year ARI then they would input 5 for the contour (and draw that contour) and 100 for max ARI. Following these are screenshots of the MRTP survey that participants completed as they drew their forecast.

Enter Max Rainfall Enter Max ARI Enter ARI Contour


Enter Flood Prob Enter Damage Prob Enter Hourly Max Rain

What is your unique username? *

Your answer _____

What is today's date? *

Date

mm/dd/yyyy 

Day 1 or Day 2? *

Day 1

Day 2

The predictability of this event(s) is: *

High

Medium

Low

Unknown

Other: _____

Briefly explain the predictability of this event: *

Your answer _____

Which forecast characteristic presents the largest challenge: *

- Total accumulation
- Rain rates
- Duration
- Initiation time
- Decay time
- Speed or stationarity
- Multiple rounds of rainfall
- Data Visualization
- aggregation of information
- probabilities too low or high
- None
- Other: _____

Briefly explain the forecast challenge(s) selected above, with this event: *

Your answer _____

Given your MRTP Forecast, would you expect: *

- Minimal impacts (brief inconsequential flooding)
- Localized flooding
- Flooding but minimal damage
- Flooding with widespread damage
- Other: _____

Which model were you assigned to evaluate? *

- RRFS P1
- RRFS P2 (GSL CTL)
- RRFS P3 (CAPS CTL)
- NSSL
- FV3-HREF
- RRFS P4
- RRFS P5
- RRFS P6
- RRFS P7
- RRFS P8
- HRRR
- NAMnest
- None
- Other: _____

Which ensemble were you assigned to evaluate? *

- HREF
- RRFSe (GSL)
- CAPS_RRFSe
- None
- Other: _____

How did you use the suite of Model or Ensemble Guidance available to you? *

	Used	Useful	Considered	Not considered	Not available
RRFSp1	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
RRFSp2 (GSL CTL)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
RRFSp3 (CAPS CTL)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
NSSL	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
RRFS P4	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
RRFS P5	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
RRFS P6	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
RRFS P7	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
RRFS P8	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
HRRR	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
NAMnest	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
HREF	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
RRFSe (GSL Ens)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
CAPS_RRFSe (CAPS Ens)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
FV3-HREF	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Please discuss the utility or usefulness of the model/ensemble you evaluated. Include primary concerns such as magnitude, timing, phenomena that may have contributed to your assessment. *

Your answer _____

Which model were you assigned to evaluate? *

- RRFS P1
- RRFS P2 (GSL CTL)
- RRFS P3 (CAPS CTL)
- NSSL
- FV3-HREF
- RRFS P4
- RRFS P5
- RRFS P6
- RRFS P7
- RRFS P8
- HRRR
- NAMnest
- None
- Other: _____

Which ensemble were you assigned to evaluate? *

- HREF
- RRFSe (GSL)
- CAPS_RRFSe
- None
- Other: _____

The phenomenon you are forecasting for the maximum rainfall involves (choose all that apply): *

- Mesoscale Convective System
- Supercells
- Multi-cell clusters
- Tropical Cyclone
- Training convective elements
- Fast moving convection
- Popcorn convection
- Stratiform rainfall
- Other: _____

C FFaIR Surveys

This appendix contains screenshots of the FFaIR Verification Survey. In some instances the whole extent of a question block is not shown. For example, for Question 1 (Fig. 69), not all the RRFSp scoring questions are shown since the sub-questions for Question 1 are repetitive. In other instances, there was not enough space to show all the possible choices for a grid; ex. Fig. 70.

C.1 2022 FFaIR Verification Survey

00z QPF	12z QPF
<p>Question 1: 00z Runs -- Using 24-h MRMS-GC QPE, evaluate the utility of the 24-h QPF forecast. Please evaluate, on a scale of 1-10, where 1 is the poor and 10 is great, the model's ability to capture precipitation location and totals. Focus on things like how the model might have informed you about the likelihood of maximum precipitation amounts, how widespread an event might be, storm propagation/mode, etc.</p>	<p>Question 2: 12z Runs -- Using 24-h MRMS-GC QPE, evaluate the utility of the 24-h QPF forecast. Please evaluate, on a scale of 1-10, where 1 is the poor and 10 is great, the model's ability to capture precipitation location and totals. Focus on things like how the model might have informed you about the likelihood of maximum precipitation amounts, how widespread an event might be, storm propagation/mode, etc.</p>
<p>HRRR</p> <p>1 2 3 4 5 6 7 8 9 10</p> <p>Poor <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Great</p>	<p>HRRR</p> <p>1 2 3 4 5 6 7 8 9 10</p> <p>Poor <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Great</p>
<p>NAMnest</p> <p>1 2 3 4 5 6 7 8 9 10</p> <p>Poor <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Great</p>	<p>NAMnest</p> <p>1 2 3 4 5 6 7 8 9 10</p> <p>Poor <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Great</p>
<p>RRFSp1</p> <p>1 2 3 4 5 6 7 8 9 10</p> <p>Poor <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Great</p>	<p>RRFSp1</p> <p>1 2 3 4 5 6 7 8 9 10</p> <p>Poor <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Great</p>
<p>RRFSp2</p> <p>.... RRFSp3 RRFSp4 RRFSp5 RRFSp6 RRFSp7....</p>	<p>General thoughts and comments.</p> <p>Your answer _____</p>
<p>RRFSp8</p> <p>1 2 3 4 5 6 7 8 9 10</p> <p>Poor <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Great</p>	
<p>General thoughts and comments.</p> <p>Your answer _____</p>	

Figure 69: Question 1 on left and Question 2 on right of the verification survey for FFaIR 2022. Rating the “goodness” of the 00z (left) and 12z (right) 24h QPF forecasts.

Data Assimilation

Question 3 - Here we will be evaluating the first 6 hours of the forecast to determine the quality of the precipitation/reflectivity from f00/f01-f06. Check the attendance sheet to see which you are assigned.

Which were you assigned? *

Precipitation/Reflectivity 1hr (forecast hour 1)

Precipitation/Reflectivity 3hr (forecast hour 3)

Precipitation/Reflectivity 6hr (forecast hour 6)

NA

Please rate each model according to the precipitation accumulation: *

	0 (lowest)	1	2	3	4	5	6	7	8
RRFSp1	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
RRFSp2	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
HRRR	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
RRFSp3	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Please rate each model according to the composite reflectivity: *

	3	4	5	6	7	8	9	10 (highest)	NA
RRFSp1	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
RRFSp2	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
HRRR	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
RRFSp3	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Briefly describe the quality of the models rated above:

Your answer _____

Figure 70: Question 3 of the verification survey for FFaIR 2022. Evaluating data assimilation impacts. The arrow shows that the question to the right followed the question to the left. The circles highlight how the whole grid of choice could not be seen without scrolling.

Prate and Pmax Evaluation

Question 4: For potential matched convective areas, how do the values of Prate or Pmax compare to MRMS? Please look at all 6hr provided.

What comparison were you assigned? *

Prate
 Pmax

In comparison to the MRMS precipitation rates please pick one or more than apply *

	NA	HRRR	NAMnest	RRFSp1	RRFSp2	RRFSp3
Overall values comparable	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Overall values higher	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Overall values lower	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Overall coverage of most intense rates are comparable	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Overall coverage of most intense rate is larger	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Overall coverage of most intense rates is smaller	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Overall coverage of rates is sporadic like MRMS	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Overall coverage of rates is sporadic unlike MRMS	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Please provide information about the differences between the models you were assigned as well as comments about the prate/pmax information.

Figure 71: Question 4 of the verification survey for FFaIR 2022. Participants were randomly assigned to evaluated either p-rate or pmax.

Model Timing

Question 5

Please select what model you were assigned. *

HRRR

RRFSp1

RRFSp2

RRFSp3

NA

Focusing on the MRTP domain and valid time, How did each forecast hour compare to observations: *

	NA	1 (Worst)	2	3	4	5 (Best)	Other
T-2hours	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
T-1hours	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
T	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
T+1hours	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
T+2hours	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Could you identify a timing error with your chosen model? Explain. *

Your answer _____

Figure 72: Question 5 of the verification survey for FFaIR 2022, evaluating the “goodness” of model timing of 6h precipitation totals. Participants were assigned which model to evaluate: HRRR, RRFSp1, RRFSp2 or RRFSp3. Due to the types of events that occurred this year, this question was usually skipped during verification.

Ensembles

Question 6: We are evaluating the probabilities of 1" and 2" in 6hrs from various ensemble systems. The MRMS QPE has had a 39km neighborhood and 39km smoother applied. Please focus on the area of the MRTF.

1" in 6hrs corresponds in some way to the observations. How?

	Too High	High	About Right	Low	Too Low
HREF	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
RRFSe	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
CAPS_RRFSe	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

2" in 6hrs corresponds in some way to the observations. How?

	Too High	High	About Right	Low	Too Low
HREF	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
RRFSe	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
CAPS_RRFSe	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Please describe the kind of event that was observed and how the probabilities do or do not correspond with that event.

Your answer _____

General thoughts and comments on the RRFSe and CAPS_RRFSe compared to the HREF. *

Your answer _____

Figure 73: Question 6 of the verification survey for FFaIR 2022. Comparison of the experimental ensembles to the HREF.

CAPS HREF+

Question 7 Evaluating the ML 0.5" in 6hr product from CAPS against HREF and CAPS_RRFSe.

Rate how the NMEP HREF+ performed in comparison to the HREF.

1 2 3 4 5

CAPS HREF+ was much worse than HREF. CAPS HREF+ was much better than HREF.

Rate how the NMEP HREF+ performed in comparison to the CAPS_RRFSe.

1 2 3 4 5

CAPS HREF+ was much worse than CAPS_RRFSe. CAPS HREF+ was much better than CAPS_RRFSe.

Focusing on the two HREF+ probabilities that differ in their neighborhood probability approach, which do you prefer?

NEP HREF+

NMEP HREF+

I like seeing both approaches

Please discuss your answer to the above questions.

Your answer _____

Figure 74: Question 7 of the verification survey for FFaIR 2022. Evaluation of the CAPS HREF+ MLP.

CSU EROs	
<p>Question 8: Using the UFV, practically perfect analysis, and the 24-h MRMS-GC Gauge-Corrected QPE, please rate the utility of the CSU-First Guess Day 1 ERO, valid from 12 UTC to 12UTC.</p>	<p>Were there noticeable differences between the GEFS EROs that used the old training for heavy rainfall and the GEFS EROs that use UFVS?</p> <p>Your answer _____</p>
<p>00z GEFSO</p> <p>1 2 3 4 5 6 7 8 9 10</p> <p>Poor <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Great</p>	<p>HRRR</p> <p>1 2 3 4 5 6 7 8 9 10</p> <p>Poor <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Great</p>
<p>00z FV3GEFSR</p> <p>1 2 3 4 5 6 7 8 9 10</p> <p>Poor <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Great</p>	<p>NSSL2</p> <p>.... BLEND....</p>
<p>12z FV3GEFSR</p> <p>1 2 3 4 5 6 7 8 9 10</p> <p>Poor <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Great</p>	<p>General thoughts and comments on the CSU-First Guess Day 1 ERO.</p> <p>Your answer _____</p>
<p>00z UFVS-FV3GEFSR 12z UFVS-FV3GEFSR</p>	<p>In realtime and in the verification images we have plotted the 2.5% exceedance (below the marginal threshold). Does the gray change your perspective on performance of the CSU MLP forecast, why or why not?</p> <p>Your answer _____</p>

Figure 75: Question 8 of the verification survey for FFaIR 2022, rating the “goodness” of the 00z and 12z CSU ML EROs.

FFaIR ERO and AERO

Question 9

Using the UFVS, practically perfect analysis, and the 20-h MRMS QPE, please rate * the FFaIR Day 1 ERO, valid from 16 UTC to 12UTC.

1 2 3 4 5 6 7 8 9 10

Poor Great

General thoughts and comments on the performance of the FFaIR ERO.

Your answer _____

Using point analysis of ARI exceedances, 20-h MRMS QPE, and NOAA Atlas 6h 2y * and 10y ARI, please rate the FFaIR Day 1 AERO, valid from 16 UTC to 12UTC.

1 2 3 4 5 6 7 8 9 10

Poor Great

Focusing specially on the LSRs, how well do you feel ARI exceedances matched * up with reports of heavy rainfall?

1 2 3 4 5

Poor Great

Please comment on your thoughts on the AERO. Include things like if you thought * it is useful for identifying heavy rainfall, if you thought the thresholds are good, etc.

Figure 76: Question 9 of the verification survey for FFaIR 2022, rating the “goodness” of the FFaIR ERO and AERO.

MRTP

Question 10

What is your MRTP username?

Your answer _____

Are you evaluating your MRTP or someone else's MRTP?

Mine

Someone else

NA

How do you feel your 1 inch contour did in comparison to observations?

	Very Poor	Poor	FFaIR	Good	Very Good	No Opinion
Area	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Location	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Orientation	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Amounts	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Timing	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
ARI	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

How do you feel your assigned models' 1 inch contour did in comparison to observations? (Please select the model cycle you feel you used the most)

	Very Poor	Poor	FFaIR	Good	Very Good	No Opinion
Area	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Location	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Orientation	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Amounts	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Timing	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
ARI	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

If you remember, please indicate what model you were assigned, along with the cycle available.

Your answer _____

Figure 77: Question 10 part 1 of the verification survey for FFaIR 2022. Evaluation of the Day 1 MRTP. The arrow shows that the question to the right followed the question to the left.

Day 2 (when applicable): How do you feel your 1 inch contour did in comparison to observations?

	Very Poor	Poor	FFaIR	Good	Very Good	No Opinion
Area	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Location	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Orientation	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Amounts	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Timing	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
ARI	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Day 2 (when applicable): How do you feel your assigned models' 1 inch contour did in comparison to observations? (Please select the model cycle you feel you used the most)

	Very Poor	Poor	FFaIR	Good	Very Good	No Opinion
Area	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Location	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Orientation	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Amounts	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Timing	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
ARI	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Day 2 v Day 1 (when applicable): How do you feel your day 2 MRTP did in comparison to your Day 1. Other thoughts and comments.

Your answer _____

Figure 78: Question 10 part 2 of the verification survey for FFaIR 2022. Evaluation of the Day 2 MRTP.

C.2 2022 End of the Week Feedback on FFaIR and its Products Survey

CSU First-Guess ML Day 1 ERO Products

This section focuses on the machine learning CSU ERO products.

Were the CSU First-Guess EROs useful in the ERO forecasting activity?

Yes
 No

You had 6 different EROs available during FFaIR. Overall, during the week did you feel one version performed better than the others?

GEFSO
 FV3GEFSR
 UFVSGEFSR
 HRRR
 NSSL2
 BLEND
 No one consistently outperformed the others
 I do not remember

Please comment on your response from the previous question.

Your answer _____

There were three versions of GEFS-trained ERO products. During your week, did you feel that either of the FV3-based GEFS products (FV3GEFSR or UFVSGEFS) outperformed the operational version of the product (GEFSO). Please explain.

Your answer _____

Focusing on the HRRR ERO, how do you feel it performed in the southwest? If you participated in FFaIR last year, do you feel the updated version of the model performed better than last year?

Your answer _____

Overall, general thoughts or comments about the First-Guess EROs.

Your answer _____

Figure 79: End of the week questions regarding the CSU MLP EROs.

The RRFs

This section focuses on the performance of the rapid refresh forecast system.

Throughout the week, overall how do you feel the RRFs performed when compared to the HRRR?

	1 (HRRR was way better)	2	3 (similar performance)	4	5 (RRFS was better)	NA
RRFS p1	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
RRFS p2	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
RRFS p3	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Throughout the week, overall how do you feel the RRFs performed when compared to the NAM?

	1 (NAM was better)	2	3 (similar performance)	4	5 (RRFS was better)	NA
RRFS p1	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
RRFS p2	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
RRFS p3	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Please comment on the above two questions.

Your answer _____

During FFaIR, did you notice any biases or challenges to using the RRFs for forecasting? Please comment. (This includes biases/challenges seen in ALL RRFSp#).

Your answer _____

Did the characteristics of precipitation accumulation seem reasonable from the RRFs models? Please explain.

Your answer _____

Did the characteristics of precipitation rate (Prate or Pmax) seem reasonable from the RRFs models? Please explain.

Your answer _____

Were products from the RRFSe or CAPS_RRFSe available for you to look at during FFaIR?

Yes

No

Only for some days

Unsure

If the ensembles were available, please comment on their utility and their performance in comparison to the HREF.

Your answer _____

Overall, would you be able to use the RRFs model or ensemble in an operational forecasting environment?

Your answer _____

Figure 80: End of the week questions regarding the RRFs.

CAPS ML 0.5" in 6hr Probabilities

Was the CAPS ML 0.5" in 6hr product available to you to use?

Yes
 No
 Unsure

How do you feel the CAPS ML product performed when compared to the HREF and the CAPS_RRFSe?

	1 CAPS MLP was worse	2	3 comparable	4	5 CAPS MLP was better	unsure	NA
HREF	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
CAPS_RRFSe	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Please provide any additional comments on the CAPS ML product.

Your answer _____

Figure 81: End of the week questions regarding the CAPS MLP HREF+.

Forecasting Activities

The AERO is a product in development. Do you feel a product like this would be useful? Please explain.

Your answer _____

As we continue to refine the AERO, do you have any suggestions on how to update the product? For example, changing the contours drawn or the ARI time (aka 6hr).

Your answer _____

Did the Maximum Rate and Timing Product (MRTP) serve its purpose to find the area/time of maximum precipitation and/or impacts? Please explain.

Your answer _____

Did you feel the MRTP forecast and verification activity helped you evaluate some of the model guidance in a focused and coherent way? Please explain.

Your answer _____

We did a Day 1 and Day 2 MRTP this year. Did you enjoy doing 2 MRTPs? Please explain the positives and negatives of doing two in a day.

Your answer _____

Please provide any additional comments on the MRTP here.

Your answer _____

Figure 82: End of the week questions regarding the forecasting activities: AERO and MRTP.

General Questions and Comments on FFaIR

Overall how did you like FFaIR? Please explain what you did and didn't like.

Your answer _____

Did the FFaIR graphics website provide you with sufficient information to make forecasts? If anything was crucially missing, please list variables that may help with future forecasting activities.

Your answer _____

Aside from creating a more dynamic website, do you have any other suggestions on ways to improve the main HMT website?

Your answer _____

Did the drawing websites provide appropriate functionality to perform the drawing tasks? Is there anything that you would add to assist in the drawing of forecasts for MRTP?

Your answer _____

Please provide any additional comments on FFaIR here.

Your answer _____

Figure 83: End of the week questions for general feedback on the 2022 FFaIR Experiment.