Spread and Skill in Mixed- and Single-Physics Convection-Allowing Ensembles

ERIC D. LOKEN

Cooperative Institute for Mesoscale Meteorological Studies, and School of Meteorology, University of Oklahoma, and NOAA/OAR National Severe Storms Laboratory, Norman, Oklahoma

ADAM J. CLARK

NOAA/OAR National Severe Storms Laboratory, Norman, Oklahoma

MING XUE

School of Meteorology, and Center for Analysis and Prediction of Storms, University of Oklahoma, Norman, Oklahoma

FANYOU KONG

Center for Analysis and Prediction of Storms, University of Oklahoma, Norman, Oklahoma

(Manuscript received 1 May 2018, in final form 12 November 2018)

ABSTRACT

Spread and skill of mixed- and single-physics convection-allowing ensemble forecasts that share the same set of perturbed initial and lateral boundary conditions are investigated at a variety of spatial scales. Forecast spread is assessed for 2-m temperature, 2-m dewpoint, 500-hPa geopotential height, and hourly accumulated precipitation both before and after a bias-correction procedure is applied. Time series indicate that the mixed-physics ensemble forecasts generally have greater variance than comparable single-physics forecasts. While the differences tend to be small, they are greatest at the smallest spatial scales and when the ensembles are not calibrated for bias. Although *differences* between the mixed- and single-physics ensemble variances are smaller for the larger spatial scales, variance *ratios* suggest that the mixed-physics ensemble generates more spread relative to the single-physics ensemble at larger spatial scales. Forecast skill is evaluated for 2-m temperature, dewpoint temperature, and bias-corrected 6-h accumulated precipitation. The mixed-physics ensemble generally has lower 2-m temperature and dewpoint root-mean-square error (RMSE) compared to the single-physics ensemble. However, little difference in skill or reliability is found between the mixed- and single-physics bias-corrected precipitation forecasts. Overall, given that mixed- and single-physics ensembles have similar spread and skill, developers may prefer to implement single- as opposed to mixed-physics convection-allowing ensembles in future operational systems, while accounting for model error using stochastic methods.

1. Introduction

Over the past decade, advances in computing power have enabled numerical weather prediction (NWP) forecasts from fine-resolution convection-allowing ensembles. As early as 2007, the Center for Analysis and Prediction of Storms (CAPS) began running an experimental 10-member, 33-h ensemble with 4-km grid spacing over the contiguous United States (CONUS) to facilitate the prediction of severe weather during the 2007 NOAA Hazardous Weather Testbed Spring Forecasting Experiment (HWT SFE; Xue et al. 2007). This convection-allowing ensemble produced skillful and useful forecasts of composite reflectivity, accumulated precipitation, and probability of precipitation (Xue et al. 2007; Schwartz et al. 2010; Clark et al. 2009). More recent HWT SFEs have studied aspects of convection-allowing ensemble design using controlled experiments based on subsets of the Community Leveraged Unified Ensemble (CLUE; Clark et al. 2018). HWT SFEs have also examined various applications of convection-allowing ensembles, including their use to create probabilistic all-hazards severe weather forecast guidance (Kain et al. 2008; Sobash et al. 2011), tornado

Corresponding author: Eric D. Loken, eric.d.loken@noaa.gov

DOI: 10.1175/WAF-D-18-0078.1

pathlength forecasts (Clark et al. 2013), and probabilistic tornado (Gallo et al. 2016) and hail (Gagne et al. 2017; Adams-Selin and Ziegler 2016) forecasts. Ultimately, the work done in past HWT SFEs led to the implementation of the High Resolution Ensemble Forecast system version 2 (HREFv2; Clark et al. 2017) as the first operational convection-allowing ensemble in the fall of 2017.

In general, ensembles can offer benefits over deterministic models because they account for uncertainties in initial conditions (ICs) and model physics (e.g., Roebber et al. 2004; Leutbecher and Palmer 2008; Clark et al. 2009). Convection-allowing ensembles show unique promise because they not only account for these uncertainties, but each of their members is able to explicitly simulate convection, which has been shown to result in better predictions of convective mode and evolution (e.g., Kain et al. 2006; Done et al. 2004). Indeed, while it has long been known that ensemble mean forecasts tend to outperform forecasts from similarly configured deterministic models at convection-parameterizing resolution (e.g., Epstein 1969; Leith 1974; Clark et al. 2009), recent evidence suggests that convection-allowing ensembles tend to outperform deterministic models at convection-allowing resolution as well (e.g., Coniglio et al. 2010; Loken et al. 2017; Schwartz et al. 2017).

Despite the promise of convection-allowing ensembles, much is still unknown about their optimal configuration (e.g., Roebber et al. 2004; Romine et al. 2014; Duda et al. 2014; Johnson and Wang 2017). One problem is that the vast majority of convection-allowing ensembles are underdispersive [i.e., observed events routinely fall outside of the forecast probability density function (PDF)], especially for precipitation fields (e.g., Clark et al. 2008, 2010; Romine et al. 2014). Many previous studies have investigated methods to increase ensemble spread at convective-parameterizing resolutions, including perturbing initial conditions (e.g., Toth and Kalnay 1993, 1997; Molteni et al. 1996) and using multiple models (e.g., Wandishin et al. 2001; Hou et al. 2001; Ebert 2001; Eckel and Mass 2005) and physics parameterizations (e.g., Stensrud et al. 2000; Gallus and Bresch 2006). More recent work has studied the impact of incorporating multiple planetary boundary layer (PBL) and/or microphysics schemes within convection-allowing ensembles (e.g., Schwartz et al. 2010; Duda et al. 2014; Johnson and Wang 2017), generally finding that mixed-microphysics and mixed-PBL ensembles result in improved ensemble spread and skill. For example, during the 2015 Plains Elevated Convection at Night (PECAN) experiment, Johnson and Wang (2017) found that both of two mixed-physics convection-allowing ensembles—which used a variety of microphysics and PBL schemes—produced better nocturnal precipitation and nonprecipitation forecasts compared to a single-physics ensemble, which used Thompson microphysics (Thompson et al. 2004) and the Mellor–Yamada–Nakanishi–Niino (MYNN; Nakanishi 2000, 2001; Nakanishi and Niino 2004, 2006) PBL. The mixed-physics ensembles in Johnson and Wang (2017) generally produced better subjective forecasts of nocturnal convection as well: relative to the single-physics ensemble, they reduced nocturnal mesoscale convective system (MCS) location errors, produced improved storm structures in nocturnal initiating convection, and had more members forecast observed nocturnal convective initiation. That multiple microphysics and PBL parameterizations can improve forecasts related to convection is unsurprising; previous research has found simulated thunderstorms to be quite sensitive to microphysics parameterizations (e.g., Gilmore et al. 2004; van den Heever and Cotton 2004; Snook and Xue 2008). However, it is currently unknown—especially for convective-allowing ensembles—whether the benefits of using multiple microphysics and PBL parameterizations are apparent only at relatively small spatial scales. Given that larger spatial scales are associated with greater predictability (Lorenz 1969), it is possible that accounting for the uncertainties in modeled microphysics and PBL may matter less for larger spatial scales, where predictability is already relatively high. For example, it is possible that, while a mixed-physics ensemble improves the precise placement of forecast convective systems and produces better forecasts of storm structure, the overall forecasts (i.e., the general location of forecast precipitation-producing systems) provided by a mixed- and single-physics convection-allowing ensemble may not be drastically different at synoptic (or larger meso) scales. It is also possible that the relative benefits (i.e., superior forecast spread and skill) of using multiple microphysics and PBL parameterizations may depend on the variable of interest (e.g., mass-related or low-level variables; Clark et al. 2010) and/or forecast hour/time of day. Given that ensembles with only one microphysics and one PBL scheme are easier for model developers to maintain, it is important to determine if and when a single-physics convection-allowing ensemble can perform nearly as well as a mixed-physics ensemble.

For this task, the present study uses data from the 2016 CLUE (Clark et al. 2018), a collection of 65 ensemble members with similar specifications and postprocessing methods contributed by a variety of organizations [e.g., the National Severe Storms Laboratory (NSSL), CAPS, the University of North Dakota, NOAA's Earth Systems Research Laboratory/Global Systems Division (ESRL/GSD), and the National Center for Atmospheric Research (NCAR)] during the 2016 HWT SFE. Forecast spread (i.e., ensemble variance) is analyzed for 2-m temperature, 2-m dewpoint temperature, 500-hPa geopotential height, and hourly accumulated precipitation at a variety of spatial scales; forecast skill is evaluated for hourly and 6-h accumulated precipitation. Up to 36-h forecasts are considered.

The remainder of this paper is organized as follows: section 2 details the methods used, section 3 presents the results, section 4 examines ensemble forecasts in four cases, section 5 summarizes and discusses the results, and section 6 concludes the paper by considering implications for ensemble design and offering suggestions for future work.

2. Methods

a. Dataset

The 65-member CLUE was run for 24 days during the 2016 NOAA HWT SFE, which spanned from early May to early June. Herein, 36-h forecast data from two 2016 CLUE subsets are analyzed for 23 days of the 2016 NOAA HWT SFE (Table 1; note that 24 May 2016 is excluded from analysis since not all members had data available on that day). Subjective analysis of archived radar reflectivity data suggests that this 23-day analysis period contained a mixture of strongly and weakly forced convective events and both discrete and linear convective modes; however, the dataset included slightly more strongly than weakly forced events and slightly more linear than discrete dominant convective modes.

The two ensemble subsets examined include a 9-member CAPS subset with multiple microphysics and PBL schemes (henceforth referred to as the mixed-physics ensemble) and a 10-member CAPS subset with only Thompson microphysics and the Mellor–Yamada–Janjić (MYJ; Mellor and Yamada 1982; Janjić 2002) PBL scheme (henceforth referred to as the single-physics ensemble). While such small ensembles provide less than optimal sampling of the forecast PDF, previous research (e.g., Clark et al. 2011; Schwartz et al. 2014) suggests that even relatively small ensembles (i.e., 10–20 members) can provide skillful precipitation forecasts. All members from both ensemble subsets use 3-km horizontal grid spacing over a domain covering the CONUS, although the analysis domain is restricted to the eastern 2/3 of the CONUS (Fig. 1). Further, all members contain 1680 grid points in the east–west direction and 1152 grid points in the north–south direction, have perturbed initial and lateral boundary conditions (LBCs), and use the Noah land surface model (Chen and Dudhia 2001) and the Advanced Research Weather Research and Forecasting dynamic core (Skamarock et al. 2008). Initialization for all members is done on weekdays using analyses from the 0000 UTC 12-km North American Mesoscale Model (NAM). Radar (WSR-88D) data, and surface and upper-air observations are assimilated using the Advanced Regional Prediction System three-dimensional variational data assimilation and cloud analysis system (ARPS 3DVAR; Xue et al. 2003; Gao et al. 2004; Clark et al. 2016). Specifications for both ensemble subsets are summarized in Table 2. Notably, both the mixed- and single-physics ensembles use one common member (core01), since it is the control member of both subsets. Further, the mixed-physics ensemble contains 9 members instead of 10 since data from core02 were unavailable throughout the analysis period. However, preliminary tests (not shown) indicate the results presented herein are similar whether the 9-member mixed-physics ensemble is compared against a 10- or 9-member (with s-phys-rad06 excluded) single-physics ensemble.

TABLE 1. Dates from the 2016 NOAA HWT SFE included in the dataset.

Month	Day
May	2–6; 9–13; 16–20; 23; 25–27; 30–31
June	2–3

b. Evaluating ensemble spread

1) ENSEMBLE VARIANCE

To determine ensemble spread, forecast ensemble variance is computed for four variables—2-m temperature, 2-m dewpoint temperature, 500-hPa geopotential height, and hourly accumulated precipitation—for forecast hours 0–36 using Eq. (B7) in Eckel and Mass (2005):

$$\text{Variance} = \frac{1}{M}\sum_{m=1}^{M}\left[\frac{1}{(n-1)}\sum_{i=1}^{n}(e_{m,i}-\overline{e}_m)^2\right], \qquad (1)$$

where M is the number of forecast–observation data pairs (which, here, includes the number of nonoverlapping spatial windows in the domain over each of the 23 days in the analysis), n is the number of ensemble members, $e_{m,i}$ is the value of the ith ensemble member at m, and $\overline{e}_m$ is the ensemble mean at m. To assess the impact of spatial scale, variance is calculated for square neighborhoods of varying sizes using the "upscaling" method (Ebert 2009), which assigns the mean of the

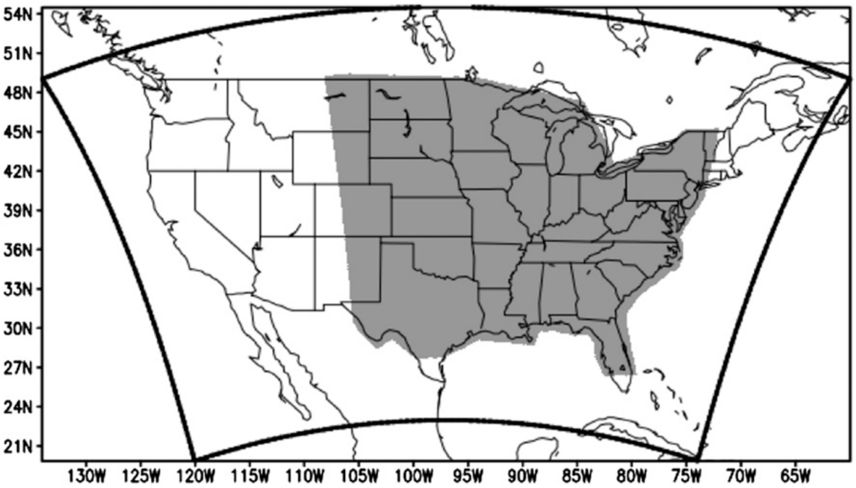

FIG. 1. 2016 CLUE domain (black contour) and analysis domain (gray shading).

finer-resolution grid boxes making up a given neighborhood to that neighborhood.

While a variety of neighborhood sizes from 3 to 720 km are analyzed, only five sizes are displayed herein. These neighborhoods contain 1, 8, 24, 48, and 96 grid boxes per side. Since all ensemble members operate at 3-km horizontal grid spacing, the five neighborhoods measure 3, 24, 72, 144, and 288 km, respectively, on each side. Only neighborhoods falling completely within the analysis domain are included in the variance calculations, and the ''upscale'' averaging is done prior to computing the ensemble mean. The difference between the mixed- and single-physics ensemble variance (i.e., mixed-physics variance minus single-physics variance) and the ratio of single-physics ensemble variance to mixed-physics ensemble variance

TABLE 2. Mixed- and single-physics ensemble member specifications (adapted from Clark et al. 2016, 2018). A superscript ''a'' denotes use in the mixed-physics ensemble, while a superscript ''b'' denotes use in the single-physics ensemble. NAMa and NAMf denote the 12-km NAM analysis and forecast, respectively. Here, 3DVAR refers to the ARPS three-dimensional variational data assimilation and cloud analysis (Xue et al. 2003; Gao et al. 2004). Elements in the IC column ending with ''pert'' are perturbations extracted from a 16-km grid-spacing 3-h Short-Range Ensemble Forecast (SREF; Du et al. 2014) member. Elements in the LBC column after the first row refer to SREF member forecasts. Ensemble microphysics schemes include Thompson (Thompson et al. 2004), Predicted Particle Properties (P3; Morrison and Milbrandt 2015), Milbrandt and Yau (MY; Milbrandt and Yau 2005), and Morrison (Morrison et al. 2005). Ensemble boundary layer schemes include MYJ (Mellor and Yamada 1982; Janjić 2002), Yonsei University (YSU; Noh et al. 2003), and MYNN (Nakanishi 2000, 2001; Nakanishi and Niino 2004, 2006).

Ensemble member	IC	LBC	Microphysics	PBL
core01[a,b]	NAMa+3DVAR	NAMf	Thompson	MYJ
core03[a]	core01 + arw-p1_pert	arw-p1	P3	YSU
core04[a]	core01 + arw-n1_pert	arw-n1	MY	MYNN
core05[a]	core01 + arw-p2_pert	arw-p2	Morrison	MYJ
core06[a]	core01 + arw-n2_pert	arw-n2	P3	YSU
core07[a]	core01 + nmmb-p1_pert	nmmb-p1	MY	MYNN
core08[a]	core01 + nmmb-n1_pert	nmmb-n1	Morrison	YSU
core09[a]	core01 + nmmb-p2_pert	nmmb-p2	P3	MYJ
core10[a]	core01 + nmmb-n2_pert	nmmb-n2	Thompson	MYNN
s-phys-rad02[b]	core01 + arw-p1_pert	arw-p1	Thompson	MYJ
s-phys-rad03[b]	core01 + arw-n1_pert	arw-n1	Thompson	MYJ
s-phys-rad04[b]	core01 + arw-p2_pert	arw-p2	Thompson	MYJ
s-phys-rad05[b]	core01 + arw-n2_pert	arw-n2	Thompson	MYJ
s-phys-rad06[b]	core01 + arw-p3_pert	arw-p3	Thompson	MYJ
s-phys-rad07[b]	core01 + nmmb-p1_pert	nmmb-p1	Thompson	MYJ
s-phys-rad08[b]	core01 + nmmb-n1_pert	nmmb-n1	Thompson	MYJ
s-phys-rad09[b]	core01 + nmmb-p2_pert	nmmb-p2	Thompson	MYJ
s-phys-rad10[b]	core01 + nmmb-n2_pert	nmmb-n2	Thompson	MYJ

(i.e., single-physics variance/mixed-physics variance) are also computed.

Because systematic biases from each ensemble member contribute to forecast spread but not to forecast uncertainty (since systematic biases are not uncertain; e.g., Eckel and Mass 2005; Clark et al. 2010, 2011), a probability matching technique (Ebert 2001; Clark et al. 2010) is used to eliminate systematic biases among the ensemble members. Conceptually, this technique assigns the PDF of one dataset to another dataset to eliminate systematic ensemble biases. Herein, because the core01 member serves as the control member of both the mixed- and single-physics ensemble, the PDF of the core01 member is assigned to each of the other ensemble members. This is done by first sorting each member's forecast precipitation values from all grid points on a given day and forecast hour from largest to smallest. Then, for each member, the grid point containing the largest forecast precipitation value is replaced with the largest forecast value from the core01 member, and so on until all of the values have been replaced. In this way, the spatial patterns of each member's original forecasts are maintained, but the amplitudes of each member's forecast are replaced with amplitudes from the core01 member (e.g., Clark et al. 2010). Hence, after probability matching, all ensemble members contain the same bias (i.e., the bias of the core01 member) for a given forecast hour on a given day, where bias is defined by

$$\text{bias} = \frac{\frac{1}{N}\sum_{i=1}^{N} F_i}{\frac{1}{N}\sum_{i=1}^{N} O_i} = \frac{\sum_{i=1}^{N} F_i}{\sum_{i=1}^{N} O_i}, \qquad (2)$$

where N is the number of grid points within the analysis domain, F_i is the forecast precipitation value at point i, and O_i is the observed precipitation value at point i. Unlike in Clark et al. (2010), the PDF of the observations is *not* assigned to each ensemble member for this portion of the study, since the primary purpose here is to evaluate ensemble spread (as opposed to skill), and using the PDF of the core01 member—which is already appropriately gridded for analysis—is more convenient than using multiple observation datasets. As with the raw dataset, bias-corrected variance differences (i.e., mixed-physics variance minus single-physics variance) and ratios (i.e., bias-corrected single-physics variance/bias-corrected mixed-physics variance) are computed.

2) Rank histograms

While ensemble variance gives a measure of agreement between ensemble members, it does not tell whether an ensemble forecast system contains an appropriate amount of spread relative to the observations. Rank histograms (e.g., Hamill 2001), which tally the rank of the observation relative to the ensemble members' forecasts, fill this role. Sloped rank histograms indicate ensemble biases, while U-shaped rank histograms can indicate ensemble underdispersion relative to the observations or conditional bias (Hamill 2001).

Herein, rank histograms are computed for the mixed- and single-physics ensemble's hourly precipitation forecasts at six forecast hours (i.e., hours 6, 12, 18, 24, 30, and 36). NCAR/EOL Stage IV precipitation data (Lin 2011) are used as the observational dataset, although the ranks are computed on the forecast grid. Rank histograms are created before and after accounting for systematic ensemble biases using the technique based on probability matching described above. While observational errors may impact the shape of rank histograms (e.g., Hamill 2001), observational errors are assumed to be small relative to the spread of the ensemble and are therefore not accounted for in the rank histograms presented herein.

c. *Evaluating ensemble skill*

1) Hourly ensemble mean 2-m temperature and dewpoint temperature

Each ensemble's mean hourly 2-m temperature and dewpoint temperature forecasts are verified against data from 2232 Automated Surface Observing Systems (ASOSs) falling within the analysis domain (Fig. 2). Specifically, the gridded mean ensemble forecasts are interpolated to the observation points shown in Fig. 2 using nearest neighbor interpolation, as performed by Model Evaluation Tools Version 6.1 (METv6.1; Developmental Testbed Center 2017). METv6.1 is then used to compute root-mean-square error (RMSE) values for each ensemble's mean 2-m temperature and dewpoint temperature at each forecast hour from 0 to 36, aggregated over the 23-day dataset.

A two-sided paired permutation test (e.g., Good 2006) is used to test for significant differences between the mixed- and single-physics RMSE at each forecast hour for ensemble mean 2-m temperature and dewpoint temperature. A paired permutation test is used in favor of a one-sample t test on the RMSE differences (e.g., Mittermaier et al. 2013) since the permutation test does not require an assumption that the data follow a normal distribution and avoids the estimation of an effective sample size. The paired permutation test uses the mixed- and single-physics RMSE values from each of the 23 individual days in the dataset. For each day, the mixed- or single-physics RMSE is randomly

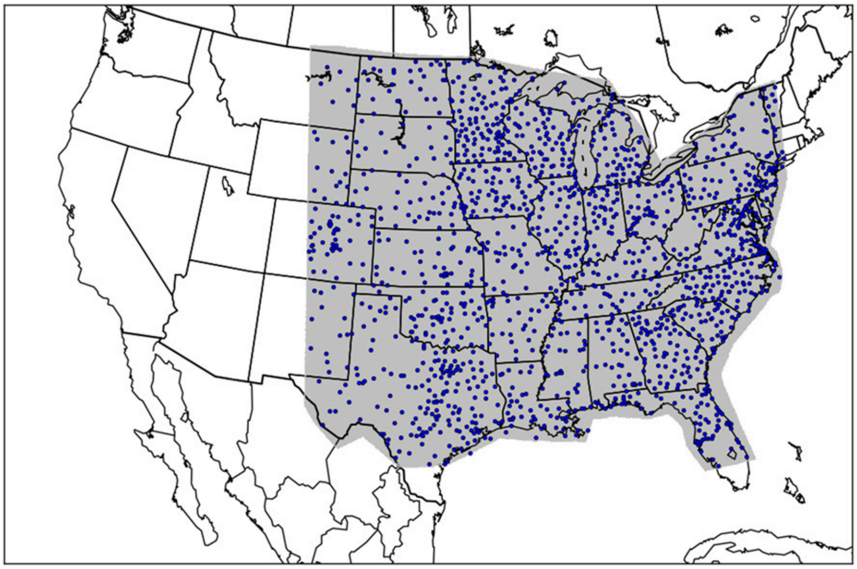

FIG. 2. ASOSs (blue dots) used for verification within the analysis domain (shaded).

assigned to list 1, while the other RMSE is assigned to list 2, and the difference between the two lists' mean RMSE is noted. This procedure is repeated 10 000 times to form a null distribution of mean RMSE differences. The actual mean RMSE difference (mixed-physics RMSE minus single-physics RMSE) is compared to the null distribution to assess significance using $\alpha = 0.05$.

2) 6-H PRECIPITATION

Six-hour ensemble precipitation forecasts are evaluated for six nonoverlapping forecast periods, which cover forecast hours 0–6,[1] 6–12, 12–18, 18–24, 24–30, and 30–36. NCAR/EOL Stage IV precipitation data (Lin 2011) are treated as ''truth'' for verification. The Stage IV data are produced on an approximately 4.8-km polar stereographic grid with 1121 east–west grid points and 881 north–south grid points; therefore, a neighbor budget method (Accadia et al. 2003) is used to remap the data to a 3-km Lambert conformal grid with 1680 east–west grid points and 1152 north–south grid points to match the grid used by the forecasts. The remapped Stage IV data are used for verification and are compared against bias-corrected precipitation forecasts from the mixed- and single-physics ensembles. Probability matching (Clark et al. 2010) is again used to calibrate each ensemble for bias. In this portion of the study, the PDF of the remapped Stage IV observation data is assigned to each ensemble member to eliminate systematic and nonsystematic biases, as in Clark et al. (2010). Metrics used for verification include: fractions skill score (FSS; Roberts and Lean 2008), area under the relative operating characteristics curve (AUC; e.g., Marzban 2004), and attributes diagrams (Hsu and Murphy 1986).

Given its design to be computed over a variety of neighborhoods, FSS is useful for determining forecast skill at a variety of spatial scales. Unlike some other forecast evaluation metrics (e.g., area under the relative operating characteristics curve), FSS depends on bias; more biased forecasts always produce lower FSS values at large spatial scales and usually produce lower FSS values at small spatial scales (Mittermaier and Roberts 2010). FSS can be expressed mathematically as

$$\mathrm{FSS} = 1 - \frac{\frac{1}{M}\sum_{m=1}^{M}(F_m - O_m)^2}{\frac{1}{M}\left(\sum_{m=1}^{M}F_m^2 + \sum_{m=1}^{M}O_m^2\right)}, \qquad (3)$$

where M is the number of forecast–observation pairs (which includes the number of overlapping spatial windows in the domain over each day in the analysis), F_m is the ensemble mean forecast fraction at m, and O_m is the observed fraction at m. Herein, FSS is computed for accumulated 6-h precipitation at each of the aforementioned 6-h forecast periods using 0.10-, 0.25-, 0.50-, 0.75-, and 1.00-in. precipitation

[1] While verification metrics are shown beginning with the first 6-h period after model initialization, it should be noted that the first several forecast hours likely fall within the spinup period for each ensemble member. Results from the early forecast periods should be interpreted accordingly.

thresholds. Forecasts (observations) meeting or exceeding the threshold are considered to be ''yes'' forecasts (observations). To determine how FSS varies with spatial scale, ten square neighborhoods are examined; these measure 3, 6, 9, 12, 18, 24, 36, 48, 72, and 144 km per side.

A skillful baseline FSS score is given by

$$\mathrm{FSS}_{\mathrm{useful}} = 0.5 + \frac{f_0}{2}, \tag{4}$$

where f_0 represents the fractional coverage of ''yes'' forecasts over the entire domain (and, in this case, over all days in the analysis; Roberts and Lean 2008). Note that $\mathrm{FSS}_{\mathrm{useful}}$, as given in Eq. (4), is equivalent to $\mathrm{FSS}_{\mathrm{uniform}}$ in Roberts and Lean (2008). The smallest scale for which $\mathrm{FSS} = \mathrm{FSS}_{\mathrm{useful}}$ is considered to be the smallest useful scale (i.e., the scale at which the forecast contains useful information; Roberts and Lean 2008). As with the 2-m temperature and dewpoint temperature skill verification, a two-sided paired permutation test (Good 2006) is used to test for significant differences between the mixed- and single-physics ensemble FSS at each spatial scale for each of the six 6-h periods.

The two ensembles' 6-h accumulated precipitation forecasts are further evaluated using AUC (e.g., Marzban 2004), which measures a forecast system's ability to discriminate between events and nonevents (e.g., Mason and Graham 2002). AUC values greater than or equal to 0.70 are considered useful in an ensemble framework (Buizza et al. 1999). The same five precipitation thresholds used in the FSS analysis are used in the AUC computations to convert the quantitative precipitation forecasts (QPF) into binary forecasts. In each ensemble member, grid points that meet or exceed the given threshold are assigned a value of 1, while all other grid points are assigned a value of 0. Next, at each point, the ratio of ensemble members containing a 1 to the number of members containing a 0 is computed. This fraction is smoothed using a two-dimensional kernel density function to create forecast probability values (e.g., Brooks et al. 1998; Sobash et al. 2011; Loken et al. 2017). Specifically, the following equation is used:

$$f = \sum_{n=1}^{N} \frac{1}{2\pi\sigma^2} \exp\left[-\frac{1}{2}\left(\frac{d_n}{\sigma}\right)^2\right], \tag{5}$$

where f is the forecast probability at a point, N is the number of points where at least one ensemble member exceeds the precipitation threshold, d_n is the distance from the current point to the nth point, and σ is the standard deviation of the Gaussian kernel [hereafter referred to as the spatial smoothing parameter as in Sobash et al. (2011) and Loken et al. (2017)]. Spatial smoothing parameter values from 2 to 144 km are tested. AUC is then computed by summing contingency table elements over all grid boxes in the domain and over all days in the analysis. As in Loken et al. (2017), probability of detection [POD; Eq. (3) in Loken et al. 2017] and probability of false detection [POFD; Eq. (4) in Loken et al. 2017)] are computed at the following levels of probability: 1%, 2%, and from 5% to 95% in increments of 5%. Grid points meeting or exceeding the given probability level are considered to be ''yes'' forecasts, while other grid points are considered to be ''no'' forecasts at the given probability level. A two-sided hypothesis test based on resampling (Hamill 1999; Loken et al. 2017) is used to test whether differences between the mixed- and single-physics AUC values are significant, using $\alpha = 0.05$.

Because AUC does not give information about forecast reliability (Wilks 2001), attributes diagrams (Hsu and Murphy 1986) are used to assess forecast reliability. Attributes diagrams, which plot observed relative frequency against forecast probability, are used to assess the impact of spatial smoothing on reliability at each of the five precipitation thresholds and at each of the six 6-h forecast periods. To determine whether statistically significant differences exist between the two ensembles' reliability, each ensemble's reliability component of the Brier score (Murphy 1973) is computed for each forecast period and value of the spatial smoothing parameter for each day in the dataset. Specifically, the reliability component of the Brier score can be expressed as

$$\mathrm{Reliability} = \frac{1}{N}\sum_{k=1}^{K} n_k(p_k - \overline{o}_k)^2, \tag{6}$$

where N is the number of grid points in the analysis domain, K is the number of forecast probability bins, n_k is the number of forecasts in bin k, p_k is the forecast probability in bin k, and $\overline{o}_k$ is the mean observed relative frequency in bin k (Wilks 1995). A paired permutation test (Good 2006) is then used to test for significance at $\alpha = 0.05$ in the same manner as previously described.

3. Results

a. Ensemble spread

1) Raw ensemble variance

For each of the four variables analyzed (i.e., 2-m temperature, 2-m dewpoint temperature, 500-hPa geopotential height, and hourly accumulated precipitation), the smallest (largest) spatial scales generally have the

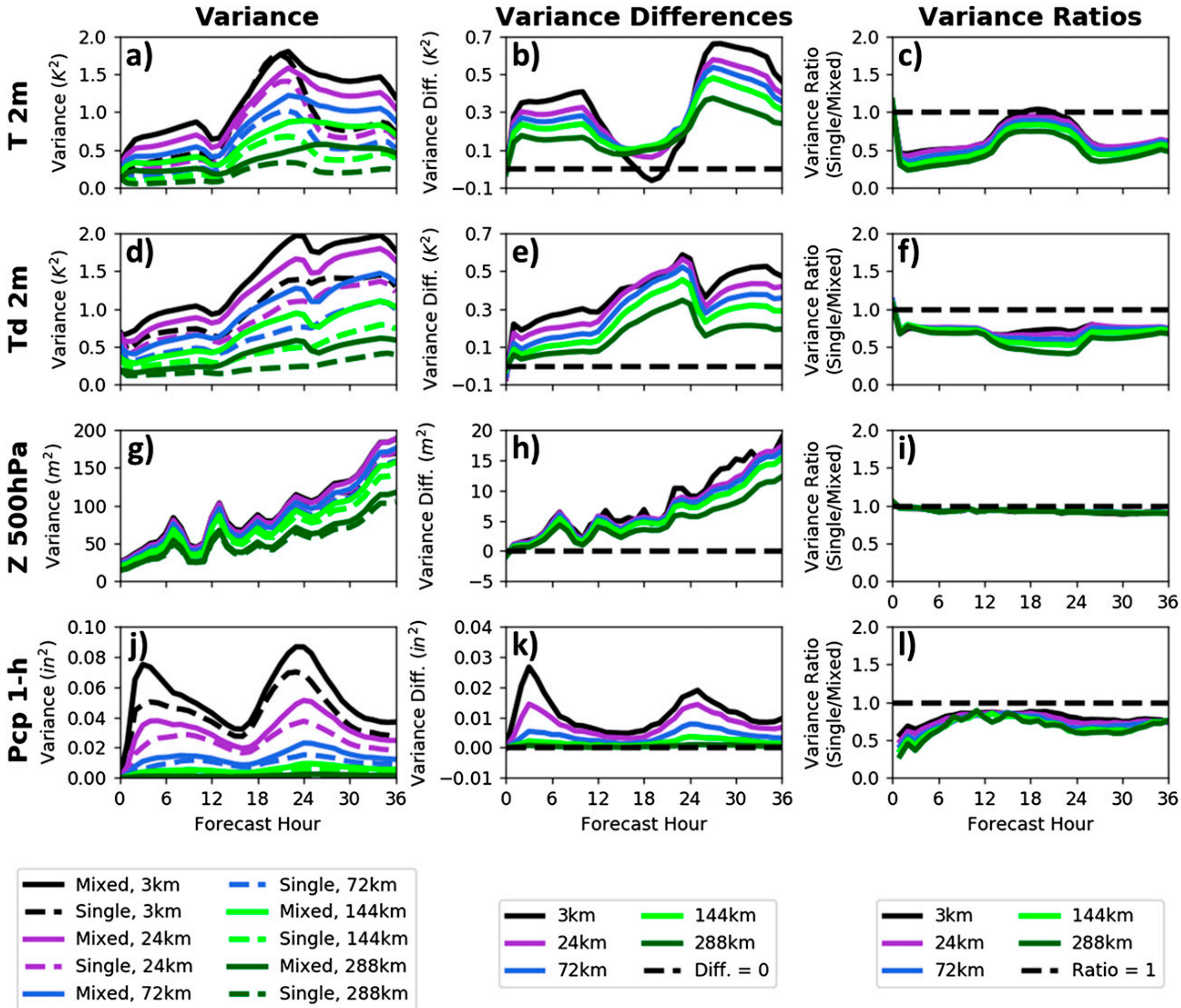

FIG. 3. Time series of (a) mixed- (solid) and single-physics (dashed) ensemble variance, (b) variance differences (mixed-physics variance minus single-physics variance), and (c) variance ratios (single-physics variance/mixed-physics variance) for 2-m temperature forecasts at spatial scales of 3 (black), 24 (purple), 72 (blue), 144 (light green), and 288 km (dark green). (d)–(f) As in (a)–(c), but for 2-m dewpoint temperature forecasts. (g)–(i) As in (a)–(c), but for 500-hPa geopotential height forecasts. (j)–(l) As in (a)–(c), but for hourly precipitation forecasts. A black dashed line denotes a variance difference of 0 in the center column and a variance ratio of 1 in the right column.

greatest (lowest) variances at a given forecast hour (Figs. 3a,d,g,j). This finding makes sense: as the spatial scale (i.e., size of the neighborhood) increases, the variance becomes less sensitive to small, local differences between ensemble members due to the increased spatial averaging. Physically, it also makes sense that the smallest scales will have the greatest variances, since smaller eddies are more difficult to predict and are therefore associated with more uncertainty (e.g., Lorenz 1969).

Consistent with the findings of Clark et al. (2010), a diurnal cycle is noted in the 2-m temperature, 2-m dewpoint, and hourly precipitation variance time series (Figs. 3a,d,j). The hourly precipitation time series (Fig. 3j) contains the most well-defined diurnal cycle; local maxima in variance exist around forecast hours 3 (i.e., 0300 UTC) and 24 (i.e., 0000 UTC the next day). Less well-pronounced diurnal cycles are seen in the 2-m temperature and 2-m dewpoint variance time series (Figs. 3a,d). Both variables have local minima in variance around forecast hours 12 and 26. As in Clark et al. (2010), the 500-hPa geopotential height variance time series does not exhibit a diurnal cycle. The 500-hPa geopotential height variance generally increases with time, with the variance increasing faster for the smaller spatial scales.

Variance differences (Figs. 3b,e,h,k) indicate that the mixed-physics ensemble nearly always generates greater variance than the single-physics ensemble at a given spatial scale and forecast hour for a given variable. This difference in variance is generally greater at the smaller spatial scales. For 500-hPa geopotential height, the difference increases steadily as forecast time increases (Fig. 3h). For the other variables, the difference depends on the diurnal cycle.

To determine how the proportion of spread generated by the mixed-physics ensemble varies with time, ratios of single-physics ensemble variance/mixed-physics ensemble variance are computed (Figs. 3c,f,i,l). While the 500-hPa geopotential height variance ratios remain approximately constant with time and do not differ dramatically with spatial scale (Fig. 3i), the variance ratios from the other fields have more noticeable variations with time and spatial scale. For example, the 2-m temperature ratios reach a local maximum at approximately forecast hour 18 (Fig. 3c), indicating that, proportionally, the mixed-physics ensemble contributes less variance at that time than at other forecast hours. The 2-m dewpoint and hourly precipitation ratios also vary with time, although with much less well-defined local maxima and minima (Figs. 3f,l). Despite these variations, the variance ratios remain below 1.0 for nearly all spatial scales and forecast hours for all four variables, signifying that the mixed-physics ensemble generally produces more spread, proportionally, relative to the single-physics ensemble.

Interestingly, for the 2-m temperature, 2-m dewpoint temperature, and hourly accumulated precipitation fields, the variance ratio is smallest—indicating that the mixed-physics generates proportionally more spread—for the largest spatial scales (Figs. 3c,f,l). Thus, even while the *difference* between the mixed- and single-physics variances is lowest for the largest spatial scales (Figs. 3a,b,d), the *proportion* of variance created by the mixed-physics ensemble is largest—at least for these three variables.

2) Bias-corrected ensemble variance

Correcting for bias preserves the general shape of the variance time series for a given variable but tends to decrease the variance from both the mixed- and single-physics ensembles (Figs. 4a,d,g,j). This result is expected given that the bias-correction procedure removes some of the ‘‘artificial’’ spread that results from systematic biases among the ensemble members (Clark et al. 2010). The reduced spread in the bias-corrected time series is most clearly seen in the 500-hPa geopotential height variances (Figs. 4g, 3g).

After bias-correction is applied, the difference between the mixed- and single-physics ensemble variance is reduced for all four variables at nearly all forecast hours and spatial scales (Figs. 4b,e,h,k and Figs. 3b,e,h,k). The precipitation variance difference after bias-correction (Fig. 4k) is especially noteworthy: the difference between the mixed- and single-physics ensemble variance after bias-correction is nearly 0 at all forecast hours and spatial scales. This result implies that the mixed-physics ensemble had more systematic biases—and therefore more ‘‘artificial’’ spread (Clark et al. 2010)—than the single-physics ensemble. Thus, removing the systematic biases from both ensembles would be expected to reduce the variance of the mixed-physics ensemble more than that from the single-physics ensemble.

For each of the four variables studied, bias-correction tends to push the single-physics ensemble variance/mixed-physics ensemble variance ratios slightly toward 1.0 (Figs. 4c,f,i,l and Figs. 3c,f,i,l). In nearly all cases, this change indicates an increased proportion of variance generated by the single-physics ensemble when bias correction is applied. The effect is seen for most spatial scales and forecast hours.

3) Rank histograms

Before correcting for systematic biases, both ensembles' rank histograms are skewed to the right for all six forecast hours examined (Fig. 5), suggesting both ensembles tend to overforecast 1-h precipitation. The mixed-physics rank histograms (Figs. 5a–f) tend to be more strongly skewed than the corresponding single-physics rank histograms (Figs. 5g–l), especially at the later forecast hours. This result suggests that the systematic biases within the mixed-physics ensemble are predominantly in one direction (i.e., positive), producing an ensemble system with more overforecasting bias than the single-physics ensemble.

Correcting for systematic biases flattens both ensembles' rank histograms (Fig. 6), a result consistent with Clark et al. (2009). However, some skewness remains at all forecast hours since the bias-correction procedure replaces the PDF of each member with the PDF of the core01 control member, which has suboptimal bias. As expected, the mixed-physics ensemble benefits more from the bias-correction technique than the single-physics ensemble due to its greater initial systematic biases. A slight U shape is noted in both ensembles after bias correction, particularly at forecast hour 24 (Figs. 6d,j), suggesting that both ensembles are underdispersive relative to the observations. Adding more members to each ensemble could potentially alleviate this underdispersion

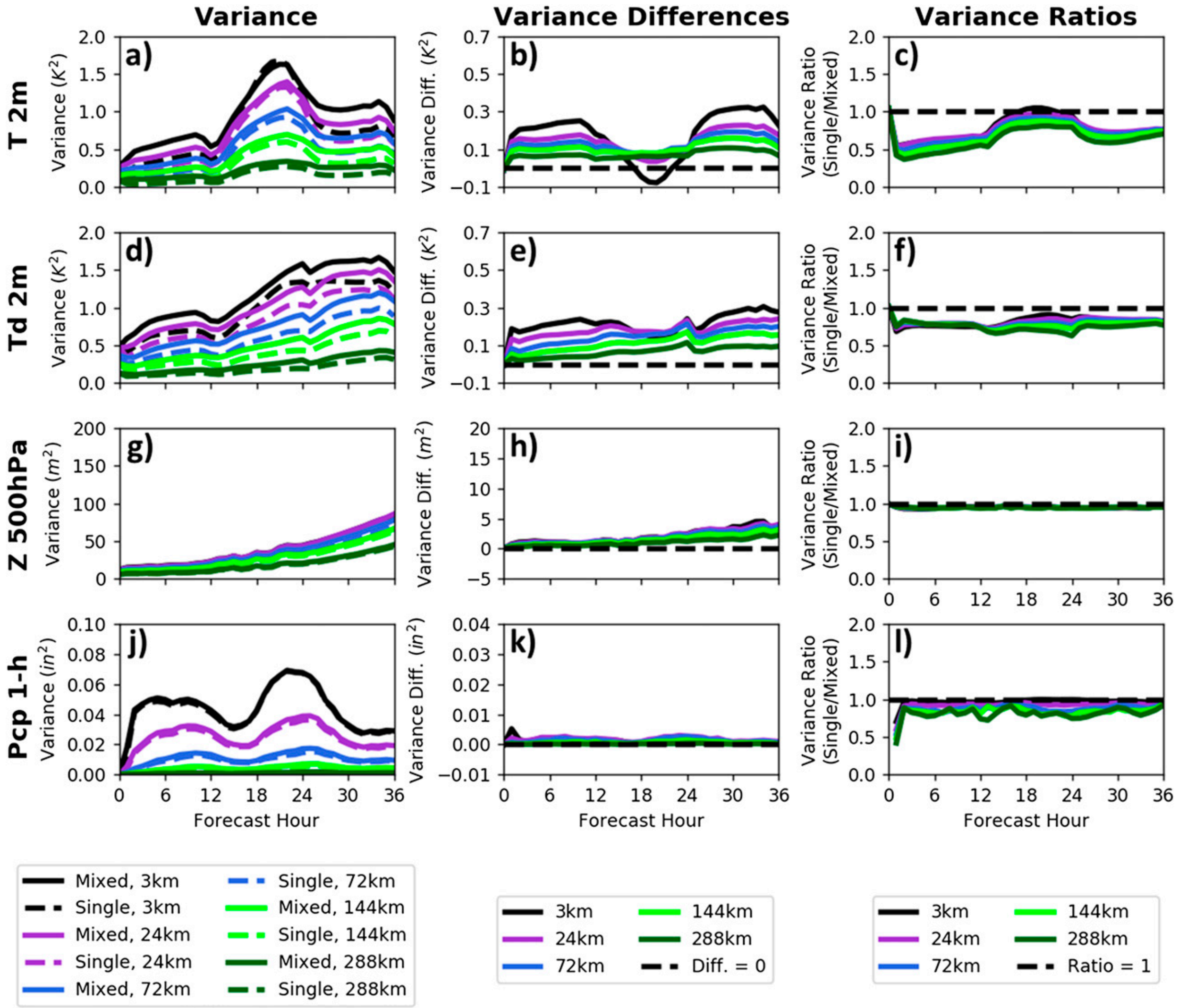

FIG. 4. Bias-corrected time series of (a) mixed- (solid) and single-physics (dashed) ensemble variance, (b) variance differences (mixed-physics variance minus single-physics variance), and (c) variance ratios (single-physics variance/mixed-physics variance) for 2-m temperature forecasts at spatial scales of 3 (black), 24 (purple), 72 (blue), 144 (light green), and 288 km (dark green). (d)–(f) As in (a)–(c), but for 2-m dewpoint temperature forecasts. (g)–(i) As in (a)–(c), but for 500-hPa geopotential height forecasts. (j)–(l) As in (a)–(c), but for hourly precipitation forecasts. A black dashed line denotes a variance difference of 0 in the center column and a variance ratio of 1 in the right column. Axis scales are identical to those in Fig. 3.

by providing a more complete sampling of the forecast PDF.

b. *Ensemble skill*

1) Hourly 2-m temperature and dewpoint temperature RMSE

The mixed- and single-physics ensembles produce forecast hourly 2-m temperatures that have subjectively similar RMSE values throughout the 36-h forecast period (Fig. 7a). Between forecast hours 14 and 27, a significant difference between the two ensembles' hourly 2-m temperature RMSE is noted at only one forecast hour (i.e., hour 24). Results from the paired permutation test show that a significant difference between the two ensembles' hourly 2-m temperature RMSE exists 22 times out of the 37 possible forecast/analysis hours (i.e., hours 0–36). In 18 of these cases, the mixed-physics ensemble has the lower RMSE.

The RMSE from the two ensembles' 2-m dewpoint temperature forecasts have greater subjective and objective differences. The mixed-physics ensemble RMSE is always less than the corresponding single-physics RMSE for all forecast hours examined (Fig. 7b).

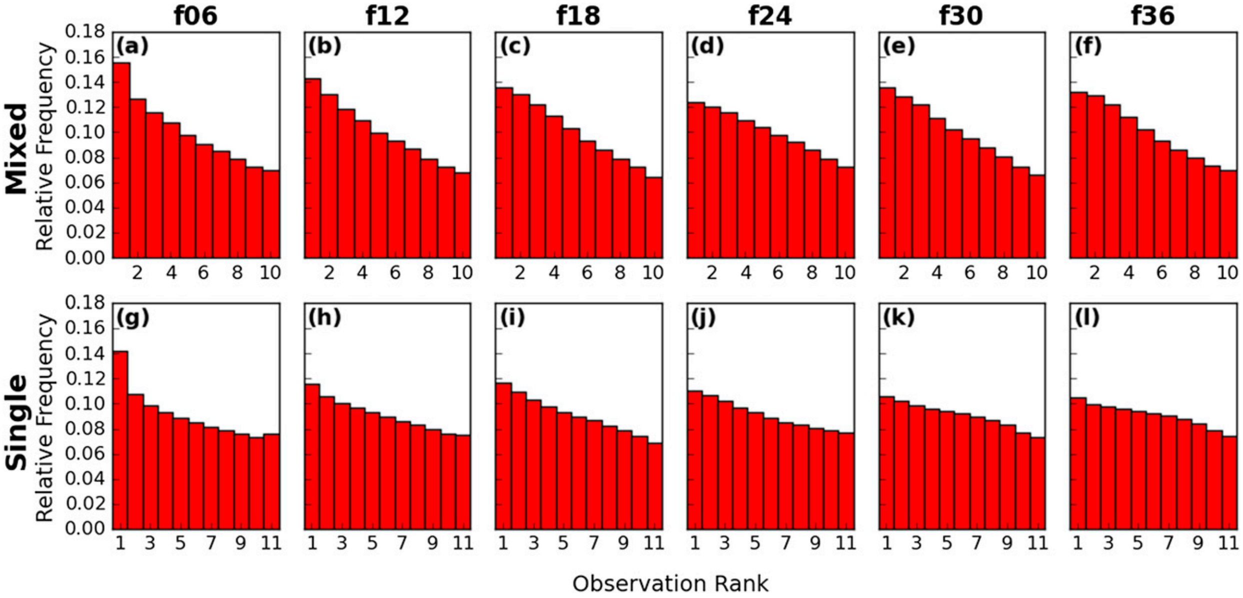

FIG. 5. (a) Rank histogram for the mixed-physics ensemble's forecast 1-h accumulated precipitation, valid for forecast hour 6. (b)–(f) As in (a), but valid for forecast hours 12, 18, 24, 30, and 36, respectively. (g)–(l) As in (a)–(f), but for the single-physics ensemble.

Moreover, a significant difference between the two ensembles' 2-m dewpoint temperature RMSE values exists for 32 of the 37 forecast hours analyzed. The greatest difference occurs between forecast hours 14 and 25 (i.e., from 1400 UTC to 0100 UTC the next day). One possible explanation for the mixed-physics ensemble's superior performance is that the systematic biases of its three PBL schemes have different signs, leading to less overall bias—and therefore less error—in its 2-m temperature forecasts compared to the single-physics ensemble.

2) 6-H PRECIPITATION

(i) FSS

For all six 6-h forecast periods, the greatest FSSs are associated with the largest spatial scale (i.e., 144 km) and the lowest precipitation threshold (i.e., 0.10 in.; Figs. 8a–f).

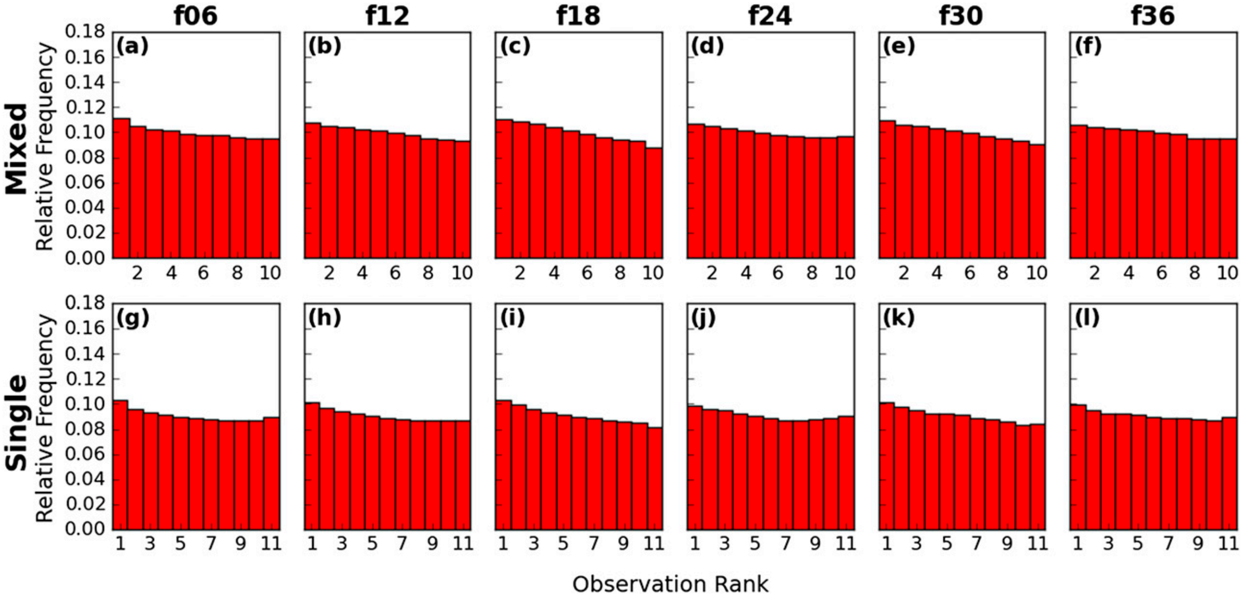

FIG. 6. As in Fig. 5, but for bias-corrected mixed- and single-physics ensemble forecasts.

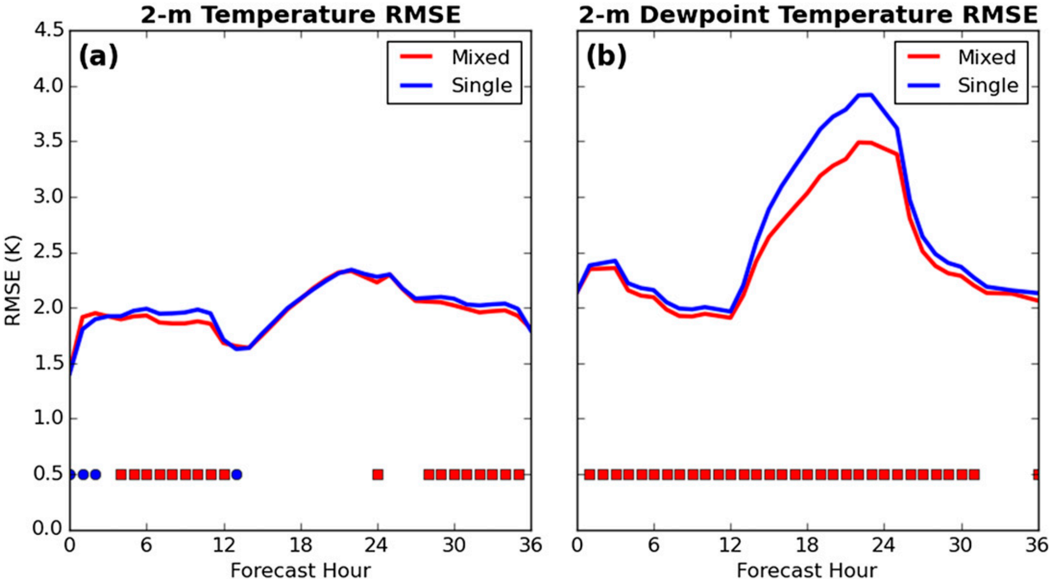

FIG. 7. Time series of mixed- (red) and single-physics (blue) ensemble RMSE for (a) forecast hourly 2-m temperature and (b) 2-m dewpoint temperature. Red squares (blue circles) denote a statistically significant difference ($\alpha \leq 0.05$) between the mixed- and single-physics RMSE values with the mixed-physics (single-physics) ensemble having the lower RMSE.

In all cases, the FSS progressively decreases as the precipitation threshold increases from 0.10 to 1.00 in. For a given threshold and spatial scale, the mixed- and single-physics ensemble forecasts produce qualitatively similar FSS values. FSS differences are not statistically significant (at $\alpha = 0.05$) at any of the spatial scales or precipitation thresholds after the first 6-h period (i.e., forecast hours 0–6; Figs. 8b–f). During the first 6-h forecast period, FSS differences are significant at one spatial scale (144 km) for the 0.75-in. forecasts, seven spatial scales (3, 18, 24, 36, 48, 72, and 144 km) for the 0.50-in. forecasts, and all 10 spatial scales for the 0.25- and 0.10-in. forecasts (Fig. 8a). Notably, in each instance of significance, the single-physics ensemble produces the greater FSS.

In general, FSS gradually decreases with increasing forecast lead time. This pattern holds for both mixed- and single-physics forecasts and is shown explicitly for the 0.10-, 0.25-, 0.50-, and 1.00-in. thresholds (Figs. 9a–d). When a forecast's FSS decreases below $\mathrm{FSS}_{\mathrm{useful}}$ depends on both the precipitation threshold and spatial scale of the forecast; higher precipitation thresholds and smaller-scale forecasts reach $\mathrm{FSS}_{\mathrm{useful}}$ faster. However, whether the ensemble contains mixed- or single-physics parameterizations does not appear to dramatically impact the time taken for its forecast to reach $\mathrm{FSS}_{\mathrm{useful}}$. Both the mixed- and single-physics ensembles have qualitatively similar FSS values for a given 6-h forecast period, precipitation threshold, and spatial scale. Statistically significant differences between the two ensembles' FSS exist only during the first 6-h forecast period, and the single-physics ensemble has the higher FSS in all cases of significance.

(ii) *AUC from 6-h probabilistic forecasts*

In general, for both the mixed- and single-physics forecasts, AUC tends to be higher for the lower threshold forecasts (e.g., Fig. 10a), perhaps because ≥1.00-in. rainfall events are rarer and more difficult for a forecast system to place precisely compared to lighter precipitation events. As more spatial smoothing is applied (Figs. 10a–f), the AUC values of all forecasts examined become increasingly similar. More spatial smoothing also increases the AUC of all forecasts examined, up to a point. For a given threshold and forecast period, the mixed-physics ensemble generally produces slightly greater AUC than the single-physics ensemble; however, the differences are small. The greatest differences between mixed- and single-physics ensemble AUC occur with the highest precipitation threshold (i.e., 1.00 in.) and during the 6-h period ending at forecast hour 30 (i.e., 0000–0600 UTC one day after the forecast is initialized; Figs. 10a–f). Notably, none of the differences between the mixed- and single-physics ensemble AUC are statistically significant at $\alpha = 0.05$.

The impact of varying the standard deviation of the Gaussian kernel (henceforth referred to as the spatial

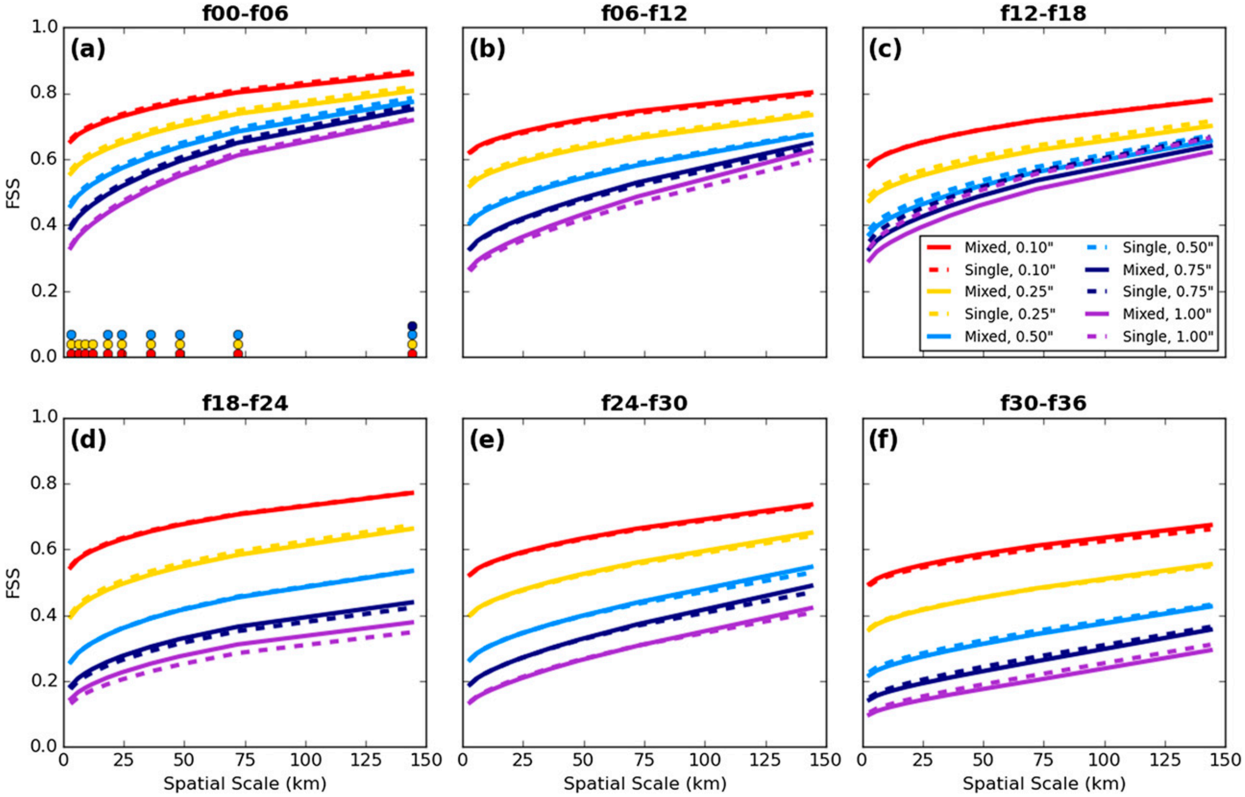

FIG. 8. Mixed- (solid) and single-physics (dashed) ensemble fractions skill score as a function of spatial scale for the 6-h forecast period spanning forecast hours (a) 0–6, (b) 6–12, (c) 12–18, (d) 18–24, (e) 24–30, and (f) 30–36. In each case, 0.10- (red), 0.25- (gold), 0.50- (light blue), 0.75- (dark blue), and 1.00-in. (purple) precipitation threshold forecasts are shown. Filled circles indicate significance at $\alpha = 0.05$.

smoothing parameter) at all six 6-h forecast periods is assessed explicitly in Figs. 11a–f. Regardless of forecast period or ensemble physics configuration (i.e., mixed or single physics), AUC increases relatively rapidly as the spatial smoothing parameter is increased from 2 to 12 km and then increases more gradually as the spatial smoothing parameter is further increased to 72 km (Figs. 7a–f). With the application of even more spatial smoothing, the AUC begins to level off or slightly decrease. The larger precipitation threshold forecasts benefit more from additional spatial smoothing relative to the lower threshold forecasts; the same amount of spatial smoothing increases the higher-threshold forecasts' AUC values more than the lower-threshold forecasts' AUC values.

(iii) Attributes diagrams

Varying the spatial smoothing parameter directly influences forecast reliability. With less spatial smoothing, the higher probabilities tend to be overforecast while the lower probabilities tend to be slightly underforecast (e.g., Fig. 12a). More spatial smoothing decreases the number of high-probability forecasts while increasing the number of low-probability forecasts. Therefore, up to a point, increasing the spatial smoothing parameter improves forecast reliability. Of the values tested, a spatial smoothing parameter of 72 or 96 km—depending on the precipitation threshold—produces the best reliability (Figs. 12a–d). As the spatial smoothing parameter is increased beyond 96 km, the forecasts tend toward an underforecasting bias at the medium and higher forecast probabilities as well as a reduction in forecast sharpness. In general, a spatial smoothing parameter of 72 km provides optimal or near optimal reliability as well as discrimination ability. This finding holds for both the mixed- and single-physics ensemble forecasts at precipitation thresholds ranging from 0.10 to 1.00 in. Statistically significant differences between the two ensembles' reliability component of the Brier score exist only at the 0.10- and 0.25-in. thresholds (Figs. 12a,b). Notably, the single-physics ensemble has the superior reliability in all cases of a statistically significant difference between the two ensembles' reliability values.

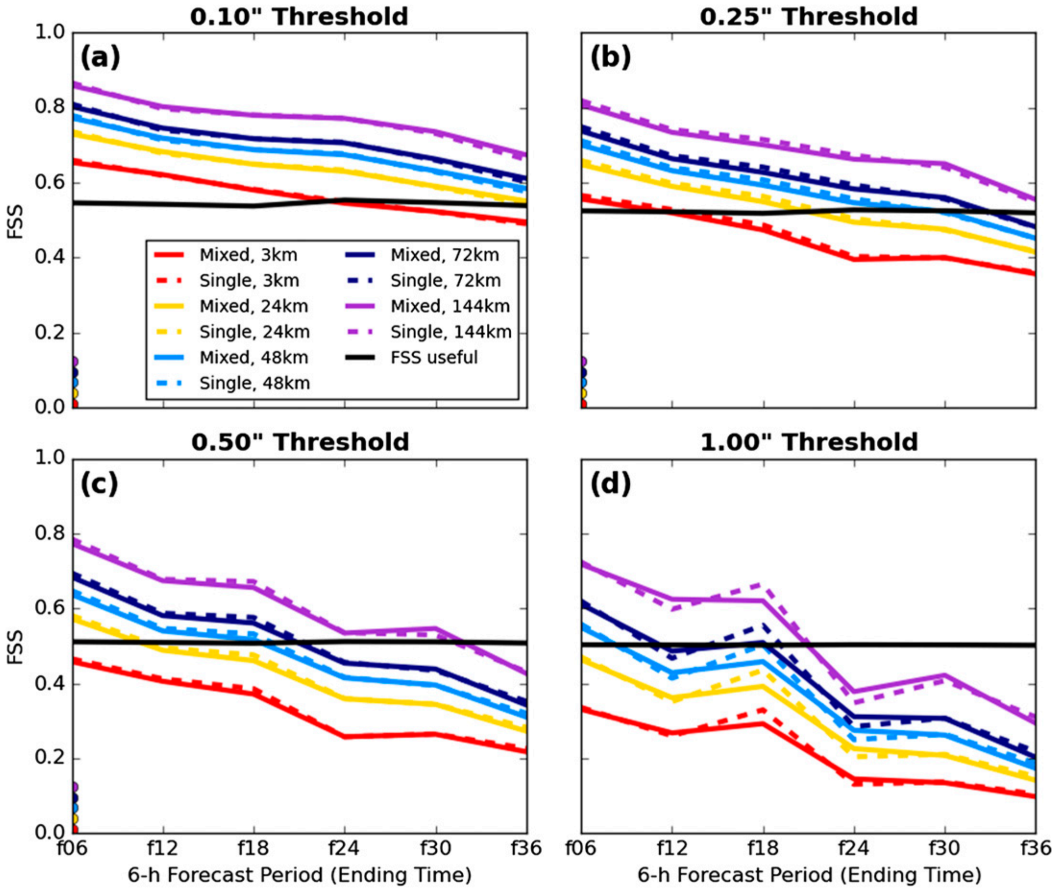

FIG. 9. Fractions skill score as a function of forecast period for mixed- (solid) and single-physics (dashed) ensemble 6-h precipitation forecasts at (a) 0.10-, (b) 0.25-, (c) 0.50-, and (d) 1.00-in. thresholds. In each case, 3- (red), 24- (gold), 48- (light blue), 72- (dark blue), and 144-km (purple) spatial scales are shown. The $\mathrm{FSS}_{\mathrm{useful}}$ value is denoted by a solid black line. Filled circles indicate significance at $\alpha = 0.05$.

Reliability is sensitive to precipitation threshold: particularly for the smaller spatial scales, the lower-threshold forecasts (i.e., the 0.10- and 0.25-in. forecasts; Figs. 12a,b) have better reliability than the higher-threshold forecasts (i.e., the 0.50- and 1.00-in. forecasts; Figs. 12c,d). The higher-threshold forecasts also suffer from a greater reduction of sharpness compared to the lower-threshold forecasts as the spatial smoothing parameter is increased, since already-rare high forecast probabilities become forecast even less often. However, for a given threshold and spatial smoothing parameter, the mixed- and single-physics ensembles have qualitatively similar forecast reliability, provided the probability bins each contain a sufficient number of forecasts. In situations when a statistically significant difference exists between each ensemble's reliability component of the Brier score, the single-physics ensemble almost always has the superior reliability. Each ensemble's reliability is not very sensitive to forecast hour; reliability curves are qualitatively similar for each of the six forecast periods examined (Figs. 13a–f).

4. Select cases

To provide a visual comparison of the mixed- and single-physics ensemble precipitation forecasts, 1-in. forecasts are examined on four case study days. These include three ''high precipitation'' cases and one ''failure'' case. All case study forecasts are valid for the 6-h period ending at forecast hour 30 (i.e., 0000–0600 UTC on the day after the forecast was initialized), since the greatest differences in mixed- and single-physics ensemble AUC were found to have occurred during this period. The first three case study days were selected by choosing the three days with the greatest number of points meeting or exceeding 1 in. of observed 6-h rainfall inside the analysis domain, while the final case was chosen subjectively as an interesting case in which both ensembles produced large forecast misses

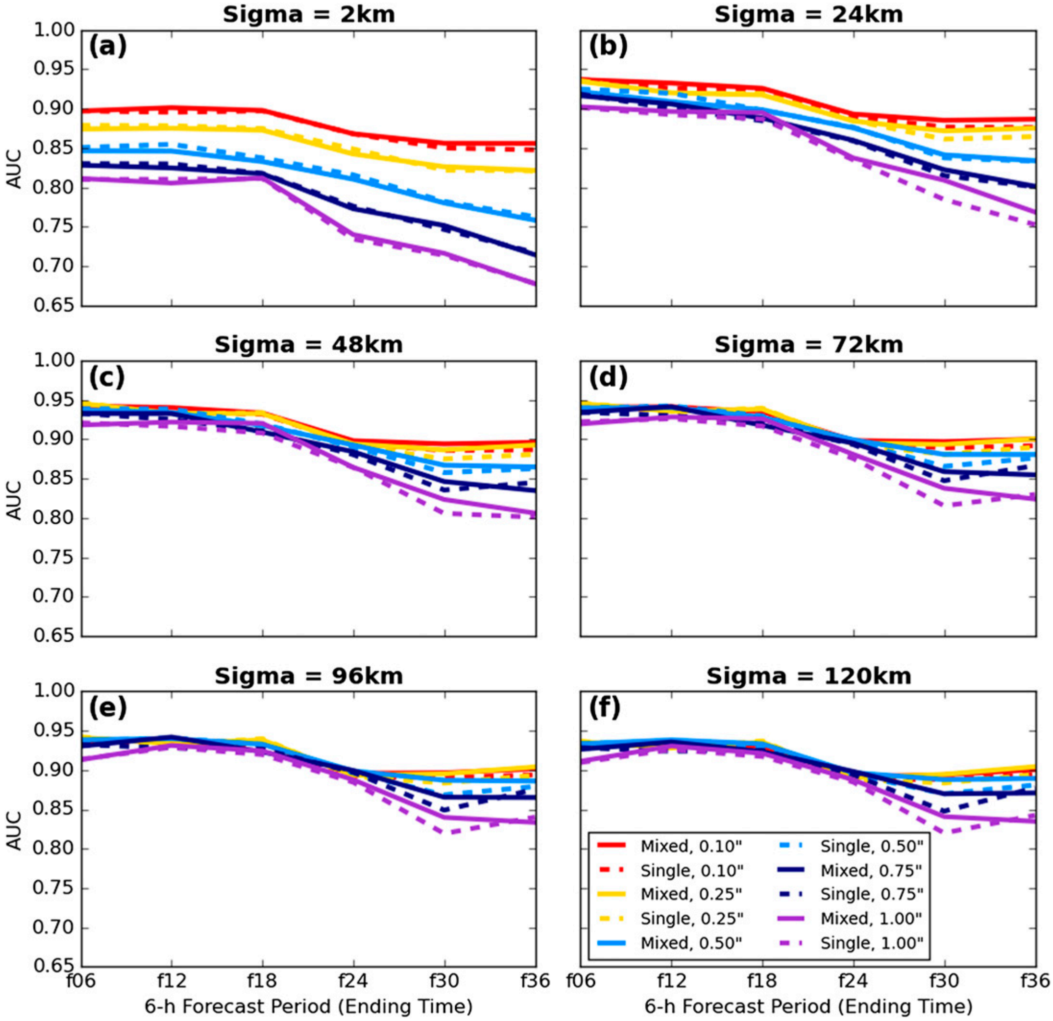

FIG. 10. AUC for mixed- (solid) and single-physics (dashed) 0.10- (red), 0.25- (gold), 0.50- (light blue), 0.75- (dark blue), and 1.00-in. (purple) 6-h accumulated precipitation threshold forecasts using a spatial smoothing parameter of (a) 2, (b) 24, (c) 48, (d) 72, (e) 96, and (f) 120 km. AUC values are plotted for the 6-h forecast periods ending at forecast hours 6, 12, 18, 24, 30, and 36.

and low objective verification scores. The 1-in. threshold was selected since that threshold gave the greatest difference between the mixed- and single-physics ensemble AUC. The 1-in. threshold was also chosen because, from an operational perspective, accurately predicting higher-impact (i.e., heavier precipitation) events is arguably more difficult and desirable for forecasters to achieve; therefore, ensemble forecasts of these events were deemed worthy of closer examination. All forecasts were created using a spatial smoothing parameter σ of 72 km since, of the values tested, σ = 72 km generally produced forecasts with the best reliability and discrimination ability. Single-day AUC and FSS are computed and displayed for each case. The FSS is calculated using a square neighborhood of 252 km (i.e., 3.5σ) per side.

a. 27 May 2016

During the day of 26 May 2016, a surface cyclone developed and strengthened in the lee of the Rocky Mountains. Storms initiated along the warm front in southern Kansas at around 1800 UTC 26 May and grew upscale as they moved to the northeast. In the late afternoon, additional storms formed along the dryline, which extended from west-central Kansas to southwestern Texas. These storms also grew upscale, bringing heavy rainfall to central Texas and western Oklahoma during the 0000–0600 UTC forecast period on 27 May. Fueled by abundant moisture

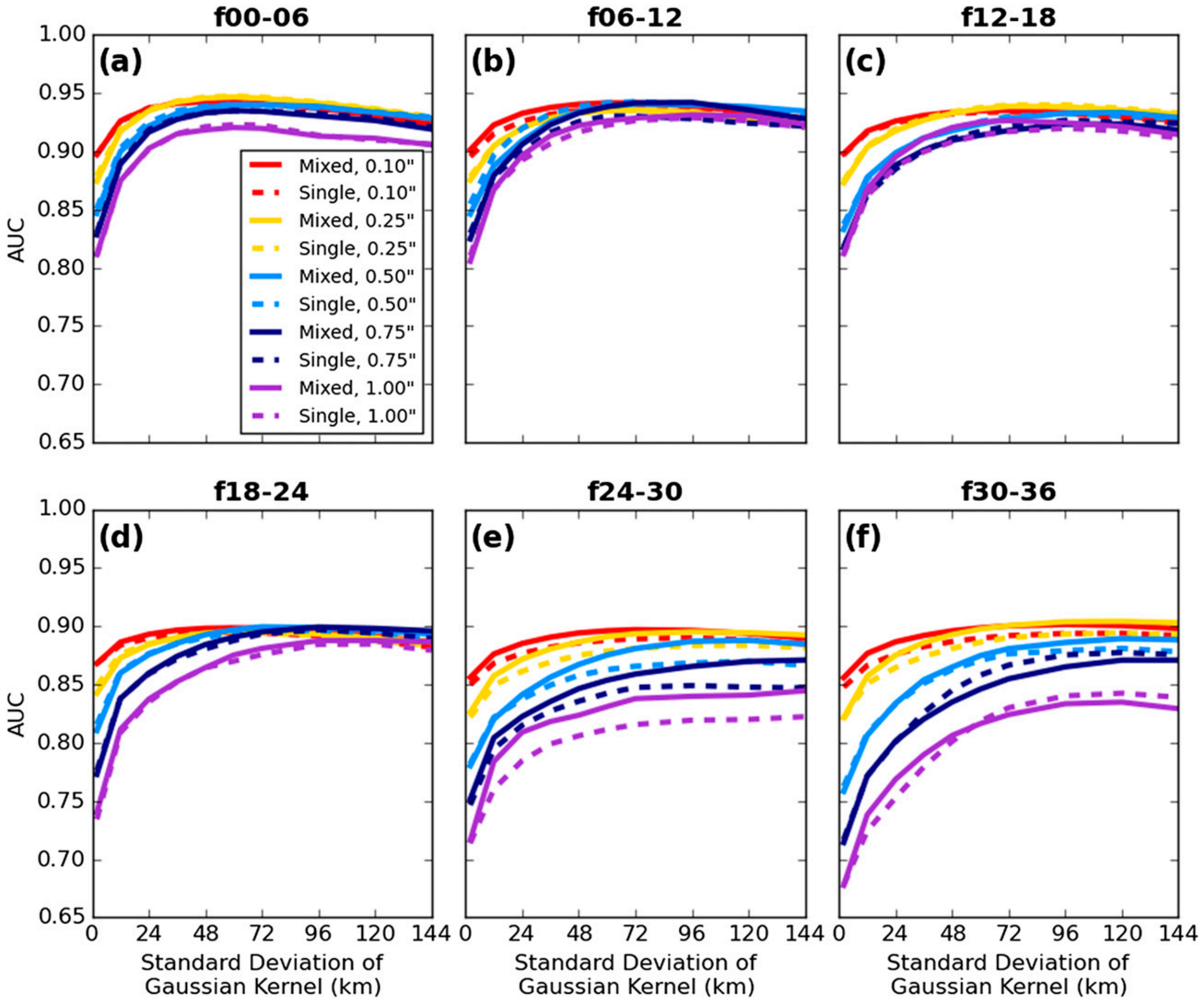

FIG. 11. AUC for mixed- (solid) and single-physics (dashed) 0.10- (red), 0.25- (gold), 0.50- (light blue), 0.75- (dark blue), and 1.00-in. (purple) 6-h accumulated precipitation forecasts as a function of the spatial smoothing parameter for the 6-h forecast period spanning forecast hours (a) 0–6, (b) 6–12, (c) 12–18, (d) 18–24, (e) 24–30, and (f) 30–36.

and instability, another complex of storms produced heavy rainfall over southeastern Texas during the period.

The mixed- and single-physics ensemble forecasts highlight the same general regions for ≥1-in. 6-h rainfall (Figs. 14a,b), and both demonstrate reasonable forecast quality, with both AUC values ≥ 0.75. Differences between the two forecasts include the mixed-physics ensemble's better prediction of heavy rainfall in central Missouri as well as in southwestern Nebraska and northeastern Colorado. Additionally, the magnitudes of the two forecasts' probabilities differ slightly in northeastern Kansas and south-central Arkansas. However, these differences are minor; the two forecasts are generally similar. Neither predicts the southeastern Texas or western Oklahoma precipitation well. Plots of individual member 1-in. forecasts (Figs. 15a–c) are also similar between the two ensembles, although the mixed-physics ensemble more accurately depicts the threat of heavy precipitation in southwestern Nebraska and central Missouri. Nevertheless, given their broad similarities, both ensembles would likely provide comparable value to forecasters.

b. 18 May 2016

Two main regions in the analysis domain recorded ≥1-in. observed 6-h precipitation totals from 0000 to 0600 UTC 18 May 2016: south-central Texas and the Florida Peninsula. In central Texas, storms initiated along a southwest–northeast oriented cold front during the midafternoon of 17 May. These storms grew

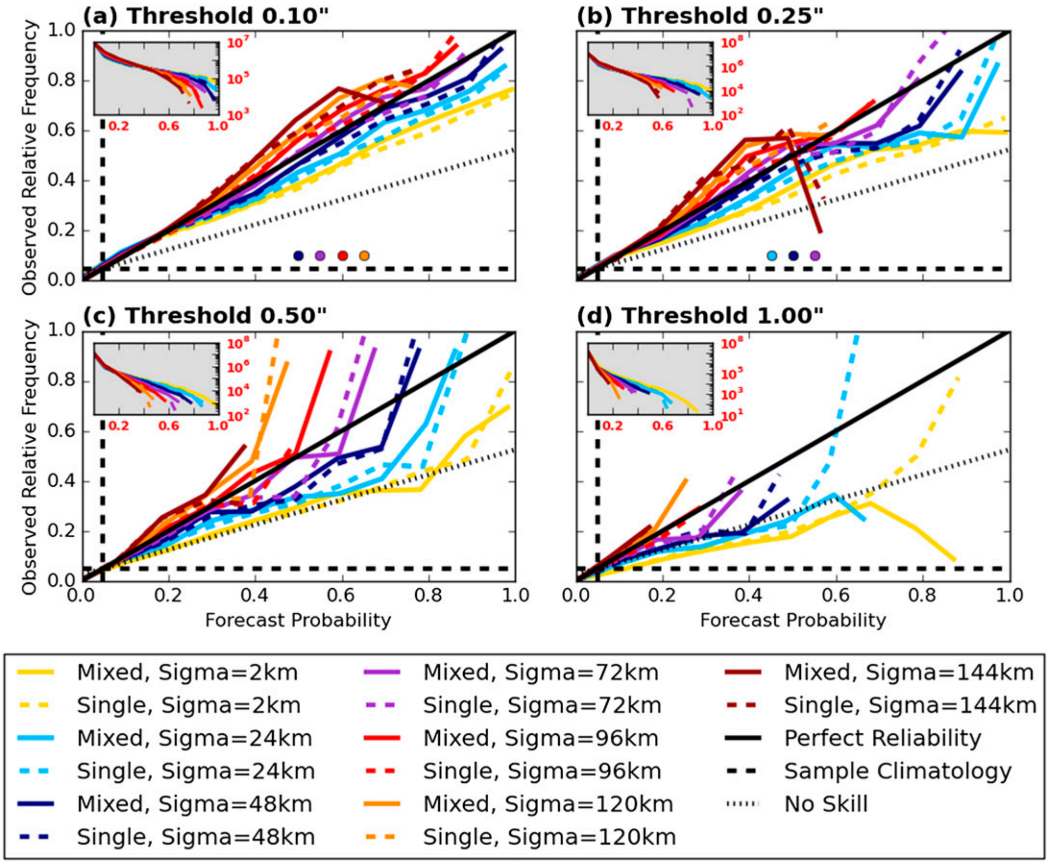

FIG. 12. Attributes diagrams for mixed- (solid) and single-physics (dashed) ensemble 6-h precipitation forecasts ending at forecast hour 30 using a threshold of (a) 0.10, (b) 0.25, (c) 0.50, and (d) 1.00 in. In each case, forecasts produced using a spatial smoothing parameter of 2 (gold), 24 (light blue), 48 (dark blue), 72 (purple), 96 (red), 120 (orange), and 144 km (dark red) are shown. The line of perfect reliability (solid black), no skill (short-dashed black), and lines of sample relative climatological frequency (long-dashed black) are also displayed. Filled circles indicate significant differences in the reliability component of the Brier score at $\alpha = 0.05$, with the single-physics ensemble having the better reliability. Insets within each panel show the number of forecasts as a function of forecast probability and use a logarithmic y axis. Note the y-scale differences in the inset plots.

upscale as they propagated south-southeastward during the forecast period, bringing heavy rainfall to portions of south-central Texas. In Florida, a broad region of storms formed as a low-amplitude 700-hPa shortwave trough moved northeastward through the Peninsula, providing forcing for ascent in an environment characterized by rich boundary layer moisture and moderate instability.

Both ensembles produce similar forecasts, which perform well (Figs. 14c,d). Each forecast assigns modest probabilities to south-central Texas and the Florida Peninsula, where heavy rainfall was observed; however, both forecasts also have a notable false alarm region extending from northeastern Texas into Louisiana and southern Arkansas. The mixed-physics ensemble has an additional small false alarm region in western North Carolina and southern Virginia, which is absent from the single-physics forecast. However, the mixed-physics ensemble has fewer members forecasting ≥1.00-in. precipitation in southern Arkansas, and it has one member forecasting ≥1.00-in. precipitation in the southern Texas Panhandle near a small region of >1.00-in. observed precipitation (Figs. 15d–f). Nevertheless, these differences are subtle, and the two ensemble forecasts are generally similar.

c. 28 May 2016

At 1200 UTC 27 May a 500-hPa shortwave trough was located in eastern Colorado. Storms began to form near the associated surface low in eastern Colorado around 1800 UTC, while storms began to initiate in central Kansas and northern Oklahoma ahead of a cold front at approximately 1900 UTC. Additional convective

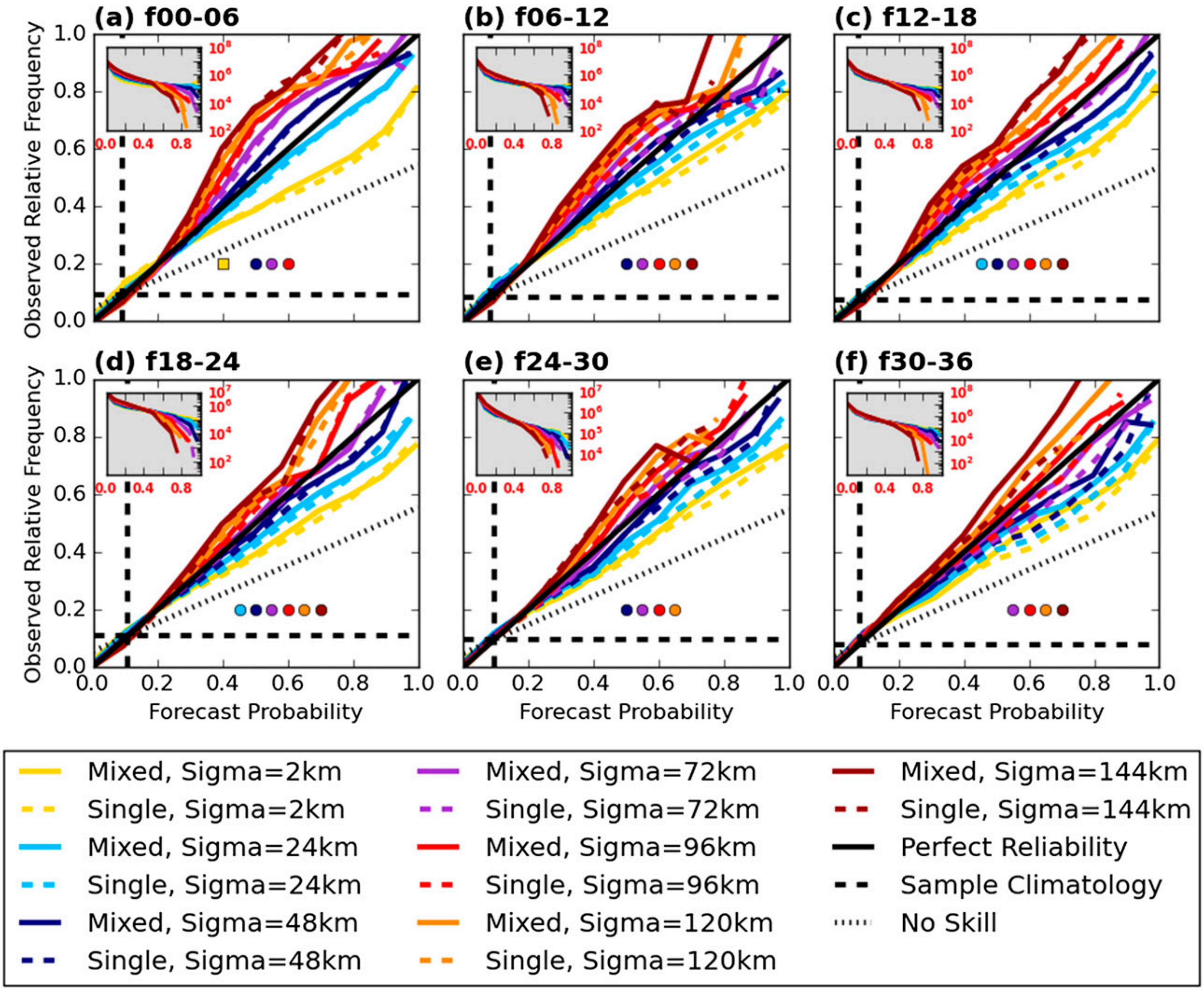

FIG. 13. Attributes diagrams for mixed- (solid) and single-physics (dashed) ensemble 0.10-in. threshold 6-h accumulated precipitation forecasts for the 6-h forecast period spanning forecast hours (a) 0–6, (b) 6–12, (c) 12–18, (d) 18–24, (e) 24–30, and (f) 30–36. In each case, forecasts produced using a spatial smoothing parameter of 2 (gold), 24 (light blue), 48 (dark blue), 72 (purple), 96 (red), 120 (orange), and 144 km (dark red) are shown. The line of perfect reliability (solid black), no skill (short-dashed black), and lines of sample relative climatological frequency (long-dashed black) are also displayed. Filled squares (circles) indicate significant differences in the reliability component of the Brier score at $\alpha = 0.05$, with the mixed-physics (single-physics) ensemble having the better reliability. Insets within each panel show the number of forecasts as a function of forecast probability and use a logarithmic y axis. Note the y-scale differences in the inset plots.

activity developed in eastern Nebraska and northern Missouri near 2200 UTC. The convection in all three areas grew upscale and moved northeastward during the 0000–0600 UTC forecast period on 28 May. Farther south, a preexisting MCS moved southeastward during the forecast period, impacting southeastern Texas and southwestern Louisiana.

The two ensemble forecasts are similar but not identical (Figs. 14e,f). Both assign nonzero probabilities to most of Wisconsin and Iowa as well as eastern portions of Nebraska, Kansas, Oklahoma, and Texas. Neither ensemble correctly predicts heavy precipitation along the Gulf coast in southeastern Texas and southern Louisiana. However, the mixed-physics ensemble arguably does a better job of representing the overall situation there compared to the single-physics ensemble. For example, the mixed-physics ensemble has multiple members forecasting long, narrow, west-southwest–east-northeast swaths of ≥1-in. precipitation, which is close to the observed scenario but displaced to the northwest (Figs. 15g–i). The mixed-physics ensemble also does a better job of depicting the threat of heavy precipitation in southern Nebraska and west-central Kansas, where the

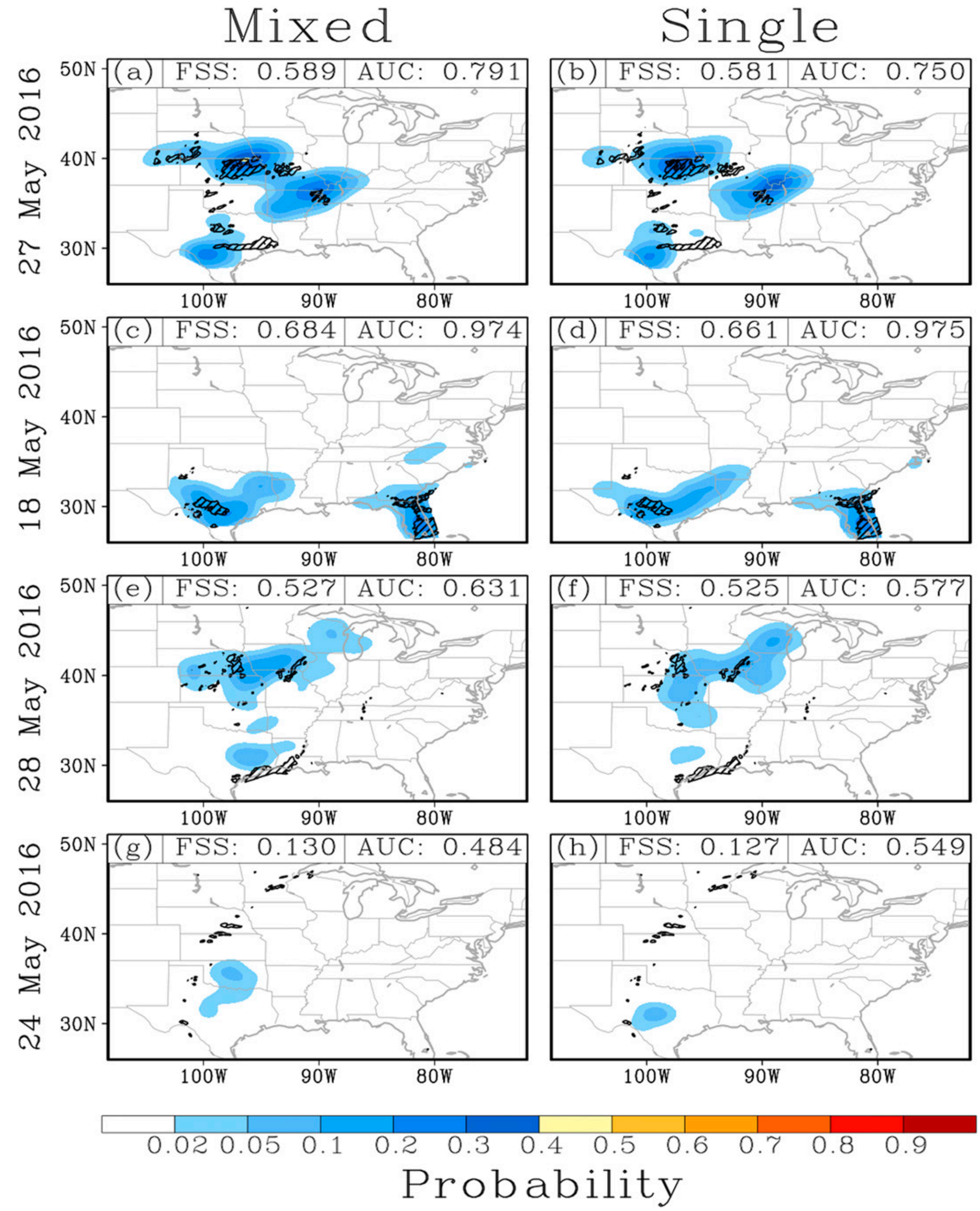
Mixed
Single
27 May 2016
18 May 2016
28 May 2016
24 May 2016
(a) FSS: 0.589 AUC: 0.791
(b) FSS: 0.581 AUC: 0.750
(c) FSS: 0.684 AUC: 0.974
(d) FSS: 0.661 AUC: 0.975
(e) FSS: 0.527 AUC: 0.631
(f) FSS: 0.525 AUC: 0.577
(g) FSS: 0.130 AUC: 0.484
(h) FSS: 0.127 AUC: 0.549
50N
40N
30N
100W
90W
80W
0.02 0.05 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9
Probability

FIG. 14. The 30-h probabilistic 1.00-in. precipitation forecast (shaded) from (a) the mixed-physics ensemble and (b) the single-physics ensemble. Forecasts are valid for 0000–0600 UTC 27 May 2016. Black hatching denotes 3-km points containing observed ≥1.00-in. precipitation over the 6-h period when the forecast is valid. Single-day AUC and FSS are displayed at the top of each plot. (c),(d) As in (a) and (b), but valid for 18 May 2016. (e),(f) As in (a) and (b), but valid for 28 May 2016. (g),(h) As in (a) and (b), but valid for 24 May 2016.

single-physics ensemble displays zero probabilities. Finally, the mixed-physics ensemble reduces the magnitude of probabilities in south-central Wisconsin and western Illinois, where ≥1-in. rainfall was not observed. Still, the two forecasts are similar enough that, in terms of forecast value, the mixed-physics ensemble likely provides only marginal benefits over the single-physics ensemble in this case.

d. 24 May 2016

Just before 2200 UTC 23 May, a line of storms extending from eastern Nebraska into northern Wisconsin formed

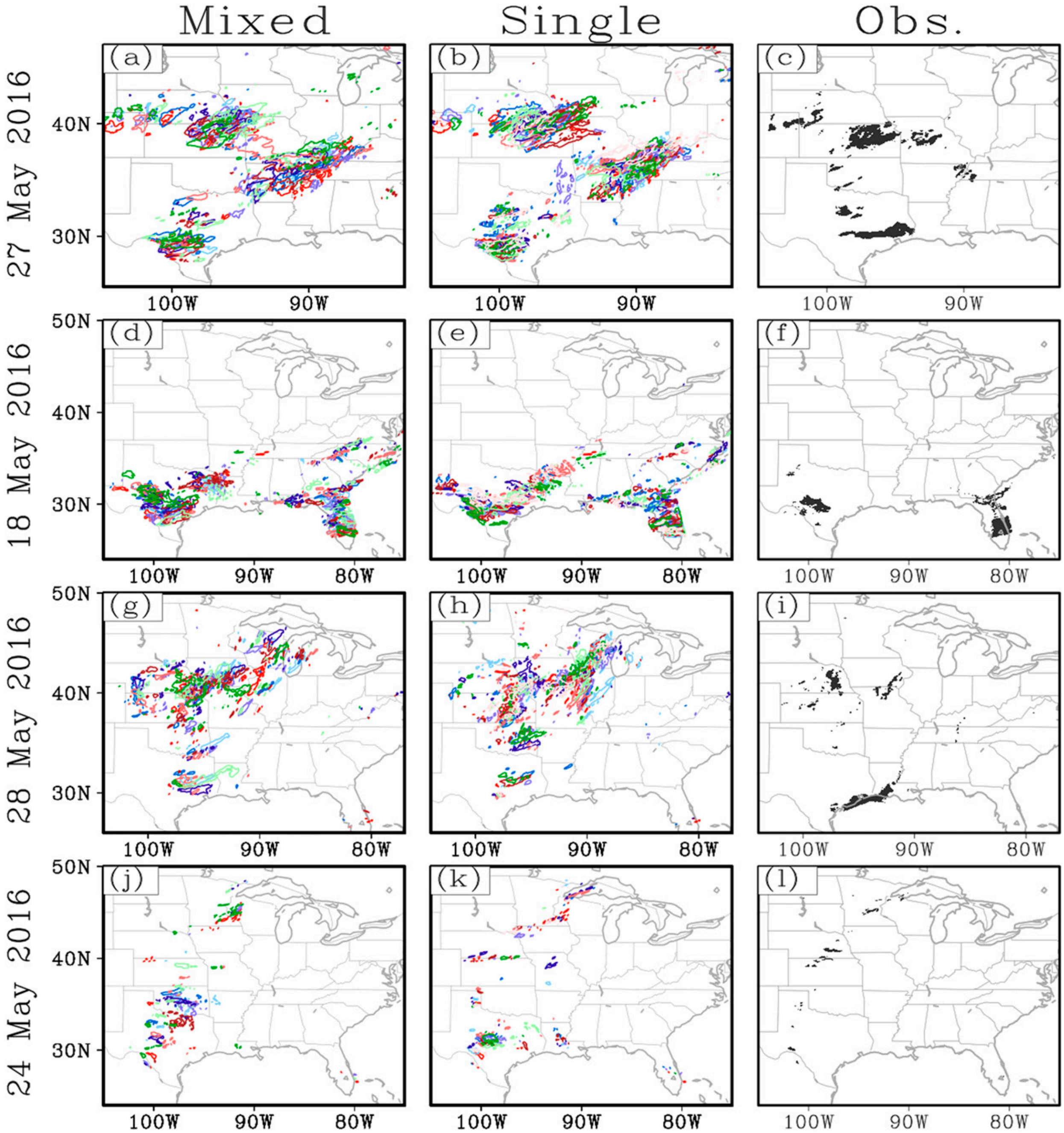

FIG. 15. Individual ensemble member 30-h 1.00-in. precipitation forecasts from (a) the mixed-physics ensemble and (b) the single-physics ensemble, valid for 0000–0600 UTC 27 May 2016. (c) Observed precipitation ≥ 1.00 in., valid for the same 6-h period as in (a) and (b). (d)–(f) As in (a)–(c), but valid for 18 May 2016. (g)–(i) As in (a)–(c), but valid for 28 May 2016. (j)–(l) As in (a)–(c), but valid for 24 May 2016.

ahead of a cold front. These storms began moving northeast while producing heavy rainfall. A 700-hPa shortwave trough provided additional forcing for ascent, helping to sustain the line of storms until approximately 0500 UTC 24 May. Around 0500 UTC, new storms began to form in northern Kansas along an outflow boundary from convection to the north; these storms led to reports of ≥1-in. rainfall before 0600 UTC. Farther south, discrete cells formed ahead of a dryline in west-central Texas around 2230 UTC 23 May. These storms moved east-northeastward and largely remained discrete, providing parts of west-central Texas with heavy rainfall before dissipating.

Interestingly, while neither ensemble performed particularly well on this day, each ensemble focused its probabilities on slightly different locations. The mixed-physics ensemble placed a local probability maximum over central Oklahoma, while the single-physics ensemble focused its probability maximum over central Texas (Figs. 14g,h). Both ensembles had individual members forecasting ≥1-in. rainfall in portions of the upper Midwest, close to where ≥1-in. rainfall occurred (Figs. 15j–l). Although both ensembles performed relatively poorly on this day, the single-physics ensemble had a greater AUC and only a slightly worse FSS. Subjectively, the single-physics ensemble outperformed the mixed-physics ensemble in this case by drastically reducing the false alarm in central Oklahoma and having more members forecast ≥1-in. rainfall in northern Kansas and the Texas Panhandle (Figs. 15j–l).

5. Summary and discussion

This study investigated how the spread and skill of mixed- and single-physics convection-allowing ensemble forecasts varied with forecast hour and spatial scale. Ensemble spread was assessed by computing ensemble variance for four variables—2-m temperature, 2-m dewpoint temperature, 500-hPa geopotential height, and hourly accumulated precipitation—using both raw and bias-corrected variance time series for forecast hours 0–36. Rank histograms were used to determine how well the spread of each ensemble's hourly precipitation forecasts corresponded to the spread of the observations. Meanwhile, ensemble skill was evaluated for forecast 2-m temperature, 2-m dewpoint temperature, and 6-h accumulated precipitation. A time series of RMSE was analyzed for 2-m temperature and dewpoint, while the 6-h precipitation forecasts were created—and assessed—in two distinct ways. First, binary (i.e., yes/no) 6-h precipitation forecasts were created using 0.10-, 0.25-, 0.50-, 0.75-, and 1.00-in. thresholds; these were evaluated for six nonoverlapping 6-h periods at spatial scales from 3 to 144 km using FSS. Additionally, probabilistic 6-h precipitation forecasts were created at each of the above five thresholds by spatially smoothing raw ensemble probabilities (i.e., the fraction of ensemble members meeting or exceeding the threshold) at each grid point. Varying values of the spatial smoothing parameter (from 2 to 144 km) were tested. Discrimination ability was measured using AUC, while reliability was assessed using attributes diagrams. Finally, 6-h, 1-in. probabilistic precipitation forecasts from the mixed- and single-physics ensembles were examined for four cases.

When the raw ensemble data were examined, the mixed-physics ensemble was found to have greater variance than the single-physics ensemble for all four variables studied at nearly all forecast hours (from 0 to 36) and spatial scales (from 3 to 288 km). However, the differences in variance were generally greatest at the smallest spatial scales and decreased as spatial scale increased. One explanation for this finding is that, as the spatial scale of the analysis is increased, precipitation systems occupy a smaller fraction of each analysis neighborhood. This is significant because the two ensembles' different representation of microphysics uncertainty only impacts each ensemble's forecast where convection exists; therefore, less fractional coverage of convection within each neighborhood implies less difference between the two ensemble forecasts. Another explanation is that localized differences in the two ensembles' forecast fields (for any of the four variables) tend to get averaged out as larger neighborhoods are considered.

Interestingly, while the variance *differences* suggested that the mixed-physics and single-physics ensemble spread became increasing similar at larger spatial scales, the variance *ratios* suggested that, proportionally, the mixed-physics ensemble provided greater spread at the larger spatial scales compared to the smaller spatial scales, at least for the 2-m temperature, 2-m dewpoint temperature, and hourly accumulated precipitation fields (the 500-hPa geopotential height variance ratios were generally quite similar at all spatial scales and forecast hours). This result was surprising. It indicated that, for the 2-m temperature, 2-m dewpoint, and hourly precipitation fields, the mixed-physics ensemble variance decreased less than the single-physics ensemble variance as spatial scale increased. Nevertheless, at large spatial scales, where the variance ratio was the lowest, the variance of both ensembles was quite small. This finding suggests that perhaps more weight should be given to the variance differences as opposed to the variance ratios when comparing the mixed- and single-physics ensemble variances at the larger spatial scales.

To remove the impact of systematic biases on the ensemble variance, a bias-correction procedure based on probability matching was applied (Ebert 2001; Clark et al. 2010); the PDF of each ensemble member was replaced with the PDF of the core01 member, since this member was present in both the mixed- and single-physics ensembles. As in Clark et al. (2010), the bias-corrected variances were generally lower than the corresponding raw variances, which makes sense given that probability matching reduces the ''artificial'' ensemble spread from systematic biases (Clark et al. 2010; Eckel and Mass 2005). Bias-correction also reduced the difference between the mixed- and single-physics ensemble variance, probably because the mixed-physics ensemble contained more systematic biases than the single-physics ensemble and therefore experienced a greater reduction in variance after calibration. Additionally, the single- to mixed-physics variance ratios moved slightly closer to 1 after bias correction at most forecast hours and spatial scales for all four variables. Thus, bias correction reduced some of the apparent spread benefits provided by the mixed-physics ensemble, suggesting that the presence of systematic biases artificially inflated spread in the raw mixed-physics ensemble. Bias correction most notably reduced the difference between the mixed- and single-physics ensembles' hourly precipitation variance. That the difference was sensitive to the bias-correction procedure suggests a large portion of the forecast precipitation variance in each ensemble (and at all spatial scales) can be attributed to the *magnitude* of the precipitation forecast and not merely the *placement* of precipitation systems.

Rank histogram analysis suggested that both the mixed- and single-physics ensembles overforecast hourly precipitation, with the mixed-physics ensemble having the greater bias. Bias correction helped flatten each ensemble's rank histogram, with the mixed-physics ensemble benefitting more from bias correction due to its greater initial systematic biases. After bias correction, both ensembles' rank histograms were slightly U-shaped for at least some forecast hours, suggesting that both ensembles were underdispersive relative to the observations. Notably, the U shape was slightly more pronounced in the single-physics ensemble. Nevertheless, the differences were small; there appeared to be only minor spread advantages to using the mixed-physics ensemble after bias correction.

Raw mixed- and single-physics ensemble forecasts had qualitatively similar hourly 2-m temperature RMSE values at all forecast hours from 0 to 36, despite the existence of statistically significant RMSE differences at 22 of those hours. Meanwhile, the mixed-physics ensemble always had a lower RMSE for forecast hourly 2-m dewpoint temperature; the RMSE differences were significant at 32 of 37 forecast hours. One possible explanation for this finding is that the biases in each member's dewpoint temperature have opposite signs due to their differing PBL schemes. Thus, when combined in an ensemble mean, the mixed-physics ensemble gave a lower RMSE than the single-physics ensemble.

Skill metrics indicated that the mixed- and single-physics ensembles had similar bias-corrected 6-h precipitation skill for most forecast periods, spatial scales, and precipitation thresholds examined. Statistically significant differences in FSS only existed within the first forecast period (i.e., forecast hours 0–6). Moreover, when they did occur, the single-physics ensemble always had the larger FSS. While the mixed-physics ensemble's 6-h probabilistic precipitation forecasts tended to have slightly greater AUC values than the corresponding single-physics forecasts, the differences were small (i.e., <0.05) and not statistically significant.

Interestingly, the degree of spatial smoothing did not have much influence on the relative skill of the mixed- and single-physics ensemble forecasts, perhaps suggesting the two ensemble forecasts differed more on the magnitude rather than location of forecast precipitation. The case studies examined herein offered some support for this idea. In the first three cases, the mixed- and single-physics forecasts assigned nonzero probabilities to similar locations, while slightly more variation was present in the forecasts' magnitudes. In the fourth case, the two ensembles had more notable differences in the placement of their nonzero probabilities, although neither ensemble performed particularly well objectively. The relative placement (and skill) of each ensemble's forecast probabilities may be due to a variety of factors, including type of convective trigger, strength of forcing for ascent, and/or dominant convective mode. For example, the mixed-physics ensemble may provide more value and skill relative to the single-physics ensemble when the large-scale forcing for ascent is weak (e.g., Stensrud et al. 2000). However, in general, across the 23 cases in the dataset, differences in the location of the two ensembles' precipitation probabilities existed but were small. Moreover, the spatial smoothing may have rendered these differences even smaller.

More spatial smoothing produced forecasts with better discrimination ability and reliability, up to a point. This result was unsurprising: smoothing reduces the magnitude of ensemble probabilities that were initially too large and spreads them spatially, thereby helping to account for ensemble underdispersion (e.g., Clark et al. 2018). Beyond 72 or 96 km, however, AUC tended to level off or diminish, and reliability started to decrease as forecast probabilities became oversmoothed. In addition to decreasing AUC and reliability, greater spatial smoothing reduced the sharpness of the higher precipitation forecasts.

6. Conclusions: Implications for convection-allowing ensemble design and future work

Overall, the mixed-physics ensemble provides slightly greater ensemble spread relative to the single-physics ensemble, especially at smaller spatial scales and if the ensemble is not calibrated for bias. This result is consistent with previous work that has found multiple microphysics and PBL parameterizations can be an important way to generate spread in convection-allowing ensembles (e.g., Johnson and Wang 2017; Clark et al. 2010). However, as the spatial scale of interest is increased, and as systematic bias is taken into account, the mixed- and single-physics ensemble variances generally become more similar.

The mixed-physics ensemble also appears to produce *slightly* more skillful precipitation forecasts than the single-physics ensemble, especially for larger precipitation thresholds at later forecast hours. Nevertheless, the differences between the mixed- and single-physics ensembles' spread and skill are generally small, especially when systematic biases are taken into account (i.e., the ensemble is well calibrated) and at larger spatial scales. Therefore, the small forecast advantages of using a mixed-physics ensemble may not outweigh other benefits of using a single-physics ensemble operationally. These benefits include: easier maintenance of a single physics suite; a more thorough, focused effort toward improving one physics package; and ensemble members generated from consistent perturbation methods, thus ensuring truly equally likely member solutions.

With that said, this study has a number of important limitations that should be considered before a final recommendation to model developers can be made. Most notably, this study examined only four variables during a single season over a subset of the United States. To be operationally useful, ensembles should function well year-round over the entire CONUS and include more than four variables. Additionally, the mixed- and single-physics forecasts should be subjectively compared more extensively and for more forecast fields than the four 1-in. precipitation forecast cases examined herein. Ideally, subjective forecaster ratings and feedback of the mixed- and single-physics ensemble forecast output could be systematically compiled over at least one full season for a variety of fields (e.g., low-level temperature, dewpoint temperature, simulated reflectivity, relative humidity). In addition to addressing these limitations, future work may wish to evaluate the individual impact of multiple microphysics and PBL parameterizations on ensemble spread and skill. Doing so would build a more complete understanding of convection-allowing ensemble design.

Acknowledgments. This work was made possible by a Presidential Early Career Award for Scientists and Engineers (PECASE). The CAPS ensemble forecasts were generated under the support of NOAA CSTAR Grant NA16NWS4680002 and HWT Grant NA15OAR4590186 on NSF Xsede Supercomputing Resources at Texas Advanced Supercomputing Center. NOAA HWT Grant NA17OAR4590186 provided additional support for data analysis. CAPS scientists Kelvin Thomas, Keith Brewster, Youngsun Jung, and Nathan Snook contributed to the projects. Additional support was provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA–University of Oklahoma Cooperative Agreement NA11OAR4320072, U.S. Department of Commerce. Stage IV precipitation data were provided by NCAR/EOL under the sponsorship of the National Science Foundation. The Stage IV data were accessed from https://data.eol.ucar.edu/. Model Evaluation Tools (MET) was developed at the National Center for Atmospheric Research (NCAR) through grants from the United States Air Force Weather Agency (AFWA) and the National Oceanic and Atmospheric Administration (NOAA). NCAR is sponsored by the United States National Science Foundation. Finally, the authors thank the three anonymous reviewers, whose feedback improved the quality of the manuscript.

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.

Adams-Selin, R. D., and C. Ziegler, 2016: Forecasting hail using a one-dimensional hail growth model within WRF. *Mon. Wea. Rev.*, **144**, 4919–4939, https://doi.org/10.1175/MWR-D-16-0027.1.

Brooks, H. E., M. Kay, and J. A. Hart, 1998: Objective limits on forecasting skill for rare events. Preprints, *19th Conf. on Severe Local Storms*, Minneapolis, MN, Amer. Meteor. Soc., 552–555.

Buizza, R., A. Hollingsworth, F. Lalaurette, and A. Ghelli, 1999: Probabilistic predictions of precipitation using the ECMWF ensemble prediction system. *Wea. Forecasting*, **14**, 168–189, https://doi.org/10.1175/1520-0434(1999)014<0168:PPOPUT>2.0.CO;2.

Chen, F., and J. Dudhia, 2001: Coupling an advanced land surface–hydrology model with the Penn State–NCAR MM5 modeling system. Part I: Model description and implementation. *Mon. Wea. Rev.*, **129**, 569–585, https://doi.org/10.1175/1520-0493(2001)129<0569:CAALSH>2.0.CO;2.

Clark, A. J., W. A. Gallus Jr., and T.-C. Chen, 2008: Contributions of mixed physics versus perturbed initial/lateral boundary conditions to ensemble-based precipitation forecast skill. *Mon. Wea. Rev.*, **136**, 2140–2156, https://doi.org/10.1175/2007MWR2029.1.

——, ——, M. Xue, and F. Kong, 2009: A comparison of precipitation forecast skill between small convection-allowing and large convection-parameterizing ensembles. *Wea. Forecasting*, **24**, 1121–1140, https://doi.org/10.1175/2009WAF2222222.1.

——, ——, ——, and ——, 2010: Growth of spread in convection-allowing and convection-parameterizing ensembles. *Wea. Forecasting*, **25**, 594–612, https://doi.org/10.1175/2009WAF2222318.1.

——, and Coauthors, 2011: Probabilistic precipitation forecast skill as a function of ensemble size and spatial scale in a convection-allowing ensemble. *Mon. Wea. Rev.*, **139**, 1410–1418, https://doi.org/10.1175/2010MWR3624.1.

——, J. Gao, P. Marsh, T. Smith, J. Kain, J. Correia, M. Xue, and F. Kong, 2013: Tornado pathlength forecasts from 2010 to 2011 using ensemble updraft helicity. *Wea. Forecasting*, **28**, 387–407, https://doi.org/10.1175/WAF-D-12-00038.1.

——, and Coauthors, 2016: Spring forecasting experiment 2016 conducted by the experimental forecast program of the NOAA/Hazardous weather testbed: Program overview and operations plan. NOAA/NSSL, 30 pp., https://hwt.nssl.noaa.gov/Spring_2016/HWT_SFE2016_operations_plan_final.pdf.

——, and Coauthors, 2017: Spring forecasting experiment 2017 conducted by the experimental forecast program of the NOAA/Hazardous weather testbed: Program overview and operations plan. NOAA/NSSL, 34 pp., https://hwt.nssl.noaa.gov/Spring_2017/HWT_SFE2017_operations_plan_FINAL.pdf.

——, and Coauthors, 2018: The community leveraged unified ensemble (CLUE) in the 2016 NOAA/Hazardous weather testbed spring forecasting experiment. *Bull. Amer. Meteor. Soc.*, **99**, 1433–1488, https://doi.org/10.1175/BAMS-D-16-0309.1.

Coniglio, M. C., K. L. Elmore, J. S. Kain, S. J. Weiss, M. Xue, and M. L. Weisman, 2010: Evaluation of WRF model output for severe weather forecasting from the 2008 NOAA Hazardous Weather Testbed Spring Experiment. *Wea. Forecasting*, **25**, 408–427, https://doi.org/10.1175/2009WAF2222258.1.

Developmental Testbed Center, 2017: MET: Version 6.1 model evaluation tools users guide. 399 pp., http://www.dtcenter.org/met/users/docs/overview.php.

Done, J., C. A. Davis, and M. Weisman, 2004: The next generation of NWP: Explicit forecasts of convection using the Weather Research and Forecasting (WRF) model. *Atmos. Sci. Lett.*, **5**, 110–117, https://doi.org/10.1002/asl.72.

Du, J., and Coauthors, 2014: NCEP regional ensemble update: Current systems and planned storm-scale ensembles. Preprints, *26th Conf. on Weather Forecasting*, Atlanta, GA, Amer. Meteor. Soc., J.1.4, https://ams.confex.com/ams/94Annual/webprogram/Paper239030.html.

Duda, J. D., X. Wang, F. Kong, and M. Xue, 2014: Using varied microphysics to account for uncertainty in warm-season QPF in a convection-allowing ensemble. *Mon. Wea. Rev.*, **142**, 2198–2219, https://doi.org/10.1175/MWR-D-13-00297.1.

Ebert, E. E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.*, **129**, 2461–2480, https://doi.org/10.1175/1520-0493(2001)129<2461:AOAPMS>2.0.CO;2.

——, 2009: Neighborhood verification: A strategy for rewarding close forecasts. *Wea. Forecasting*, **24**, 1498–1510, https://doi.org/10.1175/2009WAF2222251.1.

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.

Epstein, E. S., 1969: The role of initial uncertainties in prediction. *J. Appl. Meteor.*, **8**, 190–198, https://doi.org/10.1175/1520-0450(1969)008<0190:TROIUI>2.0.CO;2.

Gagne, D. J., II, A. McGovern, S. E. Haupt, R. A. Sobash, J. K. Williams, and M. Xue, 2017: Storm-based probabilistic hail forecasting with machine learning applied to convection-allowing ensembles. *Wea. Forecasting*, **32**, 1819–1840, https://doi.org/10.1175/WAF-D-17-0010.1.

Gallo, B. T., A. J. Clark, and S. R. Dembek, 2016: Forecasting tornadoes using convection-permitting ensembles. *Wea. Forecasting*, **31**, 273–295, https://doi.org/10.1175/WAF-D-15-0134.1.

Gallus, W. A., Jr., and J. F. Bresch, 2006: Comparison of impacts of WRF dynamic core, physics package, and initial conditions on warm season rainfall forecasts. *Mon. Wea. Rev.*, **134**, 2632–2641, https://doi.org/10.1175/MWR3198.1.

Gao, J., M. Xue, K. Brewster, and K. K. Droegemeier, 2004: A three-dimensional variational data analysis method with recursive filter for Doppler radars. *J. Atmos. Oceanic Technol.*, **21**, 457–469, https://doi.org/10.1175/1520-0426(2004)021<0457:ATVDAM>2.0.CO;2.

Gilmore, M. S., J. M. Straka, and E. N. Rasmussen, 2004: Precipitation uncertainty due to variations in precipitation particle parameters within a simple microphysics scheme. *Mon. Wea. Rev.*, **132**, 2610–2627, https://doi.org/10.1175/MWR2810.1.

Good, P. I., 2006: *Resampling Methods*. Birkhauser Boston, 228 pp.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155–167, https://doi.org/10.1175/1520-0434(1999)014<0155:HTFENP>2.0.CO;2.

——, 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.

Hou, D., E. Kalnay, and K. K. Droegemeier, 2001: Objective verification of the SAMEX '98 ensemble forecasts. *Mon. Wea. Rev.*, **129**, 73–91, https://doi.org/10.1175/1520-0493(2001)129<0073:OVOTSE>2.0.CO;2.

Hsu, W.-R., and A. H. Murphy, 1986: The attributes diagram: A geometrical framework for assessing the quality of probability forecasts. *Int. J. Forecast.*, **2**, 285–293, https://doi.org/10.1016/0169-2070(86)90048-8.

Janjić, Z. I., 2002: Nonsingular implementation of the Mellor–Yamada level 2.5 scheme in the NCEP Meso model. NCEP Office Note 437, 61 pp., http://www.emc.ncep.noaa.gov/officenotes/newernotes/on437.pdf.

Johnson, A., and X. Wang, 2017: Design and implementation of a GSI-based convection-allowing ensemble data assimilation and forecast system for the PECAN field experiment. Part I: Optimal configurations for nocturnal convection prediction using retrospective cases. *Wea. Forecasting*, **32**, 289–315, https://doi.org/10.1175/WAF-D-16-0102.1.

Kain, J. S., S. J. Weiss, J. J. Levit, M. E. Baldwin, and D. R. Bright, 2006: Examination of convection-allowing configurations of the WRF Model for the prediction of severe convective weather: The SPC/NSSL Spring Program 2004. *Wea. Forecasting*, **21**, 167–181, https://doi.org/10.1175/WAF906.1.

——, and Coauthors, 2008: Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP. *Wea. Forecasting*, **23**, 931–952, https://doi.org/10.1175/WAF2007106.1.

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 T. N. Palmer, 2008: Ensemble forecasting. *J. Comput. Phys.*, **227**, 3515–3539, https://doi.org/10.1016/j.jcp.2007.02.014.

Lin, Y., 2011. GCIP/EOP Surface: Precipitation NCEP/EMC 4KM Gridded Data (GRIB) Stage IV Data, version 1.0. UCAR/NCAR Earth Observing Laboratory, accessed 23 June 2017, https://data.eol.ucar.edu/dataset/21.093.

Loken, E., A. Clark, M. Xue, and F. Kong, 2017: Comparison of next-day probabilistic severe weather forecasts from coarse- and fine-resolution CAMs and a convection-allowing ensemble. *Wea. Forecasting*, 32, 1403–1421, https://doi.org/10.1175/WAF-D-16-0200.1.

Lorenz, E. N., 1969: The predictability of a flow which possesses many scales of motion. *Tellus*, **21**, 289–307, https://doi.org/10.3402/tellusa.v21i3.10086.

Marzban, C., 2004: The ROC curve and the area under it as performance measures. *Wea. Forecasting*, **19**, 1106–1114, https://doi.org/10.1175/825.1.

Mason, S. J., and N. E. Graham, 2002: Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Quart. J. Roy. Meteor. Soc.*, **128**, 2145–2166, https://doi.org/10.1256/003590002320603584.

Mellor, G. L., and T. Yamada, 1982: Development of a turbulence closure model for geophysical fluid problems. *Rev. Geophys.*, **20**, 851–875, https://doi.org/10.1029/RG020i004p00851.

Milbrandt, J. A., and M. K. Yau, 2005: A multimoment bulk microphysics parameterization. Part I: Analysis of the role of the spectral shape parameter. *J. Atmos. Sci.*, **62**, 3051–3064, https://doi.org/10.1175/JAS3534.1.

Mittermaier, M., and N. Roberts, 2010: Intercomparison of spatial forecast verification methods: Identifying skillful spatial scales using the fractions skill score. *Wea. Forecasting*, **25**, 343–354, https://doi.org/10.1175/2009WAF2222260.1.

——, ——, and S. A. Thompson, 2013: A long-term assessment of precipitation forecast skill using the Fractions Skill Score. *Met. Apps.*, **20**, 176–186, https://doi.org/10.1002/met.296.

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.

Morrison, H., and J. A. Milbrandt, 2015: Parameterization of cloud microphysics based on the prediction of bulk ice particle properties. Part I: Scheme description and idealized tests. *J. Atmos. Sci.*, **72**, 287–311, https://doi.org/10.1175/JAS-D-14-0065.1.

——, J. A. Curry, and V. I. Khvorostyanov, 2005: A new double-moment microphysics parameterization for application in cloud and climate models. Part I: Description. *J. Atmos. Sci.*, **62**, 1665–1677, https://doi.org/10.1175/JAS3446.1.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595–600, https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.

Nakanishi, M., 2000: Large-eddy simulation of radiation fog. *Bound.-Layer Meteor.*, **94**, 461–493, https://doi.org/10.1023/A:1002490423389.

——, 2001: Improvement of the Mellor-Yamada turbulence closure model based on large-eddy simulation data. *Bound.-Layer Meteor.*, **99**, 349–378, https://doi.org/10.1023/A:1018915827400.

——, 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.

Noh, Y., W. G. Cheon, S.-Y. Hong, and S. Raasch, 2003: Improvement of the K-profile model for the planetary boundary layer based on large eddy simulation data. *Bound.-Layer Meteor.*, **107**, 401–427, https://doi.org/10.1023/A:1022146015946.

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.

Roebber, P. J., D. M. Schultz, B. A. Colle, and D. J. Stensrud, 2004: Toward improved prediction: High-resolution and ensemble modeling systems in operations. *Wea. Forecasting*, **19**, 936–949, https://doi.org/10.1175/1520-0434(2004)019<0936:TIPHAE>2.0.CO;2.

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.

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.

——, G. S. Romine, K. R. Smith, and M. L. Weisman, 2014: Characterizing and optimizing precipitation forecasts from a convection-permitting ensemble initialized by a mesoscale ensemble Kalman filter. *Wea. Forecasting*, **29**, 1295–1318, https://doi.org/10.1175/WAF-D-13-00145.1.

——, ——, K. Fossell, R. Sobash, and M. Weisman, 2017: Toward 1-km ensemble forecasts over large domains. *Mon. Wea. Rev.*, **145**, 2943–2969, https://doi.org/10.1175/MWR-D-16-0410.1.

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.

Snook, N., and M. Xue, 2008: Effects of microphysical drop size distribution on tornadogenesis in supercell thunderstorms. *Geophys. Res. Lett.*, **35**, L24803, https://doi.org/10.1029/2008GL035866.

Sobash, R. A., J. S. Kain, D. R. Bright, A. R. Dean, M. C. Coniglio, and S. J. Weiss, 2011: Probabilistic forecast guidance for severe thunderstorms based on the identification of extreme phenomena in convection-allowing model forecasts. *Wea. Forecasting*, **26**, 714–728, https://doi.org/10.1175/WAF-D-10-05046.1.

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.

Thompson, G., R. M. Rasmussen, and K. Manning, 2004: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part I: Description and sensitivity analysis. *Mon. Wea. Rev.*, **132**, 519–542, https://doi.org/10.1175/1520-0493(2004)132<0519:EFOWPU>2.0.CO;2.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317–2330, https://doi.org/10.1175/1520-0477(1993)074<2317:EFANTG>2.0.CO;2.

——, and ——, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.*, **125**, 3297–3319, https://doi.org/10.1175/1520-0493(1997)125<3297:EFANAT>2.0.CO;2.

van den Heever, S. C., and W. R. Cotton, 2004: The impact of hail size on simulated supercell storms. *J. Atmos. Sci.*, **61**, 1596–1609, https://doi.org/10.1175/1520-0469(2004)061<1596:TIOHSO>2.0.CO;2.

Wandishin, M. S., S. L. Mullen, D. J. Stensrud, and H. E. Brooks, 2001: Evaluation of a short-range multimodel ensemble system. *Mon. Wea. Rev.*, **129**, 729–747, https://doi.org/10.1175/1520-0493(2001)129<0729:EOASRM>2.0.CO;2.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction*. Academic Press, 467 pp.

——, 2001: A skill score based on economic value for probability forecasts. *Meteor. Appl.*, **8**, 209–219, https://doi.org/10.1017/S1350482701002092.

Xue, M., and Coauthors, 2007: CAPS real-time storm-scale ensemble and high-resolution forecasts as part of the NOAA Hazardous Weather Testbed 2007 Spring Experiment. Preprints, *22nd Conf. on Weather Analysis and Forecasting/18th Conf. on Numerical Weather Prediction,* Salt Lake City, UT, Amer. Meteor. Soc., 3B, http://ams.confex.com/ams/pdfpapers/124587.pdf.

——, D. Wang, J. Gao, K. Brewster, and K. K. Droegemeier, 2003: The Advanced Regional Prediction System (ARPS), storm-scale numerical weather prediction and data assimilation. *Meteor. Atmos. Phys.*, **82**, 139–170, https://doi.org/10.1007/s00703-001-0595-6.