Evaluation of the Rapid Refresh Forecast System during the 2023 HWT Spring Forecasting Experiment

Israel Jirak, NOAA/NWS/Storm Prediction Center
Adam Clark, NOAA/OAR/National Severe Storms Laboratory
David Harrison, CIWRO/SPC
Jake Vancil, CIWOR/SPC

RRFS Evaluation during the HWT SFE

- RRFS was evaluated during the 2023 Hazardous Weather Testbed (HWT) Spring Forecasting Experiment (SFE) from May 1 - June 2, which was a ***below-average period*** climatologically for severe weather
- Deterministic Evaluations:
  - 00Z RRFS Control vs. HRRR DA (1-6 hours)
  - 00/12Z RRFS Control vs. HRRR (Day 1)
  - 00Z Flagship Models - Blinded (Day 1)
- Ensemble Evaluations:
  - 00Z RRFS vs. HREF (Day 1)
  - 12Z RRFS vs. HREF (Days 1 & 2)
  - 12Z Single-Physics RRFS vs. Multi-Physics RRFS (Days 1 & 2)
  - 12Z Single-Time RRFS vs. Time-Lagged RRFS (Days 1 & 2)
- The subjective evaluations from HWT SFE for severe weather forecasting are presented along with objective verification over daily mesoscale domains
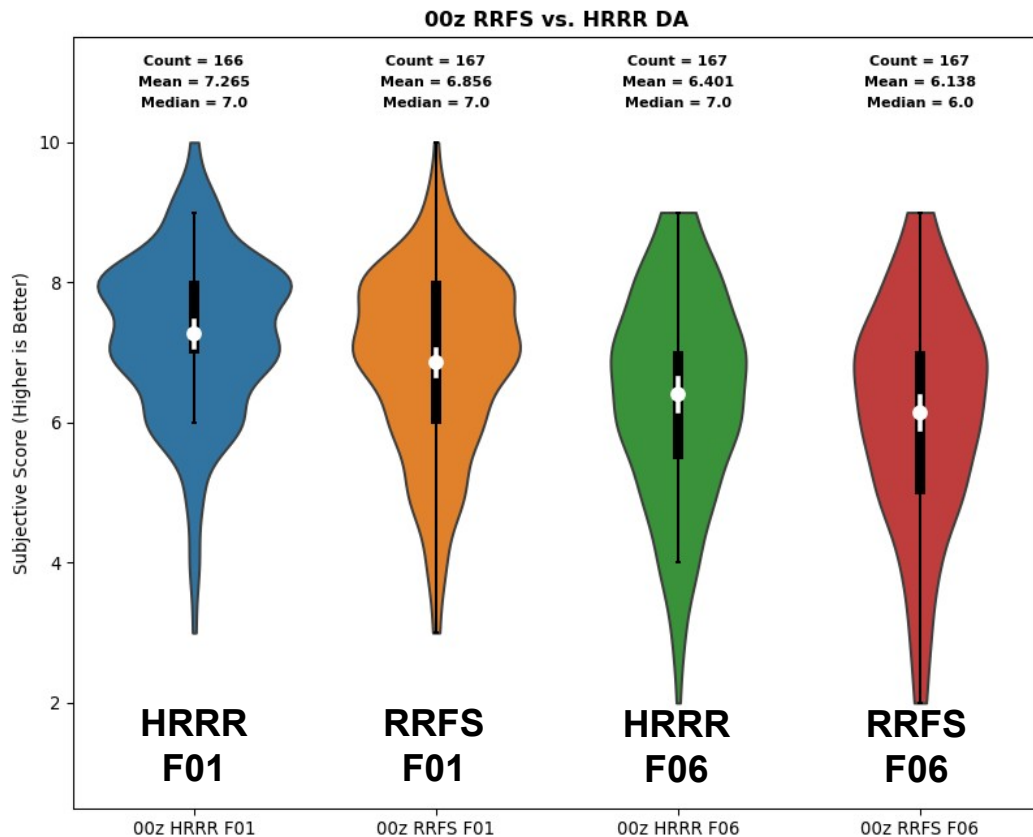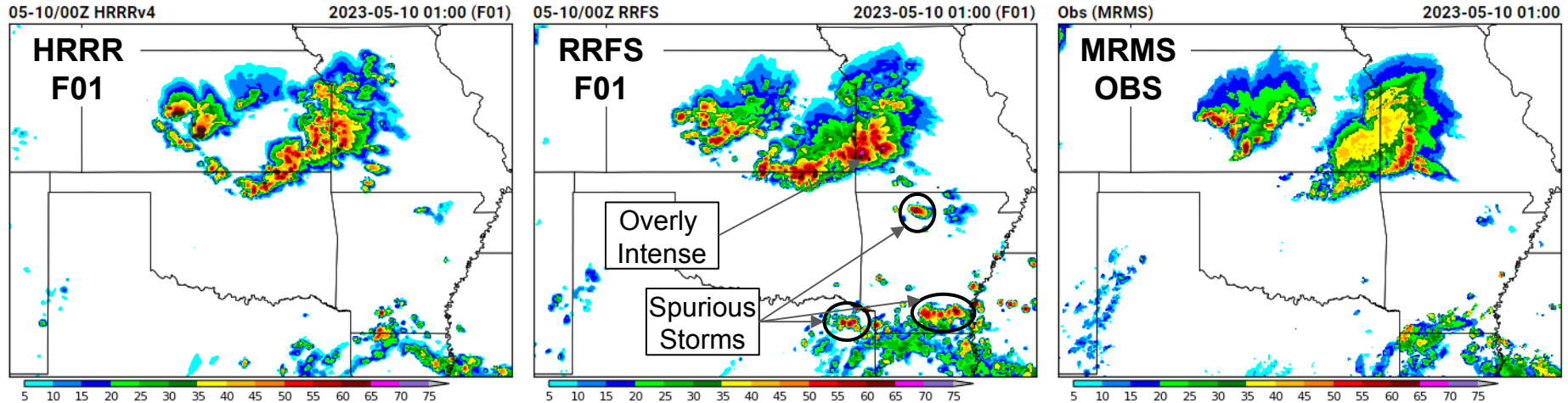
RRFS
Deterministic Evaluations

00Z RRFS Control vs. HRRR DA (1-6 hours)

- HRRR short-term composite reflectivity forecasts (1-h & 6-h) had higher ratings than those from the RRFS
- The 1-h forecasts from the RRFS had an especially high bias in intensity and coverage of reflectivity

00z RRFS vs. HRRR DA

Count = 166	Count = 167	Count = 167	Count = 167
Mean = 7.265	Mean = 6.856	Mean = 6.401	Mean = 6.138
Median = 7.0	Median = 7.0	Median = 7.0	Median = 6.0

HRRR F01   RRFS F01   HRRR F06   RRFS F06

00z HRRR F01   00z RRFS F01   00z HRRR F06   00z RRFS F06

Subjective Score (Higher is Better)

00Z RRFS Control vs. HRRR DA (1-6 hours)

- While neither HRRR nor RRFS are particularly accurate in depicting the derecho-producing MCS in this 1-h forecast, the overly intense reflectivity in the leading convective line and spurious storms in the RRFS are concerning

00Z RRFS Control vs. HRRR (Day 1): Storm Attributes

- For 00Z storm-attribute fields, the HRRR was rated slightly better for simulated reflectivity/UH, updraft speed, and QPF

- Meanwhile, the RRFS was very slightly favored for severe convective 10-m winds

- Most common comments were that the RRFS developed storms that were too intense and numerous/widespread
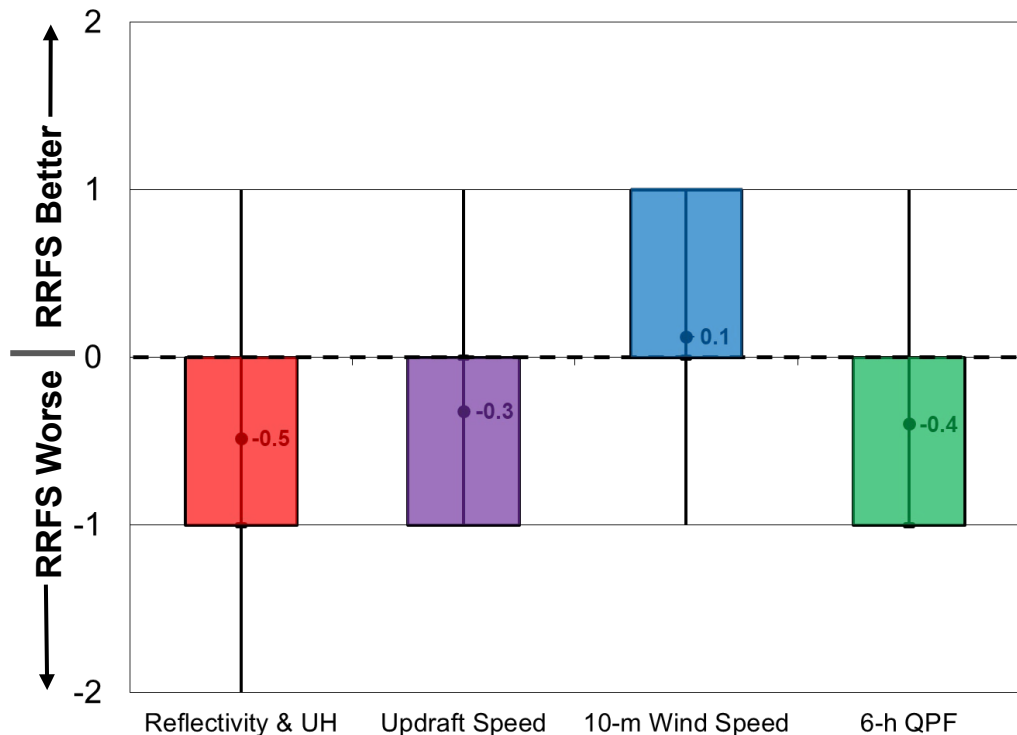
00Z RRFS Control vs. HRRR (Day 1): Storm Attributes

- HRRR has a much better forecast than the RRFS control for the derecho-producing MCS across eastern Kansas from the 00Z-initialized run on 9 May 2023
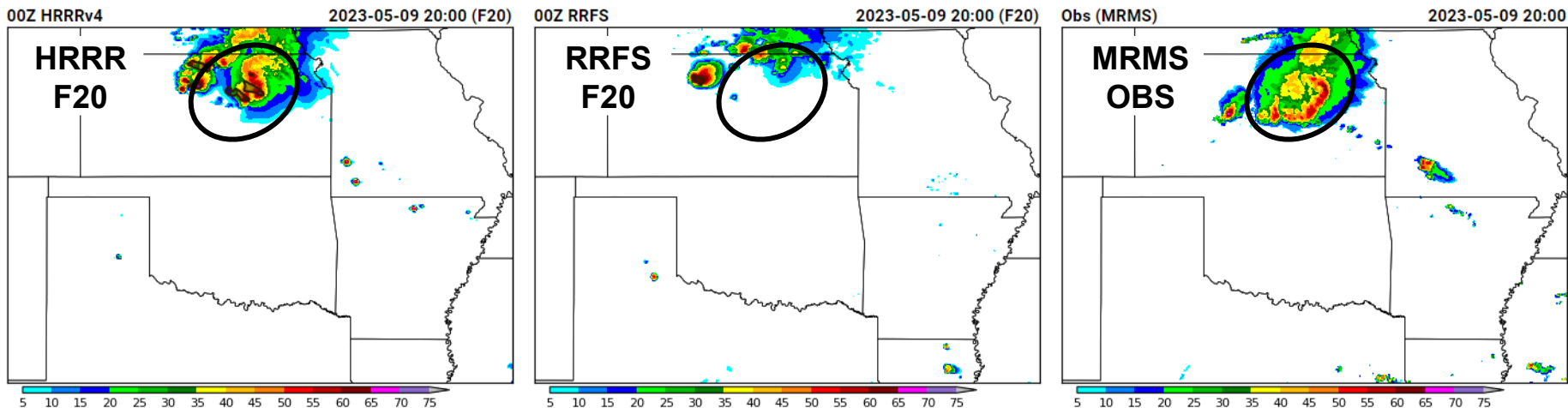
00Z RRFS Control vs. HRRR (Day 1): Environment

- For 00Z environment fields, slight edge to the HRRR for SBCAPE, 2-m temperature, and 2-m dewpoint.
- SBCAPE forecasts were the most common environmental field to be favored for the HRRR
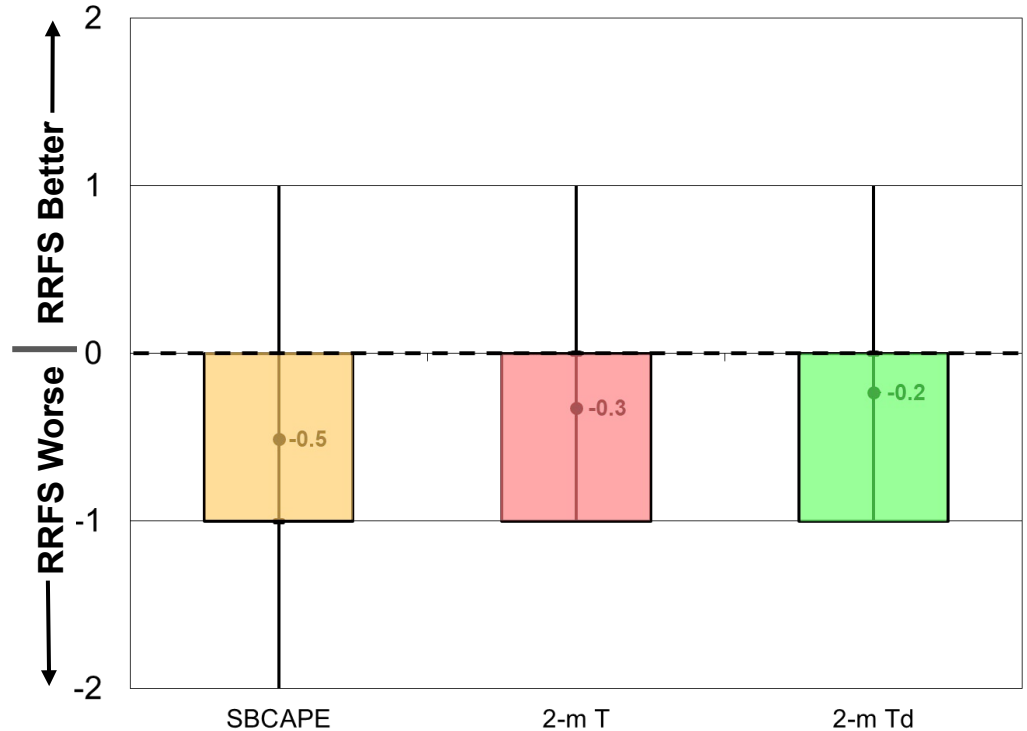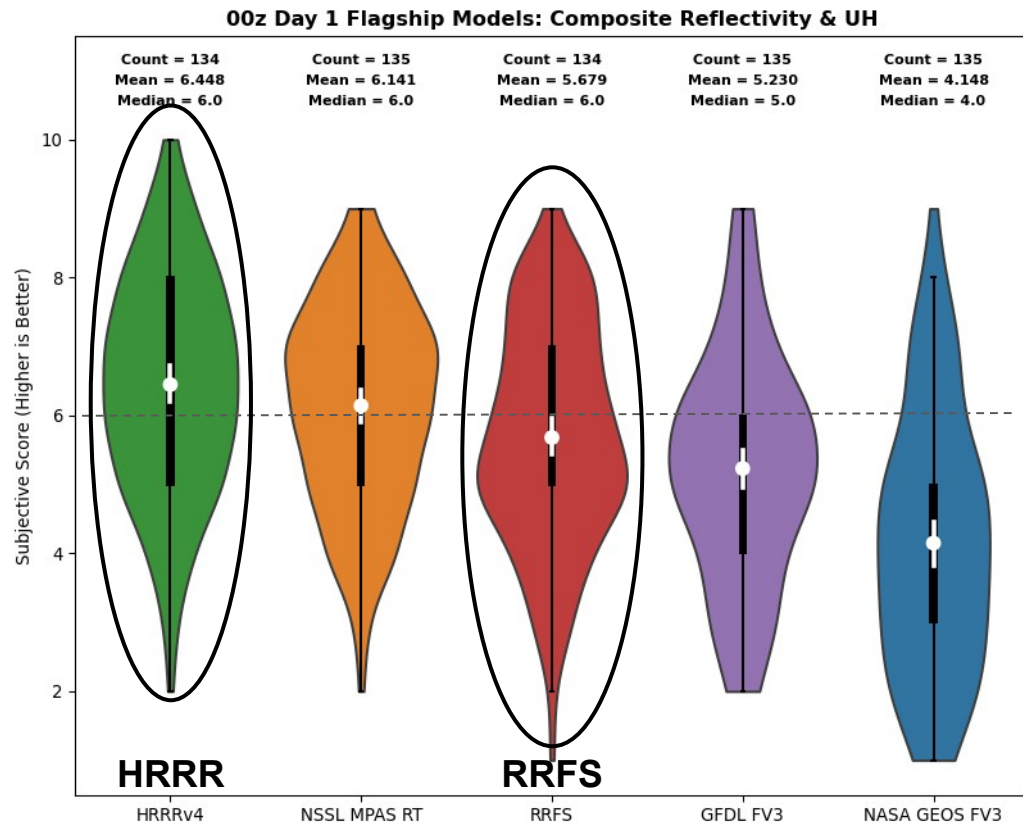
# 00Z Flagship Models - Blinded (Day 1): Reflectivity

Blinded Evaluation

- HRRR had highest mean rating for reflectivity - statistically significantly higher than RRFS (90% CI)
- NSSL MPAS actually slightly favored on average over the RRFS for these 00Z runs covering Day 1 (f12-f36)

00z Day 1 Flagship Models: Composite Reflectivity & UH
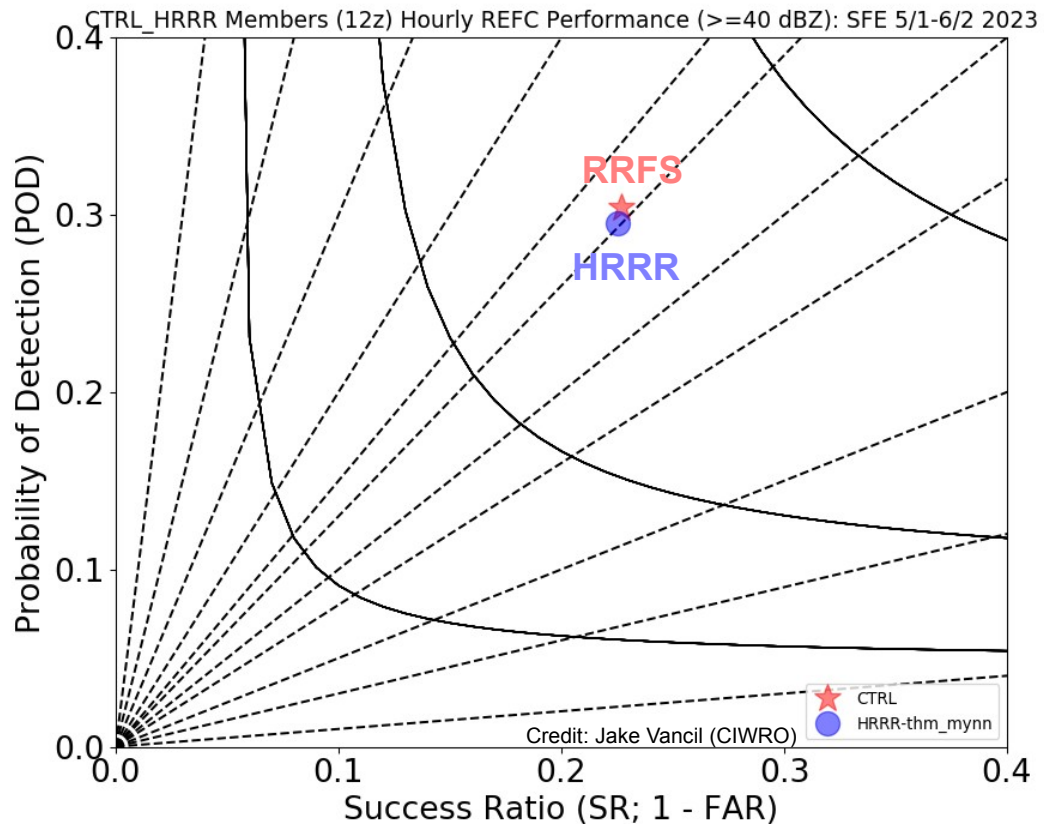
12Z RRFS Control vs. HRRR (Day 1): Storm Attributes

- For *12Z* storm-attribute fields, the distributions are shifted toward the middle where model performance is about the same
- RRFS was slightly favored for severe convective 10-m winds
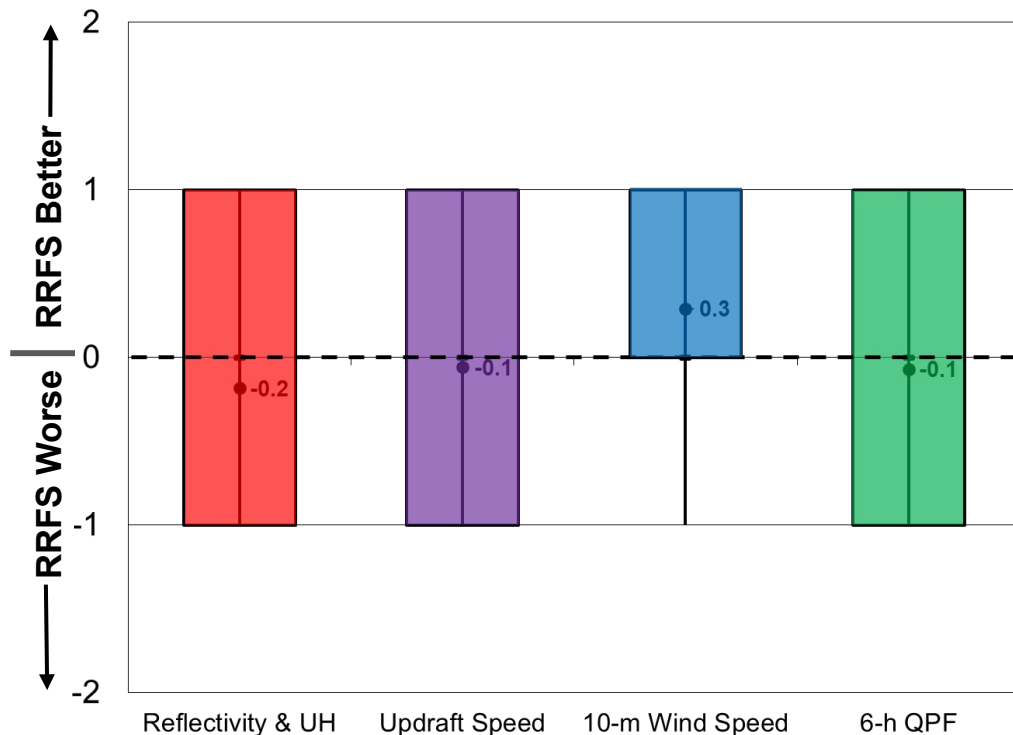
12Z RRFS Control vs. HRRR (Day 1): Reflectivity

- Objective verification of ≥40 dBZ reflectivity over the SFE daily domains confirms subjective ratings for 12Z runs
- Very similar Day 1 performance characteristics (CSI, POD, bias, etc.) for RRFS control and HRRR for 12Z cycle

* Planning to calculate objective statistics for 00Z runs for comparison

CTRL_HRRR Members (12z) Hourly REFC Performance (>=40 dBZ): SFE 5/1-6/2 2023

12Z RRFS Control vs. HRRR (Day 1): Storm Attributes

- RRFS control forecast is much improved from the 12Z cycle for the derecho-producing MCS across eastern Kansas compared to the 00Z run and better than the 12Z HRRR (though still slow)
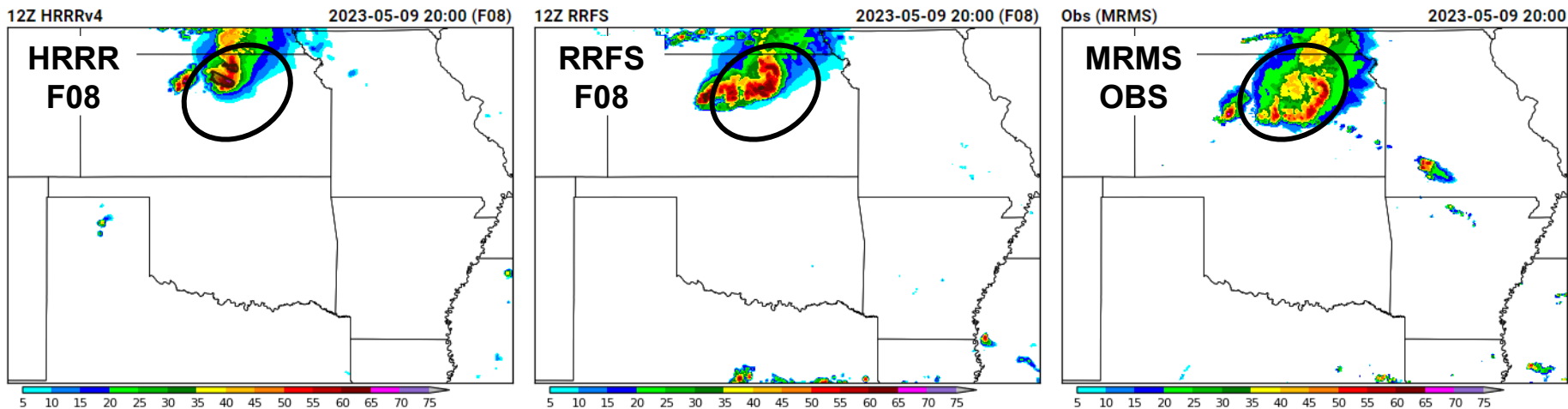
RRFS
Ensemble Evaluations

00Z RRFS Ens vs. HREF (Day 1): Storm Attributes

- Directly compared the 00Z single-physics RRFS ensemble to the HREF
- For the storm-attribute fields, the HREF has slightly higher ratings for updraft helicity & updraft speed
- The rating distributions are more neutral (about the same) for 10-m wind speed and composite reflectivity.
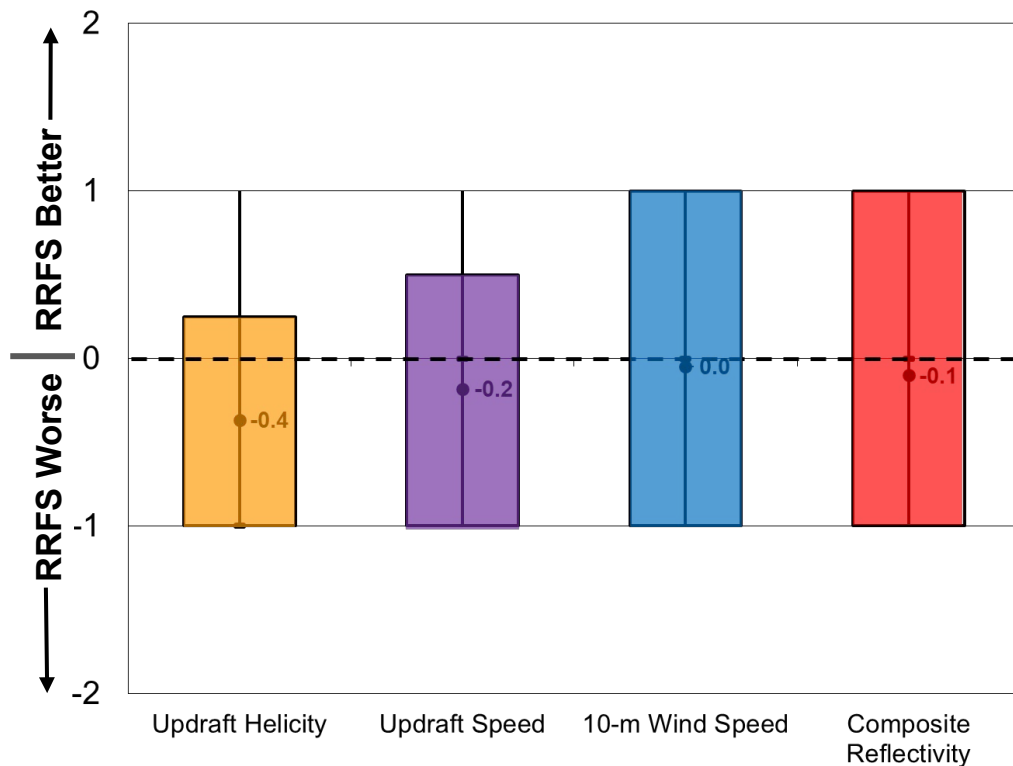
00Z RRFS Ens vs. HREF (Day 1): Storm Attributes

- For the 9 May event, the RRFS was rated slightly worse than HREF, owing to better orientation of probabilities and centering of preliminary LSRs for the HREF forecast

24-h NMEP of Updraft Helicity >99.85%
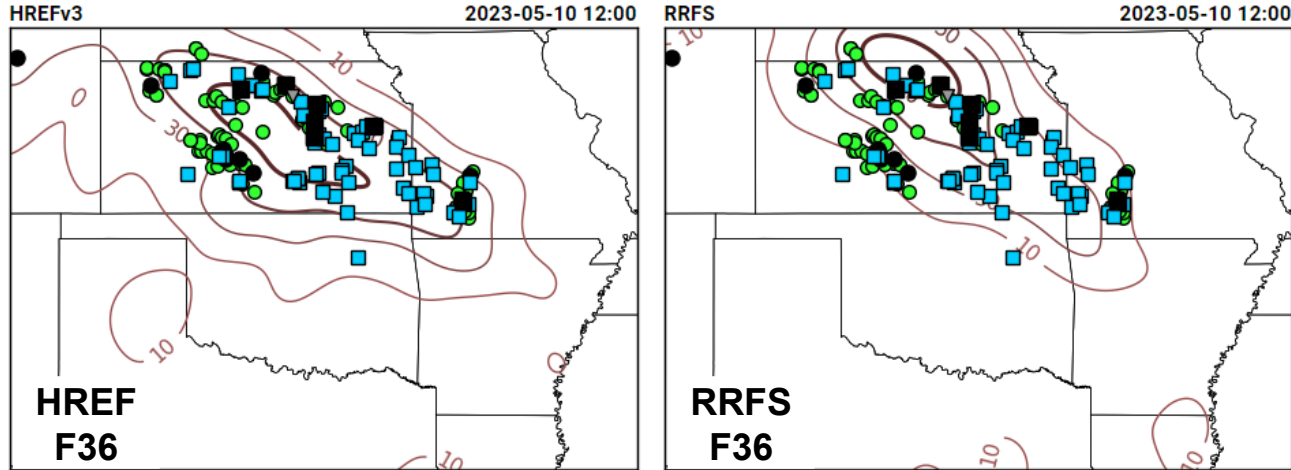
00Z RRFS Ens vs. HREF (Day 1): Environment

- For the environment fields between 16-20Z, the rating distributions are shifted toward RRFS being slightly worse
- SBCAPE forecasts were closer to "about the same"
- Later times (not shown) are similar with a subtle shift toward being about the same
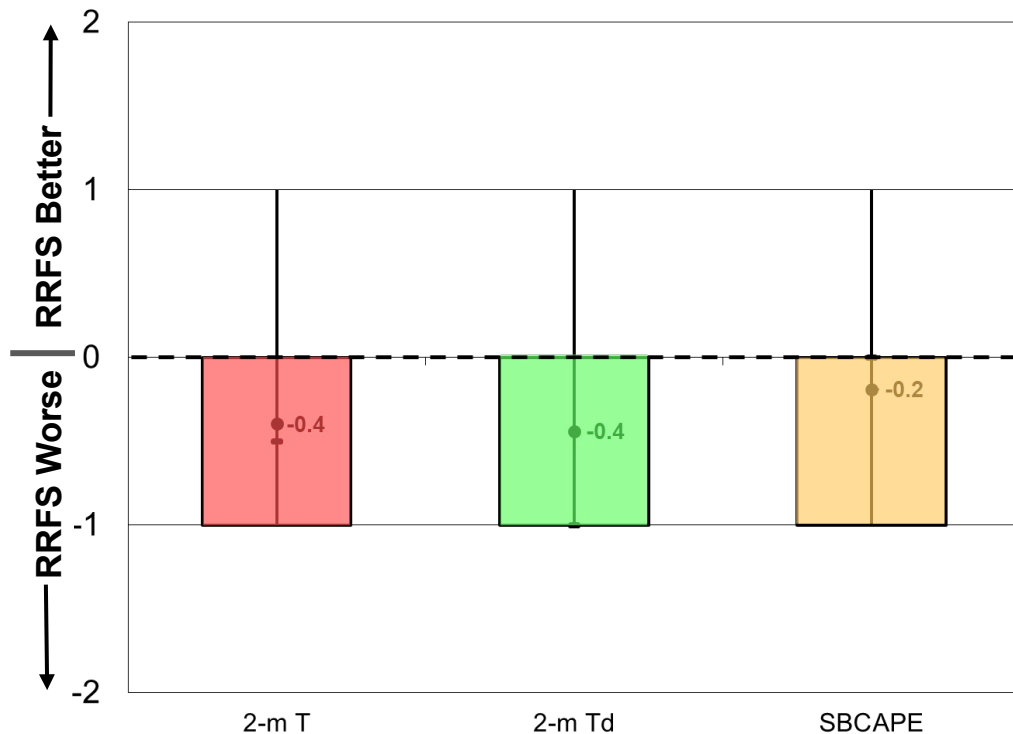
12Z RRFS Ens vs. HREF (Day 1): Storm Attributes

- Similar rating distributions for the 12Z HREF and single-physics RRFS covering the Day 1 period
- This is an encouraging result and evidence of the progress made

12z Day 1 RRFS Physics & Time-Lagging vs HREF
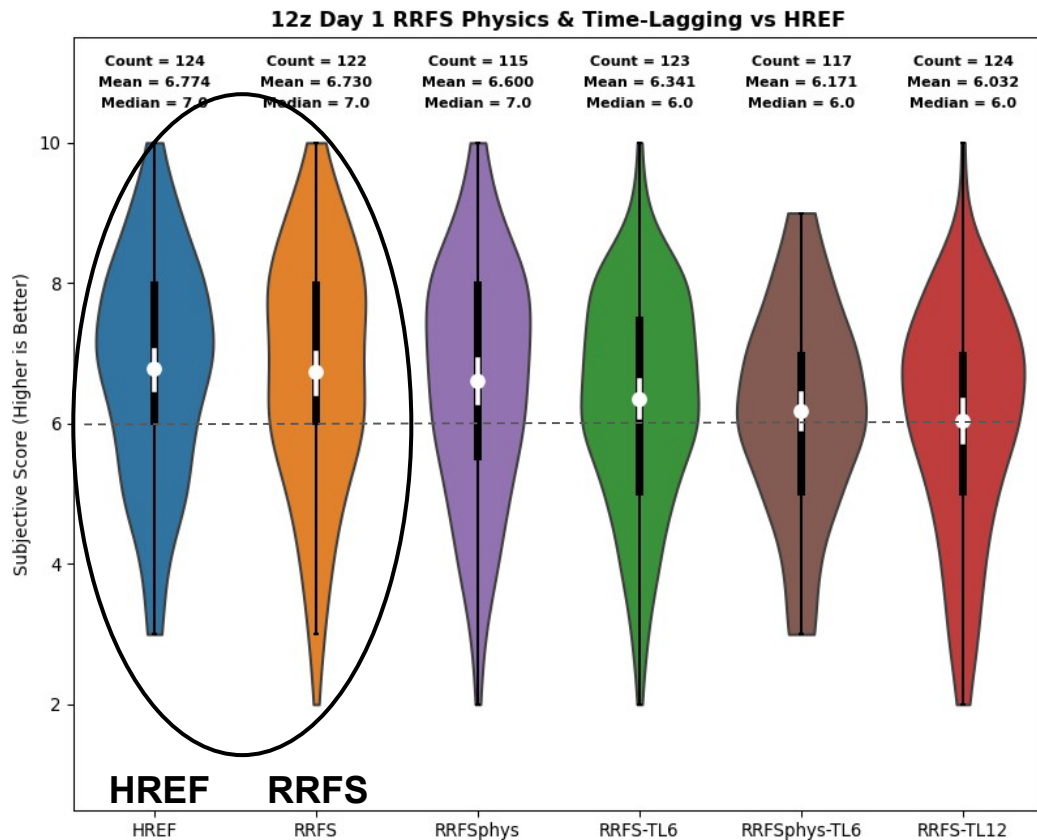
12Z RRFS Ens vs. HREF (Day 1): ROC & Reliability

RRFS and HREF (12z) Hourly REFC ROC Curve (>=40 dBZ): SFE 5/1-6/2 2023

RRFS: 0.8129
HREF: 0.8209

HREF

RRFS

Credit: Jake Vancil (CIWRO)

RRFS and HREF (12z) Hourly REFC Reliability (>=40 dBZ): SFE 5/1-6/2 2023
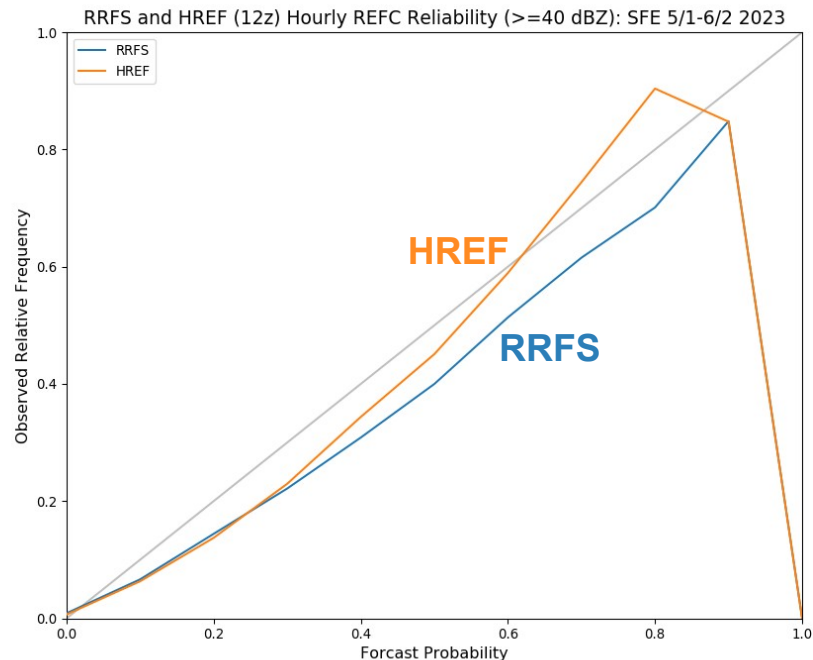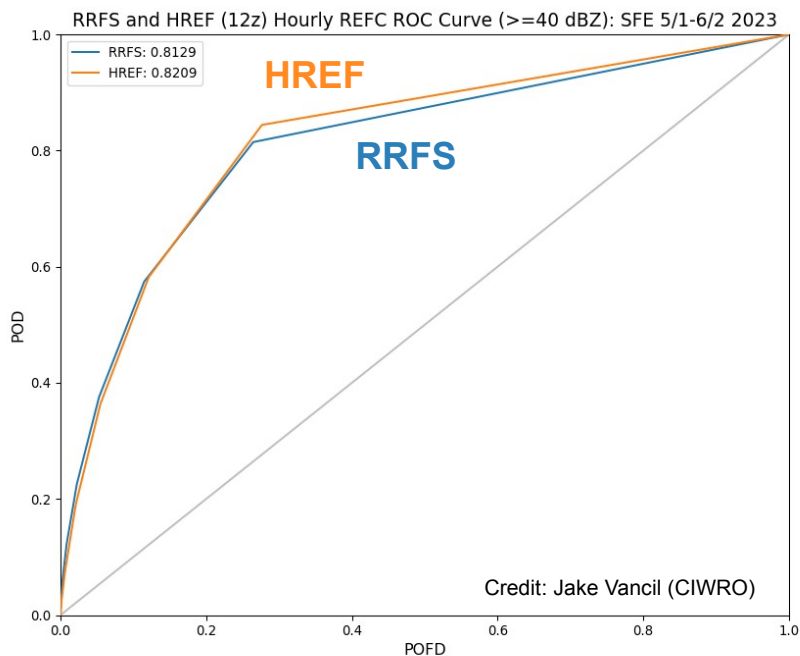
RRFS
HREF

HREF

RRFS

- Objective verification statistics over the SFE domains confirm similar ensemble performance characteristics for deep convection (>40 dBZ reflectivity) with an edge to the HREF for POD @ 10% and reliability above 30%

12Z RRFS Ens vs. HREF (Day 1): Member Perf.

- First of all, it is worth acknowledging the significant development efforts and improvements to get the FV3 core to where it is today for convective-scale forecasting
- This is a true worst-to-first story going from the HRW-FV3 to the current RRFS control member

RRFS_HREF Members (12z) Hourly REFC Performance (>=40 dBZ): SFE 5/1-6/2 2023

Legend:
- CTRL
- MEM0001
- MEM0002
- MEM0003
- MEM0004
- MEM0005
- MEM0006
- MEM0007
- MEM0008
- MEM0009
- ARW-wsm6_ysu
- FV3-gfdl_edmf
- HRRR-thm_mynn
- NAM-fa_myj
- NSSL-wsm6_myj
- ARWLAG-wsm6_ysu
- FV3LAG-gfdl_edmf
- HRRRLAG-thm_mynn
- NAMLAG-fa_myj
- NSSLLAG-wsm6_myj

RRFS Ctl

HRW-FV3

Stars: RRFS
Circles: HREF
Credit: Jake Vancil (CIWRO)

X-axis: Success Ratio (SR; 1 - FAR)
Y-axis: Probability of Detection (POD)

12Z RRFS Ens vs. HREF (Day 1): Member Perf.

- The RRFS control member and HRRR clearly stand out as the best performers
- The HRW-ARW stands alone as a next-tier performer with distinct performance characteristics
- The NAM Nest, HRW-NSSL, and perturbed RRFS members have similar performance characteristics
- The HRW-FV3 stands alone as the worst performer

RRFS_HREF Members (12z) Hourly REFC Performance (>=40 dBZ): SFE 5/1-6/2 2023

Legend:
- CTRL
- MEM0001
- MEM0002
- MEM0003
- MEM0004
- MEM0005
- MEM0006
- MEM0007
- MEM0008
- MEM0009
- ARW-wsm6_ysu
- FV3-gfdl_edmf
- HRRR-thm_mynn
- NAM-fa_myj
- NSSL-wsm6_myj
- ARWLAG-wsm6_ysu
- FV3LAG-gfdl_edmf
- HRRRLAG-thm_mynn
- NAMLAG-fa_myj
- NSSLLAG-wsm6_myj

Stars: RRFS
Circles: HREF
Credit: Jake Vancil (CIWRO)
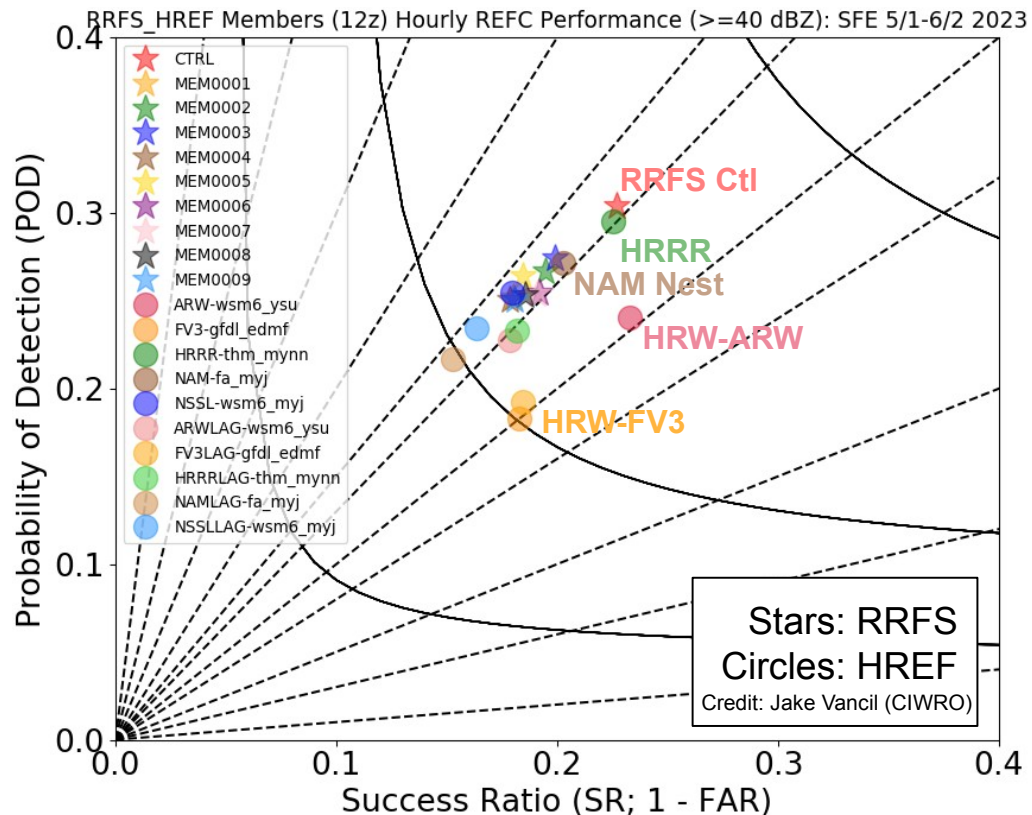
12Z RRFS Ens vs. HREF (Day 1): Member Perf.

- The perturbed RRFS members are closely clustered in terms of performance (red stars)
- The HREF members have widely varying performance characteristics (dark blue circles)
- The time-lagged HREF members (light blue circles) are clustered as lower performers
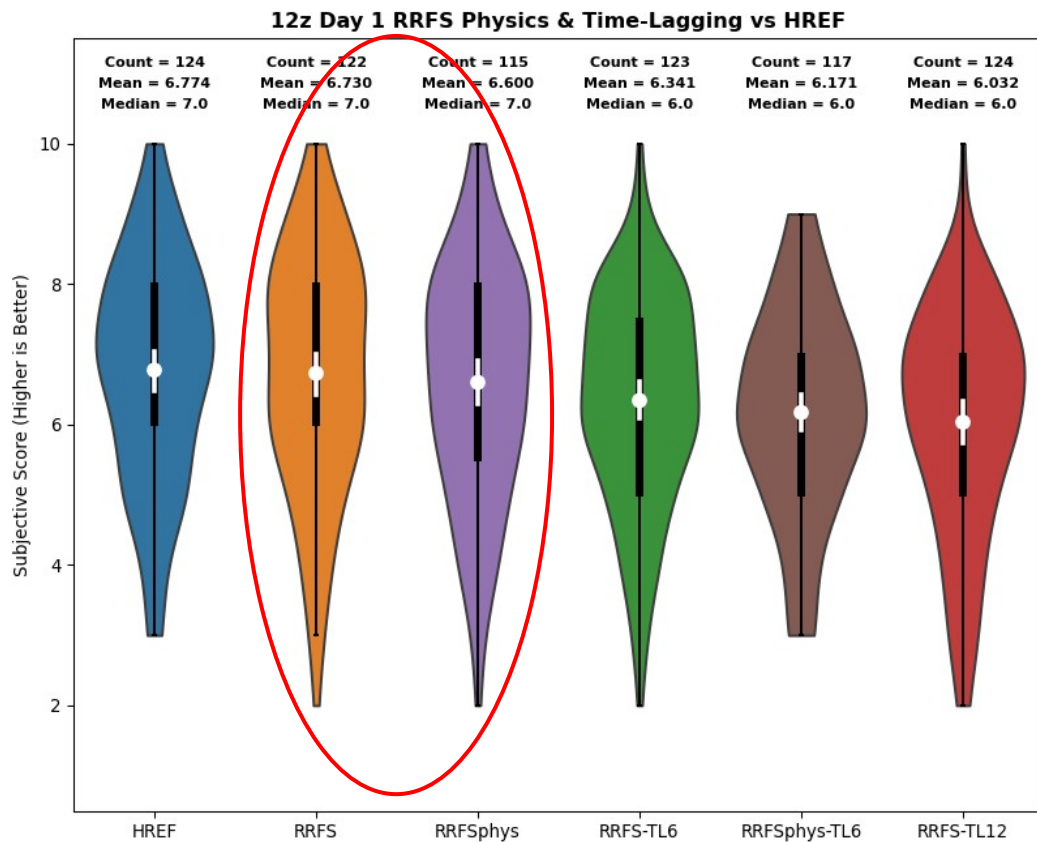- Despite being comprised of worse-performing members, the HREF **ensemble** still provides more skillful & reliable probabilistic guidance

RRFS_HREF Members (12z) Hourly REFC Performance (>=40 dBZ): SFE 5/1-6/2 2023

Legend:
CTRL, MEM0001, MEM0002, MEM0003, MEM0004, MEM0005, MEM0006, MEM0007, MEM0008, MEM0009, ARW-wsm6_ysu, FV3-gfdl_edmf, HRRR-thm_mynn, NAM-fa_myj, NSSL-wsm6_myj, ARWLAG-wsm6_ysu, FV3LAG-gfdl_edmf, HRRRLAG-thm_mynn, NAMLAG-fa_myj, NSSLLAG-wsm6_myj

Axis labels: Probability of Detection (POD) vs Success Ratio (SR: 1 - FAR)
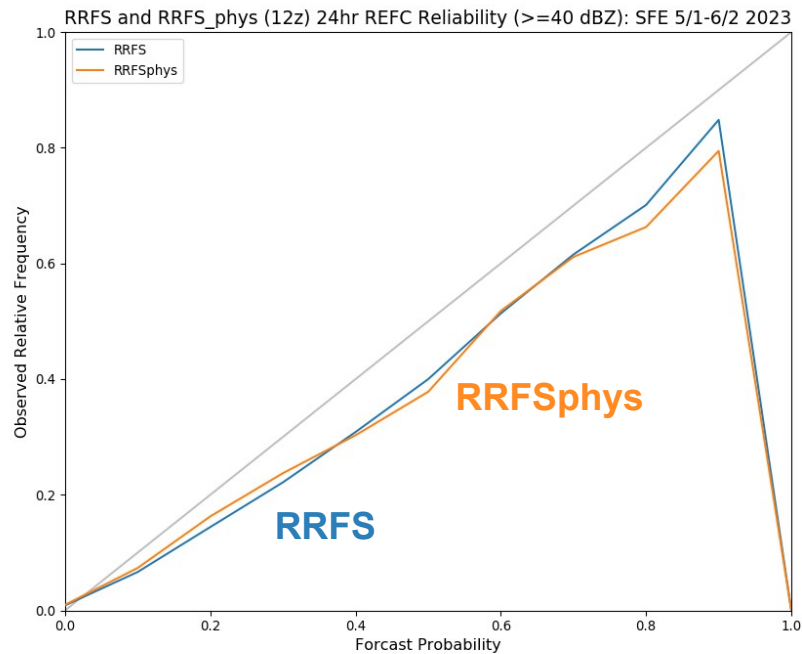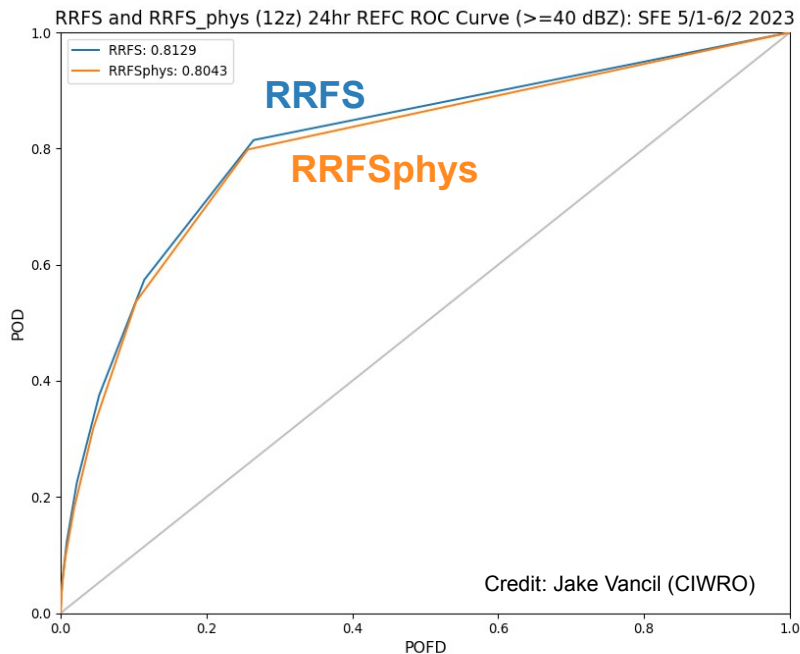
Red: RRFS
Blue: HREF
Credit: Jake Vancil (CIWRO)

12z RRFS Ensemble Physics (Day 1): Storm Attributes

- Extremely similar rating distributions for the Day 1 12Z single-physics RRFS and multi-physics RRFSphys

12z Day 1 RRFS Physics & Time-Lagging vs HREF

12z RRFS Ensemble Physics (Day 1): ROC & Reliability

RRFS and RRFS_phys (12z) 24hr REFC ROC Curve (>=40 dBZ): SFE 5/1-6/2 2023

- RRFS: 0.8129
- RRFSphys: 0.8043

RRFS

RRFSphys

POD

POFD

Credit: Jake Vancil (CIWRO)

RRFS and RRFS_phys (12z) 24hr REFC Reliability (>=40 dBZ): SFE 5/1-6/2 2023

- RRFS
- RRFSphys

RRFSphys

RRFS

Observed Relative Frequency

Forcast Probability

- Objective verification statistics over the SFE domains confirm very similar RRFS ensemble performance characteristics for deep convection (>40 dBZ reflectivity).

12z RRFS Ensemble Physics (Day 1): Member Perf.

- The control member clearly stands out as the best performer
- The perturbed single-physics members (MYNN-Thompson; stars) cluster fairly closely.
- Perturbations in members 03 (blue), 02 (green), and 05 (yellow) perform slightly better than the others.
- The MYNN physics members (03/blue and 05/yellow circles) perform the best from the mulit-physics ensemble

RRFS_RRFSPHYS Members (12z) Hourly REFC Performance (>=40 dBZ): SFE 5/1-6/2 2023

Stars: single-physics
Circles: multi-physics
Credit: Jake Vancil (CIWRO)
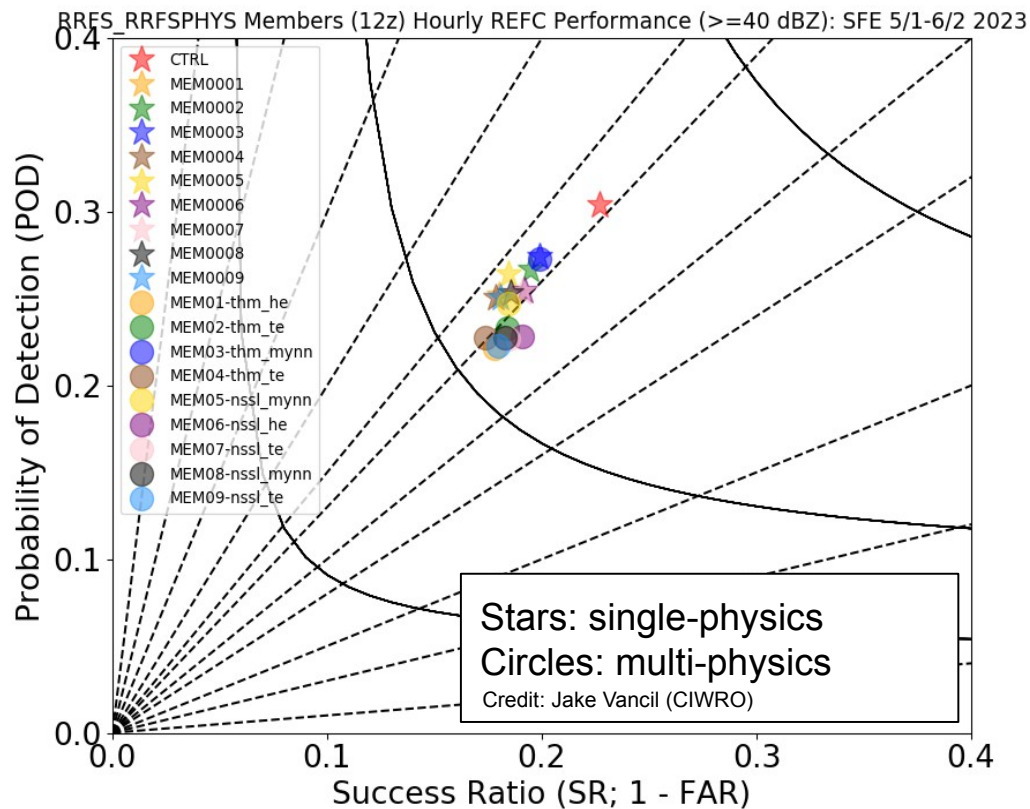
12z RRFS Ensemble Physics (Day 1): Member Perf.

- Clearly, the perturbed physics members are closely clustered as the worst performers (blue circles)

- Thus, using a poorer performing physics suite (especially GFS EDMF PBL scheme) in 6 of 10 members puts the multi-phsyics RRFS ensemble at a disadvantage

RRFS_RRFSPHYS Members (12z) Hourly REFC Performance (>=40 dBZ): SFE 5/1-6/2 2023

Red: single-physics
Blue: multi-physics
Credit: Jake Vancil (CIWRO)
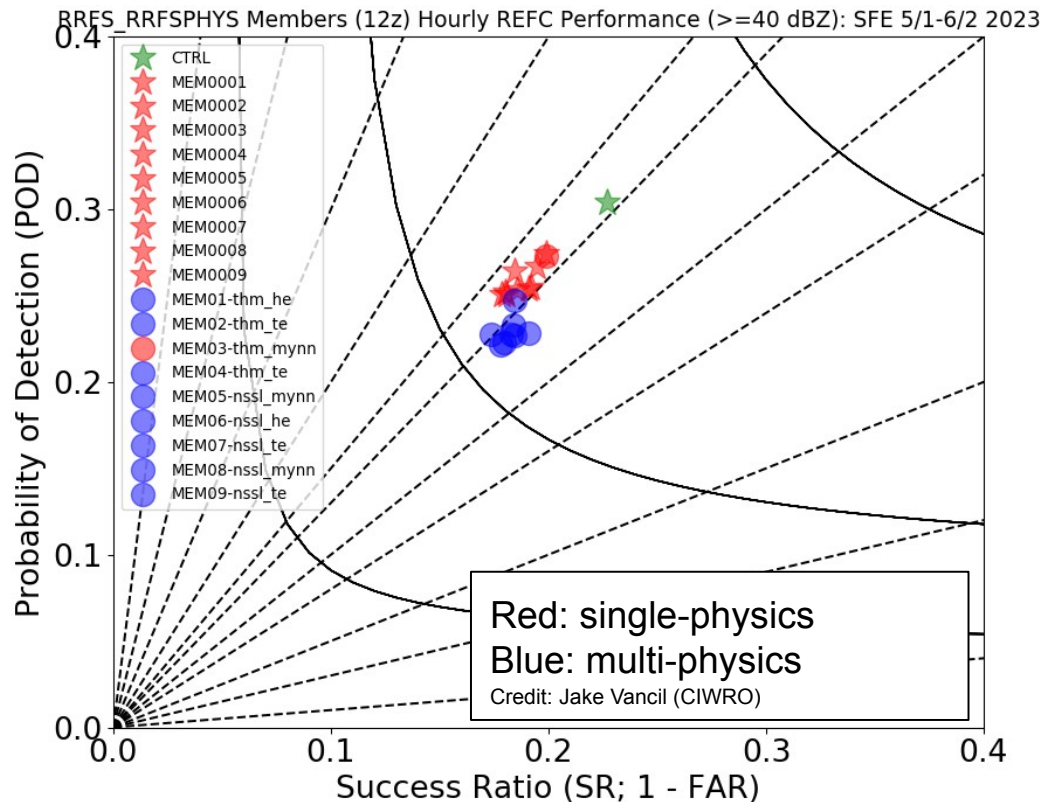
12z RRFS Time Lagging (Day 1): Storm Attributes

- The RRFS time-lagged ensembles had similar rating distributions to one another
- However, the RRFS time-lagged ensembles had lower mean and median ratings than the HREF
- With RRFS time-lagged ensembles being proposed to replace the HREF, this is concerning

12z Day 1 RRFS Physics & Time-Lagging vs HREF

Count = 124	Count = 122	Count = 115	Count = 123	Count = 117	Count = 124
Mean = 6.774	Mean = 6.730	Mean = 6.600	Mean = 6.341	Mean = 6.171	Mean = 6.032
Median = 7.0	Median = 7.0	Median = 7.0	Median = 6.0	Median = 6.0	Median = 6.0
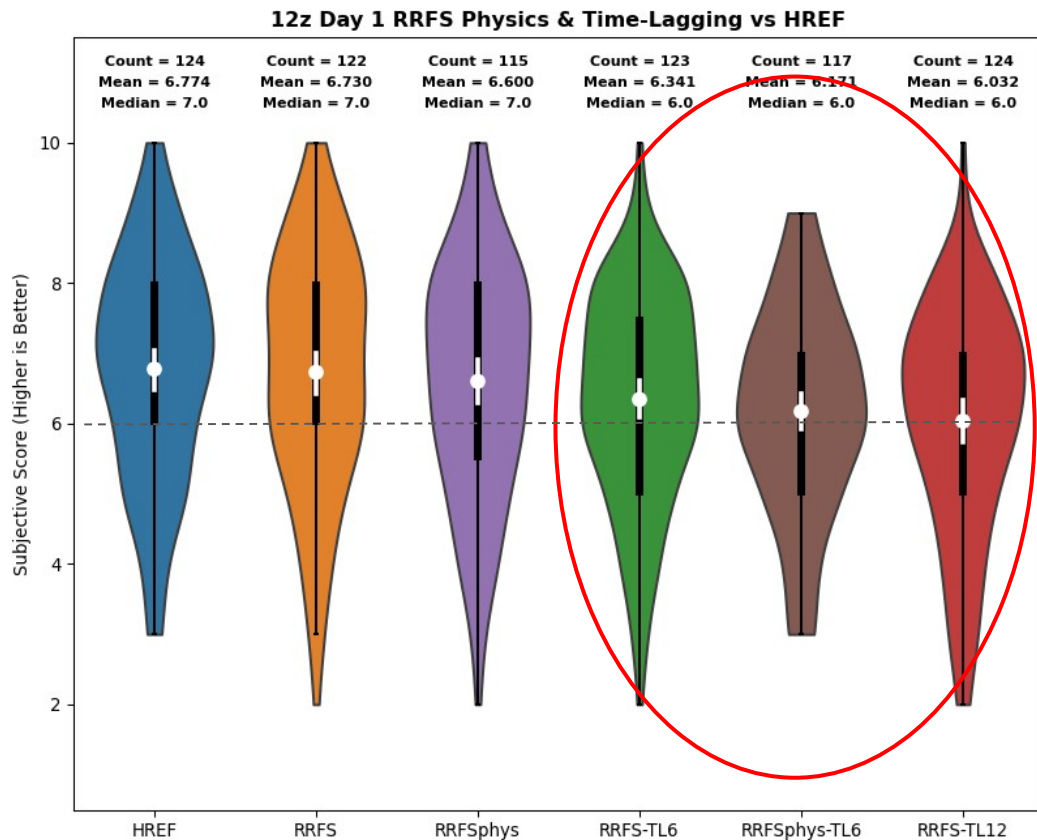
12Z RRFS Ens vs. HREF (Day 2): Storm Attributes

- One of the more surprising and strongest results occurred with Day 2 forecasts
- HREF has much higher rated Day 2 forecasts overall compared to RRFS
- All of the RRFS configurations (physics, time-lagged) had a similar distribution of ratings for Day 2

12z Day 2 RRFS Physics & Time-Lagging vs HREF
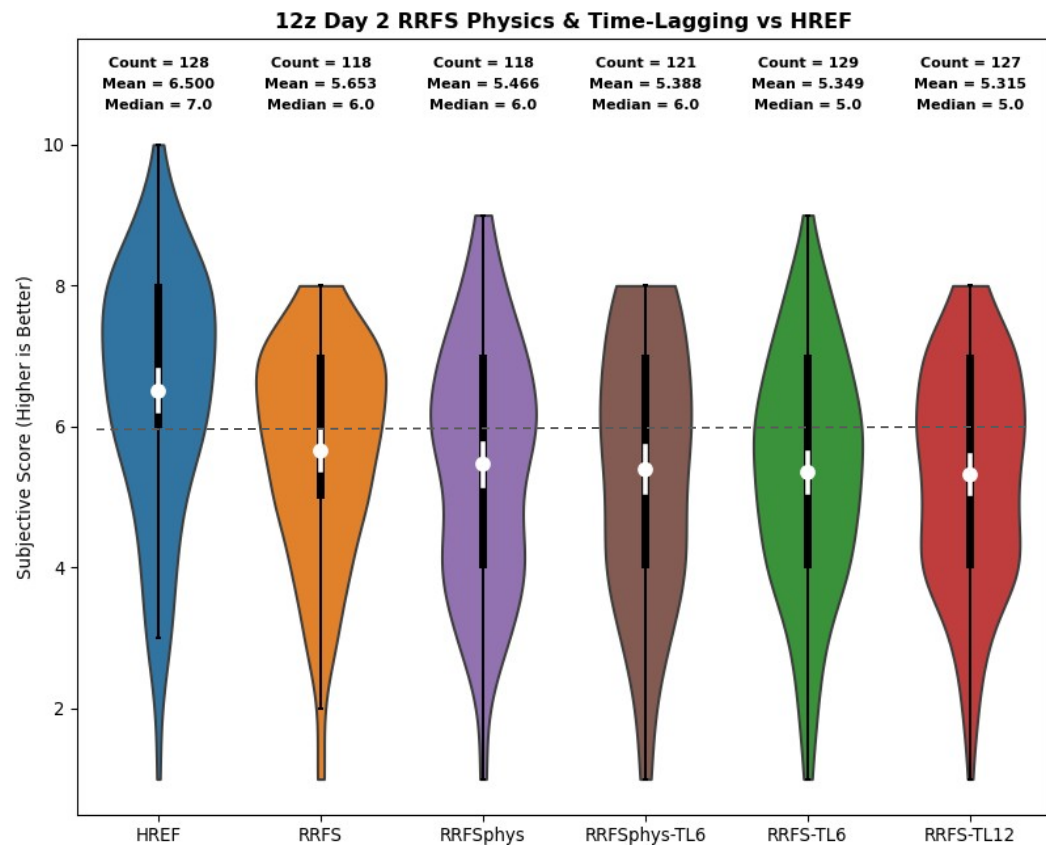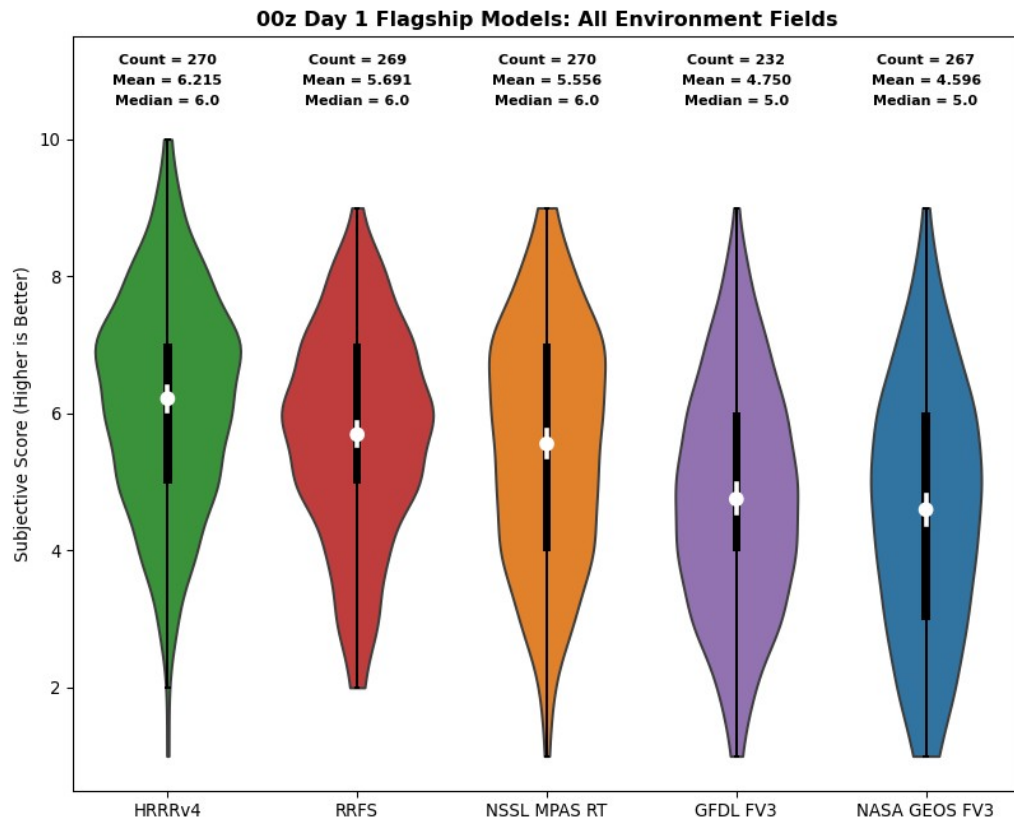
Summary and Conclusions

- There have been major strides in FV3 modeling at CAM scales
- The 00Z HRRR was rated as performing better than the 00Z RRFS control member, but the 12Z runs were much closer in terms of performance
- Despite poorer-performing members in the HREF, the probabilistic guidance is more skillful and reliable than the RRFS ensemble, which tends to be underdispersive and overconfident
- The RRFS multi-physics ensemble is likely not optimally configured, owing to being comprised of a majority of poorer-performing members (w/ GFS PBL)
- The time-lagged RRFS ensembles were rated lower on Day 1 than the ensembles initialized at a single time and the HREF
- The HREF forecasts for Day 2 were rated much higher than all of the RRFS configurations for that period

Extra Slides

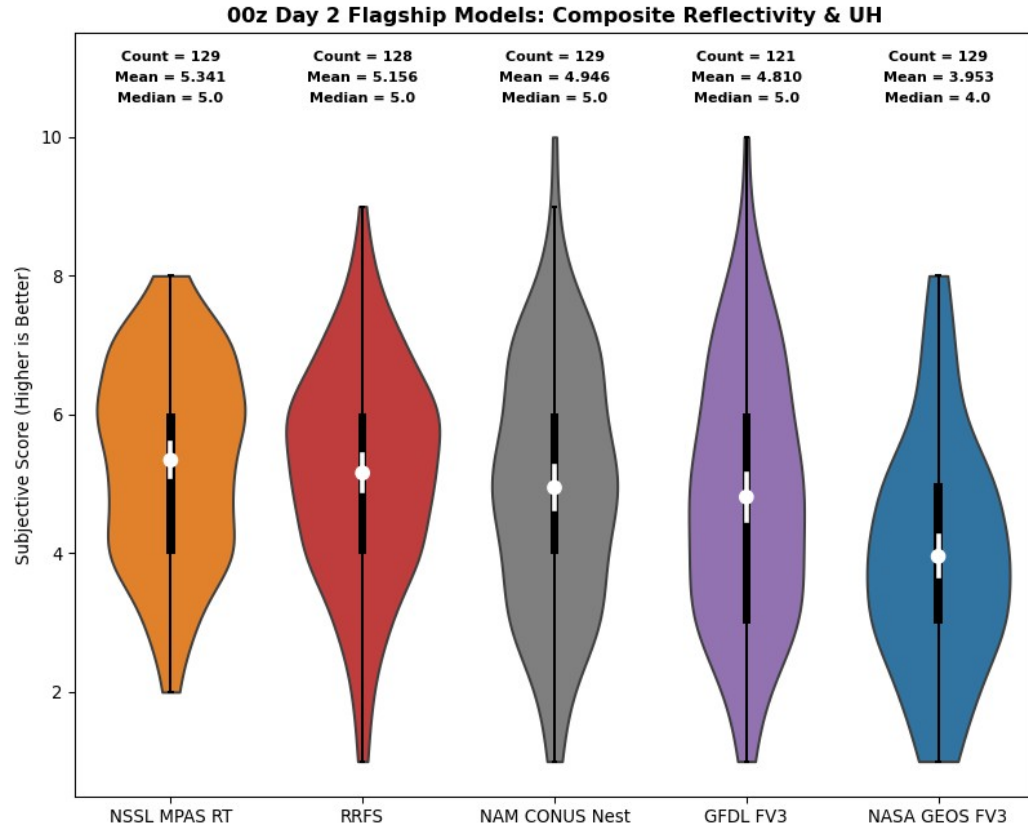# 00Z Blinded Flagships (Day 1): Environment

Blinded Evaluation

- HRRR also had highest mean rating for all environmental fields combined
- RRFS had the second-highest mean rating for the environmental fields - SBCAPE was the lowest rated RRFS environmental field (not shown)

00z Day 1 Flagship Models: All Environment Fields

Count = 270	Count = 269	Count = 270	Count = 232	Count = 267
Mean = 6.215	Mean = 5.691	Mean = 5.556	Mean = 4.750	Mean = 4.596
Median = 6.0	Median = 6.0	Median = 6.0	Median = 5.0	Median = 5.0

Subjective Score (Higher is Better)

HRRRv4   RRFS   NSSL MPAS RT   GFDL FV3   NASA GEOS FV3

# 00Z Blinded Flagships (Day 2): Reflectivity

Blinded Evaluation

- No HRRR for full Day 2, so replaced with NAM Nest
- Rating distributions were more similar across the models
- Very slight edge in the mean rating to NSSL MPAS actually over the RRFS

00z Day 2 Flagship Models: Composite Reflectivity & UH

| | Count = 129 Mean = 5.341 Median = 5.0 | Count = 128 Mean = 5.156 Median = 5.0 | Count = 129 Mean = 4.946 Median = 5.0 | Count = 121 Mean = 4.810 Median = 5.0 | Count = 129 Mean = 3.953 Median = 4.0 |

# 00Z Blinded Flagships (Day 2): Environment

Blinded Evaluation

- RRFS had highest mean rating for all environmental fields combined, but there is strong overlap in the distributions with NSSL MPAS and NAM Nest

00z Day 2 Flagship Models: All Environment Fields
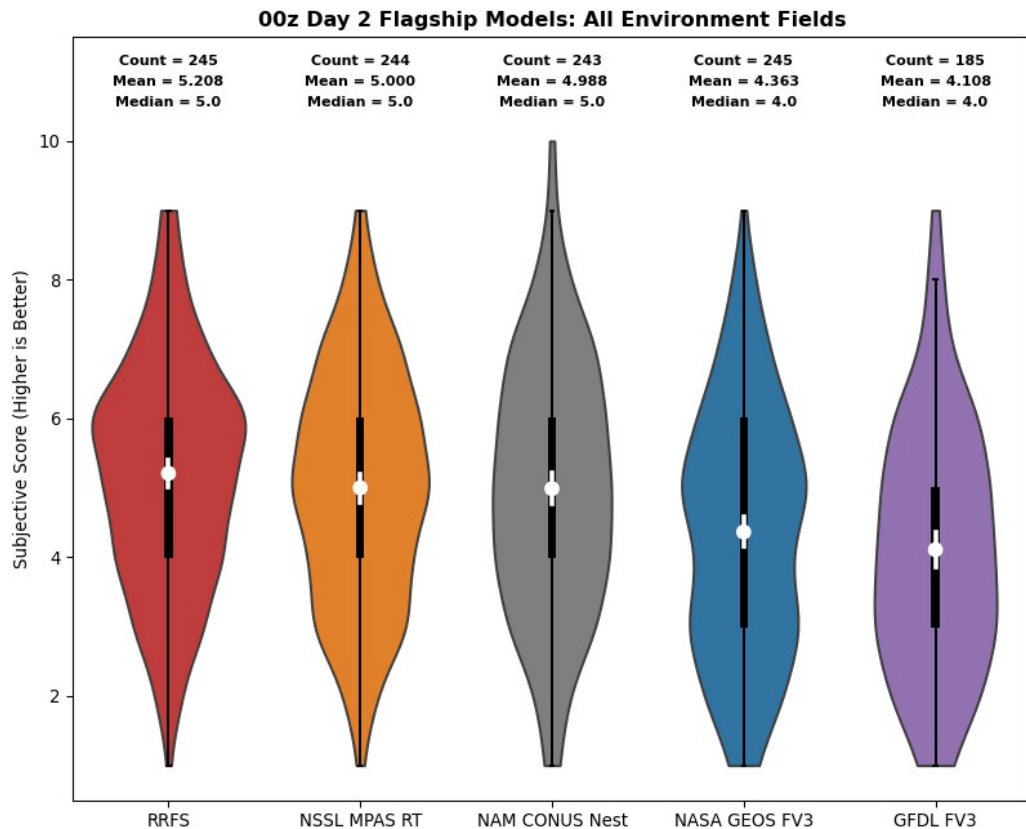
12Z RRFS Control vs. HRRR (Day 1): Environment

- For 12Z environment fields, the distributions are also shifted toward the middle ("Models performed about the same")
- SBCAPE forecasts were the most common environmental field to be favored for the HRRR
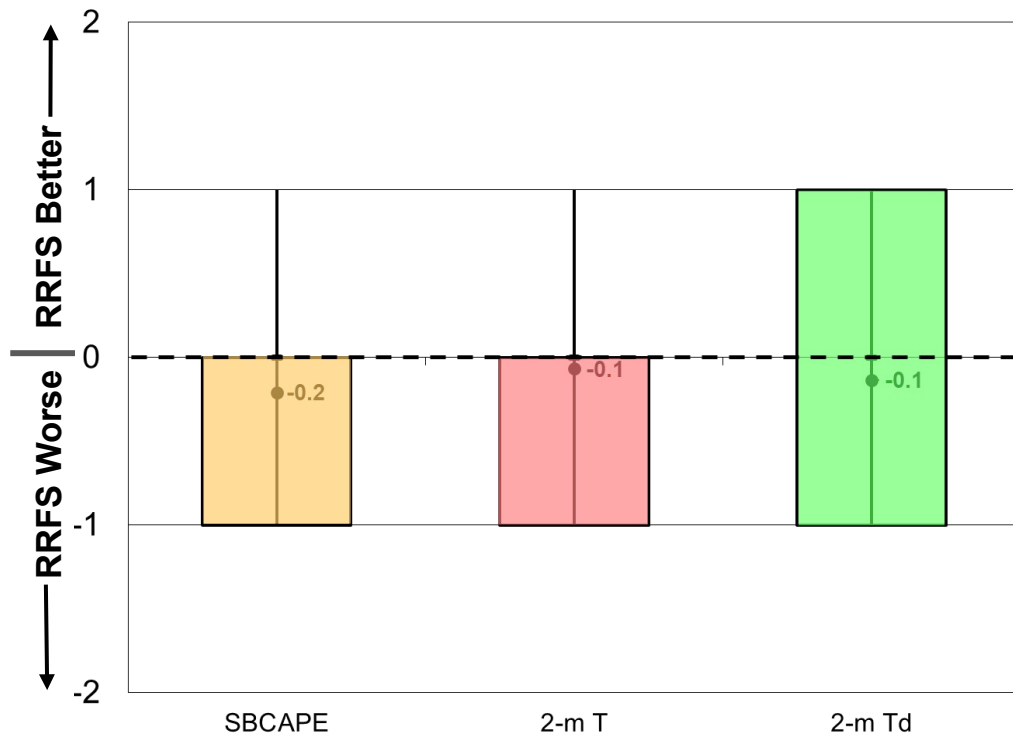
RRFS Ensemble Physics: Subjective Evaluation

- Personal subjective comparison of RRFS single-physics to multi-physics for 38 forecasts between May 1 - June 17
- Focused on 24-h **updraft helicity** and **updraft speed** probabilities for severe weather forecasting valid for the convective day from the 12Z cycle
- Overall, the forecasts were very similar - often not different enough to result in a change to a severe weather outlook
- Results using 5-Point Likert Scale shown by number of ratings:
  - **0:** Multi-Physics *Much Worse*
  - **2:** Multi-Physics *Slightly Worse*
  - **30:** Multi-Physics *About the Same* as Single-Physics
  - **6:** Multi-Physics *Slightly Better*
  - **0:** Multi-Physics *Much Better*

RRFS Ensemble Physics: Examples

- Before May 12, multi-physics RRFS (RRFSphys; bottom row) commonly had higher peak neighborhood ensemble probabilities than the single-physics RRFS (RRFS; top row)
- After May 12, the behavior seemed to change to the expected result - lower probs in the multi-physics RRFS

May 5 **May 29** **May 14**

Higher Probs   Lower Probs   Much Lower Probs
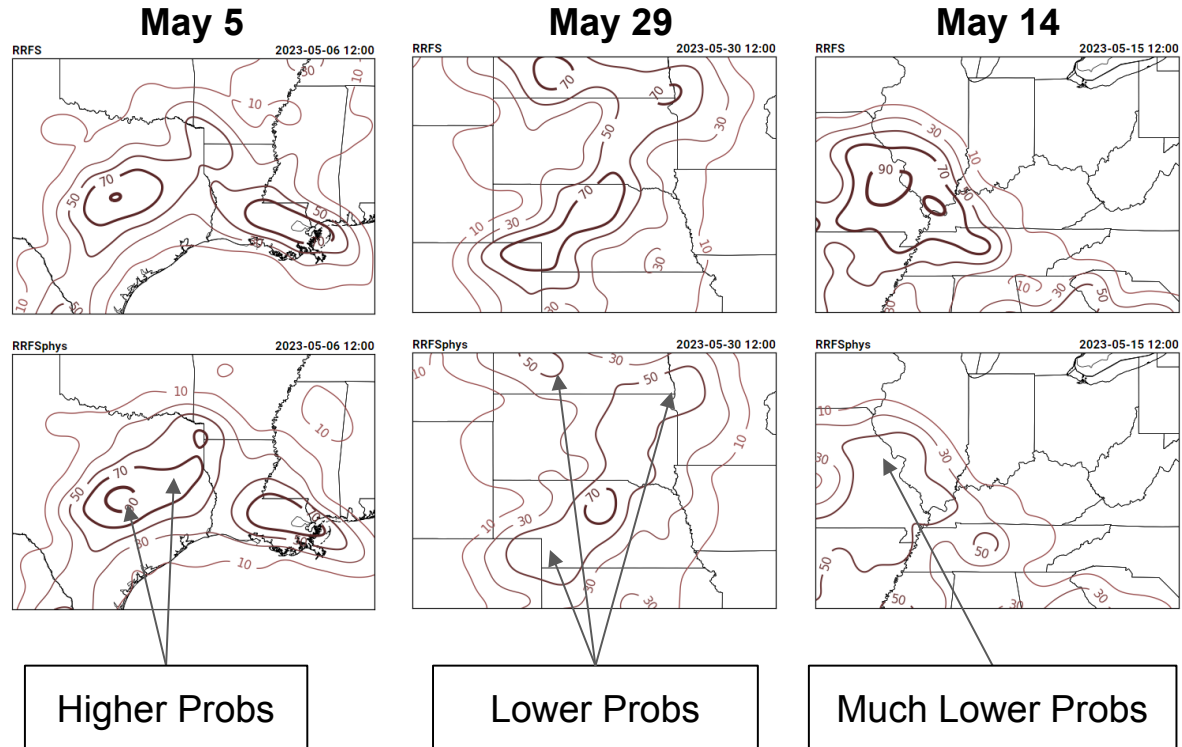
24-h NMEP of Updraft Speed >20 m/s

2023 HWT SFE: 12z Day 2 RRFS Physics & TL vs. HREF

- For Day 2 forecasts, the single-physics RRFS and multi-physics RRFSphys had similar mean ratings among SFE participants, but the multi-physics RRFSphys had a bi-modal distribution

- HREF mean rating by SFE participants was statistically significantly higher than all of the RRFS configurations for Day 2

12z Day 2 RRFS Physics & Time-Lagging vs HREF