Impact of a Stochastic Kinetic Energy Backscatter Scheme on Warm season Convection-Allowing Ensemble Forecasts

Jeffrey D. Duda
School of Meteorology and Center for Analysis and Prediction of Storms, University of Oklahoma, Norman, Oklahoma
Xuguang Wang
School of Meteorology, University of Oklahoma, Norman, Oklahoma
Fanyou Kong
Center for Analysis and Prediction of storms, University of Oklahoma, Norman, Oklahoma
Ming Xue
School of Meteorology and Center for Analysis and Prediction of Storms, University of Oklahoma, Norman, Oklahoma
Judith Berner
National Center for Atmospheric Research, Boulder, Colorado

Submitted to Monthly Weather Review
13 March 2015
Revised 2 June 2015

Corresponding author: Jeffrey D. Duda. CAPS, the University of Oklahoma, 120 David Boren Blvd., Room 2500, Norman, OK 73072-7309. Email: jeffduda319@gmail.com

Abstract

The efficacy of a stochastic kinetic energy backscatter (SKEB) scheme to improve convection-allowing probabilistic forecasts was studied. While SKEB has been explored for coarse, convection parameterizing models, studies of SKEB for convective scales are limited. Three ensembles were compared. The SKMP ensemble used mixed-physics with the SKEB scheme, whereas the MP ensemble was configured identically but without using the SKEB scheme. The SK ensemble used the SKEB scheme with no physics diversity. The experiment covered May 2013 over the central United States on a 4 km Weather Research and Forecasting model domain.

The SKEB scheme was successful in increasing the spread in all fields verified, especially mid- and upper-tropospheric fields. Additionally, the rmse of the ensemble mean was maintained or reduced, in some cases significantly. Rank histograms in the SKMP ensemble were flatter than those in the MP ensemble, indicating the SKEB scheme produces a less under-dispersive forecast distribution. Some improvement was seen in probabilistic precipitation forecasts, particularly when examining Brier scores. Verification against surface observations agree with verification against Rapid Refresh (RAP) model analyses, showing that probabilistic forecasts for 2-m temperature, 2-m dewpoint, and 10-m winds were also improved using the SKEB scheme. The SK ensemble gave competitive forecasts for some fields. The SK ensemble had reduced spread compared to the MP ensemble at the surface due to the lack of physics diversity.

These results suggest the potential utility of mixed physics plus the SKEB scheme in the design of convection-allowing ensemble forecasts.

1. Introduction

The purpose of ensemble forecasting is to address uncertainty by accounting for errors in ensemble forecast systems. Sources of error in numerical weather prediction (NWP) forecasts include initial condition error from observational error in measurement and inadequate spatiotemporal sampling of the atmospheric state and model error from physical parameterizations and the numerical schemes with associated random error from inadequately resolving subgrid-scale processes. For limited area models lateral boundary conditions also contribute to forecast error. For large-scale or global ensembles, much research exists concerning the best design practices (see, for example, Toth and Kalnay 1993; Molteni et al. 1996; Candille 2009; Wang and Bishop 2003; Wang et al. 2004). However, there is less research and understanding concerning the design of convection-allowing ensembles, although the body of work is growing (e.g., Kong et al. 2006, 2007a; Kong et al. 2007b; Mittermaier 2007; Schwartz et al. 2010; Vie et al. 2011; Xue et al. 2011; Johnson et al. 2011a,b; Bouttier et al 2012; Johnson and Wang 2012, 2013; Johnson et al. 2013, 2014; Ropnack et al. 2013; Caron 2013; Duda et al. 2014; Romine et al. 2014). Current methods used to represent model error in a convection-allowing ensemble include multi-parameter, wherein fixed parameters within a physics parameterization scheme are varied (Hacker et al. 2011; Yussouf and Stensrud 2012; Duda et al. 2014), mixed physics, where separate microphysics and boundary layer schemes are used (Johnson et al. 2011ab; Xue et al. 2011 and references therein; Duda et al. 2014), and multi-model, where separate dynamic cores are used (Ebert 2001; Wandishin et al. 2001; Kong et al. 2009; Candille 2009; Johnson and Wang 2012; and Du et al. 2014). The Storm Prediction Center also uses a multi-model ensemble in which each member is a convection-allowing forecast provided by either the National Centers for Environmental Prediction or the National Severe Storms Laboratory (see www.spc.noaa.gov/exper/sseo). This paper focuses on a different method of representing model error in an ensemble – stochastic perturbations using a stochastic kinetic energy backscatter scheme.

Research into the use of stochastic perturbations in ensemble forecasting is motivated by the results from prior research showing the benefit of using random perturbations. Buizza et al. (1999), for example, showed that merely including multiplicative random perturbations to the physical tendencies using simple spatiotemporal correlations was sufficient to increase ensemble spread and improve probabilistic precipitation forecasts. This method was based on the notion that the physical parameterizations handle subgrid-scale processes which are inherently random. The parameterizations take large-scale flow as input and thus are considered an ensemble average impact from subgrid-scale processes. The random perturbations therefore account for the variability in the subgrid-scale processes. Mason and Thomson (1992), on the other hand, used a SKEB scheme in a large eddy simulation to improve near-surface flow. Similarly, Shutts (2005) developed a cellular automaton stochastic backscatter scheme (CASBS) which inserted random perturbations into the model, but structured differently than the scheme in Buizza et al. (1999), and based on a different justification. The purpose of CASBS was to include a subgrid-scale process missing from global NWP models. The scheme injected kinetic energy (KE) into the model domain to counteract excessive energy dissipation coming from numerical diffusion and interpolation, mountain and gravity wave drag, and deep convection. Not only did CASBS correct the KE spectrum of the European Centre for Medium-Range Weather Forecasting (ECMWF) model, it also improved the spread and skill of 500 hPa geopotential height forecasts. Without CASBS the model failed to correctly simulate mesoscale circulations conforming to the observed $k^{-5/3}$ power law (Nastrom and Gage 1985). Berner et al. (2009) built off the work of

Shutts (2005) and developed a spectral stochastic kinetic energy backscatter scheme (SSBS) which was implemented in the ECMWF ensemble prediction system (EPS) in 2011 (ECMWF, 2012). The SSBS scheme was later modified for use in the Weather Research and Forecasting (WRF) model by Berner et al. (2011), who determined that the SKEB scheme gave superior ensemble mean forecasts of many fields compared to an ensemble using only physics variations. Their work was performed using a horizontal grid spacing of 45 km.

The use of SKEB schemes in operational EPSs has increased recently. Similar versions of the SSBS scheme have been introduced into the Canadian EPS (Charron et al. 2010), the Met Office Global and Regional EPS (Tennant et al. 2011), and the United States Air Force Weather Agency mesoscale ensemble (Hacker et al. 2011). The impact of SKEB has been overwhelmingly beneficial, including increased spread with a maintained or reduced root-mean square error (rmse) of the ensemble mean, and improved probabilistic forecasts of upper-level winds, temperatures, heights, and precipitation.

Prior research into the effectiveness of a SKEB scheme on probabilistic forecasts has been limited to global or otherwise coarse-grid scale EPSs. It remains to be determined how useful or valid such a scheme is for a convective-scale EPS. Such study is still limited. For example, Romine et al. (2014) compared ensemble forecasts at 3 km grid spacing using two stochastic perturbation methods, the SKEB scheme and the stochastically perturbed parametrization tendencies scheme (Buizza et al. 1999; Palmer et al. 2009). They found the SKEB scheme to provide a balance between increased ensemble spread and forecast bias change. The focus of the current study is to compare ensemble forecasts using the SKEB scheme to a mixed-physics ensemble which is typically used in convection-allowing ensemble forecast system design. The following questions are investigated: is the stochastic error representation method such as SKEB compatible with a mixed-physics approach in a convective-scale forecast? What is the impact of including SKEB perturbations on top of the typically used mixed-physics method in convection-allowing ensemble forecasts? A 4-km WRF ensemble including a portion of the United States and featuring warm season cases is adopted to achieve this goal.

The rest of this paper is organized as follows. The SKEB scheme is described briefly in section 2. The experiment design is described in section 3. Results follow in section 4. A summary and conclusions follows in section 5.

2. The SKEB scheme

a. Mathematical formulation

Stochastic parameterizations are developed to represent model-error and can generate ensemble spread by perturbing the forecast trajectory. For this purpose, a SKEB scheme is employed that adds stochastic, small-amplitude perturbations to the rotational component of horizontal wind and potential temperature tendency equations at each time step. The scheme is briefly outlined here; readers can refer to Berner et al. (2011) for details.

Let ψ(x,y,t) be a 2D-streamfunction-forcing pattern expressed in 2D-Fourier space:

$$\psi\ \ (x,y,t) = \sum_{l=-L/2}^{L/2} \sum_{k=-K/2}^{K/2} \psi_{k,l}(t) e^{2\pi i\left(\frac{kx}{X}+\frac{ly}{Y}\right)} . \quad (1)$$

Here, $\psi_{k,l}$ denotes the spectral coefficient of the perturbation field with *k* and *l* the (K-1) and (L-1) wavenumber components in the zonal *x*- and meridional *y*-direction in physical space and *t* denotes time. The representation in spectral space allows for control of the spatial length scales of the perturbations as a function of wavenumber. Subsequently, the rotational wind components are obtained by differentiation, the pattern transformed back to gridpoint space, and the perturbations added to the momentum tendency equations. The perturbations to the potential

temperature tendencies are generated analogously. The WRF implementation allows for a 2D or 3D perturbation pattern to be generated. Here, we follow Berner et al. (2011) and use the same 2D pattern to perturb the tendencies in all vertical levels.

To introduce spatial and temporal correlations, each spectral coefficient $\psi_{k,l}$ is evolved as a first-order autoregressive process:

$$\psi_{k,l}(t + \Delta t) = \alpha\psi_{k,l}(t) + \sqrt{\alpha - 1}g_{k,l}\varepsilon(t), \quad (2)$$

where α is the linear autoregressive parameter determining the temporal decorrelation time, $g_{k,l}$ is the wavenumber-dependent noise amplitude, and ε a Gaussian white-noise process with mean 0.0 and variance 1.0. The noise amplitude $g_{k,l}$ determines the variance spectrum of the forcing and is given by the power law, $g_{k,l} = bn^p$, where n is the wavenumber and p is an assumed constant slope, and b is the amplitude defined as in eqn. (4) of Berner et al. (2009), chosen so that a fixed amount of KE is injected into the flow each time the forcing is applied. The autoregressive parameter determines the decorrelation time τ of the pattern, $\tau = \Delta t/(1 - \alpha)$, where Δt is the model time step. In principle, each spectral coefficient can be associated with a different decorrelation time, but for practical reasons, the same decorrelation time is chosen for all spectral coefficients. The SKEB scheme was originally motivated by the notion that upscale and downscale cascading energy resulted in net forcing for the resolved flow from unresolved scales (Shutts 2005) and the perturbation amplitude was proportional to the instantaneous dissipation rate. Here, the simplifications in Berner et al. (2011), where it is assumed the dissipation rate is constant in space and time, are used. Since the forcing is no longer state-dependent the perturbations can be considered as additive noise with prescribed spatial and temporal correlation.

b. Tuning the scheme

The autoregressive parameter α (related to the decorrelation time of the forcing field), the slope of the power spectrum for the perturbations, and the amplitude of the perturbations are adjustable parameters which can be tuned for a specific application. Sensitivity tests were conducted to determine if the default settings of Berner et al. (2011) – obtained for forecasts in simulations with a horizontal resolution of 45 km – are also optimal for the much higher horizontal resolution used here. The cases selected, 13 April 2012, 14 June 2012, and 23 June 2013 (none of which are part of the experiment period; section 3), featured warm season precipitation episodes exhibiting a wide variety of convective modes and magnitude of large-scale dynamic forcing, thus allowing for a general tuning of the scheme over a range of scenarios. This approach also avoids issues related to in-sample tuning to generalize the results.

Berner et al. (2009) informed the slope of the forcing spectrum, *p,* using coarse-graining cloud-resolving model output (Shutts and Palmer 2007). Rather than following this strategy and coarse-grain output from large-eddy simulations, the slope of the power spectrum herein is determined empirically. For this purpose the default value of the spectral slope was perturbed in the positive and negative directions by 20%, 40%, and 80%. The decorrelation time was perturbed similarly. Additionally, the sensitivity to the amplitude of the perturbations was tested by varying the backscatter dissipation rates for wind and temperature, respectively, by several orders of magnitude. We did not test every combination of parameter values. Instead we evaluated an arbitrary set that focused on changing only one parameter at a time, although we did test a few sets of coupled parameter perturbations. This method was not intended to be comprehensive, and it is possible that other combinations of values may yield yet better ensemble statistics and may represent a better tuning of the scheme.

Sensitivity tests using larger wind and temperature amplitudes for the perturbations generated drastically increased ensemble spread relative to the default values. However, as these perturbation amplitudes increased, the member forecasts looked increasingly different from verifying precipitation analyses (not shown) and hence were subjectively judged to be degraded. Therefore we chose the default values of the wind and temperature perturbation amplitudes. An example of the structure of the perturbation tendency fields for u-wind and temperature is shown in Fig. 1.

While ensemble performance was not obviously superior for any one set of decorrelation time and spectral slope values, examination of ensemble spread helped to indicate a tendency for a 40% reduction in decorrelation time coupled with a 40% increase in the spectral slope to produce the best agreement between spread and rmse of the ensemble mean for several synoptic scale fields such as temperature, geopotential height, wind, and moisture (Figs. 2 and 3). While improvement of probabilistic quantitative precipitation forecasts (PQPF) at the convective-scale is emphasized, precipitation forecasts were not strongly sensitive to the choice of decorrelation time or spectral slope (not shown). Therefore, we used 6480 s for the decorrelation time and 2.567 for the spectral slope for both wind and temperature perturbations. This combination did not always result in the best precipitation forecasts according to Brier score of 1-hr accumulated precipitation at various thresholds, but it did not always result in the poorest precipitation forecast either (not shown). Independently, Ha et al. (2015) performed limited sensitivity tests of SKEB scheme parameters in cycled simulations with WRF at a resolution of 15 km and determined that the default parameters performed well. This together with our findings points to the generality of the parameter settings in WRF, at least for application to forecasts in the mid-latitudes.

In this study, only one set of tunable parameter values of the SKEB scheme was used for all experiments involving the use of SKEB. In other words, the same values were used in the SKEB scheme for each combination of physics options used in the ensembles tested. It is likely that the optimal parameters for the SKEB experiment may not be the same as those actually used given the potential dependence of SKEB parameters to the choice of physics. Therefore the comparison between SKMP and MP may not maximize the value of adding SKEB on top of MP. However, the conclusion of added values of including SKEB on top of MP should not change qualitatively.

c. Impact on WRF KE spectra

The original motivation for using a SKEB scheme was to counteract excessive dissipation in the ECMWF model, in part caused by the use of a semi-Lagrangian time stepping scheme (Shutts, 2005). The ECMWF ensemble is a global ensemble and is typically not run at non-hydrostatic, convection-allowing resolutions. At lower horizontal resolution the KE spectrum of NWP models typically does not capture the observed transition from the k^{-3} spectrum characterizing the synoptic scale to the shallower $k^{-5/3}$ spectrum in the mesoscale (Nastrom and Gage, 1985). The ECMWF implementations of SSBS were able to capture this transition by simulating under-represented mesoscale variability (Shutts, 2005; Berner et al., 2009). It is an area of active research if the existence of a $k^{-5/3}$ spectrum in a NWP model is necessary for reliable ensemble predictions with forecast target times of few days, since it is associated with faster error growth (e.g., Lorenz, 1969; Rotunno and Snyder, 2007).

Following Skamarock (2004) we computed KE spectra from WRF simulations at horizontal resolutions of 1 km and 4 km (Fig. 4). At both resolutions the KE spectra are characterized by a $k^{-5/3}$ slope which drops off above wavenumbers of 0.02 m^{-1} and 0.001 m^{-1} for the 4 km and 1 km

resolutions, respectively. The slope in the simulations at 1 km continues through the meso-γ scale, which indicates that the drop in the tail of the spectrum in the 4 km simulation is due to numerical truncation error rather than a characteristic of the circulation pattern. For the tuning parameters chosen the simulations using the SKEB scheme do not show any appreciable difference in KE structure from simulations that do not use the SKEB scheme which confirms that the scheme does not introduce artificial energy. Arguably, it would have been desirable to tune the scheme so that the KE spectra at coarser resolution resemble more those at 1 km, but given that WRF has already a $k^{-5/3}$ spectrum at the chosen resolution, the benefit might be small.

3. Experiment setup

To compare two methods of accounting for model error, SKEB and mixed-physics, and to evaluate the impact of adding SKEB on top of the multiple physics approach, three ensembles were constructed. Two ensembles contained mixed physics (microphysics, boundary layer, and land surface model) that only differed in use of the SKEB scheme. The ensemble that did not use the SKEB scheme is hereafter called ensemble MP, whereas the ensemble that did is hereafter called ensemble SKMP. The MP ensemble resembles the common mixed-physics design of the storm scale ensemble forecast system produced by the Center for Analysis and Prediction of Storms at the University of Oklahoma for the SPC/NSSL HWT spring forecasting experiment (see, for example, Xue et al. 2011). Additionally, to test whether stochastic error representation alone is sufficient for use in an ensemble compared to mixed-physics error representation alone, a third ensemble, the SK ensemble, was constructed that contained no physics diversity; different random number seeds were used to supply diversity in the SKEB scheme. The configuration of each ensemble is shown in Table 1. Seven members comprised each ensemble. The choice of seven members represents a balance between computational resources, availability of various physical parameterization schemes, and adequate representation of the forecast probability distribution. This size is reasonably adequate to produce precipitation forecasts at a spatial scale of 50 km (section 4c) that are statistically indistinguishable from a larger ensemble that would better populate a probability distribution (Clark et al. 2011). Since the focus of the study is on representation of model error, neither initial condition nor lateral boundary condition perturbations were used.

The Advanced Research WRF (ARW) dynamics core (Skamarock et al. 2008) of the WRF model, version 3.4.1, was used to conduct the simulations. The model domain encompassed a large portion of the central and eastern U.S. (Fig. 5) where deep convection is climatologically favored during the late spring and early summer, the period during which the experiment was conducted. We tested 31 cases spanning May 2013. There were a number of active severe weather days in the model domain during this period, so the results of this study should be representative of the overall ability of the SKEB scheme in a convective-scale EPS for warm season events. Each case was initialized at 0000 UTC and integrated for 30 hours. The grid spacing for all members was 4-km with 20 s model time step. North American Mesoscale (NAM) model analyses valid at 0000 UTC were used as the initial conditions, and NAM model forecasts from 0000 UTC for each case were used as the lateral boundary conditions.

Verification was performed on a number of fields, including temperature, wind, height, and various moisture variables at multiple atmospheric pressure levels, as well as 1-hr accumulated precipitation. Rapid Refresh (RAP; Brown et al. 2011; Weygandt et al. 2011; rapidrefresh.noaa.gov) analyses were used as verifying data for upper-air fields. METAR and mesonet observations obtained from the Meteorological Assimilation Data Ingest System

(MADIS; https://madis.noaa.gov/) were used to verify 2-m temperature and dewpoint and 10-m winds. Only observations passing quality control checks were used for verification. A motivation behind verifying surface fields using observations rather than a RAP analyses is representativeness of surface values. The fine-scale detail in the WRF can be better verified using observations than by using a 13-km grid spacing model analysis (RAP; which does not resolve features smaller than 26-km in wavelength). Even though observing stations are spread farther apart than the RAP model grid points, they are still capable of capturing small scale features such as individual thunderstorms at certain locations. The signal from a single sample would certainly be wiped out when performing a gridded analysis. However, the WRF model can capture features as small as 8 km in size. For this reason, it was not required for observations to pass spatial continuity quality control checks to be included in the verification. This freedom increases the chances of the model being rewarded for forecasting a minimally-resolved thunderstorm and associated cold pool in the right place at the right time such that the corresponding observation also sampled the storm. Verifying precipitation data were provided by the National Mosaic and Multi-Sensor Quantitative Precipitation Estimation (QPE) project (NMQ) produced by the National Severe Storms Laboratory (Zhang et al. 2011). The QPEs are the result of a combination of radar-estimated rainfall and rain gauge measurements. The NMQ QPEs (native grid at 0.01° resolution) were regridded to the verification domain using bilinear interpolation. Verifications were performed using one-hour accumulated precipitation. The NMQ QPEs have been used in earlier studies to verify storm-scale precipitation forecasts (e.g., Johnson and Wang 2012; Johnson et al. 2013, Duda et al. 2014). PQPFs were constructed without calibration or bias correction as the number of members in which a threshold value was exceeded.

4. Results

a. Skill of physics packages

The WRF-ARW offers a large set of different physics packages (mainly microphysics and PBL), and many of those schemes were used in the MP and SKMP ensembles. The skill of individual packages was examined first to facilitate the comparison between SK with MP or SKMP.

The rmse of the MP ensemble members for a large number of fields is shown in Fig. 6. Especially for upper tropospheric winds and heights (Figs. 6a-d), the scores were clustered tightly; only member m5 stood out as a poorer physics package for these fields. For lower tropospheric fields (Figs. 6e-l), there was more diversity in the rmse among the members. After about forecast hour 12 or so, two members, m1 and m3, which both used the YSU PBL scheme (Hong et al. 2006) and Noah land surface model (Ek et al. 2003), tended to perform better than the other members. This was especially apparent in near-surface fields such as wind, temperature, and dewpoint (Figs. 6j-l). In this regime (warm season over mid-latitude plains) the YSU scheme appears to be a better PBL scheme as verified here. The difference between members m1 and m3 was the choice of microphysics scheme; member m1, on which the SK ensemble is based, used the Morrison microphysics scheme (Morrison et al. 2009), noted as one of the better multi-moment microphysics schemes in Duda et al. (2014), whereas member m3 used the simpler WRF single moment 6-class microphysics scheme (Hong and Lim 2006). These members were not as skillful for 1-hr precipitation at forecast hours 6-21, and member m3 was less skillful than member m1 after forecast hour 14. However, member m1 was somewhat less biased than most other members in some fields, although it did not always have the smallest bias

(not shown). Members m5 and m6 were frequently among those with the highest rmse, especially after forecast hour 18 (Figs. 6e-l). Finally, there was a notable clustering by land-surface model in the 2-m temperature and dewpoint fields (Figs. 6k,l). Such stark contrast in rmse suggests the forecast distributions for these fields was likely bimodal in some cases. Future work will investigate ways to better account for land-surface model uncertainties to better populate the forecast distribution and to determine if such a bimodal distribution reflects the underlying truth error distribution or is an artifact that some LSMs tend to behave similarly than others.

Overall, the bias and rmse characteristics for the individual members indicates that the SK ensemble was based on a relatively skillful set of physics parameterizations and thus could be expected to provide forecasts competitive with those from the MP and SKMP ensembles, which use mixed physics.

b. Ensemble spread-error agreement and dispersion

1) Against RAP analyses

(i) Spread, rmse, and rank histograms

We first examine the spread-error relationship of the ensembles for several fields. The ensemble spread, averaged over the 31 cases and the verification domain, is shown for several fields in Fig. 7. The addition of the SKEB scheme added a large amount of spread to the upper-tropospheric fields in the SKMP ensemble. For fields such as hgt500 and u500__ (see Table 2 for field abbreviations), the amount of added spread exceeded 100% at forecast hour 36. Fig. 4b shows that, for the u-wind component, spread was added at nearly all but the finest scales, with relatively more diversity added at the largest scales where the perturbation amplitudes were also the largest. The spread difference between the SKMP and MP ensembles increased steadily with time. There was also increased spread in the SKMP ensemble at fields in the lower troposphere, although not as much as was added above. There was even an increase in spread in moisture fields (pwat__ and accppt) despite those fields not being directly perturbed. The spread in the SKMP ensemble was higher than that in the MP ensemble at all forecast hours after forecast hour 1, and the difference in spread between the MP and SKMP ensembles generally increased with time throughout the 30-hour forecast, but especially after forecast hour 6 or so, after which time model spin-up was complete. Additionally, the spread in the SK ensemble was also much larger than that in the MP ensemble for upper-tropospheric fields. At middle and lower tropospheric levels, the SK ensemble had lower spread than the MP ensemble until about forecast hour 5, after which the order was reversed.

Given the general under-dispersive nature of many ensembles (e.g., Duda et al. 2014), increased spread is an attractive result. However, increased spread does not necessarily mean the forecast error was sampled more appropriately. A large spread can result from incorrectly sampling forecast errors, which could lead to degradation of the ensemble mean forecast or member forecasts that are extremely different from one another and therefore lead to a degraded probabilistic forecast. The forecast quality is first examined via the rmse of the ensemble mean, also shown in Fig. 7. For most fields there was not a numerically large difference in the rmse among the three ensembles. However, for several fields, the rmse of the SKMP ensemble was lower than that of the MP ensemble at a large number of forecast hours (e.g., tmp500, v850__, sph850, pwat__, tmp850, and accppt), and that difference is statistically significant using a t-test at $\alpha = 0.05$. The fields showing the biggest decrease in rmse were concentrated in the lower troposphere, suggesting the SKEB scheme is quite effective in perturbing fields at lower levels

when also coupled with physics perturbations, an observation also made by Hacker et al. (2011) and Berner et al. (2011) for their studies at 45 km grid spacing. Since the magnitude of the wind and temperature tendencies for SKEB perturbations is not dependent on height, the relative magnitude of the perturbations is larger in the lower troposphere, which may have played a role in the greater improvement there. The decreased rmse in moisture fields such as pwat__ and sph850 is particularly interesting since they are not directly perturbed by the SKEB scheme. This decrease in rmse could come from improvement in precipitation and thunderstorm processes, which are directly impacted through the lower-tropospheric wind and temperature perturbations. There are a few fields for which the rmse of the ensemble mean of the SKMP ensemble was higher compared to the MP ensemble. However, the degradations were limited to winds and heights at 500 hPa and above. The increase in rmse of the hgt500 field in particular could be related to changes in the number of thunderstorms present in the forecasts caused by lower tropospheric wind and temperature perturbations. An individual thunderstorm can strongly perturb the height field through non-hydrostatic vertical accelerations.

The rmse of the ensemble mean in the SK ensemble was commonly higher than that in the MP ensemble at early lead times in most fields (black dots across the top of each panel in Fig. 7). However, in the middle of the forecast period, and for lower tropospheric fields like tmp850 and v850__, the SK ensemble had a lower rmse than the MP ensemble. In the tmp850 and sph850 fields, the SK ensemble had a lower rmse than MP ensemble generally after forecast hour 12. The SK ensemble also had a lower rmse than the MP ensemble for pwat__ between forecast hours 16 and 25. Also in the tmp850 field the SK ensemble had a lower rmse than even the SKMP ensemble for forecast hours 15-29 and in the v850__ field for forecast hours 15-22.

The increased dispersion in large-scale fields is further supported through examination of rank histograms. For nearly every large-scale field examined, the rank histograms at nearly every forecast hour were flatter in the SKMP ensemble than in the MP ensemble (Fig. 8). It should be noted that observation error was not incorporated into any verifications in this study. Error in the observation data sources is either unknown or undocumented. Therefore it is inappropriate to make claims regarding proper ensemble dispersion. Instead we can only discuss the differences in dispersion among the ensembles. It should also be noted, however, that given the broad similarities between this study and that of Berner et al. (2015), where the order of performance of various methods of representing model error was not a function of inclusion of observation error, it is not expected that the increased spread, flatter rank histograms, and lower rmse of the SKMP ensemble over the MP ensemble is conditional on inclusion of observation error.

The rank histograms for the SK ensemble were also generally flatter than those of the MP ensemble after the first few forecast hours, but not as flat as those of the SKMP ensemble, again suggesting that this method of accounting for stochastic error is at least as effective as a mixed-physics approach after the added perturbations have had time to accumulate and create diversity among the members. For accppt the rank histograms were not noticeably flatter in the SKMP ensemble than the MP ensemble, although the positive bias in that field makes dispersion characteristics less pertinent.

(ii) Case study

The increased ensemble spread and member diversity can be illustrated via some atmospheric fields from a representative case. We chose the case initialized at 0000 UTC 19 May 2013 as it contained a severe weather event associated with a mesoscale feature (a dryline) forced by a synoptic-scale upper-level trough. The impacts of the perturbations on scales ranging from

synoptic to storm scale can be seen. First we examine the 500 hPa height field (Fig. 9). Even after the model spun up convection across portions of western Kansas and Oklahoma in the early forecast hours (not shown), the 5760 m height contours in the MP ensemble at later forecast hours show little diversity in areas near and upstream of that convection, which had propagated into eastern Kansas at the valid time in Fig. 9. Compared to the same height contours in the SK and SKMP ensembles, it is clear that the perturbations in the SKEB scheme have generated some synoptic-scale diversity. The SK and SKMP ensembles have larger area-averaged ensemble standard deviations (upper right of each panel in Fig. 6) than the MP ensemble. While there are still displacement errors (bias) relative to the RAP analysis of that contour in all ensembles, there are members in the SKMP ensemble for which the contour is more accurately placed due to the increased diversity in the ensemble.

The impacts of increased diversity on mesoscale aspects of the forecast are illustrated using precipitable water in Fig 10. The 25 mm contour delineates the dryline extending generally southwestward from central Oklahoma across central Texas. In the late morning to mid-afternoon before convective initiation, the dryline surged eastward before settling at its most eastward location late in the afternoon (not shown). During one particular afternoon hour (1700 UTC, Fig. 10), each ensemble placed the dryline too far east in northern Texas and across Oklahoma. However, there was additional diversity in the SKMP ensemble compared to the MP ensemble in which one or two SKMP ensemble members contained a more westward dryline than in the MP ensemble, closer to the RAP analyzed dryline location. The MP ensemble was more biased and overconfident on the location of the dryline in eastern Oklahoma, whereas the SKMP ensemble gave a more reasonable uncertainty estimate, having members that varied more on the longitudinal placement of the dryline.

Water vapor mixing ratio at 2 m illustrates increased spread at the convective scale (Fig. 11). A contribution from the mesoscale variability in the location of the dryline combined with a contribution of diversity on the convective scale resulted in larger ensemble spread in the SKMP ensemble compared to that in the MP ensemble along the dryline in Oklahoma and Texas. There was a small region of enhanced ensemble spread in northwest Oklahoma near a bulge in the dryline. The valid time of the data in Fig. 11 is one or two hours before convection first developed near that dryline bulge. The pattern of 1-hr accumulated precipitation from each member in the SKMP and MP ensembles (not shown) suggests there is more variability in the location, coverage, and intensity of the storms that developed near this dryline bulge in the SKMP ensemble compared to the MP ensemble. This difference in variability of precipitation is likely related to the variability in 2-m mixing ratio through changes in buoyancy of surface-based parcels. While there is also mesoscale variability in the placement of the dryline and the bulge in the SK ensemble, there is generally very little spread in 2-m mixing ratio near the dryline bulge compared to that in the SKMP and MP ensembles. The 1-hr accumulated precipitation fields in the SK ensemble also appear very similar among members, thus corroborating the lower diversity compared to the SKMP and MP ensembles (see section 4b2 for a discussion on the lack of diversity in the SK ensemble).

2) Against observations from MADIS

Verifications of surface variables against METAR and mesonet observations (from MADIS) are shown in Figs. 12 and 13. Similar to other fields verified against RAP analyses, the spread was larger (a few percent to as high as 25%) in the SKMP ensemble than the MP ensemble for all surface fields by forecast hour 30. The increase in spread was accompanied by almost no

change in the rmse of the ensemble mean for 2-m temperature and dewpoint. For 10-m winds the rmse of the ensemble mean in the SKMP ensemble was lower than that in the MP ensemble for all forecast hours, and the difference is statistically significant generally after forecast hour 6. Rank histograms (not shown) for 2-m temperature and dewpoint and 10-m wind components were slightly flatter in the SKMP ensemble than the MP ensemble, but the difference was not as remarkable as for fields verified against RAP analyses. Brier scores for exceedance of certain values of dewpoint and wind speed (Fig. 13) also suggest the SKMP ensemble produced better probabilistic forecasts at nearly every forecast hour and threshold compared to the MP ensemble. In particular, the SKMP ensemble had lower Brier scores than the MP ensemble at high dewpoint thresholds, especially between forecast hours 18-24, which correspond to the mid-late afternoon, a time when convective available potential energy is likely to be at its daily maximum and convective inhibition is likely to be at its daily minimum. Therefore, the SKMP ensemble may provide an improved forecast of the thermodynamic environment in a large-scale environment overall supportive of convective storms.

The SK ensemble spread was much lower than the MP ensemble spread, especially at earlier forecast hours, and especially for 2-m temperature and dewpoint (Fig. 12). The ensemble spread difference in the 10-m wind component fields was smaller than in the 2-m temperature and dewpoint fields. The lack of physics diversity in the SK ensemble likely negatively impacted the ensemble forecasts in these fields. This was also seen in Berner et al. (2011). The rmse of the ensemble mean was significantly lower in the SK ensemble compared to the MP and SKMP ensembles for 2-m temperature and dewpoint generally between forecast hours 12 and 24 and significantly higher elsewhere. The SK ensemble had a significantly lower rmse than the MP and SKMP ensembles for 10-m wind components throughout the forecast. The SK ensemble had lower Brier scores than both of the SKMP and MP ensembles for light to moderate 10-m wind speeds and also for all but the highest 2-m dewpoint thresholds during the mid-late afternoon (Fig. 13). The physics package used for the SK ensemble is chiefly responsible for the superior near-surface wind, temperature, and dewpoint forecasts (section 4a). This result suggests the potential of simplifying the ensemble design in the future by selecting the best physics parameterization scheme and using a stochastic method to sample stochastic model errors which relieve the need of maintaining multiple physics schemes that are not necessarily independent from each other.

c. PQPF skill

The spread of 1-hr accumulated precipitation in the SKMP ensemble was greater than that in the MP ensemble after the first few forecast hours (Fig. 7). Additionally, the rmse of the ensemble mean of the SKMP ensemble was lower than that of the MP ensemble at most forecast hours, with the SKMP ensemble being significantly better from forecast hours 7-24. Each ensemble had a positive bias of around 0.01-0.05 mm during the second half of the forecast period, and the SKMP ensemble was slightly more biased than the MP ensemble (not shown). The SK ensemble had less spread than the MP ensemble from forecast hours 1-12 and 21-30. The rmse of the ensemble mean was similar between the two ensembles during these periods. In the period of forecast hour 13-19, however, the SK ensemble had increased spread and decreased rmse compared to the MP ensemble. However, this time period corresponds to the morning and midday time when precipitation is less common, so the inference that the SK ensemble provides better PQPFs during this period may not be robust. The SK ensemble was less positively biased at forecast hours 20-25.

In spite of the bias, several grid-point and neighborhood-based probabilistic and traditional verification measures suggest PQPFs were improved by the use of the SKEB scheme in the SKMP ensemble. Considering Brier scores (Fig. 14), the SKMP ensemble had a lower Brier score than the SKMP ensemble at nearly all forecast hours for light and moderate accumulation thresholds (0.254 and 6.35 mm). With a lower climatological occurrence of precipitation exceeding 25.4 mm per hour, there was less difference in the Brier score between the SKMP and MP ensembles, and due to a relatively larger standard error, the differences were less likely to be significant. The SK ensemble also had lower Brier scores than the MP ensemble during the middle portion of the forecast period. The duration in which the SK ensemble had lower Brier scores generally decreased with increasing threshold except for at the 25.4-mm threshold, where the SK ensemble had significant lower Brier scores than the MP ensemble over nearly the entire forecast range.

The fractions skill score (FSS; Roberts and Lean 2008) is a neighborhood-based verification metric that measures the squared difference in the fraction of coverage of precipitation exceeding a threshold in a neighborhood about a grid point for both the forecast and observations. It is essentially an extension of the Brier score to spatial neighborhoods. A neighborhood radius of 48 km (12 grid points as in Johnson and Wang 2012 and Duda et al. 2014; Romine et al. 2014 used a neighborhood radius of 50 km for similar precipitation verification) was used for all neighborhood based scores presented. There was very little distinction in FSSs between the SKMP and MP ensembles at the lightest threshold (Fig. 15). The SKMP ensemble had slightly higher FSSs during the middle part of the forecast (forecast hours 9-24) for moderate and heavy rain thresholds, but the FSSs of the SKMP ensemble were slightly lower than those of the MP ensemble generally after forecast hour 24 at all thresholds. The SK ensemble generally had lower FSSs than the MP ensemble except after forecast hour 23 at the lightest threshold (0.254 mm) and a few sporadic moments in the range of forecast hours 12-18 at the other thresholds. The difference in FSSs between the SK and MP ensembles at heavy rain thresholds was especially large, providing further evidence of the need to incorporate physics uncertainty into a convective-scale ensemble for heavy precipitation forecasting.

A neighborhood-based version of the Receiver Operating Characteristic (ROC) curve was also calculated (Mason 1982). The ROC curve is a plot of the probability of detection (POD) against the probability of false detection (POFD). Since a ROC curve contains the point (POFD,POD) = (0,0) and (1,1), but varies between those endpoint values, a more useful parameter to evaluate is the area under the ROC curve. Since a ROC curve describes how well a forecast discriminates between yes and no forecasts (i.e., it only forecasts “yes” when the event occurs and only forecasts “no” when the event does not occur), larger ROC areas correspond to more skillful forecasts. ROC areas as a function of 1-hr precipitation threshold for a few forecast hours are shown in Fig. 15. Each ensemble produced skillful QPFs for light rain thresholds after spin-up issues settled. Depending on forecast lead time, ROC area tended to peak at either the very lightest threshold or the light-moderate (2.54-mm or 6.35-mm) thresholds and decreased steadily with increased threshold. For example, at forecast hour 17 (the time of the minimum in domain average precipitation) the highest ROC areas occurred at the lightest threshold and decreased steadily through the highest threshold, whereas at forecast hour 25 (the diurnal peak in precipitation), the highest ROC areas occurred at the 2.54- and 6.35-mm thresholds and decreased more slowly towards both higher and lower thresholds. The shape of the plot at forecast hour 21, during a sharp increase in domain average precipitation, contained features similar to those at both forecast hours 15 and 27. As a function of forecast lead time, skill

generally increased until the late part of the forecast with some oscillation leading up to the peak around forecast hour 26-28, corresponding to evening on the next day. The exception is at the lightest threshold (0.254 mm), where skill peaked at forecast hour 16 and slowly declined afterward (not shown). In general ROC areas agreed with FSSs in that the SKMP ensemble had higher scores than the MP ensemble mostly during the middle portion of the forecast (forecast hours 6-24). Outside of that range, the ensembles had approximately the same skill. Reduced overall sample size likely explains the noisy pattern at the higher thresholds.

5. Conclusions

As a step towards improving the design of convection-allowing EPSs, the impact of a stochastic kinetic energy backscatter scheme was evaluated for a set of warm season cases over a large portion of the continental U.S. Three seven-member ensembles were constructed for the testing. The SK ensemble contained no physics diversity among the members, but the SKEB scheme was employed. Diversity in this ensemble came from the random seed used to generate the pseudo-random numbers. The MP ensemble was a mixed-physics ensemble containing variations in the microphysics, planetary boundary layer, and land surface model parameterizations. The SKEB scheme was not active in the MP ensemble. The SKMP ensemble had the same mixed-physics configuration as the MP ensemble with the SKEB scheme turned on. These ensembles were designed to answer the following questions regarding convective-scale ensemble forecasting:

1) Can a stochastic error representation scheme (SKEB in this case) add meaningful ensemble information and improve the forecast distribution?

2) Is the stochastic error representation method compatible with a mixed-physics approach? If so, does the combination of these methods further improve probabilistic forecasts on the convective-scale?

Each ensemble member had 4 km grid spacing (no convection parameterization was used) and was initialized at 0000 UTC, running for 30 hours to give a complete day 1 forecast of next-day severe weather and heavy precipitation. Both large-scale fields such as temperature, height, and winds above the boundary layer, as well as 2-m temperature, 2-m dewpoint, 10-m wind components, and 1-hr accumulated precipitation were verified using both grid-point and neighborhood probabilistic verification metrics.

The SKEB scheme is designed to (1) correct for insufficient kinetic energy near the grid scales of a forecast model and (2) add spread to the ensemble. The SKEB scheme injects kinetic energy into the model at all scales through additive perturbations to the rotational wind and temperature fields. The NWP model used in this study does not appear to suffer from excessive kinetic energy dissipation in the mesoscales, and thus (1) was not of major concern in this study. For a reasonably tuned SKEB scheme, our study found positive impact on ensemble spread and probabilistic forecasts. Neither robust nor comprehensive tests of the parameters for optimal tuning of the scheme were performed; optimal tuning of the scheme for use at the convective-scale is left for future work.

The SKEB scheme was successful in accomplishing (2). Marked gains in ensemble spread were noted in nearly every field verified, especially large-scale fields. Spread was even increased in fields that were not directly perturbed (i.e., specific humidity, dewpoint, and precipitation), although the increase in spread in those fields was reduced compared to the increase in other fields that were directly perturbed. The increase in spread is also confirmed through examining the rank histograms. Histograms in the SKMP ensemble were flatter than those from the MP

ensembles. The increased spread was accompanied by a reduction in the rmse of the ensemble mean. Some of these reductions were statistically significant.

Quantitative precipitation forecasts were also improved. Since the SKEB scheme does not correct individual grid point errors in deterministic forecasts nor does it perturb the moisture field directly, the connection between the SKEB scheme and precipitation is convoluted and indirect, occurring through changes in stability of air parcels on the convective scale as well as through changes in the wind field that provide forcing for convection initiation and affect ongoing storms. The perturbations are very small, so the changes are subtle but can accumulate over time periods long enough to impact the evolution of the near-surface wind and temperature field enough to affect the initiation or maintenance of convection. It is difficult to determine the precise factors that impact the change in PQPF skill from the use of the SKEB scheme given the random nature of the perturbations. For some individual cases ensemble forecasts are bound to be improved through a better representation of the uncertainty of the atmospheric state.

The performance of the SK ensemble was competitive with the other ensembles despite the lack of physics diversity. It contained almost as much spread as the SKMP ensemble at many forecast hours and in many fields (thus exceeding that from the MP ensemble) and also had flatter rank histograms than the MP ensemble. The rmse of the SK ensemble mean was also similar to that of the SKMP and MP ensembles. There were even forecast hours and thresholds at which precipitation skill scores in the SK ensemble were better than either of the MP or SKMP ensembles. However, in agreement with Berner et al. (2011), the spread of the SK ensemble was much lower in the boundary layer compared to the MP ensemble. Thus it seems the best choice is to combine the uncertainty in the physical processes impacting temperature, wind, and moisture in the boundary layer by using mixed physical parameterizations with the uncertainty in the dynamics and in other unparameterized subgrid-scale processes by using the SKEB scheme. Hacker et al. (2011) and Berner et al. (2015) also found the combination of a SKEB scheme and physics diversity to give the best forecasts at and below 700 hPa for convection parameterizing resolutions. Additionally, similar to Duda et al. (2014), the performance of the SK ensemble is likely sensitive to the physics parameterization options used (Morrison microphysics, YSU PBL, and Noah land-surface); this combination was shown to be more accurate than most of the other combinations used in the MP ensemble, and the SKEB scheme parameters were tuned for this particular combination and were not changed when used with other physics combinations. Since the optimal parameters for the SKEB scheme may be dependent on the choice of physics, the comparison between the SK and SKMP ensembles may not maximize quantitatively the added value of including the SKEB scheme on top of the mixed-physics approach. Tuning the SKEB scheme for individual physics packages and determining the sensitivity of similar SK ensembles based on other physics combinations is left for future work.

This study is among the first to examine the effect of combining stochastic methods with traditionally used mixed-physics methods for convection allowing ensemble design. Given the resources needed to maintain various physics schemes, future research is still needed to explore to what extent a mixed-physics method is needed in the presence of a stochastic method. It is also acknowledged that the conclusion may also be dependent on the diagnostic and verification methods adopted.

The experiment presented in this paper did not incorporate initial or lateral boundary condition perturbations to isolate the impact of the model-error representation on the ensemble forecasts. Such perturbations could further broaden the forecast probability distribution and reduce or eliminate poor forecasts (present in this study but not discussed) caused by inadequate

initial and lateral boundary conditions. Future work should incorporate such initial and lateral boundary condition error representation using advanced ensemble based data assimilation (Johnson et al. 2015) with model-error representation to further improve convective scale ensemble forecasts.

Acknowledgements. This research was primarily supported by NSF Grants AGS-1046081, AGS-0802888, and a grant from the NOAA CSTAR program. The computing for this project was performed at the OU Supercomputing Center for Education & Research (OSCER) at the University of Oklahoma (OU). OSCER Director Henry Neeman and Petascale Storage Administrator Patrick Calhoun were especially helpful in providing guidance on computing and storage procedures. The quality of this paper was improved by the helpful comments from two anonymous reviewers.

References

Berner, J., G. Shutts, M. Leutbecher, and T. Palmer, 2009: A spectral stochastic kinetic energy backscatter scheme and its impact on flow dependent predictability in the ECMWF Ensemble Prediction System. *J. Atmos. Sci.,* **66**, 603–626.

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

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

Bouttier, F., B.Vié, O. Nuissier, and L. Raynaud, 2012: Impact of Stochastic Physics in a Convection-Permitting Ensemble. *Mon. Wea. Rev.,* **140**, 3706-37.

Brown, J. M., and Coauthors, 2011: Improvement and testing of WRF physics options for application to Rapid Refresh and High Resolution Rapid Refresh. *Preprints, 14th Conf. on Mesoscale Processes/15th Conf. on Aviation, Range, and Aerospace Meteorology*, Los Angeles, CA, Amer. Meteor. Soc., 5.5. [Available online at https://ams.confex.com/ams/14Meso15ARAM/webprogram/Paper191234.html.]

Buizza, R., M. Miller, and T. N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **125**, 2887–2908.

Candille, G., 2009: The multiensemble approach: The NAEFS example. *Mon. Wea. Rev.*, **137**, 1655 – 1665.

Caron, J.-F., 2013: Mismatching perturbations at the lateral boundaries in limited-area ensemble forecasting: A case study. *Mon. Wea. Rev.*, **141**, 356–374.

Charron, M., G. Pellerin, L. Spacek, P. L. Houtekamer, N. Gagnon, H. L. Mitchell, and L. Michelin, 2010: Toward random sampling of model error in the Canadian ensemble prediction system. *Mon. Wea. Rev.*, **138**, 1877 – 1901.

Clark, A. J., and co-authors, 2011: Probabilistic precipitation forecast skill as a function of ensemble size and spatial scale in a convection-allowing ensemble. *Mon. Wea. Rev.,* **139**, 1410-1418.

Du, J., G. DiMego, B. Zhou, D. Jovic, B. Ferrier, B. Yang, and S. Benjamin, 2014: NCEP Regional Ensembles: Evolving toward hourly-updated convection-allowing scale and storm-scale predictions within a unified regional modeling system. *22nd Conf. on Numerical*

Weather Prediction and 26th Conf. on Weather Analysis and Forecasting, Atlanta, GA, Amer. Meteor. Soc., Feb. 1-6, 2014, paper J1.4.

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

Ebert, E. E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.*, **129**, 2461-2480.

ECMWF, 2012: IFS documentation – cy37r2. Operational implementation 18 May 2011. Part V: Ensemble prediction system. [Available online at http://www.ecmwf.int/research/ifsdocs/CY37r2/IFSPart5.pdf]

Ek, M. B., K. E. Mitchell, Y. Lin, E. Rogers, P. Grunmann, V. Koren, G. Gayno, and J. D. Tarpley, 2003: Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model. *J. Geophys. Res.*, **108**, 8851, doi:10.1029/2002JD003296.

Ha, S.-Y., J. Berner, C. Snyder, 2014: Model-error representation in mesoscale WRF-DART cycling. *Mon. Wea. Rev.*, submitted.

Hacker, J. P., S.-Y. Ha, C. Snyder, J. Berner, F. A. Eckel, E. Kuchera, M. Pocernich, S. Rugg, J. Schramm, and X. Wang, 2011: The U.S. Air Force Weather Agency's mesoscale ensemble: Scientific description and performance results. *Tellus,* **63A**, 1-17.

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

——— and J.-O. J. Lim, 2006: The WRF single-moment 6-class microphysics scheme (WSM6). *J. Korean. Meteor. Soc.*, **42**, 129-151.

Johnson, A., and X. Wang, 2012: Verification and calibration of neighborhood and object-based probabilistic precipitation forecasts from a multi-model convection-allowing ensemble. *Mon. Wea. Rev.*, **140**, 3054–3077, doi:10.1175/MWR-D-11-00356.1.

——, and ——, 2013: Object-based evaluation of a storm scale ensemble during the 2009 NOAA Hazardous Weather Testbed Spring Experiment. *Mon. Wea. Rev.*, **141**, 1079–1098, doi:10.1175/MWR-D-12-00140.1.

——, ——, F. Kong, and M. Xue, 2011a: Hierarchical cluster analysis of a convection-allowing ensemble during the Hazardous Weather Testbed 2009 Spring Experiment. Part I: Development of the object-oriented cluster analysis method for precipitation fields. *Mon. Wea. Rev.*, **139**, 3673–3693, doi:10.1175/MWR-D-11-00015.1.

——, ——, ——, and ——, 2011b: Hierarchical cluster analysis of a convection-allowing ensemble during the Hazardous Weather Testbed 2009 Spring Experiment. Part II: Ensemble clustering over the whole experiment period. *Mon. Wea. Rev.*, **139**, 3694–3710, doi:10.1175/MWR-D-11-00016.1.

——, ——, ——, and ——, 2013: Object-based evaluation of the impact of horizontal grid spacing on convection-allowing forecasts. *Mon. Wea. Rev.*, **141**, 3413–3425, doi:10.1175/ MWR-D-13-00027.1.

——, and Coauthors, 2014: Multiscale characteristics and evolution of perturbations for warm season convection-allowing precipitation forecasts: Dependence on background flow and method of perturbation. *Mon. Wea. Rev.*, **142**, 1053–1073, doi:10.1175/MWR-D-13-00204.1.

–––, X. Wang, J. Carley, L. Wicker, and C. Karstens, 2015: A comparison of multi-scale GSI-based EnKF and 3DVar data assimilation using radar and conventional observations for mid-latitude convective-scale precipitation forecasts. *Mon. Wea. Rev.*, **143**, 3087 – 3108.

Kong, F., K. K. Droegemeier, and N. L. Hickmon, 2006: Multiresolution ensemble forecasts of an observed tornadic thunderstorm system. Part I: Comparison of coarse- and fine-grid experiments. Mon. Wea. Rev., **134**, 807-833.

——, ——, and ——, 2007a: Multiresolution ensemble forecasts of an observed tornadic thunderstorm system. Part II: Storm-scale experiments. Mon. Wea. Rev., **135**, 759-782.

Kong, F., and Coauthors, 2007b: Preliminary analysis on the real-time storm-scale ensemble forecasts produced as a part of the NOAA Hazardous Weather Testbed 2007 spring experiment. *Preprints, 22nd Conf. on Weather Analysis and Forecasting*, Park City, UT, Amer. Meteor. Soc., Paper 3B.2.

Kong, F., and Coauthors, 2009: A real-time storm-scale ensemble forecast system: 2009 Spring Experiment. *Preprints*, *23nd Conf. on Weather Analysis and Forecasting/19th Conf. Num. Wea. Pred.,* Omaha, Nebraska, Amer. Meteor. Soc., Paper 16A3.

Lorenz, E. N., 1969: The predictability of a flow which possesses many scales of motion. *Tellus*, **21**, 289–307.

Mason, I., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.*, **30**, 291–303.

Mason, P., and D. Thomson, 1992: Stochastic backscatter in large eddy simulations of boundary layers. *J. Fluid Mech.*, **242**, 51–78.

Mittermaier, M. P., 2007: Improving short-range high-resolution model precipitation forecast skill using time-lagged ensembles. *Quart. J. Roy. Meteor. Soc.*, **133**, 1487–1500.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122**, 73-119.

Morrison, H., G. Thompson, and T. Tatarskii, 2009: Impact of cloud microphysics on the development of trailing stratiform precipitation in a simulated squall line: Comparison of one- and two-moment schemes. *Mon. Wea. Rev.*, **137**, 991–1007.

Nastrom, G. D., and K. S. Gage, 1985: A climatology of atmospheric wavenumber spectra of wind and temperature observed by commercial aircraft. *J. Atmos. Sci.*, **42**, 950–960. doi: http://dx.doi.org/10.1175/1520-0469(1985)042<0950:ACOAWS>2.0.CO;2

Palmer, T. N., R. Buizza, F. Doblas-Reyes, T. Jung, M. Leutbecher, G. J. Shutts, M. Steinheimer, and A. Weisheimer, 2009: Stochastic parametrization and model uncertainty. ECMWF technical memo 598. [Available online at http://old.ecmwf.int/publications/library/ecpublications/_pdf/tm/501-600/tm598.pdf]

Roberts, N. M., and H. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97.

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

Ropnack, A., A. Hense, C. Gebhardt, and D. Majewski, 2013: Bayesian model verification of NWP ensemble forecasts. *Mon. Wea. Rev.*, **141**, 375 – 387.

Rotunno R., and C. Snyder, 2008: A generalization of Lorenz's model for the predictability of flows with many scales of motion. *J. Atmos. Sci.*, **65**, 1063-1076.

Schwartz, C.S., J.S. Kain, S.J. Weiss, M. Xue, D.R. Bright, F. Kong, K.W. Thomas, J.J. Levit, M.C. Coniglio, and M.S. Wandishin, 2010: Toward improved convection-allowing

ensembles: Model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Wea. Forecasting*, **25**, 263‑280.
Shutts, G. J. 2005. A kinetic energy backscatter algorithm for use in ensemble prediction systems. *Quart. J. R. Meteor. Soc.* **612**, 3079–3102.
——, and T. N. Palmer, 2007: Convective forcing fluctuations in a cloud-resolving model: Relevance to the stochastic parameterization problem. *J. Climate*, **20**, 187–202.
Skamarock, W. C., 2004: Evaluating mesoscale NWP models using kinetic energy spectra. *Mon. Wea. Rev.*, **132**, 3019–3032.
----, and Coauthors, 2008: A description of the advanced research WRF version 3. NCAR technical note NCAR/TN-475+STR. 113 pp.
Tennant, W. J., G. J. Shutts, A. Arribas, and S. A. Thompson, 2011: Using a stochastic kinetic energy backscatter scheme to improve MOGREPS probabilistic forecast skill. *Mon. Wea. Rev.*, **139**, 1190–1206.
Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: the generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317-2330.
Vié, B., O. Nuisser, and V. Ducrocq, 2011: Cloud-resolving ensemble simulations of Mediterranean heavy precipitating events: Uncertainty on initial conditions and lateral boundary conditions. *Mon. Wea. Rev.*, **139**, 403–423.
Wandishin, M. S., S. L. Mullen, D. J. Stensrud, and H. E. Brooks, 2001: Evaluation of a short-range multimodel ensemble system. *Mon. Wea. Rev.*, **129**, 729–747.
Wang, X., and C. H. Bishop, 2003: A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. *J. Atmos. Sci.*, **60**, 1140–1158, doi:10.1175/1520-0469(2003)060,1140:ACOBAE.2.0.CO;2.
----, ----, and S. J. Julier, 2004: Which is better, an ensemble of positive-negative paris or a centered spherical simplex ensemble? *Mon. Wea. Rev.*, **132**, 1590–1605.
Weygandt, S. S., and Coauthors, 2011: Evaluation of the National Center for Environmental Prediction (NCEP) implementation version of the Rapid Refresh and its skill in providing short-term guidance for aviation hazards. *Preprints, 15th Conf. on Aviation, Range, and Aerospace Meteorology*, Los Angeles, CA, Amer. Meteor. Soc., 5.4. [Available online at https://ams. confex.com/ams/14Meso15ARAM/webprogram/Paper191213.html.]
Xue, M., and Coauthors, 2011: CAPS Realtime Storm Scale Ensemble and High Resolution Forecasts for the NOAA Hazardous Weather Testbed 2010 Spring Experiment. *24th Conf. Wea. Forecasting/20th Conf. Num. Wea. Pred.*, Amer. Meteor. Soc., Paper 9A.2.
Yussouf, N., and D. J. Stensrud, 2012: Comparison of single-parameter and multiparameter ensembles for assimilation of radar observations using the ensemble Kalman filter. *Mon. Wea. Rev.*, **140**, 562–586. doi: http://dx.doi.org/10.1175/MWR-D-10-05074.1
Zhang, J., and Coauthors, 2011: National Mosaic and Multisensor QPE (NMQ) system: Description, results, and future plans. *Bull. Amer. Meteor. Soc.*, **92**, 1321–1338, doi:10.1175/2011BAMS-D-11-00047.1.

Tables

Table 1. Configuration of the ensembles. The last column indicates the integer value of the random number seed used to generate pseudo-random numbers in the FORTRAN code that the WRF is built on. The SK ensemble used the same physics configuration as member 1, but the random number seed varied among the members as indicated.

Member	Microphysics	PBL	LSM	seed
m1 (control)	Morrison	YSU	Noah	2
m2	Ferrier	MYNN2.5	RUC	3
m3	WSM6	YSU	Noah	4
m4	Thompson	MYJ	Noah	5
m5	MY	ACM2/QNSE*	RUC	6
m6	WDM6	MYJ	RUC	7
m7	NSSL	QNSE	Noah	8

*-The PBL scheme for member 5 was switched from ACM2 to QNSE starting with the case initialized at 0000 UTC 8 May due to difficulties resulting from the interaction between the ACM2 PBL scheme and the other physics options in that member.

Table 2. Abbreviations for field names used for verification.

Field name	Description (units)
hgt500	500 hPa geopotential height (m)
v850__	850 hPa v-wind component (m s^{-1})
v500__	500 hPa v-wind component (m s^{-1})
u500__	500 hPa u-wind component (m s^{-1})
u250__	250 hPa u-wind component (m s^{-1})
sph850	850 hPa specific humidity (g kg^{-1})
pwat__	precipitable water (mm)
tmp850	850 hPa temperature (K)
tmp500	500 hPa temperature (K)
accppt	1-hr accumulated precipitation (mm)

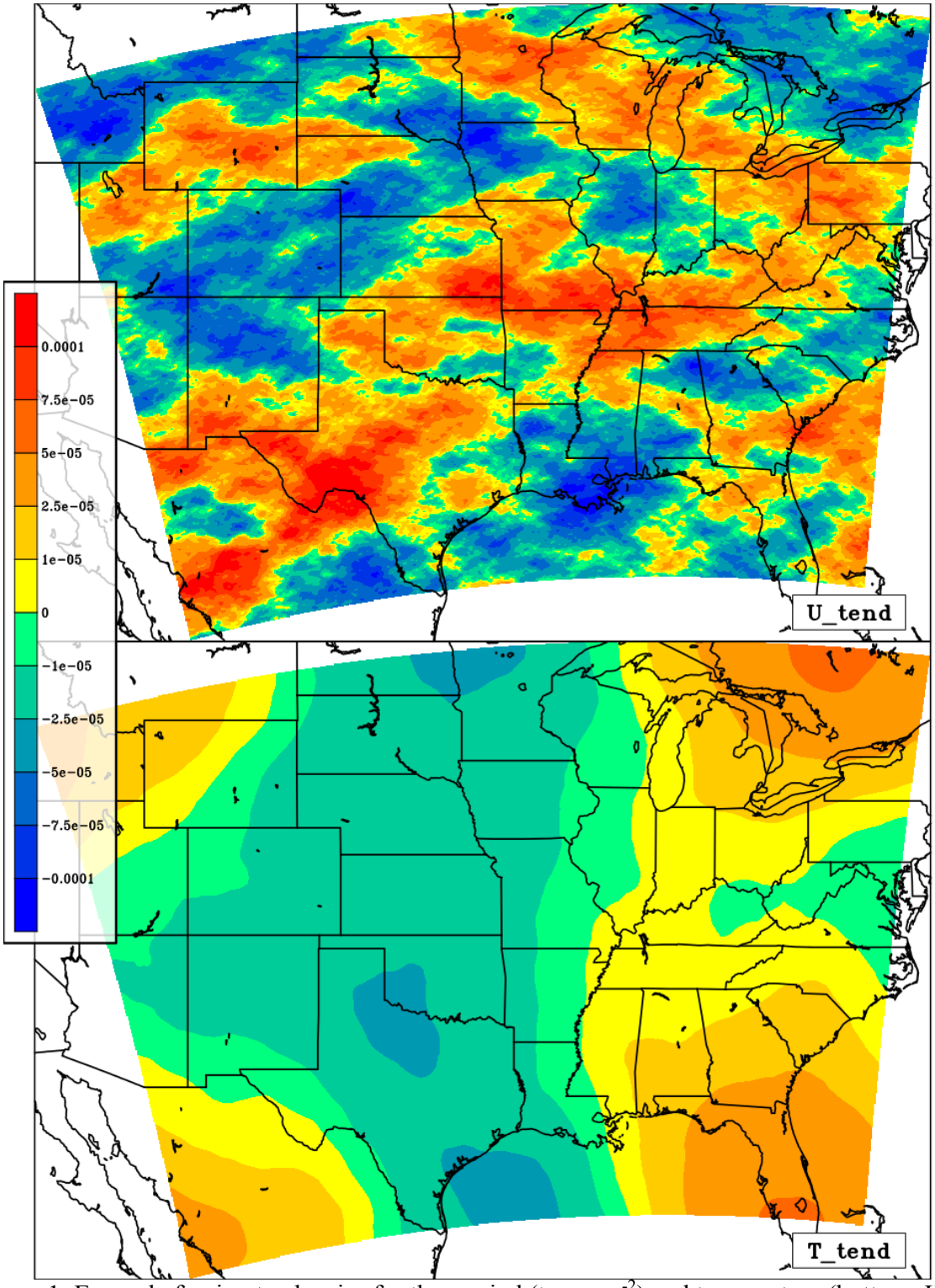

Figure 1. Example forcing tendencies for the u-wind (top; m s^{-2}) and temperature (bottom; K s^{-1}) fields.

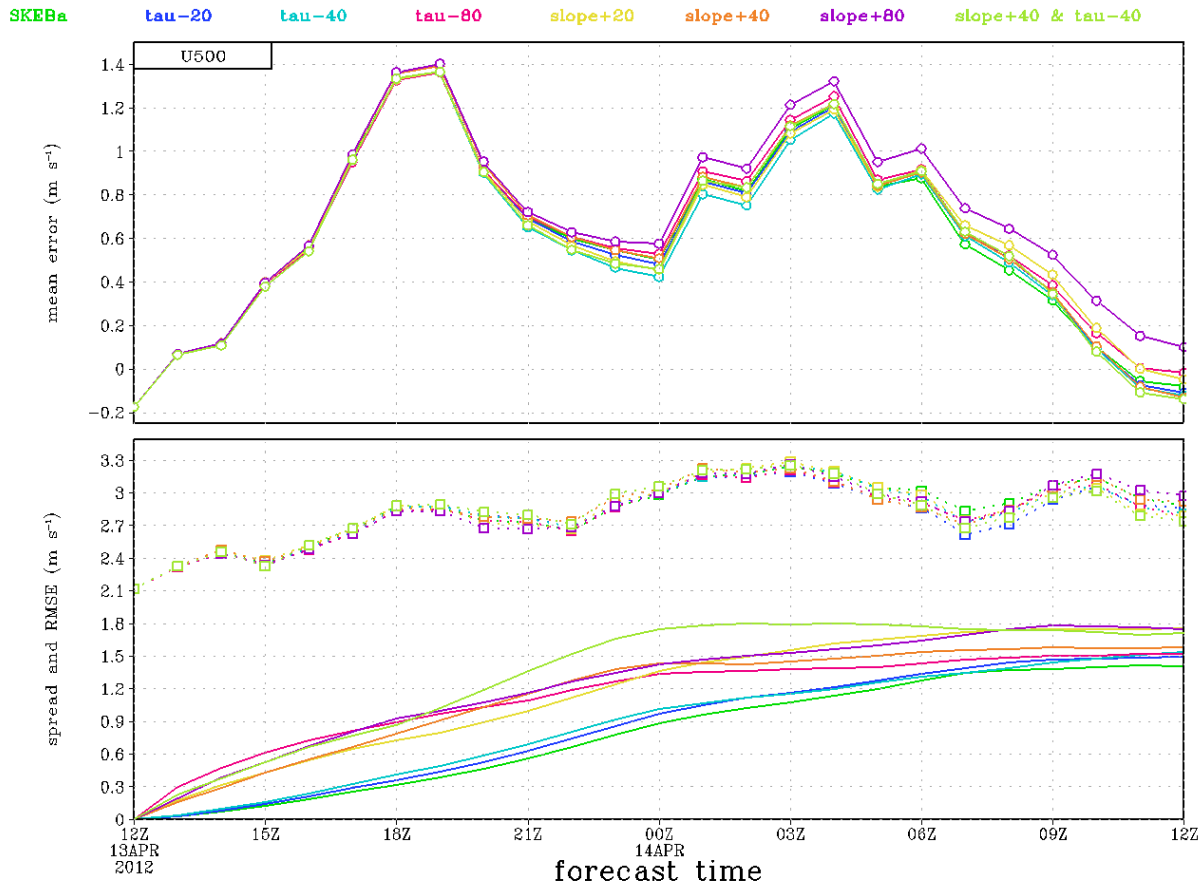

Figure 2. Bias (top), RMSE (bottom dotted), and spread (bottom solid) of the u-wind component at 500 hPa verified against RAP analyses for a test case initialized at 1200 UTC 13 April 2012. SKEB scheme settings are colored according to the key at top. Numbers indicate the percentage perturbation from the default values of the SKEB scheme (3 hr for decorrelation time tau, 1.83 for the spectral slope).

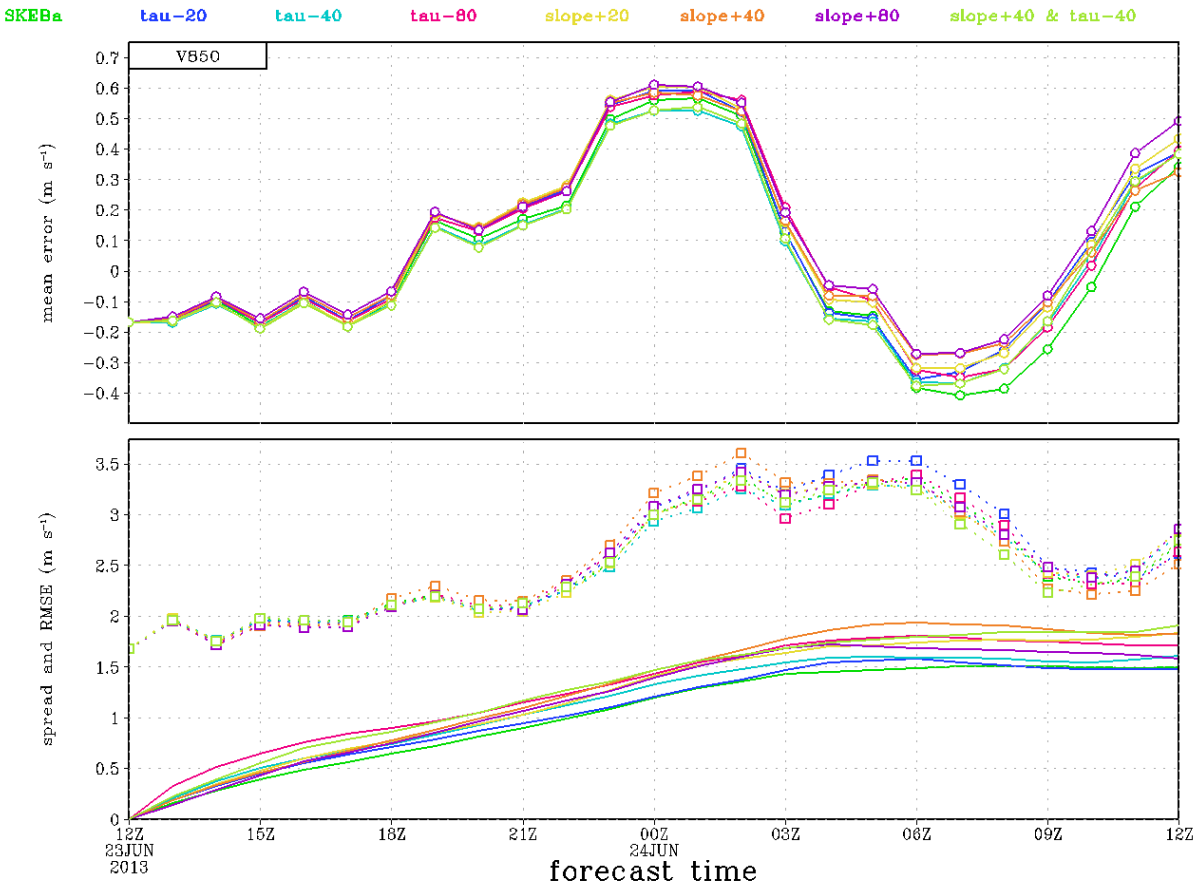

Figure 3. Same as Fig. 2 except for the v-wind component at 850 hPa for a test case initialized at 1200 UTC 23 June 2013.

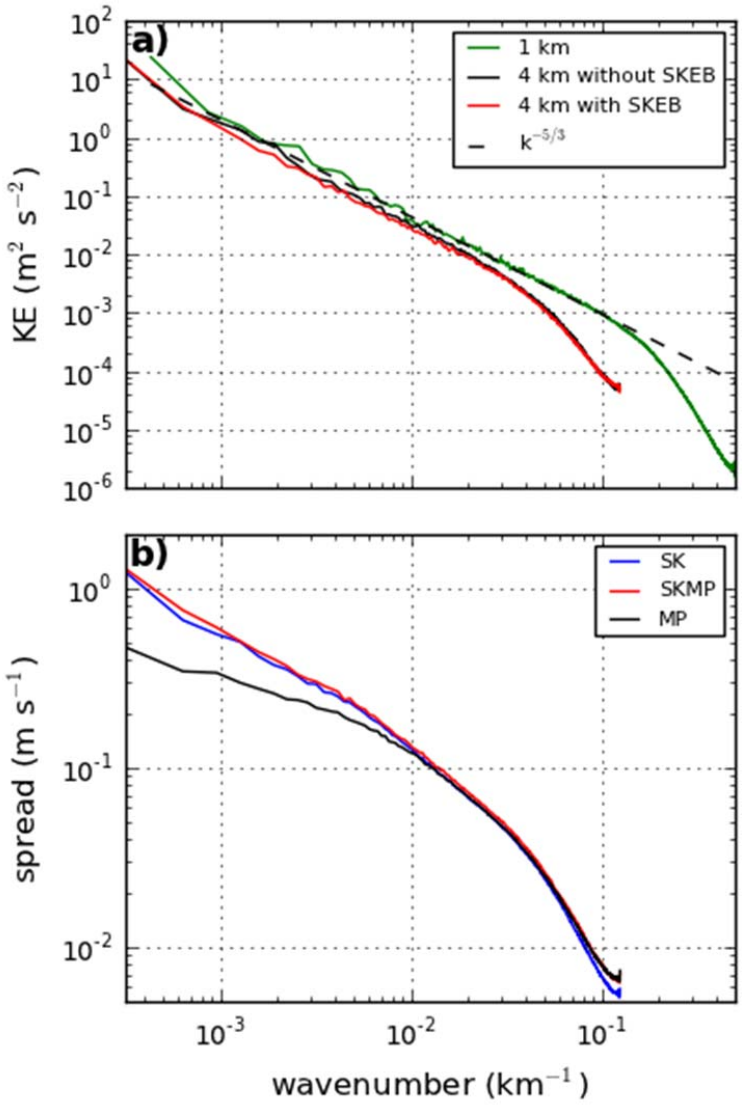

Figure 4. (a) Kinetic energy spectra from WRF simulations with grid spacings of 4 km and 1 km (24 hour forecasts from different initializations). The solid red and black spectra are from otherwise identical WRF simulations at 4 km grid spacing where one uses the SKEB scheme and the other does not. A reference $k^{-5/3}$ slope is included in dashed black. (b) Spectral decomposition of u-wind spread from one case, also a 24 hour forecast.

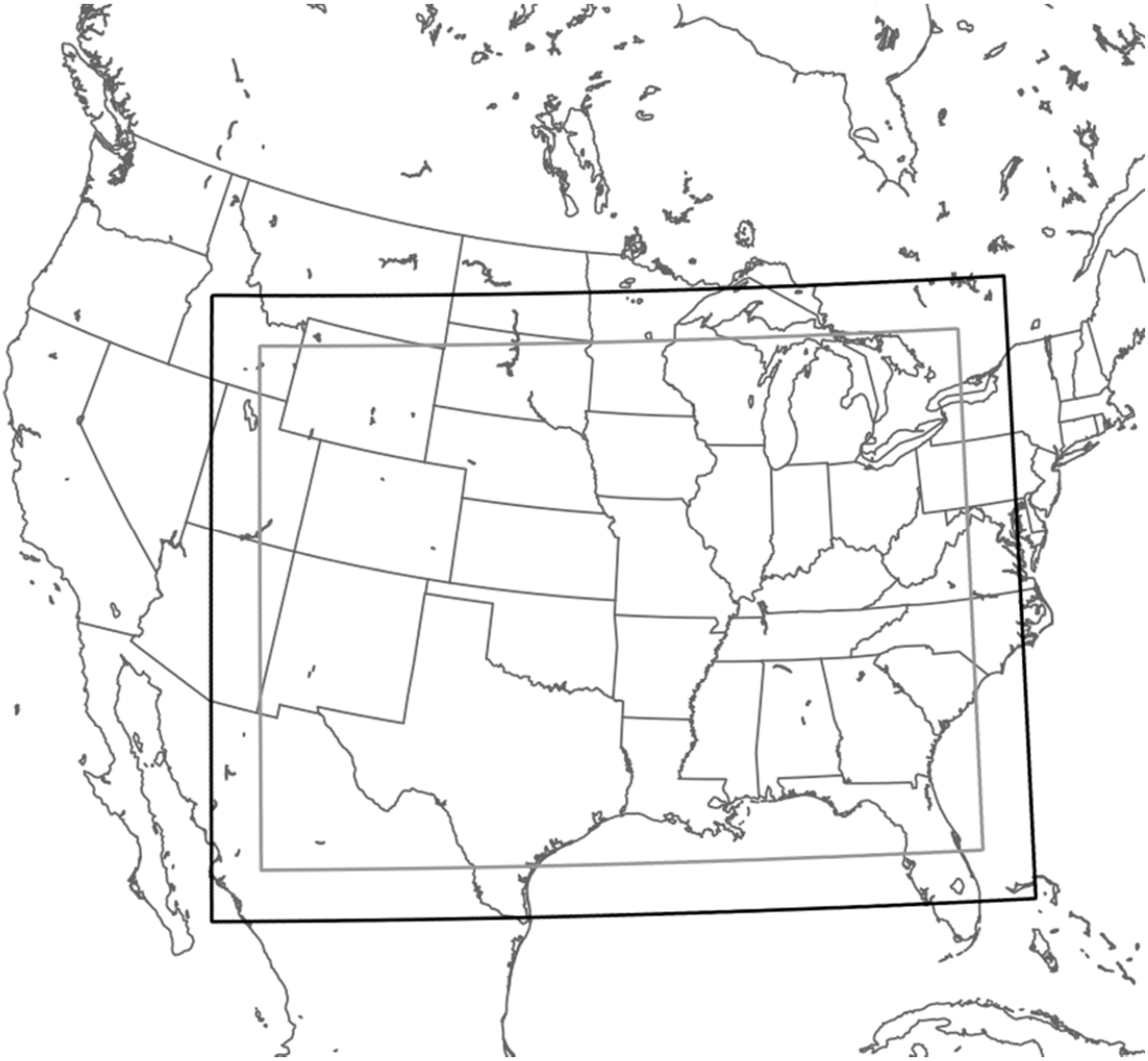

Figure 5. Model (thick black) and verification (thinner gray) domains.

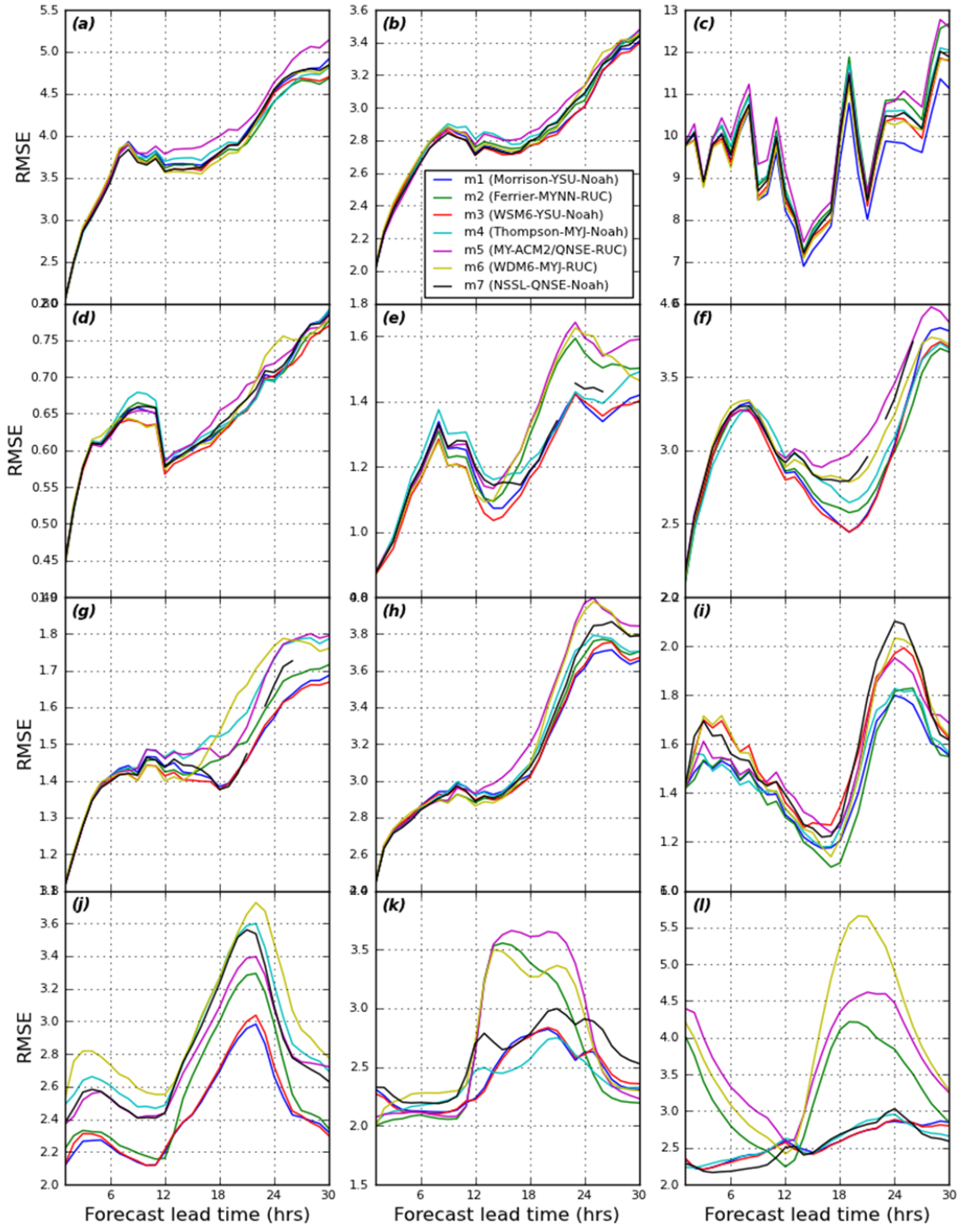

Figure 6. Member rmse for (a) u-wind at 250 hPa (m s^{-1}), (b) u-wind at 500 hPa (m s^{-1}), (c) 500-hPa geopotential height (m), (d) temperature at 500 hPa (K), (e) temperature at 850 hPa (K), (f) v-wind at 850 hPa (m s^{-1}), (g) specific humidity at 850 hPa (g kg^{-1}), (h) precipitable water (mm), (i) 1-hr accumulated precipitation (mm), (j) u-wind at 10 m (m s^{-1}), (k) temperature at 2 m (K), and (l) dewpoint at 2 m (K).

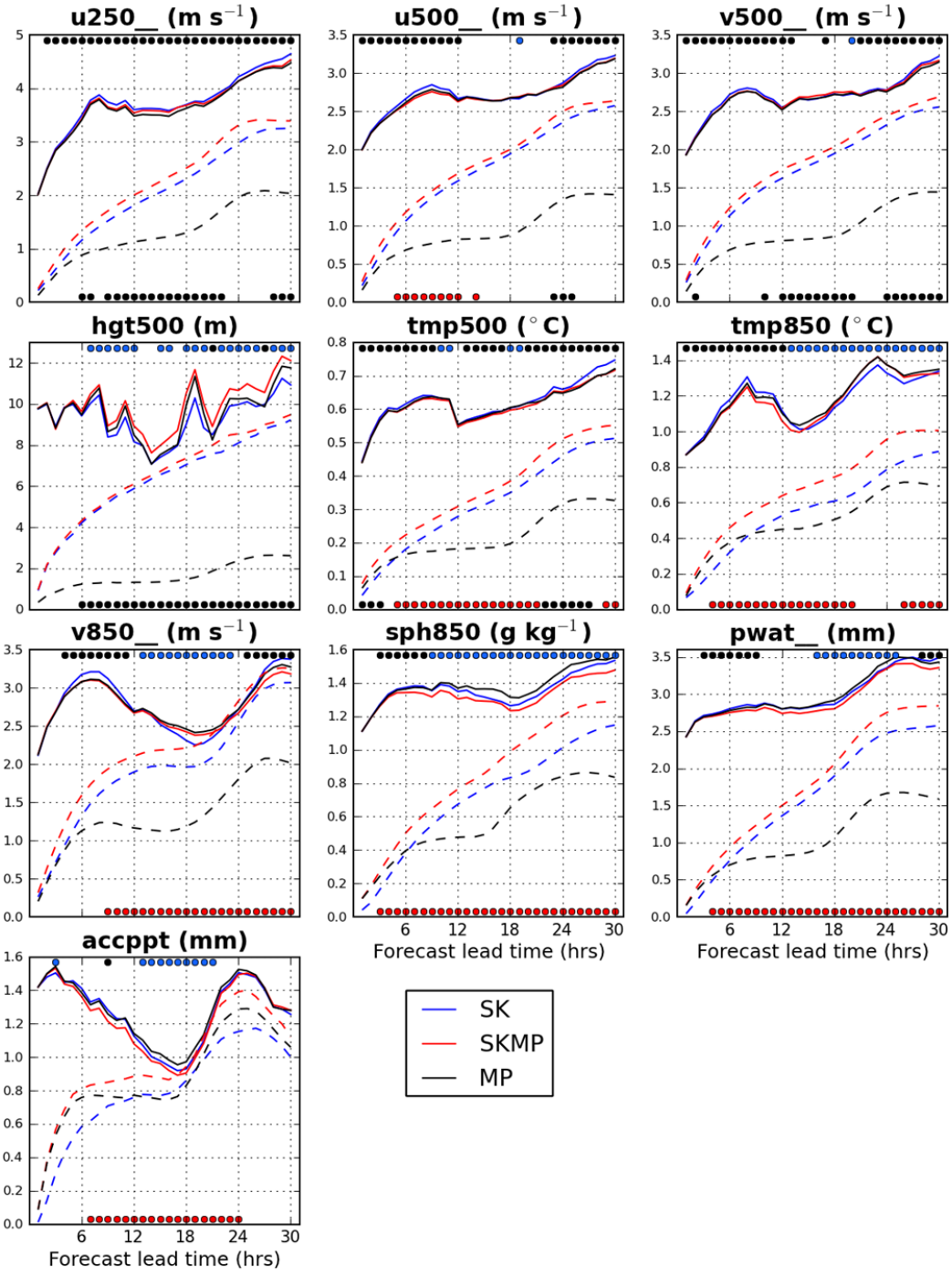

Figure 7. Verification domain-average ensemble mean rmse (solid) and ensemble spread (dashed). Red dots across the bottom indicate forecast hours at which the rmse of the SKMP ensemble was statistically significantly lower than that of the MP ensemble, whereas black dots indicate the opposite. Similarly, across the top of each panel, blue dots indicate when the SK ensemble had a significantly lower rmse than the MP ensemble, whereas black dots indicate the opposite.

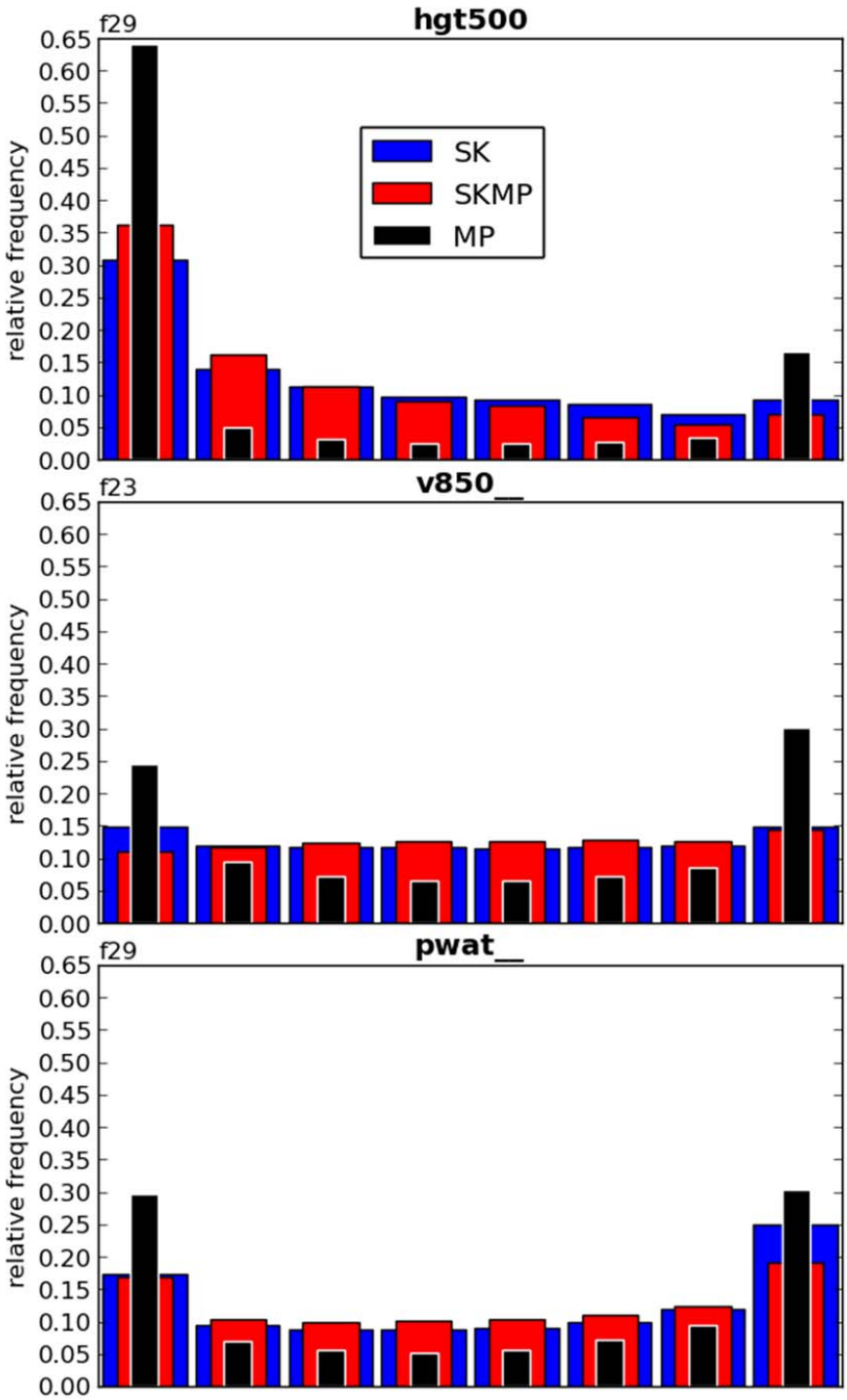

Figure 8. Rank histograms for 500 hPa geopotential height at forecast hour 29 (top), 850 hPa v-wind at forecast hour 23 (middle), and precipitable water at forecast hour 29 (bottom).

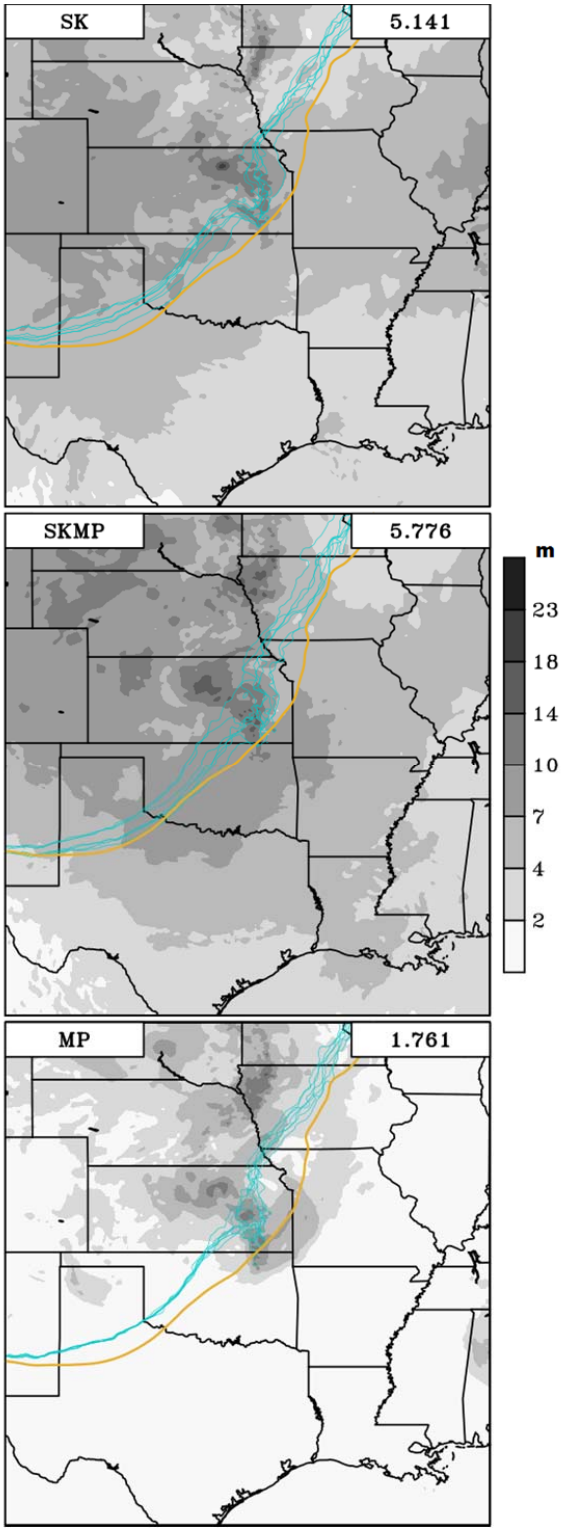

Figure 9. Ensemble standard deviation (shaded, m) of 500-hPa geopotential height valid 1100 UTC 19 May 2013. Individual member 5760-m height contours (light blue) and the analyzed 5760-m height contour from a RAP analysis (gold) are also shown. Area-averaged ensemble spread is indicated in the upper right of each panel.

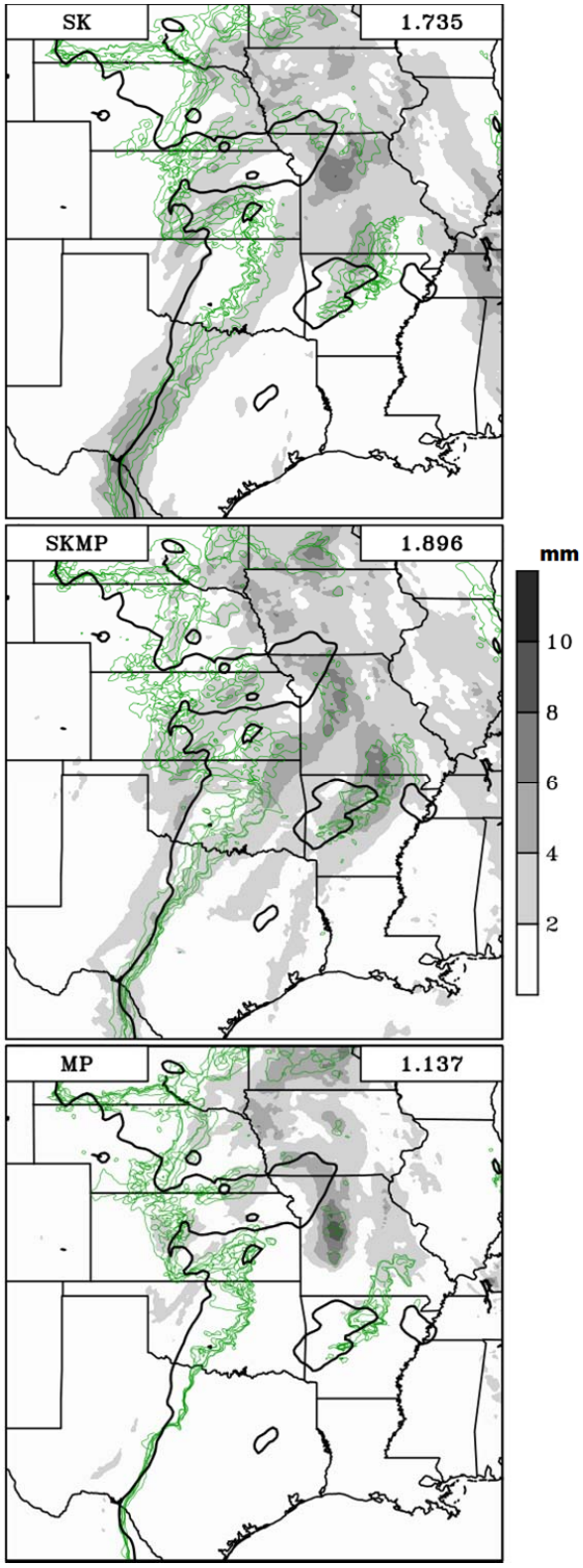

Figure 10. As in Fig. 9 except for precipitable water (mm, 25 mm contour is displayed) valid 1700 UTC 19 May 2013. Individual member contours are in green, whereas the RAP analyzed contour is shown in the thick black line.

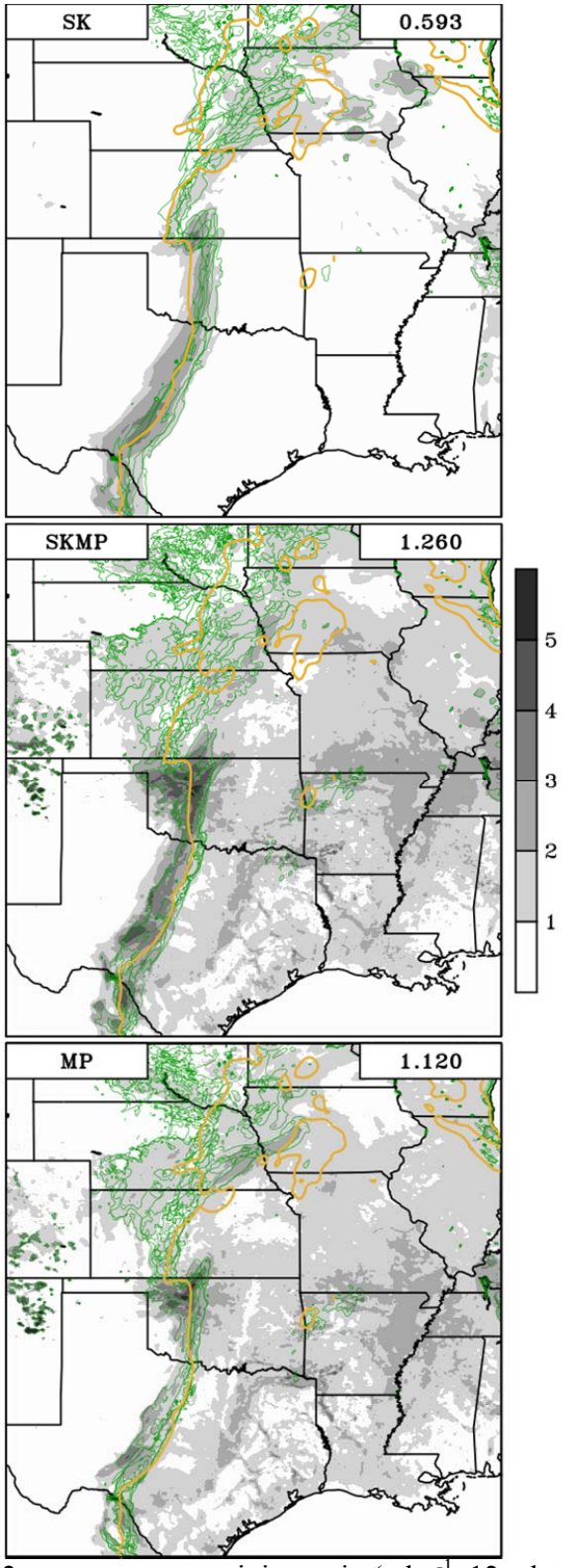

Figure 11. As in Fig. 9 except for 2-m water vapor mixing ratio (g kg^{-1}, 12 g kg^{-1} contour is displayed) at 1900 UTC 19 May 2013. Individual member contours are in green, whereas the RAP analysis contour is the thick gold line.

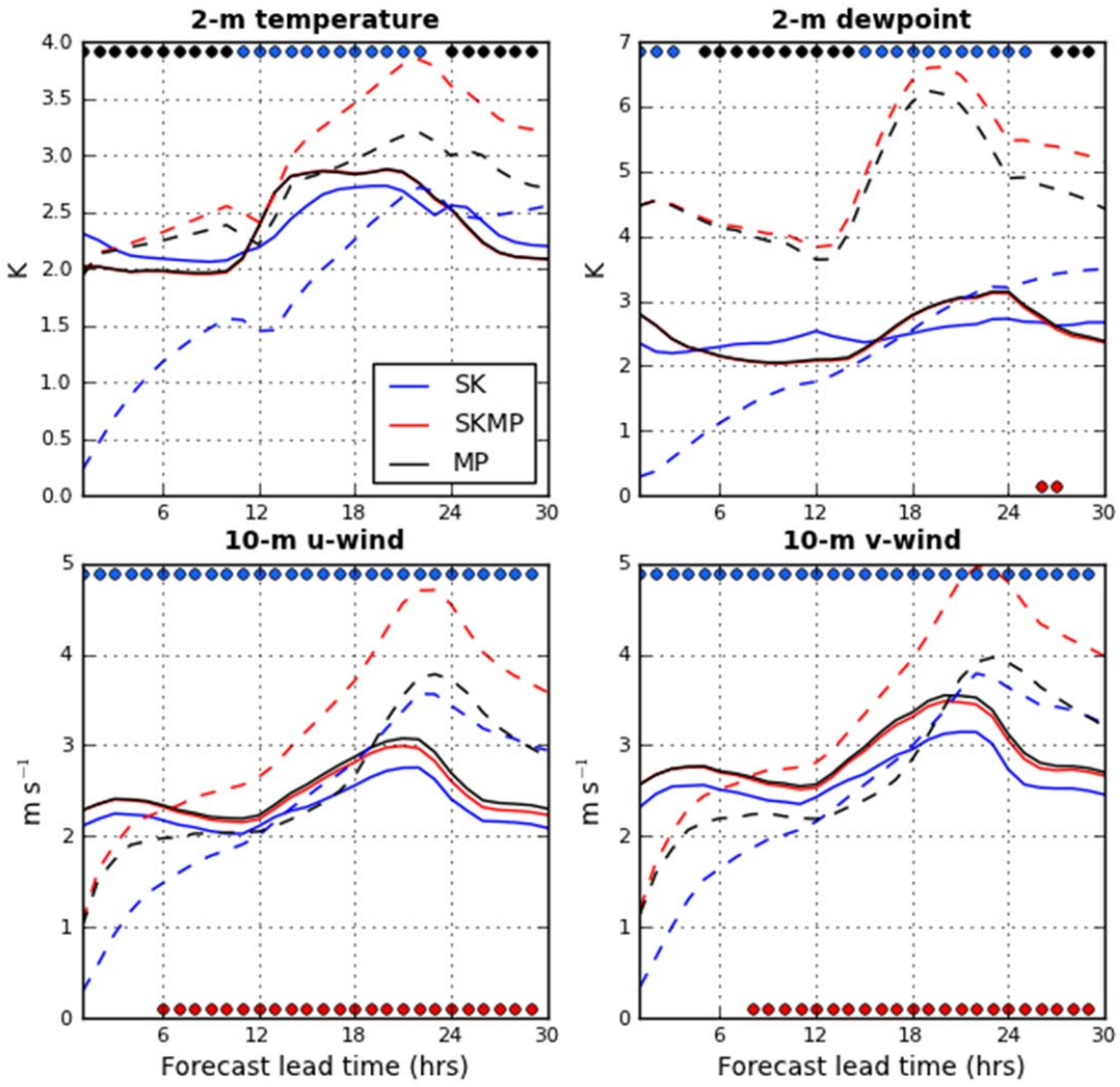

Figure 12. Same as Fig. 7 except for the indicated fields verified against METAR observations.

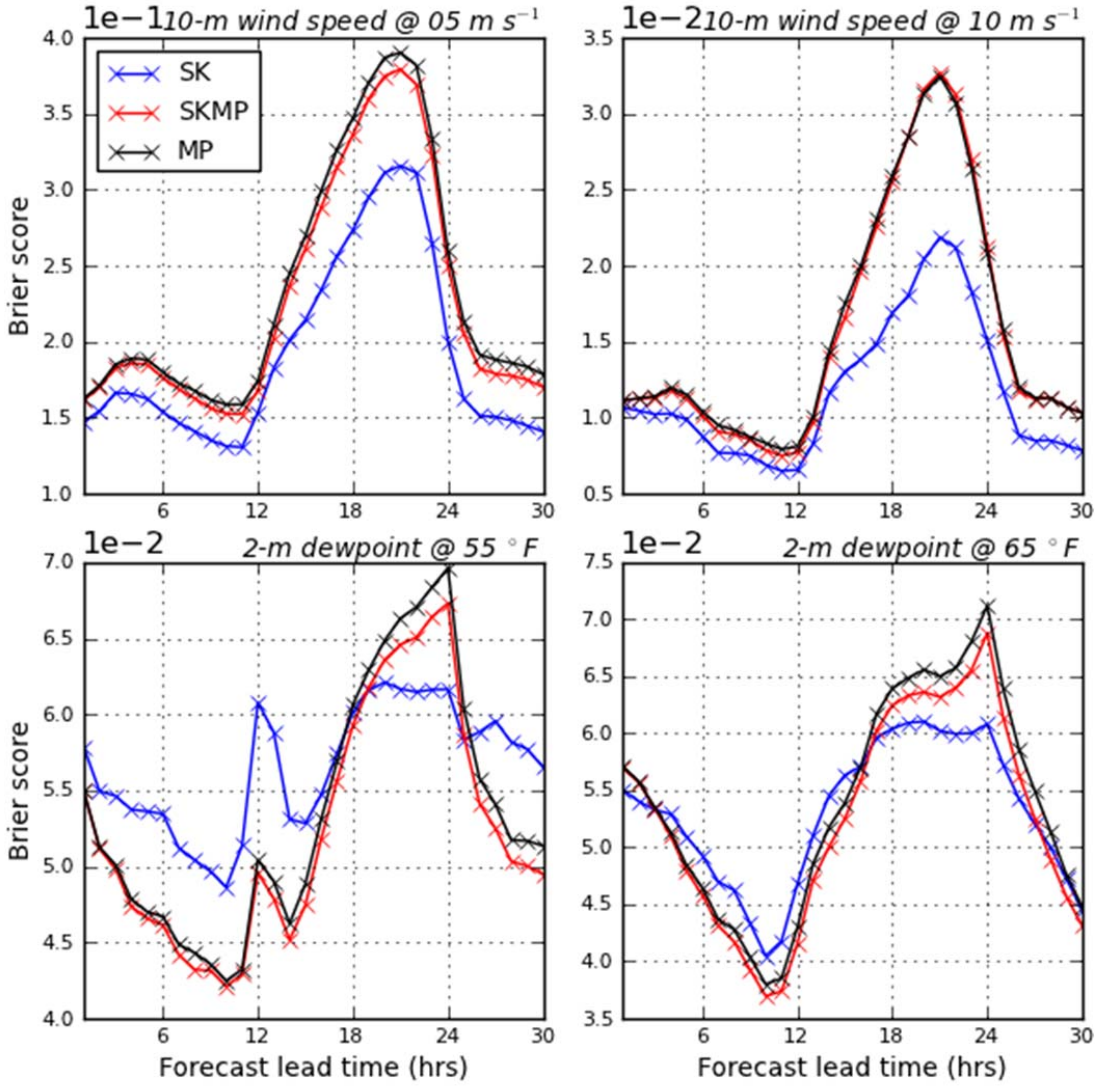

Figure 13. Brier scores for (top row) 10-m wind speed, (middle and bottom rows) 2-m dewpoint forecasts at the indicated thresholds.

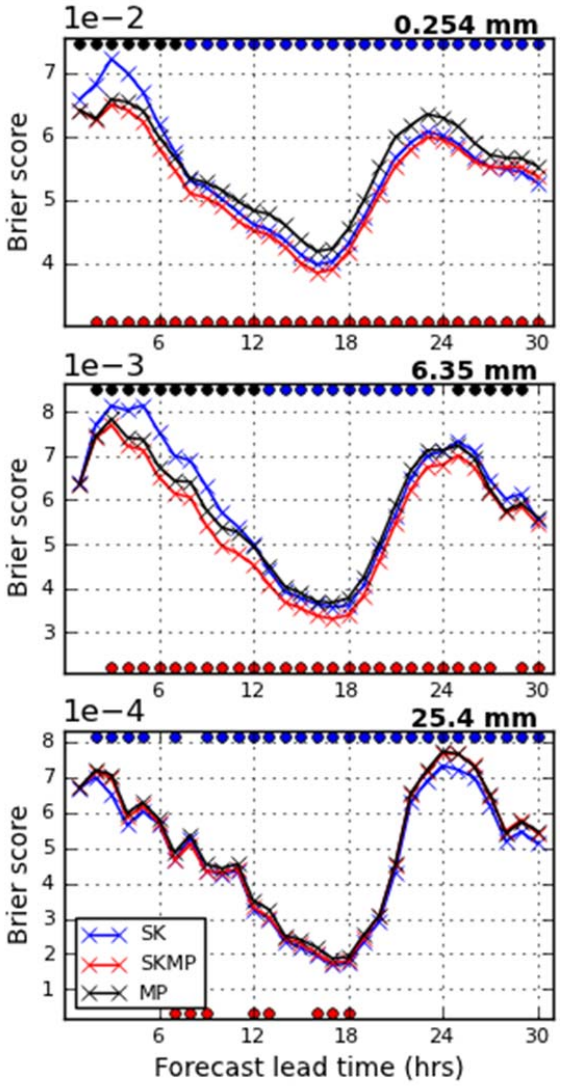

Figure 14. Brier scores for the indicated 1-hr accumulation thresholds. Colored dots represent statistically significant differences as in Fig. 7.

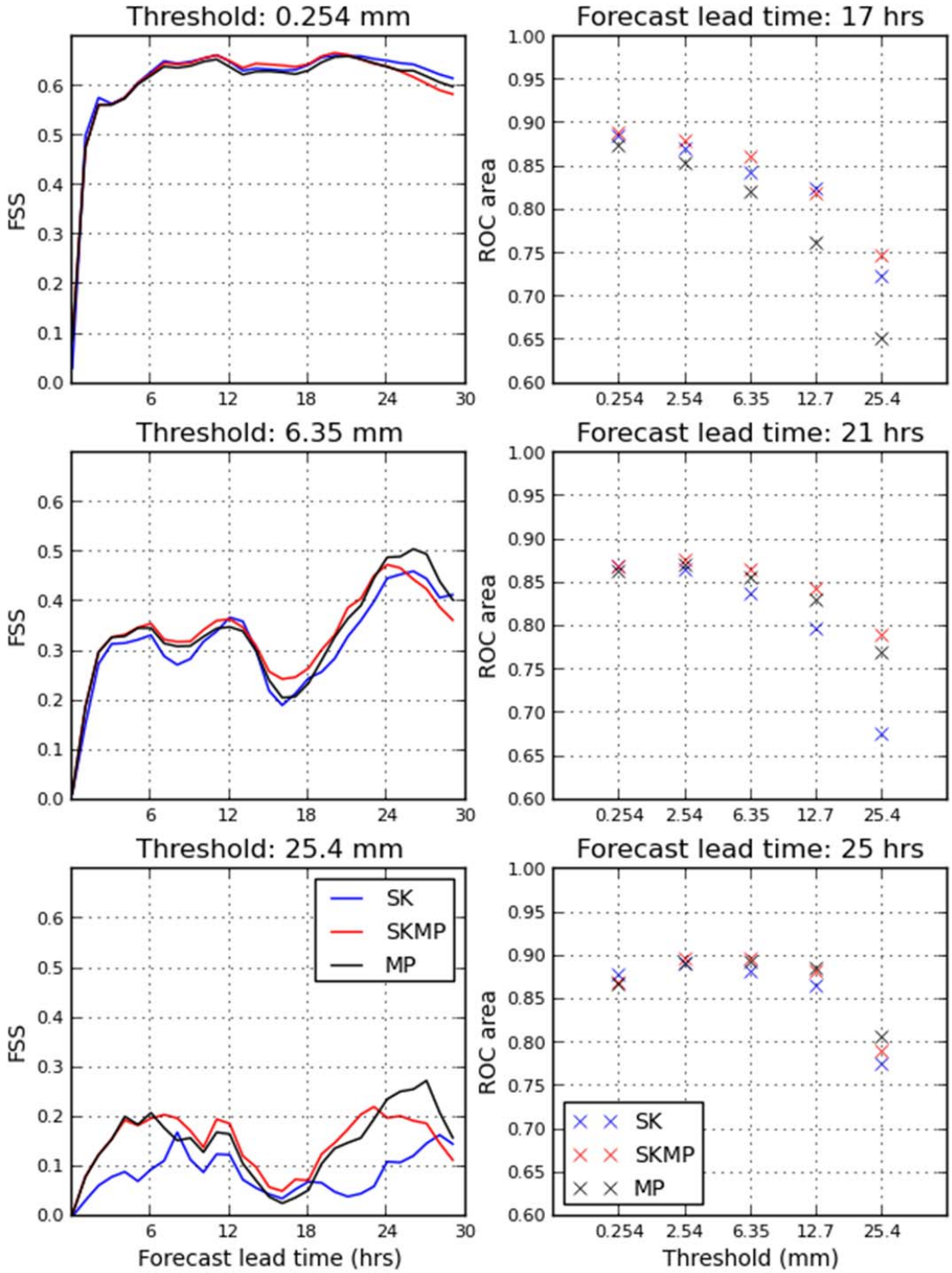

Figure 15. (left column) Fractions skill scores for the indicated 1-hr accumulation thresholds; (right column) area under the ROC curve at the indicated forecast hours for various 1-hr accumulation thresholds. No statistical significance testing was performed on the FSSs or ROC areas.