Stochastically Perturbed Parameterizations in an HRRR-Based Ensemble

ISIDORA JANKOV AND JEFFREY BECK

Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, and National Oceanic and Atmospheric Administration/Oceanic and Atmospheric Research/Earth System Research Laboratory/ Global Systems Division, and Developmental Testbed Center, Boulder, Colorado

JAMIE WOLFF AND MICHELLE HARROLD

Research Application Laboratory, National Center for Atmospheric Research, and Developmental Testbed Center, Boulder, Colorado

JOSEPH B. OLSON AND TATIANA SMIRNOVA

Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, and National Oceanic and Atmospheric Administration/Oceanic and Atmospheric Research/ Earth System Research Laboratory/Global Systems Division, Boulder, Colorado

CURTIS ALEXANDER

National Oceanic and Atmospheric Administration/Oceanic and Atmospheric Research/ Earth System Research Laboratory/Global Systems Division, Boulder, Colorado

JUDITH BERNER

Mesoscale and Microscale Meteorology Laboratory, National Center for Atmospheric Research, Boulder, Colorado

(Manuscript received 16 March 2018, in final form 22 October 2018)

ABSTRACT

A stochastically perturbed parameterization (SPP) approach that spatially and temporally perturbs parameters and variables in the Mellor–Yamada–Nakanishi–Niino planetary boundary layer scheme (PBL) and introduces initialization perturbations to soil moisture in the Rapid Update Cycle land surface model was developed within the High-Resolution Rapid Refresh convection-allowing ensemble. This work is a follow-up study to a work performed using the Rapid Refresh (RAP)-based ensemble. In the present study, the SPP approach was used to target the performance of precipitation and low-level variables (e.g., 2-m temperature and dewpoint, and 10-m wind). The stochastic kinetic energy backscatter scheme and the stochastic perturbation of physics tendencies scheme were combined with the SPP approach and applied to the PBL to target upper-level variable performance (e.g., improved skill and reliability). The three stochastic experiments (SPP applied to PBL only, SPP applied to PBL combined with SKEB and SPPT, and stochastically perturbed soil moisture initial conditions) were compared to a mixed-physics ensemble. The results showed a positive impact from initial condition soil moisture perturbations on precipitation forecasts; however, it resulted in an increase in 2-m dewpoint RMSE. The experiment with perturbed parameters within the PBL showed an improvement in low-level wind forecasts for some verification metrics. The experiment that combined the three stochastic approaches together exhibited improved RMSE and spread for upper-level variables. Our study demonstrated that, by using the SPP approach, forecasts of specific variables can be improved. Also, the results showed that using a single-physics suite ensemble with stochastic methods is potentially an attractive alternative to using multiphysics for convection allowing ensembles.

Corresponding author: Isidora Jankov, isidora.jankov@noaa.gov

DOI: 10.1175/MWR-D-18-0092.1

1. 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 multiphysics ensemble. Use of a combination of different physics schemes usually leads to a large diversity among the ensemble members, resulting in sufficient spread and improved forecast skill (e.g., Hacker et al. 2011b; Berner et al. 2011, 2015). Even though ensembles designed in this way are often characterized by good performance, there are both practical and theoretical deficiencies associated with them. Practically speaking, multiple physics parameterizations representing a single physics process (e.g., convection or boundary layer processes or microphysical processes) have to be developed and maintained in parallel, which requires extensive resources. Furthermore, for the purpose of statistical postprocessing, securing equally distributed and independent random variables is a necessity. This requirement cannot be satisfied when using the multiphysics approach. The postprocessing of a multiphysics ensemble is further complicated by the fact that each ensemble member has a different mean error and climatology, which is one possible reason that these ensembles have larger spread (e.g., Eckel and Mass 2005; Berner et al. 2015).

An alternative way to introduce ensemble spread, as discussed in Jankov et al. (2017), is to perturb physical parameterizations stochastically (Palmer 2001). The main advantage of this approach is that it results in statistically consistent ensemble distributions (e.g., Bowler et al. 2008; Berner et al. 2009; Bowler et al. 2009; Sanchez et al. 2016). The two stochastic schemes that are most commonly used and have been implemented in a variety of operational models are the stochastic kinetic energy backscatter (SKEB) and stochastic perturbations of physics tendencies (SPPT) schemes. Both were developed to better represent subgrid-scale processes. In the case of SKEB, the model uncertainty associated with subgrid-scale processes is addressed by randomly perturbing streamfunction and potential temperature tendencies (Berner et al. 2009, 2012, 2015). SPPT (Buizza et al. 1999; Palmer et al. 2009) takes subgrid-scale processes into account by perturbing the total physics tendencies such as temperature, humidity, and wind (Bouttier et al. 2012; Berner et al. 2015). It was shown that the inclusion of these stochastic schemes within an ECMWF ensemble improved the probabilistic skill by increasing reliability and reducing the ensemble mean error (Palmer et al. 2009 and Leutbecher et al. 2017).

SPPT and SKEBS account for model error in a bulk sense, where the accumulated process-level model errors are represented by a single model error term (Berner et al. 2017). To address model uncertainty in the parameters within the respective parameterization schemes, the stochastically perturbed parameterization (SPP) approach was developed (Jankov et al. 2017; Ollinaho et al. 2017). It can be applied by having the parameter and/or variable of choice unchanged throughout the integration (e.g., Hacker et al. 2011a) or by varying randomly in time and space.

Previous studies have shown that the SPP approach usually outperforms unperturbed ensembles but still does not create sufficient spread (Hacker et al. 2011b; Reynolds et al. 2011; Berner et al. 2015; Christensen et al. 2015). Comparison of the SPPT and the SPP approaches within the ECMWF ensemble forecasts demonstrated more skillful 2-m temperature associated with SPP and for shorter-range forecasts (Ollinaho et al. 2017), although this result might be influenced by the fact that SPPT in the ECMWF implementation is tapered to zero near the surface. The opposite is true for variables in the free atmosphere and longer lead times. Bouttier et al. 2012 tested SPPT within a short-range, convection-permitting ensemble prediction system. It was found that by employing SPPT, the probabilistic performance was significantly improved, particularly in terms of reliability and the spread–skill ratio. The work also pointed to the weakness of SPPT in terms of the lack of explicit perturbations for low levels and at the surface, where large model errors occur. McCabe et al. 2016 showed that by random perturbations of several parameters within microphysics and PBL schemes in a convection-permitting ensemble, an improvement in visibility and surface temperature was obtained. Also, a modest increase of spread for surface variables was detected.

SPPT implementation in the Weather Research and Forecasting (WRF) Model (Berner et al. 2015) does not use the tapering to zero near the surface and results in a bigger impact on surface variables (Romine et al. 2014). However, this beneficial impact was accompanied by increased bias. WRF adds the increments due to the microphysics after the state is updated with all other physical tendencies, possibly leading to a double counting of the microphysics perturbations in the Romine et al. (2014) implementation. Guided by these results, the microphysics tendencies were no longer explicitly perturbed (Berner et al. 2015).

To focus on the uncertainty associated with parameters as well as the initial state, the present study employs the SPP approach alone, and in combination with SKEB and SPPT. While Jankov et al. (2017) performed experiments with a Rapid Refresh (RAP)-based ensemble with parameterized convection, this study focuses on evaluating the impact of stochastic perturbations on high-resolution, convection-permitting ensemble performance. To this extent, perturbations to key parameters in the High-Resolution Rapid Refresh (HRRR) PBL scheme in addition to perturbations to the initial soil moisture state were evaluated.

2. Experiment design

a. Model

The operational HRRR configuration, which has been running at the National Centers for Environmental Prediction (NCEP), was used as a basis for all the experiments in this study. Simulations were performed over the operational HRRR continental United States (CONUS) domain (Fig. 1) with 3-km grid spacing. HRRR's horizontal integration grid is the staggered Arakawa C grid (Arakawa and Lamb 1977). The HRRR system uses the Advanced Research version of WRF (WRF-ARW) dynamic core (Skamarock et al. 2008). The physics suite includes the Mellor–Yamada–Nakanishi–Niino (MYNN; Nakanishi and Niino 2004, 2006) planetary boundary layer (PBL) parameterization, the Rapid Update Cycle (RUC; Smirnova et al. 2016) land surface model (LSM) parameterization, Thompson microphysics scheme (Thompson et al. 2008), and the Rapid Radiative Transfer Model for general circulation models (RRTMG; Mlawer et al. 1997).

HRRR lateral boundary and initial conditions are obtained by downscaling RAP (Benjamin et al. 2016) followed by an hour-long HRRR preforecast. The HRRR hourly initialization process using RAP and the preforecast is illustrated in Fig. 2. During this preforecast, 15-min reflectivity data are assimilated and used to specify latent heating rates (Benjamin et al. 2016). For observed reflectivity ≤0 dBZ, the heating rate is set to zero to suppress spurious model precipitation. For observed reflectivities between 0 and 28 dBZ the model microphysics heating rate is preserved, and for observed reflectivities ≥28 dBZ, the heating rate is positive to promote convective development. After the preforecast and radar data assimilation, Gridpoint Statistical Interpolation (GSI) hybrid data assimilation as well as GSI hydrometeor analysis are performed (Benjamin et al. 2016). In the present study, this process was applied hourly on each ensemble member for each of the experiments. The 24-h-long simulations were performed twice a day, at 0000 and 1200 UTC. The experimental dataset consists of 8 members and 10 spring season days starting on 18 May and ending on 27 May 2016. Even though the focus of this study is on uncertainties associated with model error, by following the operational process, a form of initial condition perturbations was also introduced. This resulted in some differences in statistics between the experiments at the initial times. The main motivation for employing this initialization approach was to mimic the operational systems as much as possible to facilitate evidence-based decision for transitions to operations. Recently, development of the HRRR Ensemble (HRRRE) system, which includes an ensemble-based data assimilation system, started. Based on this study, stochastically perturbed parameterizations are part of the HRRRE.

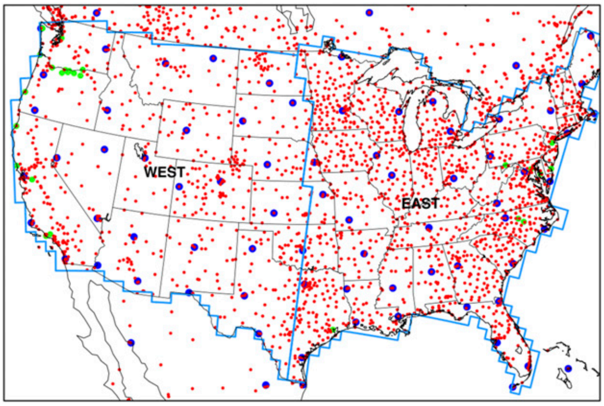

FIG. 1. Verification domain with the division between western and eastern regions presented with blue lines and surface, upper-air, and profiler observations presented with red, blue, and green dots, respectively.

Regarding the data sample, the authors recognize that the number of simulations and number of ensemble members are somewhat limited. The decision regarding experiment length was based on strict computational and storage resources available. Having these resource limitations in mind, the authors selected an active convective period across the CONUS during the 2016 Hazardous Weather Testbed (https://hwt.nssl.noaa.gov/), providing results that we believe are representative of general ensemble performance for these types of events. To alleviate small sample size issues, an analysis of simulations initialized at both 0000 and 1200 UTC is included.

The multiphysics ensemble (mixed_phys), which represents the control experiment, used different physics parameterizations for the PBL and LSM schemes (Table 1). The different PBL schemes included the Mellor–Yamada–Janjić, 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 multiphysics ensemble contained a combination of the four PBL and two LSM schemes.

All eight members of the stochastic ensemble experiments used the same physics parameterizations as the operational HRRR (Table 1). One of the stochastic experiments consisted of perturbing soil moisture

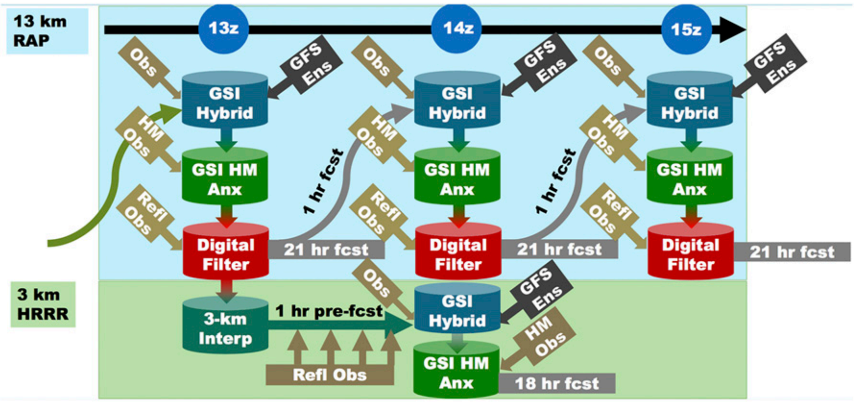

FIG. 2. HRRR initialization with RAP.

(sppLSM_IC) values at the initial time only, one perturbed multiple parameters within the MYNN PBL (sppPBL) throughout the forecast, and the final experiment combined the previous PBL perturbations with SKEB and SPPT (sppPBL_skeb_sppt). Details on SKEB and SPPT implementations in the WRF-ARW Model are provided in Berner et al. (2011) and Berner et al. (2015). Table 2 provides a summary of the experiments.

The SPP approach used here was adapted from the previously mentioned research that utilized the RAP-based ensemble system. A detailed explanation of the method and creation of the SPP perturbations can be found in Jankov et al. (2017). In summary, the spatially and temporally correlated pattern is fully determined by three namelist parameters: gridpoint standard deviation (gridpt_stddev_rand_pert, magnitude of the perturbations), length scale (length scale_rand_pert), and decorrelation time (time scale_rand_pert). Additionally, since the Gaussian distribution is unbounded, the random numbers are constrained to stay within a range, with a threshold expressed in terms of standard deviation (stddev_cutoff_rand_pert).

While initial parameter pattern values (e.g., spatial and temporal decorrelations) 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 decorrelation lengths, which were chosen based on typical spatial and temporal advective scales: 150 km and 6 h, 300 km and 12 h, and 600 km and 24 h. Experiments with a spatial and temporal decorrelation length of 150 km and 6 h, 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 soil moisture perturbations at the initial time.

As described in detail in Jankov et al. (2017), the stochastic perturbations at each grid point draw from a univariate Gaussian distribution centered on the value of the deterministic parameter. The spatial correlations guarantee that the perturbations to nearby grid points have, on average, the same sign.

The spatial correlations guarantee that the perturbations to nearby grid points have on average the same sign. Table 3 provides a summary of the targeted parameters and variables in the MYNN PBL scheme and corresponding perturbation amplitudes. Within the PBL parameterization scheme, there are two dynamically evolving parameters (Czil, Prlimit) and two variable

TABLE 1. HRRR stochastic and mixed-physics members.

Physics suite	Microphysics	Radiation (SW/LW)	Surface layer	Land surface model	PBL
HRRR (stochastic members)	Thompson aerosol aware	RRTMG	MYNN	RUC	MYNN
HRRR (mixed physics) mem0	Thompson aerosol aware	RRTMG	Revised MM5	Noah	YSU
mem1	Thompson aerosol aware	RRTMG	MYJ	Noah	MYJ
mem2	Thompson aerosol aware	RRTMG	MYNN	Noah	MYNN
mem3	Thompson aerosol aware	RRTMG	Revised MM	Noah	ACM2
mem4	Thompson aerosol aware	RRTMG	MYNN	RUC	YSU
mem5	Thompson aerosol aware	RRTMG	MYNN	RUC	MYJ
mem6	Thompson aerosol aware	RRTMG	MYNN	RUC	MYNN
mem7	Thompson aerosol aware	RRTMG	MYNN	RUC	ACM2

TABLE 2. Summary of experiments.

Expt name	Description
mixed_phys	Control mixed physics ensemble that combines different PBL and LSM schemes
sppLSM_IC	Soil moisture state perturbed at the initial time
sppPBL	Set of parameters and variables within PBL scheme stochastically perturbed throughout the simulation period
sppPBL_skeb_sppt	SKEB and SPPT combined with the PBL perturbations and applied throughout the simulation period

TABLE 3. Summary of perturbed parameters and variables in PBL MYNN and RUC LSM.

	Name	Magnitude
Perturbed parameter in MYNN PBL scheme		
Turbulent mixing length	el	30%
Subgrid cloud fraction	cldfra_bl	20%
Thermal and moisture roughness length	CZIL	30%
Prandtl number's limit set to 2.5	Prlimit	1
Perturbed parameter in RUC LSM scheme		
Soil moisture	SMOIS	20%

perturbations (el and cldfra_bl). Variations in Czil can greatly impact the size of the thermal roughness lengths, which determines the magnitude of the surface exchange coefficients for heat. The Prandtl number, Pr, is defined as Km/Kh, where Km and Kh are the eddy viscosity and eddy diffusivity, respectively. The Prandtl number limit, Prlimit, is allowed to vary between 1 and 5, which limits the amount of momentum mixing in the stable boundary layer relative to the mixing of heat. Since the value of Pr is considered to be more well known in unstable conditions, when $\mathrm{Pr} < 1$, this limit will not impact the turbulent mixing in unstable conditions. Mixing lengths are diagnosed at every model time step and are a function of the ambient stability, TKE, and surface stability parameter (z/L), where z is the height above the ground and L is the Obukhov length. The diagnosed mixing lengths are important for regulating the TKE and, therefore, the strength of the turbulent mixing in all conditions. And finally, subgrid cloud fractions, which are also diagnosed at every time step, are important for the interaction with both shortwave and longwave radiation and can greatly influence the surface energy balance.

In the PBL scheme, the turbulent mixing length and subgrid cloud fraction were directly perturbed, and thermal and moisture roughness lengths were indirectly perturbed through changing the Zilitinkevich constant, Czil. With PBL development, the smaller shallow cumulus become larger and deeper, but the total cloud fraction is not necessarily changed. On the other hand, in deep, dry boundary layers when the mixing lengths are the largest, cloud cover is small. Therefore, in fully developed PBLs, there is a negative correlation between subgrid-scale clouds and mixing length. Thus, negative correlation, which reduces subgrid-scale clouds when mixing lengths become larger, was implemented. The result of the correlation is more solar radiation reaching the surface, leading toward higher surface temperatures and larger surface heat fluxes that then result in an increase of mixing lengths, representing a positive feedback process. Increased turbulent mixing can lead to increased entrainment at the top of the PBL, which then dries the PBL and consequently further reduces the subgrid-scale clouds (Stull 2012). Czil was perturbed up to half (twice) its original value. Since Czil is in the negative exponent, decreasing (increasing) Czil on the order of half (double) its value results in a perturbation of thermal roughness length of 5%–10%.

The thermal roughness length z_t is defined as follows:

$$z_t = z_0 e^{(-\kappa \mathrm{Czil} \sqrt{\mathrm{Re}})}, \qquad (1)$$

where z_0 is aerodynamic roughness length, κ is the Von Kármán constant, and Re is the Reynolds number.

To evaluate the mechanisms by which the stochastic perturbations in the PBL scheme led to expected changes in the model state, changes in the SPP pattern and related model variables throughout the simulation period were examined at a select grid point for three ensemble members drawn from the sppPBL experiment (Fig. 3). The first member (red) was unperturbed, the second member (blue) was characterized by positive perturbation throughout much of the simulation time, while the third member (green) was characterized with negative perturbations for most of the simulation time. It can be seen that positive perturbations led to increased values of PBL height, heat flux, shortwave-down radiation, and turbulent kinetic energy (TKE), as expected. The opposite was the case for the member characterized by negative SPP pattern. The SPP implementation within the PBL scheme created physically consistent changes in the model state.

In the RAP operational system, which is used for the HRRR initialization, soil state is defined by hourly cycling and defining a soil–air forecast–error relationship. Analyses of near-surface temperature and moisture are used to make small adjustments to soil temperature and soil moisture (Benjamin et al. 2016; Smirnova et al. 2016).

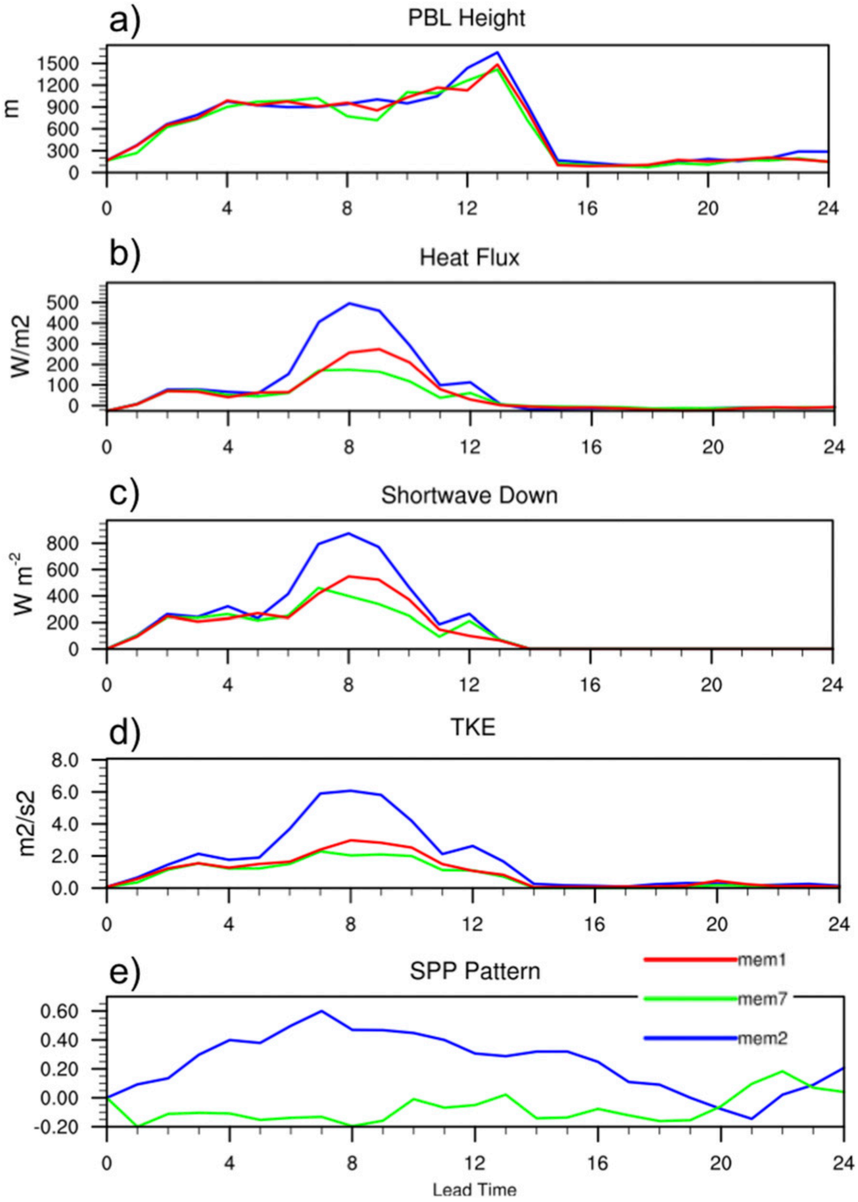

FIG. 3. Realizations of time series of (a) PBL height, (b) heat flux, (c) shortwave radiation, (d) TKE, and (e) perturbation pattern at a select grid point in perturbed and unperturbed simulations. The unperturbed simulation is presented in red and perturbed simulations characterized by persistent positive and negative perturbations are presented in blue and green, respectively.

This procedure helps retain the soil–air temperature difference between the forecast background and the analyzed state. The parameters used in this procedure were experimentally estimated for both summer and winter seasons. Given this, and the fact that there are not sufficient direct observations, soil moisture is not assimilated in RAP. However, given its direct impact on subsequent boundary layer humidity and convective initiation, soil moisture has been found to strongly influence forecast accuracy. Previous studies have shown the importance of variability in soil moisture or LSM parameterization perturbations for ensemble spread (Sutton et al. 2006; Duda et al. 2017) as well as ensemble precipitation forecasting (Aligo et al. 2007). Therefore, initial conditions of soil moisture were perturbed within the RUC LSM parameterization scheme for one of the ensemble experiments, using the same spatial and temporal decorrelation lengths as the PBL simulations.

For wider use of the SPP, specifics on how to apply it within WRF can be found in the WRF User's guide.

An example of namelist settings for a specific experiment can be found in Jankov et al. (2017). Finally, all SPP changes introduced to the code for the purpose of this work have been committed to the WRF repository and are available for public use through recent community releases.

In the present study, for assessing the impact of SPP, the focus was on precipitation and surface variables (2-m temperature, 2-m dewpoint temperature, and 10-m wind). For overall assessment, and specifically for evaluating the impact of SKEB and SPPT, additional, somewhat limited, upper-air analysis was performed. The analysis included spread/error evaluation for 500-hPa geopotential height, 850-hPa temperature, and 250-hPa wind.

b. *Observation data*

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 base 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 regridded to the 3-km integration domain using budget interpolation to allow for direct grid-to-grid comparisons. Precipitation was verified over 3-h and daily accumulations.

For conventional surface and upper-air point observations, RAP observation files in Binary Universal Form for Representation of Meteorological Data (BUFR) format were used. Verification of standard meteorological fields (temperature, dewpoint, and wind) was performed hourly for surface variables and for times valid at 0000 and 1200 UTC for upper-air variables. When compared to model output, bilinear interpolation was performed.

c. *Evaluation metrics*

Precipitation performance was assessed using a number of verification metrics for deterministic and probabilistic forecast evaluation, including rank histograms, frequency bias, fractions skill score (FSS), and reliability.

The rank histogram is a diagnostic tool that facilitates assessing the spread of ensemble forecasts, based on the assumption that the probability of occurrence of an observation in a set of forecast bins should be equally likely (Hamill 2001). These bins are determined by ranking ensemble member forecasts from lowest to highest value; thus, for an ensemble with n members, the corresponding rank histogram will have $n + 1$ bins. The rank histogram is produced by plotting the frequency of occurrence of observations in each bin. Flat rank histograms indicate an ensemble with ideal spread, while a u-shaped histogram indicates underdispersion. An asymmetric rank histogram indicates that an ensemble has bias. In terms of precipitation, higher values in the first bin on the left-hand side indicate bias in the ensemble toward heavier precipitation amounts and vice versa for the last bin on the right-hand side.

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 1, where values higher (lower) indicate that the model overpredicted (underpredicted) the exceedance of a given threshold. In the present study, frequency bias was analyzed for the two initializations and two precipitation thresholds (0.254 and 12.7 mm) as an aggregate over all members of each experiment (Fig. 5). The low precipitation threshold was selected to evaluate the experiments performance for very light precipitation, while the higher precipitation threshold was selected as a good representative threshold for the period over which the simulations were performed. Confidence intervals at the 95% level for each experiment were applied.

FSS (Roberts and Lean 2008; Schwartz et al. 2010) was evaluated for the two precipitation thresholds (0.254 and 12.7 mm) and two neighborhood sizes (9 and 45 km). Calculation of FSS includes the following steps: (i) convert all forecast (F) and observed (O) fields into binary fields for each threshold of interest, (ii) generate fractions within a square of length n that have exceeded the threshold at each grid point across the full verification domain (N_x, N_y), and (iii) compute the mean squared error (MSE) relative to a low-skill reference forecast (MSEref), which equates to the largest possible MSE that would be found if no overlap between forecast and observed events occurred. FSS for a neighborhood of length n is given by

$$\mathrm{FSS} = 1 - \frac{\mathrm{MSE}_{(n)}}{\mathrm{MSE}_{(n)\mathrm{ref}}}, \qquad (2)$$

where

$$\mathrm{MSE}_{(n)} = \frac{1}{N_x N_y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} [O_{(n)ij} - F_{(n)ij}]^2 \qquad (3)$$

and

$$\mathrm{MSE}_{(n)\mathrm{ref}} = \frac{1}{N_x N_y} \left[\sum_{i=1}^{N_x} \sum_{j=1}^{N_y} O^2_{(n)ij} + \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} F^2_{(n)ij} \right]. \qquad (4)$$

The FSS ranges from 0 to 1. A score of 1 is attained for a perfect forecast and a score of 0 indicates no skill. As the neighborhood expands and the number of grid boxes in the neighborhood increases, the FSS improves as the observed and model probability fields are generally smoothed and their overlap tends to increase.

A reliability diagram is a graphical method for assessing reliability, resolution, and sharpness of a probabilistic forecast. It includes observed frequency plotted against forecast probability of an event. Reliability is measured by proximity to the diagonal and resolution is defined as a variation from the horizontal line that represents sample base rate (Fig. 7). To obtain useful results reliability diagrams require a large dataset.

For evaluation of surface and upper-air variables, an emphasis was on root-mean-square error (RMSE), spread, bias, and reliability. Aggregate 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 the simple arithmetic average of the members. RMSE, spread, and the ratio of the two concisely summarize 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 1. In addition, bias, or more precisely mean error (ME), was computed as a function of lead times for the three surface variables.

3. Results

a. *Precipitation verification*

All simulations were performed over the CONUS domain (outer black box in Fig. 1). Verification of “raw” model output (there was no postprocessing 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 and 1200 UTC. Trends in results for CONUS-East and CONUS-West for the two initializations were very similar. Given this, results discussed here are restricted to the CONUS-East domain for both 0000 and 1200 UTC initializations. 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 4 shows rank histograms of 3-hourly precipitation accumulations aggregated between the 0- and 24-h lead times for each experiment and 0000 UTC initializations. The rank histogram for all experiments generally indicates lower relative frequency values for the middle bins and higher values for the outermost bins, thus, indicating underdispersion. Specific features differ somewhat among the experiments, however. The mixed_phys experiment indicated a bias toward heavier precipitation. Similarly, the sppPBL and sppPBL_skeb_sppt experiments exhibited bias, but were also characterized by the presence of underdispersion. The sppLSM_IC ensemble was characterized by similar values of relative frequencies in the outermost bins indicating a tendency to be underdispersed, rather than biased. Nearly identical rank histogram plots were observed for runs initialized at 1200 UTC (not shown).

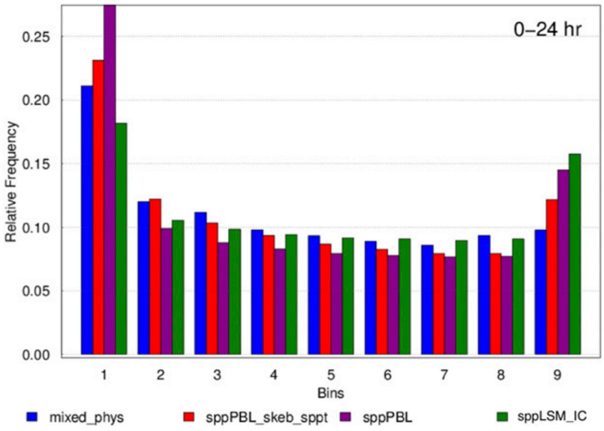

FIG. 4. Rank histograms for all experiments for 0000 UTC initializations over the eastern CONUS domain, 3-hourly accumulated precipitation aggregated between the 0- and 24-h lead times.

Frequency bias for the runs initialized at 0000 UTC, 3-h accumulated precipitation greater than 0.254 mm (Fig. 5a) 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 sppLSM_IC experiments were frequently significantly different from the sppPBL and sppPBL_skeb_sppt experiments. The sppLSM_IC ensemble was the only experiment characterized with frequency bias values lower than one for most of the lead times while the other three experiments typically had frequency bias values larger than 1. 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 sppPBL ensembles had significantly higher-frequency bias values for this threshold and the first 15 h of the forecast compared to the other two experiments, with an improved frequency bias later in the period. Similar behavior was observed for the 1200 UTC initializations (not shown).

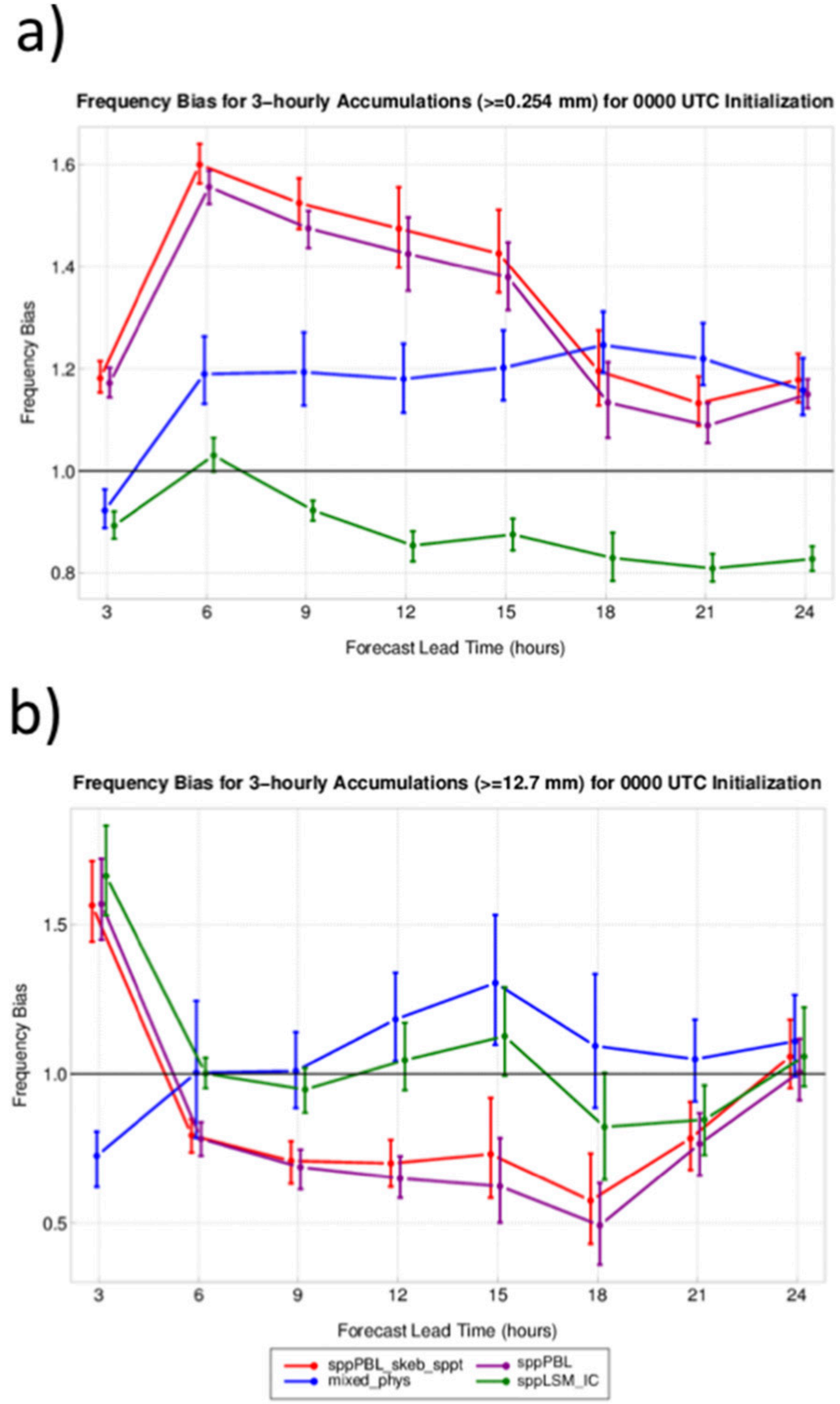

FIG. 5. Frequency bias aggregated over all members for each experiment as a function of lead time for 0000 UTC initializations over the eastern CONUS domain for a precipitation accumulation threshold of (a) >0.254 and (b) >12.7 mm. Note the *y*-axis range is different between (a) and (b).

The same type of analysis, except for 12.7-mm precipitation threshold, showed values close to 1 (confidence intervals frequently encompassed 1) for most of the lead times for the mixed_phys and sppLSM_IC experiments (Fig. 5b) for the 0000 UTC initializations. The sppPBL_skeb_sppt and sppPBL again had similar behavior among themselves, and often exhibited a significant low bias for this threshold. Similar trends were observed for the 1200 UTC initializations (not shown).

Figure 6 shows FSS for the 0000 UTC initializations, the two precipitation thresholds (0.254 and 12.7 mm), the two neighborhoods (9 and 45 km), and each experiment as a function of lead time. Generally, for the two neighborhoods, and all experiments, FSS decreased with increasing lead times. For the light precipitation threshold (Fig. 6a), a clear separation in FSS values for the two neighborhoods was indicated and the larger neighborhood was characterized by higher FSS values. At the shorter lead times, the mixed_phys experiment had somewhat lower skill compared to sppPBL_skeb_sppt and sppPBL, and significantly lower skill compared to sppLSM_IC. Interestingly, for the heavier precipitation threshold (Fig. 6b), the differences in FSS for different neighborhood sizes were not as pronounced as it was the case for the low precipitation threshold. For the occasional lead times where a statistically significant difference was noted, sppLSM_IC experiment was favored as the best. For both thresholds, similar trends were observed for the 1200 UTC initializations (not shown).

Reliability diagrams and corresponding event histograms were created for 24-h accumulations, for the 0.254- and 12.7-mm precipitation thresholds, for 0000 UTC initializations, and for each experiment (Fig. 7).

The observed frequency (i.e., the sample base rate) for the lower precipitation threshold and 0000 UTC initialization (Fig. 7a) is about 35% of the grid locations over the 10-day period, making it a somewhat common event. Reliability diagrams measure the calibration of a probability forecast. All of the ensembles at this lower threshold showed similar trend, with a generally higher observed proportion of events when the ensemble probability values were higher. Thus, all of the ensembles had some ability to discriminate these precipitation events from nonevents. Similarly, event histograms show that all of the experiments were generally sharp compared to the base rate. In other words, all of the experiments had a tendency to forecast extreme values (probabilities of 0% and close to 100%, as opposed to values around the mean). However, all of the ensembles, except the sppLSM_IC, overestimated the probability of the precipitation events (i.e., they fall below the solid gray one-to-one line). The sppLSM_IC was generally characterized by higher reliability (i.e., closer to the one-to-one line) compared to the other experiments for all forecast frequencies. Consequently, sppLSM_IC resulted in a better resolution (variance from the base rate line) compared to the other experiments. Very similar results were observed for the 1200 UTC initializations (not shown).

Reliability diagrams for 12.7-mm threshold and 0000 UTC initializations (Fig. 7b) showed overconfidence for all experiments and for all forecast frequencies. However, all of the experiments were characterized with reliability between the no skill and perfect reliability line, with exception for the lowest forecast frequency. The three stochastic experiments performed similarly to mixed_phys experiment. In terms of sharpness, all experiments failed to predict high-probability events, which was accompanied with lower reliability for those frequencies. For the same precipitation threshold but 1200 UTC initializations, the results were comparable to the 0000 UTC initialization results (not shown).

To illustrate differences in probabilities obtained for each of the ensembles, a case was selected and CONUS-wide probabilities of 24-h precipitation accumulations exceeding 25.4 mm were evaluated for each experiment initialized at 0000 UTC 24 May 2016 (Figs. 8a–d). Total precipitation accumulations for this period using MRMS measurements are also shown in Fig. 8e. 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, southeast, terminating in northern Arkansas, and southeast Missouri at the end of the period. The mixed_phys (Fig. 8a) and sppLSM_IC (Fig. 8d) experiments appear to capture the potential for this observed evolution, while only one member of the sppPBL_skeb_sppt (Fig. 8b) and no members of sppPBL (Fig. 8c) experiments produced precipitation >25.4 mm anywhere in Missouri.

When compared to the stochastic experiments, the mixed_phys (Fig. 8a) experiment generally produced probabilities covering a larger areal extent. The sppPBL (Fig. 8c) and sppPBL_skeb_sppt (Fig. 8b) experiments produced more focused probabilities, generally limited to northern Kansas, and were shifted more northwestward than probabilities from the mixed_phys experiment. The northwestward shift in the two experiments correctly highlighted the precipitation area in eastern Nebraska. In addition, probabilities over northern Kansas were somewhat higher for the two experiments as compared to mixed_phys because of smaller spread.

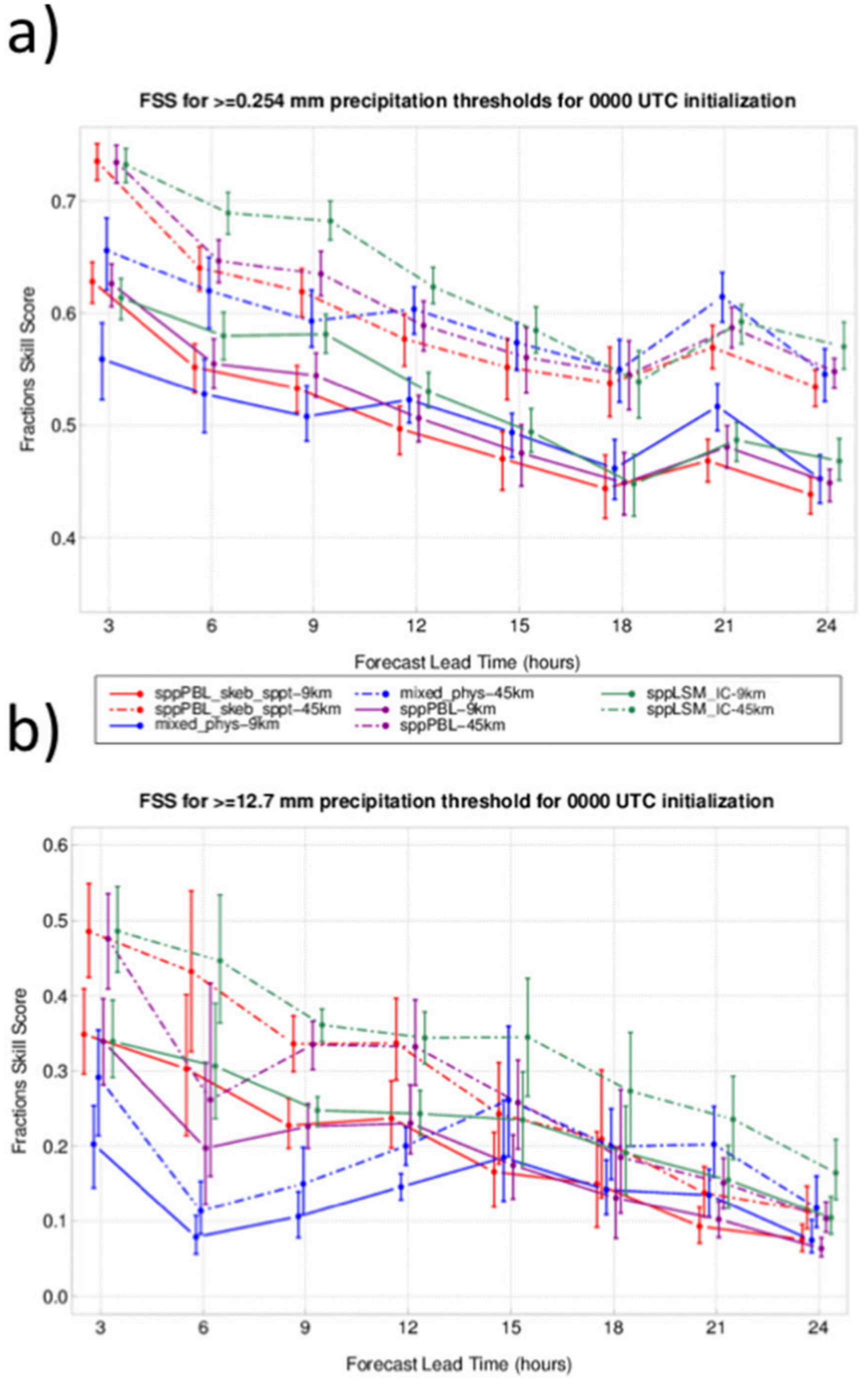

FIG. 6. Fractions skill score (FSS) for 9-km (solid) and 45-km (dot–dashed) neighborhood sizes for (a) >0.254- and (b) >12.7-mm precipitation thresholds.

Low probabilities extended farther 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_IC solution was more similar to mixed_phys with areas of high probabilities over the Missouri–Arkansas border. In the case of sppLSM_IC, 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

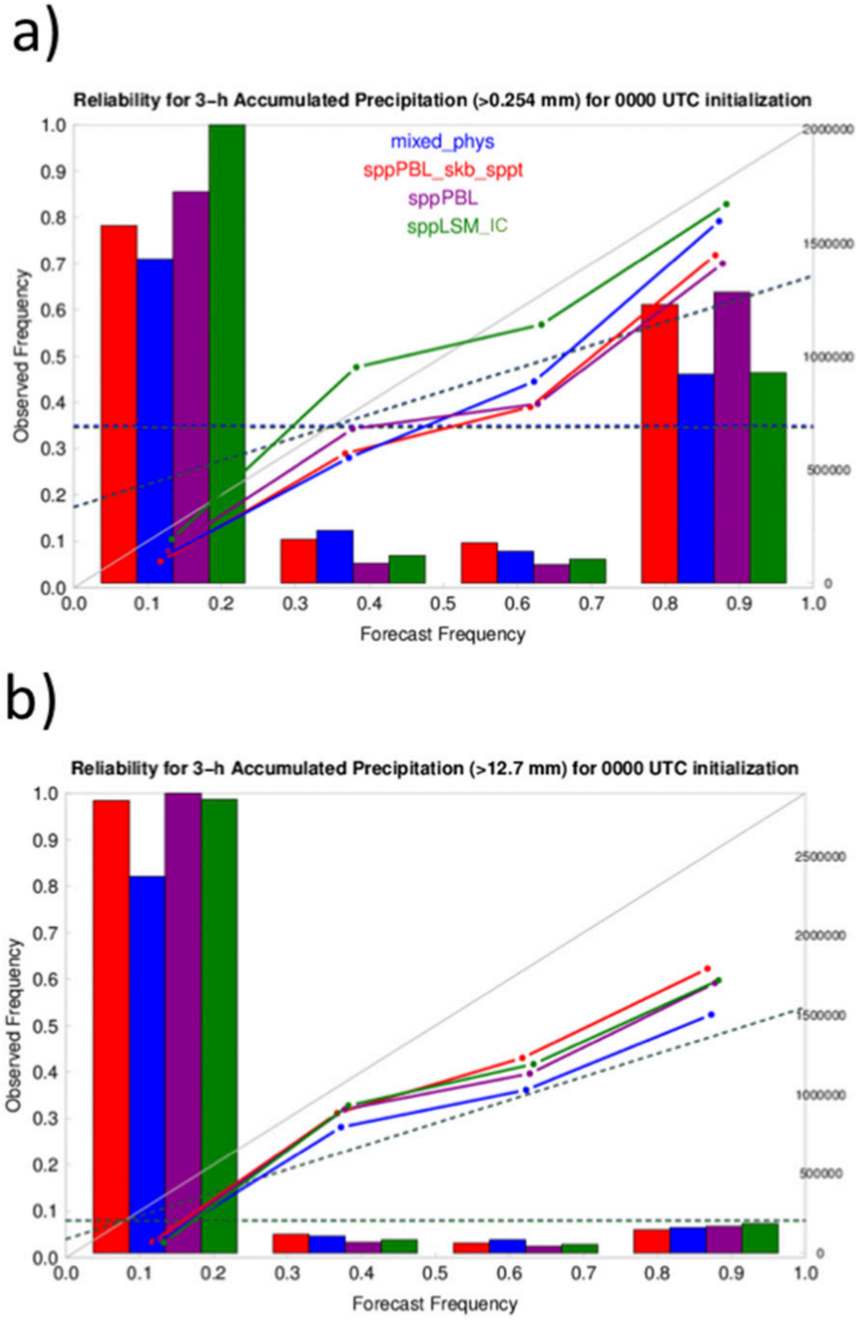

FIG. 7. Reliability diagram for 0000 UTC initializations over the eastern part of the domain and for (a) >0.254- and (b) >12.77-mm precipitation thresholds. The horizontal dotted line represents the sample base rate, the diagonal dotted line represents no skill, while the solid gray diagonal line represents perfect reliability.

the case of sppLSM_IC. More precisely, sppLSM_IC was characterized with smaller spread but higher reliability indicating a sharper and more useful forecast. It should be noted, the Kansas event was outside the CONUS-E domain and hence it was not included in the general precipitation analysis discussed elsewhere in this paper.

To summarize the 3-h accumulated precipitation verification results, the rank histograms indicated some level of underdispersion and bias for all ensemble experiments. Frequency bias results varied with threshold, initializations, and forecast lead time, where the mixed_phys experiment was characterized by positive frequency bias

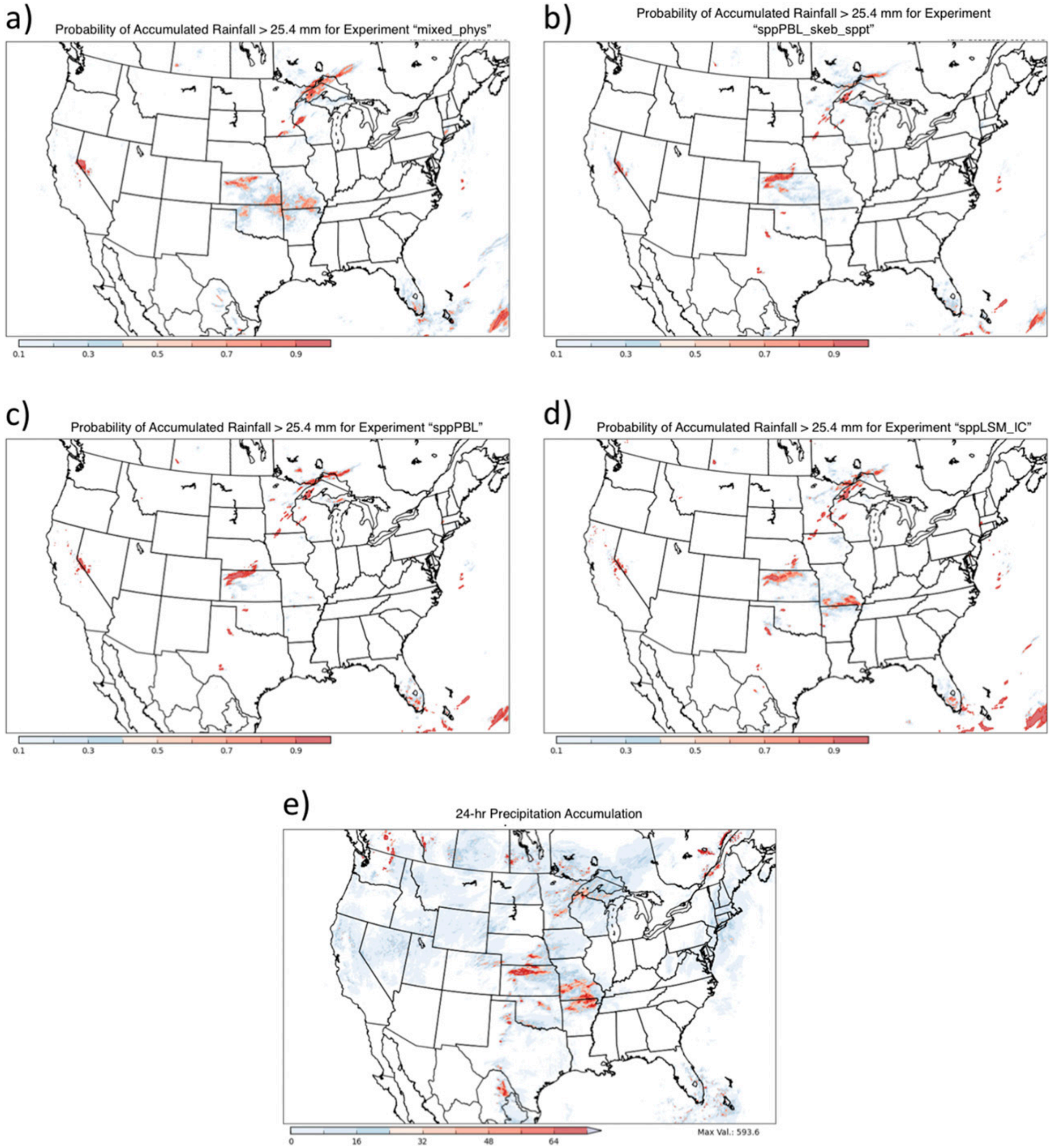
a)
Probability of Accumulated Rainfall > 25.4 mm for Experiment "mixed_phys"
0.1
0.3
0.5
0.7
0.9
b)
Probability of Accumulated Rainfall > 25.4 mm for Experiment "sppPBL_skeb_sppt"
0.1
0.3
0.5
0.7
0.9
c)
Probability of Accumulated Rainfall > 25.4 mm for Experiment "sppPBL"
0.1
0.3
0.5
0.7
0.9
d)
Probability of Accumulated Rainfall > 25.4 mm for Experiment "sppLSM_IC"
0.1
0.3
0.5
0.7
0.9
e)
24-hr Precipitation Accumulation
0
16
32
48
64
Max Val.: 593.6

FIG. 8. Probability of 24-h precipitation accumulation >25.4-mm threshold for the (a) mixed_phys, (b) sppPBL_skeb_sppt, (c) sppPBL, and (d) sppLSM_IC experiments, and (e) total estimated 24-h precipitation accumulation ending at 0000 UTC 25 May 2016.

for both thresholds and both initializations at most of the lead times. The sppLSM_IC experiment had lower-frequency bias for light precipitation and close to 1 for heavier precipitation threshold. In general, sppLSM_IC had better frequency bias when compared to other experiments. The FSS analysis for both light and heavier precipitation thresholds for the two neighborhood sizes generally showed comparable results for all experiments, with only a few statistically significant differences noted favoring the sppLSM_IC experiment. For the two evaluated precipitation thresholds (0.254 and 12.7 mm), the sppLSM_IC experiment most often had the higher reliability compared to other experiments. Finally, an evaluation of probabilities of 24-h precipitation accumulation exceeding 25.4-mm threshold for a select event and all experiments showed similarities in performance between the mixed_phys and sppLSM_IC experiments compared to the observed precipitation accumulations. The sppLSM_IC ensemble was characterized with more concentrated areas of higher probabilities, which in combination with generally higher reliability indicates sharper forecast as compared to the mixed_phys ensemble. Also, sppLSM_IC was characterized with much broader area of low to moderate probabilities compared to other stochastic experiments. In general, the soil moisture perturbations at the initial time had an overall positive impact on precipitation forecasts.

b. Surface verification

In addition to precipitation, forecasts of surface variables including 2-m temperature, 2-m dewpoint temperature, and 10-m wind speed, were analyzed. Once again, discussed results will be concentrated on the CONUS-East domain and the two initialization times. At present, 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 underdispersion (Candille and Talagrand 2008).

For 2-m temperature and 0000 UTC initialization (Fig. 9a), all experiments had comparable RMSE values early in the forecast (evening/overnight hours). While the sppLSM_IC aggregate RMSE values were lower than the other experiments overnight, the errors increased most rapidly for the sppLSM_IC experiment during the day, which led to significantly higher error when compared to sppPBL and sppPBL_skeb_sppt, but not mixed_phys. A similar trend was observed in the sppLSM_IC experiment's bias (difference between forecasts and observations) results (Fig. 9b). Mixed_phys was characterized with an increasing bias throughout the simulation period. On the other hand, sppPBL and sppPBL_skeb_sppt experiments had a positive bias during the night that reversed to a small cool bias during the day.

Spread values for all experiments for the 0000 UTC initializations were lower than RMSE values indicating underdispersion. However, spread values varied widely between the experiments. The mixed_phys and sppLSM_IC experiments had comparable spread values for most forecast hours (exception is forecast hour 12). Overnight, sppPBL_skeb_sppt and sppPBL 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 sppPBL was characterized with significantly lower spread for the duration of the forecast period. For all experiments, the same diurnal pattern in RMSE, spread (Fig. 9c), and bias (Fig. 9d) was detected in the 1200 UTC initialization.

RMSE, spread, and bias analysis as a function of lead time for 2-m dewpoint temperature is presented in Fig. 10. For the 0000 UTC initializations, during the overnight hours (Fig. 10a), mixed_physics had significantly higher RMSE values when compared to the other experiments. During the day, as was the case for 2-m temperature, sppLSM_IC was characterized by a rapid increase and significantly higher error compared to other experiments (Fig. 10a). The corresponding bias analysis (Fig. 10b) showed that the significant increase for the sppLSM_IC daytime error was associated with a dry bias. All of the other experiments were characterized by a moist bias during the day. In terms of spread, the mixed_phys ensemble had comparable spread to sppLSM_IC overnight, while during the day the sppLSM_IC 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 sppLSM_IC and mixed_phys. The sppPBL experiment was once again characterized by significantly lower spread for the duration of the forecast. Overall, the high RMSE values for sppLSM_IC were accompanied by significantly higher spread compared to other experiments leading to a spread/skill ratio close to 1 by the end of the forecast period. The same analysis for the 1200 UTC initialization showed very similar diurnal trends in RMSE and spread (Fig. 10c). The bias analysis (Fig. 10d) showed statistically significant dry bias for sppLSM_IC throughout the duration of simulations and generally no bias for mixed_phys. The other two stochastic experiments were characterized with statistically significantly positive bias, for most of the lead times. Again, the sppLSM_IC experiment was the only ensemble to have a spread/skill ratio around 1.

The 10-m wind RMSE, spread, and bias are presented in Fig. 11. For the 0000 UTC initialization, all experiments had similar RMSE values for all forecast lead times (Fig. 11a). Also, all experiments had the same trend in bias (Fig. 11b) with a high bias overnight. During the daytime, mixed_phys and sppLSM_IC exhibit a small positive bias compared to the other two experiments, which have a small negative bias. While all experiments also had generally low spread, the spread values differed notably among the experiments (Fig. 11a). The mixed_physics and sppPBL_skeb_sppt had comparable spread that was significantly larger compared to the other two experiments for most lead times, with

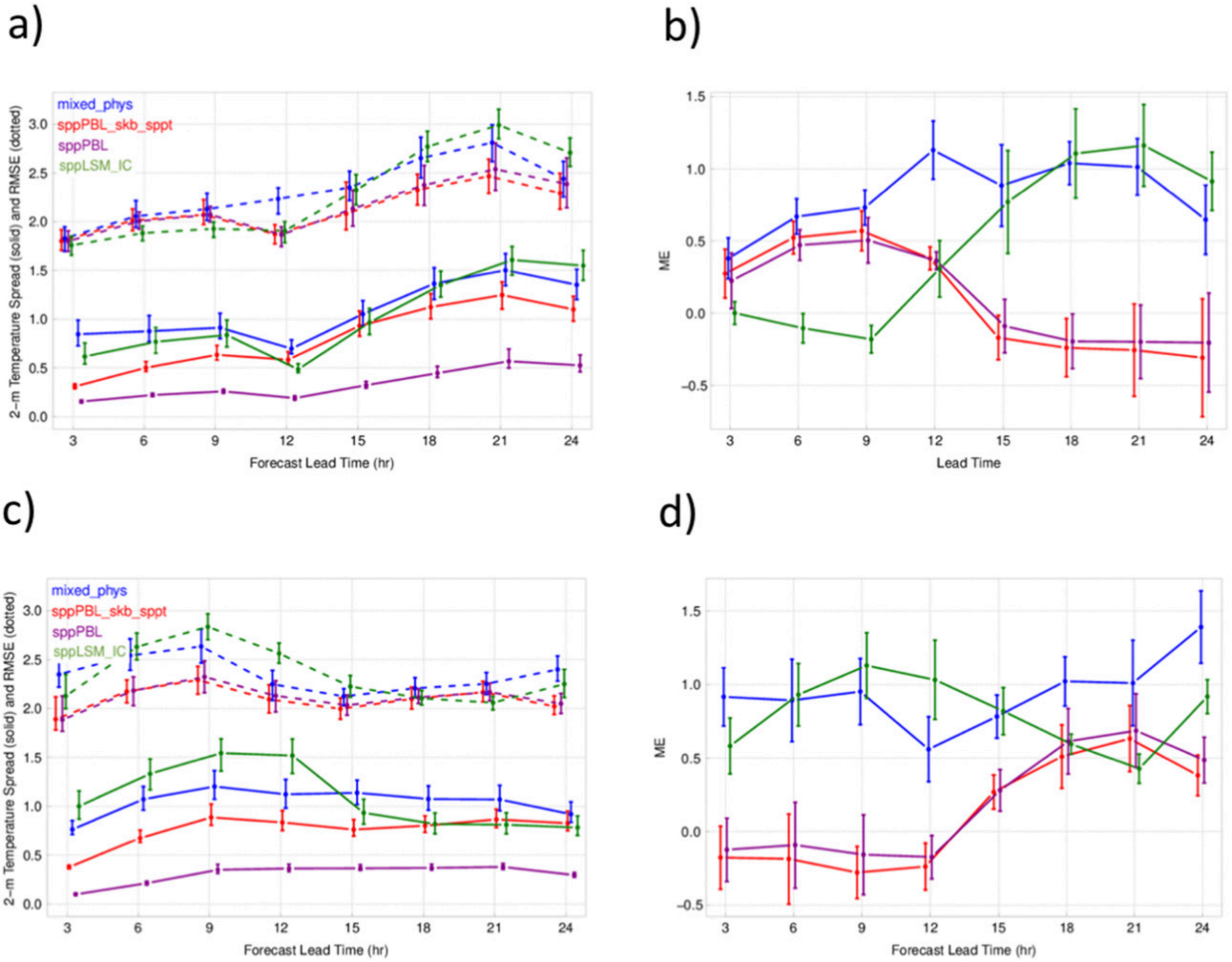

FIG. 9. The 2-m temperature (a) RMSE (dashed lines) and spread (solid lines) for 0000 UTC initialization, (b) bias for 0000 UTC initialization, (c) RMSE and spread for 1200 UTC initialization, and (d) bias for 1200 UTC initialization. The vertical bars indicate 95% confidence intervals. For 0000 UTC initialization, 12-h lead time marks approximate sunrise time, while for 1200 UTC initialization, that is an approximate sunset time.

sppLSM_IC having increasing spread toward the end of the period. The sppPBL experiment again has the lowest spread. Figures 11c and 11d show the same trend in RMSE and spread as well as bias diurnal change for the 1200 UTC initialization.

Further, reliability diagrams for surface variables aggregated over the 24-h forecast period were evaluated for select thresholds and the two initialization times. Because of similarity of the results from the two initializations, only 0000 UTC initialization results are discussed (Fig. 12). Figure 12a 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, though the stochastic experiments have better reliability for most frequencies compared to the mixed_physics ensemble, which was generally overconfident (Fig. 12a). Event histograms, for all experiments and both initializations, demonstrated the ability of the experiments to predict both low and high forecasted frequencies.

Reliability for 2-m dewpoint temperature was evaluated for the greater than 283-K threshold (Fig. 12c). The sample rate for this threshold was close to 70%, which made it a relatively common event. Generally, all experiments were underconfident. The sppLSM_IC and mixed_physics ensembles progressively became more underconfident for higher forecast frequencies. The sppPBL_skeb_sppt and sppPBL ensembles generally remained more reliable compared to the other two ensembles. The highest forecast frequency was characterized by underconfidence, no skill, and higher values in the event histogram for all ensembles.

The 10-m wind speed ensemble reliability was evaluated for wind speeds greater than 4 m s^{-1} and greater

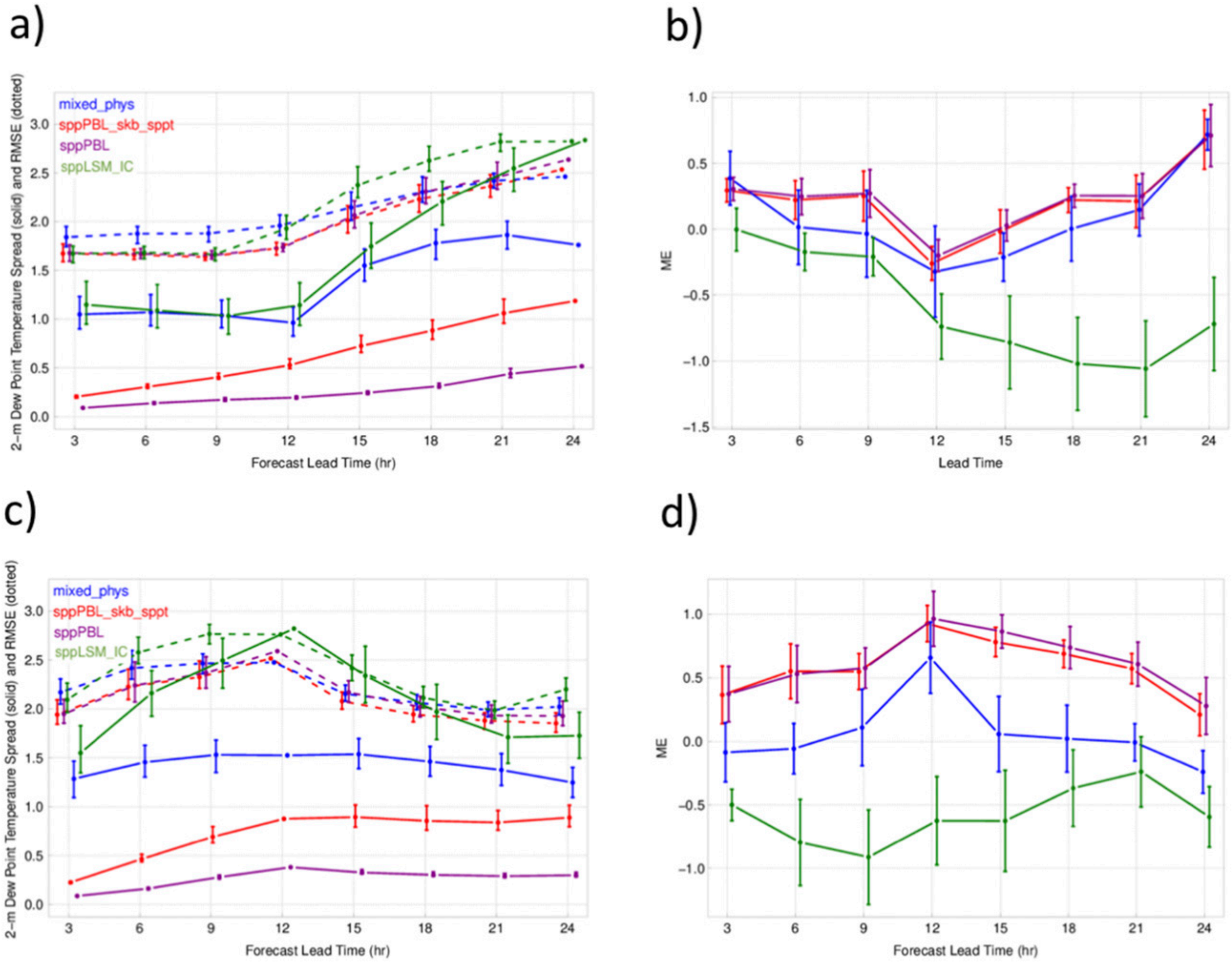

FIG. 10. As in Fig. 9, but for 2-m dewpoint temperature.

than 6 m s^{-1} (Figs. 12c and 12d). For the 4 m s^{-1} threshold (Fig. 12c) the sample rate was somewhat lower than 30%. Generally, all ensembles were overconfident, with sppPBL_skeb_sppt having better reliability. Overconfidence for higher frequencies and for all ensembles was associated with lower event histogram values. The same analysis except for the 6 m s^{-1} threshold is presented in Fig. 12d. For all frequencies sppPBL and sppPBL_skeb_sppt demonstrated better reliability compared to the other two experiments. However, all ensembles were characterized with a limited number of highest probabilities forecasts, likely related to the fact that this was a rare event (base rate <10%).

Given the fact that both sppPBL and sppPBL_skeb_sppt were characterized by small spread for 2-m temperature (Fig. 9) and 2-m dewpoint temperature (Fig. 10), this result implies that the PBL perturbations even with the addition of SKEB and SPPT, did not have much impact on spread for either of those surface variables. This agrees with other studies that suggest that solely perturbing atmospheric physics scheme parameters is currently not enough to achieve sufficient spread at surface (Jankov et al. 2017; Hacker et al. 2011b; Reynolds et al. 2011; Berner et al. 2015). Reliability diagram analysis for 2-m temperature and the 293-K thresholds showed an advantage for the stochastic experiments over mixed_physics (Fig. 12a). The 2-m dewpoint reliability analysis for the 283-K thresholds revealed similar behavior between sppLSM_IC and mixed_physics and somewhat inferior as compared to the other two ensembles (Fig. 12b). Reliability analysis for 10-m wind exceeding 4 and 6 m s^{-1} showed higher reliability for sppPBL_skeb_sppt and sppPBL compared to the other two experiments. Overall, the analysis revealed that even though sppPBL and sppPBL_skeb_sppt were generally characterized by lower spread, they resulted in more reliable forecasts for select thresholds, especially in regards to the 10-m wind forecasts.

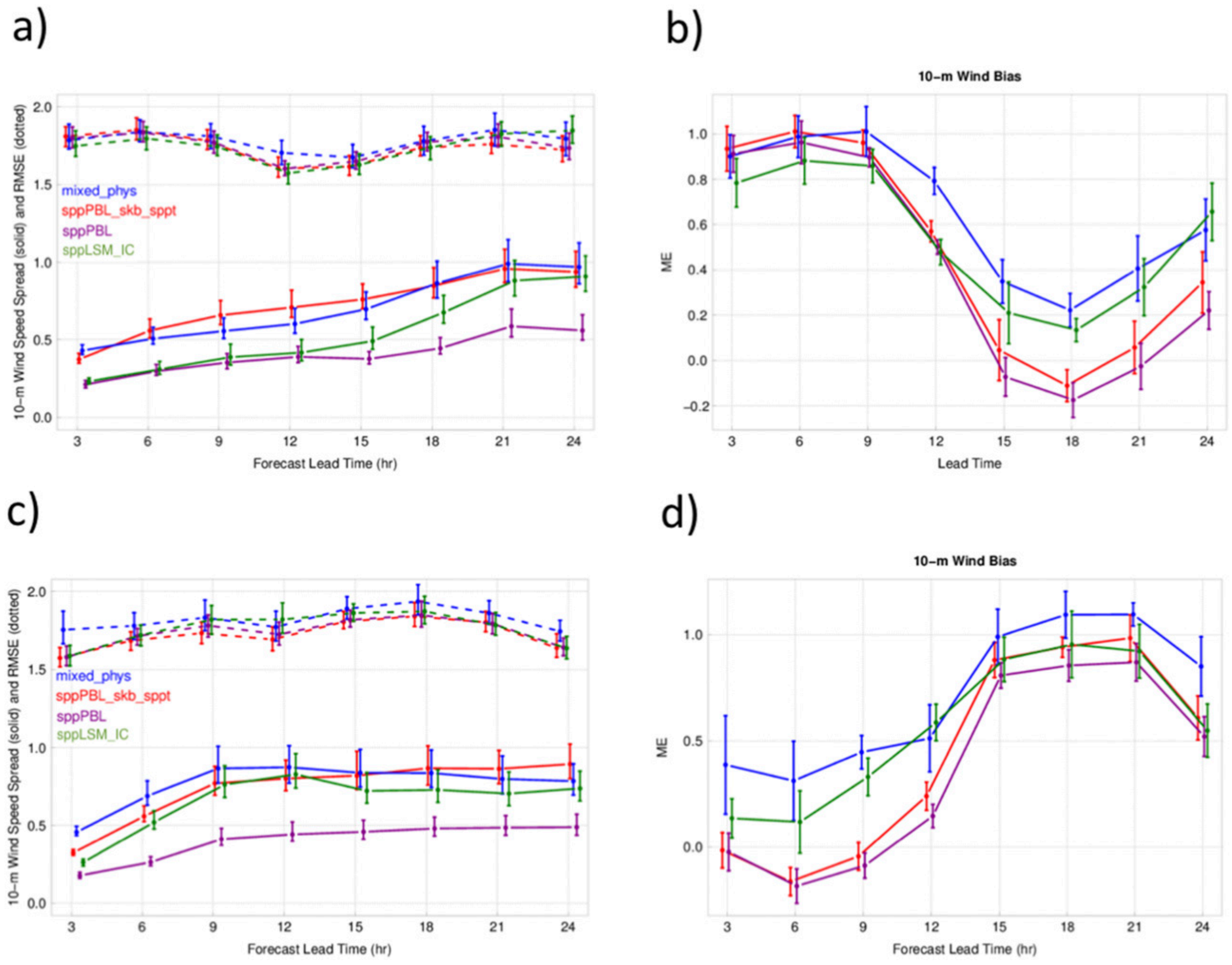

FIG. 11. As in Fig. 9, but for 10-m wind speed.

c. 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 (Fig. 13). For 850-hPa temperature and 0000 UTC initializations, all stochastic experiments had significantly lower RMSE at initialization time compared to the mixed_phys (Fig. 13a). The RMSE values generally increased with forecast lead time, with sppLSM_IC having the largest RMSE values by the end of the period (Fig. 13a). When looking at spread, the sppLSM_IC experiment had significantly larger values at initialization time, implying the initial condition perturbations, in combination with cycling of soil moisture and temperature, had a significant impact. The sppPBL_skeb_sppt experiment had the largest spread at the 1200 UTC valid time and the sppPBL was generally significantly lower than the other experiments throughout the period.

For 0000 UTC initializations,there are no statistically significant differences in RMSE values for 500-hPa geopotential height; however, spread varies significantly between each of the experiments (Fig. 13b). The sppPBL_skeb_sppt experiment had significantly larger spread, followed by mixed_phys, sppLSM_IC, and finally sppPBL. This led to the sppPBL_skeb_sppt experiment having the best spread/skill ratio.

The mixed_phys experiment had significantly larger error at the initial time for 250-hPa wind for 0000 UTC initializations, but differences in RMSE between all of the experiments were not significant at 12- and 24 h lead times (Fig. 13c). Similar spread for sppPBL_skeb_sppt and mixed_physics was noted, which was significantly larger when compared to the other two experiments. Similar trends in spread and RMSE for all variables were observed for 1200 UTC initializations (not shown).

In general, the upper-air analysis indicates that the use of SKEB and SPPT improves model performance

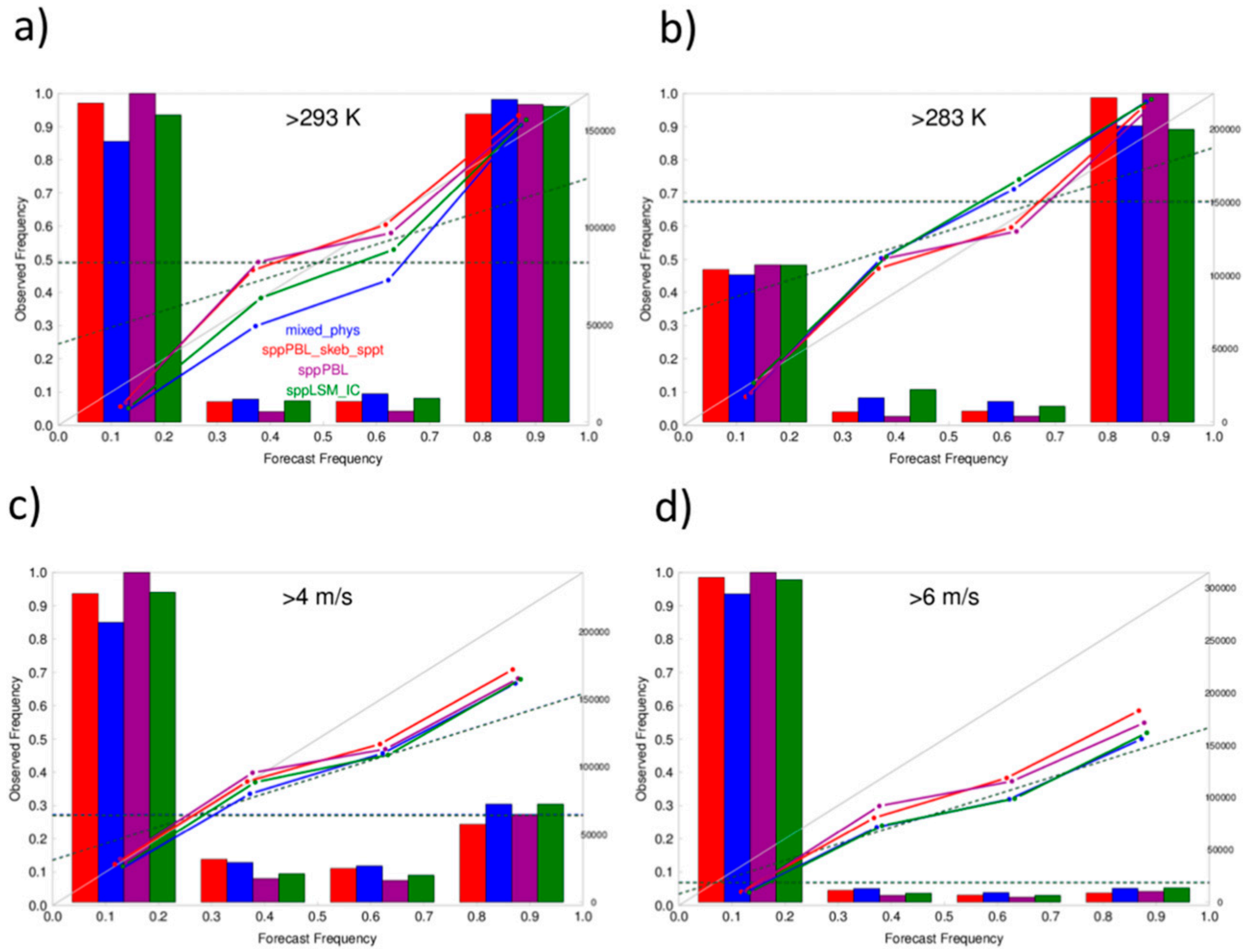

FIG. 12. Reliability diagrams for 0000 UTC initialization over the eastern part of the domain for (a) 2-m temperature for a threshold of >293 K, (b) 2-m dewpoint temperature for a threshold of >283 K, (c) 10-m wind for a threshold of >4 m s^{-1}, and (d) 10-m wind for a threshold of >6 m s^{-1}. The horizontal dotted line represents no resolution, the diagonal dotted line represents no skill, and the solid gray diagonal line represents perfect reliability.

(e.g., spread) for upper-air variables. This was especially the case for 500-hPa geopotential height. The same finding was also valid for the RAP-based ensemble (Jankov et al. 2017).

4. Summary and conclusions

The next generation of unified operational system will require a well-performing, rapid refresh, convection allowing, single dynamic core and single physics suite ensemble. Therefore, it is critical to explore stochastic approaches as a potential alternative to the current multidycore, multiphysics suites, high-resolution ensemble. A stochastically perturbed parameterization (SPP) approach was developed to represent sources of uncertainty within the HRRR physics suite. Encouraged by the performance at 15 km, as part of the RAP ensemble (Jankov et al. 2017), we present here results at a convection-allowing resolution of 3 km.

Ensemble performance using only SPP and in combination with other stochastic methods (SKEB and SPPT) was compared against a multiphysics ensemble. This is a high bar, since the multiphysics ensemble is a more effective method of representing model error than currently applied stochastic methods with a single model configuration (e.g., Berner et al. 2015).

The SPP approach introduces temporally and spatially varying perturbations to key parameters and variables in the MYNN PBL physics parameterization (turbulent mixing length, subgrid cloud fraction, thermal and moisture roughness lengths, and Prandtl's number). The SPP spatial pattern was also applied to the soil moisture field of the RUC LSM scheme at the initialization time. The detailed characteristics of these perturbations (perturbation amplitude and spatial and temporal decorrelation lengths) were determined through collaboration with physics parameterization experts. For the HRRR

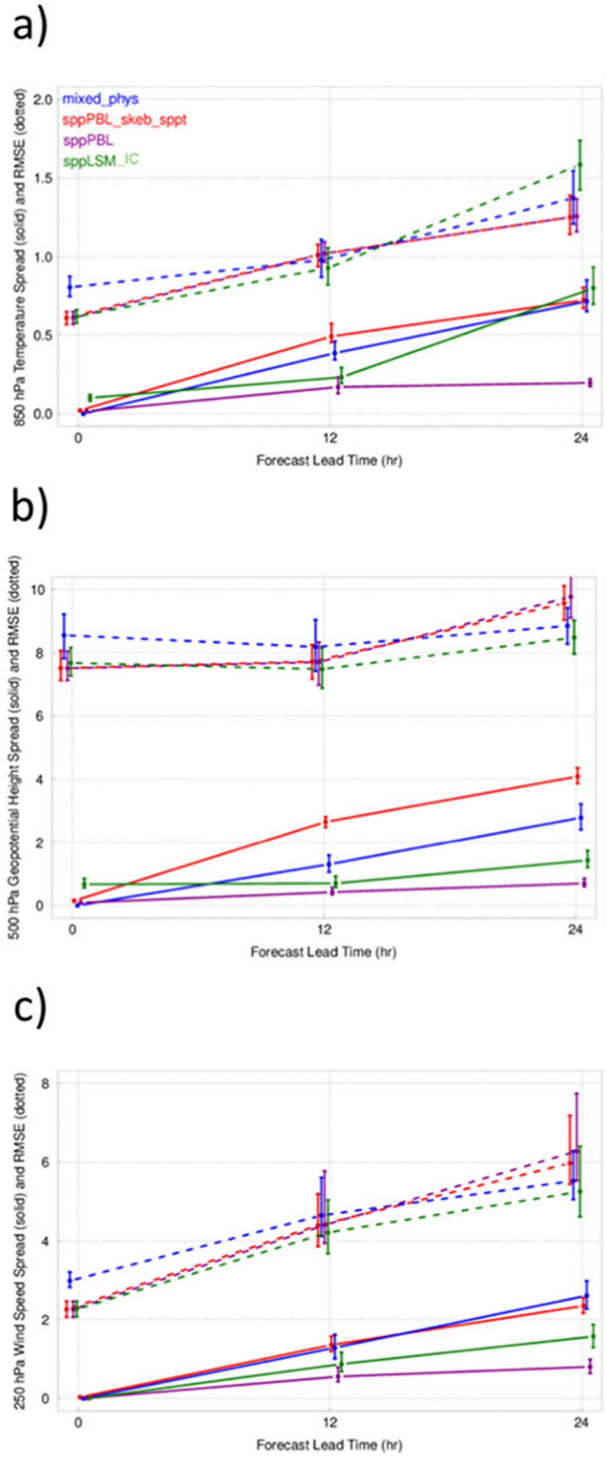

FIG. 13. RMSE and spread for eastern part of the domain for the 0000 UTC initialization for (a) 850-hPa temperature, (b) 500-hPa geopotential height, and (c) 250-hPa wind speed. The 95% confidence intervals are included.

domain, a decorrelation time and length of 6 h and 150 km, respectively, were found to be appropriate for convective scales.

An 8-member HRRR ensemble consisting of 24-h forecasts was evaluated using a variety of metrics over the 18–27 May 2016 period to assess the impact of stochastic approaches primarily on precipitation and surface variables, but also on upper levels. All model runs used RAP forecasts as initial conditions for a 1-h preforecast that included the latest observations.

Significant findings are summarized below:

- Representing uncertainty in the soil moisture initialization resulted in a generally positive impact on precipitation skill and reliability, which is consistent with the recent findings of Bouttier et al. (2016) and Schraff et al. (2016). However, these perturbations were accompanied by an increase in the RMSE of 2-m dewpoint temperature due to a dry bias. This approach should be investigated in more detail in order to effectively tune the amplitude and spatial scales of the perturbations to improve probabilistic performance without a deterioration in the error.
- Applying perturbations to different parameters within the PBL scheme did not, in itself, result in sufficient spread in near-surface variables with the exception of 10-m wind speed, where it increased reliability and sharpness.
- Perturbations from SKEB and SPPT were combined with SPP-PBL to represent other sources of model error. Ensembles using this combination of stochastic schemes showed improved skill in 10-m wind verification and all examined upper-level variables.

Our results generally confirm the findings of previous studies performed using coarser grid spacings (e.g., Jankov et al. 2017; Berner et al. 2011, 2015; Hacker et al. 2011a,b), although convection is largely resolved in our simulations and no longer dominated by convective parameterization tendencies. The latter fact implies that the ensemble is expected to represent uncertainty in convection and cannot be mimicked by perturbing convective tendencies. We found 1) perturbations of a limited number of parameters within a single physics scheme did not generate sufficient spread to remedy underdispersion for short-term ensemble forecasts, and 2) a combination of several stochastic schemes outperformed any single scheme for the dataset used in the present study.

It was generally expected that perturbations within a single scheme (in this case, PBL) would not lead to sufficient spread and—for short forecast lead times—be limited to near-surface variables. However, SPP led to frequently comparable, and in the case of the 10-m wind, generally better reliabilities. This implies that application of SPP leads toward an ensemble spread that is more effective in encompassing sources of the parameterization uncertainties. The improvement in 10-m wind speed reliability and sharpness represents a successful implementation of PBL perturbations designed to improve 10-m wind speed metrics. Therefore, an improvement in performance for targeted variables can be made when using SPP.

Our research shows that at convective-permitting resolution, 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 applying the SPP approach to a variety of schemes will account for more realistic representation of model error at the process level. In the future, SPP will be added to the Thompson microphysics scheme, additional parameters in the PBL and LSM schemes, and radiation parameterization in order to more comprehensively represent model uncertainty at its source. The use of SPP within many different physics schemes may be a valuable option for adding sufficient spread within an operational, convective-allowing, single-physics ensemble system.

Acknowledgments. This study was supported by the Developmental Testbed Center (DTC) funding for the DTC Regional Ensemble task. Special thanks go to James Frimel for his help with setting up the retrospective runs and to Pat Skinner for the use of his ensemble-based Python plotting scripts. Also, thanks to Tressa Fowler for her explanation of, and assistance with, the applied statistical methods.

REFERENCES

Aligo, E. A., W. A. Gallus Jr., and M. Segal, 2007: Summer rainfall forecast spread in an ensemble initialized with different soil moisture analyses. *Wea. Forecasting*, **22**, 299–314, https://doi.org/10.1175/WAF995.1.

Arakawa, A., and V. R. Lamb, 1977: Computational design of the basic dynamical processes of the UCLA general circulation model. *Methods Comput. Phys.*, **17**, 173–265.

Benjamin, S. G., and Coauthors, 2016: A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Mon. Wea. Rev.*, **144**, 1669–1694, https://doi.org/10.1175/MWR-D-15-0242.1.

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, https://doi.org/10.1175/2008JAS2677.1.

——, S.-Y. Ha, J. P. Hacker, A. Fournier, and C. Snyder, 2011: Model uncertainty in a mesoscale ensemble prediction system: Stochastic versus multiphysics representations. *Mon. Wea. Rev.*, **139**, 1972–1995, https://doi.org/10.1175/2010MWR3595.1.

——, 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, https://doi.org/10.1175/JCLI-D-11-00297.1.

——, 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, https://doi.org/10.1175/MWR-D-14-00091.1.

——, and Coauthors, 2017: Stochastic parameterization: Toward a new view of weather and climate models. *Bull. Amer. Meteor. Soc.*, **98**, 565–588, https://doi.org/10.1175/BAMS-D-15-00268.1.

Bouttier, F., B. Vié, O. Nuissier, and L. Raynaud, 2012: Impact of stochastic physics in a convection-permitting ensemble. *Mon. Wea. Rev.*, **140**, 3706–3721, https://doi.org/10.1175/MWR-D-12-00031.1.

——, L. Raynaud, O. Nuissier, and B. Ménétrier, 2016: Sensitivity of the AROME ensemble to initial and surface perturbations during HYMEX. *Quart. J. Roy. Meteor. Soc.*, **142**, 390–403, https://doi.org/10.1002/qj.2622.

Bowler, N. E., A. Arribas, K. R. Mylne, K. B. Robertson, and S. E. Beare, 2008: The MOGREPS short-range ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **134**, 703–722, https://doi.org/10.1002/qj.234.

——, ——, 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, https://doi.org/10.1002/qj.394.

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, https://doi.org/10.1002/qj.49712556006.

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. Developmental Testbed Center, Boulder, CO, 400 pp., https://dtcenter.org/met/users/docs/users_guide/MET_Users_Guide_v6.1.pdf.

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**, 509–521, https://doi.org/10.1002/qj.221.

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, https://doi.org/10.1175/JAS-D-14-0250.1.

Duda, J. D., X. Wang, and M. Xue, 2017: Sensitivity of convection-allowing forecasts to land surface model perturbations and implications for ensemble design. *Mon. Wea. Rev.*, **145**, 2001–2025, https://doi.org/10.1175/MWR-D-16-0349.1.

Eckel, F. A., and C. F. Mass, 2005: Aspects of effective mesoscale, Short-Range Ensemble Forecasting. *Wea. Forecasting*, **20**, 328–350, https://doi.org/10.1175/WAF843.1.

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. *J. Geophys. Res.*, **108**, 8851, https://doi.org/10.1029/2002JD003296.

Hacker, J. P., C. Snyder, S.-Y. Ha, and M. Pocernich, 2011a: Linear and nonlinear response to parameter variations in a mesoscale model. *Tellus*, **63A**, 429–444, https://doi.org/10.1111/j.1600-0870.2010.00505.x.

——, and Coauthors, 2011b: The U.S. Air Force Weather Agency's mesoscale ensemble: Scientific description and performance results. *Tellus*, **63A**, 625–641, https://doi.org/10.1111/j.1600-0870.2010.00497.x.

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550–560, https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2.

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**, 2318–2341, https://doi.org/10.1175/MWR3199.1.

Jankov, I., and Coauthors, 2017: A performance comparison between multiphysics and stochastic approaches within a North American RAP ensemble. *Mon. Wea. Rev.*, **145**, 1161–1179, https://doi.org/10.1175/MWR-D-16-0160.1.

Leutbecher, M., and Coauthors, 2017: Stochastic representations of model uncertainties at ECMWF: State of the art and future vision. *Quart. J. Roy. Meteor. Soc.*, **143**, 2315–2339, https://doi.org/10.1002/qj.3094.

McCabe, A., R. Swinbank, W. Tennant, and A. Lock, 2016: Representing model uncertainty in the Met Office convection-permitting ensemble prediction system and its impact on fog forecasting. *Quart. J. Roy. Meteor. Soc.*, **142**, 2897–2910, https://doi.org/10.1002/qj.2876.

Mlawer, E. J., S. J. Taubman, P. D. Brown, M. J. Iacono, and S. A. Clough, 1997: Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave. *J. Geophys. Res.*, **102**, 16 663–16 682, https://doi.org/10.1029/97JD00237.

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, https://doi.org/10.1023/B:BOUN.0000020164.04146.98.

——, and ——, 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, https://doi.org/10.1007/s10546-005-9030-8.

Ollinaho, P., and Coauthors, 2017: Towards process-level representation of model uncertainties: Stochastically perturbed parametrizations in the ECMWF ensemble. *Quart. J. Roy. Meteor. Soc.*, **143**, 408–422, https://doi.org/10.1002/qj.2931.

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, https://doi.org/10.1002/qj.49712757202.

——, R. Buizza, F. Doblas-Reyes, T. Jung, M. Leutbecher, G. Shutts, M. Steinheimer, and A. Weisheimer, 2009: Stochastic parametrization and model uncertainty. ECMWF Tech. Memo. 598, 44 pp., https://www.ecmwf.int/en/elibrary/11577-stochastic-parametrization-and-model-uncertainty.

Pleim, J. E., 2007: A combined local and nonlocal closure model for the atmospheric boundary layer. Part I: Model description and testing. *J. Appl. Meteor. Climatol.*, **46**, 1383–1395, https://doi.org/10.1175/JAM2539.1.

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 U.S. Navy global ensemble performance in the tropics. *Mon. Wea. Rev.*, **139**, 2145–2155, https://doi.org/10.1175/2011MWR3546.1.

Roberts, N. M., and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97, https://doi.org/10.1175/2007MWR2123.1.

Romine, G. S., C. S. Schwartz, J. Berner, K. R. Fossell, C. Snyder, J. L. Anderson, and M. L. Weisman, 2014: Representing forecast error in a convection-permitting ensemble system. *Mon. Wea. Rev.*, **142**, 4519–4541, https://doi.org/10.1175/MWR-D-14-00100.1.

Sanchez, C., K. D. Williams, and M. Collins, 2016: Improved stochastic physics schemes for global weather and climate models. *Quart. J. Roy. Meteor. Soc.*, **142**, 147–159, https://doi.org/10.1002/qj.2640.

Schraff, C., H. Reich, A. Rhodin, A. Schomburg, K. Stephan, A. Periáñez, and R. Potthast, 2016: Kilometre-scale ensemble data assimilation for the COSMO model (KENDA). *Quart. J. Roy. Meteor. Soc.*, **142**, 1453–1472, https://doi.org/10.1002/qj.2748.

Schwartz, C. S., and Coauthors, 2010: Toward improved convection-allowing ensembles: Model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Wea. Forecasting*, **25**, 263–280, https://doi.org/10.1175/2009WAF2222267.1.

Skamarock, W. C., and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Tech. Note NCAR/TN-475+STR, 113 pp., https://doi.org/10.5065/D68S4MVH.

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 Forecasting (WRF) Model. *Mon. Wea. Rev.*, **144**, 1851–1865, https://doi.org/10.1175/MWR-D-15-0198.1.

Stull, R. B., 2012: *An Introduction to Boundary Layer Meteorology*. Springer Science & Business Media, 670 pp.

Sutton, C., T. M. Hamill, and T. T. Warner, 2006: Will perturbing soil moisture improve warm-season ensemble forecasts? A proof of concept. *Mon. Wea. Rev.*, **134**, 3174–3189, https://doi.org/10.1175/MWR3248.1.

Thompson, G., P. R. Field, R. M. Rasmussen, and W. D. Hall, 2008: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part II: Implementation of a new snow parameterization. *Mon. Wea. Rev.*, **136**, 5095–5115, https://doi.org/10.1175/2008MWR2387.1.

Zhang, J., and Coauthors, 2016: Multi-Radar Multi-Sensor (MRMS) quantitative precipitation estimation: Initial operating capabilities. *Bull. Amer. Meteor. Soc.*, **97**, 621–638, https://doi.org/10.1175/BAMS-D-14-00174.1.