A Performance Comparison between Multiphysics and Stochastic Approaches within a North American RAP Ensemble

ISIDORA JANKOV,[a,b] JUDITH BERNER,[c] JEFFREY BECK,[a,b] HONGLI JIANG,[a,b] JOSEPH B. OLSON,[a,d] GEORG GRELL,[d] TATIANA G. SMIRNOVA,[a,d] STANLEY G. BENJAMIN,[d] AND JOHN M. BROWN[d]

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

(Manuscript received 29 April 2016, in final form 17 January 2017)

ABSTRACT

A stochastic parameter perturbation (SPP) scheme consisting of spatially and temporally varying perturbations of uncertain parameters in the Grell–Freitas convective scheme and the Mellor–Yamada–Nakanishi–Niino planetary boundary scheme was developed within the Rapid Refresh ensemble system based on the Weather Research and Forecasting Model. Alone the stochastic parameter perturbations generate insufficient spread to be an alternative to the operational configuration that utilizes combinations of multiple parameterization schemes. However, when combined with other stochastic parameterization schemes, such as the stochastic kinetic energy backscatter (SKEB) scheme or the stochastic perturbation of physics tendencies (SPPT) scheme, the stochastic ensemble system has comparable forecast performance. An additional analysis quantifies the added value of combining SPP and SPPT over an ensemble that uses SPPT only, which is generally beneficial, especially for surface variables. The ensemble combining all three stochastic methods consistently produces the best spread–skill ratio and generally outperforms the multiphysics ensemble. The results of this study indicate that using a single-physics suite ensemble together with stochastic methods is an attractive alternative to multiphysics ensembles and should be considered in the design of future high-resolution regional and global ensembles.

1. Introduction

Most global and regional numerical weather prediction (NWP) ensemble systems are underdispersive, producing unreliable and overconfident ensemble forecasts (e.g., Buizza et al. 2005; Charles and Colle 2009; Stensrud et al. 1999; Romine et al. 2014). With growing evidence that initial-condition uncertainties are not sufficient to entirely explain forecast uncertainty, the role of model uncertainty is receiving increasing attention. In the last decade, a number of different strategies have been proposed to represent uncertainty arising from model formulation. In the multiphysics approach, each ensemble member uses a different set of physics parameterizations to represent parameterized processes like convection, boundary layer, and land surface effects. While it can be challenging to find different sets of physics parameterizations that are physically consistent with each other (e.g., a land surface model that is consistent with the planetary boundary layer parameterization), multiple physics schemes introduce large diversity among the ensemble members, leading to improved forecast skill (e.g., Hacker et al. 2011b; Berner et al. 2011, 2015). While characterized by good performance, multiphysics schemes have several theoretical and practical disadvantages. For each physical process, several different parameterizations need to be developed and maintained, which is resource intensive. More importantly, and from a statistical perspective, multiphysics ensembles do not form consistent distributions, since some parameterization schemes are more closely related than others (e.g., Knutti et al. 2013). Statistical postprocessing generally assumes independent and identically distributed random variables, a requirement that is not

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

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

DOI: 10.1175/MWR-D-16-0160.1

met by multiphysics ensembles. Finally, each ensemble member has a different climatology and mean error. The fact that different members have different biases is one of the reasons why the multiphysics approach improves spread (Berner et al. 2015; Eckel and Mass 2005), but this result conflicts with the fundamental purpose of forecast uncertainty, which aims at representing the random—and not the systematic—component of forecast error.

A second avenue is to introduce ensemble spread by perturbing ensemble simulations stochastically (Palmer 2001). This method leads to statistically consistent ensemble distributions and has been successfully implemented in a number of operational weather models (e.g., Berner et al. 2009; Bowler et al. 2009; Sanchez et al. 2015). The two most popular stochastic parameterization schemes—the stochastic kinetic energy backscatter (SKEB) scheme and the stochastic perturbation of physics tendencies (SPPT) scheme—are formulated to represent unresolved subgrid-scale processes and to sample the distribution of the subgrid physics tendencies.

The SKEB scheme aims to represent model uncertainty arising from unresolved subgrid-scale processes by introducing random perturbations to streamfunction and potential temperature tendencies. SKEB is based on the rationale that a small fraction of the model dissipated energy interacts with the resolved-scale flow and acts as systematic forcing. Originally developed for large-eddy simulations (Mason and Thomson 1992), it was adapted to numerical weather prediction by Shutts (2005).

The SPPT scheme (Palmer et al. 2009) is a revision of the original stochastic diabatic tendency scheme of Buizza et al. (1999) and perturbs the parameterized tendency of physical processes with multiplicative noise. It is based on the notion that with decreasing horizontal grid spacing, the equilibrium assumption no longer holds and that the subgrid-scale state should be sampled rather than represented by the equilibrium mean. Consequently, SPPT multiplies the accumulated physical tendencies of temperature, zonal and meridional winds, and humidity at each grid point and time step with a random number. By design, the perturbations are large where the physical tendencies—and presumably their uncertainty—are large, and they have very little effect where and when the tendencies are small.

While the performance of stochastic parameterization schemes is very good (e.g., Berner et al. 2009, 2011), they have been criticized because they are added a posteriori to NWP models that have been independently developed and tuned. Ideally, stochastic perturbations should represent model uncertainty where it occurs and should be developed alongside physical parameterizations.

Therefore, the multiparameter approach addresses parameterization uncertainty at its source by perturbing the parameters in the physics parameterizations. There are two variants: the parameter can be fixed throughout the integration (e.g., Murphy et al. 2004; Hacker et al. 2011a) or the parameter can vary randomly in time and space (e.g., Bowler et al. 2009). The latter has the advantage that for sufficiently fast variations and/or sufficiently long integration, all ensemble members have the same climatology, although it is unclear whether typical NWP lead times fulfill this criterion (e.g., Berner et al. 2012). While multiparameter ensembles typically outperform unperturbed ensemble systems, they usually cannot account for all deficiencies in spread (Hacker et al. 2011b; Reynolds et al. 2011; Berner et al. 2015; Christensen et al. 2015), implying that parameter uncertainty is not the only source of model uncertainty. Multiparameter ensembles also do not perform with the same reliability as multiphysics or stochastically perturbed ensembles.

With the aim of representing uncertainty at its source, this study employs the stochastically varying parameter perturbation approach alone, and in combination with SKEB and SPPT. Parameters do not vary independently from grid point to grid point, but exhibit spatial and temporal correlations as described in detail in section 2. One motivation for this work is to consider the effectiveness of a simplified ensemble system without multiple physics parameterizations. For this purpose, a single-physics suite with stochastic perturbations is very appealing as an alternative to a multiphysics ensemble. This study addresses the question of whether an ensemble created by stochastic parameter perturbation (SPP), alone and/or in combination with SKEB and SPPT, can perform as well as—or better than—an ensemble designed with a multiphysics approach.

2. Experiment design and the stochastic perturbations scheme

The operational Rapid Refresh (RAP; Benjamin et al. 2016) configuration was used as a basis for all experiments. Simulations were performed over the operational RAP North American domain (see Fig. 2) with 13-km grid spacing and eight ensemble members. The experimental dataset consisted of 21 warm season days in 2013, including approximately every third day of the period starting with 23 May and ending on 23 July. The simulations were initialized at both 0000 and 1200 UTC with a simulation length of 24 h. In this study we used a "cold start" approach to model initialization—that is, there was no cycling or a rapid refresh component included. The first eight out of 20 perturbed members of the National Centers for Environmental Prediction (NCEP) Global Ensemble Forecast System

(GEFS; Toth and Kalnay 1993; Guan et al. 2015) provided initialization and lateral boundary conditions. The perturbed GEFS members use initial conditions obtained through an ensemble transform bred vector method that is designed to pick out the fastest-growing sources of uncertainty in the atmosphere. These perturbations also include stochastic total tendency perturbations.

The RAP system uses the Advanced Research version of Weather Research and Forecasting (WRF-ARW) dynamic core (Skamarock et al. 2008). Its physics suite includes the Grell–Freitas (GF; Grell and Freitas 2014) convective scheme, the Mellor–Yamada–Nakanishi–Niino (MYNN; Nakanishi and Niino 2004; Nakanishi and Niino 2006) planetary boundary layer (PBL) parameterization, and the Rapid Update Cycle (RUC; Smirnova et al. 2016) land surface model (LSM) parameterization. Several experiments were conducted to compare the performance between the multiphysics ensemble and the ensembles created by employing stochastic approaches.

The multiphysics ensemble, which represents the control experiment, uses different physics parameterizations for convection and the planetary boundary layer. The convective parameterizations used are Betts–Miller–Janjić (BMJ; Betts 1986; Janjć 1994), Grell–Freitas (Grell and Freitas 2014), and two versions (Han and Pan 2011) of the simplified Arakawa–Schubert (SAS) convective scheme. The different PBL schemes are the Mellor–Yamada–Janjić, Mellor–Yamada–Nakanishi–Niino (Nakanishi and Niino 2004; Nakanishi and Niino 2006), Bougeault–Lacarrère (BOULAC; Bougeault and Lacarrère 1989), and Yonsei University (YSU; Hong et al. 2006) parameterizations. The parameterizations for each ensemble member are summarized in Table 1. Member seven of the control ensemble uses the physics parameterizations in the operational RAP configuration, namely, GF and MYNN PBL.

All members of the stochastic ensembles use the same physics parameterizations as the operational RAP. The first four members (spp0–spp3) include perturbations to the GF parameterization, and the last four members (spp4–spp7) contain perturbations to the MYNN PBL parameterization (see Table 1). The spp experiment employs only the stochastic parameter perturbations. Experiment spp_skeb combines the SPP approach with SKEB (Shutts 2005; Berner et al. 2009, 2012, 2015). Similarly, experiment spp_sppt combines SPP with SPPT (Buizza et al. 1999; Bouttier et al. 2012; Berner et al. 2015). Finally, experiment spp_skeb_sppt includes all three stochastic approaches. The authors acknowledge that the two sets of spp members, with GF and MYNN perturbations, could result in different climatologies. However, after reviewing the bias characteristics for the two sets of four members and a variety of variables, no apparent difference was found.

TABLE 1. List of members for multiphysics and stochastic ensembles. OSAS = old SAS. NSAS = new SAS.

	Convective	PBL
Multiphysics members		
control0	OSAS	MYNN
control1	BMJ	MYNN
control2	GF	MYNN
control3	NSAS	MYNN
control4	GF	MYJ
control5	GF	YSU
control6	GF	BOULAC
control7	GF	MYNN
Stochastic members		
spp0	GF perturbed	MYNN
spp1	GF perturbed	MYNN
spp2	GF perturbed	MYNN
spp3	GF perturbed	MYNN
spp4	GF	MYNN perturbed
spp5	GF	MYNN perturbed
spp6	GF	MYNN perturbed
spp7	GF	MYNN perturbed

The perturbations used in SPP are introduced as

$$X^\star = [\mathbf{1} + r(\phi, \lambda, t)]X, \qquad (1)$$

where $X^\star$ is the perturbed and X is the unperturbed quantity. Here, X can be a parameter, such as the turbulent mixing length or a prognostic quantity, such as the closures in the GF scheme. If X is a static parameter, then the perturbations will reflect the uncertainty in this parameter.

The stochastic pattern generator generates a 2D perturbation field $r(\phi, \lambda, t)$ with spatial and temporal correlations analogous to that used at the European Centre for Medium-Range Weather Forecasts (ECMWF) to perturb physics tendencies (Palmer et al. 2009). The perturbation field in spectral space is expressed as

$$r(x, y, t) = \sum_{k=-K/2}^{K/2} \sum_{l=-L/2}^{L/2} r_{k,l}(t) e^{2\pi i(kx/X + ly/Y)}, \qquad (2)$$

where k and l denote the $(K+1)(L+1)$ wavenumber components, respectively, in the zonal x and meridional y directions, in physical space; and t denotes time. The Fourier modes $e^{2\pi i(kx/X+ly/Y)}$ form an orthogonal set of basis functions on the rectangular domain $0 < x < X$ and $0 < y < Y$.

Each spectral coefficient $r_{k,l}$ evolves as a first-order autoregressive (AR1) process,

$$r_{k,l}(t + \Delta t) = \alpha r_{k,l}(t) + g_{k,l}\varepsilon_{k,l}(t). \qquad (3)$$

Here, α is the linear autoregressive parameter, $g_{k,l}$ is the wavenumber-dependent noise amplitude, and $\varepsilon_{k,l}$ a

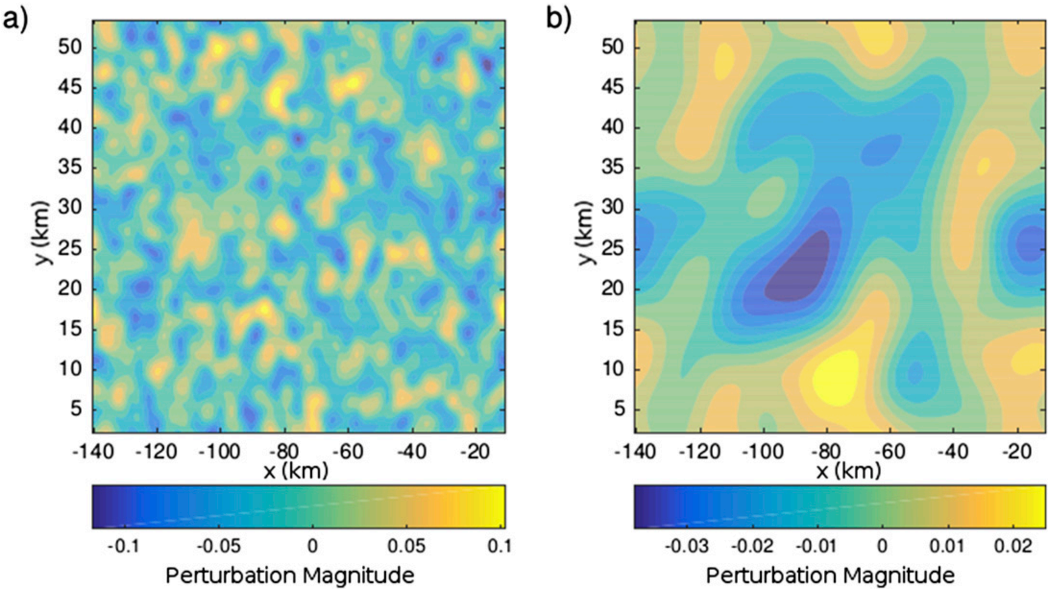

FIG. 1. The stochastic horizontal spatial stochastic perturbation pattern for the (a) GF convective parameterization and the (b) MYNN PBL parameterization. Parameters are given in Table 1.

Gaussian white-noise process with a mean of zero and a standard deviation of one. The prescribed temporal decorrelation time τ and the model time step Δt determine the autoregressive coefficient as $\alpha = \exp(-\Delta t/\tau)$. For noise amplitudes,

$$g_{k,l} = F_0 e^{-4\pi L \rho_{k,l}^2} \text{ with } F_0 = \left\{ \frac{\eta_{k,l}^2[1-(1-\alpha)^2]}{2\sum_k \sum_l e^{-8\pi\kappa\rho_{k,l}^2}} \right\}^{1/2}. \quad (4)$$

The resulting perturbation pattern will be spatially homogeneous with a horizontal length scale L and a gridpoint perturbation variance of σ^2. Here $\rho_{k,l} = \sqrt{k^2/X^2 + l^2/Y^2}$ is the effective radial wavenumber and $\eta_{k,l}^2$ is the spectral variances. The normalization constant F_0 is chosen so that the variance at any grid point, σ^2, is given by the total variance in spectral space (Weaver and Courtier 2001). At each grid point, the pattern will create perturbations drawn from a Gaussian distribution with a mean of zero and a variance of σ^2. Examples of the stochastic perturbation pattern are given in Fig. 1. For $\tau = 0$ and $\kappa = 0$, the scheme introduces noise that is white in time and space, with variance σ^2.

The pattern is fully determined by three parameters: gridpoint standard deviation (gridpt_stddev_rand_pert), length scale (length scale_rand_pert), and decorrelation time (time scale_rand_pert). Since drawing from a Gaussian distribution can result in very large values, the random numbers are capped. This capping threshold is expressed in terms of standard deviation (stddev_cutoff_rand_pert).

The perturbed parameters were chosen based on suggestions by the RAP parameterization experts and developers. Table 2 gives a summary of the targeted parameters and their perturbation amplitudes, length, and time scales.

The GF convection scheme is an ensemble scheme that offers several choices to implement stochastic perturbation methods. The ensembles are derived from the Grell–Dévényi approach, described in Grell and Dévényi (2002). One ensemble option that is used in operational applications is the dependence on the closure. The closure will determine the cloud-base mass flux, and therefore the strength and location of the convection. GF offers four choices to derive the cloud-base mass flux. While in the commonly used GF approach these are simply averaged to determine a final cloud-base mass flux, for our application we perturb each closure separately with a decorrelation time of 6 h and a spatial length scale of 150 km, before averaging to get the final cloud-base mass flux.

For the PBL scheme, three parameters were perturbed with a length scale of 700 km and a decorrelation time of 6 h. The turbulent mixing length and subgrid cloud fraction were directly perturbed, while thermal and moisture roughness lengths were perturbed indirectly through perturbations of the Zilintikevich constant Czil. The perturbations for the thermal and moisture roughness were anticorrelated with those of the mixing length and subgrid cloud fraction.

Subgrid-scale clouds (shallow cumulus clouds) are weakly positively correlated with turbulent mixing

TABLE 2. Summary of namelist parameter settings for stochastic perturbation patterns.

Perturbed parameter in GF scheme	Name
Closure	xf_ens
Namelist parameter	Value
spp_cu	1
gridpt_stddev_spp_cu	0.3
stddev_cutoff_spp_cu	3.0
length scale_spp_cu	150 000 m
time scale_spp_cu	21 600 s
Perturbed parameter in MYNN PBL scheme	**Name**
Turbulent mixing length	el
Subgrid cloud fraction (correlated with el)	cldfra_bl
Thermal and moisture roughness length (anticorrelated and twice the amplitude)	CZIL
Namelist parameter	Value
spp_pbl	1
gridpt_stddev_spp_pbl	0.15
stddev_cutoff_spp_pbl	2.0
length scale_spp_pbl	700 000 m
time scale_spp_pbl	21 600 s

(mixing lengths) during the growth of the PBL; that is, as a PBL develops, the smaller shallow cumulus become larger and deeper, but the total cloud fraction is not necessarily changed. However, when the mixing lengths are largest in the deep dry boundary layers, the cloud cover is small. So, there is a negative correlation between subgrid clouds and mixing length in fully developed PBLs. Therefore, we implemented a negative correlation, to reduce subgrid clouds when turbulent mixing (mixing lengths) become larger. Thus, more solar radiation reaches the surface, driving higher surface temperatures and higher surface heat fluxes, which then further drive the boundary layer turbulence (larger mixing lengths). The stronger turbulent mixing can result in increased entrainment at the top of the PBL, resulting in a drier PBL, which can further reduce the subgrid clouds. Tapping into this feedback is important to increasing the spread within the ensemble.

Czil was perturbed to be half/twice its original value. Since Czil is in the exponent, a halving/doubling of Czil results in a perturbation of z_t on the order of 5%–10%. The thermal roughness length z_t is defined as

$$z_t = z_0 e^{(-k\mathrm{Czil}\sqrt{\mathrm{Re}})}, \qquad (5)$$

where z_0 is the aerodynamic roughness length, k is the von Kármán constant, and Re is the Reynolds number. Perturbations of Czil are anticorrelated with mixing lengths because a reduction of Czil results in an increase in surface heat fluxes, which will drive stronger turbulent mixing (larger mixing lengths).

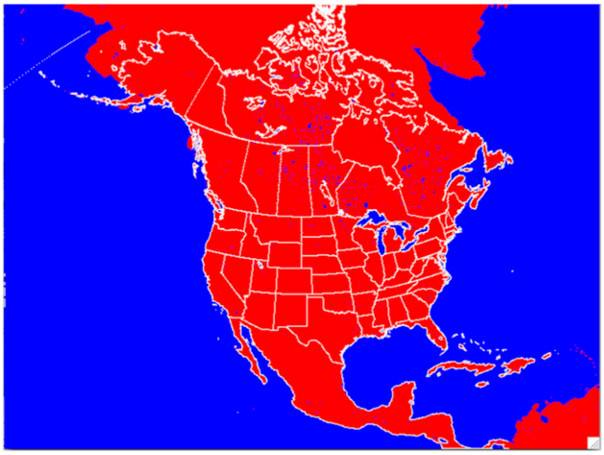

FIG. 2. Integration domain with land mask presented in red.

These choices result in perturbation patterns that are displayed for a single instance in Fig. 1 for the GF convection scheme (Fig. 1a) and the PBL scheme (Fig. 1b).

3. Results

a. Precipitation verification

All simulations were performed over the RAP North American domain (Fig. 2). Verification was performed using Model Evaluation Tools (MET; http://www.dtcenter.org/met/users/docs) software and was limited to the CONUS verification polyline applied to NCEP grid 130 (Fig. 3).

First, before performing statistical analysis, probabilities of precipitation exceeding the 50.8-mm threshold for each ensemble, for 9-h forecast and for the 1 June 2016 case, initialized at 0000 UTC (Fig. 4) were evaluated. The period was characterized with intense precipitation associated with a squall line extending from Michigan to Oklahoma.

A comparison of probabilities between the control (ctl; Fig. 4a) and spp (Fig. 4b) experiments indicated higher probabilities for spp experiments for the majority of the region characterized by precipitation exceeding the specified threshold. Higher confidence in the case of spp was associated with lower spread. This aspect was further evaluated as a part of the statistical analysis. Probabilities for the spp_skeb (Fig. 4c) experiment look very similar to the spp (Fig. 4b) experiment, implying a limited impact of SKEB on simulated precipitation. In the case of the spp_sppt experiment (Fig. 4d), the probability field was characterized with much more structure having the area of maximal probabilities along the leading edge of the system extending from northeastern corner of Oklahoma to southwestern portion of Missouri. Lower probabilities were observed for the rest

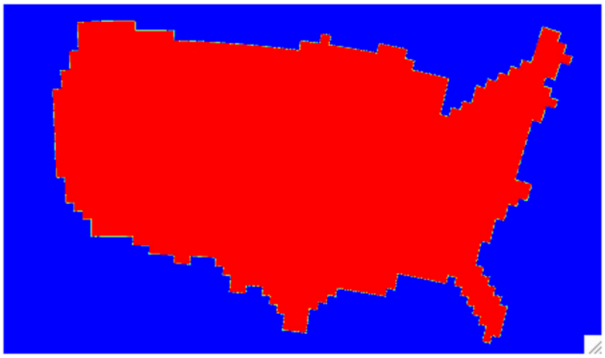

FIG. 3. Verification domain with land mask presented in red.

of the region in Missouri. Lower confidence in the case of spp_sppt may be associated with higher spread when compared to the ctl experiment. Finally, Fig. 4e shows probabilities for the spp_skeb_sppt experiment that are very similar to the spp_sppt (Fig. 4d) probabilities, implying once again the limited impact of SKEB on simulated precipitation. Figure 4f shows observed radar reflectivity valid at the same time.

Next, we focus on the verification of precipitation forecasts by comparing the ensemble output over the CONUS domain to NCEP national stage IV analyses (Lin and Mitchell 2005). Both ensemble output and stage IV data were regridded to a common 13-km NCEP grid (130). Verification was performed only over land and areas where stage IV data were available.

The performance of the five ensemble systems is assessed by examining rank histograms (Hamill 2001), ensemble mean frequency biases, and corresponding Gilbert skill scores (Schaefer 1990) and Brier scores (Brier 1950; Wilks 2011).

The rank histogram is a diagnostic tool that facilitates assessing the spread of ensemble forecasts, based on the assumption that the probability of occurrence of an observation in each of a set of forecast bins should be equally likely (Hamill 2001). These bins are determined by ranking member forecasts from the lowest to the highest value. For an ensemble with n members, the corresponding rank histogram will have $n + 1$ bins. The rank histogram is produced by plotting the frequency of occurrence of observations in each bin. Flat rank histograms indicate an ensemble with sufficient spread.

Figure 5 is a rank histogram for 6-hourly precipitation accumulation periods for 6-, 12-, 18-, and 24-h lead times for the five experiments and for the 0000 UTC initialization. The rank histograms for the 1200 UTC initialization indicated very similar performance (not shown). All ensembles and all aggregation times were characterized with a precipitation bias that increased with lead time. Also, the U-shape histograms, especially for the first two 6-hourly accumulations, suggest underdispersion in all evaluated ensembles.

To evaluate precipitation frequency bias and its diurnal change, the frequency bias was calculated for both initialization times, for three precipitation thresholds (0.254, 6.35, 12.7 mm), and for all members of the five experiments. The frequency bias is calculated as a ratio between forecast and observed precipitation coverage greater than the specified threshold. It can vary from zero to infinity. Values of the frequency bias significantly higher (lower) than one indicate that the model notably overpredicted (underpredicted) the exceedance of a given threshold.

Calculation of the frequency bias showed similar behavior between members of the four stochastic experiments (spp, spp_skeb, spp_sppt, and spp_skeb_sppt). All stochastic experiments were characterized by smaller spread when compared to the control experiment. Figure 6 illustrates the frequency bias as a function of lead time for the 0000 UTC initialization, for all eight members of the ctl and spp experiments only. For the lowest precipitation threshold of 0.254 mm, a positive frequency bias was found that increased with lead time for all members and for both experiments (Fig. 6a). In contrast, for higher precipitation thresholds (6.35 and 12.7 mm shown in Figs. 6b and 6c, respectively), frequency bias values were approximately constant with lead time and generally low (less than one) for the two ensembles.

The statistical significance of these results was assessed by employing a bootstrapping method with 1000 replications and a 95% confidence interval. Differences in the performance of two selected experiments and their members were found to be statistically significant for thresholds and lead times when pairwise differences were nonzero.

Differences in frequency bias values between the corresponding ctl and spp members (e.g., member 1 of one ensemble was compared to member 1 of the other ensemble) were statistically significant for all lead times and for the 0.254- and 6.35-mm precipitation thresholds. Results from the 1200 UTC initialization were similar (not shown).

In addition, we examined values of the Gilbert skill score (GSS) for all ensemble members. The Gilbert skill score was calculated following the equation

$$\mathrm{GSS} = \frac{\mathrm{CFA} - \mathrm{CHA}}{(F + O + \mathrm{CFA} - \mathrm{CHA})}, \quad \mathrm{CHA} = O\frac{F}{V}, \qquad (6)$$

where CFA, F, and O indicate the number of grid points at which a variable was correctly forecasted to exceed the specified threshold (CFA), a variable was forecasted to exceed the threshold (F), and a variable was observed to exceed the threshold (O); CHA is the probability

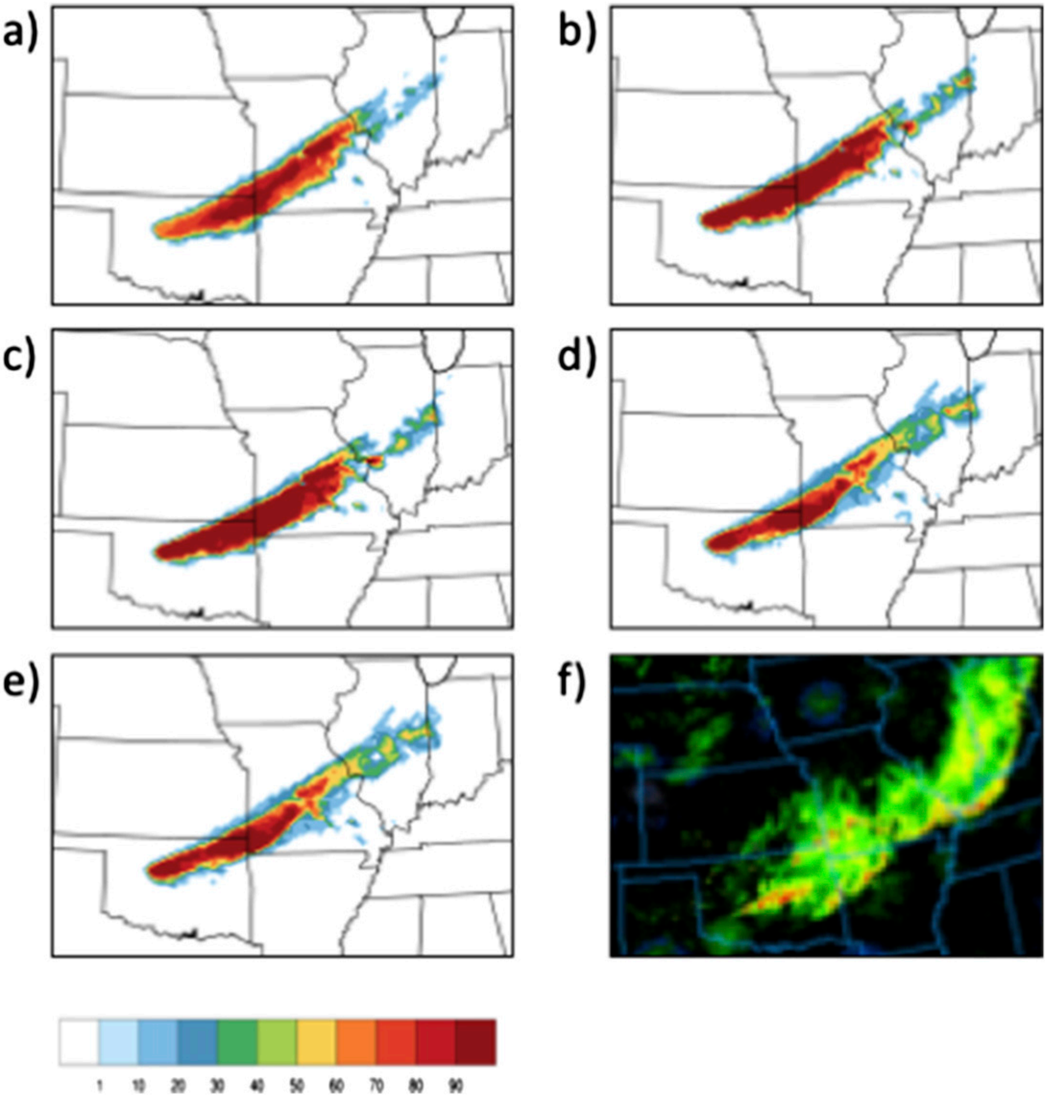

FIG. 4. (a) Probability of precipitation exceeding 50.8 mm in 3 h for the (a) ctl, (b) spp, (c) spp_skeb, (d) spp_sppt, and (e) spp_skeb_sppt experiments for the 9-h forecast, initialized at 0000 UTC valid at 0900 UTC 1 Jun 2013. (f) Radar reflectivity valid at the same time.

that a correct forecast would occur by chance, and V is the total number of evaluated grid points. A GSS of one would occur with a perfect forecast. Basically, the GSS measures how well events were predicted, taking hits associated with random chance into account.

The GSSs were computed by utilizing aggregated contingency table elements. Similarly as for frequency bias, Fig. 7 shows GSS for three different precipitation thresholds and for each member of the ctl and spp experiments only. Statistical significance testing was performed by employing a resampling technique (Hamill 1999) with 1000 replications and a significance level of 0.05.

It can be seen that the largest difference in performance between ctl and spp members was detected for the lowest precipitation threshold evaluated. For this precipitation threshold and for shorter lead times, the two experiments performed similarly. For lead times longer than 18 h, the spp ensemble was characterized with a larger number of members with higher GSS values. GSS values for the precipitation threshold larger than 6.35 mm were very similar between the experiments. None of the differences for all three precipitation thresholds were found to be statistically significant.

In addition to evaluating the ensemble mean frequency bias and GSS as measures of deterministic performance, the Brier score (BS) was used as a measure of probabilistic performance. The Brier score (Brier 1950) is a commonly used verification measure of the skill of

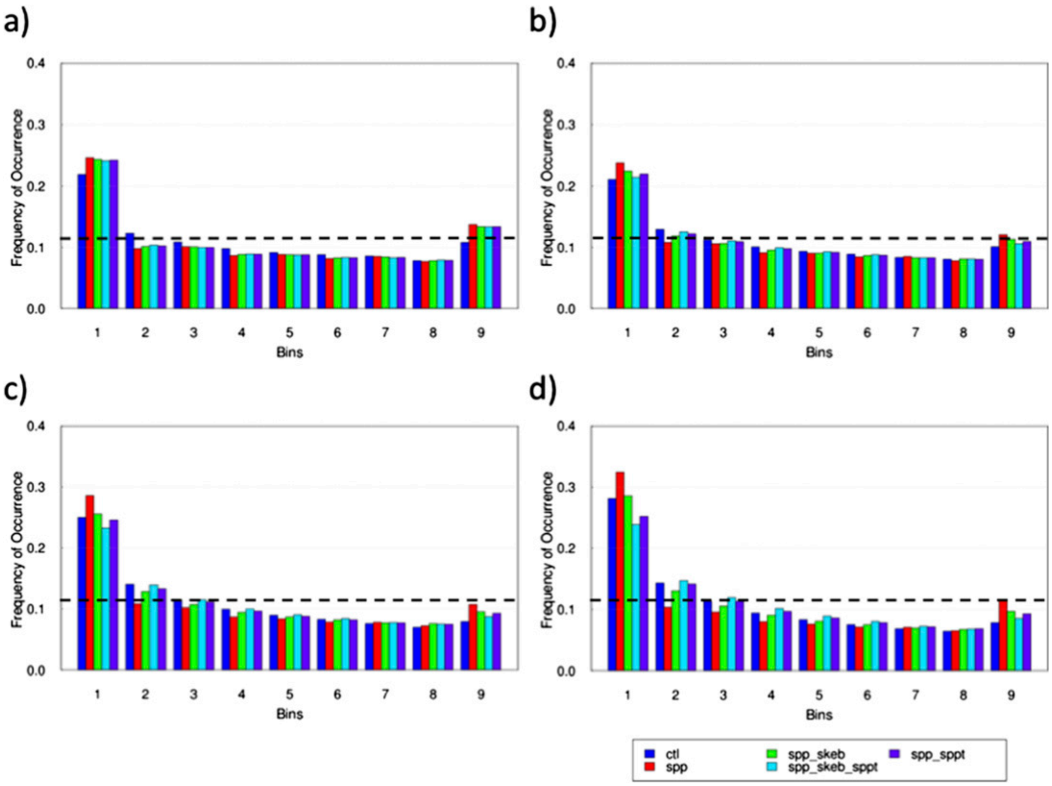

FIG. 5. Rank histogram for 0000 UTC initialization and 6-hourly accumulation periods valid at forecast hours (a) 6, (b) 12, (c) 18, and (d) 24.

probabilistic forecasts. For an event of interest, $X_e(j, k) = [X_1(j, k), \ldots, X_n(j, k)]$ is an n-member ensemble forecast for the jth of m locations and the kth of r case days, arranged from lowest to highest forecast value. This ranked ensemble is then converted into an n-member binary forecast $I_e(j, k) = [I_1(j, k), \ldots, I_n(j, k)]$ indicating whether the event was forecast (=1) or not forecast (=0) for each ensemble member. Similarly, the observed weather, $I_o(j, k)$, is converted to a binary outcome. Assuming that each member forecast is equally likely, a forecast probability $p_f(j, k)$ is calculated as

$$p_f(j,k) = \frac{\sum_{i=1}^{n} I_i(j,k)}{n}. \tag{7}$$

Then the BS of the forecast is calculated as

$$\mathrm{BS} = \sum_{k=1}^{r} \sum_{j=1}^{m} p_f(j,k) - I_o(j,k)^2. \tag{8}$$

In this study, the BS is examined for all experiments as a function of lead time for both 0000 (Fig. 8) and 1200 UTC (Fig. 9) initializations and three precipitation thresholds (Figs. 8a–c and 9a–c). This score is negatively oriented, so that lower values denote better forecast skill. Figures 8 and 9 show a clear indication of the Brier score change associated with the diurnal cycle. There is evidence that BS values increased during the day and then decreased later in the afternoon through nighttime. Overall, BS values decreased with increasing precipitation threshold. This is most likely due to large regions without intense rainfall in forecasts, rather than due to increased skill for higher precipitation thresholds.

For 0000 UTC initializations and the lowest precipitation threshold (Fig. 8a), the spp ensemble performed better than the rest of stochastic ensembles and the control during the night and early morning hours (3–15-h lead times). The opposite was true for lead times beyond 15 h. During this period, ensembles that combined stochastic approaches (spp_skeb, spp_sppt and spp_skeb_sptt) performed notably better when compared to the control and the spp ensembles. For the 2.54-mm threshold (Fig. 8b), the control ensemble outperformed all other experiments, but the differences in BS were not statistically significant.

For the 1200 UTC initialization and the 0.254-mm threshold (Fig. 9a), ensembles spp_sppt and spp_skeb_sppt were characterized by significantly lower BS when compared to the control ensemble for lead times between 6 and 12 h. For lead times beyond 12 h, the control

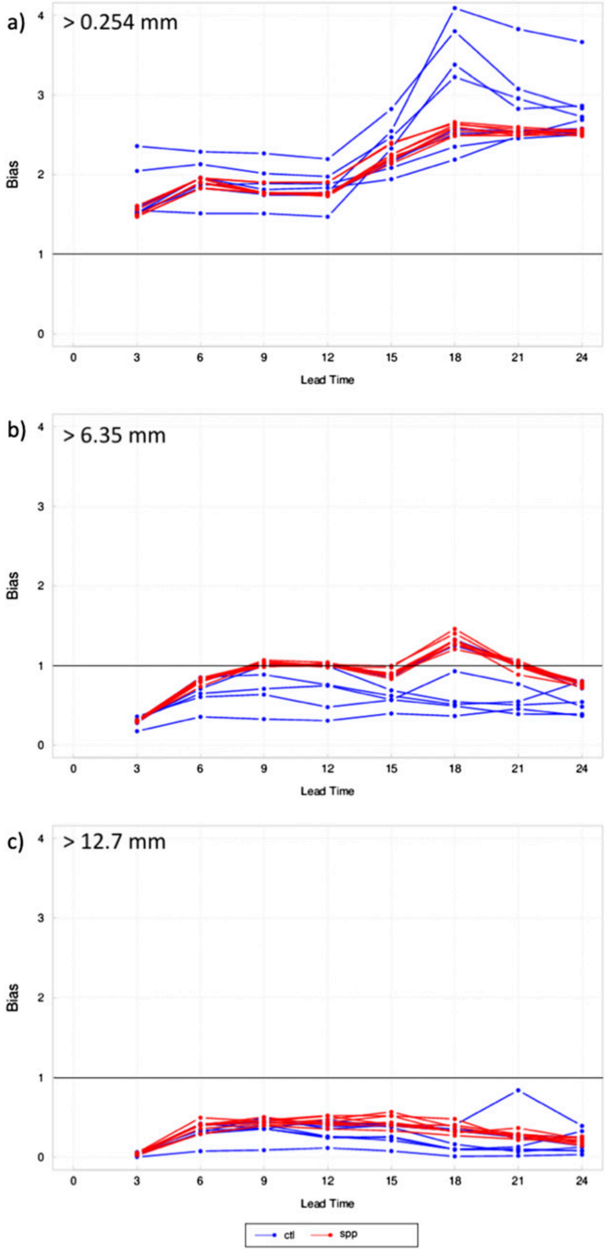

FIG. 6. Frequency bias for 0000 UTC initialization and for three precipitation thresholds: (a) 0.254, (b) 6.35, and (c) 12.7 mm for the ctl and spp experiments.

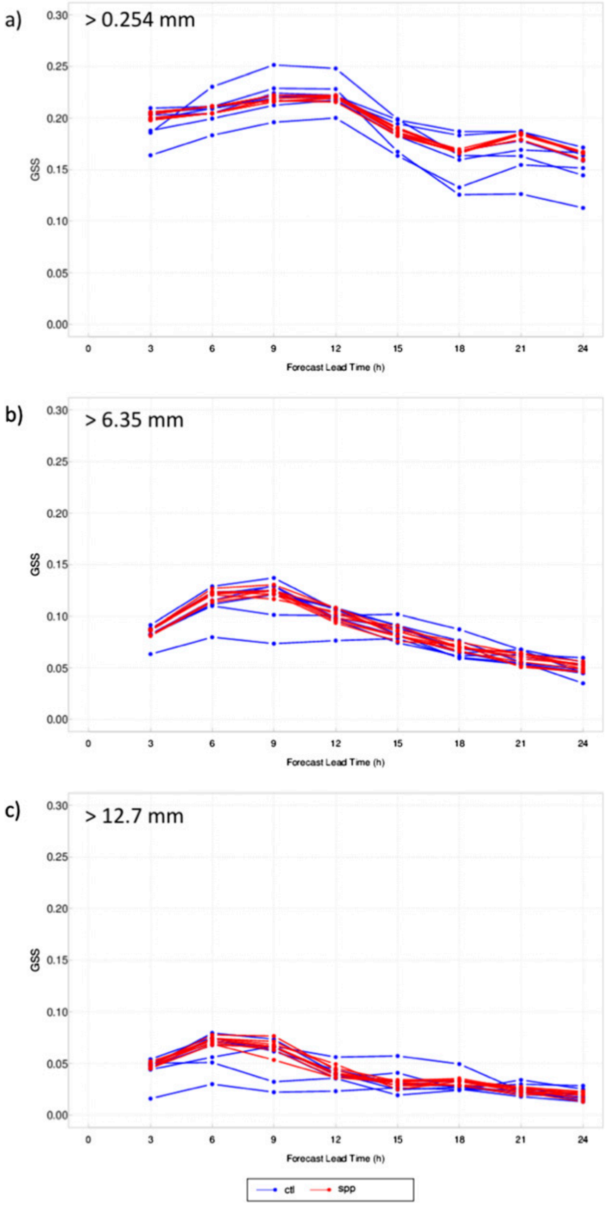

FIG. 7. GSS for 0000 UTC initialization and for three precipitation thresholds: (a) 0.254, (b) 6.35, and (c) 12.7 mm for the ctl and spp experiments.

ensemble was characterized by significantly lower BS when compared to all other experiments. Similarly to the 0000 UTC initialization, for a slightly higher threshold (Fig. 9b), the control ensemble had the lowest BS when compared to all other experiments, but the differences were not statistically significant.

The Brier skill score (BSS) measures the difference in skill of a forecast and a skill of climatology normalized by the improvement that can be achieved. Negative values of BSS indicate that the forecast is less skilled when compared to climatology. In this study the sample climatology was utilized. For all ensembles and for all evaluated precipitation thresholds, the BSS was slightly negative (not shown). This behavior was most likely a product of small ensemble size (Müller et al. 2005) and overall low reliability for all evaluated precipitation thresholds (Fig. 10). The reliability diagrams were created for all lead times (every 3 h) aggregated together. For construction of reliability diagrams, four forecast probability bins were used: 0.25, 0.5, 0.75, and 1.

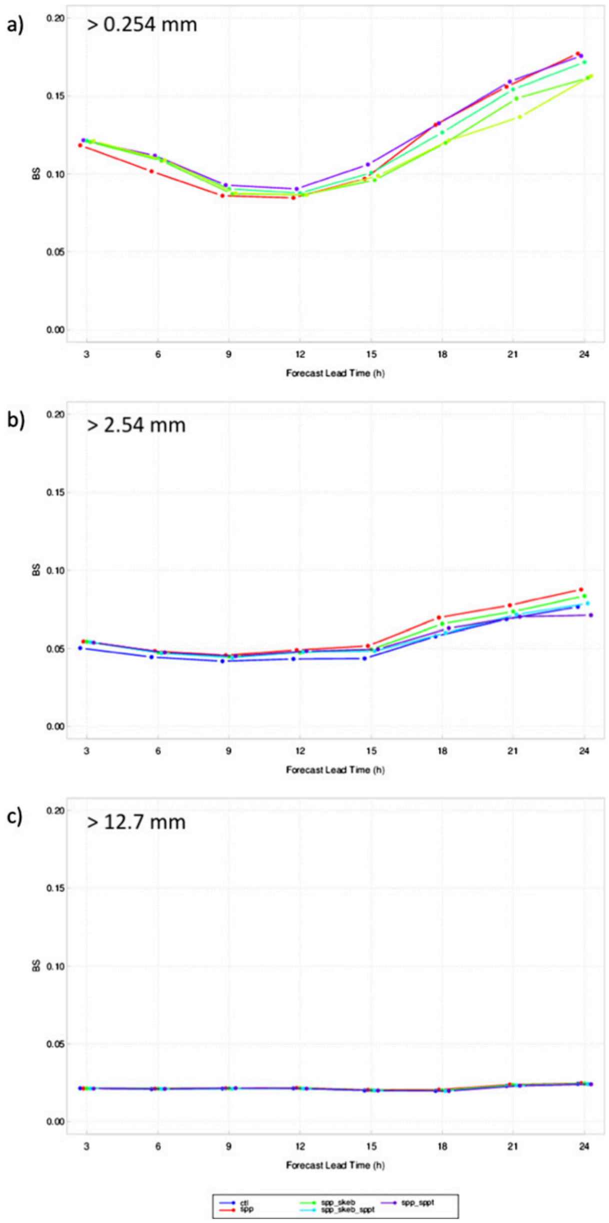

FIG. 8. BS for 0000 UTC initialization and for three precipitation thresholds: (a) 0.254, (b) 2.54, and (c) 12.7 mm for all experiments.

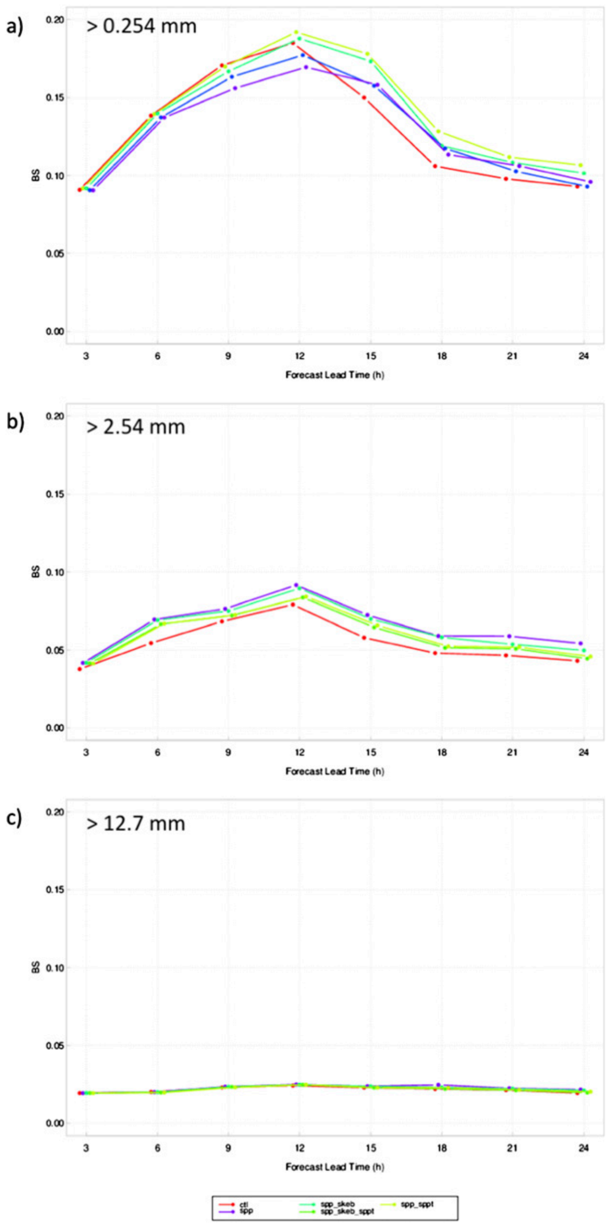

FIG. 9. Brier score for 1200 UTC initialization and for three precipitation thresholds: (a) 0.254, (b) 2.54, and (c) 12.7 mm for all experiments.

Figure 10 show overconfidence for all experiments and for all evaluated precipitation thresholds. For heavier thresholds the control ensemble was characterized with somewhat higher reliability compared to other ensembles (Figs. 10b and 10c).

In summary, the rank histograms indicate that all ensembles, in particular spp, are underdispersive. Yet when combining SPP with other stochastic methods, the resulting rank histograms are comparable to those of the control (multiphysics ensemble). All ensembles were characterized by a high-frequency bias for low thresholds and a low-frequency bias for higher thresholds. The spp ensemble had a frequency bias closest to one.

In terms of BS and for lighter precipitation thresholds, there is a tendency for the control ensemble to outperform stochastic ensembles, but in most cases, the differences in performance were not statistically significant. For higher precipitation thresholds, BS values were smaller for all ensembles, mainly due to limited areas with heavy rainfall. All experiments were characterized

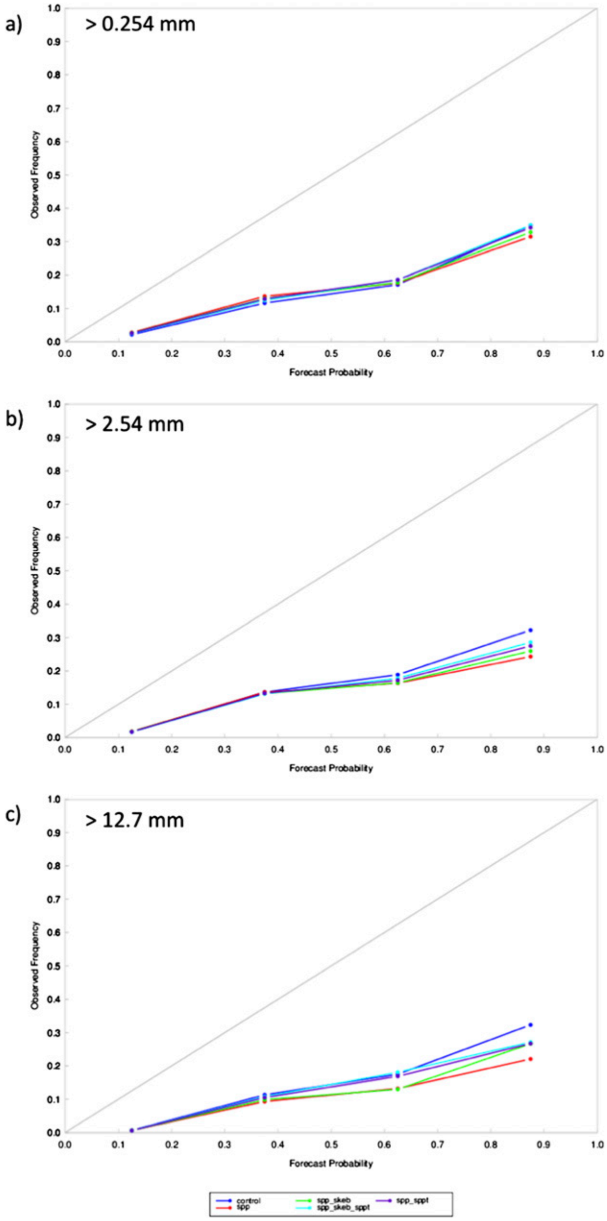

FIG. 10. Reliability diagrams for 0000 UTC initialization and for three precipitation thresholds: (a) 0.254, (b) 2.54, and (c) 12.7 mm.

with generally low reliability. For heavier thresholds, the ctl experiment was characterized with slightly higher reliability compared to the other experiments.

b. Surface and upper-air verification

In addition to precipitation, forecasts of variables such as 2-m temperature, 10-m wind, 850-hPa temperature, 250-hPa wind, and 500-hPa geopotential height were also assessed. Forecasts were compared to the North American Mesoscale Forecast System (NAM) Data Assimilation (NDAS) analysis. For this purpose, both forecasts and the analysis were interpolated to the 13-km NCEP grid (130), covering the CONUS domain (Fig. 3). First, root-mean-square error (RMSE) values of the ensemble mean and corresponding spread values were computed for all five ensembles. The spread was computed as the average ensemble standard deviation over the CONUS domain.

Figure 11 illustrates RMSE and spread values for runs initialized at 0000 UTC. Overall, the control ensemble has slightly lower RMSE values compared to the stochastic ensembles. For 2-m temperature, the control ensemble is characterized by slightly lower values of RMSE for all lead times (Fig. 11a). The same is true for 850-hPa temperature (Fig. 11c) and for the 10-m zonal component of the wind (Fig. 11e). The difference between ensembles is minimal in the case of the 250-hPa zonal wind component (Fig. 11g). For 500-hPa geopotential height, the spp_skeb ensemble, the combination of all three stochastic approaches (spp_skeb_sppt), and the control ensemble result in higher RMSE values compared to the other stochastic ensembles for longer lead times (Fig. 11i). None of the differences in RMSE between the experiments and control, for any of the variables, are statistically significant (Fig. 11).

Next, we focus on the corresponding spread values, which vary widely across the five ensembles. For all variables and all lead times, the spp ensemble is characterized by spread values well below their corresponding RMSE values. This type of behavior, insufficient spread in spp, has been documented by Hacker et al. (2011a). Combining the SPP approach with SKEB and SPPT substantially increases the spread. Interestingly, the different combinations of stochastic approaches have distinct signatures. For example, the combination of SPP with SPPT (spp_sppt) results in a notable increase in spread for the zonal component of the 10-m wind (Fig. 11f), while the combination with SKEB (spp_skeb) predominantly increased spread for 500-hPa height (Fig. 11j).

The perturbations of SPPT are proportional to the magnitude of the physical tendencies, which tend to be largest near the surface, whereas the dynamic perturbations introduced by SKEB grow fastest on the synoptic scale and in the free atmosphere, consistent with the findings of Berner et al. (2011, 2015). Importantly, the combination of all three stochastic approaches resulted in the largest spread, except in the case of 10-m zonal wind (Fig. 11f) for which spp_skeb_sppt had the same amount of spread as the spp_sppt ensemble for 12- and 18-h lead times and slightly lower spread for a 24-h lead time.

For all other variables and for most lead times (especially longer than 6 h), spp_skeb_sppt was characterized by a significantly larger spread when compared to the control ensemble. Very similar behavior was observed for

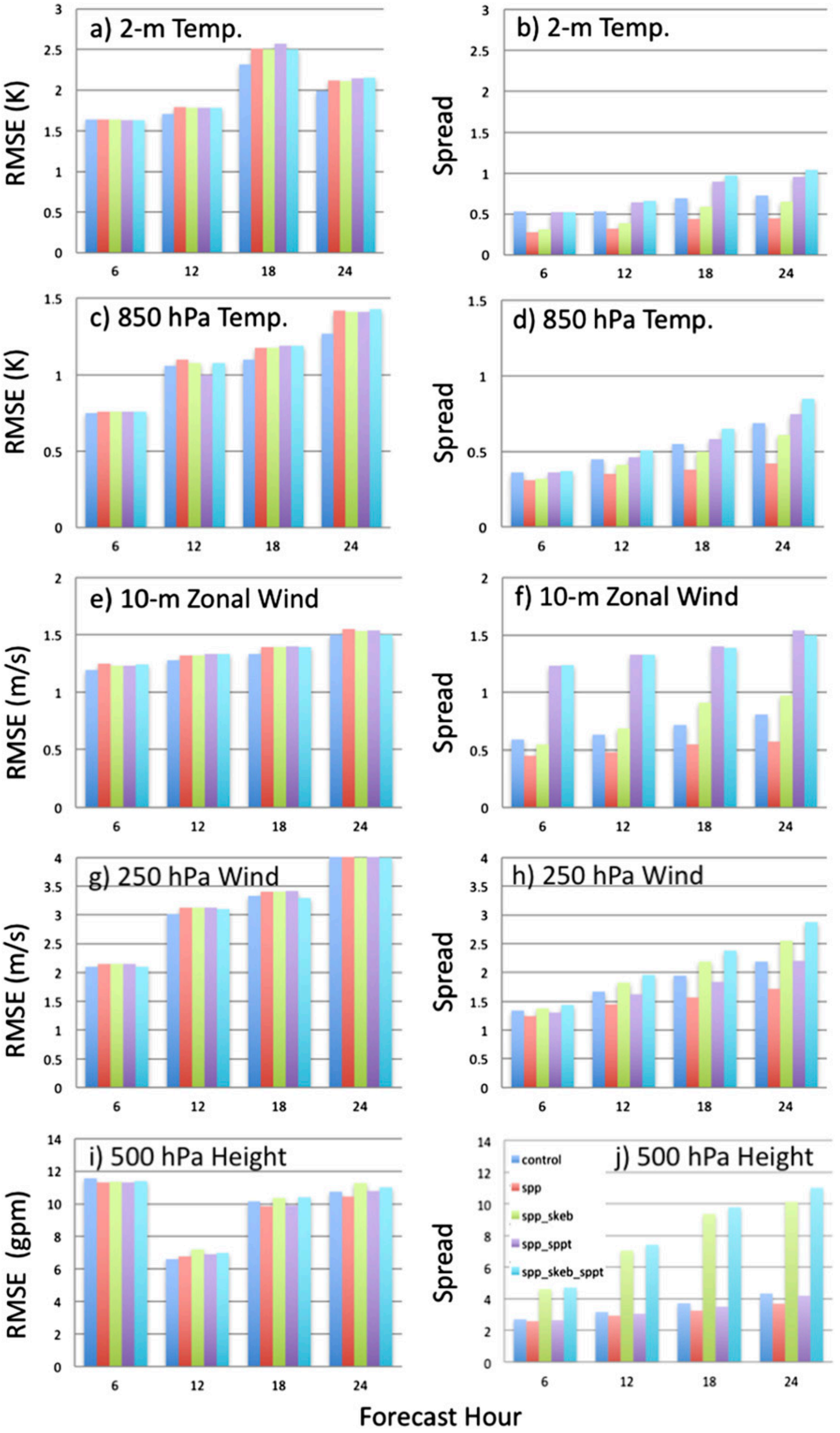

FIG. 11. (left) RMSE and (right) corresponding spread for (a),(b) 2-m temperature; (c),(d) 850-hPa temperature; (e),(f) 10-m zonal wind; (g),(h) 250-hPa wind; and (i),(j) 500-hPa geopotential height for simulations initialized at 0000 UTC.

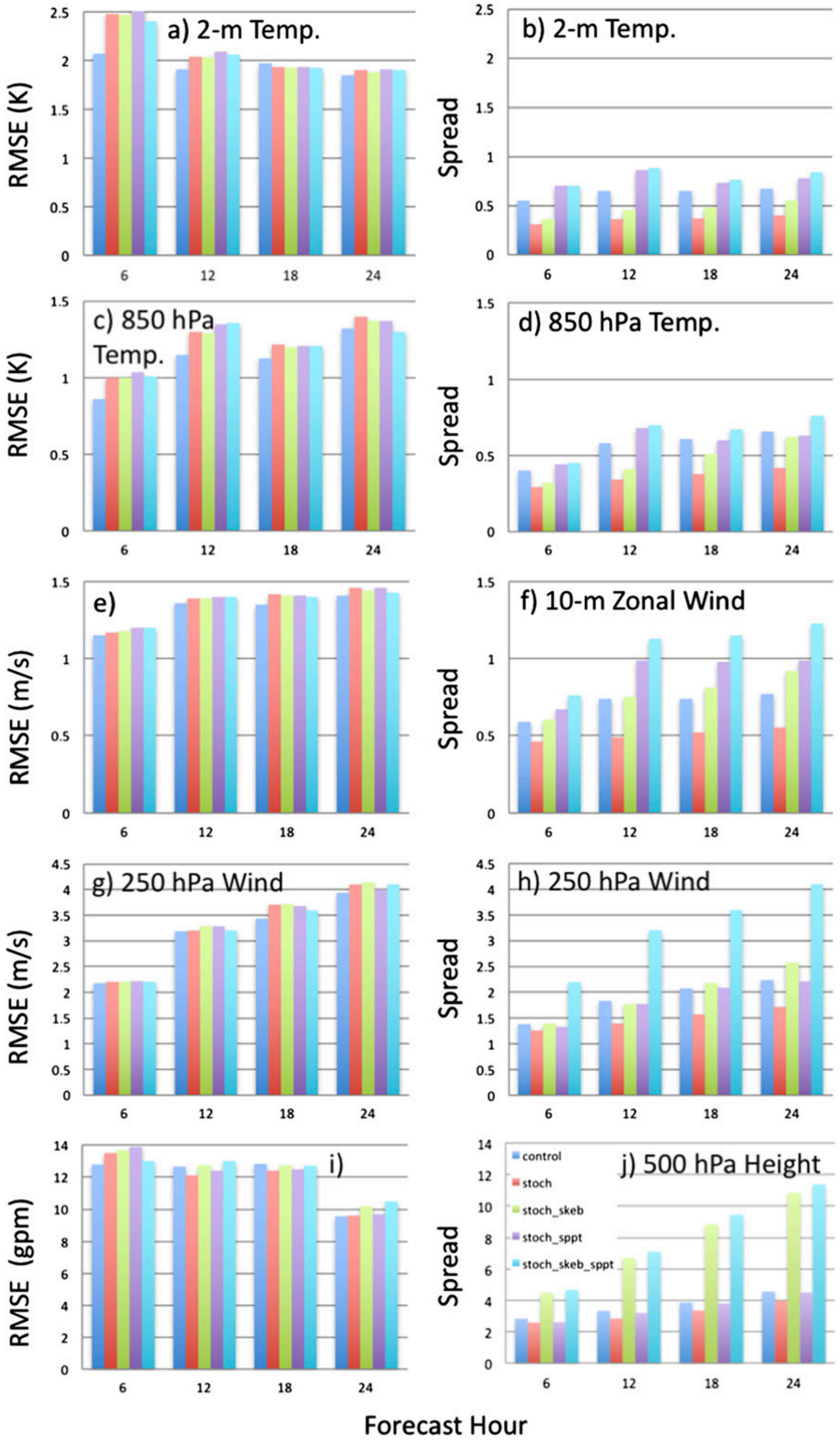

FIG. 12. As in Fig. 11, but for 1200 UTC initializations.

runs initialized at 1200 UTC (Fig. 12) in terms of both RMSE and spread.

The ratio between spread and error for the experimental ensembles is presented in Fig. 13. The ideal value for this ratio is one. For both initialization times, a smaller spread in the SPP ensemble leads to lower spread–error ratios when compared to other experiments. For lead times longer than 6 h, the ensemble that combined the

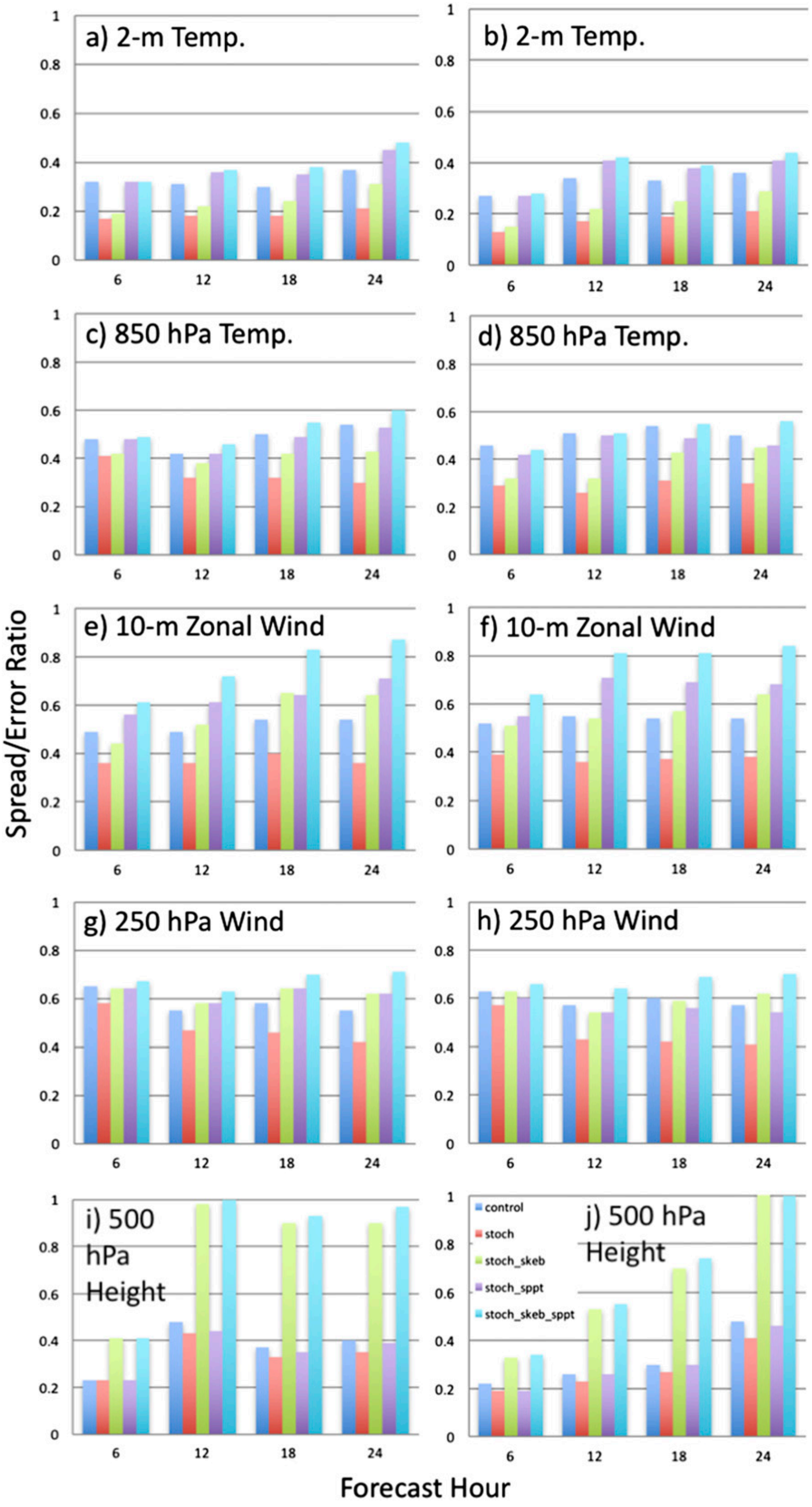

FIG. 13. Spread–error ratio for (a),(b) 2-m temperature; (c),(d) 850-hPa temperature; (e),(f) 10-m zonal wind component; (g),(h) 250-hPa zonal wind component; and (i),(j) 500-hPa geopotential height for (left) 0000 and (right) 1200 UTC initializations.

three stochastic approaches (spp_skeb_sppt) had the highest spread–error ratio (closest to one) for all variables and both initialization times.

As an additional measure of ensemble performance, the continuous ranked probability score (CRPS; Hersbach 2000; Candille and Talagrand 2005) was also computed (Fig. 14). This score is oriented such that a lower value denotes better forecast skill. If F is the cumulative distribution function of the forecast distribution and x verifies, then the CRPS is defined as

$$\mathrm{CRPS}(F,x)=\int_{-\infty}^{+\infty}[F(y)-1\{y\geq x\}]^2\,dy, \qquad (9)$$

where $1\{y\geq x\}$ denotes a step function along the line that attains the value 1 if $y\geq x$ and the value 0 otherwise.

For most of the variables, with the exception of 500-hPa geopotential height, spp has the highest CRPS (lowest skill) compared to the other ensembles (Fig. 14). For most of the variables and most lead times beyond 6 h, spp_skeb_sppt has the best forecast skill, even though differences in CRPS between spp_skeb_sppt and the control ensembles were not statistically significant. Exceptions are 2-m temperature for the 1200 UTC initialization at most lead times and 850-hPa temperature at 12- and 18-h lead times, for which the control ensemble had an advantage compared to the spp_skeb_sppt ensemble. For the 10-m zonal wind component, there are instances when the two ensembles have identical CRPS values [e.g., 0000 UTC initializations at 12- and 18-h lead times (Fig. 14e), and 6-, 12-, and 18-h lead times for the 1200 UTC initialization (Fig. 14f)]. For 500-hPa geopotential height, spp_skeb_sppt outperformed the control ensemble for most of the lead times and both initialization times (Figs. 14i and 14j).

In summary, for surface and upper-air variables, the RMSE is overall similar for all ensembles. The spread however varies widely among the experiments, with spp generally containing the least spread. When SPP is combined with other stochastic methods (SKEB and SPPT), the spread is increased with statistical significance. Therefore, the spread–error ratios are consistently larger for ensembles that combined multiple stochastic approaches (spp_skeb, spp_sppt, and spp_skeb_sppt). For most lead times and all variables, spread–error ratios closest to one were produced by the ensemble that combined all stochastic approaches (spp_skeb_sppt).

Inclusion of observational error into ensemble evaluation has been shown to affect the verification of short-term simulations (Bouttier et al. (2012)). Unfortunately, the MET software and the Environmental Modeling Center (EMC) verification system employed in this study do not have an option to take observational error into account. For a limited number of variables (not shown), we employed an additional verification method that takes observational error into account and the results showed a positive impact on spread. This result is consistent with Candille and Talagrand (2008).

c. *Added value of the stochastic parameter perturbation approach*

While the SPP approach showed some promise, especially for precipitation simulations, its general performance is hampered by limited spread. To determine whether this scheme adds value in combination with other spread-generating approaches, we assessed the performance of the spp_sppt experiment against that of using only SPPT (sppt). Figure 15 shows the percentage difference of spp over spp_sppt for a number of verification metrics. The bars in Fig. 15 are plotted only for variables, lead times, and statistics for which the difference was nonzero.

For the 0000 UTC initialization, the use of SPP results in a reduction of RMSE for most of the variables and all lead times except 18 h (Fig. 15a). The impact of SPP on RMSE for the 1200 UTC initialization is mixed (Fig. 15e). It tends to be beneficial for near-surface variable but detrimental for 250-hPa zonal wind. A general positive impact of SPP is shown in spread values (Figs. 15b and 15f). Consequently, the spread–error ratio increased and is closer to one (Figs. 15c and 15g). For the 0000 and 1200 UTC initializations, the CRPS decreased for almost all variables and lead times (Fig. 15d) and for near-surface variables (Fig. 15f), signifying added benefit from using SPP in addition to SPPT. In summary, the analysis shows that the use of SPP generally has a positive impact. This benefit is most evident in increased spread, which consequently improves the spread–error ratio. The impact on CRPS was overall beneficial (decreased values), with the exception of upper-air variables in the 1200 UTC model runs.

4. Summary and conclusions

There is strong evidence that multiphysics approaches to ensemble forecasting increase spread and improve ensemble scores relative to typically underdispersive single-physics ensembles. However, there are a number of concerns and disadvantages with the multiphysics approach related to code maintenance, consistency across parameterizations, and systematic error. This study explores whether a single-physics ensemble, employing a suite of stochastic methods, can be a viable alternative to the multiphysics approach.

To this end, a new stochastic parameter perturbation (SPP) scheme was developed and compared alone and in combination with other stochastic methods (SKEB

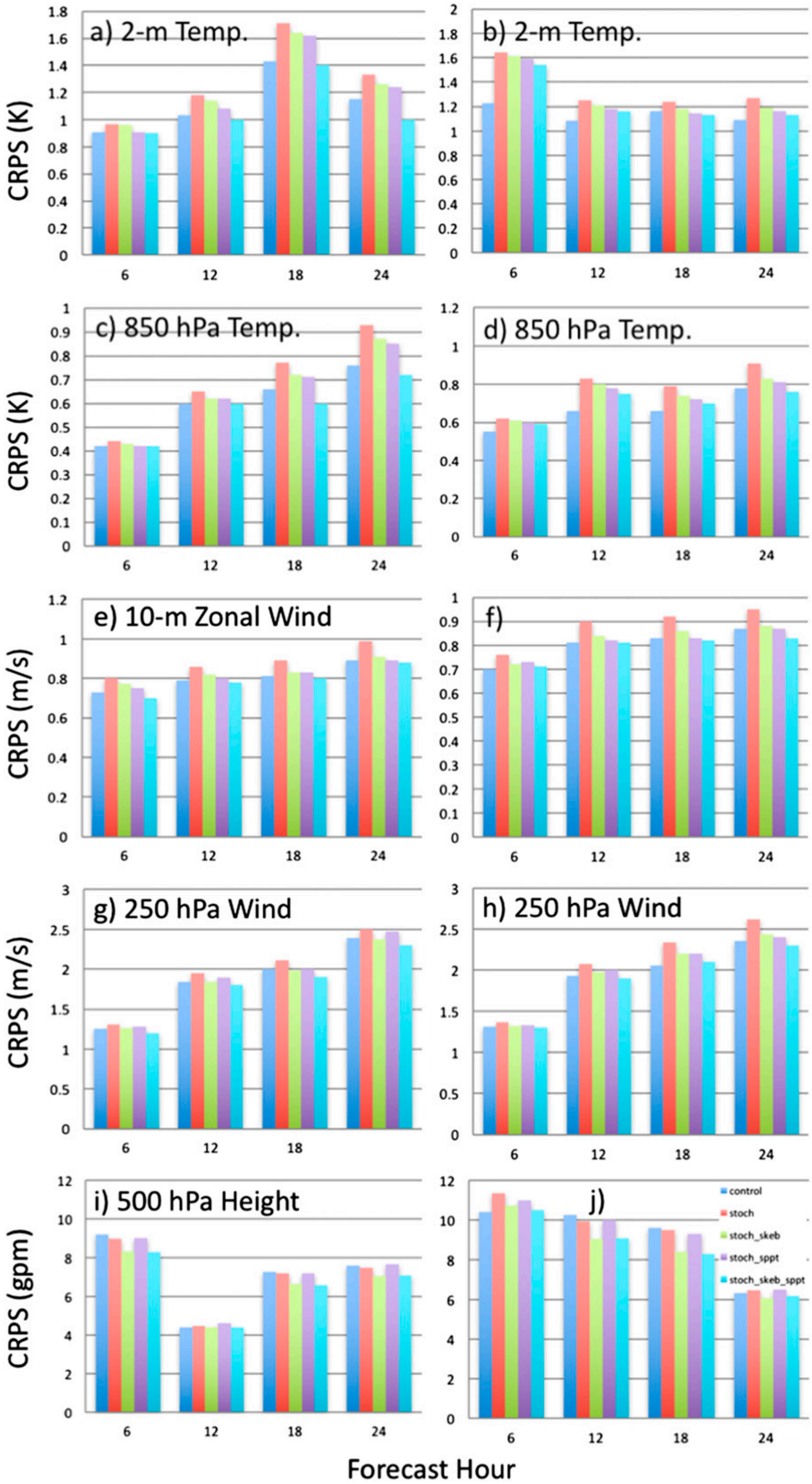

FIG. 14. CRPS for (a),(b) 2-m temperature; (c),(d) 850-hPa temperature; (e),(f)10-m zonal wind component; (g),(h) 250-hPa zonal wind component; and (i),(j) 500-hPa geopotential height for (left) 0000 and (right) 1200 UTC initializations.

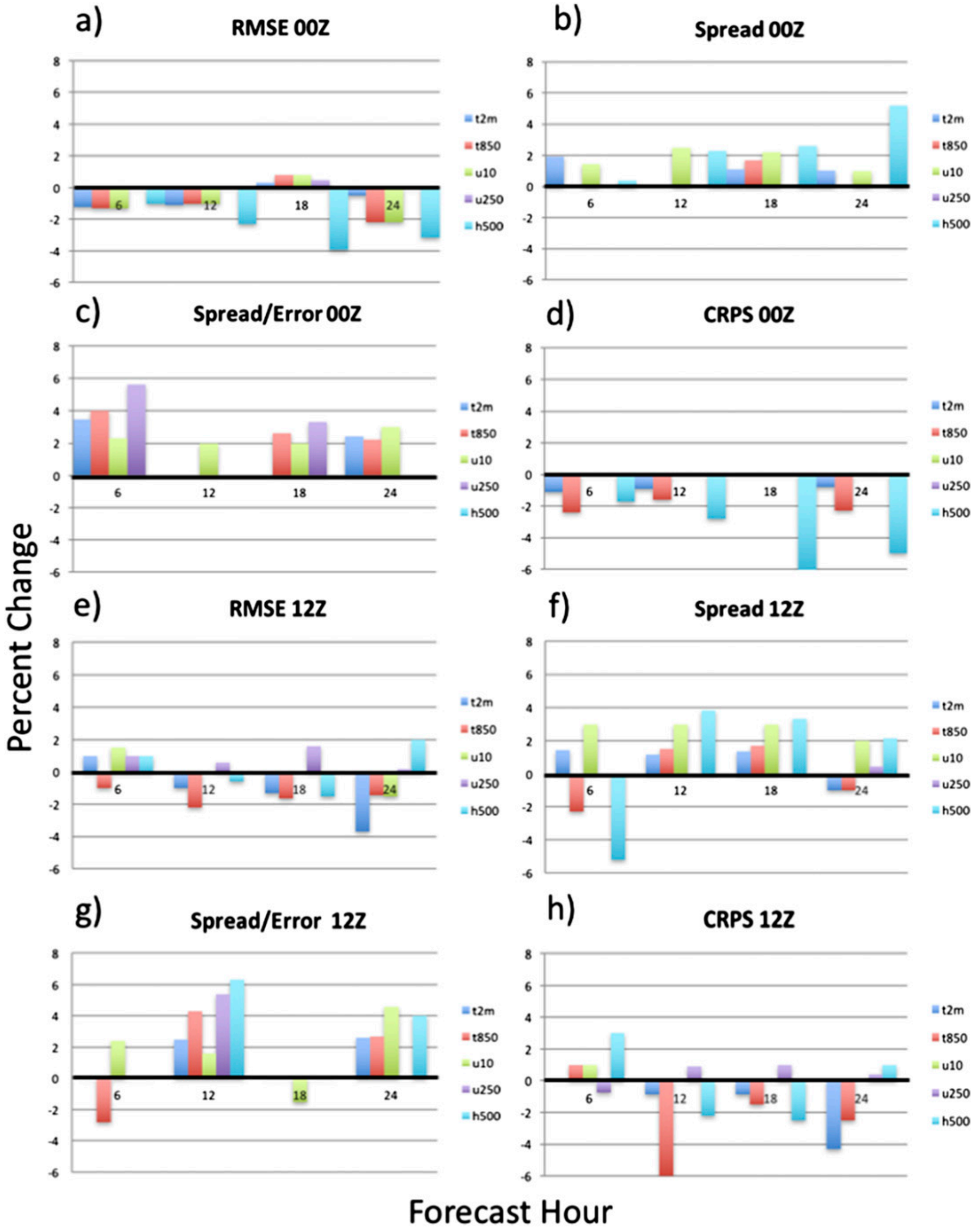

FIG. 15. Percentage change of the SPP over the spp_sppt ensemble in (a) RMSE, (b) spread, (c) spread–error ratio, and (d) CRPS for 0000 UTC initialization. Similarly, percentage change of the SPP over the spp_sppt ensemble in (e) RMSE, (f) spread, (g) spread–error ratio, and (h) CRPS for runs initialized at 1200 UTC.

and SPPT) against a multiphysics, baseline configuration. The SPP scheme introduces temporally and spatially varying perturbations to key parameters in the GF convection and MYNN PBL parameterizations. The detailed characteristics of these perturbations were determined through collaboration with physics parameterization experts. While we expect nonstatic parameter perturbations to have a smaller impact than keeping the parameter constant for each ensemble member, this method has the advantage that ensemble members have the same climatology and variance, leading to a more consistent ensemble without members that have different systematic biases. Twenty-four-hour RAP ensemble simulations were verified over 21 days during the summer of 2013, using GEFS data for initial and lateral boundary conditions. We stress that SKEB and SPPT were used in their WRF release configuration, suggested for the horizontal grid spacing used in this study.

The most important findings are summarized below:

- Alone, the parameter perturbations of SPP introduce insufficient spread. However, when combined with SKEB and/or SPPT the spread is as large and for some instances even larger than for a multiphysics ensemble. Overall, the ensemble mean error is changed only slightly.
- An ensemble created by combining three stochastic approaches (spp_skeb_sppt) performed comparably to the multiphysics control ensemble for most of the examined variables, most of the evaluated lead times, and most of the employed statistics.
- SKEB made a larger impact on spread for upper-level wind and 500-hPa geopotential height, while SPPT had a larger impact on spread for near-surface temperature.
- Combining SPP with SPPT generally yields a 2%–6% improvement in the ensemble RMSE, spread, spread–error ratio, and CRPS over an ensemble using SPPT alone. This is an important finding, since SPP represents uncertainty at its source, in a physically consistent way, without introducing systematic member biases.

Our results confirm the findings of previous studies (in particular, Berner et al. 2011, 2015; Hacker et al. 2011a,b) in that 1) parameter perturbations alone do not generate sufficient spread to remedy the underdispersion in short-term ensemble forecasts and 2) a combination of several stochastic schemes outperforms any single scheme. This result implies that a synthesis of different approaches is best suited to capture model error in its full complexity. This finding may have a large impact on the design of next-generation high-resolution regional, as well as global operational ensembles. The single-physics stochastic approach clearly provides a viable alternative to reduce code complexity and improve spread over the multiphysics approach used in some operational ensembles. The plan is to continue this work using the 3-km High-Resolution Rapid Refresh (HRRR) ensemble (Alexander et al. 2016), focusing primarily on the LSM and PBL parameterizations.

Acknowledgments. This study was supported by Developmental Testbed Center (DTC) funding for the DTC ensemble task. Special thanks go to MET and METviewer developers John Halley Gotway, Tara Jensen, Tatiana Burek and Tressa Fowler for their help with MET and METviewer. The authors would also like to thank James Kenyon for his contribution and expertise, and Trevor Alcott for a thorough internal review.

REFERENCES

Alexander, C., and Coauthors, 2016: Development of high-resolution Rapid Refresh (HRRR) ensemble data assimilation, forecasts and post processing. *Proc. Seventh Ensemble User's Workshop*, College Park, MD, NCEP, 7.4. [Available online at http://www.dtcenter.org/events/workshops16/ensembles/docs/day2/7.4_-_Alexander_Curtis_-_Development_of_High-Resolution_Rapid_Refresh_(HRRR)_Ensemble_Data_Assimilation_Forecasts_and_Post_Processing.pdf.]

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

Berner, J., G. Shutts, M. Leutbecher, and T. Palmer, 2009: A spectral stochastic kinetic energy backscatter scheme and its impact on flow-dependent predictability in the ECMWF ensemble prediction system. *J. Atmos. Sci.*, **66**, 603–626, doi:10.1175/2008JAS2677.1.

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

——, T. Jung, and T. N. Palmer, 2012: Systematic model error: The impact of increased horizontal resolution versus improved stochastic and deterministic parameterizations. *J. Climate*, **25**, 4946–4962, doi:10.1175/JCLI-D-11-00297.1.

——, K. R. Fossell, S.-Y. Ha, J. P. Hacker, and C. Snyder, 2015: Increasing the skill of probabilistic forecasts: Understanding performance improvements from model-error representations. *Mon. Wea. Rev.*, **143**, 1295–1320, doi:10.1175/MWR-D-14-00091.1.

Betts, A. K., 1986: A new convective adjustment scheme. Part I: Observational and theoretical basis. *Quart. J. Roy. Meteor. Soc.*, **112**, 677–691, doi:10.1002/qj.49711247307.

Bougeault, P., and P. Lacarrère, 1989: Parameterization of orography-induced turbulence in a mesobeta-scale model. *Mon. Wea. Rev.*, **117**, 1872–1890, doi:10.1175/1520-0493(1989)117<1872:POOITI>2.0.CO;2.

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

Bowler, N. E., A. Arribas, S. E. Beare, K. R. Mylne, and G. J. Shutts, 2009: The local ETKF and SKEB: Upgrades to the MOGREPS short-range ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **135**, 767–776, doi:10.1002/qj.394.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

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, doi:10.1002/qj.49712556006.

——, P. L. Houtekamer, Z. Toth, G. Pellerin, M. Wei, and Y. Zhu, 2005: A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Wea. Rev.*, **133**, 1076–1097, doi:10.1175/MWR2905.1.

Candille, G., and O. Talagrand, 2005: Evaluation of probabilistic prediction systems for a scalar variable. *Quart. J. Roy. Meteor. Soc.*, **131**, 2131–2150, doi:10.1256/qj.04.71.

——, and ——, 2008: Retracted and replaced: Impact of observational error on the validation of ensemble prediction systems. *Quart. J. Roy. Meteor. Soc.*, **134**, 509–521, doi:10.1002/qj.221.

Charles, M. E., and B. A. Colle, 2009: Verification of extratropical cyclones within the NCEP operational models. Part II: The Short-Range Ensemble Forecast System. *Wea. Forecasting*, **24**, 1191–1214, doi:10.1175/2009WAF2222170.1.

Christensen, H. M., I. M. Moroz, and T. N. Palmer, 2015: Stochastic and perturbed parameter representations of model uncertainty in convection parameterization. *J. Atmos. Sci.*, **72**, 2525–2544, doi:10.1175/JAS-D-14-0250.1.

Eckel, F. A., and C. F. Mass, 2005: Aspects of effective mesoscale, short-range ensemble forecasting. *Wea. Forecasting*, **20**, 328–350, doi:10.1175/WAF843.1.

Grell, G. A., and D. Dévényi, 2002: A generalized approach to parameterizing convection combining ensemble and data assimilation techniques. *Geophys. Res. Lett.*, **29**, 1693–1696, doi:10.1029/2002GL015311.

——, and S. R. Freitas, 2014: A scale and aerosol aware stochastic convective parameterization for weather and air quality modeling. *Atmos. Chem. Phys.*, **14**, 5233–5250, doi:10.5194/acp-14-5233-2014.

Guan, H., B. Cui, and Y. Zhu, 2015: Improvement of statistical postprocessing using GEFS reforecast information. *Wea. Forecasting*, **30**, 841–854, doi:10.1175/WAF-D-14-00126.1.

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

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

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

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

Han, J., and H.-L. Pan, 2011: Revision of convection and vertical diffusion schemes in the NCEP Global Forecast System. *Wea. Forecasting*, **26**, 520–533, doi:10.1175/WAF-D-10-05038.1.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting*, **15**, 559–570, doi:10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Hong, S.-Y., Y. Noh, and J. Dudhia, 2006: A new vertical diffusion package with an explicit treatment of entrainment processes. *Mon. Wea. Rev.*, **134**, 2318–2341, doi:10.1175/MWR3199.1.

Janjić, Z. I., 1994: The step-mountain Eta coordinate model: Further developments of the convection, viscous sublayer, and turbulence closure schemes. *Mon. Wea. Rev.*, **122**, 927–945, doi:10.1175/1520-0493(1994)122<0927:TSMECM>2.0.CO;2.

Knutti, R., D. Masson, and A. Gettelman, 2013: Climate model genealogy: Generation CMIP5 and how we got there. *Geophys. Res. Lett.*, **40**, 1194–1199, doi:10.1002/grl.50256.

Lin, Y., and K. E. Mitchell, 2005: The NCEP Stage II/IV hourly precipitation analyses: Development and applications. *19th Conf. on Hydrology,* San Diego, CA, Amer. Meteor. Soc., 1.2. [Available online at https://ams.confex.com/ams/Annual2005/techprogram/paper_83847.htm.]

Mason, P., and D. Thomson, 1992: Stochastic backscatter in large-eddy simulations of boundary layers. *J. Fluid Mech.*, **242**, 51–78, doi:10.1017/S0022112092002271.

Müller, W., C. Appenzeller, F. Doblas-Reyes, and M. Liniger, 2005: A debiased ranked probability skill score to evaluate probabilistic ensemble forecasts with small ensemble sizes. *J. Climate*, **18**, 1513–1523, doi:10.1175/JCLI3361.1.

Murphy, J., D. Sexton, D. Barnett, G. Jones, M. Webb, M. Collins, and D. Stainforth, 2004: Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*, **430**, 768–772, doi:10.1038/nature02771.

Nakanishi, M., and H. Niino, 2004: An improved Mellor–Yamada Level-3 model with condensation physics: Its design and verification. *Bound.-Layer Meteor.*, **112**, 1–31, doi:10.1023/B:BOUN.0000020164.04146.98.

——, and ——, 2006: An improved Mellor–Yamada Level-3 model: Its numerical stability and application to a regional prediction of advection fog. *Bound.-Layer Meteor.*, **119**, 397–407, doi:10.1007/s10546-005-9030-8.

Palmer, T. N., 2001: A nonlinear dynamical perspective on model error: A proposal for non-local stochastic-dynamic parameterization in weather and climate prediction. *Quart. J. Roy. Meteor. Soc.*, **127**, 279–304, doi:10.1002/qj.49712757202.

——, R. Buizza, F. Doblas-Reyes, T. Jung, M. Leutbecher, G. Shutts, M. Steinheimer, and A. Weisheimer, 2009: Stochastic parametrization and model uncertainty. ECMWF Tech. Memo. 598, 42 pp. [Available online at http://www.ecmwf.int/publications/.]

Reynolds, C. A., J. G. McLay, J. S. Goerss, E. A. Serra, D. Hodyss, and C. R. Sampson, 2011: Impact of resolution and design on the U.S. Navy global ensemble performance in the tropics. *Mon. Wea. Rev.*, **139**, 2145–2155, doi:10.1175/2011MWR3546.1.

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

Sanchez, C., K. D. Williams, and M. Collins, 2015: Improved stochastic physics schemes for global weather and climate models. *Quart. J. Roy. Meteor. Soc.*, **142**, 147–159. doi:10.1002/qj.2640.

Schaefer, J. T., 1990: The critical success index as an indicator of warning skill. *Wea. Forecasting*, **5**, 570–575, doi:10.1175/1520-0434(1990)005<0570:TCSIAA>2.0.CO;2.

Shutts, G. J., 2005: A kinetic energy backscatter algorithm for use in ensemble prediction systems. *Quart. J. Roy. Meteor. Soc.*, **131**, 3079–3102, doi:10.1256/qj.04.106.

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

Smirnova, T. G., J. M. Brown, S. G. Benjamin, and J. S. Kenyon, 2016: Modifications to the Rapid Update Cycle Land Surface Model (RUC LSM) available in the Weather Research and Forecast (WRF) Model. *Mon. Wea. Rev.*, **144**, 1851–1865, doi:10.1175/MWR-D-15-0198.1.

Stensrud, D. J., H. E. Brooks, J. Du, M. S. Tracton, and E. Rogers, 1999: Using ensembles for short-range forecasting. *Mon. Wea. Rev.*, **127**, 433–446, doi:10.1175/1520-0493(1999)127<0433:UEFSRF>2.0.CO;2.

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

Weaver, A., and P. Courtier, 2001: Correlation modelling on the sphere using a generalized diffusion equation. *Quart. J. Roy. Meteor. Soc.*, **127**, 1815–1846, doi:10.1002/qj.49712757518.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. International Geophysics Series, Vol. 100, Academic Press, 676 pp.