Assessment of FV3 vs WRF-ARW multi-physics ensemble performance available in the CLUE during the HWT-SFE 2018

Report compiled by:

Lindsay Blank[1], Jeff Beck[2], Jamie Wolff[1], and Michelle Harrold[1]

[1]National Center for Atmospheric Research (NCAR) Research Applications Laboratory/Joint Numerical Testbed Program and Developmental Testbed Center
[2]Cooperative Institute for Research in the Atmosphere (CIRA)/Affiliated with NOAA/ESRL/GSD and Developmental Testbed Center

Table of Contents

Introduction

Over the last several years at the Hazardous Weather Testbed Spring Forecasting Experiment (HWT-SFE), an effort to coordinate the contributed model output from participating groups around a unified setup (e.g., WRF versions, domain size, vertical levels and spacing, etc.) was undertaken to create a super-ensemble called the Community Leveraged Unified Ensemble (CLUE). The careful coordination and construction of CLUE allowed for meaningful comparisons among a variety of members to be performed. With a convection-allowing ensemble planned for operational implementation in the near future, it is critical to investigate key scientific questions related to informing the best configuration strategies for producing such an ensemble based on an evidence-driven approach.

CLUE dataset from the 2018 HWT- SFE is presented in this document. For more background on the motivation to collaborate with the HWT-SFE and specific information regarding the verification approaches used, please reference the 2016 final report located at https://dtcenter.org/eval/ensembles/hwt_collab/RE5_HWT_report_FINAL_CLUE2016.pdf. For more information on HWT-SFE 2018, the program overview and operations plan can be found at https://hwt.nssl.noaa.gov/sfe/2018/docs/HWT_SFE2018_operations_plan.pdf.

CLUE 2018 Dataset

The CLUE is a super-ensemble comprised of subset members contributed by a number of collaborating organizations, include NOAA/NWS/NSSL, University of Oklahoma Center for Analysis and Prediction of Storms (CAPS), NOAA/ESRL/GSD, NCAR, and NOAA's Geophysical Fluid Dynamics Laboratory (GFDL). Two particular subsets and two deterministic members were of interest for this analysis: the FV3 11-member ensemble, the WRF-ARW multi-physics 10-member ensemble, the nssl-fv3 (fv3-thomp-caps) deterministic member, and the gfdl-fv3 (fvgfs) deterministic member. The physics suites for each 2018 ensemble subset and deterministic member used in this analysis are presented in the tables below. Of note, the multi-physics ensemble subset includes radar assimilation while the fv3-physics ensemble does not.

Table 1: Physics suite description for FV3 ensemble subset.

FV3 Physics			
Member	**MP**	**LSM**	**PBL**
fv3-phys01	Thompson	Noah	MYNN-SA
fv3-phys02	Thompson	Noah	MYNN

fv3-phys03	Thompson	Noah	YSU-SA
fv3-phys04	Thompson	Noah	YSU
fv3-phys05	Thompson	Noah	EDMF
fv3-phys06	NSSL	Noah	MYNN-SA
fv3-phys07	NSSL	Noah	MYNN
fv3-phys08	NSSL	Noah	YSU-SA
fv3-phys09	NSSL	Noah	YSU
fv3-phys10	NSSL	Noah	EDMF
fv3-phys11*	Thompson	Noah	MYNN-SA

**uses the SA-SAS cumulus scheme instead of Tiedtke*

Table 2: Physics suite description for WRF-ARW multi-physics subset.

Mixed Physics + IC/BC perturbations			
Member	**MP**	**LSM**	**PBL**
core01	Thompson	Noah	MYJ
core02	Thompson	RUC	MYNN
core03	NSSL	Noah	YSU
core04	NSSL	Noah	MYNN
core05	Morrison	Noah	MYJ
core06	P3	Noah	YSU
core07	NSSL	Noah	MYJ
core08	Morrison	Noah	YSU
core09	P3	Noah	MYNN
core10	Thompson	Noah	MYNN

Table 3: Physics suite description for FV3 deterministic members.

FV3 deterministic Members			
Member	**MP**	**LSM**	**PBL**

nssl-fv3	Thompson	Noah	MYNN
gfdl-fv3	GFDL-6cat	Noah	YSU

Verification Results

Surface Variables

Temperature

Looking a 2-m temperature bias, the multi-physics ensemble members bias ranged between -1 - 1ºC throughout the forecast period (Fig. 1a). A value of 0 means the forecast is unbiased. The FV3 physics ensemble members exhibited a cool bias throughout the forecast period. The two deterministic FV3 members also exhibited a cool bias throughout. In terms of bias-corrected root mean square error (BCRMSE, lower is better), the multi-physics ensemble had a lower BCRMSE than both the FV3 physics ensemble members and the two FV3 deterministic members (Fig. 1b). The peak error for all members occurred between forecast lead hours 18 - 22.

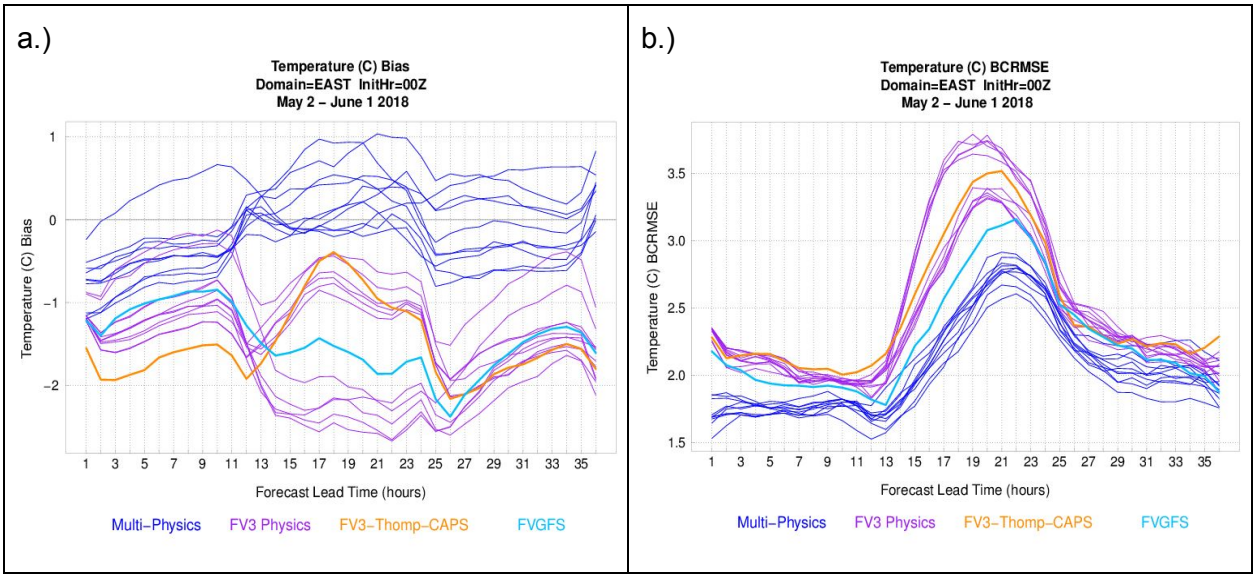

***Figure 1**: (a) Bias and (b) BCRMSE time series plots of 2-m temperature (°C) for each individual ensemble member and the two deterministic members aggregated across the eastern half of the CONUS domain for all available forecasts during the experiment. The multi-physics members are in blue, the FV3 physics members are in purple, the nssl-fv3 deterministic member is in orange, and the gfdl-fv3 deterministic member is in teal.*

Spread/skill curves (Fig. 2) show that the FV3 physics ensemble contains less spread than the mixed physics ensemble subset for all forecast lead times except 14-23 hours, during which time there is enhanced spread, likely due to convective activity. In addition, the mixed physics

ensemble generally has less error (RMSE) than the FV3 physics ensemble subset for 2-m temperature, confirming the results found in Figure 1. It can also be seen that neither ensemble subset has a spread/skill ratio close to one, signifying that they are both under-dispersive.

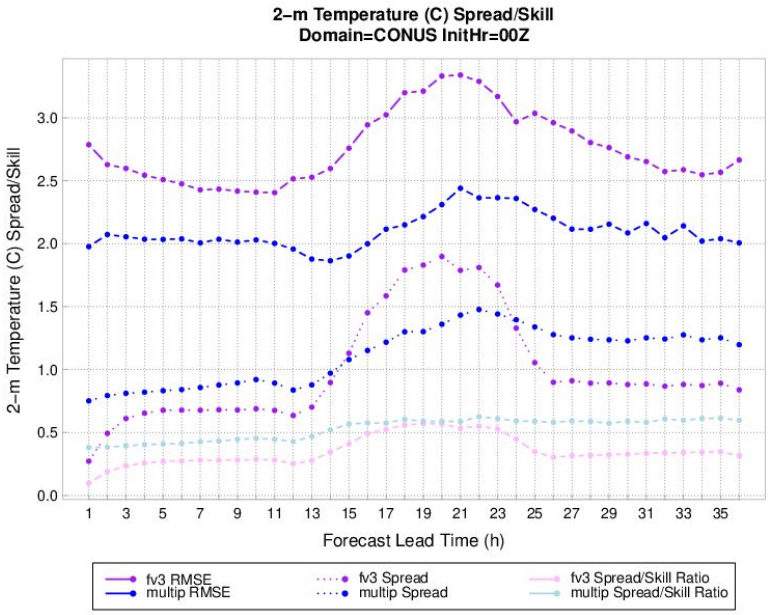

Figure 2*: Average CONUS, 2-m temperature (°C) spread (dotted), skill (RMSE; dark solid), and spread/skill ratio (light solid) lines for the multi-physics (blue) and FV3 physics ensemble (purple) subsets for 00Z initializations out to 36-hr lead times. An ideal spread/skill ratio is equal to one.*

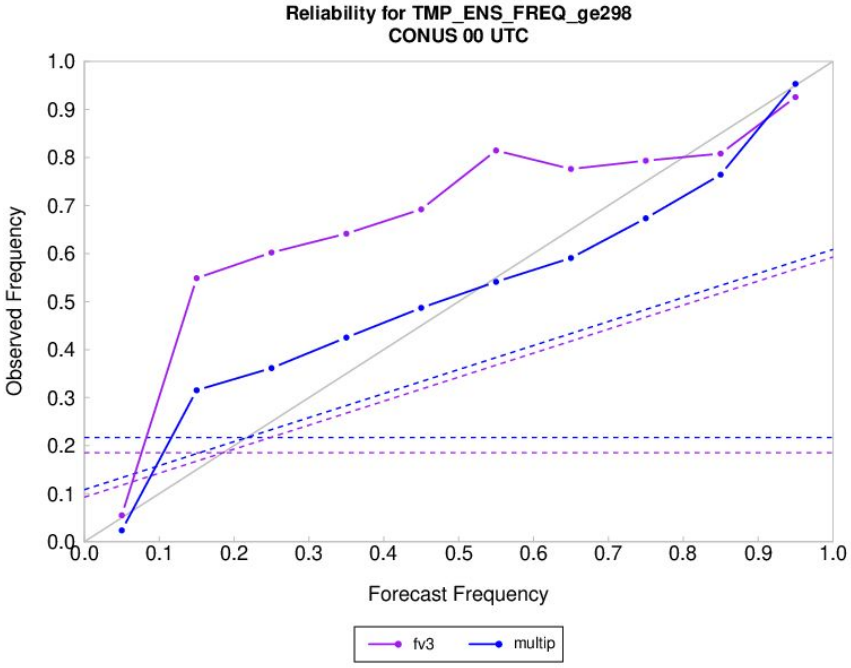

Figure 3: *Reliability plot of CONUS, 2-m temperature forecasts of greater than or equal to 298°K for the multi-physics (blue) and FV3 physics ensemble (purple) subsets for 00Z initializations, averaged spatially*

over all forecast lead times. The horizontal (diagonal) dashed lines represent the climatology of the event/limit of no resolution (limit of skillfulness) for both ensemble datasets. Ideal reliability lies along the one-to-one diagonal.

Analyzing the reliability of the FV3 physics ensemble subset, most forecast frequencies are under-forecast for the FV3. In other words, the predicted probability of the 2-m temperature equaling or exceeding 298°K was too low based on the observed frequencies, a curious finding, since this characteristic is normally reserved for over-dispersive ensembles. On the other hand, the mixed-physics ensemble is much closer to the diagonal line in the reliability diagram, indicating that forecast probabilities are more or less well calibrated with the observed frequencies of the event (forecasts of >= 298°K).

As with the spread-skill plot (Fig. 2), the probability integral transform (PIT) histograms for 24-hr forecasts also show an under-dispersiveness for both the multi-physics and FV3 physics ensemble subsets (Fig. 4). The multi-physics ensemble is more or less evenly under-dispersive, while there is a distinct negative bias to the FV3 physics ensemble subset. In other words, the FV3 physics ensemble is predicting colder temperatures at the 24-hr lead time more often than warmer temperatures (many more observations are falling in the highest rather than the lowest rank).

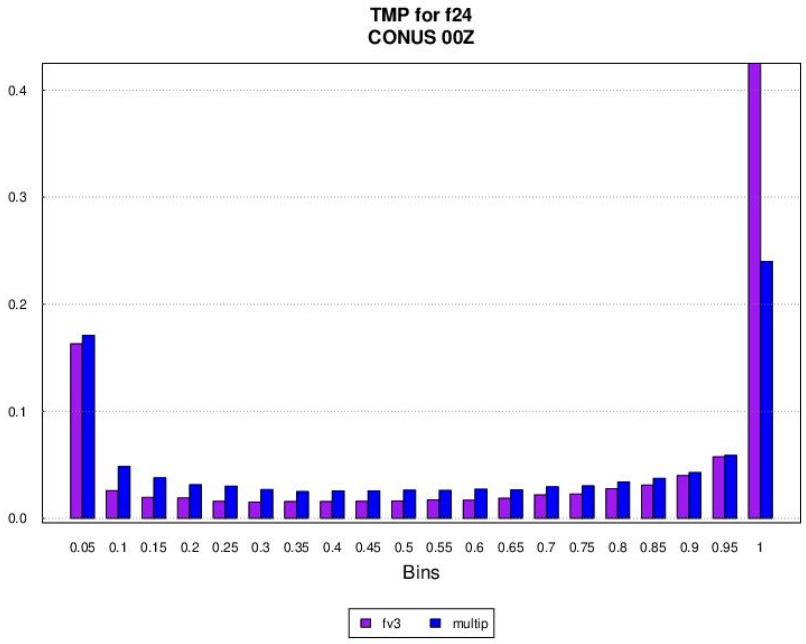

Figure 4: *Probability integral transform (PIT) histograms of 2-m temperature 24-hr forecasts for 00Z initializations, averaged over CONUS for both the mixed physics (blue) and FV3 physics ensemble (purple) subsets. An ideal PIT histogram for these two ensembles would be a flat 0.05 distribution.*

Dew Point Temperature

In terms of bias, the FV3 physics ensemble exhibited a mean error range between -1 - 1°C throughout the forecast period (Fig. 5a). The multi-physics ensemble had a warm bias between just above 0 to near 2°C throughout the forecast period. One of the multi-physics ensemble members consistently had a warmer bias than the rest of the multi-physics members throughout. Regarding BCRMSE, the multi-physics ensemble subset had a lower BCRMSE overall throughout the forecast period than the FV3 physics ensemble (Fig. 5b). There was a peak in BCRMSE around forecast hours 20 - 24 for all members. The fvgfs deterministic member generally had the highest BCRMSE throughout the forecast period. Note, there was no data from the fv3-thomp-caps deterministic member for dew point temperature throughout the experiment.

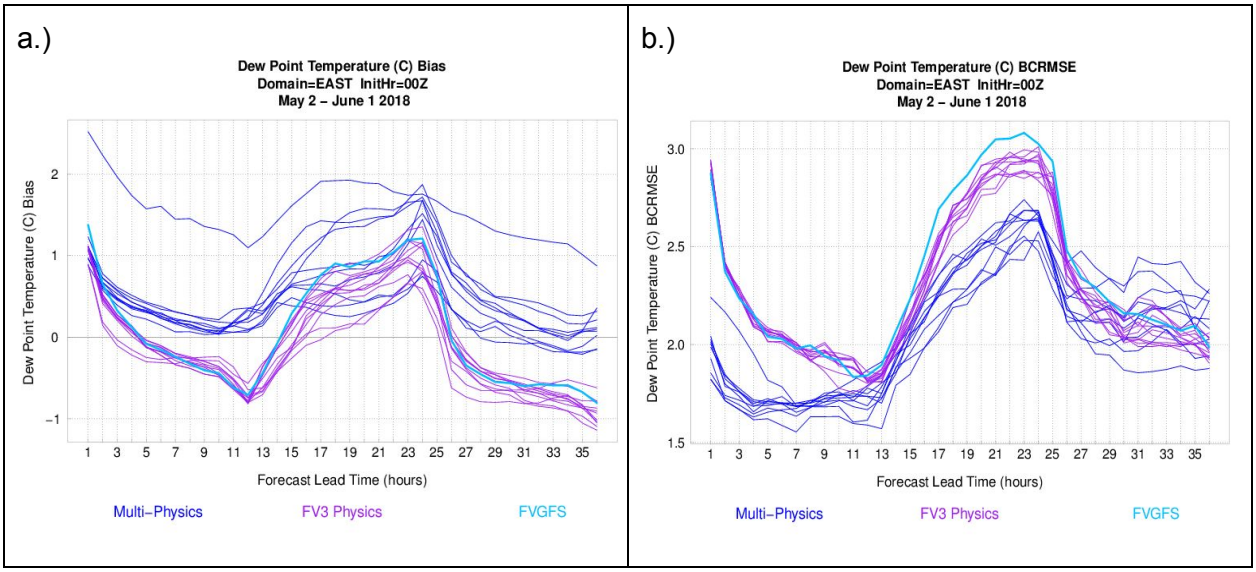

Figure 5: *Same as in Fig. 1, except for 2-m dew point temperature.*

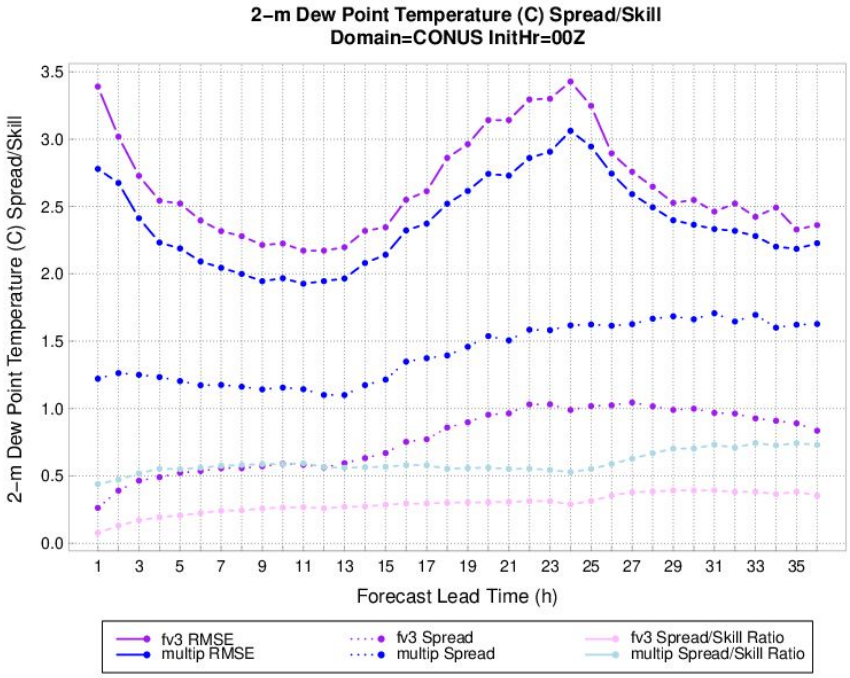

Figure 6: *Same as in Fig. 2, except for 2-m dew point temperature.*

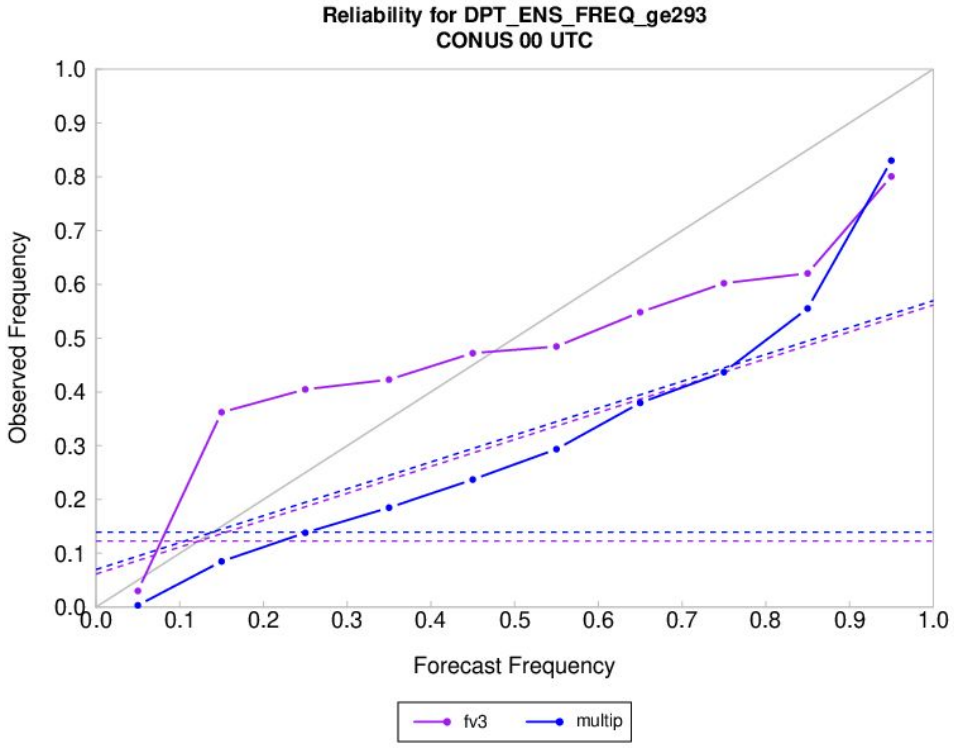

Figure 7: *Same as in Fig. 3, except for 2-m dew point temperature >= 293°K.*

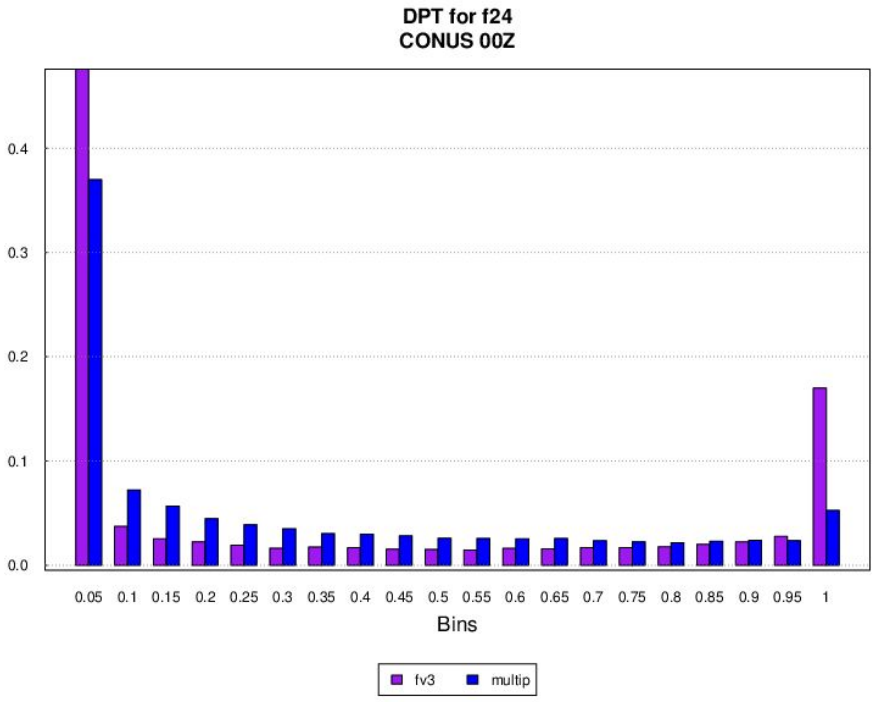

Figure 8: *Same as in Fig. 4, except for 2-m dew point temperature.*

The spread-skill plot for the FV3 physics and multi-physics ensembles for 2-m dew point temperature (Fig. 6), show an under-dispersiveness with the spread/skill ratio for both ensembles; however, it improves with forecast lead time. Errors (RMSE) increase during the daytime forecast lead times, but then decrease at night, while spread continually increases with time, helping increase the spread/skill ratio later in the forecast. As was the case for temperature, the multi-physics ensemble has the better spread/skill ratio, with smaller error and larger spread. By the end of the forecast, the multi-physics ensemble spread explains three quarters of the error (spread/skill ratio of 0.75).

The reliability diagram for dew point temperature forecasts of >= 293°K (Fig. 7) show that the multi-physics ensemble is overconfident for all forecast frequencies, with too much certainty being attributed to this events across all probabilities, a result of the under-dispersiveness of the ensemble. The FV3 physics ensemble is similar to the multi-physics ensemble for higher forecast frequencies, but at lower forecast probabilities, the model is under-forecasting, predicting probabilities that are too low for the event.

Both ensemble subsets are under-dispersive when analyzing the PIT histogram for dew point temperature (Fig. 8), but the FV3 physics ensemble is slightly more so than the multi-physics ensemble. In addition, the ensembles have a positive bias, with the FV3 physics ensemble having a stronger bias than the multi-physics subset. This positive bias also helps to explain the over-forecasting that is seen in the reliability diagram, with higher forecast frequencies than observed.

Wind Speed

The multi-physics ensemble members mean error varied from one another throughout the forecast period versus the FV3 physics ensemble members which were more clustered (Fig. 9a). Both deterministic FV3 members were within the FV3 physics ensemble member envelope (for the most part). An overall high bias existed throughout the forecast period for all members. For BCRMSE, all members across all ensemble subsets and the two deterministic members followed the same trend across the forecast period (Fig. 9b). The multi-physics ensemble subset yielded higher overall BCRMSE values and more variety across members. The FV3 physics ensemble members exhibited clustered behavior.

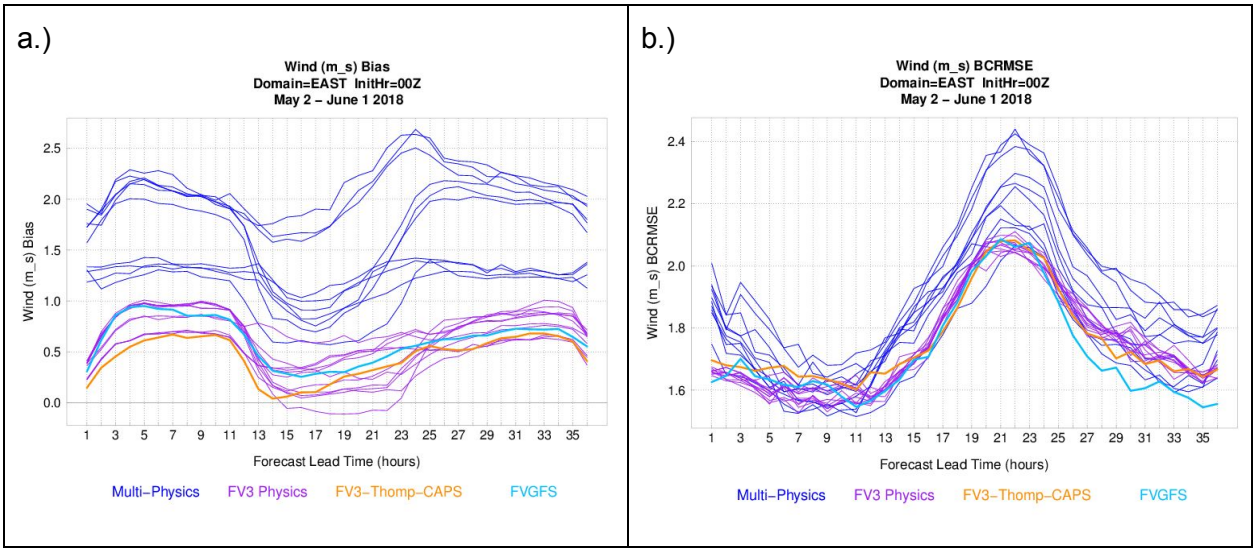

Figure 9: *Same as Fig. 1, except for 10-m wind speed (m/s).*

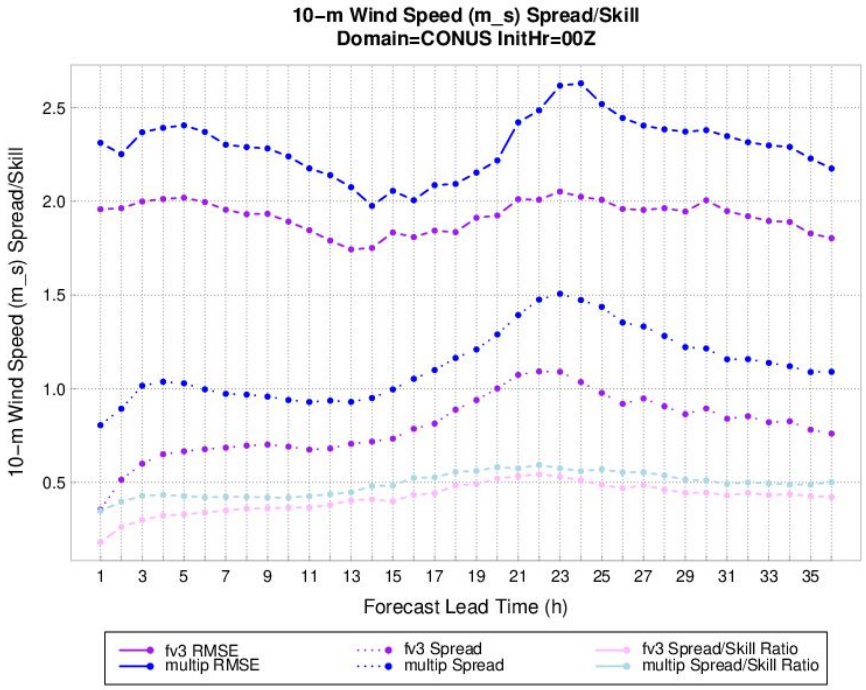

Figure 10: *Same as Fig. 2, except for 10-m wind speed (m/s).*

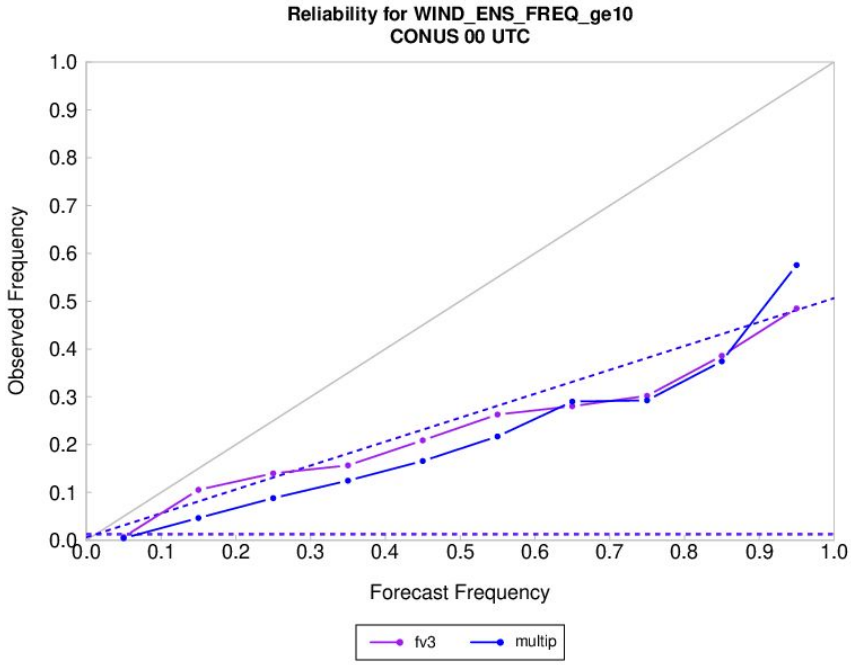

Figure 11: *Same as Fig. 3, except for 10-m wind speed >= 10 m/s.*

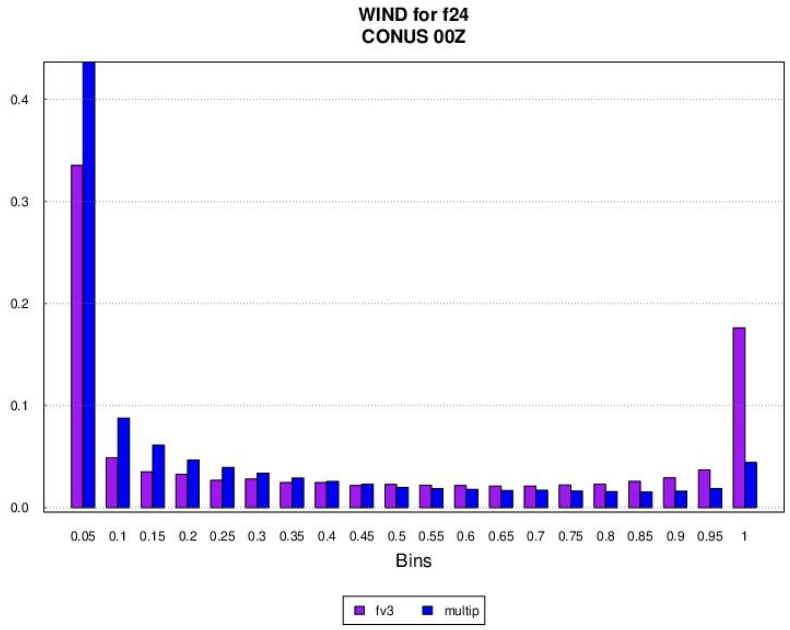

Figure 12: *Same as Fig. 4, except for 10-m wind speed (m/s).*

For 10-m wind speed forecasts, the two ensemble subsets were about equally under-dispersive based on the spread/skill plot (Fig. 10), as their spread/skill ratios were fairly similar. Overall, the multi-physics ensemble had more error (RMSE) and spread than the FV3 physics ensemble, but both ensembles peaked in error/spread around 00 UTC, near the height of convective activity for the day.

Forecasts of winds greater than or equal to 10 m/s were found to be overly-confident for both ensemble subsets (Fig. 11). Each of the ensembles produced probability forecasts that were mostly below the skill line, and, therefore, it cannot be considered that either ensemble can reliability forecast this event. It should be noted that the climatology of winds greater than or equal to 10 m/s for these simulations is very low (as shown in Figure 11). Therefore, the difficulty of accurately forecasting such a rare event should be taken into account for such an analysis.

The PIT histograms (Fig. 12) were again under-dispersive for wind speed forecasts for both ensemble subsets. This finding is in agreement with the lack of spread seen in the spread/skill plot (Fig. 10) for both ensembles. While the FV3 physics ensemble is only slightly positively biased toward higher wind speed forecasts, the mixed-physics ensemble is strongly biased in this direction. These biases shed light on the strong over-forecasting of winds equal to or greater than 10 m/s, as was seen in the reliability plot (Fig. 11).

1-Hour Accumulated Precipitation (≥ 2.54 mm)

The multi-physics and FV3 physics ensemble members exhibited similar behavior for the Gilbert Skill Score (GSS; higher is better) after forecast hour 8 (Fig. 13a.). The multi-physics ensemble subset GSS sharply decreased during the early forecast hours while the FV3 physics ensemble subset and fvgfs deterministic member increased during those same early forecast hours. The deterministic fvgfs had the overall lowest GSS score throughout the forecast period. Note, the fv3-thomp-caps deterministic members did not have accumulated precipitation data for the experiment.

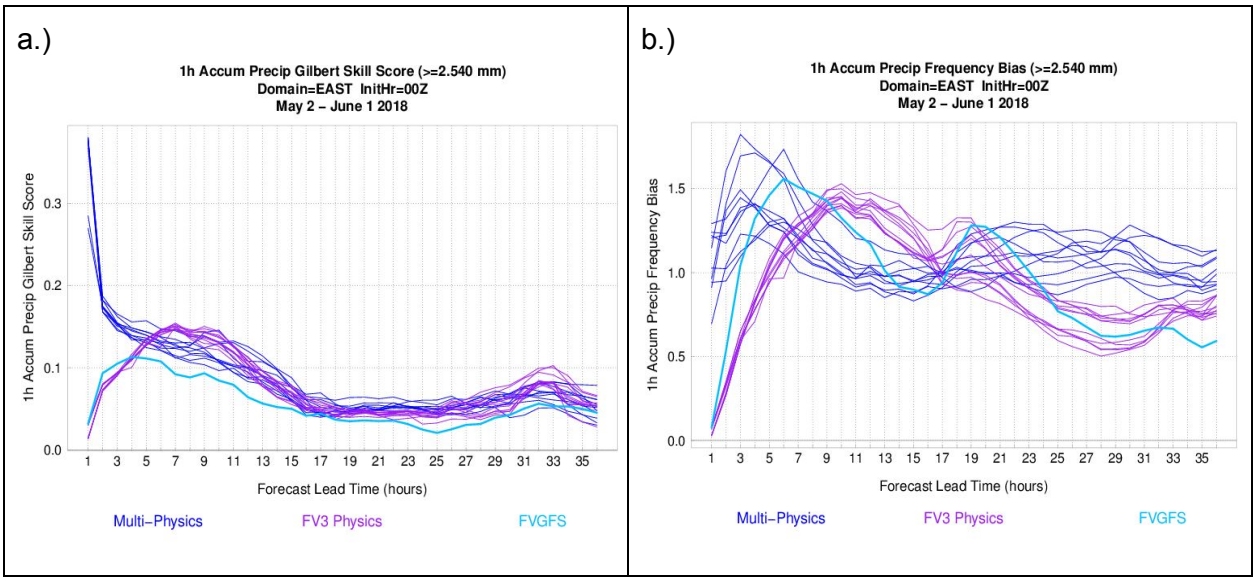

Figure 13: *(a) GSS and (b) frequency bias time series plots of 1-h accumulated precipitation >=2.54 mm for each individual ensemble member aggregated across the eastern half of the CONUS domain for all available forecasts during the experiment. The multi-physics members are in blue, the FV3 physics members are in purple, and the gfdl-fv3 member is in teal.*

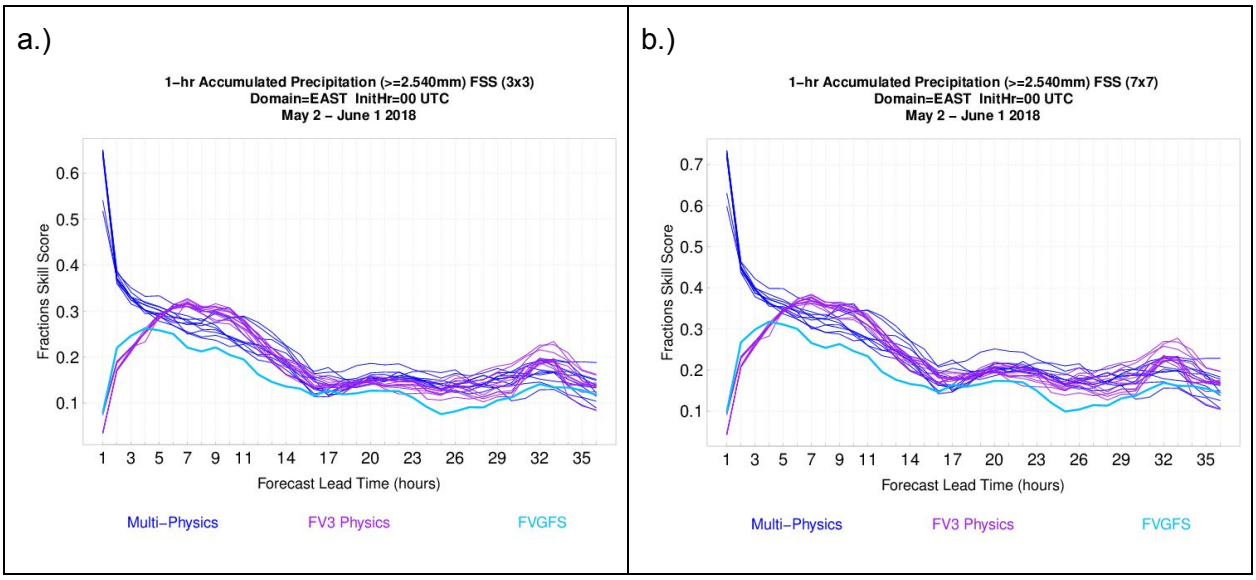

Figure 14: *FSS time series plots of 1-h accumulated precipitation >=2.54 mm for each individual ensemble member at a neighborhood width of (a) 3x3 grid squares and (b) 7x7 grid squares aggregated across the eastern half of the CONUS domain for all available forecasts during the experiment. The multi-physics members are in blue, the fv3-physics members are in purple, and the gfdl-fv3 member is in teal.*

The multi-physics ensemble members reached peak frequency bias (1 is unbiased, >1 is over-forecast, and <1 is under-forecast) values between forecast hours 3 - 6 (Fig. 13b). The members then had values generally between slightly less than 1.0 to approximately 1.25. The FV3 physics ensemble members reached peak frequency bias values later than the multi-physics ensemble, around forecast lead times 8 - 9. The FV3 physics ensemble generally decreased throughout the rest of the forecast period.

Fractions Skill Score (FSS; higher is better) was calculated over two neighborhood widths: 3x3 grid squares, or 9x9 km, and 7x7 grid squares, or 21x21 km. The FSS (3x3) followed the same temporal trend as the GSS for both the multi-physics and FV3 physics ensemble subsets (Fig. 14a). The fvgfs deterministic member exhibited the lowest FSS value throughout the forecast period. The FSS (7x7) followed the same temporal trend as the FSS (3x3) for both the multi-physics and FV3 physics ensemble subsets (Fig. 14b). The overall scores were higher which is expected with the larger neighborhood. Interestingly, the FV3 physics ensemble subset and fvgfs deterministic member exhibited the same values for the initial forecast lead times as in the (3x3). After those first forecast hours, the FV3 physics ensemble subset and fvgfs deterministic member yielded overall higher scores, as expected.

For the area of all simple MODE objects, the multi-physics ensemble subset almost always produced smaller objects than were observed throughout the forecast period (Fig. 15a). There were occasional time periods when one or a few individual ensemble members produced larger objects than were observed. Both ensemble subsets and the deterministic member followed the same temporal trend as was observed. The FV3 physics ensemble members generally kept the

observations within the envelope of solutions. At the peak times, the FV3 physics ensemble members tended to over-predict object area. The fvgfs deterministic member followed the observation fairly closely throughout.

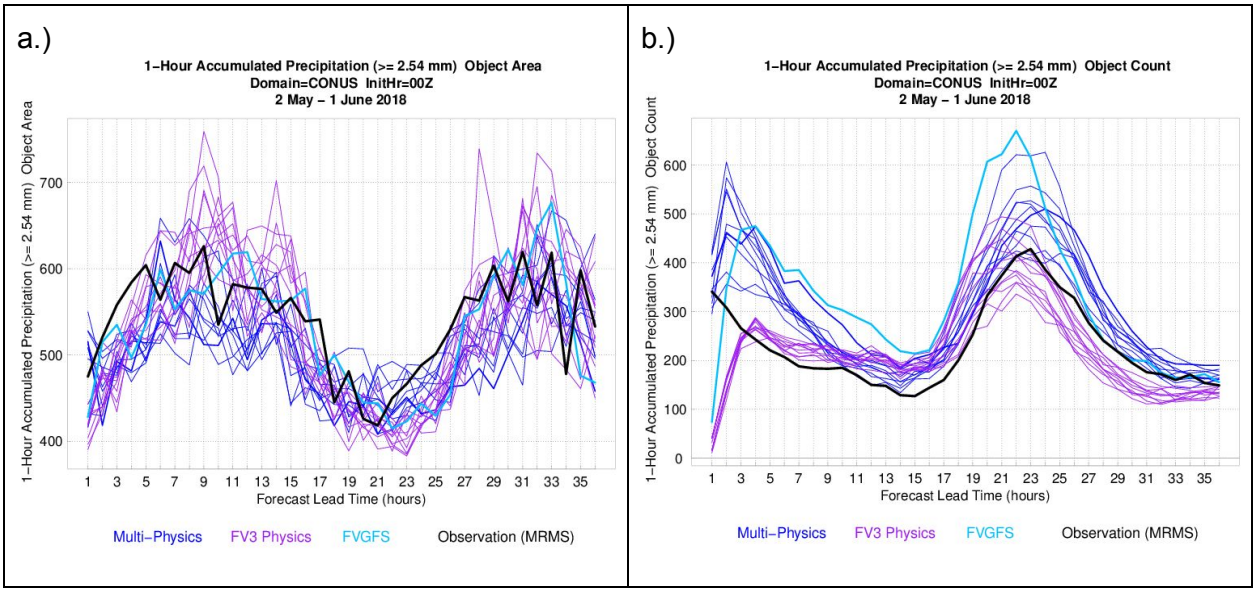

Figure 15: *(a) Median object area (grid squares) and (b) total object count of 1-h accumulated precipitation >=2.54 mm over the full CONUS domain for all available forecasts during the experiment. The observation objects are in bolded black, the multi-physics members in blue, the fv3-physics members in purple, and the gfdl-fv3 member in teal.*

For the total number of simple MODE objects, the multi-physics ensemble subset almost always produced more objects than were observed (Fig. 15b). The FV3 physics ensemble subset produced more objects than were observed between forecast hours ~4 - 18. After that time, the FV3 physics ensemble subset almost always produced less objects than were observed. The fvgfs deterministic member almost always produced more objects than were observed. All ensemble subsets and the deterministic member exhibited the same temporal trend as was observed after forecast hour 4.

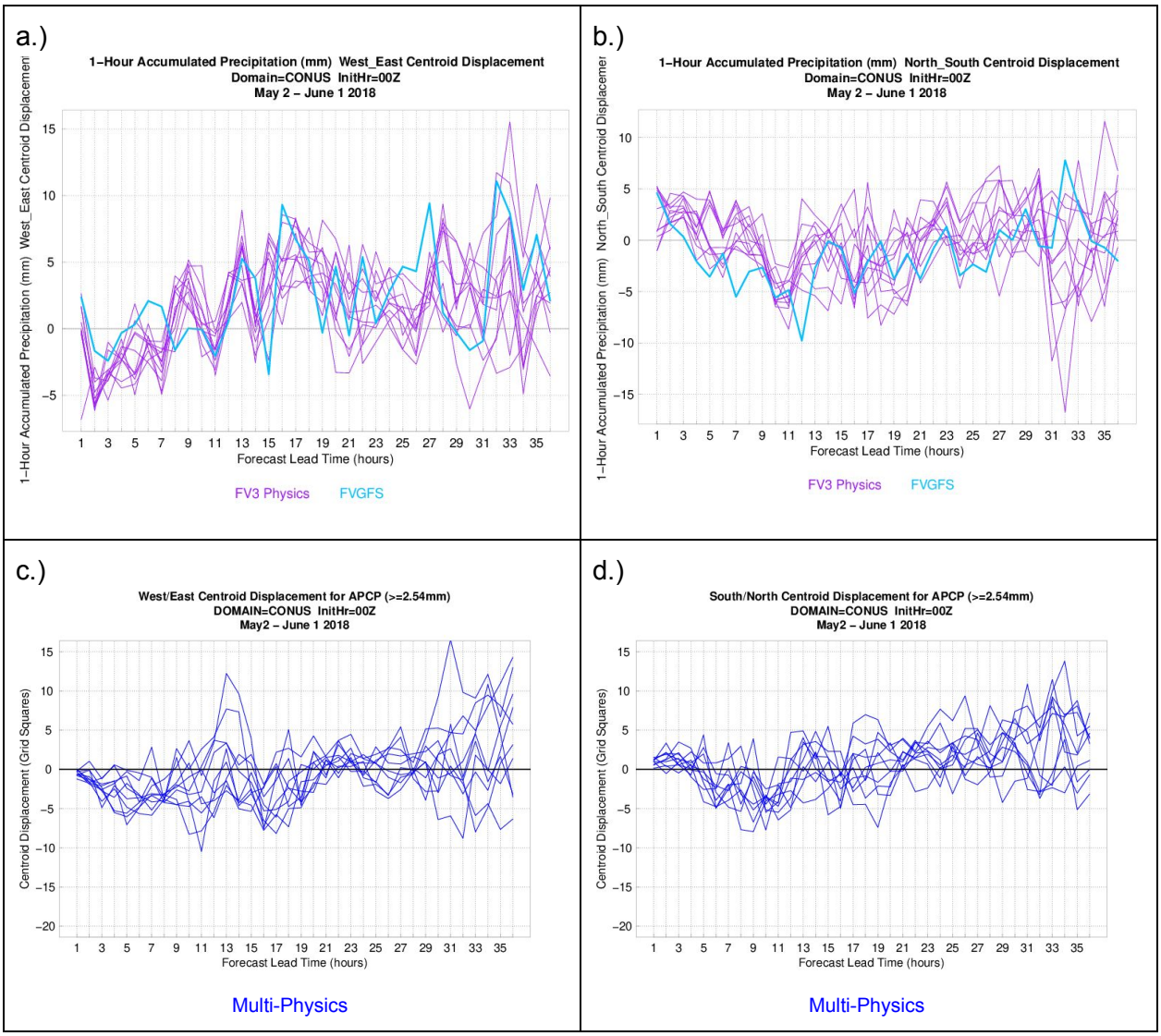

Figure 16: *Centroid displacement (grid squares) in the west-east direction (left; a., c.) and south-north direction (right) for the FV3 physics and gfdl-fv3 (top; a., b.; purple and teal) and multi-physics (bottom; c., d.; blue) for 1-h accumulated precipitation objects >=2.54 mm aggregated over the full CONUS domain for all available forecasts during the experiment.*

The displacement trends for 1-h accumulated precipitation objects was investigated with the centroid attribute derived from MODE. This is done by calculating the centroid distance between the matched forecast and the observed accumulated precipitation objects. A negative (positive) value indicates either a westerly (easterly) or southerly (northerly) displacement. For east-west direction, the FV3 physics ensemble subset and fvgfs deterministic member exhibited a westerly displacement for the first 7 forecast hours (Fig. 16a). After that time period, the majority of the displacements were easterly. At the end of the forecast period (~30) the displacements were in mixed directions. In terms of the multi-physics ensemble subset, all the ensemble subset members exhibit an overall westerly displacement for the first 11 forecast hours (Fig. 16c). At approximately forecast hour 12, there is an increase in member variability with members

exhibiting displacements in both directions until forecast hour 21 where member variability decreases. From forecast hour 24 on, there is increased variability in both directions for the remainder of the forecast period. For the north-south direction, the FV3 physics ensemble subset and fvgfs deterministic member followed the same temporal trend (Fig. 16b). The deficiency began to the north then decreased to the south, followed by a return to the north. At the end of the forecast period, the ensemble members exhibited more variance. The multi-physics ensemble subset members tend to be displaced to the north for the majority of the forecast period and displayed more variability as forecast lead time increased (Fig. 16d).

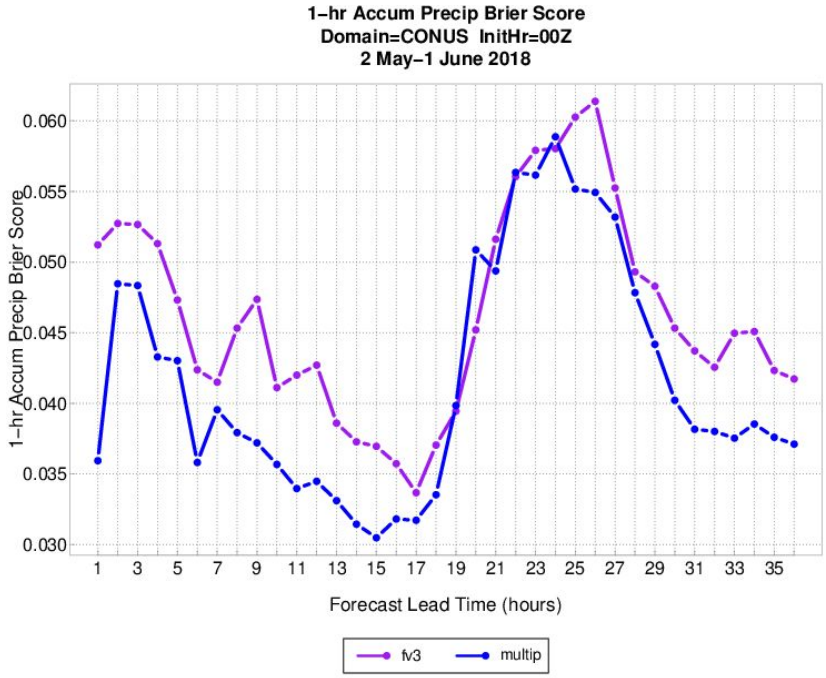

Figure 17: *One-hour precipitation accumulation Brier scores (>=0.254 mm) averaged over the CONUS for both the multi-physics (blue) and FV3 ensemble (purple) subsets as a function of lead time. Smaller values are better.*

Analyzing the Brier score as a function of forecast lead time for 1-hour precipitation accumulation (>=0.254), a clear diurnal cycle can be seen for both ensemble subsets. Brier scores drop to their lowest values around 13-17 hours into the forecast (13-17 UTC in this case), corresponding to relative minimums in precipitation during the morning hours. However, once the convective period of the day begins after 18 UTC, the uncertainty related to forecasting of precipitation accumulation becomes apparent, as Brier scores peak around 00 UTC. Neither ensemble subset performs particularly better than the other during the convective part of the day; however, the mixed-physics ensemble has lower (better) Brier scores during the nocturnal period. With average Brier scores around ~0.045, both ensemble subsets perform well for this particular event, as an ideal Brier score is 0 and the worst possible score is 1.

For the reliability plots for 1-hr precipitation accumulation >= 0.254 mm, the multi-physics and FV3 physics ensemble subsets are both slightly over-confident, particularly for higher forecast frequencies, indicating that both ensembles likely lack sufficient spread to explain the variability

inherent in the forecasts. However, the multi-physics ensemble subset is slightly less over-confident, especially for higher forecast frequencies. Finally, while the FV3 physics ensemble crosses the no-skill threshold for higher forecast frequencies, the climatological frequency is fairly low for this event.

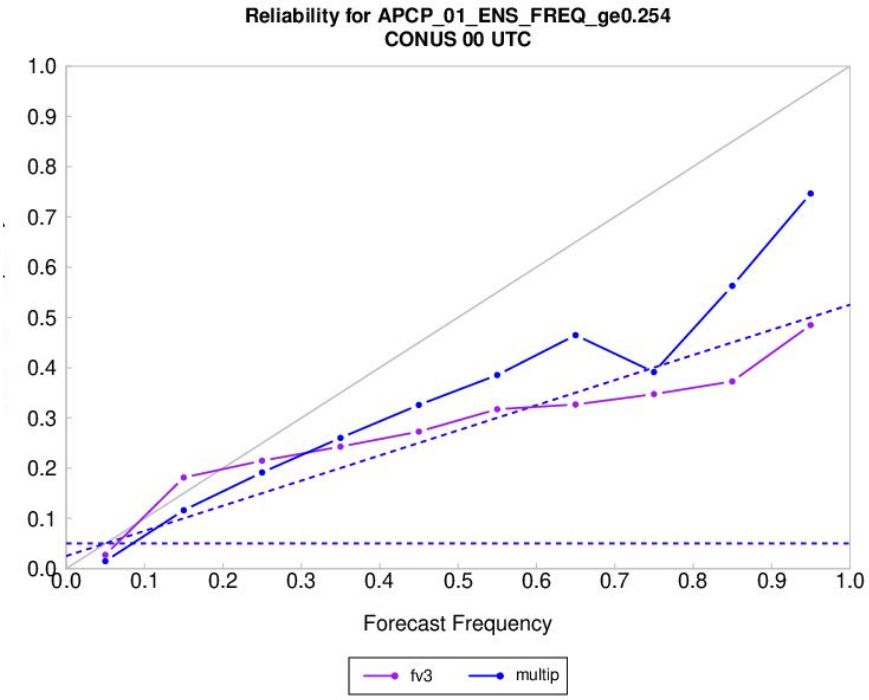

Figure 18: *Same as Figure 3, except for one-hour precipitation accumulation >= 0.254 mm.*

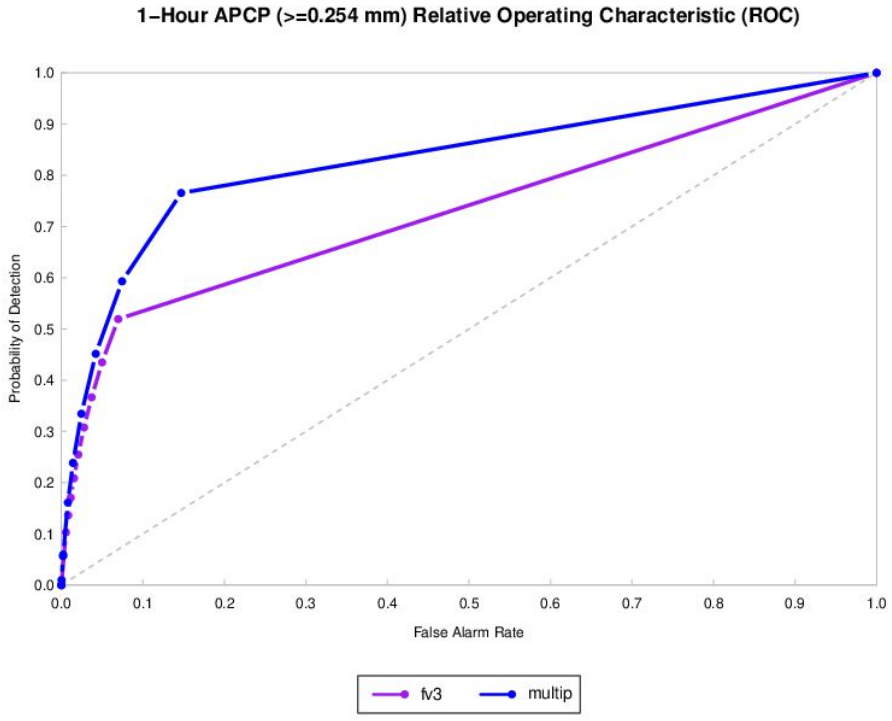

Figure 19: *Receiver Operating Characteristic (ROC) diagram for one-hour precipitation >= 0.254 mm, averaged over the CONUS for all lead times of both the multi-physics (blue) and FV3 ensemble (purple) subsets. An ideal ROC would have all probability points in the upper left-hand corner (all probability of detection values of 1, and all false alarm rates of 0). The diagonal line represents forecasts that have no skill.*

The Receiver Operating Characteristic (ROC) plot (Fig. 19) for 1-h precipitation accumulation (>= 0.254 mm) shows a clustering of probability points, which again indicates that both the mixed-physics and FV3 physics ensemble subsets are under-dispersive. However, all probability thresholds result in hit rates (y-axis) that exceed false alarm rates (y-axis), indicating that each ensemble subset is skillful in the sense that it is able to discriminate between events and non-events for forecasts of 1-hr precipitation accumulations of >= 0.254 mm.

Composite Reflectivity (≥30 dBZ)

As in 1-hour accumulated precipitation, the multi-physics and FV3 physics ensemble members exhibited similar GSS behavior after forecast hour 8 (Fig. 20a). The multi-physics ensemble subset GSS sharply decreased during the early forecast hours while the FV3 physics ensemble subset and fv3-thomp-caps deterministic member increased during those same early forecast hours. The FVGFS began with a sharp decrease and then followed the same trend as the FV3 physics ensemble subset members. The deterministic fvgfs had the overall lowest GSS score throughout the forecast period. The individual members in both ensembles exhibited more variability than for 1-hour accumulated precipitation.

For frequency bias, the multi-physics ensemble subset exhibited large inter-member variability (Fig. 20b). The members display different solutions with some members consistently under-forecasting, some over-forecasting, and some going back and forth throughout the forecast period. Despite the value differences, the multi-physics ensemble subset members followed the same temporal trend across the forecast period. The FV3 physics ensemble subset was more clustered than the multi-physics ensemble. All ensemble members start by under-forecasting, then over-forecasting for approximately the first 24 hours of the forecast period. After that time, all members generally under-forecasted again through the end of the time period. The fv3-thomp-caps deterministic member produced generally lower values than the FV3 physics ensemble as a whole, but it follows the temporal trend. Interestingly, the fvgfs deterministic member followed the temporal trend of the multi-physics ensemble subset until forecast hour 18 where it switched to following the FV3 physics ensemble subset temporal trend.

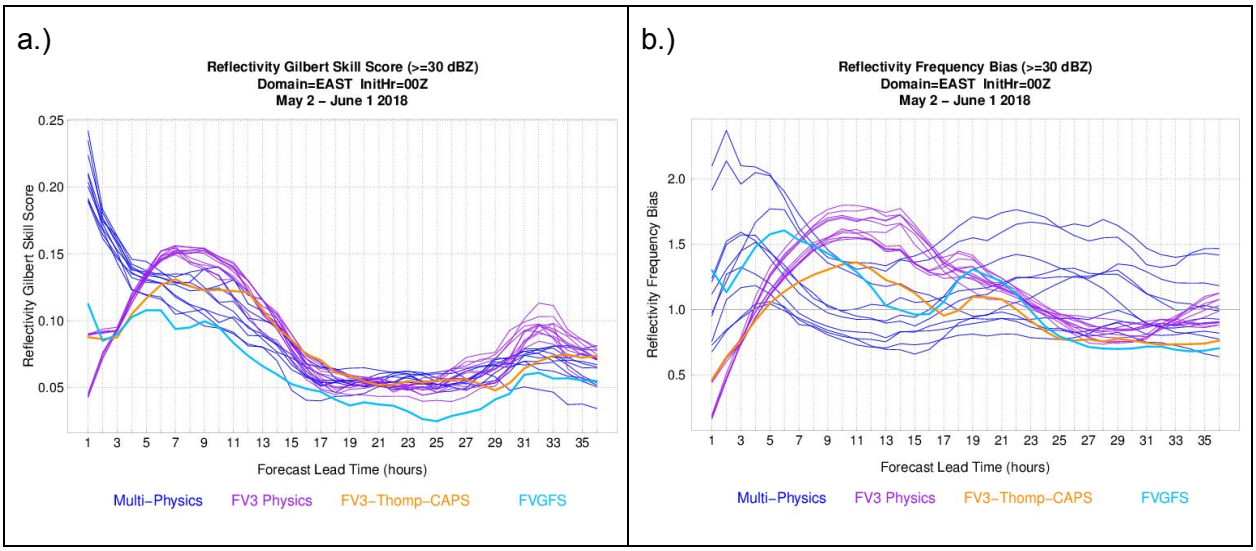

Figure 20: *Same as Fig. 13, except for composite reflectivity >= 30 dBZ.*

The FSS (3x3) followed nearly the exact same temporal trend as the GSS for both ensemble subsets and both deterministic members (Fig. 21a). The multi-physics ensemble subset was more clustered for the FSS (3x3) than in the GSS. The FSS (7x7) followed the same temporal trend as the FSS (3x3) for both the multi-physics and FV3 physics ensemble subsets (Fig. 21b). The overall scores were higher, which is expected with the larger neighborhood. The fv3-thomp-caps deterministic member tended to be closer to the higher end of the FV3 physics envelope than in the FSS (3x3). The fvgfs deterministic member exhibits the same behavior as in the FSS (3x3) solution.

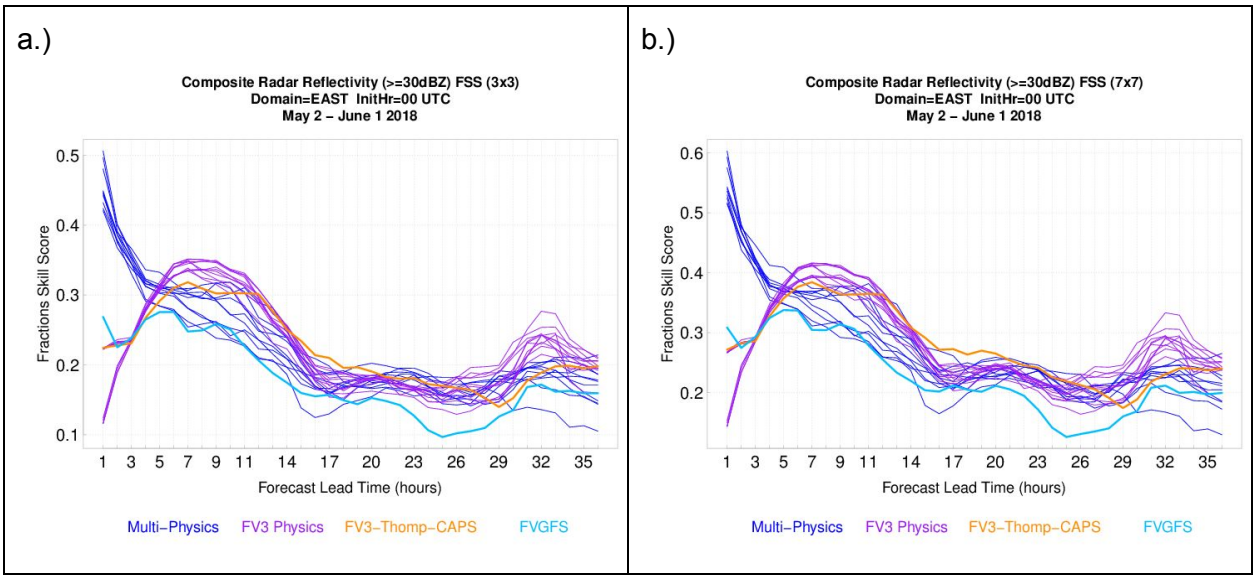

Figure 21: *Same as Fig. 14, except for composite reflectivity >= 30dBZ.*

The multi-physics ensemble subset almost always produced smaller objects than were observed throughout the forecast period (Fig. 22a). The exception to this is during forecast lead

times 16 - 24, the multi-physics ensemble subset contains the observed area within its envelope. Both ensemble subsets and deterministic members followed the same temporal trend as was observed. The FV3 physics ensemble members generally kept the observations within the envelope although towards the bottom of the envelope for the first 24 hours, with exception to the first few forecast lead times. After that, the FV3 physics ensemble members tend to produce smaller objects than were observed. The fv3-thomp-caps deterministic member tends to follow the behavior of the FV3 physics ensemble subset (generally larger objects for the first 24 hours then smaller objects for the remainder of the forecast). The fvgfs deterministic member tends to follow the behavior of the multi-physics ensemble subset (generally smaller objects than were observed for the majority of the forecast period).

The multi-physics ensemble subset almost always produced more objects than were observed (Fig. 22b). There are two individual members that nearly matched the observed number of objects from forecast hours 9 to 25. The FV3 physics ensemble subset produced more objects than were observed throughout save for the first 2 forecast hours. One individual FV3 physics member closely matched the observed object counts. As previously observed, the fvgfs deterministic member followed the multi-physics ensemble behavior and the fv3-thomp-caps deterministic member followed the FV3 physics ensemble behavior.

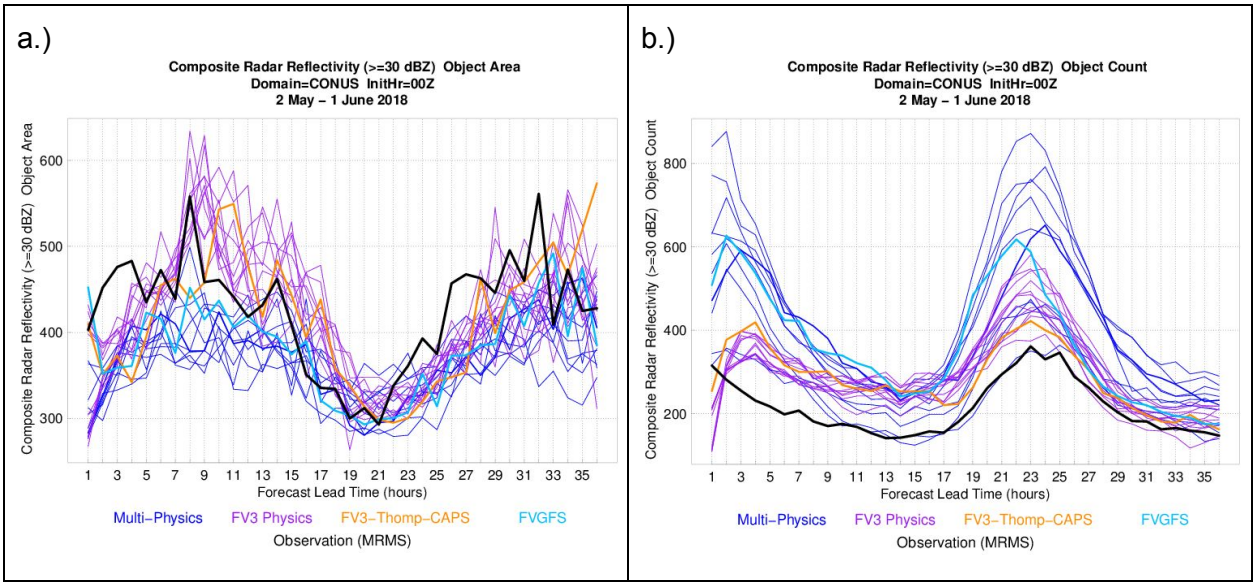

Figure 22: *Same as Fig. 16, except for composite reflectivity >= 30 dBZ.*

For east-west displacement, the FV3 physics ensemble begins with a westerly displacement for the first 8 forecast hours then switches to a majority easterly displacement for the remainder of the forecast period (Fig. 23a). The fvgfs deterministic member displays a slightly different trend of displacement. The member exhibited the same peaks and valleys as the ensemble, except they occurred at later forecast lead times. The fv3-thomp-caps deterministic member displayed a different displacement magnitude from the ensemble. The lows were not as low and the highs

were not as high as those in the ensemble subset. The majority of the multi-physics ensemble subset members are displaced to the west for the first 18 hours of the forecast period (Fig. 23c). After that point, all members switch to a majority easterly displacement until the late forecast point at which point there is variability in both direction and magnitude of the displacement from member to member. Inter-member variability remained large throughout the forecast period.

For the north-south displacement, the FV3 physics ensemble subset began with slight northerly displacement for the approximate first 6 forecast hours (Fig. 23b). From that hour to approximately forecast hour 22, the ensemble members displayed a slight southerly displacement. From that forecast hour until approximately forecast hour 24, there is a slight northerly displacement. From then until the end of the forecast period, the ensemble members display varying solutions with displacements both to the south and north. The fvgfs deterministic member displayed a similar temporal trend as the FV3 physics ensemble subset, but the displacement direction is almost always southerly. The fv3-thomp-caps deterministic more closely follows the FV3 physics ensemble behavior. The multi-physics ensemble members display the overall same trend as in the east/west displacement although there is more variability between members throughout the forecast period (Fig. 23d).

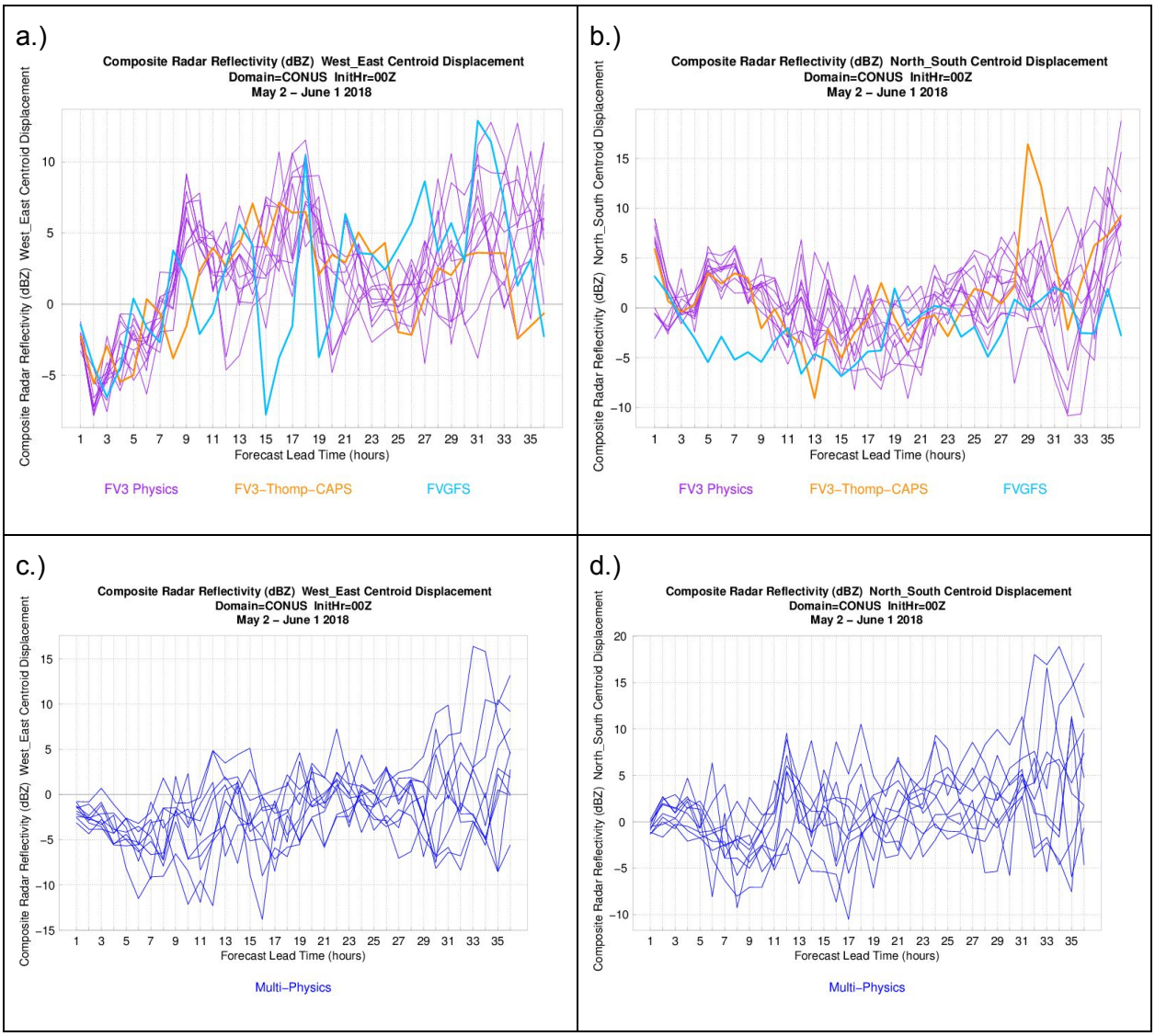

Figure 23: *Same as Fig. 17, except for composite reflectivity >= 30dBZ.*

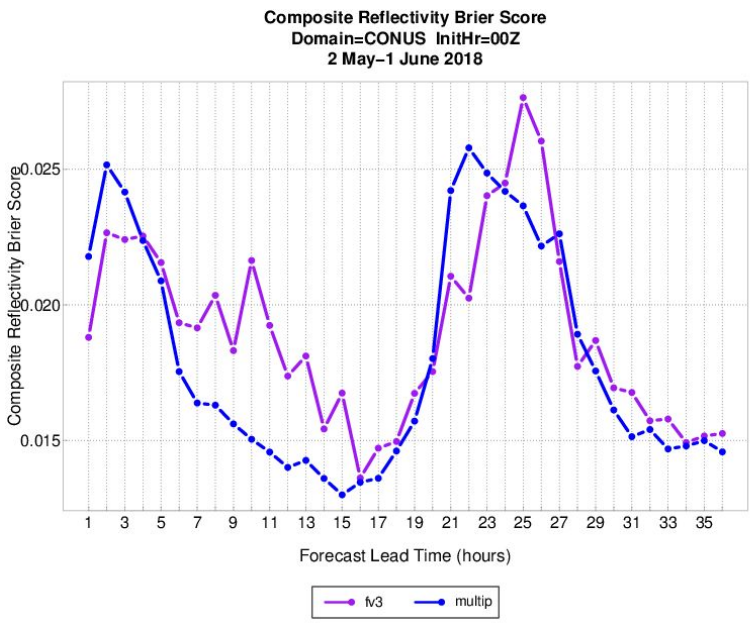

Figure 24: *Same as Fig. 17, but for composite reflectivity >= 30 dBZ.*

The Brier score plot for forecasts of reflectivity >= 30 dBZ (Fig. 24) shows a very similar diurnal signal to that seen in the 1-hr precipitation forecasts from Figure 17. Scores improve with time during the nocturnal period of the forecast until the late morning (15-17 UTC), after which they degrade due to uncertainty in convective-scale forecasting later in the day, peaking at around 00 UTC. As with precipitation forecasting, the mixed-physics ensemble does a better job during the nocturnal period of the forecast, while neither subset appears to be better during the height of the convective period in the afternoon and evening. Overall, the Brier scores for both ensemble subsets are extremely low on the scale of 0 to 1.

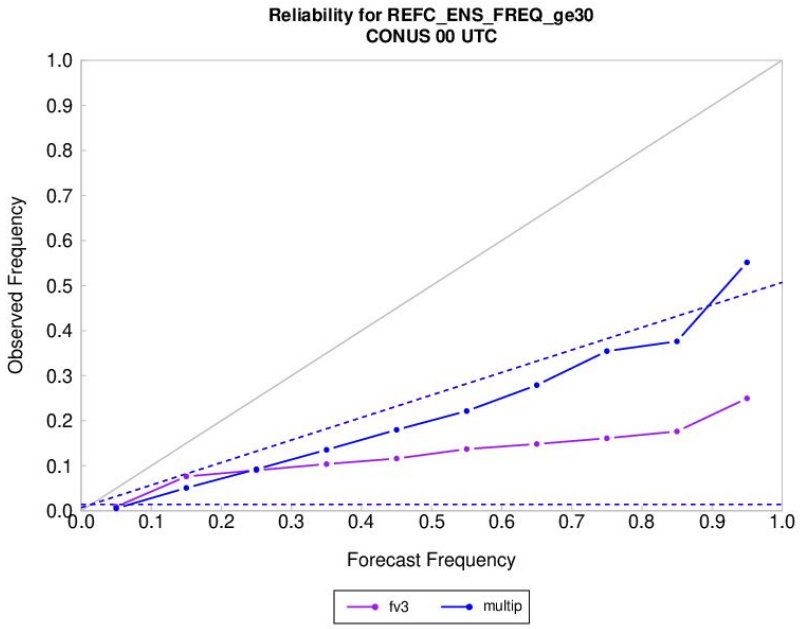

Figure 25: *Same as Fig. 3, but for composite reflectivity >= 30 dBZ.*

The reliability diagram for forecasts of reflectivity >= 30 dBZ (Fig. 25) indicates that for low forecast frequencies, both ensemble subsets show similar performance, but at higher forecast frequencies, the multi-physics ensemble subset corresponds with higher observed frequencies than the FV3 physics ensemble. However, neither ensemble subset has any real skill (based on the no-skill line), and the FV3 physics ensemble subset has nearly no resolution with an almost flat profile. The base rate for this event is extremely low, so it is important to again note that more cases would be necessary to produce a robust reliability profile for these two ensemble subsets.

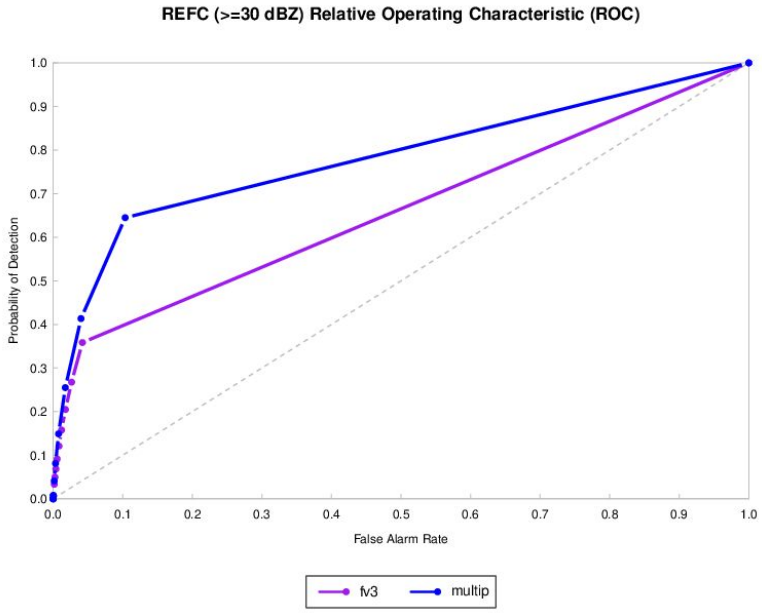

Figure 26: *Same as Fig. 19, but for composite reflectivity >= 30 dBZ.*

Similar to the ROC diagram for precipitation accumulation (Fig. 19), both ensemble subset frequency thresholds are clustered within the ROC diagram for reflectivity forecasts >= 30 dBZ (Fig. 26), with the mixed physics subset having slightly less clustering. As mentioned before, this finding is related to the under-dispersiveness of the two ensembles. However, both subsets are able to discriminate between events and non-events for forecasts of reflectivity >= 30 dBZ, as all probability thresholds correspond with hit rates that are larger than their corresponding false alarm rates.

Summary

Two ensemble subsets, FV3 physics and multi-physics, and two deterministic members, nssl-fv3 and fv3-thomp-caps, were evaluated for the 2018 Hazardous Weather Testbed Spring Forecast Experiment. Three surface variables, 2-m temperature, 2-m dew point temperature, and 10-m wind speed, were evaluated, as well as 1-hour accumulated precipitation >= 2.54 mm and composite reflectivity >= 30 dBZ.

For the surface variables, the multi-physics ensemble tended to outperform the FV3 physics ensemble for temperature and dew point temperature. The FV3 physics ensemble slightly outperformed the multi-physics ensemble for wind speed. Both deterministic fv3 members followed a similar trend as the overall FV3 physics ensemble. As a whole, both ensembles were under-dispersive for all three variables. Both ensembles displayed a tendency to under-forecast temperature and dew point temperature while exhibiting no skill when it comes to wind speed. This behavior likely stems from the under-dispersiveness of the ensembles.

In terms of 1-hour accumulated precipitation at thresholds >= 2.54 mm and composite reflectivity at thresholds >= 30dBZ, both ensemble subsets performed similarly throughout the forecast period. For traditional verification statistics, neither ensemble subset clearly outperformed the other. In terms of spatial verification, the FV3 physics ensemble subset was better overall at capturing the observed number and area of MODE objects, although both ensemble subsets tended to produce more smaller objects than observed. Both ensembles capture the observed diurnal signal. Both ensemble subsets illustrated similar temporal trends for centroid displacement although the direction of that displacement varied. The ensemble statistics for the two subsets were very similar.