Addressing model uncertainty through stochastic parameter perturbations within the HRRR ensemble

Developmental Testbed Center (DTC) Annual Operating Plan (AOP) 2016 Final Report

Regional Ensemble Team: Isidora Jankov[1], Jamie Wolff[2], Jeff Beck[1], and Michelle Harrold[2]

[1]Cooperative Institute for Research in the Atmosphere (CIRA)/Affiliated with NOAA/ESRL/GSD and Developmental Testbed Center (DTC)

[2]National Center for Atmospheric Research (NCAR) and Developmental Testbed Center (DTC)

Table of Contents

Introduction

In recent years, representation of model uncertainty within an ensemble system, both global and regional, has been receiving increasing attention. To address uncertainty associated with model formulation, a number of different strategies have been proposed. A frequently used approach is using a multi-physics ensemble. Use of a combination of different physics schemes usually leads to large diversity among ensemble members, resulting in sufficient spread and improved forecast skill (e.g. Hacker et al. 2011b; Berner et al. 2011, Berner et al. 2015). Even though ensembles designed in this way are often characterized by good performance, there are both practical and theoretical deficiencies associated with them. For example, for the purpose of statistical post-processing, securing equally distributed and independent random variables is a necessity. This requirement cannot be satisfied when using the multi-physics approach. The post-processing of a multi-physics ensemble is further complicated by the fact that each ensemble member has a different mean error and climatology, which is often the reason that these ensembles have sufficient spread (Berner et al. 2015, Eckel and Mass 2005).

Despite their satisfactory performance (e.g., Berner et al. 2009, Berner et al. 2011), the main criticism of the SKEB and SPPT schemes has been that they are an ad hoc addition to numerical weather prediction, instead of being developed and implemented within the model physics. To account for a need to address model uncertainty at its source, the Stochastic Parameter Perturbation (SPP) approach was developed. It can be applied by having the parameter of choice unchanged throughout the integration (e.g. Murphy et al. 2004; Hacker et al. 2011a) or by varying randomly in time and space (e.g. Bowler et al. 2009). For long integrations, and with the latter approach, all ensemble members are expected to have the same climatology. Previous studies have shown that the SPP approach usually outperforms unperturbed ensembles but frequently results in insufficient spread (Hacker et al. 2011b, Reynolds et al. 2011, Berner et al. 2015,Christensen et al. 2015).

Focusing on addressing uncertainty at its source, the present study employs the SPP approach alone, and in combination with SKEB and SPPT. While Jankov et al. 2017 originally performed similar experiments with a Rapid Refresh (RAP)-based ensemble, this study focuses on evaluating the impact of SPP and its combination with SKEB and SPPT on high-resolution ensemble performance. At each grid point, perturbed parameter variations were constrained with spatial and temporal correlations. Perturbations were applied within a common physics suite, applicable for inclusion in the next-generation convection-allowing ensemble.

Experiment Design

Model

The operational High Resolution Rapid Refresh (HRRR) configuration was used as a basis for all experiments. Simulations were performed over the operational HRRR Contiguous United States (CONUS) domain (Fig. 1) with 3-km grid spacing and eight ensemble members. The experimental dataset consisted of 10 spring season days starting on 18 May 2016 and ending on 27 May 2016. The simulations were initialized at both 0000 UTC and 1200 UTC with a

simulation length of 24 hours. The lateral boundary and initial conditions were provided by the Rapid Refresh (RAP; Benjamin et al. 2016). The model initialization included partial cycling (soil attributes were cycled hourly), matching the HRRR system settings running in operations at the National Centers for Environmental Prediction (NCEP). The HRRR system uses the Advanced Research version of the Weather Research and Forecasting (ARW-WRF) dynamic core (Skamarock et al. 2008). The physics suite used for both operational systems includes the Mellor-Yamada-Nakanishi-Niino (MYNN; Nakanishi and Niino 2004, Nakanishi and Niino 2006) planetary boundary layer (PBL) parameterization and the Rapid Update Cycle (RUC; Smirnova et al. 2016) land surface model (LSM) parameterization.

The multi-physics ensemble, which represents the control experiment (mixed_phys), used different physics parameterizations for the PBL and LSM schemes (Table 1). The different PB schemes included the Mellor-Yamada-Janjic, MYNN,Yonsei University (YSU; Hong et al. 2006), and Pleim-Xu (Pleim 2007) parameterizations. In terms of the LSM options, the RUC (Smirnova et al. 2016) and Noah (Ek et al. 2003) schemes were employed. The eight-member multi-physics ensemble contained a combination of the four PBL and two LSM schemes.

All members of the stochastic ensemble experiments used the same physics parameterizations as the operational HRRR (Table 1). One of the stochastic experiments perturbed initial soil moisture (spp_LSM) values only, one perturbed multiple parameters within the MYNN PBL (spp_PBL) throughout the forecast, and the final experiment combined the previous PBL perturbations with SKEB and SPPT (sppPBL_skeb_sppt).

The SPP approach used here was adapted from the previously mentioned work that utilized the RAP-based ensemble system, and a detailed explanation of the method and creation of the SPP perturbations is available in (Jankov et al. 2017). In summary, the spatially and temporally correlated pattern is fully determined by three namelist parameters: grid point standard deviation (gridpt stddev rand pert), length scale (lengthscale rand pert) and de-correlation time (timescale rand pert). Additionally, since drawing from a Gaussian distribution can result in very large values, the random numbers are constrained. This capping threshold is expressed in terms of a maximum standard deviation (stddev cutoff rand pert).

While the first guess for parameter pattern values (e.g., spatial and temporal de-correlations) were based on suggestions from HRRR developers, the final settings were chosen after a series of sensitivity tests. The sensitivity experiments included the following combinations of spatial and temporal de-correlation lengths, which were chosen based on typical spatial and temporal advective scales: 150 km and 6 hours, 300 km and 12 h, and 600 km and 24-h. Experiments with a spatial and temporal de-correlation length of 150 km and 6 hours, respectively, resulted in the greatest skill. Therefore, these values were used for each of the experiments employing the SPP approach, as well as the experiment that included initialization soil moisture perturbations.

Observations

For evaluation of accumulated precipitation, the Multi-Radar/Multi-Sensor (MRMS) local gauge bias-corrected radar quantitative precipitation estimation (QPE) analyses were used. This

dataset integrates radar data with atmospheric environmental data, satellite data, and lightning and rain gauge observations to generate a suite of severe weather and QPE products at very high spatial (1 km) resolution (Zhang et al. 2016). Prior to performing the evaluation, the MRMS gridded dataset was re-gridded to the 3-km integration domain to allow for direct grid-to-grid comparisons. Precipitation was verified over 3- and 24-h accumulations. For conventional surface and upper-air point observations, RAP observation files in Binary Universal Form for the Representation of meteorological data (BUFR) format were used. Verification of standard meteorological fields (temperature, dew point, and wind) was performed hourly for surface variables and at times valid at 00 and 12 UTC for upper-air variables. When compared to model output, bilinear interpolation was performed.

Results

Precipitation Verification

All simulations were performed over the CONUS domain (Fig. 1). Verification of ”raw” model output (there was no post processing applied to the model output such as bias removal and/or calibration) was performed using Model Evaluation Tools (MET; Bullock et al. 2017) software over the CONUS, CONUS-East, and CONUS-West 3-km verification domains (Fig. 1) for runs initialized at both 0000 UTC and 1200 UTC. Trends in results for CONUS-East and CONUS-West for the two initializations were very similar. Given this, results discussed here will be restricted to the CONUS-East domain for 0000 UTC initializations only. Confidence intervals (CIs) at the 95% level were applied to the computed statistics in order to estimate the uncertainty associated with sampling variability; however, observational uncertainty was not considered in this study. The CIs were computed using the bootstrapping technique and resampling with replacement was conducted 1500 times.

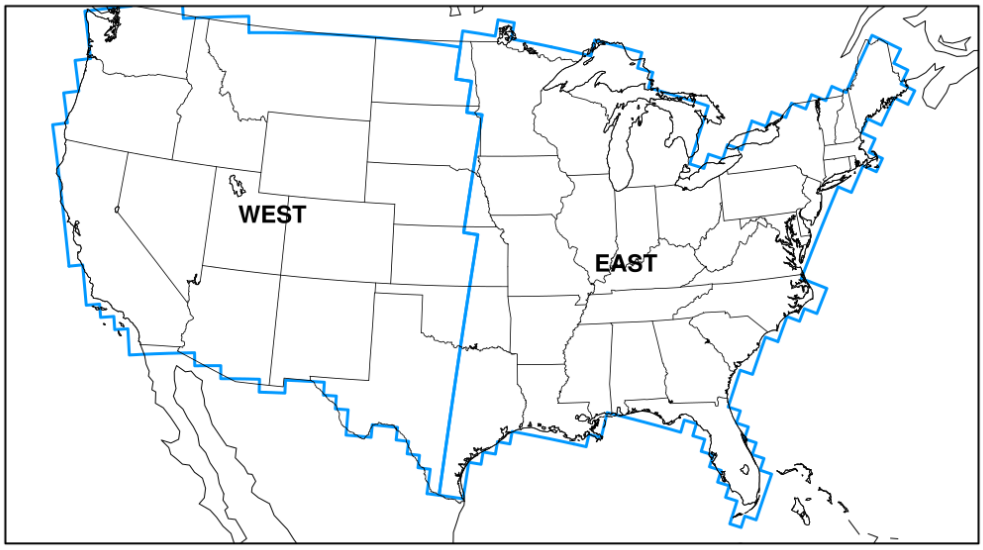

Figure 1. Verification domain with the division between western and eastern regions presented with blue lines.

Precipitation performance was assessed using a number of verification metrics for deterministic and probabilistic forecast assessment, including rank histograms, frequency bias, Gilbert Skill Score (GSS), and reliability.

Figure 2 shows rank histograms of three-hourly accumulations for 00-24-h (Fig. 2) lead time for each experiment. The rank histogram for all experiments generally indicates lower relative frequency values for the middle bins and higher values for the outermost bins. Specific features differ somewhat among the experiments, however. The mixed_phys experiments indicated bias toward heavier precipitation. Similarly, the spp_PBL and sppPBL_skeb_sppt experiments exhibited bias but were also characterized with presence of under-dispersion. The spp_LSM ensemble was characterized by similar values of relative frequencies in the outermost bins indicating a tendency to be under-dispersed, rather than biased.

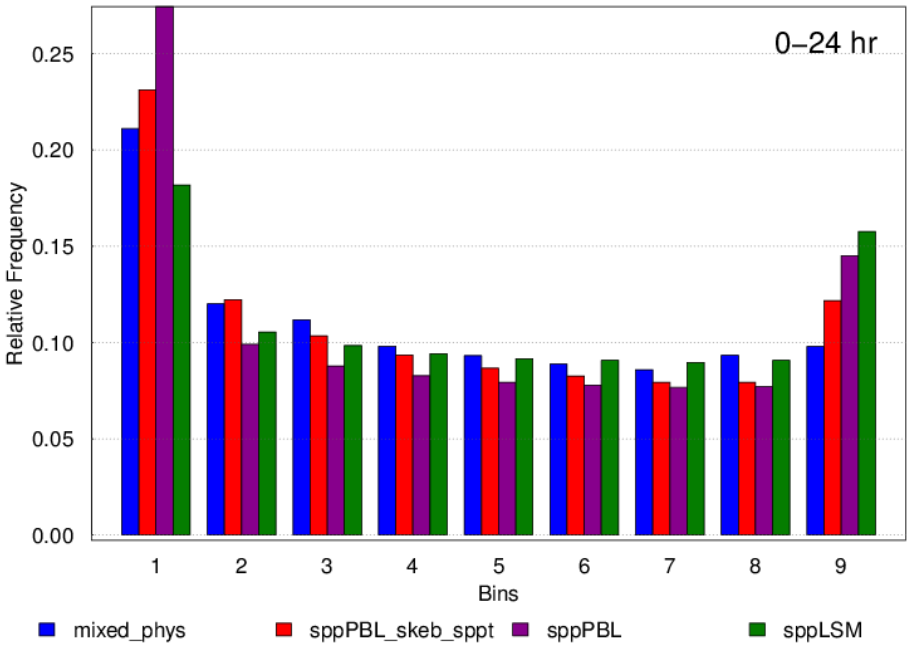

Figure 2. Rank histograms for all experiments for 0000 UTC initializations over the eastern part of the CONUS and for 0-24 h lead time.

Frequency bias was calculated as the ratio of forecast to observed grid points exceeding a specified precipitation threshold. A perfect score for frequency bias is one, where values higher (lower) indicate that the model over-predicted (under-predicted) the exceedance of a given threshold. In the present study, frequency bias was analyzed for two precipitation thresholds (0.254 mm and 12.7 mm) as an aggregate over all members of each experiment (Fig. 3). Confidence intervals at the 95% level for each experiment were applied.

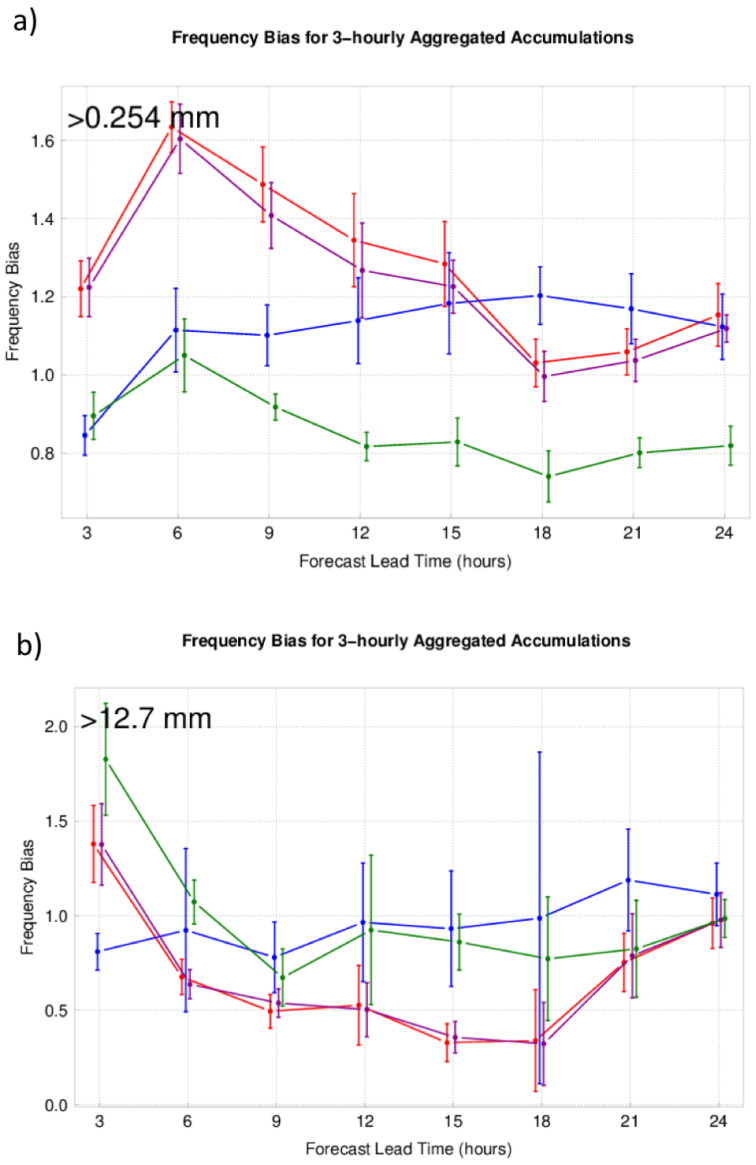

Figure 3. Frequency Bias aggregated over all members for all experiments as a function of lead time for 0000 UTC initializations over the eastern part of the CONUS for a precipitation accumulation threshold of a) >0.254 mm and b) >2.54 mm.

Frequency bias for 3-h accumulated precipitation greater than 0.254 mm (Fig. 3a) showed statistically significant differences between several experiments (i.e., confidence intervals did not overlap), although not for all lead times. In general, the mixed_phys and spp_LSM experiments were frequently statistically significantly different from the spp_PBL and sppPBL_skeb_sppt experiments. The spp_LSM ensemble was the only experiment characterized with frequency bias values lower than one for most of the lead times while the other three experiments most often had frequency bias values larger than one. While the mixed_phys frequency bias increased with lead time, the other three experiments generally decreased with lead time after an initial increase at the 6-h lead time. The sppPBL_skeb_sppt and spp_PBL ensembles had significantly higher frequency bias values for this threshold and the first 12 hours of the forecast compared to the other two experiments with an improved frequency bias

later in the period. The same type of analysis, except for 12.7 mm precipitation threshold, showed similar trend and values close to one (confidence intervals frequently encompassed one) for most of the lead times for the mixed_phys and spp_LSM experiments (Fig. 3b). The sppPBL_skeb_sppt and spp_PBL again had similar behavior among themselves, and often exhibited a significantly low bias for this threshold.

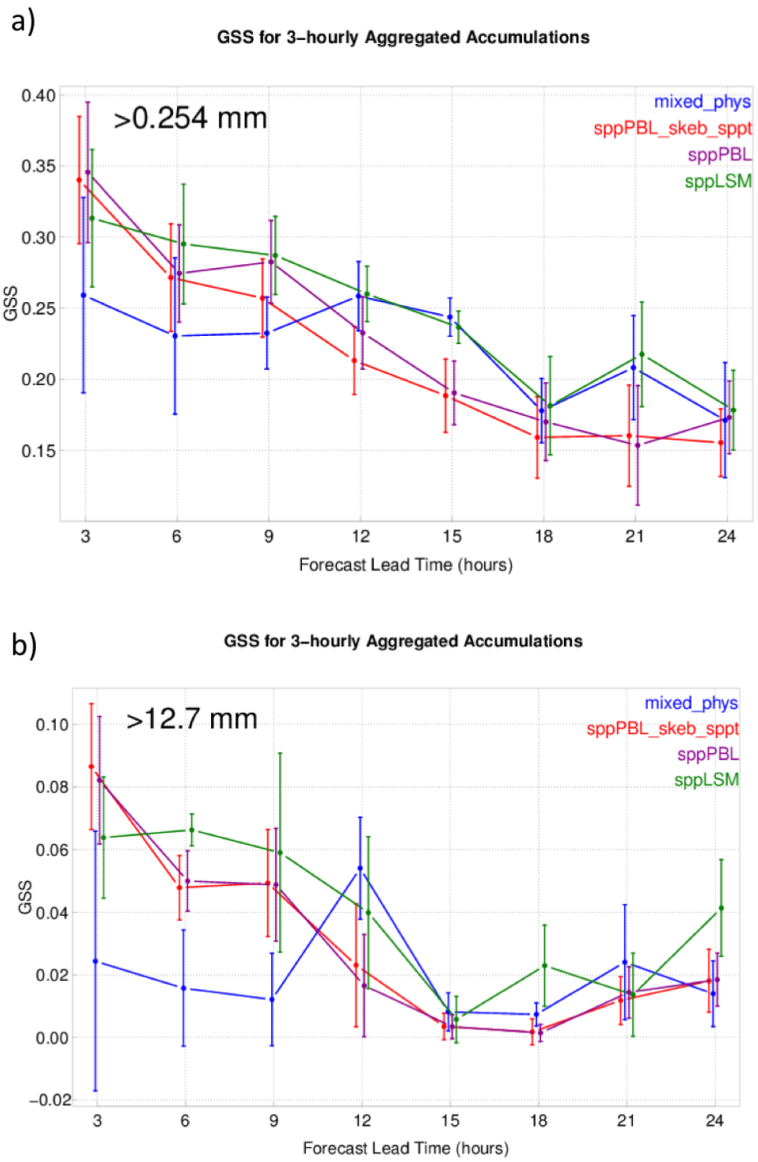

Figure 4. Same as Fig. 3 except showing Gilbert Skill Score (GSS).

In addition, for the same precipitation thresholds and for the same type of summation, Gilbert Skill Score (GSS) values were examined (Schaefer 1990). Figure 4 shows GSS for the two precipitation thresholds and for each experiment as a function of lead time. Fig. 4a shows that for the 0.254 mm precipitation threshold, the GSS values decreased with lead time for all experiments. While the median values of GSS for all of the stochastic experiments were highest during the first nine hours, the spp_PBL and sppPBL_skeb_sppt decreased more rapidly with

time and the mixed_phys increased to a comparable value of the spp_LSM experiment for the reminder of the forecast. Except at one forecast lead time the differences were not statistically significant. For the heavier precipitation threshold (Fig. 4b), the stoch_phys experiment had higher median GSS values early in the forecast period with a few being statistically significant. All experiments had similar results, with some advantage of mixed_phys and spp_LSM experiments for longer lead times.

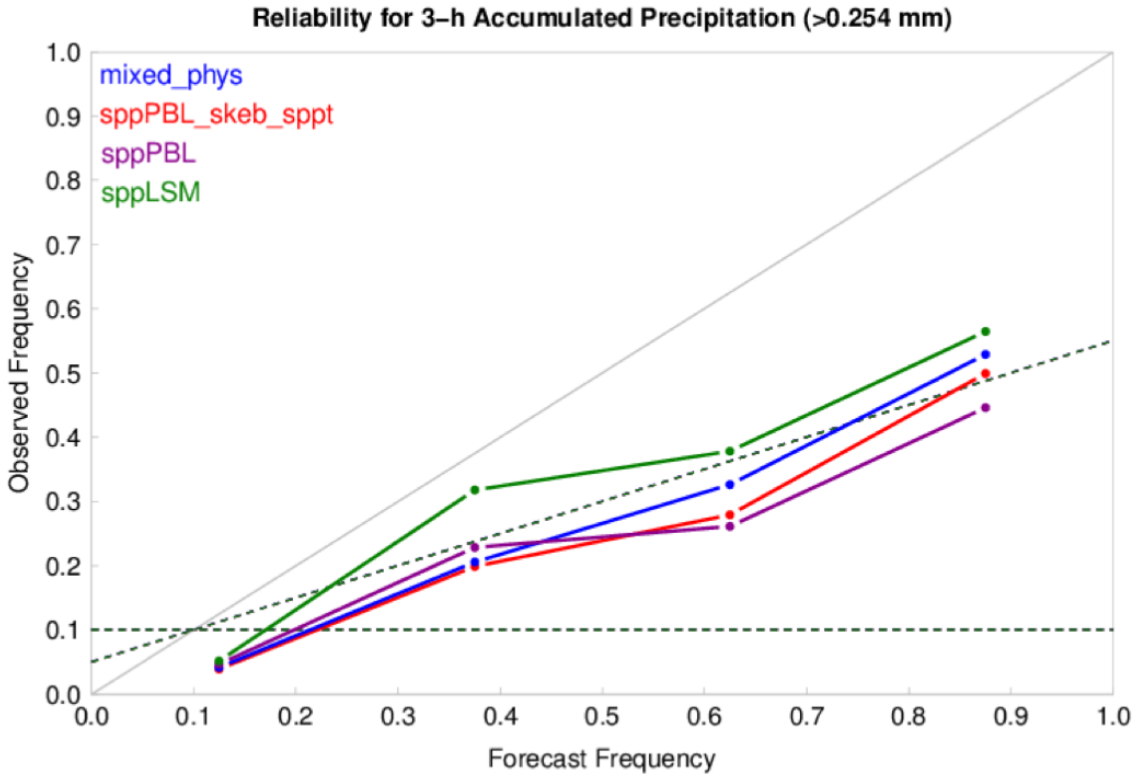

Figure 5. Reliability diagram for 0000 UTC initializations over the eastern part of the domain and for a precipitation accumulation threshold >0.254 mm. The horizontal dotted line represents the sample base rate, the diagonal dotted line represents no skill, and the solid grey diagonal line represents perfect reliability.

Reliability diagrams were created for 1-h accumulation periods aggregated together over the full 24-h period, for the 0.254 mm precipitation threshold and each experiment (Fig.5). The observed frequency (i.e. the sample base rate) of this event is about 10% of the grid locations over the 10-day period, making it a somewhat uncommon event. Reliability diagrams measure the calibration of a probability forecast. With an infrequent event and a small ensemble, even a very good forecast will not conform perfectly to the diagram and any assessment should be tempered with these expectations. All ensembles showed a similar trend, with a generally higher observed proportion of events when the ensemble probability values were higher. Thus, all ensembles had some ability to discriminate these precipitation events from non-events. However, all ensembles over-estimated the probability of the precipitation events (i.e. they fall below the solid grey one-to-one line). Additionally, the central probability categories (37.5% and 62.5%) showed little difference in observed frequency for the sppPBL experiment, suggesting that sppPBL lacked discrimination in its central forecast probabilities. The spp_LSM

was characterized by somewhat higher reliability compared to other experiments for all forecast frequencies except the lowest one.

To summarize the 3-h accumulated precipitation verification results, the rank histograms indicated some level of under-dispersion and bias for all ensemble experiments. While frequency bias results depended strongly on threshold and forecast lead time, the GSS analysis for both light and heavier precipitation thresholds generally showed comparable results for all experiments with only a few statistically significant differences noted. Though over-confident, the spp_LSM experiment most often had the higher reliability compared to other experiments.

Finally, CONUS-wide probabilities of 24-h precipitation accumulations exceeding 25.4 mm were evaluated for each experiment initialized at 0000 UTC May 24, 2016 (Fig. 6a-d). Total precipitation accumulations for this period using MRMS measurements are also shown in Figure 6e. Significant areas of precipitation were generated by a convective line on the northern border of Kansas, which formed around 0800 UTC; however, the system dissipated during south-eastward propagation through Kansas. At around 1400 UTC, a well-defined convective line reinitiated over central and southern Missouri and continued to propagate south, south-east, terminating in northern Arkansas and southeast Missouri at the end of the period. The mixed_phys (Fig. 6a) and spp_LSM (Fig. 6d) experiments appear to capture the potential for this observed evolution, while only one member of the sppPBL_skeb_sppt (Fig. 6b) and no members of spp_PBL (Fig. 6c) experiments produced precipitation >25.4 mm anywhere in Missouri.

When compared to the stochastic experiments, the mixed_phys (Fig. 6a) experiment generally produced probabilities covering a larger areal extent. The sppPBL (Fig. 6c) and sppPBL_skeb_sppt (Fig. 6b) experiments produced more focused probabilities, generally limited to northern Kansas, and were shifted more northwestward than probabilities from the mixed_phys experiment. In addition, probabilities over northern Kansas were somewhat higher for the two experiments as compared to mixed_phys due to smaller spread. Low probabilities extended further toward the southeast in Kansas and into Missouri for the sppPBL_skeb_sppt experiment when compared to sppPBL, indicating that the combination of SKEB and SPPT resulted in a more diverse solution compared to sppPBL. The sppLSM solution was more similar to mixed_phys with areas of high probabilities over the Missouri/Arkansas border. In the case of sppLSM, high probability areas were more concentrated and characterized by higher values compared to mixed_phys as a consequence of smaller spread. On the other hand, reliability analysis (for all cases aggregated together) showed higher reliability in the case of spp_LSM. More precisely, spp_LSM was characterized with smaller spread but higher reliability indicating sharper and more useful forecast.

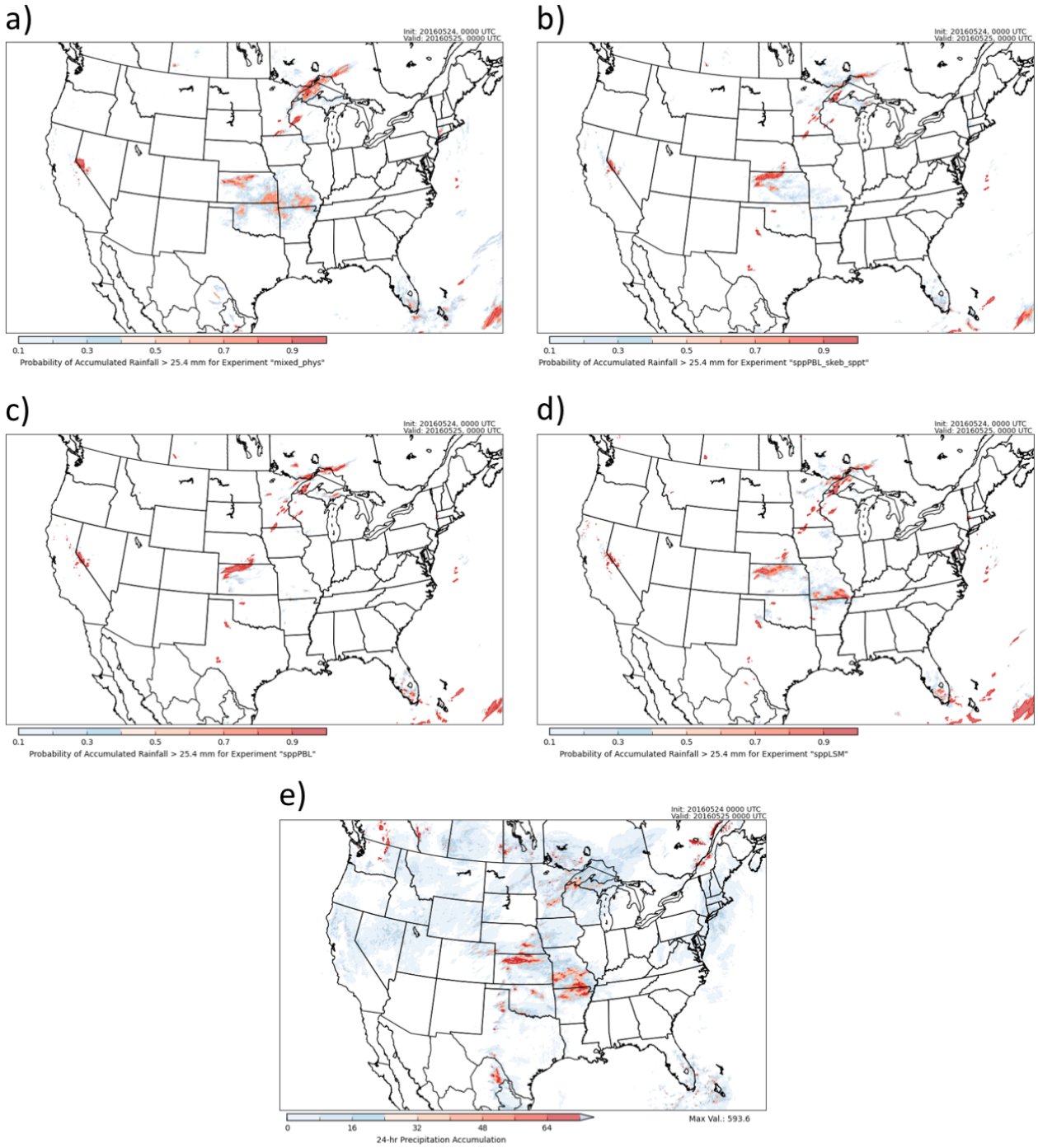

Figure 6. Probability of 24 h precipitation accumulation >25.4 mm threshold for the a) mixed phys, b) sppPBL skeb sppt, c) sppPBL, and d) sppLSM experiments, and e) total 24 h precipitation accumulation ending at 0000 UTC on 25 May 2016.

Surface Verification

In addition to precipitation, forecasts of surface variables including 2-m temperature, 2-m dew point temperature, and 10-m wind were analyzed. Once again, discussed results will be concentrated on the CONUS-East domain for 0000 UTC initializations only. At the current time, MET does not include observational error; therefore, it is not considered here. Taking

observational uncertainty into account for ensemble evaluation has been shown to affect verification of short term simulations (Bouttier et al. 2012). Inclusion of observational error would likely reduce the level of under-dispersion (Candille and Talagrand 2008).

Aggregate root mean square error (RMSE) values of the ensemble mean and corresponding spread values were computed for all experiments. The spread was computed as the average ensemble standard deviation over the domain. The ensemble mean is simple arithmetic average of the members. A ratio between the spread and error (spread/error ratio) was also assessed. RMSE, spread, and the ratio of the two concisely summarizes ensemble performance. It is desirable to have comparable spread and error values (i.e., having spread encompass the error), producing a ratio between the two near one.

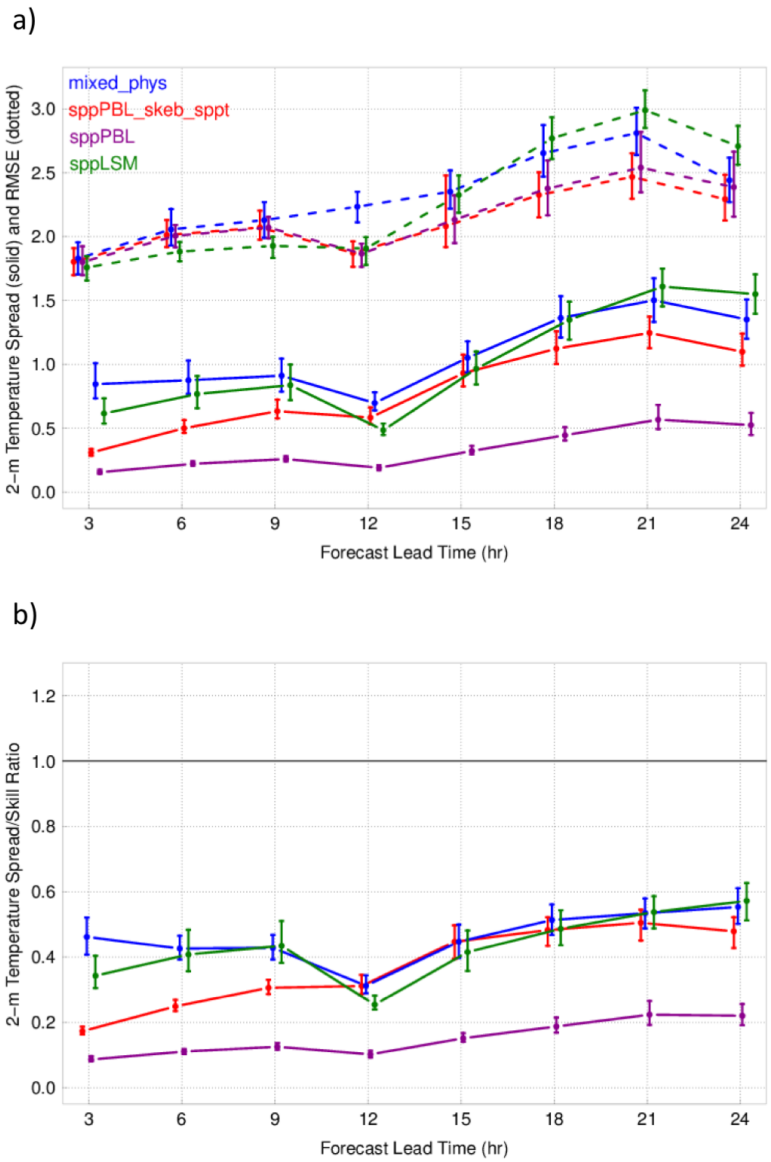

Figure 7. 2-m temperature a) RMSE and spread and b) spread/error ratio as a function of lead time for 0000 UTC initializations over the eastern part of the domain. The 95% confidence intervals are included.

For 2-m temperature (Fig. 7a), all experiments had comparable RMSE values early in the forecast (evening/overnight hours). While the spp_LSM aggregate RMSE values were lower than the other experiments overnight, the errors increased most rapidly for the spp_LSM experiment during the day which led to significantly higher error when compared to spp_PBL and sppPBL_skeb_sppt, but not mixed_phys.

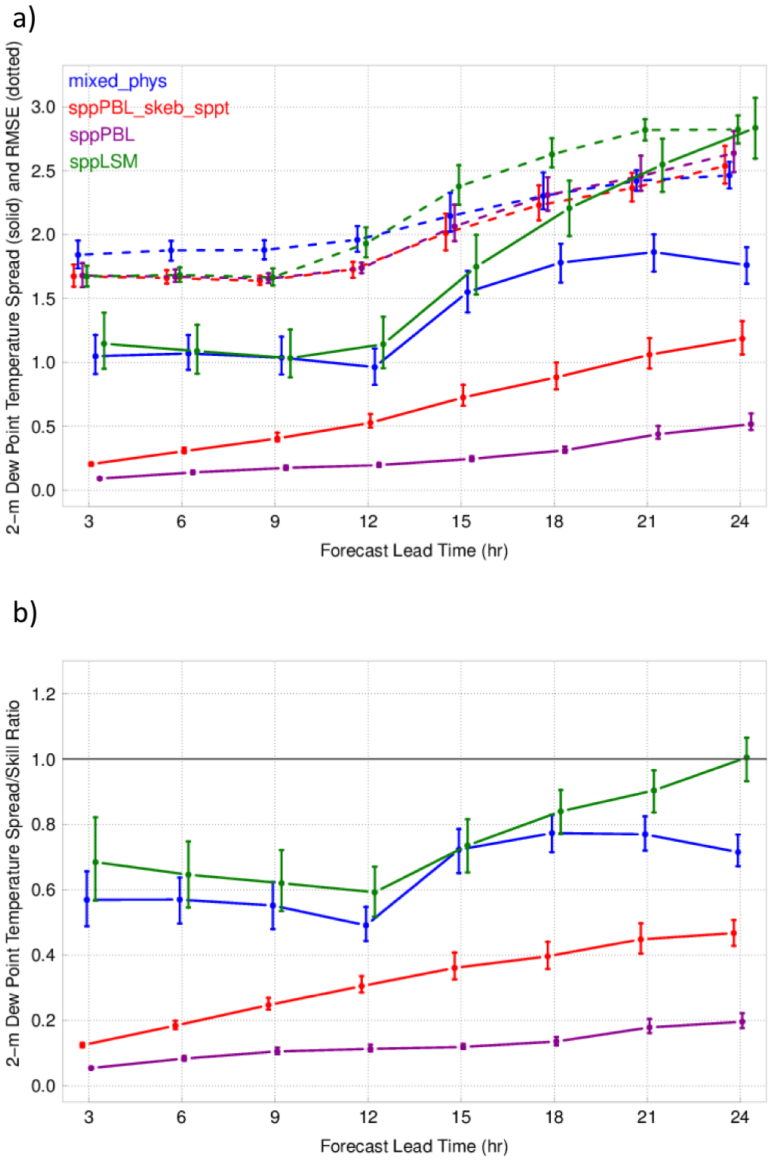

Figure 8. Same as Fig. 7 except for 2-m dew point temperature.

Spread values for all experiments were lower than RMSE values indicating under-dispersion. However, spread values varied widely between the experiments. The mixed_phys and spp_LSM experiments had comparable spread values for most forecast hours (exceptions included forecast hours 3 and 12). Overnight, sppPBL_skeb_sppt and spp_PBL had significantly lower spread compared to the other two experiments. During the day, the spread for sppPBL_skeb_sppt increased and approached the other two experiments, while spp_PBL was characterized with significantly lower spread for the duration of the forecast period.

In terms of spread/error ratio (Fig. 7b), variability between the experiments was evident during the overnight hours with mixed_phys and spp_LSM having significantly higher spread/error values (closer to one) compared to the other two experiments. However, during the day, all experiments were characterized by similar ratio values with exception to the spp_PBL experiment, which had significantly lower ratio values. Due to relatively similar RMSE values for each experiment, the spread/skill ratio was largely regulated by the spread of each experiment; this leads to the ratio having very similar characteristics to the ensemble spread results show in (Fig. 7a).

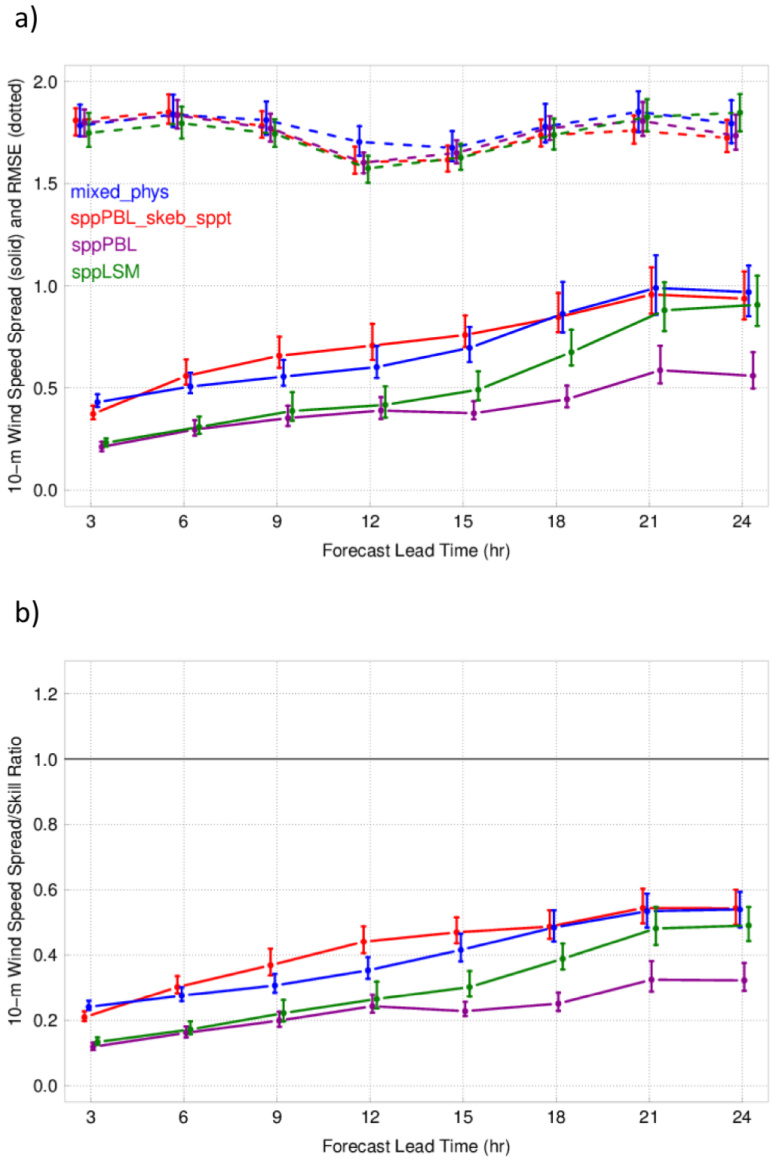

Figure 9. Same as Fig. 7 except for 10-m wind speed.

RMSE, spread, and spread/error analysis as a function of lead time for 2-m dew point temperature is presented in Figure 8. During the overnight hours, mixed_physics had significantly higher RMSE values when compared to the other experiments. During the day, as

was the case for 2-m temperature, spp_LSM was characterized by a rapid increase and significantly higher error compared to other experiments (Fig. 8a). In terms of spread, the mixed_phys ensemble had comparable spread to spp_LSM overnight, while during the day the spp_LSM spread was significantly larger than any other experiment. In the case of sppPBL_skeb_sppt, spread increased with lead time but was still significantly lower compared to spp_LSM and mixed_phys. The spp_PBL experiment was once again characterized by significantly lower spread for the duration of the forecast. For spp_LSM, the high RMSE values were accompanied by significantly higher spread that resulted in a spread/error ratio closer to one for this experiment compared to others (Fig. 8b). The low spread for both sppPBL_skeb_sppt and spp_PBL resulted in a low spread/error ratio (Fig. 8b).

For the 10-m wind, RMSE, spread, and the spread/skill ratio are presented in Figure 9. All experiments had similar RMSE values for all forecast lead times. While all experiments also had generally low spread, the spread values differed notably among the experiments (Fig. 9a). Mixed_physics and sppPBL_skeb_sppt had comparable spread that was significantly larger compared to the other two experiments for most lead times, with spp_LSM having increasing spread toward the end of the period. With comparable RMSE values between the experiments, spread/error ratio trends were again dominated by the differences in spread between the experiments (Fig. 9b); for all forecast lead times, all experiments had spread/error ratios below one.

Generally, for 2-m temperature (Fig. 7) and 2-m dew point temperature (Fig. 8), both spp_PBL and sppPBL_skeb_sppt were characterized by small spread; though an increase in spread was noted with lead time for sppPBL_skeb_sppt, which can be attributed to the addition of SKEB and SPPT. This result implies that the PBL perturbations did not have much impact on either 2-m temperature or 2-m dew point temperature and also indicates the limited impact of SKEB and SPPT on surface variables. Overall, this suggests solely perturbing physics scheme parameters is currently not enough to achieve sufficient spread (Jankov et al. 2017, Hacker et al. 2011b, Reynolds et al. 2011, Berner et al. 2015).

Further, reliability diagrams for surface variables aggregated over the full 24-h forecast period were evaluated for select thresholds (Fig. 10). Figure 10a shows reliability of 2-m temperature at a threshold greater than 293 K. The sample base rate for this threshold was 50%. It can be seen that the most reliable ensemble varied with forecast frequency with the stochastic experiments having better reliability for certain frequencies, compared to the mixed_physics ensemble, which was generally over-confident.

Reliability for 2-m dew point temperature was evaluated for the greater than 283 K threshold (Fig. 10b). The sample rate for this threshold was about 70%. All, experiments performed similarly for lower forecast frequencies, with relatively good reliability transitioning to becoming under-confident. For higher forecast frequencies, the spp_LSM and mixed_physics ensembles remained under-confident while the sppPBL_skeb_sppt and spp_PBL ensembles were again more reliable.

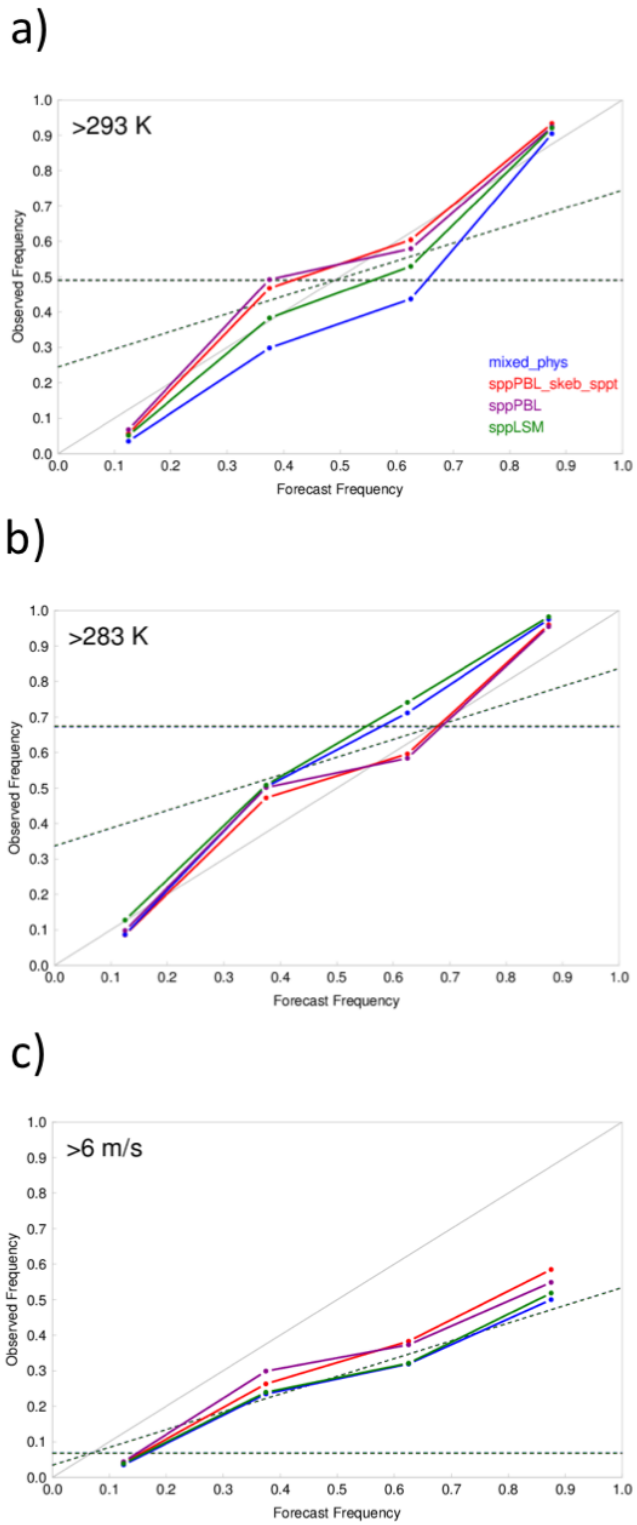

Figure 10. Reliability diagrams for 0000 UTC initializations over the eastern part of the domain for 2-m temperature for a threshold of a) >293 K, 2-m dew point temperature for a threshold of b) >283 K and 10-m wind for a threshold of c) >2 m/s. The horizontal dotted line represents no resolution, the diagonal dotted line represents no skill, and the solid grey diagonal line represents perfect reliability.

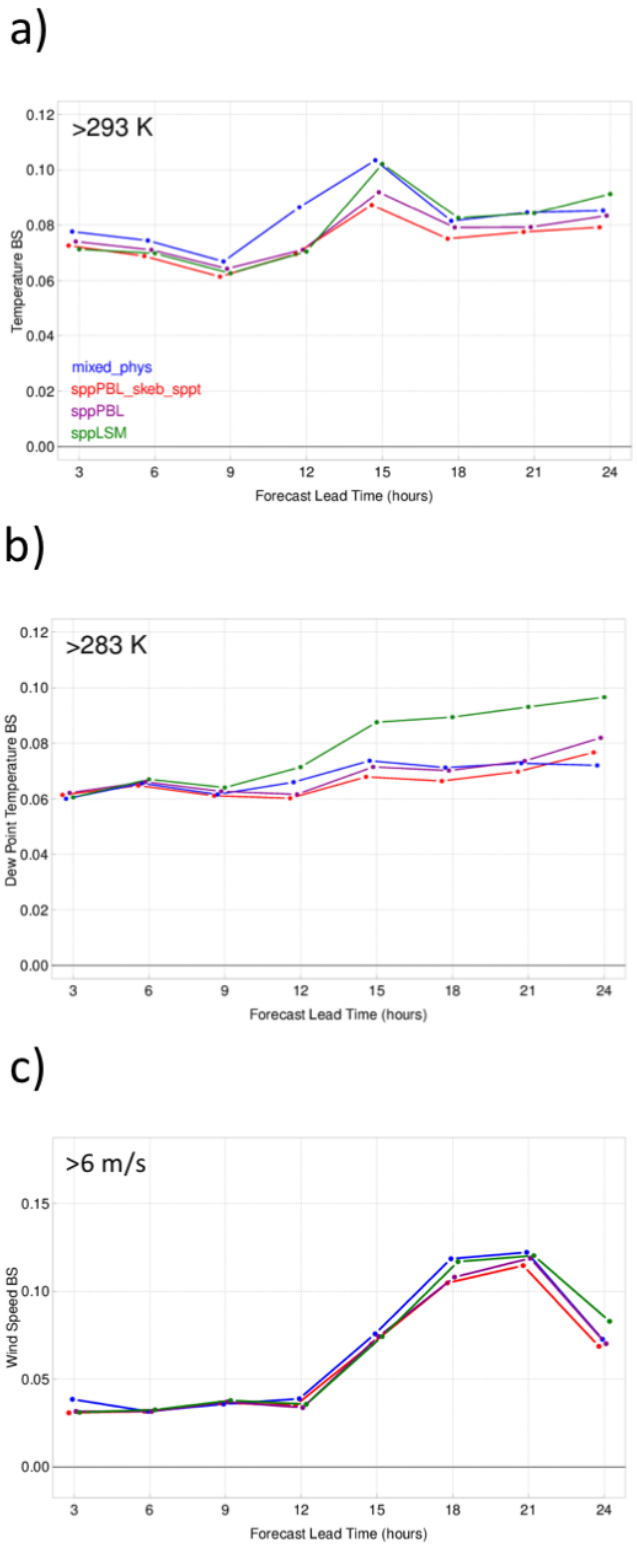

Figure 11. BS for 0000 UTC initializations over the eastern part of the domain 2-m temperature for a threshold of a) >293 K, 2-m dew point temperature for a threshold of b) >283 K and 10-m wind for a threshold of c) >2 m/s.

For 10-m wind speed, ensemble reliability was evaluated for wind speeds greater than 6 ms-1. (Fig. 10c). For this variable and threshold, the sample rate was somewhat lower than 10%. For

all forecast frequencies, all experiments were over-confident. However, the spp_PBL and the sppPBL_skeb_sppt experiments were characterized by better reliability compared to the other two experiments for all forecast frequencies. This was also the case for a number of other evaluated thresholds (not shown), implying that SPP perturbations in the PBL scheme led to improved reliability and sharpness of 10-m wind forecasts.

Next, the Brier Score (BS) was computed (Brier 1950). BS was calculated for surface variables using the same thresholds as the reliability diagrams. BS values as a function of lead time for 2-m temperature greater than 293 K threshold are presented in Fig. 11a. For this threshold, all experiments generally showed a similar trend with lower values overnight and an increase in BS during the early part of the day. Also, for lead times prior to 15 h, the mixed_phys experiment was characterized by slightly higher BS values compared to others. For a 2-m dew point temperature threshold of 283 K (Fig. 11b), similar behavior was again observed, with the exception spp_LSM which yielded higher BS values during the day. Similarly, for 10-meter wind speed greater than the 6 m/s threshold (Fig. 11c) the lowest BS values were detected 399 during the night followed by a sharp increase in BS during the early part of the day and a slight decrease in the afternoon hours. Interestingly, for all near-surface variables evaluated, BS values during the day were generally lower for the spp_PBL and the sppPBL_skeb_sppt experiments. This may imply positive impact on skill of surface variables when SPP approach was used.

Upper-Air Verification

A similar analysis to that performed for surface variables was also performed for select upper-air variables and levels, including 850-hPa temperature, 500-hPa geopotential height and 250-hPa wind (Figure 12). For 850-hPa temperature, all stochastic experiments had significantly lower RMSE at initialization time compared to the mixed_phys (Fig. 12a). The RMSE values generally increased with forecast lead time, with spp_LSM having the largest RMSE values by the end of the period (Fig. 12a). The spp_LSM experiment had significantly larger spread at initialization time, implying an impact of initial condition perturbations associated with the cycling of soil moisture and temperature. The sppPBL_skeb_sppt experiment had the largest spread at the 1200 UTC valid time. Overall, the spp_PBL experiment had the lowest spread, which was significant by the 24- forecast. This result is also clearly indicated in the spread/error ratio (Fig. 12b), with significantly lower values associated with the spp_PBL experiment.

While there are no statistically significant differences in RMSE values for 500-hpa geopotential height, spread varies significantly between each of the experiments (Fig. 12c). The sppPBL_skeb_sppt experiment had significantly larger spread, followed by mixed_phys, spp_LSM, and finally spp_PBL. This result is again highlighted in the spread/error ratio plot (Fig. 12d).

While mixed_phys had significantly larger error at the initial time for 250-hPa wind, differences in RMSE between all of the experiments were not significant at 12- and 24-h lead times (Fig. 12e). Similar spread for sppPBL_skeb_sppt and mixed_physics was noted, which was significantly larger when compared to the other two experiments. While the mixed_phys

spread/error ratio was higher (significantly for 24-h lead time) than the spp_LSM and spp_PBL experiments, it was not significantly different from the sppPBL_skeb_sppt (Fig. 12f).

In general, the upper-air analysis indicates that the use of SKEB and SPPT improves model performance (e.g. spread and reliability) for upper-air variables. This finding was also valid for the RAP-based ensemble (Jankov et al. 2017).

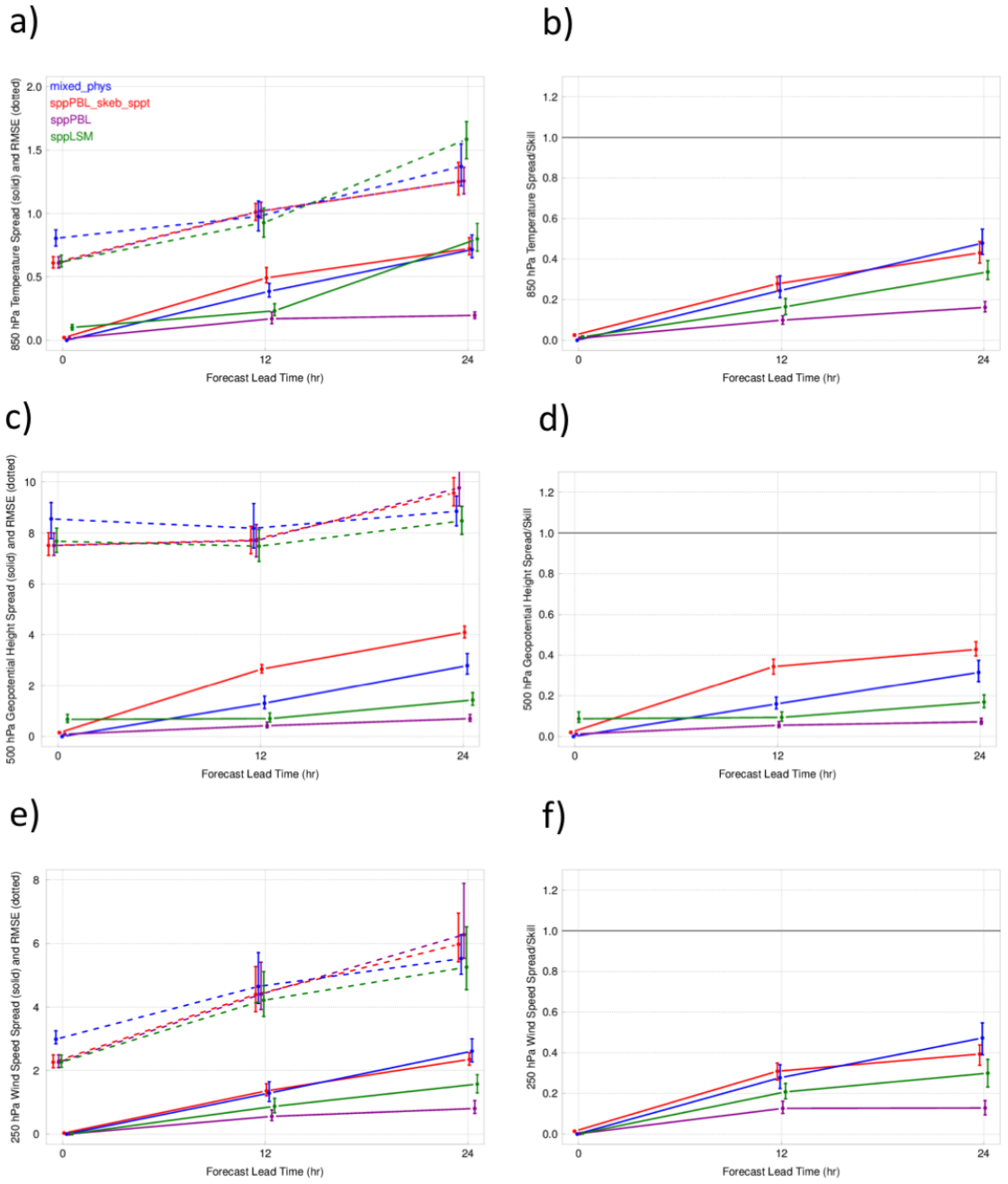

Figure 12. RMSE, spread and spread/error for a) 850-hPa temperature, b) 500-hPa geopotential height and c) 250-hPa wind and spread/error ratio for d) 850-hPa temperature, e) 500-hPa geopotential height and f) 250-hPa wind for 0000 UTC initialization and eastern part of the domain. The 95% confidence intervals are included

Summary

A new stochastic parameter perturbation (SPP) scheme was developed and tested, and ensemble performance using only SPP and in combination with other stochastic methods (SKEB and SPPT) was analyzed. All stochastic methods were assessed against a multi-physics baseline ensemble.

The SPP scheme introduces temporally and spatially varying perturbations to key parameters in the MYNN PBL physics parameterization, as well as to the soil moisture field within the RUC LSM scheme at initialization time. The detailed characteristics of these perturbations (e.g., spatial and temporal de-correlation lengths) were determined through collaboration with physics parameterization experts and by conducting a variety of sensitivity tests. For the HRRR domain, a de-correlation time and length of 6 hours and 150 km, respectively, were found to be appropriate for convective scales.

An eight-member HRRR ensemble consisting of 24-h forecasts was evaluated using a variety of metrics over the 18-27 May 2016 period. All model runs used RAP forecasts as initial conditions for a one-hour pre-forecast that included the latest observations. SKEB and SPPT were employed with the suggested configuration for the horizontal grid spacing used in this study.

Significant findings are summarized below:

Applying perturbations to initialized soil moisture resulted in a generally positive impact on precipitation forecasts. However, these perturbations created an increase in 2-m dew point temperature RMSE and BS. This approach should be investigated in more detail in order to effectively tune the magnitude and spatial scales of the perturbations to improve performance. Adding SKEB and SPPT in combination with SPP had a positive impact on 10-m wind. It also increased spread for all examined upper-level variables.

Our results generally confirm the findings of previous studies performed on coarser grid spacings (e.g., (Jankov et al. 2017), Berner et al. (2011), Hacker et al. (2011a), and Hacker et al. (2011b)), including: (1) parameter perturbations alone within a single physics scheme do not generate sufficient spread to remedy under-dispersion for short-term ensemble forecasts, and (2) a combination of several stochastic schemes outperforms any single scheme.

Regarding the first finding, it was expected that perturbations within a single scheme (in this case, PBL) would not lead to sufficient spread. Also, the general impact was limited to near surface variables. However, parameter perturbations led to an improvement in 10-m wind speed reliability and sharpness (not shown), representing a successful implementation of PBL perturbations designed specifically to improve 10-m wind speed metrics. Also, BS values for selected thresholds and near-surface variables were lower during the day for experiments that included PBL perturbations. Therefore, an improvement in performance for targeted variables can be made when using SPP.

Our research shows that a combination of several stochastic approaches outperformed any one single stochastic method. While this may suggest that a synthesis of different approaches may be best suited to capture model error in its full complexity, it is hypothesized that SPP perturbations applied to one or more parameters in a variety of schemes will lead to sufficient spread. In the future, SPP will be added to the Thompson microphysics scheme, additional parameters in the PBL and LSM schemes, and the corresponding radiation scheme in order to test this hypothesis.

Pending promising results, the use of SPP within many different parameter perturbation schemes may provide a valuable option for sufficient spread within an operational, convective-allowing, single-physics ensemble system.

References

Benjamin, S. G., and Coauthors, 2016: A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Mon. Wea. Rev.*, (144).

Berner, J., K. R. Fossell, S.-Y. Ha, J. P. Hacker, and C. Snyder, 2015: Increasing the skill of probabilistic forecasts: Understanding performance improvements from model-error representations. *Mon. Wea. Rev.*, 143, 1295–1320.

Berner, J., S.-Y. Ha, J. P. Hacker, A. Fournier, and C. Snyder, 2011: Model uncertainty in a mesoscale ensemble prediction system: Stochastic versus multi-physics representations. *Mon. Wea. Rev.*, 139, 1972–1995.

Berner, J., T. Jung, and T. N. Palmer, 2012: Systematic model error: The impact of increased horizontal resolution versus improved stochastic and deterministic parameterizations. *J. Climate*, 25, 4946–4962.

Berner, J., G. Shutts, M. Leutbecher, and T. 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.

Bouttier, F., B. Vie ́, O. Nuissier, and L. Raynaud, 2012: Impact of stochastic physics in a convection-permitting ensemble. *Mon. Wea. Rev.*, 140 (11), 3706–3721.

Bowler, N. E., A. Arribas, S. E. Beare, K. R. Mylne, and G. J. Shutts, 2009: The local ETKF and SKEB: Upgrades to the MOGREPS short-range ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, 135, 767–776.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, 78 (1), 1–3.

Bullock, R., T. Fowler, J. H. Gotway, K. Newman, B. Brown, and T. Jensen, 2017: Model evaluation tools version 6.1 (metv6. 1) user's guide.

Candille, G., and O. Talagrand, 2008: Retracted and replaced: Impact of observational error on the validation of ensemble prediction systems. *Quart. J. Roy. Meteor. Soc.*, 134 (631), 509–521.

Christensen, H. M., I. M. Moroz, and T. N. Palmer, 2015: Stochastic and perturbed parameter representations of model uncertainty in convection parametrisation. *J. Atmos. Sci.*, 72, 2525– 2544.

Eckel, F. A., and C. F. Mass, 2005: Aspects of effective mesoscale, Short-Range Ensemble Forecasting. *Wea. Forecasting*, 20 (3), 328–350.

Ek, M., K. Mitchell, Y. Lin, E. Rogers, P. Grunmann, V. Koren, G. Gayno, and J. Tarpley, 2003: Implementation of noah land surface model advances in the national centers for environmental prediction operational mesoscale eta model. *Journal of Geophysical Research: Atmospheres*, 108 (D22).

Hacker, J. P., C. Snyder, S.-Y. Ha, and M. Pocernich, 2011a: Linear and nonlinear response to parameter variations in a mesoscale model. *Tellus A*, 63, 429–444.

Hacker, J. P., and Coauthors, 2011b: The U.S. Air Force Weather Agency's mesoscale ensemble: Scientific description and performance results. *Tellus A*, 63, 625–641.

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, 129 (3), 550–560.

Hong, S.-Y., Y. Noh, and J. Dudhia, 2006: A new vertical diffusion package with an explicit treatment of entrainment processes. *Mon. Wea. Rev.*, 134 (9), 2318–2341.

Jankov, I., and Coauthors, 2017: A performance comparison between multiphysics and stochastic approaches within a north american rap ensemble. *Monthly Weather Review*, 145 (4), 1161– 1179.

Murphy, J., D. Sexton, D. Barnett, G. Jones, M. Webb, M. Collins, and D. Stainforth, 2004: Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*, 430, 768–772.

Nakanishi, M., and H. Niino, 2004: An improved Mellor-Yamada Level-3 model with condensation physics: Its design and verification. *Bound.-Layer Meteor.*, 112, 1–31.

Nakanishi, M., and H. Niino, 2006: An improved Mellor-Yamada Level-3 model: Its numerical stability and application to a regional prediction of advection fog. *Bound.-Layer Meteor.*, 119, 397–407.

Palmer, T. N., 2001: A nonlinear dynamical perspective on model error: A proposal for non-local stochastic-dynamic parameterization in weather and climate prediction. *Quart. J. Roy. Meteor. Soc.*, 127, 279–304.

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/.

Pleim, J. E., 2007: A combined local and nonlocal closure model for the atmospheric boundary layer. part i: Model description and testing. *Journal of Applied Meteorology and Climatology*, 46 (9), 1383–1395.

Reynolds, C. A., J. G. McLay, J. S. Goerss, E. A. Serra, D. Hodyss, and C. R. Sampson, 2011: Impact of resolution and design on the US Navy global ensemble performance in the tropics. *Mon. Wea. Rev.*, 139 (7).

Sanchez, C., K. D. Williams, and M. Collins, 2015: Improved stochastic physics schemes for global weather and climate models. *Quart. J. Roy. Meteor. Soc.*

Schaefer, J. T., 1990: The critical success index as an indicator of warning skill. *Wea. Forecasting*, 5 (4), 570–575.

Skamarock, W. C., and Coauthors, 2008: A Description of the Advanced Research WRF version 3. Tech. rep., NCAR Tech. Note, NCAR/TN-475+STR, 113pp., 113 pp.

Smirnova, T. G., J. M. Brown, S. G. Benjamin, and J. S. Kenyon, 2016: Modifications to the Rapid Update Cycle Land Surface Model (RUC LSM) available in the Weather Research and Forecast (WRF) model. *Mon. Wea. Rev.*, (2015).

Zhang, J., and Coauthors, 2016: Multi-radar multi-sensor (mrms) quantitative precipitation estimation: Initial operating capabilities. *Bulletin of the American Meteorological Society*, 97 (4), 621–638.