A Progress Report on the Development of the High-Resolution Rapid Refresh Ensemble

EVAN A. KALINA,[a,b] ISIDORA JANKOV,[b] TREVOR ALCOTT,[b] JOSEPH OLSON,[b] JEFFREY BECK,[b,c] JUDITH BERNER,[d] DAVID DOWELL,[b] AND CURTIS ALEXANDER[b]

[a] *Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, Boulder, Colorado*
[b] *NOAA/Global Systems Laboratory, Boulder, Colorado*
[c] *Cooperative Institute for Research in the Atmosphere, Colorado State University, Boulder, Colorado*
[d] *National Center for Atmospheric Research, Boulder, Colorado*

(Manuscript received 16 June 2020, in final form 13 February 2021)

ABSTRACT: The High-Resolution Rapid Refresh Ensemble (HRRRE) is a 36-member ensemble analysis system with 9 forecast members that utilizes the Advanced Research version of the Weather Research and Forecasting (ARW-WRF) dynamic core and the physics suite from the operational Rapid Refresh/High-Resolution Rapid Refresh deterministic modeling system. A goal of HRRRE development is a system with sufficient spread among members, comparable in magnitude to the random error in the ensemble mean, to represent the range of possible future atmospheric states. HRRRE member diversity has traditionally been obtained by perturbing the initial and lateral boundary conditions of each member, but recent development has focused on implementing stochastic approaches in HRRRE to generate additional spread. These techniques were tested in retrospective experiments and in the May 2019 Hazardous Weather Testbed Spring Experiment (HWT-SE). Results show a 6%–25% increase in the ensemble spread in 2-m temperature, 2-m mixing ratio, and 10-m wind speed when stochastic parameter perturbations are used in HRRRE (HRRRE-SPP). Case studies from HWT-SE demonstrate that HRRRE-SPP performed similar to or better than the operational High-Resolution Ensemble Forecast system, version 2 (HREFv2), and the nonstochastic HRRRE. However, subjective evaluations provided by HWT-SE forecasters indicated that overall, HRRRE-SPP predicted lower probabilities of severe weather (using updraft helicity as a proxy) compared to HREFv2. A statistical analysis of the performance of HRRRE-SPP and HREFv2 from the 2019 summer convective season supports this claim, but also demonstrates that the two systems have similar reliability for prediction of severe weather using updraft helicity.

KEYWORDS: Severe storms; Ensembles; Forecast verification/skill; Numerical weather prediction/forecasting; Parameterization

1. Introduction

All weather forecasts are characterized by some degree of inherent uncertainty (AMS 2002; NRC 2003). These uncertainties arise both from an incomplete picture of the initial state of the atmosphere (a chaotic system; Lorenz 1963) and an imperfect representation of the physical processes that govern various atmospheric phenomena in numerical weather prediction models (Stensrud et al. 2000; NRC 2006; Teixeira and Reynolds 2008). Many of these processes, including atmospheric turbulence and the interactions between individual hydrometeors, must be parameterized at the spatial scales of high-resolution operational weather prediction models ($\Delta x \sim 3$ km; $\Delta z \sim 100$ m). Modern weather forecasts increasingly attempt to account for these sources of uncertainty. As a result, today's predictions are often probabilistic (e.g., there is a 10% chance of a tornado within 25 miles of a point in central Oklahoma today) rather than deterministic (e.g., there will be a tornado in central Oklahoma today).

In recent years, model ensembles have become a valuable tool in our attempts to quantify this forecast uncertainty (Leith 1974; Grimit and Mass 2002; Keune et al. 2014). The makeup of individual ensemble systems varies widely. Each member can be identical except for small changes in the initial and/or boundary conditions provided to the forecast model (e.g., Molteni et al. 1996; Hamill and Colucci 1997), or the ensemble members can be completely different models with different physics schemes or even different dynamic cores (AMS 2002). One measure of the success of an ensemble is its reliability – whether an event forecasted by a given fraction of the ensemble members actually occurs, on average, that fraction of the time. A related desired characteristic of an ensemble is that the spread of the different member solutions (i.e., uncertainty) is comparable to the error of the ensemble mean. If the spread does not at least match the model error, then the ensemble *underestimates* the degree of uncertainty in the forecast and is said to be underdispersive. Unfortunately, this has been the state of ensemble forecasting for many years throughout the world (Berner et al. 2017).

Historically, multiphysics, multidynamics ensembles have come closest to achieving a spread that matches the error. These ensembles typically produce a large diversity of forecast solutions across the different members that comprise the ensemble, which leads to sufficient spread and improved probabilistic forecasts (e.g., Hacker et al. 2011; Berner et al. 2011, 2015). One prominent example of a multiphysics, multidynamics ensemble is version 2 of the operational High-Resolution

Denotes content that is immediately available upon publication as open access.

Corresponding author: Evan A. Kalina, evan.kalina@noaa.gov

DOI: 10.1175/WAF-D-20-0098.1

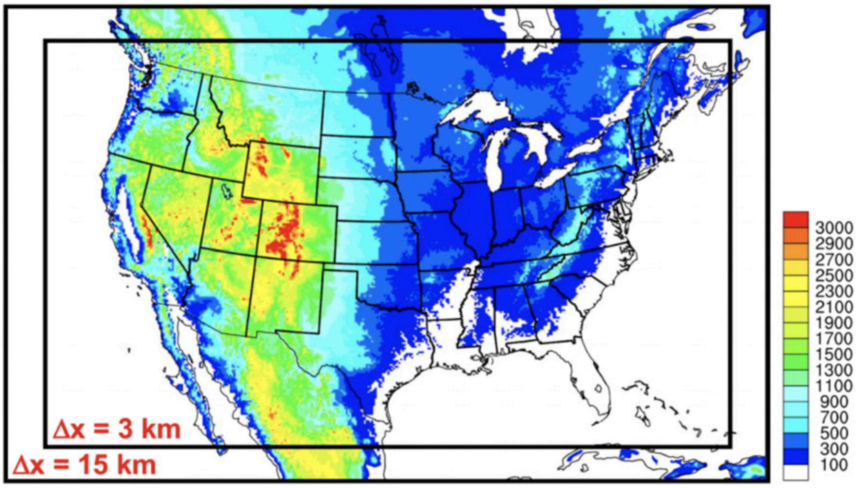

FIG. 1. The HRRRE domain configuration, which consists of a parent domain with 15-km horizontal grid spacing and a CONUS-scale nest with 3-km horizontal grid spacing. Terrain height (m) is shaded.

Ensemble Forecast (HREFv2; Roberts et al. 2019) system designed and run by the National Centers for Environmental Prediction (NCEP). HREFv2 consists of eight members that contain a mixture of dynamics and physics schemes, four of which are 12-h time-lagged versions of the other four. Clark et al. (2019) state that "the diversity in HREFv2 has proven to be a very effective configuration strategy, and it has consistently outperformed all other [convection-allowing model] ensembles examined in the [Hazardous Weather Testbed] (HWT) during the last few years." However, multiphysics, multidynamics ensembles are difficult to maintain and develop, since each ensemble member has a completely different design. Performing statistical postprocessing on such an ensemble is complicated by the fact that each member has a different climatology and a different set of error characteristics. These differences contribute to a sufficiently large ensemble spread[1] (Eckel and Mass 2005; Berner et al. 2015), but at the cost of each member's raw forecast not representing an equally likely outcome. In addition, if the biases in the individual ensemble members are removed, the spread in the model solutions generally collapses (Eckel and Mass 2005; Berner et al. 2015). The spread, in other words, is not being produced for the right reasons, but is instead a product of model biases.

To address these issues, an alternative ensemble system, the nonoperational High-Resolution Rapid Refresh Ensemble (HRRRE), was designed at NOAA's Global Systems Laboratory in 2016. Unlike HREFv2, the nine-member HRRRE uses a single dynamic core (the Advanced Research version of the Weather Research and Forecasting core; ARW-WRF) and a single physics suite. Until spring 2019, differences between ensemble members were driven entirely by initial and boundary condition perturbations. While easier to maintain with a smaller ensemble mean bias, HRRRE has suffered substantially more underdispersion relative to HREFv2.

Going forward, NCEP's goal is to adopt an ensemble system that utilizes a single atmospheric dynamic core [i.e., the Finite-Volume Cubed-Sphere dynamic core (FV3); Harris and Lin (2013) and references therein] and, potentially, a single atmospheric physics suite. An obvious challenge is to design an ensemble system that produces a comparable spread-error ratio to HREFv2 with substantially less diversity in the design of the different ensemble members. The remainder of this article describes current research efforts to accomplish this goal using HRRRE.

2. HRRRE design

The experimental HRRRE has two components: a 36-member ensemble-analysis system (HRRR Data Assimilation System, or HRRRDAS) and a 9-member ensemble forecast system. An outer domain with 15-km horizontal grid spacing (Fig. 1) exists for the purpose of adding random perturbations to the zonal and meridional winds, temperature, water vapor mixing ratio, and column dry air mass near the lateral boundaries of each ensemble member. The inner domain covers the contiguous United States with convection-allowing, 3-km horizontal grid spacing, as in the deterministic HRRR system. The physics schemes used in HRRRE are described by Benjamin et al. (2016) and Olson et al. (2019).

While there are advantages to cycling an ensemble analysis system continuously (Schwartz et al. 2019), the HRRRDAS instead uses a strategy of reinitializing members at regular intervals from parent models. This strategy has been used throughout the history of the RAP and HRRR systems (Benjamin et al. 2016). HRRRDAS members are initialized twice per day, at 0900 and 2100 UTC. RAP atmosphere and HRRR soil analyses at these times provide the initial ensemble mean for the HRRRDAS. The atmospheric states in the

[1] In this paper, the ensemble spread is defined as the standard deviation of a given variable of interest.

36 HRRRDAS members are created by adding perturbations from the first 36 members of the Global Data Assimilation System (GDAS; Parrish and Derber 1992; Derber and Wu 1998; NCEP 2004) to the RAP analysis. The land surface states are perturbed with SPP, as described in the appendix.

The HRRRDAS is cycled hourly, using an ensemble Kalman filter to assimilate conventional and radar-reflectivity observations. Nonvariational cloud clearing, based on satellite and radar observations, is also applied hourly to members individually, as in the RAP and HRRR systems (Benjamin et al. 2016). Relaxation to prior spread (RTPS; Whitaker and Hamill 2012) helps maintain ensemble spread during the hourly cycling. At times determined by testbeds (typically 0000 and 1200 UTC), the first 9 members of the HRRRDAS are advanced as a free forecast, out to lead times as much as 36 h. The other 27 HRRRDAS members do not participate in the free forecast due to computer resource constraints and the desire to produce a forecast ensemble that is similar in size to HREFv2. The 9-member HRRRE forecasts are the focus of the current study.

3. Stochastic physics in HRRRE

Beginning in late 2017, a variety of stochastic approaches were added to HRRRE and investigated. These techniques represent another way (beyond simply perturbing initial and lateral boundary conditions) to increase spread in a single-dynamic core, single-physics ensemble (e.g., Palmer 2001; Jankov et al. 2017). One clear advantage of this approach is a statistically consistent ensemble distribution (e.g., Bowler et al. 2009; Berner et al. 2009; Sanchez et al. 2015). The stochastic approaches considered for HRRRE were stochastic perturbations of physics tendencies (SPPT; Buizza et al. 1999, Palmer et al. 2009), stochastic kinetic energy backscatter (SKEB; Berner et al. 2009, 2012, 2015), and stochastic parameter perturbations (SPP). In SKEB, model uncertainty associated with subgrid-scale processes is addressed by randomly perturbing streamfunction and potential temperature tendencies (Berner et al. 2009, 2012, 2015). In contrast, SPPT (Palmer et al. 2009) considers the subgrid-scale process uncertainty by perturbing the total physics tendencies for fields such as temperature, humidity, and wind (Bouttier et al. 2012; Berner et al. 2015). The inclusion of SPPT and SKEB in the ECMWF ensemble improved probabilistic skill by reducing the ensemble mean error and by improving reliability over much of the examined 30-day forecast period (Leutbecher et al. 2017).

The main criticism of SKEB and SPPT is that they are applied in an ad hoc manner, rather than by developing and implementing them within the physics schemes to address uncertainty at its source. A third stochastic approach, SPP, targets this shortcoming. SPP is implemented by modifying select (typically uncertain) physical parameters or variables with perturbations that are either fixed in time and/or space (Hacker et al. 2011) or that evolve according to chosen decorrelation time and/or length scales (Bowler et al. 2009; Ollinaho et al. 2017; Jankov et al. 2017; Jankov et al. 2019). Unlike the SPPT and SKEB techniques, SPP directly accounts for uncertainty in individual parameters within the model physics in a physically consistent manner.

The SPP scheme in HRRRE consists of a random pattern generator that creates a vertically uniform perturbation field with prescribed spatiotemporal correlations that can be applied to two-dimensional (using the first level in the perturbation field) and three-dimensional fields. The perturbations are applied to key parameters or variables in multiple parameterization schemes. Parameters, variables, and diagnostics within the boundary layer, surface layer, gravity wave drag, radiation, microphysics, and horizontal diffusion schemes were perturbed in this work. The perturbations were also applied to parameters in the land surface model at the initial time, as described in the appendix. The perturbations were applied to a variety of fields, including fixed parameters, diagnostic variables, and prognostic state variables, but the latter were only perturbed at the initial time to limit the possibility of nonconservation. All other fields were perturbed continuously during model integration. The appendix discusses the specific parameters that were perturbed in each scheme and the rationale for the design of these perturbations.

Different spatial and temporal length scales for the perturbations were tested. We found that prescribing a correlation length scale of 150 km and a temporal scale of 6 h provided the best combination of enhanced ensemble spread without substantially increasing the error of the ensemble mean. These scales were then used to construct all of the perturbed fields. Different perturbation magnitudes were also tested; however, with a parameter/variable space as large as ours (Table 1), only approximate optimal magnitudes have been determined thus far. This testing was typically performed on individual case studies, with attention to both everyday sensible weather, such as 2-m temperature and 10-m wind speed, and also to severe convection/storm structures. The criteria for determining the optimal values were 1) attempt to generate total spread (i.e., ensemble spread plus observation error) in 2-m temperature, 2-m mixing ratio, and 10-m wind speed comparable to the standard deviation of the differences between the observed and ensemble-mean forecasted values of those same variables (after bias removal); 2) attempt to create this ensemble spread without increasing the random error and bias in the ensemble mean forecast; 3) maximize the ensemble spread without excessively perturbing the parameters/variables of interest, which would be unphysical and might result in computational instability; and 4) approximately match the size of the perturbations with the uncertainty in the parameter/variable of interest (if known).

After extensive testing, it was determined that the magnitudes of the perturbations listed in the far-right column of Table 1 were both computationally stable and provided a reasonable ensemble spread without increasing the error in the ensemble mean. We chose to perturb each parameter/variable by up to two standard deviations, which equates to twice the magnitude of the values listed in Table 1. We caution the reader that our choice of perturbation magnitude and sign (i.e., correlation to other perturbations) were chosen in the context of driving ensemble spread for short-term (<24 h) forecast

TABLE 1. List of stochastically perturbed fields, the parameterization schemes they belong to, and information on how the field was perturbed spatially, temporally, and in magnitude.

Field	Host parameterization scheme	Field type	Spatial decorrelation length scale (km)	Temporal scale (hours)	Percent magnitude perturbation (for one std dev)
K_H and K_M	MYNN-EDMF	Diagnostic	150	6	±30%
Background q_v	MYNN-EDMF	Diagnostic	150	6	±10%
Entrainment rate	MYNN-EDMF	Diagnostic	150	6	±10%
σ_H	GWD scheme	Fixed parameter	150	6	±15%
r_{ec} and r_{ei}	RRTMG radiation scheme	Diagnostic	150	6	±20%
c_s	Horizontal diffusion scheme	Fixed parameter	150	6	+50%, −25%
z_t and z_q	MYNN surface layer scheme	Diagnostic	150	6	±10%
z_0	MYNN surface layer scheme	Diagnostic	150	6	±20%
Surface emissivity	RUC LSM	Fixed parameter	150	6	±2%
Surface albedo	RUC LSM	Fixed parameter	150	6	±8%
Soil moisture	RUC LSM	Prognostic state variable	150	6	±30%
Vegetation fraction	RUC LSM	Fixed parameter	150	6	±6.6%
Graupel intercept parameter	Thompson	Diagnostic	150	6	+78%, −44%
Cloud droplet shape parameter	Thompson	Diagnostic	150	6	±1.0 (absolute)
w used in CCN activation	Thompson	Diagnostic	150	6	$+5\% \times w^{-1}$, 0%
IN concentration	Thompson	Diagnostic	150	6	+15%, 0%

applications. This configuration may not be optimal for longer-term forecast applications.

4. Results from HRRRE retrospective tests

NOAA/GSL scientists conducted several retrospective experiments to examine the behavior of SPP, SKEB, and SPPT in HRRRE. One of the most important retrospective periods was 16 July–14 August 2018, which offered an opportunity to examine the model performance during typical summertime convective weather, when synoptic constraints on sensible weather were weak and strong daytime vertical mixing was occurring within the PBL. We also ran a control experiment (HRRRE-CTRL), with no stochastic physics, to provide baseline estimates of HRRRE spread and bias under these conditions. Each set of runs consisted of 18-h forecasts that were initialized once per day at 1200 UTC. While computer resources were insufficient to conduct more varied experiments, results from additional seasons and longer lead times are of interest and may differ from the 0–18-h summer forecasts examined below.

a. *Impact of SPP on HRRRE total spread, random error, and mean bias*

Figure 2 illustrates the effect of SPP on the HRRRE total spread, random error, and ensemble-mean bias as a function of forecast lead time over the eastern United States (results over the entire United States were similar; not shown). Dowell et al. (2004) provide more details on how these metrics were calculated. For all surface variables, the total spread is the same in the two HRRRE configurations at the initial time, increases in HRRRE-SPP relative to the baseline by midmorning, reaches a peak in both configurations in the midafternoon, and then declines in both configurations (but remains greater in HRRRE-SPP than in the baseline) after dark. The inclusion of SPP in HRRRE increases the peak value of the baseline spread in 2-m air temperature and 2-m mixing ratio from 1.6° to 2.0°C (25%; Fig. 2a) and from 1.4 to 1.6 g kg^{-1} (14%; Fig. 2b), respectively. Gains in spread in the 10-m zonal wind speed (similar to the impact on the spread of the 10-m wind speed; not shown) are more modest, with an increase from 1.6 to 1.7 m s^{-1} (6%; Fig. 2c), but evolve similarly during the simulation. While the increases in total spread are not large, they result in a spread that roughly matches the random error present in HRRRE-SPP (blue lines in Fig. 2), a goal of ensemble design. The strong diurnal cycle in the ensemble spread is due to the dominant effect of the PBL perturbations, which are coupled to perturbations in the shortwave radiation scheme. PBL and shortwave radiation processes are greatly reduced or nonexistent at night, leading to a loss of ensemble spread after sunset. The addition of SPP to HRRRE slightly improves the mean bias in 2-m mixing ratio and 10-m wind speed, but has a mixed impact on the 2-m temperature bias (red lines in Fig. 2).

b. *Impact of SPP on HRRRE quantitative precipitation and surrogate severe weather forecasts*

To determine how SPP affects the precipitation and severe weather forecasts produced by the ensemble, HRRRE-CTRL and HRRRE-SPP surrogate severe and probabilistic quantitative precipitation forecasts (PQPF) from the retrospective experiment were evaluated. Thirty time-matched forecasts were compared to NCEP Stage-IV quantitative precipitation estimates, following budget interpolation (Accadia et al. 2003) of the Stage-IV (4.7-km grid spacing) values to the 3-km HRRRE grid. Grid points over water and/or those with missing values in the Stage-IV dataset were excluded from the analysis. Probabilistic forecasts depict the chance of exceeding a threshold within a 40-km (13-gridpoint) radius. Neighborhood probabilities were preferred for this analysis due to the inherently

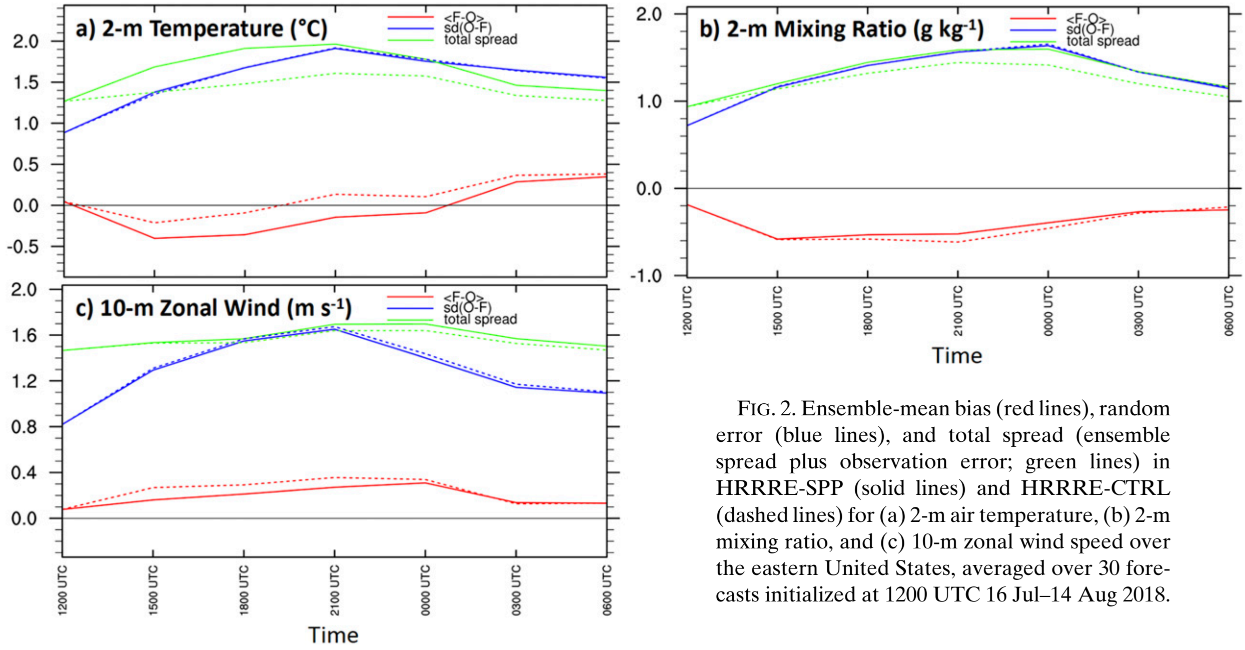

FIG. 2. Ensemble-mean bias (red lines), random error (blue lines), and total spread (ensemble spread plus observation error; green lines) in HRRRE-SPP (solid lines) and HRRRE-CTRL (dashed lines) for (a) 2-m air temperature, (b) 2-m mixing ratio, and (c) 10-m zonal wind speed over the eastern United States, averaged over 30 forecasts initialized at 1200 UTC 16 Jul–14 Aug 2018.

discontinuous nature of short-duration (6-h) precipitation accumulations. A Gaussian filter (width = 25 km; 8 × 8 grid points) was applied to the neighborhood probability fields to smooth unphysically sharp gradients. The choice of neighborhood size and filter width are likely to have some impact on the results that follow, but quantifying this effect is left to future work. In addition, although it is possible and often appropriate to perform bias correction prior to analysis of probabilistic forecast reliability, the authors find that ensemble PQPF products are not bias corrected in most operational and testbed settings, and thus this initial investigation seeks to verify ensemble output in the specific manner in which it is applied in the forecast process.

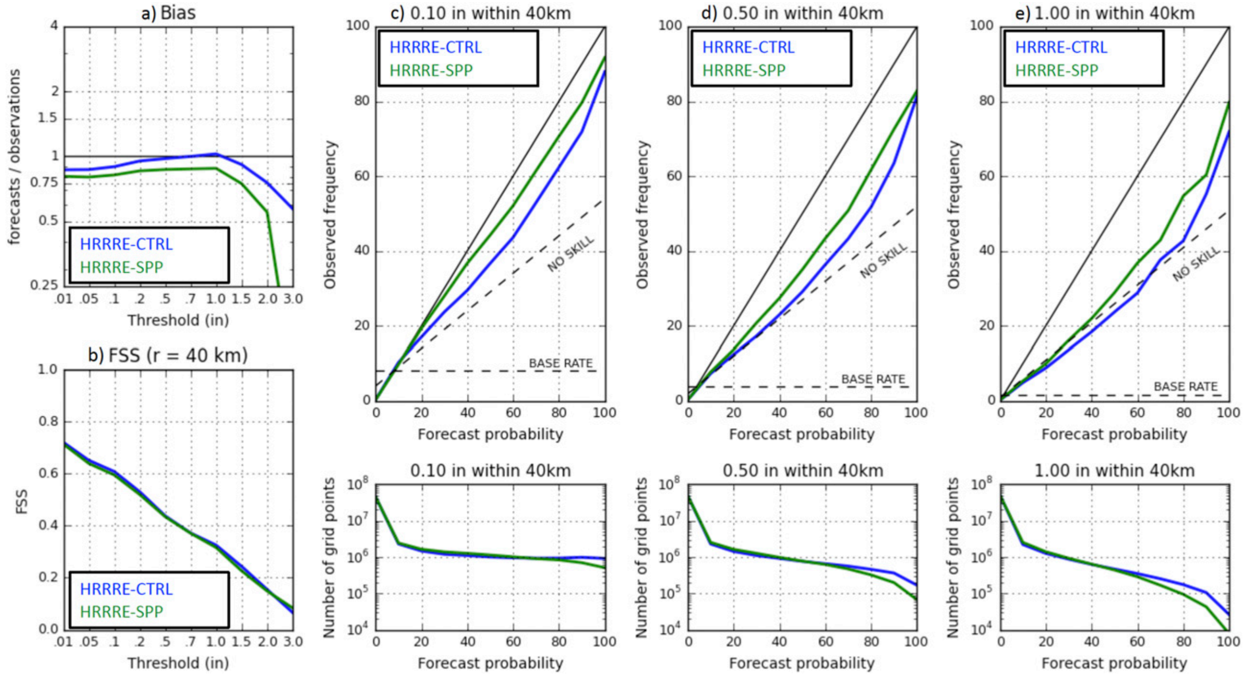

FIG. 3. Verification of HRRRE-CTRL (blue) and HRRRE-SPP (green) probabilistic quantitative precipitation forecasts relative to NCEP Stage-IV quantitative precipitation estimates at the 12–18-h lead time, consisting of (a) frequency bias (forecast/analysis), (b) fractions skill score with radius = 40 km, and reliability diagrams and forecast histograms for thresholds of (c) 0.1, (d) 0.5, and (e) 1.0 in. $(6\,\mathrm{h})^{-1}$ (1 in. = 25.4 mm).

At 12–18-h lead times, HRRRE-SPP generally produces less precipitation than HRRRE-CTRL, and the precipitation forecasts have a low-frequency bias compared to observations (Fig. 3a). This bias is especially evident at precipitation thresholds exceeding 1 in. FSS values for the two ensembles are largely indistinguishable at every threshold (Fig. 3b). Although both HRRRE-CTRL and HRRRE-SPP exhibit overconfident precipitation forecasts, HRRRE-SPP reduces this overconfidence and exhibits improved reliability across all precipitation thresholds (Figs. 3c–e). Compared to HRRRE-CTRL, HRRRE-SPP produces more forecasts with small to moderate probabilities ($<$50%) of exceeding 0.1, 0.5, and 1.0 in. of precipitation within a 40-km radius (Figs. 3c–e). In contrast, the number of high-confidence forecasts (60%–100% probability) is reduced in HRRRE-SPP, especially at the 1-in. threshold. The tendency for HRRRE-SPP to produce lower exceedance probabilities is likely a consequence of the increased spread of its ensemble members, since an increase in spread will make it less likely that two or more ensemble members produce similar forecasts at any particular grid point. This characteristic has a favorable impact on reliability, but it worsens the low bias in precipitation forecasts that was already present in HRRRE-CTRL for areas of heavy ($>$1.0 in.) precipitation. The above results are similar for other 6-hr accumulation intervals (not shown).

In addition to PQPF verification, updraft helicity (UH) forecasts were also evaluated from both ensembles over the 16 July–14 August 2018 period to determine if SPP affects the ability of HRRRE to anticipate the occurrence of severe convection. These ‘‘surrogate severe’’ forecasts and the verification dataset were constructed following the Sobash et al. (2016) method. Verification statistics were calculated using an 80-km grid, where any exceedance of a forecast threshold in a given grid box was considered a forecast of severe convection, and any local storm report [e.g., tornadoes, hail $\geq$ 25 mm in diameter, and/or wind gusts $\geq$ 50 kt (25.7 $\mathrm{m\,s^{-1}}$)] in a grid box was considered an occurrence of severe convection. Forecasts of exceeding updraft helicity thresholds from 25 to 150 $\mathrm{m^2\,s^{-2}}$ were evaluated, but reliability statistics are only presented for 75 $\mathrm{m^2\,s^{-2}}$, both for clarity and because this threshold is commonly used in operations. Binary forecast exceedance grids were smoothed using a Gaussian kernel of 120-km width. As in the PQPF evaluation, an emphasis was placed on verifying ensemble forecasts as they are typically used in operational settings (e.g., using empirically derived exceedance and smoothing thresholds, and without bias correction).

Both HRRRE-CTRL and HRRRE-SPP exhibit strong reliability for surrogate severe weather forecasts that utilize UH $\geq$ 75 $\mathrm{m^2\,s^{-2}}$ as a threshold (Fig. 4a). Both ensembles tend to underpredict (by ~5%) the likelihood of severe events that are relatively low probability (events with an observed frequency of less than 30%), and this tendency is slightly more pronounced in HRRRE-SPP. For events that occur more frequently ($>$40%), both ensembles tend to be overly confident in the event occurrence, but neither ensemble exhibits a clear advantage in terms of reliability. Regarding the forecast histogram (Fig. 4b), the two ensembles produce a similar number of forecasts that imply an event probability of $<$30%. Relative to HRRRE-CTRL, however, HRRRE-SPP is less likely to forecast higher probabilities ($>$30%) of UH $\geq$ 75 $\mathrm{m^2\,s^{-2}}$. This tendency for HRRRE-SPP to produce fewer higher-confidence forecasts of severe weather parallels the reduction in confidence for precipitation forecasts discussed above, and is likely related to the increased spread of its ensemble members.

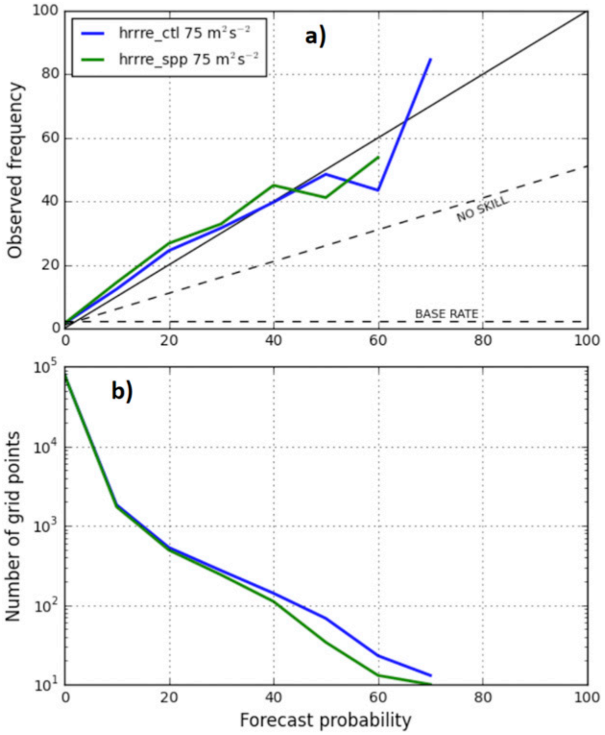

FIG. 4. Verification of 0–18-h HRRRE-CTRL (blue) and HRRRE-SPP (green) updraft helicity forecasts against local storm reports, consisting of the (a) reliability diagram and (b) forecast histogram for the 75 $\mathrm{m^2\,s^{-2}}$ updraft helicity threshold.

c. *Impact of SPP on physical realism of the ensemble*

One of the key questions that we seek to address by implementing stochastic physics in HRRRE is what stochastic approach (i.e., SPP, SKEB, or SPPT) increases ensemble spread in the most physically consistent manner. Constructing joint probability density function distributions of the spread in two related model variables in HRRRE-CTRL and comparing against those from runs with stochastic physics can help to answer this question. Figure 5 is one example of this approach. As the spread in the downward shortwave solar radiation at the surface (SWDOWN) increases from 50 to 150 $\mathrm{W\,m^{-2}}$ in HRRRE-CTRL (Fig. 5a), there is a modest increase in 2-m T spread from 0.5° to 0.75°C. However, once SWDOWN spread increases beyond 150 $\mathrm{W\,m^{-2}}$, additional 2-m T spread is gained at an increasingly rapid rate. This reflects the physical relationship between cloud fraction and 2-m T, which is nonlinear and modulated by the mixed-layer depth, and demonstrates that it is difficult to generate large 2-m T spread without a correspondingly large spread in cloud amount.

A stochasticized ensemble that preserves this relationship (among others) is more physically consistent than one that

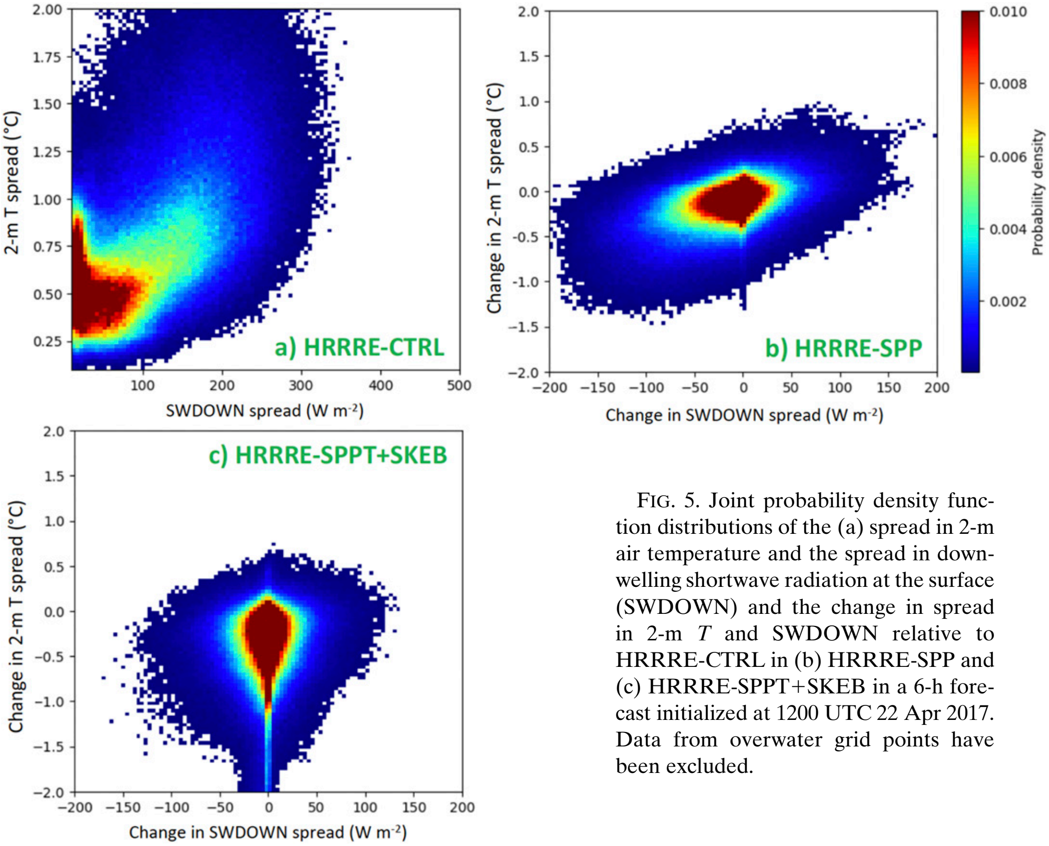

FIG. 5. Joint probability density function distributions of the (a) spread in 2-m air temperature and the spread in downwelling shortwave radiation at the surface (SWDOWN) and the change in spread in 2-m T and SWDOWN relative to HRRRE-CTRL in (b) HRRRE-SPP and (c) HRRRE-SPPT+SKEB in a 6-h forecast initialized at 1200 UTC 22 Apr 2017. Data from overwater grid points have been excluded.

does not. Figure 5b demonstrates that in HRRRE-SPP, the changes in SWDOWN spread and 2-m T spread relative to HRRRE-CTRL are positively correlated, with a coefficient of determination (R^2) of 0.23. This is not the case in HRRRE with SPPT and SKEB (HRRRE-SPPT + SKEB; Fig. 5c), which did not include SPP. In HRRRE-SPPT + SKEB, there is no apparent relationship between the relative change in SWDOWN spread and 2-m T spread ($R^2 = 0.01$). While it is not clear that the relationship between the two should be linear, since Fig. 5a implies an exponential relationship, SPPT + SKEB act to degrade the relationship between SWDOWN and 2-m T spread that is present in HRRRE-CTRL, while HRRRE-SPP comes closer to maintaining this relationship. This is a direct reflection of the physical inconsistency of perturbing model tendencies instead of perturbing the uncertain model variables that impact those tendencies.

5. Results from HRRRE-SPP in the Hazardous Weather Testbed and beyond

In recent years, NOAA/GSL has run HRRRE during the Hazardous Weather Testbed Spring Experiment (HWT-SE), which is described in detail by Clark et al. (2012). Feedback from National Weather Service (NWS) forecasters and other model users in HWT-SE has been incredibly valuable and resulted in HRRRE improvements. The retrospective test results discussed above led to the implementation of HRRRE-SPP for the first time in the 2019 HWT-SE. Collaborators at the National Severe Storms Laboratory (NSSL) also ran a nonstochasticized version of HRRRE (HRRRE-CTRL) in parallel to facilitate continued comparisons between the two systems. Graphics from both versions of HRRRE were produced in real time and uploaded to the HWT-SE website,[2] allowing forecasters to analyze ensemble output from both systems.

Figure 6 provides an example of this comparison for a heavy rainfall event on 1 May 2019 that occurred during HWT-SE. By visual inspection, HRRRE-SPP (Fig. 6b) provided an accurate forecast of this event across the northern half of the observed heavy rainfall region (hatched area in Fig. 6), although it struggled to forecast the event in Oklahoma. The HRRRE-CTRL ensemble (Fig. 6d) also placed high probabilities of heavy rainfall near the observed area, but its forecast was generally too far southeast. While HREFv2 (Figs. 6a,c) also shows the potential for heavy rainfall in the correct location, its probabilities are much smaller than those of HRRRE (~35% versus >50%), reflecting a less confident forecast. Figure 7 shows another example of the visual comparisons available to HWT-SE participants for a severe weather event on 23 May 2019. The ensemble maximum updraft helicity, a quantity that

[2] https://hwt.nssl.noaa.gov/sfe_viewer/2019/model_comparisons/.

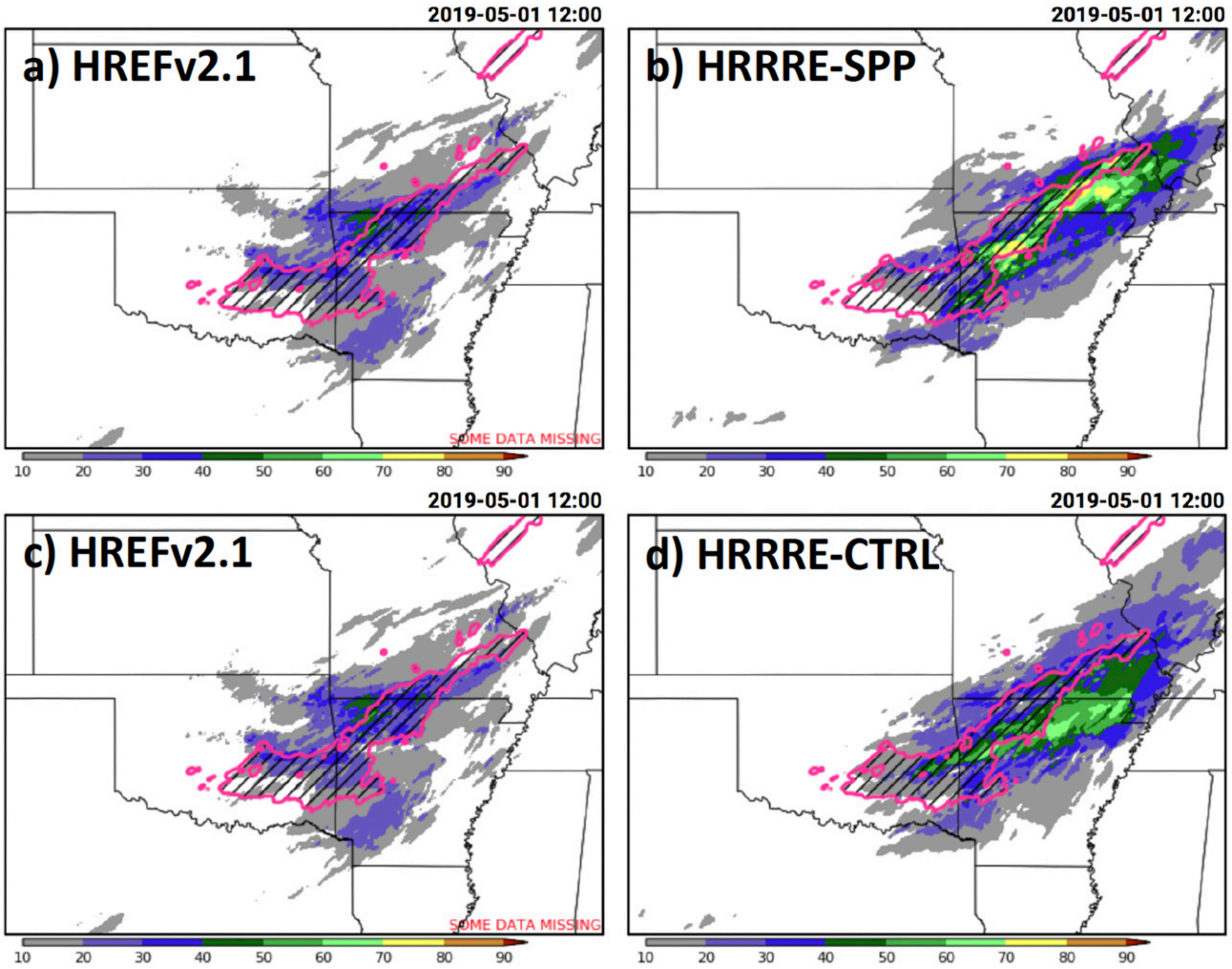

FIG. 6. Observations from the Multi-Radar Multi-Sensor (MRMS) network (hatched) and forecast probabilities (color fill) of 6-h precipitation amount exceeding 1 in. (color fill) from the (a),(c) HREFv2; (b) HRRRE-SPP; and (d) HRRRE-CTRL ensembles at 1200 UTC 1 May 2019. This figure was created from plots available on the HWT-SE website (https://hwt.nssl.noaa.gov/sfe_viewer/2019/model_comparisons/).

is commonly used by forecasters to gauge the spatial extent of a severe weather outbreak, is shown in color fill, with tornado and severe hail reports overlaid. For this event, the HREFv2 (Figs. 7a,c), HRRRE-SPP (Fig. 7b), and HRRRE-CTRL (Fig. 7d) ensembles captured the area of observed severe weather well. However, except for a small area in Kansas and Missouri, HRRRE-SPP displays somewhat lower neighborhood probabilities of 24-h maximum 2–5-km updraft helicity exceeding $75\,\mathrm{m}^2\,\mathrm{s}^{-2}$ than HREFv2, a point that will be explored later.

Despite the aforementioned increases in ensemble spread (Fig. 2) and some individual success stories (e.g., Figs. 6 and 7), subjective evaluation scores from forecasters during HWT-SE (Clark et al. 2019; their Fig. 45) suggested similar overall performance between HRRRE-SPP and HRRRE-CTRL. The median ratings of the two ensemble systems for 0000 UTC initializations were identical (6 out of a possible 10 points), while the mean rating for HRRRE-SPP was 0.25 points lower than that for HRRRE-CTRL. Clark et al. (2019) states that "subjectively, the impact of stochastic physics appeared to lower the storm-attribute probabilities (without noticeably changing the spatial ensemble envelope) by removing/weakening storms."

This feedback, while valuable, was surprising given generally favorable results from the prior retrospective experiments. A complicating factor in the diagnosis of HRRRE-SPP performance during HWT-SE was that the HRRRE and HRRRDAS configurations changed multiple times during the course of the experiment, resulting in a nonhomogeneous sample of model output. Because "HREFv2 performance is considered the baseline against which potential future operational [convection-allowing model] ensemble configurations are compared (Clark et al. 2019)," an evaluation of HRRRE-SPP performance relative to HREFv2 would help assess the ability of HRRRE-SPP to provide useful same-day forecast guidance. While the inhomogeneous nature of the HRRRE-SPP forecasts during HWT-SE made them ineligible for such an analysis, HRRRE-SPP continued to

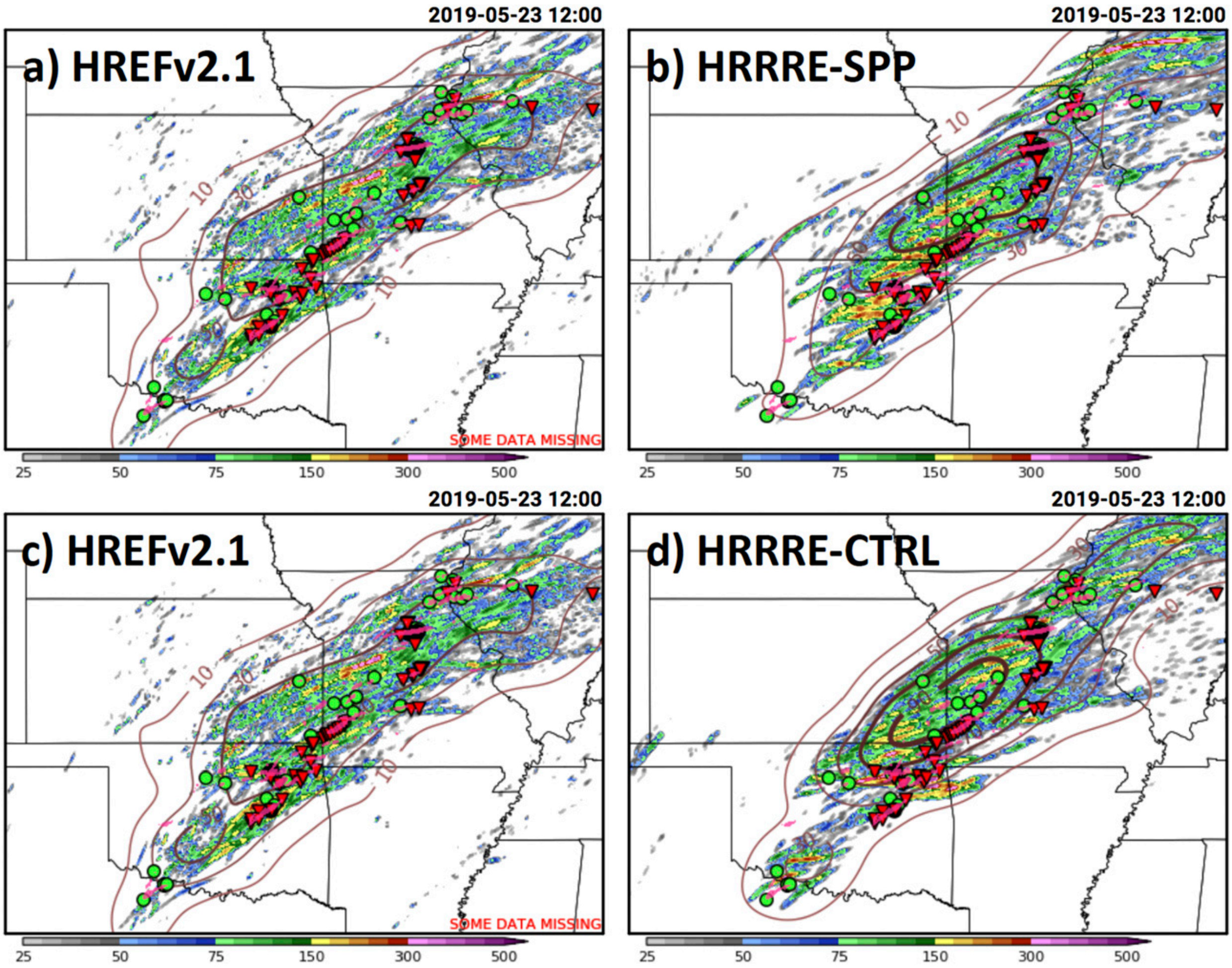

FIG. 7. Ensemble maximum forecasts of 24-h peak 2–5-km updraft helicities (color fill) and probability of the 24-h maximum 2–5-km updraft helicity exceeding 75 $m^2 s^{-2}$ in a 40-km radius (contours) from the (a),(c) HREFv2; (b) HRRRE-SPP; and (d) HRRRE-CTRL ensembles at 1200 UTC 23 May 2019. Tornado (inverted red triangles) and severe hail reports (filled green circles) are also shown. This figure was created from plots available on the HWT-SE website (https://hwt.nssl.noaa.gov/sfe_viewer/2019/model_comparisons/).

produce forecasts in real time after HWT-SE ended, and the forecast output was archived for later analysis.

The aforementioned HRRRE-SPP real-time forecasts included a lengthy period of frozen model configuration (1200 UTC 5 June–1200 UTC 31 August 2019). The HRRRE-SPP and HREFv2 surrogate severe and probabilistic precipitation forecasts from this period were subsequently evaluated using the same methods described in section 4b. The HREFv2 and HRRRE-SPP comparison consisted of 88 time-matched forecasts initialized at 1200 UTC. Prior to performing the comparison, the HREFv2 (3–3.2-km native grid spacing, distributed at 5 km) values were interpolated to the 3-km HRRRE-SPP grid. At the short lead times examined here, especially during the first six hours of the forecast, we acknowledge that HRRRE-SPP has a built-in advantage relative to HREFv2 due to the cycled HRRRDAS, whereas most of the HREFv2 ensemble members are cold-started.

HRRRE-SPP members generally produce less precipitation than HREFv2 members for thresholds up to 25 mm $(6\,h)^{-1}$, as demonstrated by the significant low-frequency bias in HRRRE-SPP for nearly all thresholds, and the near-neutral to high-frequency bias in HREFv2 (Fig. 8a). The low-frequency bias in HRRRE-SPP is modest at hours 0–6, worsens substantially at hours 6–12, and improves at hours 18–24. Both ensembles exhibit the largest biases for moderate-heavy precipitation, i.e., 25–50 mm $(6\,h)^{-1}$, with smaller biases for very light and extremely heavy rates. In spite of the low-frequency bias, HRRRE-SPP FSS values (radius = 40 km) are notably larger (better) than HREFv2 at hours 0–6 for all precipitation thresholds, but are decidedly worse at hours 6–12 and 18–24. Short-range (0–6-h) HREFv2 and longer range (6–12, 18–24-h) HRRRE-SPP probabilistic forecasts of ≥0.25 mm $(6\,h)^{-1}$ (Fig. 8c) exhibit near-perfect reliability, but substantial overconfidence is noted in 0–6-h HRRRE-SPP, and 6–12 and 18–24-h HREFv2 forecasts. Reliability of 2.5 and 25 mm $(6\,h)^{-1}$ forecasts (Figs. 8d,e) is similar and slightly overconfident for both ensembles, except for the significant overconfidence noted in 0–6-h HRRRE-SPP forecasts. HRRRE-SPP performance overall lags

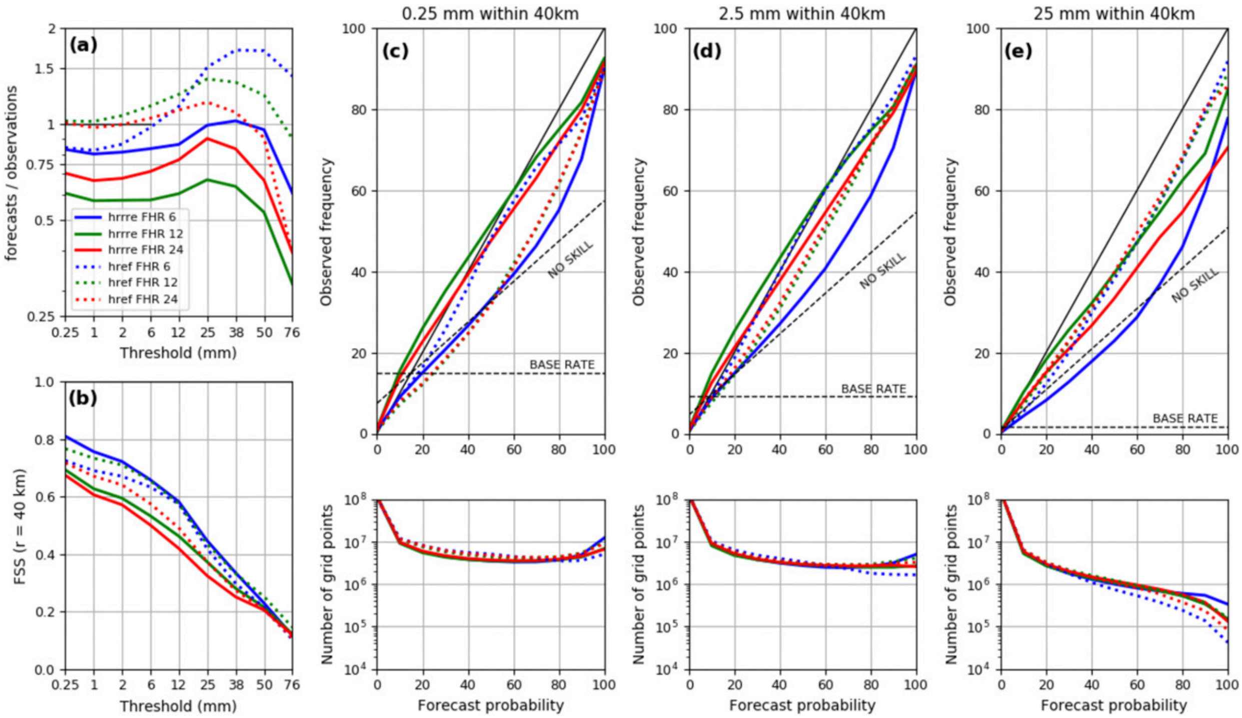

FIG. 8. Verification of HRRRE-SPP (solid) and HREFv2 (dashed) probabilistic quantitative precipitation forecasts relative to NCEP Stage-IV quantitative precipitation estimates at lead times of 0–6, 6–12, and 18–24 h (blue, green, and red, respectively), consisting of (a) frequency bias (forecast/analysis), (b) fractions skill score with radius = 40 km, and reliability diagrams and forecast histograms for thresholds of (c) 0.25, (d) 2.5, and (e) 25 mm $(6\,\mathrm{h})^{-1}$.

slightly behind that of HREFv2, although for several metrics and thresholds, scores from the two systems are quite similar.

Last, updraft helicity (UH) forecasts were evaluated from both ensembles over the 5 June–31 August 2019 frozen configuration period. Due to data availability over the period of interest, only 0–12-h maximum UH forecasts were evaluated. The HREFv2 and HRRRE-SPP 0–12-h probabilistic forecasts of UH $\geq 75\,\mathrm{m^2\,s^{-2}}$ exhibited similar reliability, with the HRRRE-SPP ensemble producing only slightly less confident forecasts than HREFv2 (Fig. 9a). High probabilities were more likely to occur in HREFv2 than in HRRRE-SPP (Fig. 9b). For a given probability, however, forecasts from both systems had a similar likelihood of being associated with an observed event, given the minor differences in reliability.

6. Summary and future work

The current implementation of SPP in HRRRE represents a proof-of-concept for designing a single-dynamic-core, single-physics ensemble. Retrospective experiments demonstrate its ability to increase the ensemble spread in near-term (0–18 h) HRRRE predictions of sensible weather in the summer season, including air temperature, dewpoint temperature, and wind speed near the surface (Fig. 2). These experiments also demonstrate that SPP increases the reliability of near-term HRRRE precipitation forecasts (Figs. 3c–e), but exacerbates a preexisting low-frequency bias in the prediction of heavy (>1.0 in.) rainfall in HRRRE (Fig. 3a). Surrogate severe weather forecasts from HRRRE-SPP have similar reliability (Fig. 4a) to those from HRRRE-CTRL, but the greater spread in HRRRE-SPP leads to fewer higher-confidence (>30%) predictions of severe weather (Fig. 4b).

Real-time experiments during and after HWT-SE 2019 also provide evidence that SPP can be used to obtain reliable forecasts of heavy rainfall (Figs. 8c–e) and severe weather (Fig. 9a). However, they also suggest that HRRRE-SPP, as currently configured, is less likely to produce higher-confidence surrogate severe weather forecasts based on updraft helicity than the HREFv2 operational system (Fig. 9b). Feedback from HWT-SE forecasters indicates that the sharpness of the resulting ensemble forecast must be considered before stochastic techniques, including SPP, can achieve operational success. It remains possible that these goals can be achieved by further refining the particular parameters and variables perturbed using SPP and the magnitudes and decorrelation scales of those perturbations (Table 1 and the appendix), but more work is needed to make this determination. It also would be worthwhile to explore how statistical postprocessing could be applied to ensemble output, including output from HRRRE-SPP, to produce more useful guidance for operations. In general, future ensemble development and/or postprocessing will need to target an optimal combination of near-neutral frequency biases and appropriate spread, i.e., a system comprised of members that, over many forecasts, accurately represent the occurrence and areal coverage of significant weather, and differ from one another such that forecast probability density functions 1) are sharper

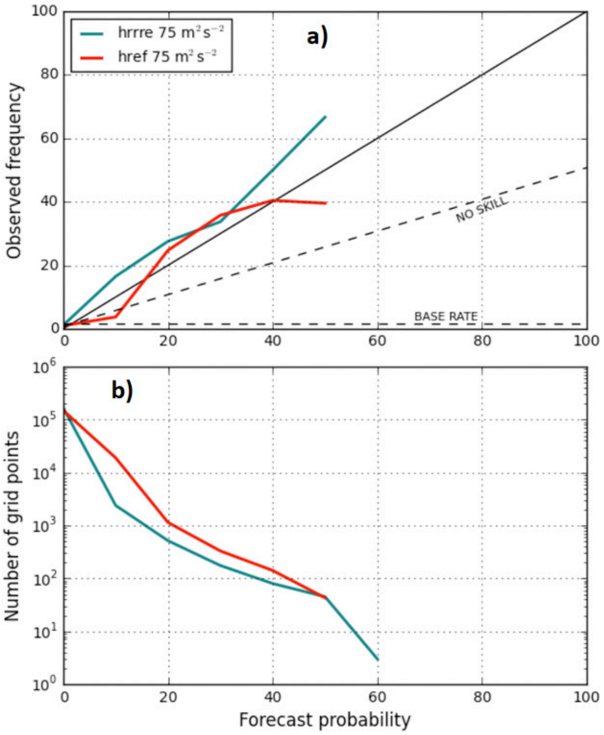

FIG. 9. Verification of 0–12-h HRRRE-SPP (teal) and HREFv2 (red) updraft helicity forecasts against local storm reports, consisting of the (a) reliability diagram and (b) forecast histogram for the 75 m^2 s^{-2} updraft helicity threshold.

than long-term climatological distributions, and 2) result in forecast exceedance probabilities consistent with observed relative frequencies.

Acknowledgments. Evan Kalina is supported by funding from NOAA Awards NA17OAR4590172 and NA17OAR4320101. Isidora Jankov and Jeffrey Beck were supported by funding from NOAA Awards NA17OAR4590181 and NA19OAR4320073. In addition to the authors, many people and organizations have contributed to the development of HRRRE and HRRRDAS. A partial list includes 1) NOAA/GSL: Stan Benjamin, John Brown, Jeff Duda, Guoqing Ge, Jeff Hamilton, Ming Hu, Eric James, Brian Jamison, Terra Ladwig, and Steve Weygandt; 2) NCAR: Glen Romine, Craig Schwartz, Kate Fossell, Ryan Sobash, and May Wong; 3) University of Oklahoma: Yongming Wang and Xuguang Wang; and 4) NSSL: Kent Knopfmeier, Dusty Wheatley, Pat Skinner, and Lou Wicker. We appreciate helpful feedback from John Brown (NOAA/GSL) and three anonymous reviewers, which improved earlier versions of this manuscript. The authors acknowledge the NOAA Research and Development High Performance Computing Program (http://rdhpcs.noaa.gov) for providing computational and storage resources that have contributed to the research results reported within this paper.

Data availability statement. The datasets used in this study have been archived on the High-Performance Storage System (HPSS) at the NOAA Environmental Security Computing Center (NESCC). An existing NOAA High Performance Computing account is required to access HPSS and to obtain the data. The authors are willing to share the location of the data on HPSS with those who have access. Due to the large size of the model output files and the lack of an appropriate public-facing institutional repository, the authors are unable to make the data publicly available at this time.

APPENDIX

More Details about HRRRE-SPP Design

The boundary layer scheme in HRRRE is the Mellor–Yamada–Nakanishi–Niino eddy diffusivity/mass-flux scheme (MYNN-EDMF; Nakanishi and Niino 2006; Nakanishi and Niino 2009; Olson et al. 2019). Within the boundary layer scheme, the eddy diffusivity K_H and eddy viscosity K_M were perturbed. These diagnostic variables govern the local mixing of all momentum and scalar variables in both stable and convective conditions. We chose to perturb K_H and K_M instead of the mixing lengths because the numerical stability constraints in the Mellor–Yamada framework may impose additional limits on the mixing lengths, voiding some fraction of the applied perturbations. The lateral entrainment rates of buoyant plumes in the mass-flux portion of the MYNN-EDMF were also perturbed, allowing the nonlocal transport in convective conditions to vary in strength stochastically. This allows plumes to penetrate higher or terminate lower in the atmosphere, which can affect both the strength of the mixing and the areal coverage of simulated shallow-cumulus clouds. Last, taking advantage of the nonconvective components of the subgrid-scale clouds within the MYNN-EDMF and their interaction with the radiation scheme, the subgrid-cloud fractions and mixing ratios were indirectly perturbed by a direct perturbation of the background water vapor specific humidity q_v. Note that this perturbation to q_v was isolated to the calculation of the subgrid-scale cloud macrophysical properties only and did not impact the magnitude of q_v (or its use) in any other component of the model.

The interaction between the subgrid-scale clouds and the radiation is dependent upon the estimated cloud water and ice effective radii, r_{ec} and r_{ei}, respectively, which are specified in the RRTMG radiation scheme. Smaller (larger) effective radii produce brighter (darker) clouds. The r_{ec} is specified as a constant over land and ocean according to Turner et al. (2007), while the r_{ei} are time-varying diagnostics that are specified according to Mishra et al. (2014). Both r_{ec} and r_{ei} are perturbed with the same spatial and temporal scales as those used in the MYNN-EDMF for perturbing the background q_v, but the sign of the perturbation is opposite with respect to the background q_v perturbations to avoid canceling out the cloud radiative impacts.

The surface exchange coefficients within the MYNN surface layer scheme are indirectly affected by the direct perturbation of aerodynamic, thermal, and moisture roughness lengths. Over land, the aerodynamic roughness lengths z_0 are specified constants according to the prescribed land use categories.

The thermal and moisture roughness lengths, z_t and z_q, respectively, are specified according to Zilitinkevich (1995) and vary according to the Reynolds number and z_0. Over water, z_0, z_t, and z_q are time-varying diagnostics specified according to the COARE 3.0 bulk surface flux algorithm (Fairall et al. 2003). Over both land and water, the constant values of z_0 (over land), the time-varying values of z_0 (over water), z_t, and z_q are perturbed with the same spatial and temporal scales as those used in the MYNN-EDMF and are perturbed with the same sign, so increased (decreased) land-atmosphere coupling is concurrent with increased (decreased) local and nonlocal diffusion in the boundary layer.

The gravity wave drag (GWD) scheme employed in the HRRR uses a topographic form drag (TOFD; Beljaars et al. 2004) and a small-scale gravity wave drag (SSGWD; Steeneveld et al. 2008). Both of these components can be applied down to grid spacing of about 1 km, unlike traditional large-scale GWD schemes, which are meant to represent the impacts of gravity waves launched by terrain wavelengths of order 10–100 km. The TOFD can be active in both stable and unstable conditions, while the SSGWD is only active in the stable boundary layer; thus, both provide an opportunity for stochastic perturbations to impact the stable boundary layer. A supplemental impact of stochastic perturbations is needed to help with ensemble spread in stable conditions because the perturbations implemented in the MYNN-EDMF scheme become much less effective as turbulent mixing diminishes in stable conditions. The behavior of both the TOFD and SSGWD is dependent upon the estimated standard deviation of the subgrid-scale terrain variations s_H, which is a two-dimensional fixed parameter valid for each surface grid cell. The quantity s_H is perturbed with the same spatial and temporal scales as used in the MYNN surface layer scheme (for perturbing z_0) and is perturbed with the same sign as used for the z_0 perturbation, so positive (negative) perturbations act to decelerate (accelerate) the low-level winds.

Including stochastic perturbations to the horizontal diffusion is another way to achieve ensemble forecast spread in both stable and unstable conditions. The Smagorinsky horizontal diffusion scheme (Smagorinsky 1963, 1993), employed in the HRRR, uses a constant known as the Smagorinsky constant c_s, which sets the horizontal length scale proportional to the horizontal grid spacing. The impact of this perturbation is typically secondary with respect to most other perturbations, but it has the ability to impact the strength of resolved- (storm-) scale updrafts and low-level winds in complex terrain. The default value of c_s is set to 0.25. Values of c_s that deviate too far from 0.25 were found to cause infrequent numerical instabilities, so asymmetric (positive/negative) perturbations were applied to stay within numerically stable bounds (see Table 1). The perturbations to c_s used the same spatial and temporal scales as those used in the MYNN-EDMF and also use the same sign, so increased (decreased) horizontal diffusion is concurrent with increased (decreased) vertical diffusion.

Stochastic perturbations were also added to fields in the RUC LSM (Smirnova et al. 2016), which ultimately provide the lower boundary conditions to drive the boundary layer scheme and resolved-scale convection. Perturbations were added to the surface emissivity, albedo, and vegetation fraction, which are all constant two-dimensional surface fields, so long as the surface characteristics do not change during the forecast (i.e., when snow cover accumulates or melts away). The perturbations to all three fields were assigned an opposite sign to the diffusion (K_H and K_M) and roughness length perturbations, so larger surface sensible heat fluxes would be concurrent with larger diffusion and land-atmosphere coupling. Perturbations were also added to the soil moisture, but these perturbations were only applied at the first model time step, since further applying perturbations to a prognostic state variable during the forecast would violate conservation of moisture. Furthermore, no attempt was made to correlate the soil-moisture perturbations with other perturbations, since the soil-moisture perturbations did not coevolve during the model forecast.

The perturbations added to the Thompson aerosol-aware microphysics scheme include the intercept parameter for the graupel size distribution, the shape parameter for the cloud droplet size distribution, the concentration of activated ice nuclei (IN), and the vertical velocity used to compute the concentration of activated cloud condensation nuclei (CCN). Because the microphysical processes being perturbed are largely independent from the processes being perturbed in the land surface and planetary boundary layer parameterizations, no attempt was made to study the impact of correlating or anticorrelating these perturbations.

REFERENCES

Accadia, C., S. Mariani, M. Casaioli, A. Lavagnini, and A. Speranza, 2003: Sensitivity of precipitation forecast skill scores to bilinear interpolation and a simple nearest-neighbor average method on high-resolution verification grids. *Wea. Forecasting*, **18**, 918–932, https://doi.org/10.1175/1520-0434(2003)018<0918:SOPFSS>2.0.CO;2.

AMS, 2002: Enhancing weather information with probability forecasts. *Bull. Amer. Meteor. Soc.*, **83**, 450–452, https://www.ametsoc.org/index.cfm/ams/about-ams/ams-statements/statements-of-the-ams-in-force/enhancing-weather-information-with-probability-forecasts/.

Beljaars, A. C. M., A. R. Brown, and N. Wood, 2004: A new parametrization of turbulent orographic form drag. *Quart. J. Roy. Meteor. Soc.*, **130**, 1327–1347, https://doi.org/10.1256/qj.03.73.

Benjamin, S., 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.

Bowler, N. E., A. Arribas, S. E. Beare, K. R. Mylne, and G. J. Shutts, 2009: The local ETKF and SKRB: Upgrades to the MOGERPS 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.

Clark, A., and Coauthors, 2019: Spring forecasting experiment 2019: Preliminary findings and results. NOAA, 77 pp., https://hwt.nssl.noaa.gov/sfe/2019/docs/HWT_SFE_2019_Prelim_Findings_FINAL.pdf.

Clark, A. J., and Coauthors, 2012: An overview of the 2010 Hazardous Weather Testbed Experimental Forecast Program Spring Experiment. *Bull. Amer. Meteor. Soc.*, **93**, 55–74, https://doi.org/10.1175/BAMS-D-11-00040.1.

Derber, J., and W.-S. Wu, 1998: The use of TOVS cloud-cleared radiances in the NCEP SSI analysis system. *Mon. Wea. Rev.*, **126**, 2287–2299, https://doi.org/10.1175/1520-0493(1998)126<2287:TUOTCC>2.0.CO;2.

Dowell, D. C., F. Zhang, L. J. Wicker, C. Snyder, and N. A. Crook, 2004: Wind and temperature retrievals in the 17 May 1981 Arcadia, Oklahoma, supercell: Ensemble Kalman filter experiments. *Mon. Wea. Rev.*, **132**, 1982–2005, https://doi.org/10.1175/1520-0493(2004)132<1982:WATRIT>2.0.CO;2.

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.

Fairall, C. W., E. F. Bradley, J. E. Hare, A. A. Grachev, and J. B. Edson, 2003: Bulk parameterization of air–sea fluxes: Updates and verification for the COARE algorithm. *J. Climate*, **16**, 571–591, https://doi.org/10.1175/1520-0442(2003)016<0571:BPOASF>2.0.CO;2.

Grimit, E. P., and C. F. Mass, 2002: Initial results of a mesoscale short-range ensemble forecasting system over the Pacific Northwest. *Wea. Forecasting*, **17**, 192–205, https://doi.org/10.1175/1520-0434(2002)017<0192:IROAMS>2.0.CO;2.

Hacker, J. P., and Coauthors, 2011: 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., and S. J. Colucci, 1997: Verification of Eta-RSM short-range ensemble forecasts. *Mon. Wea. Rev.*, **125**, 1312–1327, https://doi.org/10.1175/1520-0493(1997)125<1312:VOERSR>2.0.CO;2.

Harris, L. M., and S.-J. Lin, 2013: A two-way nested global-regional dynamical core on the cubed-sphere grid. *Mon. Wea. Rev.*, **141**, 283–306, https://doi.org/10.1175/MWR-D-11-00201.1.

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.

——, J. Beck, J. Wolff, M. Harrold, J. B. Olson, T. Smirnova, C. Alexander, and J. Berner, 2019: Stochastically perturbed parameterizations in an HRRR-based ensemble. *Mon. Wea. Rev.*, **147**, 153–173, https://doi.org/10.1175/MWR-D-18-0092.1.

Keune, J., C. Ohlwein, and A. Hense, 2014: Multivariate probabilistic analysis and predictability of medium-range ensemble weather forecasts. *Mon. Wea. Rev.*, **142**, 4074–4090, https://doi.org/10.1175/MWR-D-14-00015.1.

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.*, **102**, 409–418, https://doi.org/10.1175/1520-0493(1974)102<0409:TSOMCF>2.0.CO;2.

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.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.*, **20**, 130–141, https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

Mishra, S., D. L. Mitchell, D. D. Turner, and R. P. Lawson, 2014: Parameterization of ice fall speeds in midlatitude cirrus: Results from SPartICus. *J. Geophys. Res. Atmos.*, **119**, 3857–3876, https://doi.org/10.1002/2013JD020602.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: Methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122**, 73–119, https://doi.org/10.1002/qj.49712252905.

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

——, and ——, 2009: Development of an improved turbulence closure model for the atmospheric boundary layer. *J. Meteor. Soc. Japan*, **87**, 895–912, https://doi.org/10.2151/jmsj.87.895.

NCEP, 2004: SSI analysis system 2004. NOAA/NCEP/Environmental Modeling Center Office Note 443, 11 pp., https://www.emc.ncep.noaa.gov/officenotes/newernotes/on443.pdf.

NRC, 2003: *Communicating Uncertainties in Weather and Climate Information: A Workshop Summary*. National Academies Press, 68 pp.

——, 2006: *Completing the Forecast: Characterizing and Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts*. National Academies Press, 124 pp.

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

Olson, J. B., J. S. Kenyon, W. A. Angevine, J. M. Brown, M. Pagowski, and K. Suselj, 2019: A description of the MYNN-EDMF scheme and coupling to other components in WRF-ARW. NOAA Tech. Memo OAR GSD-61, 37 pp., https://repository.library.noaa.gov/view/noaa/19837.

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.

——, and Coauthors, 2009: Stochastic parameterization and model uncertainty. ECMWF Tech. Memo 598, 44 pp., https://www.ecmwf.int/file/23491/download?token=FkJcBDiw.

Parrish, D., and J. Derber, 1992: The National Meteorological Center's spectral statistical-interpolation analysis system. *Mon. Wea. Rev.*, **120**, 1747–1763, https://doi.org/10.1175/1520-0493(1992)120<1747:TNMCSS>2.0.CO;2.

Roberts, B., I. L. Jirak, A. J. Clark, S. J. Weiss, and J. S. Kain, 2019: Postprocessing and visualization techniques for convection-allowing ensembles. *Bull. Amer. Meteor. Soc.*, **100**, 1245–1258, https://doi.org/10.1175/BAMS-D-18-0041.1.

Sanchez, C., K. D. Williams, and M. Collins, 2015: 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.

Schwartz, C. S., G. S. Romine, R. A. Sobash, K. R. Fossell, and M. L. Weisman, 2019: NCAR's real-time convection-allowing ensemble project. *Bull. Amer. Meteor. Soc.*, **100**, 321–343, https://doi.org/10.1175/BAMS-D-17-0297.1.

Smagorinsky, J., 1963: General circulation experiments with the primitive equations. I: The basic experiment. *Mon. Wea. Rev.*, **91**, 99–164, https://doi.org/10.1175/1520-0493(1963)091<0099:GCEWTP>2.3.CO;2.

——, 1993: Some historical remarks on the use of nonlinear viscosities. *Large Eddy Simulation of Complex Engineering and Geophysical Flows*, B. Galperin and S. A. Orszag, Eds., Cambridge Press, 3–36.

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.

Sobash, R., C. Schwartz, G. Romine, K. Fossell, and M. Weisman, 2016: Severe weather prediction using storm surrogates from an ensemble forecasting system. *Wea. Forecasting*, **31**, 255–271, https://doi.org/10.1175/WAF-D-15-0138.1.

Steeneveld, G. J., A. A. M. Holtslag, C. J. Nappo, B. J. H. van de Wiel, and L. Mahrt, 2008: Exploring the possible role of small-scale terrain drag on stable boundary layers over land. *J. Appl. Meteor. Climatol.*, **47**, 2518–2530, https://doi.org/10.1175/2008JAMC1816.1.

Stensrud, D. J., J.-W. Bao, and T. T. Warner, 2000: Using initial condition and model physics perturbations in short-range ensemble simulations of mesoscale convective systems. *Mon. Wea. Rev.*, **128**, 2077–2107, https://doi.org/10.1175/1520-0493(2000)128<2077:UICAMP>2.0.CO;2.

Teixeira, J., and C. A. Reynolds, 2008: Stochastic nature of physical parameterizations in ensemble prediction: A stochastic convection approach. *Mon. Wea. Rev.*, **136**, 483–496, https://doi.org/10.1175/2007MWR1870.1.

Turner, D. D., and Coauthors, 2007: Thin liquid water clouds: Their importance and our challenge. *Bull. Amer. Meteor. Soc.*, **88**, 177–190, https://doi.org/10.1175/BAMS-88-2-177.

Whitaker, J. S., and T. M. Hamill, 2012: Evaluating methods to account for system errors in ensemble data assimilation. *Mon. Wea. Rev.*, **140**, 3078–3089, https://doi.org/10.1175/MWR-D-11-00276.1.

Zilitinkevich, S., 1995: Non-local turbulent transport: Pollution dispersion aspects of coherent structure of convective flows. *Air Pollution Theory and Simulation, Air Pollution III,* H. Power, N. Moussiopoulos, and C. A. Brebbia, Eds., Vol. I, Computational Mechanics Publications, 53–60.