A Unified Scheme of Stochastic Physics and Bias Correction in an Ensemble Model to Reduce Both Random and Systematic Errors

YU XIA

Nanjing University of Information Science and Technology, Nanjing, China

JING CHEN

Numerical Weather Prediction Center, CMA, Beijing, China

JUN DU

Environmental Modeling Center, NOAA/NWS/NCEP, College Park, Maryland

XIEFEI ZHI

Nanjing University of Information Science and Technology, Nanjing, China

JINGZHUO WANG AND XIAOLI LI

Numerical Weather Prediction Center, CMA, Beijing, China

(Manuscript received 21 February 2019, in final form 16 August 2019)

ABSTRACT

This study experimented with a unified scheme of stochastic physics and bias correction within a regional ensemble model [Global and Regional Assimilation and Prediction System–Regional Ensemble Prediction System (GRAPES-REPS)]. It is intended to improve ensemble prediction skill by reducing both random and systematic errors at the same time. Three experiments were performed on top of GRAPES-REPS. The first experiment adds only the stochastic physics. The second experiment adds only the bias correction scheme. The third experiment adds both the stochastic physics and bias correction. The experimental period is one month from 1 to 31 July 2015 over the China domain. Using 850-hPa temperature as an example, the study reveals the following: 1) the stochastic physics can effectively increase the ensemble spread, while the bias correction cannot. Therefore, ensemble averaging of the stochastic physics runs can reduce more random error than the bias correction runs. 2) Bias correction can significantly reduce systematic error, while the stochastic physics cannot. As a result, the bias correction greatly improved the quality of ensemble mean forecasts but the stochastic physics did not. 3) The unified scheme can greatly reduce both random and systematic errors at the same time and performed the best of the three experiments. These results were further confirmed by verification of the ensemble mean, spread, and probabilistic forecasts of many other atmospheric fields for both upper air and the surface, including precipitation. Based on this study, we recommend that operational numerical weather prediction centers adopt this unified scheme approach in ensemble models to achieve the best forecasts.

1. Introduction

Ensemble prediction has become one of the most important components of numerical weather prediction (NWP) (Buizza et al. 2018). Many methods have been proposed for perturbing initial conditions and models (e.g., Tracton and Kalnay 1993; Houtekamer et al. 1996; Chen et al. 2002; Shutts 2005; Ma et al. 2008, 2015; Berner et al. 2009; Kazuo et al. 2012; Ollinaho et al. 2017;

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

Corresponding author: Jing Chen, chenj@cma.gov.cn

Feng et al. 2014, 2018). For a complete review of current ensemble methods, readers are referred to Du et al. (2018). All NWP modeling systems produce two kinds of forecast errors: random and systematic errors. The ensemble methods are designed to deal with random but not the systematic errors (Du 2007). Since systematic error (bias) will inversely impact the quality of ensemble forecasts and the capability of truly assessing an ensemble prediction system (EPS; Wang et al. 2018), it would be ideal if model bias can also be reduced or removed in an EPS. Currently, the removal of model's systematic bias is usually achieved by statistical methods and done separately and independently as a post processing step following model integration (e.g., Gneiting et al. 2005; Raftery et al. 2005; Monache et al. 2006; Bakhshaii and Stull 2009; Cui et al. 2006, 2012; Du and Zhou 2011; Li et al. 2011). In other words, the treatment of random and systematic errors cannot be done at the same time using a single unified scheme within an ensemble model in current NWP, which prevents bias-corrected fields from being used in some downstream applications (see the discussions of the next paragraph and section 3c).

Very recently, we introduced a three-dimensional wholesale-like dynamical method able to remove systematic model bias for all variables at the same time, as an integral part of model integration (Chen et al. 2019, manuscript submitted to *Quart. J. Roy. Meteor. Soc.*, hereafter C19). As we discussed in C19, the advantage of this dynamical bias correction approach is threefold: 1) convenience (the two steps of "model integration and postprocessing" become one step "model integration"), 2) improvement in forecast products derived from multiple variables (due to consistency among the variables), and 3) the capability of initializing a downstream model (including two-way nested regional modeling) with dynamically consistent bias-corrected fields. This approach also makes it possible to deal with both random and systematic errors in a unified scheme during model integration. In this study, we will design and test such a unified scheme by applying the method of C19 and a stochastic physics scheme at the same time to an EPS, aiming to reduce the systematic and random errors simultaneously. To better understand how the unified scheme works, we will compare the performances of each of the components involved. Specifically, the stochastic perturbed parameterization tendency component (SPPT) alone will first be implemented on top of a base EPS (control experiment) to examine how it impacts the performance of the EPS. Then the bias correction component alone will be implemented. Finally the two components will be

implemented together as a unified scheme (i.e., the bias correction will be performed before the SPPT is performed).[1] The rest of this paper is organized as follows. Section 2 describes the model, experiment design, and data. The results are presented in section 3. A summary is given in section 4.

2. Model, experiment design, and data

a. The base model (GRAPES) and the control EPS (GRAPES-REPS)

The base model in this study is a regional version of the Global and Regional Assimilation and Prediction System (GRAPES), which is developed at the Numerical Weather Prediction Center of the China Meteorological Administration (CMA; Chen et al. 2008). The main features of GRAPES include a fully compressible dynamical core with nonhydrostatic approximation, a semi-implicit and semi-Lagrangian scheme for time integration, and a height-based terrain following coordinate. The model physics includes Rapid Radiative Transfer Model (RRTM) longwave radiation (Mlawer et al. 1997), Dudhia shortwave radiation (Dudhia 1989), WSM-6 microphysics (Hong and Lim 2006), Noah land surface model (Mahrt and Ek 1984), the MRF PBL scheme (Hong and Pan 1996), and Monin–Obukhov surface layer scheme (Noilhan and Planton 1989). Model analysis is produced by a three-dimensional variation data assimilation scheme (Zhuang et al. 2014).

The control experiment (CTL) is the GRAPES-based regional EPS (GRAPES-REPS), which has been running operationally at CMA since August 2014 (Zhang et al. 2014). It has 15 members (1 control and 14 perturbed members) covering the China domain (15°–64.35°N, 70°–145.15°E). The horizontal resolution is 15 km with 51 vertical levels (model top is 10 hPa). It runs twice a day (initialized at 0000 and 1200 UTC) out to 72 h of forecast length with 6-hourly output (model integration time step is 60 s). The lateral boundary conditions (LBCs) and initial conditions (ICs) of the GRAPES-REPS members are diversely provided (directly downscaled) from the different members of the GRAPES global EPS (Ma et al. 2008) that also runs operationally at CMA. The initial condition perturbations of the GRAPES global EPS are generated by the breeding vector (Toth and Kalnay 1997).

---

[1] To maximize the benefit of the stochastic physics, the SPPT should be carried out in an environment with as little error as possible. Therefore, the bias correction is performed prior to the SPPT.

TABLE 1. Configuration of the control EPS (GRPAES_REPS).

Member	ICs and LBCs	Convective scheme	PBL scheme
Control	Downscaling from global EPS member	Kain–Fritsch–Eta (Kain and Fritsch 1993; Kain 2004)	MRF (Hong and Pan 1996)
Member 1	Downscaling	Original Kain–Fritsch (Kain and Frisch 1990)	MRF
Member 2	Downscaling	Betts–Miller–Janjić (Betts 1986)	MRF
Member 3	Downscaling	Kain–Fritsch–Eta	MRF
Member 4	Downscaling	Original Kain–Fritsch	MRF
Member 5	Downscaling	Betts–Miller–Janjić	MRF
Member 6	Downscaling	Kain–Fritsch–Eta	MRF
Member 7	Downscaling	Original Kain–Fritsch	MRF
Member 8	Downscaling	Simplified Arakawa–Schubert (Pan and Wu 1995)	YSU (Hong et al. 2006)
Member 9	Downscaling	Betts–Miller–Janjić	YSU
Member 10	Downscaling	Original Kain–Fritsch	YSU
Member 11	Downscaling	Simplified Arakawa–Schubert	YSU
Member 12	Downscaling	Betts–Miller–Janjić	YSU
Member 13	Downscaling	Original Kain–Fritsch	YSU
Member 14	Downscaling	Simplified Arakawa–Schubert	YSU

GRAPES_REPS applies a multiphysics approach (Stensrud et al. 2000; Du et al. 2003) for its physics perturbation through a combination of two boundary parameterization and four convective cumulus parameterization schemes (Table 1). The GRAPES-REPS will be used as a reference (CTL) for the three other experiments (Table 2) in the comparisons.

b. Stochastic physics experiment (GRAPES-REPS-SPPT)

The first experiment (Exp. 1) is the stochastically perturbed parameterization tendency (SPPT). SPPT is a commonly used approach for perturbing model physics (e.g., Buizza et al. 1999, 2018; Christensen et al. 2015). Yuan et al. (2016) applied the SPPT (Charron et al. 2010) to the GRAPES-REPS as follows:

$$S_j(t) = \int_{t=0}^{t} [A(S_j, t) + P(S_j, t)] \, dt. \quad (1)$$

Equation (1) is the model integration equation. Here, A is a state variable of ensemble member j at integration time t, where $j = 0, 1, 2, ..., 14$, and $j = 0$ represents the control forecast; $\hat{B}_l(\theta_0, 0)$ is the dynamic tendency, and P is for the physics tendency.

In the SPPT scheme, the physics tendency is multiplied by a random field as follows:

$$S_j(t) = \int_{t=0}^{t} [A(S_j, t) + P(S_j, t) \times R_j(\lambda, \phi, t)] \, dt, \quad (2)$$

where $R_j(\lambda, \phi, t)$ gives the random noise for member j. Here, $R_j(\lambda, \phi, t)$ is described by the first-order Markov chain with spherical harmonics expansion, and has a time–space correlated continuous and smooth horizontal structure, defined as

$$R(\lambda, \phi, t) = \overline{r} + \sum_{l=1}^{L} \sum_{m=-l}^{l} \alpha_{l,m}(t) Y_{l,m}(\lambda, \phi). \quad (3)$$

In Eq. (3), $\overline{r}$ is the mean of $R(\lambda, \phi, t)$ and $\alpha_{l,m}(t)$ is the spectral coefficient as a function of t. The longitude and latitude are represented by λ and ϕ, respectively. The spherical harmonics is expressed by $Y_{l,m}$, where l and m indicate the total horizontal and latitudinal wavenumbers, respectively. The term L is the spatial truncation scale (i.e., wavenumber) of the random field [$L = 28$ in this study, adopted from Charron et al. (2010)]. The evolution of $\alpha_{l,m}(t)$ is obtained by the first-order Markov chain:

$$\alpha_{l,m}(t + \Delta t) = e^{-\Delta t/\tau} \alpha_{l,m}(t) + \sqrt{\frac{4\pi\sigma^2(1 - e^{-2\Delta t/\tau})}{L(L+2)}} r_{l,m}(t). \quad (4)$$

In Eq. (4), Δt is the model integration time step (60s in this case). The term τ is the time scale of temporal correlation for the random field [$\tau = 6$ h in this study; refer to Yuan et al. (2016)]. The term $r_{l,m}(t)$ is a Gaussian distribution with a mean of 0 and variance of 1. Here, σ is the standard deviation of $R(\lambda, \phi, t)$, which is also a Gaussian distribution (i.e., the uncertainty distribution of a particular parameter) [$\sigma = 0.27$ in this study, adopted from Charron et al. (2010)].

As is described in Li et al. (2008), one often needs to keep a perturbed parameter within some specified bounds. Thus, a stretching or dilation transformation $\chi(R, \overline{r})$ is applied to $R(\lambda, \phi, t)$ to obtain a new random field $R'(\lambda, \phi, t)$,

$$R'(\lambda, \phi, t) = \overline{r} + \chi(R, \overline{r})[R(\lambda, \phi, t) - \overline{r}], \quad (5)$$

which can ensure that the new random field lies within specified bounds and has the ability to modify

TABLE 2. Experiment design for the CTL, Exp. 1, Exp. 2, and Exp. 3. The integration equation is expressed in terms of potential temperature θ only. See the text for a detailed description.

Experiments	ICs and LBCs	Model perturbation	Integration equation	Description
CTL: GRAPES_REPS	Downscaling	Multiphysics	$\theta_j(t) = \int_{t=0}^{t} [A(\theta_j, t) + P(\theta_j, t)] dt$	Control EPS
Exp.1: GRAPES_REPS_SPPT	Downscaling	Multiphysics and SPPT	$\theta_j(t) = \int_{t=0}^{t} [A(\theta_j, t) + P(\theta_j, t) \times R'_j] dt$	Stochastic physics
Exp.2: GRAPES_REPS_LTBC	Downscaling	Multiphysics	$\theta_j(t) = \int_{t=0}^{t} [A(\theta_j, t) + P(\theta_j, t) - \hat{B}_j(\theta_0, 0)] dt$	Bias correction
Exp.3: GRAPES_REPS_LTBC_SPPT	Downscaling	Multiphysics and SPPT	$\theta_j(t) = \int_{t=0}^{t} \{A(\theta_j, t) + [P(\theta_j, t) - \hat{B}_j(\theta_0, 0)] \times R'_j\} dt$	Unified scheme (bias correction + stochastic physics)

its probability density function. The dilation transformation is

$$\chi(R, \bar{r}) = 2 - \frac{1 - \exp\left[\beta\left(\frac{R - \bar{r}}{R'_{\max} - \bar{r}}\right)^2\right]}{1 - \exp(\beta)}. \qquad (6)$$

Here, $\bar{r} = (R'_{\max} + R'_{\min})/2$, and $R'_{\max}$ and $R'_{\min}$ are the upper and lower thresholds for $R'(\lambda, \phi, t)$, respectively. In this case, $R'_{\max}$ and $R'_{\min}$ are set to 1.8 and 0.2, respectively, following Yuan et al. (2016). Here, β is a constant equal to -1.27 [as in Li et al. (2008)]. Thus, the SPPT scheme is implemented by multiplying the physics tendency term P by $R'(\lambda, \phi, t)$ as follows:

$$S_j(t) = \int_{t=0}^{t} [A(S_j, t) + P(S_j, t) \times R'_j(\lambda, \phi, t)] \, dt. \qquad (7)$$

This is applied to four model state variables (zonal wind u, meridional wind v, potential temperature θ, and specific humidity q) for the GRAPES model.

c. Bias correction experiment (GRAPES-REPS-LTBC)

The second experiment (Exp. 2) is the linear tendency bias correction (LTBC). This method subtracts a linear bias tendency directly in the model tendency equation at each time step during model integration (C19):

$$S_j(t) = \int_{t=0}^{t} [A(S_j, t) + P(S_j, t) - \widehat{B}_l(S_j, t)] \, dt. \qquad (8)$$

The $\hat{B}_l(S_j, t)$ is the linear bias tendency estimated from past forecasts:

$$\hat{B}_l(S_j, t) = \text{slope} \times \text{time\_step} = \frac{\hat{b}}{\Delta \times 3600} \times \delta t. \qquad (9)$$

Linear regression is used to estimate the bias increment $\hat{b}$ over a time window Δ (in hours) by linearly fitting all bias values within the window. The δt is the model integration time step (in seconds). As discussed in C19, for the feasibility of operational implementation as well as the similarity in biases for all members, the control member's bias is used for the 14 perturbed members (i.e., $j = 0$). For a detailed description of the LTBC and its test results, readers are referred to C19 where three different experiments were carried out. For the reader's convenience, an example of the estimation of the linear bias tendency as well as a performance comparison to a statistical bias correction method is provided in the appendix. Since the first experiment of C19 is the simplest with a good performance, we will use that same experimental setup in this

study. Specifically, the bias correction step is only enforced in the potential temperature field θ in the model (i.e., $S = \theta$) due to its dominance in the GRAPES-EPS bias error (the detailed discussions about the cause and the impact can be found in C19. The bias linear tendency of potential temperature is estimated from a 72-h window [with 13 (total) 6-hourly bias values in it; see the appendix] and is used for all time steps (i.e., $t = 0$) throughout the model integration [i.e., $\hat{B}_l(S_j, t) = \hat{B}_l(\theta_0, 0)$]. Therefore, the second experiment can be described by

$$\theta_j(t) = \int_{t=0}^{t} [A(\theta_j, t) + P(\theta_j, t) - \widehat{B}_l(\theta_0, 0)]\, dt. \qquad (10)$$

The other state variables are the same as in the CTL EPS.

d. Experiment combining stochastic physics and bias correction (GRAPES-REPS-LTBC-SPPT)

The third experiment (Exp. 3) is the combination of Exp. 1 and Exp. 2 to test the combined effects of stochastic physics and bias correction on EPS performance. Specifically, the bias correction is performed before the stochastic physics step:

$$S_j(t) = \int_{t=0}^{t} \{A(S_j, t) + [P(S_j, t) - \hat{B}_l(S_j, t)] \\ \times R'_j(\lambda, \phi, t)\}\, dt, \qquad (11)$$

where $\hat{B}_l(S_j, t) = \widehat{B}_l(\theta_0, 0)$ for θ and $\hat{B}_l(S_j, t) = 0$ for other model state variables (i.e., the other state variables are the same as Exp. 1). Table 2 is a summary of the control EPS and three experiments performed in this study.

e. Data

The forecast experiments were carried out over the period of a month (1–31 July 2015). The bias tendency is calculated from the prior bias. The prior bias was approximated by the average forecast error of the most recently available past 10-day forecasts. For example, the average errors of the 19–28 June forecasts are regarded as the prior biases for the 1 July forecast (6-hourly outputs during a 72-h forecast length). As mentioned above, only the control member's bias is calculated and used for the 14 perturbed members. The reason for using a 10-day period to estimate bias is that it is not too short to miss the main features of systematic error, and it is not too long to completely filter out flow-dependent error. It should be beneficial to retain some recent flow-dependent bias information in the bias tendency,

given that model bias is regime dependent (Du and DiMego 2008).

Since an analysis is the optimal estimation of truth a model can produce, the goal of the grid-to-grid model bias correction is to mimic its own model analysis, although the analysis itself may contain bias with respect to an observation (Privé et al. 2013).[2] Therefore, the GRAPES analysis is used as truth for verifying the upper air temperature, geopotential height, wind (u and v), and surface temperature and wind (2-m temperature and 10-m u and v). The CMA Multisource merged Precipitation Analysis System version 2.1 (CMPAS-V2.1) (Pan et al. 2015) is used as truth for verifying precipitation. All the verification results presented in section 3 are averaged over these 31 days and the model domain (15°–64.35°N, 70°–145.15°E). Only the verification of 0000 UTC cycle forecasts is presented since the result from the 1200 UTC cycle is similar.

The following metrics are employed in the verification: root-mean-square error (RMSE), systematic error (bias), random error, spread, spread–skill relationship (consistency, defined as the ratio of RMSE to spread), continuous ranked probability score (CRPS), reliability curve, and outlier for basic variables; and Brier score (BS) and area of relative operating characteristics (AROC) for precipitation. For a review of ensemble verification, readers are referred to Jolliffe and Stephenson (2003) and Du and Zhou (2017).

3. Results

a. Changes in ensemble spread and forecast error components

For the purpose of demonstration, 850-hPa temperature is used here. Figure 1 shows the improvements to the ensemble mean forecasts and ensemble spread for the three experiments (the blue curve) compared to the control EPS (the red curve). It clearly shows that the stochastic physics SPPT boosted the ensemble spread but only slightly improved the ensemble mean forecast accuracy (Fig. 1a). In contrast, the bias correction LTBC greatly reduced the ensemble mean forecast RMSE but made little change in the ensemble spread (Fig. 1b). When the stochastic physics and bias correction were combined, the unified scheme took the advantages from the both methods and noticeably improved both the

---

[2] Note: Correcting an analysis-like forecast to the observation at a site is a task of downscaling. The difference between the grid-to-grid bias correction and grid-to-point downscaling is often not clearly distinguished by many. These are two different steps in an NWP operation.
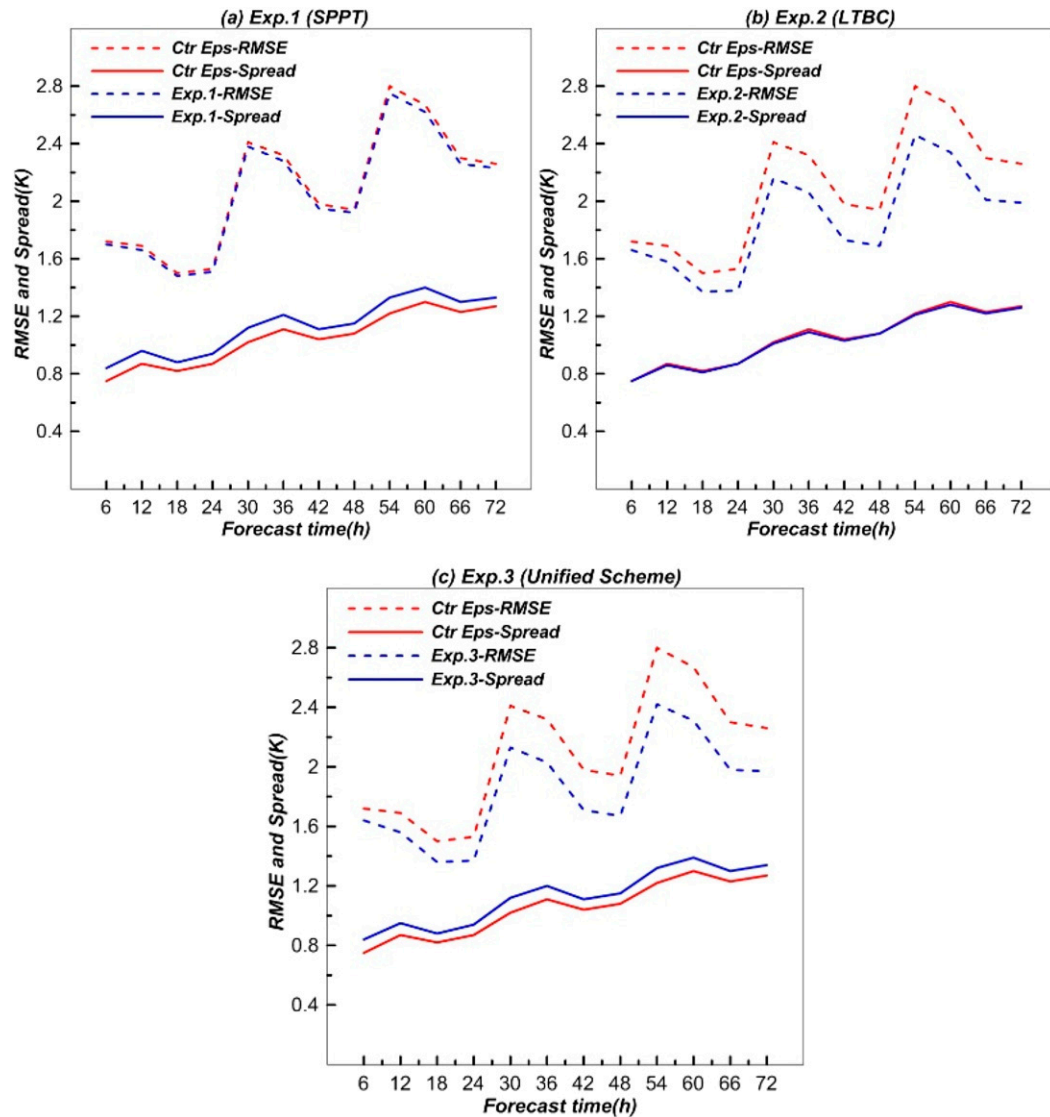
FIG. 1. The comparison of ensemble mean forecast error (RMSE; dashed curve) and ensemble spread (solid curve) between the control EPS (red) and the experiment (blue) for (a) Exp. 1 with SPPT, (b) Exp. 2 with LTBC, and (c) Exp. 3 with the unified scheme (SPPT + LTBC). The 850-hPa temperature is verified.

ensemble mean forecast accuracy and ensemble diversity (Fig. 1c). This makes the ensemble spread and the ensemble mean forecast error about 20%–40% closer to each other than in the control EPS (cf. the control EPS and Exp. 3 in Fig. 1c): an improvement in the spread–skill relationship although the underdispersion is still apparent.[3]

The spread increase (by the SPPT and the unified scheme) and the RMSE reduction (by the LTBC and the unified scheme) over the control EPS are statistically significant at the 99.99% level (Student's t test).

Figure 2 further analyzed the detailed aspects of the error reduction. The top panel (Figs. 2a1–a3) is the total error (RMSE) of the ensemble mean forecast from the three experiments (the blue curve) compared to that of the control EPS (the red curve), the middle panel (Figs. 2b1–b3) is the same but is for the systematic error (bias), and the bottom panel (Figs. 2c1–c3) is for the random error reduction. Since the ensemble averaging is a nonlinear filtering process to cancel

---

[3] Besides the insufficient sampling of model and initial condition uncertainty, model bias plays a significant role in contributing to the underdispersive nature of the GRAPES-REPS. For an in-depth analysis of how model bias adversely impacts the spread–skill relationship of an ensemble, readers are referred to Wang et al. (2018).
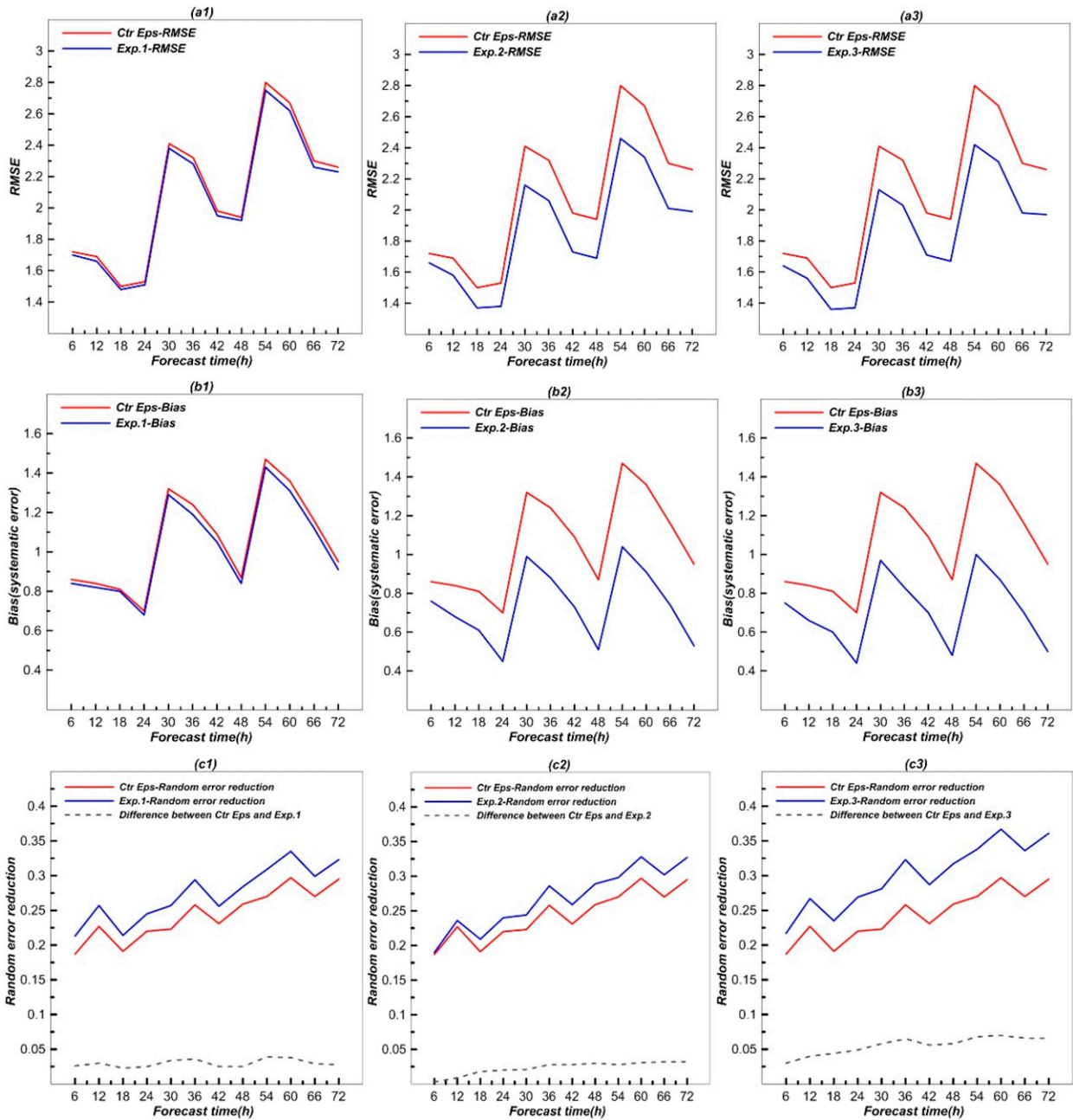
FIG. 2. The comparison of (a1)–(a3) total (RMSE) and (b1)–(b3) systematic error (bias) of the ensemble mean forecast, and (c1)–(c3) the random error reduction (with the dashed gray curve indicating the difference between the two curves) between the control EPS (red) and the three experiments (blue). (left) Exp. 1 (SPPT), (center) Exp. 2 (LTBC), and (right) Exp. 3 (the unified scheme SPPT+LTBC). The 850-hPa temperature is verified.

random errors, the random error reduction shown in Figs. 2c1–c3 is defined as the difference between the average error of individual members and the error of the ensemble mean forecast. Figure 2 shows that the systematic error was reduced very little by the SPPT (Fig. 2b1) but was significantly (99.99%) reduced by the LTBC (Fig. 2b2). As for the random error, due to

the increased diversity of ensemble members in the SPPT run, as seen in Fig. 1, the random error was reduced about 40% more by the SPPT (~0.035) than the LTBC (~0.025) (cf. the dashed curves of Fig. 2c1 and Fig. 2c2) averaged over 6–72 h. This difference is even more significant (almost double, from 0.015 to 0.03) when the forecast length is shorter (6–36 h). After

combining the SPPT and LTBC, the unified scheme can significantly (99.99%) reduce both the systematic and random errors (Figs. 2b3 and 2c3), which leads to the largest reduction in the total error of the ensemble mean forecast among the three experiments (Fig. 2a3).

To verify the above result related to the random error, we employed a different independent approach to calculate the random error [Eq. (12)]. Equation (12) is a commonly used relationship among the total (RMSE), systematic (bias), and random errors in physics (Morris and Langari 2016):

$$\text{total error}^2 = \text{systematic error}^2 + \text{random error}^2. \quad (12)$$

For each of the four EPS runs (control EPS and Exps. 1–3), we can calculate the total error and systematic error of each member and the ensemble mean forecast. Then, the random error can be derived from Eq. (12) for each member and the ensemble mean. The difference between the averaged random error of individual members and the random error of the ensemble mean forecast reflects the random error reduction through the ensemble averaging process for a given EPS. Figure 3 shows the random error reduction of the four EPSs. The result confirms that the random error reduction through ensemble averaging was significantly larger with the stochastic physics (Exp. 1) compared to the control EPS (Ctr-EPS), while it remained similar to the control EPS with the bias correction (cf. Exp. 2 and Ctr-EPS). The unified scheme (Exp. 3) performed the best (slightly better than Exp. 1).

b. Forecast improvements

In this section we will examine how the above error reductions translate into improvement of the probabilistic forecasts of various meteorological fields.

1) PROBABILISTIC FORECASTS OF TEMPERATURE, HEIGHT, MOISTURE, AND WIND

The CRPS is commonly used to measure the closeness between the forecasted and observed cumulative probability distributions. The closer a forecasted probability is to the observation (either 0 or 1), the better it is. Therefore, the CRPS becomes better as the values become smaller, with a perfect score of 0. Figure 4 compares the CRPSs of the three experiments and the control EPS for temperature, geopotential height, specific humidity, and wind at the 850-hPa level as well as for two surface fields (10-m wind and 2-m temperature) over 6–72-h forecasts. For all variables, the CRPS can be slightly improved (i.e., reduced) by Exp. 1 (SPPT) and greatly improved by Exp. 2 (LTBC). This is because the center of the
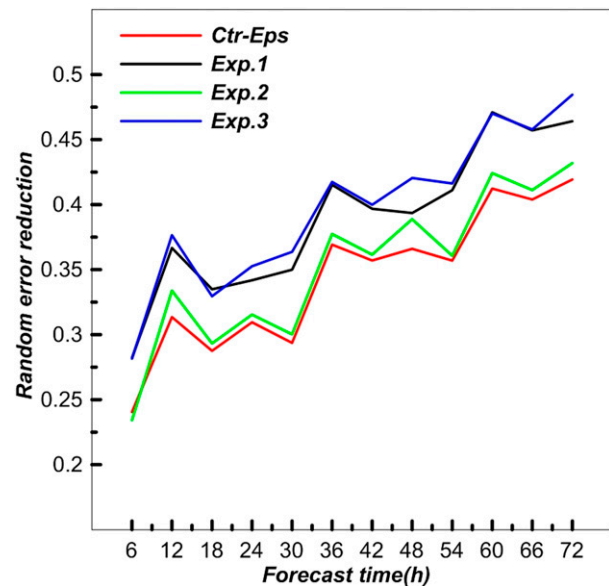

FIG. 3. Random error reduction through ensemble averaging (i.e., the difference between the averaged random error of individual members and the random error of the ensemble mean forecast), for the control EPS (black), Exp. 1 (red), Exp. 2 (green), and Exp. 3(blue). The random error is derived from Eq. (12). The 850-hPa temperature is verified.

probabilistic distribution has been correctly shifted toward the observation after the removal of systematic error in the LTBC run. The CRPS can be further improved on the top of the LTBC run by Exp. 3 (the unified scheme). The statistical significance of the improvement over the control EPS by the unified scheme exceeds the 90% level for all variables and forecast hours (most actually reached the 99.99% level). In other words, the unified scheme can significantly improve the quality of the probabilistic forecasts of all fields.

In addition to the closeness to the observation, another important property of a probabilistic forecast is the reliability. The reliability diagram graphically measures how well the predicted probabilities (the horizontal axis) of an event match its observed frequency (the vertical axis). For a perfectly reliable probabilistic forecast, the predicted probability is equal to the observed frequency, which puts the reliability curve along the 45° diagonal line. Deviation from the diagonal line suggests that the predicted probability is not reliable. If the reliability curve lies below (above) the diagonal line, the probability is overestimated (underestimated) due to either a too small (large) spread of the ensemble members or forecast biases. Figure 5 is the reliability diagram at the forecast length of 72 h for the same six variables as in Fig. 4. The probability thresholds used in Fig. 5 are, respectively, exceeding 2 K (850-hPa
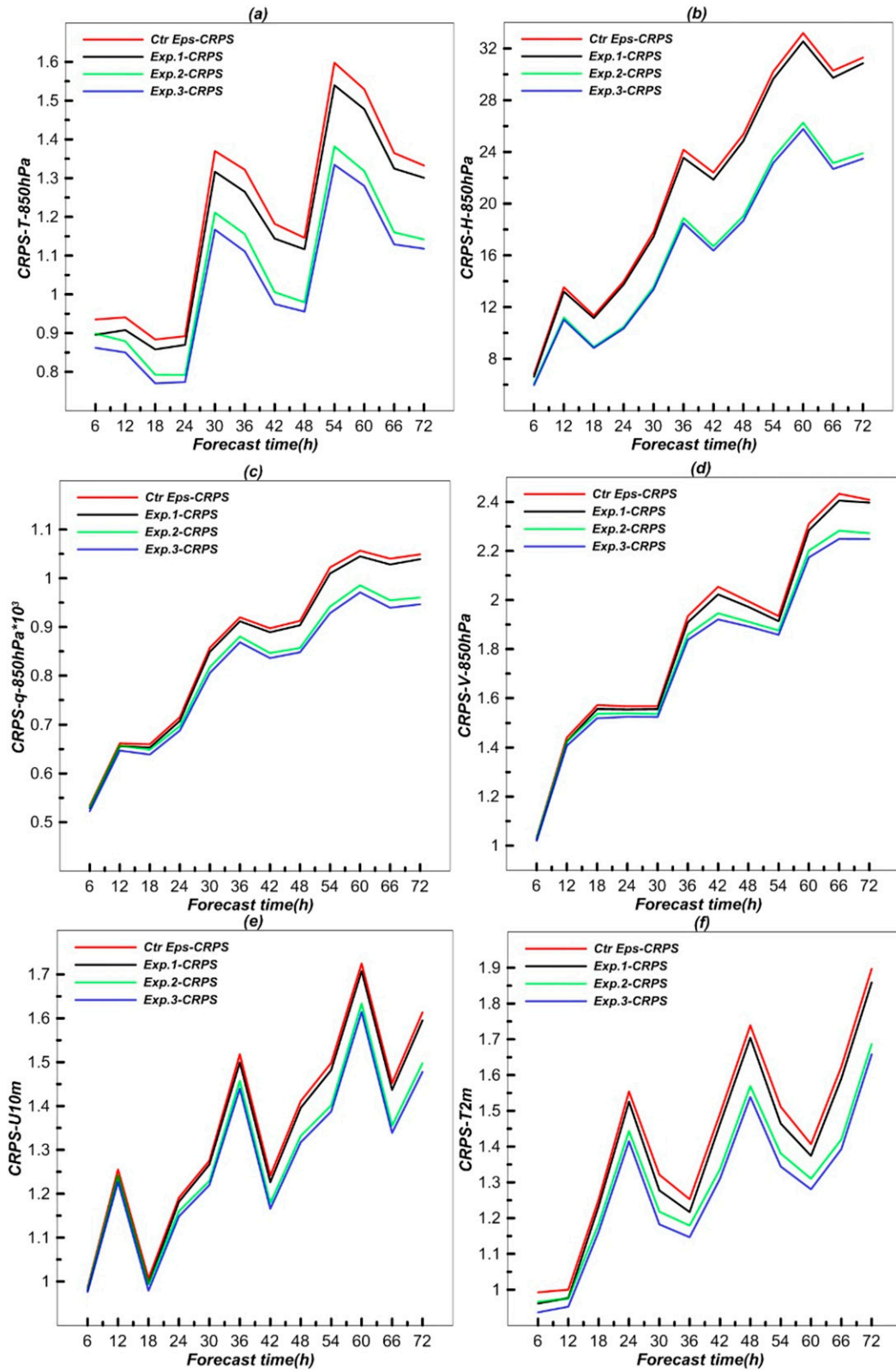
FIG. 4. CRPS of the control EPS (red), Exp. 1 (black), Exp. 2 (green), and Exp. 3 (blue) for (a) T at 850 hPa, (b) H at 850 hPa, (c) q at 850 hPa, (d) V at 850 hPa, (e) U at 10 m, and (f) T at 2 m.
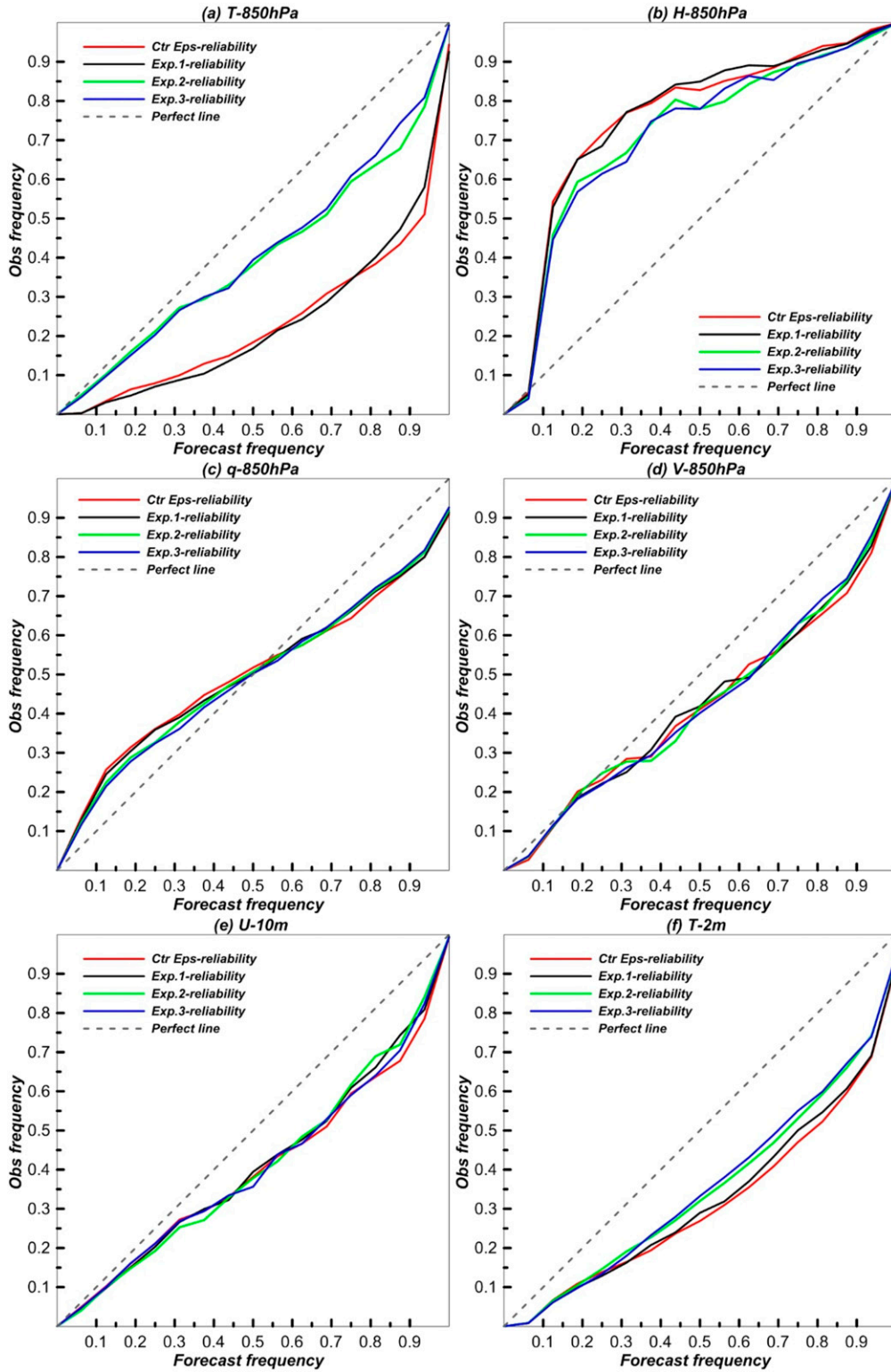
FIG. 5. Reliability diagram of the control EPS (red), Exp. 1 (black), Exp. 2 (green), and Exp. 3 (blue) for (a) T at 850 hPa, (b) H at 850 hPa, (c) q at 850 hPa, (d) V at 850 hPa, (e) U at 10 m, and (f) T at 2 m.

T), 8 gpm (850-hPa H), 0.0002 (850-hPa q), $2\,\mathrm{m\,s^{-1}}$ (850-hPa V), $1\,\mathrm{m\,s^{-1}}$ (10-m U), and 2 K (2-m T) over climatology. For example, Fig. 5a shows that the forecast probability of 850-hPa temperature exceeding 2 K over climatology is much larger than the observed frequency (i.e., overestimated) due to both an underdispersion in spread (Fig. 1) and a warm bias in the forecast (Fig. 2). The SPPT (the black curve) did not improve reliability, but the LTBC (the green curve) and the unified scheme (the blue curve) significantly (at the 95% level) improved the reliability compared to the control EPS (the red curve). Similar results were seen for the 2-m temperature (Fig. 5f) and 850-hPa geopotential height (Fig. 5b). However, the improvements from all the three experiments were not significant for the 850-hPa moisture (Fig. 5c) and wind (Fig. 5d) fields as well as the surface wind (Fig. 5e).

2) PROBABILISTIC FORECASTS OF PRECIPITATION

The CRPS is for continuous forecasts (such as a continuous range of rainfall). For single-category forecasts (such as rainfall exceeding 50 mm), the CRPS can be simplified to the Brier score. BS measures the difference (error) between the forecast probability and the observed probability (0 or 1). A smaller BS indicates a better forecast (a perfect score is 0.0). Figure 6 shows the BS of the three experiments and the control EPS for the probabilistic forecasts of 24-h-accumulated precipitation exceeding 0.1, 10, 25, 50, and 100 mm for forecast lengths of 24, 48, and 72 h. Since there is almost no skill for the 100-mm category rainfall prediction by the GRAPES-EPS, it will be excluded in the following discussion. There is no significant difference (less than 20% level) between the three experiments and the control EPS at the 24-h forecast (Fig. 6a). There is a moderately significant improvement (40% level) by the unified scheme over the control EPS for the 10- and 25-mm categories at the 48-h forecast (Fig. 6b). The moderately significant improvement (40% level) can also be seen for the 10-, 25- and 50-mm categories at the 72-h forecast (Fig. 6c).

The relative operating characteristic (ROC) measures the combined effect of the false alarm rate (FAR; the horizontal axis ranges from 0 to 1) and probability of detection (POD, the vertical axis ranges from 0 to 1) for a binary forecast. A good forecast system needs to maximize POD and minimize FAR. The area under the ROC curve (AUR or AROC) is often calculated to determine if a forecast has skill or not. It has no skill when AROC is less than 0.5 (FAR > POD). A perfect AROC is 1 (100% POD and 0% FAR). Figure 7 is the same as Fig. 6 but for the AROC. There is a significant improvement (90% level) of the unified scheme over the
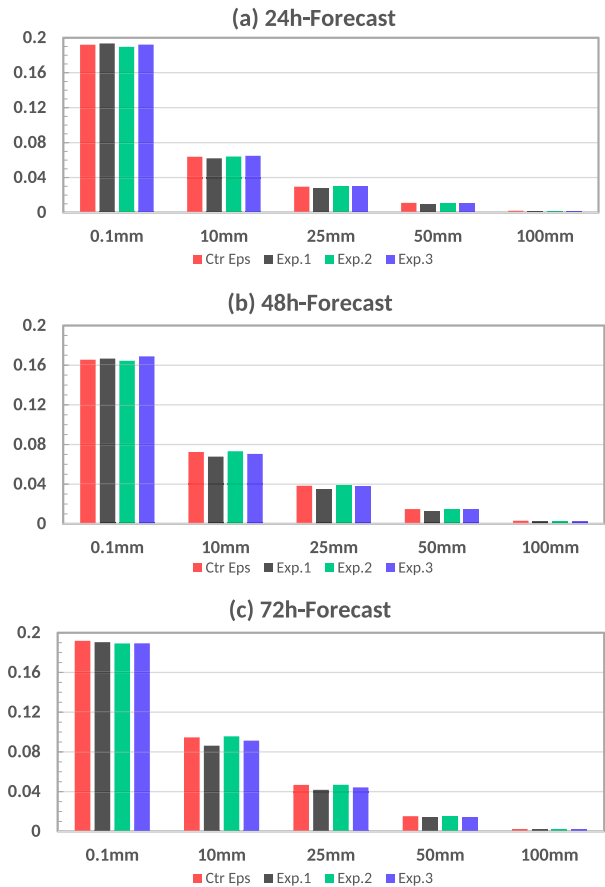

FIG. 6. Brier scores for the (a) 24-, (b) 48-, and (c) 72-h forecasts of 24-h accumulated precipitation exceeding 0.1, 10, 25, 50, and 100 mm, for the control EPS, Exp. 1 (SPPT), Exp. 2 (LTBC), and Exp. 3 (the unified scheme).

control EPS for the 50-mm rainfall category at the 24-h forecast (Fig. 7a). A moderately significant improvement (75% level) for the 10- and 25-mm rainfall and a significant degradation (90% level) for the 50-mm rainfall were seen at the 48-h forecast (Fig. 7b). All rainfall categories (10, 25, and 50 mm) were significantly improved (60%–98% level) at the 72-h forecast (Fig. 7c).

From the above results, we can see that the improvement in probabilistic precipitation forecasts from the unified scheme generally increased as the forecast length increased. However, compared to the other atmospheric variables, precipitation was much less impacted by the SPPT, LTBC, and the unified schemes. The GRAPES-REPS has very little skill in predicting the rainfall category exceeding $100\,\mathrm{mm\,day^{-1}}$.

3) OPERATIONAL IMPLEMENTATION ASSESSMENT SCORECARD

For an operational implementation decision, a scorecard is normally used, which verifies a more complete list
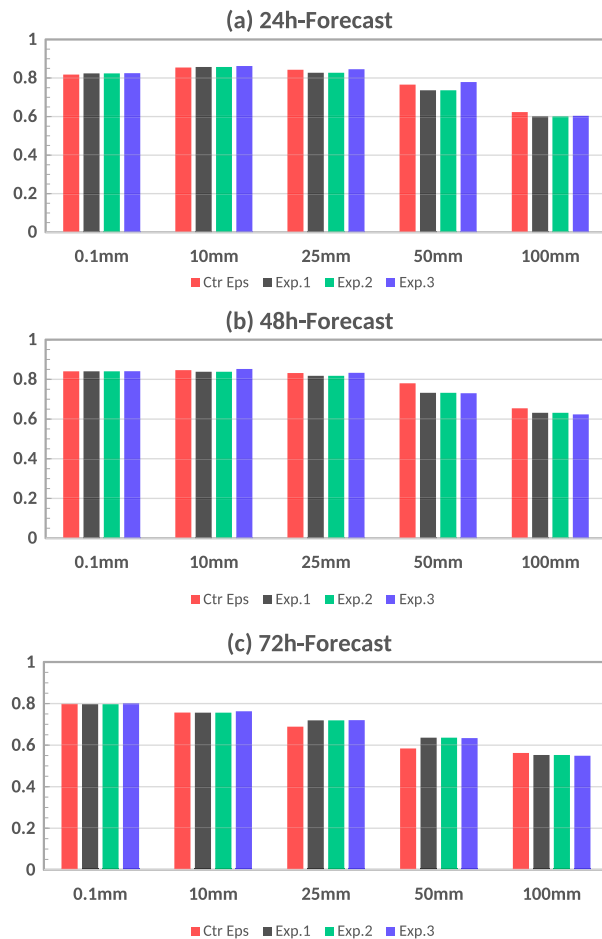
FIG. 7. AROC for the (a) 24-, (b) 48-, and (c) 72-h forecasts of 24-h accumulated precipitation exceeding 0.1, 10, 25, 50, and 100 mm, for the control EPS, Exp. 1 (SPPT), Exp. 2 (LTBC), and Exp. 3 (the unified scheme).

of fields with more measurement scores. Figure 8 shows the scorecard used for this implementation, to assess if the upgrade system (the unified scheme) can consistently outperform the current system (the control EPS). In this scorecard, six scores are computed for some isobaric fields including geopotential height H, temperature T, zonal wind U, and meridional wind V at 200-, 500-, 700-, 850-, and 1000-hPa levels, as well as some near-surface fields such as 2-m temperature (T2m), 10-m wind (U10m, V10m), and light, moderate, and heavy precipitation at 24-, 48-, and 72-h forecast lead times. These six measurement scores are selected to portray a full picture of the EPS quality for ensemble mean, forecast uncertainty and probabilistic forecasts. For nonprecipitation fields the verification metrics are RMSE, consistency (RMSE/spread), CRPS, and outlier; for precipitation AROC and BS are employed.

There are a total of 294 categories in Fig. 8. The green color indicates an improvement, the red a degradation,

and the gray is neutral (similar performance) for the unified scheme compared to the control EPS. The statistical significance level associated with an improvement or degradation is also marked on the figure. We can see that the improvement rate is 62.6% (184/294), the neutral rate is 35.4% (104/294), and the degradation rate is 2.0% (6/294). All the improvements are statistically significant. Consistent with the verification results in sections 3b(1) and 3b(2), more improvements are seen in the height and temperature fields, and less improvement in the moisture field including precipitation and upper-air wind. The improvements in surface wind and temperature are also apparent. Evidently, the overall improvement of the unified scheme over the control EPS is overwhelming, which leads to a guaranteed approval of its operational implementation.

c. A side experiment

Although the theme of this study is the unified scheme to deal with the random and systematic errors at the same time in an ensemble model, one might be interested in learning how this unified dynamical approach compares to the current two-step approach (dynamic and statistics) in performance. To shed light on this, we included a comparison between the dynamical bias correction method and a statistical postprocessing method (see Fig. A2 in the appendix). The result shows that the dynamical method outperformed the statistical method by about 64%. Given that the SPPT part is the same in both approaches, the unified scheme should also outperform the two-step approach (i.e., the SPPT in the ensemble model integration followed by a statistical postprocessing of model bias). Additionally, the statistical postprocessing method also has severe limitations. For example, only a limited number of selected variables can be processed, and each variable is processed independently with no dynamical constraints among the processed variables. These deficiencies will cause dynamical inconsistency among variables and prevent the bias corrected variables from being used in certain applications, such as initializing a downstream model. In other words, performance is not the sole motivation of our study but, more importantly, the feasibility of a certain NWP downstream application is.

4. Summary

NWP models have both random and systematic errors. Ensemble perturbation methods deal with random errors, while statistical bias correction methods deal with systematic errors. With current NWP technology, these two types of error cannot be dealt with together
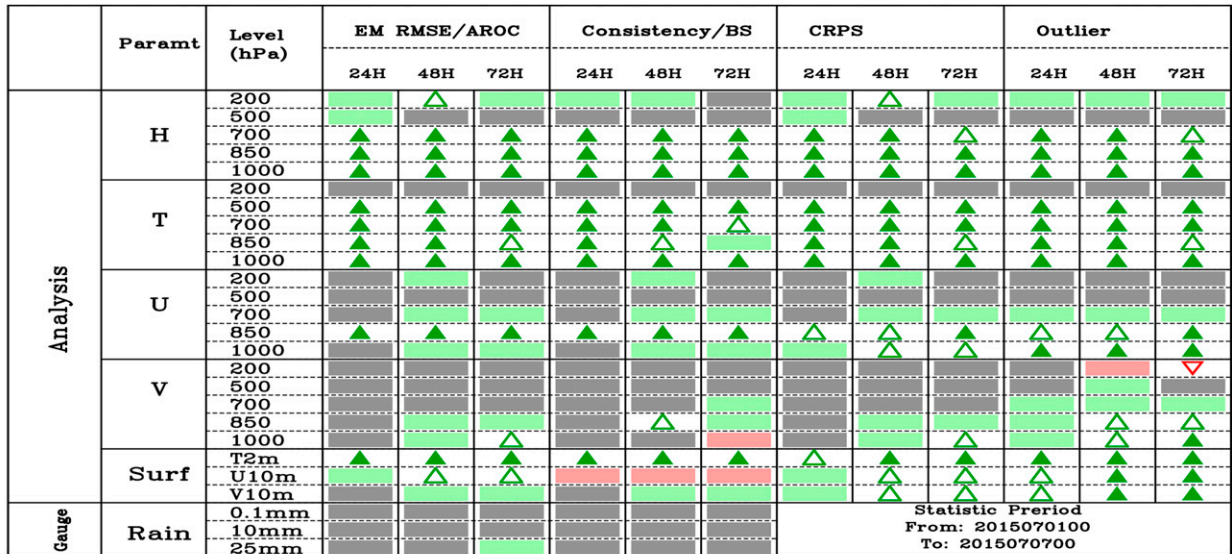
FIG. 8. The verification scorecard used for making an operational implementation decision whether an upgrade system can consistently outperform the current system, where the RMSE and consistency are for no-precipitation forecasts and the AROC and BC are for precipitation forecasts. In this case, the upgrade system is Exp. 3 (GRAPES-REPS-LTBC-SPPT) and the current system is the control EPS (GRAPES-REPS). Green indicates an improvement, gray is neutral, and red is a degradation. Filled triangles indicate that the differences are statistically significant at the 99.7% level (*t* test); unfilled triangles indicate 95% confidence; shaded rectangles indicate 75% confidence.

during model integration. Very recently, we introduced a three-dimensional wholesale-like dynamical method to remove a model's systematic bias in all variables during model integration (C19). This new model-integration-based bias correction approach has made it possible for a unified scheme to deal with both random and systematic errors at the same time as the model integrates. This study designed and tested such a unified scheme with a regional ensemble model, GRAPES-REPS, aiming to maximize the improvement in ensemble prediction skill by reducing both random and systematic errors at the same time in an operational environment.

Three experiments were performed on top of the control EPS (GRAPES-REPS). The first experiment used only the stochastic physics SPPT (GRAPES-REPS-SPPT). The second experiment used only the bias correction scheme LTBC (GRAPES-REPS-LTBC). The third experiment used both the LTBC and SPPT (GRAPES-REPS-LTBC-SPPT). The experimental period is 1–31 July 2015 (0000 UTC cycle) over the China domain. The averaged result of these 31 days showed that

1) The stochastic physics SPPT can effectively increase ensemble spread, while the bias correction LTBC

had little impact. As a result, ensemble averaging of the SPPT run can reduce random error more than the LTBC.

2) The stochastic physics SPPT had little impact on the systematic error, while the bias correction LTBC can significantly reduce it. As a result, the SPPT only slightly improved the accuracy of ensemble mean forecasts but the LTBC greatly reduced the ensemble mean forecast error.

3) By combining the stochastic physics SPPT and the bias correction LTBC into a unified scheme, both random and systematic errors can be significantly reduced at the same time. As a result, the unified scheme performed the best among the three experiments.

4) The CRPS scores show that the unified scheme can significantly increase the accuracy of probabilistic forecasts for all six selected atmospheric variables (temperature, height, moisture, and wind at the 850-hPa level, 10-m wind, and 2-m temperature). Besides improved accuracy, the reliability of the probabilistic forecasts was also significantly increased for the 850-hPa temperature, 850-hPa height, and 2-m temperature but remained about the same for moisture and wind.

5) Relative to other atmospheric variables, the improvement in probabilistic precipitation forecasts was much less with the unified scheme. However, the improvement seemed to increase as the forecast length increases. For example, there was no significant improvement in BS for all rainfall categories at the 24-h forecast; there was a moderately significant improvement for two categories (10 and 25 mm) at the 48-h forecast and for three categories (10, 25, and 50 mm) at the 72-h forecast. This trend is also observed in terms of AROC: there was a significant improvement for one rainfall category at 24 h and for two categories at the 48-h forecast, while three rainfall categories were significantly improved at the 72-h forecast.

6) Finally a scorecard was used to assess the suitability of the unified scheme for an operational upgrade. The result showed that the unified scheme can consistently outperform the control EPS. Out of the 294 categories contained in the scorecard, the improvement rate is 62.6% (184/294), the neutral rate is 35.4% (104/294), and the degradation rate is negligibly low at 2.0% (6/294). All the improvements are statistically significant.

Further improvements to both ensemble perturbation and bias correction techniques are obviously needed to improve moisture (including precipitation) and wind forecasts in a future study. Based on this study, we encourage the NWP community to adopt this unified scheme approach in their EPSs to achieve the best forecasts in operations.

APPENDIX

Estimation of the Linear Bias Tendency and a Comparison to a Statistical Bias Correction Method

For example, in our case of a 72-h model integration initialized on 1 July 2015, the bias was approximated by
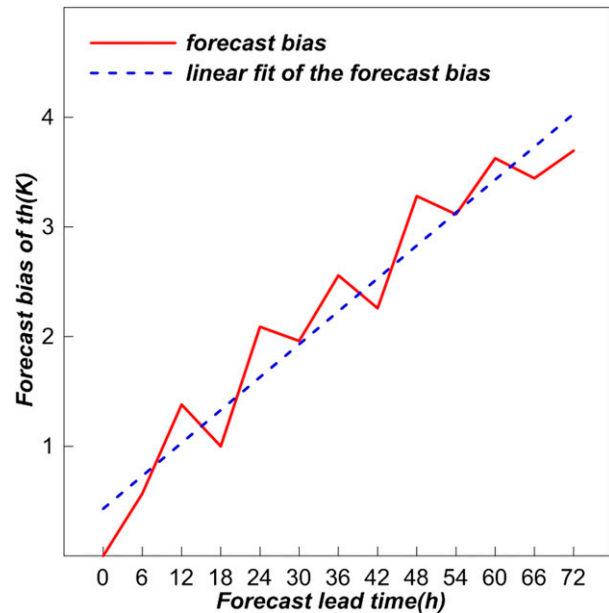

FIG. A1. The forecast bias (red curve) of the ensemble control member for potential temperature th (θ; K) on a grid point (31°N, 114°E) near the 700-hPa level over forecast hours. It is estimated from the 0000 UTC cycle forecasts during 19–28 Jun 2015 and approximated as the biases for the 72-h model integration initialized at 0000 UTC 1 Jul 2015. The blue dashed line is the linear fit of the bias error.

the average error of the past 10 days of forecasts during the period of 19–28 June 2015 at each forecast hour. To retain some recent season and flow-dependent bias information, a moving 10-day (but not longer) time period is used to estimate the bias error. Once the bias error is available, the linear bias tendency $\hat{B}_l(S_j, t)$ can be estimated from the variation of bias error with forecast time over a desired time window Δ. The red curve in Fig. A1 is the estimated bias error of potential temperature varying with forecast hour (0–72 h at 6-h intervals) at a model grid point. By applying a linear regression to the 13 bias values over this 72-h window, a linear fit of the bias tendency with time can be obtained (blue line in Fig. A1). Using the bias increment $\hat{b}$ over the time window Δ (72 h in this case), the linear bias tendency $\hat{B}_l(S_j, t)$ over a time step δt (in seconds, $\delta t = 60$ s in this case) can be obtained as the following:

$$\widehat{B}_l(S_j, t) = \text{slope} \times \text{time step} = \frac{\hat{b}}{\Delta \times 3600} \times \delta t, \quad (A1)$$

which is the same as Eq. (9) in section 2c. The linear bias tendency obtained is then used for all time steps within this time window Δ. By repeating the above steps on every model grid point at all model levels, a three-dimensional $\hat{B}_l(S_j, t)$ can be obtained at every model integration time step for a model state variable (potential
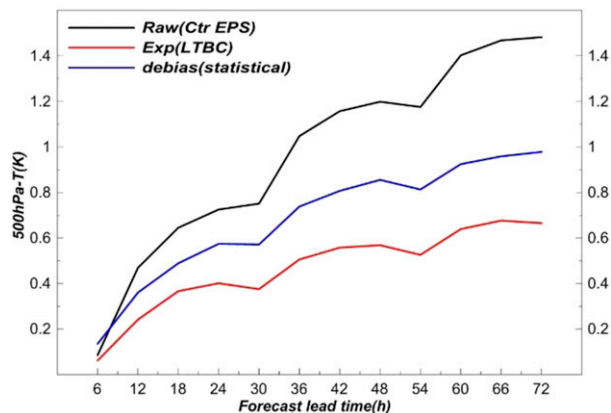
FIG. A2. The domain-averaged biases of the ensemble mean varying with forecast hour from the raw (Ctr EPS) and the experiment (LTBC) as well as the Kalman filter–based statistical method ("debias", applied to the Ctr EPS). The results are the 10-day average for the 0000 UTC cycle during 1–10 Jul 2015. The variable is 500-hPa temperature.

temperature in this case). Note that the selection of a shorter time window Δ may include more time-evolution information on the bias. However, C19 compared a 3-h window with the 72-h window and found a mixed result (improvement in some fields and degradation in other fields). C19 also compared a multistate-variable-based (potential temperature, wind, and pressure) experiment with a single-state-variable-based (potential temperature only) experiment and found no obvious benefit. As a matter of fact, we found that the single potential temperature–based method using the 72-h window performed the best overall. More research on this subject is currently under the way to understand all these findings.

To get an idea about the relative performance of this new bias correction method, we compared it with the Kalman filter–based (or decaying average) statistical bias correction method that is currently used in operations at both CMA and NCEP (Cui et al. 2012; Du and Zhou 2011). The result from this statistical method over the same time period (1–10 July 2015) for the 500-hPa temperature is shown by the blue curve in Fig. A2. Apparently, the performance of the new method (the brown curve) is significantly better (at the 99.9% level) than the current operational statistical method. For example, at a 72-h forecast, the domain-averaged bias error reduction is 33.3% for the statistical method and 54.7% for the new method, which is ~64% greater reduction achieved by the new method than by the statistical method. This is a very encouraging result.

REFERENCES

Bakhshaii, A., and R. Stull, 2009: Deterministic ensemble forecast using gene-expression programming. *Wea. Forecasting*, **24**, 1431–1451, https://doi.org/10.1175/2009WAF2222192.1.

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

Betts, A. K., 1986: A new convective adjustment scheme. Part I: Observation and theoretical basis. *Quart. J. Roy. Meteor. Soc.*, **112**, 667–691, https://doi.org/10.1002/qj.49711247307.

Buizza, R., M. Milleer, and T. N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **125**, 2887–2908, https://doi.org/10.1002/qj.49712556006.

——, J. Du, Z. Toth, and D. Hou, 2018: Major operational ensemble prediction systems (EPS) and the future of EPS. *Handbook of Hydrometeorological Ensemble Forecasting*, Q. Duan et al., Eds., Springer, 1–43, https://doi.org/10.1007/978-3-642-40457-3_14-1.

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, https://doi.org/10.1175/2009MWR3187.1.

Chen, D.-H., J.-S. Xue, and X.-S. Yang, 2008: New generation of multi-scale NWP system (GRAPES): General scientific design. *Chin. Sci. Bull.*, **53**, 3433–3445, https://doi.org/10.1007/s11434-008-0494-z.

Chen, J., D.-H. Chen, and H. Yan, 2002: A brief review on the development of ensemble prediction system. *J. Appl. Meteor. Sci.*, **13**, 497–507.

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, https://doi.org/10.1175/JAS-D-14-0250.1.

Cui, B., Z. Toth, and Y. Zhu, 2006: The trade-off in bias correction between using the latest analysis/modeling system with a short, versus an order system with a long archive. *The First THORPEX Int. Science Symp.*, Montreal, Canada, World Meteorological Organization, 281–284.

——, ——, ——, and D. Hou, 2012: Bias correction for global ensemble forecast. *Wea. Forecasting*, **27**, 396–410, https://doi.org/10.1175/WAF-D-11-00011.1.

Du, J., 2007: Uncertainty and ensemble forecast. NOAA/NWS Science and Technology Infusion Lecture Series, accessed 25 September 2019, https://www.nws.noaa.gov/ost/climate/STIP/uncertainty.htm.

——, and G. DiMego, 2008: A regime-dependent bias correction approach. *19th Conf. on Probability and Statistics,* New Orleans, LA, Amer. Meteor. Soc., 3.2, https://ams.confex.com/ams/88Annual/webprogram/Paper133196.html.

——, and B. Zhou, 2011: A dynamical performance-ranking method for predicting individual ensemble member performance and its application to ensemble averaging. *Mon. Wea. Rev.*, **139**, 3284–3303, https://doi.org/10.1175/MWR-D-10-05007.1.

——, and ——, 2017: Ensemble fog prediction. *Marine Fog: Challenges and Advancements in Observations, Modeling, and Forecasting*, D. Koracin and C. E. Dorman, Eds., Springer, 477–509, https://doi.org/10.1007/978-3-319-45229-6_10.

——, G. DiMego, M. S. Tracton, and B. Zhou, 2003: NCEP short-range ensemble forecasting (SREF) system: Multi-IC, multi-model and multi-physics approach. *Research Activities in Atmospheric and Oceanic Modelling,* J. Cote,

Ed., WMO/TD-1161, Rep. 33, CAS/JSC Working Group Numerical Experimentation (WGNE), 5.09–5.10, https://www.emc.ncep.noaa.gov/mmb/SREF/srefWMO_2003.pdf.

——, and Coauthors, 2018: Ensemble methods for meteorological predictions. *Handbook of Hydrometeorological Ensemble Forecasting,* Q. Duan et al., Eds., Springer, 1–52, https://doi.org/10.1007/978-3-642-40457-3_13-1.

Dudhia, J., 1989: Numerical study of convection observed during the winter monsoon experiment using a mesoscale two-dimensional model. *J. Atmos. Sci.,* **46,** 3077–3107, https://doi.org/10.1175/1520-0469(1989)046<3077:NSOCOD>2.0.CO;2.

Feng, J., R. Q. Ding, D. Q. Liu, and J. P. Li, 2014: The application of nonlinear local Lyapunov vectors to ensemble predictions in the Lorenz systems. *J. Atmos. Sci.,* **71,** 3554–3567, https://doi.org/10.1175/JAS-D-13-0270.1.

——, J. Li, R. Ding, and Z. Toth, 2018: Comparison of nonlinear local Lyapunov vectors and bred vectors in estimating the spatial distribution of error growth. *J. Atmos. Sci.,* **75,** 1073–1087, https://doi.org/10.1175/JAS-D-17-0266.1.

Gneiting, T., A. E. Raftery, A. H. Westfveld III, and T. Godman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.,* **133,** 1098–1118, https://doi.org/10.1175/MWR2904.1.

Hong, S.-Y., and H.-L. Pan, 1996: Nonlocal boundary layer vertical diffusion in a medium range forecast model. *Mon. Wea. Rev.,* **124,** 2322–2339, https://doi.org/10.1175/1520-0493(1996)124<2322:NBLVDI>2.0.CO;2.

——, and J.-O. Lim, 2006: The WRF Single Moment 6-Class Microphysics Scheme (WSM6). *J. Korean Meteor. Soc.,* **42** (2), 129–151.

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

Houtekamer, P. L., L. Lefaivre, J. Derome, H. Ritchie, and H. L. Mitchell, 1996: A system simulation approach to ensemble prediction. *Mon. Wea. Rev.,* **124,** 1225–1242, https://doi.org/10.1175/1520-0493(1996)124<1225:ASSATE>2.0.CO;2.

Jolliffe, I. T., and D. B. Stephenson, Eds., 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science.* Wiley Press, 543 pp.

Kain, J. S., 2004: The Kain–Fritsch convective parameterization: An update. *J. Appl. Meteor.,* **43,** 170–181, https://doi.org/10.1175/1520-0450(2004)043<0170:TKCPAU>2.0.CO;2.

——, and J. M. Fritsch, 1990: A one-dimensional entraining/detraining plume model and its application in convective parameterization. *J. Atmos. Sci.,* **47,** 2784–2802, https://doi.org/10.1175/1520-0469(1990)047<2784:AODEPM>2.0.CO;2.

——, and ——, 1993: Convective parameterization for mesoscale models: The Kain–Fritsch scheme. *The Representation of Cumulus Convection in Numerical Models, Meteor. Monogr.,* No. 46, Amer. Meteor. Soc., 165–170, https://doi.org/10.1007/978-1-935704-13-3_16.

Kazuo, S., H. Masahiro, K. Masaru, S. Hiromu, and Y. Munehiko, 2012: Comparison of initial perturbation methods for the mesoscale ensemble prediction system of the Meteorological Research Institute for the WWRP Beijing 2008 Olympics Research and Development Project (B08RDP). *Tellus,* **63A,** 445–467, https://doi.org/10.1111/j.1600-0870.2010.00509.x.

Li, L., Y.-L. Li, and H. Tian, 2011: Study of bias-correction in T213 global ensemble forecast (in Chinese). *Meteor. Mon.,* **37** (1), 31–38.

Li, X., M. Charron, L. Spacek, and G. Candille, 2008: A regional ensemble prediction system based on moist targeted singular vectors and stochastic parameter perturbations. *Mon. Wea. Rev.,* **136,** 443–462, https://doi.org/10.1175/2007MWR2109.1.

Ma, X., J. Xue, and W. Lu, 2008: Preliminary study on ensemble transform Kalman filter based initial perturbation scheme in GRAPES global ensemble prediction (in Chinese). *Acta Meteor. Sin.,* **66** (4), 526–536.

——, Y. Shi, J. He, Y. Ji, and Y. Wang, 2015: The combined descending averaging bias correction based on the Kalman filter for ensemble forecast (in Chinese). *Acta Meteor. Sin.,* **73** (5), 952–964.

Mahrt, L., and M. Ek, 1984: The influence of atmospheric stability on potential evaporation. *J. Climate Appl. Meteor.,* **23,** 222–234, https://doi.org/10.1175/1520-0450(1984)023<0222:TIOASO>2.0.CO;2.

Mlawer, E. J., S. J. Taubman, P. D. Brown, M. J. Iacono, and S. A. Clough, 1997: Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave. *J. Geophys. Res.,* **102,** 16 663–16 682, https://doi.org/10.1029/97JD00237.

Monache, L. D., T. Nipen, and X. X. Deng, 2006: Ozone ensemble forecasts: 2. A Kalman filter predictor bias correction. *J. Geophys. Res.,* **111,** D05308, https://doi.org/10.1029/2005JD006311.

Morris, A. S., and R. Langari, 2016: Statistical analysis of measurements subject to random errors. *Measurement and Instrumentation: Theory and Application*, 2nd ed. ScienceDirect, 75–130, https://doi.org/10.1016/B978-0-12-800884-3.00004-6.

Noilhan, J., and S. Planton, 1989: A simple parametrization of land surface processes for meteorological models. *Mon. Wea. Rev.,* **117,** 536–549, https://doi.org/10.1175/1520-0493(1989)117<0536:ASPOLS>2.0.CO;2.

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

Pan, H.-L., and W.-S. Wu, 1995: Implementing a mass flux convective parameterization package for the NMC Medium-Range Forecast Model. NMC Office Note 409, Washington, DC, 40 pp., http://www.emc.ncep.noaa.gov/officenotes/FullTOC.html.

Pan, Y., Y. Shen, J. Yu, and A. Xiong, 2015: An experiment of high-resolution gauge-radar-satellite combined precipitation retrieval based on the Bayesian merging method. *Acta Meteor. Sin.,* **73** (1), 177–186.

Privé, N. C., R. M. Errico, and K. S. Tai, 2013: Validation of the forecast skill of the Global Modeling and Assimilation Office observing system simulation experiment. *Quart. J. Roy. Meteor. Soc.,* **139,** 1354–1363, https://doi.org/10.1002/qj.2029.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.,* **133,** 1155–1174, https://doi.org/10.1175/MWR2906.1.

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

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

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

Tracton, S., and E. Kalnay, 1993: Operational ensemble prediction at the National Meteorological Center: Practical aspects. *Wea. Forecasting*, **8**, 379–400, https://doi.org/10.1175/1520-0434(1993)008<0379:OEPATN>2.0.CO;2.

Wang, J., J. Chen, J. Du, Y. Zhang, Y. Xia, and D. Guo, 2018: Sensitivity of ensemble forecast verification to model bias. *Mon. Wea. Rev.*, **146**, 781–796, https://doi.org/10.1175/MWR-D-17-0223.1.

Yuan, Y., X.-L. Li, J. Chen, and Y. Xia, 2016: Stochastic parameterization toward model uncertainty for the GRAPES mesoscale ensemble prediction system (in Chinese). *Meteor. Mon.*, **42** (10), 1161–1175.

Zhang, H., J. Chen, and X. Zhi, 2014: Design and comparison of perturbation schemes for GRAPES-MESO based ensemble forecast (in Chinese). *Trans. Atmos. Sci.*, **37** (3), 276–284.

Zhuang, Z., J. Xue, and H. Lu, 2014: Experiments of Global GRAPES-3DVar analysis based on pressure level and prediction system (in Chinese) *Plateau Meteor.*, **33** (3), 666–674.