Assessment of ensemble forecast performance for select members available in the CLUE during HWT-SFE 2017 and comparison to CLUE 2016 subsets

Final Report

REGIONAL ENSEMBLE TEAM: JAMIE WOLFF, LINDSAY BLANK, AND MICHELLE HARROLD

NATIONAL CENTER FOR ATMOSPHERIC RESEARCH (NCAR) RESEARCH APPLICATIONS LABORATORY/JOINT NUMERICAL TESTBED PROGRAM AND DEVELOPMENTAL TESTBED CENTER

Table of Contents

Introduction

The analysis of select subsets included in the Community Leveraged Unified Ensemble (CLUE) dataset from the 2017 Hazardous Weather Testbed (HWT) Spring Forecasting Experiment (SFE) is presented in this document as well as a comparison of performance to the CLUE 2016 data. For more background on the motivation to collaborate with the HWT-SFE and specific information regarding the verification approaches used and the CLUE 2016 ensemble subset performance, please reference the Annual Operating Plan 2017 final report, located at https://dtcenter.org/eval/ensembles/hwt_collab/RE5_HWT_report_FINAL.pdf. For more information on HWT-SFE 2017, the program overview and operations plan can be found at https://hwt.nssl.noaa.gov/Spring_2017/HWT_SFE2017_operations_plan_FINAL.pdf.

CLUE 2017 Dataset

A key difference between the CLUE 2017 and CLUE 2016 data was a change to the land surface model (LSM) and planetary boundary layer (PBL) physics parameterizations used for the control member. While Thompson microphysics was used in both 2016 and 2017, the LSM was modified from Noah in 2016 to RUC in 2017 and the PBL was changed from MYJ in 2016 to MYNN in 2017. In addition, CLUE 2017 also included a 10-member stochastic physics ensemble. The physics suites for each 2017 ensemble subset used in this analysis are presented in Table 1.

*Table 1. Physics suite description for CLUE 2017 subsets examined. A * indicates that component has been stochastically perturbed.*

Multi-physics					
Member	**IC**	**BC**	**Microphysics**	**LSM**	**PBL**
core01	**NAMa+3DVAR**	**NAMf**	**Thompson**	**Noah**	**MYJ**
core02 (control)	**RAPa+3DVAR**	**GFSf**	**Thompson**	**RUC**	**MYNN**
core03	**core01+arw-p1_pert**	**arw-p1**	**P3**	**Noah**	**YSU**
core04	**core01+arw-n1_pert**	**arw-n1**	**MY**	**Noah**	**MYNN**
core05	**core01+nmmb-p1_pert**	**nmmb-p1**	**Morrison**	**Noah**	**MYJ**
core06	**core01+nmmb-n1_pert**	**nmmb-n1**	**P3**	**Noah**	**YSU**
core07	**core02+arw-p2_pert**	**arw-p2**	**MY**	**Noah**	**MYNN**
core08	**core02+arw-n2_pert**	**arw-n2**	**Morrison**	**Noah**	**YSU**
core09	**core02+nmmb-p2_pert**	**nmmb-p2**	**P3**	**Noah**	**MYJ**
core10	**core02+nmmb-n2_pert**	**nmmb-n2**	**Thompson**	**Noah**	**MYNN**
Single Physics + IC/BC pert (10 members)					
Member	**IC**	**BC**	**Microphysics**	**LSM**	**PBL**
single-phys01 (control)	**RAPa+3DVAR**	**GFSf**	**Thompson**	**RUC**	**MYNN**
Stochastic Physics + IC/BC pert (10 members)					
Member	**IC**	**BC**	**Microphysics**	**LSM***	**PBL***
stoch-phys01 (control)	**RAPa+3DVAR**	**GFSf**	**Thompson**	**RUC**	**MYNN**

The 2017 HWT-SFE was held from 1 May – 2 Jun, 2017. 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. It is important to note that not all ensembles or individual ensemble members have data for every day during the SFE. An inventory (Table 2) of the available data

revealed that the multiple physics ensemble has data for every day during the forecast period. The single physics ensemble has data from 10 May – 2 Jun, 2017, while the stochastic physics ensemble has the least amount of data available from 10 – 24 May 2017.

Event equalization was applied to the ensemble subsets for the 2017 analysis due to the difference in data available between the datasets. This means that only data from 10 – 24 May 2017 was examined for all three subsets. It should be noted that even during that focused time period some members were missing data for individual forecast hours or days (Table 2); however, the dataset was not further limited for those additional data outages. Event equalization does not apply when comparing the datasets across the two years of the experiment; this should be kept in mind when interpreting the results in that section.

Table 2 Data inventory by forecast hour for ensemble subset members.

Multi-physics Data Inventory		
Date	core04	core07
20170510		all
20170511	fhr13-36	all
20170512		all
20170515		all

Single Physics Data Inventory				
Date	single-phys04	single-phys05	single-phys07	single-phys09
20170522	all	all	all	
20170523	all	all	all	fhr33-36

Stochastic Physics Data Inventory				
Date	stoch-phys02	stoch-phys07	stoch-phys08	stoch-phys09
20170510		all		
20170516				fhr25-36
20170522			all	
20170523	all			

Verification results

Accumulated precipitation

The Gilbert Skill Score (GSS) for each individual ensemble member was calculated by forecast lead time (Figure 2a). A perfect GSS would be a value of 1, so the larger the value, the better the member performed. As in the previously conducted CLUE 2016 analysis, the focus of this evaluation is on aggregated hourly accumulated precipitation that is greater than or equal to 2.54 mm over the eastern half of the CONUS domain for all available forecasts during the experiment (Figure 1). All ensembles have their maximum GSS values at forecast hour 1 and an exponential decrease in skill is noted for the remainder of the period. The variability in GSS values for each ensemble subset grows with forecast lead time but no one particular subset has consistently higher values than the others; all three ensemble subsets have comparable GSS.

Frequency bias was also calculated for every ensemble member by forecast lead time (Figure 2b). A frequency bias value of 1 would be a perfect value, with values less than 1 an under-forecast and values over 1 an over-forecast. A modest diurnal signal is apparent across all three ensembles. The members generally over-forecast during the first 14 hours of the forecast, transitioning to values distributed around neutral forecasts during the latter portions of the forecast. The multi-physics ensemble members tend to have the largest variability between members, with individual members both over- and under-forecasting at some forecast hours. The single and stochastic physics ensemble members frequently have a smaller envelope of GSS values than multi-physics.

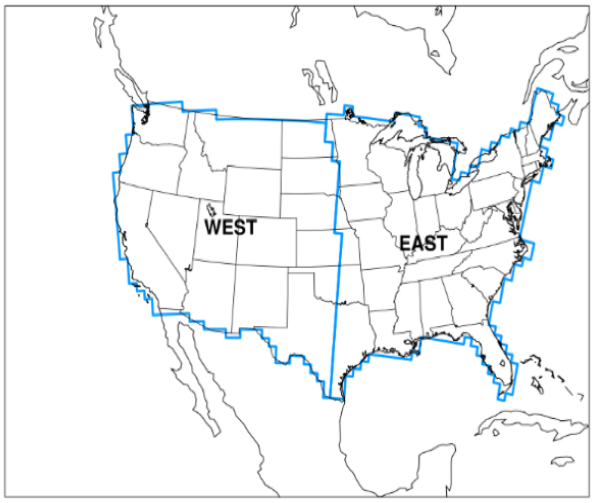


Figure 1: Verification sub-regions.

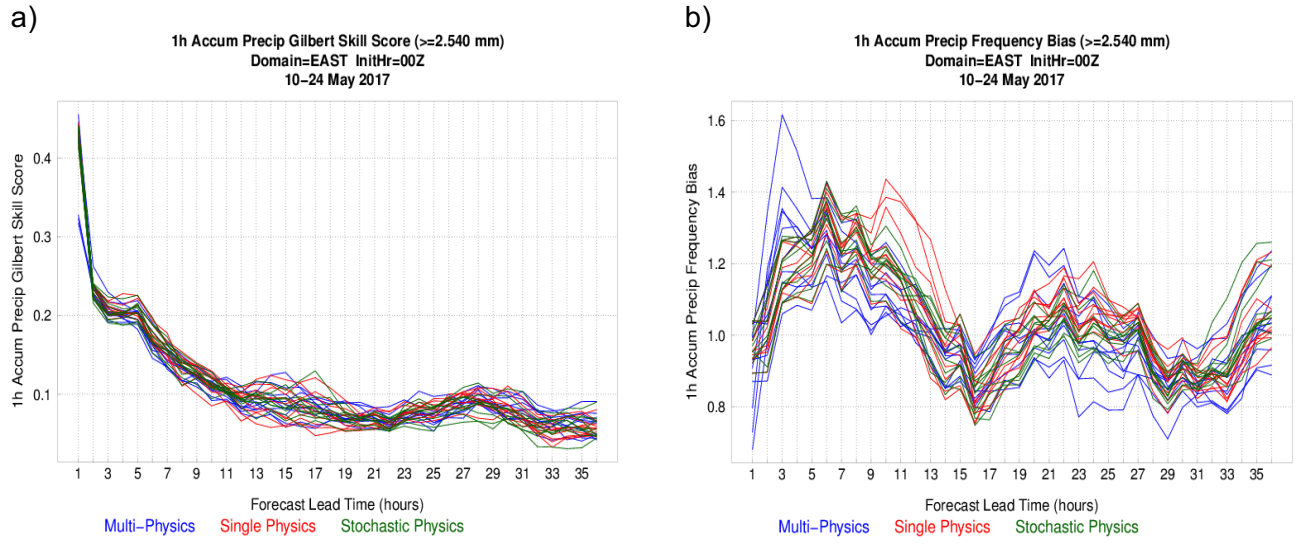


Figure 2: (a) GSS and (b) frequency bias time series plots of 1-h accumulated precipitation ≥ 2.54mm for each individual ensemble member aggregated across the eastern half of the CONUS domain for all available forecasts during the experiment. The control members used in each ensemble is shown by the dashed line, the multi-physics members are in blue, the single physics members are in red, and the stochastic physics members are in green.

Fractions Skill Score (FSS) was calculated as a function of forecast lead time for two neighborhood widths: 3x3 grid squares, or 9x9 km (Figure 3a), and 7x7 grid squares, or 21x21 km (Figure 3b). As with other skill score metrics, the higher the FSS value, the more skillful the member – with a maximum possible value of 1. FSS for both the 3x3 and 7x7 neighborhood widths display similar trends with lead time as the GSS behavior (Figure 2a). All three ensemble subsets have similar scores and variability among members for each corresponding neighborhood width. When comparing the performance between the two widths there is a shift towards higher scores for the larger neighborhoods indicating that as the spatial scale is broadened the members generally perform better. This suggests that the members generally have precipitation in the vicinity of the observational analyses, though there is some tendency for displacement error.

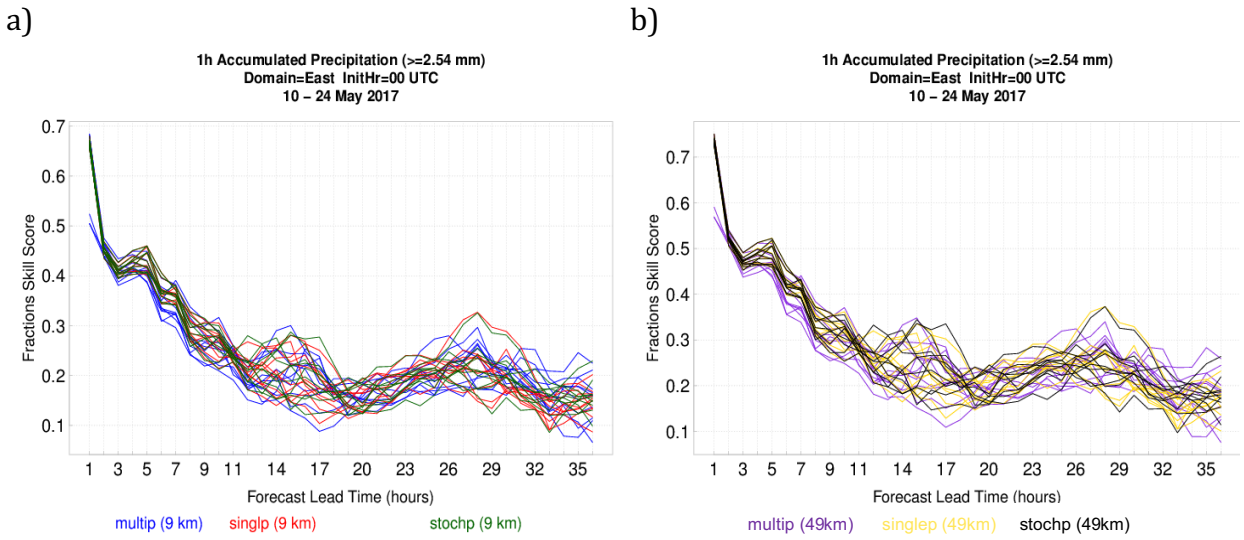


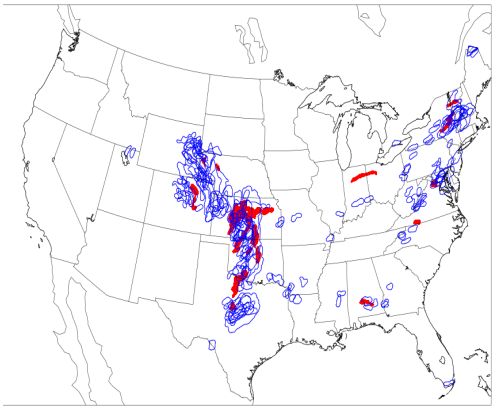
Figure 3: FSS time series plots of 1-h accumulated precipitation ≥ 2.54mm for each individual ensemble member at a neighborhood width of (a) 3x3 grid squares and (b) 7x7 grid squares aggregated across the eastern half of the CONUS domain for all available forecasts during the experiment. The multi-physics members are represented in blue (3x3) and purple (7x7), the single physics members in red (3x3) and yellow (7x7), and the stochastic physics members in green (3x3) and black (7x7).

The Method for Object-based Diagnostic Evaluation (MODE) capability of the Model Evaluation Tools (MET) software was used to identify objects of 1-hour accumulated precipitation totals of ≥ 2.54 mm. Figure 4 presents these objects for each ensemble member for the 00 UTC initialization on 18 May 2017. This date was an active convective day, featuring several supercell clusters that developed into convective squall lines with trailing stratiform across the Great Plains. A majority of the reports were oriented southwest to northeast from Texas through Illinois. This event produced 45 tornado, 246 wind, and 138 hail preliminary filtered storm reports (acquired from the SPC).

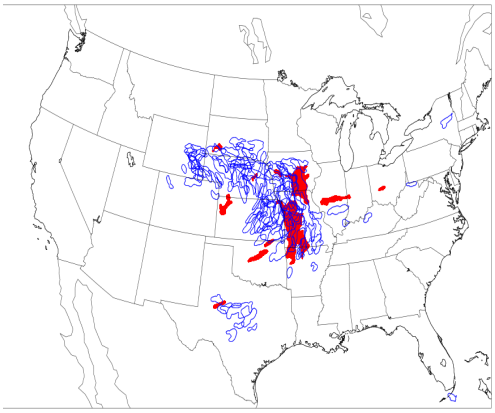
All three ensemble members capture the general placement of the observed convective objects at forecast hour 24 (Figure 4a, c, and e), though the multi-physics members had the fewest false alarms from east Texas into Louisiana/Arkansas at this time. At forecast hour 32 (Figure

4b, d, and f), the multi-physics ensemble clearly handles the eastern extent of the placement of the main precipitation line better than both the single and stochastic physics ensemble members. The majority of the single and stochastic ensemble objects are displaced to the west of the main convection at this time and there appears to be a stronger indication for additional convective initiation to the east of the main line for those two ensemble subsets as well.

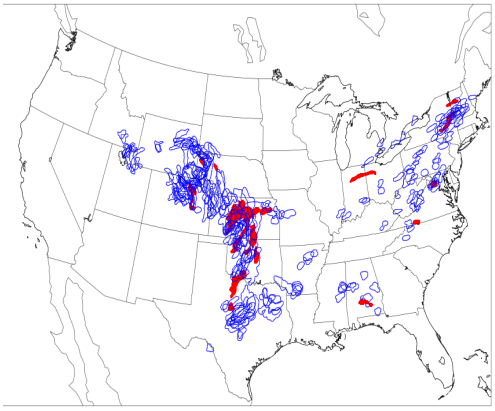
a)



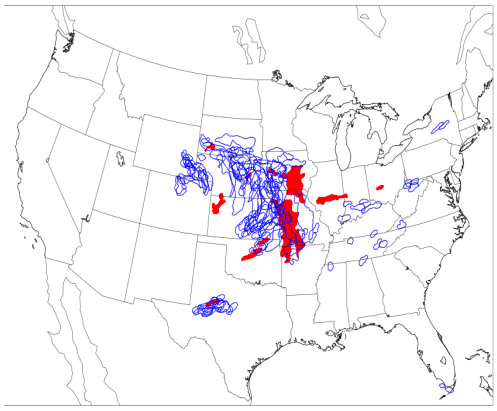
b)



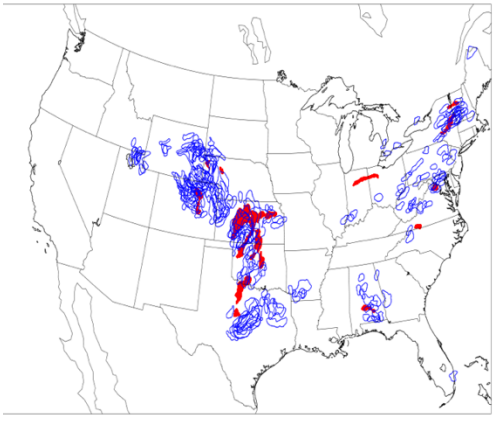
c)



d)



e)



f)

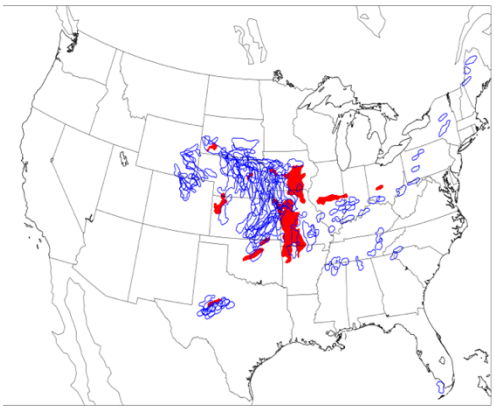


Figure 4: MODE objects identified in the observation field (red shaded) and each ensemble member (blue outlines) for the multi-physics (top row), single physics (middle row), and stochastic physics (bottom row) forecasts initialized at 00 UTC on 18 May 2017 at forecast hours 24 (left column) and 32 (right column).

The total number of MODE objects across the available set of forecasts is presented (Figure 5a). The observed number of objects is represented by the black line. A clear diurnal signal is apparent in both the observed and forecast object counts. All of the ensemble subsets capture the observed object counts reasonably well with exception to a few members in the multi- and single-physics ensembles that trend towards too few objects through the forecast period. The multi-physics ensemble subset has several members that trend on the lower side of the object counts. While some single physics members are also low throughout, there are also several that tend to be on the high side of the counts. The stochastic physics subset tends to be the most clustered and consistently closest to the observed number of objects.

The median object area by member aggregated across all available forecast times is provided in Figure 5b. Again, the observed object area is represented with a black line. While members of all three ensemble subsets frequently have smaller object areas compared to the observations, some multi-physics, and a few single physics, members fall the closest to observations. While there are forecast hours where the majority of ensemble subset members have similar object area to the observed objects, it is frequently observed throughout the forecast period that the ensemble members from each of the subsets yield smaller object areas than observed.

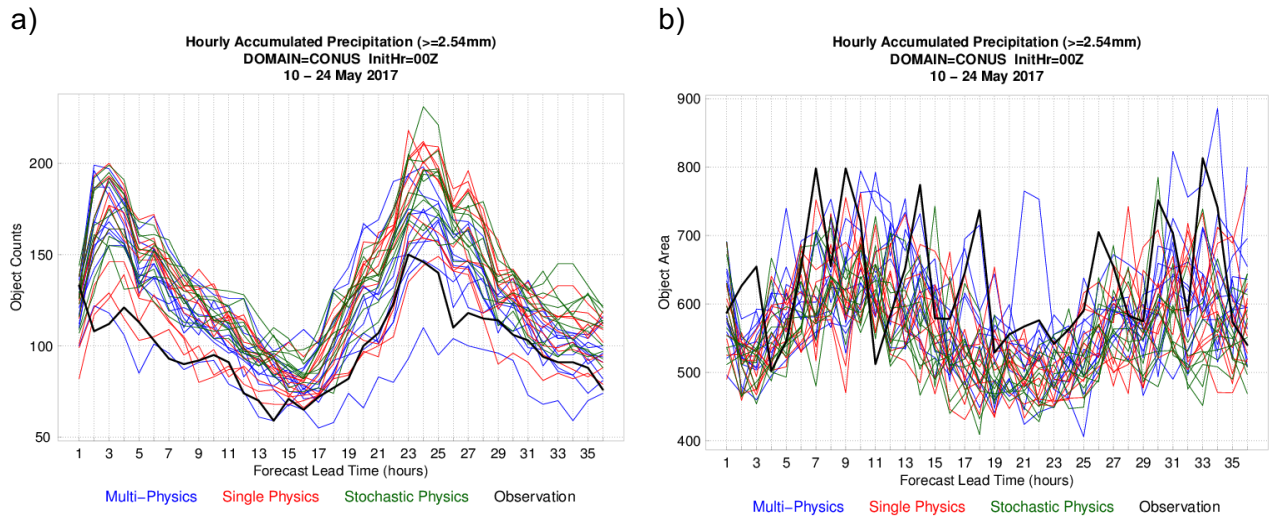
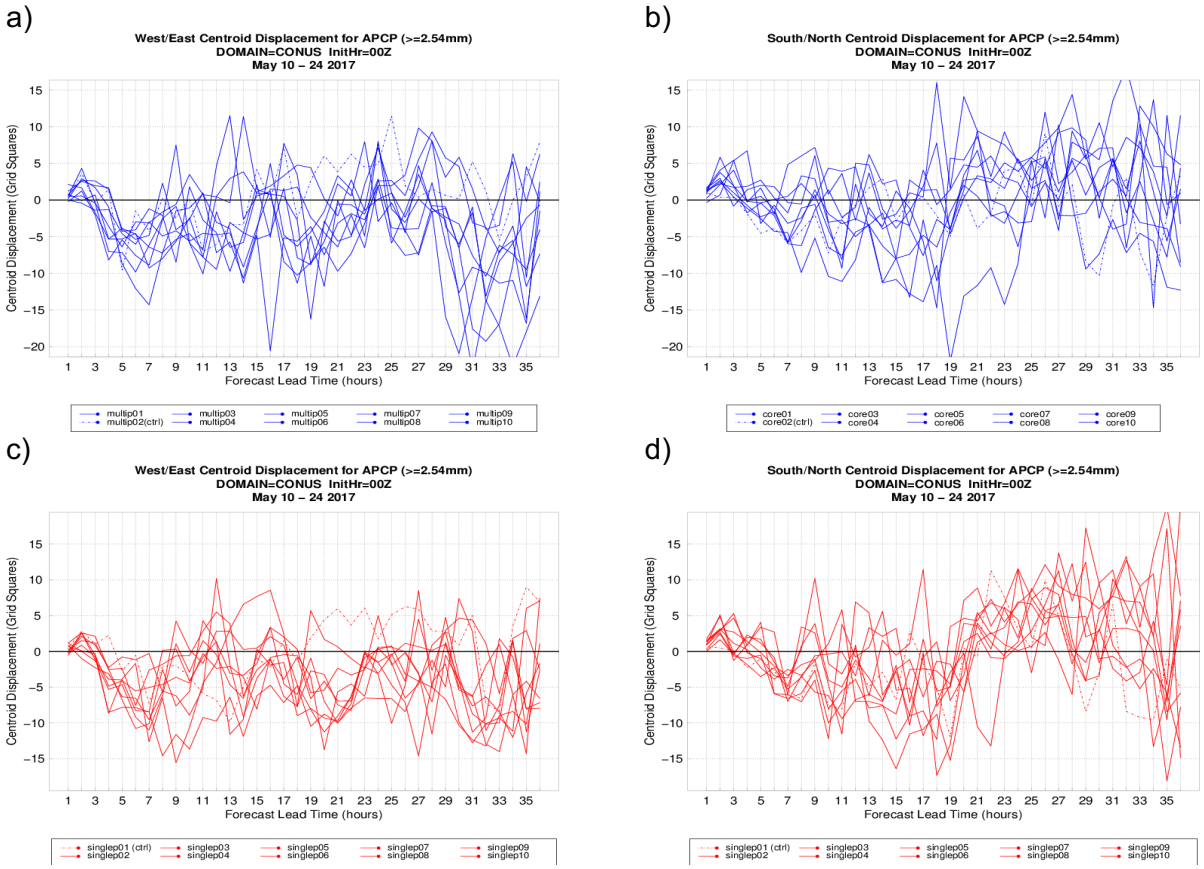


Figure 5: (a) Total object count and (b) median object area (in grid squares) of accumulated precipitation ≥ 2.54mm object counts over the full CONUS domain for all available forecasts during the experiment. The observation objects are in black, the multi-physics members in blue, the single physics members in red, and the stochastic physics members in green.

The general displacement trends for the accumulated precipitation objects can be examined using the centroid attribute derived from MODE. This is done by calculating the centroid distance between the matched forecast and the observed accumulated precipitation objects. The west-east displacement for the multi-physics, single physics, and stochastic physics

ensembles are shown in Figure 6a, Figure 6c, and Figure 6e, respectively. A negative (positive) value indicates a westerly (easterly) displacement. A majority of all examined ensemble members display westward displacement throughout the forecast period. Early in the forecast period all three ensemble subsets show a slight easterly displacement then a sharp trend toward a westerly displacement. This is potentially due to the fact that ongoing convection at the time of the 00 UTC initialization is not well assimilated in the model and lacks sufficient cold pools to translate the storms eastward. Overall, the multi-physics members exhibit the largest westward displacement, especially in the last 6 hours of the forecast (Figure 6a), while the stochastic physics members display the smallest overall westward displacement throughout the forecast period.

The north-south displacement for the multi-physics, single physics, and stochastic physics ensembles are shown in Figure 6b, Figure 6d, and Figure 6f, respectively. A negative (positive) value in this case indicates a southerly (northerly) displacement. All three ensemble subsets show an immediate northerly displacement during the first couple of hours of the forecast, then transition to a majority southerly displacement until approximately forecast hour 20. At this time, all three ensemble subsets jump to a northerly displacement for the remainder of the forecast period, likely due to day 2 convective initiation. This jump to a northerly bias is most clearly noted in the single and stochastic physics ensembles (Figure 6d and 6f). The single and stochastic physics ensemble subsets are most compare to one another.



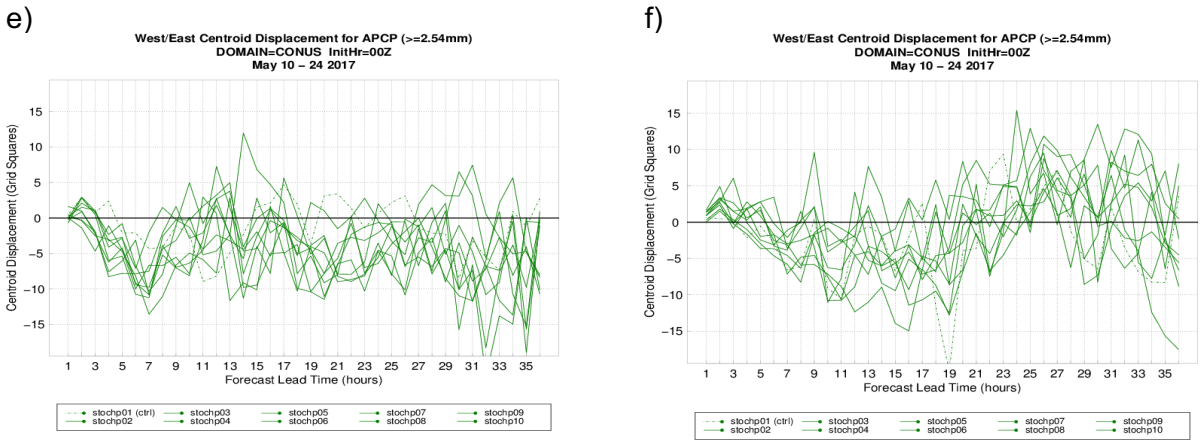


Figure 6: Centroid displacement in the west-east direction (left column) and south-north direction (right column) for the multi-physics (blue), single physics (red), and stochastic physics (green) ensemble members for accumulated precipitation objects ≥ 2.54mm aggregated over the full CONUS domain for all available forecasts during the experiment. In all plots, the control member is the dotted line.

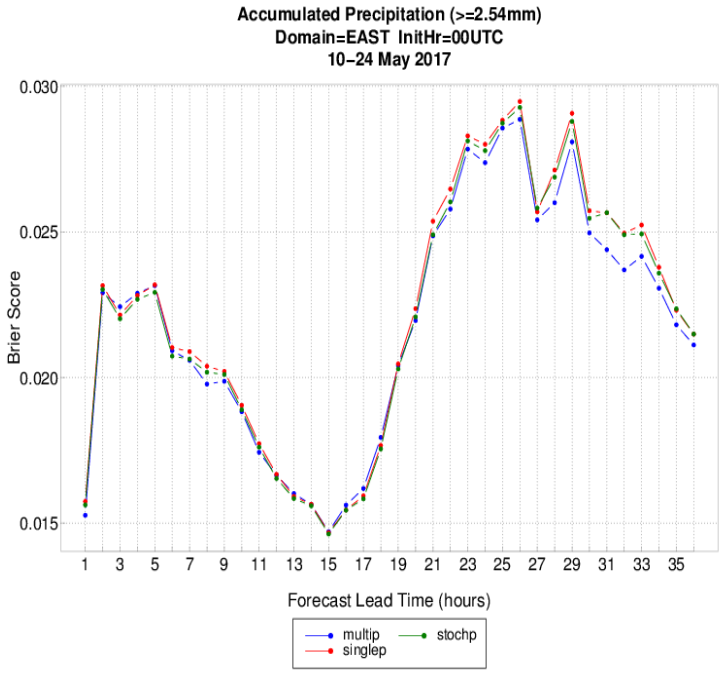
The Brier score (BS) is a tool used to analyze ensemble performance. The BS was calculated for all three ensemble subsets as a function of lead time. The BS is a measure of the mean squared probability error and can be split into three terms: reliability, resolution, and uncertainty. A perfect BS is 0. An important note is that this statistic is sensitive to the climatological frequency of the event, so the rarer the event in question, the easier it is to achieve a good BS without having any real skill. The three ensembles' BS perform similarly with regards to the temporal trend: a slight peak at forecast hour 2, then decreasing to a minimum at forecast hour 15, after which, the BS increases to reach a maximum at forecast hours 26 and 29, and then decreasing throughout the rest of the forecast (Figure 7a). The ensemble subsets demonstrate an extremely similar trend in values until the latter forecast hours where some separation is observed and the multi-physics ensemble has a slight performance edge.

A further examination of the ensemble mean accumulated precipitation is performed by looking at the individual components of the BS. The first term is reliability, which is displayed by a reliability diagram. This diagram is conditioned on the forecasts (i.e., given that an event was predicted, what was the outcome?) and gives information on forecast probability performance. As in the CLUE 2016 analysis, the reliability diagram examined here was created from 1-hour accumulated precipitation ≥ 2.54 mm aggregated over a 24-hour time period between forecast hours 12 – 36 (Figure 7b). All three ensembles are at or below the no-skill line (dotted diagonal line) for the majority of forecast frequencies. For forecast frequencies at 0.65 and 0.75, the multi-physics performs slightly better than no skill. All three ensembles display the most skill, though still over-confident, at the highest forecast frequency bin.

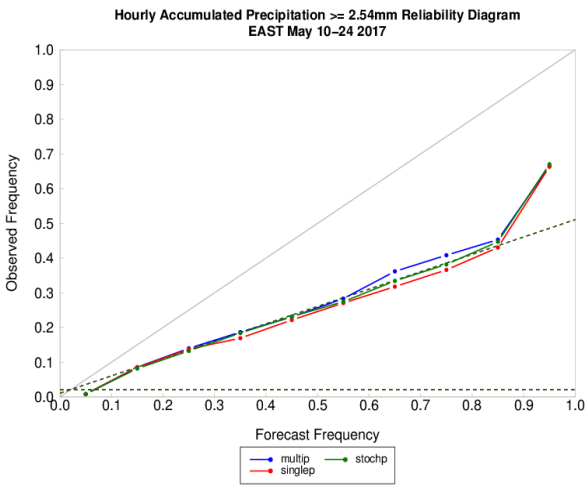
The second BS term is resolution, which can be examined via a Relative Operating Characteristic (ROC) curve. The ROC curve measures the ability of a forecast to discriminate between two alternate outcomes. Figure 7c looks at the same aggregated precipitation as examined in figure 7b. As seen in the reliability diagram, all three ensembles perform very

similarly where the multi-physics ensemble has the largest area under the ROC curve, followed by stochastic then single physics ensembles.

a)



b)



c)

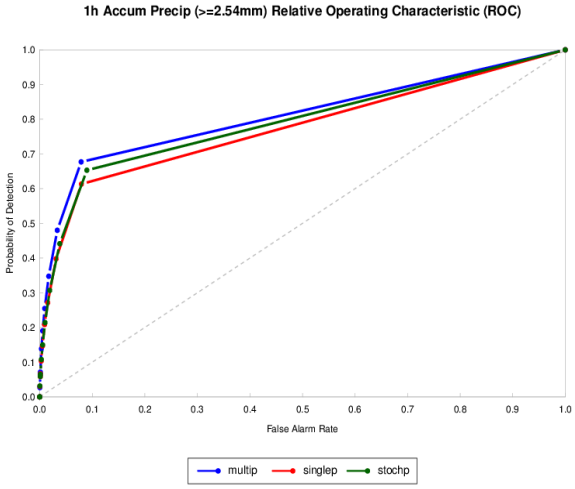


Figure 7: (a) BS time series plots, (b) reliability diagram, and (c) ROC curve of 1-h accumulated precipitation ≥ 2.54mm aggregated over a 24-hr time period between forecast hours 12 - 36 for each ensemble subset over the eastern half of the CONUS domain for all available forecasts during the experiment. The multi-physics ensemble is in blue, the single physics ensemble is in red, and the stochastic physics ensemble is in green.

The rank histogram results for each ensemble are shown in Figure 8. The rank histogram measures the frequency of observation in each bin. The forecasts are sorted from highest to lowest and then the observations are placed into one of the bins. An ideal rank histogram is flat, indicating good spread across the ensemble. A u-shaped histogram is indicative of an under-

dispersive ensemble and an inverse u-shape is indicative of an over-dispersive ensemble. A right skewed histogram indicates that the forecast has a high bias and a left skewed histogram indicates that the forecast has a low bias.

Figure 8 displays the aggregated accumulation precipitation results for the 24-hour time period from forecast hour 12 – 36. All three ensembles display a right-skewed histogram, indicating a high bias. This means that the observed values of accumulated precipitation are lower than what all of the ensemble members forecasted and thus placed in the first bin. The bias appears largest in the single physics ensemble, followed by the stochastic physics and finally the multi-physics ensemble.

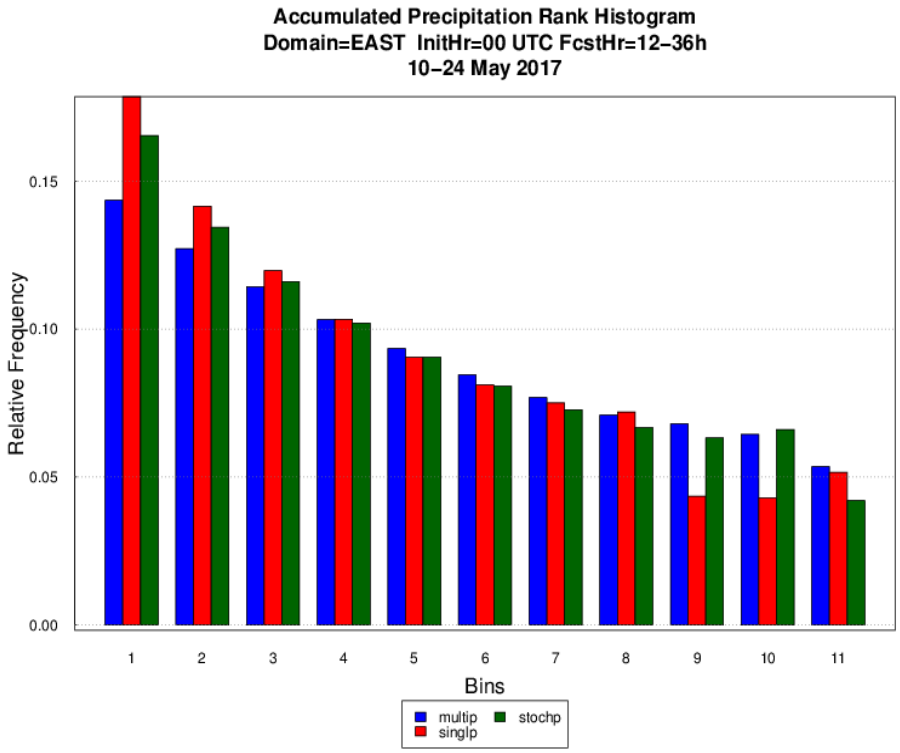


Figure 8: Rank histogram aggregated over a 24-hr time period between forecast hours 12-36 for each ensemble subset over the eastern half of the CONUS domain for all available forecasts during the experiment. The multi-physics ensemble is in blue, the single physics ensemble is in red, and the stochastic physics ensemble is in green.

A scorecard verification of the ensemble mean statistics was performed on the multi-physics and single physics ensemble subsets (Figure 9). This scorecard synthesizes the analysis in the same manner as the CLUE 2016 report (Wolff et al. 2017). The metrics compared were the Critical Success Index (CSI), GSS, bias-corrected GSS, and frequency bias. For the majority of the metrics, the multi-physics ensemble mean outperforms the single physics ensemble mean for CONUS and the West. In terms of the EAST domain, the multi-physics ensemble outperforms the single physics ensemble subset at a statistically significant level for the 24 and 30-hour forecasts.

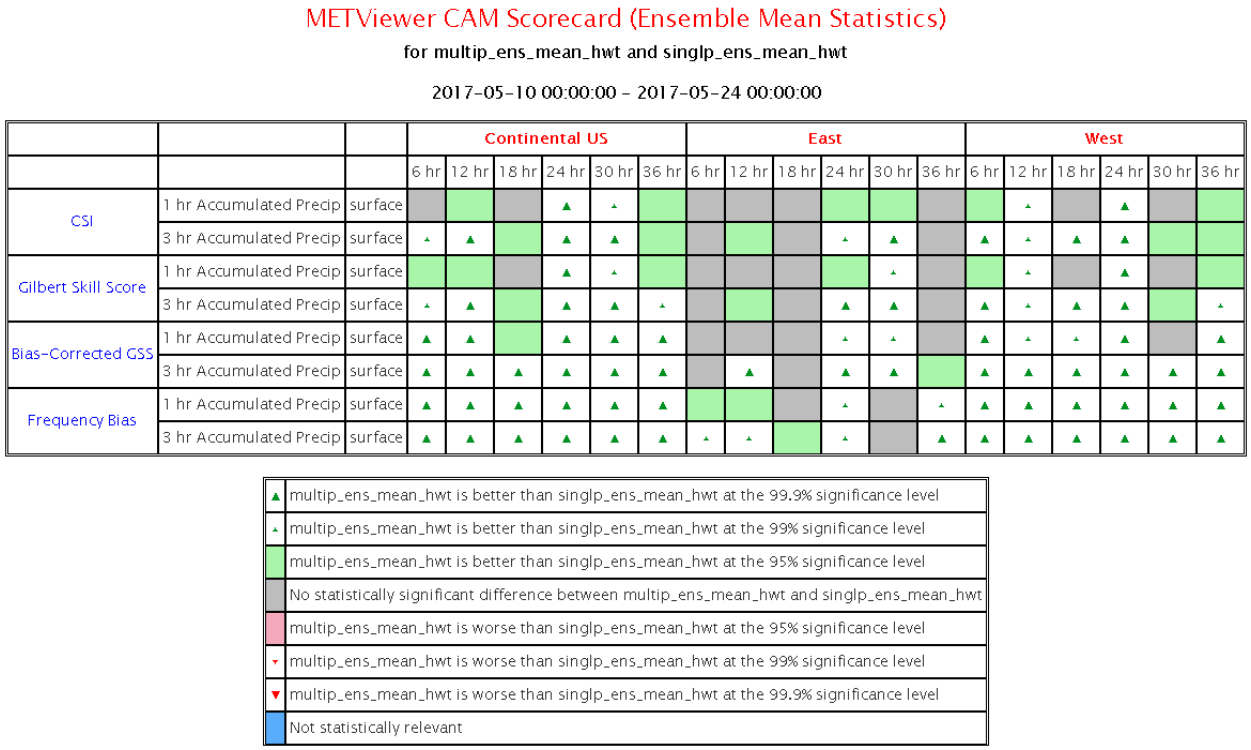


Figure 9: Precipitation accumulation scorecard comparing scores for the multi-physics and single physics ensemble mean. Scores are broken out among multiple domains and different precipitation accumulation periods, with arrows and colors indicating different levels of statistical significance.

Composite Radar Reflectivity

The same analysis performed on accumulated precipitation was performed for composite radar reflectivity $\geq$ 30 dBZ. GSS was calculated for each ensemble member as a function of forecast lead time (Figure 10a). Again, a similar temporal trend is noted across all ensemble subsets. The multi-physics subset displays the lowest GSS values for approximately the first 12 hours of the forecast. The single physics and stochastic physics subsets perform similarly throughout the forecast period.

Frequency bias was calculated for each ensemble member as a function of lead time, aggregated across all available forecasts (Figure 10b). There is a slight diurnal signal apparent in all subsets, with maxima in the evening/overnight hours and minima during the morning to afternoon hours. Overall, there is a slight over-forecast after the initial spin-up that transitions to an under-forecast for a majority of members for the remainder of the forecast period. The multi-physics ensemble subset displays the most variability among members. Two of the multi-physics members, core01 and core10 substantially under forecast composite reflectivity compared to the other members; these members use the Thompson microphysics scheme and the Noah land surface model. On the other hand, two multi-physics members substantially over forecast compared to the other members; these members are core04 and core07, which both use the Milbrandt-Yau microphysics scheme. The single and stochastic physics ensemble subsets exhibit less variability across individual members and perform similarly to one another.

As a reminder, the single and stochastic ensemble subsets use the Thompson microphysics scheme with RUC land surface model.

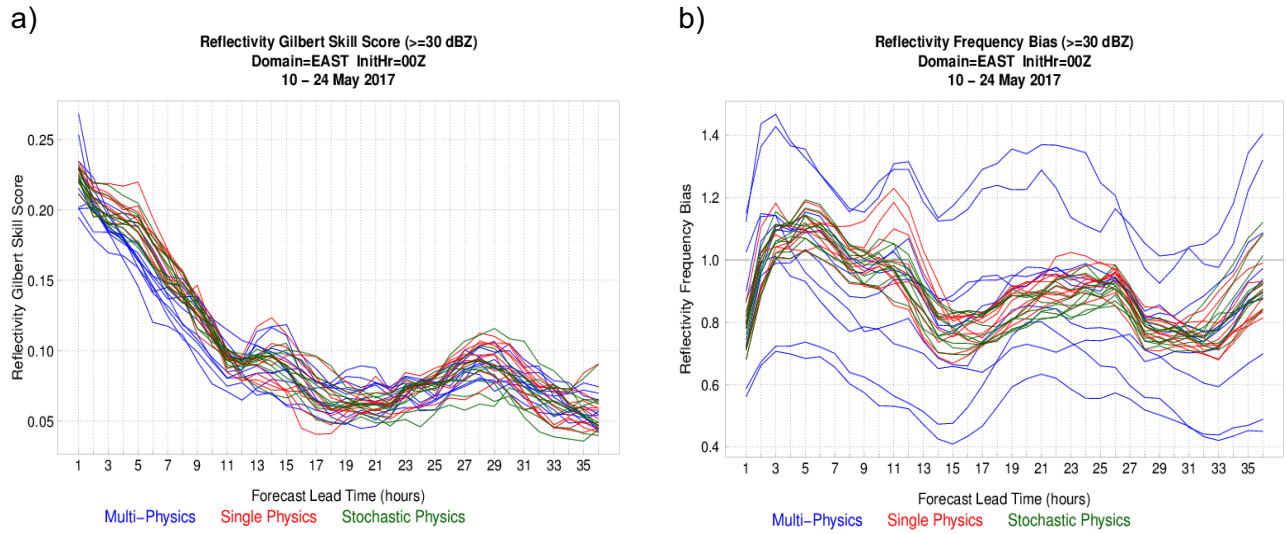


Figure 10: Same as in figure 2 except for composite reflectivity ≥ 30 dBZ.

FSS was calculated as a function of lead time for the same two neighborhood widths as for accumulated precipitation (3x3 grid squares and 7x7 grid squares; Figure 11). All three subsets display a similar temporal trend. Similar to the GSS results, the multi-physics ensemble members tend to perform the worst during the first 12 hours of the forecast for both neighborhood sizes. The single and stochastic physics ensemble members are clustered closer together. Again, skill is higher with increased neighborhood size.

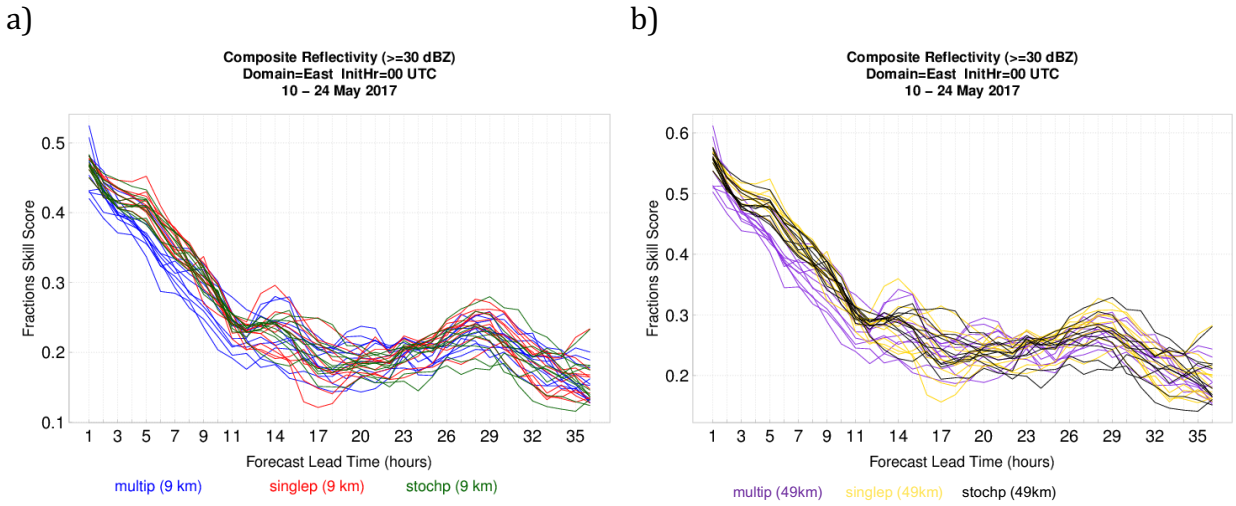


Figure 11: Same as in figure 3 except for composite reflectivity ≥ 30 dBZ.

MODE was applied to create composite radar reflectivity objects ≥ 30 dBZ across the forecast period, and the number of objects as a function of forecast lead time is displayed in Figure 12a.

A clear diurnal signal is apparent across all three ensembles, with maxima in the late afternoon/evening hours and minima during the overnight/morning hours. With exception to a few multi-physics members, all ensemble subset members have a high bias in the number of objects identified in the forecast field. As seen in previous metrics, the multi-physics ensemble has the largest member variability, though a few of the multi-physics members most often have the closest object count to the observations at any given forecast lead time. While both the single and stochastic physics members over-forecast, the stochastic physics tends to be clustered toward a higher bias, in general, compared to the single physics members.

A diurnal signal is apparent during the forecast period when assessing median object area in all three ensembles (Figure 12b). The observed object area has two elevated peaks occurring during the late night/early morning hours, where the majority of the ensemble members exhibit similar behavior during that general time period as well. No one ensemble subset outperforms the others for this metric.

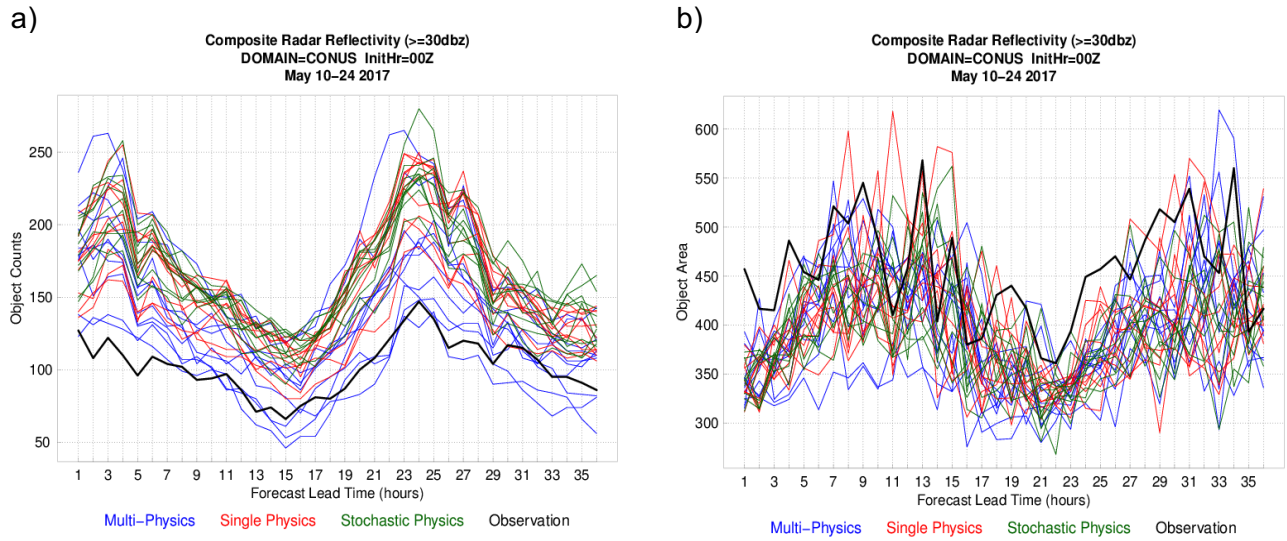


Figure 12: Same as in figure 4 except for composite reflectivity ≥30 dBZ.

The west-east centroid displacement for composite reflectivity objects ≥ 30 dBZ for multi-physics, single physics, and stochastic physics ensemble members are displayed in Figure 13a, 13c, and 13e respectively. The majority of all three ensemble subset members display a westward displacement throughout the forecast period. The multi-physics subset displays higher variability than single and stochastic physics members. The north-south displacement for the ensemble subsets are displayed in Figure 13b, 13d, and 13f. All three subsets start with a slight northerly displacement, shifting to members being fairly well distributed about 0 until approximately hour 25 where there is a shift back to northerly displacement, especially for the single and stochastic ensemble subsets, through much of the remainder of the forecast.

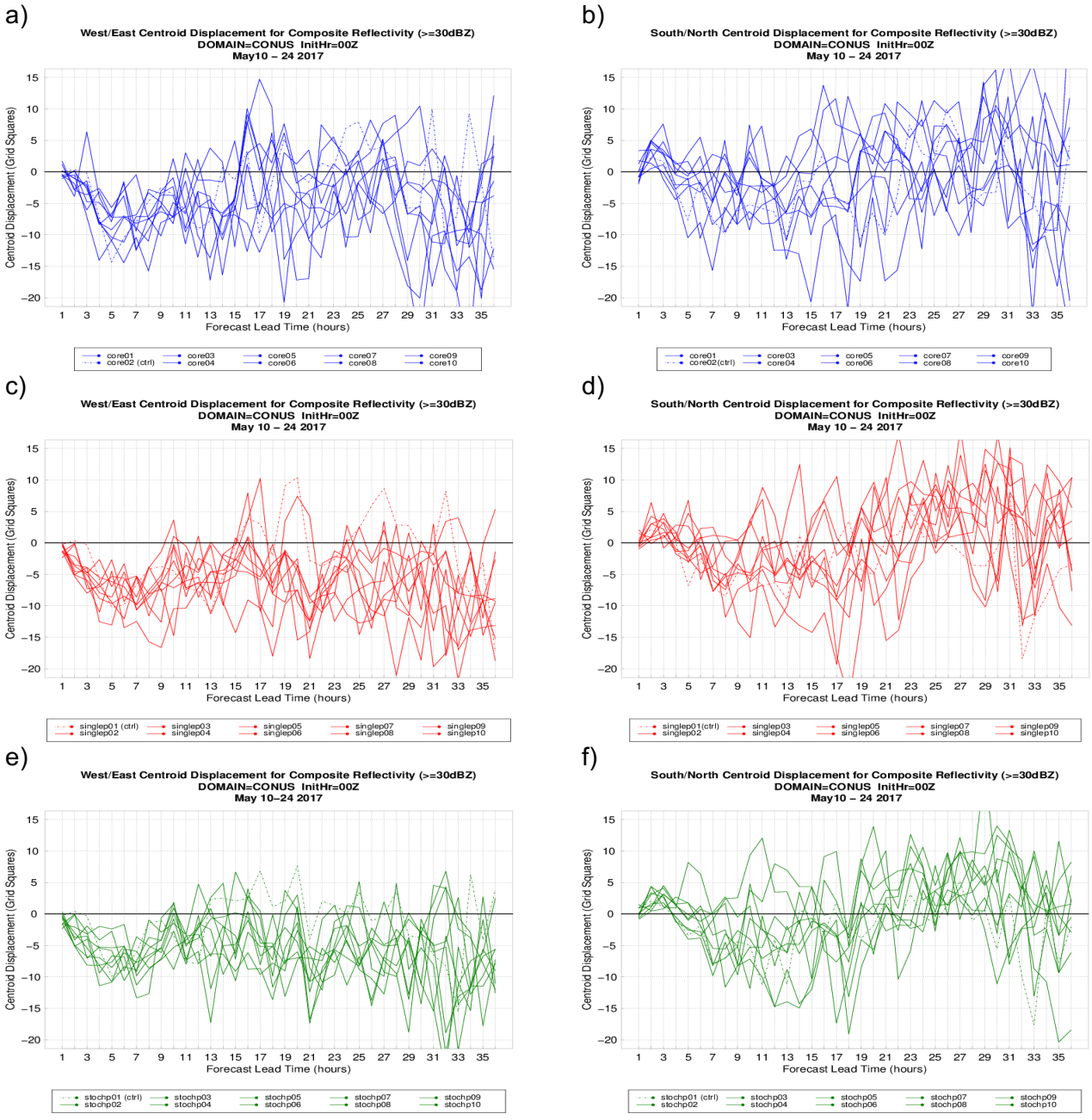
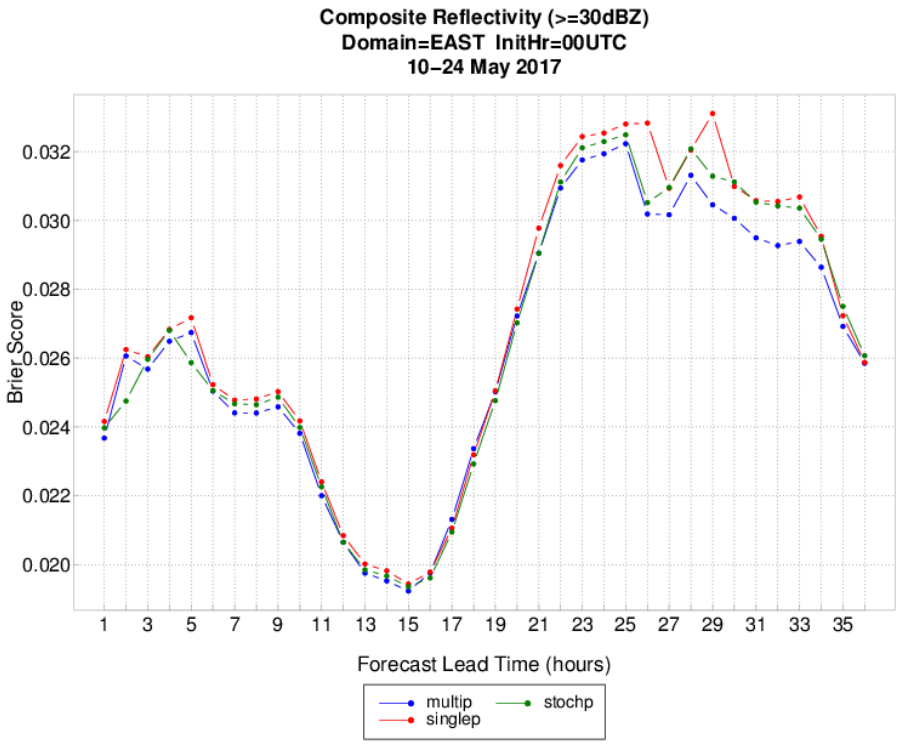


Figure 13: Same as in figure 6 except for composite radar reflective ≥ 30 dBZ.

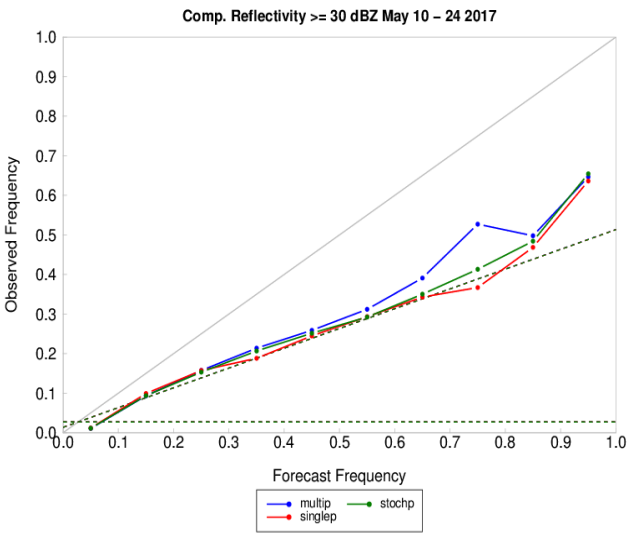
The Brier Score is examined for all three ensembles to look at ensemble performance as a function of lead time (Figure 14a). Overall, the multi-physics, single physics, and stochastic physics ensembles perform similarly to one another with the largest differences at the longer forecast hours, favoring the multi-physics ensemble.

When examining the reliability (Figure 14b), all three ensembles are at or just above the no-skill line. They all have similar performance with the largest difference apparent above the 50% forecast probability favoring the multi-physics ensemble. Similar to the accumulated precipitation assessment, the ROC curves are very similar between all three ensembles with the multi-physics ensemble having a slight edge, while still over-confident.

a)



b)



c)

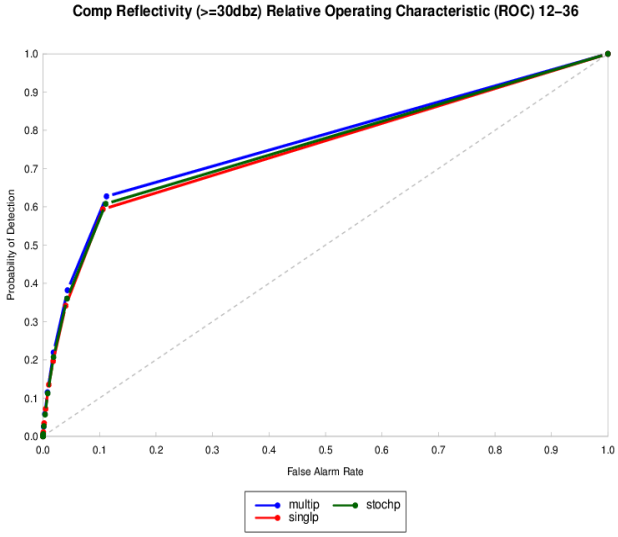


Figure 14: Same as in figure 7 except for composite reflectivity ≥30 dBZ.

Finally, the rank histogram is examined for composite radar reflectivity aggregated over the forecast hours 12 – 36 (Figure 15). Again, all ensembles display a right skewed histogram indicating a high bias. For this variable, the bias is largest for the multi-physics ensemble and smallest for the stochastic ensemble.

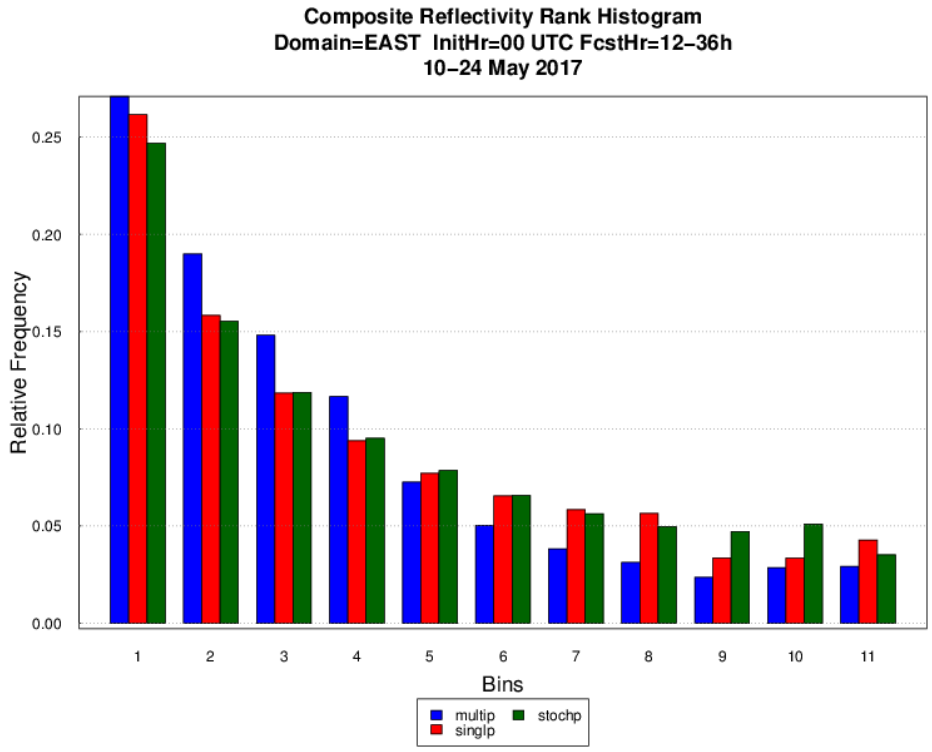


Figure 15: Same as in figure 8 except for composite reflectivity ≥ 30 dBZ.

Comparison to CLUE 2016 Dataset

This aim of this section is to compare the results presented above using the CLUE 2017 dataset with the results of the CLUE 2016 analysis. The previous evaluation can be found at: https://dtcenter.org/eval/ensembles/hwt_collab/RE5_HWT_report_FINAL.pdf. The stochastic physics ensemble was not included in CLUE 2016, so the comparisons made herein will only refer to the multi-physics and single physics ensemble performance between the two years.

Accumulated Precipitation

The GSS behavior for CLUE 2017 (Figure 2a) is very similar to the behavior for CLUE 2016 (Wolff et al., Figure 1a). The range of values is the same, and the same temporal trend occurs. The frequency bias behavior is also similar for the two datasets with the same temporal trend, and the multi-physics ensemble exhibiting more member variability than the single physics ensemble. The main difference is the range of values for frequency bias; in CLUE 2017 (Figure 2b) the range is slightly smaller than in CLUE 2016 (Wolff et al., Figure 1b). This indicates that the CLUE 2017 ensemble subsets over- and under-forecasted less than the CLUE 2016 ensemble subsets.

When comparing spatial verification metrics, the same neighborhood values were used in CLUE 2017 as in CLUE 2016. The CLUE 2017 3x3 grid square and 7x7 grid square neighborhood FSS (Figure 3a-b) values exhibit more member variability and a slight diurnal signal as compared to the CLUE 2016 counterparts (Wolff et al., Figure 2a-b).
Looking at the MODE object counts between the two years, a diurnal cycle is seen in CLUE 2017 (Figure 5a) as in CLUE 2016 (Wolff et al., Figure 4a). The 2017 ensembles display more member variability than their 2016 counterparts, and unlike in 2016, the multi-physics ensemble members have lower object counts overall, while the single physics members have overall higher object counts. Notably, there is not as much of an over-forecast in CLUE 2017 as there was in CLUE 2016. The MODE object areas for the two years (Figure 5b and Wolff et al. Figure 4b) display the same temporal trend with pronounced minima and variability among ensemble members.

The west-east centroid displacement between the two years is nearly the same in terms of the temporal behavior and the westward displacement (Figure 6a and 6c; Wolff et al. Figure 5a and 5c). Both the multi-physics and single physics ensemble for CLUE 2017 are displaced slightly more westward than their CLUE 2016 counterparts and have more ensemble member variability. The north-south centroid displacement (Figure 6b and 6d; Wolff et al. Figure 6b and 6d) displays a somewhat different trend between the two years with the shift to the north persisting throughout most of the forecast period after it occurs in CLUE 2017 while it is more transient in CLUE 2016.

The Brier Scores between the two datasets are both exhibit lower (better) values during times with climatologically lower rates of convection (early morning into early afternoon) and higher values (worse) during periods of active convection (Figure 7a; Wolff et al. Figure 6a). The most apparent difference between the two years is that while there are very small differences in behavior between the two ensembles for 2016, there is some separation in the latter forecast hours 2017 subsets.

In terms of reliability and resolution, the 2017 multi-physics and single physics ensembles display slightly higher reliability and probability of detection, respectively, than their 2016 counterparts (Figure 7b and Figure 7c; Wolff et al. 2017 Figure 6b and 6c).

The rank histograms for both years are right skewed (Figure 8; Wolff et al. Figure 7). The CLUE 2017 ensemble subsets display more evenly distributed bins than the CLUE 2016 ensembles, indicating that the CLUE 2016 dataset has a higher bias than the CLUE 2017 dataset.

Finally, the CLUE 2017 CAM scorecard (Figure 9) is drastically different than the CLUE 2016 CAM scorecard (Wolff et al. Figure 14). In CLUE 2016, the majority of the comparisons result in no statistically significant differences between the multi-physics and single physics ensemble means. In fact, the majority of the statistically significant differences favor the single physics ensemble mean. In contrast, the CLUE 2017 scorecard only favors the multi-physics ensemble

mean and the majority of the comparisons result in statistically significant differences. One contributing factor for the stark difference is the reduction in the size of the data set for 2017.

Composite Reflectivity

The GSS trends for both datasets are the same although the 2017 multi-physics ensemble members have a lightly larger maximum value (Figure 10a) than the 2016 multi-physics ensemble members (Wolff et al., Figure 8a). In both years, the multi-physics ensemble subset displays more member variability than the single physics ensemble subset. Additionally, in 2017, the multi-physics ensemble members display the lowest GSS score for the first 12 hours of the forecast while the 2016 counterpart displays larger member variability with members consistent in their position.

The frequency bias behavior is also similar with regards to member variability in both years. The multi-physics ensembles demonstrate large member variability, and the single physics ensemble members are clustered closer together. More of the 2016 members (Wolff et al., Figure 8b) tend to over-forecast throughout the forecast period while the 2017 members (Figure 10b) tend to under-forecast for the majority of forecast lead times. In addition, the outliers under-forecasting multi-physics members differ between 2016 and 2017. The 2016 members are the P3 microphysics members (multip03, 06, and 09 – all members use Noah LSM) while the 2017 members are the Thompson microphysics with the Noah LSM members (multip01 and 10).

The 2017 fractions skill scores for both neighborhood sizes (Figure 11) yield higher maximum values than their 2016 counterparts (Figure 9a-b), indicating that the 2017 ensemble subset performed better. Both subsets exhibit similar temporal trends to their FSS counterparts.

The MODE object behavior between the two subsets is very different. The CLUE 2016 single physics subset captures the observed object counts (Wolff et al., Figure 10a), while the CLUE 2017 counterpart consistently over predicts the object count (Figure 12a). In contrast, the 2017 multi-physics subset best captures the observed object counts. One similarity between the two years is the diurnal signal present in the observations of which the models are able to accurately capture the timing. In terms of MODE object area, the CLUE 2017 subsets (Figure 12b) better capture the observed object area compared to the CLUE 2016 subsets (Wolff et al., Figure 10b), especially in regards to the overnight maximum.

In terms of west/east centroid displacement, both the CLUE 2016 and 2017 ensemble subsets exhibit an overall westerly displacement, especially in the first part of the forecast period. The CLUE 2016 multi-physics subset (Wolff et al., Figure 11a) is displaced more to the west than its 2017 counterpart (Figure 13a). In terms of the single physics subset, the 2017 ensemble displays larger displacement values (Figure 13c) than its CLUE 2016 counterpart (Wolff et al., Figure 11b). Both 2017 subsets exhibit more member variability than their 2016 counterparts.

The CLUE 2017 multi-physics and single physics ensemble subsets (Figure 13b and 13d respectively) exhibit larger overall north/south centroid displacement than their 2016 counterparts (Wolff et al., Figure 11c-d). The temporal trends are the same, with both years

featuring a slight and immediate northerly displacement at the beginning of the forecast period when convection is initiating, transitioning to a southerly displacement during period where existing convection is propagating.

Finally, the ensemble statistics are compared, starting with the Brier Score. The CLUE 2016 Brier Score (Figure 13a) features a peak around forecast hour 24, and a general minimum during the overnight hours, with both the ensembles performing very similar to one another. The CLUE 2017 Brier Score (Figure 14a) features a clear minima near forecast hour 15 followed by a maximum during the evening hours (forecast hours 22-29). As in 2016, the single and multi-physics ensembles perform very similarly to one another with regards to their temporal trend and values. The main difference here is that the 2017 ensembles exhibit increased separation from one another in the later forecast hours.

Reliability diagrams for CLUE 2016 (Wolff et al., Figure 12b) and CLUE 2017 (Figure 14b) show similarities, with both ensembles near the no-skill line for a majority of forecast probabilities. Some skill is gained as the forecast frequency increases, especially for 2017, and the multi-physics ensemble subset performs slightly better than the single physics ensemble for both years.

While the 2016 (Wolff et al., Figure 12c) and 2017 ROC (Figure 14c) curves have a similar shape, there is more separation between the ensembles for 2017, and both the single and multi-physics ensembles have a higher maximum probability of detection in 2017 as compared to 2016, indicating an increase in skill.

The final comparison is between the rank histograms, where it is noted that the 2016 (Wolff et al., Figure 13) and the 2017 (Figure 15) plots, are both right-skewed, indicating a high bias. In addition, there is some indication of an increase in under-dispersion in 2016 that is not as prevalent in 2017.

Summary

Thorough deterministic and probabilistic evaluations of three ensemble subsets (multi-, single, and stochastic physics) from the CLUE 2017 super ensemble, and a comparison to the corresponding subsets from the CLUE 2016 was presented. Many verification techniques are available to evaluate ensemble performance and all of the ones utilized in this analysis were accomplished using the MET verification software system.

Overall, no one ensemble subset was the clear winner in terms of hourly accumulated precipitation ($\geq$ 2.54 mm) prediction when consulting the deterministic or spatial metrics. In general, the multi-physics ensemble subset often performed better when assessed with the ensemble metrics and this was confirmed when looking at the ensemble metric scorecard.

Similarly, there was no ensemble subset that was definitively the best performer in terms of composite radar reflectivity ($\geq$ 30 dBZ) performance. The single and stochastic physics ensemble subsets outperform the multi-physics ensemble subset in terms of deterministic

metrics. No ensemble subset was a clear winner in terms of spatial metrics. As in accumulated precipitation, the multi-physics ensemble tended to perform the best in terms of ensemble metrics.

Given these results, it is promising that the single and stochastic ensemble subsets perform comparably throughout this analysis to the multi-physics ensemble approach. This suggests that with additional testing and tuning it is plausible for an easier to maintain and less resource intensive approach to be considered for operational implementation in the future.

References

Wolff, J., M. Harrold, J. Beck, I. Jankov, T. Hertneky, and L. Blank, 2018: Assessing ensemble forecast performance for select members available in CLUE during 2016 HWT-SFE. Available at: https://dtcenter.org/eval/ensembles/hwt_collab/RE5_HWT_report_FINAL.pdf