2023 Flash Flood and Intense Rainfall (FFaIR) Final Report: *Part 2 - Non-RRFS Related Results and Findings*

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

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

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

Contents

1 Introduction

This is part 2 of the 2023 Flash Flood and Intense Rainfall (FFaIR) Experiment report. For a detailed summary of the FFaIR Experiment this year, please refer to Part 1 of the Final Report (Trojniak and Correia, Jr., 2023b). As a reminder, the weeks FFaIR were in session were:

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

The list of participants and the seminars given can be found in Appendix A of Part 1 of the report.

This portion of the Final Report will summarize the results and findings of the products evaluated in FFaIR that were not related to the Rapid Refresh Forecast System (RRFS); the Colorado State University (CSU) First-Guess Excessive Rainfall Outlooks (EROs) and the satellite products provided by the Cooperative Institute for Research in the Atmosphere (CIRA) at CSU. It will also cover analysis of the two Day 1 forecasting activities done in FFaIR: the Excessive Rainfall Outlook (ERO) and an Average Recurrence Interval (ARI) based ERO called the AERO. Those interested in results related to the RRFS, products provided by the Center for Analysis and Prediction of Storms (CAPS) at the University of Oklahoma (OU), or the Maximum Rainfall and Timing Product (MRTP) forecasting activity should refer to Part 1 of the Report.

2 Science and Operations

This section will provide a brief summary of the forecasting activities, products and verification methods as it relates to Part 2. An in-depth summary of the

daily operations of FFaIR and the data and products can be found in the 2023 FFaIR Operations Plan (Trojniak and Correia, Jr., 2023a).

2.1 Daily Schedule

As stated in Part 1, the day-to-day operations of the 2023 FFaIR Experiment were similar to those in 2022. In the morning, after a weather briefing by a WPC forecaster, participants were broken into two groups to do either a Day 1 (16-12 UTC) ERO or Day 1 AERO forecast. This consisted of them creating an individual outlook and a collaborative one. This differed from last year where participants had the option to create an individual forecast using Google Sheets but it was not required. Furthermore, this year participants had access to the HMT ERO drawing tool that allowed them to create geoJSON files to be exported and sent to the FFaIR team, rather than just drawing polygons on a static map image in a Google Slides deck. An example of how the individual and collaborative ERO and AERO might differ can be seen in Figs. 1 and 2 respectively. Although no formal analysis was done on the individual ERO and AERO products for this report, the FFaIR team hopes to use the forecasts collected for future analysis and development of ERO and AERO verification and methodology.

Once the individual outlooks were completed, volunteers were asked to share their screen to discuss their ERO/AERO. This included explaining their thought process and what data they looked at (models, ensembles, tools). This helped to quickly show when participants agreed on regions of interest and what areas might warrant more in-depth discussion during the collaborative process. The goal was to complete both the individual and collaborative products by their valid start time, 16 UTC, like would be required in operations. However, due to the experimental nature of the models/products (ex. late data) and other circumstances (ex. teaching how to use the tool on the first day), the collaborative ERO/AEROs were not always completed by 16 UTC. When this was the case, participants were instructed they could no longer look at real-time data like radar reflectively or satellite products. Once both groups completed their respective forecasts, the two groups would summarize their forecast to the other group, highlighting things such as where they had the most/least confidence in the forecast and products

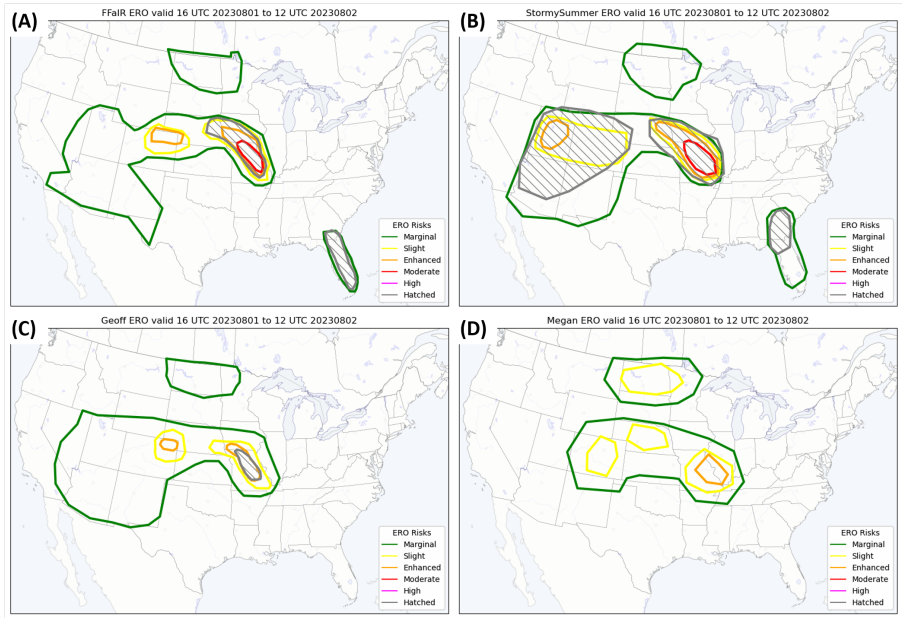

Figure 1: The Day 1 ERO valid 16 UTC 01 Aug. to 12 UTC 02 Aug. 2023 for the (A) FFaIR ERO and (B)-(D) three of the participants' individual EROs that helped lead to the collaborated FFaIR ERO in (A). The ERO Risk contours are - Marginal: 5%-15% (green), Slight: 15%-25% (yellow), Enhanced: 25%-40% (orange), Moderate: 40%-70% (red) and High: >70% (purple/pink). The Hatched (intensity) contour is grey with hatching.

used in the forecast. The daily collaborative EROs and AEROs can be found in Section 2.3, organized by week and product.

After a short break, the verification portion of the experiment took place. This flanked either side of the lunch break. The verification session, as it relates to Part 2, will be discussed in Section 2.5. After this, a weather briefing was done on the near to short term heavy rainfall chances for the MRTP activity. Once a domain and time were picked for the MRTP participants spent the remainder of the day working on their individual MRTP.

2.2 Description of the Day 1 Forecasting Activities

Since the forecasting activities closely followed those done in the 2022 FFaIR Experiment and are explained in great detail in the 2023 FFaIR Operations Plan (Trojniak and Correia, Jr., 2023a), the activities will only be briefly explained. The

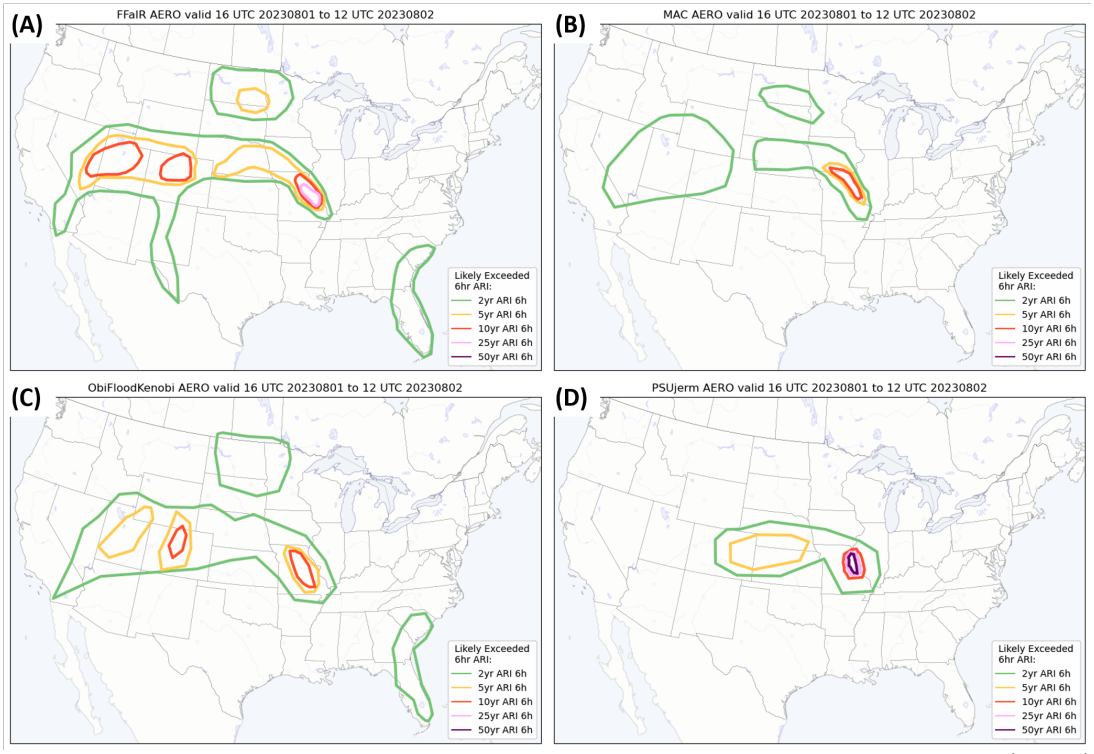

Figure 2: Similar to Fig. 1 but for the AERO. The AERO contours are - 2-y (green), 5-y (yellow), 10-y (red), 25-y (pink), and 50-y (purple) 6-h ARI; contours indicate when ARI is likely to be exceeded during the 20-h the AERO is valid.

two Day 1 products, the ERO and AERO, were both valid 16 UTC to 12 UTC to mimic when the last scheduled update to the Day 1 WPC ERO is valid. The FFaIR ERO followed the same definition as the operational ERO, "the probability that rainfall will exceed Flash Flood Guidance (FFG) within 40 kilometers (25 miles) of a point." However, there were some deviations from the operational ERO. Rather than four risk categories defined as Marginal (5-15%), Slight (15-40%), Moderate (40-70%), and High (>70%), the FFaIR ERO had five risk categories, testing a category set between the Slight and Moderate[1], referred to as the Enhanced Slight (hereafter Enhanced). Therefore, for the FFaIR ERO the risk categories were: Marginal (5-15%), Slight (15-25%), Enhanced (25%-40%) Moderate (40-70%), and High (>70%).

[1]WPC operations is also currently testing the addition of an Enhanced risk in realtime. ERO forecasters create both an ERO with 4 categories and an ERO with additional risk categories.

Also tested this year for the ERO product was an Intensity contour. Initially the contour was loosely defined as highlighting where the 10-y 6-h ARI could be exceeded. The idea for this definition was based on feedback in previous FFaIR experiments about how aspects of the AERO could be combined with the ERO to better convey the excessive rainfall risk based on precipitation amounts alone. However, once put into practice, the Intensity contour was also used to try and highlight where other sorts of rainfall rates might be intense, for instance if the 15 min rain rate was likely to be 0.5 in 15 min^{-1}. The Intensity contour is also referred to as the Hatched area, since the area was displayed in grey hatching.

The methodology of the AERO followed closely to the 2022 FFaIR Experiment, with the product defined as the 6-h ARI that is most likely to be exceeded within 25 miles of a point, for any six-hour time period within the valid time of the product. Unlike the ERO, there is no set probability of exceedance needed for a contour to be drawn; it is up to the participants to determine for what probability they want to draw. The ARI exceedance values that were included in the AERO were: the 6-h 2-y, 5-y, 10-y, 25-y, and 50-y ARIs. In general, the best way to distinguish what the AERO is trying to highlight versus what the ERO is highlighting is thinking in terms of intensity and coverage. The ERO risk categories are used to highlight the coverage of FFG being exceeded while the AERO is trying to highlight the likely rainfall intensity. Refer to Figs. 1A and 2A to see how these forecasts might highlight different areas.

2.3 Overview of Data and Products

This section will serve as a brief summary of the data and products evaluated in FFaIR that will be discussed in Part 2 of the 2023 Final Report. For an in-depth description of the products please refer to the 2023 FFaIR Operations Plan (Trojniak and Correia, Jr., 2023a).

2.3.1 Colorado State University's First-Guess EROs

Russ Schumacher and his team at Colorado State University (CSU) once again provided Day 1 First-Guess ERO (also referred to as MLP ERO) products for

evaluation during FFaIR. Three versions were trained on the GEFS and one on the HRRR. Last year, the FFaIR team recommended that one of the versions, referred to as the FV3GEFSR, be transitioned into operations at WPC. This occurred and has been used in WPC operations alongside the original GEFS version (called the GEFSO). These MLP EROs vary only in which version of the GEFS the model was trained on, the GEFSO used GEFSv11 for training while the FV3GEFSR used GEFSv12[2]. Even though both are operational, they were evaluated to make sure the GEFSO could be transitioned out of operations and be replaced by the FV3GEFS.

The third GEFS MLP, called the UFVSGEFSR was also evaluated last year. This version was trained on the GEFSv12 but uses the Unified Flooding Verification System (UFVS) as its observational training dataset rather than rainfall exceedances of 1-y and/or 2-y ARIs. The HRRR MLP ERO has also been evaluated in past FFaIRs and has undergone numerous updates to how its forecasts are produced over the years. This year's changes included using hourly predictors instead of 3-h predictors and training solely on HRRRv4, rather than a combination of the HRRRv3 and HRRRv4.

2.3.2 Cooperative Institute for the Atmosphere Satellite Products

The Cooperative Institute for the Atmosphere (CIRA) at CSU provided two new derived satellite products centered around the analysis of vapor content and advection. An Hourly Percentile Ranking of Advected Layer Precipitable Water (ALPW) by layer which uses monthly LPW fields dating back to 2013 are used to rank the current ALPW field in terms of percentiles. The other product was a Layered water Vapor Transport (LVT), which was designed to mimic Integrated Vapor Transport (IVT) and is used to help identify the strength of atmospheric rivers. Both of these products were created hourly and were available by :40 past the hour. The products were shown daily in the weather briefings so participants got a sense of how useful they were in real time to help diagnose moisture availability. These were not formally evaluated daily like the other products and tools but general feedback questions were asked at the end of the week about the products.

[2]Both versions use the GEFSv12 as its forecast.

2.4 Science Questions and Goals for Part 2

The science questions and goals that will be addressed in Part 2 of the Final Report are listed below.

- Analyze the Colorado State University (CSU) ERO MLPs.
- Explore the addition of an ERO risk category between a Slight and Moderate risk.
- Explore including an intensity contour on the ERO, defined by exceeding some ARI threshold.
- Continue to analyze the the utility of using 6-h ARI QPF exceedances as a proxy to identify rainfall intensity via the AERO and work to develop a verification methodology for the product.
- Gather feedback on the two new CIRA satellite products.

2.5 Verification

Like with the verification questions discussed in Part 1, "goodness" questions were utilized during subjective verification, using a scale of goodness from 1 (poor) to (10) great. For some questions, "preference" questions were also asked to help understand what participants liked or didn't like about certain products or verification techniques. Additionally, as is usually the case with FFaIR verification questions, participants were able to provide additional insight to the forecasts both in written and verbal form.

As noted earlier, evaluation of the Satellite products from CIRA was done via real time use and feedback and via an end of the week survey. The questions asked can be seen in Fig. 3. Not all of the questions asked will addressed individually, but rather a general summary of the feedback will be provided.

Evaluation of the performance of the CSU and FFaIR ERO risk areas was done using practically perfect derived from the UFVS (see Erickson et al. (2019) for more information) and 20-h MRMS; see Fig. 4 for an example of the verification graphic. These were scored using the goodness scale from 1 to 10. Since the FFaIR ERO

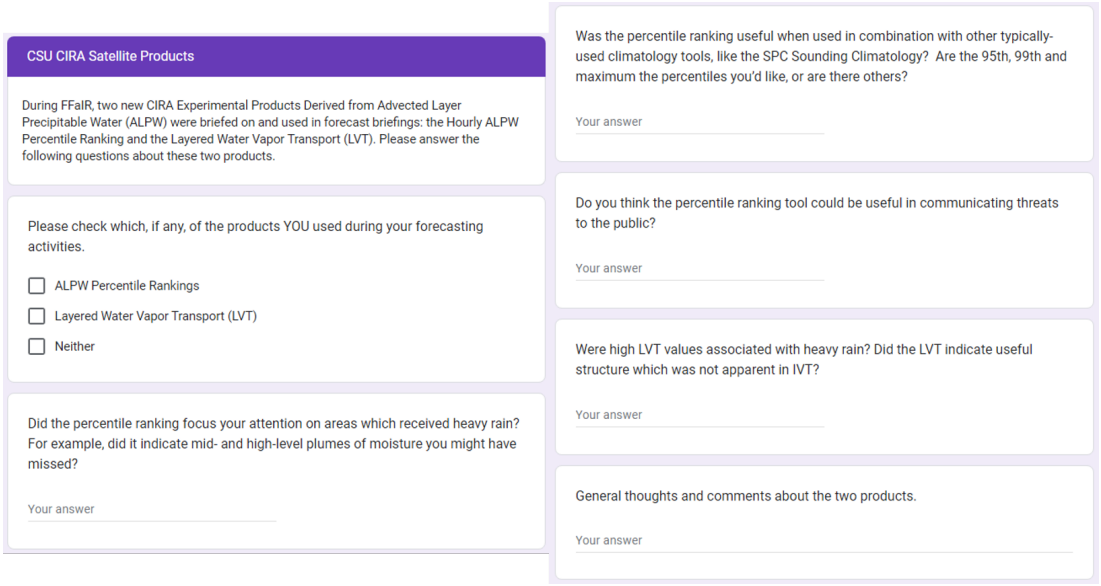

Figure 3: The end of the week survey questions ask for the CIRA Hourly Percentile Ranking of ALPW and LVT Satellite Products for the 2023 FFaIR Experiment.

was testing a new risk category between Slight and Moderate, called the Enhanced Risk, the probabilities for the risk category were shown for the CSU MLP EROs as well. For the FFaIR ERO, the verification graphic also included the instances of the 6-h 10-y ARI exceedance to aid in the analysis of the Hatched contour; Fig. 4D. In addition to the goodness question, participants were also asked about the Enhanced Risk and Hatched area. For the Enhanced Risk they were asked "If applicable, was having the ability to draw an Enhanced risk useful? Do you feel it helped better convey the flash flooding risk? Please explain and comment on anything else about the Enhanced risk you would like us to know." There were two questions for the Hatched area, Fig. 5 shows the first question. The second was a written response question to "Please provide feedback on the Hatched risk if you would like."

Verification of the FFaIR AERO was accomplished in two parts. The first part followed the same format as last year and was a goodness question, with scoring from 1 to 10. To help with evaluation of the FFaIR AERO performance, the participants were shown the MRMS 6-h ARI exceedances for the AERO thresholds, 2-y, 5-y, 10-y, 25-y, and 50-y, with the FFaIR AERO overlayed. Also included was the 20-h MRMS and the values of the 2-y and 25-y 6-h ARIs. The second part

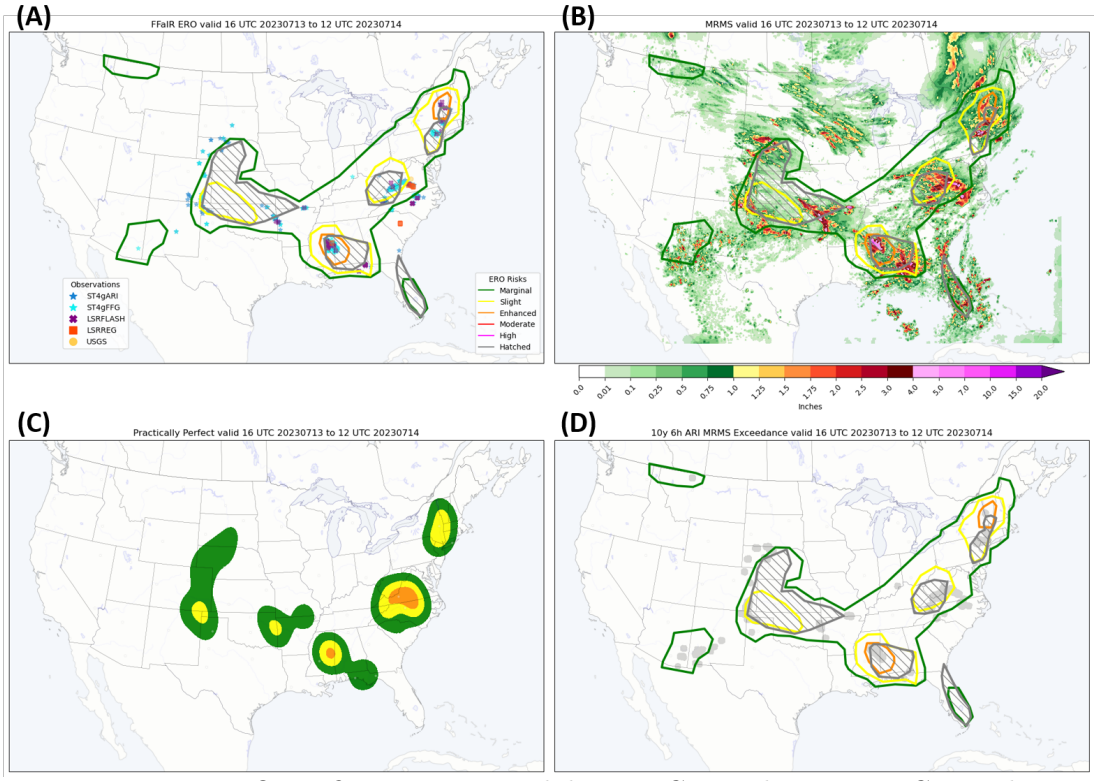

Figure 4: FFaIR ERO verification image valid 16 UTC 12 July to 12 UTC 13 July 2023. (A) FFaIR ERO with UFVS overlayed. Risk categories - Marginal: 5%-15% (green), Slight: 15%-25% (yellow), Enhanced: 25%-40% (orange), Moderate: 40%-70% (red) and High: >70% (purple/pink). (B) the 20hr QPE, (C) is the practically perfect verification, and (D) plots where the 6-h 10-y ARI exceedances.

was an exploratory question for the development of a practically perfect verification product for the FFaIR AERO. The practically perfect was created using 6-h MRMS QPE exceedances of the ARI's used in the AERO definition. A radius of influence (ROI) of 40 km was applied to the locations of the ARI exceedances. Two different Gaussian smoothing radii (70 km and 105 km) were used for the subjective evaluation. Participants were shown the practically perfect for the 2-y and 10-y 6-h ARI. The practically perfect was contoured at the values of the ERO risks (5%, 15%, 25%, 40% and 70%) for simplicity's sake. Figure 6 shows what the participants used to evaluate the 2-y 6-h ARI practically perfect. A similar graphic was provided for the evaluation of the 10-y 6-h ARI, but based on the 10-y 6-h ARI exceedances. Participants were asked if they preferred a specific ROI/smoothing

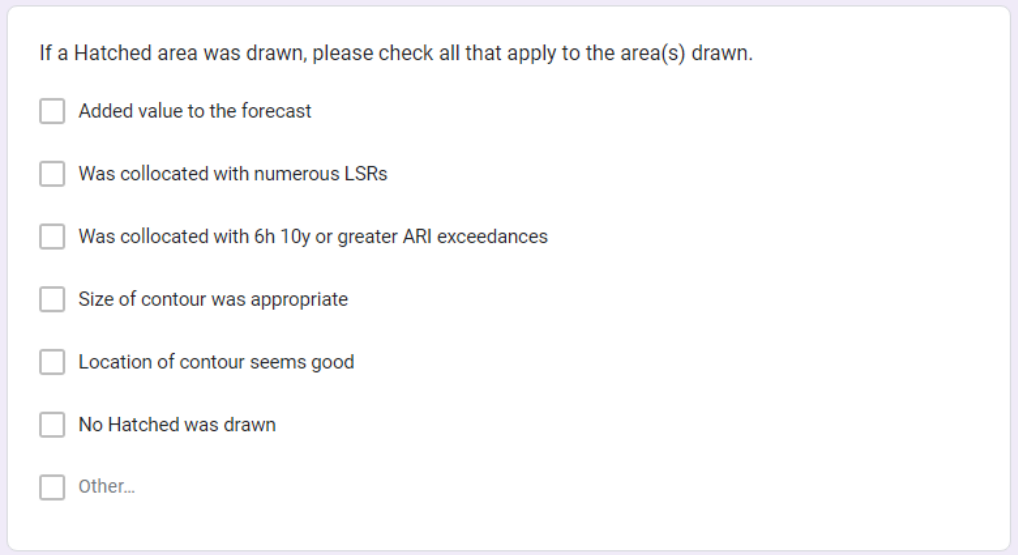

Figure 5: The question setup for the FFaIR ERO's Hatched area.

combination and at what practically perfect value(s) they felt best conveyed the risk of ARI exceedance within 25 miles of point given what occurred; see Fig. 7. They were allowed to choose multiple options. For instance, on the day shown in Fig. 6 one participant picked 40% for the 70 km smoothing and 25% for the 105 km, while another participant picked 15% and 25% respectively. The goal is to use this information to understand what forecasters would find useful in a product like this, what they want it to "look like", and at what probability of exceedance threshold the AERO contours should be drawn at.

3 A Brief Discussion on the Weather for FFaIR

A section with the same title can be found in Part 1 of the 2023 Final Report (Trojniak and Correia, Jr., 2023b). It provides a general summary of the synoptic patterns/total precipitation seen during FFaIR and how this differs from previous years. It also touched on some of the events that occurred during FFaIR, specifically as they related to the MRTP activity. In this section, we will review one additional event. Also, a summary of the daily extreme precipitation risks shown by the collaborative EROs and AEROs can be seen in Appendix A.

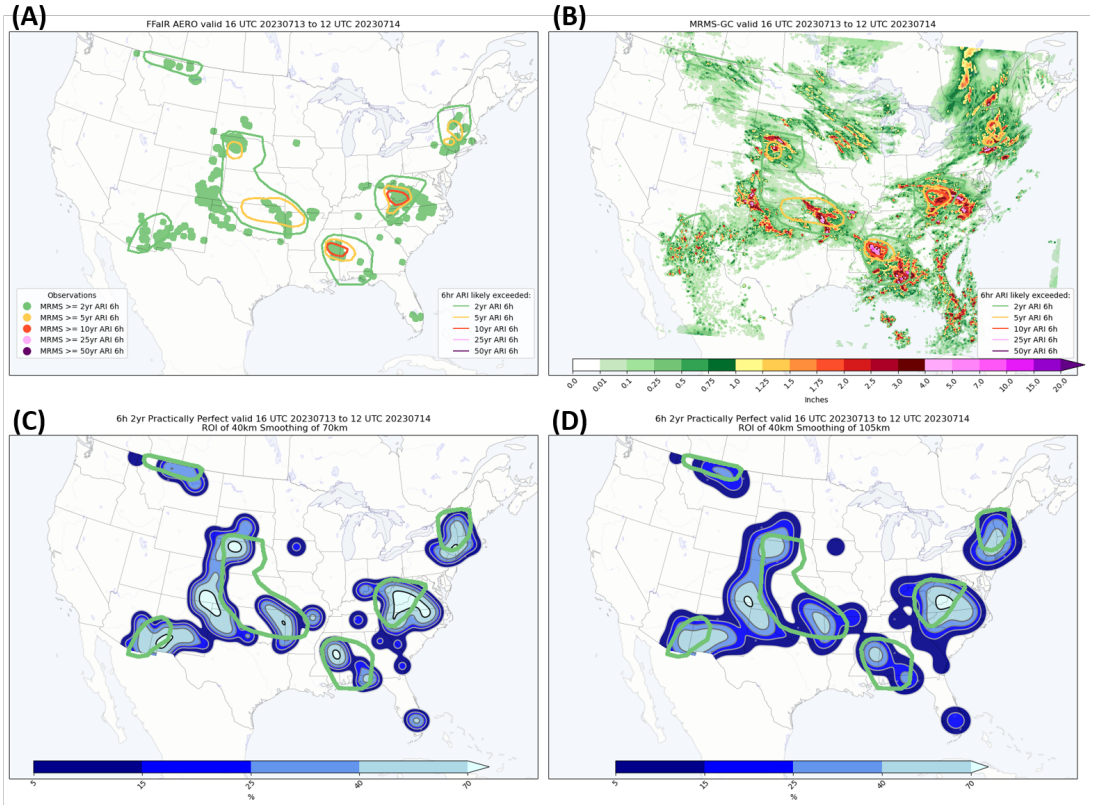

Figure 6: FFaIR AERO verification image valid 16 UTC 12 July to 12 UTC 13 July 2023. (A) FFaIR AERO with the 6-h 2-y (green), 5-y (yellow), 10-y (red), 25-y (pink), and 50-y (purple) ARI exceedances and (B) the 20hr QPE. Practically perfect for the 6-h 2-y ARI based on (C) ROI 40 km Smoothing 70 km and (D) ROI 40 km Smoothing 105 km. The practically perfect is contoured for 5%, 15%, 25%, 40% and 70% from dark to light blue.

One of the heaviest rainfall events to occur during FFaIR happened in Pensacola, FL, in the evening to early morning hours of June 15-16 2023. Nearly continuous convection trained over the area beginning around 00 UTC 16 June 2023 until finally moving out of the area around 1430 UTC and resulted in a Flash Flood Emergency in the city. In the 6-h period ending at 08 UTC on June 16 15.69" fell, with an impressive 11+ inches in 3 hours ending at 05 UTC; Fig.8A-C. For 5 hours the maximum 1h precipitation remained above 3" and for 2 of those hours exceed 4.4 and 5.4" (exceeding a 1h 100y ARI) per hour in and around the Pensacola area; an example can be seen in Fig.8D. This resulted in a Flash Flood Emergency for the area and a significant and locally life-threatening tag for the

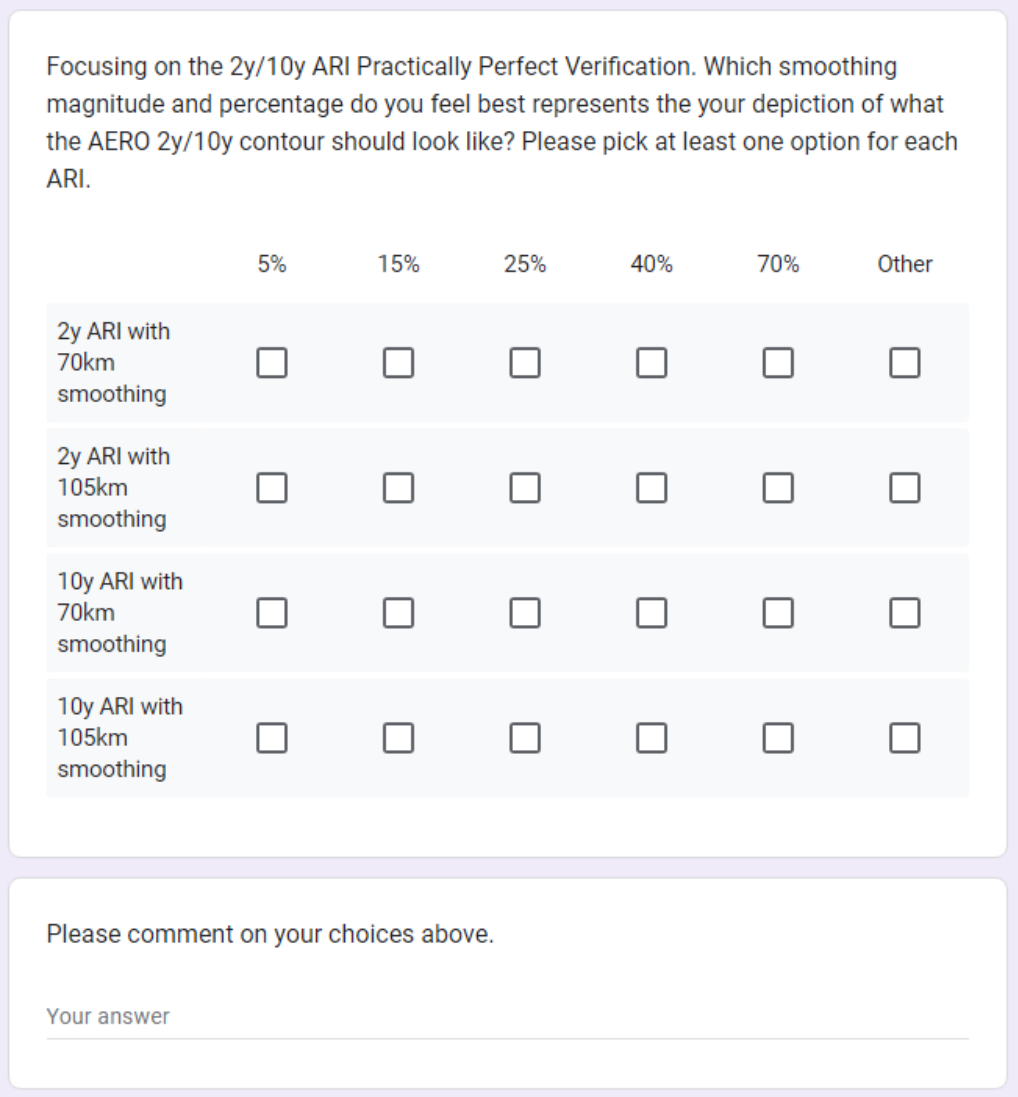

Figure 7: The question setup for the FFaIR AERO's exploratory question on practically perfect development.

WPC MPD #0497 (Fig.8E). One CoCoRaHS[3] observer recorded 11.14" (Fig.8F) and wrote in their observations notes, "Largest rainfall amount ever collected by station FL-ES-58. All but 0.02 were collected in a precipitation training event from about 7:45 PM to 4:00 AM CDT. "

[3]Community Collaborative Rain, Hail and Snow Network. The full precipitation report highlighted can be found here.

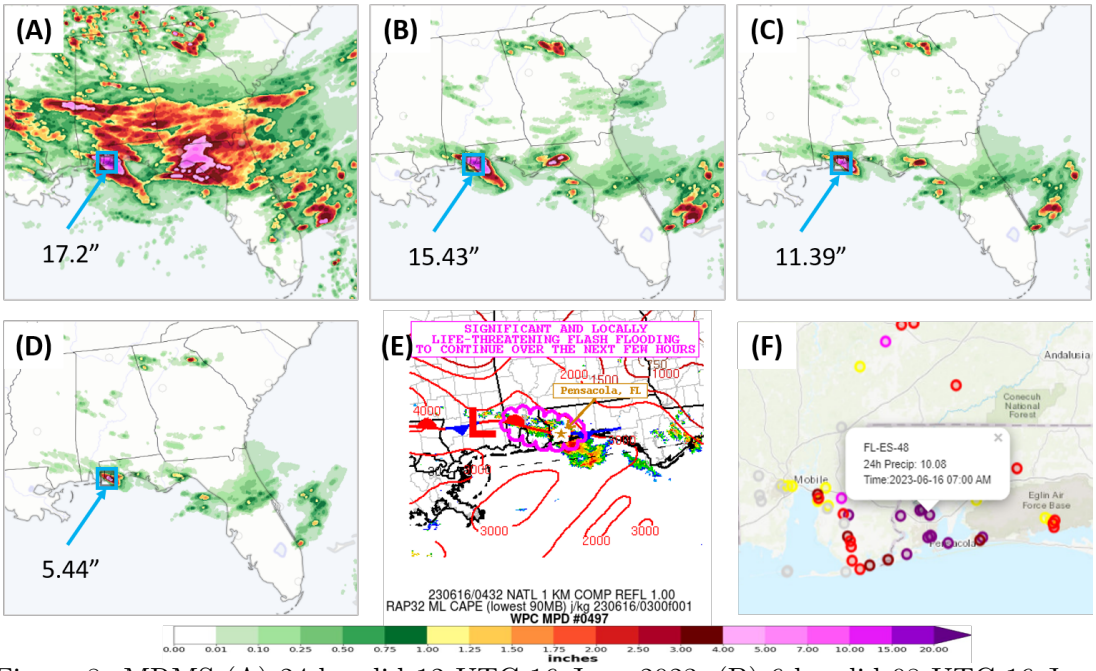

Figure 8: MRMS (A) 24-h valid 12 UTC 16 June 2023, (B) 6-h valid 08 UTC 16 June 2023, (C) 3-h valid 05 UTC 16 June 2023 and (D) 1-h valid 03 UTC 16 June 2023. (A)-(D) the maximum MRMS for each valid time/duration occurred within the blue box in and the accumulated maximum is noted. (E) Image used for the WPC MPD #0497. (F) Screen capture of the 24-h CoCoRaHS precipitation reports; purple indicates 10" or more was recorded.

Model guidance hinted at some extreme totals occurring along the Gulf Coast from MS to Gainesville FL, though the magnitude and duration was underdone. Often the models had a progressive line that only stalled for a short period of time. Additionally, across cycles the maximum location varied significantly; for instance the NAMnest, HRRR, and RRFSp1 (aka RRFSa) 06z run on June 15 all had a 24-h maximum near Gulf Shores AL but the 12z runs shifted the maximum rainfall east, into north-central FL. The progression of the 24-h rainfall over the region from the 00z, 06z, and 12z cycles of the aforementioned models can be seen in Fig.9. How this threat was handled in the ERO and the use of the Hatched region will be discussed in Section 4.2.3.2.

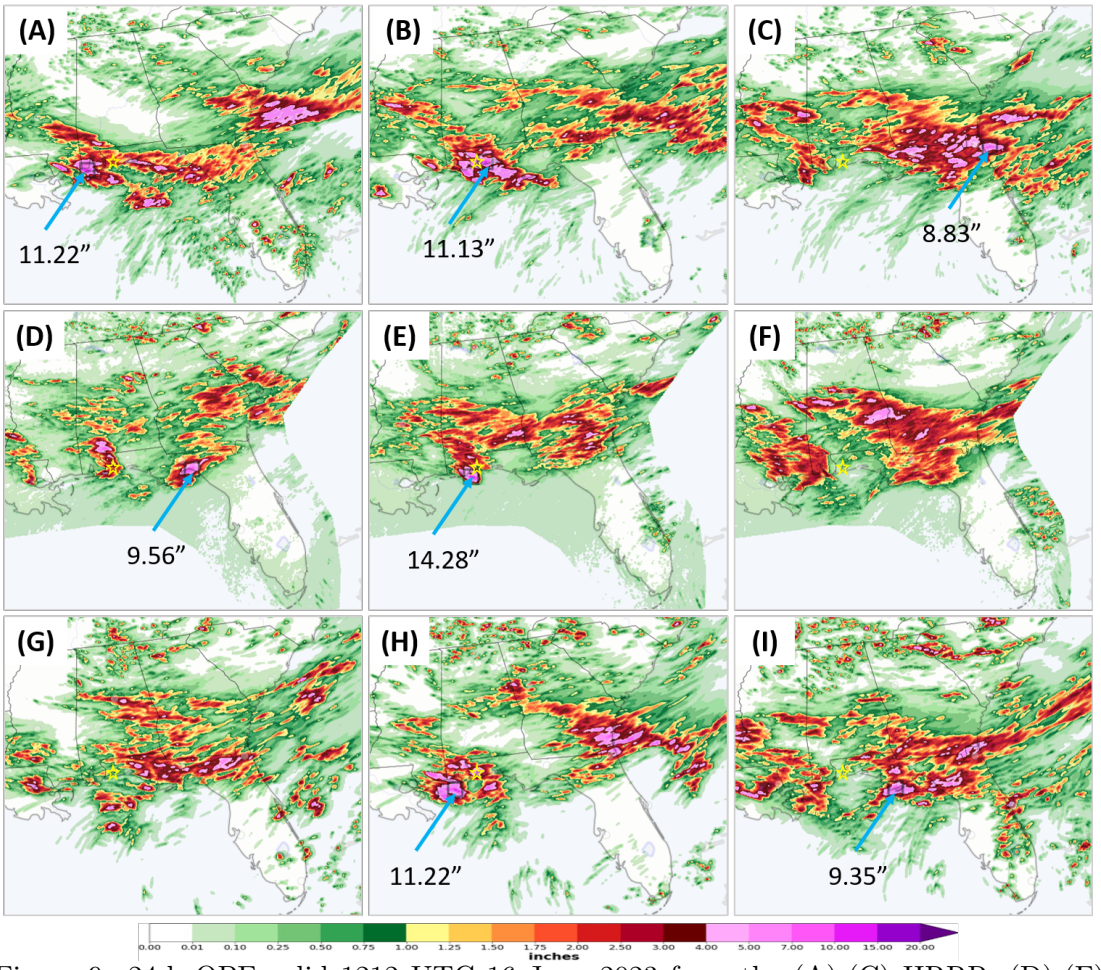

Figure 9: 24-h QPF valid 1212 UTC 16 June 2023 from the (A)-(C) HRRR, (D)-(F) NAMnest, and (G)-(I) RRFSp1 from their 00z [left], 06 [middle], and 12z [right] cycles. If the forecasted CONUS 24-h maximum QPF fell within the zoomed in domain, a blue arrow points to the location of the maximum and includes the accumulation value. The yellow indicates approximately where Pensacola, FL is.

4 Results

The results section will encompass both subjective and objective results from the 2023 FFaIR experiment. One exception is the analysis of the CIRA Satellite products where only subjective analysis was performed.

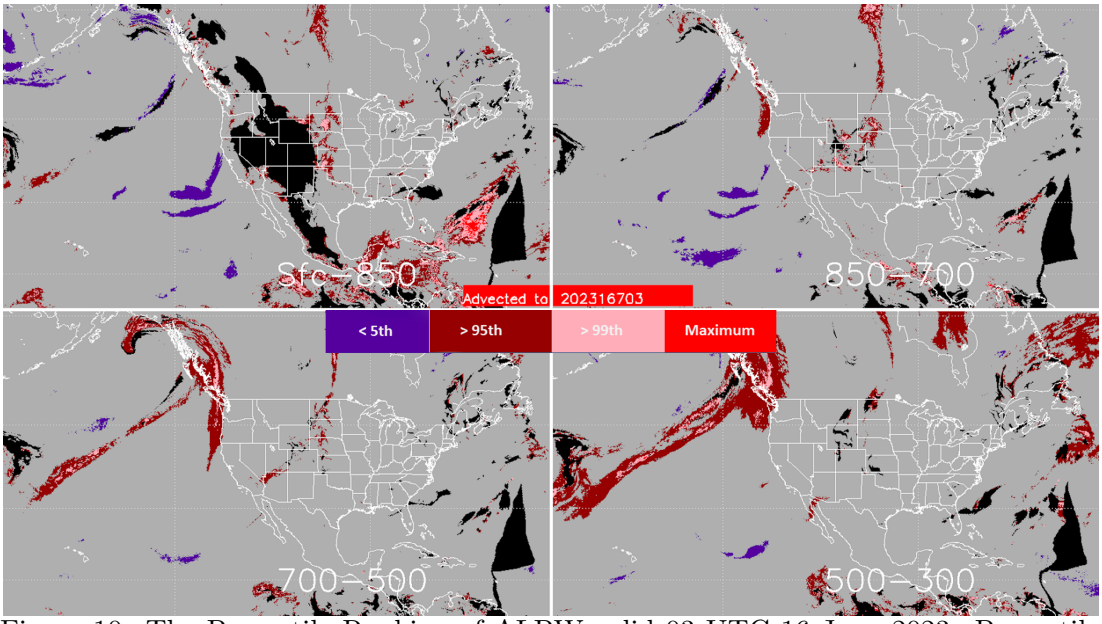

Figure 10: The Percentile Ranking of ALPW valid 03 UTC 16 June 2023. Percentiles shown are: <5th, >95th, >99th, and maximum.

4.1 CIRA at CSU Satellite Products

The CSU CIRA team provided two experimental precipitable water (PW or PWAT) based satellite products: an Hourly Percentile Ranking of Advected Layer Precipitable Water (ALPW) product and a Layered water Vapor Transport (LVT) product. An example of these two products for the Pensacola FL event can be seen in Figs. 10 and 11. As stated in the Section 2.5, these were used by the WPC forecaster in their briefings, often along with CSU CIRA's operational PWAT products, to help demonstrate the possible utility of the products in identifying heavy rainfall risk. Participants were also encouraged to use the products during their forecasting process. At the end of the week participants provided feedback on the two products.

Overall the feedback was positive about the products. Participants felt both the percentile and LVT products were a good addition to the already operational PWAT products created by CIRA. For the percentile ranking, they liked that the product helped provide some insight as to how anomalous the ALPW was, though they wished that the climatology was longer than 13 years. They felt the product

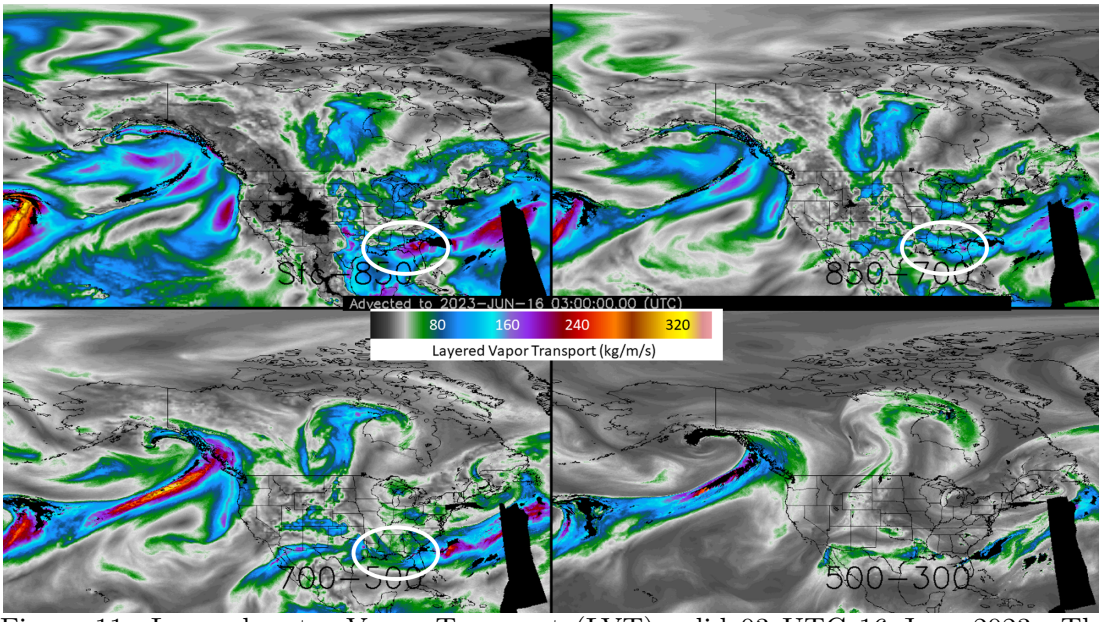

Figure 11: Layered water Vapor Transport (LVT) valid 03 UTC 16 June 2023. The layers from top left to bottom right are: surface-850mb, 850-700mb, 700mb-500mb and 500mb-300mb. The black areas missing due to terrain or precipitation. Circled are the areas of enhanced LVT over the Pensacola FL region.

was useful at highlighting where more extensive analysis for the forecast might need to be spent and helped to put PWAT values into perspective. For instance, one participant stated that "it allowed for a unique look at how impactful X inches of PWATs are for the region. 2 inches of PWATs in the Southeast is a lot different than 2 inch PWATs in the northern plains." Participants felt that adding a 90th percentile PWAT product would be of interest.

The LVT product was well liked as a synoptic analysis tool. There were numerous comments similar to this one: "LVT was really useful to get a synoptic scale and overview of the moisture for the events." Participants felt it also helped them better visualize the moisture being transported by the low/upper level jets. Perhaps more interesting was that they also saw the utility of it for post-event analysis, with one participant stating: "It sure seemed like LVT values were absolutely associated with heavy rain. In fact, I think LVT was a good post-event analysis tool in which heavy rainfall occurred (i.e., Pensacola flooding) that was not well forecast. It clearly explained the $*why*$."

For the Pensacola event, summarized in Section 2.3, post-event analysis provided by Sheldon Kusselson of the CIRA team showed that the ALPW wasn't necessarily impressive, hovering around 2", during this event. This was supported by analysis of the Percentile AWAT product, which, aside from some spotty locations, did not have percentile rankings $\geq 95^{th}$ for the duration of the event. Overall the percentile ranking looked similar to what was seen in Fig. 10 at 03 UTC 16 June. However, examination of the LVT product showed that prior to and during the event (roughly from 00 UTC 15 to 21 UTC 16 June 2023) there was nearly continuous vapor transport into the area from multiple locations and over all levels[4]. An example LVT for 03 UTC June 16 (during the middle of the event) can be seen in 11 while in Fig. 12 the 00 UTC June 15 analysis of the LVT can be seen, showing that even 24h before the onset of the event, moisture was being funneled into the region, helping to prime the environment. As can be seen by the white circles in the two Figures, over the Gulf region multiple sources of moisture transport were converging across the area at various levels. Even at 500-300mb (bottom right in Fig. 11) there is a weak signature of vapor transport from the sub-tropical jet.

4.2 Excessive Rainfall Outlook (ERO)

This section will discuss the subjective and objective evaluation of the CSU ML Day 1 (12-12 UTC) EROs and the FFaIR Day 1 (16-12 UTC) ERO. The first portion will mostly focus on the CSU ML EROs' performance against each other and the FFaIR ERO. The second part will focus on the two experimental portions of the FFaIR ERO, the Enhanced Risk and the Hatched Area.

4.2.1 CSU First-Guess Day 1 EROs and the FFaIR ERO

This is the second year that FFaIR has evaluated the FV3GEFSR and UFVS-GEFSR ERO to replace the GEFSO ERO. Last year the FV3GEFS was recommended for transition since it outperformed the GEFSO and the training method was identical aside from the version of the GEFS used; refer to the 2023 FFaIR

[4]All levels refers to the four levels the product analyse: surface-850mb, 850-700mb, 700mb-500mb and 500mb-300mb. Every layer at every hour did not always show moisture transport but for the majority of the event at least 3 of the 4 levels had vapor transport for any given hour, though often it was seen at all 4 levels.

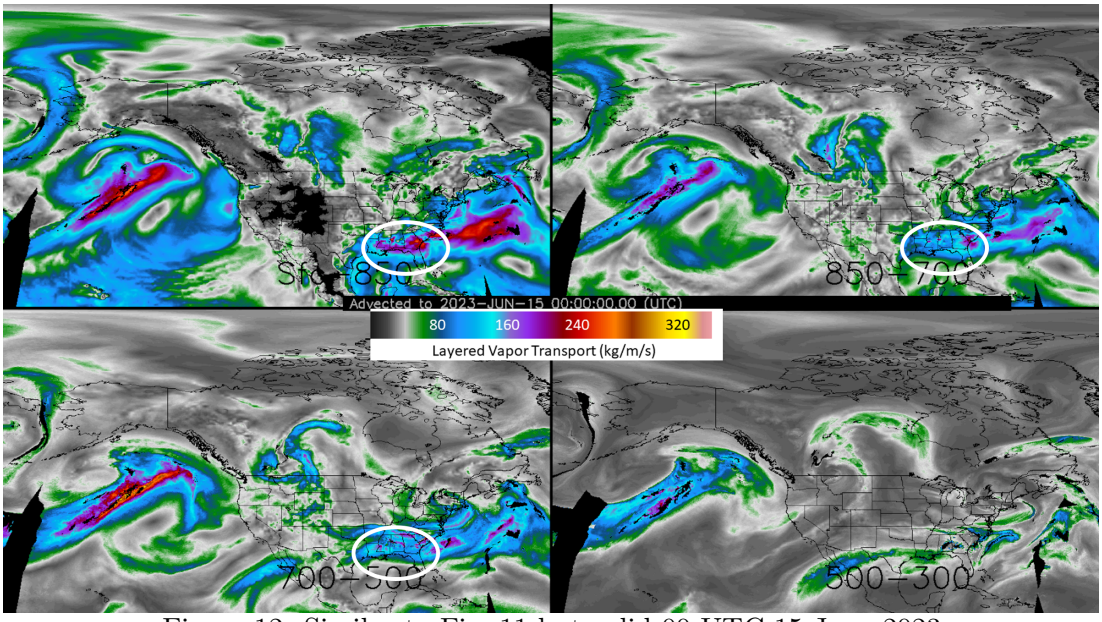

Figure 12: Similar to Fig. 11 but valid 00 UTC 15 June 2023.

Operations Plan (Trojniak and Correia, Jr., 2023a) for which versions were used. The UFVSGEFS version would have been recommended for transition as well but the FFaIR team was unsure of how the new method of observation training would perform under Moderate to High risk days since none were observed during the 2022 FFaIR Season. Thus, despite the positive feedback from participants on the product, it was determined to review the product for one more summer before making a decision on recommending to operations.

As has been the trend the past couple of FFaIR Experiments, the CSU GEFS-based EROs were well liked by participants, who found them extremely useful in the forecasting process to create their EROs. The UFVSGEFSR and FV3GEFSR were usually preferred over the GEFSO (which has been operational at WPC since 2020) and the HRRR-based ERO by participants. Figure 13 shows the percent of times each of the EROs received a score from 1 (poor) to 10 (great) over the course of FFaIR. Since the GEFSO and HRRR ERO were only available for the 00z cycle, subjective scores for the this cycle will be discussed first.

UFVSGEFSR had the highest average score of the CSU First-Guess EROs for the 00z cycles with an average of 5.45 followed by the FV3GEFRS (5.32), then

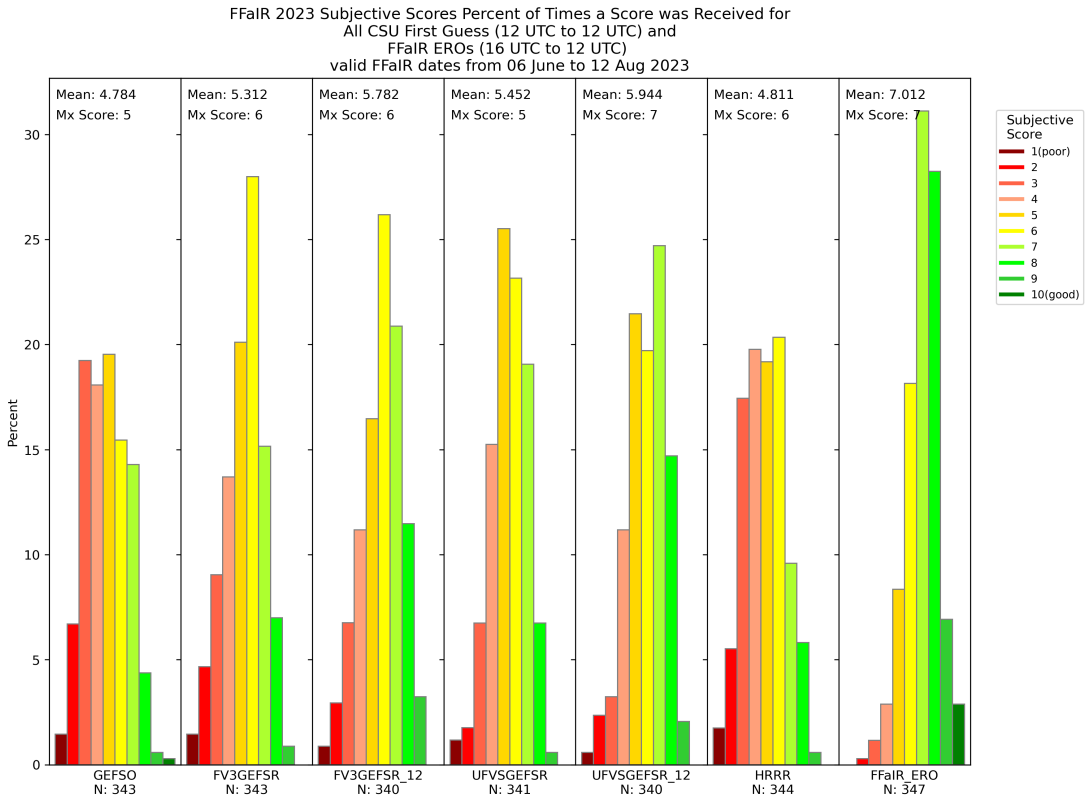

Figure 13: Results from the subjective verification for the Day 1 CSU ML (12z-12 UTC) and FFaIR (16z-12 UTC) EROs showing the percent of the time each model received a score from 1 (dark red) to 10 (dark green) during the 2023 FFaIR Experiment. The EROs are ordered from left to right: GEFSO, FV3GEFSR, FV3GEFSR 12z, UFVSGEFS, UFVSGEFS 12z, HRRR and FFaIR. The number of scores received (N) is plotted below the ERO name. The score received the most and the mean score for each model is plotted long the top.

the HRRR-based version (4.81) and lastly the GEFSO (4.78). The score most likely to be received by the FV3GEFSR was 6 (28% of the time) while for the UFVSGEFSR it was a score of a 5 (26%). However, the UFVSGEFS received slightly more scores of 7 or greater and less scores of 4 or lower (∼4% in both instances) than the FV3GEFS, thus resulting in the marginally higher average score. The HRRR also had a score of 6 as its score most chosen at 20.5% of the time. However, a score of 4 was chosen 20% of the time, followed by a 5 (19.5%) and a 3 (17.5%). This resulted in the lower average score than FV3GEFSR and the UFVSGEFSR and a score distribution that is skewed left. The GEFSO was

also skewed left, though 5 was the most likely score to be received (∼20%) followed by a score of 3 at ∼19.5%.

The FV3GEFSR and UFVSGEFSR (which in tandem will now be referred to as the FV3/UFVS-GEFSR) saw an increase in their average scores from their 00z to 12z forecasts to 5.78 and 5.94 respectively. Additionally the UFVSGEFSR experienced a jump of the value of the goodness score from a 5 to a 7, receiving a score of 7 25% of the time; it also increased in the percent of scores that were an 8 or 9 from the 00z forecast. A score of 6 remained the most likely score to be received by the FV3GEFSR, though the number of times it received a 5 or 6 decreased from the 00z forecast and increased for all scores valued 7 or higher. Finally, the FFaIR ERO was the most liked ERO, with an average score of 7.01, with over 65% of the scores being a 7 or higher. The FFaIR ERO outperforming the CSU MLP EROs subjectively is consistent with past FFaIRs and was expected.

During discussion, participants generally talked highly of the FV3/UFVS-GEFSR EROs. Participants often noted that they felt the two aforementioned GEFS-based ERO methods were extremely useful in helping to highlight areas of concern for excessive rainfall and were better than the soon to be retired GEFSO. For example, one participant wrote "the FV3 and UFVS versions are noticeable improvements over the GEFSO". Comments like "The GEFSO was underdone with its depiction of the threat. Missed out on the Slight and Enhanced areas." and "GEFSO was "OK" with coverage and location, but did terrible with magnitude of probabilities (missing two higher categories compared to PP[5])" were commonly made or written by the participants suggesting that the overall sentiment towards the GEFSO was that it would highlight the correct risk locations but usually had too low of probabilities. An example of this can be seen in Fig. 14 and by looking at the coverage of Slight Risk during FFaIR from the GEFSO compared to the FV3/UFVS-GEFSR and FFaIR EROs in Fig. 15. The preference of the FV3/UFVS-GEFSR EROs over the GEFSO was further supported by the end of the week survey question that asked the participants which First-Guess ERO they felt performed the best. None of the respondents picked the GEFSO; see Fig. 16. Although the GEFSO is slowly being phased out to be replaced with the

[5]PP: Practically Perfect

FV3GEFSR based on the results from last year's FFaIR, it is encouraging that the FV3GEFSR continues to outperform its predecessor.

As highlighted above, subjective scoring showed that participants felt the FV3/UFVS-GEFSR EROs were comparable to one another. This was further emphasized in the oral and written comments, with remarks like "they were pretty similar" or "only slight differences" often seen. Although there were days when participants noted that one version was superior, there didn't appear to be notably more days when the participants felt the UFVSGEFSR or the FV3GEFSR outperformed the other. Despite this, the goodness scores for both cycles and in the end of the week survey question mentioned above (Fig. 16) suggest that the UFVSGEFSR is preferred by the participants over the FV3GEFSR.

One driving factor to this was likely the look of the UFVSGEFSR probability contours compared to the FV3GEFSR's. Throughout the duration of FFaIR, participants commented that despite similar forecasts from the FV3/UFVS-GEFSR EROs, they leaned towards the UFVSGEFSR's forecast because it looked more coherent and less jagged than the FV3GEFSR's. An example of what they were referring to can be seen in Fig. 17. When looking at these verification graphics, the forecasts from the two systems are relatively similar but the contours look smoother for the UFVSGEFS's forecasts (middle two images) compared to the FV3GEFSR's forecasts (top two images). Additionally, rather than having multiple, spatially small contours of Enhanced risk in close proximity to one another in MT like the FV3GEFSR has, the UFVSGEFSR had a single, larger area of Enhanced risk over the same area (implying coherency). The following comments were made about these features on this day:

- "The FV3 continues to have jagged lines, along with too many distinct MRGL "bubbles" in central CONUS on the 12Z which may be best to be smoothed out."
- "FV3 had a smaller enhanced area than UFVS, though again had funky contours. UFVS smoother and more realistic, and seemed to capture the heavier rain area better."

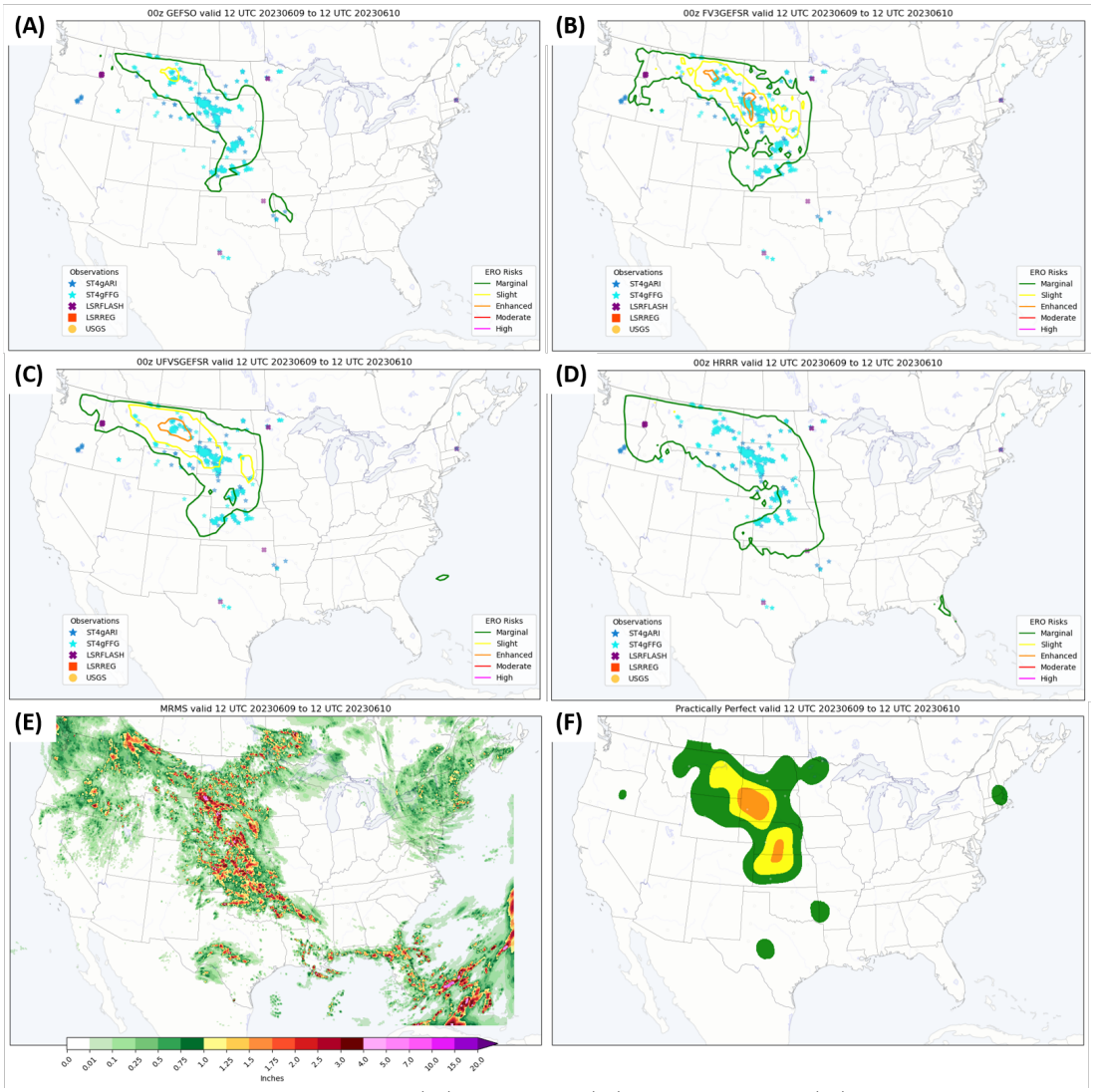

Figure 14: CSU First-Guess ML (A) GEFSO, (B) FV3GEFSR, (C) UFVSGEFSR, and (D) HRRR Day 1 EROs, initialized at 00z valid 12 UTC 09 June to 12 UTC 10 June 2023. (E) MRMS QPE and (F) Practically Perfect valid for the same time. Overlaid on (A)-(D) are the UFVS Observation Dataset. Risk categories - Marginal: 5%-15% (green), Slight: 15%-25% (yellow), Enhanced: 25%-40% (orange), Moderate: 40%-70% (red) and High: >70% (purple/pink).

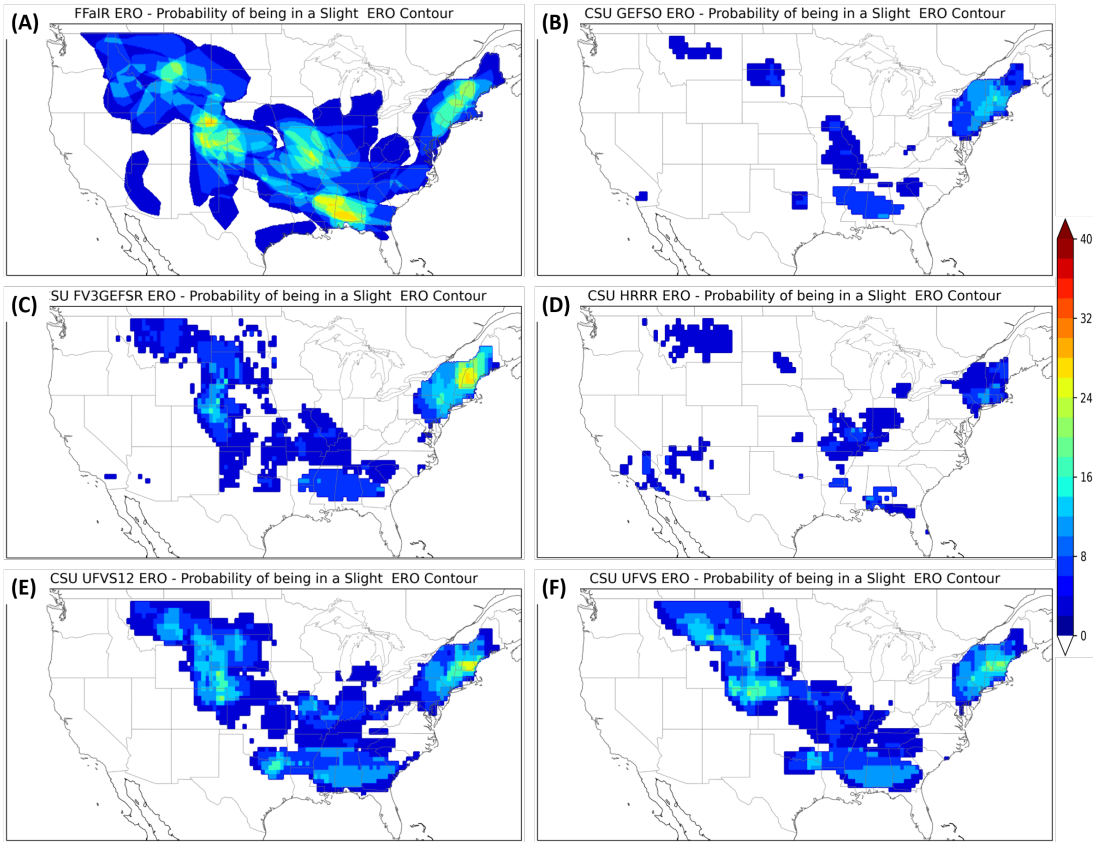

Figure 15: Probability of being in a Day 1 ERO Slight Risk during FFaIR 2023 for the (A) FFaIR, (B) GEFSO, (C) 00z FV3GEFSR, (D) HRRR, (E) 00z UFVSGEFSR, and (F) 12z UFVSGEFSR EROs.

- "The UFVSGEFS did better at delineating separate higher prob areas across ND and MT than in the FV3GEFSR. However, both the 12z FV3GEFSR and UFVSGEFSR had a noticeable improvement over ND (higher probs) than their respective 00z runs for this forecast day."
- "UFVS had better areal coverage in Montana which was better."

Comments like these, especially those similar to the first comment listed, were constantly made both during verification and when the MLP EROs were being used in the forecasting process. Even in the end of the week comments, participants noted that they didn't like the jagged look of the FV3GEFSR. This perhaps summed up best by this comment: "UFVSGEFSR was much much smoother (I think that was noted in last year's evaluation too), and I liked that from just an interpretation

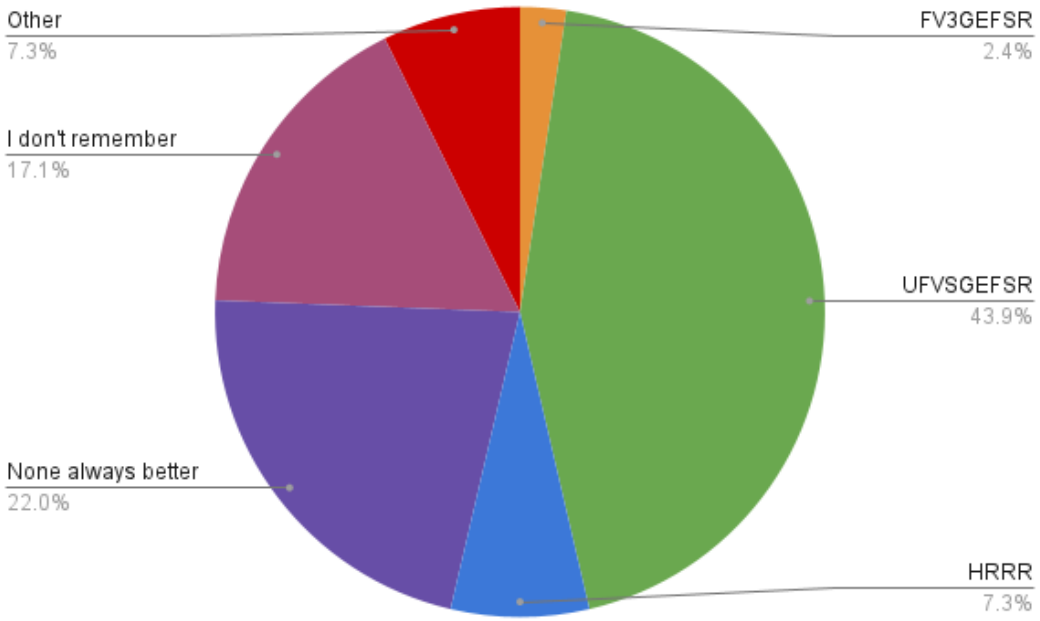

Figure 16: Results from the end of the week survey question which asked participants: "Did you feel there was one MLP ERO configuration that performed the best (most consistent) throughout the week?". Participants were allowed to write in a comment rather than pick a MLP ERO, this is referred to as other.

point of view (e.g., were the small-scale "features" of the FV3GEFSR meaningful or just artifacts of the forecast process? I wasn't sure how to evaluate those small-scale features). But generally, the two systems seemed very comparable."

The differences in the observation training set between the two also likely played a role. The UFVSGEFSR is trained on the entire verification dataset used to create the practically perfect that the ML EROs are verified against rather than just ARI exceedance and flood reports. Therefore, the UFVSGEFSR version, in theory, has a better representation of "truth" since the question addressed is "what is the probability that there will be a report in the UFVS" rather than "what is the probability that there will be an ARI exceedance or flash flood report" given the same meteorological input for training and the same forecast dataset[6].

[6]Thank you Russ Schumacher for this great summary of the differences.

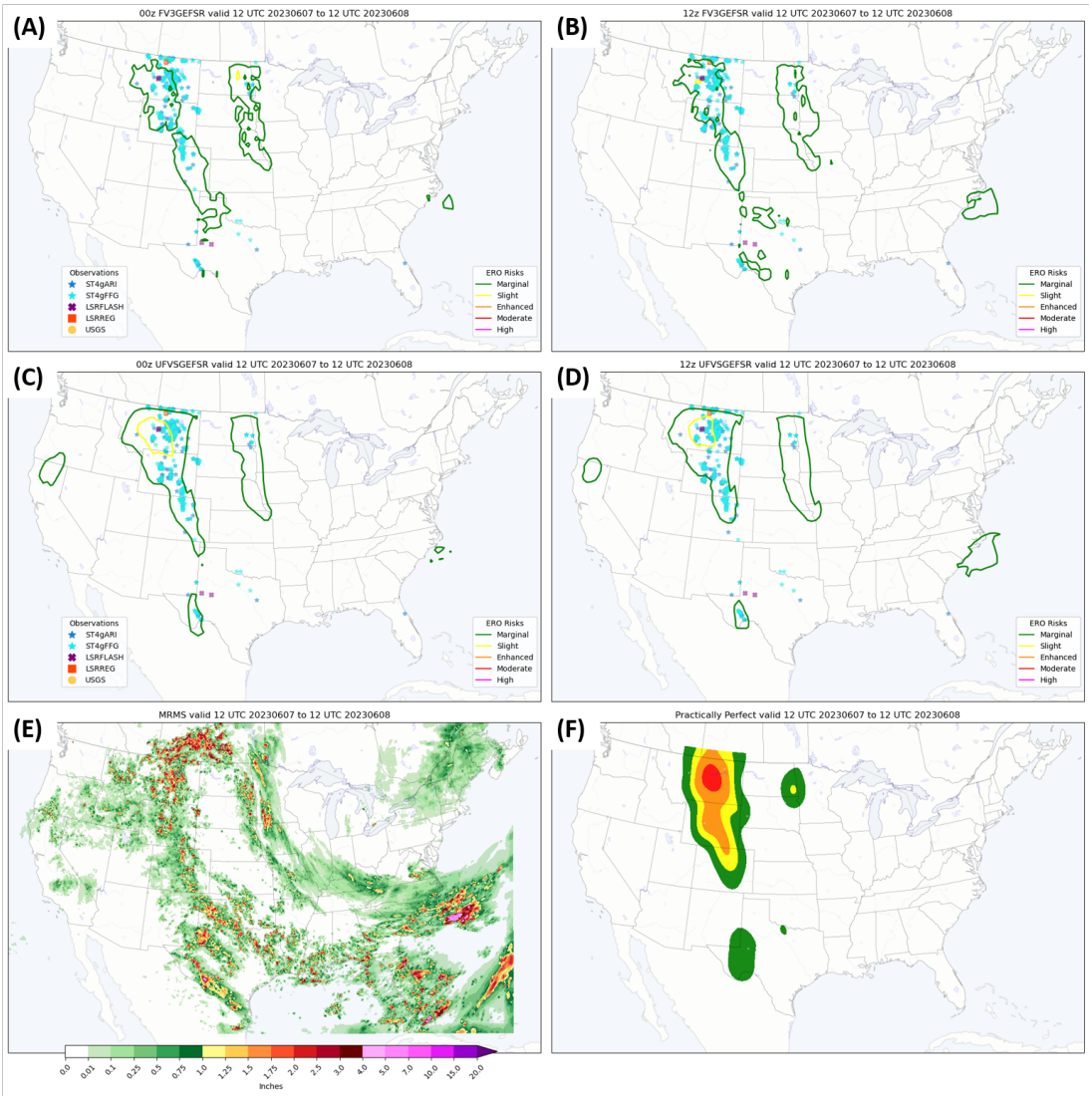

Figure 17: Similar to Fig. 14 but for the (A) 00z FV3GEFSR, (B) 12z FV3GEFSR, (C) 00z UFVSGEFSR, and (D) 12z UFVSGEFSR EROs valid 12 UTC 07 June to 12 UTC 08 June 2023.

In regards to the weekly survey results, it can be seen in Fig. 16 that the UFVSGEFSR was the overwhelming favorite at 42% while the FV3GEFSR only had 2.3% of "the votes". In fact, the HRRR-based ML ERO had a higher percentage of preferred votes at 7.1%. When looking solely at these results it might seem as though the FV3GEFSR was not well liked but the daily subjective verification results and discussions suggest otherwise. What might be the cause of

this misleading result was found when reading the responses of the participants for why they chose the model they did. A majority of the responses that picked the UFVSGEFSR wrote that they felt the FV3/UFVS-GEFSR MLPs performed similarly but that overall the UFVSGEFSR was a bit better. Some did not explain why they felt it was better while others mentioned the look of the UFVSGEFSR compared to the look of the FV3GEFSR as the reason. For example, a participant that picked the UFVSGEFSR as the preferred First-Guess product wrote "UFVSSGEFSR seemed to have the smoothest, most realistic ERO feel." Another wrote "FV3GEFSR and UFVSGEFSR were my clear favorites. I think notably both of those outperformed the GEFSO version routinely. On the more "marginal" days the HRRR version was OK, but it did not seem to do as well on the "higher end" rainfall days. So that is complicated ranking of sorts." Finally, participants were allowed an "Other" option that was a write in. There were only a couple of these but they generally followed this theme: FV3 and UFVS were best and comparable.

As noted in the summary of the subjective scores, the HRRR-based ERO just barely edged out the GEFSO version in terms of average score (4.81 v 4.78). A large complaint about the HRRR-based ERO by the participants was that, like the GEFSO, it rarely forecasted risks higher than a Marginal. As can be seen in Fig. 15 the probability of being under a Slight Risk from the HRRR ERO was 0 across the majority of the country. However, when looking at the probability of being under a Marginal Risk (Fig. 18) the HRRR appears to slightly overforecast the risk in certain areas of the country with respect to the FFaIR ERO, such as in FL and the Southern Mississippi Valley into the Southeast.

In FL, it is hard to say if this is actually a high bias or rather an artifact of forecast methodology by participants and WPC forecasters. Often during the creation of the ERO, participants would acknowledge the nearly daily risk there but chose not to draw for it because it rarely verified due to the tendency of urban street flooding to not be logged as a flood or flash flood report in the LSR dataset. In fact, many of the participants noted that they liked that the HRRR was identifying the risk in FL that often was not shown in the GEFS MLPs. The bias across the Southern Mississippi Valley into the Southeast might also be misleading,

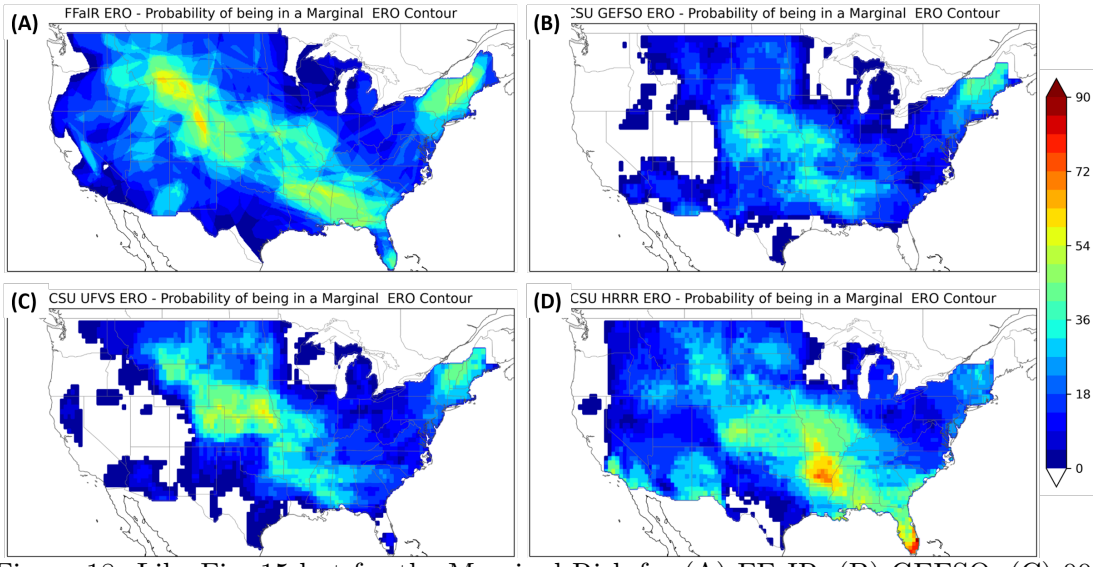

Figure 18: Like Fig. 15 but for the Marginal Risk for (A) FFaIR, (B) GEFSO, (C) 00z UFVSGEFSR, and (D) HRRR.

but for a different reason. The HRRR-based ERO, like the GEFS-based ones, are valid 12 UTC-12 UTC while the FFaIR and Operational EROs are valid 16 UTC-12 UTC. This means that if there is a risk of excessive rainfall between 12 UTC and 16 UTC, the FFaIR and Operational EROs would not forecast for it but the MLP EROs would. Therefore it is possible the the HRRR ERO was identifying this risk. It would be beneficial for the developers to look into this further.

Finally, subjectively the participants felt the HRRR-based ERO has improved from last year. Fig 45 in the 2022 FFaIR Final Report (Trojniak and Correia, Jr., 2022) shows that in last year's experiment the average score for the HRRR ML ERO was 3.41 vs 4.81 this year. Additionally, the distribution of scores shifted from a strong skew to the left (poorer forecasts) towards a more uniform distribution centered around a score of 4/5. Last year, the score most likely to be received was a 2 and this year it was a 6, closely followed by a 4 then a 5. Furthermore, the HRRR ML ERO went from roughly 6% of the score received being a 7 or higher to 16% . Therefore, although the HRRR did not perform as well as the FV3/UFVS-GEFS ML EROs, it did perform better than last year's version, at least subjectively. That said, like last year, participants still noted that the Marginal Risk seemed too large and over done.

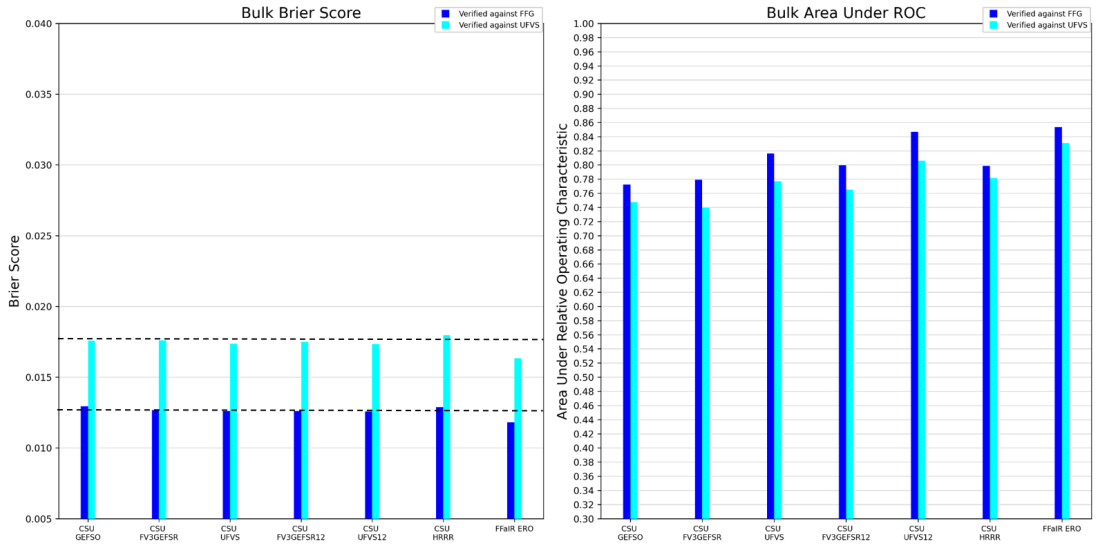

Figure 19: CSU ML and FFaIR EROs' [left] Brier Score (BS) and [right] Area Under the Curve (AUC) Receiver Operating Characteristics (ROC) verified against FFG exceedances only (dark blue) and the UFVS (light blue) across the 2023 FFaIR experiment days.

The Brier Score (BS) and Area Under the Curve Receiver Operating Characteristics (AUROC) for the CSU and FFaIR EROs can be seen in Fig. 19. Agreeing with the subjective scores, the UFVSGEFSR ERO overall was the best performer among the CSU MLPs. In fact, the 12z UFVSGEFSR ERO had a similar AUROC to the FFaIR ERO when verified against FFG, around 0.86. When verified against the UFVS dataset, the FFaIR ERO was better. This makes sense, given that although the UFVSGEFSR is trained using the UFVS as its verification dataset, the MLP itself does not know anything about current conditions like ongoing flooding, which are factored into forecasters' analysis for risk for the day. Interestingly, the HRRR-based ML ERO performance when using these two metrics does not lag behind the models as one would have expected based off the subjective results and feedback. In fact, when looking at the AUROC, the HRRR ERO slightly outperforms the 00z FV3GEFSR and the GEFSO for both verification datasets.

Figure 20 shows the fractional coverage for the operational (left) and experimental (right) risk categories. Overall the CSU and FFaIR EROs are well calibrated to the operational Marginal and Slight risks, thus reliable. Although the GEFS-based EROs approach the upper threshold for the Marginal risk while

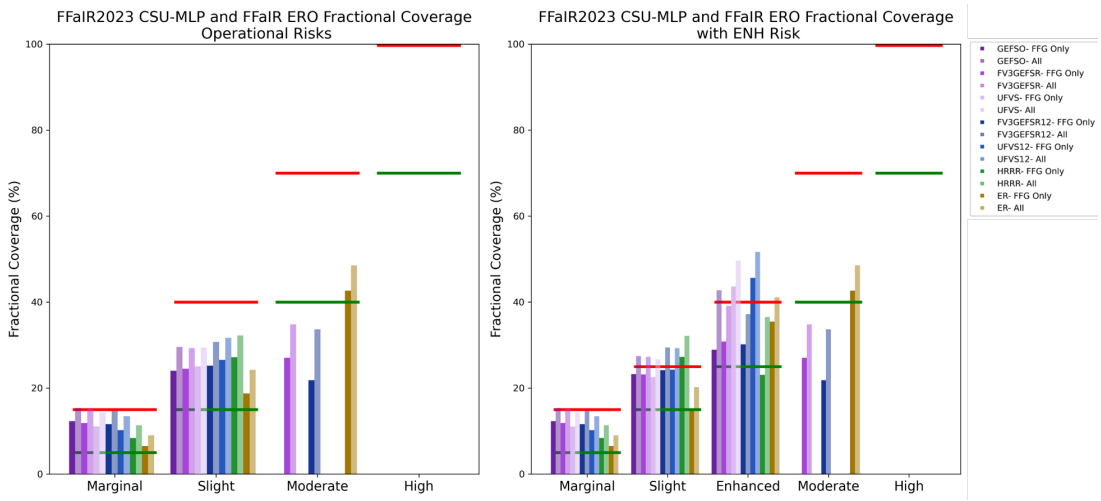

Figure 20: CSU ML and FFaIR EROs' fractional coverage using the (left) operational risk categories and (right) the experimental risk categories (aka the inclusion of the Enhanced Risk) verified against FFG exceedances only (darker shade) and the UFVS (lighter shade) across the 2023 FFaIR experiment days. The 00z GEFS-based EROs are in the purples, the 12z GEFS-based EROs are in the blues, the HRRR-based ERO is in green and the FFaIR ERO is in gold.

the FFaIR EROs fall on the lower end. This could suggest that the GEFS-based Marginals are too small, while FFaIR EROs have Marginals that generally are too large. A similar conclusion could be made for the FFaIR Slight Risk, though to a lesser extent. The 00z UFVSGEFSR and the 12z FV3GEFSR were the only CSU ML EROs that had some success in forecasting a Moderate Risk, though their fractional coverage forecasts tended to be too low; meanwhile, the FFaIR ERO outperformed the CSU EROs, proving to be more reliable for this risk category.

Focusing on the fractional coverage with the Enhanced Risk included (right side of Fig. 20), the reliability of all the EROs shifts dramatically when the Slight Risk (15%-40%) is split into the experimental Slight (15%-25%) and Enhanced (25%-40%) Risks[7]. The shift in the performance of the FFaIR ERO is not surprising, at least in the team's eyes. Generally we have noticed that participants' perception of probabilities can vary greatly between individuals. However, the more often that a probability space is seen, the more likely participants are to

[7]Unless noted, the discussion will focus on verification using the UFVS dataset (labeled as All in Fig. 20).

correctly perceive the probability; here a probability space refers to the lower and upper bound probabilities for each ERO risk category. Therefore, changing the probability space, in this instance making it smaller and adding a new category and thus a new probability space, disrupts their mental picture of what the risk should look like. Another driving factor was likely that the participants were overly cautious at upgrading the Slight Risk, wanting to draw smaller and more precise than they would for the operational Slight. This is shown by the fractional coverage of the FFaIR ERO exceeding 40% when the UFVS dataset is used.

The factors that impact how the FFaIR ERO was created, such as participants being overly cautious at upgrading the Slight Risk, do not factor into the CSU MLPs. Despite that, when the Slight Risk was broken into the Slight and Enhanced Risk definitions, the reliability tended to decrease; i.e. the MLPs were less calibrated to the new definitions. This was seen in all the versions of the MLPs but the 00z/12z UFVSGEFS reliability seems to be the most impacted by this redefining of the ERO categories. The 12z UFVSGEFSR has a fractional coverage around 50%, suggesting that when the UFVS MLP output probabilities fell within the Enhanced Risk range, the area they covered verified as a Moderate. This is interesting, given that the GEFS-based MLP have been noted to be "hot" by participants in the past so one would expect the fractional coverage of the MLPs to fall below the high end of the risk definition (the red lines in Fig. 20). Overall this suggests that the GEFS-based EROs, and to some extent the HRRR-based ERO, are underforecasting the areal extent of the risk or missing events entirely.

In general, the overarching results fall in line with what has been noted in the past; the participants like the CSU ML EROs and find them useful in guiding them in the ERO forecasting process. The HRRR-based and GEFSO EROs subjectively performed worse compared to the FV3/UFVS-based EROs, though the difference is muted in the objective results. This suggests perhaps that having the guidance provide general regions to focus upon is more important than the MLP's ability to accurately predict the exact risk category. The UFVSGEFSR ERO overall outperformed the GEFSO and FV3GEFSR in both verification methods, thus the FFaIR team supports its transition into operations.

4.2.2 Additional Discussion on the FFaIR ERO's Enhanced Risk Areas

The introduction of an additional risk contour for the ERO, which followed the internal WPC experiment being done, was relatively welcome for the participants. The collaborative FFaIR ERO issued 23 enhanced risk areas over the course of 15 days. From the WPC practically perfect observations there were 26 enhanced risk areas over the course of 18 days. The collaborative ERO overlapped 12 of the areas with observations. There were 7 days where an Enhanced Risk was observed but none were forecast. This suggests that not drawing enough Enhanced Risks could have helped drive the fractional coverage distribution described above and seen in Fig 20).

The dominant theme in the collaborative discussion was simply specifying the higher potential for flash flooding with more confidence in location. Most participants drew Enhanced Risk areas and expressed the desire to see this addition in operations. The operational gap between Slight and Moderate is large and also perceived as such. According to almost all (35/38 responses) participants, the Enhanced Risk added information and value to the ERO forecast. Some comments were:

- "It was good having an additional option - with moderate has the second highest contour, you have to be pretty confident there's going to be some flooding within your moderate to draw it. At the same time a slight is basically isolated flash flooding. Enhanced slight came in handy for perhaps less confident moderate or perhaps greater coverage than a slight."
- "(i)t provides higher precision to the risk while still providing a wide enough range to recognize a difference."
- "I like the idea of the enhanced. There's a huge gap between the slight and moderate and I like the gap being bridged. I don't really like the name due to its ambiguity, but I don't like it in regards to SPC outlooks either."

4.2.3 The FFaIR ERO's Hatched Areas

The so-called Hatched risk was intended to be an Intensity based area focused uniquely on the heaviest rainfall potential. This area need not be specified inside an ERO risk category, since this category is exploratory and experimental. The precedence for this area dates back to older ERO products from HPC/WPC as a 5" rainfall contour on the ERO which was removed in 2018[8]. The hope was to highlight areas where the intensity of the rain could be impactful but not necessarily lead to flooding/flash flooding. An example would be rain falling faster than it can runoff into the storm drains, causing a thin layer of water on roads increasing the likelihood of hydroplaning. Once the rates decrease, the water clears from the roads.

The initial starting point definition was exceedance of the 6-h 10-y ARI, as previous experiments in the AERO group have shown this to be relatively rare and also intense enough to be of concern for flash flooding. However, this definition was seen as limiting since not all heavy rainfall events can be encapsulated uniquely by 6 hour rainfall. With that in mind, the facilitator also entertained the idea of using rainfall "rates" as a proxy for Intensity at multiple time scales. This was a device deployed to keep this activity experimental and exploratory, which had many positive and negative effects on the participants, these are discussed below.

4.2.3.1 Bulk Evaluation

The first critique, consistent across all 6 weeks, was the conflation of the hatched graphical depiction to the SPC significant severe (also referred to as hatched) criteria. Participants raised two objections: the similarity between hatched graphical depictions and the need for stringent definitions, similar to SPC. For the latter, communication of what hatched areas mean and what they should convey on the ERO were frequent topics of conversation. Many participants believed hatched, as SPC does, should convey increased risk for severe weather. In the ERO depiction some believed it should only convey some aspect of "Rarity" via Impacts (in this case flash flood potential). In so raising this concern, other participants

[8]We could not find an exact date in which this practice ended.

expressed the concern that the existing risk categories already convey such risk and that hatched then should be used as a higher category or only for the higher categories. These conversations about Intensity had transformed back into flash flooding risk, which the facilitators were hoping to keep the Hatched contour from being associated with.

The second critique laid into the facilitators use of hatcheing outside of the ERO risk areas. This occurred very often in Florida where ARIs are high but the corresponding flash flooding is seen as either nuisance (or temporary and brief) or extremely unlikely outside of urban corridors. The use of the hatching could also be envisioned in the intermountain western US where precipitation rates are relatively low for even 6h 10y ARIs but where any general thunderstorm rainfall extremes could form in climatologically unfavorable areas. While these could be envisioned as areas of concern, there was much trepidation about using the hatched in a way that did not conform to the rarity of Impacts as previously mentioned.

The third critique was mostly concerned with the areal coverage of the hatched areas and thus by extension its specificity of the threat. On most summer days any thunderstorm is capable of producing rainfall that is intense. The two ingredients approach offered by Doswell et al. (1996) can be summarized as: "..the heaviest precipitation occurs where the rainfall rate is the highest for the longest time". And indeed for a singular event this rings true, and for other events such as training the only additional component would be the number of high intensity long duration events that can move over an area in some finite time window. This leaves forecasters in a dilemma since thunderstorms in summer are very numerous and encompassing any and every thunderstorm seems an impossible task. To narrow down the search, the group turned to using the rainfall rates, here defined as the rainfall accumulation over a specific time-period such as 15-min or an hour-h. The facilitator of the activity used this terminology, again as an exploratory device, to use more than the 6-h 10-y ARI (e.g. 1 and 3-h 10-y ARI) and also the sub-hourly ARI. To simplify matters, rainfall rate thresholds like 1.5-2" per hour or 4" per 3 hours were also used to make the forecasting task simpler so as to not have to consult every time window ARI map depiction in every forecasting session. In retrospect, while this worked for the facilitator, who had many weeks

of preparation and many more weeks of practice than the participants, this was not a skill easily picked up in days for the participants.

In regard to the ARI, post experiment work examined the Atlas-14 ARIs available during the experiment (Fig. 21) and the difference between the time windows, i.e. 1-h, 6-h, 24-h. One way to identify extremes based ARIs is to use a particular duration, say 1-h, and use the least common threshold across the CONUS. For example, 1-h 10-y ARI has the smallest eastern US areal coverage for 2.5 in. So for 1-h precipitation accumulations we use 2.5 in as a proxy for intensity and the evaluation of the Hatched area. Likewise we have chosen 4 and 5 in in 3 and 6-h using the same basic principle. In addition, the difference between the temporal windows shows that for a large part of the central and southern US, the rates differ by about 0.75 - 1.5" (25-50%) for three and six hour with a background near 3". Thus for any random event with a large magnitude, the 10-y ARI can be met across a range of temporal windows. Herman and Schumacher (2018) found that higher thresholds of precipitation rates and ARIs do not correspond well to flash flooding and its impacts, essentially that there are many ways (intensities and durations) to cause flash flooding. So, even though a high precipitation rate or ARI were exceeded, no individual threshold could be used to described flash flood occurrence. In fact, verification was better as thresholds were lowered.

The number of hatched areas for all 30 EROs are shown in Fig. 22A. The number of areas drawn on any given day via consensus forecasting was around three to five. Florida had the most number of days, with 14 polygons drawn, mostly due to the persistence of heavy rainfall associated with sea breezes. Elsewhere, almost all parts of US had some expectation of heavy rainfall through the 30 events we forecast for and matches the accumulated precipitation for the FFaIR dates.

A number of practically perfect like probabilities of rainfall exceedances can be seen in Fig. 22B-D, meant to represent the MRMS data across a range of thresholds. To summarize the practically perfect depictions the areas meeting a threshold of 25% are summed to show occurrence frequency over the FFaIR experiment. The chosen thresholds for each depiction match both the rarity of events and are larger than the scale of the rainfall footprints. However, these

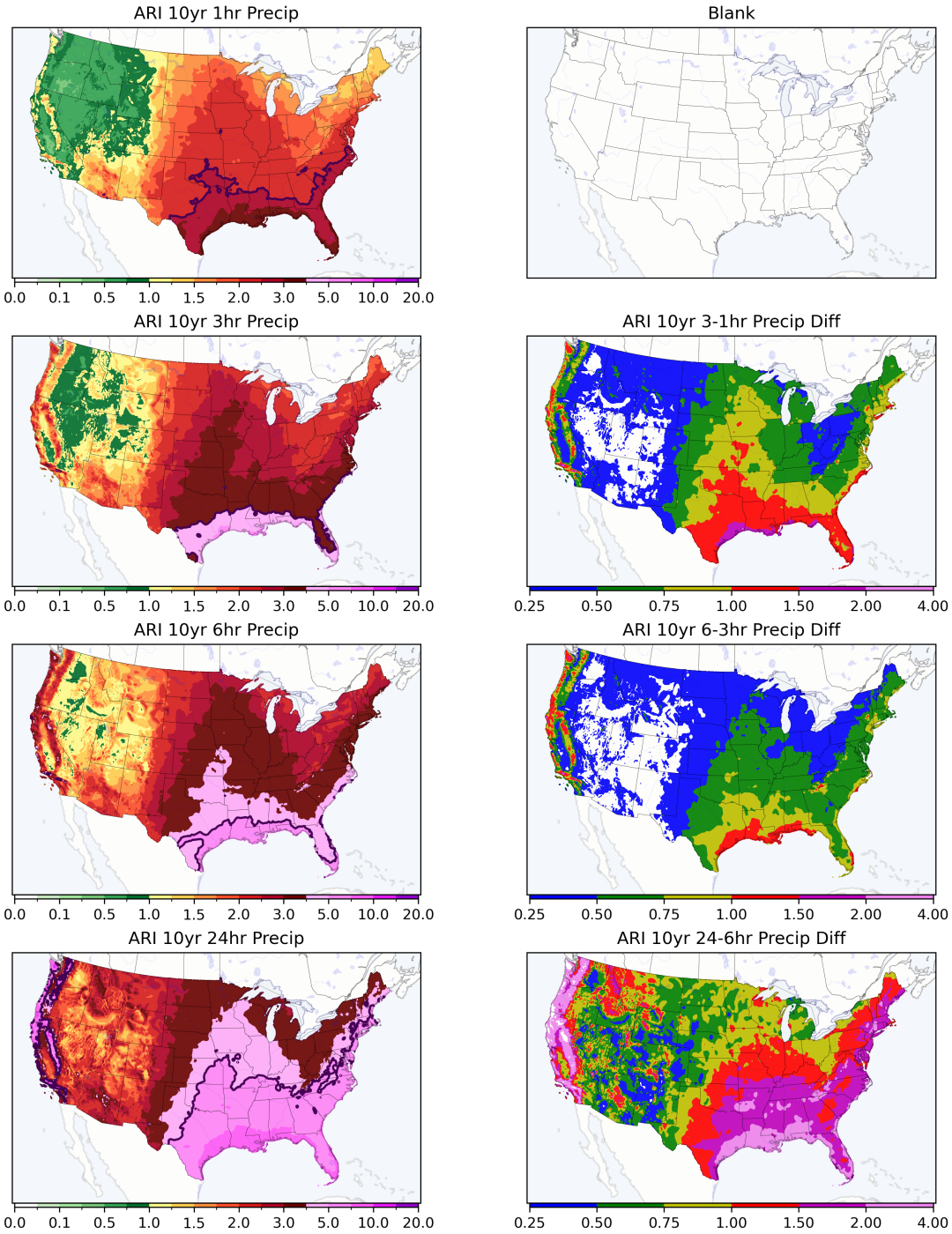

Figure 21: [Left] A depiction of the 4 10-y ARI for temporal windows of 1,3,6, and 24 hours along with [right] the difference between three and one hour, six and three hour, and twenty four and six hour windows.

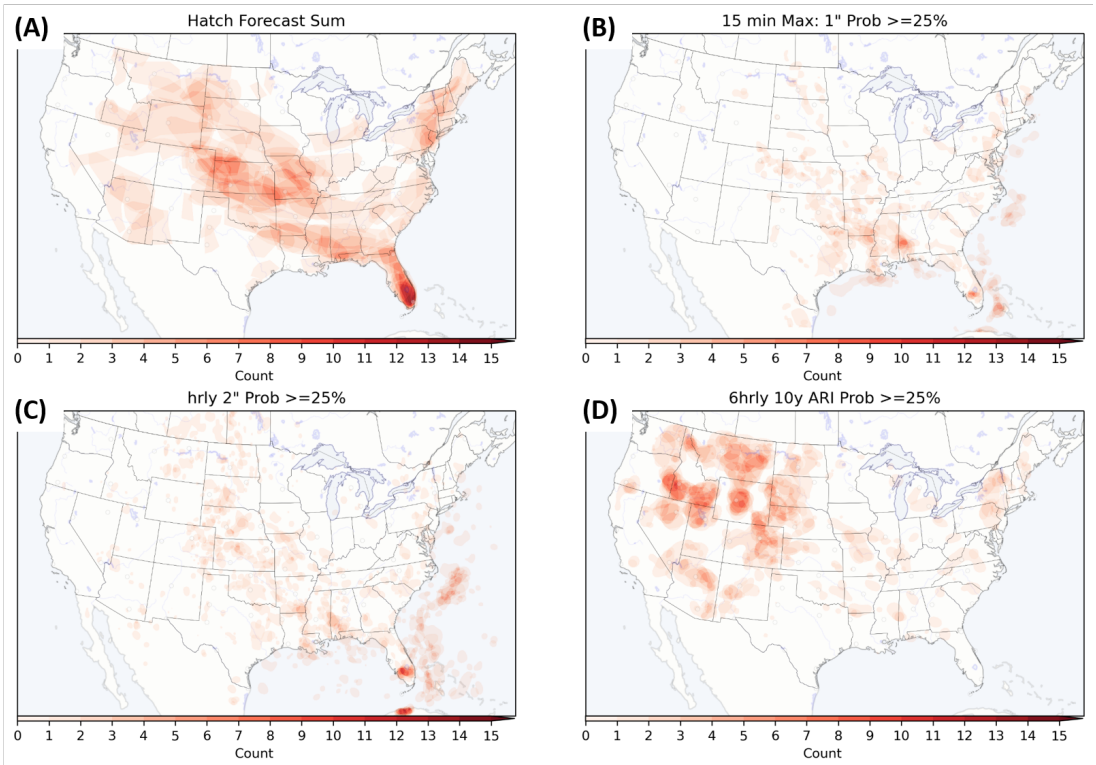

Figure 22: The frequency of (A) hatched areas drawn for the FFaIR ERO, (B) 1 in 15 min^{-1} prob $> 25\%$ areas, (C) 2 in hr^{-1} prob $> 25\%$ areas, and (D) 6-h 10-y ARI prob $> 25\%$ areas.

choices were tested in terms of gaussian smoothing and radius of influence, and while sensitive, the choices made were based on discussed rainfall rates during the experiment for sub and hourly rainfall rates as well as ARI.

None of the observational aggregated quantities match particularly well to the aggregate forecasts as predictability of the events was somewhat low throughout the experiment. The key points are that even the observational data do not resemble each other en masse, rather they each emphasize something unique about the events. That said, individual events can have all 3 probabilities overlap for the most extreme events - lasting the longest and having the highest rainfall rates across a range of accumulation windows. With this in mind, we proceed to evaluate the forecasts of a few noteworthy events and graphically display the critiques mentioned above.

4.2.3.2 15-16 June 2023

One of the heaviest rainfall events to occur during FFaIR happened in Pensacola, FL and is summarized in Section 3. Farther east, multiple 5 in 6 h^{-1} events throughout the early evening and early morning contributed to creating a larger and higher magnitude probability area as well. These events were encapsulated in a WPC slight risk area and all ERO FFaIR participants were in agreement for both a Slight and Hatched Risk across the Florida Panhandle. The main driver for the Pensacola event was convection training along a remnant boundary from a series of mesoscale convective clusters that eventually formed multiple lines into northern and eastern FL. The practically perfect summary in Fig. 23 shows that all rate depictions had higher probabilities across the FL panhandle which is of no surprise since hourly, 3 hourly and 6 hourly precipitation accumulations were all impressive and occurred across a range of times during this remarkable event.

Elsewhere, multiple areas of convection formed in CO/KS and OK, with at least 2 MCSs evolving though the overnight hours, almost reaching the FL panhandle by 12 UTC. While pockets of 4-5"+ did accumulate over 6 hours, it was mostly on the lower duration high intensity side as depicted by the 1 in 15 min^{-1} and 2 in 1 h^{-1} rainfall threshold probabilities. The exact opposite was true across eastern WY, western NE and SW SD where the nearly stationary storms generated heavy and spotty rainfall that was reflected in the 6-h 10-y ARI but low probability in the 15-min and 1h rainfall depictions.

4.2.3.3 2-3 August 2023

Multiple mesoscale convective vortices formed overnight on 2 August in Nebraska/Kansas and moved across Iowa and Missouri on the evening of the 3^{rd}. As many as 4 precipitation bands formed across MO and IL after 05 UTC, yielding precipitation maxima over 5" in each band with the most falling in IL with 10.15" in a 6-h period. Figure 24 shows the practically perfect depictions for the event. It can be seen that the probabilities approach or exceed 50% for each of the metrics.

Through the intermountain west, a hatched area was well depicted according to the probabilities from the 6-h 10-y ARI but had spotty correspondence with the

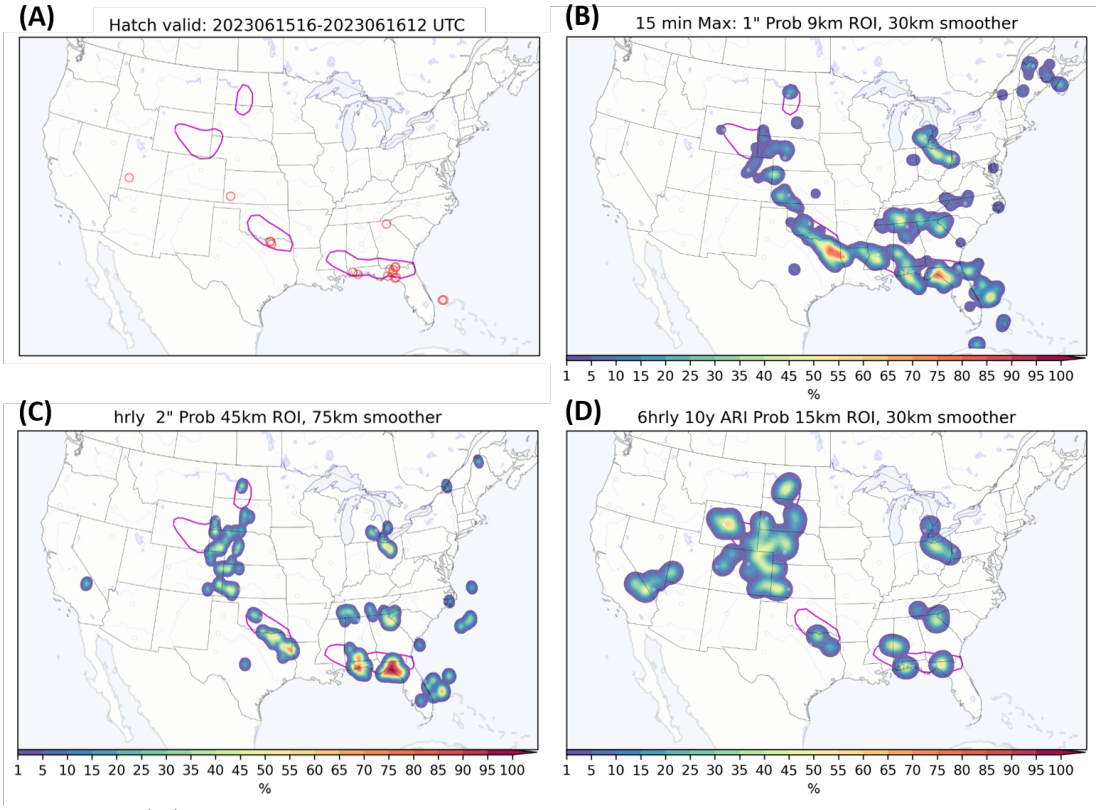

Figure 23: (A) The Hatched areas drawn for the Day 1 ERO valid 16 UTC 15 June to 12 UTC 16 June 2023 along with points above 5" in 20 hrs. Practically perfect for (B) 1 in 15 min^{-1} with a 9 km ROI and 30 km smoother, (C) 2 in hr^{-1} with a 45 km ROI and 75 km smoother and (D) 6-h 10-y ARI exceedances with a 15 km ROI and 30 km smoother

high thresholds in the smallest time windows. The opposite was true in FL where 2 in h^{-1} had probabilities over 50% and near 50% for the 1 in 15 min^{-1} but the 6-h 10-y ARI probabilities were around 15%. Another notable area was Colorado, where a single storm produced over 5.4" of rain in a 6h period, but probabilities were relatively low for this singular event. These small scale singular events (621 km^2) are very difficult to forecast.

4.2.3.4 Participant Verification and End of Week Comments

Roughly half of the participants noted that the hatched polygons added value to the forecast (not shown). As participants were shown the 6-h 10-y ARI as part

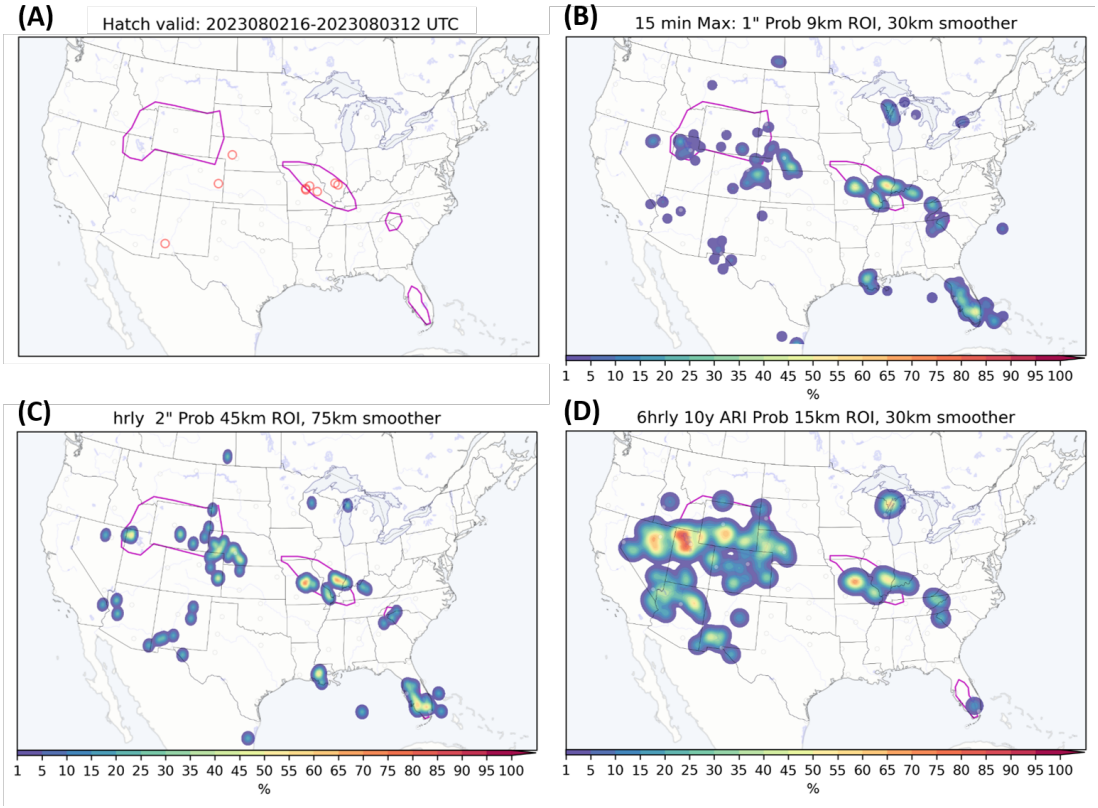

Figure 24: Same as Fig. 23 but valid 16 UTC 2 Aug. to 12 UTC 3 Aug. 2023.

of the evaluation, they found that hatched areas corresponded well to the LSRs (46%), ARI (39.5%), and had a decent size (41%) and location (44%) throughout the experiment.

Comments from end-of-week surveys showed the many philosophies that stuck with participants through the whole week, as opposed to individual days. Of the 40 comments received, 7 had a flash flood based definition, 10 had a rates definition and 7 had an ARI definition. Others wanted to denote hatched areas as vulnerable (2 people), depicting low probability high impact (3 people), and confidence in a high impact event (2 people). The mixed results suggest that it will be hard to develop a concrete definition of hatched areas given the pre-existing use of the graphical method by SPC (i.e. hatched contours). Generally, about half of the comments were supportive of some depiction highlighting intensity; of flash flooding or confidence thereof, of timing, of intense rainfall rates, or of severity.

4.2.3.5 Summary

The hatched area experiment generated many reasonable critiques, each put to real-time testing and evaluation. The idea that hatched areas can be confusing to communicate without a concrete definition is a fair point, however it was not tested here. We did demonstrate that intensity, through a range of definitions, can be depicted with the ERO. In this way intense rainfall, either expressed in easy to see 15-min, or 1,3,6-h accumulations or transformed into ARIs, has different characteristics across the US and across events. Thus we concur with the literature that heavy rainfall is not uniquely described by specific or singular metrics. However, there need not be a single metric deployed and perhaps keeping it general for somewhat rare events can aid in forecasting. This complicates the desire to communicate and deliver impact based decision support clearly. Future experiments should continue to examine this approach in light of these findings.

4.3 ARI-based Excessive Rainfall Outlook (AERO)

The other Day 1 forecasting activity was the AERO, which was designed by the FFaIR team in the 2021 FFaIR Experiment as a counterpart to the ERO. The goal was to design an excessive rainfall outlook that was based on rainfall intensity rather than on coverage of rainfall impacts. To do this, ARI exceedances rather than FFG exceedances were used as the base of the product. Unlike the ERO, the AERO does not forecast risks but rather the highest 6-h ARI to be exceeded, with options for 2-y, 5-y, 10-y, 25-y, and 50-y 6-h ARI to be exceeded. An example of how the ERO and AERO might differ can be seen in Figs. 1A and 2A.

The subjective goodness scores for the FFaIR AERO compared to the FFaIR ERO can be seen in Fig. 25. Like the ERO, the subjective scores evaluating the participant-created AERO were generally positive. Comparing the subjective evaluation to the ERO, the AERO average score was lower than the ERO (6.483 vs 7.012). Both had a score of 7 being the most likely score to be received[9] but the ERO percent of scores that were a 7 was 10% greater than the AERO's. Finally, the AERO received more instances of a score of 5 or less than the ERO did, 23.5%

[9]Reminder a 1 is the worst score and a 10 is the best score.

vs 12.5%. This could suggest that the participants overall felt the ERO did a better job at highlighting the excessive rainfall threat it was trying to identify better than the AERO. However, for the reasons discussed below it is difficult to actually make this conclusion.

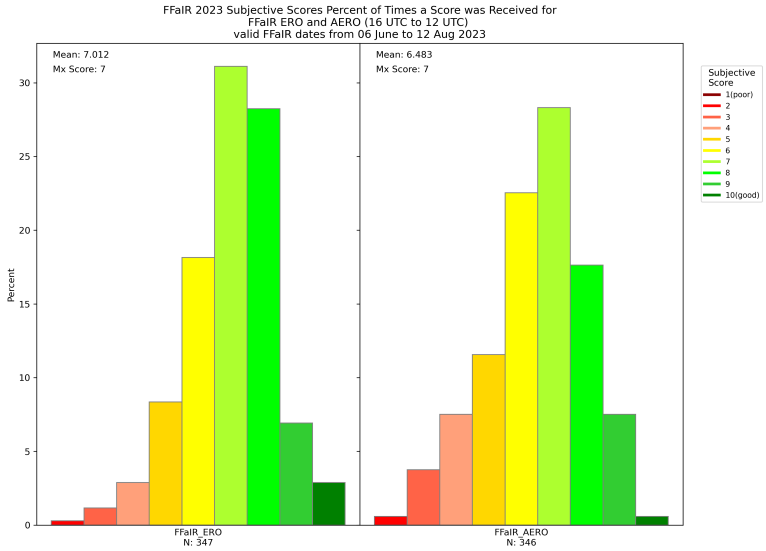

Figure 25: Like Fig. 13 but for the FFaIR ERO and AERO.

Aside from the fact that the two products identify excessive rainfall differently and have different approaches (intensity vs coverage), the verification method, the ARI database itself and the loose definition for drawing the contours all seemed to have impacted how the participants evaluated the AERO and their view of the utility of the product. The first point, about the ERO and AERO trying to convey different things, is something that the participants seemed to struggle with. Based on their comments in the AERO collaboration process and in the daily written verification comments, participants had difficulty with the product not inferring impacts (i.e. flooding). Often, comments like the following were made[10]:

- "I will say I still don't think the ARI exceedance contours match well with heavy rainfall LSRs, but I don't know how well those LSRs match impacts either. Clearly there was lots of rainfall, but where were the impacts?"

[10]Note: the AERO subjective verification question did not mention to compare the AERO to the flood or flash flood LSRs.

- "In this case the LSRs didn't really correspond well to the highest ARI regions."
- "The AERO is useful but may fall short in terms of context. A recurrence rate that isn't also tied to flood frequency may not be particularly helpful."
- "ARI consideration is meaningful, but its ability to consistently represent flood threat should be understood for a given geographic region."

The verification and ARI dataset impacts to some extent go hand-in-hand. As has been the case in past FFaIR experiments, a large portion of the discussion revolved around the western CONUS. This was focused on two things, the lack of MRMS[11] coverage across the region and the outdated ARIs. Speaking to the latter first, since the ARIs in the Pacific Northwest and Northern Rockies have not been updated since Atlas-2, it is likely that the ARIs for these regions are not representative of the current climate. The general feeling was that the values were too low, though this can not be confirmed until Atlas-15 is released with updated ARIs for the region[12]. This perception often made it difficult initially for them to draw ARI exceedances greater than 5-y over the region.

The method used to interpolate the MRMS grid to the ARI grid was the same as described in Section 2.6 of Part 1 of the Final Report (Trojniak and Correia, Jr., 2023b). This method retains the maximum value of the 9 grid point neighborhood and was used since our forecasting activities involve forecasting extreme events, thus we don't want the extremes that occur to be "washed out". Although this method is a good representation of the observed rainfall east of the Rockies, it appears to have shortcomings in the west. This is possible because, even though the gauge-corrected MRMS is used, errors due to insufficient radar coverage, beam blockage and slow gauge reporting are present. These errors could be exacerbated by the maximum value approach and thus lead to an over representation of 6h ARI exceedance.

One such example of this was discussed in depth in FFaIR. The ARI exceedances for AERO verification incorporate MRMS using the above method while

[11]This is the MRMS Gauge Corrected product.
[12]For more information about NOAA Atlas please visit https://www.weather.gov/owp/hdsc

the 5-y ARI exceedances (1-h, 3-h, and 6-h) utilized as part of the UFVS for the ERO incorporate Stage IV QPE. Fig. 26 shows the differences in ARI exceedances based on the two methods along with the MRMS and STAGE IV QPE and data from the gauge network over parts of CA and NV. Using MRMS-QPE widespread 50-y 6-h (purple) exceedances are present, thus, not surprisingly, when focusing just on the 5-y 6-h ARI, exceedances are extensive as well. However, using the ARI exceedance dataset that goes into the UFVS (Fig. 26C), the number of exceedances is significantly lower. Most notably, from ID to CA the MRMS indicated numerous instances of exceedance, while STAGE IV showed only a handful of exceedances over the same region. When comparing the 24-h QPE[13] from both methods (Fig. 26D-E), Stage IV has barely any exceedances of 1-in while in MRMS there are numerous locations with over 2-in over the region. For reference, the gauge network over CA/NV had no 24-h totals exceeding 1-in while the MRMS QPE has pockets of 1-in or more along the northern Sierra Nevada's. To add to the to the issue of the likely over-identification of 5-y 6-h ARI exceedances, the participants noted that the symbols (filled, colored circles) used for the AERO verification graphic were overly large. This they felt lead to a false representation of the true coverage of the ARI exceedances. This could be addressed by changing the radius of influence used.

Whether it was the fact that the old ARIs no longer are representative of the precipitation climate in the Pacific Northwest and Northern Rockies, the verification dataset for identifying ARI exceedances, or a combination of both, during the experiment it seemed as though if it rained over the aforementioned area, 50-y 6-h ARI exceedances were likely. After a few days of forecasting, this would result in the participants saying things that followed the theme of "draw higher than you think because it will hit in verification" or "if you are confident it will rain in the ID, draw a 50-y contour."

The loose definition of "draw for the highest 6-h ARI that is **likely** to be exceeded for any 6-h period over the 20-h period the outlook is valid" did not provide a probability of exceedance but rather relied on the participants to deter-

[13]ARI exceedances are over the 20-h that the AERO/ERO are valid for. 24-h QPE was shown here to match the gauge network.

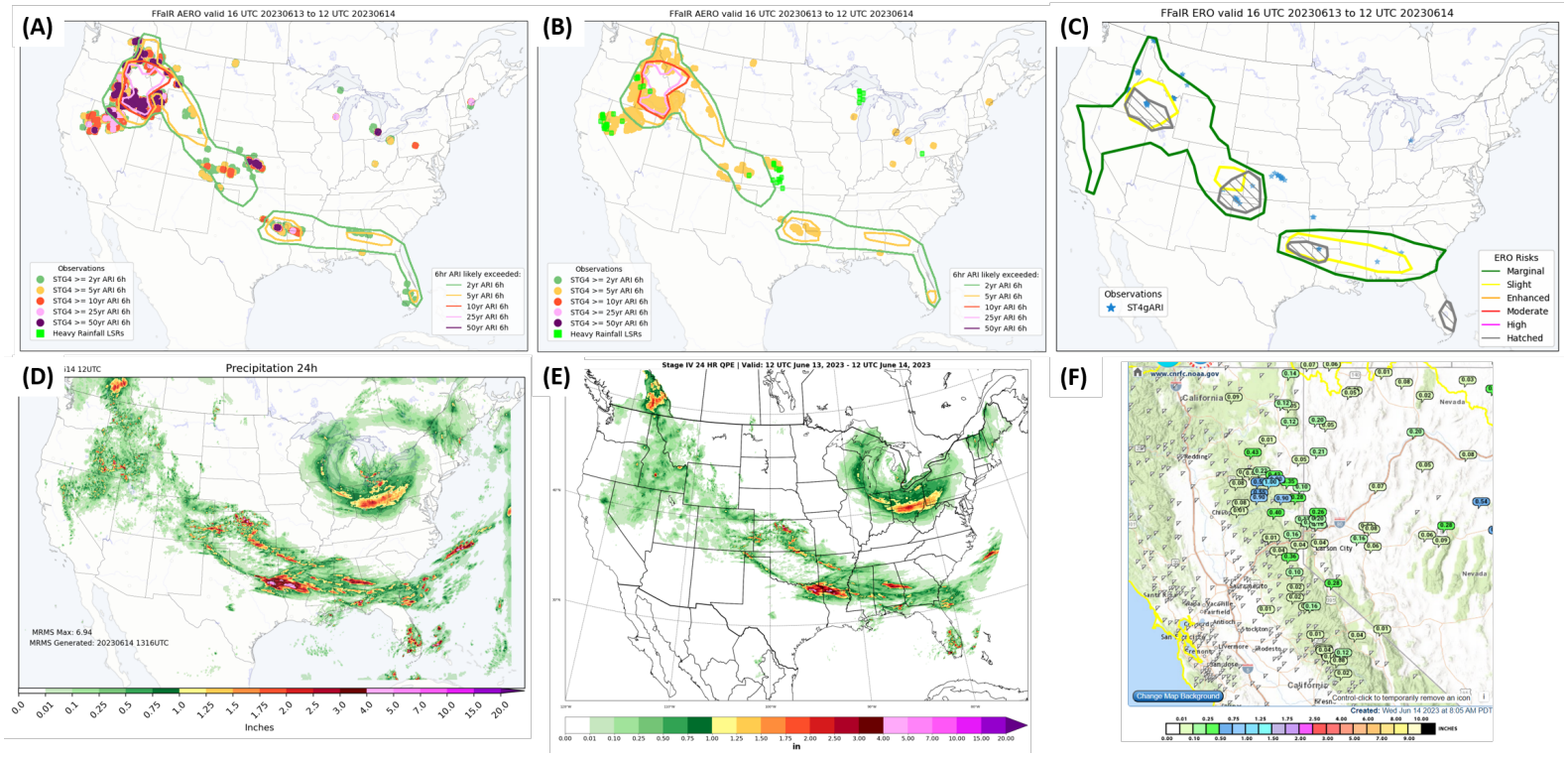

Figure 26: (A) and (B) the FFaIR AERO with (A) the 2-y (green), 5-y (yellow), 10-y (red), 25-y (pink), and 50-y (purple) and (B) only 5-y 6-h ARI exceedances based on MRMS QPE; note the legend is mislabeled in image. (C) The FFaIR ERO with 5-y (star) 1-h, 3-h, and 6-h ARI. 24-h QPE from (D) MRMS-GC and (E) Stave IV. (F) Zoomed in screen capture of the gauge network for NCRFC. (A)-(C) valid 16 UTC 13 June to 12 UTC 14 June 2024. (D)-(F) valid 12 UTC 13 June to 12 UTC 14 June 2024.

mine the probability they felt would be useful in the forecasting process. This led to much confusion and discussion, which to some extent was the goal of having a loose definition for the outlook. The participants found this loose definition not only made it challenging to create the product but also verify it. For instance, one participant wrote: "Its a challenging product to grade as such localized heavy rainfall points of high ARI can be somewhat randomly placed, and thus matching that to a smooth contoured field is hard to do visually. Its also difficult to grade as the definition of what the smoothed ARI contours are supposed to represent aren't clearly defined."

These remarks are similar to the ones given about the AERO activity last year and were expected. The reasons to perform this activity included: (1) the team was curious what would happen if forecasters were given free reign to define the product based on how they would like to use it, (2) exploring what a product like this might look like, and (3) exploring variations in Practically Perfect (P-P) to determine what probabilities of exceedances should be used in the definition. Reason 1 was not well liked by a majority of the participants in 2022 or 2023. Despite not liking having a strict definition for the AERO, overall feedback about the premise of the AERO leaned towards more positive than negative, with many participants stating that they liked the idea of having a product specific to heavy rainfall that didn't include flooding.

Combining reasons 1 and 2, by first understanding the how and why the participants draw for the various thresholds, P-P methods can be used to try and determine what probability of exceedance participants are actually drawing for (aka reason 3). A brief examination of possible radius of influence (ROI) and gaussian smoothing combinations applied to the MRMS 6-h ARI exceedances was done in the 2022 FFaIR Final Report (Trojniak and Correia, Jr., 2022). To examine P-P further and to work towards determining what probabilities might work best to define the AERO, participants were asked to provide feedback on P-P combinations. These P-P used a ROI of 40 km with either a gaussian smoothing of 70 km or 105 km and will be referred to as lower and higher smoothing respectively. They could pick from probability thresholds created by the P-P method of 5%,

15%, 25%, 40% and 70%. These were chosen for simplicity's sake since they matched the ERO risk probabilities[14].

Fig. 27 shows the number of times the various P-P combinations and probabilities were chosen[15]. Overall, probabilities from the lower smoothing (70 km) were more likely to be picked over those from the higher smoothing (105 km) for both the 2-y and 10-y ARI exceedances; 394 vs 311 and 395 vs 276 respectively. However, 77.64% and 74.84% of the time respectively, participants indicated that they liked percentage thresholds from both the 70 km and 105 km smoothing for P-P. It appears that the reason participants picked percentages from both smoothing methods was because they felt different modes of convection as well as different parts of the country warranted different "looks" to the P-P. This is well summarized by this comment "Widely spaced ARI in the practically perfect seems to suggest lower percentages would be best to verify against, but I more organized events, the 5% is probably too big! So this is a challenging problem."

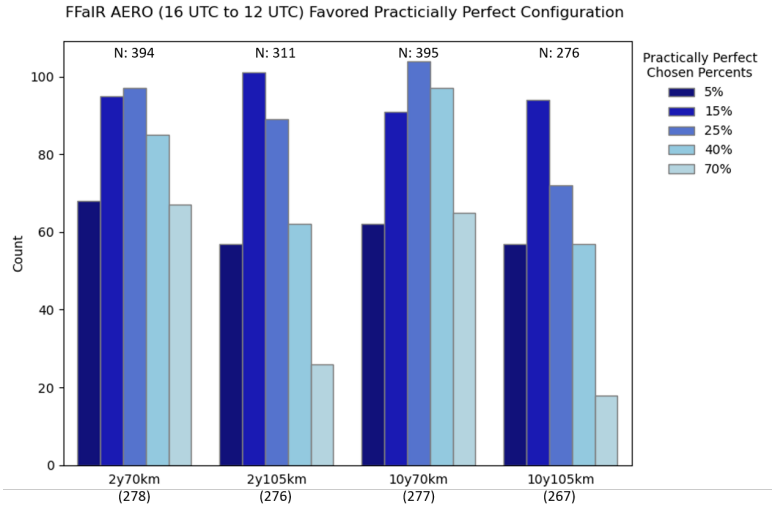

Figure 27: Summary of the preferred P-P probabilities using an ROI of 40 km and smoothing of either 70 km or 105 km for the 2-y and 10-y 6-h ARI exceedances. From dark blue to light blue the probability options were: 5%, 15%, 25%, 40% and 70%. N is the total number of percentage thresholds picked; participants could pick multiple probabilities if they wanted. In the () is the number of times at least one percentage threshold was chosen for the ARI smoothing combination.

[14]Refer to Section 2.5 for more information and examples

[15]Remember that participants were allowed to pick multiple combinations and probabilities.

An example of a day in which the precipitation patterns, convective modes and coverage of exceedances was a topic of discussion can be seen in Figs. 28 and 29. Below is the written feedback provided by the participants while Table 1 shows the P-P probability choices from the participants for this day.

- "These were all drawn too small or missed lots of areas, even at the 2y level, so they didn't capture some areas of high percentage ARI exceedance."
- "On a day like this when there are a lot of 2y ARI exceedances kind of sporadically spread around the country, I didn't like either. It kind of really bothered me to have large areas that had small discontinuities between them, I would have rather had a larger continuous area."
- "Generally liked the 25% contour on the various smoothing levels."
- "I'm okay with lower probability thresholds to increase POD, even though FAR may be inflated with this approach."
- "It's honestly so subjective and difficult to choose. The areas are essentially bullseyes and each contour is very close to the others, so which one is chosen really doesn't make much difference in this case. I wouldn't put much weight onto my answers on this."
- "Given the number of areas to be covered, it is impossible to choose an ideal percentage contour for 10y ARI with 105km smoothing as strictly adhering to any of these percentages would amount to overestimation in some areas and underestimation in other areas."
- "The small-scale areas of heavy rain yesterday made it difficult to draw a detailed enough AERO over the whole CONUS, so I think the larger radius of smoothing is beneficial for the lower prob thresholds for these kinds of days."

Reading through the comments made by the participants, it can be seen that they were struggling to determine what approach they wanted to take; capture all the reports but have a high false alarm or have a low false alarm at the expense of catching reports. They also were thinking about what they wanted the product to

look like; continuous areas or event specific areas. These go hand in hand to some extent. If a product would be more helpful by giving a general idea of the risk without being event specific then false alarm rate (FAR) is increased. For instance, using the 2-y P-P with 105 km smoothing (Figs. 28D) one could debate whether the 5% chance of exceedance contour should be used since the events along the Front Range to western NE, along the OK/AR border, in MS/AL and over AZ/NM are connected. This gives the continuous look some participants mentioned above but doing so would lead to a perceived high FAR. If the 25% threshold was used, this area would be separated into 4 regions, with only the events the Front Range to western NE included together. This would lower FAR but is less appealing to look at and, based on forecaster feedback, would be something they would be less likely to draw; i.e. they don't like to draw multiple areas with only a small space between them.

Table 1: The results for evaluation of the preferred probabilities for P-P combinations using an ROI of 40 km and smoothing of either 70 km or 105 km for the 2-y and 10-y 6-h ARI exceedances for the FFaIR AERO valid 16 UTC 13 July to 12 UTC 14 July 2023.

2y 70km	2y 105km	10y 70km	10y 105km
Other	Other	Other	Other
5%, 25%, 70%	15%, 25%, 40%	25%, 40%, 70%	5%, 25%, 40%
Other	Other	5%	
25%	15%	25%	15%
25%	25%	25%	25%
25%	15%	40%	25%
5%	15%	5%	15%
40%	25%	40%	25%
15%	15%	15%	15%
	25%	15%	
15%	25%	40%	Other
25%	15%	40%	25%
15%	15%	40%	40%
25%, 40%, 70%	5%, 15%, 25%, 40%	25%, 40%, 70%	5%, 15%, 25%

Problems like that described above, along with other issues like participants wanting to know the product's end user or not understanding what information the question was trying to tease out, makes it difficult to come to a precise conclusion

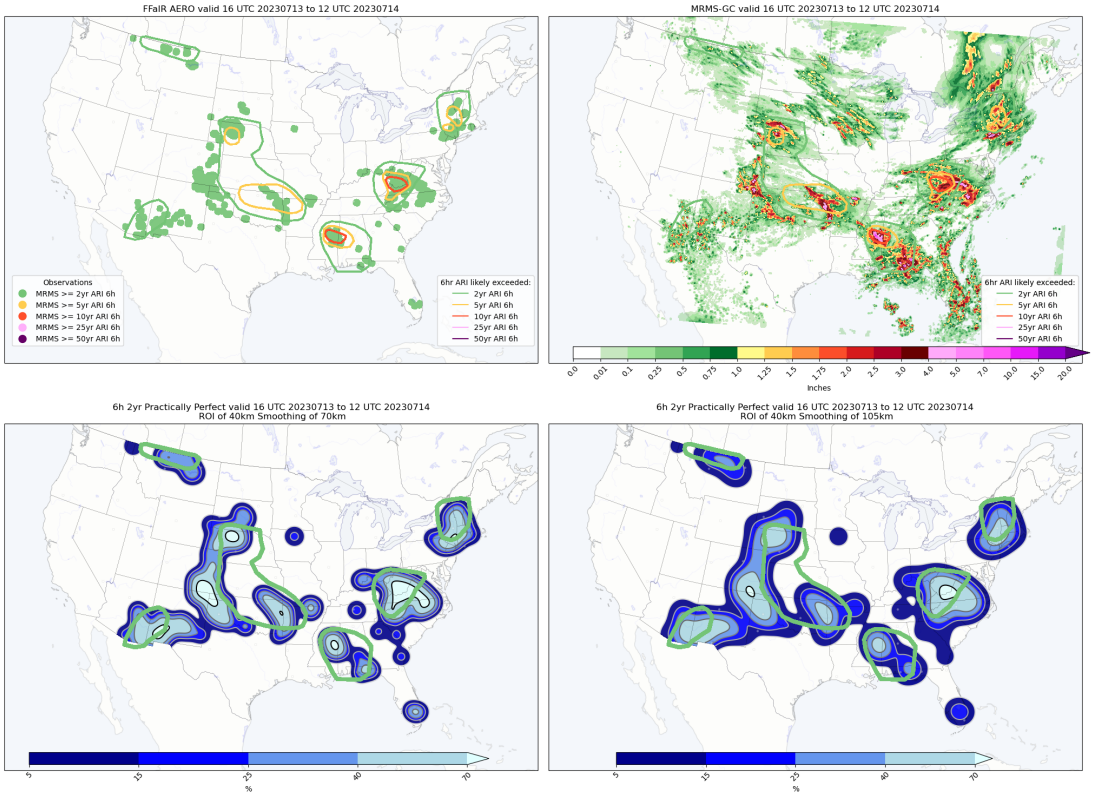

Figure 28: FFaIR AERO verification image valid 16 UTC 13 July to 12 UTC 14 July 2023. (A) FFaIR 2-y AERO contours overlayed with with the 2-y 6-h ARI exceedances and (B) the 20hr QPE overlayed with the FFaIR AERO; 2-y (green), 5-y (yellow), 10-y (red), 25-y (pink), and 50-y (purple). Practically perfect for the 2-y 6-h ARI based on (C) ROI 40 km Smoothing 70 km and (D) ROI 40 km Smoothing 105 km. The practically perfect is contoured for 5%, 15%, 25%, 40% and 70% from dark to light blue.

about this exercise. That said, over the three years that some form of the AERO forecasting activity has be been done, some general conclusions can be made:

1. Forecasters tend to not think in the ARI realm when forecasting for heavy rainfall.

2. Forecasters are very focused on impacted-based forecasts. Because of this, without a clear end user or clear connection between ARI exceedances and flash flood or flood reports, they found it difficult to both create the AERO and determine what thresholds to use.

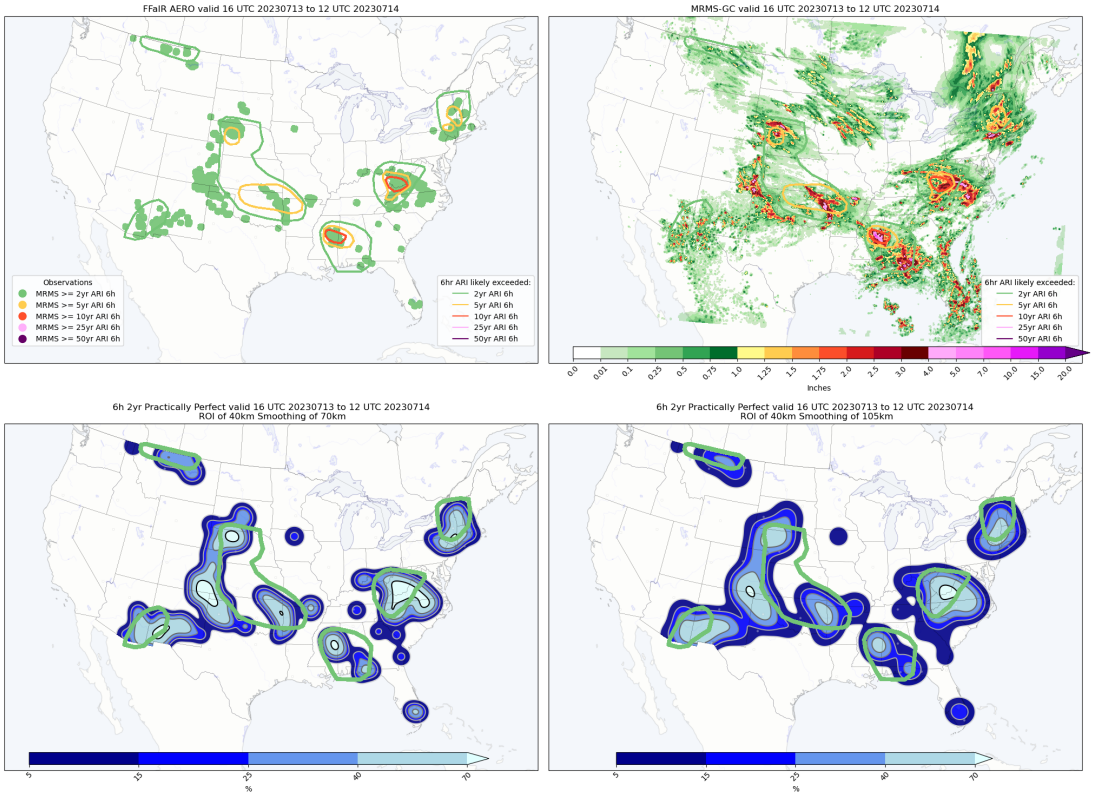

Figure 29: Same as Fig. 28 but for the 10-y 6-h ARI analysis.

3. Participants felt the AERO had utility but didn't think it was worth creating daily like the ERO. Many suggested some sort of automated system.

4. Although it was a challenging activity, the participants liked the activity because it made them think about excessive rainfall in a different light.

5. Participants seem to prefer the 70 km smoothing option for P-P verification over the 105 km smoothing.

5 Summary and Conclusions

In terms of the experimental products discussed in Part 2 of the 2023 Final Report, most of the feedback was positive. The CSU CIRA Hourly Percentile Ranking of Advected Layer Precipitable Water (ALPW) product and a Layered water Vapor Transport (LVT) product were well liked by the participants. They

felt that both derived satellite products were a good addition to the other, already operational, PWAT products generate by CSU CIRA. Participants did note that they would like to have a 90th percentile for the Percentile Ranking product. The CSU MLP EROs continue to be highly beneficial to the forecasting process for the ERO. Participants felt that the two MLPs that were trained on the GEFv12 re-analysis dataset, FV3GEFSR and the UFVSGEFSR, were better than the original operational version (GEFSO). For the UFVSGEFSR specifically, they liked that this version tended to have a smoother look to its contours than the FV3GEFSR, which was recommended to operations last year. The UFVSGEFSR has performed well and had overwhelmingly positive feedback during the past two FFaIR experiments and thus is recommended for operations. The products/models that are recommended for transition can be seen in Fig. 2 and was also shown in Fig. of Part 1 of the Final Report (Trojniak and Correia, Jr., 2023b).

Table 2: Research to Operations Transition Metrics for the 2023 FFaIR Experiment. Models/Products that are in **bold** were discussed in Part 1 of the Final Report while those that are *italicized* were discussed in Part 2.

Models, Ensembles, and Products Evaluated	Recommended for transition to operations	Recommended for further development and testing	Rejected for further testing	Provider/ Funding Source
RRFSp1 (aka RRFS_a)		X		EMC
RRFS Ensemble	Unable to evaluate			
CAPS Spatial-Aligned Mean (SAMs) Products		X		OU/CAPS Funding: Testbed Program
CAPS MLP HREF+		X		OU/CAPS Funding: Testbed Program
CSU ML Day 1 ERO UFVSFV3GEFSR	X			CSU Funding: JTTI
CSU ML Day 1 ERO HRRR		X		CSU Funding: JTTI

The addition of an Enhanced Risk for the ERO was applauded by the majority of the participants. Many felt that the probability space between the operational Slight and Moderate Risks was too large and resulted in a wider range of impacts

than those that fall in the Marginal or Moderate Risk. One caution might be that it will take time for forecasters to adjust to this new probability space. This was shown by the comparison of the fractional coverage with and without the addition of the Enhanced Risk in Fig. 20. Since WPC was also drawing an Enhanced Risk for internal evaluation during FFaIR, it will be interesting to see if they have the same difficulty, either drawing too small or not drawing enough Enhanced Risks. As for the addition of the Hatched contour, no clear consensus was found. Participants disagreed on how the Hatched area should be defined, how large it should be and what the coverage of exceedances within the Hatched area would look like. Furthermore, they still wanted to attach it to impacts.

Finally, the AERO was once again full of much discussion, both while drawing the product and during verification. The fact that the product is loosely defined continued to be an issue for the participants. And when asked what probability to use as the threshold via P-P comparisons, there was no agreement. Preferences changed depending on location and mode of precipitation. Because of this, the AERO will be tabled next year and the team will think on how they might want to incorporate what they have learned into the ERO activity.

References

Doswell, C. A., H. E. Brooks, and R. A. Maddox, 1996: Flash flood forecasting: An ingredients-based methodology. *Weather and Forecasting*, **11 (4)**, 560 – 581, https://doi.org/https://doi.org/10.1175/1520-0434(1996)011⟨0560:FFFAIB⟩2.0.CO;2, URL https://journals.ametsoc.org/view/journals/wefo/11/4/1520-0434_1996_011_0560_fffaib_2_0_co_2.xml.

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.

Herman, G. R., and R. S. Schumacher, 2018: Flash flood verification: Pondering precipitation proxies. *Journal of Hydrometeorology*, **19 (11)**, 1753 – 1776, https://doi.org/https://doi.org/10.1175/JHM-D-18-0092.1, URL https://journals.ametsoc.org/view/journals/hydr/19/11/jhm-d-18-0092.1.xml.

Trojniak, S., and J. Correia, Jr., 2022: 2022 flash flood and intense rainfall (ffair) final report: Findings and results. Tech. rep., NCEP WPC-HMT. URL https://www.wpc.ncep.noaa.gov/hmt/Reports/FFaIR/2022_FFaIR_Final_Report.pdf.

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

Trojniak, S., and J. Correia, Jr., 2023b: 2023 flash flood and intense rainfall experiment: Part 1 rrfs related findings and results. Tech. rep., NCEP WPC-HMT. URL https://www.wpc.ncep.noaa.gov/hmt/Reports/FFaIR/2023_FFaIR_Final_Report_Part1.pdf.

Appendices

A Daily Collaboration ERO and AEROs for FFaIR

The daily FFaIR EROs and AEROs drawn in collaboration by the participants during FFaIR can be seen in Figs. 30-35, grouped by weeks. The EROs are on the left and the AEROs are on the right.

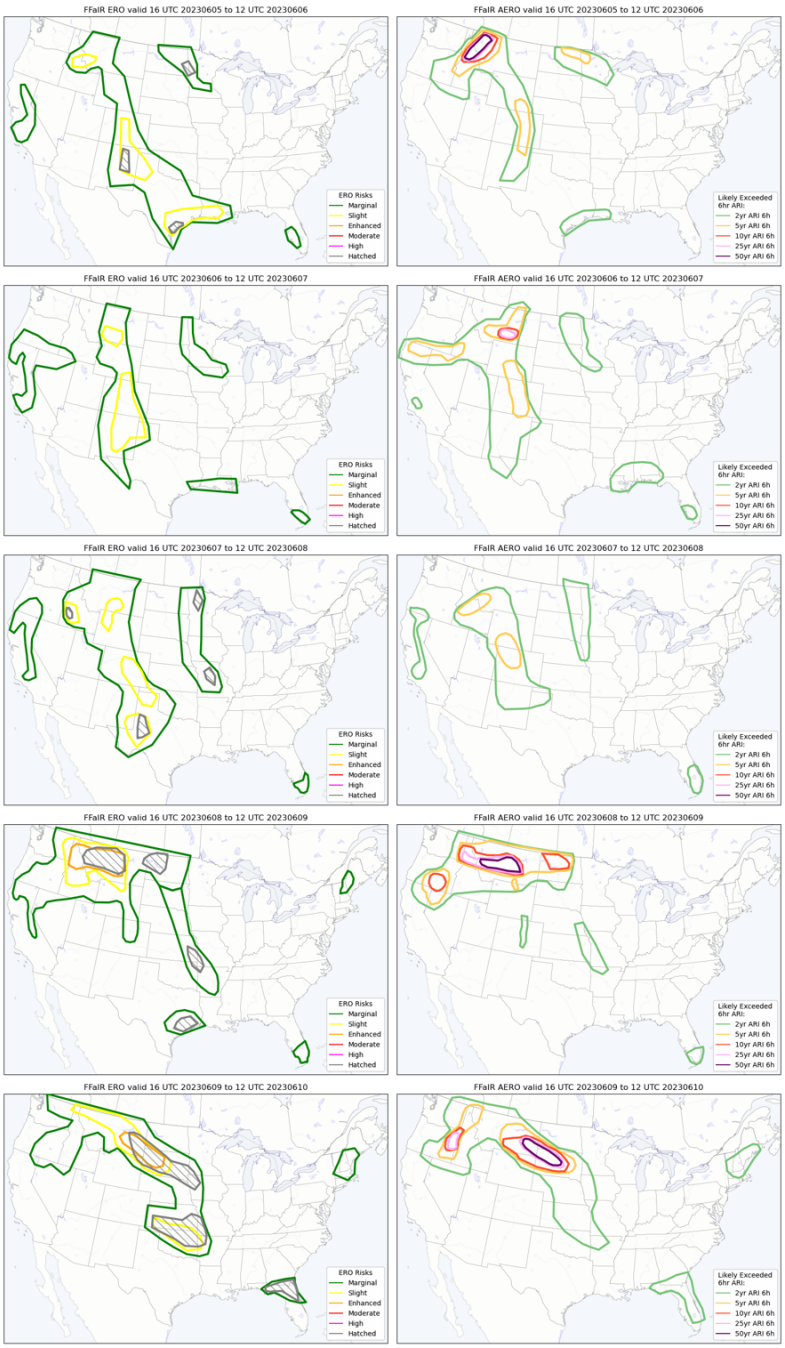

Figure 30: The Day 1 (valid 16 UTC to 12 UTC) FFaIR EROs (left) and AEROs (right) for each day of Week 1 of FFaIR which was from June 5-9 2023. The ERO Risk contours are - Marginal: 5%-15% (green), Slight: 15%-25% (yellow), Enhanced: 25%-40% (orange), Moderate: 40%-70% (red) and High: >70% (purple/pink). The Hatched (intensity) contour is grey with hatching. The AERO contours are - 2-y (green), 5-y (yellow), 10-y (red), 25-y (pink), and 50-y (purple) 6-h ARI; contours indicate when ARI is likely to be exceeded during the 20-h the AERO is valid.

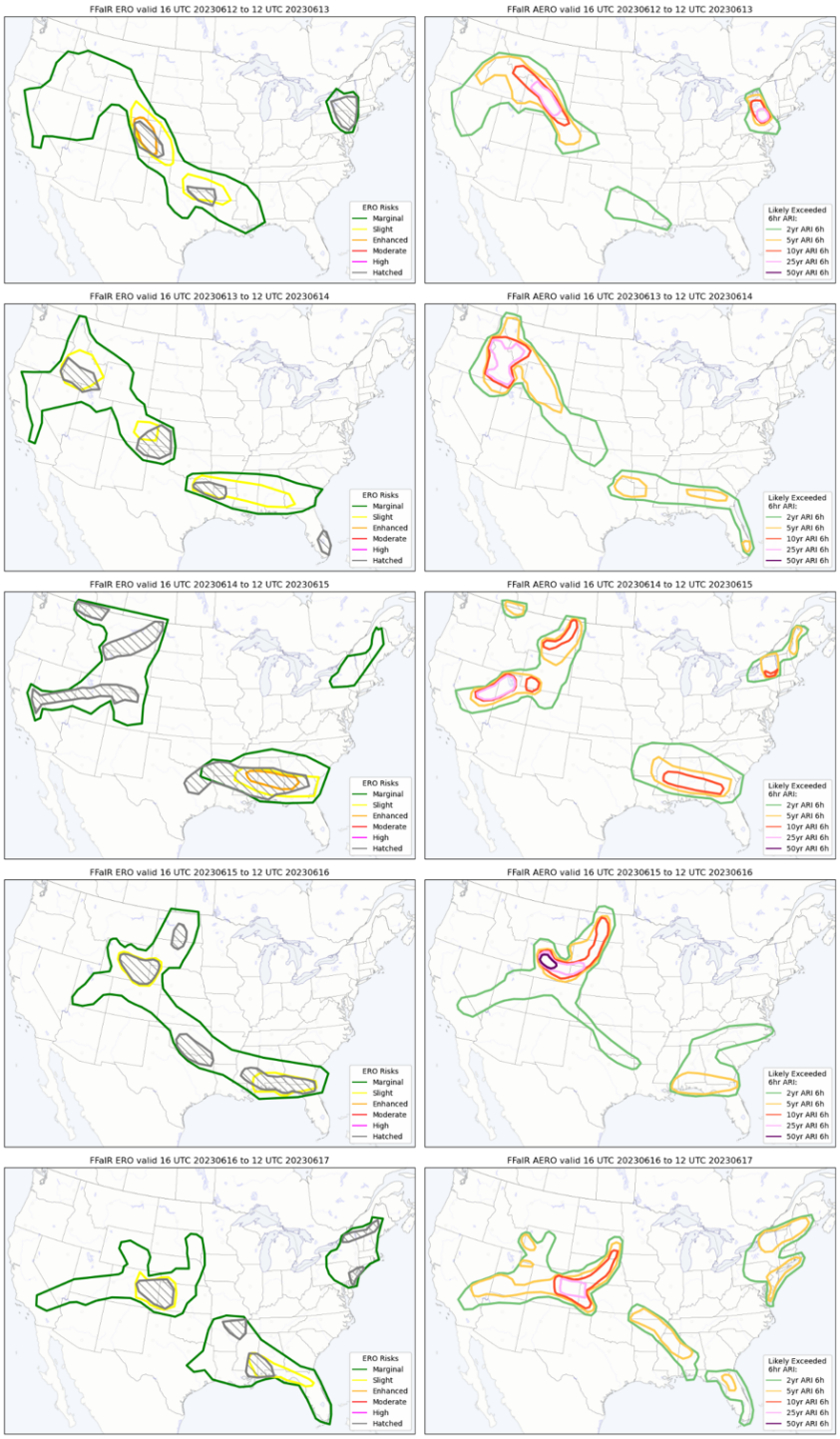

Figure 31: Same as Fig. 30 but for Week 2 of FFaIR which was from June 12-16 2023.

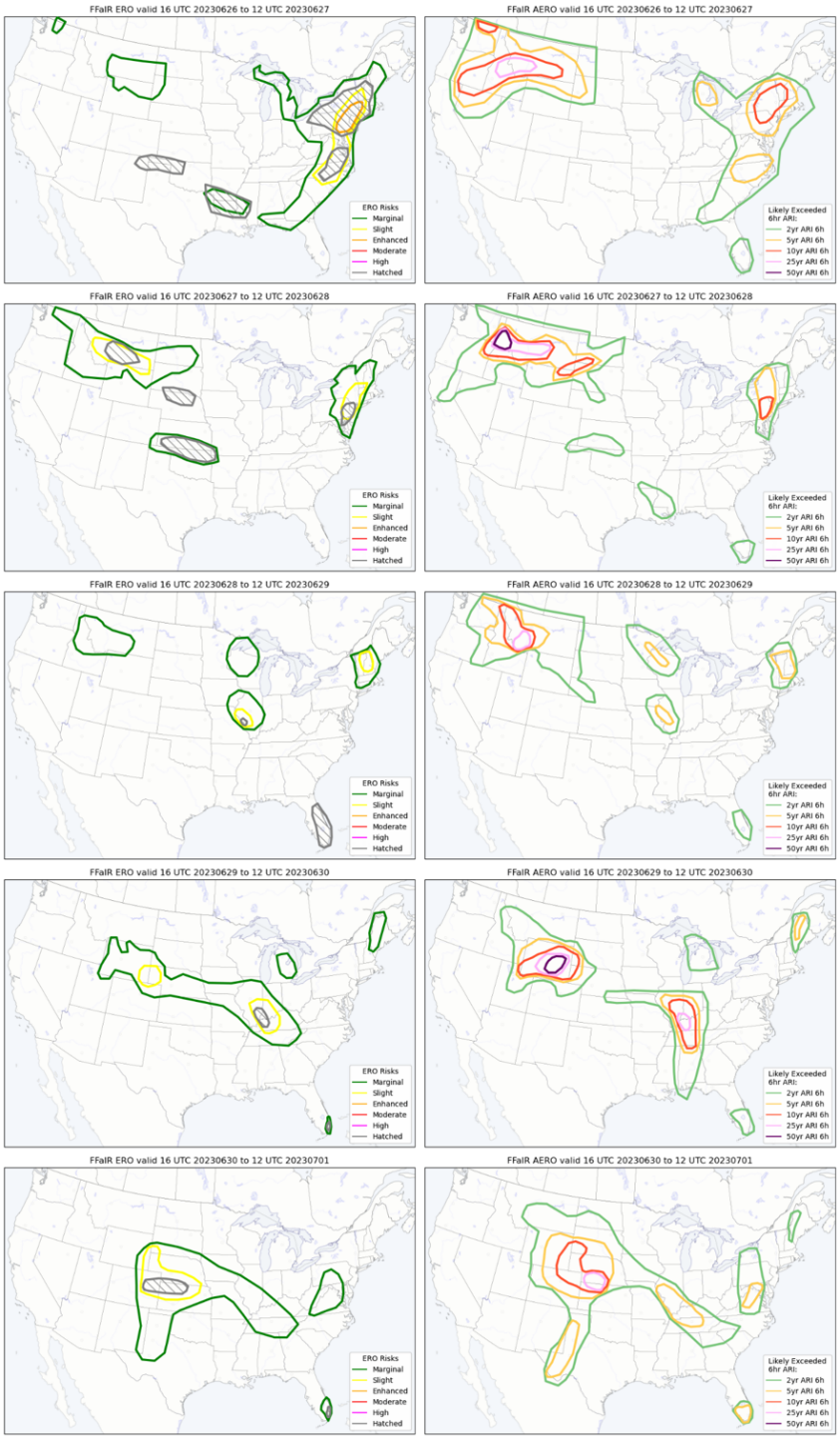

Figure 32: Same as Fig. 30 but for Week 3 of FFaIR which was from June 26-30 2023.

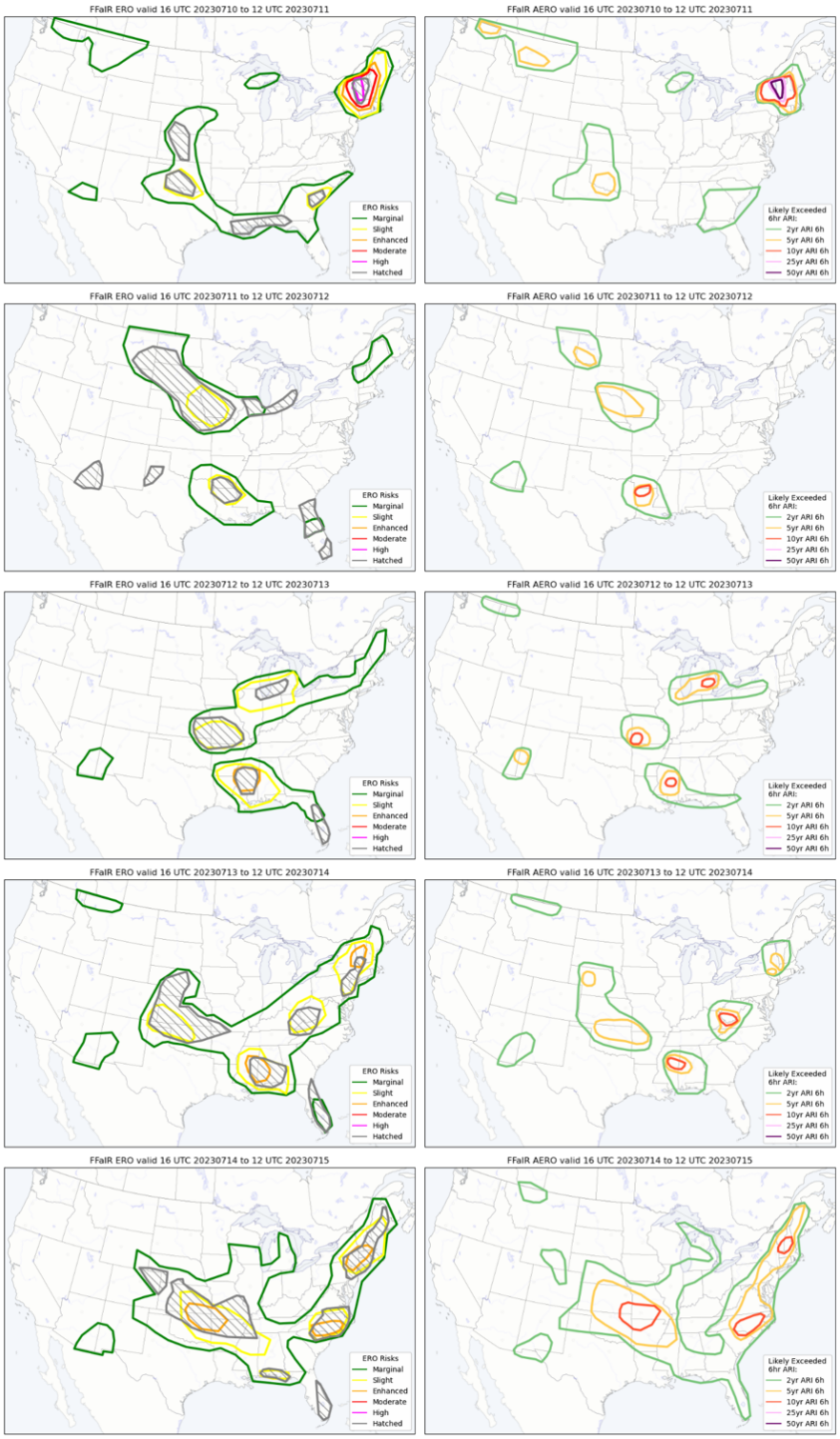

Figure 33: Same as Fig. 30 but for Week 4 of FFaIR which was from July 10-14 2023.

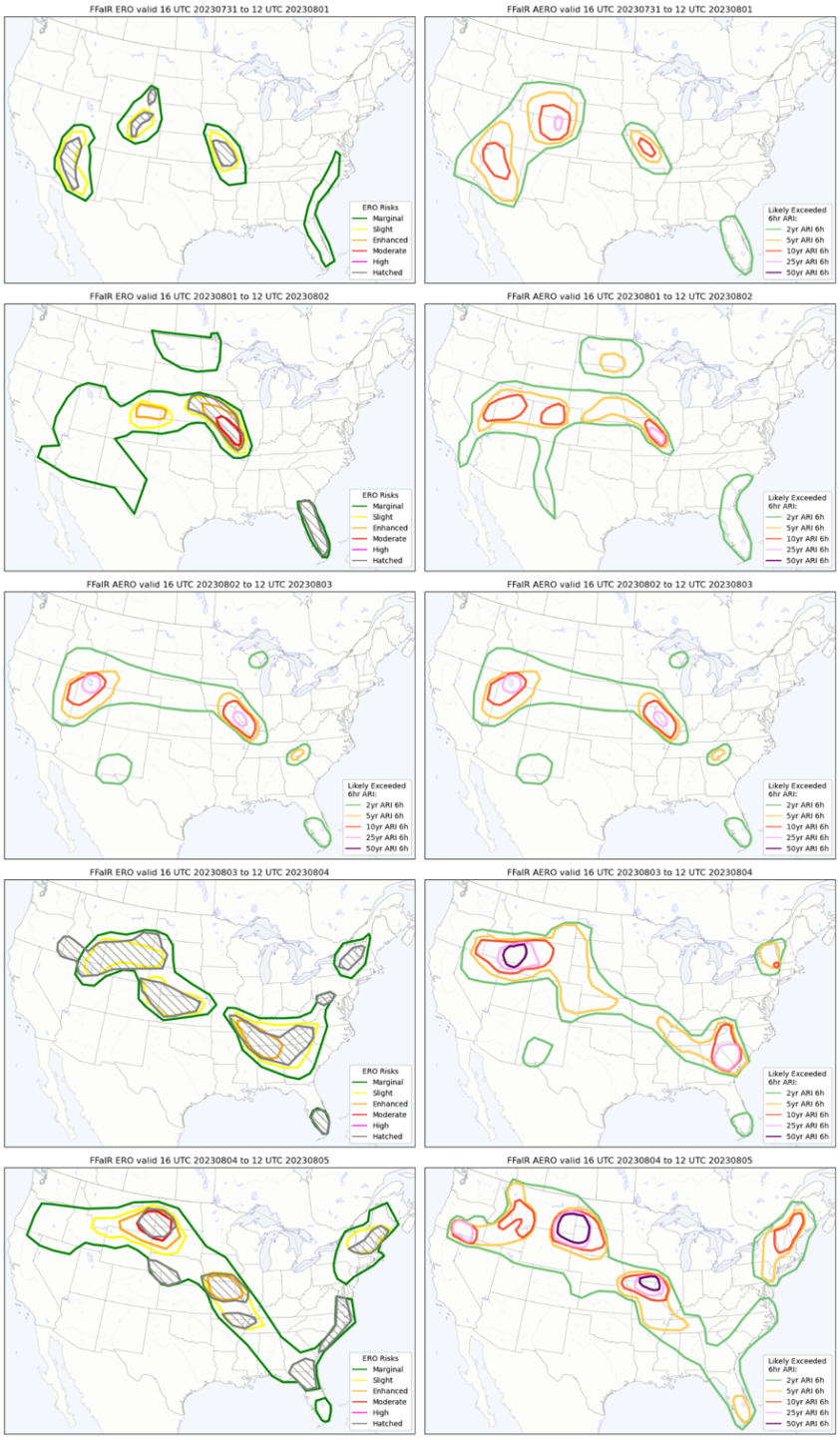

Figure 34: Same as Fig. 30 but for Week 5 of FFaIR which was from July 31 - Aug 4 2023.

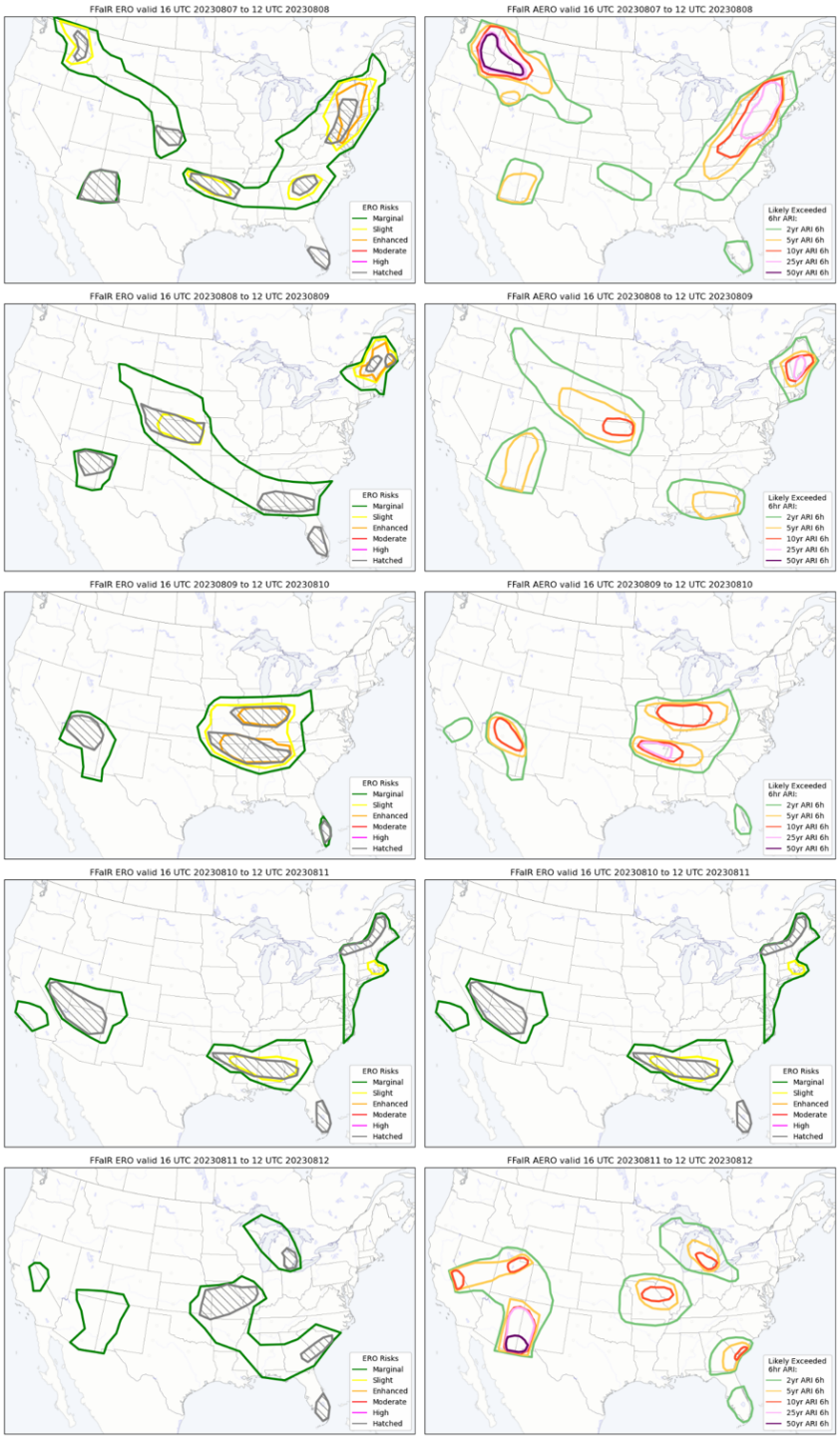

Figure 35: Same as Fig. 30 but for Week 6 of FFaIR which was from Aug 7 - 11 2023.