Assessing ensemble forecast performance for select members available in CLUE during 2016 HWT-SFE

Developmental Testbed Center (DTC) Annual Operating Plan (AOP) 2017 Final Report

REGIONAL ENSEMBLE TEAM: JAMIE WOLFF[1], MICHELLE HARROLD[1], JEFF BECK[2], ISIDORA JANKOV[2], TRACY HERTNEKY[1], AND LINDSAY BLANK[1]

[1]NATIONAL CENTER FOR ATMOSPHERIC RESEARCH (NCAR) AND DEVELOPMENTAL TESTBED CENTER (DTC)

[2]COOPERATIVE INSTITUTE FOR RESEARCH IN THE ATMOSPHERE (CIRA)/AFFILIATED WITH NOAA/ESRL/GSD AND DEVELOPMENTAL TESTBED CENTER (DTC)

Table of Contents

Introduction

During NOAA Testbed experiments that occur throughout the year [e.g., Hazardous Weather Testbed (HWT), Weather Prediction Center (WPC) Hydrometeorological Testbed (HMT)], a plethora of experimental model data is produced to support the typically several weeks-long events. While this data is subjectively assessed daily during the experiments, there is often times a lack of extensive objective verification after the experiment to thoroughly investigate the contributed model configuration strengths and weaknesses. The large datasets produced during these testbed experiments provide an excellent opportunity to help identify and begin to answer the most pressing scientific questions that need to be addressed.

During the 2016 and 2017 HWT Spring Forecasting Experiment (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.

Many questions remain regarding the best approach to constructing a convection-allowing model (CAM) ensemble system. For example, should model uncertainty be addressed through multiple dynamic cores, multiple physics parameterizations, stochastic physics, or some combination of these? CLUE provides the datasets necessary to begin to explore this question; the methods targeted for this work included examining single physics vs. multi-physics approaches. During AOP 2018, a retrospective evaluation to investigate the probabilistic performance of ensembles constructed from each targeted ensemble subset of CLUE, along with deterministic forecasts from individual members to understand their contribution to the overall ensemble spread, was conducted and the results are described below.

Experiment design

Community Leveraged Unified Ensemble (CLUE) dataset

A large number of experimental model data is produced to support the five-weeklong experiment as part of CLUE. CLUE is a thoughtfully constructed super-ensemble of over 60 members contributed by a number of modeling groups. It provides a substantial amount of data that is subjectively assessed during the experiment and provides a great dataset for retrospective evaluation as well.

During AOP 2018, DTC staff examined two 9-member ensemble subsets from CLUE 2016 (Clark et al. 2016). The first was a single physics ensemble with variations to the members coming from initial condition (IC) and boundary condition (BC) perturbations. The physics suite used included the Thompson microphysics, RRTMG long-wave and short-wave radiation, Noah land surface model, and the Mellor–Yamada–Janjic (MYJ) planetary boundary layer (PBL) scheme. The second CLUE subset examined was a multi-physics ensemble initialized with the same IC and BC perturbations as used in the single physics; however, the physics suite was not fixed and included variations in the microphysics scheme [using Thompson, Predicted Particle Property (P3), Millbrandt-Yau (MY), and Morrison] and the PBL [including MYJ, Yonsei University (YSU), and Mellor-Yamada Nakanishi and Niino (MYNN)]. One member (the control member) was included in both ensemble subsets (Table 1).

Table 1. Physics suite description for CLUE 2016 subsets examined.

Single physics + IC/BC pert (9 members)			
MEMBER	**MP**	**LSM**	**PBL**
control	Thompson	Noah	MYJ
Multi-physics + IC/BC pert (9 members)			
MEMBER	**MP**	**LSM**	**PBL**
control	Thompson	Noah	MYJ
core03	P3	Noah	YSU
core04	MY	Noah	MYNN
core05	Morrison	Noah	MYJ
core06	P3	Noah	YSU
core07	MY	Noah	MYNN
core08	Morrison	Noah	YSU
core09	P3	Noah	MYJ
core10	Thompson	Noah	MYNN

The data evaluated for this test came from the 2016 HWT-SFE, which was held from 4 May – 3 June. Model output was available for weekdays during that time period for a minimum of 36-hour forecasts initialized at 00 UTC over a 3-km CONUS domain.

Observations

The Multi-Radar/Multi-Sensor (MRMS) dataset, first developed at the National Severe Storms Laboratory (NSSL) and now run operationally at NCEP, was used as the observational analysis product for comparison. Radar-based data are integrated with atmospheric environmental data, satellite data, and lightning and rain gauge observations to generate a suite of severe weather and quantitative precipitation estimation (QPE) products at very high spatial (1 km) resolution (Zhang et al. 2016). The analyses used in this comparison included the local gauge bias-corrected radar QPE and the composite reflectivity. The MRMS data was regridded to the model integration domain to allow for grid-to-grid comparisons. Budget interpolation was used for the QPE field, while the nearest neighbor approach was used for the surface precipitation product.

Model verification approaches

A variety of methods can be used to conduct an evaluation of both deterministic and probabilistic forecasts. To support this analysis, the DTC developed and supported Model Evaluation Tools (MET) verification software system was utilized. Metrics examined included both traditional methods commonly used in the community (spread, skill, error, reliability, etc.) and newer approaches that provide additional diagnostic information, especially at higher resolution, including the Method for Object-based Diagnostic Evaluation (MODE) and Fractions Skill Score (FSS). More information on all of the metrics used for this work can be found in the MET Users' Guide (2017).

The goal of this research is to help answer significant scientific questions that remain regarding the best approach to constructing a convection-allowing ensemble system using the MET package and the CLUE dataset.

Traditional verification metrics

In terms of traditional verification metrics, two key statistics were used in this study. The first was Gilbert Skill Score (GSS), which is the fraction of observed events that were correctly predicted (or hits over the total forecast and observed area) and adjusted for random hits. The second was frequency bias, which is the ratio of the frequency of forecast events to observed events (or total forecast area divided by the total observed area).

Spatial verification metrics

One of the advanced spatial verification techniques we used was MODE. This method identifies objects and then merges and matches those objects in the forecast and observation fields. MODE has many attributes associating the forecast and observation objects, including things like centroid distance, angle difference, and area ratio, in order to expand the diagnostic investigation into forecast performance.

We also computed FSS, which is a neighborhood method that is used to obtain a measure of how forecast skill varies with spatial scale. While traditional grid-to-grid comparisons can be used to see that our forecast misses an event, FSS can help identify whether or not a forecast has good skill at larger spatial scales. This approach may be particularly relevant as models move to higher resolution.

Ensemble verification metrics

In the course of this evaluation, several ensemble verification metrics were also applied to assess the ensemble performance. These included: (a) spread, the standard deviation of the individual member forecasts compared to the ensemble mean, (b) Brier score, a measure of the mean squared probability error, (c) reliability diagram, showing observed frequency of events versus the forecast probability of those events, (d) Relative Operating Characteristic (ROC) curve, a measure of resolution given by the ability of the forecast to discriminate between two alternative outcomes, and (e) rank histogram, to compare the rank of the observations to all members of the ensemble forecast.

Verification results

Ultimately, the question being addressed is: is there an advantage to using multiple microphysics/PBL parameterizations compared to one common physics suite within an ensemble. Using the CLUE dataset we address this question by examining the forecasts of accumulated precipitation and composite reflectivity. To focus on areas where there is more organized convection, the analysis below is restricted to the Eastern CONUS.

Accumulated precipitation

GSS was calculated for each individual ensemble member by forecast lead time (Figure 1a). A perfect GSS is 1, so the higher the value, the more skillful the member. For purposes of the analysis presented here, we focused on the quantitative precipitation forecasts (QPF) greater than or equal to 2.54 mm accumulated over 1 hour aggregated across the eastern half of the CONUS domain for all available forecasts during the experiment. At most lead times, members from both ensemble subsets are clustered together with similar performance throughout the forecast period. In general, the control member (black line) trends towards the top of the cluster.

Frequency bias was also calculated for every ensemble member by forecast lead time. A perfect frequency bias score is 1, which is highlighted by the black line in Figure 1b. Any value below 1 is an under-forecast and any value above 1 is an over-forecast. There is an obvious diurnal signal across both ensembles with a pronounced high bias during the afternoon/evening hours, followed by a neutral to somewhat low bias overnight and through the morning hours. The single physics members generally have a smaller envelope of frequency bias values and are more often closer to 1 (unbiased) than the multi-physics members.

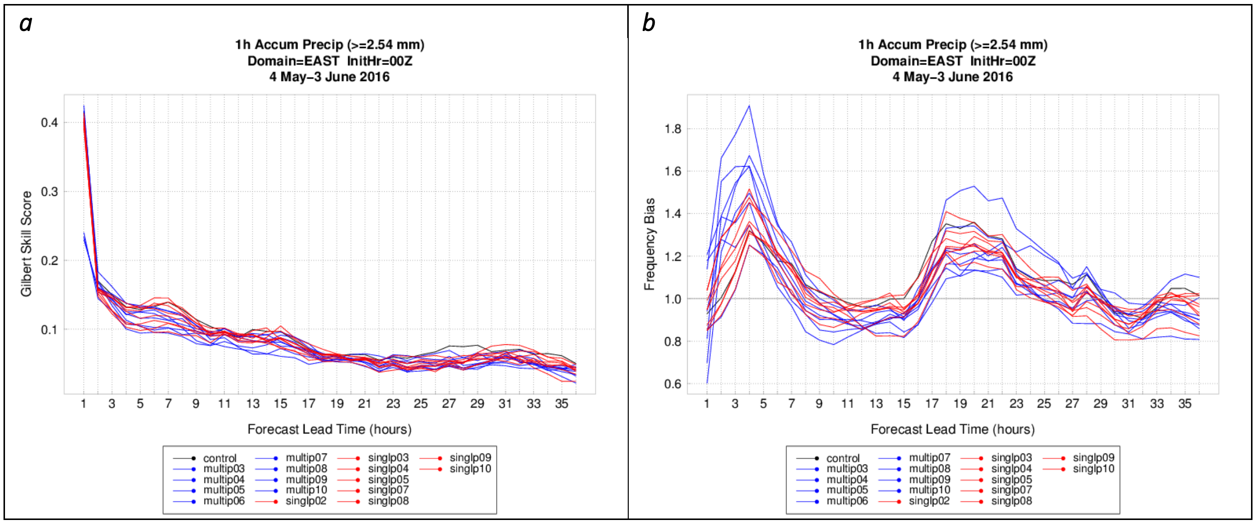


Figure 1. (a) GSS and (b) frequency bias time series plots of 1-h accumulated precipitation ≥2.54 mm for each individual ensemble member aggregated across the eastern half of the CONUS domain for all available forecasts during the experiment. The control member (used in both ensembles) is shown in black, the multi-physics members in blue, and the single physics members in red.

FSS was computed as a function of lead time for two neighborhood widths [3x3 grid squares = 9x9 km or 81km2 (Figure 2a) and 7x7 grid squares =21x21 km or 441km2 (Figure 2b)]. The FSS also has the characteristic that higher is better. Similar to that shown for GSS, members from both ensemble subsets are clustered together with similar performance throughout the forecast period. When comparing the performance at the two neighborhood widths, we see a shift toward higher scores for all members from the 3x3 to 7x7 grid square neighborhood, as we might expect. This indicates that as we broaden the spatial comparison area the model performs better indicating it often has storms in the general region of the observations.

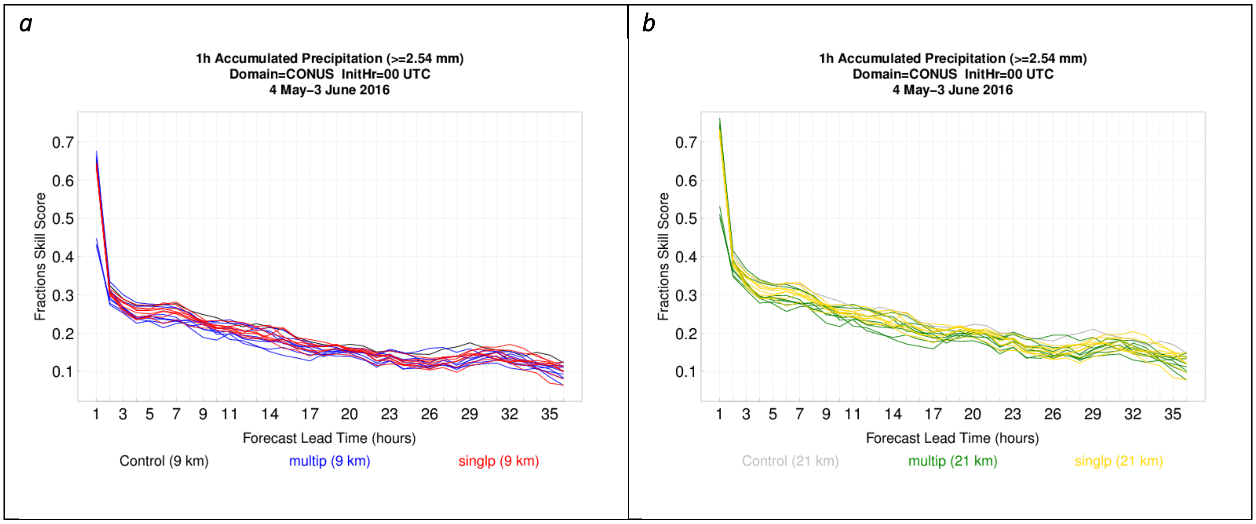


Figure 2. FSS time series plots of 1-h accumulated precipitation ≥2.54 mm for each individual ensemble member at a neighborhood width of (a) 3x3 grid squares and (b) 9x9 grid squares aggregated across the eastern half of the CONUS domain for all available forecasts during the experiment. The control member (used in both ensembles) is shown in black, the multi-physics members in blue, and the single physics members in red.

Using MODE, 1-h accumulated precipitation objects at a threshold of ≥2.54 mm were defined using the MODE approach. Figure 3 provides a visual example of the MODE objects created in the observed field and for each forecast member for one particular 00 UTC initialization on 25 May 2016. While the single physics members encompassed the observation objects better later in day 1, the multi-physics handled the overnight convective initiation and mesoscale convective system (MCS) better.

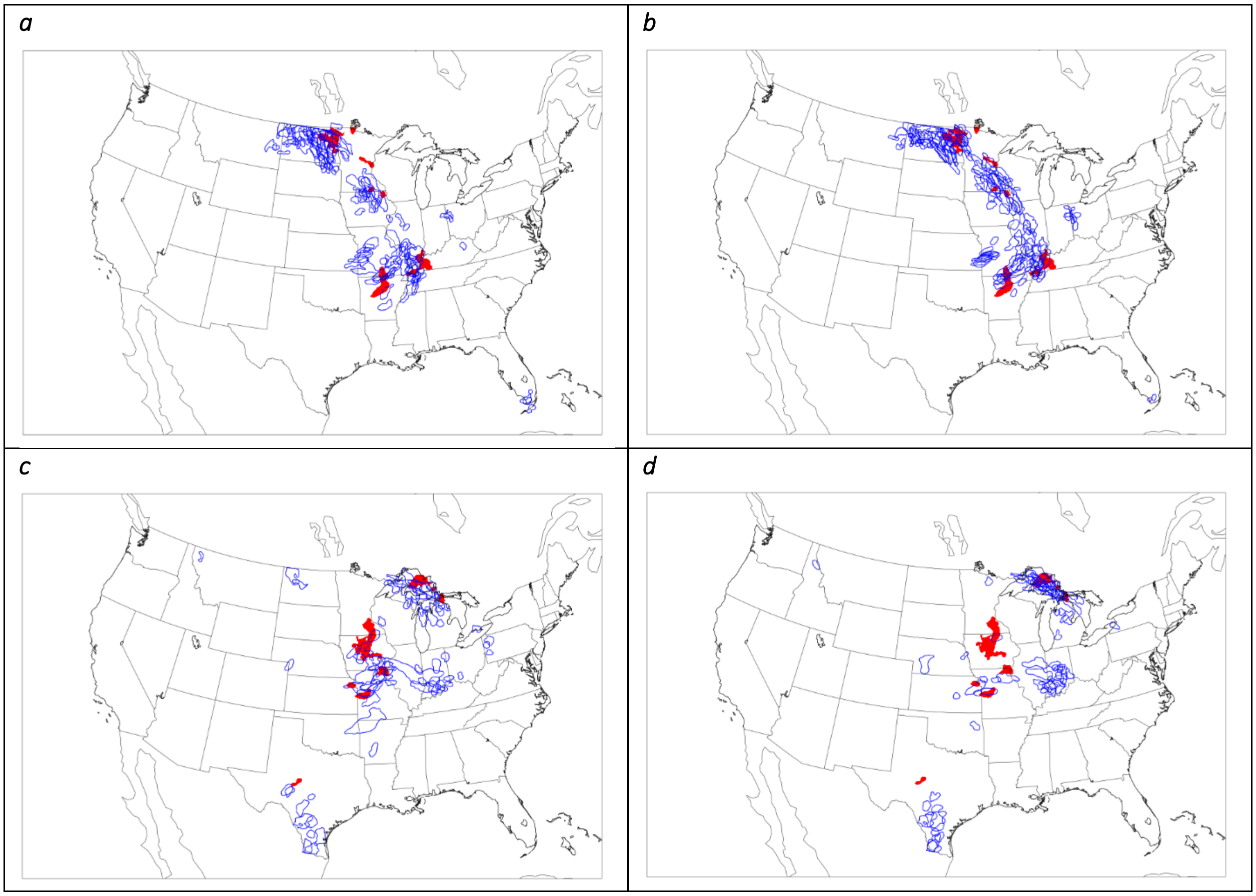


Figure 3. MODE objects identified in the observation field (red shaded) and each ensemble member (blue outlines) for the multi-physics (left column) and single physics (right column) forecasts initialized at 00 UTC on 25 May 2016 at forecast hours 18 (top row) and 33 (bottom row).

When aggregated over the entire set of forecasts available during the experiment, all ensemble members had more forecast objects identified than observed, especially during the afternoon/evening hours (Figure 4a). In addition, several multi-physics ensemble members had more objects than the single physics members and, in general, the multi-physics ensemble had a larger spread in counts compared to the single physics ensemble, which has a tighter cluster of values. All ensemble members follow the temporal maxima and minima of the observed object counts after the initial spin-up time; however, there is an offset in the timing of convective initiation with the models generally being a few hours too early.

When examining the median object area aggregated across all available forecasts (Figure 4b), two more prominent peaks are seen in the observations at times valid around 05 and 13 UTC with a smaller peak near 18 UTC. The overnight/early morning peaks are potentially associated with larger nocturnal MCSs. The distribution of median values for each ensemble members forecast object area is particularly variable during the first half of the forecast with little trend among them; however, during the second half of the forecast the single physics ensemble members tend to have somewhat lower object areas than the multi-physics members. In general, members within both ensemble subsets encompass the median observed object counts for a majority of the forecast lead times.

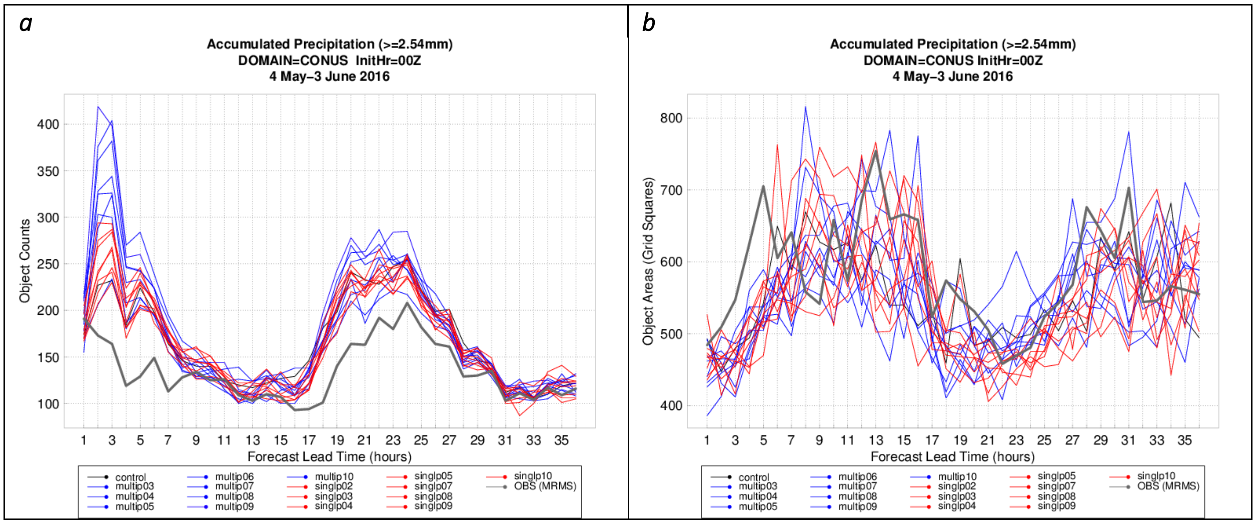


Figure 4. (a) Total count and (b) median area of accumulated precipitation objects > 2.54 mm object counts over the full CONUS domain for all available forecasts during the experiment. The observation objects are in black, the multi-physics members in blue, and the single physics members in red.

Using the centroid attribute from MODE, we can look at general displacement trends for each of the members by computing the centroid distance between forecast and observed accumulated precipitation objects. The west-east displacement for the multi-physics and single physics members is displayed in Figure 5a and Figure 5b, respectively. Negative values indicate a westerly displacement while positive values indicate an easterly displacement. Overall, the multi-physics and single physics members are both displaced to the west for a majority of the forecast period. The north-south displacement for the multi-physics and single physics members is displayed in Figure 5c and Figure 5d, respectively. In this case, the multi-physics ensemble members have larger variability in centroid displacement than the single physics ensemble, although both exhibit a similar temporal trend. For the first 18 hours, the displacement is predominantly to the south. During convective initiation during day 2 there is an indication that a majority of ensemble members from both subsets tend to be displaced a bit too far north. In the later forecast hours, the single physics members in particular trend back towards a southerly displacement again.

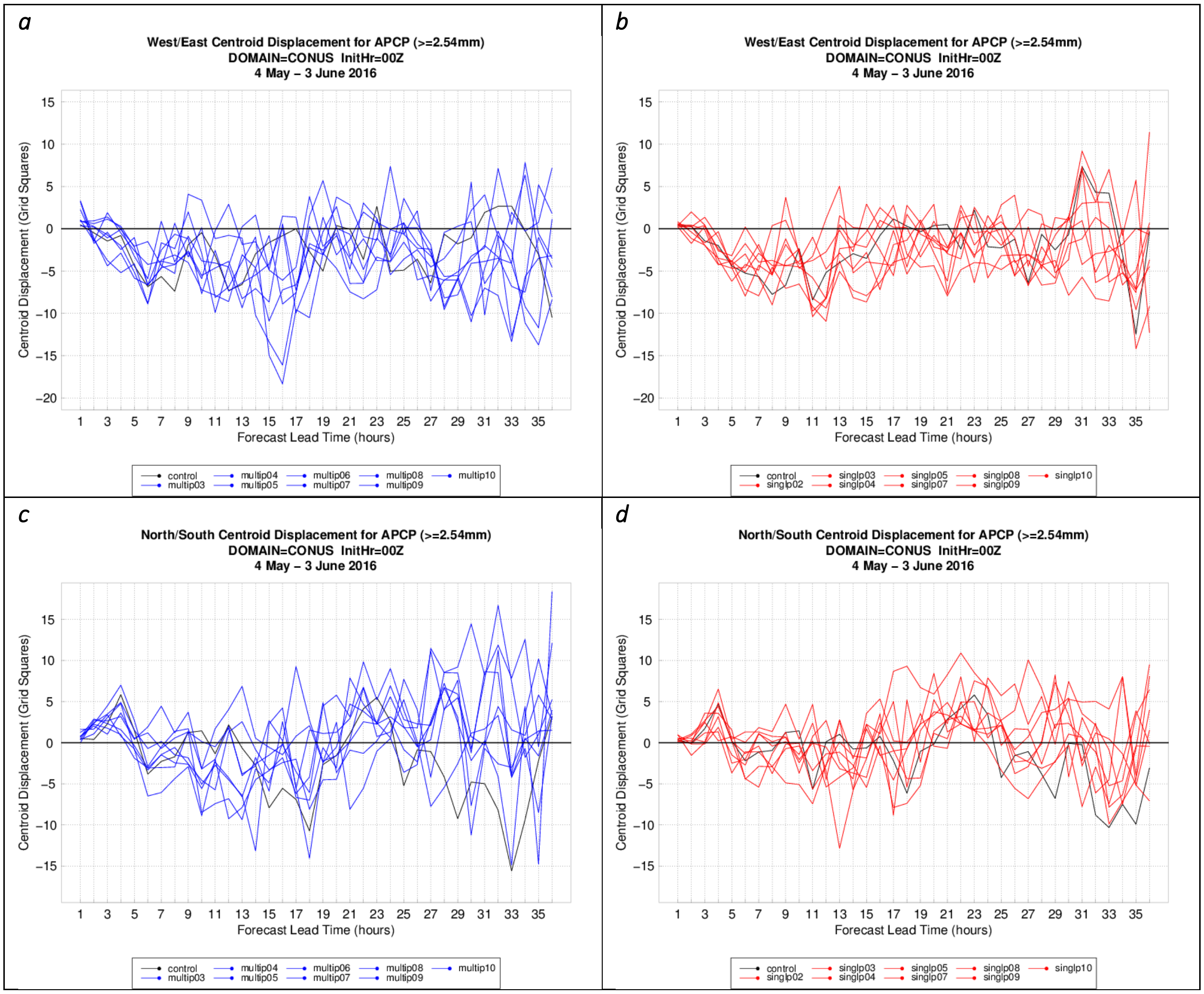
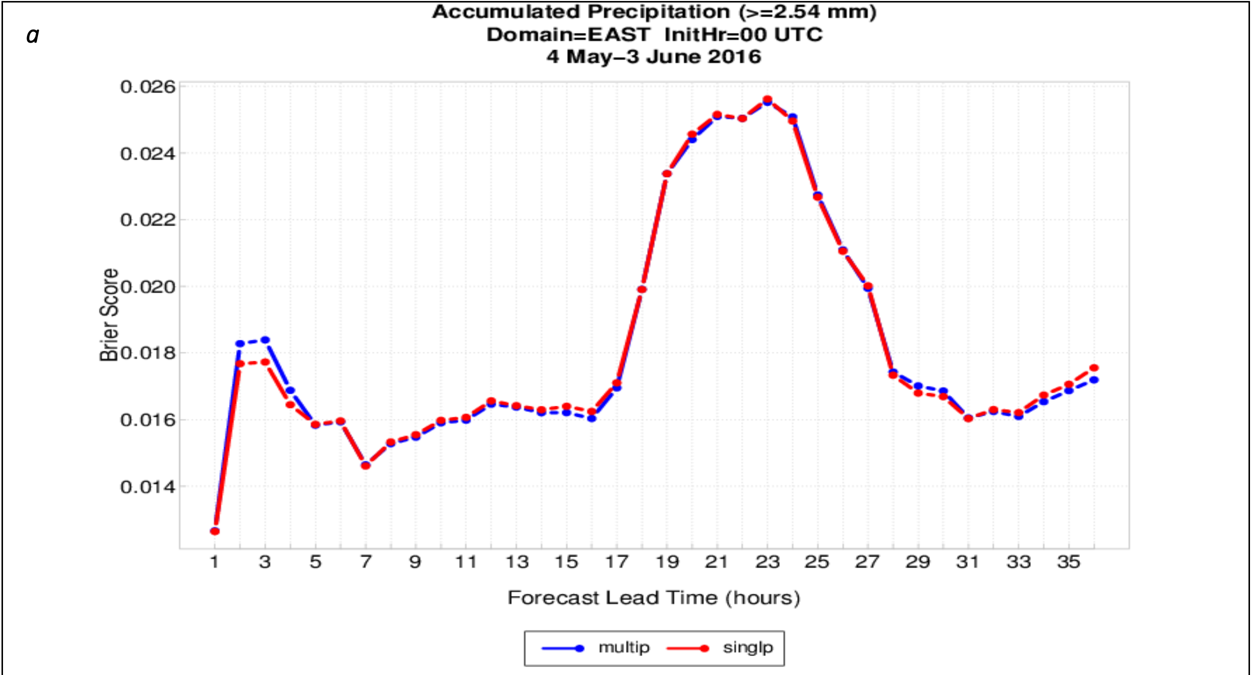


Figure 5. Centroid displacement in the west-east direction (top) and south-north direction (bottom) for the multi-physics (blue; left) and single physics (red; right) ensemble members for accumulated precipitation objects ≥ 2.54 mm aggregated over the full CONUS domain for all available forecasts during the experiment.

To look at the ensemble performance, we first examined the Brier score (BS) for both ensemble subsets as a function of lead time. The BS measures the mean squared probability error that can be partitioned into three terms: reliability, resolution, and uncertainty. A perfect BS is 0, so the lower the BS value, the better. It is worth noting that this metric is sensitive to the climatological frequency of the event so the rarer an event, the easier it is to get a good BS without having any real skill.

Both ensembles perform extremely similarly throughout the 36-hr forecast (Figure 6a) with no meaningful difference in forecast performance between the two when looking at 1-hr accumulated precipitation ≥2.54 mm. We can further examine the ensembles by looking at the components of the BS, the first of which is reliability. The reliability diagram, which is conditioned on the forecasts (i.e., given that an event was predicted, what was the outcome?), can be expected to give information on the real meaning of the forecast probability performance. For this analysis, the reliability diagram was derived from 1-hr accumulated precipitation >2.54 mm aggregated over a 24-hr time period between forecast hours 12-36 (Figure 6b). Both ensembles have very low reliability, near the no skill line, except at the highest forecast probability where the single physics ensemble has slightly higher reliability than the multi-physics ensemble.

The second term of BS is resolution. The ROC curve measures the ability of a forecast to discriminate between two alternative outcomes, thus measuring their resolution. The ROC is conditioned on the observations (i.e., given that an event occurred, what was the corresponding forecast?). For Figure 6c we continue to look at the 12-36 hour forecasts of 1-hr accumulated precipitation >2.54 mm aggregated together and we see that the performance is essentially indistinguishable between the two ensembles. The final term of BS is uncertainty, which in this case is the climatology of the observations in the sample and is equivalent between the two ensemble subsets.



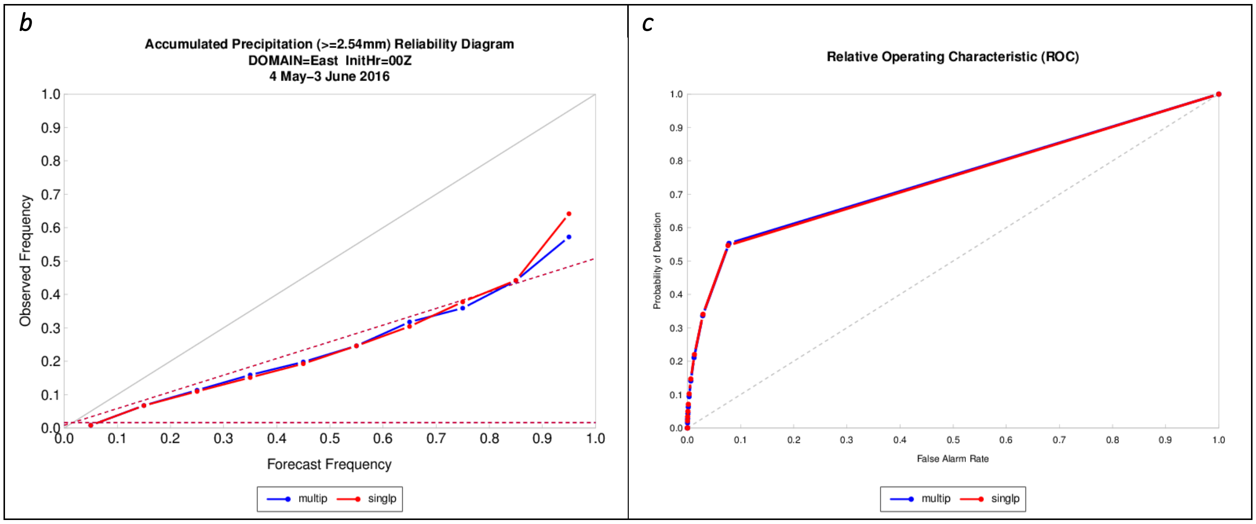


Figure 6. (a) BS time series plots (b) reliability diagram, and (c) ROC curve of 1-h accumulated precipitation >2.54 mm aggregated over a 24-hr time period between forecast hours 12-36 for each ensemble subset over the eastern half of the CONUS domain for all available forecasts during the experiment. The multi-physics ensemble is in blue, and the single physics ensemble in red.

Finally, we examined the rank histogram results for each of the ensembles. The rank histogram measures the frequency of observation in each bin. Forecasts are sorted from highest to lowest then the observations are placed in one of the bins. For this metric, a flat histogram is desirable as it indicates good spread. A u-shaped histogram indicates under-dispersion and an inverse u-shape is indicative of an over-dispersive ensemble. If the histogram is skewed right, the forecast has high bias, and if the histogram is skewed left, the forecast has a low bias.

When aggregating the accumulated precipitation results for the 24-hr period between 12-36 hour forecasts, both the single and multi-physics ensembles display a high bias (right skewed) histogram, meaning the observed values are typically lower than all of the members (Figure 7). There is also a general trend for both ensembles to be under-dispersive as well.

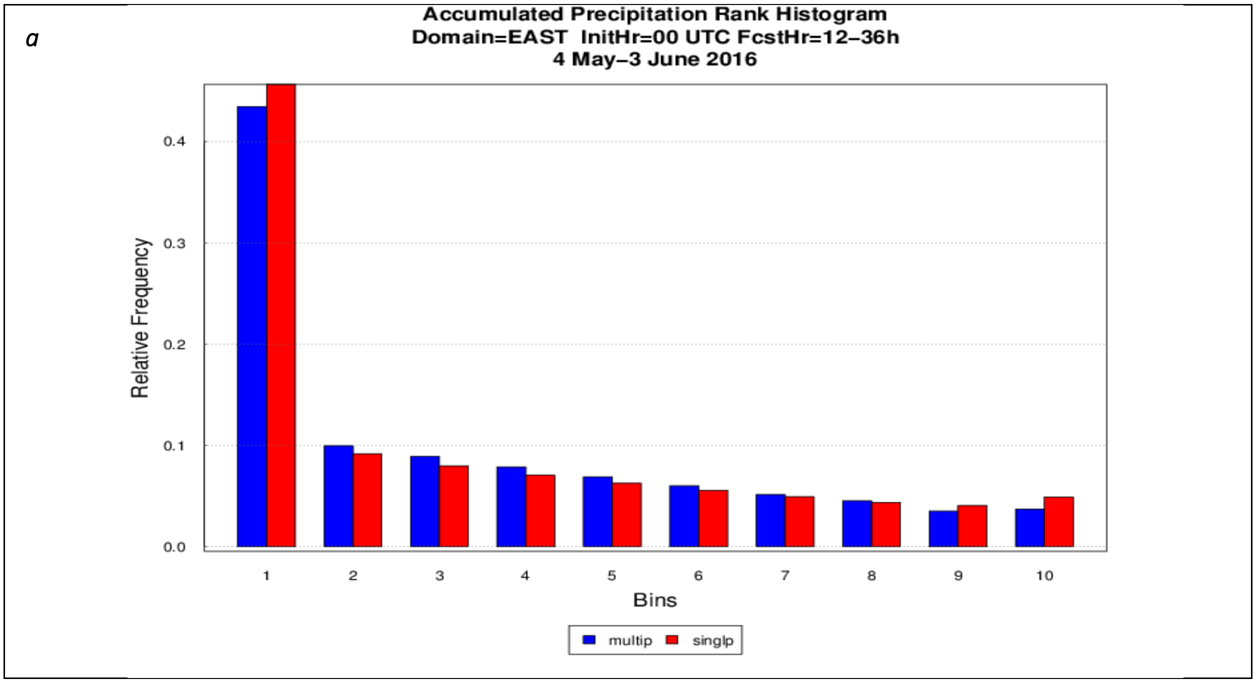


Figure 7. Rank histogram for accumulated precipitation aggregated over a 24-hr time period between forecast hours 12-36 for each ensemble subset over the eastern half of the CONUS domain for all available forecasts during the experiment. The multi-physics ensemble is in blue, and the single physics ensemble in red.

Composite reflectivity

A similar analysis to that done for accumulated precipitation was completed for composite reflectivity. Starting with GSS as a function of lead time, the single physics members are generally clustered at higher GSS values than the multi-physics members, which has a few members that have consistently lower GSS values than the other ensemble members (Figure 8a). The control member (the member that is present in both the multiple and single physics ensembles) is frequently one of the better performing ensemble members.

In terms of frequency bias, it is immediately evident that the multi-physics ensemble has a large variation in bias values between its members, where core04 and core07 (MY microphysics) have a significant high bias and cores 03, 06, and 09 (P3 microphysics) have a low bias (Figure 8b). The single physics members are clustered closer to a value of one; however, they also exhibit a high bias during the afternoon hours when we would except to see more convective initiation.

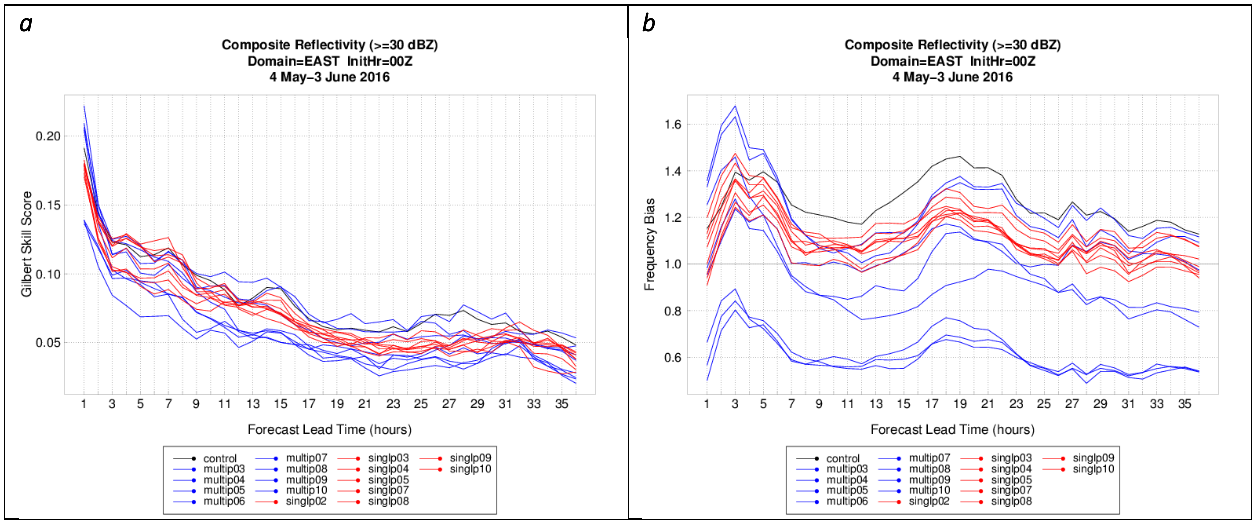


Figure 8. Same asFigure 1 except for composite reflectivity >30 dBZ.

Using FSS to assess composite reflectivity performance, it is evident that while the multi-physics members have more variability in scores, the single physics members tend to be clustered at an overall higher score, regardless of neighborhood size. Again, the control member tends to be one of the better performing members. When looking at scores for composite reflectivity as compared to accumulated precipitation, there tends to be a larger envelope of values for several of the verification metrics, in general.

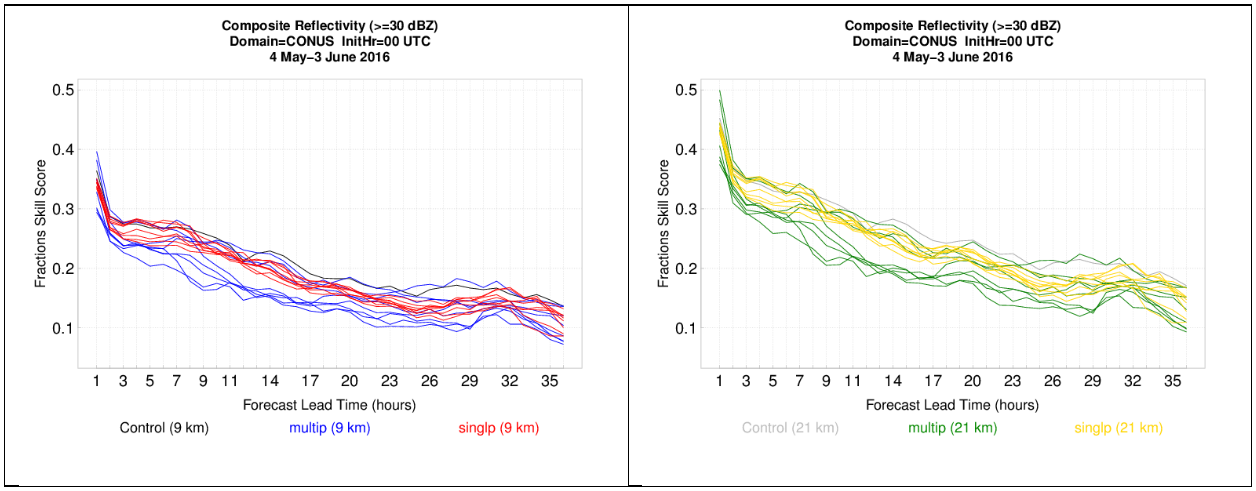


Figure 9. Same as Figure 2 except for composite reflectivity >30 dBZ.

When applying MODE to create composite reflectivity objects using a threshold of >30 dBZ, there is an expected diurnal signal in the observations with low object counts during the overnight/early morning hours and elevated object counts in the afternoon/evening hours during the time of convective initiation (Figure 10a). Overall, the single physics members are generally clustered together and are relatively close to the observation object counts at most forecast lead times. For the multi-physics members, there is a distinct separation in values into three groups depending on the microphysics used in the physics suite. While all members are generally able to capture the peaks and valleys seen in the observations, cores 04 and 07 severely overestimate the number of objects. In contrast, core members 03, 06, and 09 tend to have too few objects. This difference leads to a large variability in the multi-physics object counts. The observed object counts always fall within the range of the multi-physics member solutions, although it sometimes falls outside the narrower solution spread of the single physics members.

When examining object area (Figure 10b), a pronounced peak can be seen in the observations at times valid at 05 and 11 UTC, possibly related to the stronger nocturnal MCSs that persist through that time period. A smaller increase is observed around 20 UTC, associated with strengthening daytime convection. Looking at the ensemble members, there is some spread between them, but all members generally have similar distributions to the observations, with larger object areas present during the overnight/morning hours and smaller areas during the late morning through early afternoon. Overall, the ensemble members have difficulty capturing the overnight peaks in observed object areas. It is also interesting to note the loose relationship between object counts and object areas: when the counts are high, during the afternoon and evening, the object areas are generally lower and vice-versa.

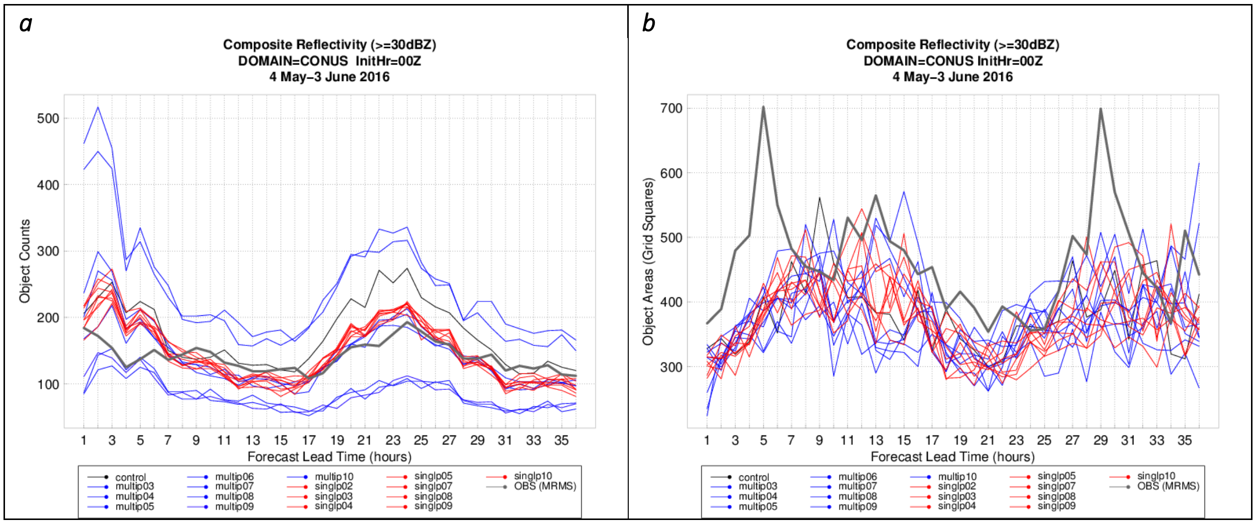


Figure 10. Same as Figure 4 except for composite reflectivity ≥30 dBZ.

Overall, the displacement in composite reflectivity objects ≥30 dBZ is similar to the result for accumulated precipitation objects ≥2.54 mm discussed above. Members of both ensemble subsets frequently have displacement to the west and south, especially in the first half of the forecast period (Figure 11).

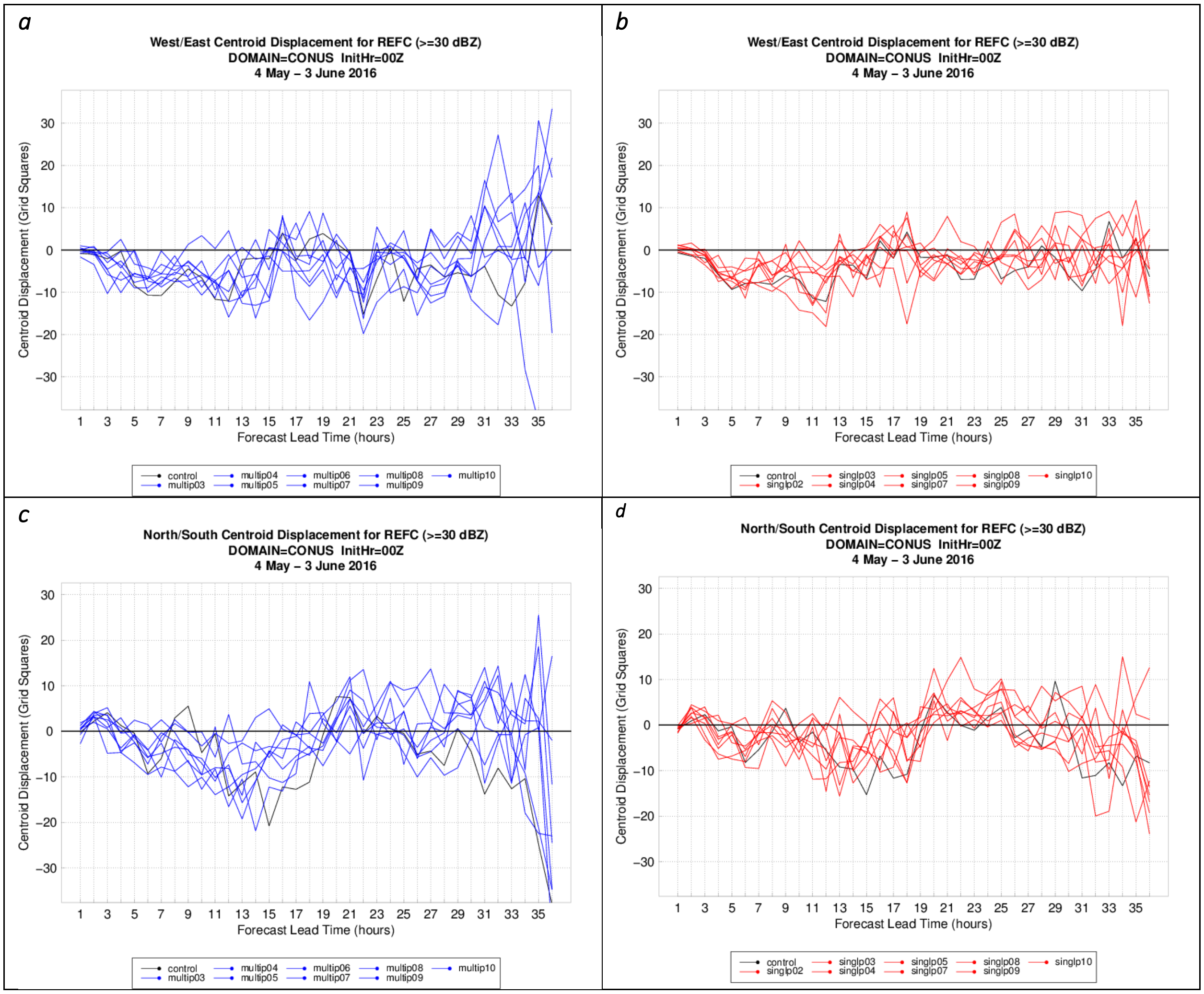


Figure 11. Same as Figure 5 except for composite reflectivity ≥30 dBZ.

While still very similar, the median BS values of the multi-physics ensemble are lower (better) at most forecast lead times compared to the single physics ensemble (Figure 12a). Very little difference is noted between the two ensembles when examining composite reflectivity using a reliability diagram (very little skill at ≥30 dBZ but significant improvement for both ensembles at the ≥20 dBZ threshold – not shown) and ROC curve (Figure 12a/b).

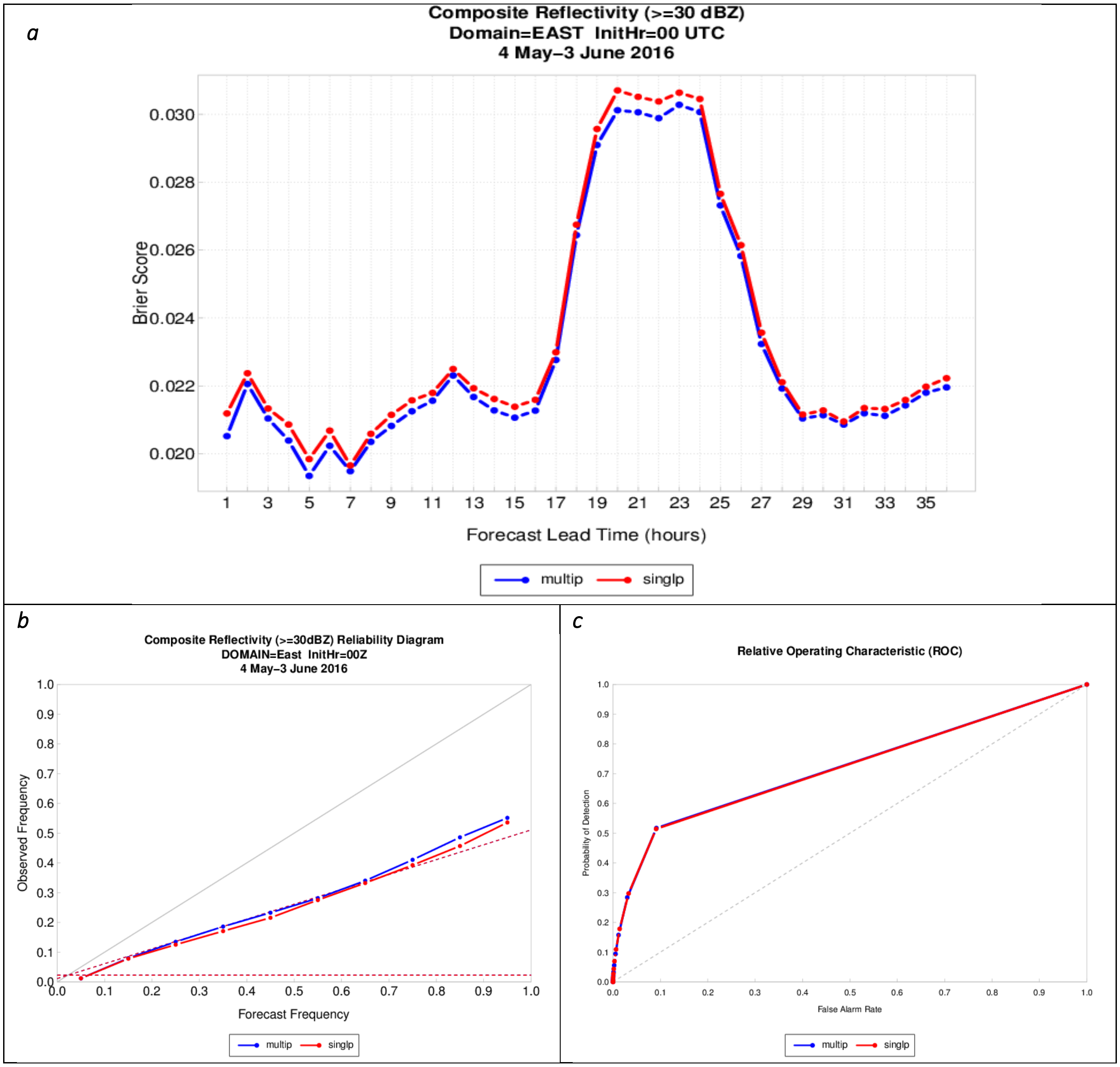


Figure 12. Same as Figure 6 except for composite reflectivity ≥30 dBZ.

Finally, we examined the rank histogram results for composite reflectivity, aggregating the results for the 24-hr period between 12-36 hour forecasts (Figure 13). The single physics displays an under-dispersive ensemble and also exhibits a somewhat high bias. The multi-physics has a large high bias and has an interesting discontinuity at the center of the histogram. This discontinuity may be due to the fact that the members within the multi-physics ensemble have very different behaviors between them with a couple very high and a few low leading to this distribution.

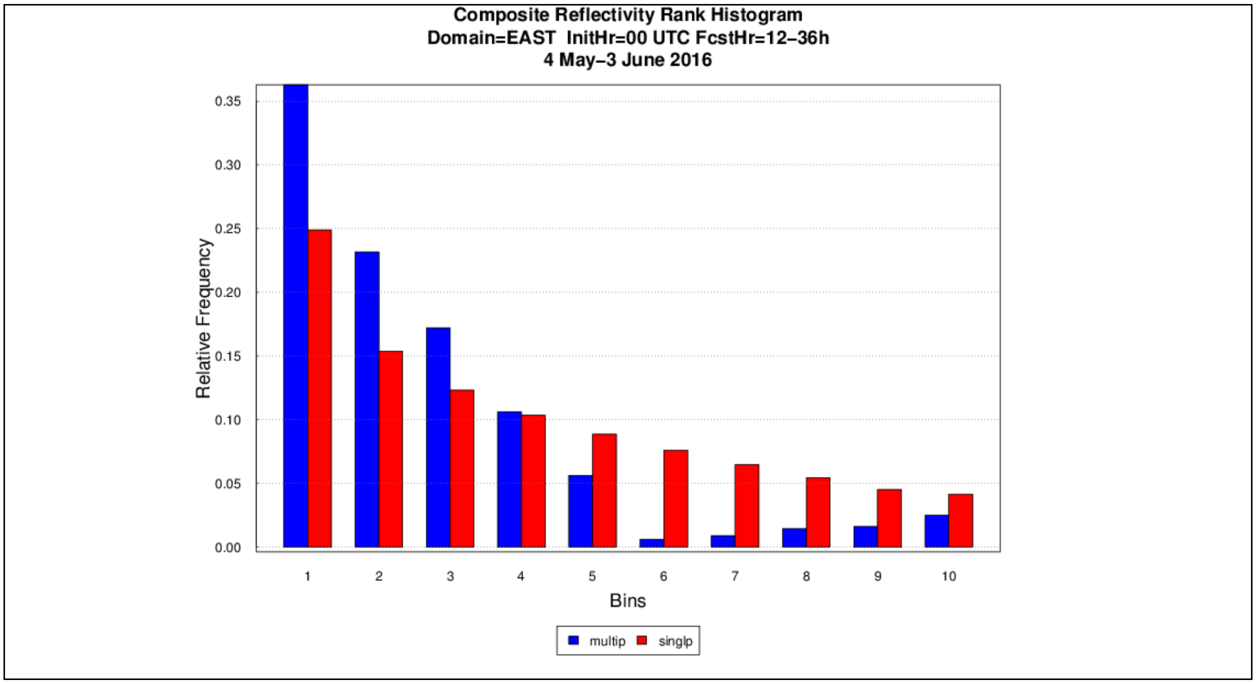


Figure 13. Same as Figure 7 except for composite reflectivity ≥30 dBZ.

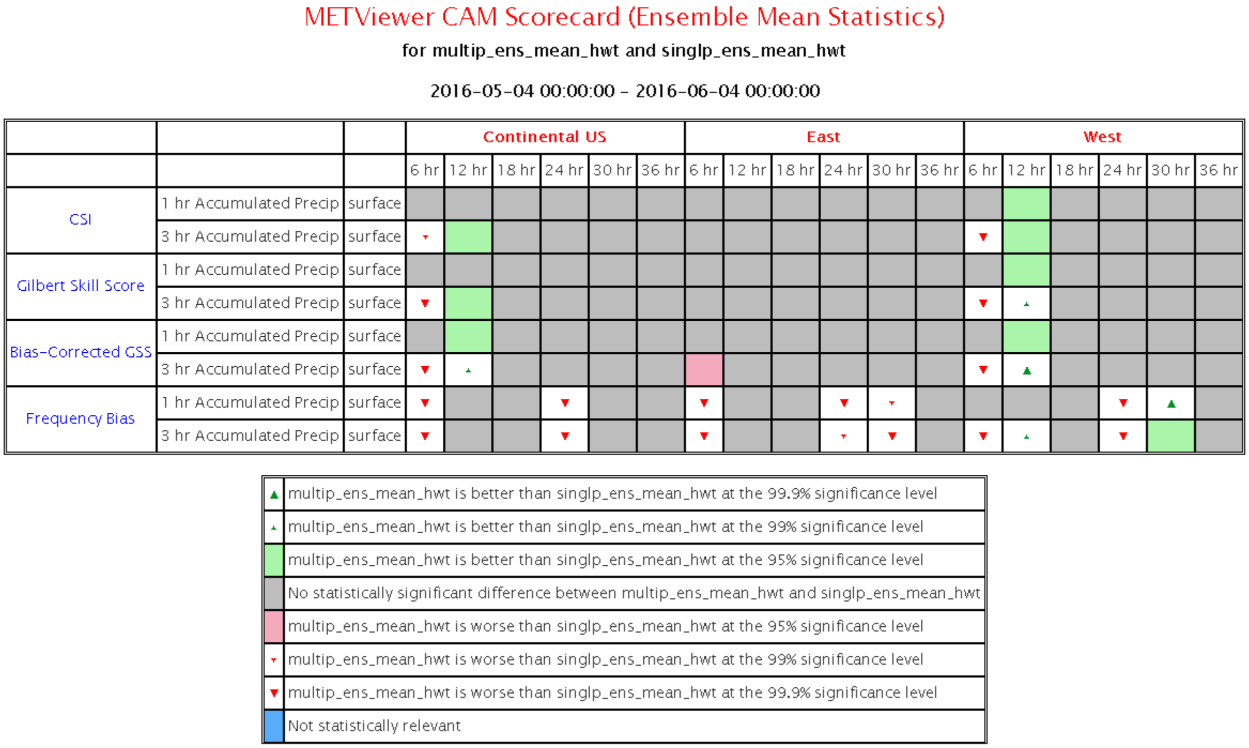
Scorecard verification

An enormous amount of information is available from these kinds of retrospective analyses, especially when numerous verification metrics are evaluated for multiple ensembles. In order to consolidate this information in a concise way, a quantitative comparison of scores can be compiled in a scorecard, providing a general overview of metrics and easy comparison between different ensembles. In addition, statistical significance can be incorporated into the scorecard, to allow for different levels of certainty regarding differences found between ensemble scores.

Given the ongoing development of a number of convective-allowing ensembles, and the need to compare these systems, the DTC has developed a preliminary ensemble scorecard using the MET framework. Figure 14 shows an example of this scorecard for the 2016 HWT data comparing the multi- and single-physics ensembles. For the purpose of this comparison, ensemble mean values were compared for Critical Success Index (CSI), GSS, bias-corrected GSS, and frequency bias.

For the majority of metrics, domains, and lead times, the multi- and single-physics ensembles had differences in 1- and 3-h precipitation accumulation scores that were statistically insignificant (grey boxes). An exception to this rule was for 3-h precipitation accumulation scores for the western US (and the CONUS as well), where 6-hour forecasts indicated the multi-physics ensemble was statistically better for all metrics than the single-physics ensemble. However, the opposite was true at the 12-hour lead time, with the single-physics ensemble showing a number of improved scores at a 95-99% significance level when compared to the multi-physics ensemble.

Currently, these scorecards only compare metrics using ensemble mean values. However, one of the main conclusions that came from the NCEP Ensemble User's Workshops was the need for a convective-allowing ensemble scorecard which includes traditional probabilistic measures, especially as efforts ramp up to choose a replacement for the HREF v2. With this requirement in mind, work is continuing to advance the MET scorecard utility to include these probabilistic measures, and will be available in a future release of MET.



METViewer CAM Scorecard (Ensemble Mean Statistics)

for multip_ens_mean_hwt and singlp_ens_mean_hwt

2016-05-04 00:00:00 - 2016-06-04 00:00:00

			Continental US						East						West					
			6 hr	12 hr	18 hr	24 hr	30 hr	36 hr	6 hr	12 hr	18 hr	24 hr	30 hr	36 hr	6 hr	12 hr	18 hr	24 hr	30 hr	36 hr
CSI	1 hr Accumulated Precip	surface	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	🟩	▒	▒	▒	▒
	3 hr Accumulated Precip	surface	▾	🟩	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▼	🟩	▒	▒	▒	▒
Gilbert Skill Score	1 hr Accumulated Precip	surface	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	🟩	▒	▒	▒	▒
	3 hr Accumulated Precip	surface	▼	🟩	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▼	▴	▒	▒	▒	▒
Bias-Corrected GSS	1 hr Accumulated Precip	surface	▒	🟩	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	🟩	▒	▒	▒	▒
	3 hr Accumulated Precip	surface	▼	▴	▒	▒	▒	▒	🟥	▒	▒	▒	▒	▒	▼	▲	▒	▒	▒	▒
Frequency Bias	1 hr Accumulated Precip	surface	▼	▒	▒	▼	▒	▒	▼	▒	▒	▼	▾	▒	▒	▒	▒	▼	▲	▒
	3 hr Accumulated Precip	surface	▼	▒	▒	▼	▒	▒	▼	▒	▒	▾	▼	▒	▼	▴	▒	▼	🟩	▒

▲	multip_ens_mean_hwt is better than singlp_ens_mean_hwt at the 99.9% significance level
▴	multip_ens_mean_hwt is better than singlp_ens_mean_hwt at the 99% significance level
🟩	multip_ens_mean_hwt is better than singlp_ens_mean_hwt at the 95% significance level
▒	No statistically significant difference between multip_ens_mean_hwt and singlp_ens_mean_hwt
🟥	multip_ens_mean_hwt is worse than singlp_ens_mean_hwt at the 95% significance level
▾	multip_ens_mean_hwt is worse than singlp_ens_mean_hwt at the 99% significance level
▼	multip_ens_mean_hwt is worse than singlp_ens_mean_hwt at the 99.9% significance level
🟦	Not statistically relevant

Figure 14. Precipitation accumulation scorecard comparing scores for the multi-physics and single-physics ensemble mean. Scores are broken out among multiple domains and different precipitation accumulation periods, with arrows and colors indicating different levels of statistical significance.

Summary

Detailed evaluations of the probabilistic performance of two targeted subset ensembles of the CLUE super ensemble from the 2016 Springe Forecast Experiment are presented along with the results of deterministic forecasts from the ensemble individual members.

The benefit to a single physics ensemble over a multi-physics ensemble are that underlying model climatologies are similar for calibration purposes, and thus easier to maintain. While the single physics ensemble displays less ensemble spread overall, the skill between the multiple and single physics is comparable.

Accumulated Precipitation Findings

Both single physics and multi-physics ensemble members over- and under-forecast 1-hour accumulated precipitation at different forecast hours. Both ensemble members tend to be clustered closer together during the early forecast hours and have more skill at these times (ex. figures 1a, 2a, 5c). Ensemble spread generally remains the same for both ensembles across all forecast hours except for object displacement (figure 5) where it increases with forecast hour. While both the multi-physics and single physics ensembles exhibit extremely similar trends for this variable across most metrics, the single physics ensemble tends to have a tighter ensemble spread overall.

Specifically examining the MODE object counts and object displacement for the 1-hour accumulated precipitation, the object counts are typically too high before observed convective initiation. This coincides with results from the traditional metric frequency bias, where there is a high bias. There is a westward displacement at similar times for both ensembles, signaling that in addition to initiating too early, the objects are initiating too far west.

When examining the CAM Scorecard values, the majority of the differences are not statistically significant. However, it is apparent that the single physics ensemble mean outperforms the multi-physics ensemble mean when the differences are statistically significant. Of the 33 occurrences when the differences are statistically significant, the single physics ensemble mean performs better 20 of those times.

Composite Reflectivity Findings

Both the multi-physics and single physics ensembles over- and under-forecast composite reflectivity depending on the forecast hour and individual ensemble member. Although both ensembles exhibit similar temporal trends across most metrics (ex. figures 8, 9, and 11), the single physics ensemble generally has a smaller spread and is closer to the observed object counts. Unlike hourly accumulated precipitation, the single physics members have more overall skill for all metrics except for the overall Brier Score and associated reliability diagram.

Model Evaluation Tools

This comprehensive evaluation, which consisted of both traditional and probabilistic metrics as well as more advanced spatial metrics, was completed using the verification tools available in the MET package. This evaluation highlights the spectrum to tools within the powerful MET package – from preprocessing and re-gridding data to outputting a range of traditional and more advanced metrics, all available within one software package.

References

Clark, A., I. Jirak, C. Melick, J. Correia, S. J. Weiss, J. Kain, A. Dean, K. Knopfmeier, C. Karstens, and B. Twiest, 2016: Spring Forecasting Experiment 2016 Program Overview and Operations Plan. Available at: https://hwt.nssl.noaa.gov/Spring_2016/HWT_SFE2016_operations_plan_final.pdf

Developmental Testbed Center, 2017: MET: Version 6.0 Model Evaluation Tools Users Guide. Available at http://www.dtcenter.org/met/users/docs/overview.php. 348 pp.

Zhang, J., K. Howard, C. Langston, B. Kaney, Y. Qi, L. Tang, H. Grams, Y. Wang, S. Cocks, S. Martinaitis, A. Arthur, K. Cooper, J. Brogden, and D. Kitzmiller, 2016: Multi-Radar Multi-Sensor (MRMS) quantitative precipitation estimation: Initial operating capabilities. *Bull. Amer. Meteor. Soc.*, **97**, 621–638, doi:10.1175/BAMS-D-14-00174.1.