Addressing model uncertainty through stochastic parameter perturbations within the HRRR ensemble

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

Regional Ensemble Team: Jamie Wolff[1], Jeff Beck[2], Greg Thompson[1], Isidora Jankov[2], Michelle Harrold[1], Mike Kavulich Jr.[1], Tressa Fowler[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 2

Experiment design 2

Model 2

SPP applied to graupel treatment 3

SPP applied to cloud water distribution 4

SPP applied to CCN activation 5

SPP experiments 5

Observations 6

Model verification approaches 6

Deterministic verification metrics 6

Ensemble verification metrics 7

Verification results 7

Surface verification 7

Traditional 7

Ensemble 9

Upper-air verification 13

Traditional 13

Ensemble 15

Precipitation and composite reflectivity verification 17

Traditional 17

3-h Accumulated Precipitation 17

Composite Reflectivity 21

Ensemble 23

3-h Accumulated Precipitation 23

Comparison of Ensemble Members to Control Member 25

Summary 28

References 31

Introduction

In most existing regional ensemble systems, model-related forecast uncertainty is addressed by using multiple dynamic cores, multiple physics suites, or a combination thereof. While such multi-model ensembles have demonstrated potential, their maintenance is resource-intensive. More importantly, probabilistic forecasts from multi-model ensembles do not have consistent distributions since each member can have a different mean error and variance. Post-processing generally assumes independent and identically distributed random variables, a requirement that is not met by multi-model ensemble systems. To facilitate a sustainable and unified operational forecasting system, we propose to extensively test an alternative option for creating desirable spread and reliability by perturbing the members stochastically within a storm-scale ensemble. The stochastic-dynamic approach results in statistically consistent ensemble distributions. Two widely used stochastic schemes are the Stochastic-Kinetic Energy Backscatter (SKEB; Shutts 2005, Berner et al. 2009) and the Stochastic Perturbations of Physics Tendencies (SPPT; Buizza et al. 1999, Palmer et al. 2009). These methods are formulated to represent the effect of unresolved subgrid-scale variability and are added a posteriori to independently tuned models. An additional approach is the Stochastic Parameter Perturbation (SPP) scheme, which targets parameter uncertainty in the physical parameterization schemes directly.

Experiment design

Model

In 2016, the DTC Regional Ensemble Task successfully built the necessary infrastructure for execution of a High Resolution Rapid Refresh (HRRR) ensemble for prolonged retrospective runs [the HRRR represents a specific configuration of the Weather Research and Forecasting (WRF; Skamarock et al. 2008) model using a highly tuned set of physical parameterizations]. The developed framework also took advantage of an option in WRF to generate random perturbations that can be applied to a variety of parameters in physics schemes.

Based on promising results from 2016, where the SPP option was implemented in both the HRRR Planetary Boundary Layer (PBL) and Land Surface Model (LSM) schemes (results described in Jankov et al. 2017), the DTC Regional Ensemble Task proposed continued testing of stochastic physics in 2017. Research for this performance period focused on implementing SPP within the Thompson microphysics scheme (Thompson and Eidhammer 2014). In particular, the SPP method was employed to to treat known parameter uncertainties for three aspects within the scheme: (a) the relationship used to specify the Y-intercept parameter of the assumed inverse exponential size distribution for graupel, (b) the shape factor of the generalized gamma size distribution of cloud water droplets, and (c) cloud condensation nuclei activation via consideration of sub-grid scale eddies with higher supersaturation.

SPP applied to graupel treatment

The first stochastic parameter perturbation applied relates to parameters that fundamentally control the size spectra of the graupel/hail hybrid category. The assumed number density function for this category follows a generalized gamma distribution of the following form:

$$N(D) = N_0 \; D^{\mu} \; e^{-\lambda D}, \qquad \text{(Eq. 1)}$$

where D is the droplet diameter, N_0 is the intercept parameter, μ is the shape parameter, and λ is the slope.

The Thompson bulk microphysics parameterization was specifically designed to predict only one free variable of a mixed or hybrid graupel/hail category, its mass mixing ratio, in order to reduce computational cost as compared to fully double-moment schemes. One-moment schemes typically assume an inverse-exponential size distribution (μ = 0 in Eq. 1) with an a priori assigned and constant Y-intercept parameter. Numerous observations from aircraft and surface measuring campaigns (e.g., McFarquhar and Black, 2004; Knight et al, 1982) generally support this distribution shape, although the intercept parameter has been known for decades to vary by as many as 2-3 orders of magnitude.

Since using a fixed intercept parameter was known to be a problem when G. Thompson developed the scheme, he used a relationship combining graupel mass mixing ratio and amount of supercooled liquid water to compute a space/time-varying Y intercept parameter diagnostically during a simulation. From prior observational studies, the intercept parameter is permitted to vary from 10^4 to 10^6 m^{-4} consistent with overall observations, but the diagnostic relationship itself was ad hoc and not well tested. The McFarquhar and Black (2004) observations contradict the scheme's existing diagnostic relation for decreasing intercept parameter as a function of higher graupel mixing ratio (Fig. 1); although their observations were collected in tropical storms so their applicability to mid-latitude deep convection is unknown.

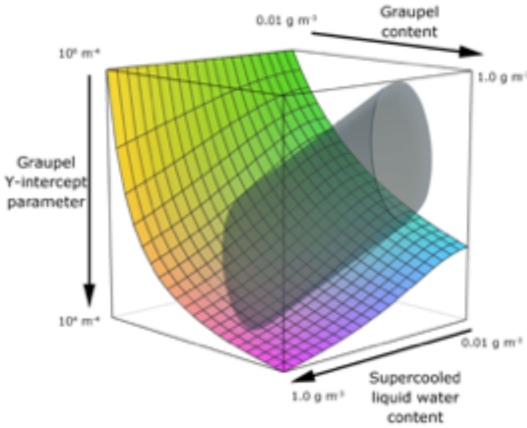


Figure 1. *The current graupel/hail Y-intercept relationship is shown as the colored parabolic surface as a function of both graupel mixing ratio and supercooled water content. Note that as graupel mixing ratio increases, the intercept parameter decreases. The SPP technique will result in a "cloud" of points represented by the semi-transparent gray cylinder.*

Therefore, the SPP technique together with a pre-determined probability density function (PDF) that aligns well with the variability found in observations to choose a variable Y-intercept parameter was employed. Rather than choosing an entirely randomly generated number that is directly scaled to become a value of intercept parameter, an ellipsoid of values that produce the highest probability of occurrence to match the McFarquhar and Black (2004) observations was created.

SPP applied to cloud water distribution

The SPP technique was also applied to the value of the shape parameter (μ) of the generalized gamma distribution of the cloud water variable to vary between 2 and 15. In the case of the standard code without the use of SPP, the value of μ is determined diagnostically from the cloud droplet number concentration, which was chosen to follow observations by Martin et al (1994). Their data also showed considerable spread, so we choose to permit a variety of possible outcomes of the diagnosed value of μ using SPP while centering the perturbation near zero but permitting deviations of 6 integer values. In other words, if the diagnosed μ resulted in 8, then it was perturbed to vary as far as 2 and 14, but the nature of the Gaussian perturbations would rarely result in values reaching those extremes. Using μ to alter the size distribution assumption greatly alters the mean diameter of the particles (Fig. 2), which will have an immediate impact on the initial formation of rain from the scheme's autoconverson. Other potential side effects of changing cloud droplet sizes include snow riming (collection of cloud water by falling snow), rate of ice nucleation (which is a function of drop size), and the potential for lofting the smaller cloud droplets higher in the vertical within convective updrafts. As such, the total changes to the entire system are highly non-linear and include initial rain formation, cloud growth, ice formation, etc. and, therefore, have the potential for greatly impacting evaporative cooling and resulting convective cold pools (Morrison et al., 2012). By altering cold pool strength and propagation speed, there will be impacts to convective forecasts, especially after convective initiation.

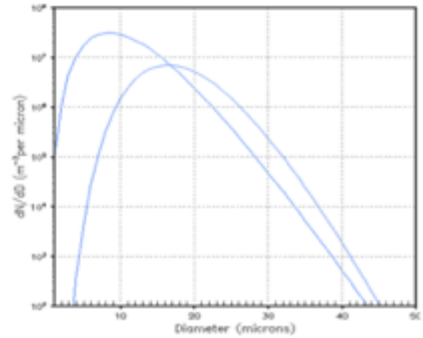


Figure 2. *Number distribution of particles as a function of diameter (Eq. 1) using a generalized gamma function with shape parameter μ = 2 (left curve) and μ = 12 (right curve).*

SPP applied to CCN activation

The final application of SPP was designed to affect the activation of water-friendly aerosols as cloud condensation nuclei (CCN). In general, the Thompson and Eidhammer (2014) aerosol-aware microphysics parameterization is likely to produce too few CCN because it utilizes only the WRF model grid-scale vertical velocity, *w*. Since even convection-allowing models at approximately 3-km grid spacing do not fully predict all scales of vertical motions in the atmosphere, we decided to account for some of this uncertainty by perturbing *w* **only** for the purpose of determining what number of CCN activate from a look-up table. In order to avoid a lack of CCN activation, we only applied positive perturbations to *w* since negative perturbations could inhibit proper CCN activation entirely. Also, due to comparisons of cloud water drop number and size in a climatological sense by Thompson et al. (2017), we determined that boosting the CCN activation might generally lead to a better match to observations in general.

SPP experiments

For all experiments, the latest version of the WRF-based HRRR system was used. The Contiguous U.S. (CONUS) domain contained 1536 x 1024 grid points and 50 vertical levels with a model top of 20 hPa. The numerics of WRF are well documented in other publications and, thus, not described herein.

The forecasts were conducted as cold starts, producing 36-hour forecasts in hourly increments. Tests were initialized every-other-day during the 2017 warm season from 2 May - 31 July 2017 at 00 UTC. Also, since the microphysics scheme is used year-round and needs to perform equally well in the cool season, cases initialized daily at 00 UTC from 1-31 December 2016 were also run to ensure the proposed ensemble system does not deteriorate forecasts in the cool season when deep convection is essentially absent from the CONUS.

In order to assess whether including SPP within microphysics improves the ensemble forecast spread, an eight-member ensemble was run. The first member of the subset was identical to the deterministic HRRR setup with no stochastic technique applied (considered the "control" member), while the remaining seven members each employed SPP within the Thompson microphysics scheme for the three aspects of interest described above.

A second test for future comparison is also being run that utilizes the SPP technique within the PBL parameterization scheme (for parameters including turbulent mixing length, sub-grid cloud fraction, thermal and moisture roughness lengths, and Prandtl number) in addition to the microphysics scheme perturbations to assess the cumulative impact on ensemble forecast spread.

Observations

RAP observation files in BUFR format, which include conventional surface and upper-air data, were used for verification of point-based fields, including temperature, dew point temperature, and wind speed. Bilinear interpolation was applied to match the point observations with the gridded model output. The Multi-Radar/Multi-Sensor (MRMS) dataset was used as the observational analysis product for the precipitation accumulation and composite reflectivity gridded comparisons. The MRMS data was regridded to the model integration domain to allow for grid-to-grid comparisons. Budget interpolation was used for the quantitative precipitation estimate (QPE) field, while the nearest neighbor approach was used for the composite reflectivity.

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 used. Metrics applied for this study included traditional methods commonly used in the community for both deterministic and probabilistic forecast performance assessment. More information on all of the metrics used for this work can be found in the MET Users' Guide (2017) and Wilks (2011).

Deterministic verification metrics

Standard verification statistics were computed for surface (including precipitation) and upper-air variables over the 3-km HRRR domain and were aggregated over the CONUS, CONUS-East, and CONUS-West verification domains. Focus for this report will mainly be placed on the CONUS-East domain and the summer season runs for brevity. Metrics calculated for both surface and upper-air temperature, dew point temperature, and wind speed consisted of mean error, or the measure of overall bias for continuous variables, and (bias corrected) root-mean-square error ((BC)RMSE). For upper-air results, verification statistics were computed for times valid at 00 and 12 UTC for the mandatory levels and surface verification results were computed for the full 36-h forecast length in 1-h increments. When evaluating precipitation and composite reflectivity, 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). Due to conducting the experiment with cold starts, for precipitation and composite reflectivity analysis, the first 6 hours of the forecast are discarded.

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. A subset of these results will be described in detail for this report.

Verification results

Surface verification

Traditional

Figure 3 displays the summary median bias for surface temperature, dew point temperature, and wind speed across the summer forecast period (May 2 - July 31 2017). Overall, there are relatively small bias values and minimal ensemble spread at all forecast hours for all three surface variables. In all cases, the bias is positive - typically less than 0.6 °C for temperature, 1.2 °C for dew point temperature, and 1 ms^{-1} for wind speed. Ensemble spread is negligible in the first 12 forecast hours; however, ensemble spread increases in the latter half of the forecast, with the largest spread (though still relatively small) occurring over the longer forecast lead times (e.g. forecast hours 30 and greater). Ensemble spread increasing with forecast length is expected as the effects of the stochastic perturbations on the environment are more apparent with time.

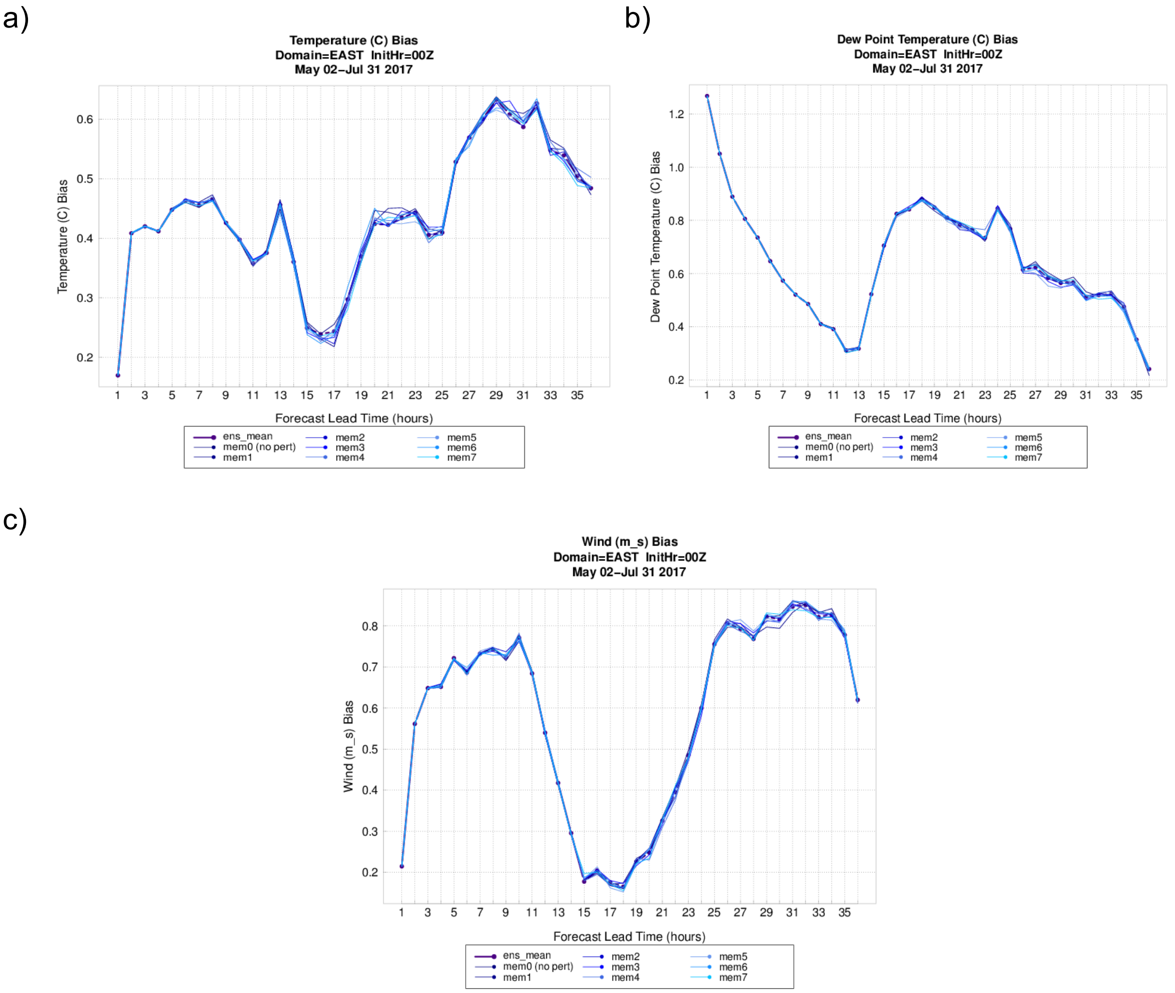


Figure 3. *Bias by forecast lead time for a) 2-m temperature, b) 2-m dew point temperature, and c) 10-m wind speed for 0000 UTC initializations over the summer cases aggregated across the CONUS-East verification domain.*

Figure 4 illustrates the summary median RMSE for the three surface variables. Overall, the behavior of this metric is similar to bias: minimal ensemble spread that increases with forecast lead time. Additionally, the values of RMSE are also small, on the order of 1 - 2 degrees/ms^{-1}. Unlike temperature and dew point temperature, the individual ensemble members for 10-m wind speed are clustered together with higher error than the ensemble mean (Fig 4c). The BCRMSE for the surface parameters exhibits the same trends as in the RMSE (not shown).

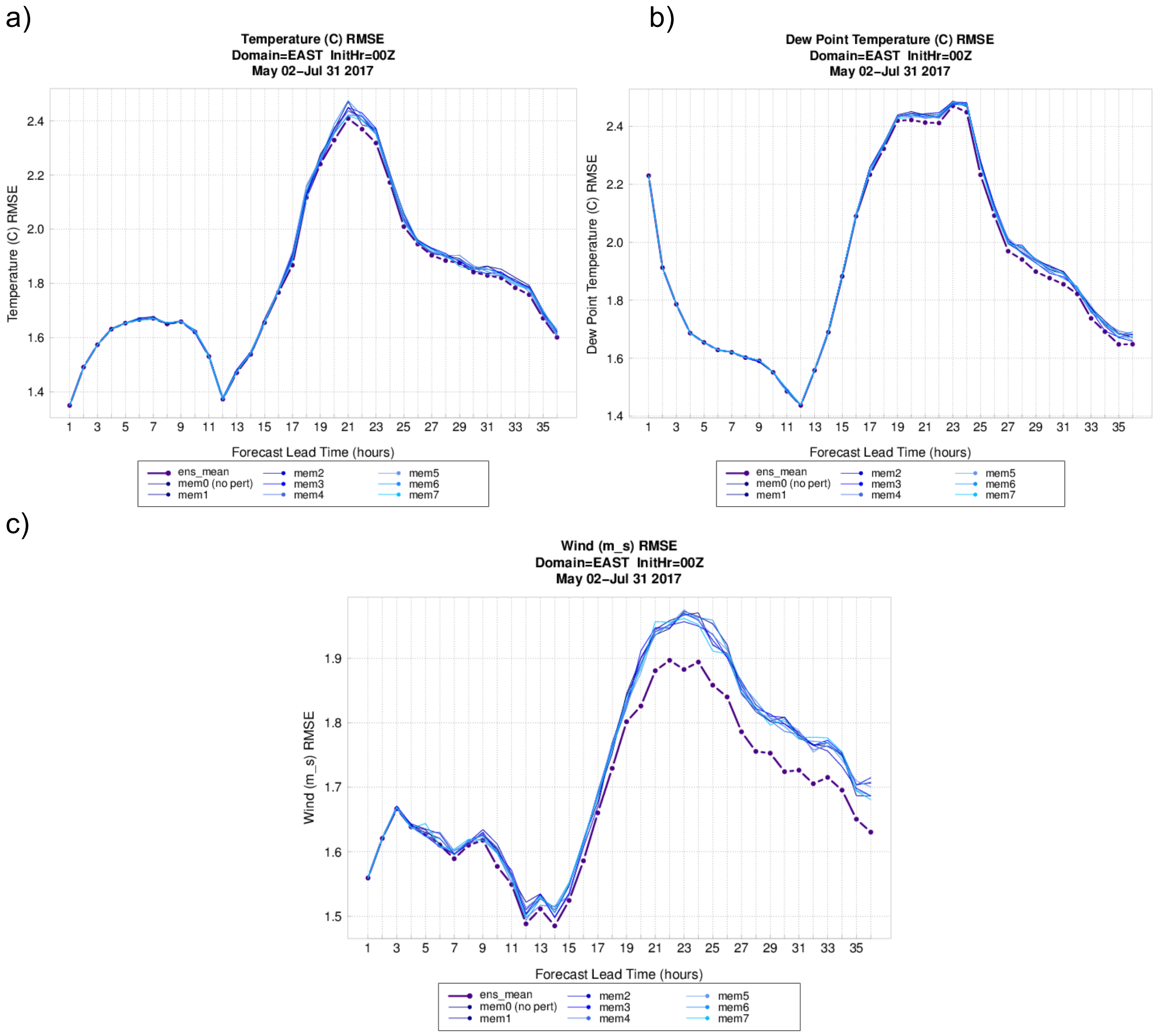


Figure 4. *RMSE by forecast lead time for a) 2-m temperature, b) 2-m dew point temperature, and c) 10-m wind speed for 0000 UTC initializations over the summer cases aggregated across the CONUS-East verification domain.*

Ensemble

Surface verification revealed very little spread compared to the measured RMSE during the three month summer period for the variables analyzed over the CONUS-East domain. Spread/skill plots for 2-m temperature (Fig. 5a), dew point temperature (Fig. 5b), and 10-m wind speed (Fig. 5c) indicate that, while spread increases slowly with forecast lead time, the ensemble lacks sufficient spread to account for the amount of error. Ideally, the spread/skill ratio would be equal to one; it is significantly lower than one in all cases due to the insufficient spread. This result was somewhat expected, given that the sole source of ensemble spread is coming from the microphysics perturbations and, therefore, only has an indirect impact on the these particular surface variables.

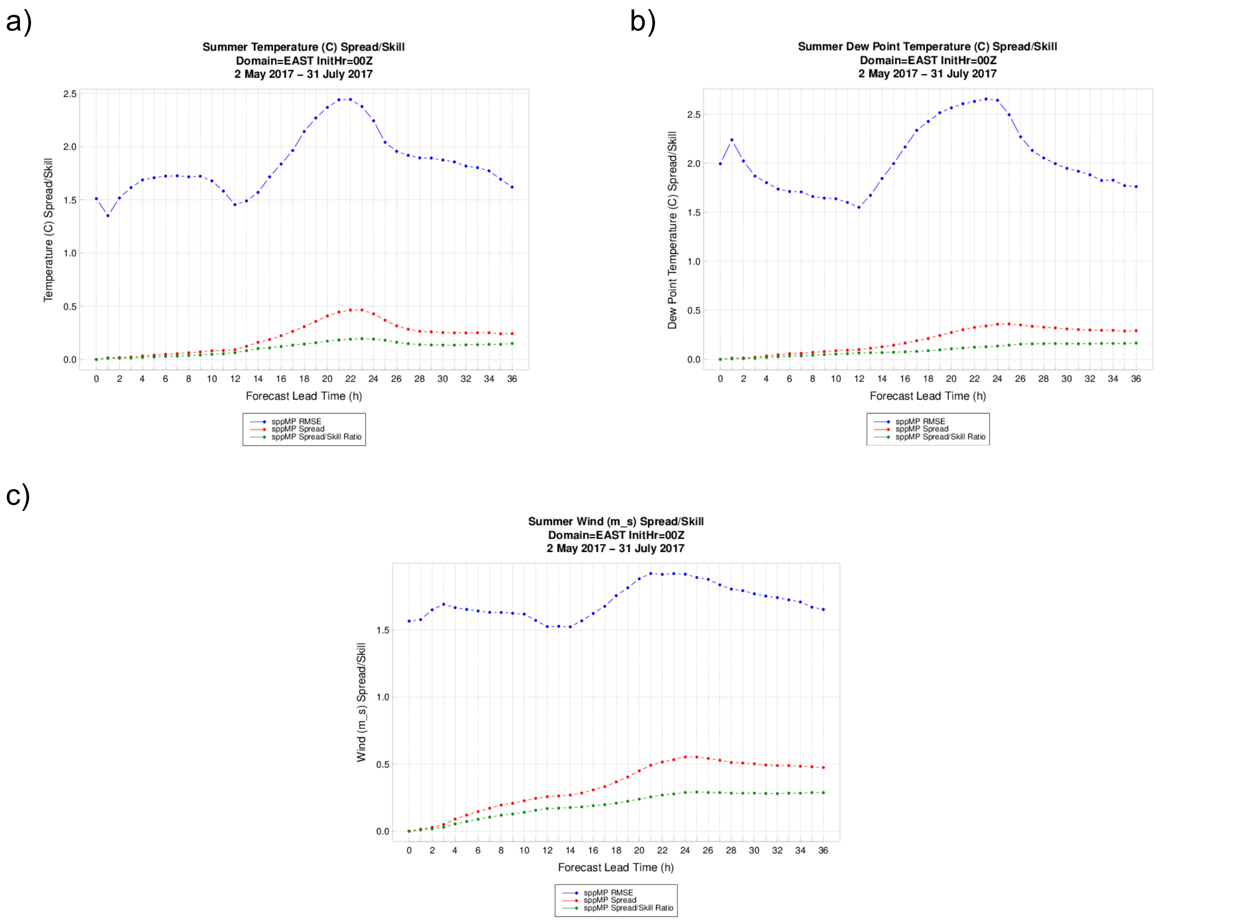


Figure 5. *Spread/skill plots by forecast lead time for a) 2-m temperature, b) 2-m dew point temperature, and c) 10-m wind speed for 0000 UTC initializations over the summer cases aggregated across the CONUS-East verification domain. Skill (RMSE) is designated by the blue line, spread by the red line, and spread/skill ratio by the green line.*

Rank histograms for the same surface variables over the CONUS-East domain are shown in Figure 6, averaged over the first 24 hours of the forecast. 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. 6a), 2-m dew point temperature (Fig. 6b), and 10-m wind speed (Fig. 6c). There is also a slight hint of a high bias for each surface variable, as more observations fall in the first bin, rather than the last.

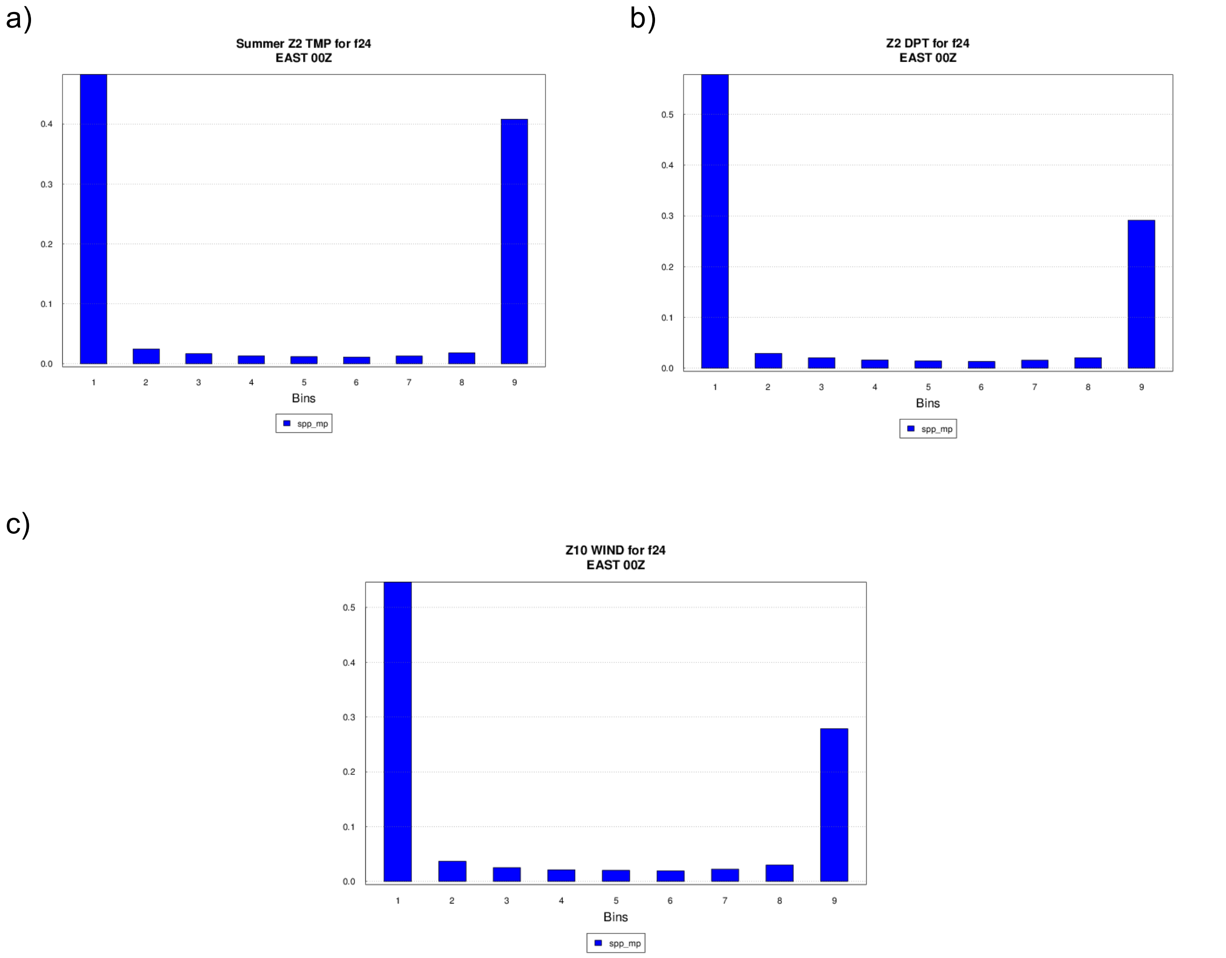


Figure 6. *Rank histogram plots for a) 2-m temperature, b) 2-m dew point temperature, and c) 10-m wind speed averaged over the first 24 hours of the forecast.*

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. Given these parameters, the area where the probabilistic forecast is skillful is indicated by the green shading in Figure 7. To illustrate the performance for 2-m temperature, 2-m dew point temperature and 10-m wind speed, thresholds that yielded a sample climatology of approximately 60% were chosen leading to a temperature threshold of ≥293 K, dew point temperature threshold of ≥288 K, and wind speed threshold of ≥2 ms^{-1}. In all cases the reliability curve has a positive slope, indicating that, to at least a certain degree, as the forecast probability increase, so too does the observed frequency; however, the slope is less than the diagonal. In general, for the lower forecast probabilities the ensemble tends to under-forecast the event probability transitioning to over-forecasting the event probability. This is a common feature of under-dispersive ensembles. When assessing thresholds with lower sample base rates (e.g., higher temperature, dew point temperature, and wind speed values; not shown), the forecasts tend to over-forecast observed event frequencies, with a lack of skill evident at the higher probabilities.

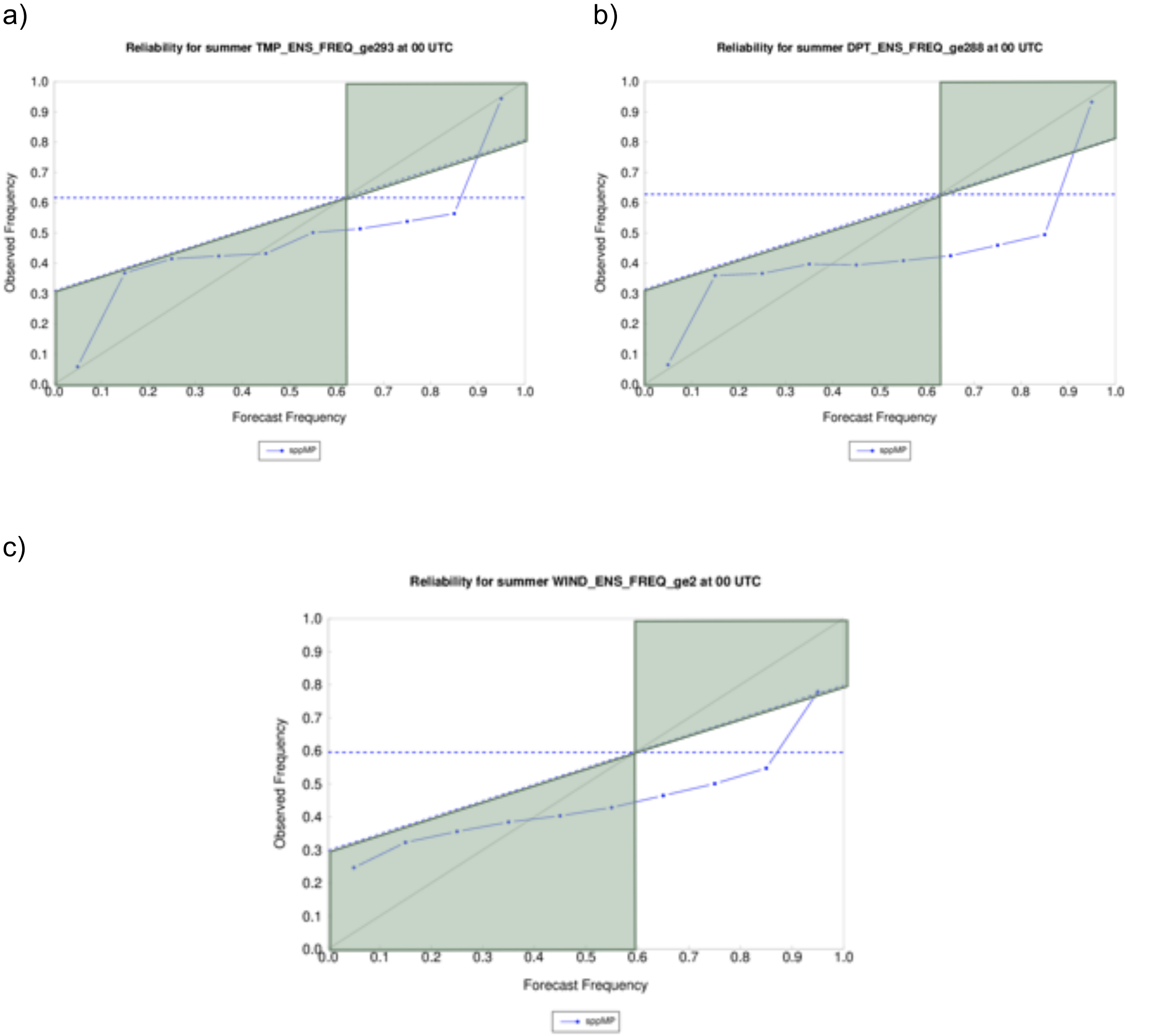


Figure 7. *Reliability diagrams for 0000 UTC initializations over the summer cases aggregated across the CONUS-East verification domain for a) surface temperature at a threshold of ≥293 K, b) surface dew point temperature at a threshold of ≥288 K, and c) surface wind speed at a threshold of ≥2 ms^{-1}. The horizontal dotted line represents no resolution, the diagonal dotted line represents no skill, the solid grey diagonal line represents perfect reliability, and the green shaded areas indicate skillful forecasts.*

Upper-air verification

Traditional

Figure 8 illustrates the vertical profile of mean error at the 24 hour forecast lead time. As exhibited by their surface counterparts, the biases for temperature and wind speed are relatively small (less than 1 degree/ms^{-1}). On the other hand, dew point temperature exhibits positive biases increasing with height to 3 °C at 500 hPa. Ensemble spread is relatively small near the surface between 925 - 850 hPa and relatively larger in the mid- to upper-levels. Temperature and wind speed have a general cool/slow bias that increases with height while the opposite is true for dew point temperature.

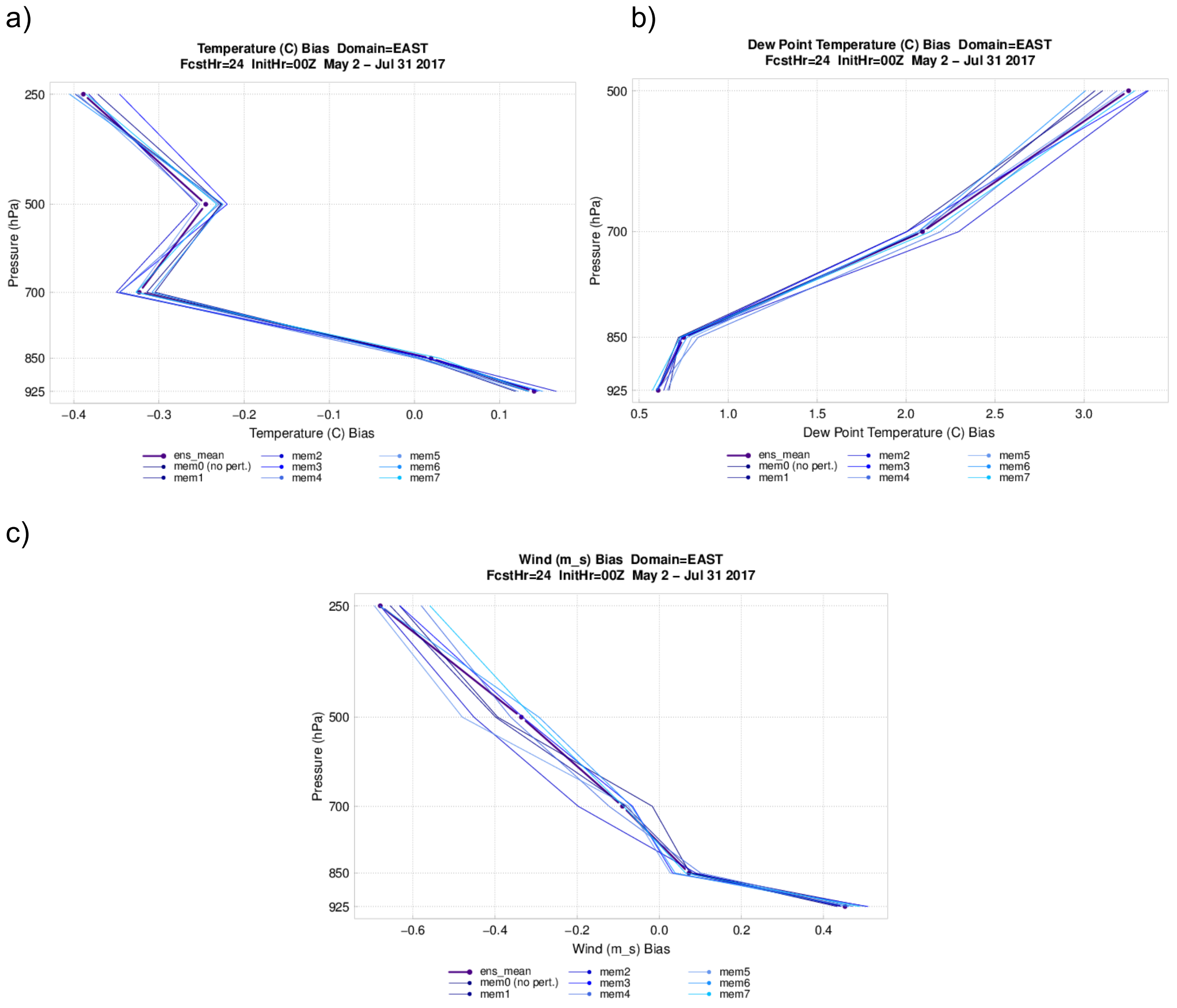


Figure 8. *Vertical profile of bias at forecast hour 24 for a) temperature, b) dew point temperature, and c) wind speed for the 0000 UTC initializations over the summer cases aggregated across the CONUS-East verification domain.*

Figure 9 displays the RMSE as a function of pressure level for each variable at the 24 hour forecast lead time. Temperature and dew point temperature exhibit very minimal ensemble spread from 925 - 850 hPa and up to 700 hPa for dew point (Fig. 9b). Wind speed exhibited the most spread among ensemble members, especially in the lower- to mid-levels (Fig. 9c). Vertical profiles of BCRMSE display very similar trends to RMSE for the three variables discussed (not shown). The value of RMSE increase with forecast lead time for all variables. The ensemble spread does not always increase, however, unlike the surface counterparts. The RMSE for temperature and dew point temperature reach peak values at forecast hour 24 (Fig. 9a-b). RMSE for wind speed reaches its maximum value at forecast hour 36 (not shown).

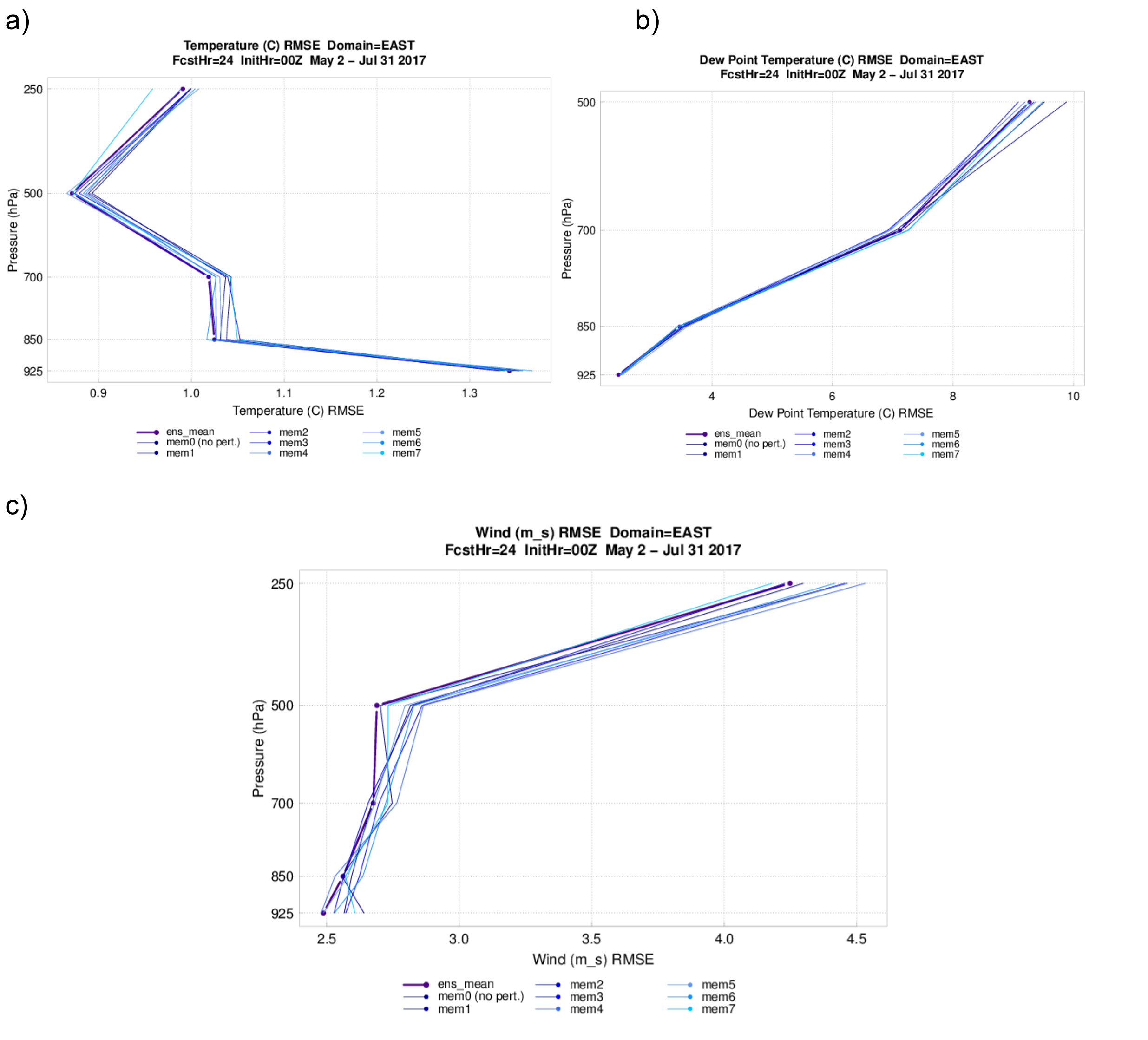


Figure 9. *Vertical profile of RMSE at forecast hour 24 for a) temperature, b) dew point temperature, and c) wind speed for the 0000 UTC initializations over the summer cases aggregated across the CONUS-East verification domain.*

Ensemble

Upper-air spread/skill plots for 500 hPa temperature, dew point temperature, and wind speed are shown in Figure 10. Insufficient spread aloft results in very low spread/skill ratios for these variables, although spread does increase with forecast lead time. Overall, it is clear that the microphysics parameter perturbations alone are not sufficient to produce the necessary spread in upper-air variables, following along the lines of spread/skill plots for surface variables (Fig. 5).

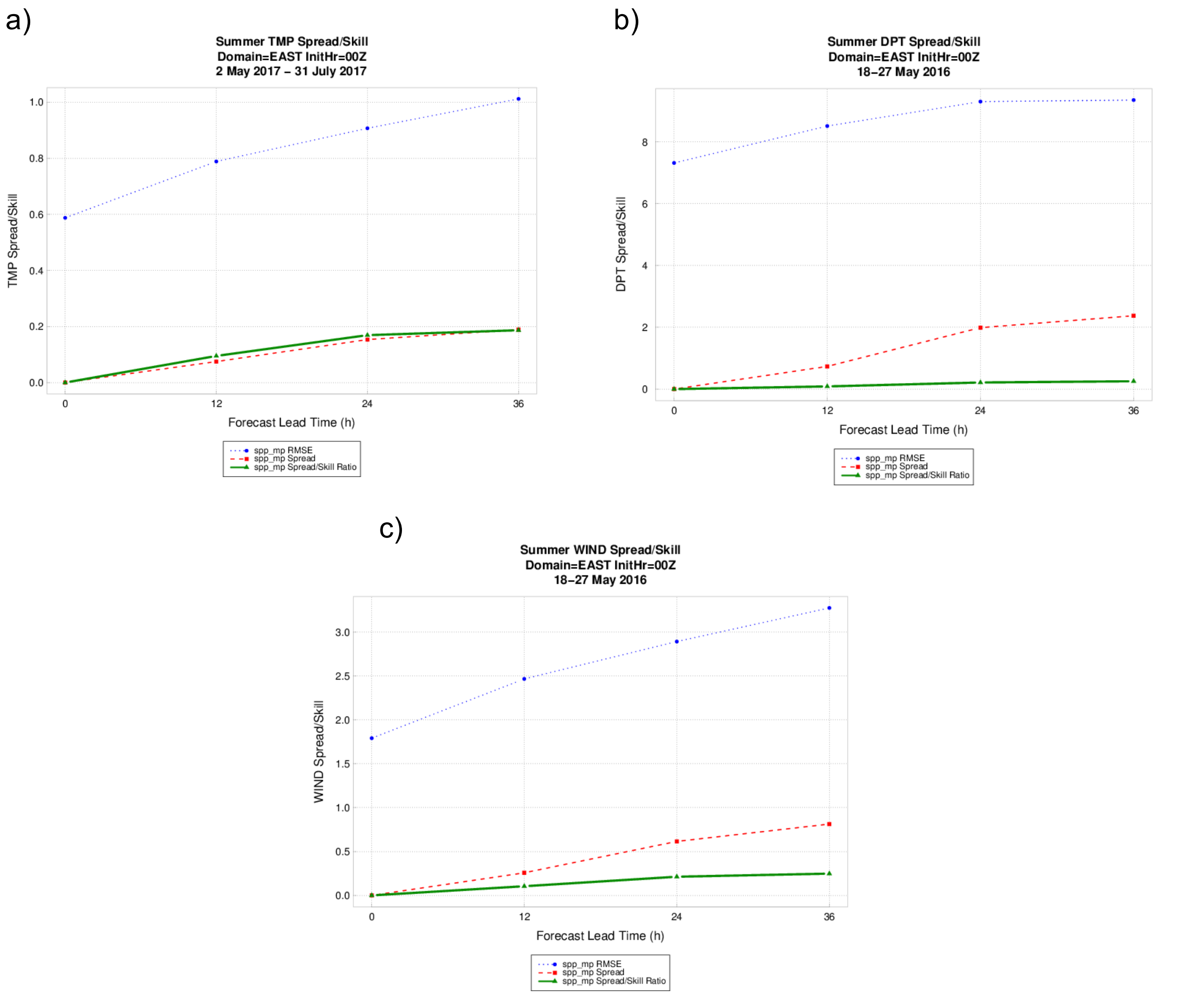


Figure 10. *Upper-air (500 hPa) spread/skill plots by forecast lead time for a) temperature, b) dew point temperature, and c) wind speed for 0000 UTC initializations over the summer cases aggregated across the CONUS-East verification domain. Skill (RMSE) is designated by the blue line, spread by the red line, and spread/skill ratio by the green line.*

Given the lack of spread aloft, the rank histograms for 500 hPa temperature (Fig. 11a), dew point temperature (Fig. 11b), and wind speed (Fig. 11c) also illustrate the under-dispersive nature of the ensemble for these variables. The same U-shape as was seen for the surface variable rank histograms (Fig. 6) is noted in these plots. There is again some indication of bias for these variables, with 500-hPa temperature (Fig. 11a) and wind speed (Fig. 11c) values showing a tendency to be too low with more observations falling in the last bin than in the first. The rank histogram for 500-hPa dew point temperature (Fig. 11b) values indicate an opposite trend, with more observations falling in the lowest bin indicating the forecast values are too high.

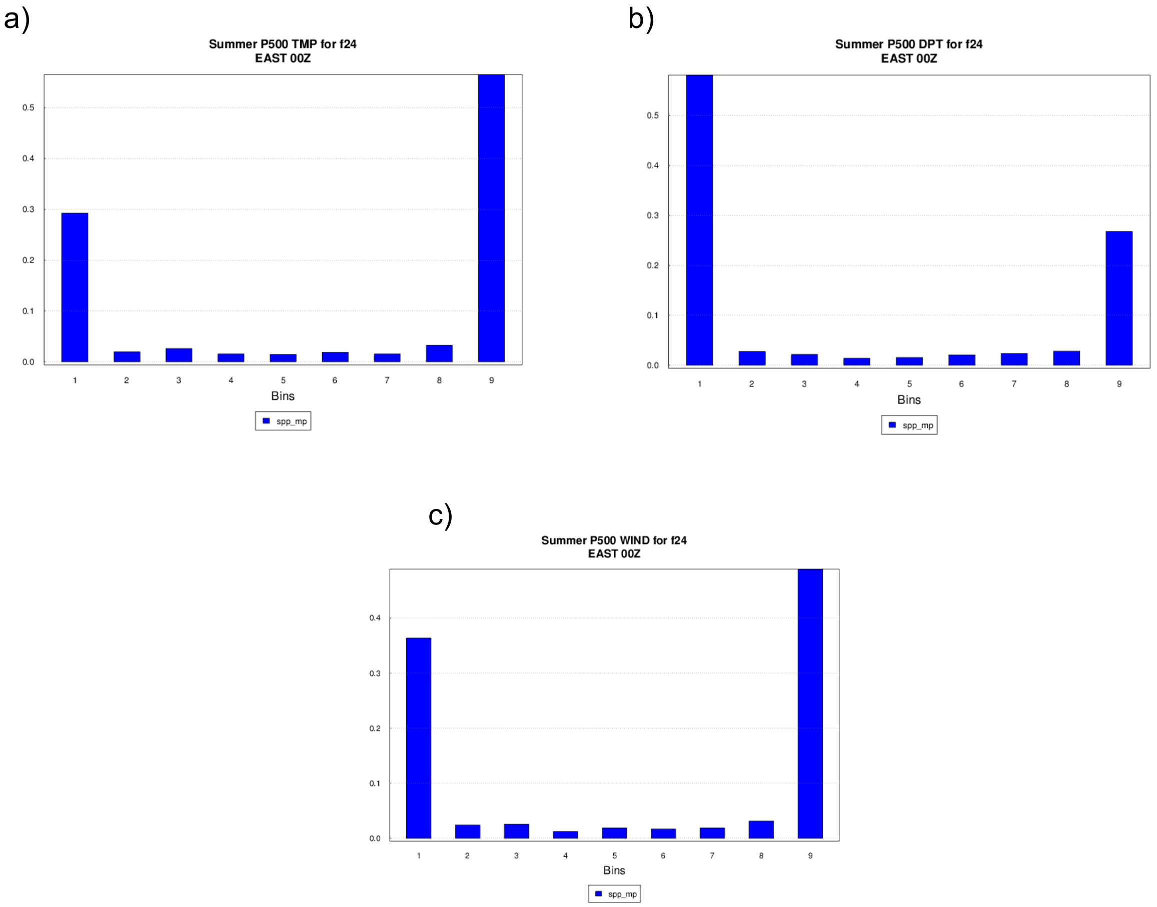


Figure 11. *Rank histogram plots for 500 hPa a) temperature, b) dew point temperature, and c) and wind speed averaged over the first 24 hours of the forecast.*

Reliability diagrams for upper-air temperature, dew point temperature, and wind speed at a variety of thresholds and pressure levels were examined. In most cases the vast majority of data were in the 0 and 0.9 bins (lowest and highest, respectively), both of which had very reliable forecasts; however, there were an insufficient number of cases in the middle bins, so they are excluded from this analysis.

Precipitation and composite reflectivity verification

Traditional

3-h Accumulated Precipitation

Figure 12a-e shows GSS for 3-h accumulated precipitation at 5 thresholds (≥0.254 mm, ≥2.54 mm, ≥6.35 mm, ≥12.7 mm, and ≥25.4 mm) aggregated over the summer cases. When interpreting precipitation verification, it is necessary to consider the base rate (i.e., ratio of total observed grid-box events to the total number of grid boxes summed over all cases; black line on y-2 axis in Fig. 12) to better understand the underlying sample and how it affects the statistical results. The base rate is highest at the lowest thresholds and from 21 – 00 UTC and decreases to near zero at the higher thresholds. Overall, skill for the individual ensemble members and the ensemble mean is generally highest at the smallest thresholds and at the earlier forecast lead times, with a gentle decrease in skill by the end of the forecast integration period. For all thresholds except ≥25.4 mm, a bimodal diurnal signal is observed with peak values at 06 – 09 UTC, when the base rate is lowest, along with a smaller secondary peak at 21 UTC. All members are generally clustered together showing minimal spread; however, a slight increase in spread is observed as forecast lead time increases and as precipitation threshold increases. The growth in spread with forecast lead time is likely attributed to the cumulative effects from the stochastic perturbations throughout the model integration (e.g., convection during day 1 may have effects on boundary locations up during day 2). In general, the aggregate ensemble mean has more skill than the envelope of individual members, but the gain in skill from the ensemble mean decreases as threshold increases. The increase in skill from the pure arithmetic ensemble mean is likely due to the individual forecasts being averaged together resulting in a smoother forecast with more spatial coverage; these types of coarse fields generally verify better than high-resolution fields when using traditional verification metrics. While not shown, results for the aggregation of winter cases share similar outcomes as the summer season. Generally, there is small spread among the members, which increases as forecast lead time and threshold increases. One noted difference in the winter season is at the lower thresholds, where the ensemble mean is closer to the individual members than in the summer cases. This may be attributed to synoptic-scale features more often seen in winter precipitation, which are typically smoother and larger in spatial coverage than more discrete mesoscale features commonly seen in the summer.

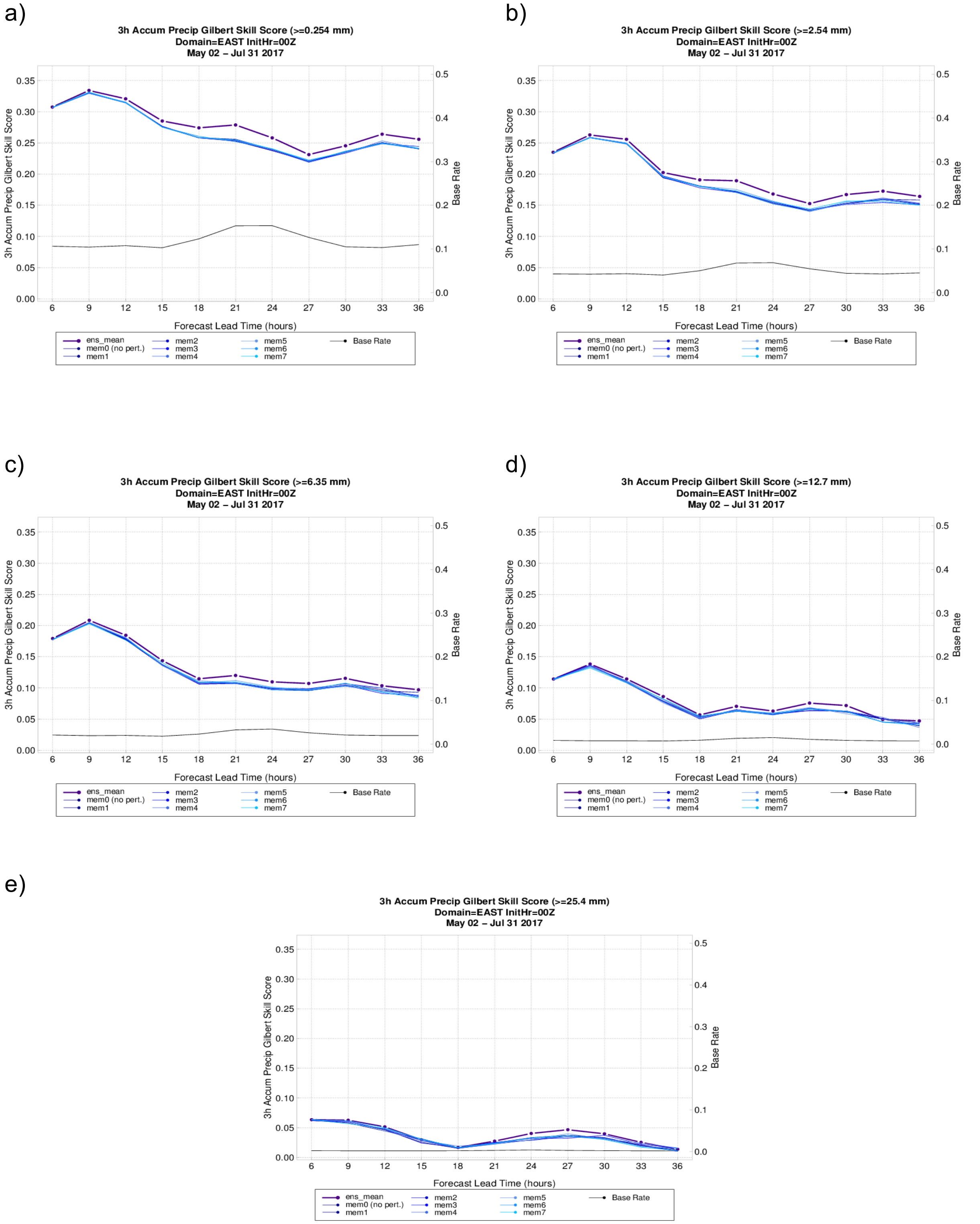
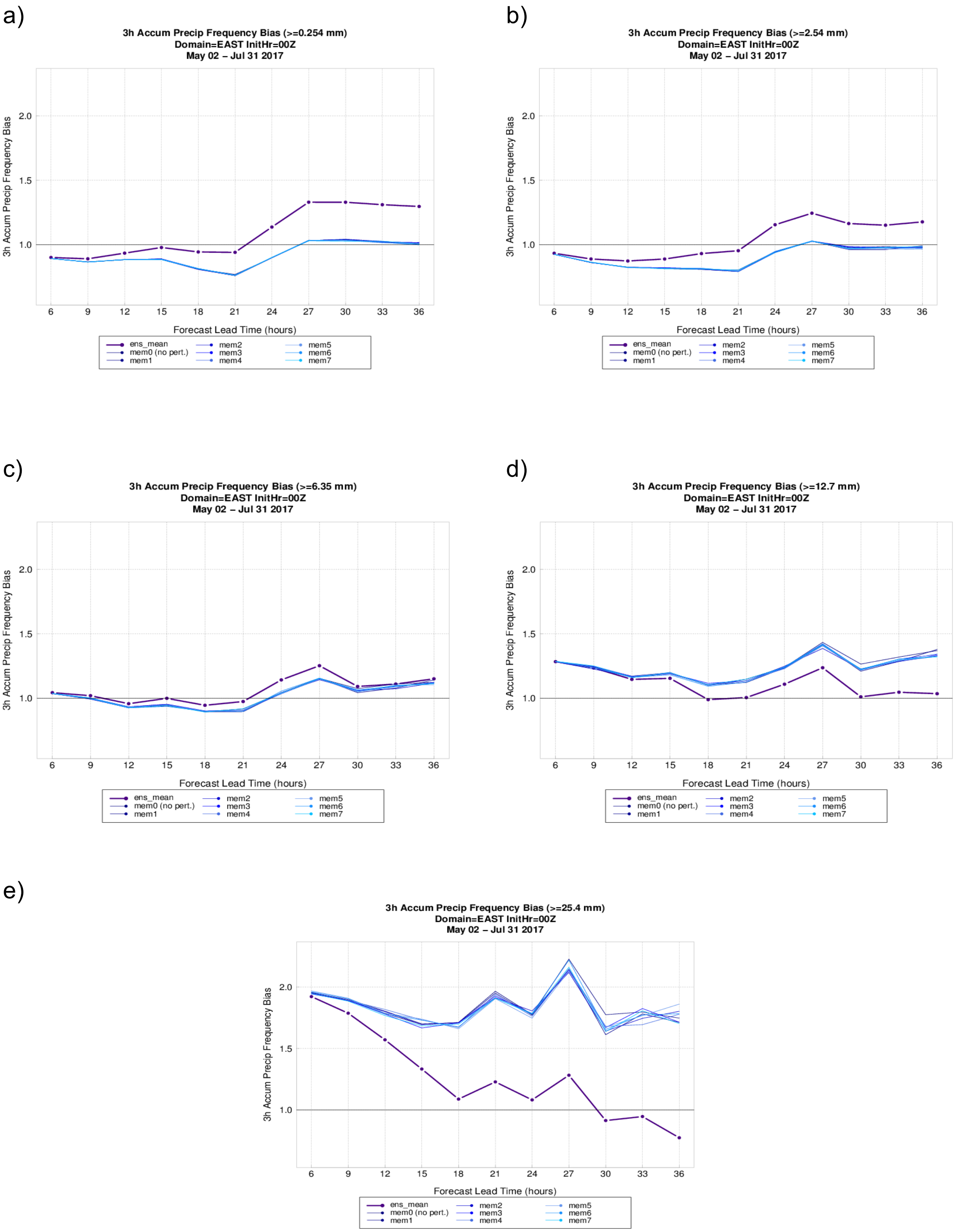


Figure 12. *Gilbert Skill Score by forecast time time for 3-h accumulated precipitation for 0000 UTC initializations over the summer cases aggregated across the CONUS-East verification domain for a) ≥0.254 mm, b) ≥2.54 mm, c) ≥6.35 mm, d) ≥12.7 mm, and e) ≥25.4 mm. The black line on the y-2 axis is the base rate.*

When evaluating 3-h accumulated precipitation frequency bias over the summer cases, forecasts generally perform well with most thresholds having bias values close to 1 (i.e., unbiased; Fig. 13a-e). For the individual members, there is a small under-forecast at the early-to-middle forecast lead times for the lower thresholds; this transitions to an over-forecast at the highest thresholds. A slight diurnal signal is also noted for the ≥0.254 to ≥12.7 mm thresholds, with the smallest bias values around forecast hours 18 – 21 and a small peak in high bias at the 27-h forecast (valid at 03 UTC). At the lowest thresholds, the aggregate ensemble mean typically has higher values than the individual members; as threshold increases, this pivots to the ensemble mean having lower bias than the ensemble members. In general, this leads to lower bias for the ensemble mean, likely due to the coarser field. Similar to GSS, there is minimal spread among the individual ensemble members for all evaluated thresholds; spread does slightly increase throughout the model integration period and at higher thresholds. The winter season (not shown) also displays minimal spread, but the individual members as well as ensemble mean generally have frequency bias values near 1, with exception to the ≥25.4 threshold.



a) 3h Accum Precip Frequency Bias (>=0.254 mm) Domain=EAST InitHr=00Z May 02 – Jul 31 2017

b) 3h Accum Precip Frequency Bias (>=2.54 mm) Domain=EAST InitHr=00Z May 02 – Jul 31 2017

c) 3h Accum Precip Frequency Bias (>=6.35 mm) Domain=EAST InitHr=00Z May 02 – Jul 31 2017

d) 3h Accum Precip Frequency Bias (>=12.7 mm) Domain=EAST InitHr=00Z May 02 – Jul 31 2017

e) 3h Accum Precip Frequency Bias (>=25.4 mm) Domain=EAST InitHr=00Z May 02 – Jul 31 2017

Figure 13. *Frequency bias by forecast time time for 3-h accumulated precipitation for 0000 UTC initializations over the summer cases aggregated across the CONUS-East verification domain for a) ≥0.254 mm, b) ≥2.54 mm, c) ≥6.35 mm, d) ≥12.7 mm, and e) ≥25.4 mm.*

Composite Reflectivity

GSS for composite reflectivity was evaluated for three thresholds (≥20 dBZ, ≥30 dBZ, and ≥40 dBZ; Fig. 14a-c) over the summer season. The ensemble mean for reflectivity was withheld, as statistically speaking, it is questionable to take a pure arithmetic mean of instantaneous and potentially discrete fields. As threshold increases, skill as well as the base rate decreases. At all thresholds, skill is highest at the earliest lead times, with an overall decrease in GSS as forecast lead time increases. Diurnal trends are also observed, with GSS decreasing throughout the daytime into early evening period, where a broad minimum in skill coincides with the highest base rate values. Skill then increases overnight, perhaps due to discrete convection evolving into larger, more organized MCSs; it is worth noting that peak in skill shifts toward earlier times as the threshold increases. As seen with accumulated precipitation, the ensemble has minimal overall spread, but a small increase in spread is observed at the longer lead times as the effects of the stochastic perturbations grow through the model integration. The winter season (not shown) has a general shift toward higher skill at the ≥20 dBZ and ≥30 dBZ thresholds. Minimal spread is also noted, but compared to the summer season, there is a slight increase in spread.

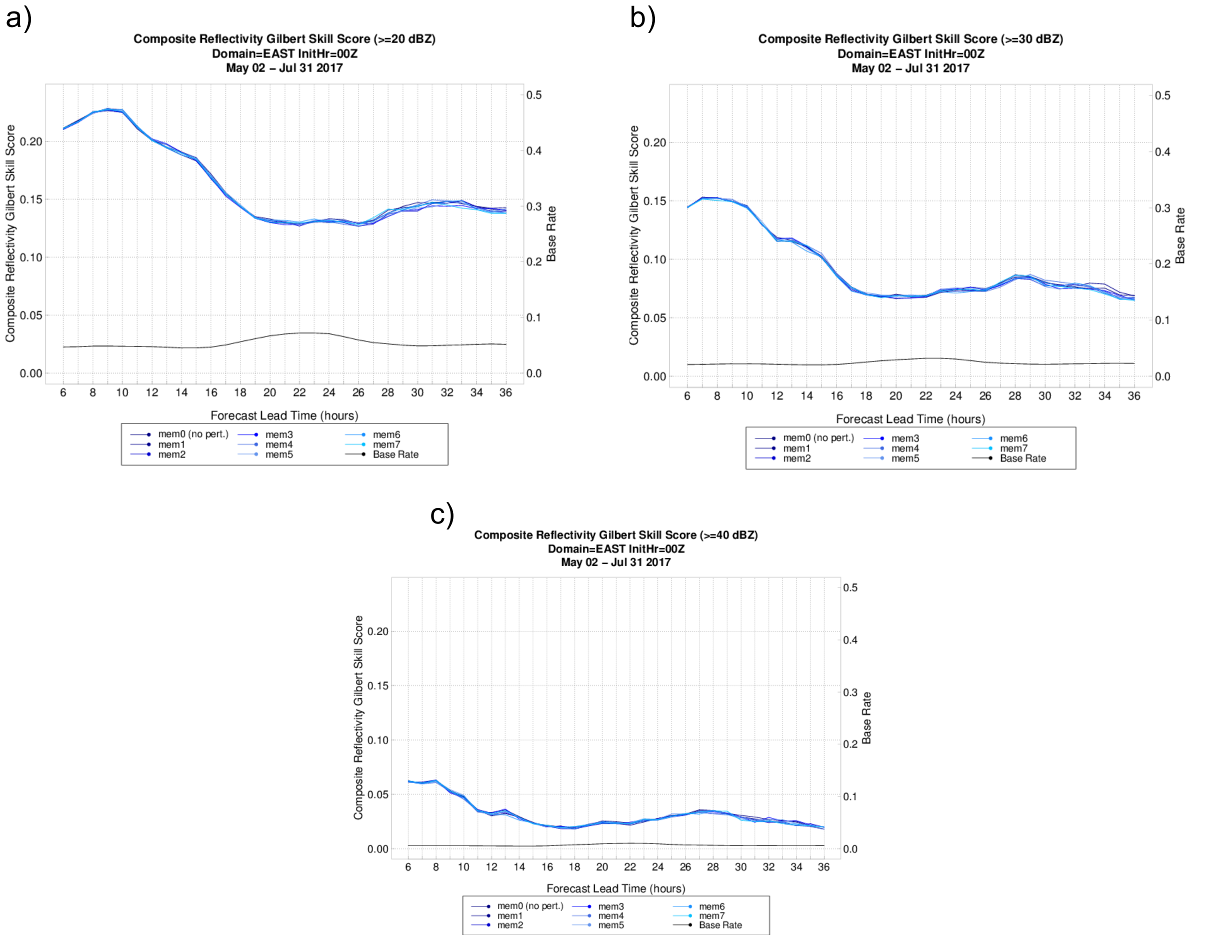


Figure 14. *Gilbert Skill Score by forecast time time for composite reflectivity for 0000 UTC initializations over the summer cases aggregated across the CONUS-East verification domain for a) ≥20 dBZ, b) ≥30 dBZ, c) ≥40 dBZ. The black line on the y-2 axis is the base rate.*

At all evaluated composite reflectivity thresholds, forecasts perform quite well when considering frequency bias (Fig. 15a-c). At the ≥20 dBZ threshold, all forecast lead times have a slight low bias (i.e., under-forecast). As threshold increases, there is a shift toward higher frequency bias values, which is more amplified at longer forecast lead times, where at the ≥40 dBZ threshold, all forecast lead times at and beyond the 21-h forecast transition to having a small high bias. Diurnal variations in bias values are noted at all thresholds; a bimodal distribution of frequency bias is observed with minimum values late in the morning and early afternoon period (17 - 20 UTC, depending on the threshold) and a smaller minimum in the early morning (9 - 12 UTC, depending on the threshold). The timing of when the minimum in frequency bias transitions to higher values (17 - 20 UTC, depending on the threshold) coincides with the broad area of low GSS and high base rate values, potentially signalling an issue with the timing and or location of convective initiation. Maximum frequency bias values peak around 03 - 06 UTC at the ≥20 dBZ threshold and around 01 - 02 UTC at the ≥30 and ≥40 dBZ thresholds. The timing shifts in peak bias values for each threshold may provide insight on how the model evolves the convection in terms of spatial coverage and intensity and could be further explored using spatial verification techniques. When considering the ensemble membership, there is minimal overall spread, with any increase in spread seen as forecast lead time and threshold increase. The winter season (not shown) has a general shift toward higher biases at all thresholds. Minimal spread is also noted, but compared to the summer season, there is a slight increase in spread.

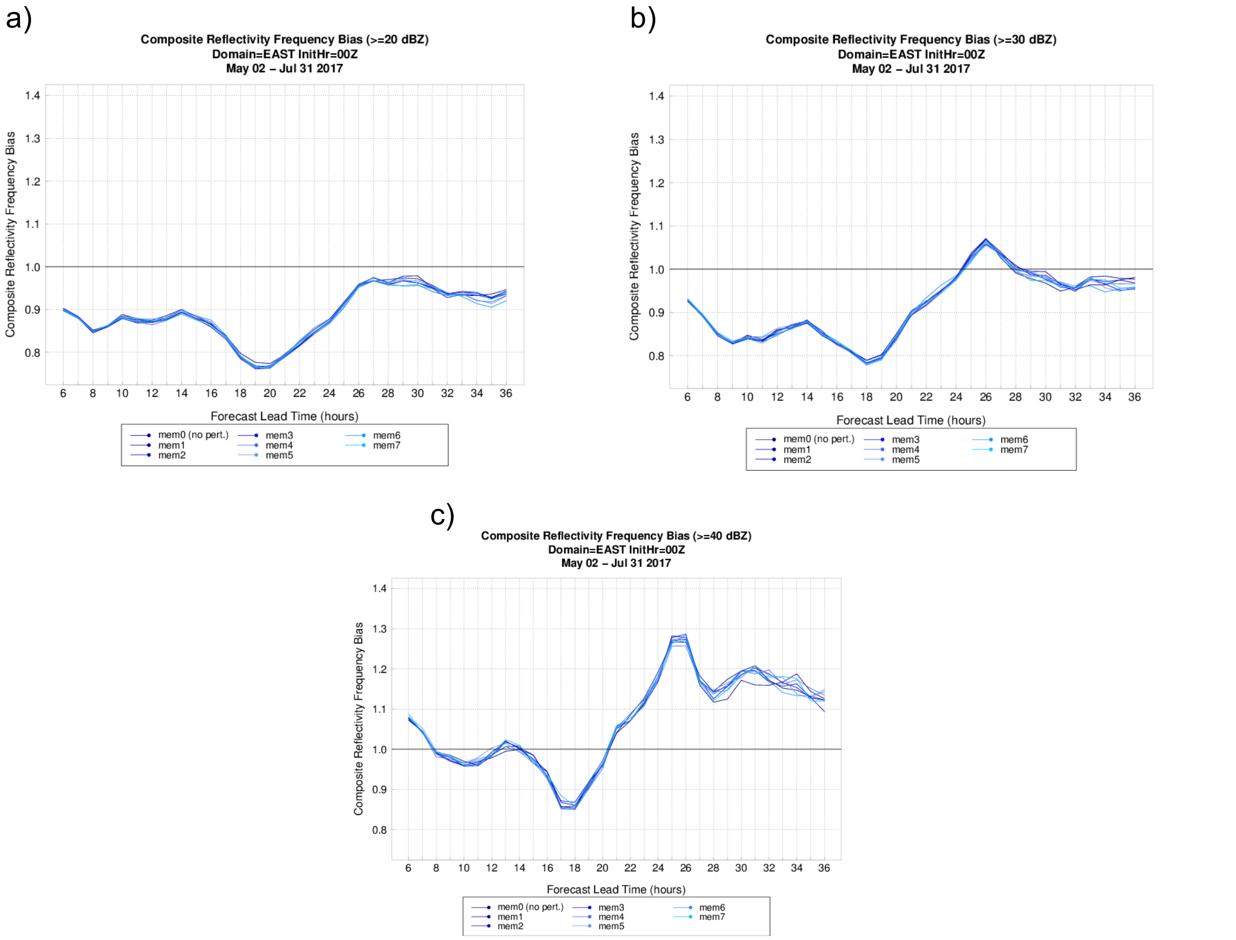


Figure 15. *Frequency bias by forecast time time for composite reflectivity for 0000 UTC initializations over the summer cases aggregated across the CONUS-East verification domain for a) ≥20 dBZ, b) ≥30 dBZ, c) ≥40 dBZ.*

Ensemble

3-h Accumulated Precipitation

While the traditional surface variable spread/skill plots showed a general lack of spread, the three-hour precipitation accumulation spread/skill plot shows a substantial impact from the Thompson MP perturbations (Fig. 16a). Spread quickly increases with lead time, and by 24 hours into the forecast, reaches a point where the spread/skill ratio is around 0.8, indicating that most of the RMSE is now being accounted for through the spread produced by the MP perturbations.

Although the spread/skill ratio for three-hour precipitation accumulation has improved by the 24-hour forecast, compared to the initialization, the rank histogram (Fig. 16b) still shows that the ensemble is under-dispersive, indicated by the classic U-shape of the plot. However, the additional spread in the middle bins is evident when the three-hour precipitation accumulation rank histogram is compared that of the 2-m temperature (Fig. 6a), dew point (Fig. 6b), and wind speed (Fig. 6c) rank histograms.

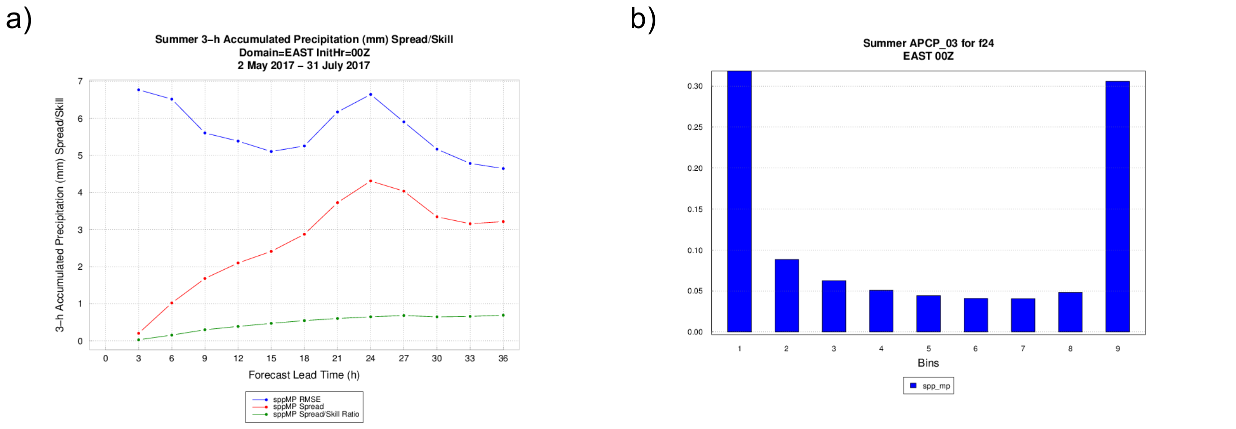


Figure 16. *Spread/skill plot (a) as a function of forecast lead time and rank histogram (b) for three-hour precipitation accumulation. Lines for (a) are the same as in Fig. 5.*

For 3-h accumulated precipitation, the most reliable (i.e. calibrated) forecasts are at the lowest threshold of ≥0.254 mm where the sample climatology is just over 10% (Fig 17a). At the higher forecast probabilities, and especially for the higher thresholds (not shown above 2.54 mm), the events are rare and the ensemble probabilities are poorly calibrated. There is a lack of skill at the highest thresholds examined (≥12.7 and 25.4 mm) with a tendency to over-forecast the probability of precipitation occurrence. As indicated for other variables, the ensemble tendency to under-forecast at the lower probabilities is noted for 24-accumulated precipitation forecasts at thresholds of ≥12.7mm and lower (i.e., ≥2.54 and ≥6.35 mm shown in Figure 18c,d) when the sample climatology is 10% or higher. For several thresholds, the 24-h accumulated forecast precipitation is skillful for the mid- to high-probabilities. At the highest threshold examined (≥2.54 mm, not shown), the forecasts are only skillful at the lowest probabilities. All of these reliability diagrams show that the ensemble distinguishes between precipitation events and non-events in the highest and lowest probability categories. In each case, a small proportion of events are observed when the ensemble probability is low, and a much higher proportion of precipitation events occur when the ensemble probability is high. This is a desirable outcome. However, the graphs are very flat in the middle. This indicates the inability of the ensemble to effectively distinguish the middle probability categories from one another. Specifically, precipitation occurs with almost the same frequency whether the ensemble probability is 40% or 70%. A good deterministic forecast can distinguish between precipitation events and non-events. The goal of the ensemble is to provide information in those central probabilities. This one just barely accomplishes that goal.

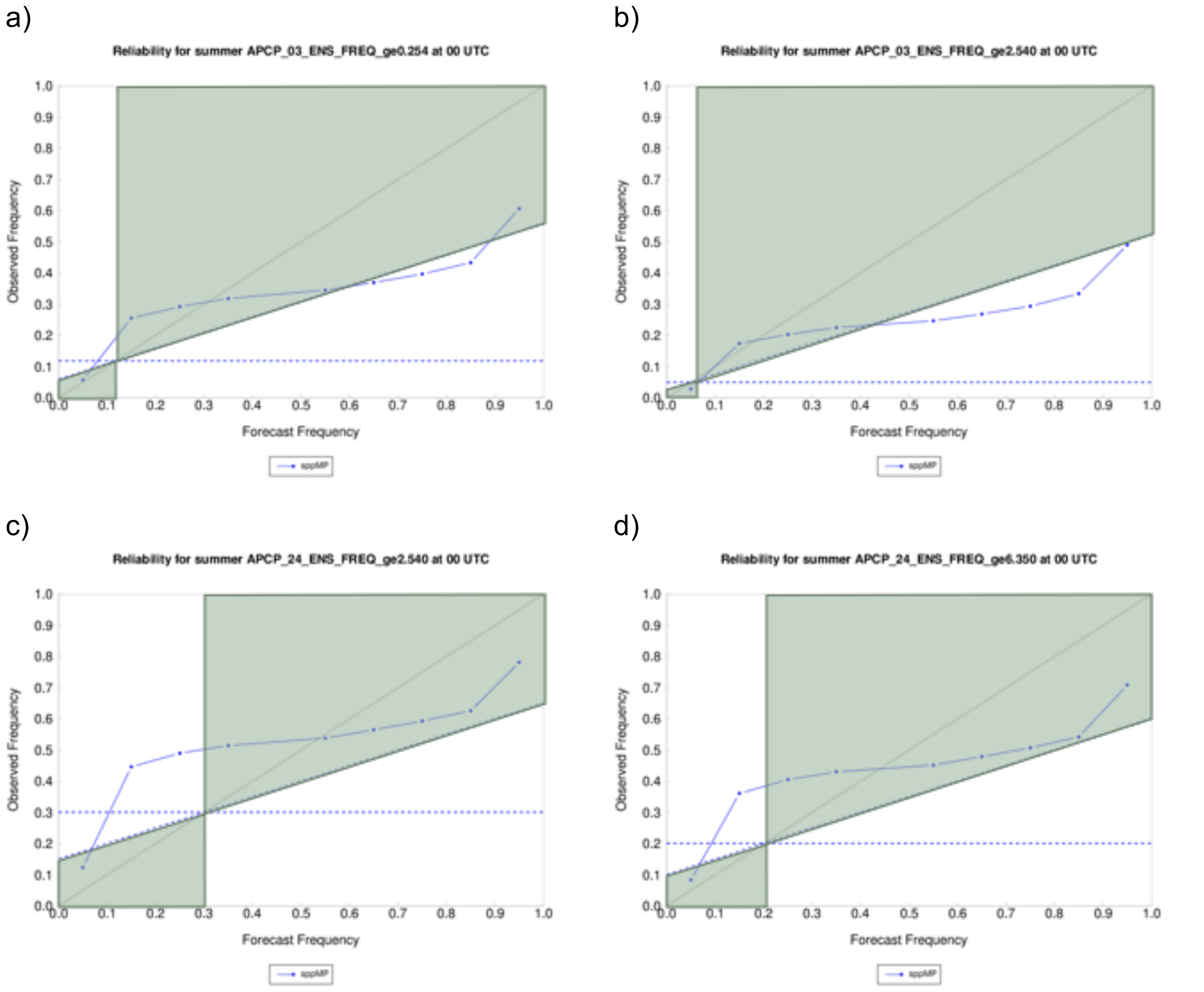


Figure 17. *Reliability diagrams for 0000 UTC initializations over the summer cases aggregated across the CONUS-East verification domain for a) 3-h accumulated precipitation at a threshold of ≥0.254 mm, b) 3-h accumulated precipitation at a threshold of ≥2.54 mm, c) 24-h accumulated precipitation at a threshold of ≥2.54 mm, and d) 24-h accumulated precipitation at a threshold of ≥6.35 mm. The horizontal dotted line represents no resolution, the diagonal dotted line represents no skill, the solid grey diagonal line represents perfect reliability, and the green shaded areas indicate skillful forecasts.*

Comparison of Ensemble Members to Control Member

Comparison of perturbed ensemble members with the control member provides information about the effects of the perturbation. In particular, it is possible to assess how much the precipitation forecasts change due to the perturbations applied in the microphysics scheme. Standard precipitation verification metrics are applied, but the interpretation differs since this is a comparison. In a true verification, we look for errors and hope they are small. Here, we look for differences and hope to find some of reasonable size, indicating that our ensemble has some spread. If the control and perturbed ensemble members are very similar, then the ensemble has little spread. In this case, the ensemble and control members are nearly identical. The greatest differences are seen at the longest lead times and for the higher thresholds of precipitation, but even these are quite small. Differences in perturbed ensemble members are small in both the summer and winter seasons. One example plot comparing the control and ensemble members for each of the seasons is included.

Figure 18 is a performance diagram (Roebber, 2009), which shows the balance of detection and success (1-false alarm) rates. Additionally, it shows the frequency bias in dashed straight lines and curves of constant critical success index (CSI; Wilks, 2011). The red line shows the performance of the control forecast to detect 3-hr precipitation as observed by the MRMS for a variety of thresholds during the winter season over the eastern CONUS. The 18h lead time is used for this example, though other lead times have similar results. Overall, the frequency bias for each threshold is near one, which is very good. Skill, measured by CSI, ranges from above 0.4 to near 0.1, decreasing with precipitation threshold. The perturbed ensemble members are nearly identical to each other and the control member for the lower thresholds, as indicated by the layering of points atop each other near the upper right of the diagram. Again, here we are not measuring skill, but agreement. These agree quite well. At higher thresholds, they match less well to the ensemble control and exhibit more spread. However, they are still very close to each other. Similar plots for the summer season results (not shown) exhibit even smaller differences in ensemble members, even at higher thresholds.

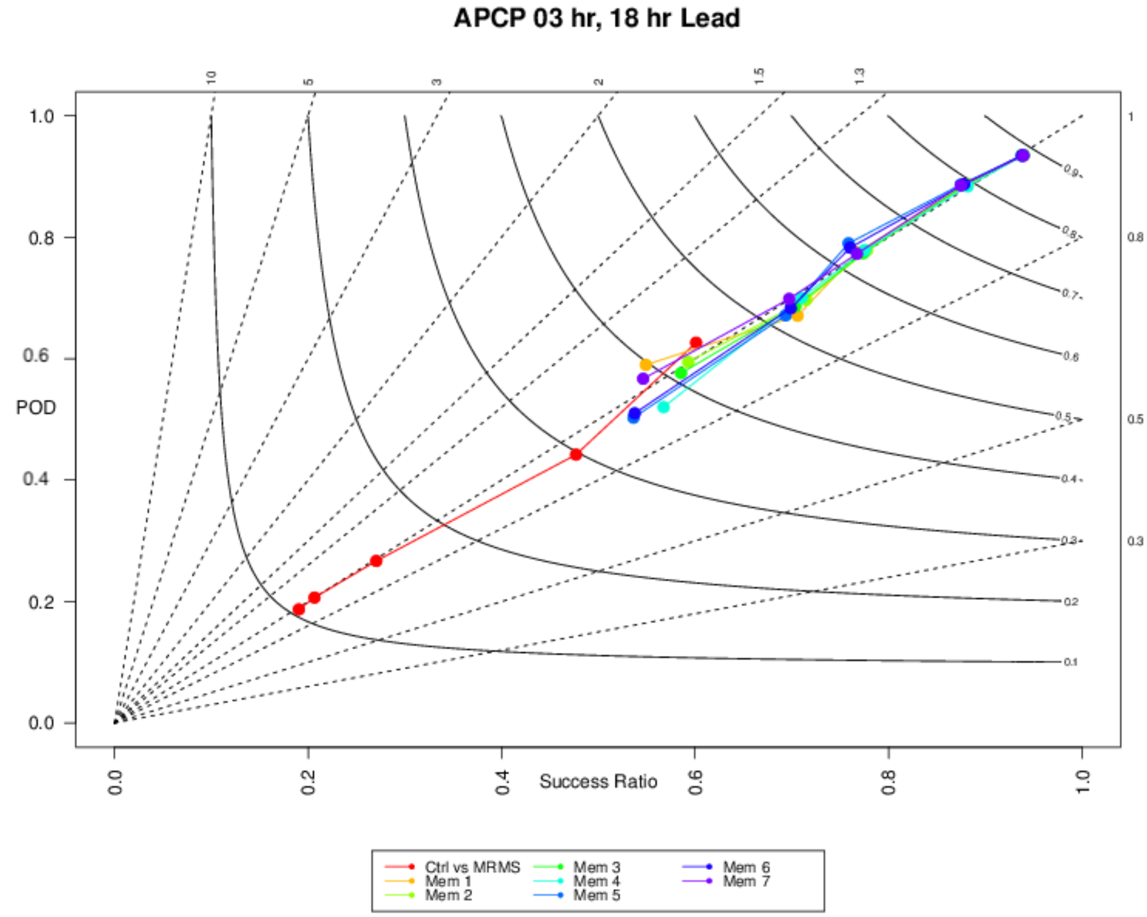


Figure 18. *Performance diagram showing CONUS-East 3hr accumulated precipitation agreement with an 18h lead for all thresholds between control and MRMS (red) and ensemble members and control (all other colors) during winter.*

To determine if the precipitation spatial coverage is similar between the perturbed ensemble members and the control, we examine the frequency bias statistic. This measure is the ratio of the counts of locations with forecast precipitation above the threshold for the member divided by that of the control. When the total precipitation coverage has an identical number of locations, the ratio is one. However, these locations need to be coincident. Because the comparison is with the control rather than the observation, the information here is not how accurate is the spatial precipitation coverage, but how similar is it to the control. For all thresholds and lead times, ensemble members produced very similar spatial coverage of precipitation (i.e. frequency biases near one). Figure 19 shows one example from the summer season in CONUS-East, for 24-hour accumulated precipitation with a 24-hour lead time at a variety of thresholds. Each ensemble member differs no more than about half of one percent (<0.5%) from the control in the number of grid locations where precipitation is forecast to be greater than the threshold. The differences are least at the lower thresholds with slightly greater differences at the highest thresholds.

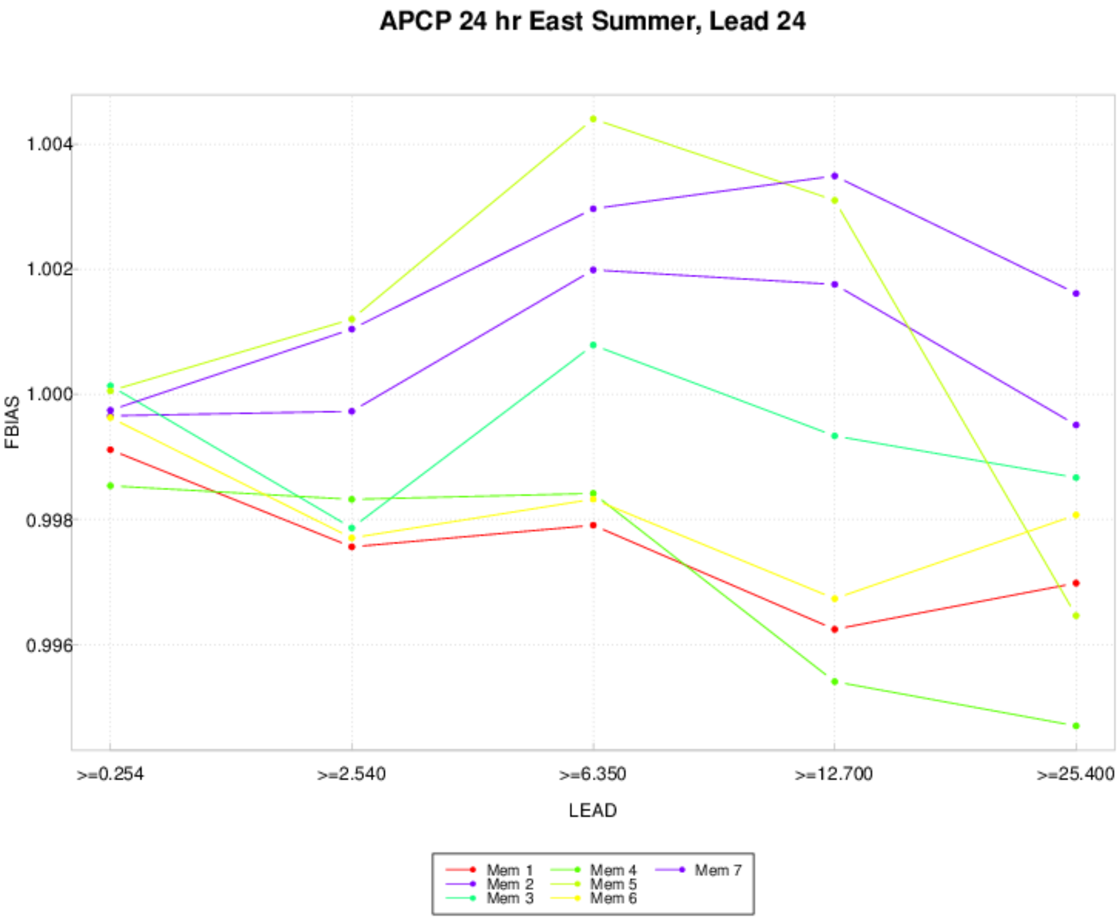


Figure 19. *Series of frequency bias by threshold for perturbed ensemble members vs. control for 24h accumulated precipitation over CONUS-East for the summer season.*

Summary

Using the previously established infrastructure to run a HRRR ensemble, a comprehensive testing and evaluation effort was undertaken to assess the implementation of SPP within the Thompson microphysics scheme. This report focused on evaluating traditional and ensemble verification for the 2017 warm season from model runs with all three stochastic parameter perturbations activated within the Thompson scheme (i.e., SPP applied to grapul treatment, cloud water distribution, and CCN activation).

Surface Verification Findings

In evaluating traditional metrics, temperature, dew point temperature, and wind speed exhibit little to no ensemble spread for approximately the first 12 - 16 hours of the forecast. However, ensemble spread does increase with forecast lead time, as anticipated. Bias, RMSE, and BCRMSE are not proportionally related to forecast lead time. Bias is minimal for all ensemble members as is RMSE. Wind speed exhibits the largest member spread of the metrics in question.

Evaluation of ensemble statistics revealed similar results to traditional methods with spread-skill plots, rank histograms, and reliability illustrating the small spread and under-dispersiveness of the ensemble members for temperature, dew point temperature, and wind speed.

Vertical Verification Findings

Evaluation of traditional metrics show that temperature, dew point temperature, and wind speed behave similarly to their surface counterparts, displaying small overall ensemble spread. A positive result is that all of the ensemble members displayed very low values of bias at all forecast hours examined. The behavior differs, however, with respect to ensemble spread and values of error with time. Temperature and dew point temperature reach peak RMSE values at forecast hour 24 while wind speed reaches its peak at forecast hour 36.

In evaluating ensemble statistics, the same under-dispersive and low spread behavior was observed in the spread/skill plots and rank histograms as was observed for the surface counterparts of the variables.

Accumulated Precipitation Findings

The highest GSS values for the individual members were seen at the lowest thresholds and at the earlier forecast lead times. In addition, the ensemble mean has more skill than the envelope of individual members, but the skill differential between the mean and individual members decreased as threshold increased.

All individual members produced forecasts that had minimal bias. As precipitation threshold increased, a general shift toward higher bias values was observed. At the lowest precipitation thresholds, the ensemble mean typically had smaller low bias values than the individual members; as the threshold increased, the ensemble mean had a smaller high bias than the ensemble members. With exception to the longer lead times at the lower precipitation thresholds, the ensemble mean generally performed better than the individual members.

For evaluating traditional metrics such as GSS and frequency bias, the individual members are generally clustered closely together showing minimal spread; however, a slight uptick in spread is observed as forecast lead time increases and as precipitation threshold increases.

Composite Reflectivity Findings

For all evaluated thresholds and individual members, GSS is highest at the earliest lead times, with an overall decrease in skill throughout the forecast period. Distinct diurnal trends were noted, with skill decreasing throughout the daytime before rebounding in the overnight hours.

Frequency bias values at all thresholds and for all members are often near 1, indicating minimally biased forecasts. In general, as threshold increases, there is a small increase in bias values. Diurnal variations are noted, with maximums and minimums depending on the threshold being evaluated, but, overall, lower bias is seen during the morning and early afternoon with higher biases in the evening and overnight.

Similar to accumulated precipitation, minimal spread was observed for both GSS and frequency bias. A slight increase in spread is seen with increasing forecast lead time and threshold.

Ensemble Verification Findings

Overall, ensemble spread of traditional metrics was hindered by the minimal impact that microphysics perturbations had on surface and upper-air temperature, dew point temperature, and wind speed. However, it is important to note that the MP perturbations did provide valuable spread to 3-hr precipitation verification, accounting for a large portion of the RMSE by 24 hours into the forecast. While these results were to be somewhat expected, they illustrate that the MP perturbations should ideally be used in concert with perturbations to other physics parameterizations in order to provide adequate spread.

The rank histogram plots mirrored what was found in the spread/skill diagrams, showing a lack of spread for both surface and upper-air variables (somewhat less so for precipitation accumulation). A number of rank histograms indicated a slight bias, but there was no systematic trend when considering all plots. Instead, a consistent under-dispersive signal was seen, with too many observations falling in the lowest and highest bins.

Reliability plots of surface temperature, dew point temperature, wind speed, and precipitation accumulation thresholds indicated a general tendency for the ensemble to underestimate occurrences of low probability events, and overestimate high probability events. Consistent with an under-dispersive ensemble, these findings match the other verification metrics showing that insufficient spread hindered the ability of the ensemble to forecast over a wide range of events.

Ensemble vs Control Member Findings

A comparison of precipitation accumulation thresholds for the control and ensemble members indicated that they all forecast about the same amount of precipitation over the domain, regardless of season, threshold, or lead time. The ensemble members that contain MP perturbations are quite skilled at replicating the precipitation from the control, even when the control lacks skill in replicating observed precipitation from MRMS. Spread in skill/frequency bias increases with increasing precipitation accumulation threshold and lead time. Direct comparison of the perturbed ensemble members with the control member provides complementary information to traditional information available via evaluations versus observations.

References

Berner J., G. Shutts, M. Leutbecher, and T. N. Palmer, 2009: A Spectral Stochastic Kinetic Energy Backscatter Scheme and its Impact on Flow-dependent Predictability in the {ECMWF} Ensemble Prediction System, *J. Atmos. Sci.,* **66,** 603-626.

Buizza R., M. Miller and T. N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF Ensemble Prediction System, *Quart. J. Roy. Meteor. Soc.,* **125,** 2887-2908.
Gilmore, M. S., J. M. Straka, and E. N. Rasmussen, 2004: Precipitation uncertainty due to variations in precipitation particle parameters within a simple microphysics scheme. *Mon. Wea. Rev.,* **132,** 2610–2627.

Jankov, I., J. Beck, J. Wolff, M. Harrold, J. B. Olson, T. Smirnova, C. Alexander, J. Berner, 2017: Stochastic Parameter Perturbations in a HRRR-Based Ensemble. Submitted to *Mon. Wea. Rev.*

Knight, C. A., W. A. Cooper, D. W. Breed, I. R. Paluch, P. L. Smith, and G. Vali, 1982: Microphysics. Hailstorms of the Central High Plains, C. Knight and P. Squires, Eds., Vol. 1, Colorado Associated University Press, 151–193.

Martin, G. M., D. W. Johnson, A. Spice, 1994: The measurement and parameterization of effective radius of droplets in warm stratocumulus clouds. *J. Atmos. Sci.,* **51,** 1823-1842.

McFarquhar, G. M., and R. A. Black, 2004: Observations of particle size and phase in tropical cyclones: Implications for mesoscale modeling of microphysical processes. *J. Atmos. Sci.,* **61,** 422– 439.

Model Evaluation Tools Users' Guide, Version 6.1 (METv6.1), 2017. Available at: https://dtcenter.org/met/users/docs/users_guide/MET_Users_Guide_v6.1.pdf

Morrison, H., S. Tessendorf, K. Ikeda, and G. Thompson, 2012: Sensitivity of a simulated mid-latitude squall line to parameterization of raindrop breakup. *Mon. Wea. Rev.,* **140,** 2437–2460.

Palmer, T. N., R. Buizza, F. Doblas-Reyes, T. Jung, M. Leutbecher, G. Shutts, M. Steinheimer, and A. Weisheimer, 2009: Stochastic parametrization and model uncertainty. *ECMWF Technical Memorandum*, 598, available at http://www.ecmwf.int/publications/.

Roebber, P.J. , 2009: Visualizing Multiple Measures of Forecast Quality, Weather and Forecasting. 24, pp - 601 - 608.

Shutts, G. J., 2005: A kinetic energy backscatter algorithm for use in ensemble prediction systems. *Quart. J. Roy. Meteor. Soc.*, 612, 3079–3102.

Skamarock W. C., J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, M. G. Duda, X.-Y. Huang, W. Wang and J. G. Powers, 2008: A Description of the Advanced Research {WRF} version 3, *NCAR Tech. Note, NCAR/TN-475+STR, 113pp.*

Thompson, G. and T. Eidhammer, 2014: A Study of Aerosol Impacts on Clouds and Precipitation Development in a Large Winter Cyclone. *J. Atmos. Sci.,* **71**, 3636–3658, doi: 10.1175/JAS-D-13-0305.1.

Thompson, G., M. K. Politovich, and R. M. Rasmussen, 2017: A numerical weather model's ability to predict the characteristics of aircraft icing environments. *Wea. and Forecasting,* **32,** 207-221, doi:10.1175/WAF-D-16-0125.1.

Tong, M., and M. Xue, 2008: Simultaneous estimation of microphysical parameters and atmospheric state with simulated radar data and ensemble square root Kalman filter. Part I: Sensitivity analysis and parameter identifiability. *Mon. Wea. Rev.,* **136,** 1630–1648.

Wilks, D. S., 2011: Statistical methods in the atmospheric sciences, International geophysics series, Vol. v. 100. 3rd ed., Elsevier/Academic Press, Amsterdam.