AMERICAN METEOROLOGICAL SOCIETY

Weather and Forecasting

EARLY ONLINE RELEASE

The DOI for this manuscript is doi: 10.1175/WAF-D-18-0078.1

The final published version of this manuscript will replace the preliminary version at the above DOI once it is available.

If you would like to cite this EOR in a separate work, please use the following full citation:

Loken, E., A. Clark, M. Xue, and F. Kong, 2018: Spread and Skill in Mixed- and Single-Physics Convection-Allowing Ensembles. Wea. Forecasting. doi:10.1175/WAF-D-18-0078.1, in press.

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

Eric D. Loken[1,2,3], *Adam J. Clark*[3], *Ming Xue*[2,4], *Fanyou Kong*[4]

[1]*Cooperative Institute for Mesoscale Meteorological Studies, The University of Oklahoma, Norman, Oklahoma*
[2] *School of Meteorology, University of Oklahoma, Norman, Oklahoma*
[3]*NOAA/OAR National Severe Storms Laboratory, Norman, Oklahoma*
[4]*Center for Analysis and Prediction of Storms, The University of Oklahoma, Norman, Oklahoma*

Submitted to
Weather and Forecasting
April 2018

*Corresponding author's address:
Eric D. Loken
National Weather Center, NSSL/FRDD
120 David L. Boren Blvd.,
Norman, OK, 73072
Email: eric.d.loken@noaa.gov
Phone: 405-325-3041

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-hour 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-hour 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 pathlength forecasts (Clark et al. 2013), and probabilistic tornado (Gallo et al. 2016) and hail (Gagne et al. 2017; Adams-Selin 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

78 long been known that ensemble mean forecasts tend to outperform forecasts from similarly-
79 configured deterministic models at convection-parameterizing resolution (e.g., Epstein 1969;
80 Leith 1974; Clark et al. 2009), recent evidence suggests that convection-allowing ensembles tend
81 to outperform deterministic models at convection-allowing resolution as well (e.g., Coniglio et
82 al. 2010; Loken et al. 2017; Schwartz et al. 2017).

83 Despite the promise of convection-allowing ensembles, much is still unknown about their
84 optimal configuration (e.g., Roebber et al. 2004; Romine et al. 2014; Duda et al. 2014; Johnson
85 and Wang 2017). One problem is that the vast majority of convection-allowing ensembles are
86 under-dispersive [i.e., observed events routinely fall outside of the forecast probability density
87 function (PDF)], especially for precipitation fields (e.g., Clark et al. 2008, 2010; Romine et al.
88 2014). Many previous studies have investigated methods to increase ensemble spread at
89 convective-parameterizing resolutions, including perturbing initial conditions (e.g., Toth and
90 Kalnay 1993, 1997; Molteni 1996) and using multiple models (e.g., Wandishin et al. 2001; Hou
91 et al. 2001; Ebert 2001; Eckel and Mass 2005) and physics parameterizations (e.g., Stensrud et
92 al. 2000; Gallus and Bresch 2006). More recent work has studied the impact of incorporating
93 multiple planetary boundary layer (PBL) and/or microphysics schemes within convection-
94 allowing ensembles (e.g., Schwartz et al. 2010; Duda et al. 2014; Johnson and Wang 2017),
95 generally finding that mixed-microphysics and mixed-PBL ensembles result in improved
96 ensemble spread and skill. For example, during the 2015 Plains Elevated Convection at Night
97 (PECAN) experiment, Johnson and Wang (2017) found that both of two mixed-physics
98 convection-allowing ensembles—which used a variety of microphysics and PBL schemes—
99 produced better nocturnal precipitation and non-precipitation forecasts compared to a single-
100 physics ensemble, which used Thompson microphysics (Thompson et al. 2004) and the Mellor-
101 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 Community Leveraged Unified

Ensemble (CLUE; Clark et al. 2018), a collection of 65 ensemble members with similar specifications and post-processing methods contributed by a variety of organizations [e.g., the National Severe Storms Laboratory (NSSL), the Center for Analysis and Prediction of Storms (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-hour accumulated precipitation. Up to 36-hour 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-hour forecast data from two 2016 CLUE subsets is 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 (ICs/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 was 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.

174

175 *(b) Evaluating ensemble spread*

176 *1) ENSEMBLE VARIANCE*

177 To determine ensemble spread, forecast ensemble variance is computed for four
178 variables—2-m temperature, 2-m dewpoint temperature, 500-hPa geopotential height, and hourly
179 accumulated precipitation—for forecast hours 0-36 using equation (B7) in Eckel and Mass
180 (2005):

181 $$\text{Variance} = \frac{1}{M}\sum_{m=1}^{M}\left[\frac{1}{(n-1)}\sum_{i=1}^{n}(e_{m,i} - \bar{e}_m)^2\right] \qquad (1),$$

182 where M is the number of forecast-observation data pairs (which, here, includes the number of
183 non-overlapping spatial windows in the domain over each of the 23 days in the analysis), n is the
184 number of ensemble members, $e_{m,i}$ is the value of the *i*th ensemble member at *m*, and $\bar{e}_m$ is the
185 ensemble mean at *m*. To assess the impact of spatial scale, variance is calculated for square
186 neighborhoods of varying sizes using the "upscaling" method (Ebert 2009), which assigns the
187 mean of the finer-resolution grid boxes making up a given neighborhood to that neighborhood.
188 While a variety of neighborhood sizes from 3- to 720-km are analyzed, only 5 sizes are displayed
189 herein. These neighborhoods contain 1, 8, 24, 48, and 96 grid boxes per side. Since all ensemble
190 members operate at 3-km horizontal grid spacing, the 5 neighborhoods measure: 3-, 24-, 72-,
191 144-, and 288-km, respectively, on each side. Only neighborhoods falling completely within the
192 analysis domain are included in the variance calculations, and the "upscale" averaging is done
193 prior to computing the ensemble mean. The difference between the mixed- and single-physics
194 ensemble variance (i.e., mixed-physics variance – single-physics variance) and the ratio of
195 single-physics ensemble variance to mixed-physics ensemble variance (i.e., single-physics

196 variance/mixed-physics variance) are also computed.

197 Because systematic biases from each ensemble member contribute to forecast spread but
198 not to forecast uncertainty (since systematic biases are not uncertain; e.g., Eckel and Mass 2005;
199 Clark et al. 2011; Clark et al. 2010), a probability matching technique (Ebert 2001; Clark et al.
200 2010) is used to eliminate systematic biases among the ensemble members. Conceptually, this
201 technique assigns the probability distribution function (PDF) of one dataset to another dataset to
202 eliminate systematic ensemble biases. Herein, because the core01 member serves as the control
203 member of both the mixed- and single-physics ensemble, the PDF of the core01 member is
204 assigned to each of the other ensemble members. This is done by first sorting each member's
205 forecast precipitation values from all grid points on a given day and forecast hour from largest to
206 smallest. Then, for each member, the grid point containing the largest forecast precipitation value
207 is replaced with the largest forecast value from the core01 member, and so on until all of the
208 values have been replaced. In this way, the spatial patterns of each member's original forecasts
209 are maintained, but the amplitudes of each member's forecast are replaced with amplitudes from
210 the core01 member (e.g., Clark et al. 2010). Hence, after probability matching, all ensemble
211 members contain the same bias (i.e., the bias of the core01 member) for a given forecast hour on
212 a given day, where bias is defined by:

213

214 $$\text{bias} = \frac{\frac{1}{N}\sum_{i=1}^{N} F_i}{\frac{1}{N}\sum_{i=1}^{N} O_i} = \frac{\sum_{i=1}^{N} F_i}{\sum_{i=1}^{N} O_i} \qquad (2),$$

215

216 where N is the number of grid points within the analysis domain, F_i is the forecast precipitation
217 value at point *i*, and O_i is the observed precipitation value at point *i*. Unlike in Clark et al. (2010),
218 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 – 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 under-dispersion 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 2,232 Automated Surface Observing Systems (ASOS) 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 0-36, 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 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 – single-physics RMSE) is compared to the null distribution to assess significance using $\alpha = 0.05$.

2) SIX-HOUR PRECIPITATION

6-hour ensemble precipitation forecasts are evaluated for six non-overlapping 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 non-systematic 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:

[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 spin-up period for each ensemble member. Results from the early forecast periods should be interpreted accordingly.

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

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-hour precipitation at each of the aforementioned 6-hour forecast periods using 0.10-, 0.25-, 0.50-, 0.75-, and 1.00-inch precipitation 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:

$$\mathrm{FSS_{useful}} = 0.5 + \frac{\mathrm{f_0}}{2} \qquad (4),$$

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 FSS_{useful}, as given in equation (4), is equivalent to $FSS_{uniform}$ in Roberts and Lean (2008). The smallest scale for which $FSS = FSS_{useful}$ 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-hour periods.

The two ensembles' 6-hour accumulated precipitation forecasts are further evaluated using area under the relative operating characteristics curve (AUC; e.g., Marzban 2004), which measures a forecast system's ability to discriminate between events and non-events (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 (QPF) forecasts 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 2-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:

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

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 *n*th 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; equation 3 in Loken et al. 2017) and probability of false detection (POFD; equation 4 in Loken et al. 2017) are computed at the following levels of

probability: 1, 2, and 5 to 95% by 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-hour 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:

$$\text{Reliability} = \frac{1}{N}\sum_{k=1}^{K} n_k(p_k - \bar{o}_k)^2 \qquad (6),$$

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 $\bar{o}_k$ 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 greatest (lowest) variances at a given forecast hour (Fig. 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 (Fig. 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 (Fig. 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. 500-hPa geopotential height variance generally increases with time, with the variance increasing faster for the smaller spatial scales.

Variance differences (Fig. 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 $\frac{\text{single-physics ensemble variance}}{\text{mixed-physics ensemble variance}}$ are computed (Fig. 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 (Fig. 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 (Fig. 3c,f,l). Thus, even while the *difference* between the mixed- and single-physics variances is lowest for the largest spatial scales (Fig. 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

403 (Fig. 4a,d,g,j). This result is expected given that the bias-correction procedure removes some of
404 the "artificial" spread that results from systematic biases among the ensemble members (Clark et
405 al. 2010). The reduced spread in the bias-corrected time series is most clearly seen in the 500-
406 hPa geopotential height variances (Fig. 4g; Fig. 3g).

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

416 For each of the four variables studied, bias-correction tends to push the
417 $\frac{\text{single-physics ensemble variance}}{\text{mixed-physics ensemble variance}}$ ratios slightly toward 1.0 (Fig. 4c,f,i,l; Fig. 3c,f,i,l). In nearly all
418 cases, this change indicates an increased proportion of variance generated by the single-physics
419 ensemble when bias-correction is applied. The effect is seen for most spatial scales and forecast
420 hours.

421

422 *3) RANK HISTOGRAMS*

423 Before correcting for systematic biases, both ensembles' rank histograms are skewed to
424 the right for all six forecast hours examined (Fig. 5), suggesting both ensembles tend to over-
425 forecast 1-h precipitation. The mixed-physics rank histograms (Fig. 5a-f) tend to be more

strongly skewed than the corresponding single-physics rank histograms (Fig. 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 over-forecasting 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 (Fig. 6d,j), suggesting that both ensembles are under-dispersive relative to the observations. Adding more members to each ensemble could potentially alleviate this under-dispersion 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 14-27, 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). 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 14-25 (i.e., 1400 UTC – 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 errors—in its 2-m temperature forecasts compared to the single-physics ensemble.

2) SIX-HOUR PECIPITATION

(i) FSS

For all six 6-hour 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-inch; Fig 8a-f). In all cases, the FSS progressively decreases as the precipitation threshold increases from 0.10- to 1.00-inch. 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 six-hour period (i.e., forecast hours 0-6; Fig. 8b-f). During the first six-hour forecast period, FSS differences are significant at one spatial scale (144-km) for the 0.75-inch forecasts, seven spatial scales (3-, 18-, 24-, 36-, 48-, 72-, and 144-km) for the 0.50-inch forecasts, and all ten spatial scales for the 0.25- and 0.10-inch 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-inch thresholds (Fig. 9a-d). When a forecast's FSS decreases below FSS_{useful} depends on both the precipitation threshold and spatial scale of the forecast; higher precipitation thresholds and smaller-scale forecasts reach FSS_{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 FSS_{useful}. Both the mixed- and single-physics ensembles have qualitatively similar FSS values for a given 6-hour forecast period, precipitation threshold, and spatial scale. Statistically significant differences between the two ensembles' FSS exist only during the first 6-hour forecast period, and the single-physics ensemble has the higher FSS in all cases of significance.

(ii) AUC From 6-Hour 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 $\geq$ 1.00-inch 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 (Fig. 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-inch) and during the 6-hour period ending at forecast hour 30 (i.e., 0000-0600 UTC one day after the forecast is initialized; Fig. 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 smoothing parameter) at all six 6-hour forecast periods is assessed explicitly in Fig. 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 (Fig. 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 over-forecast while the lower probabilities tend to be slightly under-forecast (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 (Fig. 12a-d). As the spatial smoothing parameter is increased beyond 96-km, the forecasts tend toward an under-forecasting 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-inch. Statistically significant differences between the two ensembles' reliability component of the Brier score exist only at the

0.10- and 0.25-inch thresholds (Fig. 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.

Reliability is sensitive to precipitation threshold: particularly for the smaller spatial scales, the lower-threshold forecasts (i.e., the 0.10- and 0.25-inch forecasts; Fig. 12a-b) have better reliability than the higher-threshold forecasts (i.e., the 0.50- and 1.00-inch forecasts; Fig. 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 (Fig. 13a-f).

4. Select cases

To provide a visual comparison of the mixed- and single-physics ensemble precipitation forecasts, 1-inch 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-hour 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-inch of observed 6-hour rainfall inside the analysis domain, while the final case was chosen subjectively as an interesting case in which both ensembles produced large forecast misses and low objective verification scores. The 1-inch threshold was selected since that threshold gave the greatest difference between mixed- and single-physics ensemble AUC. The 1-inch 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 on 26 May and grew upscale as they moved to the northeast. In the late afternoon, additional storms formed along the dry line, 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 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 $\geq$ 1-inch 6-hour rainfall (Fig. 14a-b), and both demonstrate reasonable forecast quality, with

570 both AUC values ≥ 0.75. Differences between the two forecasts include the mixed-physics
571 ensemble's better prediction of heavy rainfall in central Missouri as well as in southwestern
572 Nebraska and northeastern Colorado. Additionally, the magnitudes of the two forecasts'
573 probabilities differ slightly in northeastern Kansas and southcentral Arkasas. However, these
574 differences are minor; the two forecasts are generally similar. Neither predicts the southeastern
575 Texas or western Oklahoma precipitation well. Plots of individual member 1-inch forecasts (Fig.
576 15a-c) are also similar between the two ensembles, although the mixed-physics ensemble more
577 accurately depicts the threat of heavy precipitation in southwestern Nebraska and central
578 Missouri. Nevertheless, given their broad similarities, both ensembles would likely provide
579 comparable value to forecasters.
580
581 *(b) 18 May 2016*
582 Two main regions in the analysis domain recorded ≥ 1-inch observed 6-hour
583 precipitation totals from 0000-0600 UTC on 18 May 2016: southcentral Texas and the Florida
584 Peninsula. In central Texas, storms initiated along a southwest–northeast oriented cold front
585 during the mid-afternoon of 17 May. These storms grew upscale as they propagated south-
586 southeastward during the forecast period, bringing heavy rainfall to portions of southcentral
587 Texas. In Florida, a broad region of storms formed as a low-amplitude 700-hPa shortwave trough
588 moved northeastward through the Peninsula, providing forcing for ascent in an environment
589 characterized by rich boundary layer moisture and moderate instability.
590 Both ensembles produce similar forecasts, which perform well (Fig. 14c-d). Each forecast
591 assigns modest probabilities to southcentral Texas and the Florida Peninsula, where heavy
592 rainfall was observed; however, both forecasts also have a notable false alarm region extending
593 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-inch precipitation in southern Arkansas, and it has one member forecasting ≥ 1.00-inch precipitation in the southern Texas Panhandle near a small region of > 1.00-inch observed precipitation (Fig. 15d-f). Nevertheless, these differences are subtle, and the two ensemble forecasts are generally similar.

(c) 28 May 2016

At 1200 UTC on 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 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. Further south, a preexisting mesoscale convective system (MCS) moved southeastward during the forecast period, impacting southeastern Texas and southwestern Louisiana.

The two ensemble forecasts are similar but not identical (Fig. 14e-f). Both assign non-zero 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-inch precipitation, which is close to the observed scenario but displaced to the northwest (Fig. 15g-i). The mixed-physics ensemble

618 also does a better job of depicting the threat of heavy precipitation in southern Nebraska and
619 west-central Kansas, where the single-physics ensemble displays zero probabilities. Finally, the
620 mixed-physics ensemble reduces the magnitude of probabilities in southcentral Wisconsin and
621 western Illinois, where ≥ 1-inch rainfall was not observed. Still, the two forecasts are similar
622 enough that, in terms of forecast value, the mixed-physics ensemble likely provides only
623 marginal benefits over the single-physics ensemble in this case.
624
625 *(d) 24 May 2016*
626 Just before 2200 UTC on 23 May, a line of storms extending from eastern Nebraska into
627 northern Wisconsin formed ahead of a cold front. These storms began moving northeast while
628 producing heavy rainfall. A 700-hPa shortwave trough provided additional forcing for ascent,
629 helping to sustain the line of storms until approximately 0500 UTC on 24 May. Around 0500
630 UTC, new storms began to form in northern Kansas along an outflow boundary from convection
631 to the north; these storms led to reports of ≥ 1-inch rainfall before 0600 UTC. Further south,
632 discrete cells formed ahead of a dryline in west-central Texas around 2230 UTC on 23 May.
633 These storms moved east-northeastward and largely remained discrete, providing parts of west-
634 central Texas with heavy rainfall before dissipating.
635 Interestingly, while neither ensemble performed particularly well on this day, each
636 ensemble focused its probabilities on slightly different locations. The mixed-physics ensemble
637 placed a local probability maximum over central Oklahoma, while the single-physics ensemble
638 focused its probability maximum over central Texas (Fig. 14g-h). Both ensembles had
639 individual members forecasting ≥ 1-inch rainfall in portions of the upper Midwest, close to
640 where ≥ 1-inch rainfall occurred (Fig. 15j-l). Although both ensembles performed relatively
641 poorly on this day, the single-physics ensemble had a greater AUC and only a slightly worse

642 FSS. Subjectively, the single-physics ensemble outperformed the mixed-physics ensemble in this
643 case by drastically reducing the false alarm in central Oklahoma and having more members
644 forecast ≥ 1-inch rainfall in northern Kansas and the Texas Panhandle (Fig. 15j-l).
645
646 **5. Summary and discussion**
647 This study investigated how the spread and skill of mixed- and single-physics
648 convection-allowing ensemble forecasts varied with forecast hour and spatial scale. Ensemble
649 spread was assessed by computing ensemble variance for four variables—2-m temperature, 2-m
650 dewpoint temperature, 500-hPa geopotential height, and hourly accumulated precipitation—
651 using both raw and bias-corrected variance time series for forecast hours 0-36. Rank histograms
652 were used to determine how well the spread of each ensemble's hourly precipitation forecasts
653 corresponded to the spread of the observations. Meanwhile, ensemble skill was evaluated for
654 forecast 2-m temperature, 2-m dewpoint temperature, and 6-hour accumulated precipitation. A
655 time series of RMSE was analyzed for 2-m temperature and dewpoint, while the 6-hour
656 precipitation forecasts were created—and assessed—in two distinct ways. First, binary (i.e.,
657 yes/no) 6-hour precipitation forecasts were created using 0.10-, 0.25-, 0.50-, 0.75-, and 1.00-inch
658 thresholds; these were evaluated for six non-overlapping 6-hour periods at spatial scales from 3-
659 to 144-km using FSS. Additionally, probabilistic 6-hour precipitation forecasts were created at
660 each of the above five thresholds by spatially smoothing raw ensemble probabilities (i.e., the
661 fraction of ensemble members meeting or exceeding the threshold) at each grid point. Varying
662 values of the spatial smoothing parameter (from 2- to 144-km) were tested. Discrimination
663 ability was measured using AUC, while reliability was assessed using attributes diagrams.
664 Finally, 6-hour, 1-inch probabilistic precipitation forecasts from the mixed- and single-physics
665 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 0-36) 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.

690 To remove the impact of systematic biases on the ensemble variance, a bias-correction
691 procedure based on probability matching was applied (Ebert 2001; Clark et al. 2010); the PDF of
692 each ensemble member was replaced with the PDF of the core01 member, since this member
693 was present in both the mixed- and single-physics ensembles. As in Clark et al. (2010), the bias-
694 corrected variances were generally lower than the corresponding raw variances, which makes
695 sense given that probability matching reduces the "artificial" ensemble spread from systematic
696 biases (Clark et al. 2010; Eckel and Mass 2005). Bias-correction also reduced the difference
697 between the mixed- and single-physics ensemble variance, probably because the mixed-physics
698 ensemble contained more systematic biases than the single-physics ensemble and therefore
699 experienced a greater reduction in variance after calibration. Additionally, the single- to mixed-
700 physics variance ratios moved slightly closer to 1 after bias-correction at most forecast hours and
701 spatial scales for all four variables. Thus, bias-correction reduced some of the apparent spread
702 benefits provided by the mixed-physics ensemble, suggesting that the presence of systematic
703 biases artificially inflated spread in the raw mixed-physics ensemble. Bias-correction most
704 notably reduced the difference between the mixed- and single-physics ensembles' hourly
705 precipitation variance. That the difference was sensitive to the bias-correction procedure suggests
706 a large portion of the forecast precipitation variance in each ensemble (and at all spatial scales)
707 can be attributed to the *magnitude* of the precipitation forecast and not merely the *placement* of
708 precipitation systems.
709 Rank histogram analysis suggested that both the mixed- and single-physics ensembles
710 over-forecast hourly precipitation, with the mixed-physics ensemble having the greater bias.
711 Bias-correction helped flatten each ensemble's rank histogram, with the mixed-physics ensemble
712 benefitting more from bias-correction due to its greater initial systematic biases. After bias-
713 correction, both ensembles' rank histograms were slightly U-shaped for at least some forecast

hours, suggesting that both ensembles were under-dispersive 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 0-36, 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-hour 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-hour 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 non-zero probabilities to similar locations, while slightly

738 more variation was present in the forecasts' magnitudes. In the fourth case, the two ensembles
739 had more notable differences in the placement of their nonzero probabilities, although neither
740 ensemble performed particularly well objectively. The relative placement (and skill) of each
741 ensemble's forecast probabilities may be due to a variety of factors, including type of convective
742 trigger, strength of forcing for ascent, and/or dominant convective mode. For example, the
743 mixed-physics ensemble may provide more value and skill relative to the single-physics
744 ensemble when the large-scale forcing for ascent is weak (e.g., Stensrud et al. 2000). However,
745 in general, across the 23 cases in the dataset, differences in the location of the two ensembles'
746 precipitation probabilities existed but were small. Moreover, the spatial smoothing may have
747 rendered these differences even smaller.

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

755

756 **6. Conclusion: Implications for convection-allowing ensemble design and future work**

757 Overall, the mixed-physics ensemble provides slightly greater ensemble spread relative to
758 the single-physics ensemble, especially at smaller spatial scales and if the ensemble is not
759 calibrated for bias. This result is consistent with previous work that has found multiple
760 microphysics and PBL parameterizations can be an important way to generate spread in
761 convection-allowing ensembles (e.g., Johnson et al. 2017; Clark et al. 2010). However, as the

762 spatial scale of interest is increased, and as systematic bias is taken into account, the mixed- and
763 single-physics ensemble variances generally become more similar.

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

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

understanding of convection-allowing ensemble design.

Acknowledgements

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 was 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 would like to thank the three anonymous reviewers, whose feedback improved the quality of the manuscript.

804 **References**

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

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

812 Buizza, R., A. Hollingsworth, F. Lalaurette, and A. Ghelli, 1999: Probabilistic predictions of
813 precipitation using the ECMWF ensemble prediction system. *Wea. Forecasting*, **14**, 168–
814 189.

815 Chen, F., and J. Dudhia, 2001: Coupling an advanced land surface–hydrology model with the
816 Penn State–NCAR MM5 modeling system. Part I: Model description and
817 implementation. *Mon. Wea. Rev*., **129**, 569–585.

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

821 Clark, A. J., W. A. Gallus, M. Xue, and F. Kong, 2009: A comparison of precipitation forecast
822 skill between small convection-allowing and large convection-parameterizing ensembles.
823 *Wea. Forecasting*, **24**, 1121–1140, doi: 10.1175/2009WAF2222222.1.

824 Clark, A. J., W. A. Gallus Jr., M. Xue, and F. Kong, 2010: Growth of spread in convection-
825 allowing and convection-parameterizing ensembles. *Wea. Forecasting*, **25**, 594–612, doi:
826 10.1175/2009WAF2222318.1.

827 Clark, A.J., and Coauthors, 2011: Probabilistic precipitation forecast skill as a function of

828 ensemble size and spatial scale in a convection-allowing ensemble. *Mon. Wea. Rev.*, **139**,
829 1410–1418, doi: 10.1175/2010MWR3624.1.
830 Clark, A. J., J. Gao, P. Marsh, T. Smith, J. Kain, J. Correia, M. Xue, and F. Kong, 2013: Tornado
831 pathlength forecasts from 2010 to 2011 using ensemble updraft helicity. *Wea.*
832 *Forecasting*, **28**, 387–407.
833 Clark, A., and Coauthors, 2016: Spring forecasting experiment 2016 conducted by the
834 experimental forecast program of the NOAA/Hazardous weather testbed: Program
835 overview and operations plan. Available online at:
836 https://hwt.nssl.noaa.gov/Spring_2016/HWT_SFE2016_operations_plan_final.pdf.
837 Accessed 23 Jun 2017.
838 Clark A., and Coauthors, 2017: Spring forecasting experiment 2017 conducted by the
839 experimental forecast program of the NOAA/Hazardous weather testbed: Program
840 overview and operations plan. Available online at:
841 https://hwt.nssl.noaa.gov/Spring_2017/HWT_SFE2017_operations_plan_FINAL.pdf.
842 Accessed 17 January 2018.
843 Clark A. J., and Coauthors, 2018: The community leveraged unified ensemble (CLUE) in the
844 2016 NOAA/Hazardous weather testbed spring forecasting experiment. *Bulletin of the*
845 *American Meteorological Society*, **99**, 1433-1488, https://doi.org/10.1175/BAMS-D-16-
846 0309.1.
847 Coniglio, M. C., K. L. Elmore, J. S. Kain, S. J. Weiss, M. Xue, and M. L. Weisman, 2010:
848 Evaluation of WRF model output for severe weather forecasting from the 2008 NOAA
849 Hazardous Weather Testbed Spring Experiment. *Wea. Forecasting*, **25**, 408–427, doi:
850 10.1175/2009WAF2222258.1.
851 Developmental Testbed Center, 2017: MET: Version 6.1 Model Evaluation Tools Users Guide.

852 Available at http://www.dtcenter.org/met/users/docs/overview.php. 399 pp.
853 Done, J., C. A. Davis, and M. Weisman, 2004: The next generation of NWP: Explicit forecasts
854 of convection using the Weather Research and Forecasting (WRF) model. *Atmos. Sci.*
855 *Lett.*, **5**, 110–117, doi: 10.1002/asl.72.
856 Du, J., and Coauthors, 2014: NCEP regional ensemble update: Current systems and planned
857 storm-scale ensembles. Preprints, *26th Conf. on Wea. Forecasting*, Atlanta, GA, Amer.
858 Meteor. Soc., J.1.4.
859 Duda, J. D., X. Wang, F. Kong, and M. Xue, 2014: Using varied microphysics to account for
860 uncertainty in warm-season QPF in a convection-allowing ensemble. *Mon. Wea. Rev.*,
861 **142**, 2198–2219, doi: 10.1175/MWR-D-13-00297.1.
862 Ebert, E. E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of
863 precipitation. *Mon. Wea. Rev.*, **129**, 2461–2480, doi: 10.1175/1520-
864 0493(2001)129<2461:AOAPMS>2.0.CO;2.
865 Ebert, E. E., 2009: Neighborhood verification: A strategy for rewarding close forecasts. *Wea.*
866 *Forecasting*, **24**, 1498–1510.
867 Eckel, F. A., and C. F. Mass, 2005: Aspects of effective mesoscale, short-range ensemble
868 forecasting. *Wea. Forecasting*, **20**, 328–350, doi: 10.1175/WAF843.1.
869 Epstein, E. S., 1969: The role of initial uncertainties in prediction. *J. Appl. Meteor.*, **8,** 190–198.
870 Gagne, D. J., II, A. McGovern, S. E. Haupt, R. A. Sobash, J. K. Williams, and M. Xue, 2017:
871 Storm-based probabilistic hail forecasting with machine learning applied to convection-
872 allowing ensembles. *Weather and Forecasting*, **32**, 1819-1840.
873 Gallo, B. T., A. J. Clark, and S. R. Dembek, 2016: Forecasting tornadoes using convection-
874 permitting ensembles. *Wea. Forecasting*, **31**, 273–295, doi: 10.1175/WAF-D-15-0134.1.

875 Gallus, W. A., Jr., and J. F. Bresch, 2006: Comparison of impacts of WRF dynamic core, physics
876 package, and initial conditions on warm season rainfall forecasts. *Mon. Wea. Rev.*, **134**,
877 2632–2641.
878 Gilmore, M. S., J. M. Straka, and E. N. Rasmussen, 2004: Precipitation uncertainty due to
879 variations in precipitation particle parameters within a simple microphysics scheme. *Mon.*
880 *Wea. Rev.*, **132**, 2610–2627, doi: 10.1175/MWR2810.1.
881 Gao, J., M. Xue, K. Brewster, and K. K. Droegemeier, 2004: A three-dimensional variational
882 data analysis method with recursive filter for Doppler radars. *J. Atmos. Oceanic Technol*.,
883 **21**, 457–469.
884 Good, P. I., 2006: *Resampling Methods*. Birkhauser Boston, 228 pp.
885 Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea.*
886 *Forecasting*, **14**, 155–167, doi:https://doi.org/10.1175/1520-
887 0434(1999)014<0155:HTFENP>2.0.CO;2.
888 Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon.*
889 *Wea. Rev.*, **129**, 550–560, https://doi.org/10.1175/1520-
890 0493(2001)129<0550:IORHFV>2.0.CO;2.
891 Hou, D., E. Kalnay, and K. K. Droegemeier, 2001: Objective verification of the SAMEX '98
892 ensemble forecasts. *Mon. Wea. Rev.*, **129**, 73–91, doi: 10.1175/1520-
893 0493(2001)129<0073:OVOTSE>2.0.CO;2.
894 Hsu W.-R., and A. H. Murphy, 1986: The attributes diagram: A geometrical framework for
895 assessing the quality of probability forecasts. *Int. J. Forecasting*. **2**, 285-293.
896 Janjić, Z. I., 2002: Nonsingular implementation of the MellorYamada level 2.5 scheme in the
897 NCEP Meso Model. NCEP Office Note 437, 61 pp.
898 Jirak, I. L., S. J. Weiss, and C. J. Melick, 2012: The SPC storm-scale ensemble of opportunity:

899 Overview and results from the 2012 Hazardous Weather Testbed Spring Forecasting
900 Experiment. *Proc. 26th Conf. Severe Local Storms*, Nashville, TN, Amer. Meteor. Soc.,
901 137. [Available online at
902 https://ams.confex.com/ams/26SLS/webprogram/Paper211729.html].
903 Johnson, A., and Xuguang Wang, 2017: Design and implementation of a GSI-based convection-
904 allowing ensemble data assimilation and forecast system for the PECAN field
905 experiment. Part I: Optimal configurations for nocturnal convection prediction using
906 retrospective cases. *Wea. Forecasting*, **32**, 289-315.
907 Kain, J. S., S. J. Weiss, J. J. Levit, M. E. Baldwin, and D. R. Bright, 2006: Examination of
908 convection-allowing configurations of the WRF Model for the prediction of severe
909 convective weather: The SPC/NSSL Spring Program 2004. *Wea. Forecasting*, **21**, 167–
910 181, doi: 10.1175/WAF906.1.
911 Kain, J. S., and Coauthors, 2008: Some practical considerations regarding horizontal resolution
912 in the first generation of operational convection-allowing NWP. *Wea.*
913 *Forecasting*, **23**, 931–952, doi: 10.1175/WAF2007106.1.
914 Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.*, **102,** 409–418.
915 Leutbecher, M., and T. N. Palmer, 2008: Ensemble forecasting. *Journal of Computational*
916 *Physics,* **227**, 3515-3539.
917 Lin, Y. 2011. GCIP/EOP Surface: Precipitation NCEP/EMC 4KM Gridded Data (GRIB) Stage
918 IV Data. Version 1.0. UCAR/NCAR - Earth Observing Laboratory.
919 https://data.eol.ucar.edu/dataset/21.093. Accessed 23 Jun 2017.
920 Loken, E., A. Clark, M. Xue, and F. Kong, 2017: Comparison of Next-Day Probabilistic Severe
921 Weather Forecasts from Coarse- and Fine-Resolution CAMs and a Convection-Allowing
922 Ensemble. *Wea. Forecasting*. doi: 10.1175/WAF-D-16-0200.1.

923 Lorenz, E. N., 1969: The predictability of a flow which possesses many scales of motion. *Tellus*,
924 **21**, 289–307.
925 Marzban, C., 2004: The ROC curve and the area under it as performance measures. *Wea.*
926 *Forecasting*, **19**, 1106–1114.
927 Mason, S. J., and N. E. Graham, 2002: Areas beneath the relative operating characteristics
928 (ROC) and relative operating levels (ROL) curves: Statistical significance and
929 interpretation. *Quart. J. Roy. Meteor. Soc.*, **128**, 2145–2166.
930 Mellor, G. L., and T. Yamada, 1982: Development of a turbulence closure model for geophysical
931 fluid problems. *Rev. Geophys*., **20**, 851–875.
932 Milbrandt, J. A., and M. K. Yau, 2005: A multimoment bulk microphysics parameterization. Part
933 I: Analysis of the role of the spectral shape parameter. *J. Atmos. Sci.*, **62**, 3051-3064.
934 Mittermaier, M., and N. Roberts, 2010: Intercomparison of spatial forecast verification methods:
935 Identifying skillful spatial scales using the fractions skill score. *Wea. Forecasting*, **25**,
936 343–354, doi: 10.1175/2009WAF2222260.1.
937 Mittermaier, M., N. Roberts, and S. A. Thompson (2013): A long-term assessment of
938 precipitation forecast skill using the Fractions Skill Score. *Met. Apps.*, **20**, 176-186.
939 doi:10.1002/met.296.
940 Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF Ensemble Prediction
941 System: Methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122**, 73–119, doi:
942 10.1002/qj.49712252905.
943 Morrison, H., J. A. Curry, and V. I. Khvorostyanov, 2005: A new double-moment microphysics
944 parameterization for application in cloud and climate models. Part I: Description. *J.*
945 *Atmos. Sci*., **62**, 1665–1677.

946 Morrison, H., and J. A. Milbrandt, 2015: Parameterization of Cloud Microphysics Based on the
947 Prediction of Bulk Ice Particle Properties. Part I: Scheme Description and Idealized Tests.
948 *J. Atmos. Sci.*, **72**, 287–311. doi: http://dx.doi.org/10.1175/JAS-D-14-0065.1.
949 Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595–
950 600, doi:https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.
951 Nakanishi, M., 2000: Large-eddy simulation of radiation fog. *Bound.-Layer Meteor.*, **94**, 461–
952 493.
953 Nakanishi, M., 2001: Improvement of the Mellor-Yamada turbulence closure model based on
954 large-eddy simulation data. *Bound.-Layer Meteor.*, **99**, 349–378.
955 Nakanishi, M., and Niino, H., 2004: An improved Mellor-Yamada level-3 model with
956 condensation physics: Its design and verification. *Bound.-Layer Meteor.*, **112**, 1–31.
957 Nakanishi, M., and Niino, H., 2006: An improved Mellor-Yamada level-3 model: Its numerical
958 stability and application to a regional prediction of advection fog. *Bound.-Layer Meteor.*,
959 **119**, 397–407.
960 Noh, Y., W. G. Cheon, S.-Y. Hong, and S. Raasch, 2003: Improvement of the K-profile model
961 for the planetary boundary layer based on large eddy simulation data. *Bound.-Layer*
962 *Meteor.*, **107**, 401–427.
963 Roberts, N. M., and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations
964 from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97, doi:
965 10.1175/2007MWR2123.1.
966 Roebber, P. J., D. M. Schultz, B. A. Colle, and D. J. Stensrud, 2004: Toward improved
967 prediction: High-resolution and ensemble modeling systems in operations. *Wea.*
968 *Forecasting*, **19**, 936–949, doi: 10.1175/1520-0434(2004)019<0936:TIPHAE>2.0.CO;2.
969 Romine, G. S., C. S. Schwartz, J. Berner, K. R. Fossell, C. Snyder, J. L. Anderson, and M. L.

970 Weisman, 2014: Representing forecast error in a convection-permitting ensemble system.
971 *Mon. Wea. Rev.*, **142**, 4519–4541, doi: 10.1175/MWR-D-14-00100.1.
972 Schwartz, C. S., and Coauthors, 2010: Toward improved convection-allowing ensembles: Model
973 physics sensitivities and optimizing probabilistic guidance with small ensemble
974 membership. *Wea. Forecasting*, **25**, 263–280, doi: 10.1175/2009WAF2222267.1.
975 Schwartz, C. S., G. S. Romine, K. R. Smith, and M. L. Weisman, 2014: Characterizing and
976 optimizing precipitation forecasts from a convection-permitting ensemble initialized by a
977 mesoscale ensemble Kalman filter. *Wea. Forecasting*, **29**, 1295–1318,
978 https://doi.org/10.1175/WAF-D-13-00145.1.
979 Schwartz, C. S., G. S. Romine, R. A. Sobash, K. R. Fossell, and M. L. Weisman, 2015: NCAR's
980 experimental real-time convection-allowing ensemble prediction system. *Wea.*
981 *Forecasting*, **30**, 1645–1654, doi: 10.1175/WAF-D-15-0103.1.
982 Schwartz, C., G. Romine, K. Fossell, R. Sobash, and M. Weisman, 2017: Toward 1-km ensemble
983 forecasts over large domains. Mon. Wea. Rev. doi: 10.1175/MWR-D-16-0410.1.
984 Skamarock, W. C., and Coauthors, 2008: A description of the Advanced Research WRF version
985 3. NCAR Tech. Note NCAR/TN-475+STR, 113 pp., doi:10.5065/D68S4MVH.
986 Snook, N., and M. Xue, 2008: Effects of microphysical drop size distribution on tornadogenesis
987 in supercell thunderstorms. *Geophys. Res. Lett.*,**35,** L24803, doi:
988 10.1029/2008GL035866.
989 Sobash, R. A., J. S. Kain, D. R. Bright, A. R. Dean, M. C. Coniglio, and S. J. Weiss, 2011:
990 Probabilistic forecast guidance for severe thunderstorms based on the identification of
991 extreme phenomena in convection-allowing model forecasts. *Wea. Forecasting*, **26**, 714–
992 728.

993 Stensrud, D. J., J.-W. Bao, and T. T. Warner, 2000: Using initial condition and model physics
994 perturbations in short-range ensemble simulations of mesoscale convective
995 systems. *Mon. Wea. Rev.*, **128**, 2077–2107, doi: 10.1175/1520-
996 0493(2000)128<2077:UICAMP>2.0.CO;2.
997 Thompson, G., R. M. Rasmussen, and K. Manning, 2004: Explicit forecasts of winter
998 precipitation using an improved bulk microphysics scheme. Part I: Description and
999 sensitivity analysis. *Mon. Wea. Rev.*, **132**, 519–542, doi: 10.1175/1520-
1000 0493(2004)132<0519:EFOWPU>2.0.CO;2.
1001 Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of
1002 perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317–2330, doi:
1003 10.1175/15200477(1993)074<2317:EFANTG>2.0.CO;2.
1004 Toth, Z., and E. Kalnay, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon.*
1005 *Wea. Rev.*, **125**, 3297–3319, doi: 10.1175/1520-
1006 0493(1997)125<3297:EFANAT>2.0.CO;2.
1007 van den Heever, S. C., and W. R. Cotton, 2004: The impact of hail size on simulated supercell
1008 storms. *J. Atmos. Sci.*, **61**, 1596–1609, doi: 10.1175/1520-
1009 0469(2004)061<1596:TIOHSO>2.0.CO;2.
1010 Wandishin, M. S., S. L. Mullen, D. J. Stensrud, and H. E. Brooks, 2001: Evaluation of a short-
1011 range multimodel ensemble system. *Mon. Wea. Rev.*, **129,** 729–747.
1012 Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction.* Academic
1013 Press, 467 pp.
1014 Wilks, D. S., 2001: A skill score based on economic value for probability forecasts. *Meteor.*
1015 *Appl*, **8,** 209–219.
1016 Xue, M., D. Wang, J. Gao, K. Brewster, and K. K. Droegemeier, 2003: The Advanced Regional

1017 Prediction System (ARPS), storm-scale numerical weather prediction and data
1018 assimilation. *Meteor. Atmos. Phys*., **82**, 139–170.
1019 Xue, M., and Coauthors, 2007: CAPS real-time storm-scale ensemble and high-resolution
1020 forecasts as part of the NOAA Hazardous Weather Testbed 2007 Spring Experiment.
1021 Preprints, *22nd Conf. on Weather Analysis and Forecasting/18th Conf. on Numerical*
1022 *Weather Prediction,* Salt Lake City, UT, Amer. Meteor. Soc., 3B. [Available online at
1023 http://ams.confex.com/ams/pdfpapers/124587.pdf].

List of Tables

List of Figures

1050 Figure 6 As in Figure 5, but for bias-corrected mixed- and single-physics ensemble forecasts.

1051 Figure 7 Time series of mixed- (red) and single-physics (blue) ensemble root mean square error
1052 (RMSE) for (a) forecast hourly 2-m temperature and (b) 2-m dewpoint temperature. Red
1053 squares (blue circles) denote a statistically significant difference ($\alpha \leq 0.05$) between the
1054 mixed- and single-physics RMSE values with the mixed-physics (single-physics)
1055 ensemble having the lower RMSE.

1056 Figure 8 Mixed- (solid) and single-physics (dashed) ensemble fractions skill score as a function
1057 of spatial scale for the 6-hour forecast period spanning forecast hours (a) 0-6, (b) 6-12,
1058 (c) 12-18, (d) 18-24, (e) 24-30, and (f) 30-36. In each case, 0.10- (red), 0.25- (gold), 0.50-
1059 (light blue), 0.75- (dark blue), and 1.00-inch (purple) precipitation threshold forecasts are
1060 shown. Filled circles indicate significance at $\alpha = 0.05$.

1061 Figure 9 Fractions skill score as a function of forecast period for mixed- (solid) and single-
1062 physics (dashed) ensemble 6-hour precipitation forecasts at (a) 0.10-, (b) 0.25-, (c) 0.50-,
1063 and (d) 1.00-inch thresholds. In each case, 3- (red), 24- (gold), 48- (light blue), 72- (dark
1064 blue), and 144-km (purple) spatial scales are shown. The FSS_{useful} value is denoted by a
1065 solid black line. Filled circles indicate significance at $\alpha = 0.05$.

1066 Figure 10 Area under the relative operating characteristics curve (AUC) for mixed- (solid) and
1067 single-physics (dashed) 0.10- (red), 0.25- (gold), 0.50- (light blue), 0.75- (dark blue), and
1068 1.00-inch (purple) 6-hour accumulated precipitation threshold forecasts using a spatial
1069 smoothing parameter of (a) 2-, (b) 24-, (c) 48-, (d) 72-, (e) 96-, and (f) 120-km. AUC
1070 values are plotted for the 6-hour forecast periods ending at forecast hours 6, 12, 18, 24,
1071 30, and 36.

1072 Figure 11 Area under the relative operating characteristics curve (AUC) for mixed- (solid) and
1073 single-physics (dashed) 0.10- (red), 0.25- (gold), 0.50- (light blue), 0.75- (dark blue), and

1.00-inch (purple) 6-hour accumulated precipitation forecasts as a function of the spatial smoothing parameter for the 6-hour forecast period spanning forecast hours (a) 0-6, (b) 6-12, (c) 12-18, (d) 18-24, (e) 24-30, and (f) 30-36.

Figure 12 Attributes diagrams for mixed- (solid) and single-physics (dashed) ensemble 6-hour precipitation forecasts ending at forecast hour 30 using a threshold of (a) 0.10-, (b) 0.25-, (c) 0.50-, and (d) 1.00-inch. 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. Smaller plots 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.

Figure 13 Attributes diagrams for mixed- (solid) and single-physics (dashed) ensemble 0.10-inch threshold 6-hour accumulated precipitation forecasts for the 6-hour 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. Smaller plots 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.

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

Figure 15 Individual ensemble member 30-hour 1.00-inch precipitation forecasts from (a) the mixed-physics ensemble and (b) the single-physics ensemble, valid for 0000-0600 UTC on 27 May 2016. (c) Observed precipitation ≥ 1.00-inch., valid for the same 6-hour 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.

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

Month	**Day**
May	02-06; 09-13; 16-20; 23; 25-27; 30-31
June	02-03

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. 3DVAR refers to the Advanced Regional Prediction System 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 BC 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: Mellor-Yamada-Janji**ć** (MYJ; Mellor and Yamada 1982; Janji**ć** 2002), Yonsei University (YSU; Noh et al. 2003), and Mellor-Yamada-Nakanishi-Niino (MYNN; Nakanishi 2000, 2001; Nakanishi and Niino 2004, 2006).

Ens. Member	**IC**	**BC**	**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

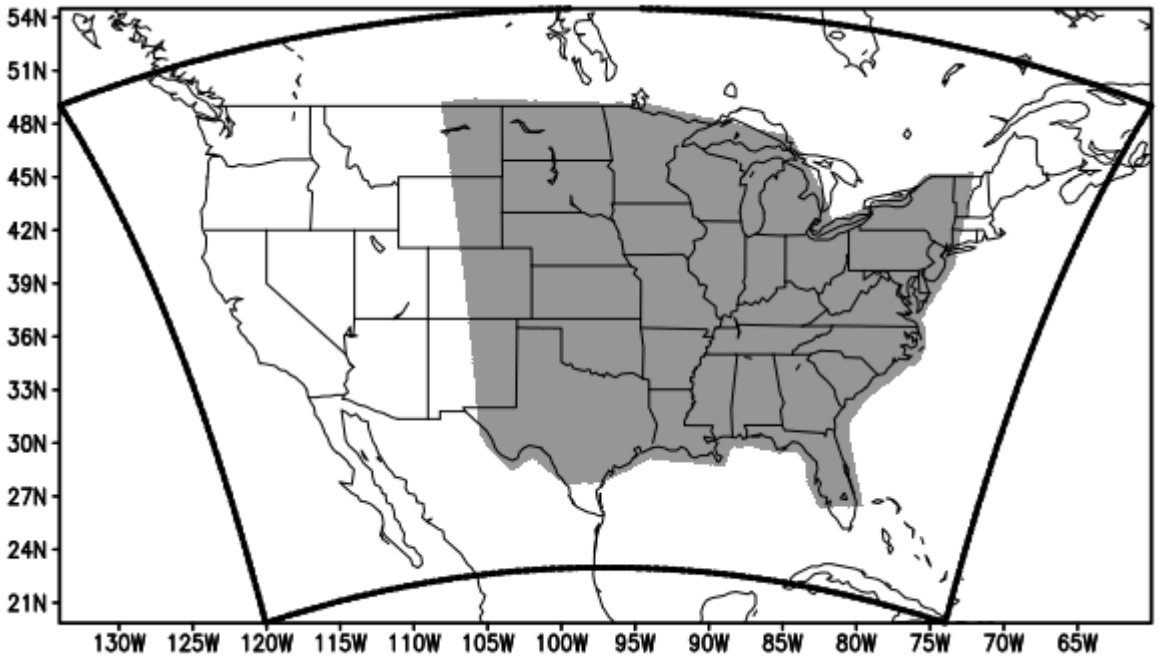

Figure 1 2016 CLUE domain (black contour) and analysis domain (gray shading).

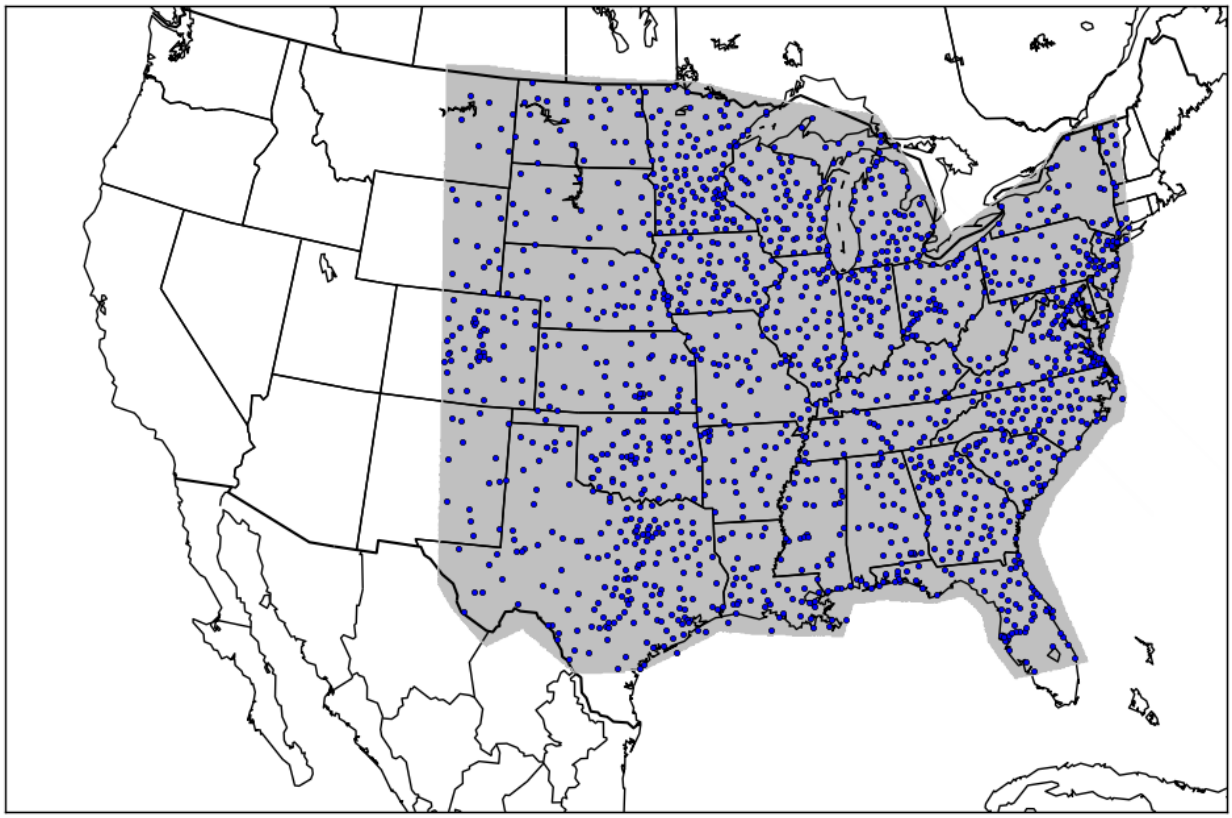

Figure 2 Automated Surface Observing Systems (blue dots) used for verification within the analysis domain (shaded).

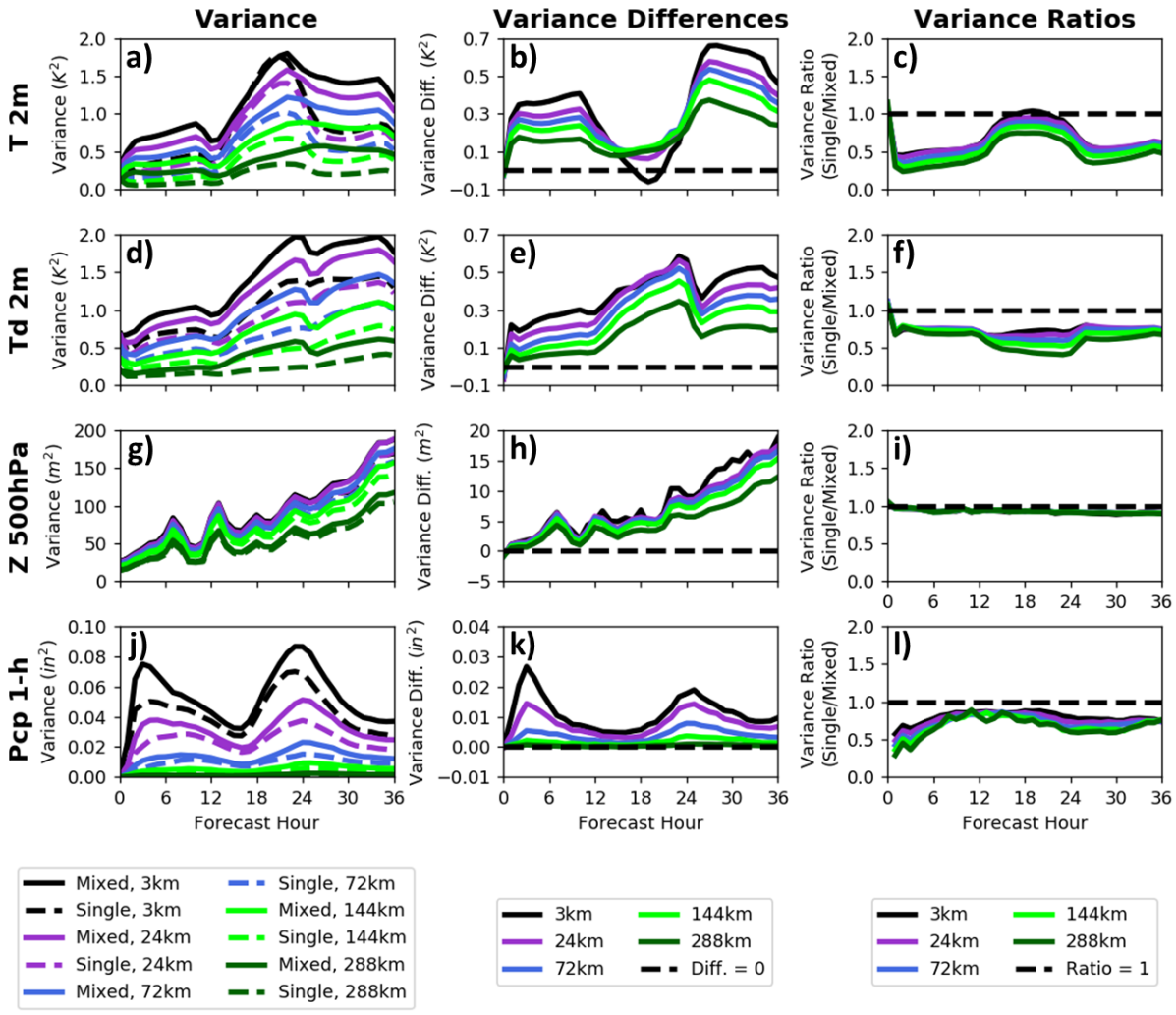

Figure 3 Time series of (a) mixed- (solid) and single-physics (dashed) ensemble variance, (b) variance differences (mixed-physics variance – single-physics variance), and (c) variance ratios (single-physics variance/mixed-physics variance) for 2-m temperature forecasts at spatial scales of 3- (black), 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 second column and a variance ratio of 1 in the third column.

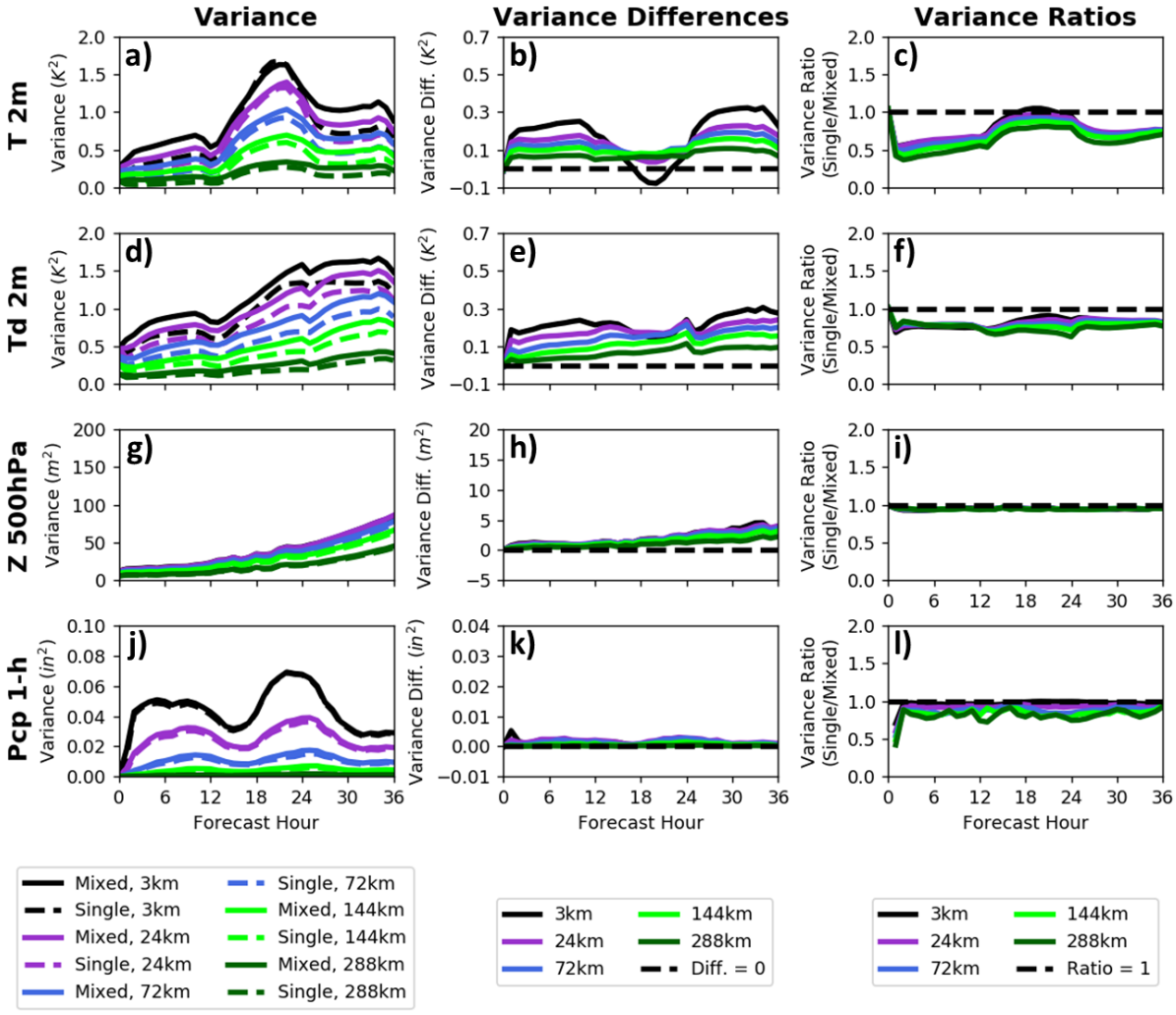

Figure 4 Bias-corrected time series of (a) mixed- (solid) and single-physics (dashed) ensemble variance, (b) variance differences (mixed-physics variance – single-physics variance), and (c) variance ratios (single-physics variance/mixed-physics variance) for 2-m temperature forecasts at spatial scales 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 second column and a variance ratio of 1 in the third column. Axis scales are identical to those in Fig. 3.

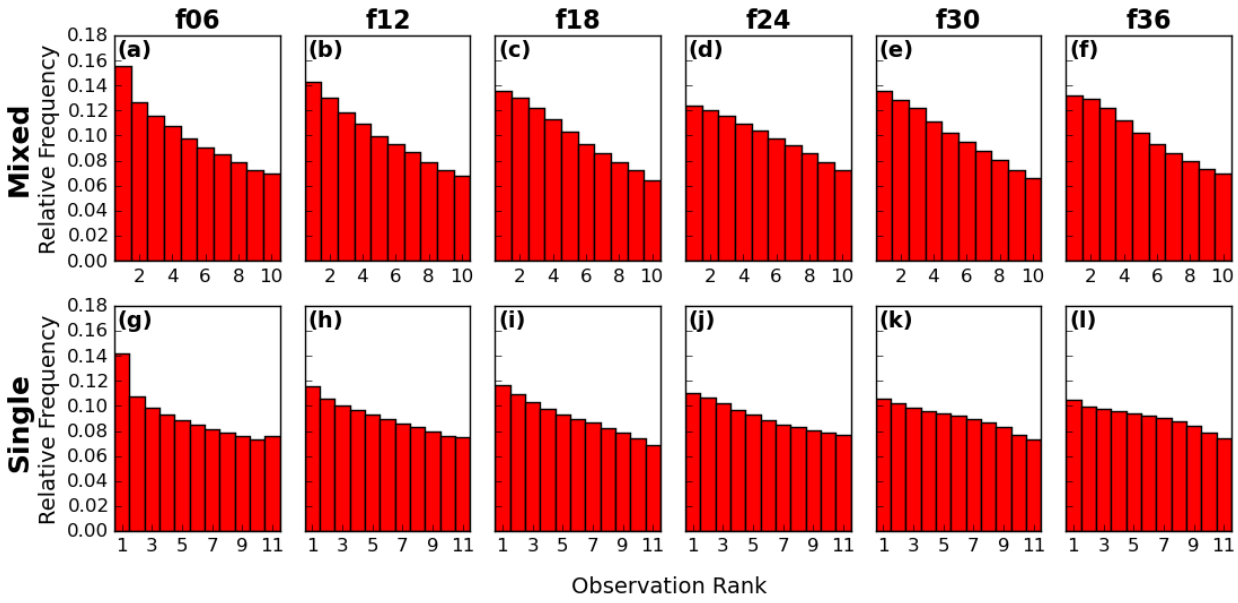

Figure 5 (a) Rank histogram for the mixed-physics ensemble's forecast 1-hour accumulated precipitation, valid for forecast hour 6. (b)-(f) As in (a) but valid for forecast hours 12, 18, 24, 30, and 36, respectively. (g)-(l) As in (a)-(f) but for the single-physics ensemble.

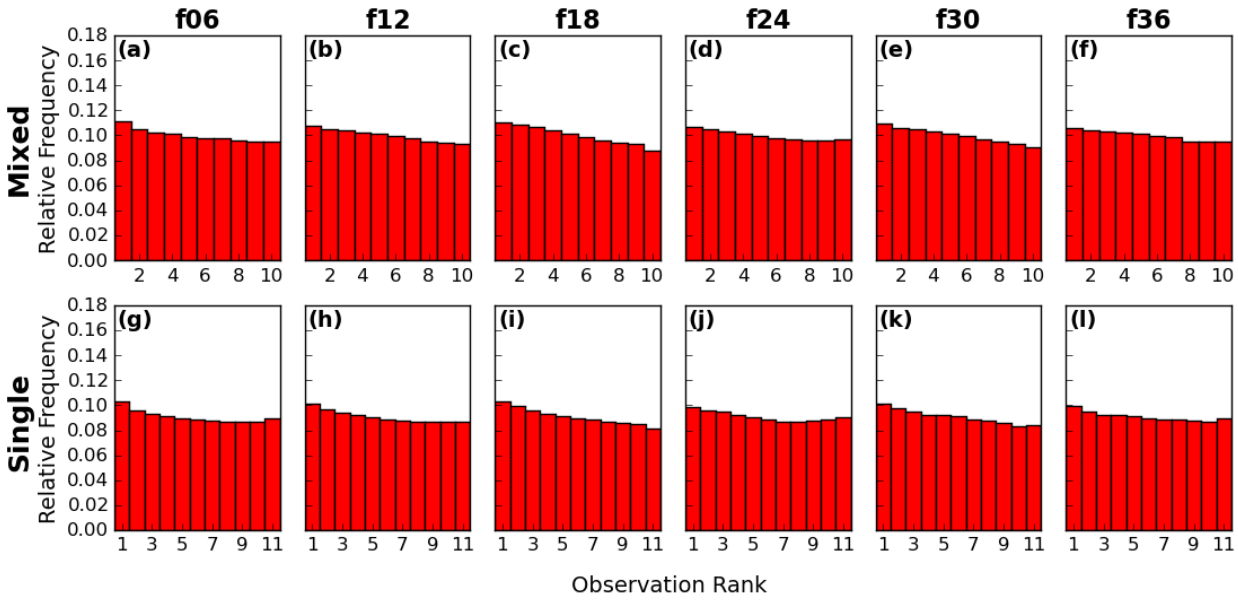

Figure 6 As in Figure 5, but for bias-corrected mixed- and single-physics ensemble forecasts.

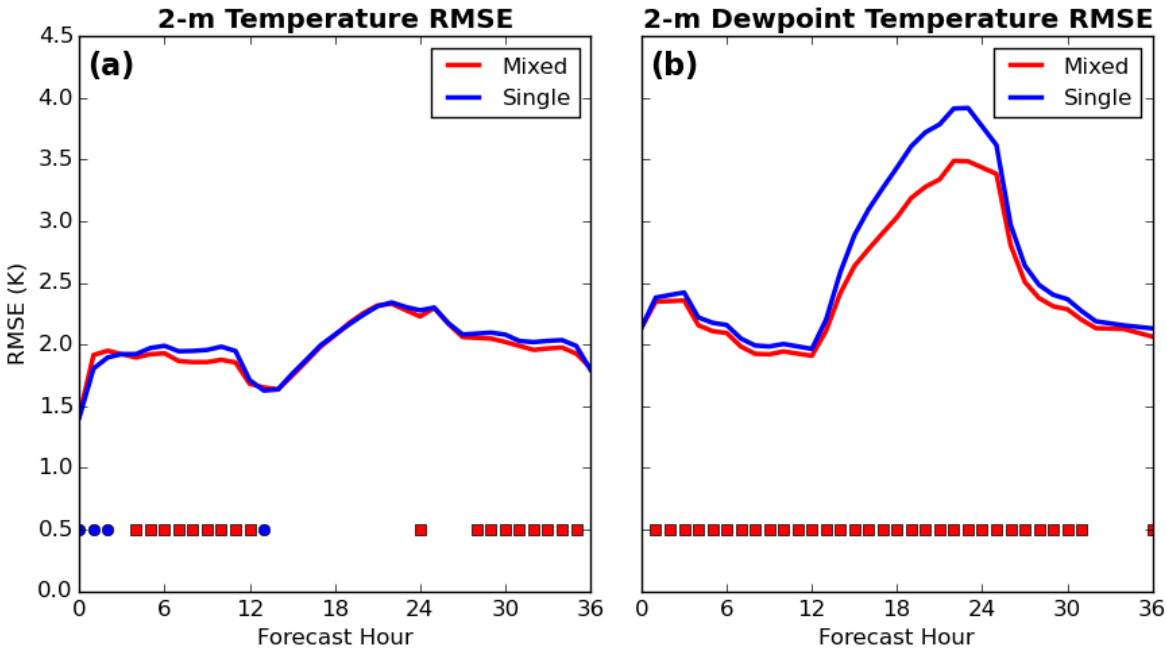

Figure 7 Time series of mixed- (red) and single-physics (blue) ensemble root mean square error (RMSE) for (a) forecast hourly 2-m temperature and (b) 2-m dewpoint temperature. Red squares (blue circles) denote a statistically significant difference ($\alpha \leq 0.05$) between the mixed- and single-physics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

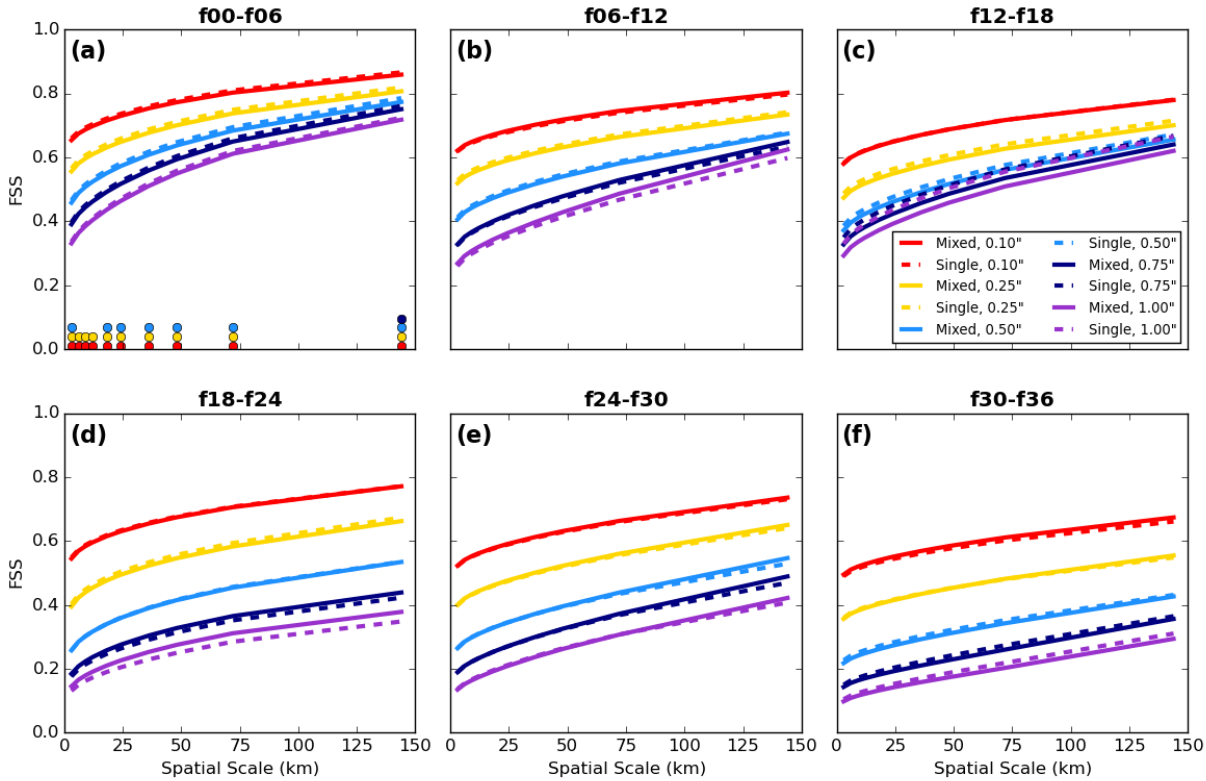

Figure 8 Mixed- (solid) and single-physics (dashed) ensemble fractions skill score as a function of spatial scale for the 6-hour forecast period spanning forecast hours (a) 0-6, (b) 6-12, (c) 12-18, (d) 18-24, (e) 24-30, and (f) 30-36. In each case, 0.10- (red), 0.25- (gold), 0.50- (light blue), 0.75-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

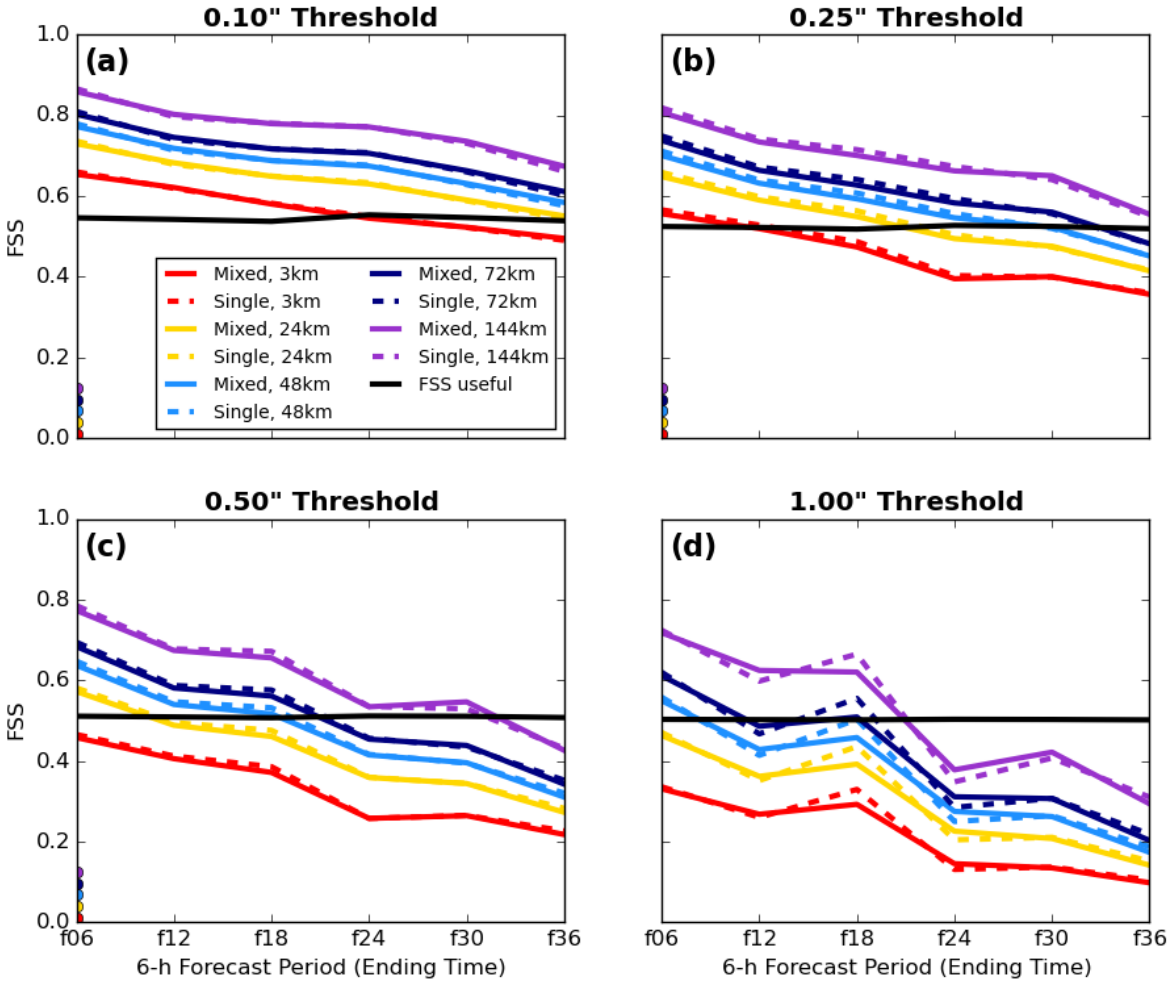

Figure 9 Fractions skill score as a function of forecast period for mixed- (solid) and single-physics (dashed) ensemble 6-hour precipitation forecasts at (a) 0.10-, (b) 0.25-, (c) 0.50-, and (d) 1.00-inch thresholds. In each case, 3- (red), 24- (gold), 48- (light blue), 72- (dark blue), and 144-km (purple)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

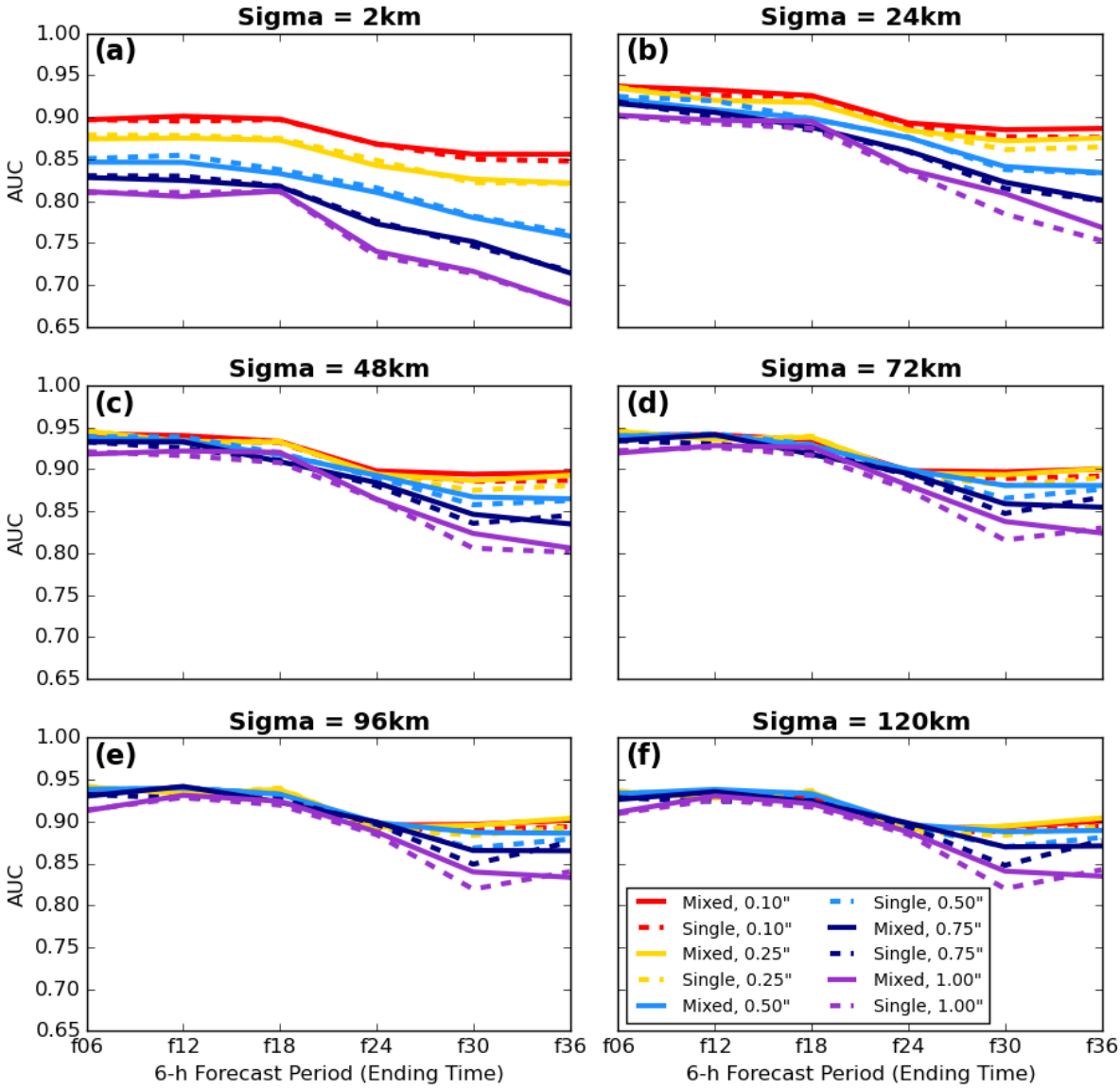

Figure 10 Area under the relative operating characteristics curve (AUC) for mixed- (solid) and single-physics (dashed) 0.10- (red), 0.25- (gold), 0.50- (light blue), 0.75- (dark blue), and 1.00-inch (purple) 6-hour accumulated precipitation threshold forecasts using a spatial smoothing parameter of (a)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

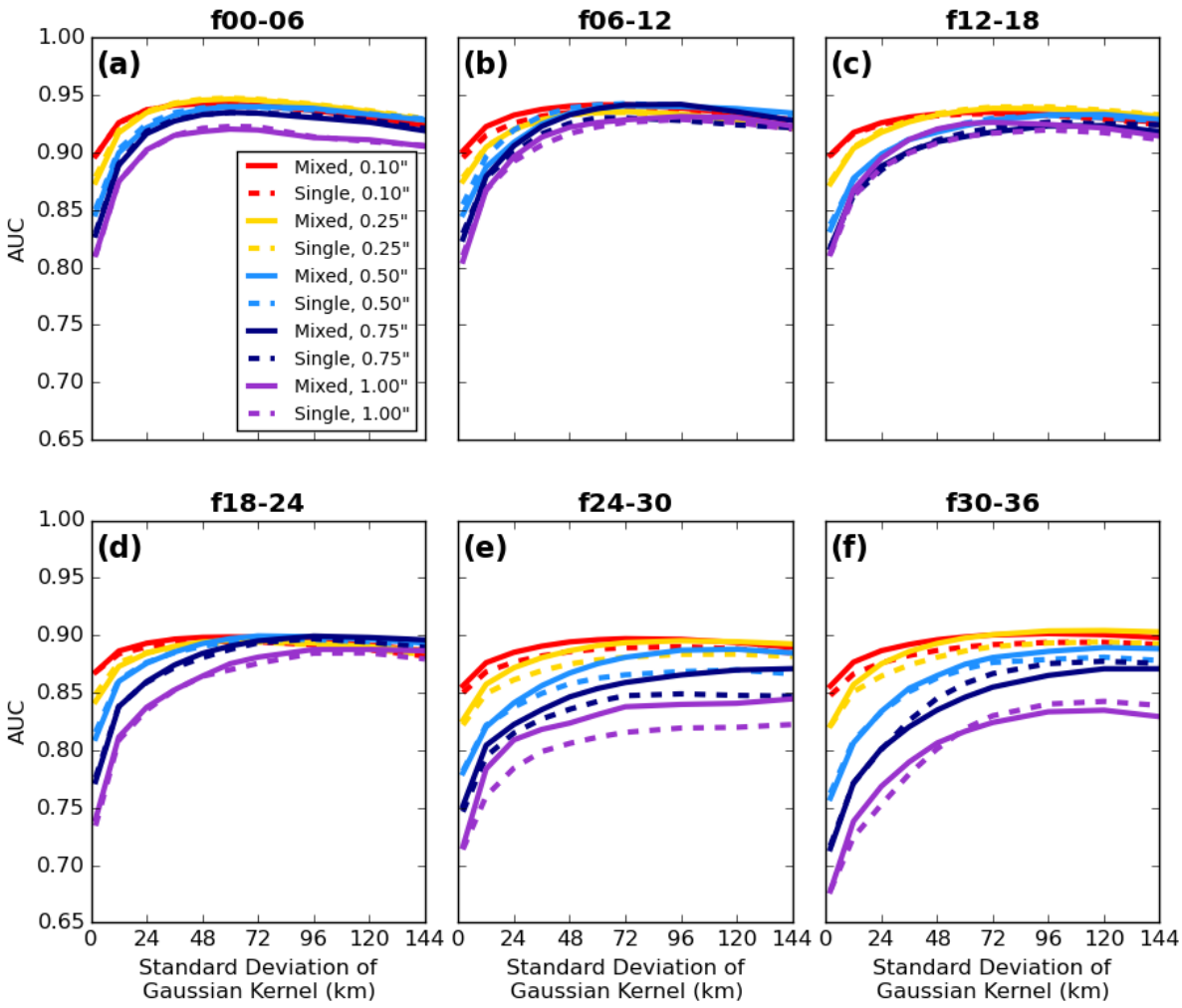

Figure 11 Area under the relative operating characteristics curve (AUC) for mixed- (solid) and single-physics (dashed) 0.10- (red), 0.25- (gold), 0.50- (light blue), 0.75- (dark blue), and 1.00-inch (purple) 6-hour accumulated precipitation forecasts as a function of the spatial smoothing parameter for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

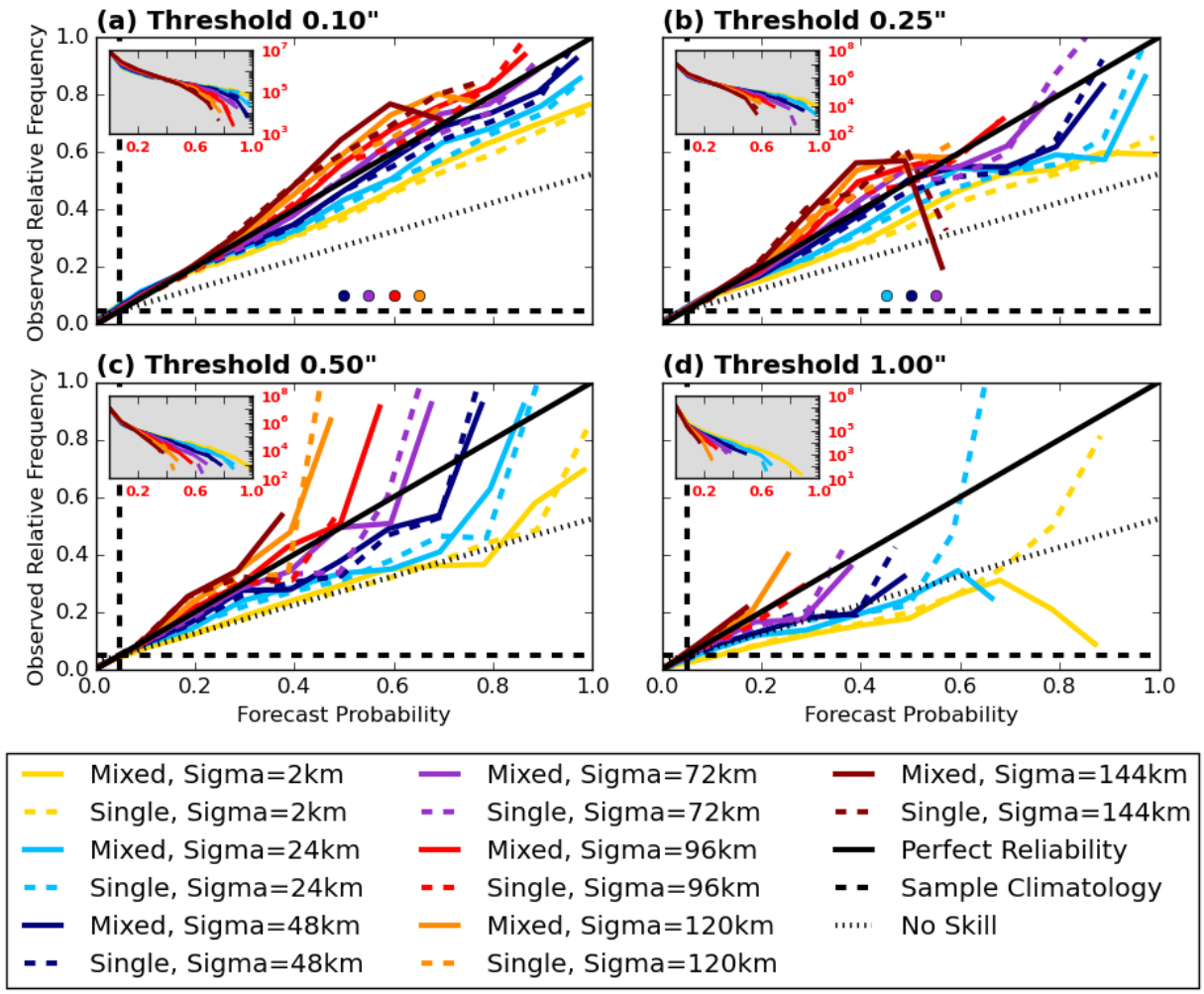

Figure 12 Attributes diagrams for mixed- (solid) and single-physics (dashed) ensemble 6-hour precipitation forecasts ending at forecast hour 30 using a threshold of (a) 0.10-, (b) 0.25-, (c) 0.50-, and (d) 1.00-inch. In each case, forecasts produced using a spatial smoothing parameter of 2- (gold), 24- (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. Smaller plots 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.

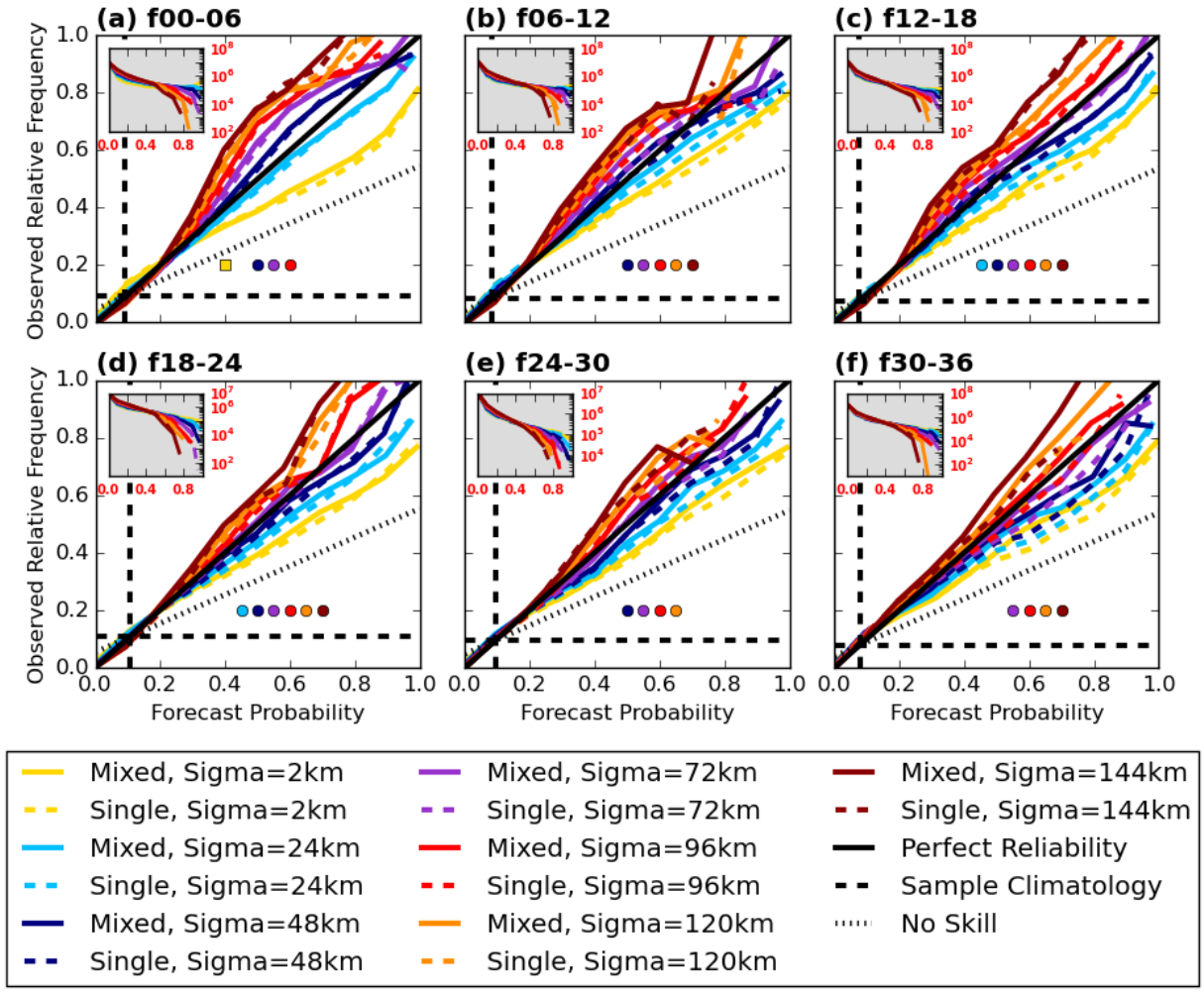

Figure 13 Attributes diagrams for mixed- (solid) and single-physics (dashed) ensemble 0.10-inch threshold 6-hour accumulated precipitation forecasts for the 6-hour forecast period spanning forecast hours (a) 0-6, (b) 6-12, (c) 12-18, (d) 18-24, (e) 24-30, and (f) 30-36. In each case, forecasts produced 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. Smaller plots 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.

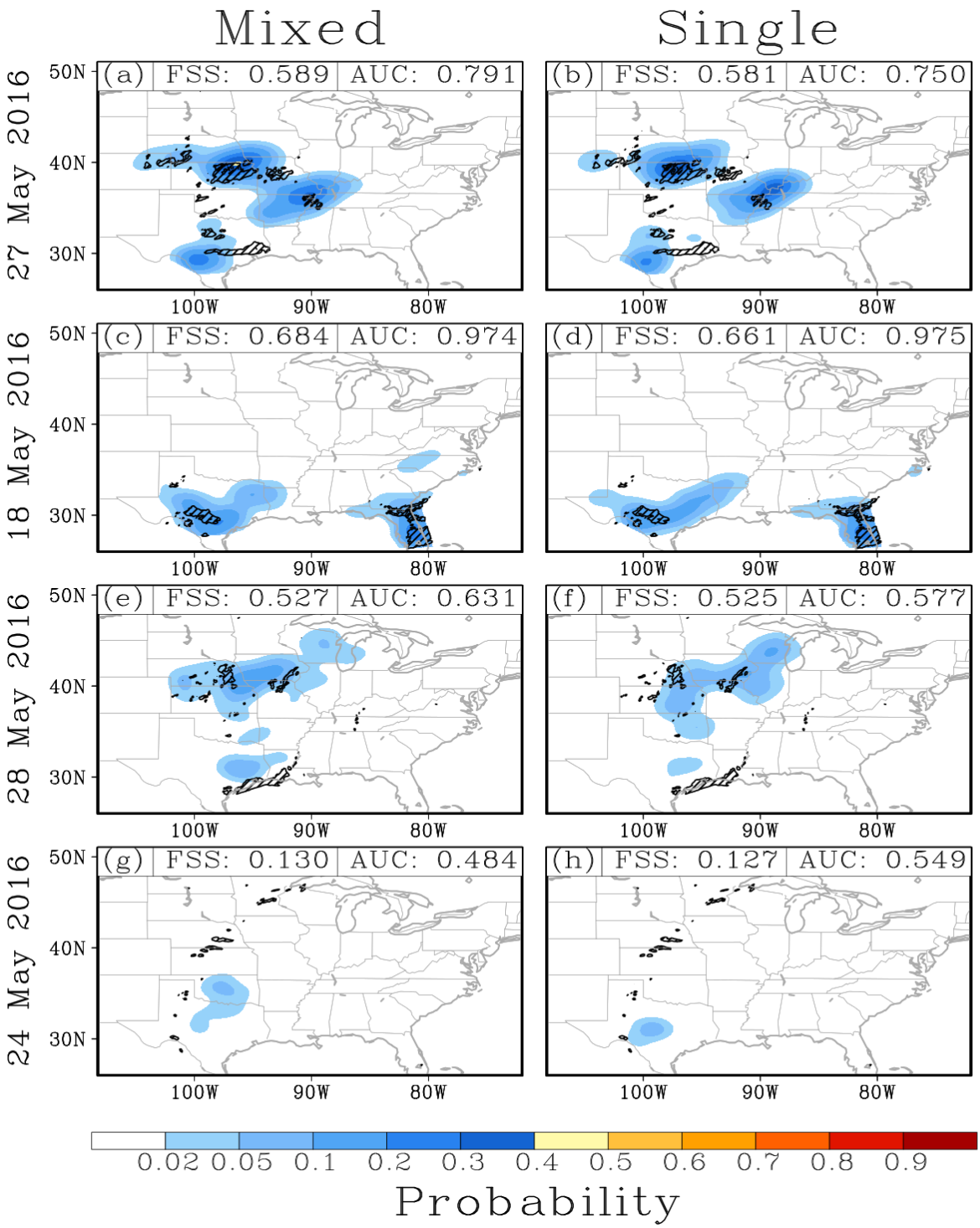

Figure 14 30-hour probabilistic 1.00-inch precipitation forecast (shaded) from (a) the mixed-physics ensemble and (b) the single-physics ensemble. Forecasts are valid for 0000-0600 UTC on 27 May 2016. Black hatching denotes 3-km points containing observed ≥ 1.00-inch precipitation over the 6-hour period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

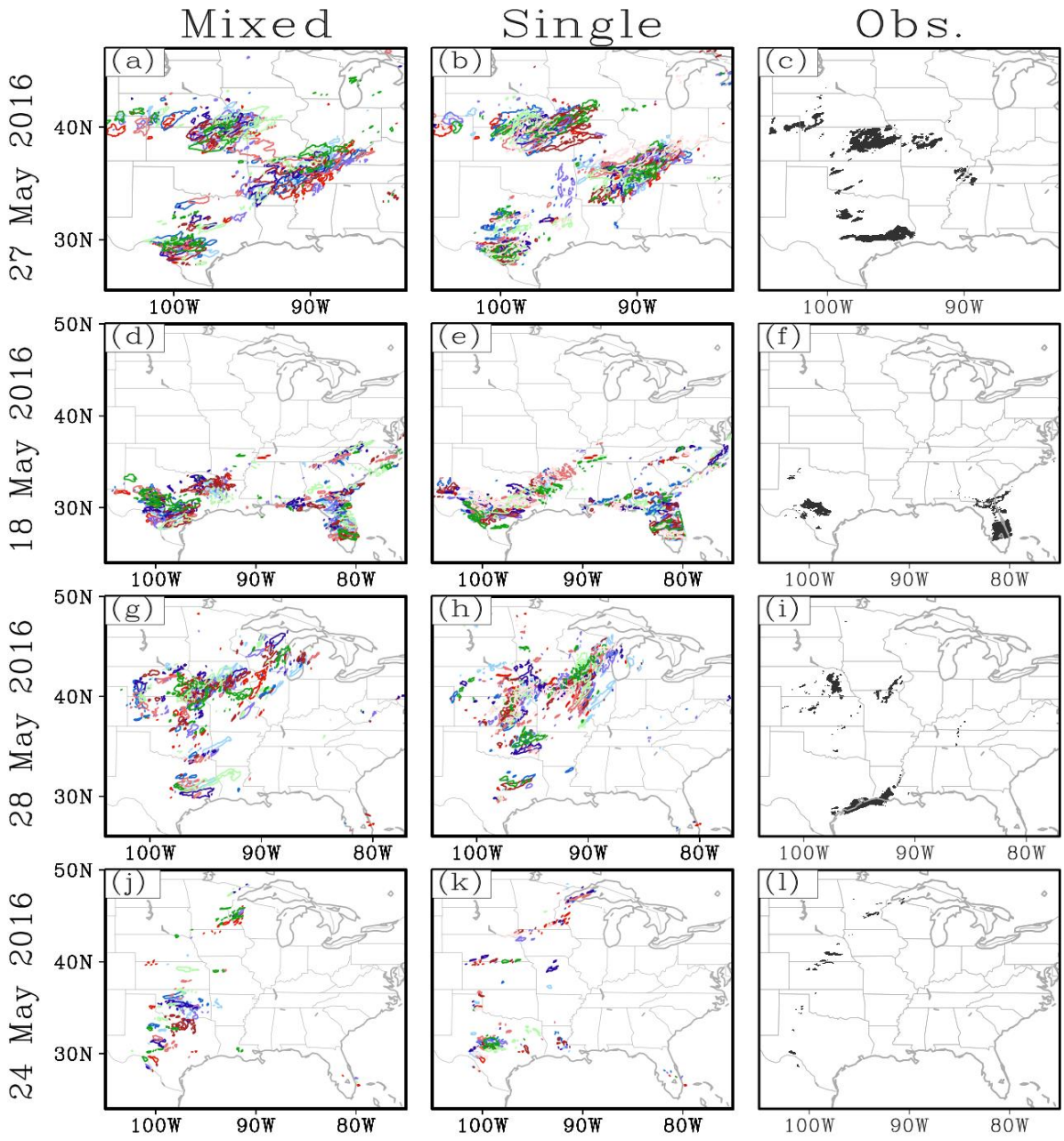

Figure 15 Individual ensemble member 30-hour 1.00-inch precipitation forecasts from (a) the mixed-physics ensemble and (b) the single-physics ensemble, valid for 0000-0600 UTC on 27 May 2016. (c) Observed precipitation ≥ 1.00-inch., valid for the same 6-hour period as in (a) and (b). (d)-(f) As in (a)-(c) 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.