Assessment of ensemble forecast performance for select members available in the CLUE during HWT-SFE 2018

Report compiled by:

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

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 previous final report located at https://dtcenter.org/eval/ensembles/hwt_collab/RE5_HWT_report_FINAL.pdf. For more information on HWT-SFE 2018, the program overview and operations plan can be found at https://hwt.nssl.noaa.gov/Spring_2018/HWT_SFE2018_operations_plan_FINAL.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). Three particular subsets were of interest for this analysis, including the mixed-physics 10-member ensemble, the single physics 8-member ensemble, and the stochastic physics 8-member ensemble, all using the WRF-ARW dynamic core. The physics suites for each 2018 ensemble subset used in this analysis are presented in Table 1. To utilize a consistent number of ensemble members between the three subsets described here, only members 1-8 were used for the mixed-physics ensemble.

*Table 1. Physics suite description for CLUE 2018 subsets examined. A * indicates that component has been stochastically perturbed.*

Multi-physics (10 members)					
Member	**IC**	**BC**	**Microphysics**	**LSM**	**PBL**
mixed-phys01	**NAMa+3DVAR**	**NAMf**	**Thompson**	**Noah**	**MYJ**
mixed-phys02 (control)	**RAPa+3DVAR**	**GFSf**	**Thompson**	**RUC**	**MYNN**
mixed-phys03	**mixed-phys01+ arw-p1_pert**	**arw-p1**	**NSSL**	**Noah**	**YSU**
mixed-phys04	**mixed-phys01+ arw-n1_pert**	**arw-n1**	**NSSL**	**Noah**	**MYNN**
mixed-phys05	**mixed-phys01+ nmmb-p1_pert**	**nmmb-p1**	**Morrison**	**Noah**	**MYJ**
mixed-phys06	**mixed-phys01+ nmmb-n1_pert**	**nmmb-n1**	**P3**	**Noah**	**YSU**
mixed-phys07	**mixed-phys01+ arw-p2_pert**	**arw-p2**	**NSSL**	**Noah**	**MYNN**
mixed-phys08	**mixed-phys01+ arw-n2_pert**	**arw-n2**	**Morrison**	**Noah**	**YSU**
mixed-phys09	**mixed-phys01+ nmmb-p2_pert**	**nmmb-p2**	**P3**	**Noah**	**MYNN**
mixed-phys10	**mixed-phys01+ nmmb-n2_pert**	**nmmb-n2**	**Thompson**	**Noah**	**MYNN**

Single Physics + IC/BC pert (8 members)					
Member	**IC**	**BC**	**Microphysics**	**LSM**	**PBL**
single-phys02 (control)	**RAPa+3 DVAR**	**GFSf**	**Thompson**	**RUC**	**MYNN**

Stochastic Physics + IC/BC pert (8 members)					
Member	**IC**	**BC**	**Microphysics***	**LSM**	**PBL***
stoch-phys02 (control)	**RAPa+ 3DVAR**	**GFSf**	**Thompson**	**RUC**	**MYNN**

The 2018 HWT-SFE was held from 30 April – 1 Jun, 2018. 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. For this report, the dates used in the evaluation included 2 May - 1 June, 2018. A full inventory of the available data revealed that that the most common date to be missing was 14 May 2018. It is also noted that the single physis member 2 was missing throughout the experiment. Overall, data availability was reasonably covered during the period of interest (Table 2) .

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

Multi-physics										
Date	*mp01*	*mp02*	*mp03*	*mp04*	*mp05*	*mp06*	*mp07*	*mp08*	*mp09*	*mp10*
4/30-5/02						*all*				
5/03									*fhr18*	
5/04			*fhr00*							
5/14	*fhr24-36*	*fhr6, 18-36*	*fhr24-36*	*fhr22-36*	*fhr28-36*	*fhr14, 18, 22-36*	*fhr20, 23, 26-36*	*fhr28, 34-36*	*fhr27-36*	*fhr24, 27-36*

Single physics								
Date	*mp01*	*mp02*	*mp03*	*mp04*	*mp05*	*mp06*	*mp07*	*mp08*
4/30	*fhr00*	*Missing for all*						

5/14	*fhr12, 18, 22-36*	*dates of the exper.*	*fhr17, 23-36*	*fhr23-36*	*fhr23-36*	*fhr23-36*	*fhr23-36*	*fhr21, 23-36*
5/15			*fhr 10*					

Stochastic physics								
Date	*mp01*	*mp02*	*mp03*	*mp04*	*mp05*	*mp06*	*mp07*	*mp08*
5/14	*fhr23-36*	*fhr29-36*	*fhr18, 23-36*	*fhr23-36*	*fhr23-36*	*fhr10, 23-36*	*fhr13, 16, 23-36*	*fhr22-36*
6/01	*fhr00*	*fhr00*	*fhr00*	*fhr00*	*fhr00*	*fhr00*	*fhr00*	*fhr00*

Verification results

Surface variables

Temperature

When examining 2-m temperature bias (0 indicates unbiased forecast) of individual members of each ensemble subset, the least variability among subsets is seen in the stochastic and single physics ensemble subsets, while the most variability is seen in the multi-physics subset (Fig. 1a). The stochastic physics and single physics ensembles have a prominent diurnal signal in bias, with most members having a small cold-to-neutral bias for approximately the first 12 hours of the forecast period before transitioning to a strengthened cold bias in the daytime hours. At this time, the cold bias is stronger for most stochastic members than in the single physics. The multi-physics ensemble displays a weak diurnal signal in error curves. At the beginning of the forecast period, members generally have a weak cold bias before transitioning to a neutral-to-warm bias during the daytime hours. Day 2 of the forecast period has members ranging between a weak warm to weak cold bias, with an overall ensemble mean close to 0 (i.e., unbiased). For bias corrected root mean square error (BCRMSE) (lower is better), all ensemble subsets display similar distributions, with variability among members being very tight during the Day 1 forecast period; more variability is seen in the Day 2 forecast period, with the stochastic physics members having marginally higher overall BCRMSE values compared to the other physics ensemble subsets (Fig. 1b). In general, the BCRMSE of the ensemble means are lower than the individual members throughout the forecast period, with the multi-physics generally having the lowest BCRMSE.

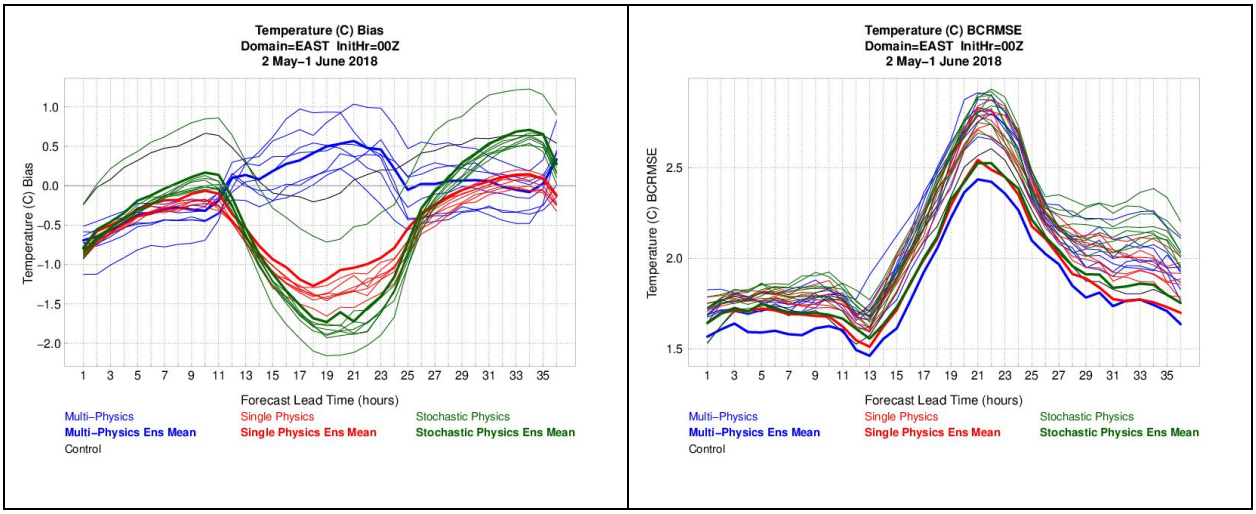


Figure 1. *(a) Bias and (b) BCRMSE time series plots of 2-m temperature (°C) for each individual ensemble member aggregated across the eastern half of the CONUS domain for all available forecasts during the experiment. The control member is in black, the multi-physics members are in blue, the single physics members are in red, and the stochastic physics members are in green. The thick blue, red, and green lines represent the ensemble mean for the multi-physics, single physics, and stochastic physics ensembles, respectively.*

Overall skill, as assessed by root mean square error (RMSE) (lower is better), is the best for the multi-physics ensemble, followed by the single physics, with the stochastic physics having the largest RMSE values, particularly during the afternoon hours (Fig. 2). In terms of spread, the multi-physics also has the largest values; however, the next largest spread is seen for the stochastic physics with the single physics consistently having the lowest spread values. It is interesting to note late in the forecast period that the stochastic ensemble spread does increase, though this also corresponds to an increase in RMSE at the same time. Ideally, the spread/skill ratio would be equal to one. The end results in the spread/skill ratio is that while the values are closest to one for the multi-physics ensemble a majority of the time, all of the ensembles are less than one indicating a of lack sufficient spread to account for the amount of error.

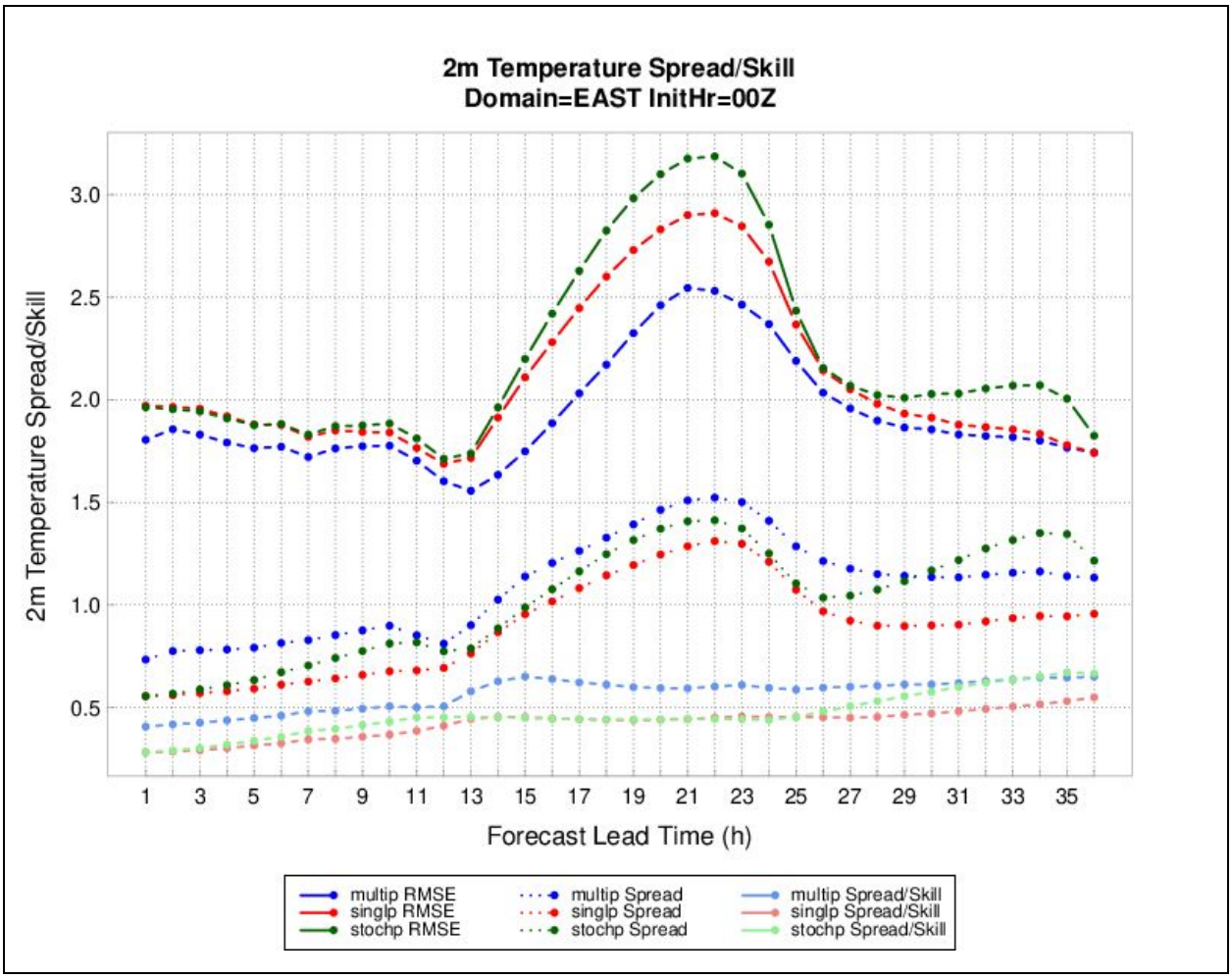


Figure 2. *Spread/skill by forecast lead time for 2-m temperature (°C) for 0000 UTC initializations aggregated across the eastern half of the CONUS domain for all available forecasts during the experiment. Skill (RMSE) is designated by the solid line, spread by the dotted line, and spread/skill ratio by the light dashed line. The multi-physics ensemble is in blue, the single physics ensemble in red, and the stochastic physics ensemble in green.*

Reliability is a measure of conditional frequency bias. Within each probability category for the forecast, we examine the frequency of occurrence of the observed events. When assessing ensembles using reliability diagrams, the forecast probabilities are binned and assessed against the observed frequency. Thus, perfect reliability would be when the forecast and observed frequencies in each category are equal and lie along the 1-to-1 line. In addition, the “no resolution” line (or sample base rate) is plotted as the horizontal dashed line and corresponds to a uniform forecast of the climatological frequency of the event. The “no skill” line is indicated by the diagonal dashed line that lies halfway between the climatology and perfect reliability lines. 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.

For 2-m temperature using a threshold of ≥298K results in a sample climatology of about 30% for each of the ensembles (Fig. 3). The most reliable ensemble for this test period is the

multi-physics ensemble. In general, the curve has a positive slope, indicating that as the forecast probability increases, so too does the observed frequency. For the lower forecast probabilities the multi-physics ensemble tends to under-forecast the event probability transitioning to over-forecasting the event probability at the higher forecast probabilities. This is a common trend for under-dispersive ensembles. A generally different behavior is noted for the single and stochastic ensembles where it is seen that both tend to under-forecast the events at a the lower forecast probabilities but are very close to the one-to-one line for the higher forecast probabilities, with a slight edge noted with the stochastic ensemble.

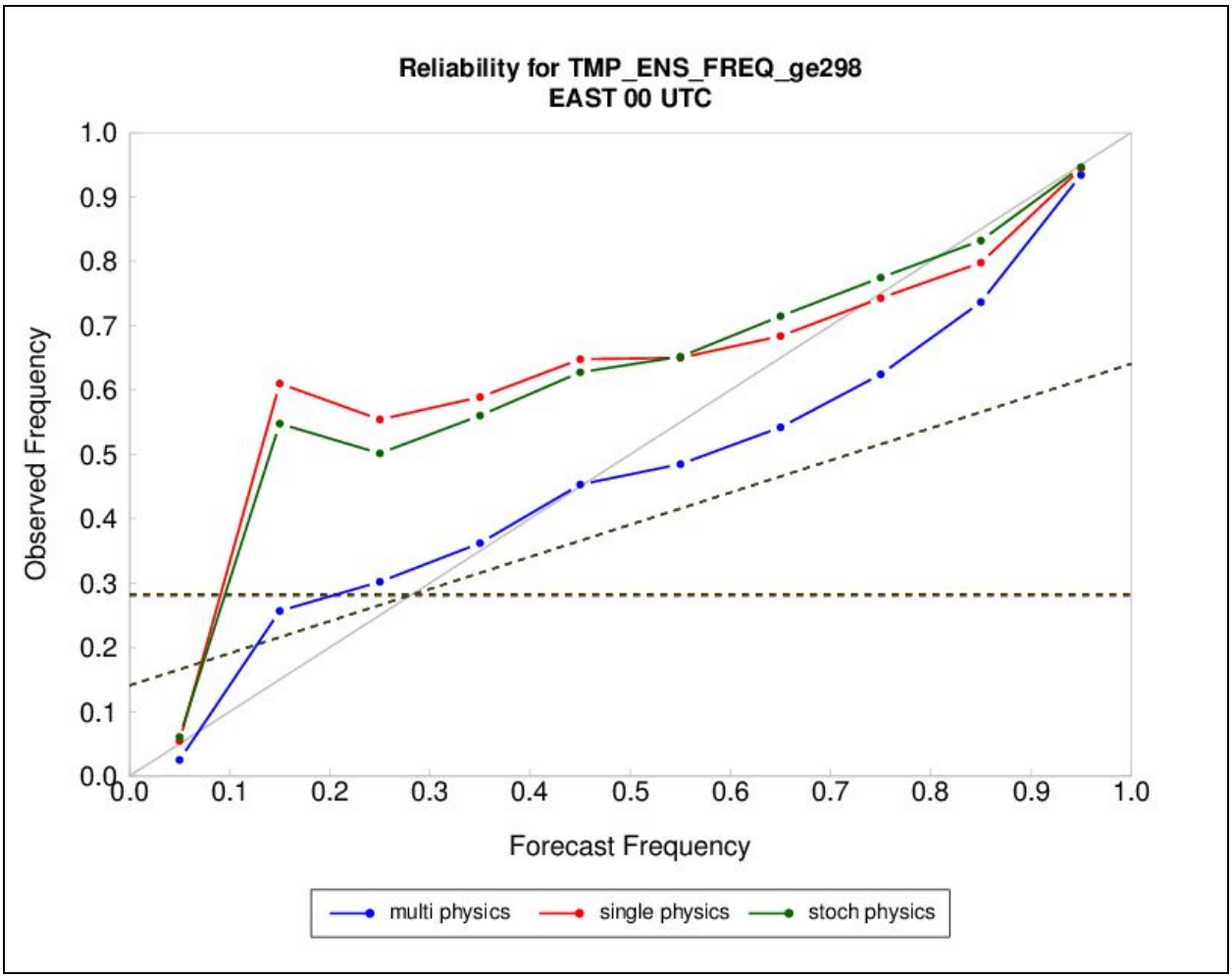
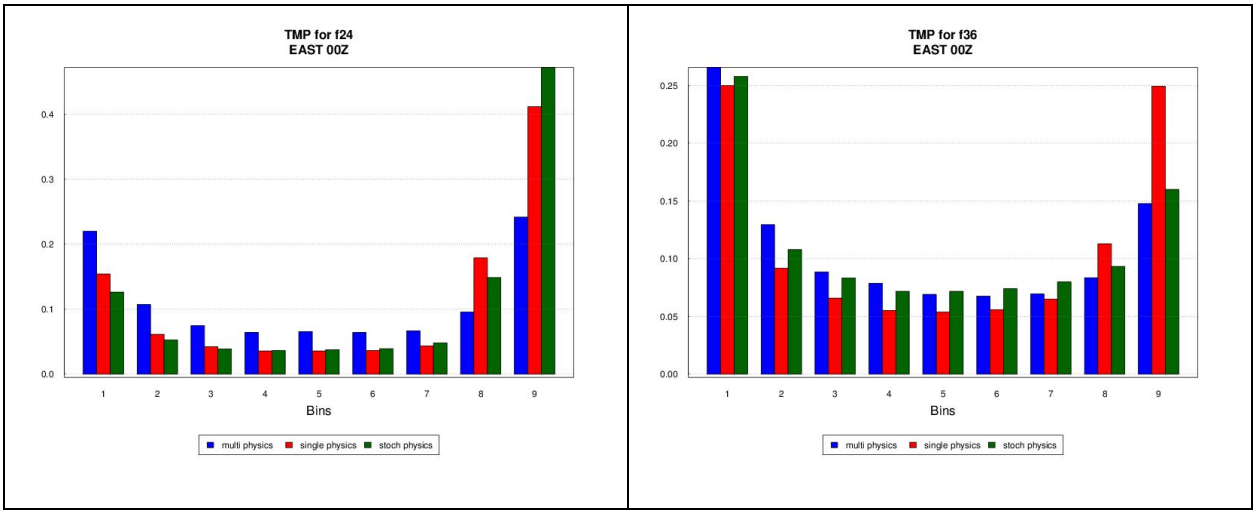


Figure 3. *Reliability diagrams for 0000 UTC initializations aggregated across the eastern half of the CONUS domain for all available forecasts during the experiment for 2-m temperature at a threshold of ≥298 K. The horizontal dotted line represents no resolution, the diagonal dotted line represents no skill, and the solid grey diagonal line represents perfect reliability. The multi-physics ensemble is in blue, the single physics ensemble in red, and the stochastic physics ensemble in green.*

The under-dispersiveness seen from the spread/skill plots is also seen in the rank histograms, denoted by the U-shaped plot for 2-m temperature (Fig. 4). For the single and stochastic ensembles at forecast hour 24 (valid at 00 UTC), there is also an indication of a low bias as

more observations fall in the last bin, rather than the first (Fig. 4a). Interestingly, the opposite is generally noted for the multi-physics and stochastic ensembles at forecast hour 36 (valid at 12 UTC) where we see an indication of a high bias (Fig. 4b). Overall, when forecast hours 12-36 are aggregated together, all three ensembles are under-dispersive, with the multi-physics the least so and a low bias noted in the single and stochastic ensembles (Fig. 4c).



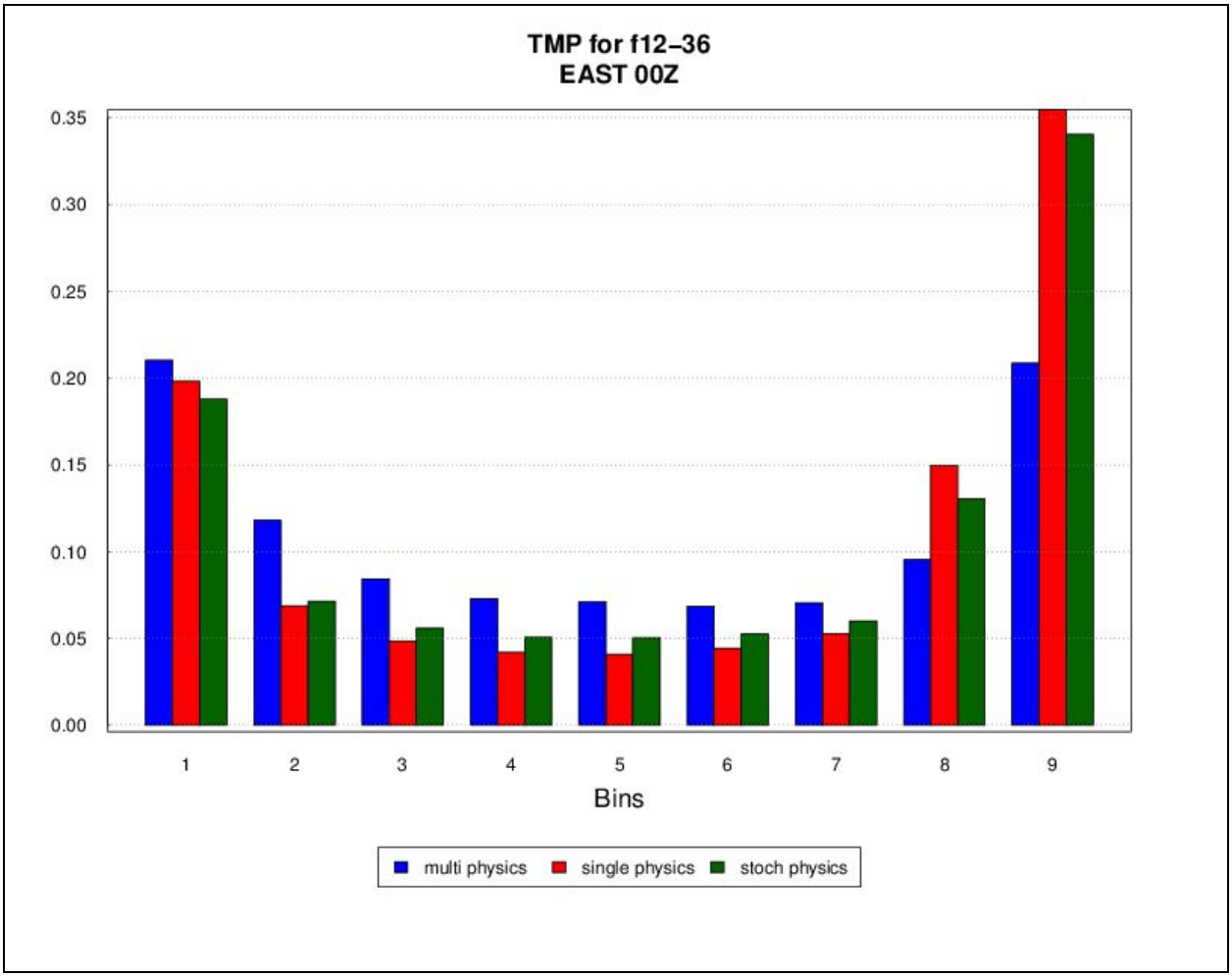


Figure 4. *Rank histogram plots for 2-m temperature for 0000 UTC initializations aggregated across the eastern half of the CONUS domain for all available forecasts during the experiment for (a) the 24 hour forecast lead time, (b) 36 hour forecast lead time, and (c) for the 12-36 hour forecast lead times aggregated together. The multi-physics ensemble is in blue, the single physics ensemble in red, and the stochastic physics ensemble in green.*

Moisture

Similar to 2-m temperature bias, variability of bias values among members is lowest in the stochastic and single physics ensemble subsets, while more variability is noted in the multi-physics subset (Fig. 5a). In addition, a wet bias is noted for all members in all subsets at a majority of forecast lead times; while the shape of the bias curves are generally the same for all ensemble subsets the multi-physics members generally have bias values on the order of 0.5-2 °C lower (i.e., smaller wet bias) than the stochastic and single physics subsets. A diurnal signal in bias is noted for all 3 subsets, with maximum wet biases in the afternoon to evening (~16 - 00 UTC) and minimum bias values in the morning (~12 UTC). When considering BCRMSE, all ensemble subsets have a diurnal signal in error values, with peak values in the afternoon and evening (Fig. 5b). Generally, the multi-physics has BCRMSE values similar or lower than the stochastic and single physics ensemble subsets. Similar to 2-m temperature, the ensemble

mean BCRMSE values are lower than the individual members and variability among members increases at the end of the forecast period.

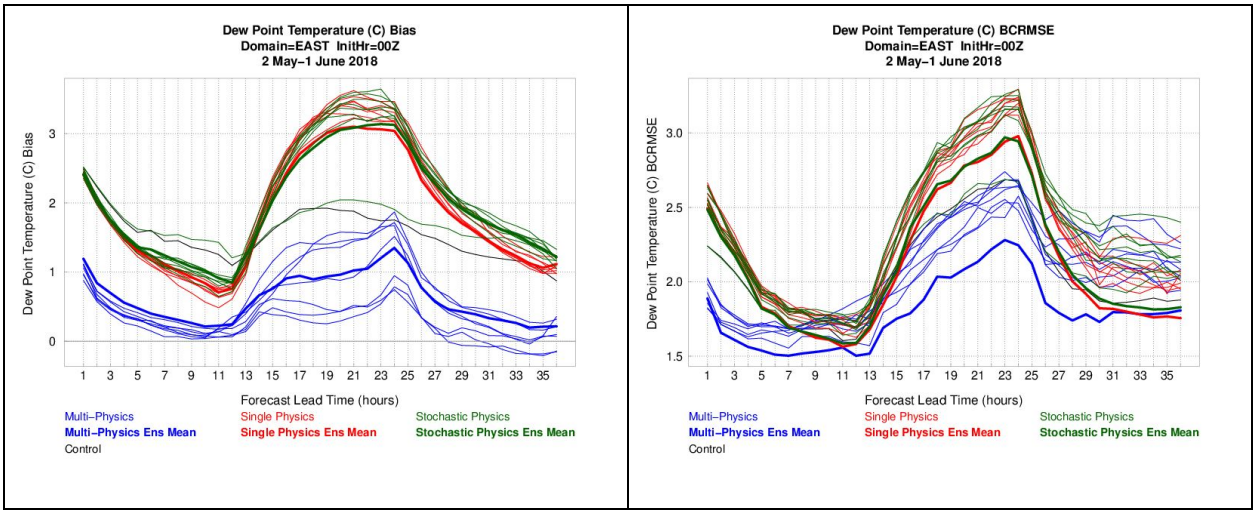


Figure 5. *Same as Fig. 1, except for 2-m dew point temperature.*

As noted for 2-m temperature, the RMSE values are lowest and the spread values are the highest for the multi-physics ensemble at all forecast lead times for 2-m dew point temperature (Fig. 6). This leads to the highest spread/skill ratio. Very little differences are noted in the values for the single and stochastic ensembles for this variable. All three ensembles show a general trend to higher spread as the forecast lead time increases.

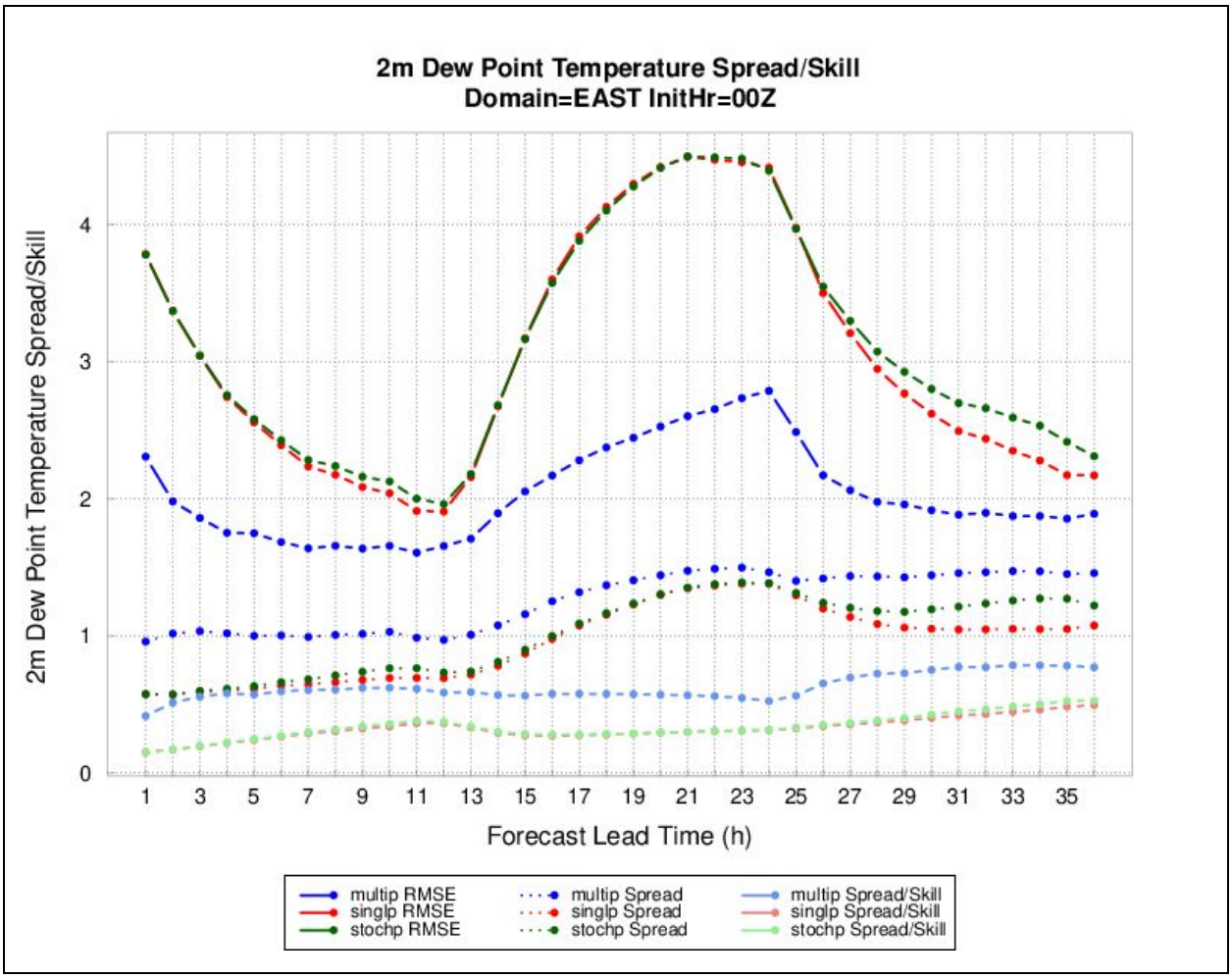


Figure 6. *Same as Fig. 2, except for 2-m dew point temperature.*

When looking at reliability, a threshold of ≥293K for 2-m dew point temperature results in a sample climatology of just over 20% (Fig. 7). In this case, all three ensembles over-forecast the observed frequencies for all forecast probabilities. While the multi-physics ensemble is closer to the one-to-one line, the only noted skill for any of the ensembles is at the very lowest and highest forecast probabilities.

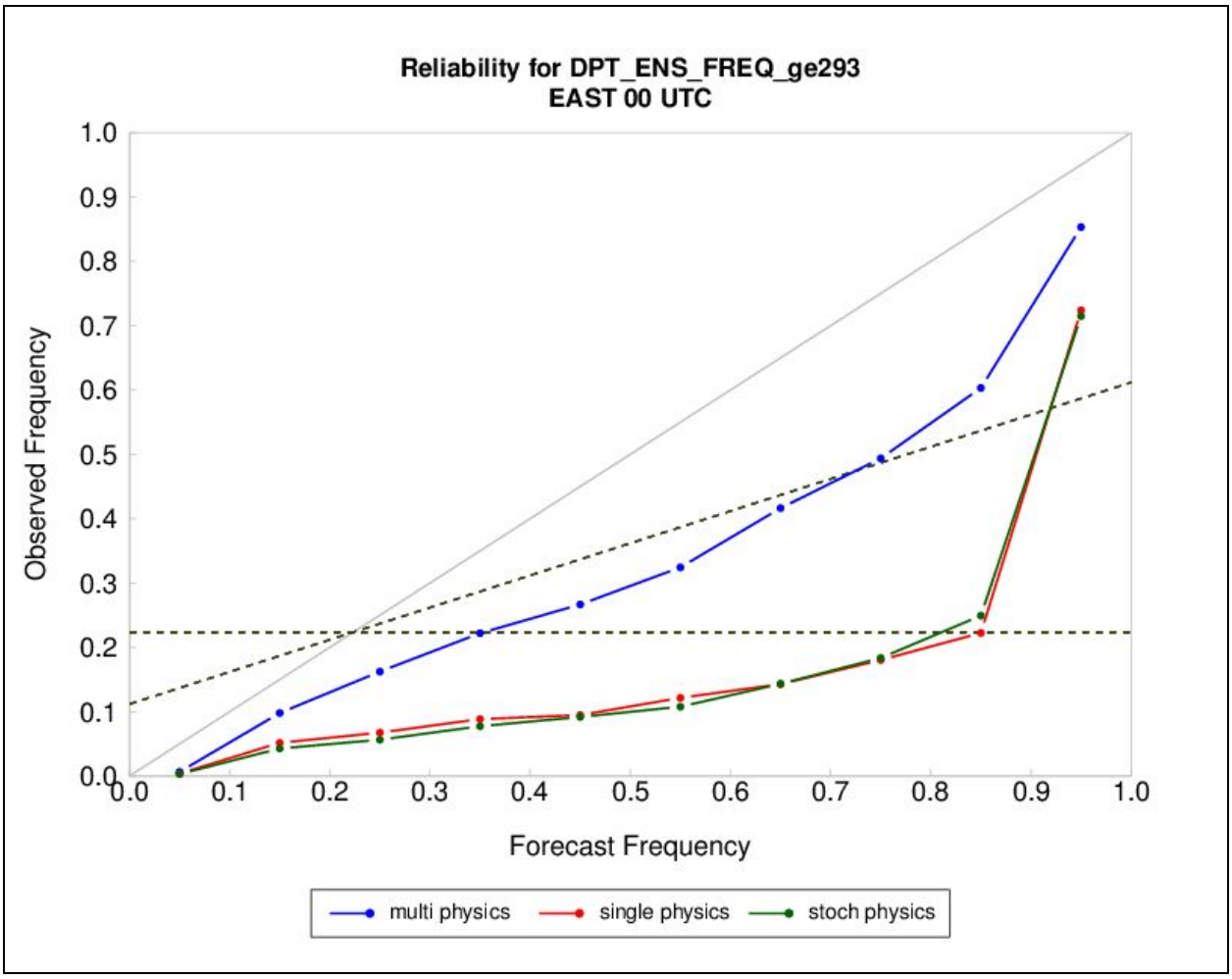


Figure 7. *Same as Fig. 3, except for 2-m dew point temperature.*

When plotting the rank histogram for 2-m dew point temperature at all forecast hours from 12 - 36, while there is not an obvious under-dispersiveness, all three ensembles are skewed left and exhibit a high bias (Fig. 8). This bias is notably worse for the single and stochastic physics ensembles.

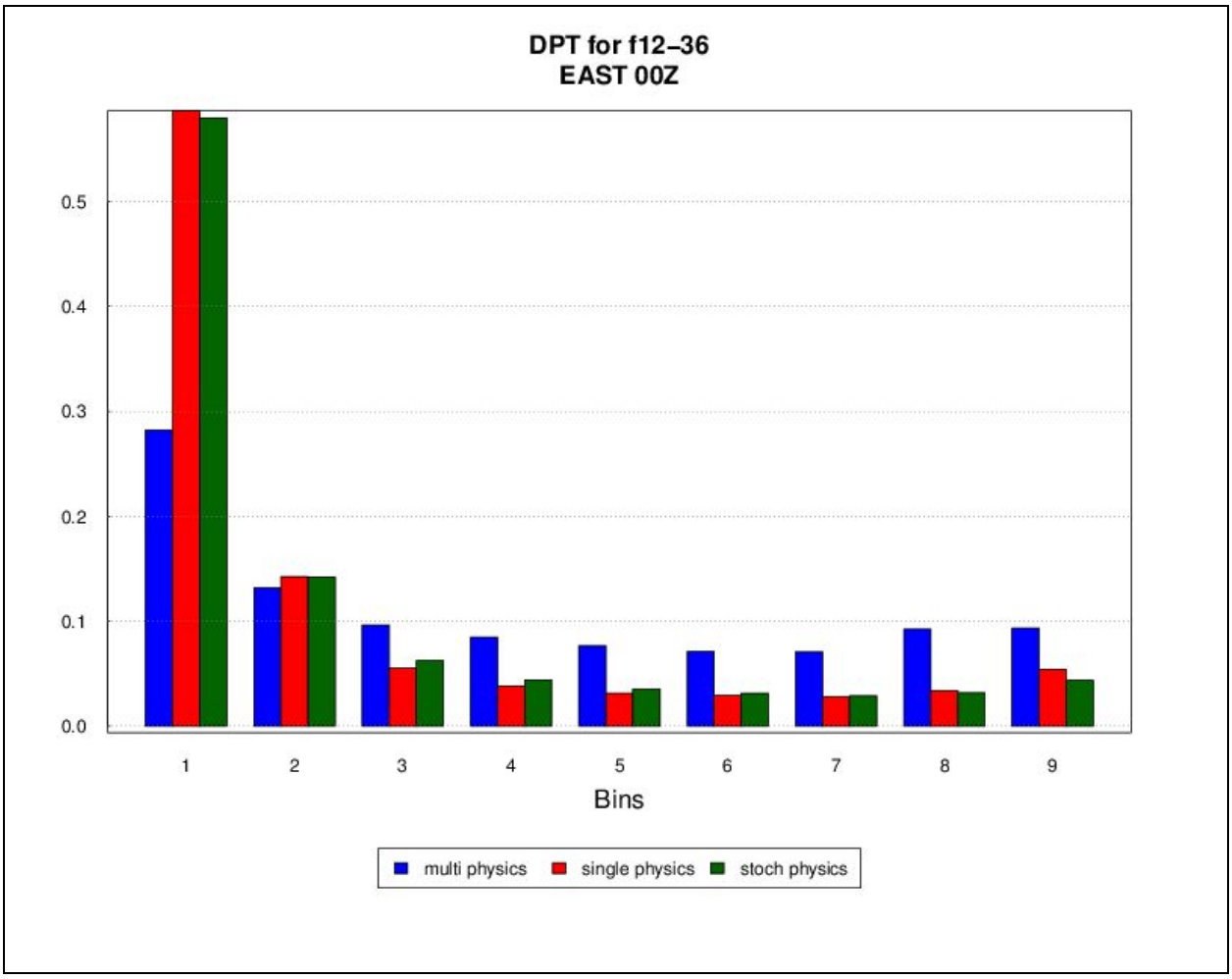


Figure 8. *Same as Fig. 4c, except for 2-m dew point temperature.*

Wind

For all members of all physics subsets and for all forecast lead times, a high wind speed bias is noted, with the multi-physics ensemble subset having higher overall bias than the stochastic physics and single physics ensembles (Fig. 9). Minimal variability is seen among both the stochastic physics and single physics ensembles, while the multi-physics ensemble displays a higher degree of variability. Similar to the 2-m temperature and dew point temperature, a diurnal signal is noted for all 3 physics subsets; the smallest high biases are seen during the daytime (~13 - 22 UTC), while higher positive wind speed biases occur in the evening and overnight hours (01 - 11 UTC). For wind speed BCRMSE, while all ensemble subsets have a similar error distribution pattern, the multi-physics ensemble has higher overall BCRMSE values than the stochastic physics and single physics subsets. Similar to bias, the variability among members is larger in the multi-physics subset. A diurnal signal is also noted, with higher BCRMSE values in the evening (20 - 02 UTC) and lowest values overnight and morning (07 - 13 UTC). Overall, the BCRMSE for the ensemble mean is lower than the individual members for each subset.

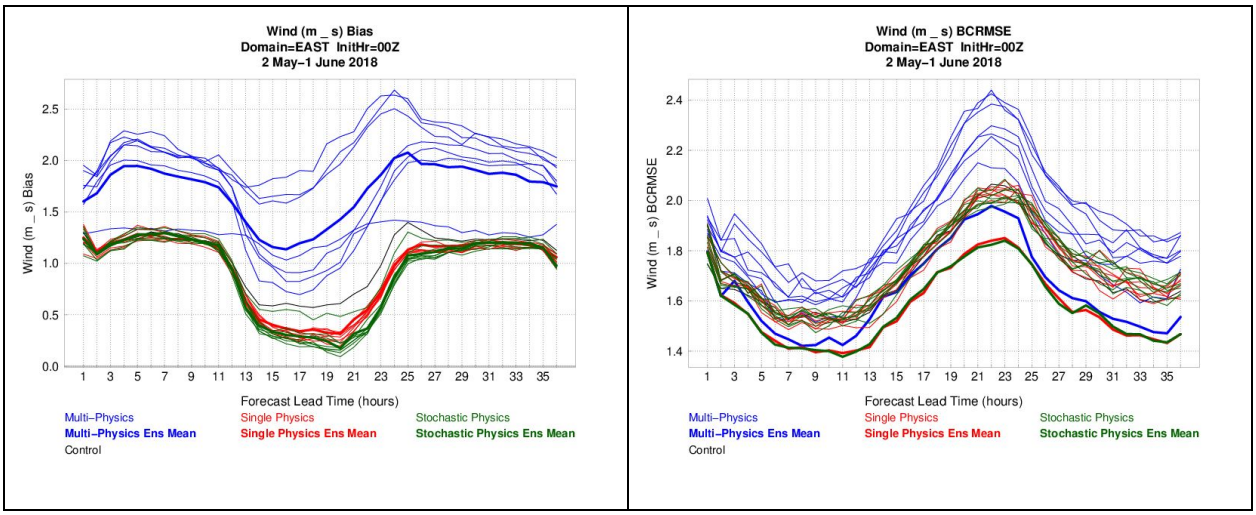


Figure 9. *Same as Fig. 1, except for 10-m wind speed (m s^{-1}).*

While the RMSE values are the lowest (and nearly identical) for the single and stochastic ensembles compared to the multi-physics ensemble, they also have the lowest spread by a similar margin resulting in very similar spread/skill ratios for all three ensembles (Fig. 10).

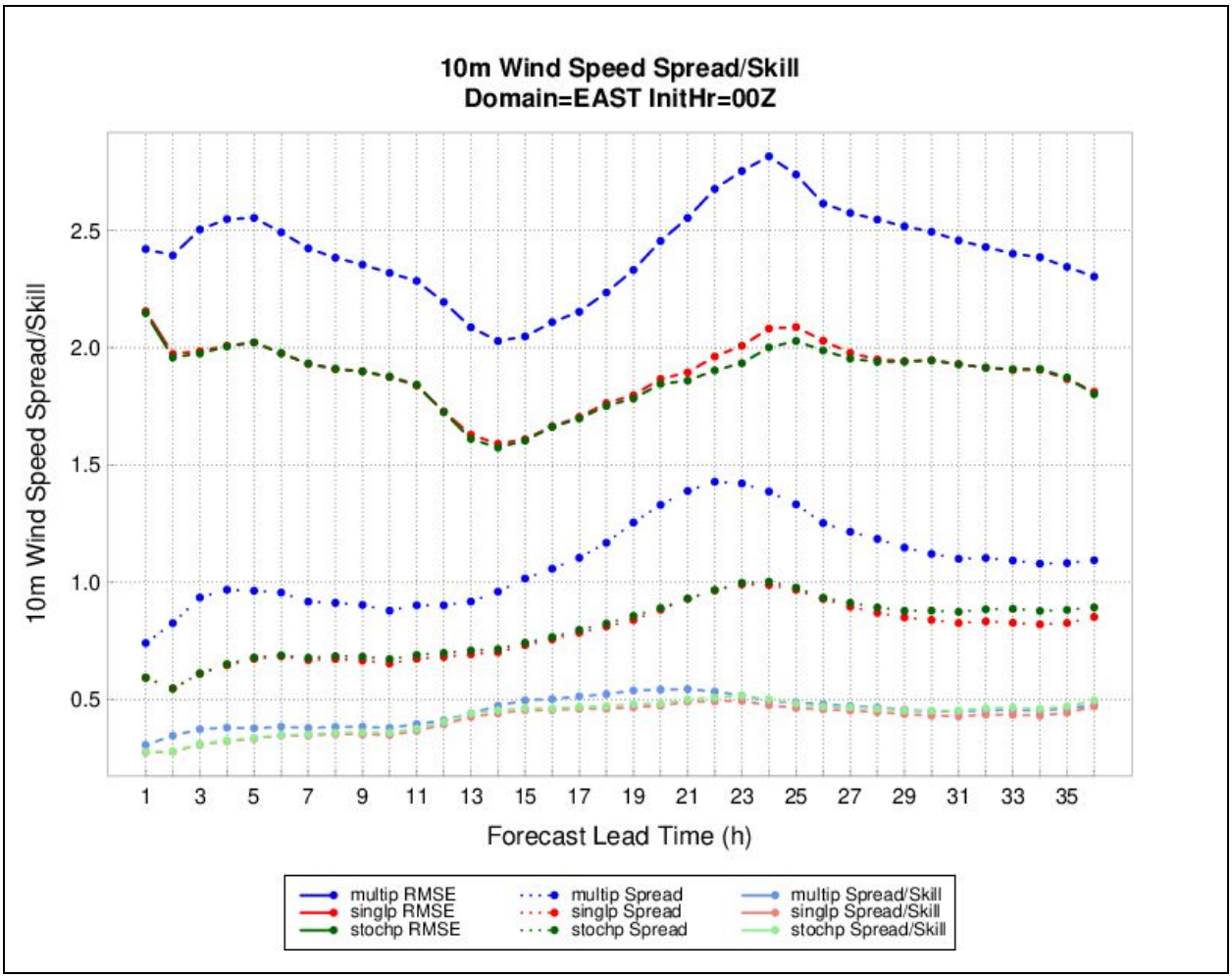


Figure 10. *Same as Fig. 2, except for 10-m wind speed (m s^{-1}).*

The reliability diagram for 10-m wind speed at a threshold of 5 m/s (resulting in a sample climatology of under 20%) indicates that, similar to 2-m dew point temperature none of the ensembles have skill (Fig. 11). In this case the single and stochastic ensembles are closer to the one-to-one line, but the only skill for any of the ensembles comes at the very lowest and highest forecast probabilities. For all other forecast probabilities, the ensembles over-forecast the observed frequency.

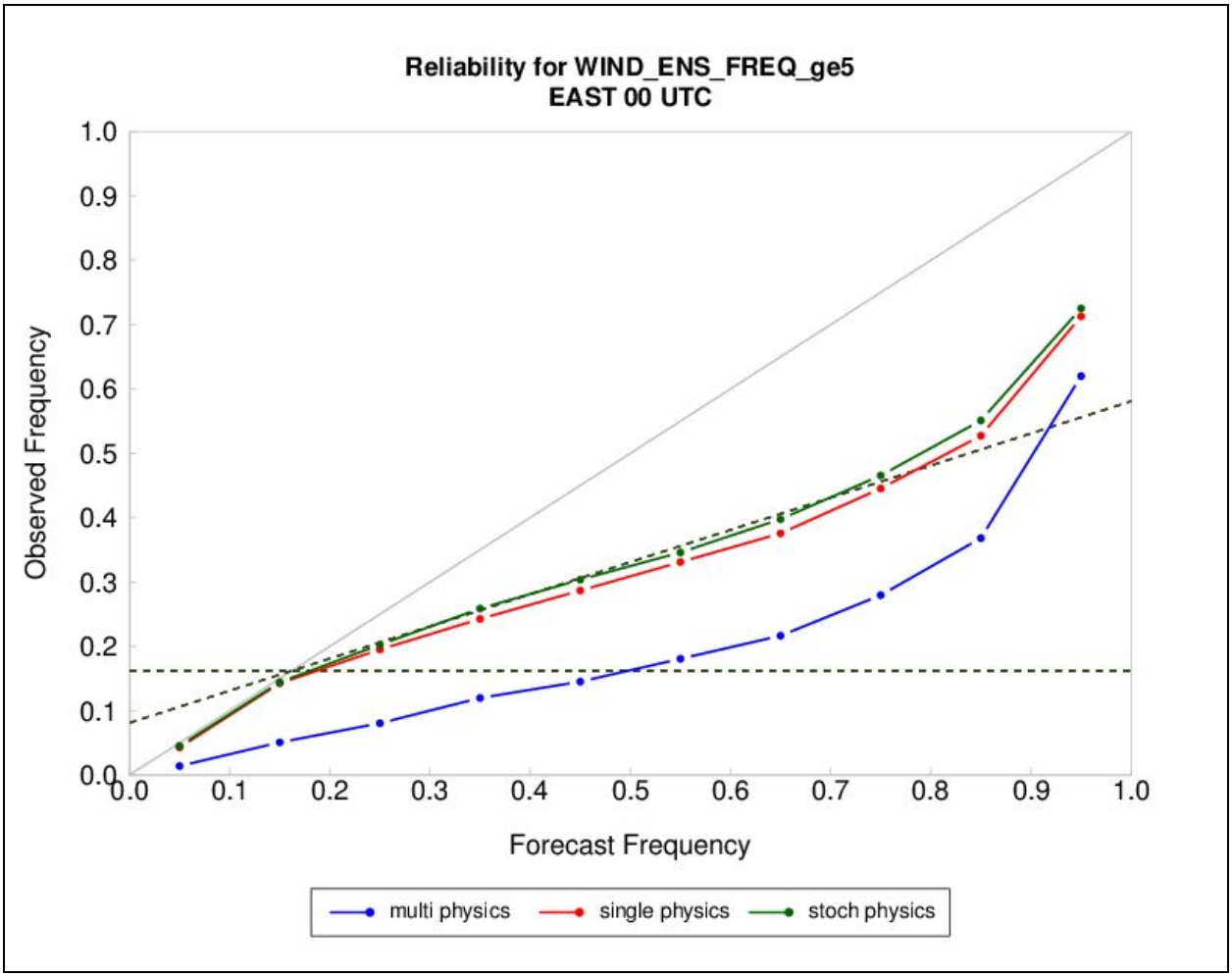


Figure 11. *Same as Fig. 3, except for 10-m wind speed (m s^{-1}).*

Finally, the rank histogram for wind speed aggregated over forecast hours 12 - 36 exhibits a high bias for all three ensembles. This trend is slightly worse for the multi-physics ensemble compared to the stochastic and single ensembles.

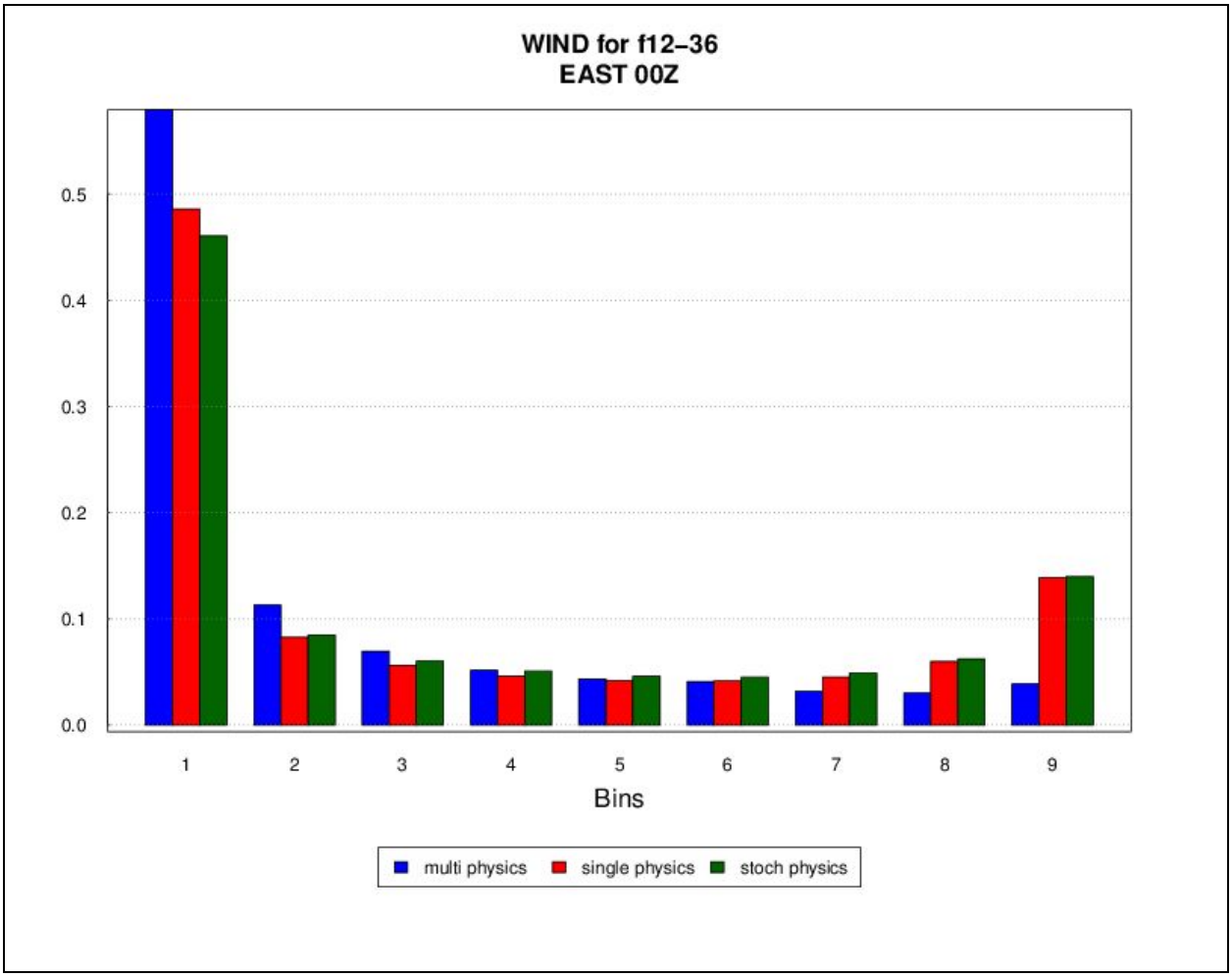


Figure 12. *Same as Fig. 4c, except for 10-m wind speed (m s^{-1}).*

Precipitation-related fields

Accumulated precipitation

For both 1-h and 3-h accumulated precipitation, focus was placed on >=2.54 mm threshold. All members for 1-h accumulated precipitation have maximum GSS (higher is better) at forecast hour 1 before a sharp decrease in skill at forecast hour 2, where the remainder of the period sees a gentle decrease in skill (Fig. 13a). The increase in skill at forecast hour 1 can likely be attributed to the benefits of data assimilation. A minimum in skill is noted from approximately 18 - 04 UTC, which is also coincident with the smallest variability between members. All three ensemble subsets have comparable GSS with no particular subset having superior skill.

When considering frequency bias (1 is unbiased, >1 is over-forecast, and <1 is under-forecast), a diurnal signal is noted in all ensemble subsets, with signals being slightly out-of-phase between the multi-physics ensemble and the stochastic physics and single physics ensembles (Fig. 13b). The multi-physics ensemble has peak and minimum frequency bias values occurring

3-6 hours earlier than the stochastic-physic and single-physics ensembles. A majority of all members have a high bias for the first several hours of the forecast period. Multi-physics typically has overall higher bias values than stochastic-physics and single-physics ensembles. The variability in the stochastic-physics and single-physics ensembles is limited in the first part of the forecast before becoming more variable in the middle-to-later part of the forecast. Overall, the variability in the frequency bias values is greatest in the multi-physics ensemble.

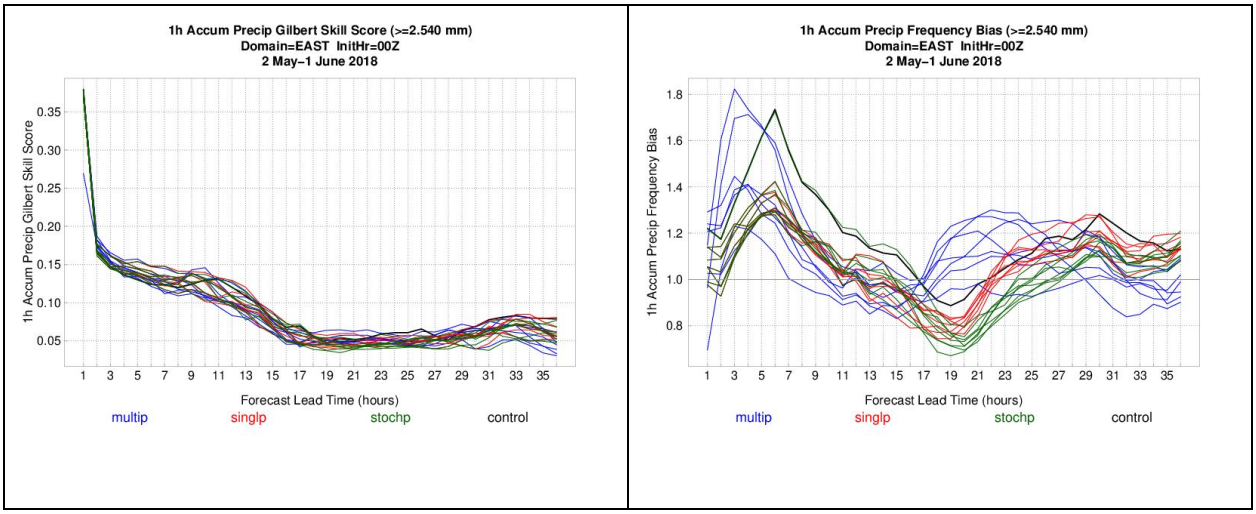


Figure 13. *(a) GSS and (b) frequency bias time series plots of 1-h accumulated precipitation >=2.54 mm for each individual ensemble member aggregated across the eastern half of the CONUS domain for all available forecasts during the experiment. The control member is in black, the multi-physics members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

Fractions Skill Score (FSS) (higher is better) for 1-h accumulated precipitation was calculated as a function of forecast lead time for two neighborhood widths: 3x3 grid squares, or 9x9 km, and 7x7 grid squares, or 21x21 km. FSS for both neighborhood widths display similar trends as seen in GSS (Fig. 14). All three ensemble subsets have similar performance and variability among members. As the spatial scale broadens (i.e., neighborhood width increases), there tends to be a shift toward higher scores, indicating members may have small displacement errors but generally have precipitation in the vicinity of the observational analyses.

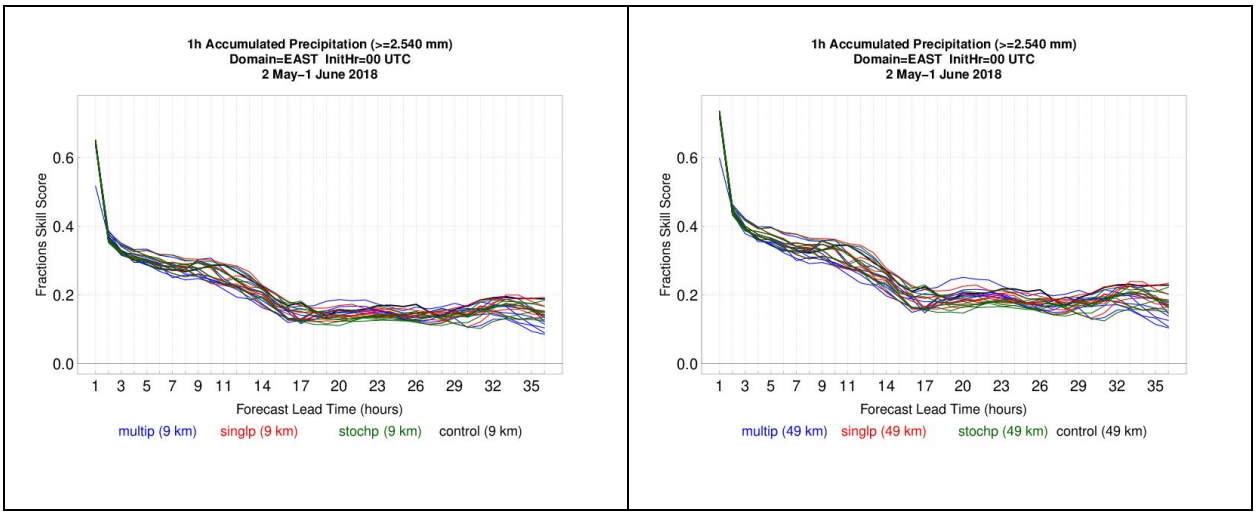


Figure 14: *FSS time series plots of 1-h accumulated precipitation >=2.54 mm for each individual ensemble member at a neighborhood width of (a) 3x3 grid squares and (b) 7x7 grid squares aggregated across the eastern half of the CONUS domain for all available forecasts during the experiment. The control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

For 3-hr accumulated precipitation, all members have maximum Gilbert Skill Score (GSS) values at forecast hour 3, before a sharp decrease at forecast hour 6 (Fig. 15a). From 6-h onward, there is an overall decrease in GSS values, with a modest diurnal signal in all ensemble subsets. Maximum GSS values are seen around 09 - 12 UTC while a minimum in GSS is noted between 21 - 03 UTC. Decent variability is noticed in all three ensemble subsets; however, all three subsets have comparable GSS with no particular subset having superior skill.

As seen in the 1-hr accumulated precipitation frequency bias, 3-hr accumulated precipitation also sees a pronounced diurnal signal in frequency bias (Fig. 15b). While all three subsets have peak values initially at forecast hour 6, a shift in peak values for day 2 is noted between the multi-physics (forecast hours 21 - 26) and stochastic physics and single physics (forecast hours 28 - 32). A low bias is noted in the multi-physics ensemble centered near forecast hour 15 in day 1 and between forecast hours 33 - 36 in day 2. The stochastic physics and single physics ensembles have a low bias centered on forecast hour 21. The multi-physics ensemble has the most variability among members, but the stochastic physics and single physics ensembles have increasing variability in the second half of the forecast period.

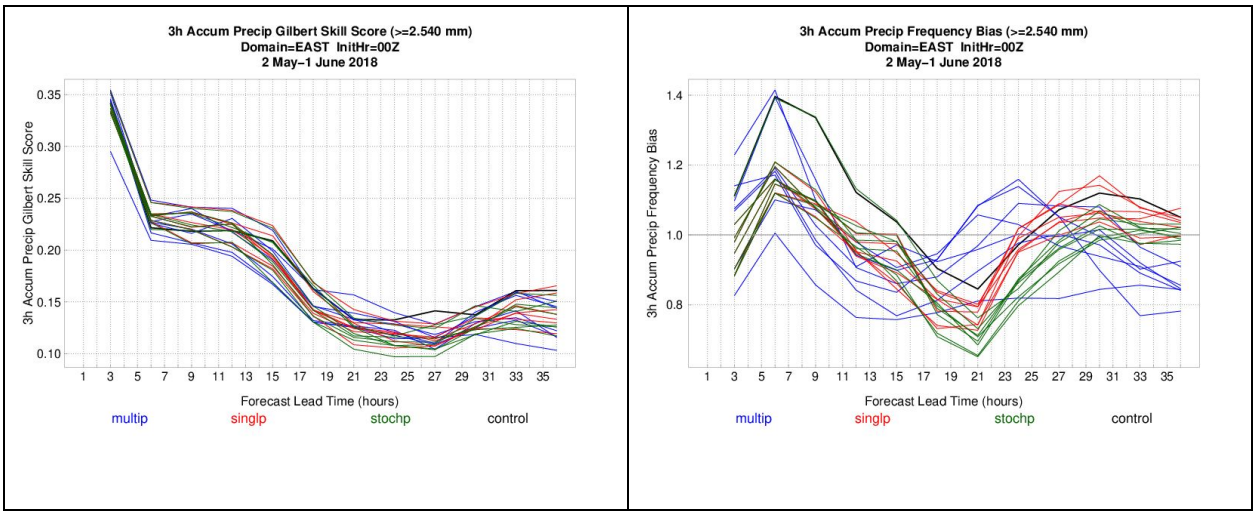


Figure 15. *(a) GSS and (b) frequency bias time series plots of 3-h accumulated precipitation >=2.54 mm for each individual ensemble member aggregated across the eastern half of the CONUS domain for all available forecasts during the experiment. The control member is in black, the multi-physics members are in blue, the single physics members are in red, and the stochastic physics members are in green.*

The total number of all simple MODE objects across the available set of forecasts (Fig. 16a) shows a clear diurnal signal and is apparent in both the observed and forecast object counts, with maximum (minimum) counts occurring 21 - 02 UTC (12 - 17 UTC). The diurnal trend is not surprising given typical convective climatologies for warm-season convection; when convection initiates, there are often a larger number of small convective features that eventually evolve to form larger mesoscale systems. While all of the ensemble subsets capture the diurnal signal in the object counts reasonably well, the ensembles trend towards too many objects through the forecast period, with the largest divergence occurring in the first 10 hours of the forecast period. In this time period, the stochastic physics and single physics ensembles are closely clustered, while the multi-physics ensemble has more variability. When the number of objects begins to increase around forecast hour 18, the ensemble subsets become more striated, with stochastic physics producing the least amount of objects, followed by single physics, and then a majority of the mixed-physics members producing the largest number of objects. At the end of the forecast period when the object counts experience a minimum, multi-physics members have less objects (closest to observations) than the stochastic and single physics ensembles, which are closely clustered.

The median object area (in grid squares) also displays a prominent diurnal trend (Fig 16b), with the largest (smallest) objects in the overnight night (daytime) hours from 5 - 8 UTC (18 - 22 UTC). All three ensembles capture the diurnal trend well; however, the ensembles overall have smaller objects areas compared to observations, with exception to when there is a minimum in object areas, and the ensembles capture the observations in the envelope of membership. No one ensemble subset outperforms the others when considering object areas.

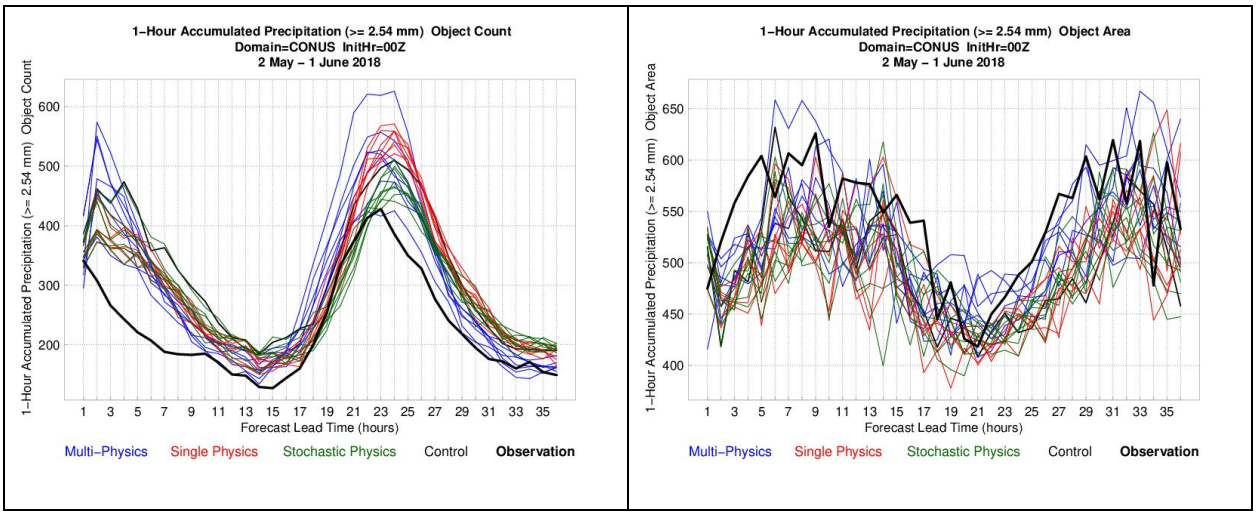


Figure 16: *(a) Total object count and (b) median object area (in grid squares) of 1-h accumulated precipitation >=2.54 mm object counts over the full CONUS domain for all available forecasts during the experiment. The observation objects are in bolded black, the multi-physics members in blue, the single physics members in red, the stochastic physics members in green, and the control in black.*

The displacement trends for 1-h accumulated precipitation objects can be investigated with the centroid attribute derived from MODE. This is accomplished 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. All three ensemble subsets show an immediate westerly bias at the beginning of the forecast period before having select members transition to an easterly bias from forecast hour 11 - 15. In addition, from approximately forecast hour 21 - 27, all three ensemble subsets are clustered tightly around zero, with minimal variability among members. At the end of the forecast period, the stochastic and single physics ensembles have members equally distributed about the zero line, while the multi-physics ensemble has a shift toward a more easterly displacement bias. When considering the north/south displacement, all ensemble subsets have a small northerly bias right at the beginning of the forecast period before a general shift toward a southerly bias. This gently transitions to an overall northerly bias in a majority of members for all three ensemble subsets from approximately the 21 hour forecast onward.

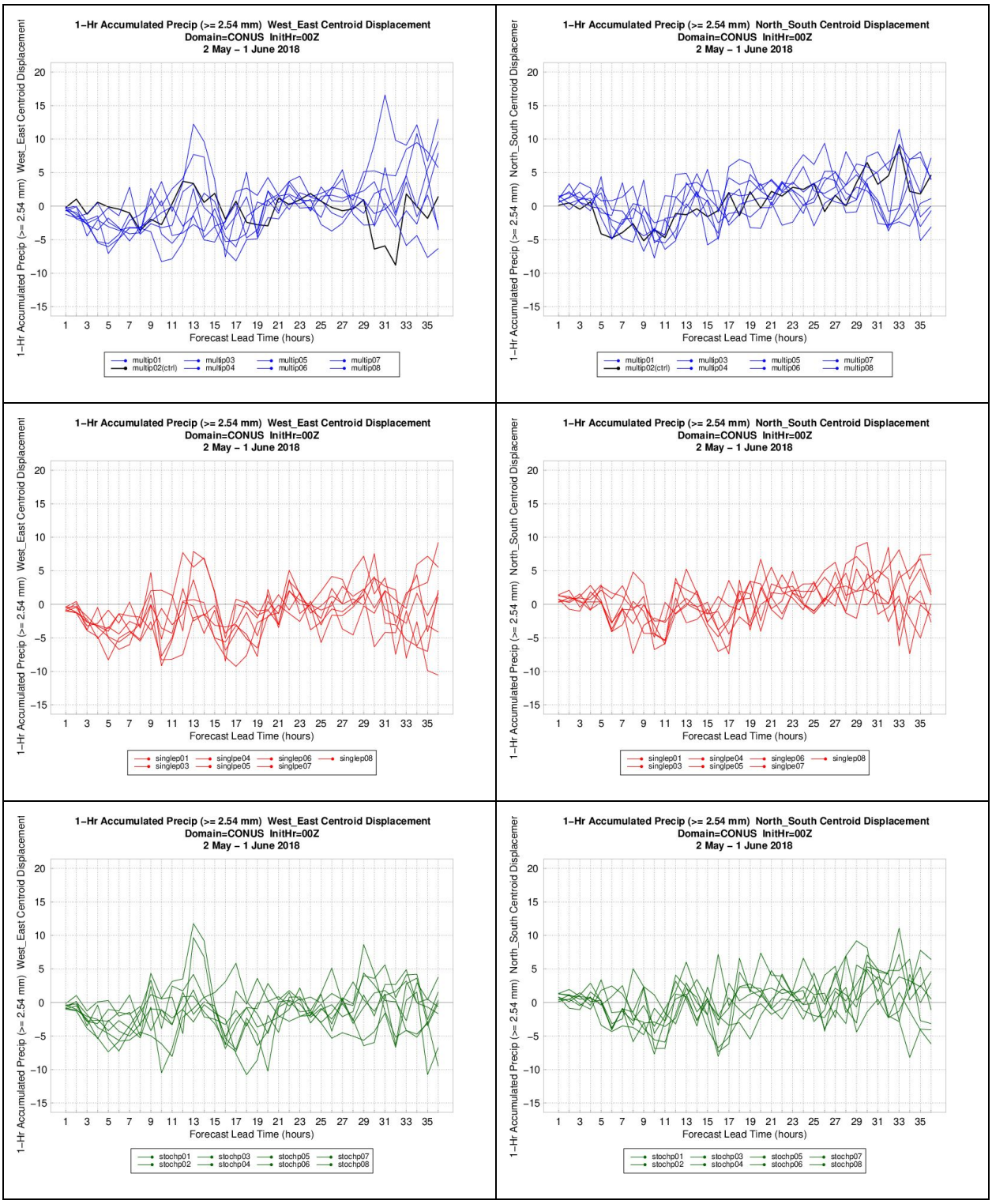


Figure 17: *Centroid displacement in the west-east direction (left) and south-north direction (right) for the multi-physics (top; blue), single physics (middle; red), and stochastic physics (bottom; green) ensemble members for 1-h accumulated precipitation objects >=2.54 mm aggregated over the full CONUS domain for all available forecasts during the experiment.*

Overall, the three ensembles have comparable RMSE and spread values for most lead times. The spread for each ensemble in terms of 1-h accumulated precipitation actually exceeds the RMSE values. This leads to a spread/skill ratio greater than one and indicates that for this variable there is sufficient spread to account for the amount of error.

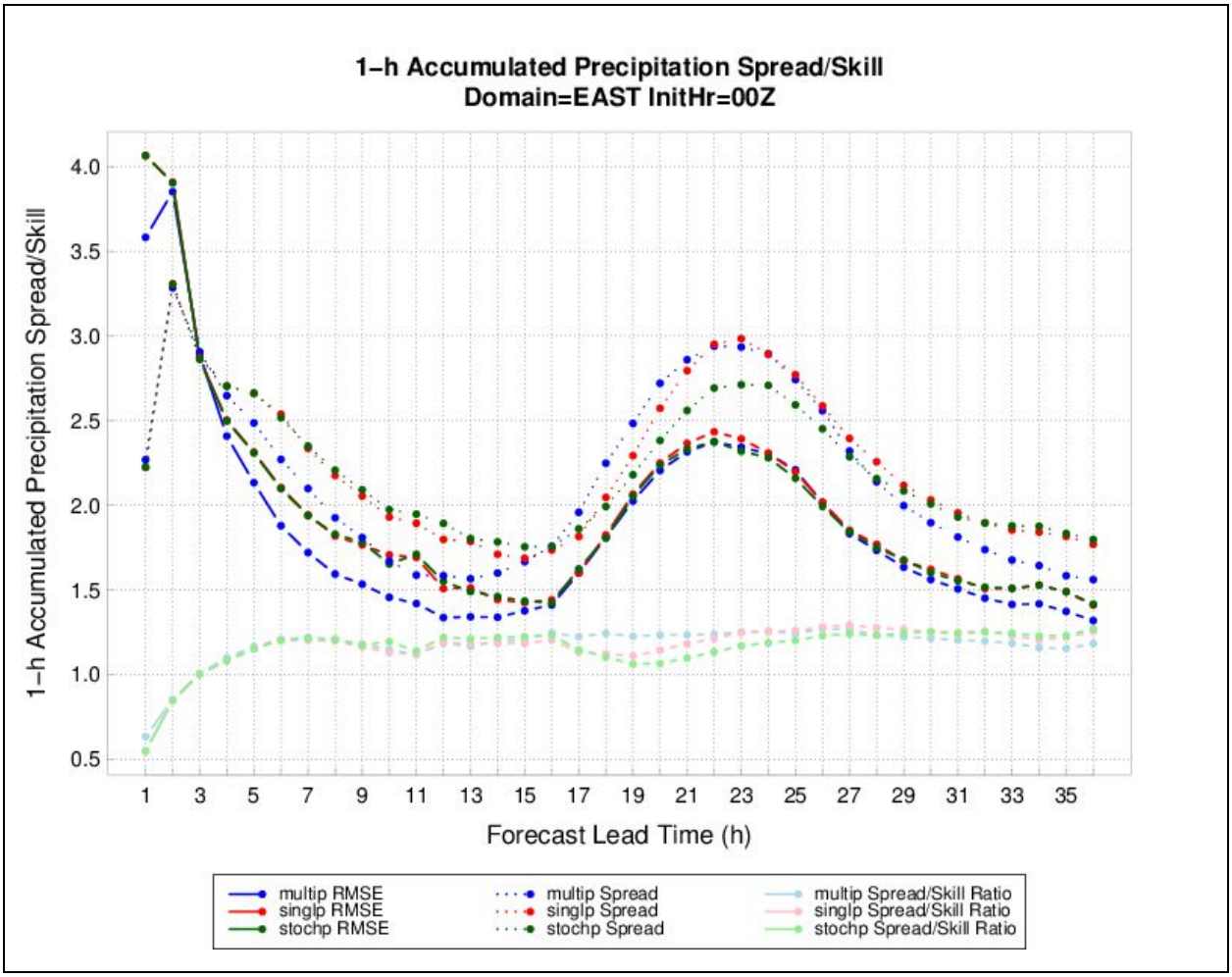


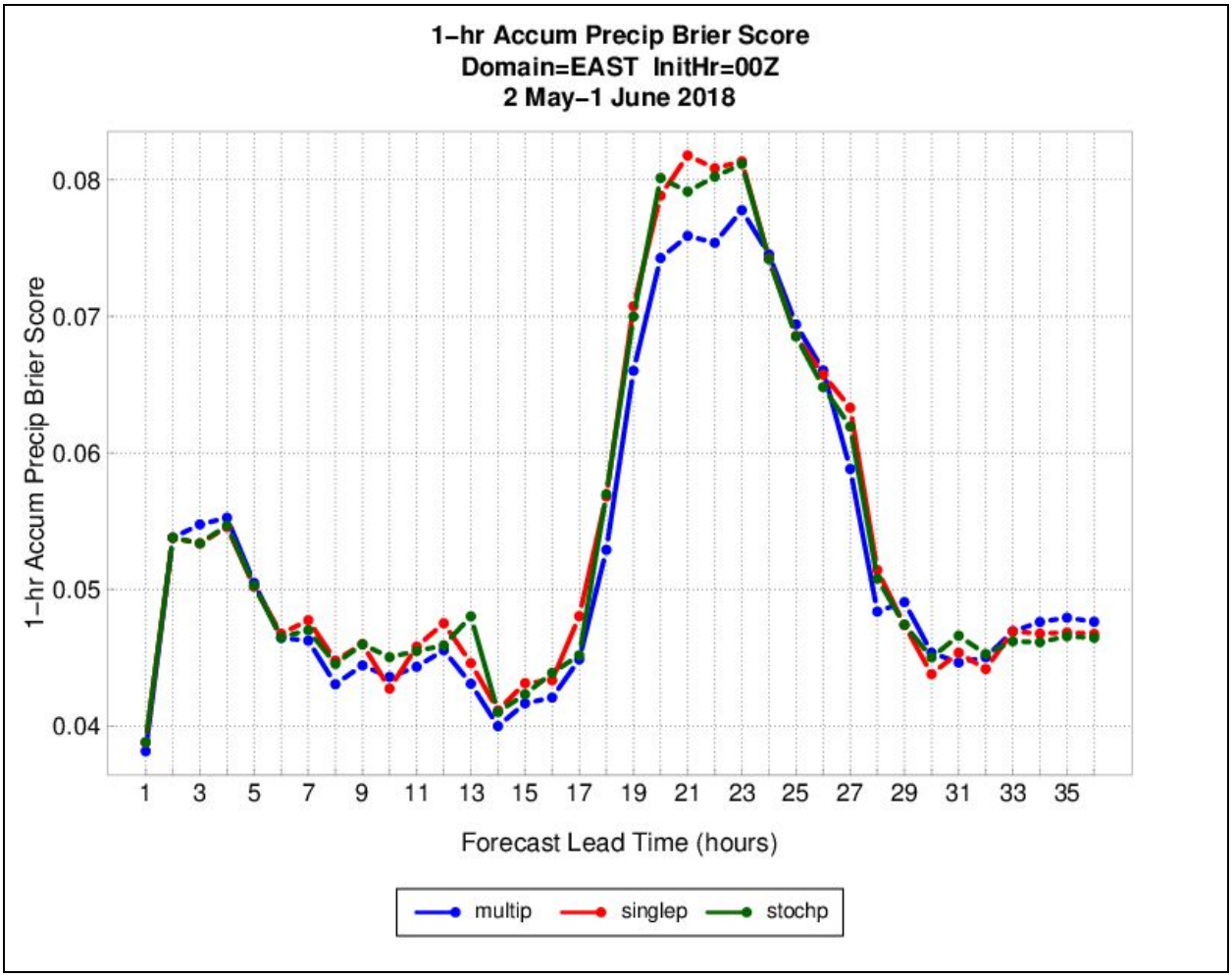
Figure 18. *Same as Fig. 2, except for 1-h accumulated precipitation.*

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 4, then decreasing to a minimum at forecast hour 14, after which, the BS increases to reach a maximum at forecast hours 20 through 23, and then decreasing throughout the rest of the forecast (Fig. 19). The ensemble subsets

demonstrate a similar trend in values with the main difference occurring during the peak, 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. The reliability diagram examined here was created from 1-hour accumulated precipitation $\geq$ 0.254 mm aggregated over a 24-hour time period between forecast hours 12 – 36 (Figure 19b). All three ensembles follow the one-to-one diagonal with near perfect reliability for forecast frequencies below 40%. Above that threshold, all three ensembles fall below the diagonal indicating overconfidence in forecasting an event.

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 19c looks at the same aggregated precipitation as examined in figure 19b. 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 single then stochastic physics ensembles.



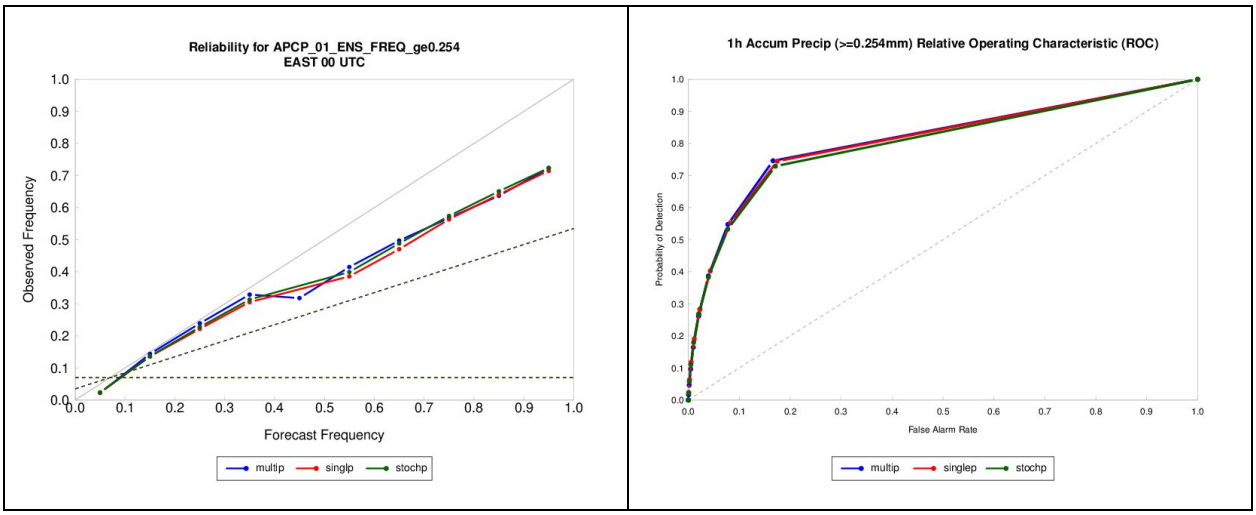


Figure 19. *(a) BS time series plot of 1-h accumulated precipitation >=0.254 mm for each ensemble member aggregated across the eastern half of the CONUS domain for all available forecasts during the experiment. (b) Reliability diagram for 1-h accumulated precipitation >=0.254 mm aggregated over forecast hours 12-26 across the eastern half of the CONUS domain for all available forecasts during the experiment. (c) ROC curve for 1-h accumulated precipitation >=0.254 mm aggregated over forecast hours 12-26 across 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 in red, and the stochastic physics ensemble in green.*

Composite radar reflectivity

For composite reflectivity, focus was placed on >=30 dBZ threshold. Similar to accumulated precipitation, GSS and frequency bias were calculated for each ensemble member as a function of lead time, aggregated over all available forecasts. All three ensemble subsets have comparable GSS with no particular subset having superior skill (Fig. 20a). Maximum GSS values are seen at the first forecast hour, with an overall decrease in skill throughout the remainder of the forecast period. A very gentle diurnal signal is noted, with a minimum approximately from forecast hour 17 - 28.

When looking at frequency bias, a diurnal signal is noted in all ensemble subsets, with signals being slightly out-of-phase between the multi-physics ensemble and the stochastic physics and single physics ensembles (Fig. 20b). The multi-physics ensemble has peak and minimum frequency bias values occurring 3-6 hours earlier than the stochastic-physic and single-physics ensembles; this behavior is similar to results seen for accumulated precipitation as well, potentially pointing to differences in convective initiation and evolution between the subsets. The stochastic and single physics ensemble subsets have minimal variability among membership and perform similarly to one another, with high biases noted overnight into the morning hours. The multi-physics ensemble displays far greater variability among membership, with a majority of members having a neutral-to-high bias.

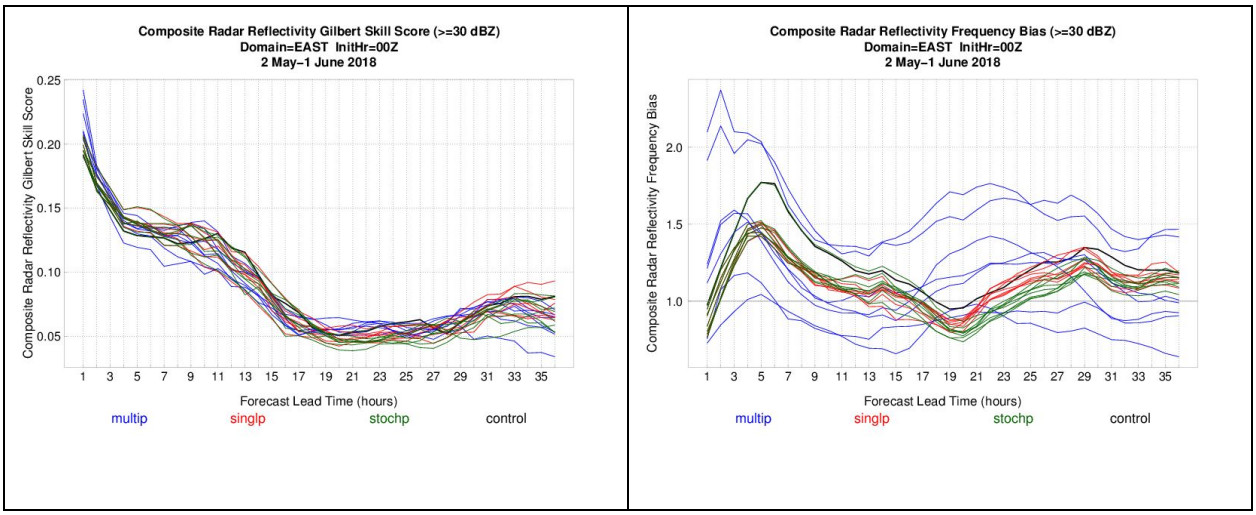


Figure 20. *Same as Fig. 13, except for composite reflectivity >= 30dBZ.*

Similar to 1-h accumulated precipitation, FSS for composite reflectivity >=30 dBZ was calculated as a function of forecast lead time for two neighborhood widths: 3x3 grid squares, or 9x9 km, and 7x7 grid squares, or 21x21 km. FSS for both neighborhood widths and for all three ensemble subsets have similar temporal trends to GSS, with highest skill at the earliest lead times and a general decrease in skill throughout the forecast period (Fig. 21). All three ensemble subsets have similar performance and variability among members. As neighborhood width increases, there tends to be a shift toward higher scores (Fig. 21b), indicating members may have small displacement errors but generally have precipitation in the vicinity of the observational analyses.

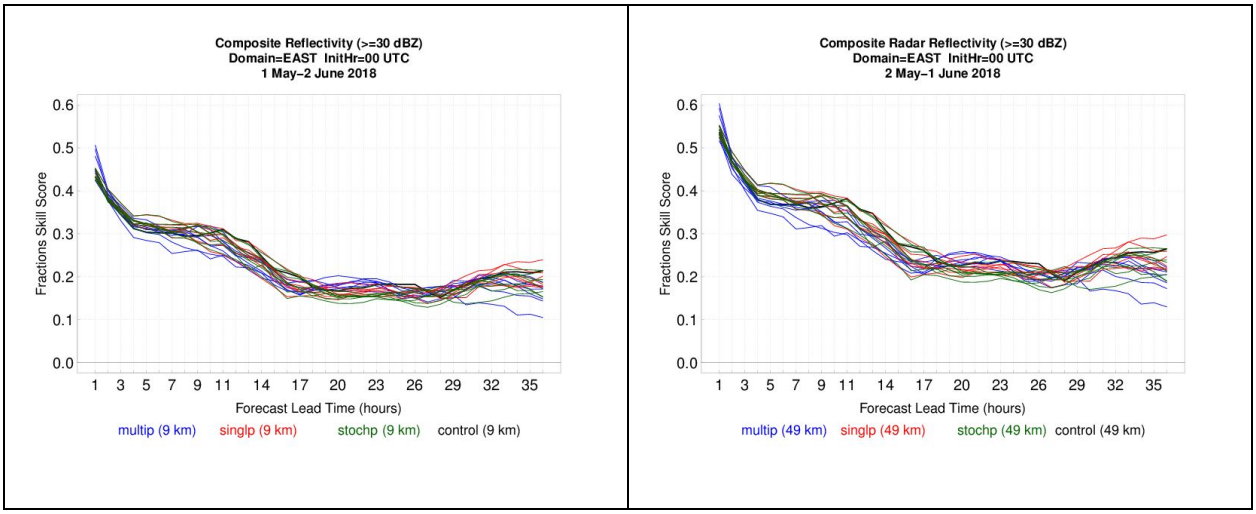


Figure 21: *Same as Fig. 14, except for composite reflectivity >= 30dBZ.*

The total number of all simple MODE objects for composite reflectivity >=30 dBZ shows diurnal signal in both the observed and forecast object counts, with maximum (minimum) counts occurring 22 - 01 UTC (12 - 17 UTC). This result is similar to the object counts seen for the 1-h accumulated precipitation. While all of the ensemble subsets capture the diurnal signal in the

object counts reasonably well, the ensembles, overall, produce too many objects, with the largest divergence from observations occurring when there is a peak in observed objects. The multi-physics ensemble displays a large variability among members, typically having the smallest and largest object counts, with the stochastic and single physics ensembles having much less variability among members. The median object area also displays a prominent diurnal trend, with the largest (smallest) objects in the evening/overnight night (afternoon) hours. All three ensembles capture the diurnal trend well; however, the ensembles overall have smaller objects areas compared to observations, with exception to when there is a minimum in object areas, and the ensembles capture the observations in the envelope of membership. No one ensemble subset outperforms the others when considering object areas.

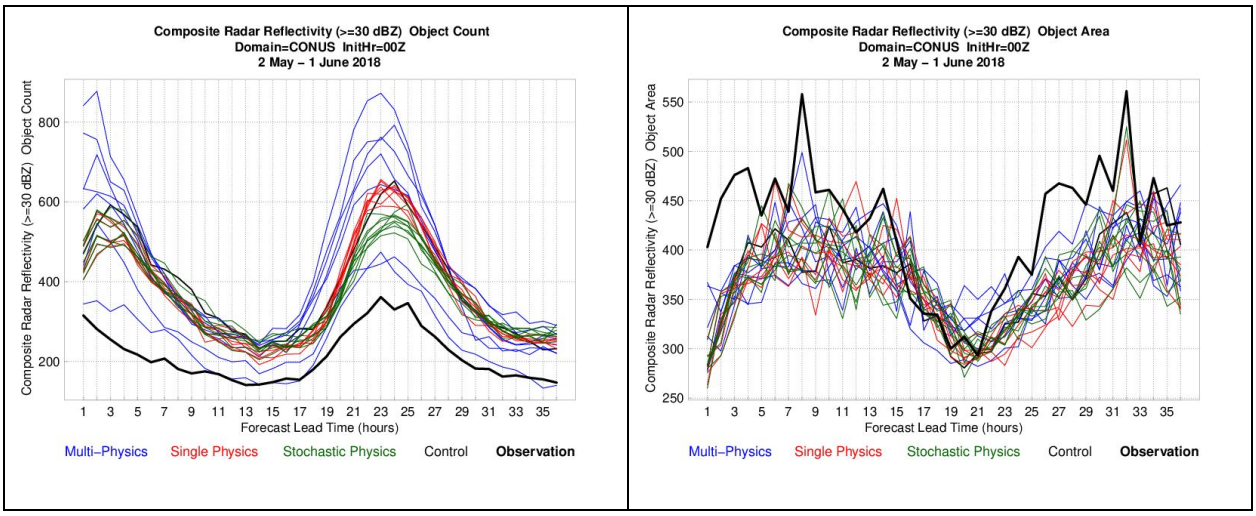


Figure 22: *Same as Fig. 16, except for composite reflectivity >= 30dBZ.*

Similar to the 1-h accumulated precipitation objects, the composite reflectivity objects >=30 dBZ for all three ensemble subsets have a westerly bias at the beginning of the forecast period. In general, the stochastic and single physics ensembles have similar behavior and variability among members, with most members having a westerly bias throughout the forecast period. The multi-physics ensemble, however, has members spread about the zero line in the middle of the forecast period before transitioning to an overall easterly bias by the end of the forecast. When considering the north/south displacement, the stochastic and single physics ensemble subsets again have similar behavior, with a more pronounced northerly bias at the end of the forecast period. The multi-physics ensemble subset has a number of members with a northerly bias in the middle of the forecast period before a shift at the end of forecast period to having much greater variability around the zero line. This difference may be associated with the phase shift noted in other metrics previously between the multi-physics and the single/stochastic physics members.

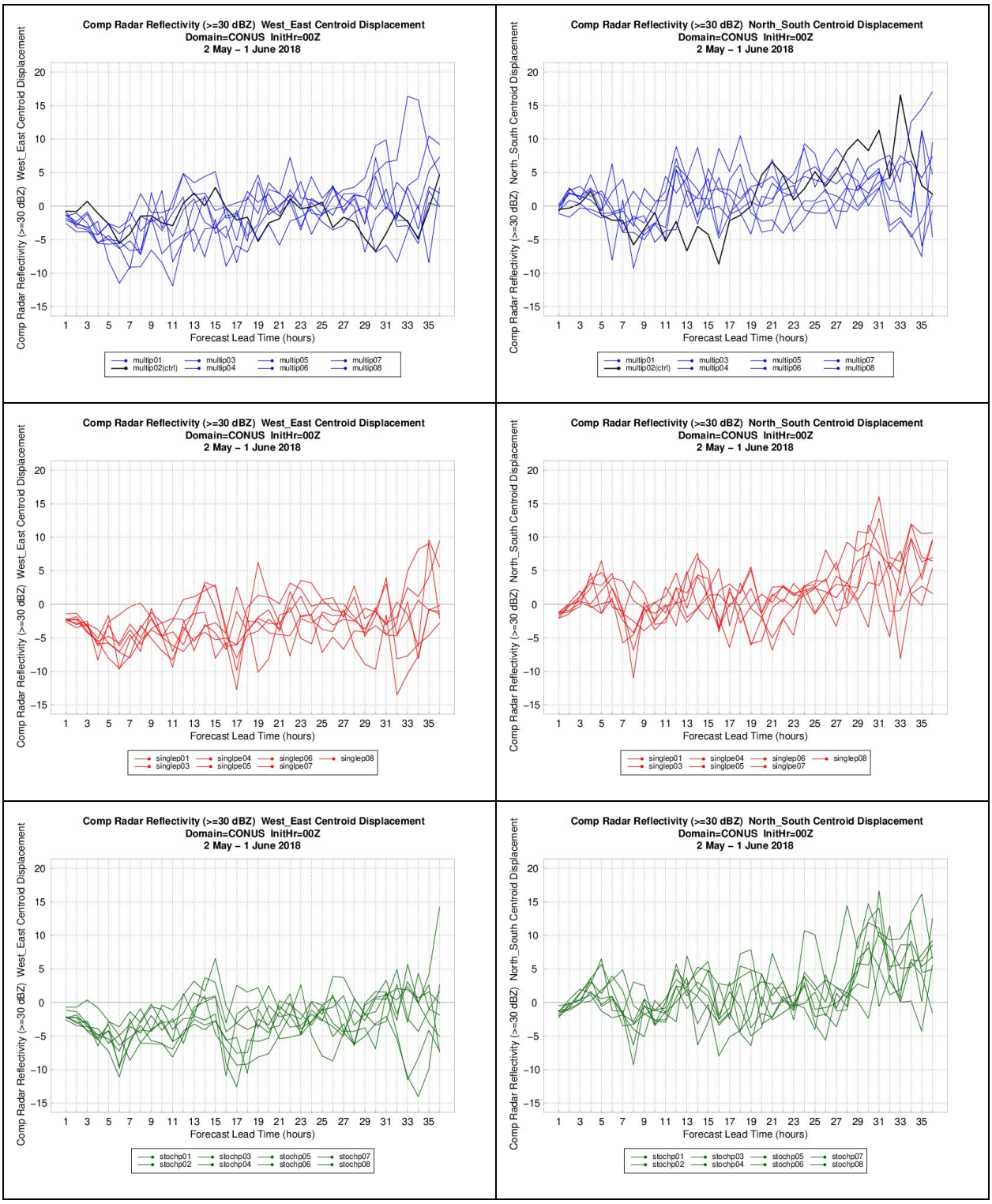


Figure 23: *Same as Fig. 17, except for composite reflectivity >= 30dBZ.*

While the single and stochastic ensembles have very similar RMSE, they do differ from the multi-physics ensemble at certain lead times (during the afternoon/evening time periods).

The spread for the multi-physics ensemble is considerably higher than that for the single and stochastic ensembles leading to a spread/skill ratio closer to one for that ensemble.

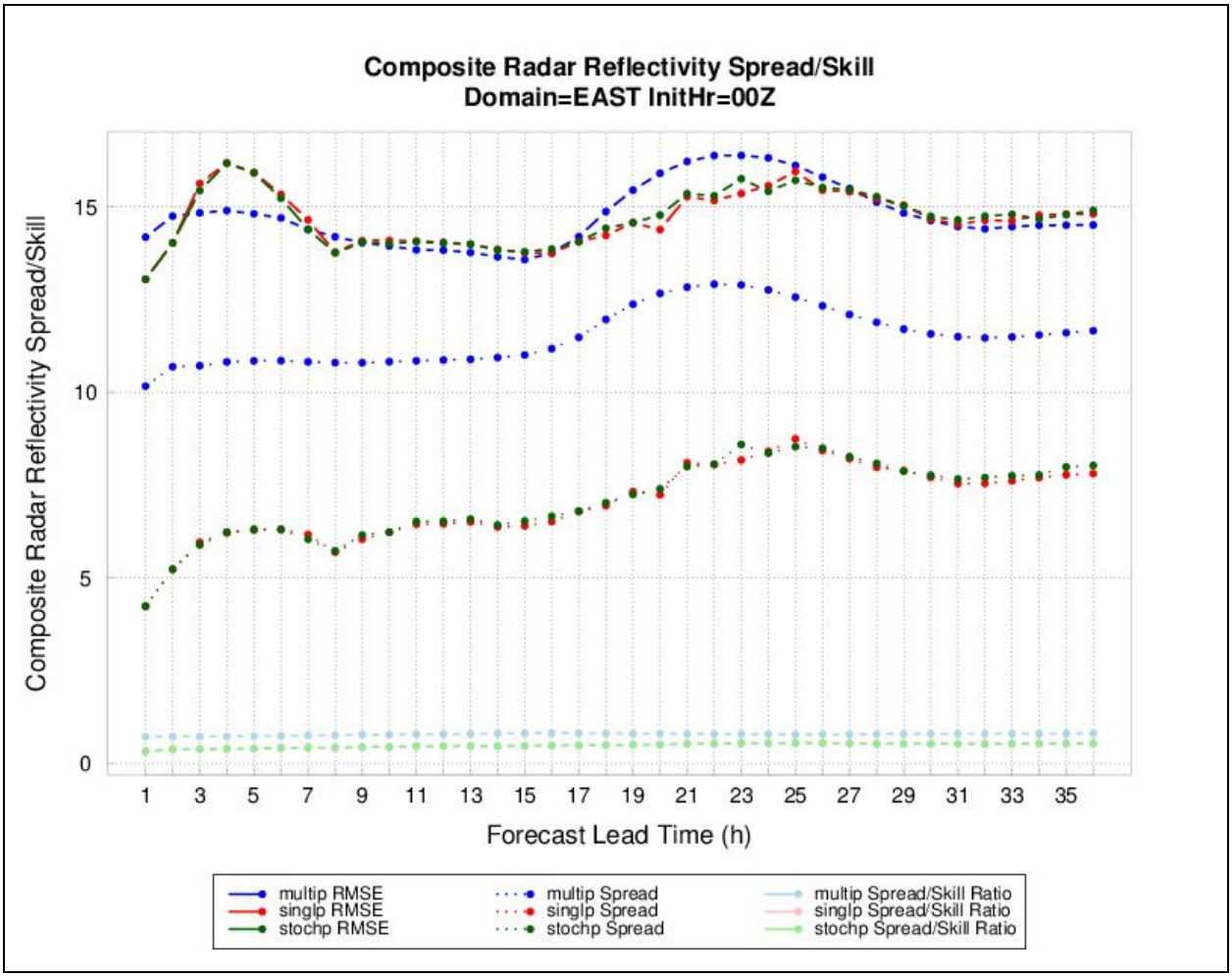


FIgure 24. *Same as Fig. 2, except for composite reflectivity >= 30dBZ.*

While we see similar temporal trends in the BS for all three ensembles, in the case of composite reflectivity, the stochastic and single physics ensembles perform slightly better than the multi-physics ensemble during the convectively active afternoon/evening hours (Fig. 25a). During other time periods, the performance between the three ensemble subsets is very similar.

When looking at reliability, all three ensembles are at or below the no-skill line (dotted diagonal line) for the majority of forecast frequencies (Fig. 25b). There is a slight increase in area on the ROC for the multi-physics ensemble as seen in Fig 25c.

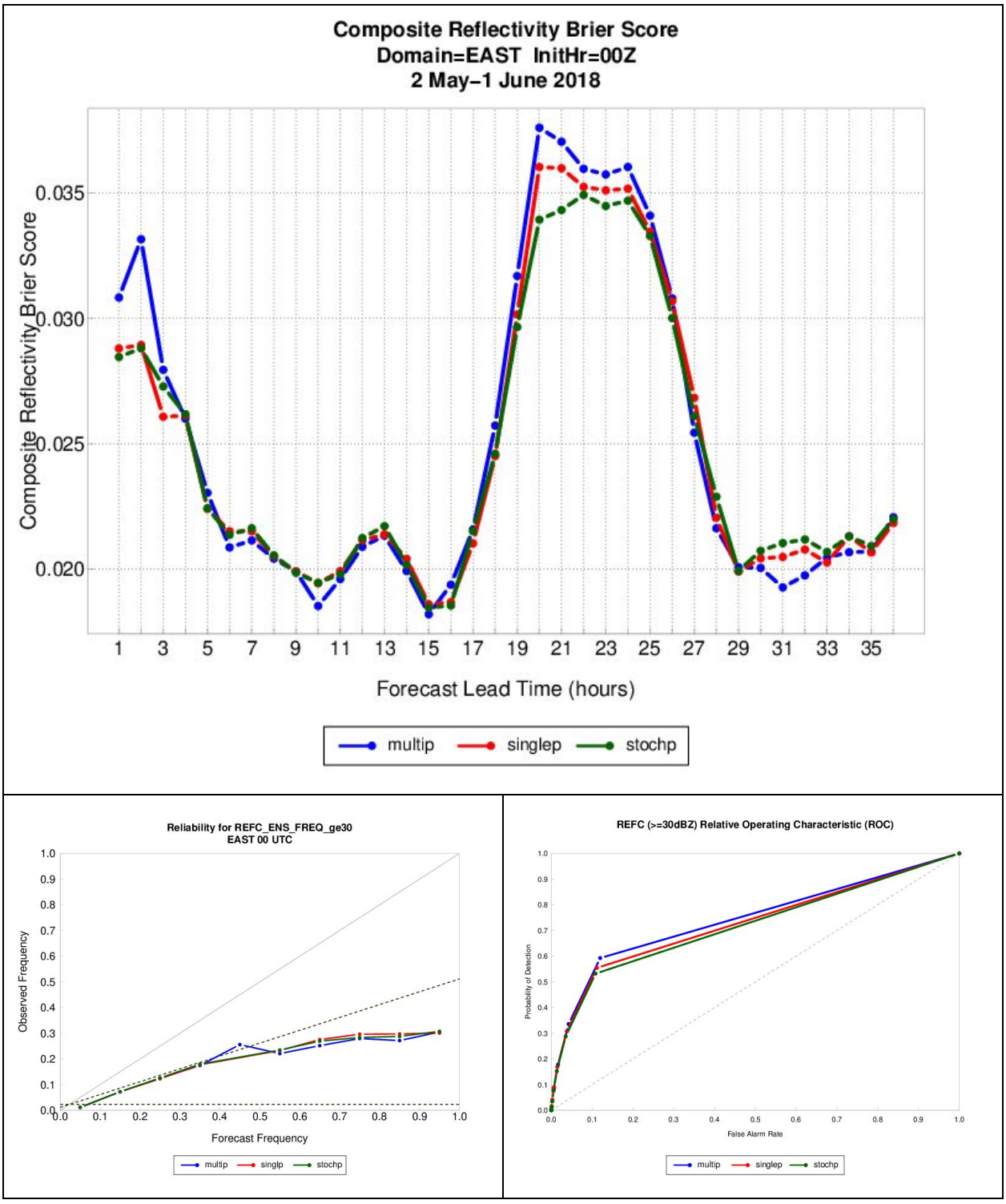


FIgure 25. *Same as Fig. 19, except for composite reflectivity >= 30dBZ.*

Summary

Overall, when looking at near surface temperature and dew point temperature, the multi-physics ensemble outperforms the stochastic and single ensembles for most metrics examined. This result does not hold for near-surface wind speed, where the stochastic and single physics ensembles generally have lower high bias and better performance across the board. While the ensembles are fairly reliable when predicting 2-m temperature they are all under-dispersive, as indicated by the rank histograms. When assessing 2-m dew point temperature and 10-m wind speed, the ensembles lack reliability, likely due to the large biases that exist for those variables.

For accumulated precipitation and composite reflectivity, traditional deterministic and probabilistic verification metrics show little difference in overall performance between the three ensemble subsets. One interesting trend to keep in mind is the phase shift in timing of frequency bias results between the multi-physics and two other ensembles. When using MODE to assess the performance of precipitation-related fields through object-based measures, a few additional trends can be teased out. The ensemble subsets capture the diurnal signal reasonably well, though they trend towards too many objects throughout the forecast period. The maximum offset is during the peak of convective initiation. In general, the ensemble members forecast objects that are too small compared to the observations, though, again, they are able to capture the temporal trend well. A fairly consistent result among the members of all ensembles is also an immediate westerly bias at the beginning of the forecast period. While the stochastic ensemble members generally trend closest to the observed counts, the other MODE attributes perform similarly between the different ensembles.