Received: 11 April 2019 | Revised: 24 October 2019 | Accepted: 2 December 2019 | Published on: 27 January 2020

DOI: 10.1002/qj.3730

Quarterly Journal of the Royal Meteorological Society RMetS

RESEARCH ARTICLE

Forecast bias correction through model integration: A dynamical wholesale approach

Jing Chen[1] | Jingzhuo Wang[1] | Jun Du[2] | Yu Xia[3] | Fajing Chen[1] | Li Hongqi[1]

[1]Numerical Weather Prediction Center/CMA, Beijing, China

[2]Environmental Modeling Center/NCEP/NWS/NOAA, College Park, Maryland

[3]College of Atmospheric Sciences, Nanjing University of Information Science and Technology, Nanjing, China

Correspondence
J. Chen, Numerical Weather Prediction Center, China Meteorological Administration, Beijing 100081, China.
Email: chenj@cma.gov.cn

Funding information
National Key R&D Program of China, 2018YFC1507405

Abstract

Unlike the retail-like (for selected variables) statistical post-processing methods, a wholesale-like (for all variables) dynamical approach is proposed to correct forecast bias during model integration. Subtracting a bias tendency from the model total tendency is intended to de-bias all variables at once to better (i.e. more dynamically consistent) couple with downstream applications. Three experiments were tested using an ensemble prediction system since the method is intended for an ensemble model. The verification was carried out over China for a period of 31 days (1–31 July 2015). The verification of 500 hPa temperature indicates that all three experiments have significantly improved the raw ensemble forecasts with reduced bias error, a more accurate ensemble mean, a better spread-skill relationship, and more reliable and sharper probabilities. The performance is better than or comparable to the current operational statistical method. When the verification was expanded to include more variables, a summary scorecard shows that the three experiments also had a general positive or neutral impact on both upper-air and surface variables, especially the height and temperature fields. Precipitation forecasts remained relatively unchanged. There were only a few categories that were degraded. The comparison between the three experiments yielded a mixed result: the most sophisticated approach often performed the best for 500 hPa temperature, while the simplest approach worked the best when verifying a mixture of variables. The degradation of the wind forecasts by the third experiment was discussed. These are the two challenges: how to accurately describe the bias tendency and how to add internally coherent bias tendencies to multiple variables. Given its advantages, this approach could be a promising approach for correcting biases in a numerical model.

KEYWORDS

bias correction, ensembles, numerical model

1 | INTRODUCTION

Our recent study shows that ensemble performance and verification is very sensitive to model bias (Wang *et al.*, 2018). Therefore, systematic model bias has to be removed from each ensemble member in order to maximize the forecast utility as well as allow for a correct assessment of an ensemble prediction system (EPS). Current methods to remove bias are retail-like (only for some selected variables), mainly statistical, and done separately from the model integration as a post-model correction (Roulston and Smith, 2003; Gneiting *et al.*, 2005; Monache *et al.*, 2005; Raftery *et al.*, 2005; Bakhshaii and Stull, 2009; Du and Zhou, 2011; Cui *et al.*, 2012; Satterfield and Bishop, 2014). This requires the addition of an extra step to remove the bias before utilizing and verifying an ensemble of forecasts. Besides the inconvenience, this causes a serious problem or even a stoppage to some downstream applications if one wants to use bias-corrected fields, due to dynamical inconsistency among variables. The current methods correct only a small subset of all model output variables (it is almost impossible to correct everything in an operational environment) and correct each of those variables independently. As a consequence, the final forecast products (often multi-variable based) sometimes behave erroneously due to the inconsistency among the ingredient variables (i.e. a mixed use of some independently corrected and other uncorrected variables). For example, an inconsistency between winter precipitation type and surface temperature was seen (National Centers for Environmental Prediction (NCEP) Weather Prediction Center, personal communication). A poor replication of the convective environment (such as convective available potential energy, CAPE) might be produced due to the inconsistently corrected temperature and moisture fields (NCEP Storm Prediction Center, personal communication). In other cases, one simply cannot use the bias'-corrected variables in one's application. For example, it is impossible to use bias-corrected fields to initialize a downstream model like a nested domain (including a concurrently running nest like the NCEP FV3 nested run), an air-quality or a dispersion model, due to the inconsistency among input variables. Therefore, we need a new type of bias correction approach to overcome all these difficulties in numerical weather prediction (NWP). An ideal approach would be: (a) de-biasing all the model output fields (not only a few selected fields), including all derived fields like precipitation and clouds, together in a dynamically consistent way, (b) doing the de-biasing during the model integration with no extra step needed following the model integration, and (c) having the capability to remove a big portion of the bias errors. This study will design and test such a new type of wholesale-like approach: bias correction through model integration. Results will be assessed through the verification of ensemble forecasts. It will demonstrate that this approach can perfectly satisfy the first two criteria. As a first attempt with this method, we do not expect that it will remove all model biases but hope that it will perform better than or at least be comparable to a commonly used statistical method. Therefore, a comparison with a current operational statistical method will also be given in this study. More importantly, issues we encountered in this study will be presented and discussed for future research to improve the method.

The dynamical de-biasing concept is not entirely new in literature. It can be traced all the way back to the work of Leith (1978). The method has already been applied to a simplified weather model with promising results by Danforth *et al.* (2007). Following the work of Danforth *et al.* (2007), work is ongoing to apply the method to an operational NWP model as documented in Bhargava *et al.* (2018). Independently, a group at the UK Met Office (UKMO) developed a similar methodology, to be used in their global EPS, which tries to replace the model's stochastic physics perturbations where the model error tendency is estimated from archived analysis increments (Piccolo and Cullen, 2016; Piccolo *et al.*, 2019), assuming that analysis increments are a good proxy to diagnose model errors. The UKMO method is the same concept as this study, although the approach proposed by Piccolo and Cullen (2016) differs from this work in that it accomplishes de-biasing and spread inflation at the same time, whereas, in this study, spread inflation is accounted for by using different parametrization schemes in different members (Table 1). As an extension of the work of this study, a unified scheme has also been proposed to accomplish de-biasing and spread inflation at the same time by applying this bias-correction method and a stochastic physics scheme together (Xia *et al.*, 2019). The UKMO work might be in alignment with this extended work of ours.

With the merging of the dynamical de-biasing approach, we invite other researchers to join us to further refine this technique to improve its effectiveness in removing the bias. Based on this study, some possible areas for future improvement will be discussed at the end of this article. The scope of this study is to design and demonstrate this new type of method to float an idea to the NWP community rather than complete a systematic verification study. The rest of this article is organized as follows. The model and methodology are described in Section 2. In Section 3 the verification results are given through three experiments. A summary and discussion are presented in Section 4.

TABLE 1 Configuration of the GRAPES_REPS

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

2 | MODEL AND METHODOLOGY

A regional version of the Global and Regional Assimilation and Prediction Enhanced System (GRAPES) is the base model employed in this study, which is developed at the Numerical Weather Prediction Center of China Meteorological Administration (CMA) (Chen *et al.*, 2008). The main features of GRAPES include a fully compressible dynamical core with non-hydrostatic approximation, a semi-implicit and semi-Lagrangian scheme for time integration, and a height-based terrain-following sigma coordinate. The model physics includes RRTM (Rapid Radiative Transfer Model) long-wave radiation (Mlawer *et al.*, 1997), Dudhia short-wave radiation (Dudhia, 1989), the Weather Research and Forecasting Single-Moment 6-class (WSM-6) microphysics (Hong and Lim, 2006), Noah land surface model (Mahrt and Ek, 1984), Medium Range Forecast (MRF) planetary boundary-layer (PBL) scheme (Hong and Pan, 1996), and Monin–Obukhov surface layer scheme (Noilhan and Planton, 1989). Model analysis is produced by a 4-dimensional variable data assimilation scheme, available every 6 hr. Based on the GRAPES model, a Regional Ensemble Prediction System (GRAPES-REPS, Table 1) was also developed and is running operationally at CMA (Zhang *et al.*, 2014). It has 15 ensemble members (1 control and 14 perturbed members) with 51 vertical levels (model top is 10 hPa) and a horizontal resolution of 15 km. Initial and boundary condition uncertainties are provided by different members of a global EPS also operationally running at CMA. The initial condition perturbations of the global EPS are generated by the breeding vector (Toth and Kalnay, 1997). Model perturbation of the GRAPES-REPS is represented by multiple physics schemes (Stensrud *et al.*, 2000; Du *et al.*, 2015). GRAPES-REPS runs twice a day, initialized at 0000 and 1200 UTC, respectively, out to a 72 hr forecast lead time. The model integration time step is 60 s. There is no perturbation in the control member. It is the GRAPES-REPS that will be used in this study.

Following a stochastic physics perturbation approach (Houtekamer *et al.*, 1996; Buizza *et al.*, 1999; Shutts, 2005; Berner *et al.*, 2009; Ollinaho *et al.*, 2017), a bias-correction forcing is added to the model total tendency term of a state variable S at every time step during a model's integration, with the intent that it will produce bias-free forecasts for

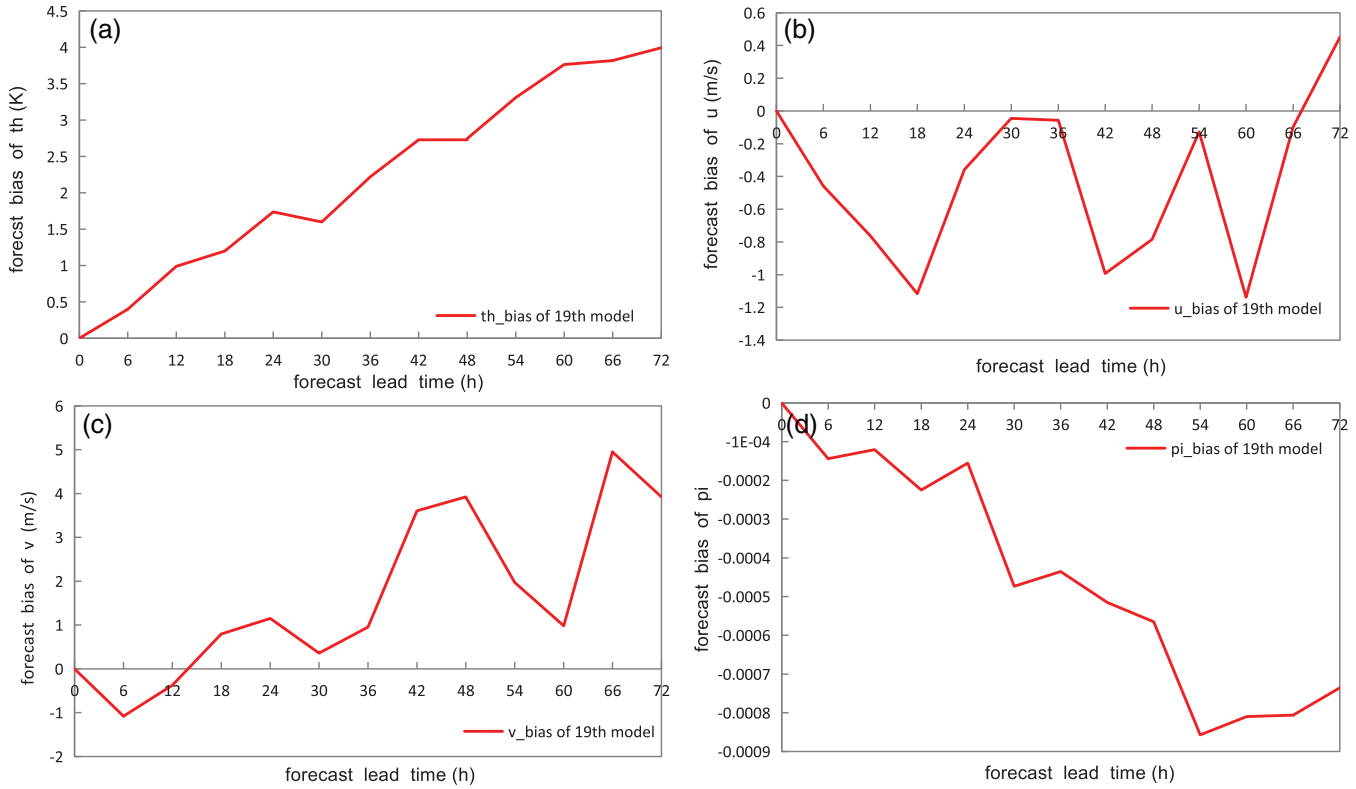

FIGURE 1 The forecast biases of the ensemble control member for (a) potential temperature th (θ, K), (b) zonal wind u (m·s^{-1}), (c) meridional wind v (m·s^{-1}), and (d) dimensionless pressure pi (π) on a grid point (114°E, 31°N) near 700 hPa level over forecast hours. It is estimated from the 0000 UTC cycle forecasts during 19–28 June 2015 and approximated as the biases for the 72 hr model integration initialized at 0000 UTC 1 July 2015

all variables. Equation 1 is the model integration formula:

$$S_j(t) = \int_{t=0}^{t} \{A(S_j, t) + P(S_j, t)\} dt, \tag{1}$$

where $S_j(t)$ is a state variable of the jth ensemble member at model integration time t, $j = 0$ is the control forecast, and $j = 1, 2, \ldots, n$ represent n perturbed ensemble members ($n = 14$ in this study). A is the model dynamic tendency term, and P is the physical process tendency term. De-biasing is realized by subtracting a bias tendency B from dynamic and physical process tendencies during model integration, shown in Equation 2:

$$S_j(t) = \int_{t=0}^{t} \{A(S_j, t) + P(S_j, t) - B(S_j, t)\} dt. \tag{2}$$

Equation 2 is the theoretical formula for this new approach.

Bias tendency B can be estimated from the variation in available bias error with forecast time. For example, in our case, for a 72 hr model integration initialized on 1 July 2015, the bias was approximated by the average error of the old forecasts for the period of 19–28 June 2015 at each forecast hour. 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 also not too long to completely filter out flow-dependent error. Since bias is regime dependent (Du and DiMego, 2008), it should be beneficial to retain some recent flow-dependent bias information in the bias tendency. A period of about 10–20 days has been proven to be optimal for correcting regime-dependent bias in short-range weather forecasts, as shown by the experience of the US NCEP's Short Range Ensemble Forecast (Du *et al.*, 2015). To mimic an operational environment, the estimation of forecast bias was done directly on the GRAPES model native grid and levels using the GRAPES analysis (f00 files) as truth for the four state variables (potential temperature θ, zonal wind u, meridional wind v, and dimensionless pressure π), so that there was no horizontal or vertical grid interpolation error introduced. In spite of the fact that the estimated bias here is only relative to the GRAPES analysis but not to observations, this configuration is the only implementable way in real-time production. However, for forecast verification, the independent and best available ECMWF analysis will be used in the next section. Once we have bias on model grid, the needed bias tendency B can be derived from it. Figure 1 is the estimated biases varying with forecast hour (0–72 hr

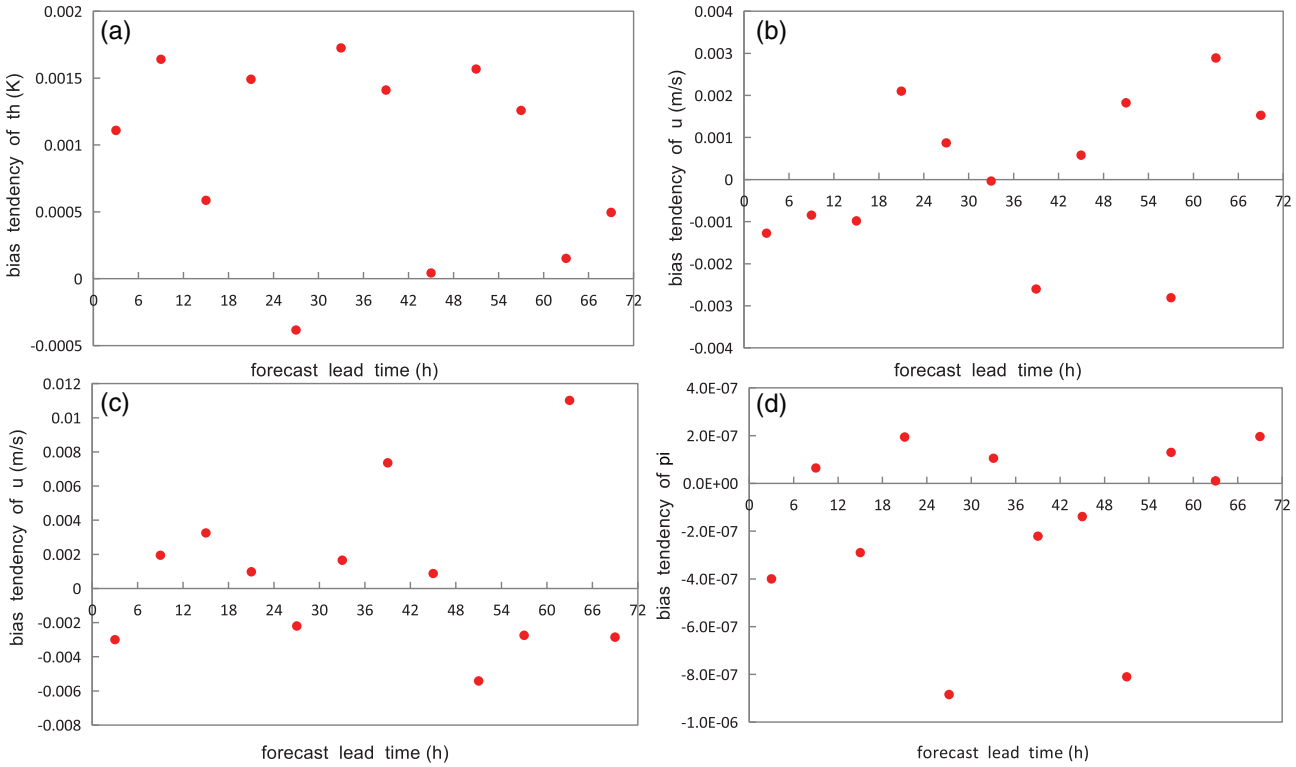

FIGURE 2 The 6 hr bias tendencies corresponding to Figure 1. Note that although the red dots are plotted only at the middle point of a 6 hr window, they really represent the bias tendency values over the entire 6 hr window

at 6 hr intervals) at a model grid point for the four state variables. Based on Figure 1, bias tendency can be calculated at any time interval. For example, the 6 hr bias tendency (Figure 2) is the change in bias between the two consecutive forecast hours in Figure 1. A 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. Therefore, bias tendency over a time step δt (in seconds, $\delta t = 60$ s in this case), denoted as $\hat{B}_l(S_j, t)$, can be obtained as the following:

$$\hat{B}_l(S_j, t) = slope \times time\ step = \frac{\hat{b}}{\Delta \times 3600} \times \delta t. \quad (3a)$$

Equation 3a becomes Equation 3b if there are only two bias values available within Δ:

$$\hat{B}_l(S_j, t) = \frac{B(S_j, t+\Delta) - B(S_j, t)}{\Delta \times 3600} \times \delta t. \quad (3b)$$

Thus, by repeating the above steps on every model grid point at all model levels within the model domain, a three-dimensional $\hat{B}_l(S_j, t)$ can be obtained at every model integration time step. With this, Equation 2 can be approximated into Equation 4,

$$S_j(t) = \int_{t=0}^{t} \{A(S_j, t) + P(S_j, t) - \hat{B}_l(S_j, t)\} dt, \quad (4)$$

which is the practical version of this proposed bias correction approach we are going to test in this study. This method can be applied to any NWP model. The difference between Equations 2 and 4 has two aspects: the omission of nonlinear bias tendency (unresolved temporal structures) and at least partially flow-dependent bias (unresolved spatial structures) by Equation 4. The omission of nonlinear bias tendency is due to the linear fitting used to interpret a large time interval (Δ) value into a time-step value (anything less than Δ time-scale is not resolved). Obviously, the smaller the interval Δ is, the closer $\hat{B}_l$ will be to B. The omission of flow-dependent bias is due to the time averaging in the bias estimation process over a past time period where only the systematic component (spatial structure) is retained. How to include full flow-dependent bias effect in Equation 4 is a challenging issue, where the singular value decomposition (SVD) method has been tried by Danforth *et al.* (2007). By the way, if bias estimation is done on a different grid and at levels other than the model native grid and levels, the process of spatial interpolation (both horizontal and vertical to model native grid and level) will introduce errors too.

The ensemble forecast verification metrics are selected based on Du and Zhou (2017) and Jolliffe and Stephenson (2003). In addition to ensemble mean, ensemble spread and probability distribution are the two important features

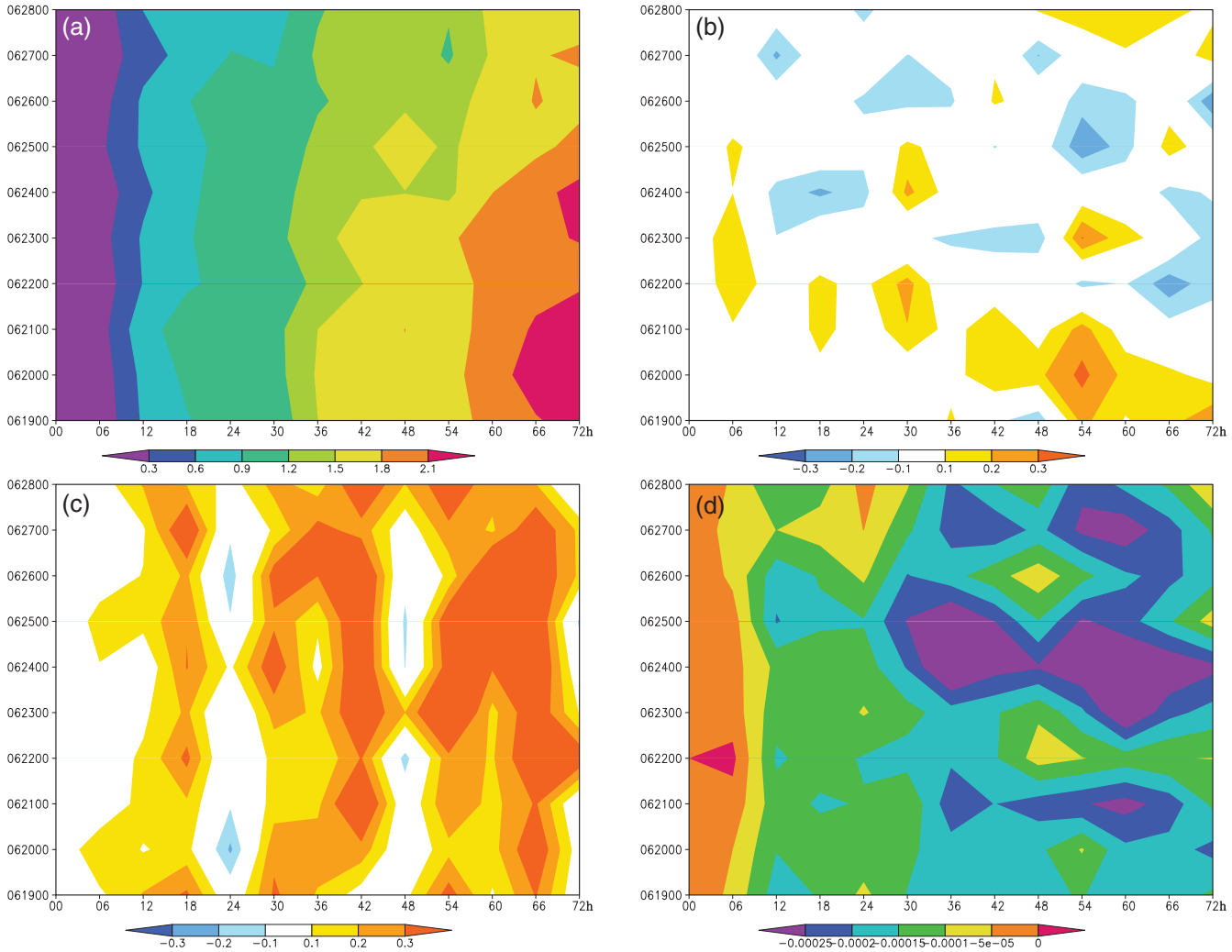

FIGURE 3 The 10-day (19–28 June 2015) evolution of the model domain-averaged forecast biases of the control member at different forecast hours for (a) potential temperature th (θ, K), (b) zonal wind u (m·s^{-1}), (c) meridional wind v (m·s^{-1}), and (d) dimensionless pressure pi (π) near 700 hPa level

to be verified in this study. Model bias is calculated separately for the 0000 and 1200 UTC cycles and obtained by averaging forecast errors over a time period of 10 days (for the reason already mentioned above) immediately prior to model integration. For example, for a 72 hr model forecast initialized at 0000 UTC on 1 July 2015, the model bias is obtained from the 0000 UTC cycle forecasts from 19 to 28 June 2015. In this study, the experimental period is 31 days from 1 to 31 July 2015 for forecasts initialized at 0000 UTC (i.e. a total of 31 72-hr forecasts) over China as a demonstration. The averaged results of these 31 days should be robust enough, and will be presented in the next section. The 6 hr ECMWF gridded analysis (https://apps.ecmwf.int/datasets/data/interim-full-daily/levtype=pl) is used as truth (except for precipitation) in verification (Sections 3.2 and 3.3), while the CMPAS-V2.1 (CMA Multi-source merged Precipitation Analysis System: Pan *et al.*, 2015) is used as truth for verifying precipitation.

3 | RESULTS

3.1 | Forecast bias analysis

Let us examine the forecast bias situation over a period of 10 days prior to our test period (1–31 July 2015). While Figure 1 is the 10-day (19–28 June 2015) averaged biases for a grid point, Figures 3 and 4 are for the domain average. Figure 3 shows the 10-day time evolution of the control member's biases at each forecast hour for the four state variables, at a level near 700 hPa. Apparently, persistent biases exist in all variables. The strongest bias is in the potential temperature (Figure 3a, it is about 50–80% relative to its total forecast error); a moderate bias is in both the dimensionless pressure (Figure 3d, about 10–20%) and the meridional wind (Figure 3c, about 10%); and a weak bias is in the zonal wind (Figure 3b, about 2%). For the zonal wind, moderate bias also exists at other levels as

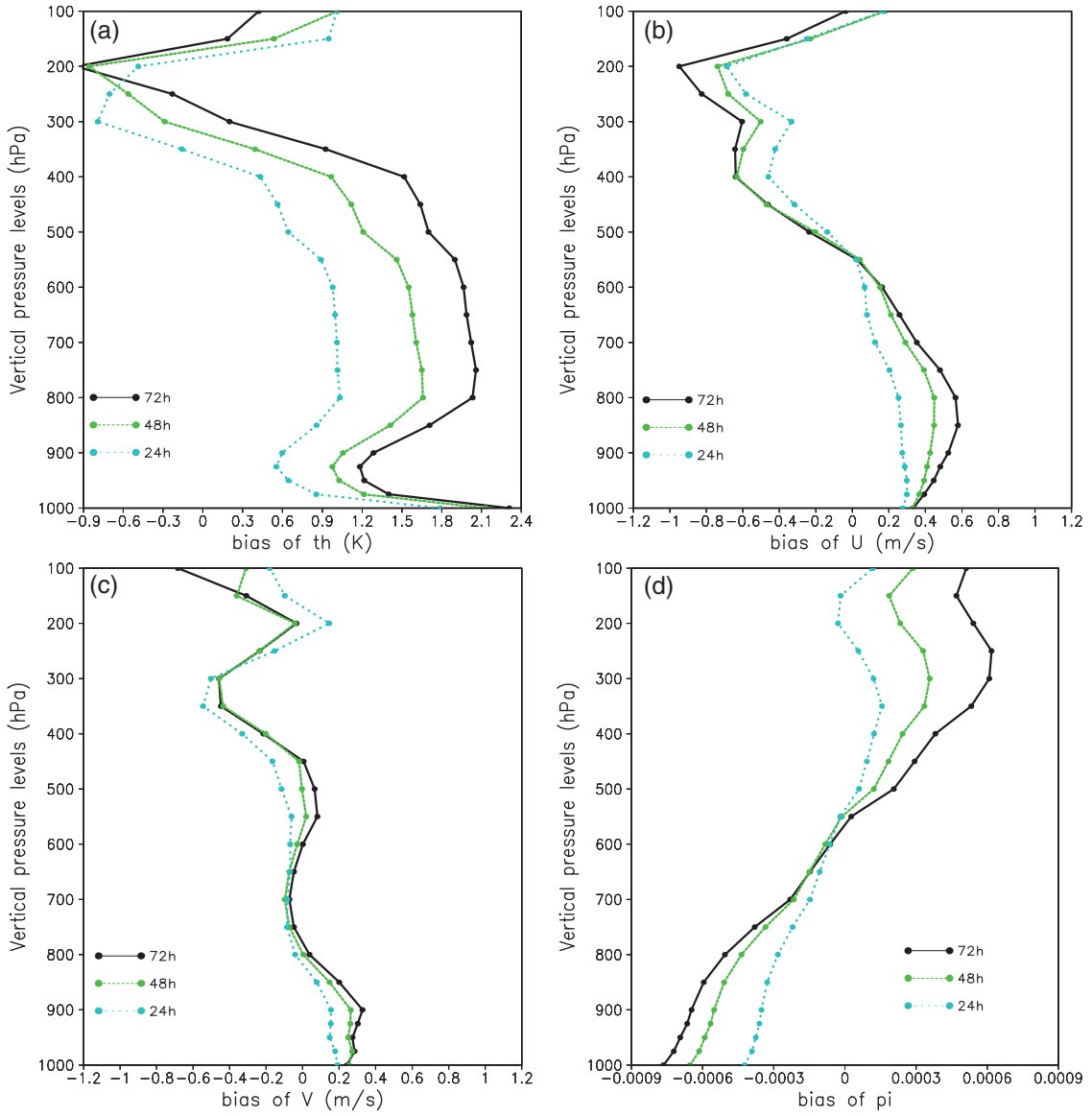

FIGURE 4 The vertical distribution of the domain-averaged biases at the forecast lead time of 24 (blue), 48 (green), and 72 h (black) for (a) potential temperature th (θ, K), (b) zonal wind u (m·s^{-1}), (c) meridional wind v (m·s^{-1}), and (d) dimensionless pressure pi (π). The results are the 10-day average for the 0000 UTC cycle control member during 19–28 June 2015

shown in Figure 4. Both Figures 1 and 3 clearly show that potential temperature has the most dominant warm bias with a linear upward trend with forecast time. The maximum bias reaches about 4 K at 72 hr forecast lead time (Figure 1a). The dimensionless pressure also shows an obvious negative bias with a linear downward trend, i.e. becoming stronger with forecast time (Figures 1d and 3d). The zonal and meridional winds exhibit apparent diurnal biases: too strong during the daytime and normal or slightly too weak during the night-time (Figures 1b,c and 3c). While Figures 1–3 show the biases at one particular level (700 hPa), Figure 4 shows the vertical distribution of the control member's biases. For potential temperature (Figure 4a) it had a warm bias in the entire atmosphere except in a layer near 200 hPa where a cold bias existed. For zonal wind (Figure 4b) a westerly bias was observed below the 600 hPa level, and an easterly bias between 600 and 200 hPa. For meridional wind (Figure 4c) a southerly bias was observed below 800 hPa, a northerly bias existed above 550 hPa, and little bias was seen between 800 and 550 hPa. For the dimensionless pressure (Figure 4d) it was negatively biased below 550 hPa and positively biased above 550 hPa. For all the four variables, their biases increased as the forecast length increased.

These biases are similarly present in the experimental period (1–31 July 2015). For instance, Figure 5 is the domain-averaged biases at the 700 hPa level, derived from the first 10 days of the experimental period (1–10 July

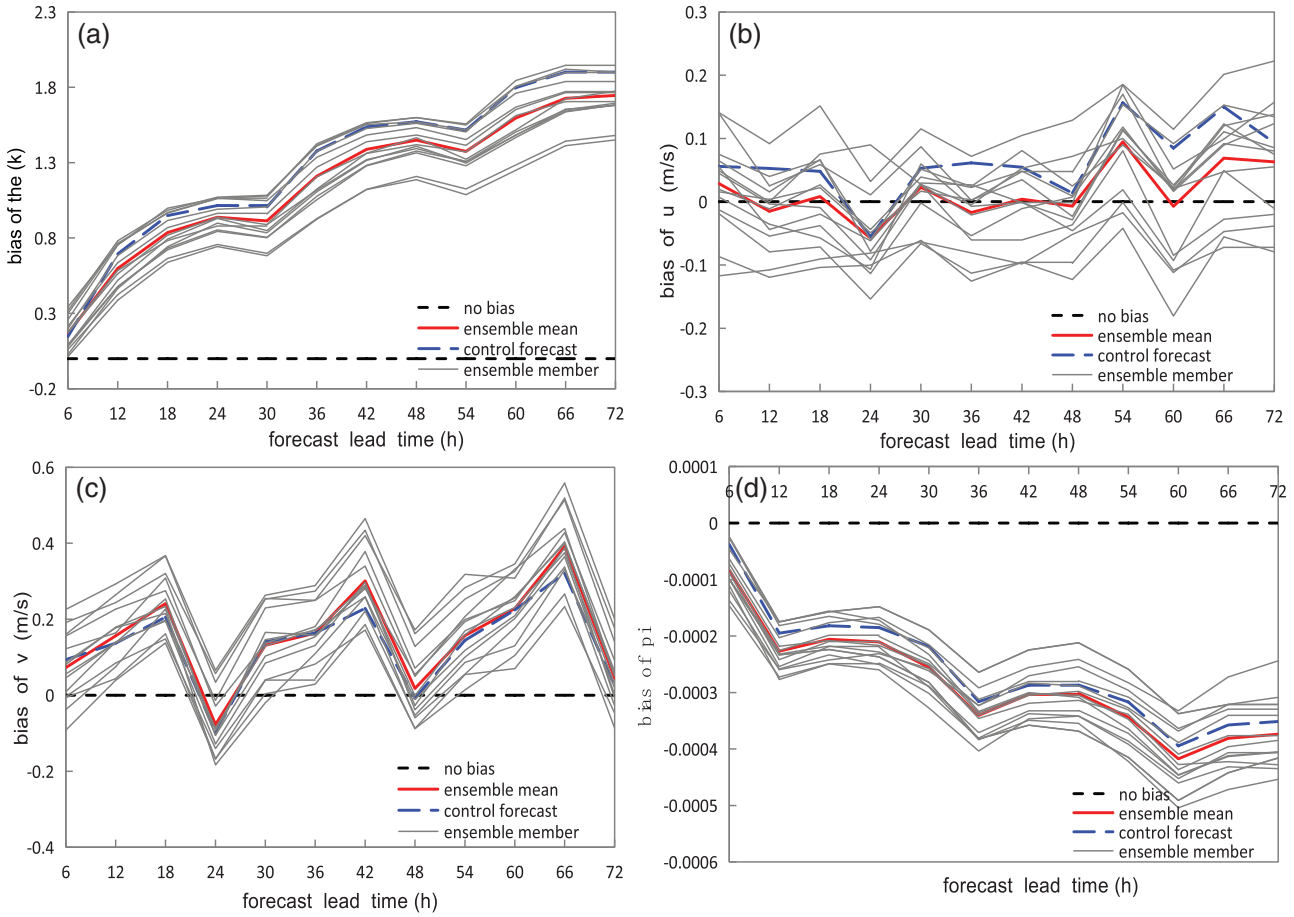

FIGURE 5 The model domain-averaged biases of the 15 ensemble members plus the ensemble mean for (a) potential temperature th (θ, K), (b) zonal wind u (m·s^{-1}), (c) meridional wind v (m·s^{-1}), and (d) dimensionless pressure pi (π) near 700 hPa level, averaged over the 10 forecasts (1–10 July 2015)

2015). Besides the control member, Figure 5 also shows the biases of the 14 perturbed members and the ensemble mean. It suggests that all members have similar bias tendency to that of the control member (less so in the zonal wind due to its weaker bias). Given this similarity, it is reasonable to use the bias tendency of the control member to correct the 14 perturbed ensemble members. Therefore, Equation 4 can be further simplified into

$$S_j(t) = \int_{t=0}^{t} \{A(S_j, t) + P(S_j, t) - \hat{B}_l(S_0, t)\} dt. \qquad (5)$$

Since forecast bias could be different in a multi-physics EPS, using the control member's bias tendency for all members is an approximation. As stochastic physics is more preferable than multi-physics and becomes more popular for perturbing a model (Du *et al.*, 2018), this approximation should be eased with time.

For an implementation of this method into operations, the control member runs without bias correction to establish a historical raw forecast dataset for the bias tendency estimation of other perturbed ensemble members. Although this method might not be practical (resource costs) in operations for a single deterministic model (since it requires the same model to run twice, once with bias correction for the actual application and again without bias correction for the bias tendency calculation), it can be implemented at almost no cost in an EPS environment. It takes about 7 min to estimate past bias and prepare bias tendency for model integration in our IBM Flex-460 computer. The model integration time is almost unchanged after the bias tendency term is added to the model. Keep in mind that an EPS has become a standard prediction system nowadays at all major NWP centres in the world (Buizza *et al.*, 2018).

3.2 | Three experiments

Based on the above forecast bias analysis, we have designed three experiments to examine the effectiveness of ways to incorporate the bias tendency term in a model.

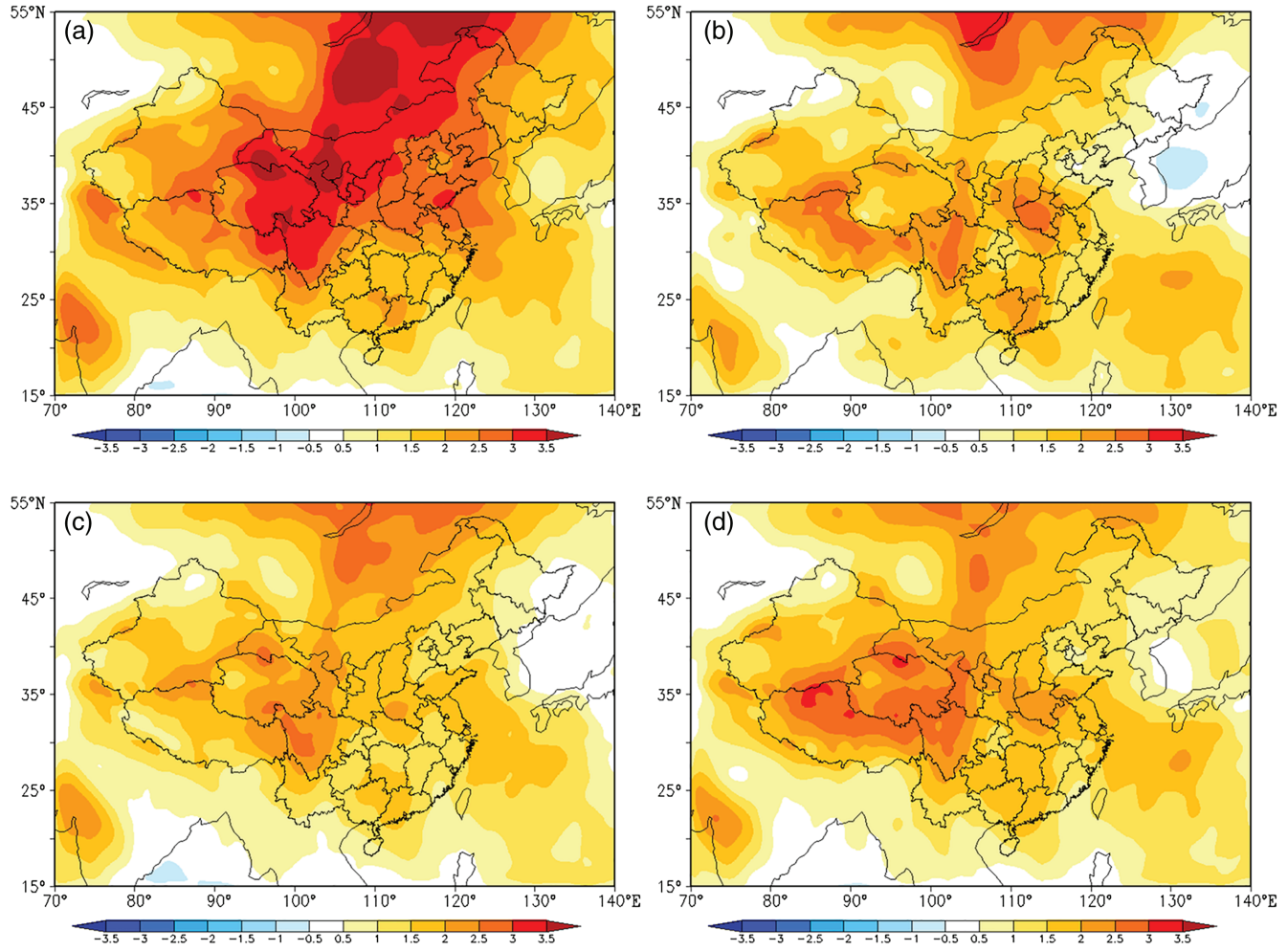

FIGURE 6 The ensemble mean forecast biases at 72 hr forecast lead time for (a) the raw forecast, (b) Exp. 1, (c) Exp. 2, and (d) Exp. 3. The results are the 31-day average for the 0000 UTC cycle during 1–31 July 2015. The variable is 500 hPa temperature

The first experiment (Exp. 1) is the most simplified one: since the potential temperature bias is the most dominant and has a roughly linear increasing trend over all forecast times, only the potential temperature's bias tendency of the first time step (Equation 3a was used with $\Delta = 72$ hr) will be used in the bias correction term throughout the entire model integration, i.e. the bias tendency forcing is fixed at all time steps,

$$\theta_j(t) = \int_{t=0}^{t} \{A(\theta_j, t) + P(\theta_j, t) - \hat{B}_l(\theta_0, 0)\} dt. \quad (6)$$

No bias tendencies of other variables are used in Exp. 1. The second experiment (Exp. 2) is the same as Exp. 1 but the potential temperature's bias tendency varies at every time step during the model integration (Equation 3b was used with $\Delta = 6$ hr),

$$\theta_j(t) = \int_{t=0}^{t} \{A(\theta_j, t) + P(\theta_j, t) - \hat{B}_l(\theta_0, t)\} dt. \quad (7)$$

The third experiment (Exp. 3) is the most sophisticated one, directly using Equation 5 with no simplifications. The time-varying bias tendencies (Equation 3b was used with $\Delta = 6$ hr) are added to four state variables (θ, u, v and π) in the model integration (note: other two model state variables, vertical velocity w and moisture q, were not perturbed),

$$S_j(t) = \int_{t=0}^{t} \{A(S_j, t) + P(S_j, t) - \hat{B}_l(S_0, t)\} dt, \quad (8)$$

$$S \rightarrow \{\theta, u, v, \pi\}. \quad (9)$$

Figure 6 compares the 500 hPa temperature biases of the ensemble mean forecasts at 72 hr forecast lead time for the original (raw, Figure 6a) and the three experiments (Figure 6b–d). The results show that all three experiments can greatly reduce the bias compared to the raw run, although the warm bias still exists in all runs.

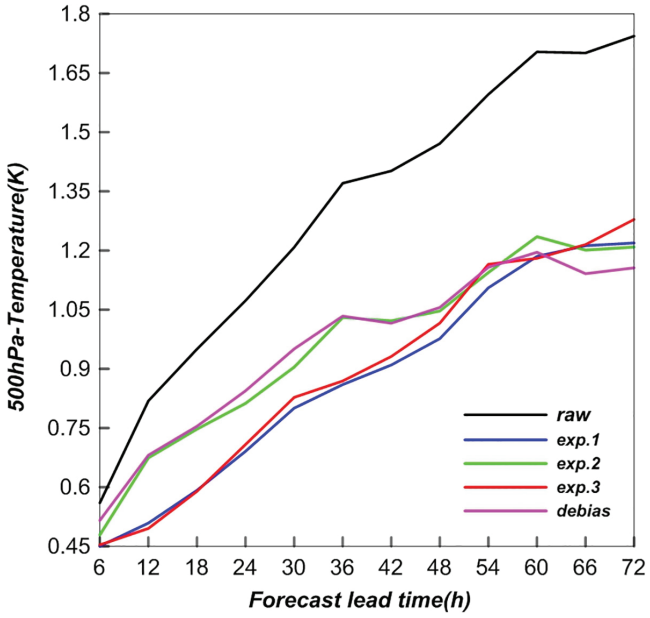

FIGURE 7 The domain-averaged biases of the ensemble mean varying with forecast hour from the raw and three experimental runs as well as the Kalman-filter based statistical method ("debias"). The results are the 31-day average for the 0000 UTC cycle during 1–31 July 2015. The variable is 500 hPa temperature

The domain-averaged biases (Figure 7) show that this reduction increases with forecast lead time and is about 30–40% on average for all forecast hours. The reduction is statistically significant at the 99.95% level based on Student's *t*-test (which will be used elsewhere throughout this article as the statistical significance test). Exp. 3 and Exp. 1 worked noticeably better than Exp. 2 before 48 hr and similarly afterwards. To get an idea about the relative performance of this new method, we compared it with the Kalman-filter (or decaying average) based statistical bias-correction method, which is currently used in operations at both CMA and NCEP (Du and Zhou, 2011; Cui *et al.*, 2012). The result of this statistical method for the same time period is shown by the magenta curve in Figure 7. Apparently, the performance of the new method is better or comparable to the current operational statistical method, which is encouraging.

Although the under-dispersive nature still exists in all runs, the three experimental runs have greatly improved the ensemble's spread-skill relationship (Whitaker and Loughe, 1998; Du, 2012; Fortin *et al.*, 2014; Du and Zhou, 2017). Figure 8 shows that the ensemble spread (black dashed line) and the root-mean-squared error of the ensemble mean forecast (RMSE, the black solid line) are much closer to each other in the experimental runs (Figure 8b–d) than in the raw run (Figure 8a) for 500 hPa temperature. An average improvement of about 20–27% has been achieved in Consistency score (the ratio of the RMSE of ensemble mean forecast to ensemble spread, the red solid line): from 2.24 (raw) to 1.65 (Exp. 1), 1.82 (Exp. 2), and 1.64 (Exp. 3). This improvement is statistically significant at all forecast lead times for all three experiments. Exp. 3 worked the best, followed by Exp. 1 and Exp. 2.

Rank histogram is another common metric to verify ensemble spread (Talagrand *et al.*, 1997; Hamill, 2001; Candille and Talagrand, 2005; Jolliffe and Primo, 2008; Du and Zhou, 2017). Figure 9a compares the rank histograms of 500 hPa temperature at 72 hr forecast lead time for the four runs. Although all runs have a strong warm bias ("L" shape), the three experimental runs noticeably reduced the warm bias, showing a reduced extent of the left-skewness. This warm bias reduction significantly (at 99.9% level) reduced the outlier (Figure 9b), from the original 50% to about 35% (a 30% improvement) for the experimental runs at 72 hr forecast lead time. Therefore, the experimental ensembles have a greater chance of encompassing the truth in their forecasts than the raw ensemble. Consistent with the result of Figure 7, Exp. 1 and Exp. 3 had slightly outperformed Exp. 2 in the first 2 days, while Exp. 1 and Exp. 3 had similar performance over all forecast hours.

Improved ensemble mean and spread should result in better probabilistic forecasts. Figure 10 shows the continuous ranked probability score (CRPS: Hersbach, 2000; Grimit *et al.*, 2006) of 500 hPa temperature at 72 hr forecast lead time. CRPS is a negatively oriented score, the smaller the better (with more reliable and higher-resolution information). The overall reduction in CRPS can be clearly observed over the entire domain from the raw run (Figure 10a) to the experimental runs (Figure 10b–d). The improvement occurs at all forecast hours and increases with the increase in forecast length (Figure 11). The experimental runs reduced the CRPS by about 33% from 1.5 to 1 at 72 hr forecast lead time. This reduction is statistically significant (99.9%) at all forecast lead times for all experiments. In general, the three experiments performed similarly, with Exp. 2 slightly behind Exp. 3 and 1.

Reliability is an important characteristic of probabilistic forecasts, providing a key factor in the cost/loss ratio based decision-making process (Du and Deng, 2010). Figure 12 shows the reliability diagrams of 500 hPa temperature at various forecast hours for the four runs. The event defined for producing probability is selected as exceeding 1 °C over climatology. Although all runs are overconfident due to the warm bias, the improvement of the experimental runs over the raw run is obvious: the reliability curves of the three experimental runs are closer to the diagonal line (perfect reliability). This improvement is statistically significant at all forecast lead times for all experiments.

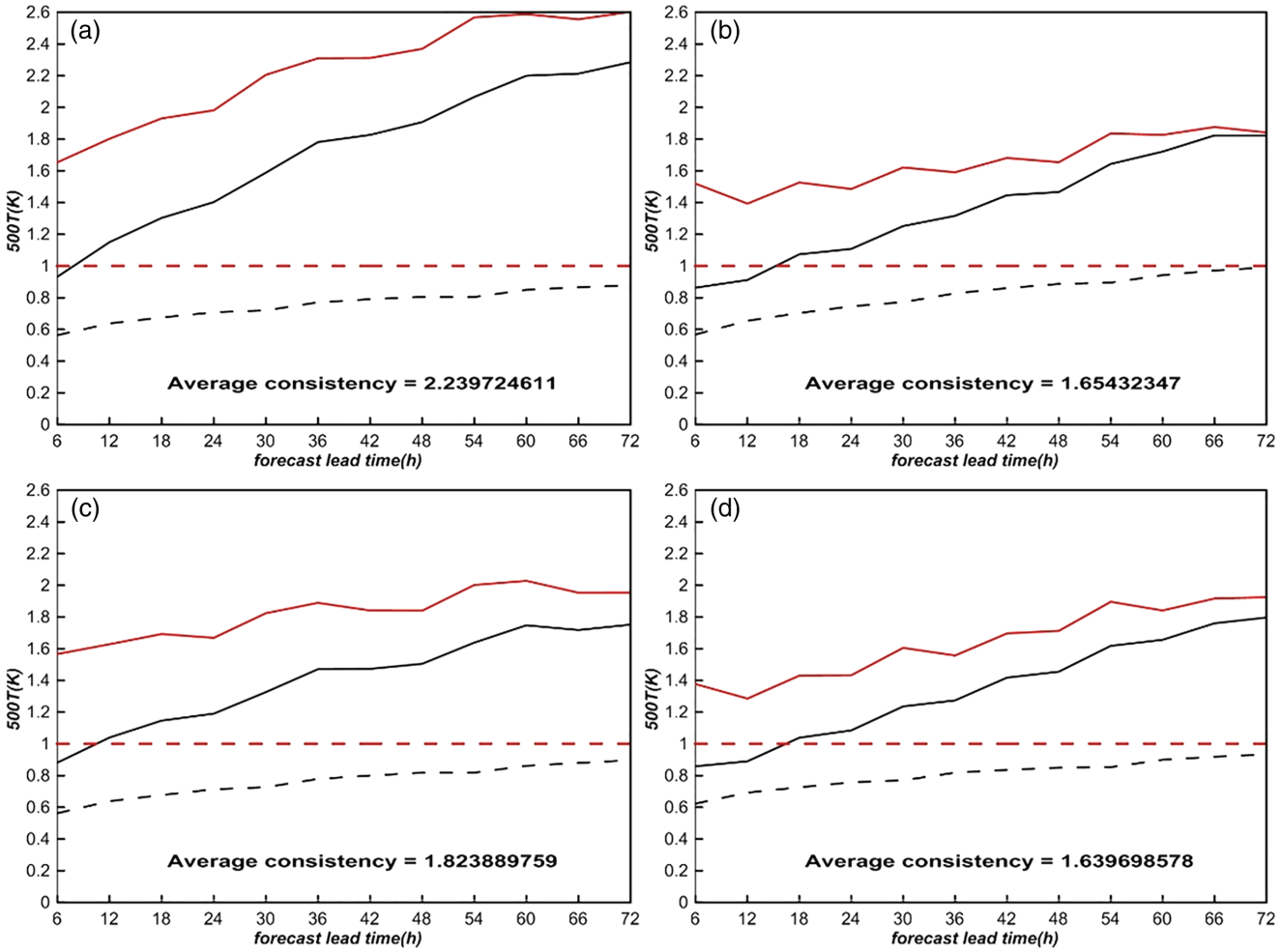

FIGURE 8 The domain-averaged ensemble mean forecast RMSE (black solid line), ensemble spread (black dashed line) as well as consistency score (RMSE/spread, red solid line) and the perfect consistency line (red dashed line) varying with forecast hour for (a) the raw forecast, (b) Exp. 1, (c) Exp. 2 and (d) Exp. 3. The averaged consistency score value over all forecast hours for each run is shown on each subplot. The results are the 31-day average for the 0000 UTC cycle during 1–31 July 2015. The variable is 500 hPa temperature

Among the three experiments, Exp. 2 and 3 are slightly more reliable than Exp. 1.

Unlike the CRPS and reliability diagram, the relative operating characteristics (ROC) is a score that is less sensitive to model bias. Figure 13 is the ROC diagrams of 500 hPa temperature at various forecast hours for the four runs. The threshold to define an event for producing probability is the same as was used in the calculation of reliability. Figure 13 shows that there is still an improvement (higher hitting rate and lower missing rate) in the three experimental runs over the raw run, although the score is less sensitive to model bias. The improvement increases with the increase in forecast length. This improvement is also statistically significant at all forecast lead times for all experiments. The three experimental runs performed very similarly to each other in terms of ROC.

All the verification above is based on 500 hPa temperature, which has a strong warm bias. To determine if the proposed approach can also calibrate other variables including surface variables and derived fields such as precipitation, a scorecard approach is applied. A scorecard is a summary of statistics from many variables and can easily show which variables or aspects have been improved, worsened or were neutral (remaining unchanged) by a new method. In our scorecard, several forecast skill 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 1,000 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 hr, 48 hr and 72 hr forecast lead times. For non-precipitation fields the verification metrics are RMSE, Consistency (RMSE/spread), CRPS and outlier; for precipitation the metrics AROC (area under ROC curve) and BS (Brier score) are used. There is a total of 294 categories in the scorecard. Figure 14 shows

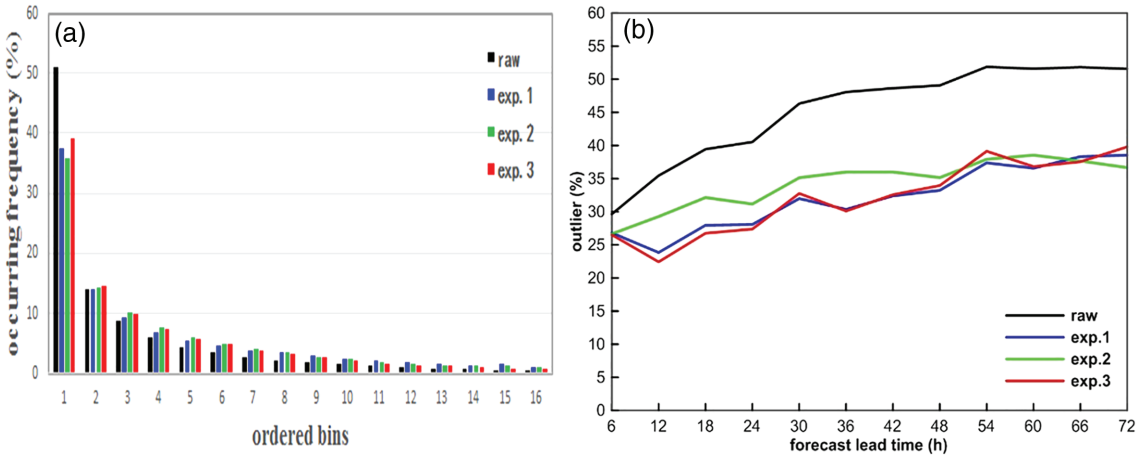

FIGURE 9 (a) The rank histograms at 72 h forecast lead time, and (b) the outliers over forecast hour. Black bar is for the raw forecast, blue for Exp. 1, green for Exp. 2 and red for Exp. 3. The results are the 31-day average for the 0000 UTC cycle during 1–31 July 2015. The variable is 500 hPa temperature

the three scorecards, respectively, for Exp. 1, Exp. 2 and Exp. 3 improving upon the raw run. Consistent with the verification results of 500 hPa temperature, all three experimental runs had a generally positive or neutral impact on both upper-air and surface variables, especially the height and temperature fields. Precipitation forecasts generally remain unchanged. The improvement (green), neutral (grey) and degradation (red) rates are, respectively, 59% (172/294), 40% (119/294) and 1% (3/294) for Exp. 1; 46% (135/294), 53% (155/294) and 1% (4/294) for Exp. 2; and 34% (101/294), 40% (116/294) and 26% (77/294) for Exp. 3. Particular attention needs to be paid to the behaviour of the most sophisticated or least simplified approach Exp. 3. As with Exp. 1 and Exp. 2, Exp. 3 has greatly improved height and temperature forecasts in most categories. However, it unfortunately degraded many upper-air wind forecasts, resulting in the highest degradation rate (26%) among the three experimental runs. On the other hand, our investigation reveals that Exp. 3 increased ensemble spread the most, while the other two experiments had little impact on ensemble spread (Figure 15). This implies that model forecasts were more sensitive to bias tendency in wind than in temperature. Adding bias tendency to the wind fields results in larger variations among ensemble members (larger spread), which is a welcome change for an under-dispersive EPS such as this one.

3.3 | A challenging issue to be investigated

A failure can be the mother of future success if we can learn a lesson from it. Why were the upper-level wind forecasts degraded in Exp. 3? One possible reason is the violation of the linear assumption in estimating bias tendencies for the wind field, given the obvious diurnal variation of u and v biases as shown in Figures 1,3 and 5. Another reason might be the inconsistency in the wind field when u and v biases were processed separately. Biases in u and v are related; they might need to be dealt with together. Generally speaking, since wind bias is likely to be more flow dependent than temperature bias, it is more challenging to "correctly" incorporate wind bias tendency in a model than a thermal or mass field. In addition to wind, the internal consistency among all model state variables should also be carefully investigated in the multi-variable approach (as Exp. 3).

Although a thorough in-depth study of this challenge certainly needs a separate study, a preliminary investigation was performed to shed a light. Since the 24 hr forecasts of 200 hPa u were improved in Exp. 1 and 2 but degraded in Exp. 3 (Figure 14), this variable is chosen for investigation. Figure 16 is the bias error of the ensemble mean forecasts for the four runs, while Figure 17 is the absolute value of the bias error. It is insightful to see that the bias magnitude was reduced but the spatial pattern or bias sign remained the same (mainly easterly bias) in Exp. 1 and 2 (Figure 16a–c), while the spatial pattern or bias sign changed (e.g. easterly bias was replaced by westerly bias over a large portion of the area) in Exp. 3 (Figure 16d). This result can be confirmed by the domain-averaged bias: $-0.54\,\mathrm{m\cdot s^{-1}}$ for Raw run, $-0.49\,\mathrm{m\cdot s^{-1}}$ for Exp. 1, $-0.5\,\mathrm{m\cdot s^{-1}}$ for Exp. 2, and $0.028\,\mathrm{m\cdot s^{-1}}$ for Exp. 3. This means that the bias was over-corrected in Exp. 3, which leads to a larger absolute bias error (Figure 17) and total forecast error (figure not shown). Figure 17 shows that the domain-averaged absolute bias error is $2.61\,\mathrm{m\cdot s^{-1}}$ for Raw run, $2.56\,\mathrm{m\cdot s^{-1}}$ for Exp. 1, $2.47\,\mathrm{m\cdot s^{-1}}$ for Exp. 2, and

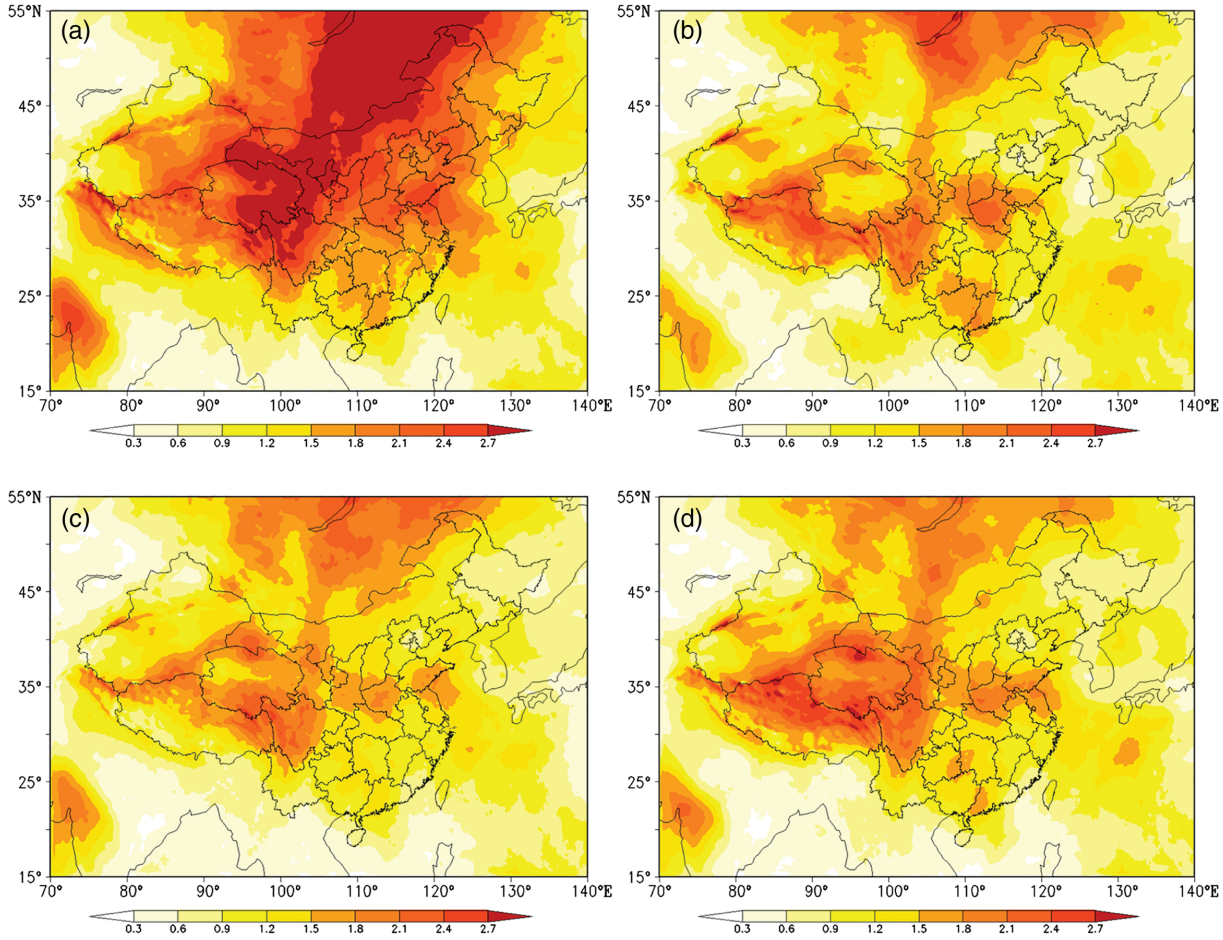

FIGURE 10 The CRPS of the ensemble-based probabilistic forecasts at 72 hr forecast lead time for (a) the raw forecast, (b) Exp. 1, (c) Exp. 2, and (d) Exp. 3. Probability of exceeding 1 °C over climatology is used. The results are the 31-day average for the 0000 UTC cycle during 1–31 July 2015. The variable is 500 hPa temperature

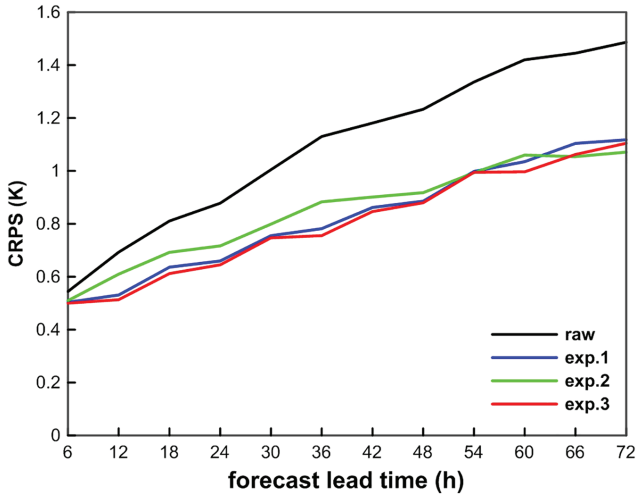

FIGURE 11 The domain-averaged CRPS of the ensemble-based probabilistic forecasts varying with forecast hour. Black line is for the raw forecast, blue for Exp. 1, green for Exp. 2 and red for Exp. 3. Probability of exceeding 1 °C over climatology is used. The results are the 31-day average for the 0000 UTC cycle during 1–31 July 2015. The variable is 500 hPa temperature

2.72 $m \cdot s^{-1}$ for Exp. 3. The degradation could be caused by either one of the two reasons mentioned above (i.e. the linear assumption violation or inconsistency among variables).

Given the fact that this problem might not exist in the one-variable approach as in Exp. 1, the simplest one-variable (such as temperature) approach is recommended to use in production for now. Since temperature has a dominant bias in this case, if this is the reason that Exp. 1 worked well or not is also a question to answer. For example, for a model where no single variable has a dominant bias, will this one-variable approach still be superior to a multi-variable approach?

4 | SUMMARY AND DISCUSSIONS

Unlike the retail-like (for selected fields) statistical post-processing methods commonly used to calibrate forecast biases, this study proposed and tested a wholesale-like

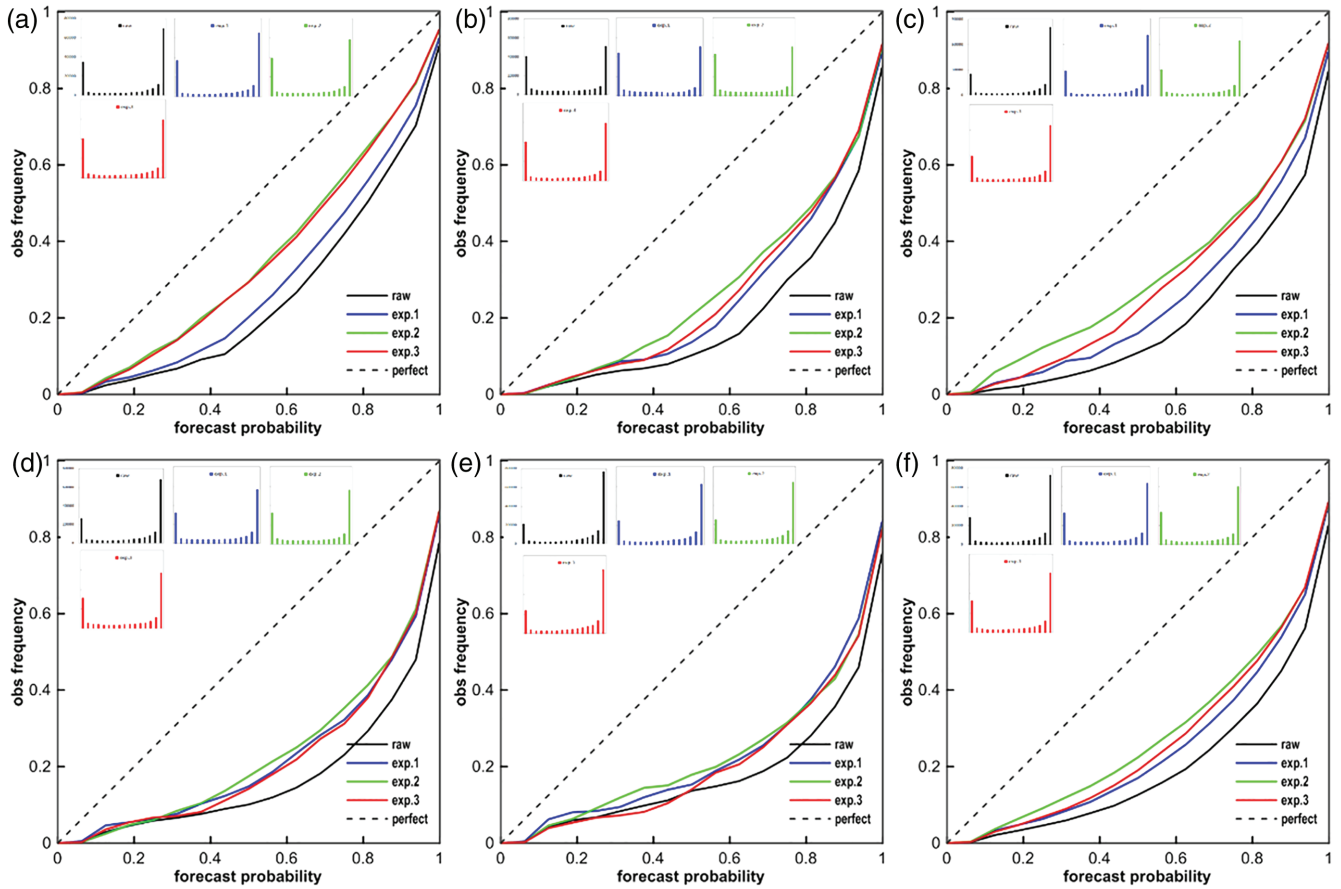

FIGURE 12 Reliability diagrams of the ensemble-based probabilistic forecasts for the raw and three experimental forecasts at the forecast lead time of (a) 12, (b) 24, (c) 36, (d) 48, and (e) 72 hr as well as (f) the average of all forecast hours (6–72 hr). Probability of exceeding 1 °C over climatology is used. The results are the 31-day average for the 0000 UTC cycle during 1–31 July 2015. The variable is 500 hPa temperature

(for all fields) dynamical approach to correct forecast bias during model integration. The method is not only more convenient (two steps are consolidated into one step) but also makes the downstream products more consistent and makes some downstream applications such as model initialization possible. Following the idea of the ensemble stochastic physics perturbation, this approach subtracts a bias tendency from the model's total tendency term of a state variable. The bias tendency is updated at every time step until the end of model integration. The bias tendency can be estimated from the bias error of past forecasts. During this bias tendency estimation, two approximations have been introduced: one is the omission of at least a partial flow-dependent bias (unresolved spatial structures) caused by averaging over a period of time to obtain the bias error. Another is the omission of nonlinear bias tendency (unresolved temporal structures) caused by the linear fitting to interpret a large time interval value into a time-step value (e.g. any bias for less than a 6 hr time-scale is not resolved in this study). Since these approximations reduce the accuracy of this method, how to accurately describe the bias tendency term should be one of the main tasks to improve the approach. Given the advantages of this approach, we believe that it represents the future of correcting biases in a numerical weather prediction model. The computing resource needed for this method is almost negligible in a major operational NWP centre.

With this proposed approach, three experiments (Exp. 1–3) were carried out and compared with each other to examine the effectiveness of ways to incorporate bias tendency into a model. Exp. 1 tests the most simplified setting by adding the bias tendency term only to the most biased variable (potential temperature in this case), where the bias tendency does not vary with forecast time (i.e. the first time-step value is used for all time steps). Exp. 2 is the same as Exp. 1 but the potential temperature's bias tendency varies at every time step during the model integration. Exp. 3 is the most sophisticated, where the time-varying bias tendencies were added to four state variables (θ, u, v and π) in the model integration. To mimic operational implementation of this method, the control member ran without bias correction to provide a historical dataset of raw forecasts

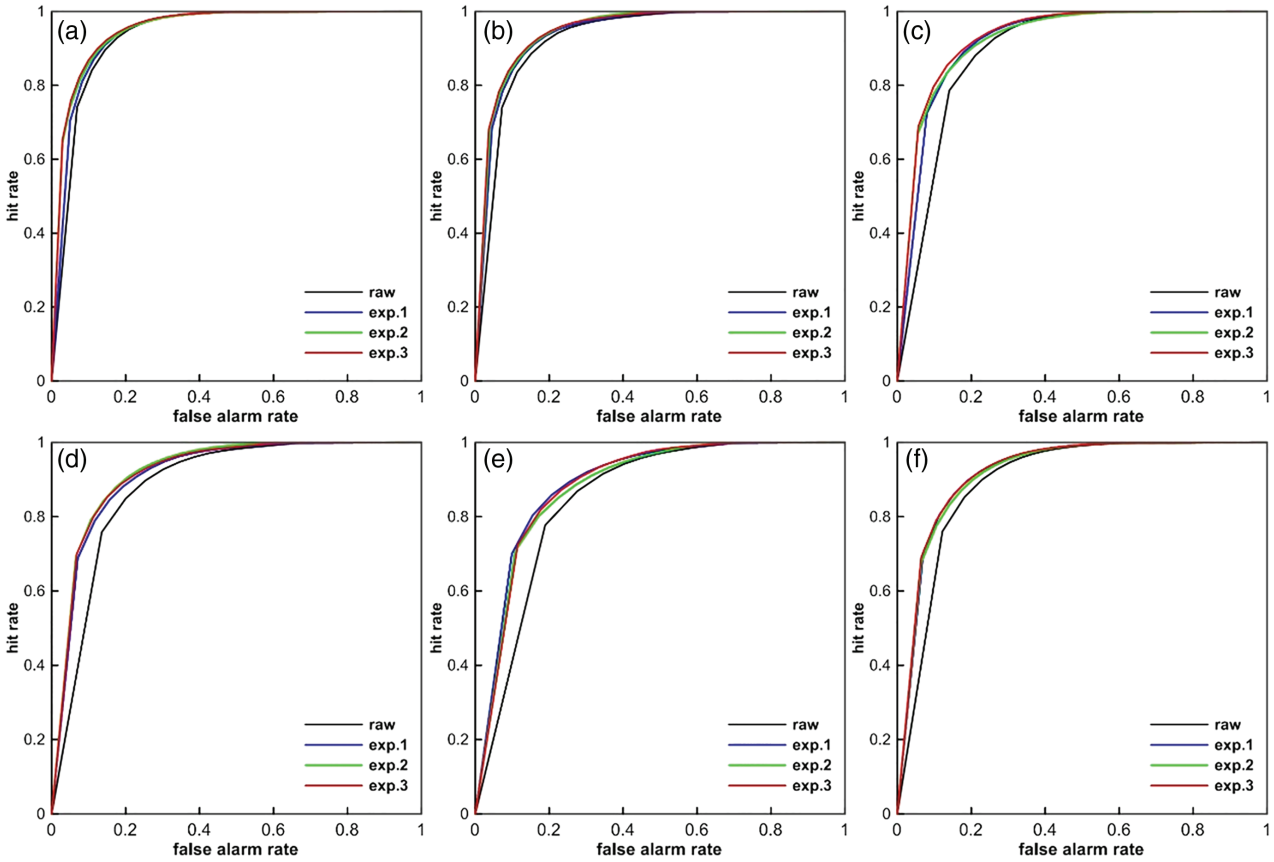

FIGURE 13 Same as Figure 12 but for ROC diagrams

for the bias tendency estimation. The experiments were performed on each of the 14 perturbed members of the GRAPES-REPS. Given the similarity of the bias tendencies of the 14 perturbed members to the control member, the use of the control member's bias tendency for all 14 perturbed members is reasonable, although this is another approximation of the method.

The verification was carried out in the framework of ensemble forecasts over China for a period of 31 days (0000 UTC 1–31 July 2015). It was done for 500 hPa temperature first and then expanded to many other variables including near-surface variables and precipitation. RMSE and bias score were used for ensemble mean and individual members; spread-skill relationship (Consistency score), rank histogram and outlier for ensemble spread; and CRPS, BS, reliability, ROC diagrams and AROC for probabilistic forecasts. A scorecard was created to summarize multiple variables and multiple scores of a variable. From the verification of 500 hPa temperature, results indicate that all three experiments significantly improved the raw ensemble forecasts in all aspects with reduced bias error, more accurate ensemble mean, better spread-skill relationship, and more reliable and sharper probabilities. The improvement normally increased as forecast length increased. Among the three experiments, Exp. 1 and 3 generally performed better than Exp. 2. A comparison of the new method with the Kalman-filter based statistical method currently used in the operations showed better or at least similar performance. This is very encouraging.

When the verification was expanded to include more variables, the scorecard (containing 294 categories) shows that the three experiments also had a general positive or neutral impact on both upper-air and surface variables, especially the height and temperature fields. Precipitation forecasts remained relatively unchanged. There were only a few aspects that were degraded. For example, the improvement rates are 59, 46 and 34% out of the 294 categories for Exp. 1–3, respectively; the degradation rates are 1, 1 and 26%; and the neutral rates are 40, 53 and 40%. An unexpected result is that the most sophisticated approach, Exp. 3, while being a superior player for 500 hPa temperature, became the worst performer overall among the three experimental runs. Exp. 3 degraded many u and v components of upper-air wind forecasts, resulting in the highest degradation rate (26%). Our preliminary investigation suggests that the bias of wind forecasts was overly corrected. On other hand, Exp. 3 increased ensemble spread the most, while the other two experiments had little impact. This implies that model forecasts were more sensitive to bias tendency in wind than in temperature.

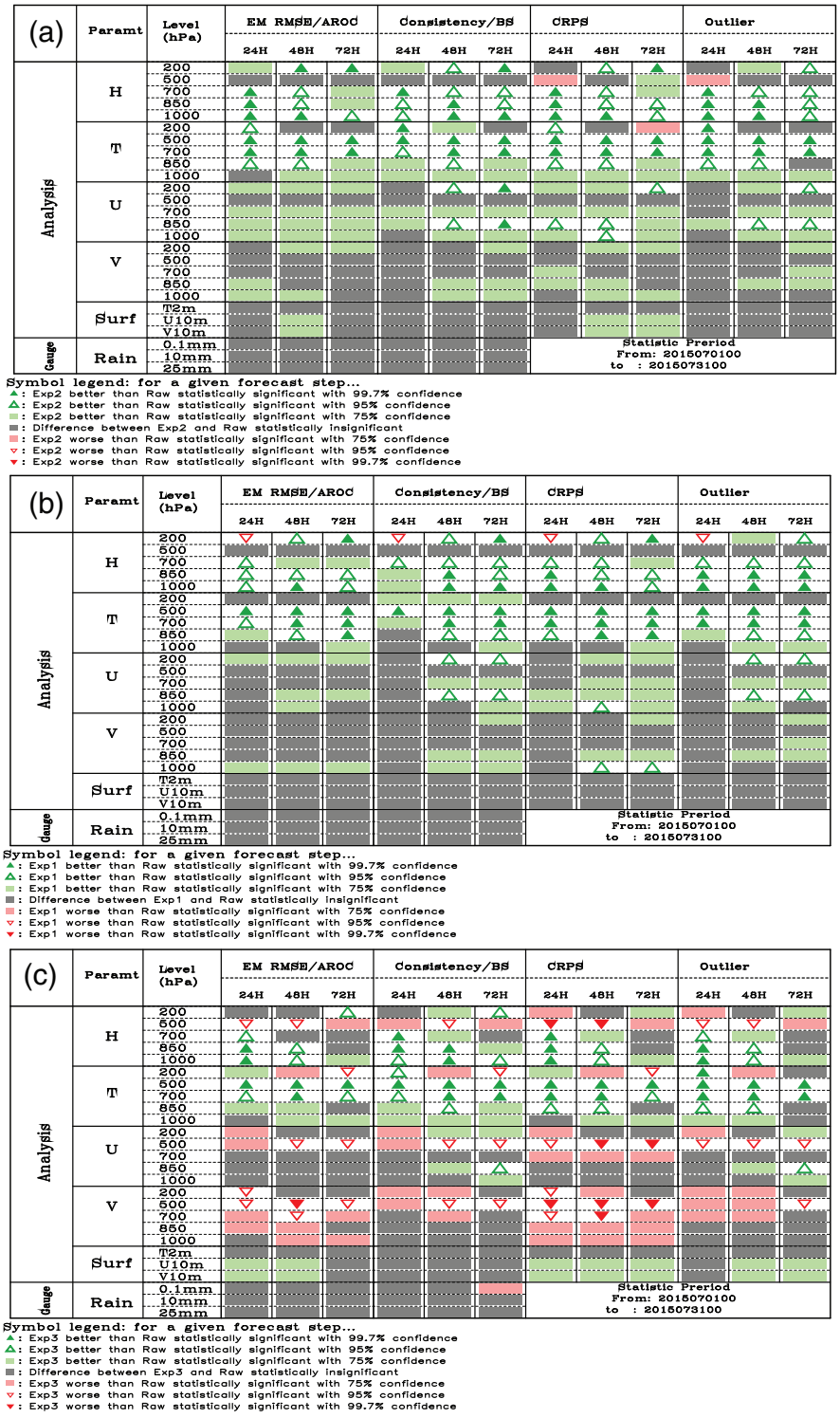

FIGURE 14 Scorecards for (a) Exp. 1, (b) Exp. 2 and (c) Exp. 3. Green indicates an improvement, red a degradation, and grey a no-change (neutral) with respect to the raw run. Different symbols are associated with different level of statistical significance of *t*-test (see the legend for the details). The results are the 31-day average for the 0000 UTC cycle during 1–31 July 2015. Note that the scores AROC and BS were used for precipitation forecasts only

Thus, adding bias tendency to wind fields could result in larger variations among ensemble members, a desired property for an under-dispersive EPS.

The scope of this study is to propose and demonstrate this new type of method but it is not a thorough verification study. To fully understand and refine this new approach, more case-studies using different numerical weather prediction models are needed. Many questions have not been answered yet. For example, why did Exp. 3 overly correct the wind bias? Can the inconsistency among model state variables, especially u and v after independently adding bias tendencies, play a role? Or is it due to the imperfect treatment of the bias tendency in wind (i.e. a violation of the linearity assumption)? These issues need to be further investigated using different models with different scenarios, such as "no dominant bias by one single field" and

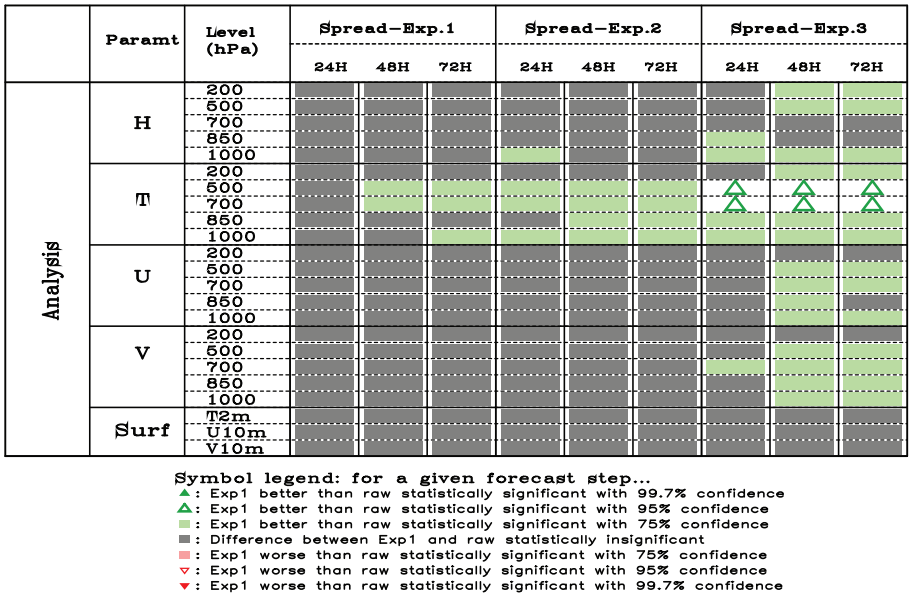

FIGURE 15 Scorecard of Exp. 1–3 for ensemble spread only

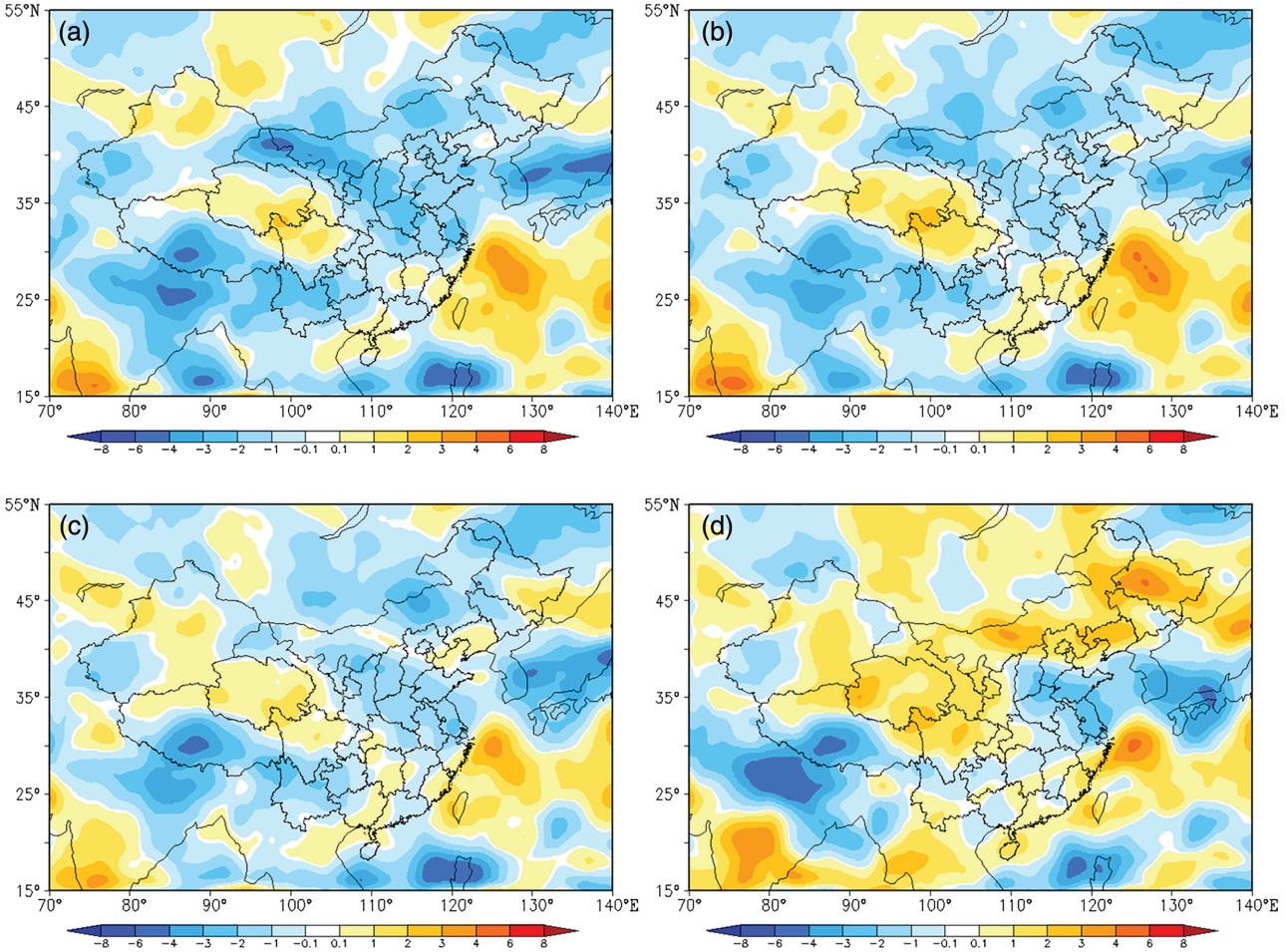

FIGURE 16 The biases of 200 hPa u at 24 hr lead time for (a) raw ($-0.54\ m{\cdot}s^{-1}$), (b) Exp. 1 ($-0.49\ m{\cdot}s^{-1}$), (c) Exp. 2 ($-0.5\ m{\cdot}s^{-1}$), and (d) Exp. 3 ($0.028\ m{\cdot}s^{-1}$). The values inside the brackets are the domain-averaged bias

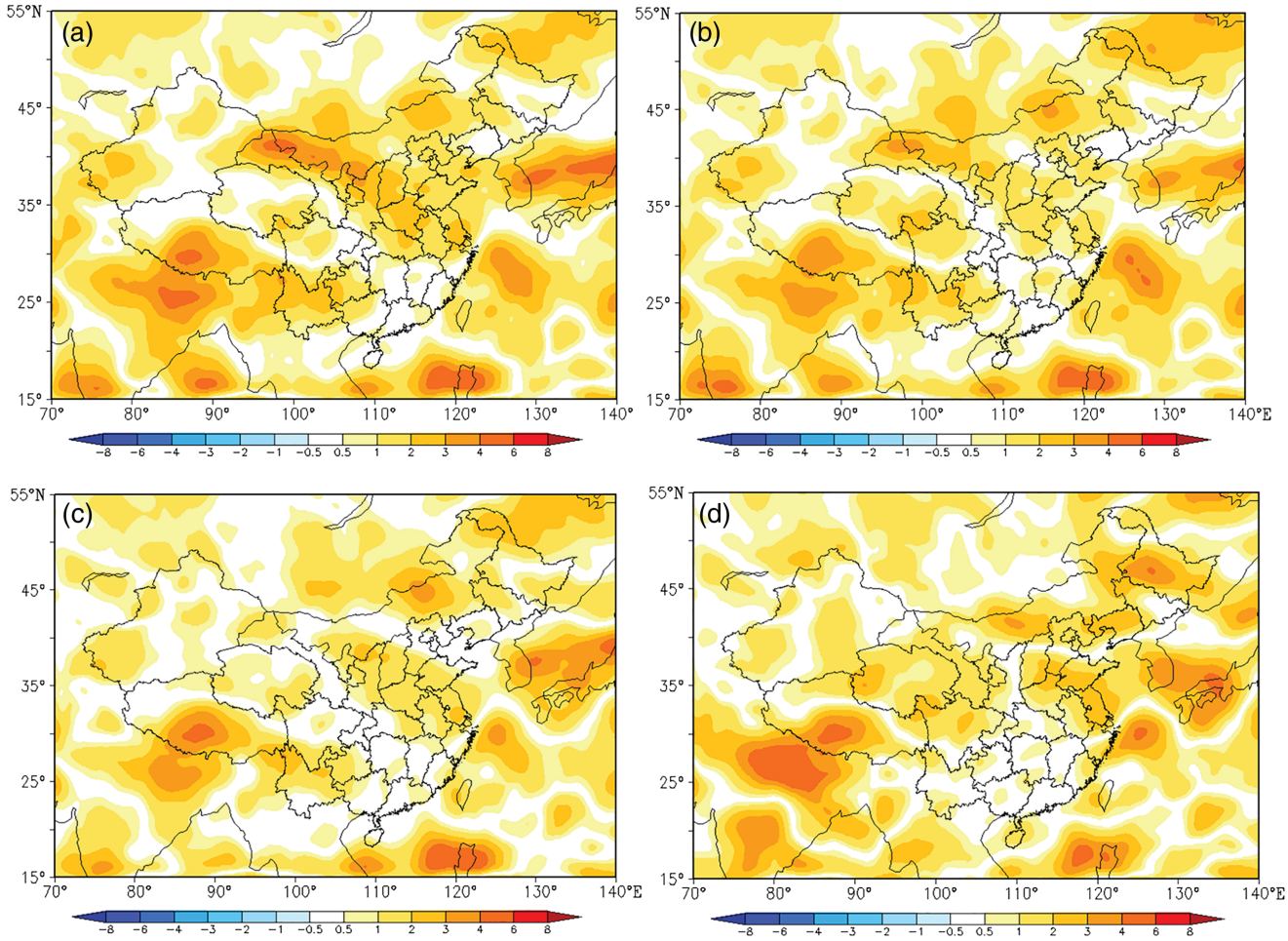

FIGURE 17 The absolute biases of 200 hPa u at 24 h lead time for (a) raw (2.61 $m \cdot s^{-1}$), (b) Exp. 1 (2.56 $m \cdot s^{-1}$), (c) Exp. 2 (2.47 $m \cdot s^{-1}$), and (d) Exp. 3 (2.72 $m \cdot s^{-1}$). The values inside the brackets are the domain-averaged absolute bias

"weak bias for all variables". Last but not least, the approximation of the use of the control member's bias tendency for all other members should also limit the effectiveness of this method. This approximation will fortunately be eased with time as stochastic physics becomes more popular than multi-physics for perturbing a model in an EPS.

ACKNOWLEDGEMENTS

We appreciate Mary Hart of NCEP for checking the readability of our manuscript as well as Yong Wang of ZAMG and two other anonymous reviewers for their constructive suggestions to significantly improve the manuscript. This work is sponsored by National Key R&D Program of China (2018YFC1507405).

REFERENCES

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

Berner, J., Shutts, G.J., Leutbecher, M. and Palmer, T.N. (2009) A spectral stochastic kinetic energy backscatter scheme and its impact on flow-dependent predictability in the ECMWF ensemble prediction system. *Journal of the Atmospheric Sciences*, 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. *Quarterly Journal of the Royal Meteorological Society*, 112, 667–691. https://doi.org/10.1002/qj.49711247307

Bhargava, K., Kalnay, E., Carton, J.A. and Yang, F. (2018) Estimation of systematic errors in the GFS using analysis increments. *Journal of Geophysical Research: Atmospheres*, 123, 1626–1637. https://doi.org/10.1002/2017JD027423.

Buizza, R., J. Du, Z. Toth, and D. Hou, (2018) Major operational ensemble prediction systems (EPS) and the future of EPS. *Handbook of Hydrometeorological Ensemble Forecasting* Duan, Q.Y., Pappenberger, F., Wood, A., Cloke, H.L. and Schaake, J.C., Springer, Berlin, 151–193. https://doi.org/10.1007/978-3-642-40457-3_14-1.

Buizza, R., Milleer, M. and Palmer, T.N. (1999) Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 125, 2887–2908. https://doi.org/10.1002/qj.49712556006.

Candille, G. and Talagrand, O. (2005) Evaluation of probabilistic prediction systems for a scalar variable. *Quarterly Journal of the*

Royal Meteorological Society, 131, 2131–2150. https://doi.org/10.1256/qj.04.71.

Chen, D.H., Xue, J.S., Yang, X.S., Zhang, H.L., Shen, X.S., Hu, J.L., Wang, Y., Ji, L.R. and Chen, J.B. (2008) New generation of multi-scale NWP system (GRAPES): general scientific design. *Chinese Science Bulletin*, 53(22), 3433–3445. https://doi.org/10.1007/s11434-008-0494-z.

Cui, B., Toth, Z., Zhu, Y. and Hou, D. (2012) Bias correction for global ensemble forecast. *Weather and Forecasting*, 27, 396–410. https://doi.org/10.1175/WAF-D-11-00011.1.

Danforth, C.M., Kalnay, E. and Miyoshi, T. (2007) Estimating and correcting global weather model error. *Monthly Weather Review*, 135, 281–299. https://doi.org/10.1175/MWR3289.1.

Du, J. (2012) New metrics for evaluating ensemble spread. In: *21st Conference on Probability and Statistics in the Atmospheric Sciences*. New Orleans, LA: American Meteorological Society, pp. 22–26. Available at: https://ams.confex.com/ams/92Annual/webprogram/Paper205576.html.

Du, J., Berner, J., Buizza, R., Charron, M., Houtekamer, P., Hou, D., Jankov, I., Mu, M., Wang, X., Wei, M. and Yuan, H. (2018) Ensemble methods for meteorological predictions. In: Duan, Q.Y., Pappenberger, F., Wood, A., Cloke, H.L. and Schaake, J.C. (Eds.) *Handbook of Hydrometeorological Ensemble Forecasting*. Berlin: Springer, pp. 99–149. https://doi.org/10.1007/978-3-642-40457-3_13-1

Du, J. and Deng, G. (2010) The utility of the transition from deterministic to probabilistic weather forecasts – verification and application of probabilistic forecasts. *Meteorological Monthly (in Chinese)*, 36(12), 10–18.

Du, J. and DiMego, G. (2008) A regime-dependent bias correction approach. In: *19th Conference on Probability and Statistics*, 20–24 January 2008, New Orleans, LA, American Meteorological Society, paper 3.2. Available at: https://ams.confex.com/ams/88Annual/webprogram/19PROBSTAT.html.

Du, J., DiMego, G., Jovic, D., Ferrier, B., Yang, B. and Zhou, B. (2015) Short-range ensemble forecast (SREF) system at NCEP: recent development and future transition. In: *27th Conference on Weather Analysis and Forecasting and 23rd Conference on Numerical Weather Prediction*. Chicago, IL: American Meteorological Society, paper 2A.5. Available at: https://ams.confex.com/ams/27WAF23NWP/webprogram/Paper273421.html.

Du, J. and Zhou, B. (2011) A dynamical performance-ranking method for predicting individual ensemble member performance and its application to ensemble averaging. *Monthly Weather Review*, 139, 3284–3303. https://doi.org/10.1175/MWR-D-10-05007.1.

Du, J. and Zhou, B. (2017) Ensemble fog prediction. In: Koracin, D. and Dorman, C.E. (Eds.) *In Marine Fog: Challenges and Advancements in Observations, Modeling and Forecasting*. Switzerland: Springer, pp. 477–509. https://doi.org/10.1007/978-3-319-45229-6_10

Dudhia, J. (1989) Numerical study of convection observed during the winter monsoon experiment using a mesoscale two-dimensional model. *Journal of the Atmospheric Sciences*, 46(20), 3077–3107. https://doi.org/10.1175/1520-0469(1989)046<3077:NSOCOD>2.0.CO;2

Fortin, V., Abaza, M., Anctil, F. and Turcotte, R. (2014) Why should ensemble spread match the RMSE of the ensemble mean? *Journal of Hydrometeorology*, 15, 1708–1713. https://doi.org/10.1175/JHM-D-14-0008.1.

Gneiting, T., Raftery, A.E., Westveld, A.H., III and Goldman, T. (2005) Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133, 1098–1118. https://doi.org/10.1175/MWR2904.1.

Grimit, E.P., Gneiting, T., Berrocal, V.J. and Johnson, N.A. (2006) The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Quarterly Journal of the Royal Meteorological Society*, 132, 2925–2942. https://doi.org/10.1256/qj.05.235.

Hamill, T.M. (2001) Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129, 550–560. https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2.

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

Hong, S.-Y. and Lim, J.-O. (2006) The WRF single moment 6-class microphysics scheme (WSM6). *Journal of the Korean Meteorological Society*, 42(2), 129–151.

Hong, S.-Y., Noh, Y. and Dudhia, J. (2006) A new vertical diffusion package with an explicit treatment of entrainment processes. *Monthly Weather Review*, 134(9), 2318–2341. https://doi.org/10.1175/MWR3199.1

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

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

Jolliffe, I.T. and Stephenson, D.B. (2003) *Forecast Verification: A practitioner's guide in atmospheric science*. New York, NY: Wiley and Sons.

Jolliffe, I.T. and Primo, C. (2008) Evaluating rank histograms using decompositions of the chi-square test statistic. *Monthly Weather Review*, 136, 2133–2139. https://doi.org/10.1175/2007MWR2219.1.

Kain, J.S. (2004) The Kain–Fritsch convective parameterization: an update. *Journal of Applied Meteorology*, 43, 170–181. https://doi.org/10.1175/1520-0450(2004)043<0170:tkcpau>2.0.co;2.

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

Kain, J.S. and Fritsch, J.M. (1993) Convective parameterization for mesoscale models: the Kain–Fritsch scheme. In: Emanuel, K.A. and Raymond, D.J. (Eds.) *The Representation of Cumulus Convection in Numerical Models. Meteorological Monograph*, Vol. 24. Boston: American Meteorological Society, pp. 165–170. https://doi.org/10.1007/978-1-935704-13-3_15

Leith, C.E. (1978) Objective methods for weather prediction. *Annual Review of Fluid Mechanics*, 10, 107–128.

Mahrt, L. and Ek, M. (1984) The influence of atmospheric stability on potential evaporation. *Journal of Climate and*

Applied Meteorology, 23, 222–234. https://doi.org/10.1175/1520-0450(1984)0232.0.CO;2

Mlawer, E.J., Taubman, S.J., Brown, P.D., Iacono, M.J. and Clough, S.A. (1997) Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave. *Journal of Geophysical Research*, 102(D14), 16663–16682. https://doi.org/10.1029/97JD00237

Monache, L.D., Nipen, T., Deng, X.X., Zhou, Y.M. and Stull, R. (2005) Ozone ensemble forecasts. Part II: A Kalman-filter predictor bias correction. *Journal of Geophysical Research*, 111, D05308. https://doi.org/10.1029/2005JD006311.

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

Ollinaho, P., Lock, S.J., Leutbecher, M., Bechtold, P., Beljaars, A.C.M., Bozzo, A., Forbes, R.M., Haiden, T., Hogan, R.J. and Sandu, I. (2017) Towards process-level representation of model uncertainties: stochastically perturbed parametrizations in the ECMWF ensemble. *Quarterly Journal of the Royal Meteorological Society*, 143(702), 408–422. https://doi.org/10.1002/qj.2931.

Pan, H.-L. and Wu, W.-S. (1995) *Implementing a mass flux convective parameterization package for the NMC medium-range forecast model*. NMC Office Note 409,. Washington, DC: National Centers for Environmental Prediction, pp. 1–40. Available at: http://www.emc.ncep.noaa.gov/officenotes/FullTOC.html.

Pan, Y., Shen, Y., Yu, J.J. and Xiong, A.Y. (2015) An experiment of high-resolution gauge–radar–satellite combined precipitation retrieval based on the Bayesian merging method. *Acta Meteorologica Sinica*, 73(1), 177–186. https://doi.org/10.11676/qxxb2015.010

Piccolo, C. and Cullen, M.J.P. (2016) Ensemble data assimilation using a unified representation of model error. *Monthly Weather Review*, 144, 213–224. https://doi.org/10.1175/MWR-D-15-0270.1.

Piccolo, C., Cullen, M.J.P., Tennant, W.J. and Semple, A.T. (2019) Comparison of different representations of model error in ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 145(718), 15–27. https://doi.org/10.1002/qj.3348

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

Roulston, M.S. and Smith, L.A. (2003) Combining dynamical and statistical ensembles. *Tellus A*, 55, 16–30. https://doi.org/10.3402/tellusa.v55i1.12082.

Satterfield, E.A. and Bishop, C.H. (2014) Heteroscedastic ensemble postprocessing. *Monthly Weather Review*, 142, 3484–3502. https://doi.org/10.1175/MWR-D-13-00286.1.

Shutts, G.J. (2005) A kinetic energy backscatter algorithm for use in ensemble prediction systems. *Quarterly Journal of the Royal Meteorological Society*, 131, 3079–3102. https://doi.org/10.1256/qj.04.106.

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

Talagrand, O., Vautard, R. and Strauss, B. (1997) Evaluation of probabilistic prediction systems. In: *Proceedings of ECMWF Workshop on Predictability*. Reading: ECMWF, pp. 1–25. Available from ECMWF, Shinfield Park, Reading, RG2 9AX, United Kingdom.

Toth, Z. and Kalnay, E. (1997) Ensemble forecasting at NCEP and the breeding method. *Monthly Weather Review*, 125, 3297–3319.

Wang, J.Z., Chen, J., Du, J., Zhang, Y., Xia, Y. and Deng, G. (2018) Sensitivity of ensemble forecast verification to model bias. *Monthly Weather Review*, 146, 781–796. https://doi.org/10.1175/MWR-D-17-0223.1.

Whitaker, J.S. and Loughe, A.F. (1998) The relationship between ensemble spread and ensemble mean skill. *Monthly Weather Review*, 126, 3292–3302. https://doi.org/10.1175/1520-0493(1998)126<3292:TRBESA>2.0.CO;2.

Xia, Y., Chen, J., Du, J., Zhi, X., Wang, J. and Li, X. (2019) A unified scheme of stochastic physics and bias correction in an ensemble model to reduce both random and systematic errors. *Weather and Forecasting*, 34, 1675–1691. https://doi.org/10.1175/WAF-D-19-0032.1.

Zhang, H.B., Chen, J., Zhi, X., Li, Y. and Sun, Y. (2014) Study on the application of GRAPES regional ensemble prediction system. *Meteorological Monthly*, 40, 1076–1087.

How to cite this article: Chen J, Wang J, Du J, Xia Y, Chen F, Hongqi L. Forecast bias correction through model integration: A dynamical wholesale approach. *Q J R Meteorol Soc*. 2020;146:1149–1168. https://doi.org/10.1002/qj.3730