U.S. Department of Commerce
National Oceanic and Atmospheric Administration
National Weather Service
National Centers for Environmental Prediction
5830 University Research Court
College Park, MD 20740-3818

Office Note 516

https://doi.org/10.25923/ccgj-7140

Mitigation Efforts to Address Rapid Refresh Forecast System (RRFS) v1 Dynamical Core Performance Issues and Recommendations for RRFS v2

Jacob Carley[1], Curtis Alexander[2], Louis Wicker[3], Christiane Jablonowski[4], Adam Clark[3], James Nelson[5], Israel Jirak[6], Kevin Viner[1]

NOAA/NWS/NCEP Environmental Modeling Center

College Park, Maryland

December 2023

[1] NOAA/NWS/NCEP/EMC
[2] NOAA/OAR/GSL
[3] NOAA/OAR/NSSL
[4] University of Michigan
[5] NOAA/NWS/NCEP/WPC
[6] NOAA/NWS/NCEP/SPC

E-mail: Jacob.Carley@noaa.gov

Review Team[1]

Name	Affiliation
Brian Gross	NWS/NCEP/EMC
Vijay Tallapragada	NWS/NCEP/EMC
Jennifer Mahoney	OAR/GSL
Young-Joon “YJ” Kim	NWS/AFS
Kevin Garrett	NWS/OSTI
Youngsun Jung	NWS/OSTI

[1] NOAA/GFDL did not respond to a request for review.

Table of Contents

Executive Summary

The RRFS is being developed as the flagship convective-scale ensemble-based prediction system for NWS operations to replace legacy limited-area modeling systems currently in operations in early calendar year 2025. The RRFS is an application based on the Unified Forecast System (UFS) and, like other UFS-based applications, uses the FV3 dynamical core. At this time, both deterministic and ensemble components of RRFS lag behind their operational counterparts, the High-Resolution Rapid Refresh (HRRR) and High-Resolution Ensemble Forecast (HREF) systems, respectively. HRRR and HREF are currently the principal applications used by forecasters for convective-scale prediction to meet the mission of the NWS[2]. The RRFS features forecasts of convective-storms that are too-intense which leads to a significant high bias in the warm season.

After approximately 5+ years of development, the principal developers of the RRFS are in consensus that the FV3 dynamical core has a limited future at the convective[3] (and finer) scales. Addressing the problem would require a significant overhaul of at least two of the key components of the FV3: the C-D grid discretization and the Lagrangian vertical coordinate. This would involve an effort spanning several years, culminating in what would effectively be a second dynamical core in the UFS. Most importantly, a successful outcome is not guaranteed.

There is an alternative dynamical core that has been tested at the convective-scale and shown to produce results within expectations for modern-era convective-scale NWP applications, the Model for Prediction Across Scales (MPAS). These tests featured physics configurations matching what is used in the current RRFS. Results showed that the MPAS-based runs outperformed the FV3-based runs using the same initial conditions for convective-applications.

While convective performance in the first version of RRFS lags behind operational systems and has reached an asymptote, RRFSv1 does present substantial improvements in terms of product availability and modeling system unification. Some of these improvements include:

- RRFSv1 uses a bigger domain, longer forecast length, and includes ensemble forecasts as well high resolution smoke and dust over North America.
- RRFSv1 unifies many CAM-scale operational modeling systems for the NWS within the UFS, enhances the NWS's ability to upgrade and maintain the NCEP Production Suite, and more effectively uses the scientific talents of the NOAA and NWS workforce by unifying development efforts on a single system.

[2] While comparison relative to the HRRR is of primary focus in this document there exist other convective-scale prediction systems that are in regular operational use, such as the North American Mesoscale forecast system's 3 km nests and the suite of High Resolution Window applications.

[3] Convective-scale refers to grid-spacings of approximately 3-km.

- RRFSv1 enhances service equity by bringing rapidly updating, high resolution NWP to historically underserved regions (Alaska, Hawaii, and Puerto Rico).

We therefore recommend the following two paths forward:

- Continue limited development of the first version of RRFS with the FV3 dynamical core in order to realize the above improvements in operations. Development will end following a beta evaluation near the end of March 2024. The outcome of this evaluation will determine how, or if, we proceed with the implementation process of RRFS (targeting early calendar year 2025). The likely outcome is one which involves retaining the HRRR in the present production suite to ameliorate risks associated with poor convective-storm prediction.
- For version 2 of RRFS, the team replaces the FV3 dynamical core with MPAS, which has already demonstrated to have superior performance at the convective-scale relative to FV3 in the 2023 NOAA Hazardous Weather Testbed (HWT) Spring Forecasting Experiment. Replacing the dynamical core will require a multi-year effort to integrate into the UFS and ensure as-good-or-better performance relative to RRFSv1 beyond elements of hazardous convective weather.

This approach allows the current and future versions of RRFS to best serve the NWS mission to protect lives, property, and enhance the national economy in a way that realizes the benefits of the current system, manages its flaws, and gets the system on track for sustainable improvements with a proven dynamical core for these applications and finer. Finally, it must be stressed that the scope of this paper only covers the RRFS. Other applications, such as global and hurricane, are beyond scope.

1. Introduction

Development of what is the Rapid Refresh Forecast System (RRFS) began soon after the Next Generation Global Prediction System (NGGPS) project's selection of the Finite-Volume Cubed-Sphere (FV3) dynamical core in 2016 (Ji 2016) as the atmospheric core in the Unified Forecast System[4](UFS). During the next seven years the development teams for the North American Mesoscale forecast system (NAM),the Rapid Refresh (RAP)/High Resolution Rapid Refresh (HRRR), and the High Resolution Ensemble Forecast (HREF) completed the final upgrades of their respective systems[5] and formed a unified, collaborative development team around the UFS-based RRFS. Their vision was to make it NOAA's flagship high-resolution ensemble prediction system underpinned by the transition to a new atmospheric dynamical core, the FV3 (Harris and Lin 2013, 2014; Lin 2004; Putman and Lin 2007).

The RRFS development team is a multi-organizational, cross-line-office group with well over 50+ years of combined experience in convective-scale NWP development, implementations, and applications. One primary goal identified by the Team is that the RRFS is to facilitate the retirement of the rather large number of disparate high-resolution models in the NCEP production suite (Fig. 1.1) and consolidate all regional convective-scale applications under RRFS.

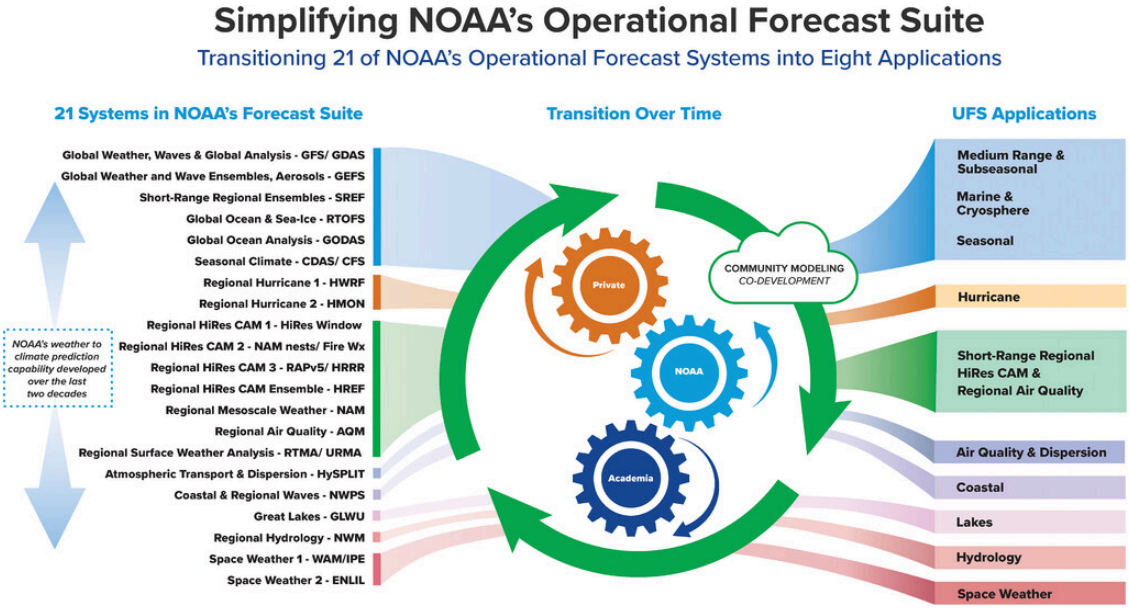

Fig. 1.1. From Uccellini et al. (2022), "Unification of the NWS operational modeling suite through community-driven UFS, reducing its 21 stand-alone operational forecast models into eight major forecast systems, or forecast applications, that share the same community modeling framework." In this diagram the RRFS is represented by the green shading in the right column labeled, "Short-Range Regional HiRes CAM & Regional Air Quality".

[4] https://ufscommunity.org/

[5] The NAM was last upgraded in March of 2017. The RAP/HRRR were last upgraded in December of 2020. The HREF was last upgraded in May 2021.

Unification of the regional suite around the RRFS will provide a greater level of service to NWS stakeholders that is both more sustainable and has greater potential for future improvements. In order to effectively consolidate the regional suite the RRFS must deliver a comparable or better level of service *and* performance relative to the systems it will replace. Service refers to attributes like update cadence, grid-spacing, and domain size (i.e., coverage), and products tailored to meet stakeholder requirements. Performance refers to characteristics of forecast quality, like skill and bias.

To meet service requirements the RRFS covers North America at 3-km grid-spacing, features hourly updates with a hybrid 3DEnVar data assimilation system that cycles its own 3-km ensemble, provides forecasts every hour out to 18 hours, and provides ensemble forecasts to 48+hrs every 6 hours (i.e., at synoptic times). The RRFS configuration is based upon the RAP/HRRR physics and data assimilation components - a suite that is known to perform well in operations and provides a proven foundation to build upon. This design allows the RRFS to match and exceed the level of service currently provided by the NAM nests, HRRR, HiRes Window, and HREF systems (Fig. 1.2).

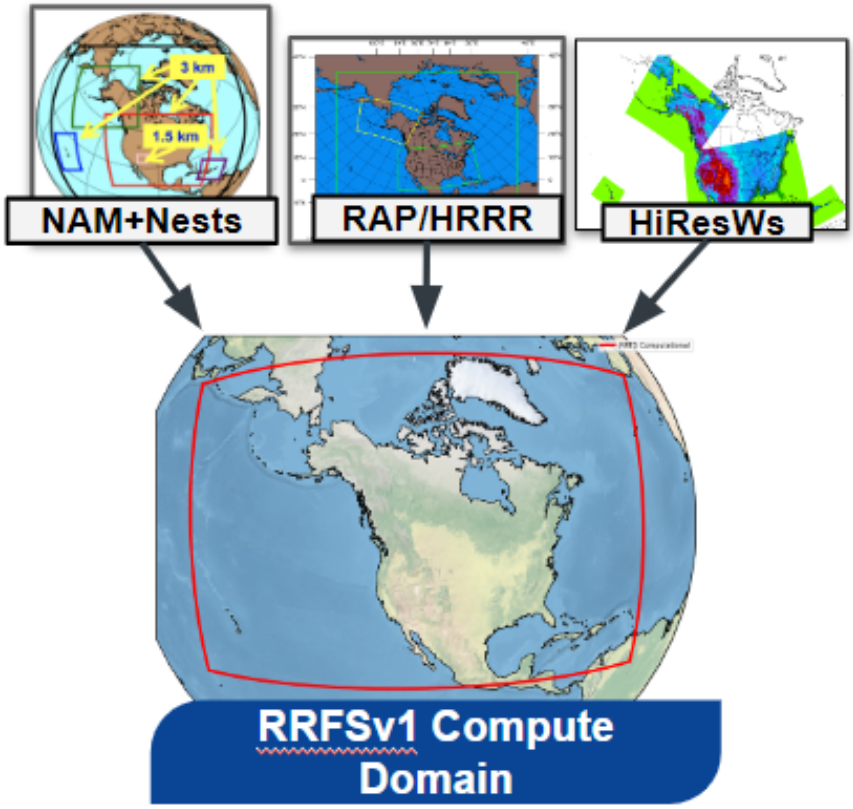

Fig. 1.2 The computational domains of the NAM and nests, RAP/HRRR, and HiRes Window systems along with the RRFS domain (bottom). Note that all legacy domains fit within the RRFS.

In a broad sense, the performance of RRFS must meet or exceed the skill of the models it is planned to replace. In order to assess the performance of RRFS, the development team had to introduce and connect several fundamental capabilities to the FV3 dynamical core that were necessary for operational convective-scale NWP. Some of those features included: a limited area model capability (Black et al. 2021), a refined computational grid that minimizes variance in cell size across the domain, integration of physics appropriate for convective-scales through CCPP, and connection to data assimilation infrastructure. While not an exhaustive list, each effort took 2-3 years to accomplish.

As development progressed, it became apparent that the RRFS system suffered from problems with representing convective storms with the same fidelity as the present operational suite, including the HRRR and HREF. This was both clear in subjective assessments, feedback from NOAA's flagship testbeds, as well as in objective verification (Figs. 1.3 and 1.4). Specifically, the RRFS was found to predict storms that are too intense in both rainfall and reflectivity. These storms also have a tendency to be too widespread. Further, individual convective cells tend to be larger in size with atypical shapes (e.g., circular and symmetrical) than operational counterparts.

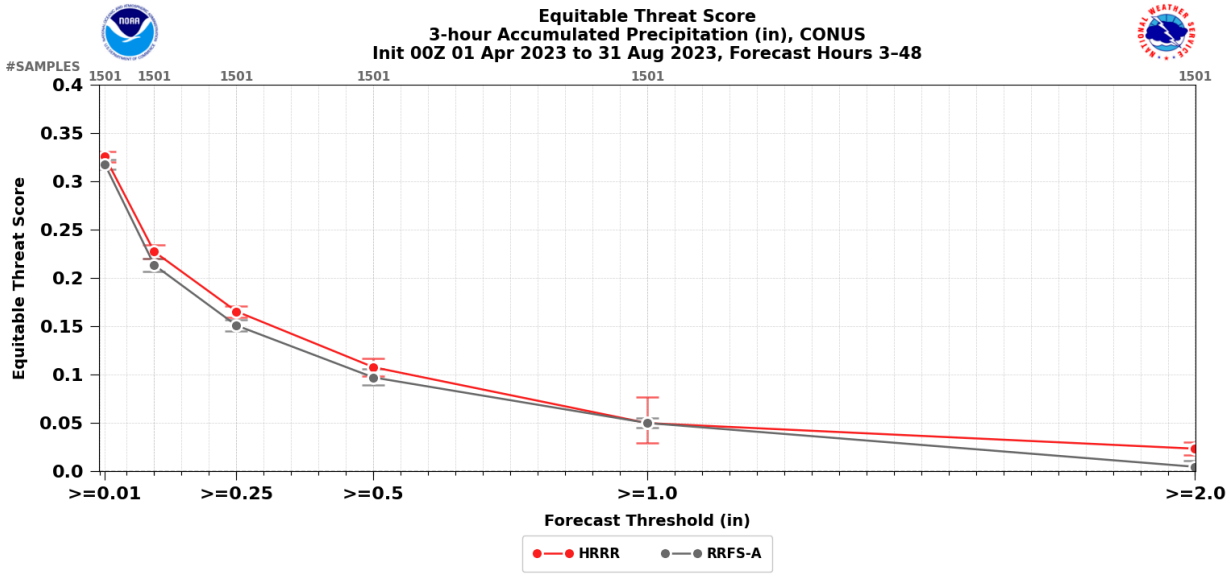

Fig. 1.3. Equitable Threat Score by precipitation threshold comparing HRRR (red) and RRFS-A (gray) for 3-hr accumulation intervals for a 48-hr forecast period from 1 April to 31 August 2023. The 'A' in RRFS_A refers to its status as the primary experimental system running in real-time.

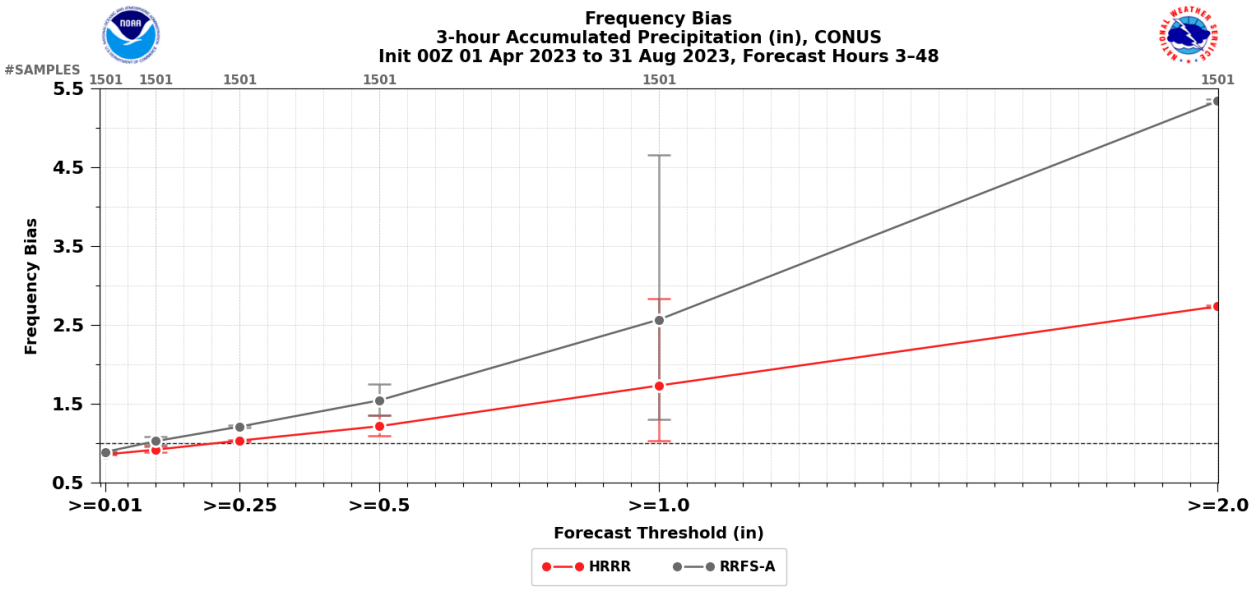

Fig. 1.4 Frequency Bias by precipitation threshold comparing HRRR (red) and RRFS-A (gray) for 3-hr accumulation intervals for a 48-hr forecast period from 1 April to 31 August 2023.

Both developers and stakeholders have noted that these storms tend to be most prevalent in convectively unstable environments with relatively weak forcing and vertical wind shear. Such conditions are relatively common in the southeastern United States in the summer (Fig. 1.5). This issue is especially problematic as quality model guidance to support forecasts of convective hazards, such as flash flooding and severe storms, are critical to support the NWS mission and Impact-Based Decision Support Services (IDSS).

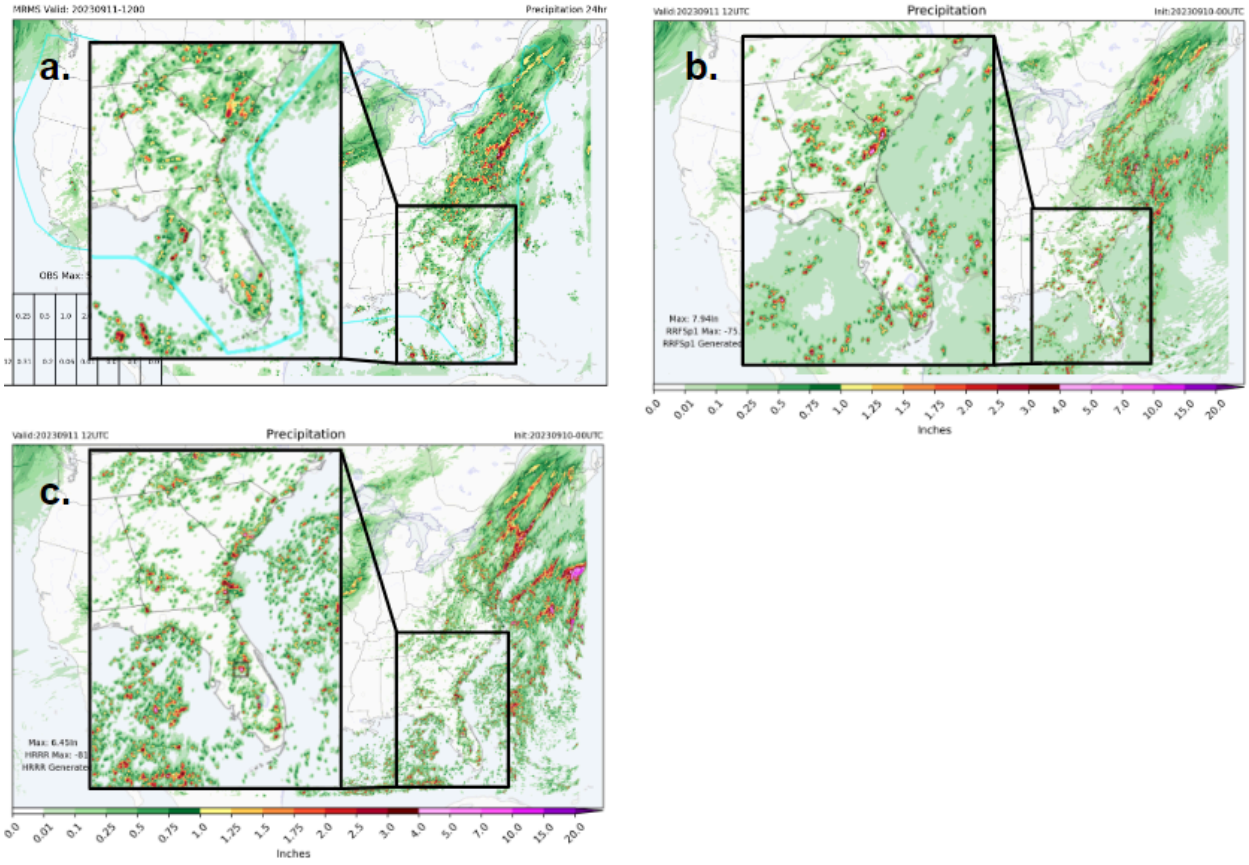

Fig. 1.5. Accumulated precipitation over a 24-hr period valid 11 September 2023 at 1200 UTC from (a) MRMS analyzed precipitation, (b) a RRFS prototype forecast, and (c) a HRRR forecast. Both models were initialized on 10 September 2023 at 0000 UTC. This configuration of RRFS was running with parameterized deep convection, yet still exhibits the issue of too-intense convection with storms that are too large in size - especially over Florida. Subjectively the precipitation forecast by the HRRR, which runs no parameterized deep convection, is finer scale and exhibits a wider spectrum of precipitation intensities.

Developers have spent considerable time and effort attempting to (1) understand why the RRFS exhibits this pathology, despite the fact that it runs the same physics package as found in the operational HRRR and (2) from that understanding, improve the forecast. We have employed idealized testing, analysis of real-time forecasts with expert analysis via NOAA testbeds, performed case studies, and run experiments with multiple dynamical cores (FV3, CM1, ARW, and MPAS). While the combination of these efforts have led to improvements in the RRFS, it still lags behind operational baselines (Fig. 1.1 and 1.2). We have reached the difficult conclusion that the FV3 dynamical core is the issue - the same pathologies are not seen in experiments with other cores. Also note that the data assimilation framework and physics package closely aligns with the operational HRRR system - the primary difference is the dynamics. For example, OAR/NSSL ran a RRFS-like configuration with a different dynamical core, MPAS, at the 2023 Hazardous Weather Testbed Spring Forecast Experiment. Of the three configurations evaluated in the testbed, one was configured to use RRFS-generated initial conditions and an RRFS-like physics package. When compared to actual RRFS

forecasts of convection, the MPAS-based configuration was a considerable improvement over the FV3-based RRFS.

It is with this abundance of testing and evaluation among the development team and key stakeholders, all represented on this paper, that we recommend that future versions of the RRFS adopt a new atmospheric dynamical core, i.e., MPAS. It should be noted that there exists some risk that MPAS-based RRFS, although showing better performance, may encounter unexpected challenges since it is presently less mature in NWS operational forecast applications than the FV3 dynamical core. We believe, however, it is a reasonable risk given the circumstances. Moreover, adopting MPAS for RRFS while retaining FV3 for other applications (e.g., global and hurricane, respectively) would introduce an additional cost for maintenance in the UFS and a discrepancy from the FV3-driven global model that provides lateral boundary conditions[6] . These issues will need to be discussed and resolved separately.

This paper is organized as follows: Section 2 describes the testing, development, and evaluation efforts by the RRFS development team. Section 3 provides hypotheses regarding the source of the convective-storm issues in FV3 and provides an estimated scope and cost of a hypothetical project that would be required to attempt to address the problems in the dynamical core. It also provides an estimated scope and cost of a project to switch dynamical cores from FV3 to MPAS. Section 4 provides a recommended path forward for RRFSv1 and beyond.

2. Record of Past Efforts and Outcomes

2.1. Idealized Analyses of Convective Storms

Section 2.1 describes efforts to identify the underlying causes of the differences in the RRFS solutions (from the HRRR) using idealized storm simulations.

2.1.1. Initial Analysis of HRRR Versus RRFS Storm Characteristics

After forecasters and quantitative assessments identified important differences between the HRRR and RRFS solutions from the 2019-2020 spring forecast experiments with the horizontal size and spatial characteristics of the simulated storm reflectivity fields as well as excessive precipitation accumulations, NSSL began to examine the storm-scale attributes using object- (i.e., storm-) based methodologies developed for the Warn on Forecast system's error metrics. One analysis is depicted in Fig. 2.1 that displays the differences between the updrafts profiles and storm sizes

[6] We emphasize that runs using MPAS, a different dynamical core than the global, outperform those from RRFS, even when using RRFS initial conditions. This suggests that discrepancies from the global model are comparatively less important than the suitability of the underlying dynamical core. These results are discussed in Section 2.3.1.

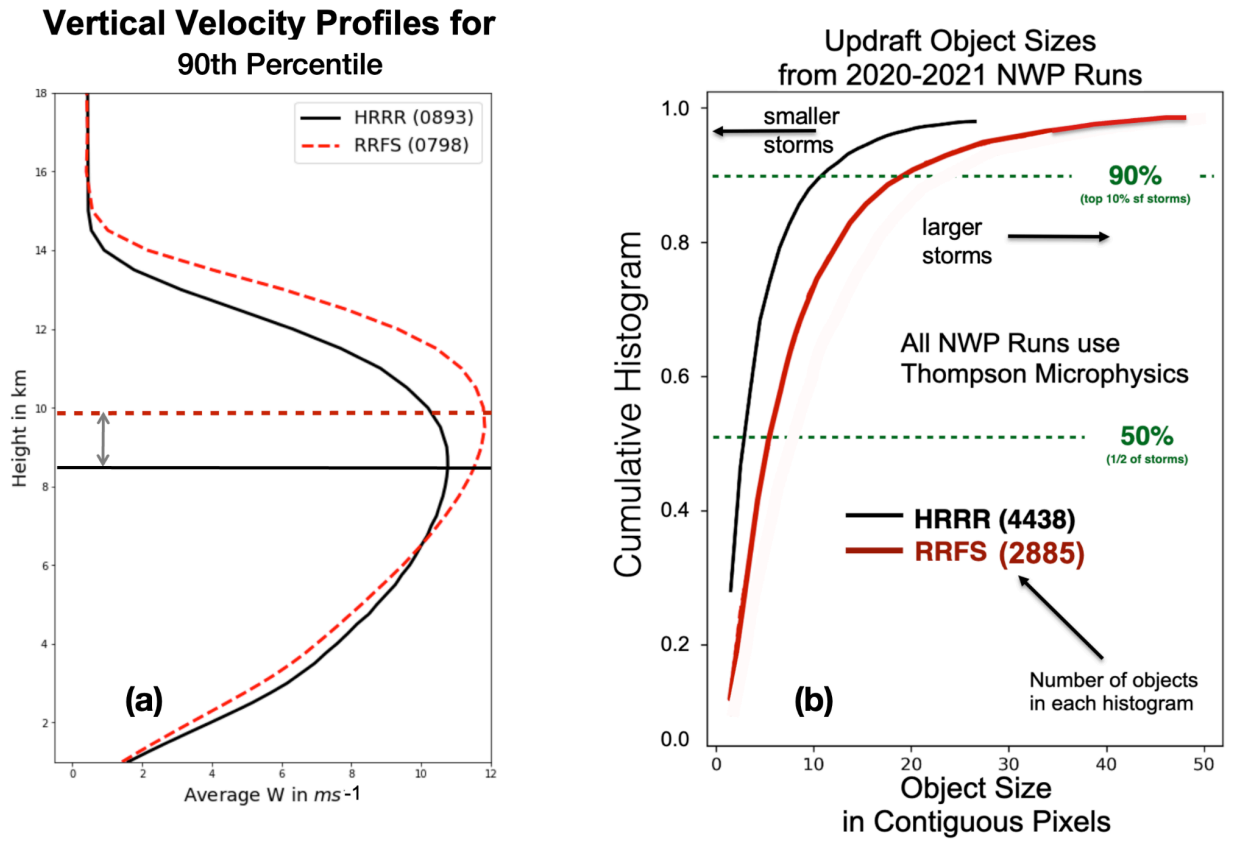

Figure 2.1: (a) Aggregated profiles of vertical velocity. The solid black (red dashed) line indicates the height of the maximum w for the HRRR (RRFS) (b) Cumulative frequency diagram shows the number of pixels of each storm object from the HRRR and RRFS 18 hours of forecast during the 2020-2021 Spring Forecast Experiments.

obtained from thousands of storm objects from those forecasts. This shows that the RRFS model has stronger and deeper updrafts than HRRR (Fig. 2.1a), and that storms in the RRFS are larger with the 90th percentile objects having twice the area coverage than HRRR objects (Fig. 2.1b). After some initial efforts, it was determined that the complexity associated with full physics runs would make uncovering the source of the differences unlikely. If the systematic differences between models can be reproduced using a simpler version of the model, the physics, the environment, and the initial conditions, this likely would provide a better opportunity to determine the source of these behaviors. As will be shown, most of the important differences (primarily associated with deep convection) between the HRRR and the RRFS can be reproduced using idealized squall line simulations. The source of large precipitation biases seen in the RRFS (already discussed in Sec. 1, Fig. 1.5) appear to be closely related to these solution differences and reproducible using simplified experiments.

Since most continental convection occurs in lines or clusters, it seemed appropriate to next examine convective storm characteristics within idealized squall line simulations using simple homogeneous environments. Unlike supercell simulations, squall line simulations provide a much larger number of storm objects which can be analyzed and then can provide meaningful statistics. Given that there are few direct (or

indirect) observations of updraft profiles available to validate the predicted solutions from real data cases, there are also no "observations" to validate the idealized simulations. Therefore, we chose the best available model(s) to help determine the potential sources of the differences. A simplified version of the RRFS dynamical core is available from GFDL, the FV3 "SOLO" model (hereafter, SOLO[7]). SOLO is used to simulate convective storms in an idealized environment. SOLO solutions are compared to two models that are well known to produce skillful and realistic simulations for convective storms, the Weather Research and Forecasting model (WRF, Skamarock et al. 2021) and the Cloud Model 1 (CM1, Bryan and Fritsch 2002). The WRF model has been NCAR's community model for the past 20 years and is the basis for the operational HRRR model at NCEP. CM1 is a state-of-the-art research non-hydrostatic simulation code used to simulate convection, hurricanes, turbulence, as well as other atmospheric phenomena and has been used by hundreds of researchers over the past decade[8].

2.1.2. Idealized Simulations Experimental Design

A large number of simulations are conducted using three different unidirectional vertical shear profiles and five different CAPE environments. The results shown here focus on two CAPE environments and two profiles that are representative of the general results for the full experimental set of solutions. In order to minimize differences between the various models, the experiments are constructed using the following criteria and code changes:

- Vertical diffusion was reduced considerably (WRF) or removed (CM1) since vertical diffusion in the SOLO (and FV3) is minimal due to the Lagrangian vertical coordinate.
- The numerical approximations used in each model are set to either operational settings (RRFS & HRRR) or research settings which are close to the HRRR (CM1).
- A warm-rain Kessler microphysics scheme was used for all experiments. No other physical parameterizations were used (e.g., no PBL or radiation physics). The Kessler scheme was chosen because the algorithm is simple enough to check to ensure all coefficients within the three models are essentially identical.
- Reflectivity is computed using the SOLO algorithm for all three models.
- The vertical grid spacing in all models is nearly identical to that used in the RRFS, with the exception that the top of the model is at 10 mb, not 2 mb. The domain is 25 km deep. Vertical grid spacing at the surface is ~12 m.

[7]FV3 SOLO is GFDL's simplified version of the dynamical core. SOLO is nearly identical to the full FV3 (or RRFS) model. SOLO can be easily used to simulate idealized convection. The version used here was downloaded in March 2022 from GFDL's github repository.

[8]CM1 has been cited in more than 350 peer-reviewed published articles in the last decade in over 30 different journals.

- Horizontal grid spacing is 3 km. Doubly periodic domains are used. The horizontal domain is 768 km x 768 km.
- Due to known sensitivities associated with the precipitation fallout scheme, the physics time step is 20 seconds for all simulations.
- In order to limit differences in storm propagation effects for various CAPE environments that can be associated with various boundary layer relative humidities, the environmental soundings have identical relative humidity profiles below 1.6 km (McCaul and Weisman 2000). Different CAPE environments are created via increasing the lapse rates above the level of free convection.

The initial condition includes seven bubbles with a 3 K temperature perturbation that are aligned in the north-south direction along the western portion of the domain. The bubbles are centered 40 km apart. Simulations are integrated up to 6 hours with history output every 15 minutes. Figure 2.2 displays the squall line solutions after 5 hours for an environment using low-shear (6 m/s westerly shear below 2 km) and a moderate CAPE (2000 J/kg) environment. The two most noticeable features are the differences in cold pool size (which can be related back to precipitation fallout and evaporation) and the size of the color-filled "objects" (see caption for description).

To determine whether the idealized experiments "reproduce" the differences seen in the full-physics runs, Figure 2.3 shows the updraft profiles from the 2000 J/kg CAPE environment with two different vertical shear profiles (6 or 18 m/s shear over the lowest 2 km). Objects are identified by choosing the 99th percentile reflectivity value for each simulation and where vertical velocity is greater than 2 m/s above 700 hPa.

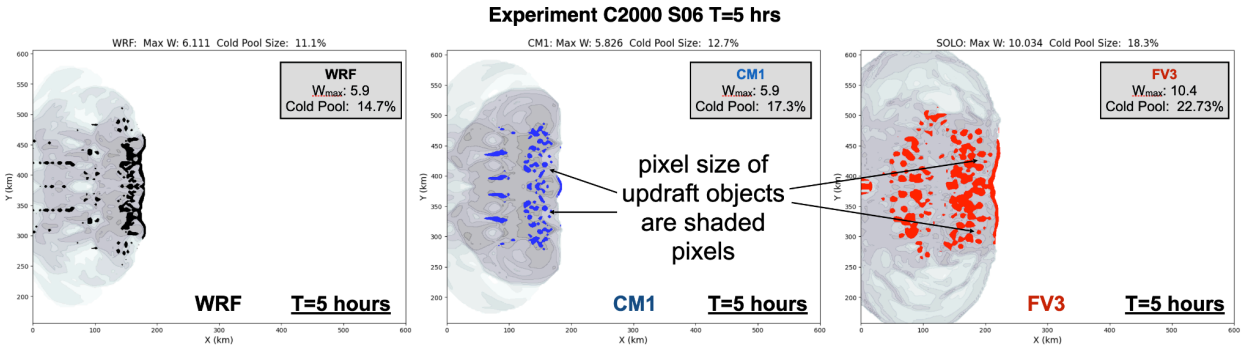

Figure 2.2: Horizontal cross-sections from squall line at 5 hours. The gray shaded regions are the cold pool (perturbation theta less than -1 K) and the solid colored regions indicate storm objects identified as regions where the composite reflectivity > 35 dBZ and vertical velocity above 700 mb is at least 2 m/s.

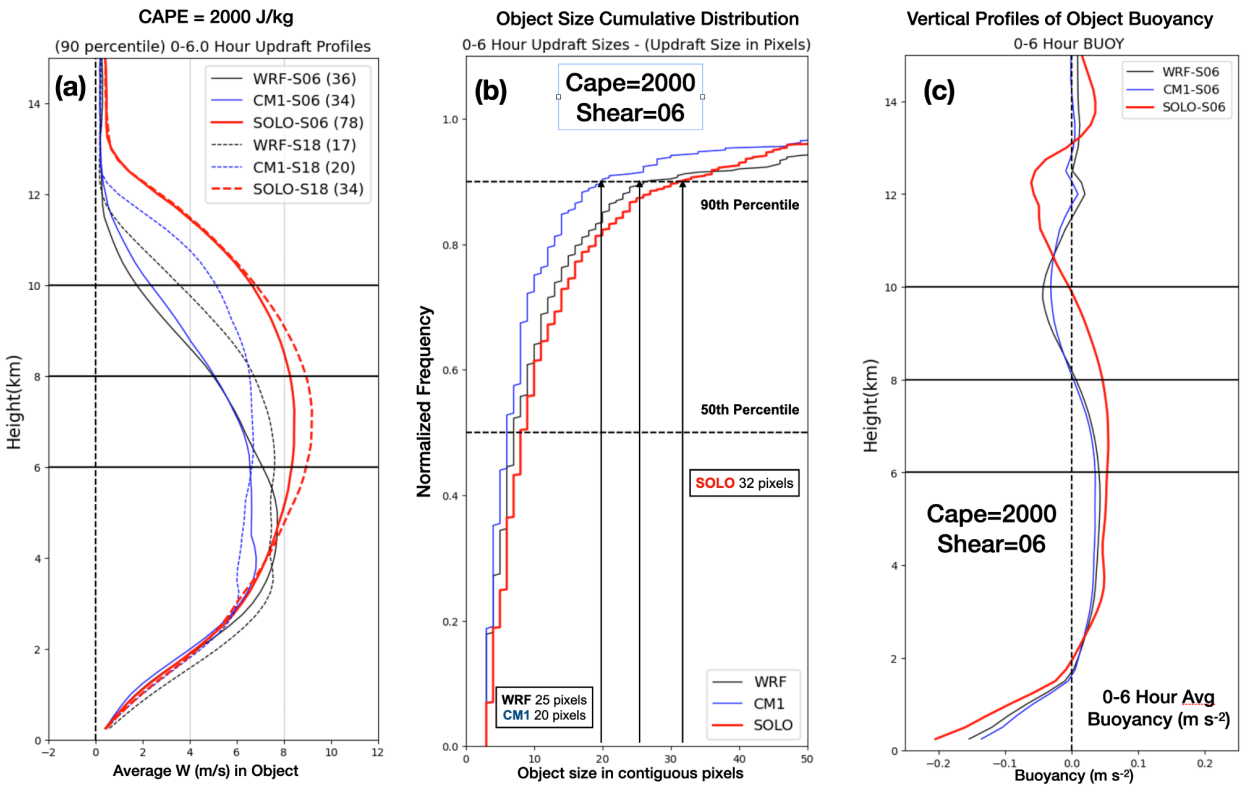

Figure 2.3: Storm object updraft profiles from the 2000 J/kg CAPE runs and for 6 (solid line) and 18 m/s (dashed line) shear over the lowest 2 km. Line colors indicate the model: SOLO (red), WRF (black), and CM1 (blue). (a) The average of the top 10% (90th percentile and above) for the entire 6 hour period. The number of objects used for the averaging is shown in parentheses within each line label. The solid and dashed lines represent the 6 and 18 m/s shears, respectively. (b) Cumulative frequency diagram showing the average number of pixels for each storm object over the 6 hour period for the 2000 CAPE and 6 m/s shear simulations. (c) Mean vertical profiles of buoyancy across all storm objects over 6 hours for the 2000 CAPE and 6 m/s shear simulations.

Fig. 2.3 shows an analysis of the resulting updrafts. The SOLO updrafts are always deeper and usually stronger than either the WRF or CM1 solutions (Fig. 2.3a). The updraft profiles are very similar to those shown from the full physics simulations shown previously (Fig. 2.1a). Figure 2.3b shows the cumulative distribution of pixel size from the storm objects. As in the full-physics runs (Fig. 2.1b), the SOLO storm objects are always larger than the WRF or CM1 objects, although the differences are somewhat smaller at the 90th percentile level than the full-physics runs. SOLO storm objects at the 90% level have an area approximately 40-50% larger than either the WRF or CM1 model. The vertical distribution of buoyancy within all objects (Figure 2.3c) shows that the SOLO simulation has larger negative buoyancy in the cold pool below 2.5 km, and then switches to be warmer over a significant depth of the troposphere above that level. The larger buoyancy in the SOLO simulation is likely due to less lateral entrainment due to the larger storm sizes. They are also associated with deeper and stronger updrafts within SOLO between 6-10 km as shown in Fig. 2.3a. At higher shears (e.g., the 18 m/s), the differences are somewhat smaller between the three models, as lateral entrainment from the environmental shear becomes more important. This effect can be

seen in Fig. 2.3a, as the differences between the updraft profiles are larger in the 6 m/s simulations (solid lines) than in the 18 m/s profiles (dashed lines).

Figure 2.4 shows the kernel density estimates of accumulated precipitation for any gridpoint over the 6 hour period from all three models in two environments. In the 2000 CAPE and 6 m/s shear (Fig. 2.4a), both the WRF and CM1 models have fairly similar precipitation distributions. The SOLO solution has lower accumulations below 25 mm, but has a very long "tail" relative to the WRF and CM1 distributions indicating large amounts of precipitation has fallen. For a significant number of grid points, SOLO produces rainfall exceeding 150 mm over 6 hours (~6 inches), 50% larger than the WRF or CM1 distributions. This SOLO "tail" likely represents extreme precipitation as seen in the full-physics RRFS discussed in Sec. 1 (Fig. 1.4). As in the dynamics, differences between the precipitation distributions is also a function of the environmental CAPE and shear. In the 3500 CAPE and 18 m/s shear runs (Fig. 2.4b) precipitation differences between model simulations are reduced considerably when compared to Fig. 2.4a. Even though reduced, Fig. 2.4b shows that the largest SOLO accumulations in the large CAPE and high shear environment still are 35-40% larger than WRF or CM1 in the middle of the density. Of course, grid point accumulations are smaller in all simulations for the high-shear case as individual storm cells are moving more rapidly eastward due to larger westerly winds aloft.

A number of experiments, from both baroclinic wave tests (Jablonowski, 2023 personal communication) and squall line experiments, suggest that the horizontal divergence damping in SOLO (and the RRFS) is important toward explaining these different model behaviors. Neither WRF or CM1 filter the horizontal divergence. Both RRFS and SOLO use a horizontal divergence filter to maintain solution stability (see Section 7.1, Harris et al. 2021). The non-dimensional coefficient used for divergence damping is large (between 0.1 and 0.16, see Whitehead et al. 2011). Horizontal divergence damping[9] (hereafter, HDD) traditionally is used to remove gravity waves and other grid scale vertical motion in hydrostatic models. However, when HDD is used in non-hydrostatic models a much smaller coefficient is employed. The NAM-3km and MPAS use a non-dimensional coefficient value 5 and 10 times smaller, respectively, than the RRFS (Janjic and Gall 2012, Skamarock personal communication 2023). To help reduce the impact on scales of interest in the solution, HDD in RRFS and SOLO use very high-order filtering (8th order) to be more scale selective. HDD using an 8th-order filter is a unique capability within the FV3 dynamical core and is not available in other non-hydrostatic models. However, the use of a 8th-order filter with a large coefficient for the HDD appears to then *generate wider updrafts that have less lateral entrainment and subsequently are warmer and deeper. This results in stronger storms and excessive precipitation.* The differences in the precipitation profiles in Fig. 2.4 already support part of this hypothesis. The SOLO solutions from the environment having less vertical shear (and less background entrainment) produce more

[9]Horizontal divergence damping should not be confused with 3D divergence damping used in WRF or CM1 to remove spurious sound waves that do become unstable due to the time splitting algorithm (Skamarock 1992). Filtering the compressible sound waves does not impact the meteorological solution.

precipitation relative to the other two models than precipitation in the higher shear environment.

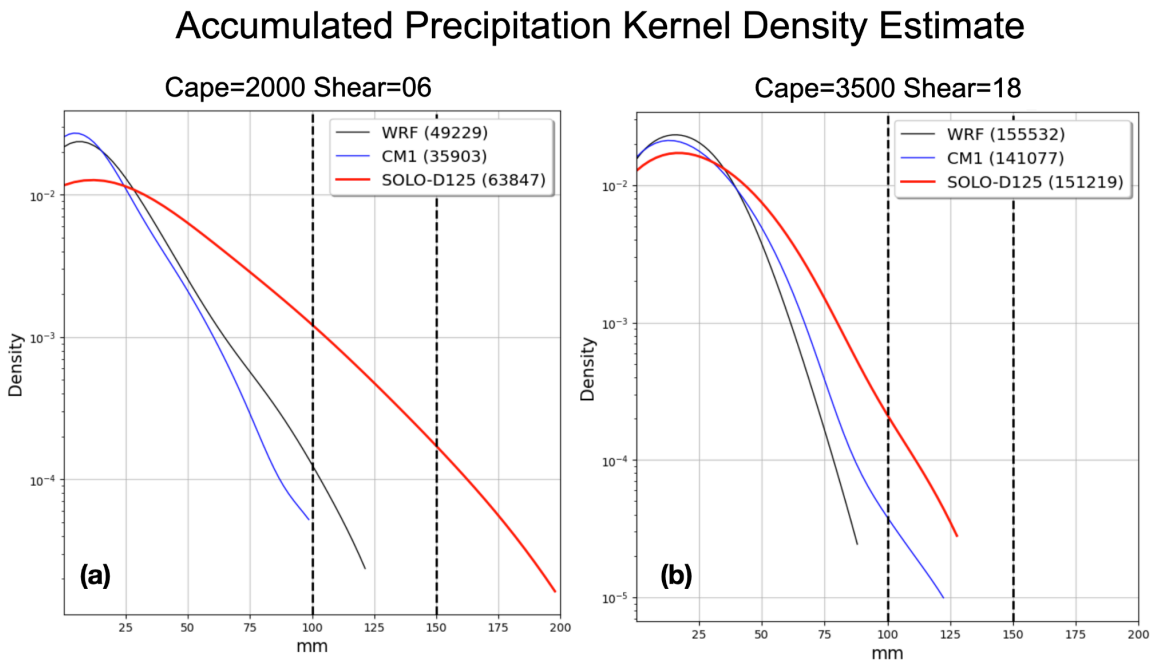

Figure 2.4: Kernel density estimates of the accumulated precipitation over the six hour period from the squall lines using the three models in two different environments (low-shear, moderate CAPE (left), high-shear, high CAPE (right)). The line colors indicate the model (SOLO (red), WRF (black), and CM1 (blue)). Note the y-axis scales for each panel are different.

The HDD hypothesis is tested in two ways. First, using SOLO, squall line simulations are performed using a value of the nondimensional HDD coefficient which is ten times smaller. The value of the reduced coefficient is similar to what is used in other NWP models. Second, 8th-order horizontal divergence damping is implemented in the CM1 model and squall line simulations are performed using the SOLO HDD coefficient (e.g., the RRFS value). *If the hypothesis is correct, the reduced HDD SOLO solutions should look more like the unfiltered CM1 solutions, and the HDD CM1 solutions should look more like the SOLO experiments*.

The results from both experiments are shown in Fig. 2.5. Figure 2.5a shows that updrafts in both model solutions are sensitive to the HDD filter coefficient. Updraft intensities are reduced when the background filter coefficient is reduced in SOLO, and increased when the HDD is applied in the CM1 simulation. Figure 2.5b shows the mean buoyancy profiles from the storms' updrafts. The differences are less obvious than the updraft profiles, but definitely have the correct trend in the middle troposphere. Changing the filter coefficient will result in warmer or cooler updraft cores when the HDD coefficient is larger or smaller. Figure 2.5c is the same as Fig. 2.4a, except that precipitation statistics from low-HDD SOLO-D012 and high-HDD CM1-D125 simulation are added. Reduction of HDD in SOLO results in less precipitation, consistent with the weaker updraft profile shown in Fig. 2.5a. However, the high HDD simulation from CM1 has slightly less precipitation. This is attributed to the fact that HDD in CM1 is

computed on the C-grid, which has effectively twice the resolution for divergent modes than the SOLO D-grid configuration. The larger filtering from HDD in CM1-D125 has less impact on the physical solution because the filtering is acting on an effectively finer grid. Further analysis is being conducted on this issue. A power spectra analysis of the v-wind at 1.3 km is shown in Fig. 2.5d. V-wind is chosen because initially it is zero in the environment, so this is the spectra of the perturbation winds with a zero mean. Spectra from other variables (pressure, u) also have similar structures. Note that CM1's solution has no accumulation of excess energy at 2 Δx as the slope in the CM1 spectra between 6 Δx and 2 Δx is nearly constant. This result is a consequence from computing the divergence and pressure gradient on the C-grid. This also explains why there is only a minor difference in the spectra from the high-HDD CM1-D125 simulation. In contrast, when the HDD coefficient is reduced by a factor of ten in the SOLO-D012 simulation, energy begins to accumulate at the grid scale, e.g., the "bump" in Fig. 2.5d for the SOLO-D012 spectra line. There is also much more energy in the SOLO-D012 solution between 4 Δx and 2 Δx than in the SOLO-D125. Overall, the SOLO-D012 spectra have ten times the energy of the SOLO-D125 spectra at 2 Δx. The buildup of energy at the smallest grid scale demonstrates why large values of HDD are required to suppress the buildup of 2 Δx noise. As suggested by Harris et al. (2021, Sec. 7.1), without a sufficiently large value of the HDD coefficient, real data simulations often become unstable due to the effects of terrain forcing and data assimilation.

These results may also explain the Zhao et al. (2012) "counterintuitive" result that increasing the HDD filter coefficient resulted in 40% more tropical cyclones in simulations using FV3. The authors expected that increasing the amount of HDD would reduce the number of tropical cyclones. Those results are consistent with this analysis, as increasing the HDD filtering can result in intensifying marginal convective structures. It seems likely that the marginal tropical convective structures (the convection that ordinarily would not organize) are intensified by the increased HDD, enabling deeper and stronger updrafts, which subsequently develop into tropical cyclones.

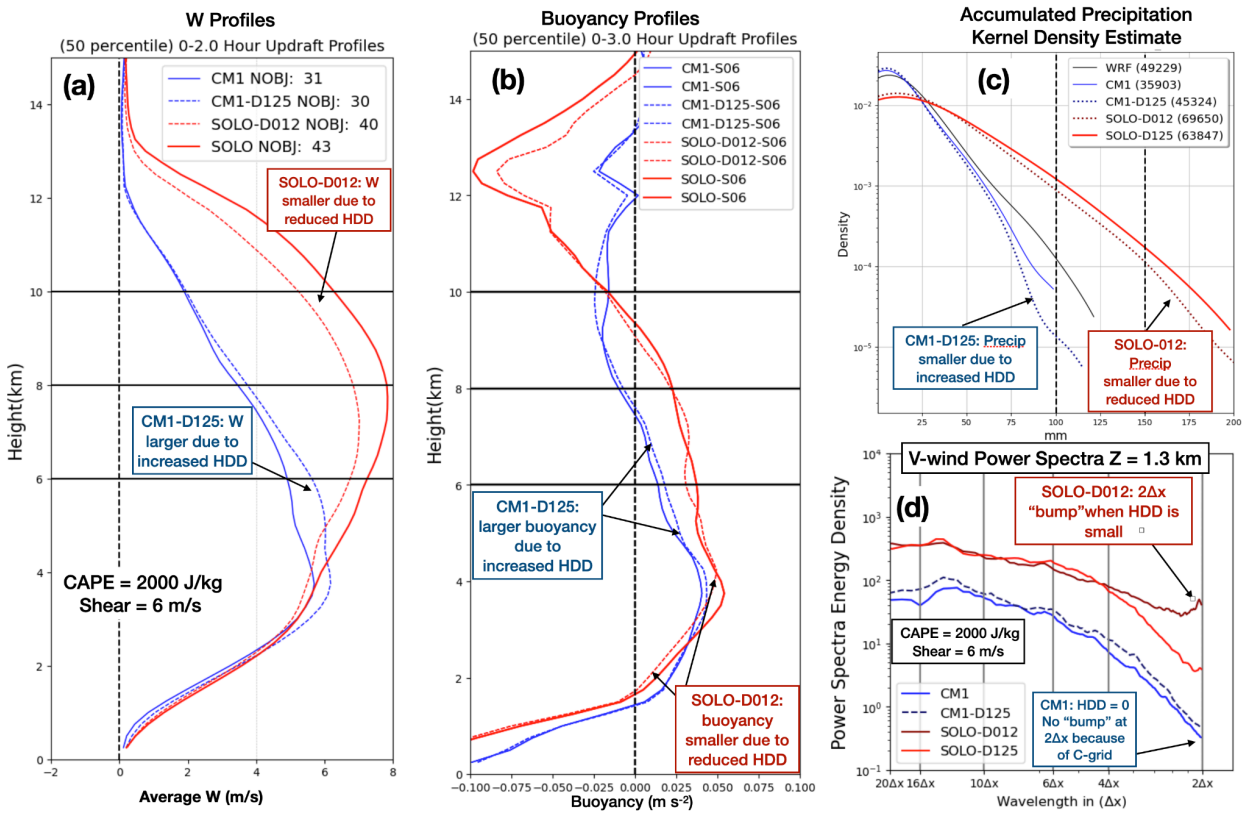

Figure 2.5: Storm object plots from the 2000 J/kg CAPE and 6 m/s (dashed line) shear over the lowest 2 km. The number of objects used for the averaging is shown in parentheses within each line label. (a) Average w profiles from the 50th percentile during 0-2 hours. The SOLO solid/dashed (red) lines have horizontal divergence damping coefficients of 0.12 and 0.012, respectively. CM1 solid/dashed lines are reversed, the HDD coefficient values are 0.0 and 0.125, respectively. (b) Average buoyancy profiles for the four simulations. (c) Same as Fig 2.4, with the addition of the SOLO simulation with reduced horizontal damping and the high-HDD simulation from CM1. (d) Horizontal power spectra for the V-wind field at 1.3 km over 0-6 hours for the four runs.

To summarize, the idealized experiments demonstrate similar behaviors to those seen in the full-physics simulations from HRRR and RRFS (Fig. 1.4, Fig. 2.1). Storm size, updraft profiles, and precipitation biases all mirror their full physics counterparts. This strongly suggests that, even with the simplicity of the experimental design, that the solution differences between the HRRR and RRFS or the WRF/CM1 and SOLO models are associated with the fundamental design of each dynamical core rather than the physics tuning or numerical filtering settings. The results demonstrate that the FV3 dynamical core has a fundamentally different behavior at the grid scale which is almost certainly due to use of the C-D-grid (the grid behaves, at least to a leading order, as a D-grid; Konor and Randall 2018a,b; Skamarock 2008). The C-D grid prioritizes a more accurate representation of the rotational motions and relegates effective resolution of the divergent motion to approximately half that of the C-grid models. Consequently, as shown in Fig. 2.5d, excess energy at 2 Δx is only removed using large values of horizontal divergence damping. This will be even more important in the full-physics simulations where external forcing (e.g., terrain forcing, data assimilation increments) will generate high-frequency forcing. Even with an 8th-order filter for the divergence,

the results show that large horizontal divergence damping results in wider updrafts that have warmer and deeper thermodynamic cores. The warmer and deeper updrafts are then extremely efficient at producing precipitation. These effects are particularly apparent in low-shear environments which are similar to those in the U.S. summer warm season where the RRFS biases are largest. These results are still valid at higher resolutions as initial squall line comparisons using 1.5 km grid spacing *still* show the differences between SOLO and WRF/CM1 comparing storm object sizes and updraft intensities (not shown). Finally, because the dynamical core requires large amounts of HDD filtering, the subsequent impacts cannot be appreciably reduced in the RRFS without making the integrations unstable. These limitations are a consequence of the design choice for the underlying dynamical core. *This greatly limits what developers can do to ameliorate the persistent biases seen in the RRFS's convective structures and precipitation fields.*

2.2. Efforts to Improve RRFS Performance.

Real-data experiments are those which stem from case studies or cycled, retrospective experiments. Often cases are chosen that represent forecast phenomena of particular importance to stakeholders. Here we first provide a summary table (Table 2.1) followed by a more descriptive list of what was tested, and the outcome.

Effort	Area	Impact	Promoted to RRFS? (Y/N)
Less diffusive horizontal advection	Dynamics	Led to storms that had faster motions, introducing phase errors into the forecast.	N
Positive definite tracer advection	Dynamics	Exacerbated the existing high precipitation bias.	N
8th order divergence damping	Dynamics	Modest improvement to storm structure and precipitation forecasts.	Y
Height dependent divergence damping	Dynamics	No significant impact on RRFS forecasts beyond idealized testing.	N
Tuning coefficient for the damping of "other variables"	Dynamics	Reducing this dimensionless parameter from 0.075 to 0.03 eliminated problems associated with frequent model failures.	Y

Tuning fundamental timestep	Dynamics	Shorter timesteps (i.e., dt_atmos=36s) led to overall improvements in forecast precipitation skill and bias.	Y
Linearized non-hydrostatic solver	Dynamics	Improved consistency between the w-field and motion of the Lagrangian surfaces at upper levels. Unaffected precipitation.	Y
Inner looping microphysics	Physics	Testing revealed that this increased model expense and did not improve precipitation bias.	N
Semi-Lagrangian sedimentation	Physics	This update did allow for a somewhat longer timestep, precipitation and forecast reflectivity scores were degraded.	N
Condensation timescale adjustment	Physics	Reduced precipitation and intensity bias, generated supersaturations aloft, degraded upper air and near surface forecast scores. Not well-justified on a scientific basis.	N
Limiter on microphysics temperature tendency	Physics	A fairly restrictive value between 0.02 K/s and 0.05 K/s would be needed to correct the bias. While individual storms may be less intense they will still remain too widespread in unstable, weakly forced regimes. Owing to the lack of scientific justification this approach was not pursued further.	N
Saturation adjustment delay	Physics	Testbed results suggest it introduces a dry bias.	N
Parameterized deep convection	Physics	Improved precipitation bias primarily via a reduction in the frequency of false alarms.	Y

Table 2.1: Summary of experiments tested to address the storm and precipitation intensity bias, the impact, and whether or not the experiment was promoted to the RRFSv1 package.

2.2.1. Dynamics

2.2.1.1. Impact of Less Diffusive Horizontal Advection

In response to feedback that individual convective-storms tended to be too smooth and large, we conducted a series of real-data and idealized testing comparing the current FV3 advection option in RRFS (so-called hord*=6) with a less diffusive option (hord*=5). The less diffusive option was considered as a way to improve convective storm

structure. Unfortunately the less diffusive option led to storms that had faster motions, introducing phase errors into the forecast. It also increased the intensity of convective storms, exacerbating the current bias. This option was not pursued further.

2.2.1.2. Positive Definite Tracer Advection

Upon recommendation from colleagues at GFDL, we tested the positive definite advection scheme described in Gao et al. (2021) which had been shown to improve tropical cyclone forecasts and may improve convective storm structures. This option was tested in RRFS and was shown to exacerbate the existing high precipitation bias in a series of retrospective experiments (Fig. 2.6). This was not pursued further.

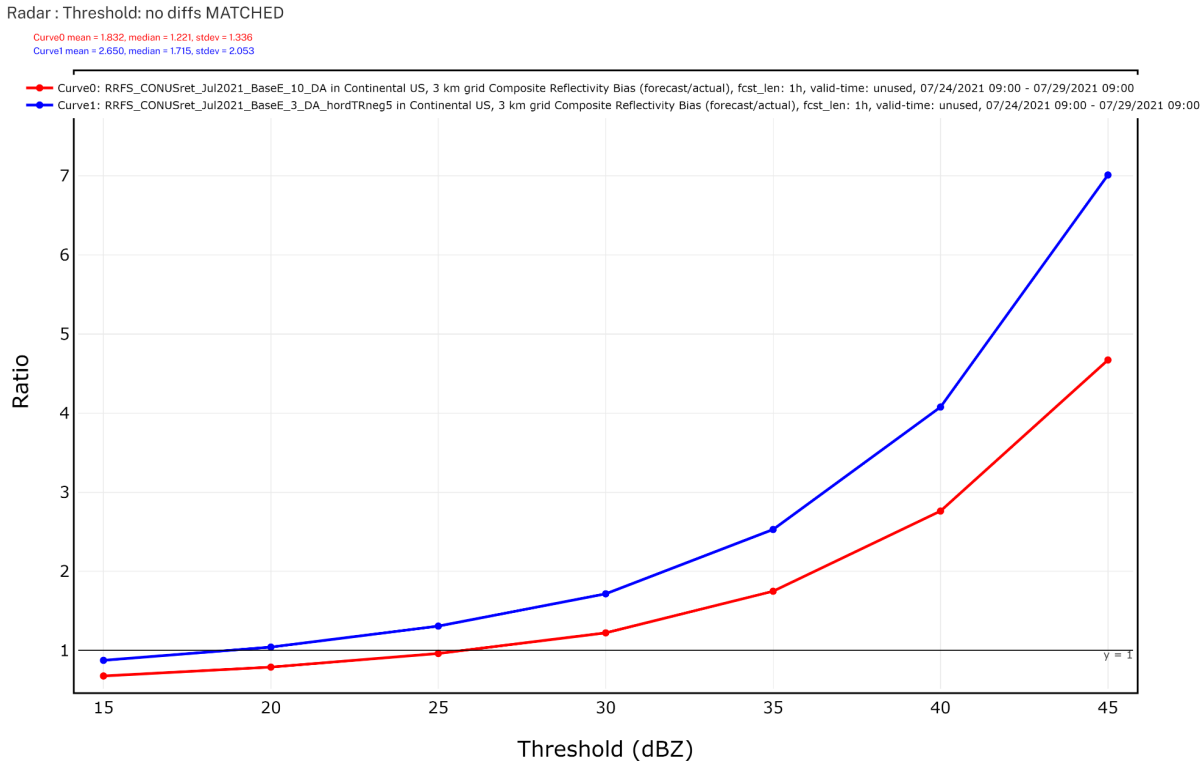

Fig. 2.6 1-hr forecast radar reflectivity frequency bias by threshold for the period of 24-28 July 2021 without positive definite advection (hord_tr=10; red line) and with (hord_tr=-5; blue line).

2.2.1.3. 8th Order Divergence Damping

In an effort to improve forecast storm structure and precipitation, it was suggested by GFDL to try a higher order divergence damping option. RRFS had been using 6th order (nord=2) and the recommendation was to try 8th order (nord=3). Enabling higher-order damping also required careful adjustment to the associated non-dimensional damping coefficient (d4_bg) to keep the model stable. Furthermore, an additional 2nd order divergence damping mechanism with a Smagorinsky-type flow-dependent coefficient was activated (dddmp). Subsequent tests were successful and the change to higher order damping had the intended effect. Forecast storms featured improved structure and precipitation forecasts improved as well (Fig 2.7). While

the improvement was encouraging, and this was promoted to the RRFS package, it was not sufficient to fully address the convective storm bias.

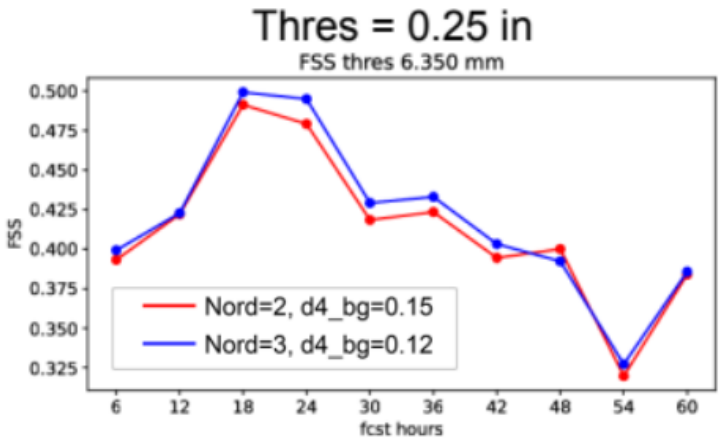

Fig. 2.7. Fractions Skill Score (FSS) with a neighborhood size of 5 grid boxes for the 0.25 inch threshold over 6-h accumulation periods. Statistics cover seven cases featuring a variety of severe and hazardous weather, including flooding, supercells, and tropical cyclone remnants. The 8th order option is in blue while the 6th order option is in red. Note the changes required to the damping coefficient required to keep the model stable are also annotated.

2.2.1.4. Height Dependent Divergence Damping

Modifications to divergence damping have led to improvements in the forecast of convective storms (see Fig. 2.7) with RRFS. Therefore work was done to see if further improvements could be obtained via adjustments to how divergence damping is applied through the column. In this approach a height dependence was added such that the value could be reduced in the lower troposphere and then gradually increased aloft. Hypothetically, reducing the damping in the troposphere would retain more desirable convective-storm characteristics. Idealized tests showed some promise, reducing the amount of storm total precipitation and weakening too-strong updrafts. However, follow-on tests with real-data revealed no significant impact on RRFS forecasts (Fig. 2.8). This option was not pursued further.

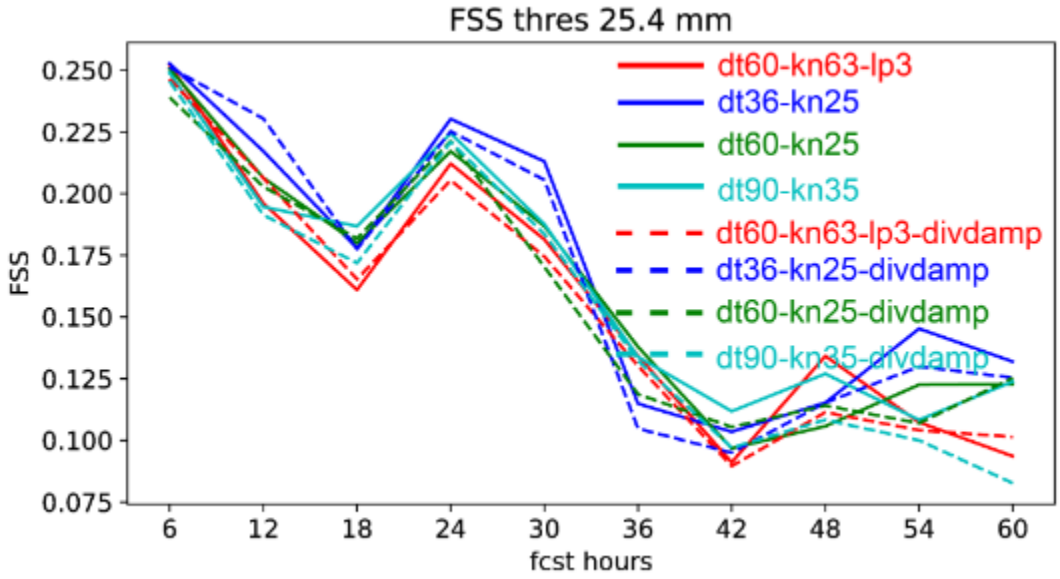

Fig. 2.8. Fractions Skill Score (FSS) with a neighborhood size of 45 km for a threshold of 25.4 mm (1 inch) over 6-h accumulation periods. Statistics are derived from 33 forecasts 03 May 2021 to 04 June 2021. While many experiments are shown here, the ones evaluating height dependent divergence damping are the blue lines. The dashed blue line corresponds to the experiment that includes the height dependent scheme while the solid blue line does not.

2.2.1.5. Coefficient for the Damping of "Other Variables"

The "other variable" damping process damps the fluxes of vorticity, air mass, and the non-hydrostatic vertical velocity and can be selected via a namelist switch (do_vort_damp). This damping mechanism is typically switched on in RRFS. The order of the "other variable" damping mechanism depends on the order of the divergence damping which is determined by the namelist parameter nord. However, the maximum order is capped at 6th order in case the (current RRFS default) 8th order divergence damping is selected. Analogous to the divergence damping, the "other variable" damping coefficient needs to obey stability constraints in a narrow window and can even become the trigger of instabilities. The damping coefficient (vtdm4) was therefore adjusted in an effort to address model instability issues. It was found that values that were too large led to a model that was less stable. Reducing this dimensionless parameter from 0.075 to 0.03 eliminated problems associated with frequent failures. Slight improvements were seen in precipitation forecast skill though the frequency bias did not exhibit statistically significant changes (not shown). Since this modification improved model stability it was promoted to the RRFS package.

2.2.1.6. Fundamental Timestep

RRFS developers have noticed a high degree of forecast sensitivity to model timestep and therefore tested a variety of timestepping options. It was found that shorter timesteps (i.e., dt_atmos) led to overall improvements in forecast precipitation skill and bias (Figs 2.9 and 2.10). While a shorter timestep is more computationally expensive - it

was less so than with using a longer timestep combined with inner looping of microphysics (see sec. 2.2.2.1).

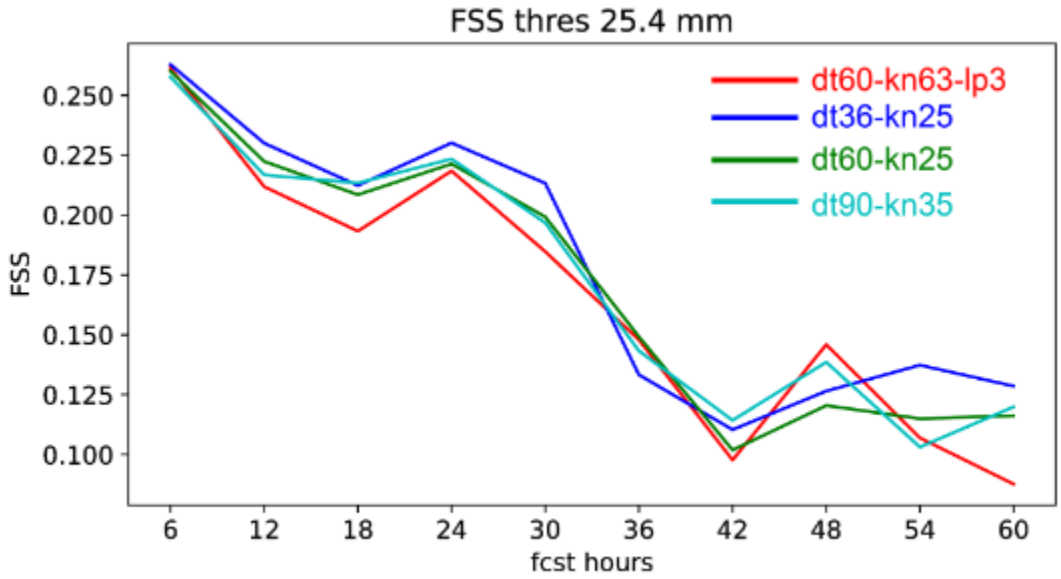

Fig 2.9. Fractions Skill Score (FSS) with a neighborhood size of 45 km for a threshold of 25.4 mm over 6-h accumulation periods. Statistics are derived from 33 forecasts 03 May 2021 to 04 June 2021. The experiments are as follows: microphysics inner looping at 20s interval with a fundamental timestep of 60 s (red, dt60-kn63-lp3), 36 s fundamental timestep (blue, dt36-kn25), 60 s fundamental timestep (green, dt60-kn25), 90 s fundamental timestep (cyan, dt90-kn35).

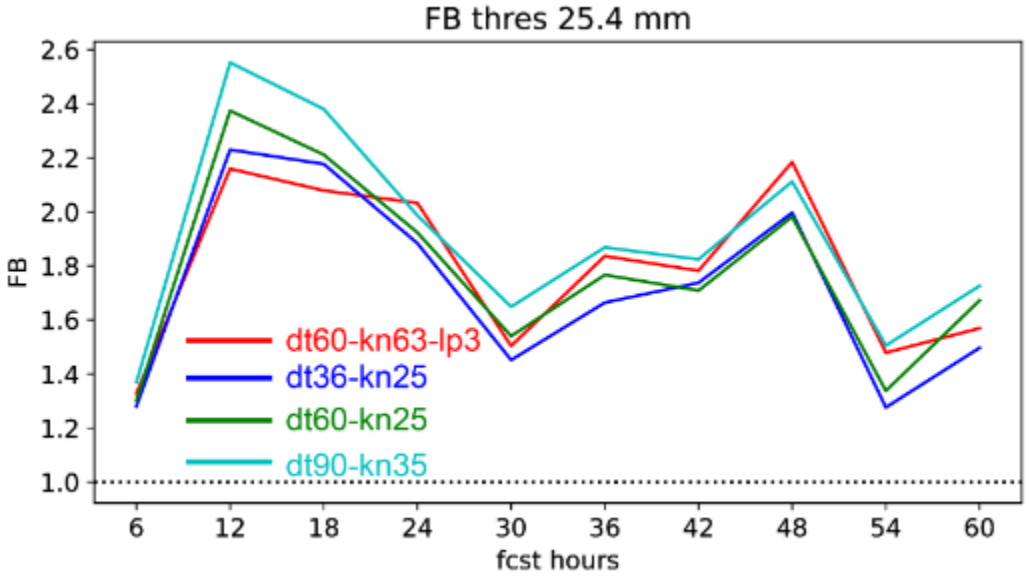

Fig. 2.10. As in Fig. 2.9 except frequency bias is shown.

2.2.1.7. Linearized Non-hydrostatic Solver

The non-hydrostatic semi-implicit solver contained a nonlinear term in the vertical equation of motion associated with the use of total pressure used in the explicit forcing. Due to the accumulation of vertical integration errors in the algorithm in the semi-implicit solver, these errors would manifest above developing deep convection as unphysical negative vertical motions in the stratosphere (the use of Rayleigh damping on the w-field in upper model layers generally masked the error). This error resulted in the w-field and motion of the Lagrangian surfaces to be inconsistent at upper levels. To eliminate this inconsistency it was suggested by GFDL to replace the full pressure with the hydrostatic pressure in the semi-implicit solver. This change effectively eliminated the error, resulting in much better consistency between the two fields. Importantly, none of these effects impacted precipitation scores. This change was promoted to the RRFS package due to its general neutral impact on other fields and its positive impact on the vertical motion field in RRFS.

2.2.2. Physics

2.2.2.1. Inner Looping Microphysics

It was hypothesized that calling the microphysics more frequently may help address the issue of too-intense convective storms. In this approach, the full microphysics routine is called 2 to 3 times more frequently than other, non-radiation schemes (i.e. 2 to 3 times more frequent than dt_atmos). Testing revealed that this increased model expense and did not improve precipitation bias. While inner-looping was present in the RRFS prototype system for several months, it was later removed when it was found that a slightly shorter fundamental timestep was both more efficient and improved the precipitation forecast (see Sec. 2.2.1.6).

2.2.2.2. Semi-Lagrangian Sedimentation

A revised hydrometeor sedimentation algorithm for Thompson microphysics, originally developed to enable a stable long timestep for the coarser global model, was tested in RRFS to assess potential impacts on precipitation. While this update did allow for a somewhat longer timestep, precipitation and forecast reflectivity scores were degraded. This option was not pursued further.

2.2.2.3. Parameterized Deep Convection

While somewhat non-traditional, the advent of scale-aware parameterized convection schemes have made re-introducing these schemes a possibility (e.g., Bengtsson et al. 2022; Grell and Freitas 2014; Han et al. 2017). As the precipitation and storm intensity bias became increasingly difficult to address, developers decided to test parameterized deep convection. Initial results indicate improvement in the precipitation bias, though it appears to largely do so by reducing the frequency of false alarms. While

certainly an improvement, when storms do form they still remain too intense and too large. Further, the introduction of a deep convection scheme has now also introduced a high bias in light precipitation. Owing to the general improvement by reducing false alarms, parameterized convection has since been added to the RRFS package. Further tuning and development is ongoing.

2.2.3. R2O Transition Plan Agreement: Improvements to Rapid Refresh Forecast System (RRFS) and FV3 Interoperability to Address Model Biases

In October of 2022 a Research to Operations Transition plan was developed between OAR/GFDL, OAR/GSL, and NWS/EMC to work toward improving the precipitation and storm intensity bias issue in the RRFS[10]. As the goal of the project was to correct some significant biases with convective-storms, a request was made to have GFDL work with GSL and EMC directly on the RRFS. GFDL expressed preference to only work in their own system, known as C-SHiELD, owing to additional burden on an already-busy staff associated with learning and running RRFS. GFDL would, of course, share results based upon their testing with C-SHiELD. While not ideal for RRFS development, it was felt that there would still be benefit to working with the dynamical core experts at GFDL even if all participants were not working on the same system. A compromise was reached and the transition plan was finalized. Below are three options that were discussed during the time period covered by the transition plan.

2.2.3.1. Condensation Timescale [Physics]

It was suggested that adding a timescale to condensation can help reduce the precipitation and storm intensity bias. The approach effectively limits the amount of total condensation produced by the microphysics, and therefore reduces the associated amount of heating and moistening. This method was added to the Thompson-Eidhammer microphysics scheme used in RRFS and several timescale values were examined. Testing revealed that large reductions in condensation (40-65%) were required to see an impact. The change did reduce precipitation and intensity bias. However it also led to supersaturations aloft and degraded upper air and near surface forecast scores. Further, the addition of the timescale was considered not well-justified on a scientific basis. Since degradation was present in upper air and near surface fields, this approach was not pursued further.

2.2.3.2. Limiter on Microphysics Temperature Tendency [Physics]

The use of a limiter on the microphysics temperature tendency was recommended. When activated, this places an upper limit on the maximum value of the temperature tendency from the microphysics. This limiter will reduce updraft intensity and therefore ameliorate the precipitation and storm intensity biases. However, the approach lacks scientific rigor. The limiter was originally designed to allow a large time step in early versions of the HRRR (using a 0.07 K/s limit) so it could fit within operational runtime requirements without violating vertical CFL, i.e. a safety valve. This is no longer used in operations as of HRRRv4 with the inclusion of an implicit-explicit

[10] https://drive.google.com/file/d/1_oBzPImJIhgKTtEQy7Pu6pR3h8X9-YK_/view?usp=sharing

vertical advection scheme. If used in RRFS, this tendency limiter would act as a bias corrector rather than as the originally intended safety-valve. Despite these concerns, the limiter was thoroughly tested. It was found that a fairly restrictive value between 0.02 K/s and 0.05 K/s would be needed to correct the bias. Further, while individual storms may be less intense they will still remain too widespread in unstable, weakly forced regimes. Owing to the lack of scientific justification, this approach was not pursued further.

2.2.3.3. Saturation Adjustment Delay [Physics]

GFDL noted a sensitivity to how saturation adjustment is done, which later led to the introduction of the Saturation Adjustment Delay technique in Spring of 2023 and was demonstrated to be a factor in reducing a high precipitation bias in GFDL's C-SHiELD system (Cheng et al. 2023). However, the overall C-SHiELD performance during the 2023 HWT SFE and 2023 HMT FFaIR experiments was not favorable for severe weather prediction (Sec 2.3.1 and Appendix B) or heavy precipitation forecasting (Fig. 2.11, lower CSI and dry bias), respectively. This ultimately led the RRFS team to decline pursuing this technique in favor of alternate strategies, such as parameterized convection (discussed later).

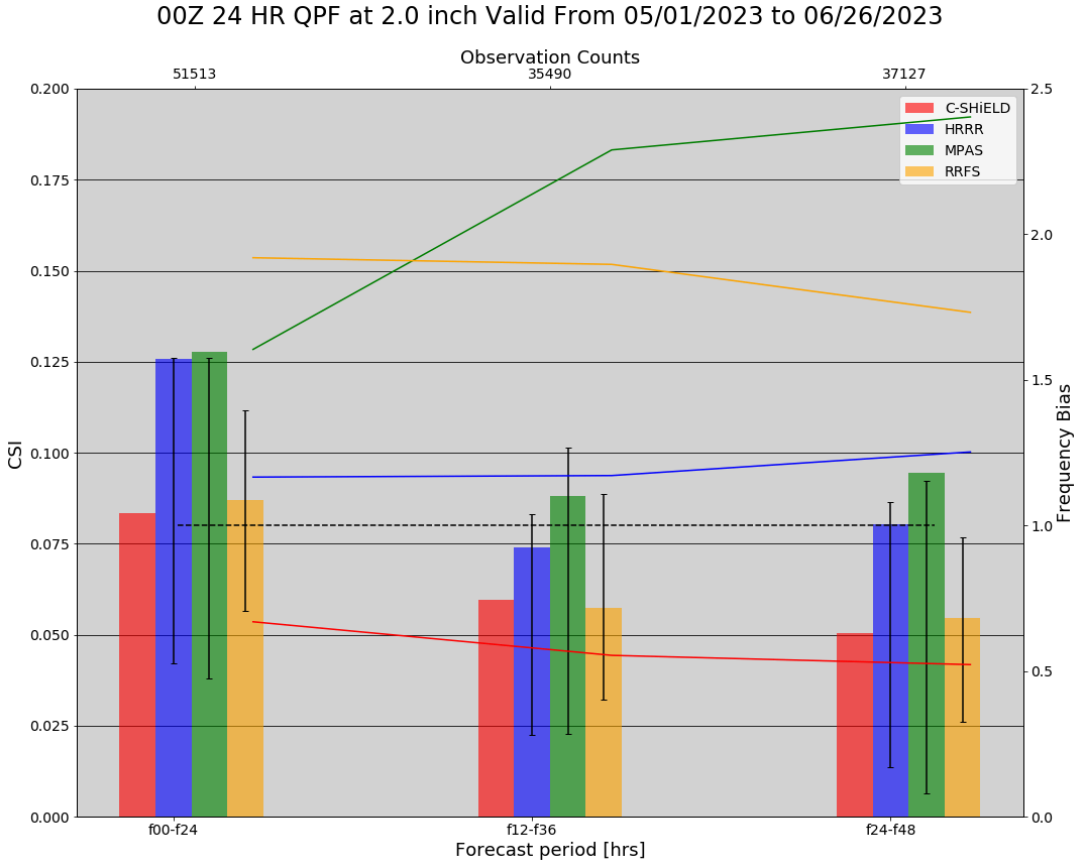

Fig 2.11 24hr QPF CSI (bars) and frequency bias (lines) for 12Z/00Z cycles of CSHiELD (red), HRRR (blue), MPAS (green), and RRFSP1 (orange). Statistics are shown for the 2 inch threshold. Verification is performed on a common 5 km grid and covers the period of 1 May 2023 through 26 June 2023.

2.3. Efforts to Evaluate the RRFS - NOAA Testbeds

There are a number of testbeds within NOAA that help to facilitate the transition of research to operations. Two of these NOAA testbeds in particular, the Hazardous Weather Testbed (HWT) and Hydrometeorological Testbed (HMT), play an active and important role in evaluating experimental CAM prototypes against current operational CAM systems. These testbeds bring in diverse stakeholders (i.e., operations, research, and academia) from their respective communities to offer an honest and unbiased perspective on the performance of CAM systems in a real-time, real-world setting. Collecting, assessing, and addressing feedback from these testbed experiments has proven to be an effective strategy for model developers to build into the annual development cycle of next-generation CAM systems. The subsections below discuss the evaluation results of FV3-based models at convection-allowing scales from HWT and HMT experiments.

2.3.1. HWT Spring Forecasting Experiments

Each spring, the NOAA/NWS/Storm Prediction Center and the NOAA/OAR/National Severe Storms Laboratory jointly conduct a Spring Forecasting Experiment (SFE) in the HWT with participation from severe-weather experts across the world spanning multiple sectors of the community. The HWT SFE is a five-week, real-time collaborative experiment to test emerging concepts and technologies designed to improve the prediction of hazardous convective weather. One of the primary goals of the HWT is to accelerate the transfer of research to operations by evaluating and documenting the performance of state-of-the-art experimental CAM systems. The HWT SFE has been evaluating the performance of the FV3 at convection-allowing scales since 2017.

During the first two years of FV3 evaluation in 2017 and 2018, global configurations with ~3-km CONUS nests were evaluated in the HWT SFE. These early FV3 CAM runs generally produced reasonable forecasts compared to operational CAMs. Despite these encouraging early results, all versions of FV3 still lagged clearly behind the operational CAM standard at the time (i.e., HRRRv3) for convective forecasting applications, largely owing to an over-forecast of coverage and intensity of convective storms (e.g., Fig. 2.12). These subjective ratings and impressions were confirmed with the extensive objective analysis and verification performed and documented by Gallo et al. (2021).

In subsequent years (2019-2021), a number of sensitivity tests were conducted with regional versions of FV3 CAMs, including exploring the impact of advanced physics packages, number of vertical levels, initial conditions, diffusivity settings, and land-surface model. Using an advanced physics suite (MYNN PBL and Thompson MP) instead of the standard physics suite (EDMF PBL and GFDL MP) tended to have the greatest positive impact on model performance for convective applications. Although the deterministic FV3 CAMs showed improvement relative to prior years, they still lagged behind the operational HRRRv4 in terms of storm-attribute and environment forecasts for convective applications. Participant comments often noted the tendency of FV3-based CAMs to have too much storm coverage with the storms having overly

intense circular updrafts. Ironically, these overly abundant and intense storms often occurred in an environment characterized by having a low bias in instability.

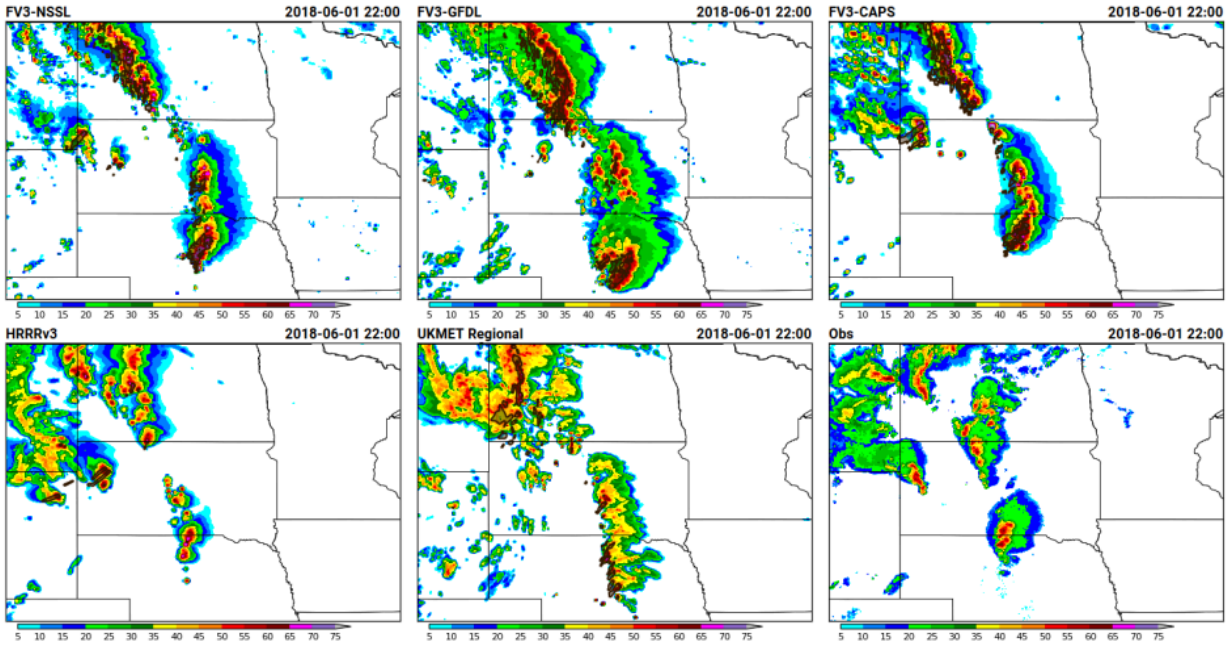

Figure 2.12. Example of subjective comparison plots used for rating CAM performance at convective scales valid at 2200 UTC on 1 June 2018. The 21-h forecasts of composite reflectivity are shown for the a) FV3 NSSL (upper-left panel), b) FV3-GFDL (upper-middle panel), c) FV3-CAPS (upper-right panel), d) HRRRv3 (lower-left panel), and e) UK Met Office UM. The bottom-right panel shows the observed composite reflectivity at 2100 UTC on 1 June 2018.

In 2022, the first end-to-end RRFS prototypes, including conventional (i.e., no radar) data assimilation, were evaluated during the HWT SFE. In blinded subjective evaluations, deterministic and ensemble RRFS prototypes showed skill approaching that of the operational standards (HRRR and HREF, respectively). This was a noteworthy result as this was the closest any deterministic (ensemble) CAM had been rated to the HRRR (HREF) in previous HWT SFEs. The results in 2023 were not as encouraging for the RRFS prototype, which included radar data assimilation for the first time. One contributing factor to these results was the weakly forced and weakly sheared pattern in 2023. This type of environment accentuates the issues noted in the FV3 core, as documented in Sections 2.1.1 and 2.1.2.

In the 2023 HWT SFE data assimilation evaluation, the HRRR had superior performance in short-term forecasts compared to the RRFS control member. The RRFS control run systematically produced overly intense convection in the one-hour forecast with an abundance of spurious storms (Fig. 2.13 middle panel), as noted in participant ratings and comments. This radar DA issue seemingly impacted the resultant RRFS forecast also, as the 0000 UTC RRFS runs were rated lower than the HRRR in blinded Day 1 forecast evaluations and verified worse in terms of objective verification statistics. Figure 2.14 depicts the performance of all flagship deterministic forecast systems from the 2023 HWT SFE and shows that the three MPAS configurations had skill most comparable to the operational HRRR. Among the three

FV3-based applications the RRFS was the best performing, though it markedly lagged behind the MPAS systems. The remaining two FV3-based systems, GFDL-CSHiELD and NASA GEOS, were the two poorest performing CAMs.

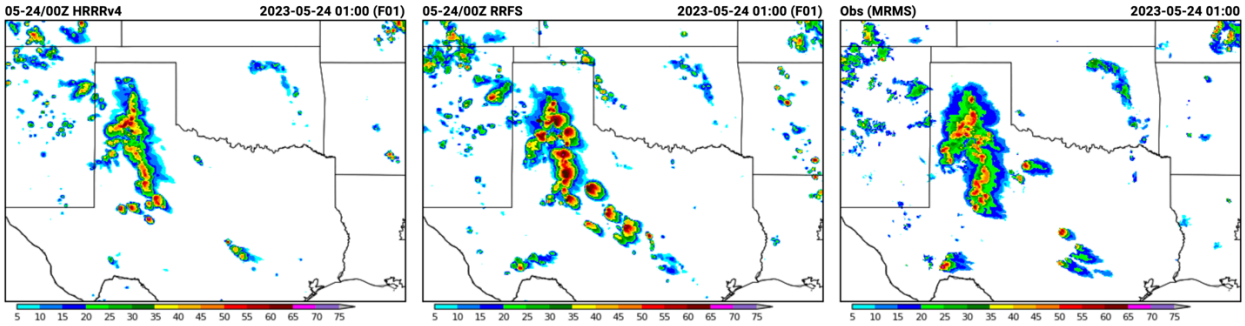

Figure 2.13. Example of multi-panel comparison webpage for the RRFS vs. HRRR DA evaluation. The top row displays simulated composite reflectivity from 2100 UTC initializations of HRRRv4 (left) and RRFS (middle) valid at 0100 UTC compared to MRMS observations (right). The bottom row displays the same as the top, except for 0000 UTC initializations.

Interestingly, the 1200 UTC-initialized RRFS control forecasts performed much better than the 0000-UTC initialized runs, indicating that the RRFS performance was strongly dependent on the initialization time during the 2023 HWT SFE, much more so than the operational HRRR. The RRFS ensemble evaluation followed the same pattern as the deterministic evaluation results, where the HREF had an overall edge over the RRFS in 0000-UTC initialized forecasts while the 1200-UTC initialized forecasts of RRFS and HREF were rated similarly. Objective statistics for composite reflectivity (>40 dBZ), however, still give a distinct edge to the 1200 UTC HREF over the single-physics RRFS. It is also worth noting that the Day 2 subjective evaluations strongly favored the HREF over any 1200 UTC RRFS configuration. Overall, the RRFS results from the 2023 HWT SFE are concerning given the limited time remaining for RRFS development before the expected code freeze.

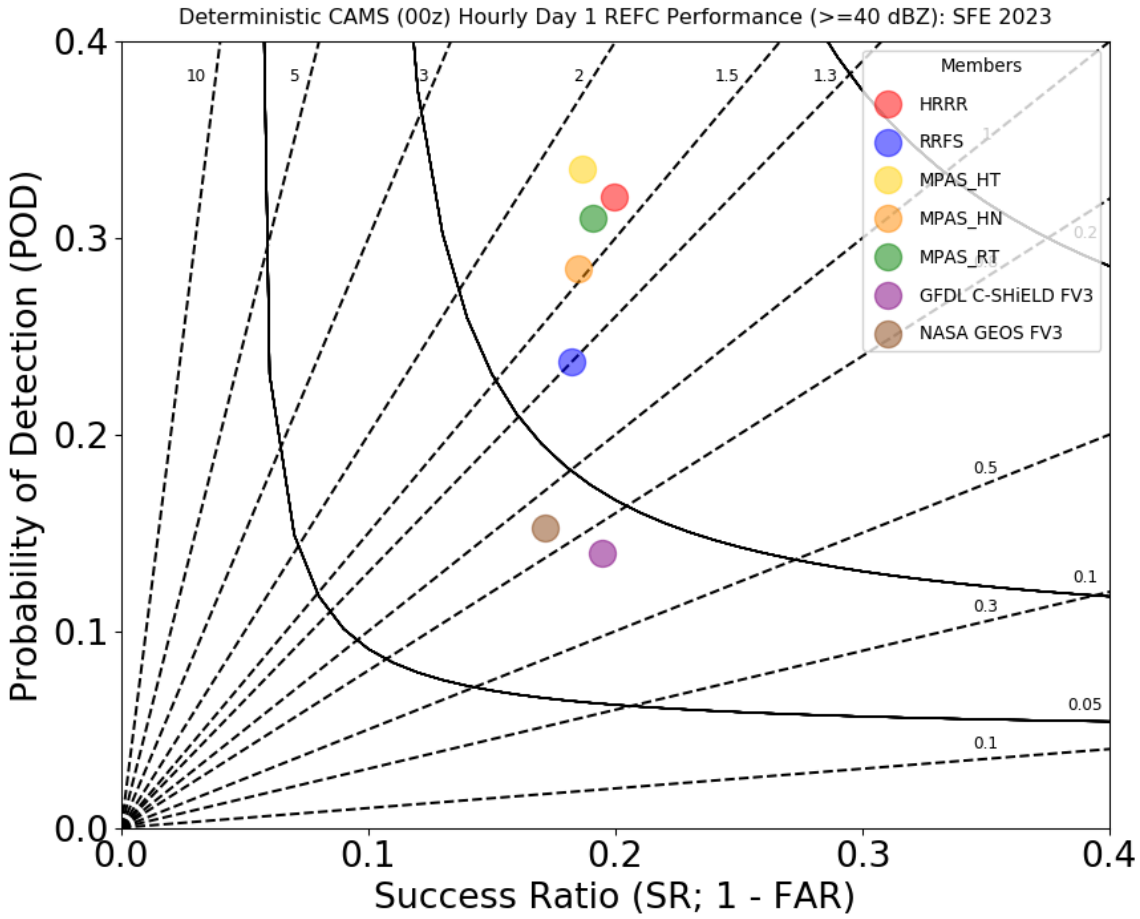

Figure 2.14. Performance diagram for hourly forecasts of simulated composite reflectivity ≥ 40 dBZ within 40-km neighborhood computed over SFE 2023 domains during the Day 1 forecast period (i.e., f12-36). The abbreviations for the MPAS configurations indicate the initialization dataset and microphysics scheme used: “HT” - HRRR and Thompson, “HN” - HRRR and NSSL, and “RT” - RRFS and Thompson.

2.3.2. HMT Flash Flood and Intense Rainfall Experiments

Each summer the NOAA/NWS/Weather Prediction Center and the NOAA/OAR/Physical Science Laboratory jointly conduct the Flash Flood and Intense Rainfall Experiment (FFaIR) in the HMT with participation from heavy rainfall experts across the country with occasional participation from other International operational units. The groups span many areas: from numerical modeling to operational forecasters to partners who use the information. The HMT FFaIR is a multi-week, quasi real-time collaborative experiment to test emerging concepts and technologies designed to improve the prediction of heavy rainfall. Like the HWT, the HMT evaluates and documents the performance of CAMs.

Beginning in 2018, FV3 CAM runs were being evaluated in the HMT FFaIR. These early FV3 CAM runs generally produced reasonable forecasts that were comparably good at times relative to operational CAMs, but displacement errors of precipitation events led to some concerns and forecaster distrust in the FV3-based models. In 2019, both the FV3 configurations from EMC performed well based on the

subjective scores given by the participants. Despite these high subjective ratings, neither of the FV3 runs were as impressive when evaluated using objective metrics. Most notable was the high wet bias seen in both models, which at the 1 inch threshold was higher than the wet bias seen in the NAM-Nest.

In 2020, it was repeatedly discussed by the participants, both verbally and in written comments, that there was a pronounced wet bias in the QPF from all the FV3 configurations. This wet bias was noted in both synoptically forced and mesoscale events but what was most concerning to the participants was the precipitation forecast for single cell convection, often referred to as “popcorn” convection (Figure 2.15). Each configuration appears to be creating a gridscale convection that “rains out” all the moisture available in the grid cell column. The FV3 runs were forecasting hourly rates exceeding 3 inches for nearly every single-cell storm in the model, which was much higher than in MRMS-based quantitative precipitation estimates.

In 2021, it was apparent based on both the subjective and objective evaluation of the FV3 CAM runs that the wet bias noted previously was still evident. Additionally, like the previous year, there continued to be an overdevelopment of “popcorn” convection in the FV3 runs. From investigating the hourly averaged maximum precipitation rate, the HRRR average maximum precipitation rate is consistent with the maximum average from MRMS, hovering around 8 in/hr. The FV3 CAM runs have a much higher average maximum precipitation rate that are generally consistent with one another, roughly staying between 20-30 in/hr throughout the forecast.

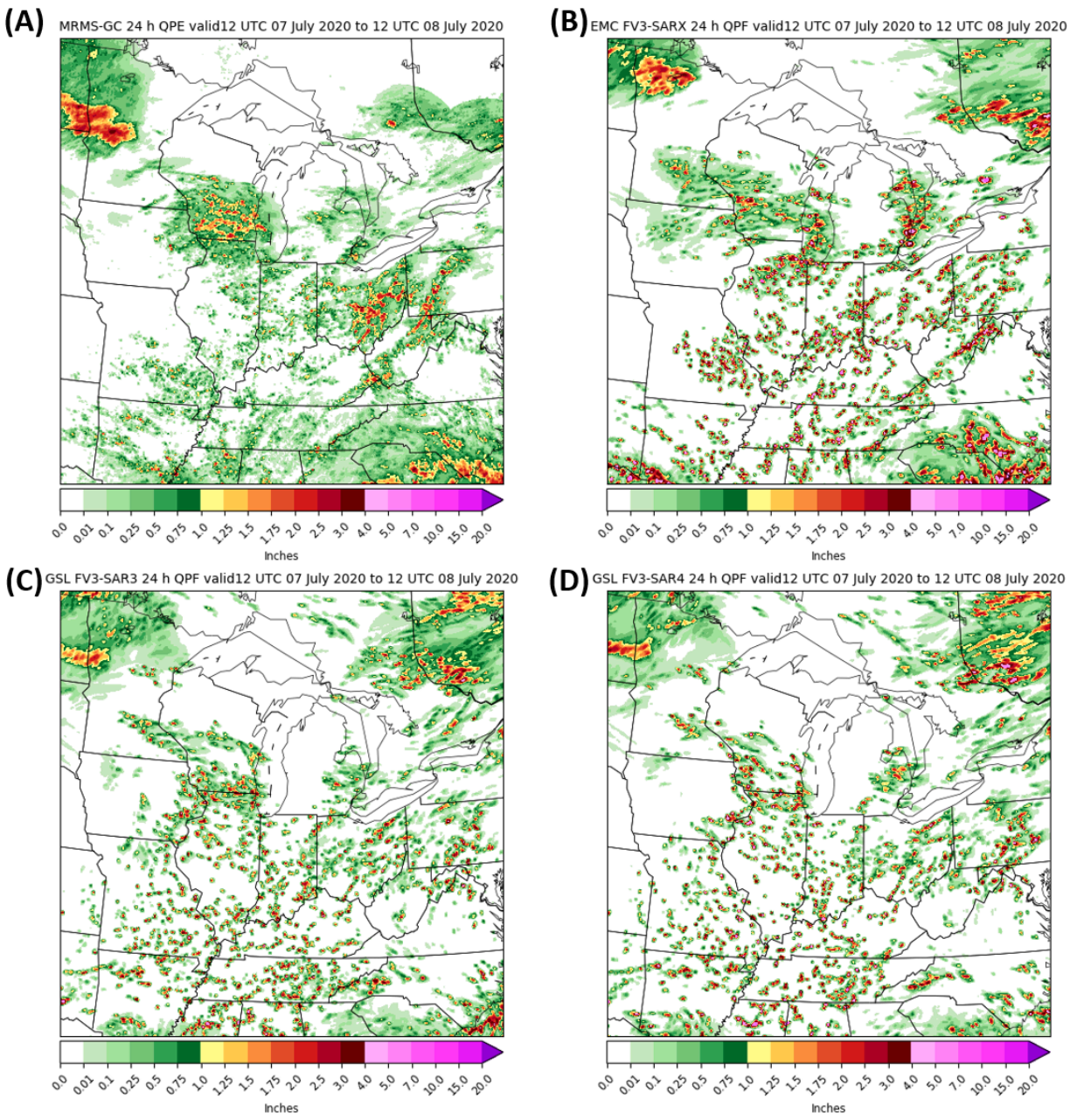

Figure 2.15. (A) 24 h MRMS-GC QPE and 24 h QPF from (B) EMC FV3-SARX, (C) GSL FV3-SAR3, and (D) GSL FV3-SAR4 valid 12 UTC 07 July to 12 UTC 08 July 2020.

For 2022, which featured a relatively inactive pattern for extreme rainfall relative to 2021, the FV3 CAM wet bias was still present in the FV3-based RRFS prototypes. This bias appeared to be most extreme over the southeast United States. Participants once more commented on the prolific simulation of “popcorn” storms in the RRFS models. When analyzing the hourly precipitation (Figure 2.16), the RRFS precipitation forecasts outpaced MRMS precipitation estimates in terms of coverage at 1.5” across the CONUS and 1.25” over the Southeast. When compared to the operational models

and MRMS precipitation rates, both instantaneous and hourly maximum precipitation rates, the RRFS deterministic and ensemble members had high-to-extreme rates. This included some simulated hourly maximum precipitation rates of over 100 in h^{-1}. Often these high rates were collocated with “popcorn” storms. When looking at rainfall diurnally, both RRFS prototypes had convective initiation too early (by roughly 2 hours) as compared to MRMS. Additionally, comparing coverage to intensity showed that although the average hourly rainfall coverage is more similar to MRMS for the RRFS runs than the operational models; when focusing on intensity, both models have higher averages than MRMS for all hours of the day. This suggests that the RRFS models are more likely to simulate rainfall than the operational models but when it does rain, the intensity is too high.

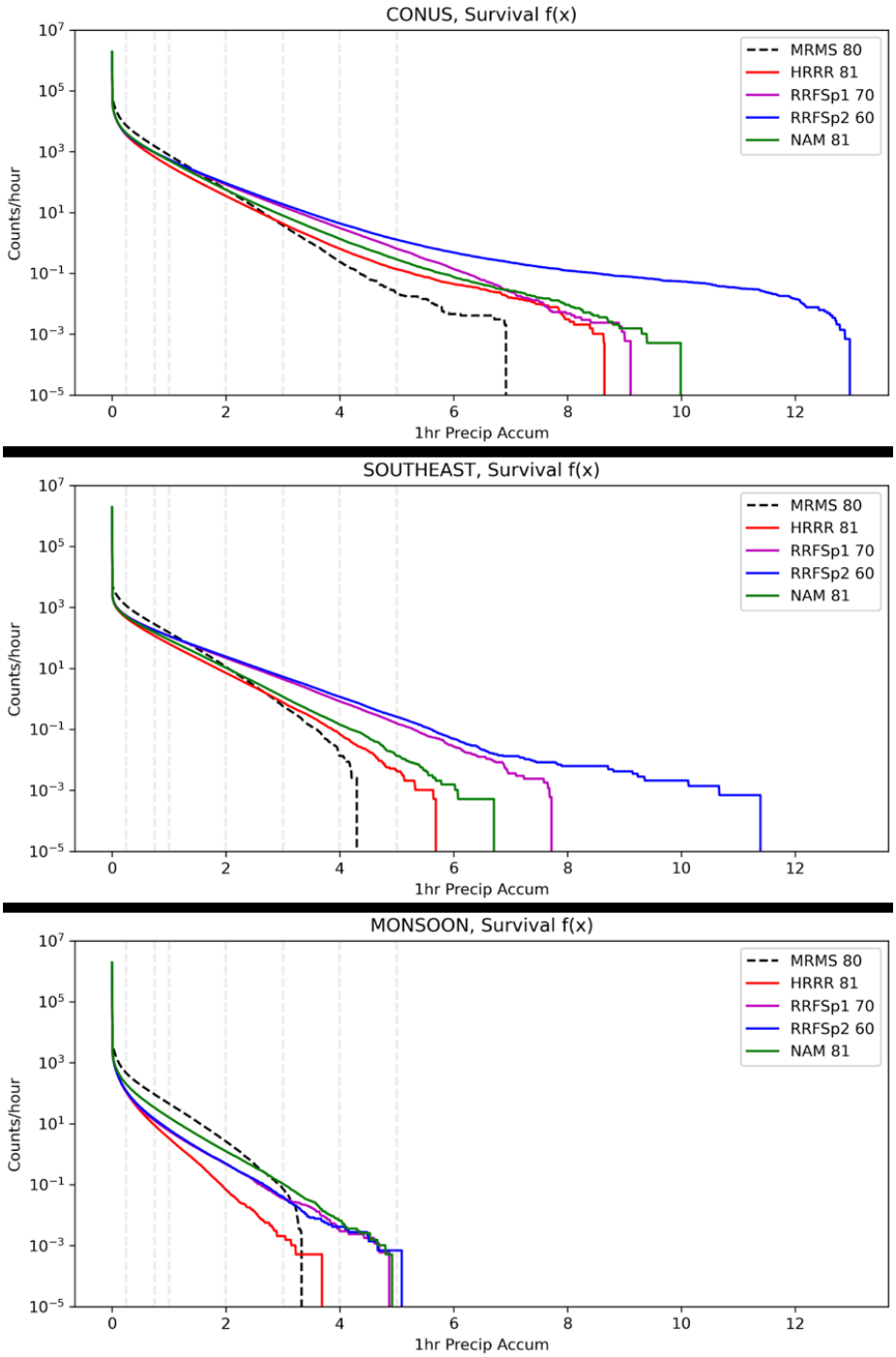

Figure 2.16: Hourly QPF survival function for the Testbed Season over the CONUS (top), the Southeast (middle), and the Monsoon (bottom; Southwest), comparing MRMS (dashed), HRRR (red), NAMnest (green), RRFSp1 (purple), and RRFSp2 (blue). On the y-axis is the counts per hour and on the x-axis is the 1 hr precipitation accumulation. The numbers in the figure legends indicate the number of cases for MRMS and the models.

In 2023, the RRFS transitioned to fully cycled data assimilation and split the FFaIR season between running over a CONUS (7 weeks) and a North American domain (3 weeks). To further show how the precipitation bias manifests in the RRFS,

the 1”+ coverage peaks about 1 and 2 hours prior to MRMS (00 and 12 UTC) respectively. This behavior is opposite to the HRRR and NAMnest performance. The survival function depiction (Figure 2.17) also shows opposite behavior to the operational models in that the 12 UTC initialization produces fewer extremes than its 00 UTC counterpart. In this depiction, the coverage bias up to the ~5.5” exceedance threshold is greater on average than the operational models. We can summarize that the RRFS has larger areas of heavier precipitation and are more intense and occur earlier (~22Z) in the diurnal cycle than operational CAMs.

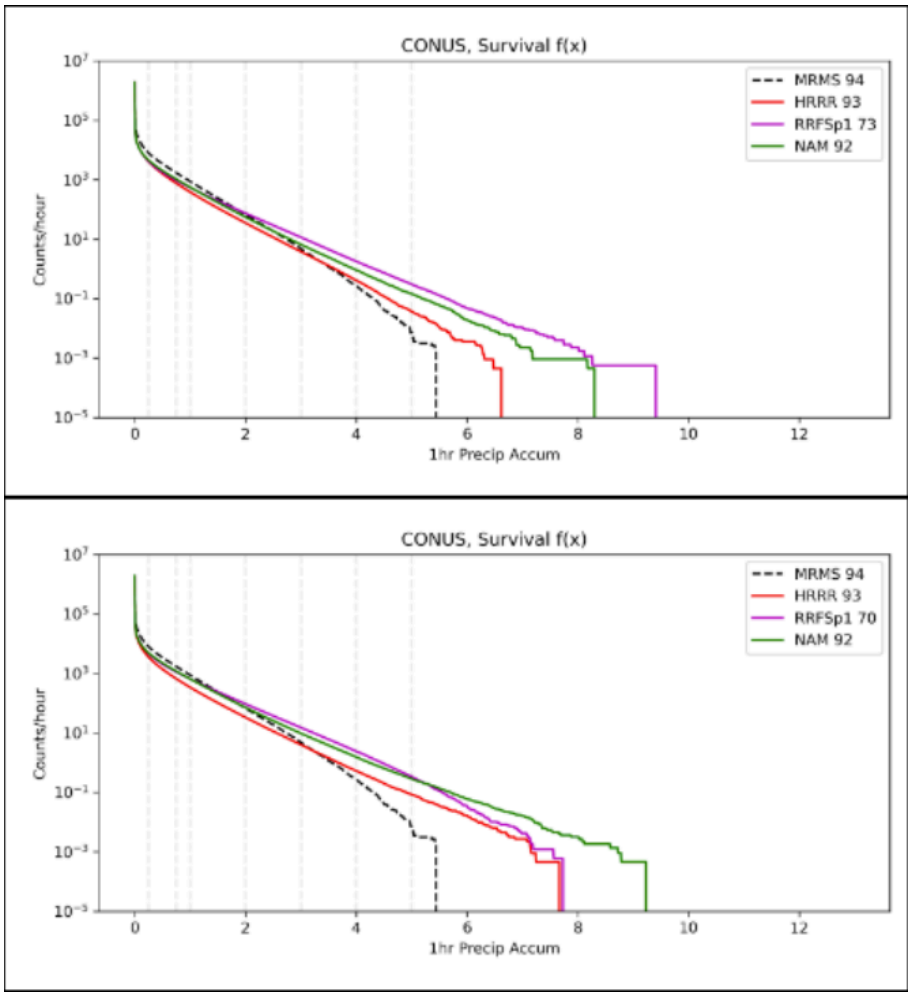

Fig 2.17: The survival function depiction of 1hr rainfall for 00z initialization (top) and 12z initialization (bottom) from the 2023 HMT-FFaIR.

3. Dynamical Core Design at the Convective-Scale

3.1. Overview of the Current FV3 Design

The numerical design of the FV3 dynamical core is complex. The non-hydrostatic design was largely informed by FV3’s performance at large hydrostatic scales which are dominated by rotational motions. These benefitted from the early design decisions (Lin and Rood 1997; Lin 2004), such as the choice of the horizontal C-D-grid discretization. The latter behaves, at least to a leading order, as a D-grid (Konor and Randall 2018a, b;

Skamarock 2008) and is well suited for rotational modes. The design of the non-hydrostatic FV3 extension on the cubed-sphere grid was later retrofitted into the existing hydrostatic dynamical core software framework, which led to more options for dissipation. This is reflected by, e.g., (a) the 12+ runtime choices for the numerical characteristics and accuracy of the horizontal advection and vertical remapping schemes, (b) a broad portfolio of dissipation mechanisms that include various variants of horizontal divergence damping, vorticity damping, sponge-layer diffusion processes, external mode damping, filters in the vertical direction, artificial limiters, and fixers that convert the dissipation effects into heat, and (c) user-selected subcycling decisions as certain processes, such as the advection of dynamic fields, the advection of tracers, or the vertical remapping step, are computed with tailored time steps. However, the optimal tailoring of time steps depends on the flow conditions. This makes the FV3 dynamical core susceptible to numerical instabilities and noise (Zhou and Juang 2023), especially as more extreme conditions are represented at convection-allowing grid scales. These instabilities were partly linked to FV3's floating Lagrangian coordinate in the vertical direction and the chosen remapping strategy. Related stability problems with vertical Lagrangian coordinates have also been found in other nonhydrostatic dynamical cores as discussed in Kavčič and Thuburn (2018a,b). These authors reached the conclusion that Lagrangian vertical coordinates are rather less robust than their height-based alternatives.

All these numerical design decisions are intertwined. For example, the choice of the horizontal C-D-grid discretization necessitates the use of horizontal divergence damping and other damping schemes to guarantee numerical stability. However, horizontal divergence damping is known to aggressively damp the desirable divergent and gravity wave motions at convective scales. In addition, divergence damping can become a source of instability itself as the choice of its coefficient needs to obey stability constraints in a narrow window (Whitehead et al. 2011). Furthermore, the simulated atmospheric flow is highly sensitive to the choice of the FV3 divergence damping coefficient. For example, changing the coefficient changes tropical cyclone forecasts in counterintuitive ways (Zhao et al. 2012). In addition, the representation of tropical storms is highly dependent on the user-chosen options for FV3's horizontal advection scheme that vary in their dissipation strength (Harris et al. 2020, Gao et al. 2021). Therefore, FV3's numerical characteristics, both related to its accuracy and stability, are complex to control, are flow- and scale-dependent, and can trigger unexpected flow responses. Recommendations for optimal FV3 configurations cannot be generalized as they rely on the whole portfolio of the numerous run-time options. As a generic rule of thumb, dissipation controls the appearance of numerical artifacts. Its overuse eliminates or distorts desirable physical signals, especially at convection-allowing scales. Its underuse allows the growth of instabilities.

A consequence of FV3's broad numerical design portfolio with its typical parameter settings is that FV3's effective resolution is generally lower than that of other non-hydrostatic dynamical cores. The lower effective resolution of FV3's default version is mostly determined by the explicit diffusion settings which are connected to the choice of the C-D grid. This leads to forecast degradations for small-scale phenomena, especially for divergent motions, updrafts, and precipitation. Reduced grid spacings

would be needed to match the flow characteristics of other models with higher effective resolutions. Examples of overdiffusive FV3-based configurations are shown in Lin et al. (2023). Such overdiffusive simulations are characterized by steeply declining kinetic energy spectra at small scales as demonstrated for a wide range of grid spacings down to 6 km.

Early prototypes of the non-hydrostatic FV3 configuration date back to 2006 and participated in the NGGPS comparisons in the early-to-mid 2010s. However, so far there are no peer-reviewed publications that survey and explain the non-hydrostatic FV3 design, and no inlined documentation is available in the codebase. The non-hydrostatic FV3 documentation mostly resides as a limited technical report (Harris et al. 2021) and has not been routinely updated. Therefore the description of the FV3 design remains incomplete.

In the longer term, a more rigorous understanding of the current RRFS dynamical core, or any additional core added to the UFS, is needed so that the development of new forecast systems includes a clear comprehension of the constraints associated with filtering and damping required for stability and accuracy. This would include detailed documentation of the model equations and the discretization, as well as more extensive testing across the parameter space for the various damping/filtering schemes to understand where the various limitations and/or instabilities might occur.

3.2. Options for Moving Forward

3.2.1. Scope of Effort Required to Improve the FV3 Dynamical Core for Convection

It is important to realize that the changes needed for the FV3 dynamical core, as described below, are significant enough that the resulting code would effectively **create a second dynamical core within the UFS**. This new, modified FV3 dynamical core would require its own code base, as the needed changes could not be integrated into the current FV3 dynamical code base to be "turned on or off" during compilation. Subsequently, the physics and data assimilation will have to be retuned and reworked for the new dynamical core. New interfaces will be needed for pre- and post-processing as well as for data assimilation to accommodate the model's new grid structure.

Our experiences have shown that undertaking any major code rewrite is tricky (e.g., Appendix C), and especially difficult without the original developers' support for the project and, as is the case with the FV3 dynamical core, without detailed documentation of the current model equations to work from. It is estimated that development and testing will take ~5 years (project estimates follow), with no guarantee of success.

3.2.1.1. Summary of Major Code Changes Needed Leading to a New FV3 Dynamical Core Variant (In Order of Importance)

- Conversion of D-grid for horizontal velocity to C-grid
- Removal of the Lagrangian vertical coordinate

- Incorporation of physics tendencies using a higher-order time scheme
- Conversion from a pressure to height coordinate

3.2.1.2. Conversion of D-grid for Velocity to C-grid

(EFFORT=significant, RISK=high)

Why: The calculation of pressure gradient and divergence on the D-grid is less accurate than on a C-grid (Mesinger and Arakawa, 1976). Despite FV3 using accurate 4th order differencing for these terms, with D-grid staggering the 2 Δx horizontal mode is stationary. In order to remove this mode, FV3 uses high-order horizontal divergent damping. Unfortunately, the amount of damping needed for good solutions has been found to significantly alter the updraft characteristics of the model, leading to other biases with the size and shape of storms and a very high bias in precipitation in certain convective environments (Sec. 2.1). As far as we can determine, the only solution for this issue is to change the grid staggering to better handle these important three-dimensional divergent motions.

What: Changing the D-grid staggering to a C-grid would require invasive and extensive changes to the current model code. Essentially, one would need to create a new solver. E.g., every U, V and Vorticity array will have to be redeclared, e.g., going from U(NX,NY+1) → U(NX+1,NY), V(NX+1,NY) → V(NX,NY+1), and VORT(NX,NY) → VORT(NX+1,NY+1), etc. While this looks like a small change, every place in the code where variables from multiple locations are averaged or differentiated will require new code to reflect the change in grid staggering. A large percentage of loop indices will have to be changed. It's unclear whether the predictor-corrector algorithm would need to change, as well as other details of the integration algorithm. Documentation of the current model equations are non-existent at the level of detail needed to scope out the project. Any developer will need to have or develop a great deal of knowledge about the existing code. The C-grid will also require reworking the interfaces between faces of the cubed sphere. A completely different set of velocity variables will be located on the cubed-sphere interfaces.

Finally, the use of a C-grid will reduce the dynamics time step[11] by a factor of 2 for the new model to be stable for the gravity and sound waves (Skamarock 2008). This issue may be of minor importance, because in order to remain stable, the current RRFS is using a time step that is 40% smaller than the stability limit (for explanation see 3.2.1.2 below). Most likely, the current time steps needed for stability will be roughly the maximum time step for the new dynamical core dynamics to be stable. This time step will be similar to other current C-grid models (e.g., WRF, MPAS).

[11] The dynamics time step in FV3 is dt_atmos / (nsplit*ksplit). So while dt_atmos could remain close to the same (this is the physics time step), the amount of work needed will increase (not by a lot as it turns out, because we are already using a much reduced dynamic time step due to issues with the Lagrangian coordinate - see Sec. 3.2.1.2).

3.2.1.3. Removal of the Lagrangian Vertical Coordinate

(EFFORT=moderate/significant, RISK=high)

Why: Runs of the RRFS show that in strong convection, the Lagrangian surfaces can intersect (i.e., "collapse") even during the dynamics time step before a vertical remapping occurs. This causes the model to crash. This issue was diagnosed by EMC's Kevin Viner during the past year. There is logic in the code which prevents the surfaces from intersecting by setting a minimum layer thickness, but it appears there are unanticipated effects when this minimum layer thickness logic is used during the next dynamics time step on the velocity field, which often results in a crash. This is why current RRFSv1 runs use a time step which is 40% smaller than the anticipated stability limit, to avoid the collapse of the Lagrangian layers. The Lagrangian vertical coordinate is very useful when (U,V) >> W (e.g., when $\Delta x > 10$ km). In these regimes, the dynamics are very different and the horizontal motions are effectively only weakly coupled to the vertical. This permits the model to run in pancake-like layers nearly independently. For CAM applications, where (U,V) ~ W (within mid-latitude convective updrafts), the larger vertical divergences within updrafts collapse the Lagrangian surfaces together much more quickly. To prevent this, one either has to reduce the dynamics time step (as in RRFSv1) or needs to call the vertical remapping step much more frequently to prevent the collapse. Either way, efficiency is reduced. While the Lagrangian vertical coordinate is not necessarily the primary cause of the excessive storm intensity and precipitation bias, it does need to be removed to improve convective-scale applications and thus is within scope of this effort.

What: Removing the vertical Lagrangian coordinate will then require a vertical advection scheme to be implemented and a new semi-implicit scheme for the vertical propagation of sound and gravity waves. The speed of the new model will be different, but given that the current Lagrangian collapse problems have already reduced the current UFS FV3 core's efficiency, it is unclear whether replacing the Lagrangian surfaces with an Eulerian vertical advection scheme will significantly change the efficiency

3.2.1.4. Incorporation of Physics Tendencies using a Higher-order Temporal Integration

(EFFORT=low/moderate, RISK=low)

Why: This problem was first discovered by EMC's Kevin Viner (Viner 2023). Almost all physics tendencies in other NWP models (e.g., IFS, UKMET, WRF, COAMPS, etc.) are added into the state variables during the subsequent time step incrementally (WRF) or use a second-order time integration on the physics tendencies (IFS, UKMET). The UFS with the FV3 dynamical core simply adds these tendencies instantaneously and reflects a first-order physics-dynamics coupling approach. This approach can create abrupt changes in the state fields and the vertical coordinate (the pressure field is also impacted) which leads to imbalances and an increased sensitivity to the physics time step. EMC's Viner has experimentally implemented a second-order

method in the code. The second-order methodology appears to help decrease the sensitivity of the model to the chosen time step. By upgrading the use of physics increments to second-order, the imbalances should be reduced and the solutions improved.

What: Find a way to rewrite the physics interface such that the computed increments can be added to the state fields using a 2nd-order time integrator. This will require a major rethinking about how the physics is called and stored, and how to add tendencies more accurately.

3.2.1.5. Conversion to a Height Coordinate System (From Pressure)

(EFFORT=significant, RISK=med/high)

Why: As model resolutions increase, a significant numerical issue arises in the presence of very steep terrain. In these cases, the transformed coordinate has a significant horizontal component across very thin layers associated with the vertical grid. This imposes a more strict timestep size than for shallower terrain. A number of researchers have suggested using a cut-cell coordinate (Adcroft 1997; Steppler 2002; Good et al. 2012) which is for the z-coordinate system in non-hydrostatic models. Conversely, Shaw and Heller (2016) found no advantage in either terrain-following or cut-cells in simplified model tests. Moving forward, the most flexible approach would be to adopt the z-coordinate. Also, as the UFS moves forward toward the end of the decade to include fire-prediction (either implicitly or explicitly), the use of a thermodynamic variable as the vertical coordinate where the near ground temperatures can exceed 600-900 °C would appear to be very problematic. The pressure surface would be responding to the fire temperatures and would then be changing continuously and chaotically. This would impact all the transport and dynamics, and could require a very small time step due to the vertical coordinate. Therefore we believe a z-coordinate is required for these types of problems.

What: The FV3 dynamical core does have a geometric z-coordinate available in the hydrostatic version. It is unknown whether it was ever successfully tested in the non-hydrostatic version. It is important to consider that the other changes listed above will have an impact on the code development, as the dependent variables would need to change when using the z-coordinate. It is unclear whether moving to a z-coordinate would complicate or simplify the other changes listed above. No matter what, almost any detailed fire weather application (where $\Delta x < 100$ m) will need to use a geometric, not thermodynamic, vertical coordinate.

3.2.1.6. Project Estimates for Developing a New Variant of the FV3 Dynamical Core

We have divided efforts to address the outstanding deficiencies in FV3 into 5 tasks. The first 4 focus on development within the dynamical core. The 5th task encompasses the time and effort required to test and validate the collection of changes

in the RRFS system. Cost estimates are based upon past efforts where large portions of a dynamical core were overhauled (Appendix C).

Task	Time (years)	FTE	Total Cost ($200k per FTE does not include HPC)
1. Conversion of D-grid for horizontal velocity to C-grid	3	2 - dynamics experts 3 - testing and validation	$3 M
2. Removal of the Lagrangian vertical coordinate	3	2 - dynamics experts 3 - testing and validation	$3 M
3. Incorporation of physics tendencies using a higher-order time scheme.	2	1 - dynamics expert 2 - testing and validation	$1.2 M
4. Conversion from a pressure to height coordinate.	3	2 - dynamics experts 3 - testing and validation	$3 M
5. Integration and testing for RRFSv2	2	10	$4 M

Table 3.1. Scope and cost for primary activities to address FV3 deficiencies.

We anticipate a period of about 5 years and estimate $14.2M will be required to accomplish this work, with a timeline noted below in Fig. 3.1. We emphasize that this work is only scoped to meet the needs for the RRFS. Any other application may require additional scoping and resources. Approaches to resource allocation and staffing are at the discretion of the performing organization(s).

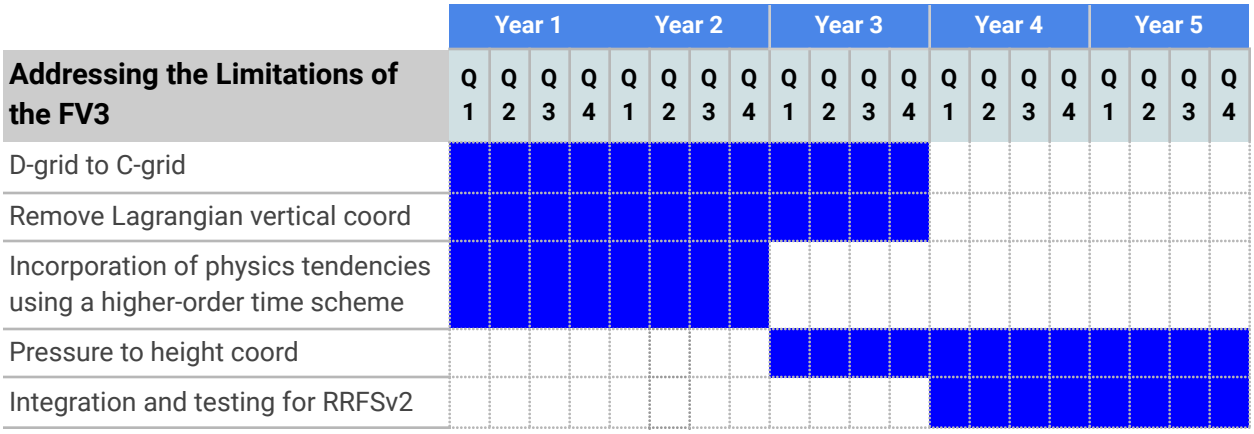

Fig. 3.1. Project timeline to address FV3 deficiencies. See Table 3.1 for costs.

Table 3.2 contains a list of anticipated, high-level risks should a project be undertaken to address gaps in the FV3 dynamical core for convective-scales.

Risk	Level (Low, Medium, High)	Mitigation
Insufficient expertise in FV3 dynamics. There are few in the community with sufficient expertise in the model dynamics such that the same people will likely be responsible for all work, this means that tasks that could otherwise be done in parallel may need to be done serially.	High	Extend timeline, train new staff (added cost)
Insufficient documentation of model dynamical core and associated code	High	Extend timeline, train new staff to write documentation (added cost)
May not sufficiently address convective issues	High	Switch to an alternate dynamical core following completion of efforts (many years)
Fall further behind science as time spent fixing current model is time not spent advancing science	High	Switch to an alternate dynamical core

Table 3.2. Anticipated risk and mitigation profile for primary activities to address FV3 deficiencies.

3.2.2. Move to a New Dynamical Core: MPAS

3.2.2.1. Overview of MPAS

The atmospheric component of MPAS (mpas-dev.github.io, Skamarock et al., 2012) has been developed by NCAR's Mesoscale & Microscale Meteorology Laboratory (MMM) during the last decade. MPAS contains a set of WRF-based physical parameterizations and a non-hydrostatic finite-volume dynamical core on a staggered C-grid with a height-based orography-following vertical coordinate. The MPAS dynamical core has been developed for a hexagonal, unstructured spherical centroidal Voronoi mesh (Thuburn et al., 2009; Ringler et al., 2010, Skamarock and Gassmann, 2011) and can either utilize a global mesh, a variable-resolution option (Park et al., 2014) or a limited-area configuration (Skamarock et al., 2018). MPAS is designed for convection-allowing regional and global applications. It is intended to replace the limited-area model WRF in future years as NCAR/MMM has discontinued further development of the WRF model as of summer of 2023.

So far, MPAS has been primarily used for research purposes and experimental numerical weather prediction applications. It also has become an optional non-hydrostatic dynamical core in NCAR's Community Atmosphere Model for a potential use in subseasonal to seasonal (S2S) or global high-resolution storm-resolving applications. The inclusion into NCAR's Community Atmosphere Model allows MPAS to be coupled to an interactive ocean that is available in the Community Atmosphere Model via the Community Earth System Model (CESM) framework. Experimental versions with variable-resolution grids down to 3 km and 12 km mesh spacings have also been tested for precipitation assessments during the wet season over the western United States (Huang et al., 2022) and for climate applications over the CONUS domain (Sakaguchi et al., 2023), respectively. In general, these studies concluded that the MPAS-based CESM climate model configurations are able to reproduce the regional climate characteristics in the refined regions. In particular, Huang et al. (2022) showed that the spatial patterns and mean intensity of the precipitation over the western US closely matches observations during the wet winter season. In addition to the variable-resolution experiments, MPAS has also been used across a wide range of uniform resolutions down to convection-allowing grid spacings of 3.75 km for the assessments of tropical rainfall, tropical waves as well as tropical cyclones (e.g., Judt and Rios-Berrios, 2021). Furthermore, MPAS has been employed for medium-range convection-allowing (3 km) ensemble forecasts with a variable-resolution global configuration (Schwartz 2019) as well as a recent study on the low predictability of a Derecho event (Ribeiro et al. 2022).

The MPAS model developments are mature, but undergo continued modernizations at NCAR/MMM, such as the extension to a deep-atmosphere configuration (Skamarock et al., 2021; Klemp and Skamarock, 2021) or the exploration of alternative model lids (Klemp and Skamarock, 2022). At the current time, MPAS is the core model for IBM's Global High-Resolution Atmospheric Forecasting System with enabled GPU capabilities, is planned to be used for operational NWP in Taiwan, has

been selected for operational NWP and research applications in Brazil (Skamarock 2023, personal communication), and has been selected by the Air Force for regional convective-scale NWP. In addition, the Environmental Protection Agency has recently announced the adoption of MPAS[12]. Further, various MPAS-based forecast model prototypes developed by NSSL were tested at convection-allowing resolutions during NOAA's HWT in 2023. NSSL is currently transitioning their WRF-based Warn on Forecast system (WoFS) to an MPAS-based WoFS to be tested in parallel during the spring of 2024 in the HWT SFE.

During the NGGPS evaluation process in 2016, MPAS was found to require twice as much computational resources as the FV3 system (Ji, 2016). Over the past seven years, the computational efficiency of the MPAS code has been improved, and a capability to sub-cycle the dynamical equations relative to the scalar transport has been added. The latter further increases the efficiency. MPAS can now use time steps which are ~40-50% larger than the WRF-ARW model. It is important to note that the current RRFS, in operational testing, remains stable only when the time step is 25-40% smaller than the time step suggested possible from the 2016 NGGPS tests (Viner 2023). Therefore the overall difference between model performance is much smaller than in 2016.

Below are timings from a recent benchmark of the FV3, WRF-ARW, and regional MPAS models from 48 hour 3 km CONUS forecasts using NOAA's RDHPCS Jet supercomputer. The case is run for 24 March 2023 which had a large outbreak of severe weather in the southeast U.S. The IO costs are included in the timings. Wallclock times are normalized to reflect xJet computational nodes using 1200 cores. The timings have **not been** normalized to the number of total grid points, which differ slightly. Regional MPAS's computational performance is now similar to the RRFS and WRF-ARW models over the CONUS. This close equivalency between models aligns with NSSL's experience after running three MPAS simulations each day during the past 10 months.

Model	**Wall Clock** (hh:mm)	**Grid points** (nx*ny, nz)	**Macro (Physics) Time Step** (NSSL 2-Moment microphysics)
FV3	2:46	1761280 x 60	36 sec (ksplit/nsplit = 5,2)
WRF	3:13	1584000 x 60	15 sec
MPAS	2:56	1894063 x 59	25 sec (dynamics substep = 3)

[12] https://www.epa.gov/cmaq/next-generation-air-quality-model

3.2.2.2. Project Estimates for Transitioning to the MPAS Dynamical Core

MPAS presently has many capabilities that both it and FV3 lacked in the 2017 and 2018 time frame. While considerable time and expense was spent in the FV3 transition to introduce the necessary limited-area (Black et al. 2021) and convective-scale capabilities, a similar degree of effort is not anticipated for MPAS for the following reasons:

- MPAS has an established Limited-Area Model capability (Skamarock et al. 2018) which eliminates the significant effort that had been required to develop such a capability for FV3, which was not completed until mid 2019.
- MPAS has basic pre- and post- processing capabilities in place that will allow developers to get started immediately, while more operationally viable components are developed in parallel within the UFS architecture.
- The RAP/HRRR physics suite is available for immediate use, although fully integrating CCPP would need to be scoped.
- A robust effort has been underway for some time to connect MPAS to JEDI (Liu et al. 2022; Jung et al. 2023; Guerrette et al. 2023); JEDI is the target for RRFSv2 regardless of the atmospheric dynamical core.
- MPAS has already demonstrated promising performance for RRFS convective applications (e.g., Fig. 2.11, as well as 2023 HWT SFE and Fig. 2.14).

The scope and effort required to adopt and support MPAS in the UFS was recently studied in a report by a Tiger Team under the UFS System Architecture and Infrastructure Cross Cutting Team (Wang et al. 2023). All but the final task, and level of effort estimates, were taken directly from this report and placed into a hypothetical timeline to support a RRFSv2 implementation. The final task encompasses all scientific testing and development work, which includes testing the system across all relevant seasons and weather phenomena (e.g. severe convective as well as winter hazards).

Task	**Time (years)**	**FTE**	**Total Cost ($200k per FTE does not include HPC)**
Code Mgmt and Governance at NOAA	3.5	0.5	$0.35 M
Generalizing ATM: FV3ATM → UFSATM[13]	0.5	2	$0.2 M

[13] Generalization and refactoring of the atmospheric component of the UFS Weather Model.

Build system and software stack	0.6	1	$0.12 M
Post-processing/IO	1.5	3	$0.9 M
NUOPC Cap[14] (inter-component coupling)	1	1	$0.2 M
Connection to CCPP (physics dynamics coupling)	2	1	$0.4 M
Workflow	4	1	$0.8 M
Integration and testing for RRFSv2	3	10	$6 M

Table 3.3. Scope and cost for primary activities to transition RRFS to MPAS. All numbers, with the exception of the final row, were derived from the *Integration of MPAS Dycore into UFS* whitepaper by Wang et al. (2023).

We anticipate a period of about 3.5 years and estimate $8.97M (about 65% of the cost needed to redesign the FV3 dynamical core) will be required to accomplish this work, with a timeline noted below in Fig. 3.2. We emphasize that this work is only scoped to meet the needs for the RRFS. Any other application may require additional scoping and resources. Approaches to resource allocation and staffing are at the discretion of the performing organization(s).

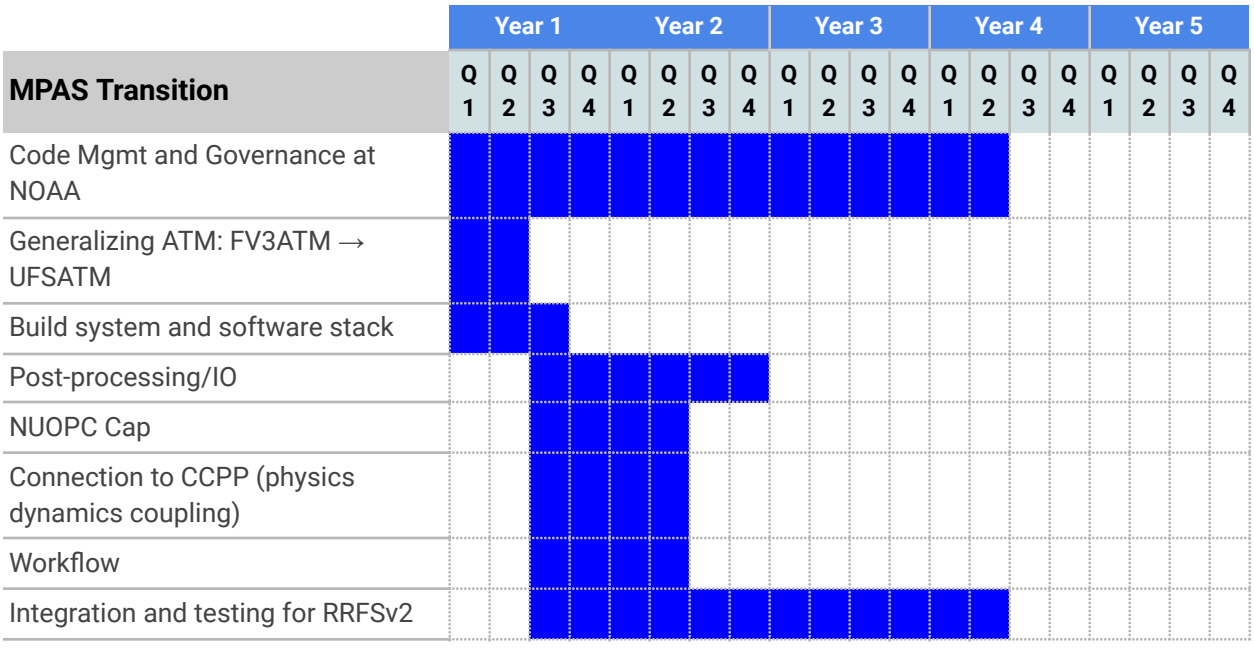

Fig. 3.2. Project timeline to transition RRFS to MPAS. See Table 3.3 for costs.

[14] https://earthsystemmodeling.org/nuopc/

Table 3.4, noted below, contains a list of anticipated, high-level risks should a project be undertaken to transition to MPAS for the RRFS dynamical core.

Risk	Level (Low, Medium, High)	Mitigation
Staff and resource constraints associated with having multiple atmospheric dynamical cores in UFS	High	Adjust resources, timeline and/or scope.
Model is too expensive to run	Medium	Engage HPC scientists to assess what can be accelerated. Consider adjusting scope of RRFSv2+ upgrade components to moderate computational burden. Benchmarking (Appendix D) shows performance on par with WRF-ARW (i.e., dynamical core in HRRR).
Underestimate of time and expense	Medium	Prioritize core capabilities for RRFSv2, consider rescoping and/or timeline modifications if needed.
Loss of MPAS expertise at NCAR due to retirements in 2-5 years	Medium	Work with NCAR to develop expertise within NOAA and develop an agreement for rigorous MPAS documentation and support from MMM
Performance may not meet or exceed operational baseline performance for all weather categories (e.g., winter).	Medium	Monitor. Early prototype indications are favorable for warm season convection.

Table 3.4. Anticipated risk and mitigation profile for primary activities associated with transitioning to MPAS.

4. Recommendations for the RRFS

4.1. RRFSv1

Neglecting the convective-storm problem for a moment, RRFSv1 does include a number of benefits, some of which include:

- A first-of-its-kind, high resolution 3-km domain covering North America.
- Improved service equity by expanding rapidly updated data assimilation and forecasts to Alaska, Hawaii, and Puerto Rico.
- High resolution smoke and dust.
- Cool season synoptic performance superior to HRRR (Fig. 4.1)

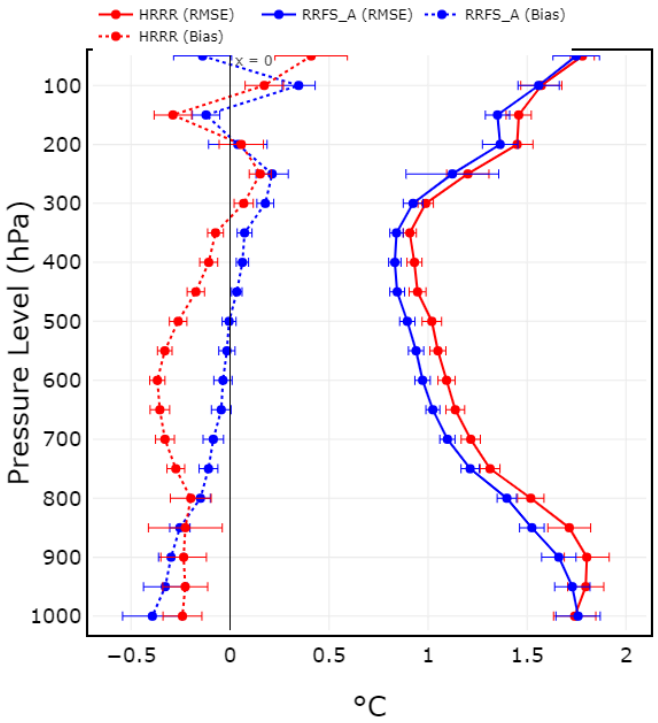

Fig. 4.1. Averaged 24-h temperature forecast RMSE (solid) and bias (dotted) for RRFS_A (blue) and HRRR (red) for December, January, and February of 2022-2023 winter season.

However the convective-storm biases and errors are too extreme to ignore. As has been demonstrated, there is little room for improvement in the FV3 dynamical core absent an expensive multi-year effort to overhaul the dynamics. The ability to forecast realistic convective hazards is a critically important attribute for the NWS's flagship high resolution operational NWP system. The current limitations with the model will impact the NWS's ability to meet its mission and Impact-Based Decision Support Services (IDSS) needs associated with convective hazards if, upon RRFSv1's implementation, all operational regional high-resolution NWP systems are retired. We therefore recommend an approach that minimizes stakeholder exposure to risk associated with convective-storm biases present in RRFS by keeping the HRRR in the production suite when RRFS is implemented. Keeping the HRRR also allows us to add some skillful

diversity to the RRFS ensemble through the inclusion of time-lagged HRRR members[15], and brings ensemble performance near HREF (Fig. 4.2). This allows for the retirement of the operational NAM nests (CONUS, Alaska, Hawaii, Puerto Rico, and Fire Weather) along with the operational HiRes Window FV3, ARW, and ARW2 domains (CONUS, Alaska, Hawaii, Puerto Rico)[16]. This allows for HREF to be retired in favor of the RRFS ensemble when augmented with HRRR membership. To summarize:

- Implement RRFSv1, subject to meeting project requirements (summarized in Appendix A).
- Keep HRRRv4 in operations and include it as a member in the RRFS ensemble.
- Retire the remaining regional high resolution modeling suite, which includes the NAM nests, HiRes Window systems, and the HREF.

Such a compromise ensures all stakeholders keep access to the best performing operational high-resolution NWP system, the HRRR, while allowing stakeholders to still get the benefits offered by the RRFS. Further, such an approach still allows us to retire much of the complex regional suite of models. And finally, keeping the HRRR in operations allows for its inclusion as an additional time-lagged member in the RRFS ensemble, which will add much needed skillful ensemble spread to approach the skill of the HREF.

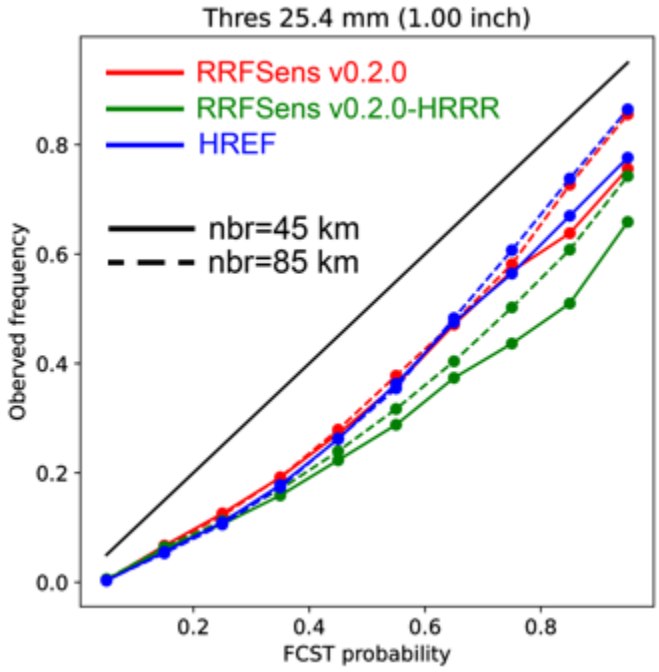

Fig. 4.2. Reliability diagram comparing the RRFS ensemble *with* the HRRR (red) to *without* (green) inclusion of the HRRR in its ensemble. The HREF (blue) is the operational baseline. Statistics are calculated for precipitation at the 1 inch threshold using 6 hour accumulation intervals and are averaged in the averaged 6-36 hour forecast period every 6 hours. The forecast period encompasses May 10th at 1200 UTC to May 25th at 1200 UTC, 31 total cycles are verified.

[15] The HRRR is currently also a member within HREF.

[16] Combining NAM nest and HiRes Window applications yields 17 distinct domains/systems that will still be retired with this proposal. Including HREF makes this 18.

4.2. RRFSv2

The extensive experience demonstrated by the RRFS developers and stakeholders outlined in this document, make it clear that a transition away from the FV3 dynamical core is a necessity for the future success of NOAA flagship high-resolution NWP system. Unless a multi-year overhaul of the FV3 is undertaken, with no guarantee of success, we have likely reached the upper limit on the performance offered by FV3 at convective-scales. **A transition to a new dynamical core is necessary for long-term, sustained innovation for NOAA's flagship high-resolution NWP system. Based upon the subject matter expertise within this document, we recommend that the dynamical core for the second version of RRFS (RRFSv2) be MPAS,** which has demonstrated potential to overcome the convective-issues identified with the FV3 dynamical core (Fig. 2.11). Finally, we emphasize that the discussion and recommendations within this document pertain only to the RRFS application.

Acknowledgements

The authors thank the reviewers and the full RRFS development team for their constructive comments on earlier drafts of this work. We also thank our respective leadership for their support throughout this process, including DaNa Carlis, Brian Gross, Jennifer Mahoney, David Novak, and Russell Schneider. We also thank Pam Heinselman for her unwavering support.

References

Adcroft A., C. Hill, J. Marshall,, 1997: Representation of topography by shaved cells in a height coordinate ocean model. *Monthly Weather Review*, **125**, 2293–2315. https://doi.org/10.1175/1520-0493(1997)125%3C2293:ROTBSC%3E2.0.CO;2

Bengtsson, L., L. Gerard, J. Han, M. Gehne, W. Li, and J. Dias, 2022: A Prognostic-Stochastic and Scale-Adaptive Cumulus Convection Closure for Improved Tropical Variability and Convective Gray-Zone Representation in NOAA's Unified Forecast System (UFS). *Monthly Weather Review*, **150**, 3211-3227, https://doi.org/https://doi.org/10.1175/MWR-D-22-0114.1.

Black, T. L., J. A. Abeles, B. T. Blake, D. Jovic, E. Rogers, X. Zhang, E. A. Aligo, L. C. Dawson, Y. Lin, E. Strobach, P. C. Shafran, and J. R. Carley, 2021: A Limited Area Modeling Capability for the Finite-Volume Cubed-Sphere (FV3) Dynamical Core and Comparison With a Global Two-Way Nest. *Journal of Advances in Modeling Earth Systems*, **13**, e2021MS002483, https://doi.org/10.1029/2021ms002483.

Bryan, G. H., and J. M. Fritsch, 2002: A benchmark simulation for moist nonhydrostatic numerical models. *Monthly Weather Review*, **130**, 2917–2928, https://doi.org/10.1175/1520-0493(2002)130%3C2917:ABSFMN%3E2.0.CO;2

Cheng, K.-Y., L. M. Harris, L. Zhou, A. Kaltenbaugh, and M. Morin, 2023: Development of a Global-nested Severe Weather Forecast Model for CONUS (C-SHiELD). *28th Conf. on Numerical Weather Prediction*. Madison, WI, Amer. Meteor. Soc., 8.4, https://ams.confex.com/ams/WAFNWPMS/meetingapp.cgi/Paper/424943.

Clark, A. J., and Coauthors, 2018: The Community Leveraged Unified Ensemble (CLUE) in the 2016 NOAA/Hazardous Weather Testbed Spring Forecasting Experiment. *Bulletin of the American Meteorological Society*, **99**, 1433–1448, https://doi.org/10.1175/BAMS-D-16-0309.1.

Flora, M. L., P. S. Skinner, C. K. Potvin, A. E. Reinhart, T. A. Jones, N. Yussouf, and K. H. Knopfmeier, 2019: Object-Based Verification of Short-Term, Storm-Scale Probabilistic Mesocyclone Guidance from an Experimental Warn-on-Forecast System. *Weather and Forecasting*, **34**, 1721-1739, https://doi.org/https://doi.org/10.1175/WAF-D-19-0094.1.

Gallo, B. T., and Coauthors, 2021: Exploring convection-allowing model evaluation strategies for severe local storms using the finite-volume cubed-sphere (FV3) model core. *Weather and Forecasting*, **36**, 3–19, https://doi.org/10.1175/WAF-D-20-0090.1.

Gao, K., L. Harris, L. Zhou, M. Bender, and M. Morin, 2021: On the Sensitivity of Hurricane Intensity and Structure to Horizontal Tracer Advection Schemes in FV3. *Journal of the Atmospheric Sciences*, **78**, 3007-3021, https://doi.org/10.1175/Jas-D-20-0331.1.

Grell, G. A., and S. R. Freitas, 2014: A scale and aerosol aware stochastic convective parameterization for weather and air quality modeling. *Atmospheric Chemistry and Physics*, **14**, 5233-5250, https://doi.org/10.5194/acp-14-5233-2014.

Good B., A. Gadian, S.J. Lock, A. Ross, 2014: Performance of the cut-cell method of representing orography in idealized simulations. *Atmos Sci Lett* **15**, 44–49. https://journals.ametsoc.org/view/journals/mwre/144/6/mwr-d-15-0226.1.xml

Guerrette, J. J., Z. Liu, C. Snyder, B. J. Jung, C. S. Schwartz, J. Ban, S. Vahl, Y. Wu, I. H. Banos, Y. G. Yu, S. Ha, Y. Tremolet, T. Auligne, C. Gas, B. Menetrier, A. Shlyaeva, M. Miesch, S. Herbener, E. Liu, D. Holdaway, and B. T. Johnson, 2023: Data assimilation for

the Model for Prediction Across Scales -- Atmosphere with the Joint Effort for Data assimilation Integration (JEDI-MPAS 2.0.0-beta): ensemble of 3D ensemble-variational (En-3DEnVar) assimilations. *Geoscientific Model Development Discuss.*, 2023, 1-34, https://doi.org/10.5194/gmd-2023-54.

Han, J., W. Wang, Y. C. Kwon, S.-Y. Hong, V. Tallapragada, and F. Yang, 2017: Updates in the NCEP GFS Cumulus Convection Schemes with Scale and Aerosol Awareness. *Weather and Forecasting*, **32**, 2005-2017, https://doi.org/https://doi.org/10.1175/WAF-D-17-0046.1.

Harris, L., X. Chen, W. Putman, L. Zhou, and J.-H. Chen, 2021: A Scientific Description of the GFDL Finite-Volume Cubed-Sphere Dynamical Core. NOAA Technical Memorandum OAR GFDL, 2021, 109 pp., https://doi.org/10.25923/6nhs-5897.

Harris, L. M., and S.-J. Lin, 2013: A Two-Way Nested Global-Regional Dynamical Core on the Cubed-Sphere Grid. *Monthly Weather Review*, **141**, 283-306, https://doi.org/10.1175/MWR-D-11-00201.1.

Harris, L. M., and S.-J. Lin, 2014: Global-to-Regional Nested Grid Climate Simulations in the GFDL High Resolution Atmospheric Model. *Journal of Climate*, **27**, 4890-4910, https://doi.org/10.1175/Jcli-D-13-00596.1.

Harris, L., Zhou, L., Lin, S.-J., Chen, J.-H., Chen, X., Gao, K., et al., 2020: GFDL SHiELD: A unified system for weather-to-seasonal prediction. *Journal of Advances in Modeling Earth Systems*, 12, e2020MS002223. https://doi.org/10.1029/2020MS002223

Huang, X., A. Gettelman, W. C. Skamarock, P. H. Lauritzen, M. Curry, A. Herrington, J. T. Truesdale, and M. Duda, 2022: Advancing precipitation prediction using a new-generation storm-resolving model framework – SIMA-MPAS (V1.0): a case study over the western United States, Geoscientific Model Development, 15, 8135–8151, https://doi.org/10.5194/gmd-15-8135-2022

Janjić, Z. I., and R. Gall, 2012: Scientific documentation of the NCEP nonhydrostatic multiscale model on the B grid (NMMB). Part 1: Dynamics. *NCAR Tech. Note*, **74**, https://doi.org/https://doi.org/10.5065/D6WH2MZX.

Ji, M. 2016: Dynamical Core Evaluation Test Report for NOAA's Next Generation Global Prediction System (NGGPS). NGGPS Dycore Phase 2 Test Report website.pdf (weather.gov).

Judt, F., and R. Rios-Berrios, 2021: Resolved convection improves the representation of equatorial waves and tropical rainfall variability in a global nonhydrostatic model. *Geophysical Research Letters*, **48**, e2021GL093265, https://doi.org/10.1029/2021GL093265.

Jung, B. J., B. Ménétrier, C. Snyder, Z. Liu, J. J. Guerrette, J. Ban, I. H. Baños, Y. G. Yu, and W. C. Skamarock, 2023: Three-dimensional variational assimilation with a multivariate background error covariance for the Model for Prediction Across Scales–Atmosphere with the Joint Effort for data Assimilation Integration (JEDI-MPAS 2.0.0-beta). *Geoscientific Model Development Discuss.*, 2023, 1-29, https://doi.org/10.5194/gmd-2023-131.

Kavčič, I., and J. A. Thuburn, 2018a: Lagrangian vertical coordinate version of the ENDGame dynamical core. Part I: Formulation, remapping strategies, and robustness. *Q. J. R. Meteorol. Soc*, **144**, 1649–1666, https://doi.org/10.1002/qj.3368

Kavčič, I., and J. A. Thuburn, 2018b: Lagrangian vertical coordinate version of the ENDGame dynamical core. Part II: Evaluation of Lagrangian conservation properties. *Q. J. R. Meteorol. Soc.*, **144**, 2620–2633, https://doi.org/10.1002/qj.3375

Klemp, J. B., and W. C. Skamarock, 2021: Adapting the MPAS dynamical core for applications extending into the thermosphere. *Journal of Advances in Modeling Earth Systems*, **13**, e2021MS002499. https://doi.org/10.1029/2021MS002499.

Klemp, J. B., and W. C. Skamarock, 2022: A Constant Pressure Upper Boundary Formulation for Models Employing Height-Based Vertical Coordinates. *Monthly Weather Review*, **150**, 2175–2186, https://doi.org/10.1175/MWR-D-21-0328.1.

Konor, C. S., and D. A. Randall, 2018a: Impacts of the horizontal and vertical grids on the numerical solutions of the dynamical equations – Part 1: Nonhydrostatic inertia–gravity modes. *Geoscientific Model Development*, **11**, 1753-1784, https://doi.org/10.5194/gmd-11-1753-2018.

Konor, C. S., and D. A. Randall, 2018b: Impacts of the horizontal and vertical grids on the numerical solutions of the dynamical equations – Part 2: Quasi-geostrophic Rossby modes. *Geoscientific Model Development*, **11**, 1785-1797, https://doi.org/10.5194/gmd-11-1785-2018.

Lin, P., Y. Ming, and T. Robinson, 2023: On the resolution sensitivity of equatorial precipitation in a GFDL global atmospheric model. *Journal of Advances in Modeling Earth Systems*, 15, e2022MS003300, https://doi.org/10.1029/2022MS003300

Lin, S.-J., 2004: A "Vertically Lagrangian" Finite-Volume Dynamical Core for Global Models. *Monthly Weather Review*, **132**, 2293-2307, https://doi.org/10.1175/1520-0493(2004)132<2293:AVLFDC>2.0.CO;2.

Lin, S.-J., and R. B. Rood, 1997: An explicit flux-form semi-lagrangian shallow-water model on the sphere. *Q. J. R. Meteorol. Soc.*, **123**, 2477-2498, https://doi.org/10.1002/qj.49712354416.

Liu, Z., C. Snyder, J. J. Guerrette, B. J. Jung, J. Ban, S. Vahl, Y. Wu, Y. Trémolet, T. Auligné, B. Ménétrier, A. Shlyaeva, S. Herbener, E. Liu, D. Holdaway, and B. T. Johnson, 2022: Data assimilation for the Model for Prediction Across Scales – Atmosphere with the Joint Effort for Data assimilation Integration (JEDI-MPAS 1.0.0): EnVar implementation and evaluation. *Geoscientific Model Development*, **15**, 7859-7878, https://doi.org/10.5194/gmd-15-7859-2022.

Mesinger, F., and A. Arakawa, 1976: Numerical Methods Used in Atmospheric Models. Vol. I. GARP Publications Series No. 17, WMO, Geneva, x + 64 pp.

Park, S.-H., J. B. Klemp, and W. C. Skamarock, 2014: A Comparison of Mesh Refinement in the Global MPAS-A and WRF Models Using an Idealized Normal-Mode Baroclinic Wave Simulation. *Monthly Weather Review*, **142**, 3614–3634, https://doi.org/10.1175/MWR-D-14-00004.1.

Putman, W. M., and S.-J. Lin, 2007: Finite-volume transport on various cubed-sphere grids. *Journal of Computational Physics*, **227**, 55-78, https://doi.org/10.1016/j.jcp.2007.07.022.

Ribeiro, B. Z., S. J. Weiss, and L. F. Bosart, 2022: An Analysis of the 3 May 2020 Low-Predictability Derecho Using a Convection-Allowing MPAS Ensemble. *Weather and Forecasting*, **37**, 219-239, https://doi.org/https://doi.org/10.1175/WAF-D-21-0092.1.

Ringler, T., J. Thuburn, J. B. Klemp, and W. C. Skamarock, 2010: A unified approach to energy conservation and potential vorticity dynamics for arbitrarily-structured C-grids. *Journal of Computational Physics*, **229,** 3065–3090, https://doi:10.1016/j.jcp.2009.12.007.

Sakaguchi, K., L. R. Leung, C. M. Zarzycki, J. Jang, S. McGinnis, B. E. Harrop, W. C. Skamarock, A. Gettelman, C. Zhao, W. J. Gutowski, S. Leak, and L. Mearns, 2023: Technical descriptions of the experimental dynamical downscaling simulations over North America by the CAM–MPAS variable-resolution model. *Geoscientific Model Development,* **16**, 3029–3081, https://doi.org/10.5194/gmd-16-3029-2023.

Schwartz, C. S., 2019: Medium-Range Convection-Allowing Ensemble Forecasts with a Variable-Resolution Global Model. *Monthly Weather Review*, **147**, 2997-3023, https://doi.org/https://doi.org/10.1175/MWR-D-18-0452.1

Shaw J., H. Weller, (2016: Comparison of terrain-following and cut-cell grids using a nonhydrostatic model. *Monthly Weather Review*, **144**, 2085–2099. https://doi.org/10.1175/MWR-D-15-0226.1

Skamarock, W. C., 2008: A Linear Analysis of the NCAR CCSM Finite-Volume Dynamical Core. *Monthly Weather Review*, **136**, 2112–2119, https://doi.org/10.1175/2007MWR2217.1.

Skamarock, W. C., and A. Gassmann, 2011: Conservative Transport Schemes for Spherical Geodesic Grids: High-Order Flux Operators for ODE-Based Time Integration. *Monthly Weather Review*, **139**, 2962–2975, https://doi.org/10.1175/MWR-D-10-05056.1.

Skamarock, W. C., J. B. Klemp, M. G. Duda, L. D. Fowler, S.-H. Park, and T. D. Ringler, 2012: A Multiscale Nonhydrostatic Atmospheric Model Using Centroidal Voronoi Tesselations and C-Grid Staggering. *Monthly Weather Review*, **140**, 3090–3105, https://doi.org/10.1175/MWR-D-11-00215.1.

Skamarock, W. C., M. G. Duda, S. Ha, and S.-H. Park, 2018: Limited-Area Atmospheric Modeling Using an Unstructured Mesh. *Monthly Weather Review*, **146**, 3445–3460, https://doi.org/10.1175/MWR-D-18-0155.1.

Skamarock, W. C., H. Ong, and J. B. Klemp, 2021: A Fully Compressible Nonhydrostatic Deep-Atmosphere Equations Solver for MPAS. *Monthly Weather Review*, **149**, 571–583, https://doi.org/10.1175/MWR-D-20-0286.1.

Skamarock, W.C., and coauthors, 2021: A Description of the Advanced Research Version of the WRF Version 4. NCAR Technical Note 556. http://dx.doi.org/10.5065/1dfh-6p97

Steppeler J, H.W. Bitzer, M. Minotte, L. Bonaventura, 2002: Nonhydrostatic atmospheric modelling using a z-coordinate representation. *Monthly Weather Review*, **130**, 2143–2149. https://doi.org/10.1175/1520-0493(2002)130%3C2143:NAMUAZ%3E2.0.CO;2

Thuburn, J., T. Ringler, W. C. Skamarock, and J. B. Klemp, 2009: Numerical representation of geostrophic modes on arbitrarily structured C-grids. *Journal of Computational Physics*, **228**, 8321–8335, https://doi.org/10.1016/j.jcp.2009.08.006.

Uccellini, L. W., R. W. Spinrad, D. M. Koch, C. N. McLean, and W. M. Lapenta, 2022: EPIC as a Catalyst for NOAA's Future Earth Prediction System. *Bulletin of the American Meteorological Society*, **103**, E2246-E2264, https://doi.org/https://doi.org/10.1175/BAMS-D-21-0061.1.

UXarray Organization. (2023). UXarray (version 0.9) [Software]. Project Raijin & Project SEATS. https://github.com/UXARRAY/uxarray.

Viner, K., 2023: A new physics-dynamics coupling paradigm for NOAA's Unified Forecast System (UFS). *Unified Forecast System (UFS) Physics Workshop*, May 16-18, 2023, Boulder, CO.

Wang, J., K. Viner, D. Swales, D. Rosen, D. Heinzeller, and L. Bernardet, 2023: Integration of MPAS Dycore into UFS. Technical Report.

Whitehead, J. P., C. Jablonowski, R. B. Rood, and P. H. Lauritzen, 2011: A Stability Analysis of Divergence Damping on a Latitude–Longitude Grid. *Monthly Weather Review*, **139**, 2976–2993, https://doi.org/10.1175/2011MWR3607.1.

Zhao, M., I. M. Held, and S.-J. Lin, 2012: Some Counterintuitive Dependencies of Tropical Cyclone Frequency on Parameters in a GCM. *Journal of the Atmospheric Sciences*, **69**, 2272–2283, https://doi.org/10.1175/JAS-D-11-0238.1.

Zhou, L., and L. Harris, 2022: Integrated Dynamics‐Physics Coupling for Weather to Climate Models: GFDL SHiELD With In‐Line Microphysics. *Geophysical Research Letters*. **49**, e2022GL100519, https://doi.org/10.1029/2022GL100519.

Zhou, X., and H.-M. H. Juang, 2023: A model instability issue in the National Centers for Environmental Prediction Global Forecast System version 16 and potential solutions. *Geoscientific Model Development,* **16**, 3263–3274, https://doi.org/10.5194/gmd-16-3263-2023.

Appendix A — Requirements for RRFSv1

The following is a list of requirements for RRFSv1, reproduced from the RRFSv1 Project Plan.

> The primary goal for version 1 of the RRFS is to *unify the multitude of regional, high resolution forecast systems in the production suite under a single system*. Accordingly, the system must therefore meet the performance of the existing systems it seeks to replace. Unification of several disparate systems into a single, unified application will enable future versions of RRFS to continue to meet requirements identified and proposed by the NOAA/NWS Analyze, Forecast and Support (AFS) Office in Capabilities and Requirements Decision Support (CaRDS) document 21-012[17] as well as CaRDS materials on high resolution ensembles (CaRDS 22-023)[18].
>
> This goal will be achieved by meeting the following high level requirements for RRFSv1.

1. RRFSv1 ensemble performance is as good or better than HREFv3 as measured by:
    a. Objective statistical verification
    b. Expert analysis
    c. Equivalent products, with exceptions for removal of those which are no longer necessary
    d. Services (e.g., latency, additional forecast cycles, changes to forecast range, coverage)
2. RRFSv1 deterministic performance is as good or better than 3 km NAM nests and HRRR (CONUS and Alaska domains) as measured by:
    a. Objective statistical verification
    b. Expert analysis
    c. Equivalent products, with exceptions for removal of those which are no longer necessary
    d. Services (e.g., latency, additional forecast cycles, forecast range, coverage)
3. RRFSv1 provides functionally equivalent products and services for the NAM 1.5 km Fire Weather domain.

---

[17] CaRDS (21-012) on HRRR: https://nsdesk.servicenowservices.com/cards?sys_id=9c44c12d1b403cd021dd20eae54bcb53&view=sp&id=cards_request&table=x_anno2_cards_nws_card [Requires NOAA login]

[18] CaRDS (22-023) on HREF: https://nsdesk.servicenowservices.com/cards?sys_id=054f83e51b3e9d1021dd20eae54bcb74&view=sp&id=cards_request&table=x_anno2_cards_nws_card [Requires NOAA login]

4. HiRes Window Guam will be updated to be consistent with the RRFSv1 model configuration.
5. RRFSv1 will not exceed computational resources as permitted by NWS Central Operations and the NCEP High Performance Computing Resource Allocation Council.
6. Dissemination of RRFSv1 products will conform to all bandwidth constraints (e.g. NOAA Operational Model Archive Distribution System and the Satellite Broadcast Network).

Appendix B — Detailed Summaries of Past NOAA Testbeds

1.1. HWT Spring Forecasting Experiments

Expanding upon Section 2.3.1, the subsections below document the chronological evaluation of the FV3 at convection-allowing scales over the past several years in the HWT SFE[19].

1.1.1. 2017 HWT SFE

The FV3 was examined for the first time in the SFE at convection-allowing scales in 2017. There were two versions of the FV3 that were similarly configured from GFDL and OU CAPS. They both used a 13-km global grid configuration with a 3-km CONUS nest and were initialized from the GFS. The only differences between the two runs were the microphysics scheme (GFDL and Thompson) and PBL scheme (MRF and YSU). With this being the first year of evaluating the FV3 at convection-allowing scales, the expectations were low just to see if reasonable storm-scale forecasts could be made with FV3. While objective verification metrics indicated that the FV3 forecasts exhibited lower skill for storm-attributes than WRF-based CAM forecasts, the subjective ratings from participants indicated that reflectivity forecasts were often comparable to operational CAMs (Fig. 1). This was an encouraging result in the first year that supported further research and development to improve the FV3 for storm-scale applications.

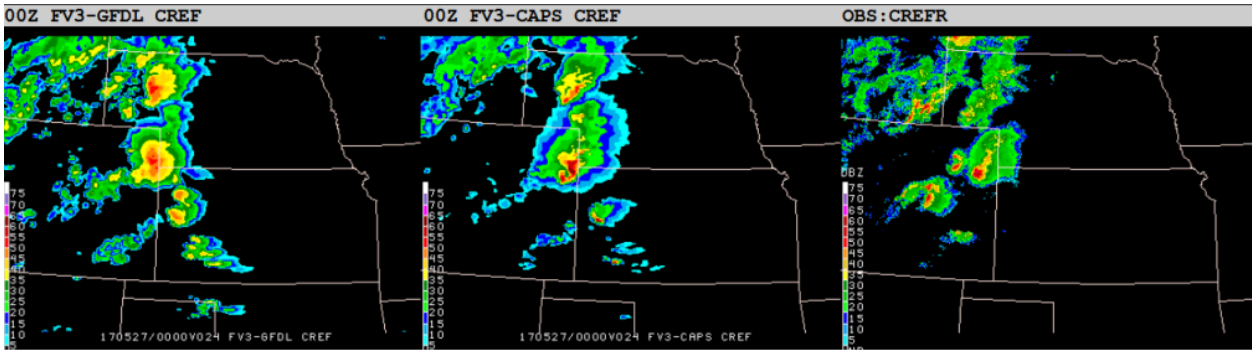

Figure 1. Example of subjective comparison plots used for rating FV3 performance at convective scales in the first year. The left panel shows a 24-h forecast of composite reflectivity of the FV3-GFDL, the middle panel shows the 24-h forecast of composite reflectivity of the FV3-CAPS, and the right panel shows the observed composite reflectivity at 0000 UTC on 27 May 2017.

1.1.2. 2018 HWT SFE

During the 2018 HWT SFE, there was an expansion of evaluation activities of FV3 ~3-km CONUS nests within a global configuration, as NSSL joined GFDL and CAPS in generating a 0000 UTC run initialized from the GFS (Fig. 2). The NSSL and CAPS configurations were very similar with both using the MYNN PBL scheme and Thompson microphysics, so the subjective ratings from participants from those two runs were nearly identical (Fig. 3). The GFDL version of the FV3 was rated lower than the

[19] The full HWT reports may be found here: https://hwt.nssl.noaa.gov/efp/archive/.

NSSL and CAPS versions (Fig. 3) during the 2018 HWT SFE, largely owing to an overforecast of coverage and intensity of convective storms (e.g., Fig. 2). Note that all versions of FV3 still lagged clearly behind the operational CAM standard at the time (i.e., HRRRv3) for convective forecasting applications. These subjective ratings and impressions are confirmed with the extensive objective analysis and verification performed and documented by Gallo et al. (2021).

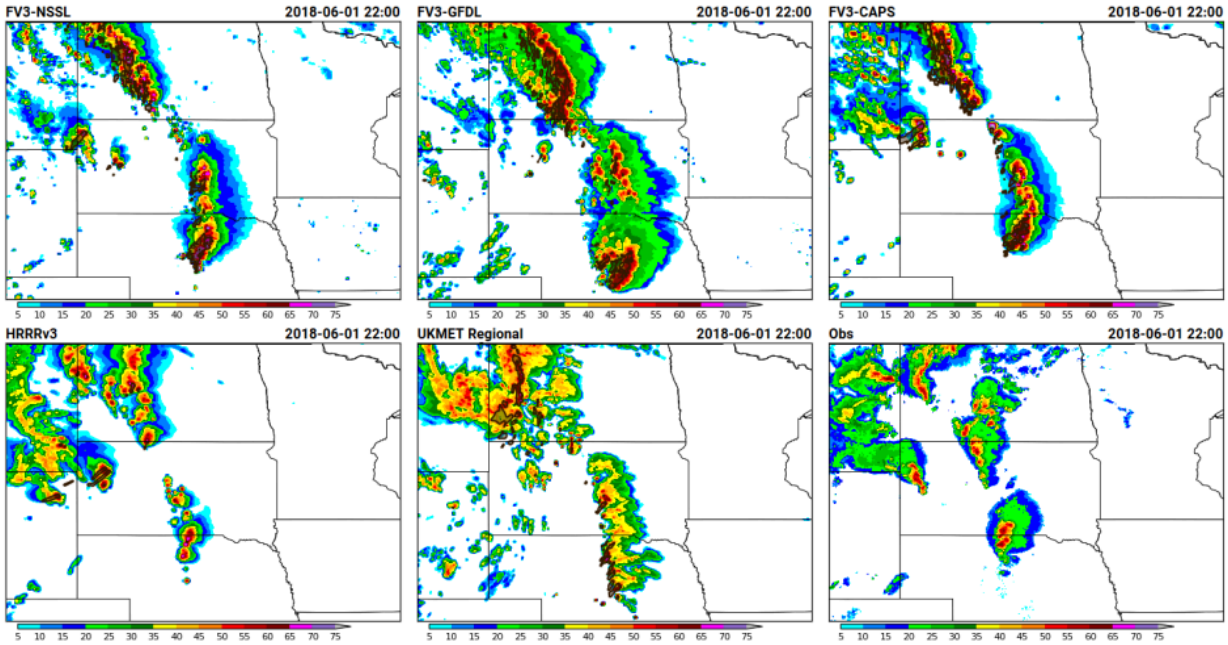

Figure 2. Example of subjective comparison plots used for rating CAM performance at convective scales valid at 2200 UTC on 1 June 2018. The 21-h forecasts of composite reflectivity are shown for the a) FV3 NSSL (upper-left panel), b) FV3-GFDL (upper-middle panel), c) FV3-CAPS (upper-right panel), d) HRRRv3 (lower-left panel), and e) UK Met Office UM. The bottom-right panel shows the observed composite reflectivity at 2100 UTC on 1 June 2018.

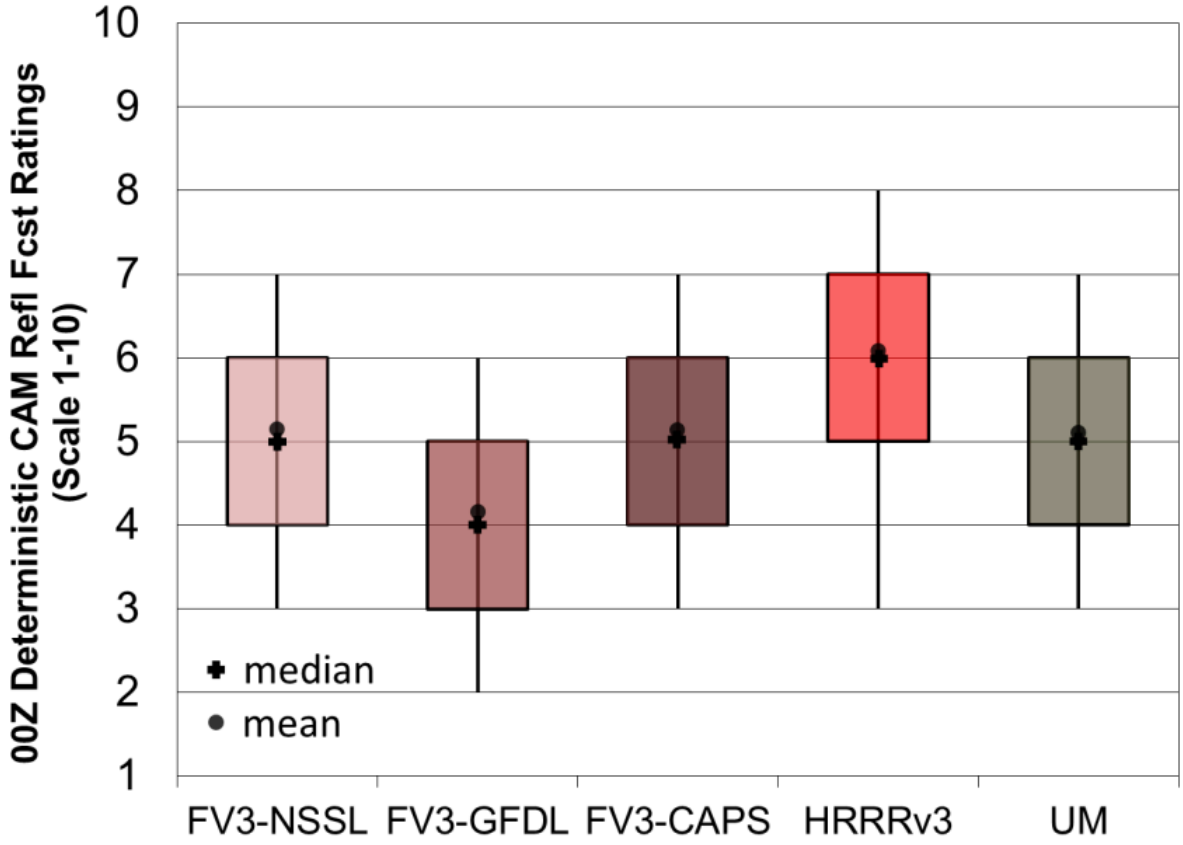

Figure 3. Box-and-whiskers plot of subjective ratings (1-10) for deterministic CAM reflectivity forecasts from 0000 UTC during the five-week 2018 HWT SFE. The boxes represent the interquartile range, and the whiskers represent the 10th and 90th percentiles. The crosses represent the median ratings, and the circles represent the mean ratings.

1.1.3. 2019 HWT SFE

A regional CAM version of FV3 was introduced in 2019. These runs were initially called stand-alone regional (SAR) runs, but the name was later changed to limited-area model (LAM) runs. Both EMC and NSSL ran global-with-nest and regional versions of FV3 at CAM scales with comparable configurations. In terms of subjective evaluation by SFE participants for Day 1 forecasts, the regional versions of FV3 were rated at least as high as their respective global-with-nest versions (not shown). This result validated the capability of running the FV3 in a regional, limited-area mode for convective applications. SFE participant comments about these FV3 runs frequently mentioned overly intense convection and a 2-m cold bias. GSL (ESRL/GSD at the time) also ran a regional FV3 with the same ICs/LBCs and physics configurations as HRRRv4. However, configuration changes of the HRRRv4 during the SFE and limited availability of this FV3 run until the end of the experiment limits the generalizability of the results. Nevertheless, the subjective ratings of this FV3 run (i.e., HRRR-FV3) for storm attributes and environment were notably lower than both HRRRv3 and HRRRv4 (Fig 4), indicating that significant development work was still needed for the regional FV3 run. CAPS also ran an ensemble of FV3 runs with different physics schemes. The runs

using Thompson and NSSL microphysics schemes received similar subjective ratings while the run using the Morrison scheme received notably lower ratings. Meanwhile, all of the runs with different PBL schemes (MYNN, Shin-Hong, and EDMF) received similar ratings and displayed a persistent 2-m moist bias.

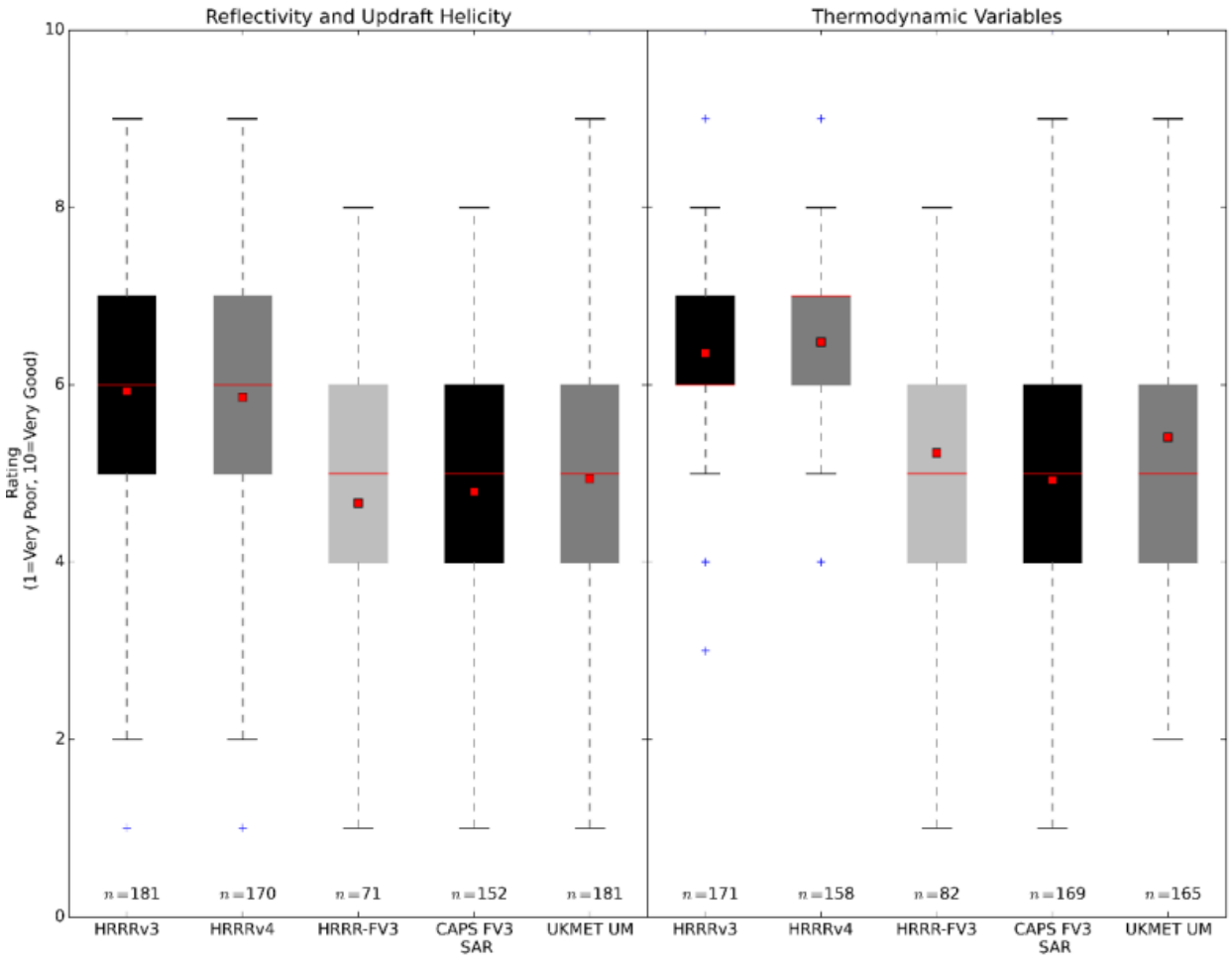

Figure 4. Participant subjective ratings of five deterministic model configurations provided by three organizations. Red squares indicate the mean rating, and response sample size is located directly above the label. Reflectivity and UH ratings are on the left, and thermodynamic variable ratings are on the right.

1.1.4. 2020 HWT SFE

A number of sensitivity tests were performed for the regional FV3 CAMs in 2020, including exploring the impact of advanced physics packages, number of vertical levels, initial conditions, diffusivity settings, and land-surface model. A summary of these sensitivity tests are listed below in terms of the magnitude on the impact on model performance for convective applications:

- An advanced physics suite (MYNN PBL and Thompson MP) offered a large improvement over the standard physics suite (EDMF PBL and GFDL MP).
- Using the NOAH land-surface model resulted in improved storm-attribute and environmental forecasts over runs with the RUC land-surface model.

- Increasing the number of vertical levels from 50 to 80 offered a small forecast improvement that was most evident in early forecast hours.
- In terms of the horizontal advection method, a more accurate and less diffusive setting (i.e., HORD=5) revealed a small improvement in storm-attribute forecasts.
- While the initial conditions used (i.e., GFS or HRRR) had a large impact on subsequent forecasts, there was not a set of ICs that performed consistently better.

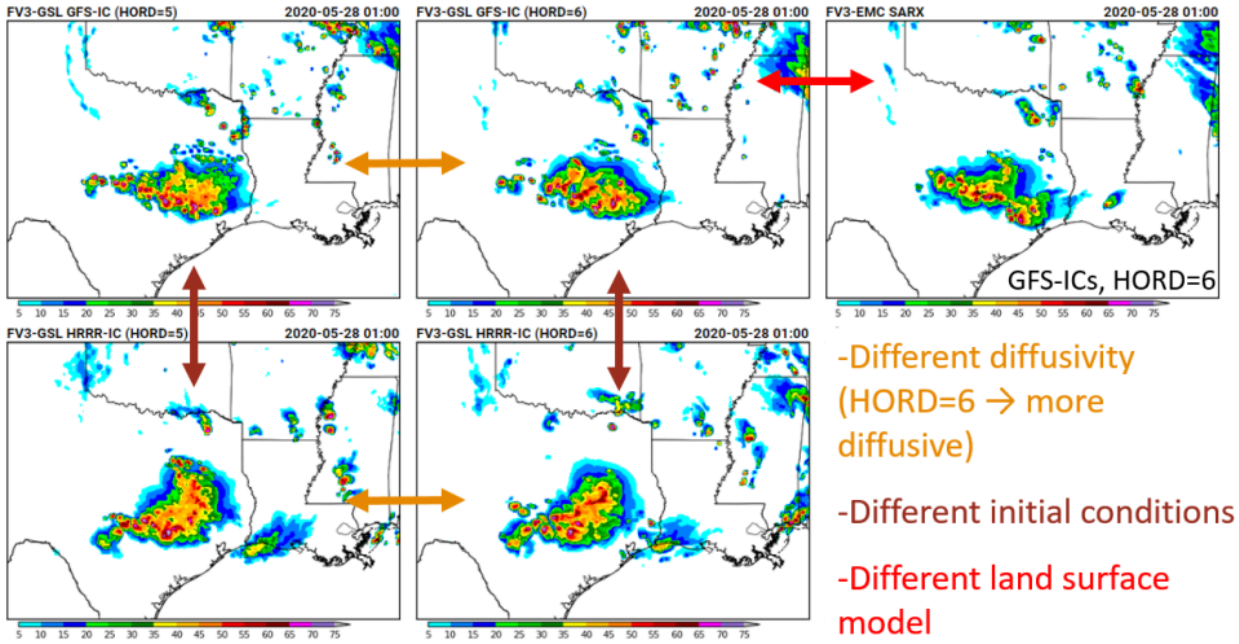

Figure 5. An annotated example of the panels that 2020 HWT SFE participants evaluated. Colored arrows indicate the differences between pairs of models. Annotations cover a sixth panel, which showed radar observations during participant evaluations.

1.1.5. 2021 HWT SFE

A large number of regional FV3 CAM runs, including CAM ensembles, were evaluated during the 2021 HWT SFE comprising 46 of the 64 Community Leveraged Unified Ensemble (CLUE; Clark et al. 2018) members. Regarding the deterministic FV3 CAMs, they showed improvement relative to prior years, but still lagged behind the operational HRRRv4 in terms of storm-attribute and environment forecasts for convective applications (Fig. 6). Participant comments often noted the tendency of FV3-based CAMs to have too much storm coverage with the storms having overly intense circular updrafts. Ironically, these overly abundant and intense storms often occurred in an environment characterized by having a low bias in instability, as noted by participants.

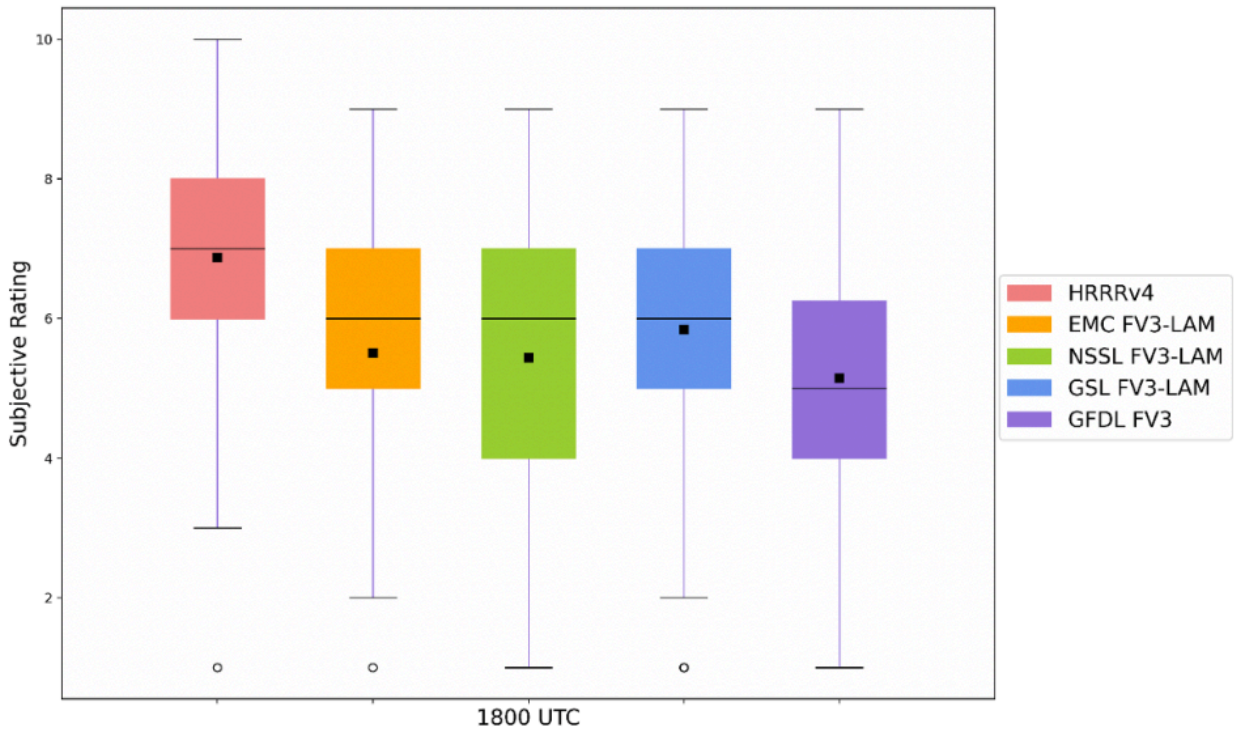

Figure 6. Deterministic flagship environmental results at 1800 UTC during the 2021 HWT SFE. Mean subjective evaluation ratings are shown in the black square on each bar.

This was also the first year in which the regional FV3 CAM runs were performed over the North American domain. These runs fared well relative to their CONUS counterparts, especially for Day 2, validating the move to the North American domain. Different physics suites were also examined in FV3 as part of the RRFS cloud runs. The control members (initialized from the GFS) for three different physics suites were compared, and the run with MYNN PBL and Thompson MP schemes generated the highest-rated forecasts while the runs from the other physics suites (TKE-EDMF/GFDL and Hybrid-EDMF/NSSL) were rated much lower, especially for environmental fields.

Regarding CAM ensembles, this was the first time that FV3-based CAM ensembles were evaluated in the HWT SFE. GSL ran an FV3-based ensemble forecast (GSL RRFS) that was initialized from the HRRRDAS. EMC and GSL also ran a multi-physics, multi-IC (GFS and GEFS) FV3 CAM ensemble using cloud computing (RRFS Cloud). While the HREF was still rated as the best-performing CAM ensemble, the GSL RRFS, in particular, performed well and just slightly lagged behind the HREF in terms of subjective ratings (Fig. 7). Indicative of including poorer-performing physics members (as noted above), the RRFS Cloud configuration received lower subjective ratings than the HREF and GSL RRFS.

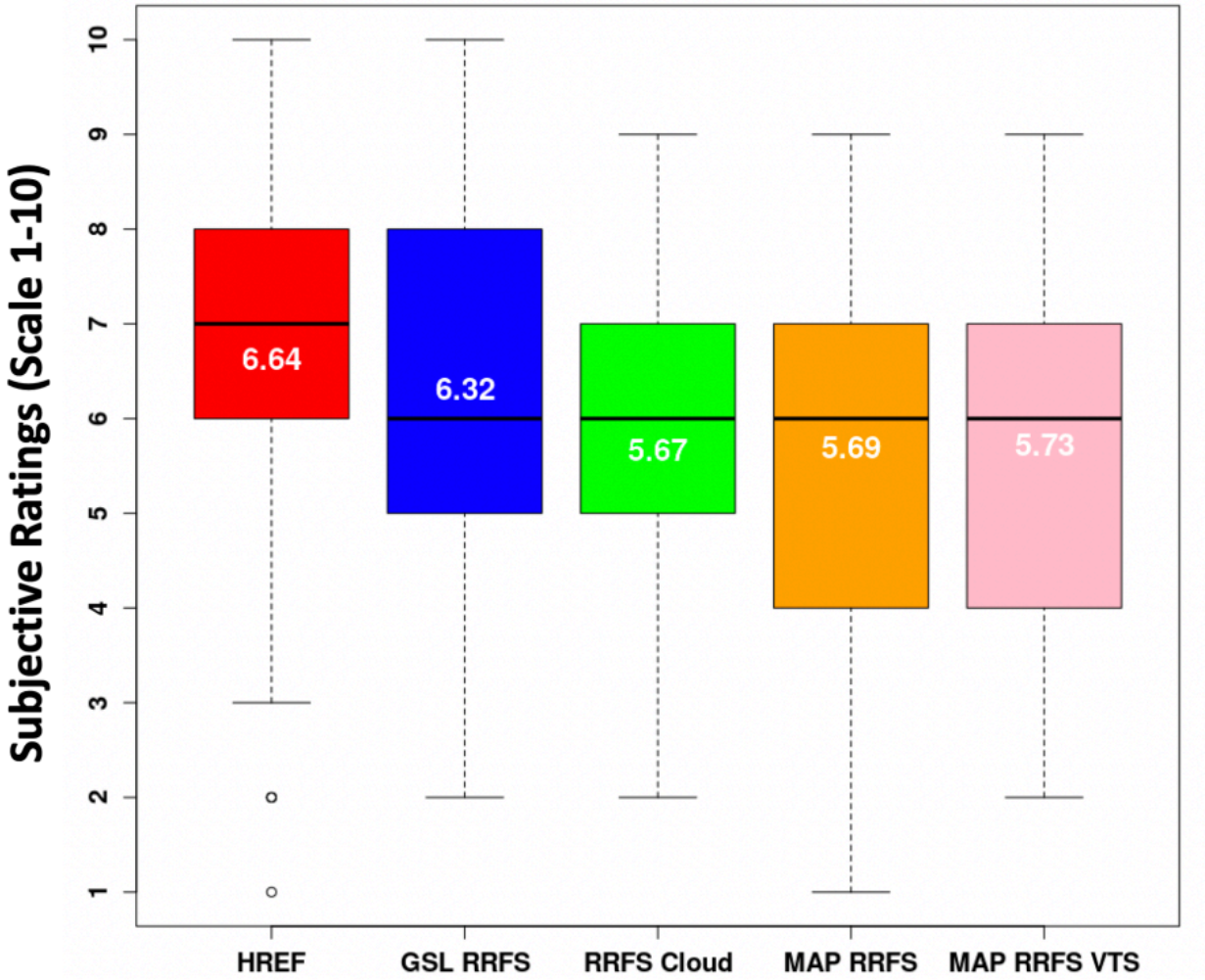

Figure 7. Distributions of subjective ratings (1-10; with 10 being best) by SFE participants of hourly maximum fields for severe weather forecasting over a mesoscale area of interest for the forecast hours 13-36 for the C1: CLUE 00Z Ensembles evaluation (HREF – red; GSL RRFS – blue; RRFS Cloud – green; MAP RRFS – orange; MAP RRFS VTS – pink). The numbers in white text indicate mean ratings, while the horizontal black lines indicate the median.

1.1.6. 2022 HWT SFE

The first end-to-end RRFS prototypes, including conventional (i.e., no radar) data assimilation, were evaluated during the 2022 HWT SFE. In blinded evaluations, the two deterministic RRFS prototypes showed skill approaching that of the operational HRRRv4, though the HRRR was still more frequently ranked as the best CAM for most storm-attribute and environmental variables than any other model. The only field where the RRFS prototype had a clear advantage was for 2-m dewpoint, owing to the systematic afternoon dry bias in the HRRR during the warm season. For the CAM ensemble evaluations, the RRFS prototype ensemble compared very favorably to the HREF. This was a noteworthy result as no other CAM ensemble in previous HWT SFEs had received subjective ratings as close to the HREF as this RRFS prototype. For

some cases, the HREF was noted as having more spread that was beneficial in capturing the severe weather event (e.g., Fig. 8).

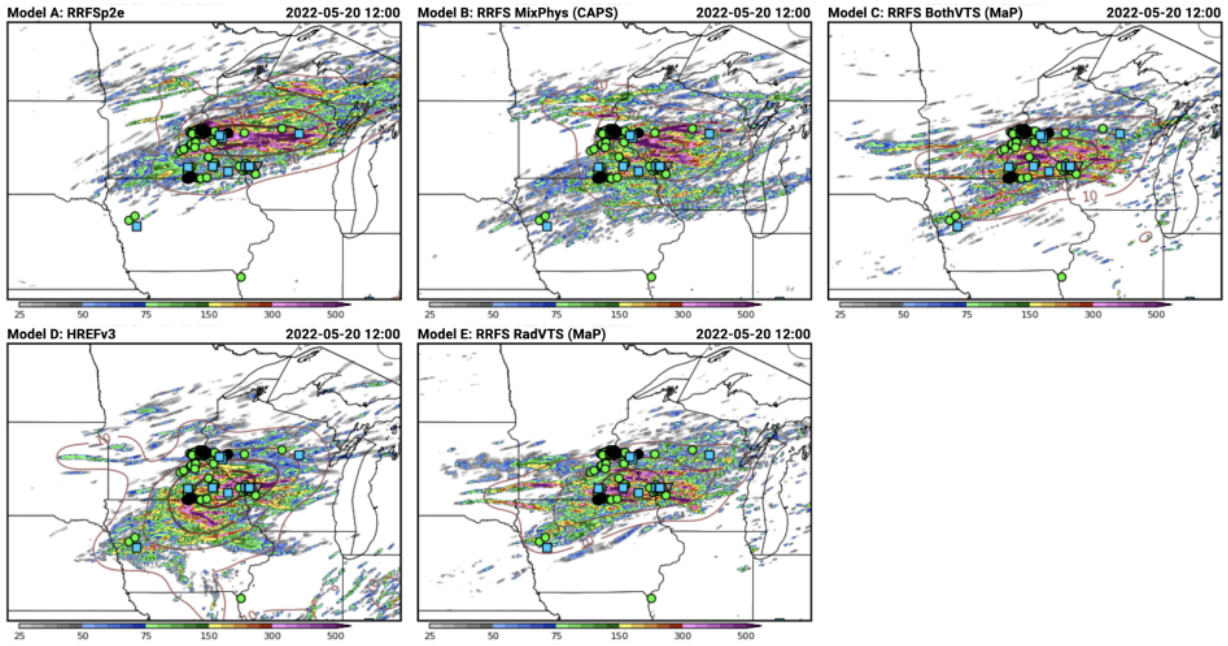

Figure 8. Example of multi-panel comparison webpage for the 0000 UTC CAM ensemble evaluation during the 2022 SFE. The 24-h ensemble maximum UH (shaded) and neighborhood probability of UH>99.85$_{th}$ percentile (contoured) is displayed for RRFSp2e (upper left), RRFS MixPhys (upper middle), RRFS BothVTS (upper right), HREFv3 (lower left), and RRFS RadVTS (lower middle) for 19 May 2022. Preliminary severe storm reports are also overlaid (wind – blue squares, hail – green circles, and tornado – red upside-down triangles. Significant reports are filled in black). Note, only the "Model A", "Model B", etc., labels were displayed during evaluations.

1.1.7. 2023 HWT SFE

During the 2023 HWT SFE, the full RRFS prototype with radar data assimilation was in place for both deterministic and ensemble evaluations. One of the more notable results was in the data assimilation evaluation, where the HRRR had superior performance in short-term forecasts compared to the RRFS control member. The RRFS control run systematically produced overly intense convection in the one-hour forecast with an abundance of spurious storms (e.g., Fig. 9), as noted in participant ratings and comments. This radar DA issue seemingly impacted the resultant RRFS forecast also, as the 0000 UTC RRFS runs were rated lower than the HRRR and MPAS in blinded forecast evaluations (Fig. 10; Table 1). The three FV3-based CAMs were the poorest performers with the RRFS receiving the highest ratings of the FV3 runs. Interestingly, the 1200 UTC-initialized RRFS control forecasts performed much better than the 0000-UTC initialized runs (Fig. 11), indicating that the RRFS performance was strongly dependent on the initialization time during the 2023 HWT SFE, much more so than the operational HRRR.

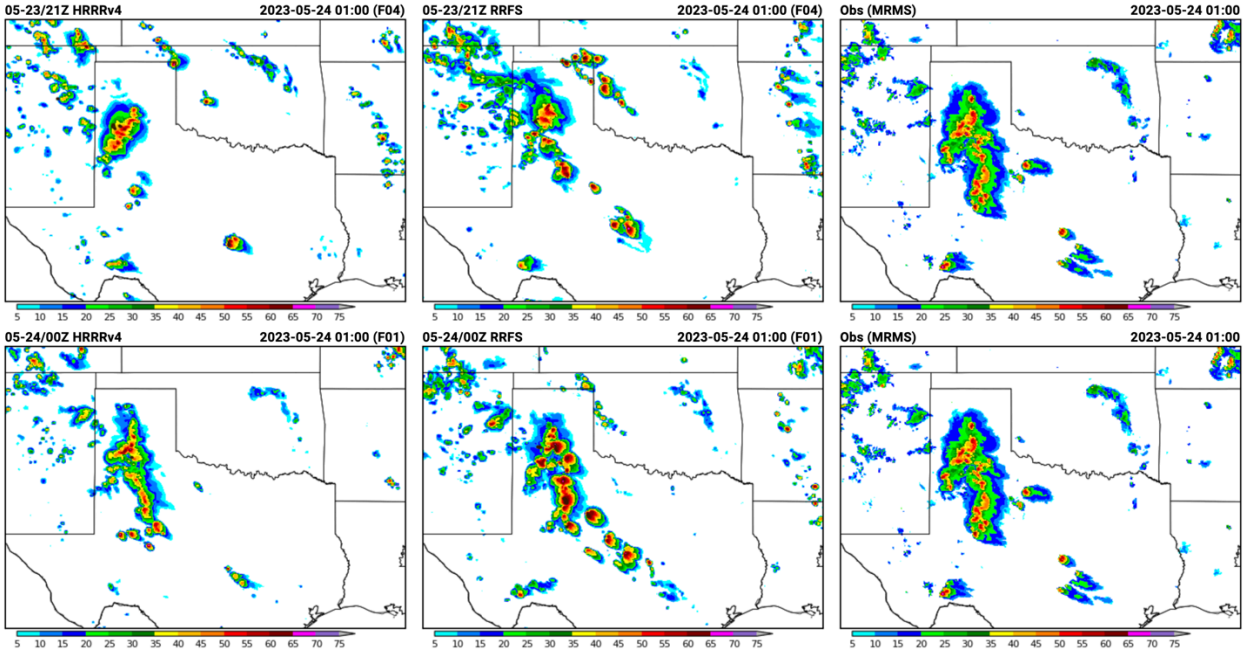

Figure 9. Example of multi-panel comparison webpage for the RRFS vs. HRRR DA evaluation. The top row displays simulated composite reflectivity from 2100 UTC initializations of HRRRv4 (left) and RRFS (middle) valid at 0100 UTC compared to MRMS observations (right). The bottom row displays the same as the top, except for 0000 UTC initializations.

The RRFS ensemble evaluation followed the same pattern as the deterministic evaluation results, where the HREF had an overall edge in 0000-UTC initialized forecasts while the 1200-UTC initialized forecasts of RRFS and HREF were rated similarly (Fig. 12). Objective statistics for composite reflectivity (>40 dBZ), however, still give a distinct edge to the 1200 UTC HREF over the single-physics RRFS. Other RRFS configurations (multi-physics and time-lagging) did not offer an improvement in subjective ratings over the single-physics, single-initialization-time RRFS (Fig. 12), though sub-optimal physics options (i.e., GFS PBL scheme) and performance dependence on initialization times hinder those specific configurations. It is also worth noting that the Day 2 subjective evaluations strongly favored the HREF over any 1200 UTC RRFS configuration. Overall, the RRFS results from the 2023 HWT SFE are concerning given the limited time remaining for RRFS development before the expected code freeze.

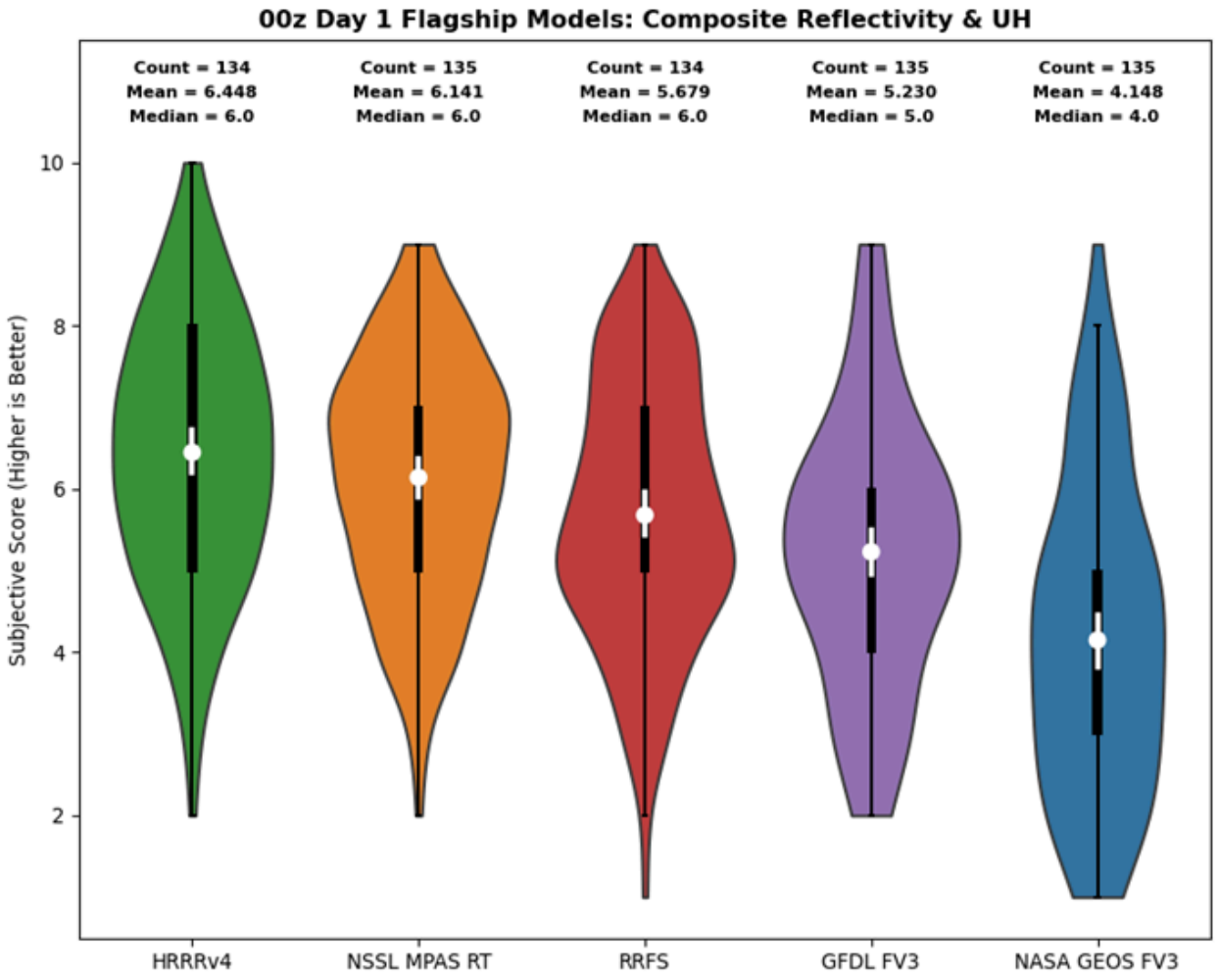

Figure 10. Distribution of subjective scores received by each deterministic flagship model at Day 1 lead times during the 5-week experiment. The white dots represent the mean scores for each ensemble, and the white bars indicate the 95% confidence intervals for each mean.

Table 1. Configuration details for the 0000-UTC initialized flagship models during the 2023 HWT SFE.

Model	Data Assimilation	IC/LBCs	DyCore	Microphysics	PBL	LSM	Shallow Cumulus
HRRR	Hybrid 3DEnVar	Own/RAP	WRF	Thompson	MYNN	RUC	none
RRFS	Hybrid 3DEnVar	Own/GFS	FV3	Thompson	MYNN	RUC	none
NSSL-MPAS	none	RRFS/RRFS	MPAS	Thompson	MYNN	RUC	Grell-Freitas
GFDL-FV3	none	GFS/n/a	FV3	GFDL	TKE-EDMF	NOAH-MP	none
NASA-FV3	GEOS-DA	Own/n/a	FV3	GEOS-GFDL	Lock-Louis	NASA	none

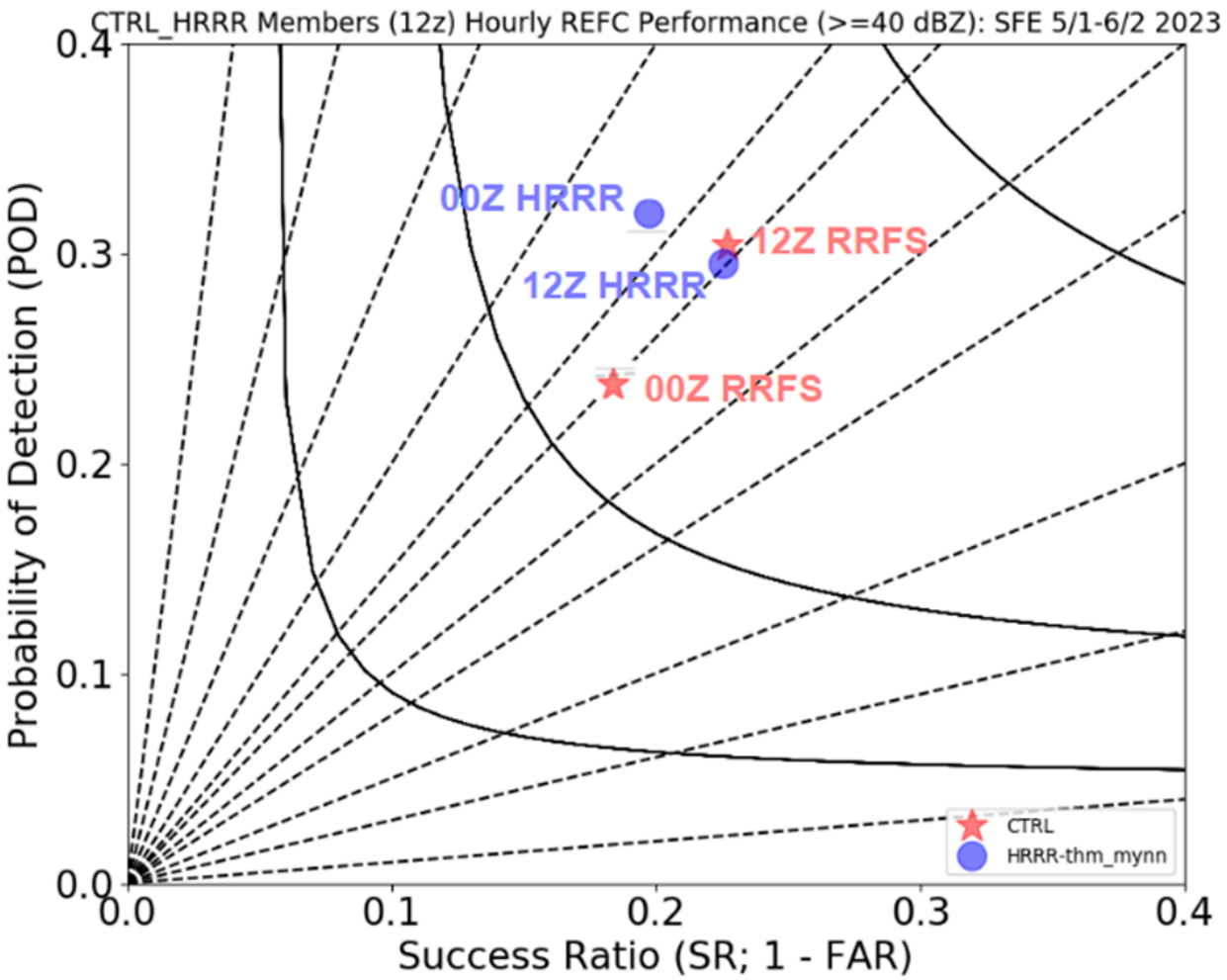

Figure 11. Performance diagram for hourly composite reflectivity ≥40 dBZ covering the 24-h convective day (i.e., 12-12Z) over the five-week period of the HWT SFE. The 00Z and 12Z HRRR (blue circle) and RRFS (red star) performance characteristics are labeled on the diagram. The statistics are only calculated over the primary mesoscale domain used each day for evaluation activities.

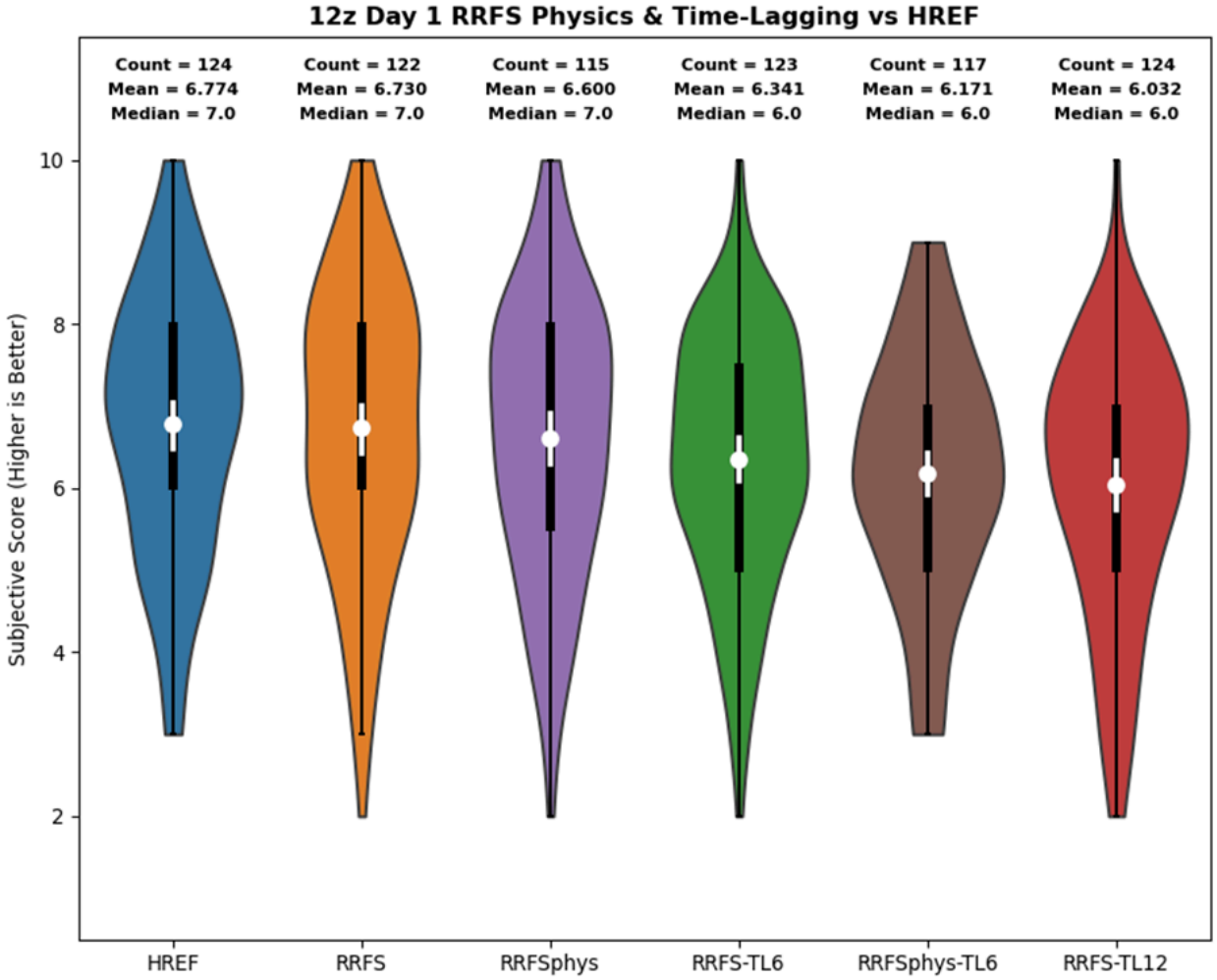

Figure 12. Distribution of subjective scores received by each 12Z ensemble at Day 1 lead times during the 5-week experiment. The white dots represent the mean scores for each ensemble, and the white bars indicate the 95% confidence intervals for each mean..

1.2. HMT Forecasting Experiments

Expanding upon Section 2.3.2, the subsections below document the chronological evaluation of the FV3 at convection-allowing scales over the past several years in the HMT[20].

1.2.1. 2018 FFaIR Experiment

The experiment featured several deterministic high-resolution models, including the High-Resolution Rapid Refresh (HRRRv3) provided by GSD and two 3-km FV3 CAM variants provided by OU/CAPS. Two 3-km deterministic FV3-based models were provided by the OU CAPS team, one with NSSL microphysics and the other with Thompson microphysics.

[20] The full HMT reports may be found here: https://www.wpc.ncep.noaa.gov/hmt/experimentsummaries.shtml.

The FV3-NSSL regularly struggled to produce organized precipitation over the CONUS. Participants frequently commented on the low magnitude and scattered nature of the precipitation. Figure 13 shows a MODE analysis of 24 hour QPF from both the FV3-Thompson and FV3-NSSL at 0.5 in. threshold with three separate areas highlighted in the red circles where MRMS-GC QPE was not matched by MODE to QPF from either model. At times, erroneously high QPF maxima would occur related to certain features which the participants found distracting and could be misleading in the forecast process. The FV3-Thompson, likewise, often struggled to produce precipitation leading to low magnitudes and scattered representations but did slightly better than the FV3-NSSL in organization, timing and location accuracy. Figure 14 shows that the FV3-Thompson had slightly higher CSI values at the 0.5/1.0/2.0 in. thresholds when compared to the FV3-NSSL. There were days when the FV3-based models performed well, as both models were given scores as high as a 9 out of 10 for some events, but the erratic errors in the rainfall prediction tended to create forecaster distrust in both FV3-based models over the experiment.

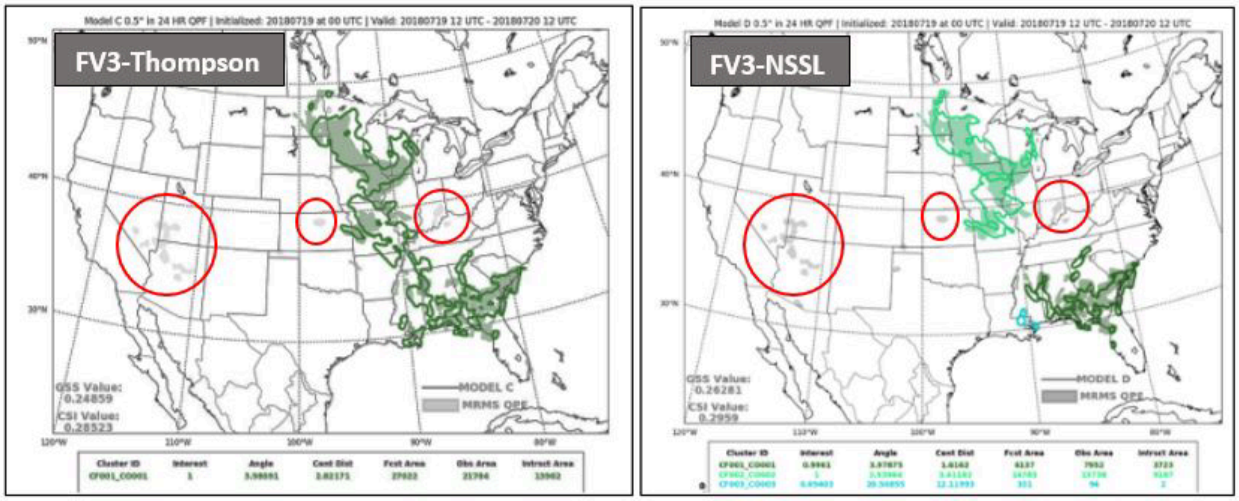

Figure 13. Objective MODE performance of the 24-hour 0.5” QPF from the FV3-Thompson (left) and FV3-NSSL (right) both valid at 12 UTC July 20, 2018. Noted in the red circles are areas of MRMS QPE that were not matched by MODE to the model QPF.

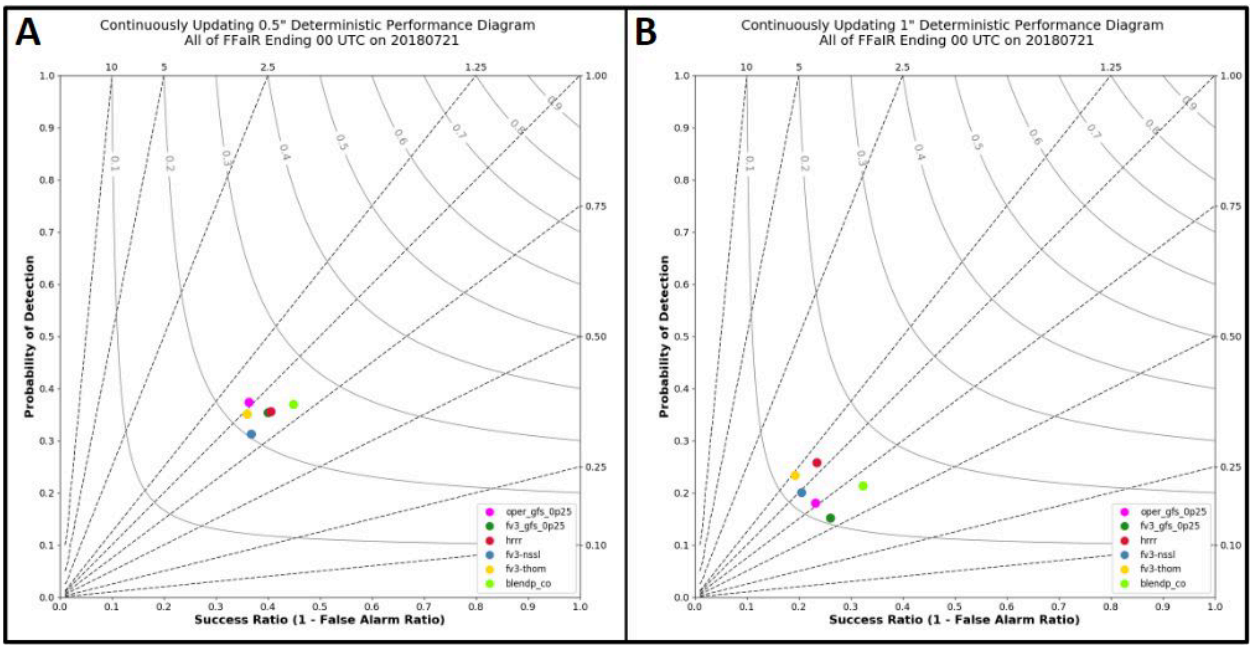

Figure 14. Performance diagram displaying the FV3-GFS (dark green), HRRRv3 (red), FV3-NSSL (blue), FV3-Thompson (yellow), NBMv3.1 (light green), and the operational GFS (magenta) over the four weeks of the 2018 FFaIR Experiment.

The HRRRv3 had the highest subjective and objective scores of the high-resolution CAMs. It tended to have a low bias in the Southwest U.S. and be too heavy at times in strongly forced patterns. Both high-resolution FV3-based models had the lowest CSI values at 0.5 in. and low 1.0 in. scores as well. WPC-HMT recommended that work continue on the FV3-Thompson and FV3-NSSL to investigate whether the microphysics schemes were causing the large QPF differences sometimes seen. WPC-HMT also recommended further study to improve HRRRv3 QPF in the Southwest U.S. associated with monsoon moisture.

1.2.2. 2019 FFaIR Experiment

In addition to evaluating the newer version of the deterministic High-Resolution Rapid Refresh model (HRRR), the 2019 FFaIR experiment included numerous experimental deterministic and ensemble model guidance that utilized the FV3 core. For instance, the Environmental Modeling Center (EMC) provided deterministic runs of their CAMs that use the FV3 core; the FV3-Nest and the FV3 stand-alone regional model. Likewise, the University of Oklahoma's Center for Analysis and Prediction of Storms (CAPS) research team provided their CAM ensemble, the storm-scale ensemble forecast (SSEF), which is composed of 14 members, all with the FV3 core. To further assess the performance of the FV3 core in CAMs three members of the SSEF were treated as deterministic runs and analyzed individually as well.

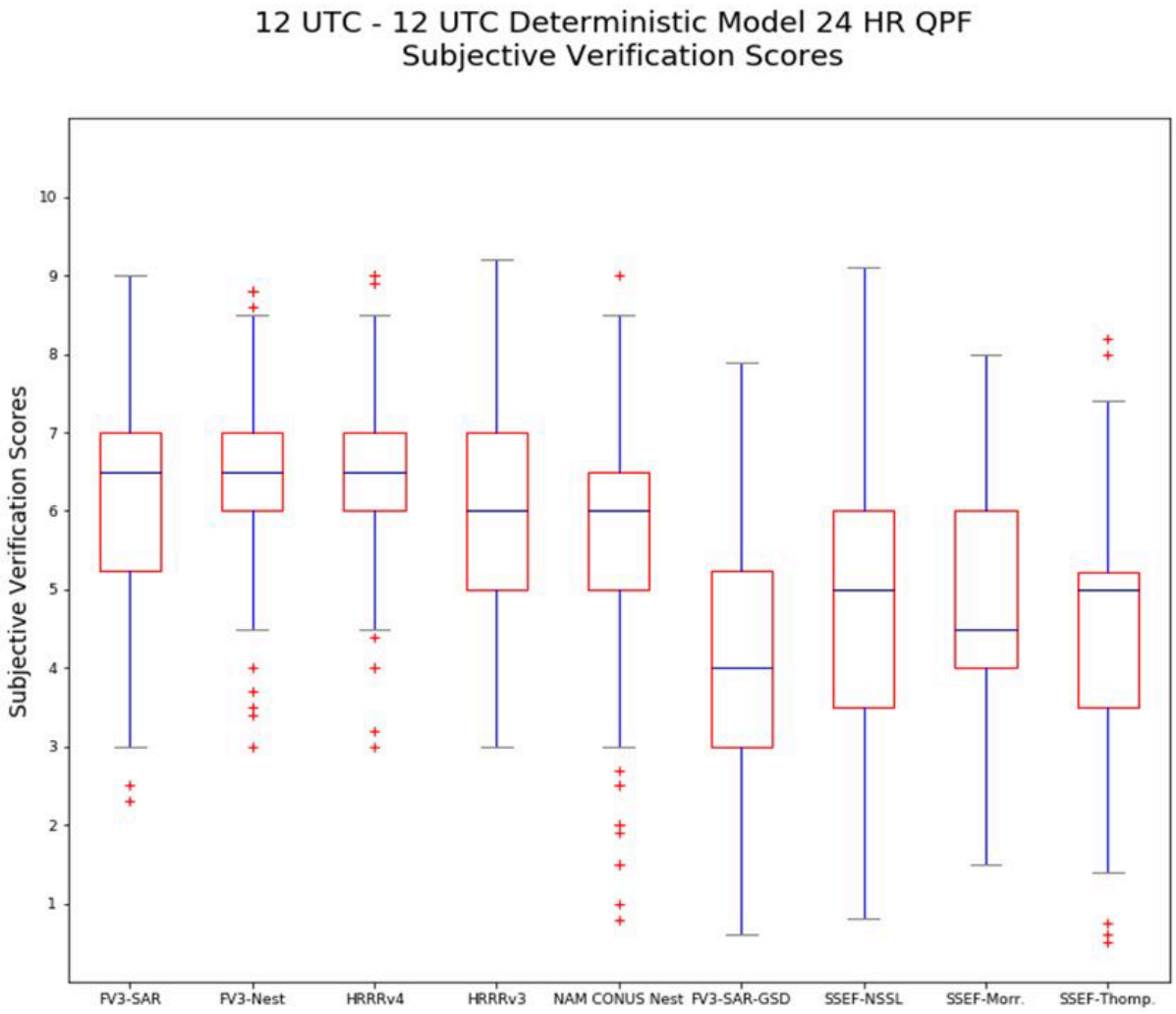

Figure 15: Box and whisker plot of all the subjective scores given for the deterministic models evaluated in the 2019 FFaIR experiment. Subjective scores were evaluating the 24 h QPF model guidance across the CONUS. All models were initialized at 0000 UTC and the 24 h forecast was from 1200 UTC to 1200 UTC.

The general consensus of the participants during the experiment was that the individual members of the OU FV3 runs (Figure 15), which only differed in the microphysics scheme used, routinely performed fairly poorly. The FV3 run from GSL performed the worst of all the deterministic models during the experiment according to the subjective scores from the FFaIR participants while the EMC FV3 runs performed comparably to the HRRRv3 and HRRRv4.

Differing from the subjective evaluation, performance diagrams suggest that every model except for the GSL FV3 and the OU FV3s provided a 24h precipitation forecast that was as good or better than the forecasts from the EMC FV3 runs.

The NAM-Nest had a higher CSI and a smaller wet bias than the EMC FV3. This was extremely noteworthy because the participants regularly commented on the NAM-Nest having a reputation for a high wet bias and the NAM-Nest is being used by

EMC as a baseline comparison for the FV3. This wet bias and low CSI was observed at all the thresholds evaluated in the experiment for the EMC FV3s.

Another takeaway from the performance diagrams was that both versions of the HRRR outperformed every FV3 CAM that was evaluated, which, as stated earlier, was not the case for the subjective verification.

1.2.3. 2020 FFaIR Experiment

The goal of the 2020 experiment was to evaluate various configurations for the FV3 from both the Environmental Modeling Center (EMC) and Global Systems Laboratory (GSL), focusing on things like timing, amount, and location of QPF.

Repeatedly discussed by the participants, both verbally and written, was the pronounced wet bias in the QPF from all the FV3 configurations. This wet bias was noted in both synoptically forced and mesoscale events but what was most concerning to the participants was the precipitation forecast for single cell convection, often referred to as “popcorn” convection (Figure 16). For instance one of the participants wrote: “The 'popcorn' storms in all of the models have too much rain compared to observations - they all have 2-3" of rain in each little blob. That makes it hard to distinguish between everyday storms and more organized threats.”

When this type of convection was being simulated by the FV3-SARs, nearly every convective cell was forecasted to be a heavy rain event. Furthermore, the structure of the “popcorn” convection looked gridded in nature.

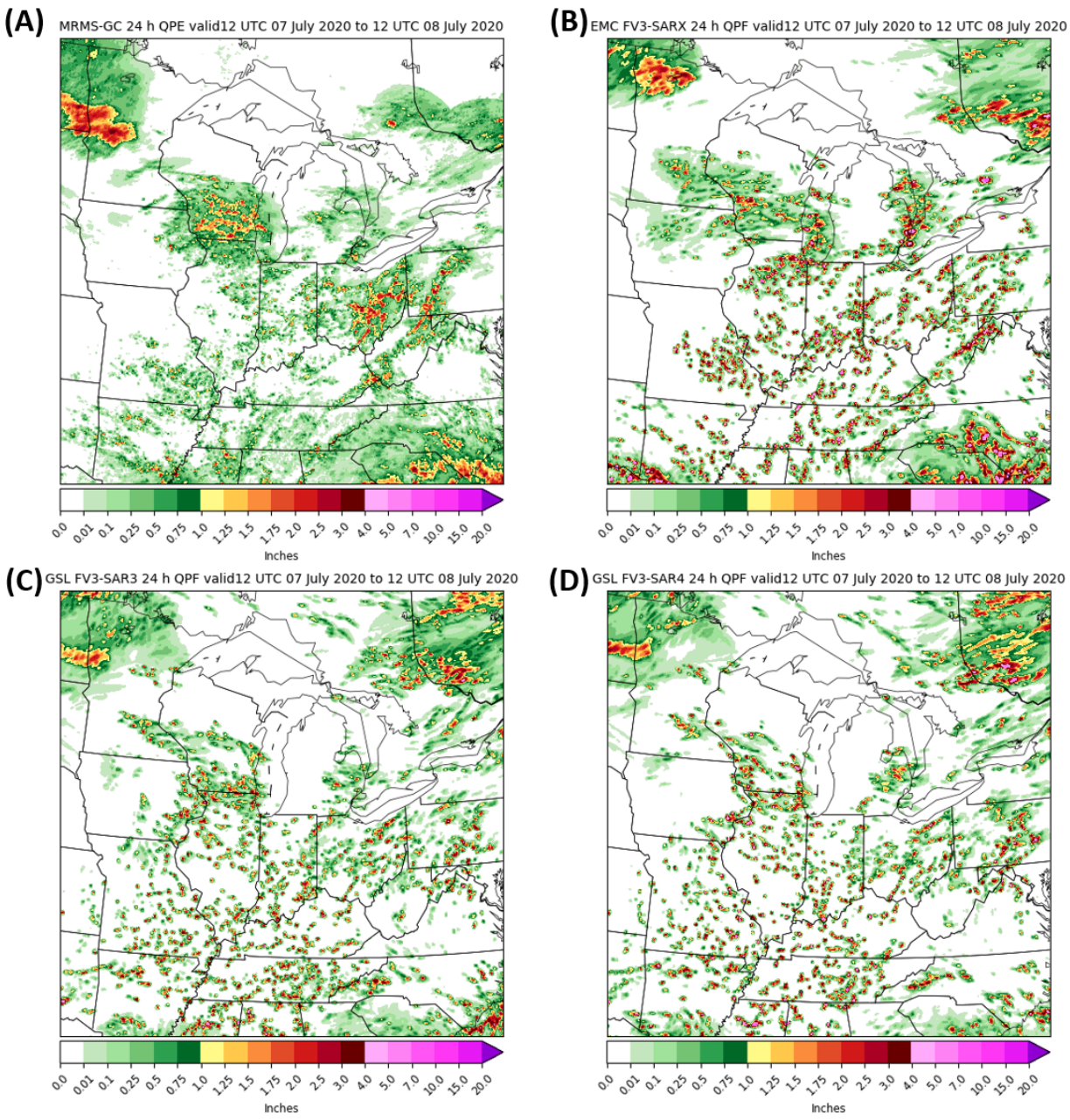

Figure 16: (A)24 h MRMS-GC QPE and 24 h QPF from (B) EMC FV3-SARX, (C) GSL FV3-SAR3, and (D) GSL FV3-SAR4 valid 12 UTC 07 July to 12 UTC 08 July 2020.

Over the course of the experiment the overall performance of the 12z run of the deterministic models was better than the 00z. All the GSL FV3 configurations were generally clustered together on the 24 h QPF grid-based performance diagrams, so it is difficult to determine if one was constantly better than the other. Overall the configurations with the same ICs and LBCs were more similar to one another than those with the same dynamics suite. This suggests that the ICs/LBCs are more influential in the model performance than the parameters forcing the dynamics.

Overall, it was noted that all FV3 configurations had a notable wet bias. Each configuration appears to be creating a gridscale convection that “rains out” all the moisture available in the grid cell column.

1.2.4. 2021 FFaIR Experiment

During the 2021 FFaIR experiment numerous versions of the FV3 were evaluated from EMC, OU, and GSL (Table 2).

Model	ICs	LBCs	DA	Domain	Micro-physics	LSM
EMC FV3-LAM (LAM)	GFS	GFS	no	CONUS	Thompson	Noah
EMC FV3-LAMX (LAMX)	GFS	GFS	no	North American	Thompson	Noah
EMC FV3-LAMDAX (LAMDAX)	Own	GFS	yes	CONUS	Thompson	Noah
GSL RRFS-dev 1 (RRFS1)	Cycled	13km FV3LAM	yes	CONUS	Thompson	RUC
OU-CAPS M0B0L0 (SSEF Cntl member)	GFS	GFS	no	CONUS	Thompson	Noah
OU-CAPS M0B2L1 (SSEF RRFS-like member)	GFS	GFS	no	CONUS	Thompson	Noah-MP
OU-CAPS M0B0L2 (SSEF HRRR-like member)	GFS	GFS	No	CONUS	Thompson	RUC
OU-CAPS M1B0L0 (SSEF WoFS-like member)	GEFS	GEFS	No	CONUS	NSSL	Noah

Table 2: Model configuration for the FV3-CAMs provided by EMC, GSL and the four members from the OU-CAPS SSEF that were evaluated deterministically.

On the over-development of single cell convection by the FV3s, referred to as popcorn convection, participants who looked at the FV3s last year stated that when it came to the EMC FV3s the over-development of popcorn convection was less pronounced, though still noticeable, than last year. They also noted a wet bias in the FV3s, though again, this appeared to be subdued compared to last year. As for the RRFS, participants commented almost daily on the excessive QPF values that were being forecasted by the model, as well as the widespread, high precipitation popcorn storms. This problem was exacerbated by the fact that often, especially with 00z cycles, the forecasted pattern did not appear correct. The SSEF members, though not commented on as much by the participants, also had the tendency to overproduce popcorn convection.

The wet bias noted by the participants was not seen in the low-end QPF amounts when performance diagrams (not shown) were analyzed for both 00Z and 12Z cycles, specifically at the half inch threshold. However, the performance diagrams at QPF thresholds of 2 inches or greater showed a clear wet bias from the FV3s that increases with each threshold at a greater rate than the wet bias seen in the NAMnest.

The transition from a dry (near-zero) bias at low thresholds to a high wet bias in the FV3-CAMs suggests that the models might be under forecasting precipitation area but over forecasting magnitude.

Overall there was a slight decrease in the wet bias in the 2021 FV3-CAMs at all precipitation thresholds compared to those in 2020. The EMC FV3 configurations in 2020 and 2021 were similar, with the exception of the ICs/LBCs being from the GFSv15 in 2020 and GFSv16 in 2021, and an increase from 60 to 65 vertical layers. It is difficult to say if the increase in vertical resolution and the updated GFS fully explain the reduced bias, as it is possible other changes were introduced between 2020 and 2021 that were unknown to the FFaIR team. Either way, the wet bias and aggressive forecasting of popcorn convection from the FV3-CAMs remains a concern of the FFaIR team, and one that must be addressed before proceeding forward with implementation of the RRFS system.

Precipitation-Rate Investigation:

Subjective verification of precipitation-rate found that the experimental models have higher, in some cases significantly higher, precipitation-rate values compared to MRMS and the HRRR.

It was found that the HRRR average maximum precipitation-rate is consistent with the maximum average from MRMS, hovering around 8 in/hr. The FV3s from EMC have a higher average maximum precipitation-rate but are generally consistent with one another, roughly staying between 20-30 in/hr throughout the forecast. However, unlike the HRRR and MRMS, the diurnal cycle was notable in the maximum averages from the EMC FV3s, peaking between 18z and 22z for both the 00z and 12z model initializations. A similar diurnal cycle was in the RRFS average maximum precipitation-rate, with the average maximum precipitation-rate ranging from 40-60 in/hr for the 00z model run and 35-55 in/hr for the 12z run. Additionally, for the 00z RRFS run, there are noticeably higher precipitation-rates at model initialization compared to the rest of the model forecast hours. This could suggest shortcomings in the radar DA processes used for the RRFS.

1.2.5. 2022 FFaIR Experiment

The RRFS configurations provided by EMC, GSL, and CAPS for this experiment were designed to be interconnected. RRFSp1 provided the initial conditions for the RRFSp2 at 18 UTC and the baseline for the GEFS perturbations to re-center around, which were used to start the EnKF “RRFSDAS” ensemble. Six hours of cycling with the RRFSDAS, which provided flow-dependent information in the EnVar cost function for a hybrid analysis, then followed. The RRFSp2 is the central state of the RRFSDAS, is used to

recenter the EnKF ensemble mean each hour, and also serves as the control member for the RRFS ensemble (referred to as RRFSe during the experiment). The first 8 members of the RRFSDAS were perturbed to create the members of the RRFSe, which also included the RRFSp2 as a member. The RRFSe mean was used as the initial conditions for RRFSp3-8, which differ among themselves via their parameterizations.

Finally, like the RRFSe, the second ensemble evaluated, referred to as CAPS RRFSe, used members of the RRFSDAS to initialize its members. RRFSe had 9 members (including RRFSp2) while CAPS RRFSe had 11 (including RRFSp3) members. An additional difference was that RRFSe had no mixed physics while CAPS RRFSe had mixed physics.

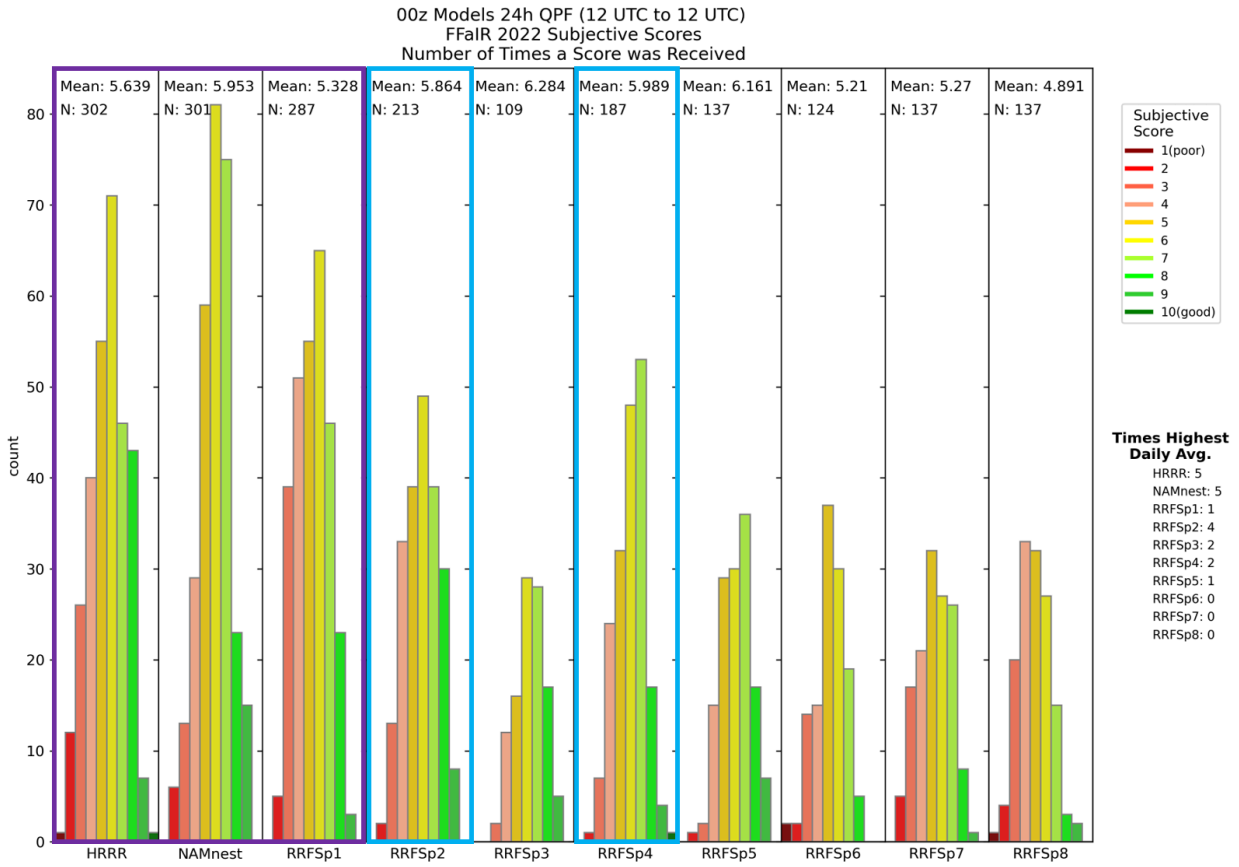

Figure 17: Results from the subjective verification for 24 h QPF for 00z model initialization showing number of times each model received a score from 1 (dark red) to 10 (dark green) during the duration of FFaIR. The number of scores received (N) and the mean score for each model is plotted along the top. Outlined in dark purple are the models in Group 1, outlined in blue is Group 2, and the ones not outlined are in Group 3. On the right, below the legend, is the number of times each model had the highest daily average score.

Echoing what has been noted in previous FFaIR Experiments, participants once again commented on the over abundance of weakly forced (AKA “popcorn”) convection, along with a wet bias at higher thresholds (Figure 17). They often noted that the RRFS was wetter than the NAMnest. Furthermore, when it came to the weakly forced convection, they were hesitant to trust the evolution of storms since the RRFS kept

these as isolated, strong storms rather than clustering them into more realistic areas of storms.

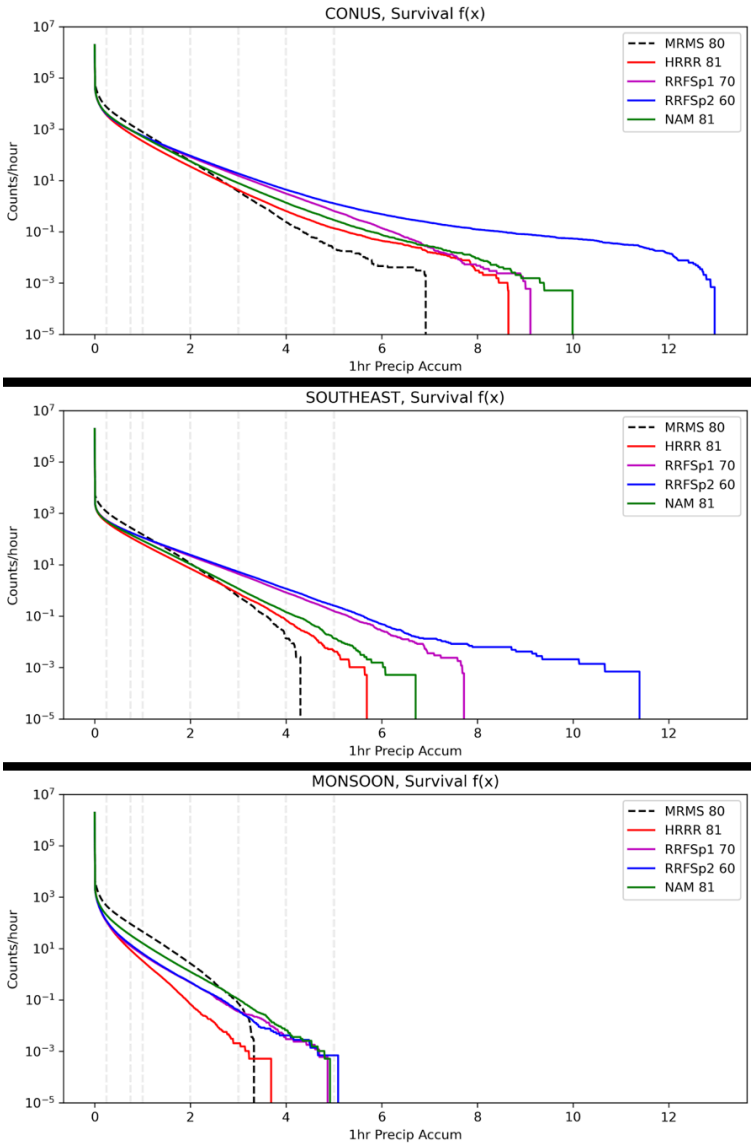

Figure 18: Hourly QPF survival function for the Testbed Season over the CONUS (top), the southeast (middle), and the southwest (bottom), comparing MRMS (dashed), HRRR (red), NAMnest (green), RRFSp1 (purple), and RRFSp2 (blue). On the y-axis is the counts per hour and on the x-axis is the 1 hr precipitation accumulation.

Focusing on the CONUS (Figure 18), all the models have a wet bias at higher thresholds, while slightly under forecasting the coverage of hourly totals <=1.5 in comparison to MRMS. Above 1.5 in, RRFS runs begin to differ from the operational models and MRMS, having greater coverage of hourly precipitation totals ranging from 1.5 to ~6.75 inches. The differences in the coverage of hourly rainfall between the RRFS prototypes versus the operational models and MRMS are more pronounced when the domain is narrowed to the southeast. At about 1.25 in, the RRFS runs coverage exceeds MRMS without their curve ever resembling the MRMS curve. On the other hand, the HRRR and NAMnest have a survival curve that is more similar to

MRMS, though both do begin to have a wet bias around 2.5 in. In contrast to both the CONUS and southeast, across the southwest (bottom image) all the models have a dry bias for thresholds below about 3.25 in, with the NAMnest, RRFSp1, and RRFSp2 all performing similarly to one another.

When looking at rainfall diurnally (Figure 19), both RRFS models have convective initiation too early (by roughly 2 hours). Additionally, comparing coverage to intensity showed that although the average hourly rainfall coverage is more similar to MRMS for RRFS runs than the operational models, when focusing on intensity both models have higher averages than MRMS for all hours of the day. This suggests that the RRFS models are more likely to simulate rainfall than the operational models but when it does rain, the intensity is too high.

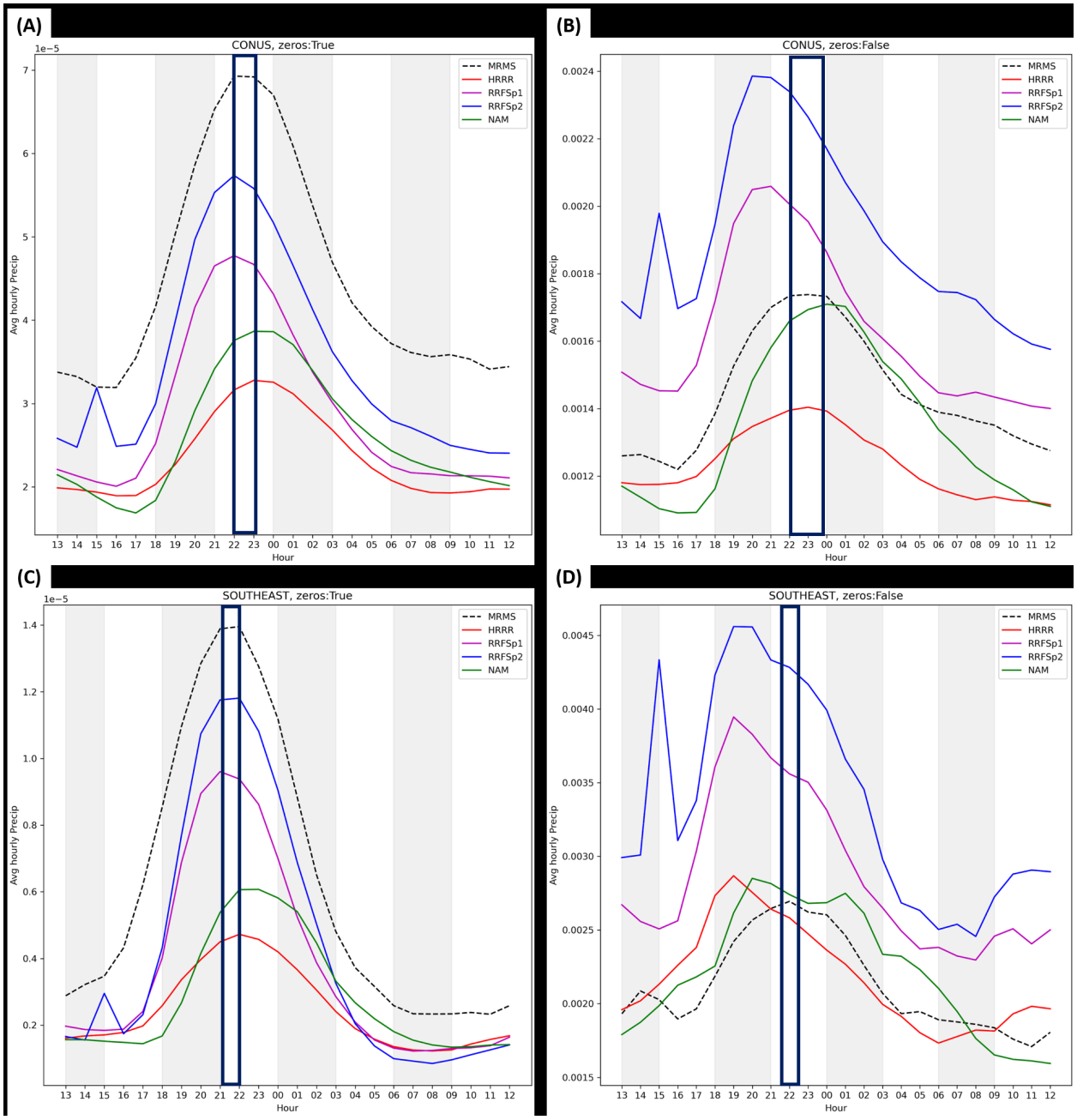

Figure 19: Diurnal analysis of hourly QPF for the Testbed Season over the CONUS (top) and the southeast (bottom). (A) and (C) show coverage (zeros included in average) of hourly precipitation. (B) and (D) show the intensity (zeros NOT included in average). The black rectangles indicate the time in which MRMS is at its maximum; the width of the rectangle varies among the plots based on the duration of the MRMS maximum. Note: the average hourly precipitation is scaled differently for each of the images.

1.2.6. 2023 FFaIR Experiment

In 2023, the RRFS transitioned to fully cycled data assimilation and split the FFaIR season by running over CONUS (7 weeks) then over a North American domain (3 weeks). The diurnal cycle of the RRFS has a positive intensity bias relative to MRMS,

most prominent in the early afternoon when convection is initiating. While the intensity bias is large, it has a better depiction of the diurnal cycle than other models for both the 00 and 12 UTC initializations. While the areal coverage of precipitation is larger than other models, coverage at higher amounts is driving the bias and improving the amplitude of the diurnal cycle.

To further show how the bias manifests, the 1"+ coverage peaks about 1 and 2 hours prior to MRMS (00 and 12 UTC) respectively (Figure 20). This behavior differs from the HRRR and NAMnest, which peak later than MRMS. Again, comparing coverage to intensity showed that although the average hourly rainfall coverage is more similar to MRMS for the RRFS runs than the operational models; when focusing on intensity, both models have higher averages than MRMS for all hours of the day. The survival function depiction for RRFS also shows differences with the operational models in that the 12 UTC initialization produces fewer extremes than its 00 UTC counterpart. In this depiction, the coverage bias up to the ~5.5" exceedance threshold is greater on average than for the operational models. We can summarize that the RRFS has larger areas of heavier precipitation, which are more intense than the operational models early in the diurnal cycle.

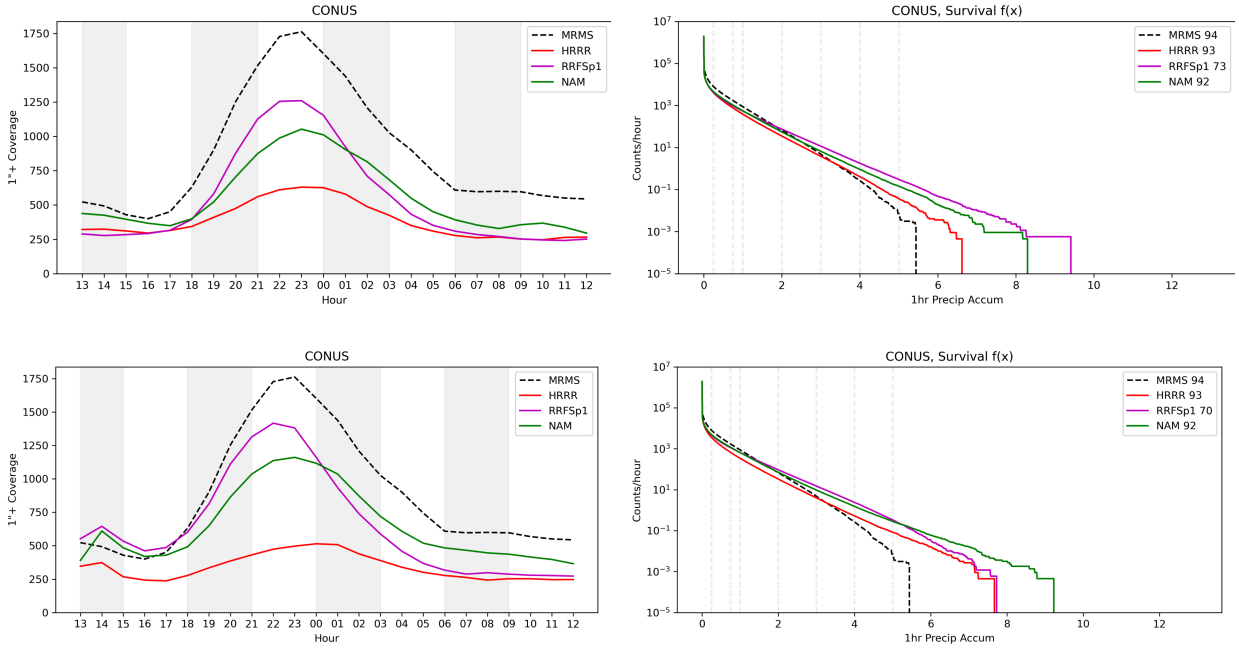

Figure 20: The diurnal cycle of 1hr 1"+ coverage (left) and survival function depiction of 1hr rainfall (right) for 00z initialization (top) and 12z initialization (bottom).

After examining the conus maximum hourly-maximum precipitation rates, FFaIR 2023 noted lower values relative to previous years. However, the perturbed single physics members still produced values in excess of 50-70 in/hr.

Appendix C — Costs of Past Major Dynamical Core Development Efforts

To assist in estimating time and resources required to execute the work required to update the FV3, or switch to MPAS, we have obtained estimates from similar efforts in the past at EMC. Numbers are approximate as not all historical data are available.

Task	Time (years)	FTE (average over the time period)	Total Cost ($200k per FTE does not include HPC)
1. Convert hydrostatic Eta model to Nonhydrostatic Mesoscale Model (NMM) 2.Conversion from Eta coordinate to hybrid sigma-p 3. Inclusion in WRF	10 years (1996-2006)	10	$20 M
1. Convert WRF-NMM from E grid to B grid 2. Add global capability 3. Convert WRF-NMM to ESMF framework, i.e. the Nonhydrostatic Multiscale Model on the B grid	4 years (2007-2011)	10	$8 M
1. Develop Semi-Lagrangian dynamics for global spectral model 2. Testing and evaluation 3. Implementation in operational GFS	6 years (2009-2015)	30	$36 M

Appendix D — MPAS Computational Performance

During the NGGPS testing in 2016, MPAS ran ~2x slower than FV3 for global configurations (Ji 2016). Much of that difference is due to a larger time step FV3 can employ for non-convection allowing grids (e.g., $\Delta x \geq 12$ km) facilitated by its vertical Lagrangian coordinate. With the recent doubling of the vertical levels in the GFS, the maximum stable time step has had to be reduced by as much as 40% (despite the Lagrangian coordinate). Therefore the relative performance noted in 2016 may not be true for current application configurations and requirements

Similarly, for convection permitting resolution in the RRFSv1, the FV3 dynamical core is unable to use a time step that is 2x of MPAS because the much larger vertical velocities (and as important, the vertical divergence) collapse the Lagrangian surfaces together much more rapidly than simulations using coarser grids. Avoiding the collapse between the vertical levels then limits the time step for the dynamics. The incorporation of two and three moment microphysical parameterizations (e.g., Thompson and NSSL 2-moment), required when assimilating radar reflectivity to reduce the O-B bias, require the physics time step to be much more comparable to the dynamics/transport time step given the strength of the updrafts and downdrafts[21].

During 2023, NSSL has run three different configurations of MPAS from 0000 UTC over CONUS running on GSL's Jet facility (xJet). Currently MPAS runs about 10% slower than the FV3. Table 1 below shows the relative performance of CONUS-scale forecasts, with similar (but not identical) physics using WRF, MPAS, and FV3 on a 3 km grid. The CONUS grids are also not identical either due to the different projections used, but they are similar in the area they cover. NSSL's experience is that MPAS is close to the speed of the WRF model, consistent with NCAR's experience. Obviously, further computational testing is needed for regional CAM simulations, but it appears that MPAS's speed relative to the FV3/RRFS is not a significant issue.

Model	**Wall Clock** (hh:mm)	**Grid points** (nx*ny, nz)	**Macro (Physics) Time Step** (NSSL 2-Moment microphysics)
FV3	2:46	1761280 x 60	36 sec (ksplit/nsplit = 5,2)
WRF	3:13	1584000 x 60	15 sec
MPAS	2:56	1894063 x 59	25 sec (dynamics substep = 3)

Table 1: Comparison of timings for a 48-hour forecast initialized at 0000 UTC 24 March 2023.

[21] The GFDL FV3 SHiELD applications avoid this problem by splitting the "fast" microphysical processes from the slower process within their GFDL microphysics, placing a subset of these on the dynamics time step (Zhou and Harris 2022). This enables a larger "physics time step" which increases integration efficiency. For more complex microphysical schemes, too many processes would need to be split to the dynamics time step which would then trade a larger time step for more costly computations on the dynamics time step.

Appendix E — Acronyms

3DEnVar	Three-Dimensional Ensemble Variational
ARW	Advanced Research WRF
CAM	Convection-Allowing Model
CAPE	Convective Available Potential Energy
CAPS	Center for Analysis and Prediction of Storms
CCPP	Common Community Physics Package
CESM	Community Earth System Model (NCAR)
CFL	Courant–Friedrichs–Lewy condition
CM1	Cloud Model 1
CONUS	Contiguous United States
DA	Data Assimilation
DAS	Data Assimilation System
EDMF	Eddy-Diffusivity/Mass-Flux (PBL scheme)
EMC	Environmental Modeling Center
EnKF	Ensemble Kalman Filter
EPIC	Earth Prediction Innovation Center
FV3	Finite-Volume Cubed-Sphere dynamical core
GEFS	Global Ensemble Forecast System (NOAA)
GPU	Graphics Processing Unit
GFDL	Geophysical Fluid Dynamics Laboratory
GFS	Global Forecast System
GSL	Global Systems Laboratory
HPC	High performance computing
HRRR	High-Resolution Rapid Refresh
HMT FFaIR	Hydrometeorological Testbed Flash Flood and Intense Rainfall Experiment
HWT SFE	Hazardous Weather Testbed Spring Forecasting Experiment
IC	Initial Conditions
IDSS	Impact-Based Decision Support Services
JEDI	Joint Effort for Data assimilation Integration
LAM	Limited Area Model
LBCs	Lateral Boundary Conditions
LSM	Land Surface Model
MMM	NCAR Mesoscale & Microscale Meteorology Laboratory
MPAS	Model for Prediction Across Scales
MRMS	Multi-Radar/Multi-Sensor
MYNN	Mellor-Yamada Nakanishi and Niino (PBL scheme)
NAM	North American Mesoscale forecast system
NCAR	National Center for Atmospheric Research
NCEP	National Centers for Environmental Prediction
NGGPS	Next Generation Global Prediction System
NOAA	National Oceanic and Atmospheric Administration
NSSL	National Severe Storms Laboratory
NUOPC	National Unified Operational Prediction Capability
NWP	Numerical weather prediction

NWS	National Weather Service
OAR	Oceanic and Atmospheric Research
PBL	Planetary Boundary Layer
QPF	Quantitative Precipitation Forecast
R2O	Research to Operations
RAP	Rapid Refresh (system)
RDHPCS	Research and Development High Performance Computing System
RRFS	Rapid Refresh Forecast System
RUC	Rapid Update Cycle (NOAA LSM)
SAR	Stand-Alone Regional
UFS	Unified Forecast System
UPP	Unified Post Processor (NOAA)
WoFS	Warn on Forecast System
WRF	Weather Research and Forecasting Model