Sensitivity of Ensemble Forecast Verification to Model Bias

JINGZHUO WANG,[a,b] JING CHEN,[b] JUN DU,[c] YUTAO ZHANG,[b] YU XIA,[d] AND GUO DENG[b]

[a] *Chinese Academy of Meteorological Sciences, China Meteorological Administration, Beijing, China*
[b] *Numerical Weather Prediction Center, China Meteorological Administration, Beijing, China*
[c] *Environmental Modeling Center, NOAA/NWS/NCEP, College Park, Maryland*
[d] *Nanjing University of Information Science and Technology, Nanjing, China*

(Manuscript received 1 August 2017, in final form 1 February 2018)

ABSTRACT

This study demonstrates how model bias can adversely affect the quality assessment of an ensemble prediction system (EPS) by verification metrics. A regional EPS [Global and Regional Assimilation and Prediction Enhanced System-Regional Ensemble Prediction System (GRAPES-REPS)] was verified over a period of one month over China. Three variables (500-hPa and 2-m temperatures, and 250-hPa wind) are selected to represent ‘‘strong’’ and ‘‘weak’’ bias situations. Ensemble spread and probabilistic forecasts are compared before and after a bias correction. The results show that the conclusions drawn from ensemble verification about the EPS are dramatically different with or without model bias. This is true for both ensemble spread and probabilistic forecasts. The GRAPES-REPS is severely underdispersive before the bias correction but becomes calibrated afterward, although the improvement in the spread's spatial structure is much less; the spread–skill relation is also improved. The probabilities become much sharper and almost perfectly reliable after the bias is removed. Therefore, it is necessary to remove forecast biases before an EPS can be accurately evaluated since an EPS deals only with random error but not systematic error. Only when an EPS has no or little forecast bias, can ensemble verification metrics reliably reveal the true quality of an EPS without removing forecast bias first. An implication is that EPS developers should not be expected to introduce methods to dramatically increase ensemble spread (either by perturbation method or statistical calibration) to achieve reliability. Instead, the preferred solution is to reduce model bias through prediction system developments and to focus on the quality of spread (not the quantity of spread). Forecast products should also be produced from the debiased but not the raw ensemble.

1. Introduction

Prediction of predictability is the primary mission of an ensemble prediction system (EPS). An EPS is designed to quantify the predictability of the atmosphere [measured by absolute forecast error such as root-mean-square error (RMSE)], estimate the uncertainties, and identify the range of possible solutions associated with a numerical weather prediction (NWP) model forecast (measured by ensemble spread and ensemble distribution) (Du 2007; Du and Chen 2010; Garcia-Moya et al. 2011; Hopson 2014). Ensemble spread (defined by the standard deviation, unless otherwise specified, of ensemble members with respect to ensemble mean) is, therefore, used to simulate the RMSE of ensemble mean forecast in both magnitude and spatial structure. In a perfect EPS the ensemble spread and RMSE of the ensemble mean forecast should be the same (i.e., of equal magnitude and a perfect correlation in space), and therefore, have the same statistical distributions between forecast error and spread when averaged over many cases (Whitaker and Loughe 1998; Du 2012; Du et al. 2014; Fortin et al. 2014). Such a relationship between spread and forecast error is called the ‘‘spread–skill relationship.’’ For the current operational EPSs around the world, it is reported that underdispersion (i.e., ensemble spread is significantly smaller than ensemble mean forecast error) is unfortunately a common problem. Even for a

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

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

DOI: 10.1175/MWR-D-17-0223.1

multimodel EPS, this problem remains. For example, McCollor and Stull (2009) reported that the North American Ensemble Forecast System suffers underdispersion for day-1 through day-6 forecasts although spread improves with the increase in forecast hour. The multimodel short-range ensemble prediction system at the Spanish Meteorological Service is also reported to be underdispersive (Garcia-Moya et al. 2011). The spatial correlation between spread and forecast error is also very low (Stensrud et al. 1999). People normally argue that the reason why an EPS tends to be underdispersive is because it does not incorporate all sources of uncertainty (Jolliffe and Stephenson 2003; Buizza et al. 2005; Ho et al. 2013) including not accounting for errors in observations (e.g., Saetra et al. 2004). In this paper, we demonstrate that model bias is another major contributor to this "underdispersion" phenomenon in terms of verification metrics.

Model bias is common and unavoidable. As Toth et al. (2003) indicated, ensemble distribution is biased (shifted) because of inherent deficiencies in the model and in initial conditions. Therefore, the error of the ensemble mean forecast contains a large portion of systematic error (bulk bias) and does not truly reflect the predictability of the atmosphere (random portion of error); and the ensemble distribution is shifted in its mean position and does not cover the true uncertainties (possibilities) of a forecast. One can imagine that an inflated ensemble mean forecast error will lead to a poorer spread–skill relationship; and a biased or shifted ensemble distribution will lead to a poorer performance by probabilistic forecasts. In the past, many studies have shown that removing model bias can enhance ensemble forecast performance (Hamill and Colucci 1997; Eckel and Mass 2005; Reynolds et al. 2011; Cui et al. 2012; Berner et al. 2015). Berner et al. (2015) quantify the impact of debiasing on forecast skill and conclude that bias rather than random forecast error is the leading source of forecast error. They further demonstrated that including a model-error representation (stochastic physics) in an EPS can reduce model bias. Rodwell et al. (2016) even tried to understand how systematic and random errors contribute to forecast reliability. However, all these past studies have focused on improving forecast performance, especially reliability, rather than verifying an EPS.

The purpose of this study is focusing on how to correctly verify an EPS. Since an EPS is designed to deal with a random error only, not a systematic error of a forecast system (Du 2007), verifying an EPS using its full error will not truly reveal what it is intended for (like comparing an apple with an orange) in principle and can often draw a wrong conclusion. For example, if an event was missed by EPS A but captured by EPS B, people often say "EPS B is better than EPS A" although the truth could be the opposite in terms of ensembling quality. The real reason for the result in this example could simply be that model B is superior to model A and has nothing to do with the ensembling technique. To faithfully verify what one really intended or to have an apples-to-apples comparison, we argue that an EPS should be verified against a random error only and not the full error, which is contaminated by systematic error (first-moment error). A systematic error only contributes to error but not to diversity, which is the focus of an EPS. Since most NWP models have biases, incorrectly assessing EPS quality is believed to be a common problem. Therefore, it is important to explicitly address and emphasize this pitfall in verifying an EPS. A true assessment of an EPS is important for EPS developers to focus on real problems related to ensembling techniques but not to the model itself.

The shift of the mean position of an ensemble distribution will not only affect the ensemble mean but also the spread–skill relationship and probabilistic forecasts. Model bias cannot only negatively impact verification metrics related to the ensemble mean, but also those related to spread–skill relationships, probabilistic metrics, or any metrics that involve a comparison to truth. Since model bias impacts the distribution of forecast errors of the members, even metrics that do not involve the ensemble mean will be sensitive to first-moment errors, although the sensitivity might be less. We attempt to demonstrate all of these points in our study. In this study we use a real-world operational EPS to systematically demonstrate how model bias can severely impact the assessment of the EPS quality by comparing the results of raw and debiased ensemble forecasts. The paper is organized as follows. In section 2 the EPS configuration, the data used for verification, as well as the bias correction method are described. The verification results are presented in section 3. A summary and discussion are given in section 4.

2. EPS configuration, bias correction method, and verification data

The Global and Regional Assimilation and Prediction Enhanced System-Regional Ensemble Prediction System (GRAPES-REPS) (Zhang et al. 2014) is used in this study. It is a mesoscale EPS, first developed at the NWP center of the China Meteorological Administration (CMA) in 2008 and implemented for operations in August 2014. GRAPES-REPS has 15 members (1 control and 14 perturbed) with a horizontal resolution of 15 km and 50 vertical levels. The boundary conditions are provided by different members of a global EPS, which is also running operationally at CMA. Model uncertainty is taken into account by applying multiple physics schemes (Stensrud et al. 2000;

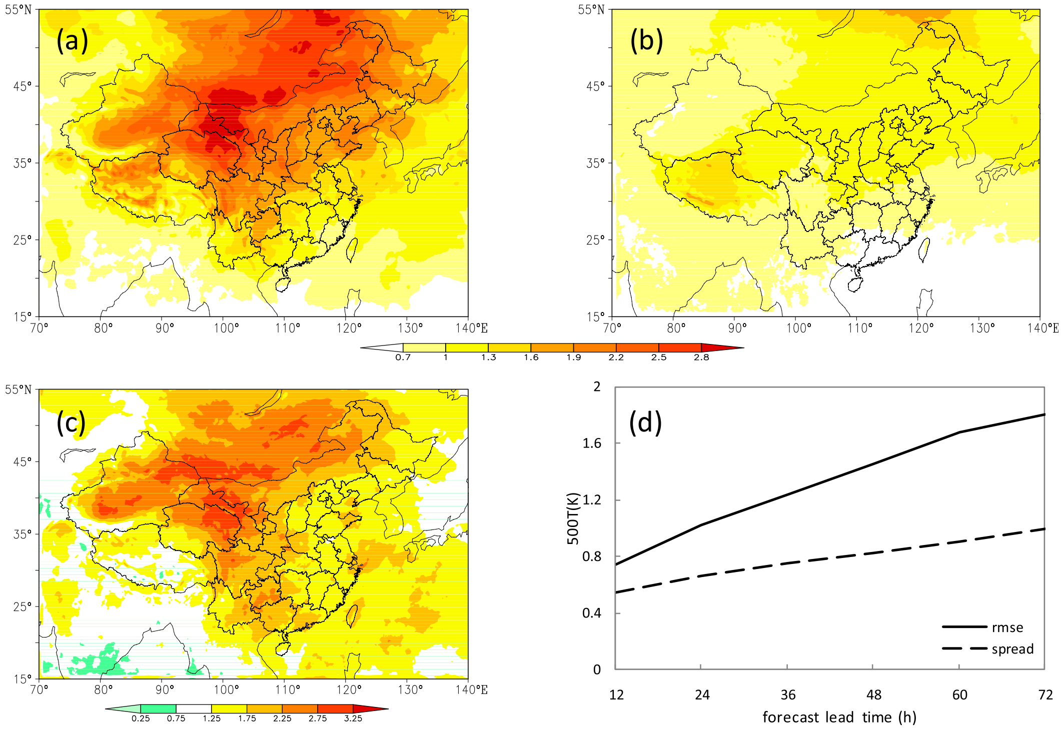

FIG. 1. The horizontal distribution of (a) the ensemble mean forecast RMSE, (b) ensemble spread, and (c) the *consistency* (defined as the ratio of the RMSE to the spread) at 72-h forecast lead time. (d) The domain-averaged ensemble mean forecast RMSE (solid line) and ensemble spread (dashed line) varying with forecast hour. The results are the monthly average for the 0000 UTC cycle during 1–31 Jul 2016. The variable is 500-hPa temperature.

Jones et al. 2007; Du et al. 2015). Initial condition (IC) uncertainty is addressed by an ensemble transform Kalman filter (ETKF) scheme (Bishop et al. 2001; Wang and Bishop 2003; Wang et al. 2004; Bowler et al. 2008; Wei et al. 2008; Kay and Kim 2014). ETKF not only offers a framework to assess the influence of observations on forecast error covariance, but also considers the fastest-growing perturbations that evolve during ensemble forecast cycling. The perturbations derived from ETKF were added to the GRAPES control analysis to form perturbed ICs for the perturbed members. It runs twice a day (at 0000 and 1200 UTC) out to 72 h of forecast length. For a review of ensemble methods for meteorological predictions, readers are referred to Du et al. (2018).

The statistical decaying-average method (Du and Zhou 2011; Li et al. 2011; Cui et al. 2012) is used to remove biases in a forecast. This method uses a simple decaying average to estimate bias, shown in Eq. (1):

$$B_{i,j}(t) = (1-w)B_{i,j}(t-1) + w \times [f_{i,j}(t) - a_{i,j}(t)], \quad (1)$$

where $B_{i,j}(t-1)$ is an accumulated bias from the past, $f_{i,j}(t) - a_{i,j}(t)$ is the current forecast error (forecast minus analysis), and w is a selected weight to partition the two terms. Generally, w is smaller for longer-range forecasts and larger for shorter-range forecasts because forecast error at shorter ranges is more flow dependent (i.e., weighted more on the current error term). In this study, w is set to 2% following Cui et al. (2012), which approximates to a bias accumulated over a 50-day (2% = 1/50) period from the immediate past. The bias estimation [Eq. (1)] is performed twice daily (at 0000 and 1200 UTC). To have a realistic accumulated "past bias" ready for our verification period (1–31 July), the bias estimation step starts on 1 May, so it gives us 60 days (1 May–30 June) prior to our verification starting point (1 July). A debiased forecast $F_{i,j}(t)$ can be obtained by subtracting the decaying averaged bias from a raw forecast as shown in Eq. (2):

$$F_{i,j}(t) = f_{i,j}(t) - B_{i,j}(t). \quad (2)$$

To mimic the operational environment and maximize the benefit from a bias correction, this bias correction method is independently applied to each ensemble member at each forecast lead time and each model grid point for any meteorological variable. By the way, for a comparison we have also tested using the common bias in the ensemble mean to correct each member and

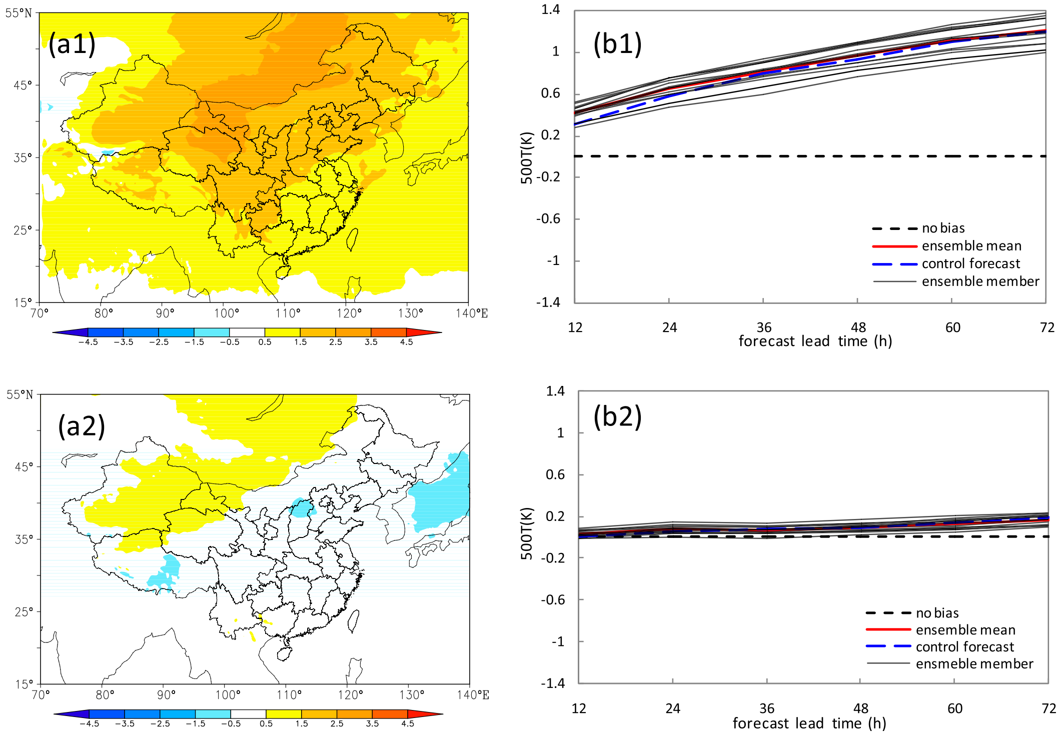

FIG. 2. (a1) The horizontal distribution of ensemble mean forecast bias at 72-h forecast lead time. (b1) The domain-averaged biases of ensemble mean (red solid line), control forecast (blue dashed line), and perturbed members (gray dotted line) varying with forecast hour. (a2),(b2) As in (a1),(b1), but for the debiased ensemble forecasts. The results are the monthly average for the 0000 UTC cycle during 1–31 Jul 2016. The variable is 500-hPa temperature.

found that the same result has been reached from the two approaches. Since the bias is strong and similar among members in this case, there is not much difference in the first-moment (ensemble mean) correction between these two approaches, while it has a benefit of slightly correcting ensemble spread (being smaller or less uncertain in forecasts) when correcting the individual members differently. The three variables (500-hPa and 2-m temperatures, and 250-hPa wind) are shown as examples in this paper.

Verification is performed daily from 1 to 31 July 2016 for forecasts initialized both at 0000 and 1200 UTC (a total of 62 forecasts) over China (15°–55°N, 70°–140°E). These daily results are then averaged to obtain the July monthly average, which will be presented in the next section. Since the bias might be different for different cycles, the monthly averaging is done separately for the 0000 and 1200 UTC cycles. Because the results are almost the same for both cycles, only the 0000 UTC cycle will be presented in this paper. The GRAPES 15-km gridded analysis is used as truth. Given that an analysis itself could contain errors or biases, verification normally favors a model if it is verified against its own analysis. Therefore, the underdispersive nature of the GRAPES EPS could be even more severe than is revealed here if it is verified against station observations.

3. Results

The scores for measuring an EPS's quality were selected based on Du and Zhou (2017). Since ensemble spread and the probability distribution are the two main aspects to portray how an EPS performs, spread–skill relationships and rank histograms are used to verify ensemble spread, and the continuous ranked probability score (CRPS), reliability, and relative operating characteristic (ROC) curve are used for the probability distribution. For a description of each of the scoring rules the reader is referred to Jolliffe and Stephenson (2003), Du (2007), and Du and Zhou (2017).

a. *Ensemble spread*

Similarity between ensemble spread and ensemble mean forecast error is a desired feature. Figures 1a–c

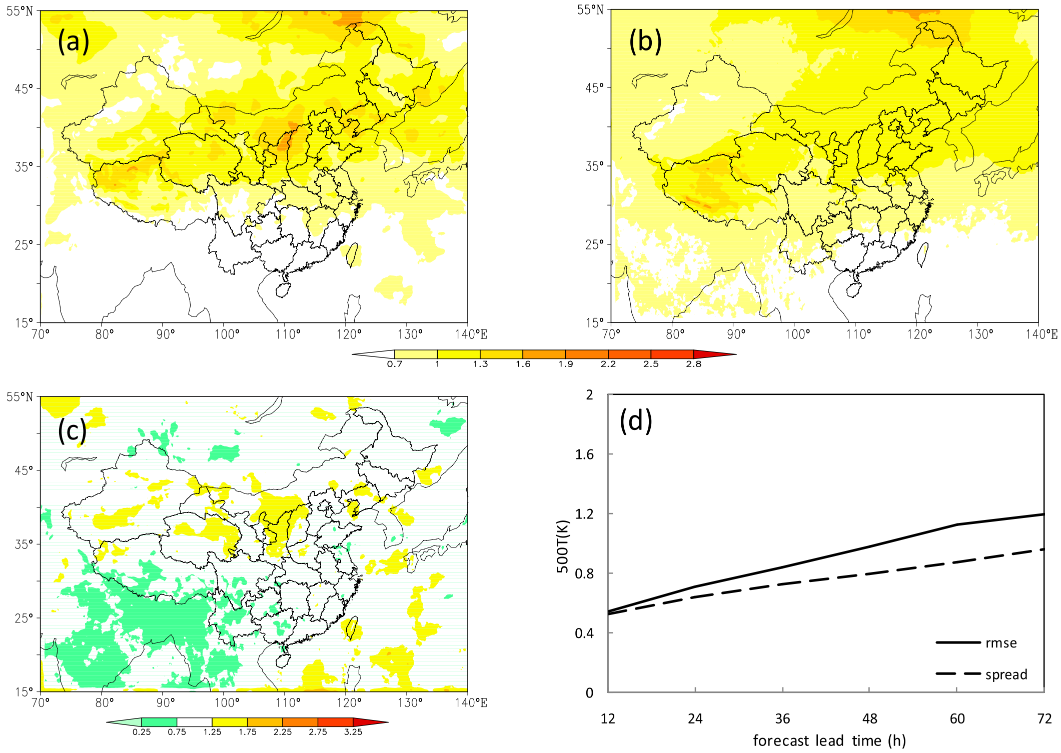

FIG. 3. As in Fig. 1, but for the debiased ensemble forecasts.

compare the spatial distributions of the ensemble spread and the ensemble mean forecast RMSE for 500-hPa temperature at 72-h lead time. The magnitude of ensemble spread (Fig. 1b) is obviously too small compared to the ensemble mean forecast RMSE (Fig. 1a) over all of China for GRAPES-REPS. For example, the maximum RMSE is 2.8, while the maximum spread is only 1.6. To quantitatively compare spread and RMSE grid point by grid point, the monthly averaged "consistency" (defined as the ratio of RMSE to ensemble spread with a perfect *consistency* of 1.0) is shown in Fig. 1c, which shows severe underdispersion almost everywhere over the domain (the ratio $\gg$ 1.0). Figure 1d shows the evolution of domain-averaged RMSEs and spreads with forecast hour. First, we can see that this severe underdispersion is true not only at the 72-h lead time (1.8 in

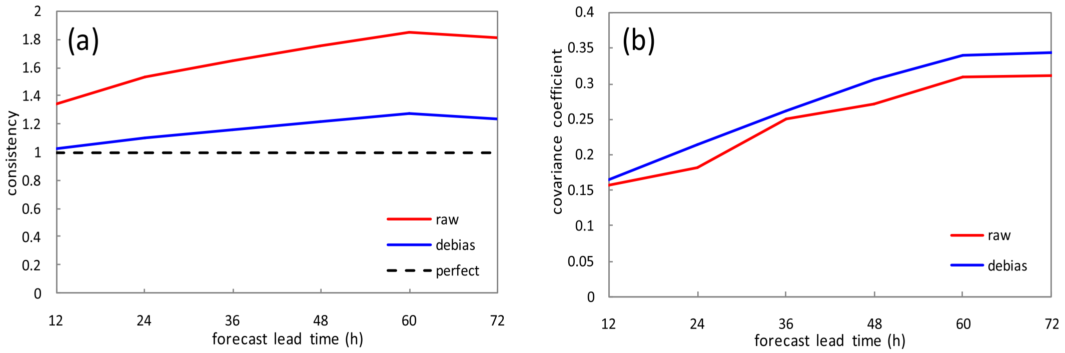

FIG. 4. The domain-averaged values of (a) *consistency* and (b) spatial correlation between the ensemble mean forecast RMSE and ensemble spread before (red) and after (blue) the bias correction, varying with forecast hour. The results are the monthly average for the 0000 UTC cycle during 1–31 Jul 2016. The variable is 500-hPa temperature. The improvement in *consistency* is statistically significant at 99.8% level (*t* test). The improvement in covariance coefficient is statistically not significant at 12 and 36 h but significant at 90%, 50%, 50%, and 50% levels at 24-, 48-, 60-, and 72-h lead times, respectively.

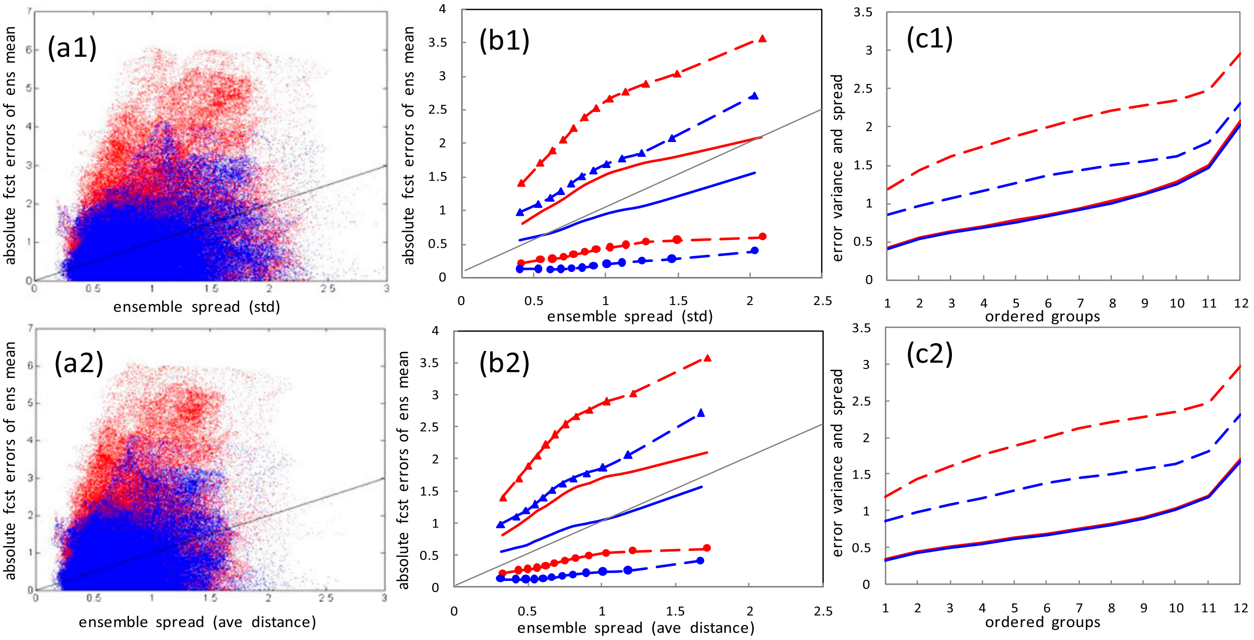

FIG. 5. Spread–skill relationship at 72-h forecast lead time for 500-hPa temperature. (a1),(a2) The scatter diagrams of "absolute forecast error of ensemble mean vs ensemble spread" for 0000 UTC 1 Jul 2016; (b1),(b2) the average errors (solid lines) of (a1),(a2) (but for the month of 1–31 Jul 2016) plotted together with the error variance (+1 and −1 standard deviation in dashed lines); and (c1),(c2) derived from (b1),(b2), but directly showing the error variance (2 × error standard deviation, dashed lines) vs ensemble spread (solid lines). All the horizontal axes are 12 spread bins (each bin has the same sample size: 13 805 grid points), expressed by the average spreads of each bin in (a1),(a2) and (b1),(b2) and by bin numbers in (c1),(c2). The raw ensemble is in red, the debiased is in blue. In (a1),(b1),(c1) "ensemble spread" is defined as the standard deviation of members with respect to the ensemble mean, while in (a2),(b2),(c2) it is defined as the average distance of members to the ensemble mean to better match the absolute forecast error. The black straight lines in (a1),(a2) and (b1),(b2) indicate where "spread is equal to forecast error in magnitude."

RMSE vs 0.95 in spread) but for all forecast lead times. Second, the growth rate of ensemble spread is lower than that of forecast error, with the result that the underdispersion becomes more severe as forecast lead time increases.

The above scores show that the GRAPES-REPS is severely underdispersive. However, does this result really reflect a problem in the ensembling technique used by this EPS or is it mainly a deficiency of the base model? Figures 2a1 and 2b1 show the forecast bias of 500-hPa temperature, where Fig. 2a1 is the horizontal distribution of the ensemble mean forecast bias at the 72-h lead time and Fig. 2b1 is the domain-averaged biases of individual members and the ensemble mean. It clearly shows that a strong systematic warm bias is present everywhere for all members. For example, in the ensemble mean, the maximum monthly average warm bias exceeds 2.5 K (Fig. 2a1). Since model bias stems primarily from a deficiency of the model but not from the ensembling technique (although in certain circumstances model bias could also stem from an unrealistic ensemble technique such as "noise-induced drift"; Weisheimer et al. 2014), a verification truly revealing EPS quality should prevent the results from being contaminated by model bias. Therefore, we applied the decaying-average bias correction method [Eqs. (1) and (2)] to each ensemble member to remove its forecast bias from the raw ensemble data (see section 2). After removing the bias, both the ensemble mean (Fig. 2a2) and individual members (Fig. 2b2) are indeed nearly bias free. Figure 3 repeated the results of Fig. 1 but was based on the debiased data. Because of the significant error reduction in debiased ensemble forecasts, ensemble spread (remaining similar before and after the bias correction) is now very close to the ensemble mean forecast RMSE in magnitude for the entire domain (cf. Figs. 3a and 3b) and all forecast hours (Fig. 3d). Most areas show a nearly perfect match between ensemble spread and ensemble mean forecast RMSE in magnitude (the *consistency* value ranges from 0.75 to 1.25, the white area in Fig. 3c). The domain-averaged *consistency* (Fig. 4a) is now reduced from severe underdispersion (1.4–1.9) to the nearly perfect value "1.0" (1.0–1.3) over all lead times. This improvement in *consistency* is statistically significant at the 99.8%

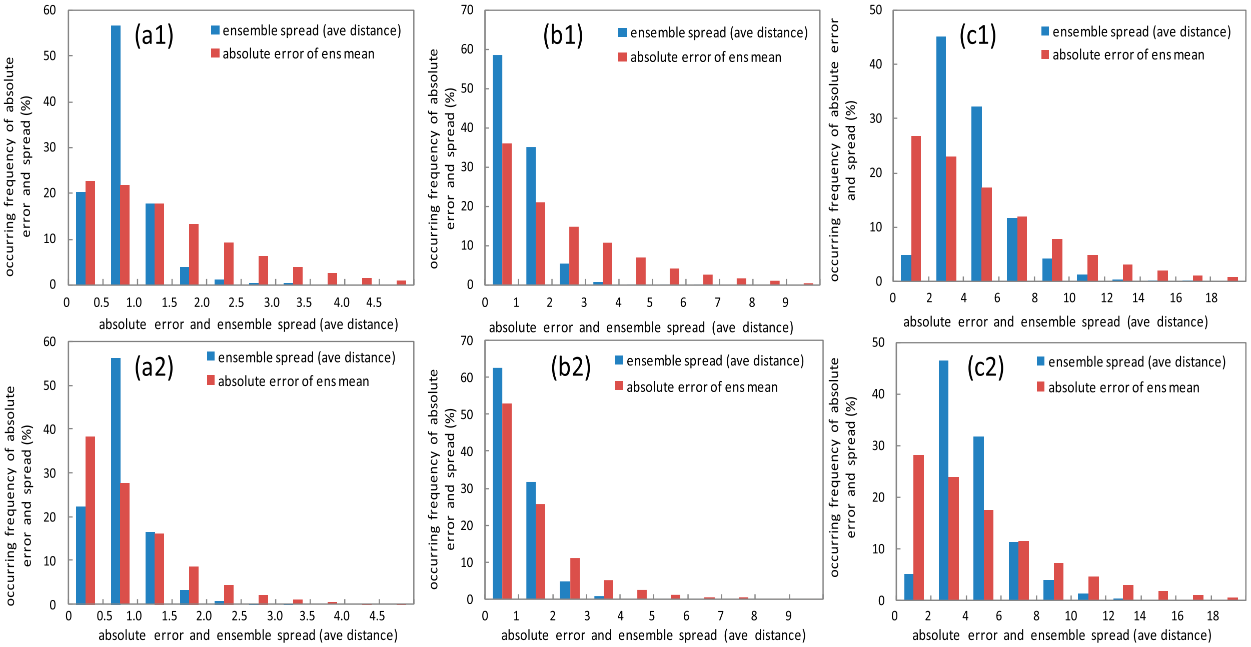

FIG. 6. The occurrence frequency distributions of absolute forecast error of the ensemble mean (blue) and ensemble spread at the same magnitude (a1),(b1),(c1) before and (a2),(b2),(c2) after the bias correction at 72-h forecast lead time for (a1),(a2) 500-hPa and (b1),(b2) 2-m temperatures, and (c1),(c2) 250-hPa zonal wind. Statistics are derived from the entire model forecast domain over the entire month of July 2016. Here the spread is defined as the average distance of members to the ensemble mean. Note that the same result is seen if spread is defined as the standard deviation of members with respect to the ensemble mean (not shown).

level (*t* test). In other words, without the model bias, the GRAPES-REPS has a nearly perfect spread size that can estimate the magnitude of ensemble mean forecast RMSE quite well.

Besides the size match between spread and ensemble mean forecast error, ensemble spread is also expected to simulate the spatial structure of the ensemble mean forecast error (Du et al. 2014). Figure 4b is the spatial correlation between ensemble spread and the ensemble mean forecast RMSE of 500-hPa temperature before and after the bias correction. Unlike the magnitude improvement, the improvement in spatial structure matching is not dramatic but only slight. For example, the correlation is increased by about 13% from 0.31 (raw forecast) to 0.35 (debiased forecast) at the 72-h lead time. The improvement is statistically not significant at the 12- and 36-h lead times but significant at 90%, 50%, 50%,

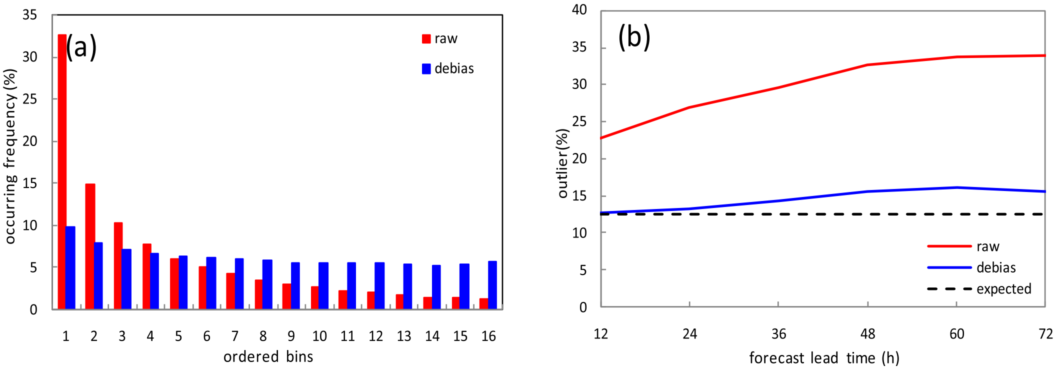

FIG. 7. (a) The rank histograms at 72-h forecast lead time, and (b) the outliers over forecast hour before (red) and after (blue) the bias correction. The results are the monthly average for the 0000 UTC cycle during 1–31 Jul 2016. The variable is 500-hPa temperature. The improvement in the outlier is statistically significant at 99.9% level (*t* test).

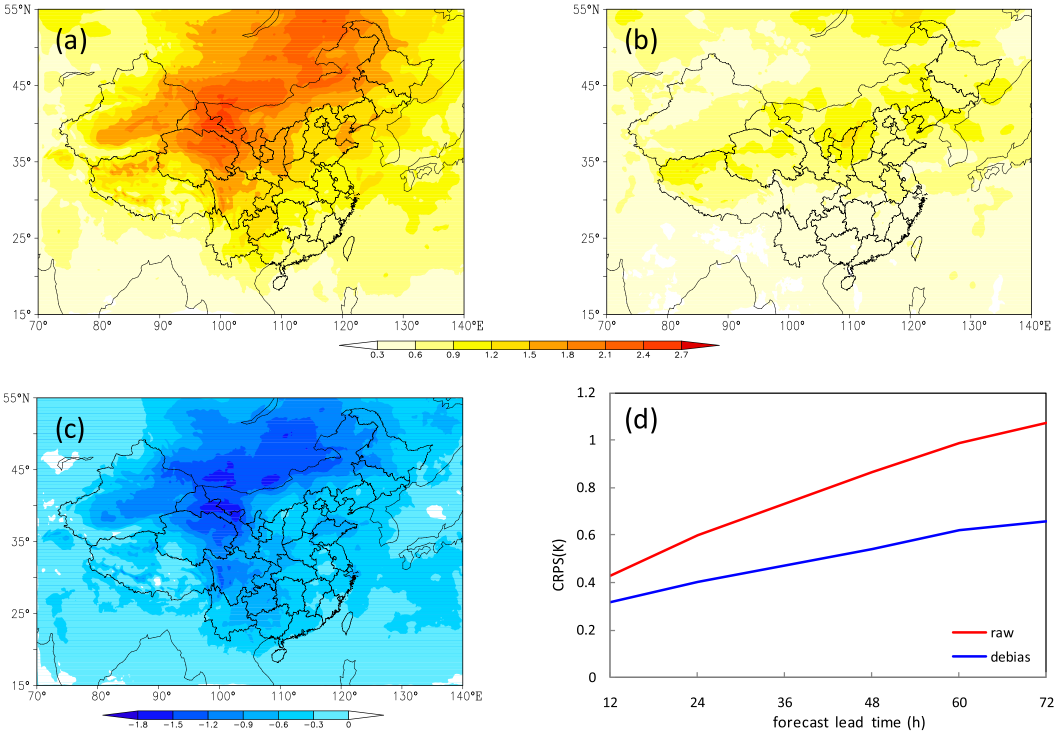

FIG. 8. The horizontal distribution of the CRPS for (a) raw and (b) debiased ensemble forecasts, as well as (c) the difference in CRPS between the two (debiased − raw) at 72-h forecast lead time. (d) The domain-averaged CRPSs before (red) and after (blue) the bias correction, varying with forecast hour. The results are the monthly average for the 0000 UTC cycle during 1–31 Jul 2016. The variable is 500-hPa temperature. CRPS is a negatively oriented score, which is the smaller the better. The reduction in the CRPS is statistically significant at 99.8% level (*t* test) for all forecast hours.

and 50% levels at 24-, 48-, 60-, and 72-h lead times, respectively.

Since the one-to-one relationship between spread and forecast skill is currently not strong (Whitaker and Loughe 1998), researchers turned to look at the statistical relationship between error variance and ensemble variance (spread-to-spread aspect) aiming to use ensemble variance to predict error variance but not the error itself (e.g., Wang and Bishop 2003; Kolczynski et al. 2011; Du 2012). Figure 5 shows the error variance of the ensemble mean forecast over binned ensemble spread. Ensemble spread is divided into 12 bins between its minimum and maximum values, where each bin contains the same number of error samples (13 805 grid points at each bin). Ensemble spread is defined in two formats: one is the commonly used standard deviation (“std”) of members with respect to the ensemble mean (Figs. 5a1, 5b1, and 5c1) and another is the average distance (“ave distance”) of members to the ensemble mean (Figs. 5a2, 5b2, and 5c2) particularly defined to better match the error definition (absolute error). The left panel scatter diagrams (Figs. 5a1 and 5a2) show the variation of individual error points with spread for one cycle (0000 UTC 1 July 2016), the middle panels (Figs. 5b1 and 5b2) are the averages of the left panel (but for the entire month of 1–31 July 2016) together with the error variance (+1 and −1 error standard deviation in dashed lines), and the right panels (Figs. 5c1 and 5c2) are the error variance (2 × error standard deviation, dashed lines) versus ensemble spread (solid lines) over the 12 bins. Although the general statistical relationships are the same for both raw (red) and debiased (blue) ensembles [i.e., forecast error (error variance) increases as spread (ensemble variance) increases on average (Figs. 5b1, 5b2, 5c1, and 5c2)], the debiased ensemble is better at matching between error and spread. For the debiased ensemble, not only is the average error closer to the spread (Figs. 5b1 and 5b2), the error variance is also closer to the ensemble variance (Figs. 5c1 and 5c2). After the removal of bias, forecast error variance becomes smaller indicating a better EPS, which provides a sharper and more reliable forecast. Figures 6a1 and 6a2 are the frequency distributions of the forecast error and spread occurrence at the same magnitude before and after the bias correction for 500-hPa temperature. The error

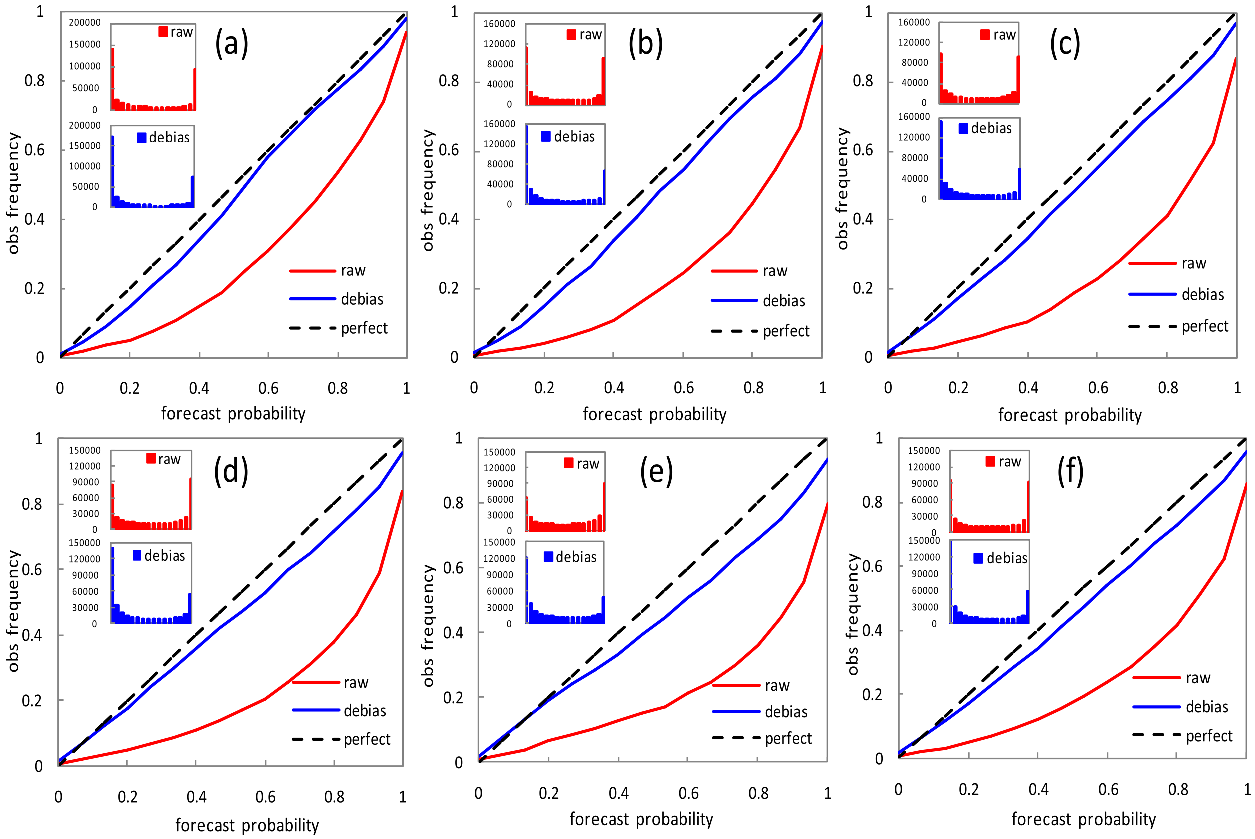

FIG. 9. Reliability diagrams before (red) and after (blue) bias correction for the forecast lead time at (a) 12, (b) 24, (c) 36, (d) 48, and (e) 72 h. (f) The average of all forecast hours (12–72 h). Probability of exceeding 2° over climatology is used. The results are the monthly average for the 0000 UTC cycle during 1–31 Jul 2016. The variable is 500-hPa temperature. The difference between the two is statistically significant at 99.99% level (t test) for all forecast hours.

distribution (red) becomes closer to the spread distribution (blue) because of the increase of "smaller error" and decrease of "larger error" occurrence. In other words, the ensemble spread represents forecast error better after the systematic error is removed. Note that the spread is defined as the average distance of members to the ensemble mean in Fig. 6 but the same result is seen if spread is defined as the standard deviation of members with respect to the ensemble mean (not shown).

Another common verification metric for ensemble spread is the rank histograms (Talagrand et al. 1997; Hamill and Colucci 1997; Hamill 2001; Candille and Talagrand 2005; Du and Zhou 2017). For an ideal EPS that can sample forecast uncertainty well, a flat distribution should be expected over $n + 1$ sorted bins ranging from the smallest to the largest values (n is the total number of ensemble members). The sum of the two end bins (the "outlier") indicates how often an observed event falls outside of the ensemble envelope (being either smaller than the ensemble minimum or greater than the ensemble maximum at a given location and time). For a perfect EPS with n members, the theoretically expected outlier is $2/(n + 1)$. Therefore, the expected outlier for the GRAPES-REPS should be around 12.5% since it has 15 members. Figure 7a compares the rank histograms before (red) and after (blue) the bias correction for 500-hPa temperature at a 72-h lead time. Before the bias correction, the distribution is severely skewed to the left (L shaped) indicating a strong warm bias, resulting in a much too high outlier of 34% (Fig. 7b). After the bias correction, the distribution becomes almost flat (Fig. 7a) and the outlier is close to the expected value 12.5% over all forecast lead times (slightly increasing with forecast hour, Fig. 7b). This improvement is statistically significant at the 99.9% level based on a t test. The dramatic change from very bad to near perfect of the GRAPES-REPS spread (magnitude) before and after the bias correction demonstrates the importance of removing forecast bias prior to verifying an ensemble of forecasts. Otherwise, the conclusions drawn could be very misleading.

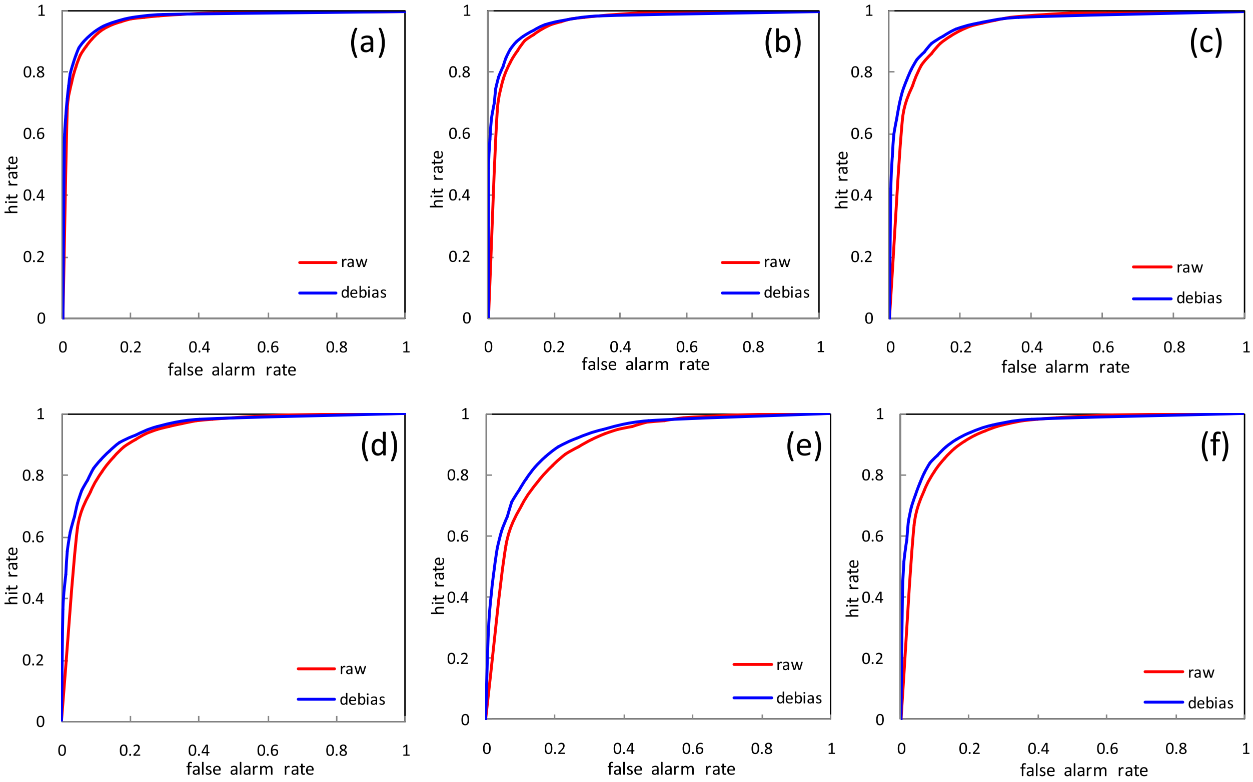

FIG. 10. ROC diagrams before (red) and after (blue) bias correction for the forecast lead time at (a) 12, (b) 24, (c) 36, (d) 48, and (e) 72 h. (f) The average of all forecast hours (12–72 h). Probability of exceeding 2° over climatology is used. The results are the monthly average for the 0000 UTC cycle during 1–31 Jul 2016. The variable is 500-hPa temperature. The difference between the two is statistically significant at 80%, 95%, 98%, 99%, 99.9%, and 99.5% levels (*t* test) for 12, 24, 36, 48, 72, and the average of all forecast hours, respectively.

b. Probabilistic forecast

One of the main reasons for employing an ensemble method to forecast weather is to provide flow-dependent probabilistic information on a range of possible solutions. It is reported that an ensemble forecast can provide skillful probabilistic information valuable to multiple users related to a particular weather event (Stensrud and Yussouf 2005). Since the bias in forecasts will impact the mean position of an ensemble distribution, probabilistic forecasts should be also significantly improved if bias is removed from raw ensemble forecasts. CRPS is used to measure the absolute error between forecast probability and observations (either 0% or 100%) (Hersbach 2000; Grimit et al. 2006; Zhu and Toth 2008). The smaller the CRPS score, the better the probabilistic forecast is by exhibiting higher resolution (sharper) and being more reliable. Figure 8 compares the CRPS of 500-hPa temperature at the 72-h lead time before and after the bias correction. The probabilistic forecast error is indeed much reduced after the bias is removed (cf. Figs. 8b and 8a). This error reduction is almost everywhere within the entire domain (all negative values in Fig. 8c). It is true not only for 72 h but for all forecast lead times (Fig. 8d). Since the bias grows with the increase of forecast lead time (Fig. 2b1), the improvement in probabilistic forecasts also becomes more prominent with forecast time. This CRPS reduction is statistically significant at the 99.8% level (*t* test) for all forecast hours.

Statistical reliability is another important property of probabilistic forecasts and key information for cost–lost-based decision-making (Du and Deng 2010). Reliability measures whether probabilistic forecasts are statistically coherent with observations over a large number of cases for a forecasting system. For a perfectly reliable system its forecast probabilities should exactly match observed frequencies. Therefore, in a reliability diagram the diagonal line represents perfect reliability. Figure 9 shows the reliability curves of 500-hPa temperature at different forecast lead times before and after the bias correction (probability threshold is defined as 2° over climatological value, viz., +2° anomaly threshold). Before the bias correction, forecast probabilities apparently exceed the corresponding observed frequencies, indicating that the GRAPES-REPS is greatly overconfident for all ranges

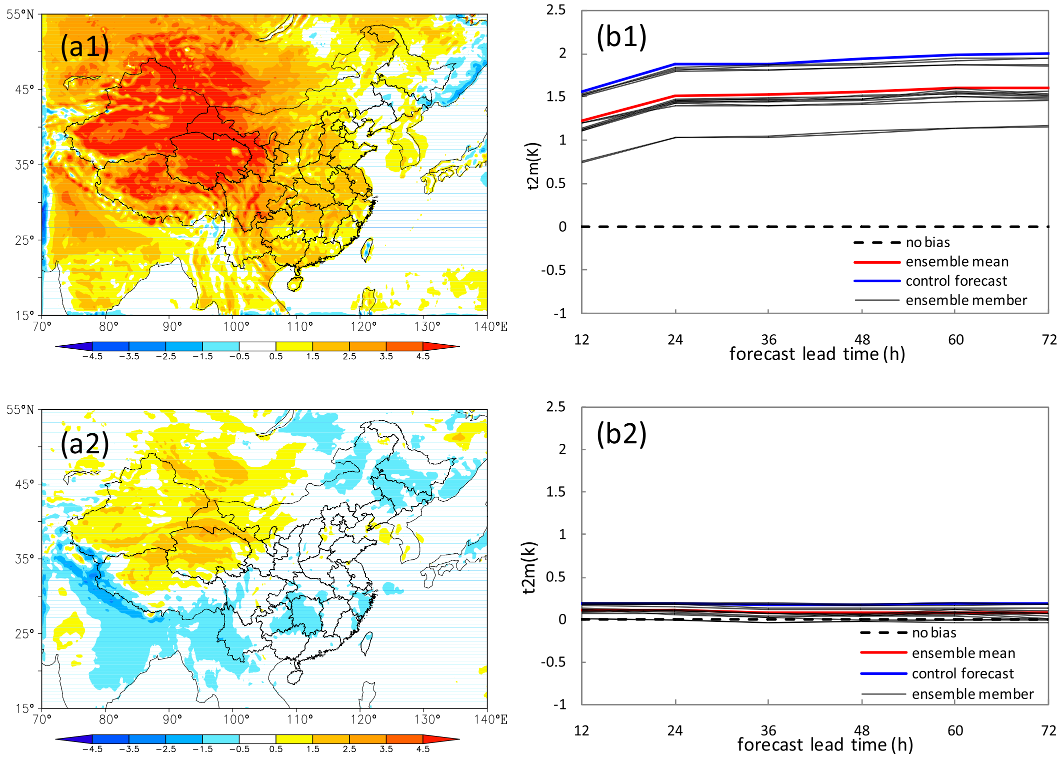

FIG. 11. As in Fig. 2, but for 2-m temperature.

of probabilities because of the strong warm bias. However, after removing the warm bias the reliability curves almost coincide with the perfect diagonal line. In other words, originally overconfident probability becomes perfectly reliable. This improvement is statistically significant at the 99.99% level for all forecast hours (*t* test).

Since relative operating characteristics (ROC) is less sensitive to forecast bias, Fig. 10 compares the ROC curves of 500-hPa temperature at different forecast lead times before and after the bias correction (the probability threshold is the same as used for the reliability calculation). It shows that the improvements with the bias removed are statistically significant at all forecast lead times in a *t* test (80%, 95%, 98%, 99%, and 99.9% for 12, 24, 36, 38, and 72 h, respectively). All the CRPS, reliability, and ROC results above demonstrate that a poor probabilistic forecast derived from the raw ensemble forecasts is not due to the deficiency of EPS design but the deficiency of the model itself (strong warm bias). Therefore, without removing model bias, ensemble verification metrics cannot truly reflect the quality of an EPS.

Since the bias of surface variables is normally less uniform in space than that of upper-air variables, we have repeated the above experiment with 2-m temperature that possesses an even stronger bias. The results are presented in Figs. 11 and 12. Qualitatively speaking, the result is the same as what we have seen for the 500-hPa temperature. The strong warm bias of the raw forecasts (Figs. 11a1 and 11b1) has been noticeably reduced by the bias correction procedure for all individual members including the ensemble mean (Figs. 11a2 and 11b2). The comparison of verification results between the raw and debiased ensembles is summarized by Fig. 12 and Figs. 6b1 and 6b2. Similar to the 500-hPa temperature, significant improvements have been seen in both ensemble spread quality (Figs. 12a–c and Figs. 6b1 and 6b2) and probabilistic forecasts (Figs. 12d–f) after the bias is corrected. As a result one's view toward the GRAPES EPS will also be very different, as it changes from a poor EPS to a reasonably good EPS. Once again this suggests that without removing model bias, ensemble verification metrics will give us a wrong impression about an EPS's performance.

c. *Ensemble verification with little model bias*

The above results have demonstrated that without removing model bias, the conclusion drawn from verification will be very misleading about the quality of an EPS. However, one can imagine that removing the bias should have little impact on the verification result if a field has no or little bias. To prove this we chose 250-hPa

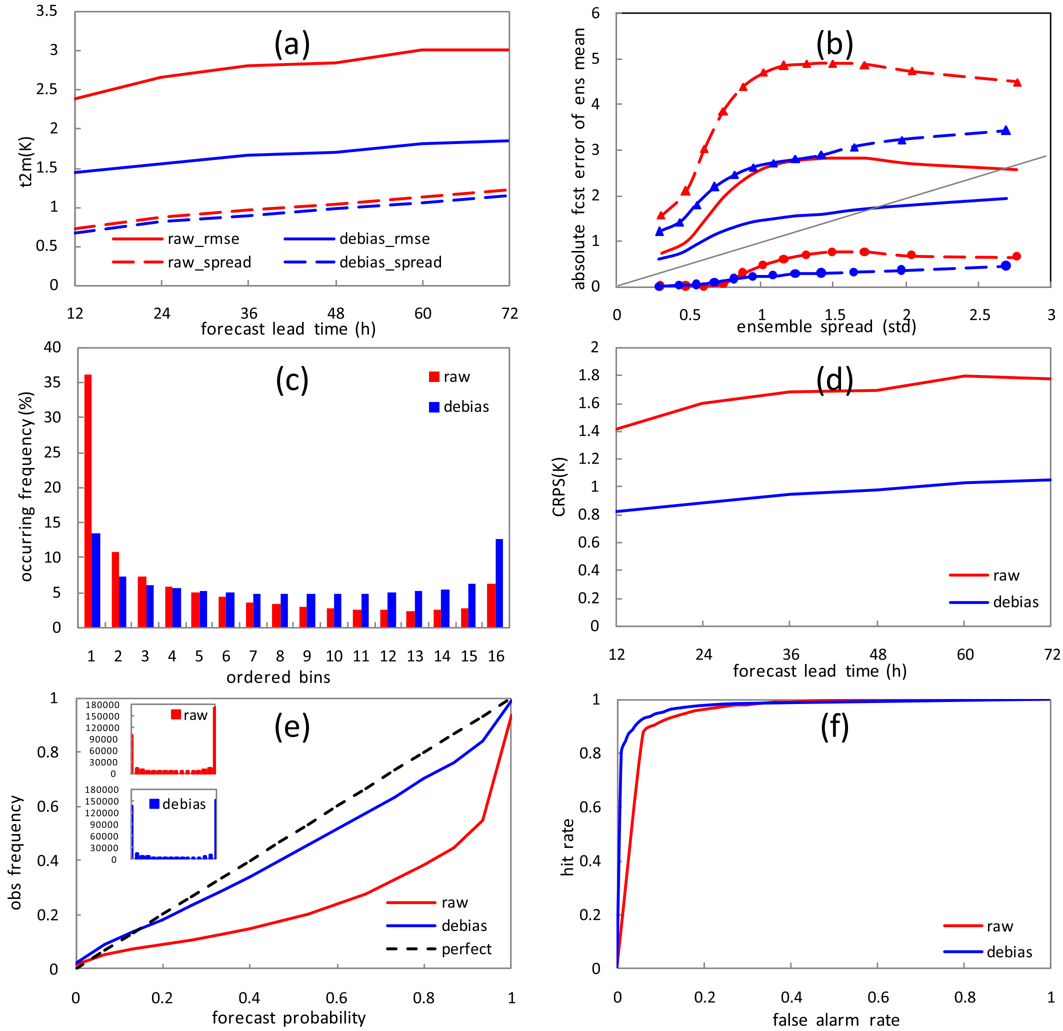

FIG. 12. (a) The domain-averaged ensemble mean RMSE (solid line) vs ensemble spread (dashed line), (b) rank histograms (72 h), (c) spread–skill relation (72 h), (d) the domain-averaged CRPS, (e) reliability diagram (72 h), and (f) ROC diagram (72 h) before (red) and after (blue) the bias correction. Probability of exceeding 20°C is used for (e),(f). The results are the monthly average for the 0000 UTC cycle during 1–31 Jul 2016. The variable is 2-m temperature.

zonal wind *U* to demonstrate since it has little or weak systematic bias in the raw forecasts, as shown in Figs. 13a1 and 13b1 (slight low bias in wind speed). Unlike the 500-hPa and 2-m temperature forecasts (Figs. 2 and 11), the bias correction procedure has much less impact on either the ensemble mean forecast or individual members of 250-hPa *U* (Figs. 13a2 and 13b2). The comparisons of verification summary statistics before and after the bias correction are presented in Fig. 14 and Figs. 6c1 and 6c2. It is shown that the bias correction indeed has only a minor impact on the assessment of all aspects of ensemble quality including the spread–skill relationship (Figs. 14a,b and Figs. 6c1 and 6c2), spread distribution and outlier (Fig. 14c), and the sharpness and reliability of probabilistic forecasts (Figs. 14d–f). The probability threshold “exceeding 5 m s^{-1} over climatology” is used in Figs. 14e and 14f. Only slight improvements are seen given the weak wind bias in the raw forecasts. This slight difference in verification metrics certainly will not change the assessment conclusion about GRAPES-REPS quality for 250-hPa *U* forecasts.

4. Summary and discussion

This study demonstrates how model bias can adversely affect the assessment about the quality of an EPS using common ensemble verification metrics. The 15-member GRAPES-REPS was verified twice daily

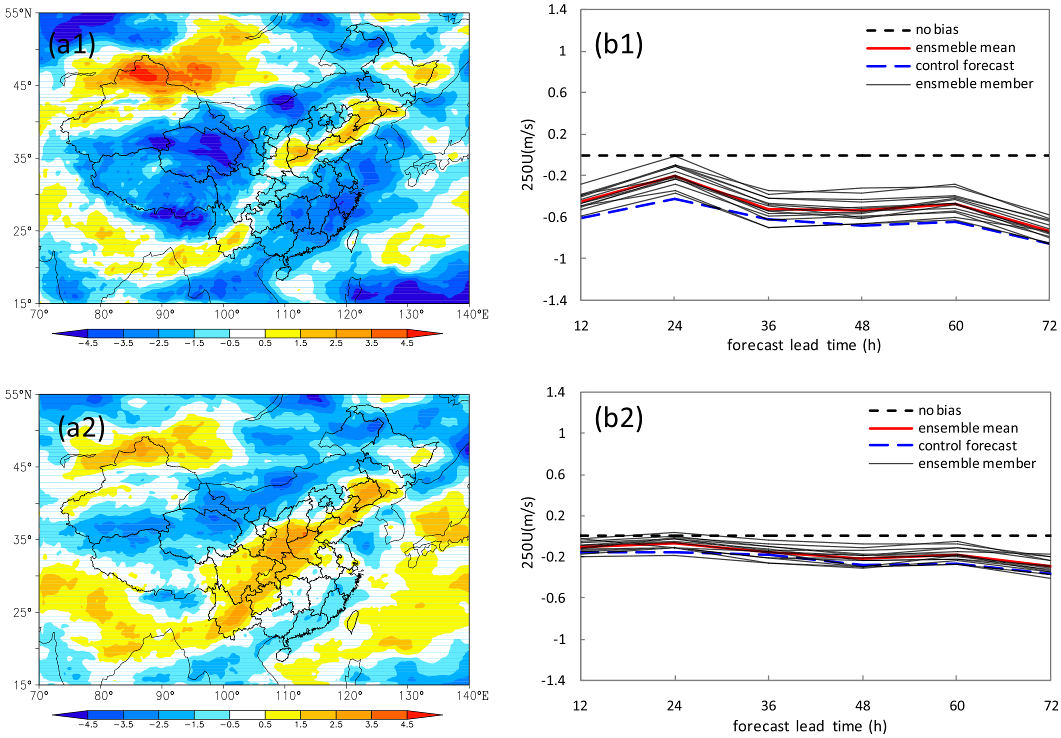

FIG. 13. As in Fig. 2, but for 250-hPa zonal wind *U*.

(at 0000 and 1200 UTC) over China for a period of one month (1–31 July 2016). Three variables (500-hPa and 2-m temperatures, and 250-hPa wind) are selected to represent "strong bias" (both at upper air and surface) and "weak bias" situations, respectively. Ensemble spread and probabilistic forecast are assessed and compared before and after a bias correction. The spread–skill relationship (ensemble mean forecast error vs spread and error variance vs ensemble variance) and rank histograms are used to verify the quality of ensemble spread; CRPS, reliability diagrams, and ROC curves are used to measure the quality of probabilistic forecasts. The decaying-average method is used in the bias correction, which is done for each ensemble member and each forecast hour separately to mimic an operational environment as well as to maximize its benefit.

The results show that the conclusions drawn from ensemble verification about the EPS quality are dramatically different with or without model bias. This is true for both ensemble spread and probabilistic forecasts. For example, for 500-hPa and 2-m temperatures the GRAPES-REPS is severely underdispersive before the bias correction but becomes nearly perfect after it although the improvement in the spread's spatial structure is much less. The spread–skill relationship is noticeably improved too. For the debiased ensemble, not only is the average error closer to the spread, the error variance is also closer to the ensemble variance; the error distribution becomes closer to the spread distribution in their occurrence frequency over the entire forecast domain. The probabilistic forecasts become much sharper and almost perfectly reliable after the bias is removed. All these differences are statistically significant based on a *t* test. Since the forecast's systematic error or bulk bias stems primarily from a model deficiency but not from the ensembling technique, it mainly reflects the quality of a model but not that of an EPS. Therefore, it is necessary to remove systematic (bulk) forecast biases before one can accurately evaluate an EPS. Only when an EPS has no or little systematic forecast bias can ensemble verification metrics reliably reveal the true quality of an EPS without having removed the forecast bias first. This is proved by the 250-hPa wind forecasts. In principle, since an EPS is designed to deal with random error only and not the systematic error of a forecast system, verifying an EPS using its full error cannot truly reveal for what an EPS is intended. Instead, it is the random error component that needs to be used for verifying an EPS. Note that not all bias can be easily removed in a forecast. The bias that can be removed is normally the systematic bulk bias, not the flow-dependent bias. Bulk bias mainly impacts the mean position of an ensemble distribution, while flow-dependent bias could impact the ensemble distribution

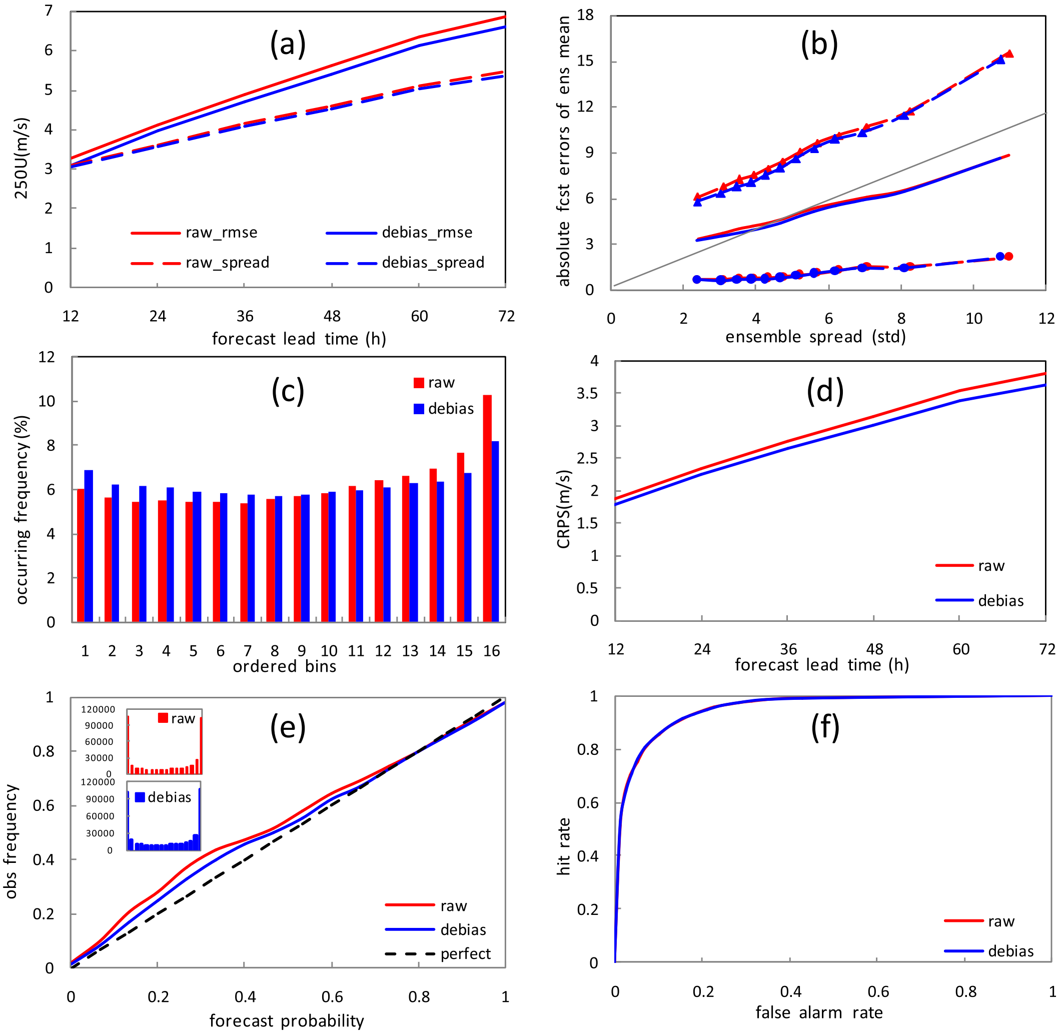

FIG. 14. As in Fig. 12, but for 250-hPa zonal wind U. Probability of exceeding 5 m s^{-1} over climatology is used for (e),(f).

itself. By the way, it needs to be pointed out that a bias correction approach is used to remove systematic error as an approximation in this study, which is suitable and probably the only way for real-time forecasts in an operational environment. Since this bias correction approach worked very effectively for this case (as shown by Figs. 2 and 11), it has served well as a demonstration in this study. However, for the purpose of a pure EPS verification, it is recommended that the total forecast error should be decomposed into random and systematic errors before we verify an EPS. An EPS should be strictly verified against the random component but not the systematic component. This study confirms that model bias will not only negatively impact verification metrics related to ensemble mean, but also those related to spread–skill relation and probabilistic forecasts. Negative impacts have been also seen even for those verification metrics that do not directly involve mean position but compare ensemble distributions such as ensemble variance versus forecast error variance, although their sensitivity might be in a lesser degree.

We are confident that the conclusion from this study can be generalized to a multimodel EPS. Removing forecast bias might be even more important for a multimodel EPS situation. For example, opposite biases could present in different participating models, which leads to large but spurious and even overdispersive spread such as in the NCEP SREF (Du et al. 2015). However, a bias correction should correct this fake "overdispersion" result. Given our results, the implication seems to be that unrealistic EPS probabilities may be less due to imperfections in EPS methodologies and more due to model bias. And, hence, both deterministic and EPS guidance will be improved by addressing those biases. In other words, EPS developers should not be expected to introduce methods to dramatically increase ensemble spread (either by perturbation method or statistical calibration to its second moment) to achieve reliability. Instead, the preferred solution is to reduce the model's first-moment bias through prediction system developments including better model-perturbing methods to enhance the quality of spread (not the quantity of spread). Stochastic physics is one such model-perturbing method. Since some model biases can in part be due to suboptimal methods of treating model uncertainty like "noise-induced drift," better model uncertainty treatments such as stochastic physics will help to not only increase ensemble spread but also reduce model bias (Berner et al. 2015). Another implication of this study is that forecast products should be produced from debiased ensembles rather than raw ensembles when the model bias is substantial. Otherwise, bias in the ensemble mean will lead to errors in the ensemble estimation of the probability distribution of possible true states.

Finally, one needs to keep in mind that even though a perfect reliability has been achieved with no model bias for an EPS, improving ensemble technique is still needed given currently low and imperfect spread–skill relation as shown by Figs. 4b, 5, and 6. As Whitaker and Loughe (1998) indicated that we want an EPS to predict flow-dependent variations in spread. The more an EPS is capable of predicting large variations in spread while remaining reliable, the more useful and skillful it is. Therefore, EPS development that contributes to this is still important.

Acknowledgments. Ms. Mary Hart of NCEP is appreciated for her help to improve the readability of the manuscript. We thank Mr. Fajing Chen and Dr. Hanbin Zhang for providing us with their verification codes. Constructive comments from five reviewers including Drs. Thomas Hamill and Carolyn Reynolds and other three anonymous reviewers have significantly improved this revised manuscript. This work is sponsored by The Special Fund for Meteorological Scientific Research in the Public Interest (GYHY201506005), National Natural Science Foundation of China (NSFC) (91437113 and 41475097), and National Key Technology Research and Development Program of the Ministry of Science and Technology of China (2015BAC03B01).

REFERENCES

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

Bishop, C. H., B. J. Etherton, and S. J. Majumdar, 2001: Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects. *Mon. Wea. Rev.*, **129**, 420–436, https://doi.org/10.1175/1520-0493(2001)129<0420:ASWTET>2.0.CO;2.

Bowler, N. E., A. Arribas, K. R. Mylne, K. B. Robertson, and S. E. Beare, 2008: The MOGREPS short-range ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **134**, 703–722, https://doi.org/10.1002/qj.234.

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

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

Cui, B., Z. Toth, Y. Zhu, 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, 42 pp., http://www.nws.noaa.gov/ost/climate/STIP/uncertainty.htm.

——, 2012: New metrics for evaluating ensemble spread. *21st Conf. on Probability and Statistics in the Atmospheric Sciences*, New Orleans, LA, Amer. Meteor. Soc., https://ams.confex.com/ams/92Annual/webprogram/Paper205776.html.

——, and J. Chen, 2010: The corner stone in facilitating the transition from deterministic to probabilistic forecasts—Ensemble forecasting and its impact on numerical weather prediction. *Meteor. Mon.*, **36**, 1–11.

——, and G. Deng, 2010: The utility of the transition from deterministic to probabilistic weather forecasts—Verification and application of probabilistic forecasts. *Meteor. Mon.*, **36**, 10–18.

——, 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.

——, R. Yu, C. Cui, and J. Li, 2014: Using a mesoscale ensemble to predict forecast error and perform targeted observation. *Acta Oceanol. Sin.*, **33**, 83–91, https://doi.org/10.1007/s13131-014-0426-5.

——, G. DiMego, D. Jovic, B. Ferrier, B. Yang, and B. Zhou, 2015: Short Range Ensemble Forecast (SREF) system at NCEP: Recent development and future transition. *27th Conf. on Weather Analysis and Forecasting and 23rd Conf. on Numerical Weather Prediction* , Chicago, IL, Amer. Meteor. Soc., 2A.5, https://ams.confex.com/ams/27WAF23NWP/webprogram/Paper273421.html.

——, and Coauthors, 2018: Ensemble methods for meteorological predictions. NCEP Office Note 493, in press.

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

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

Garcia-Moya, J. A., A. Callado, P. Escribà, C. Santos, D. Santos-Munoz, and J. Simarro, 2011: Predictability of short-range forecasting: A multimodel approach. *Tellus*, **63A**, 550–563, https://doi.org/10.1111/j.1600-0870.2010.00506.x.

Grimit, E. P., T. Gneiting, V. J. Berrocal, and N. A. Johnson, 2006: The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Quart. J. Roy. Meteor. Soc.*, **132**, 2925–2942, https://doi.org/10.1256/qj.05.235.

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

——, and S. J. Colucci, 1997: Verification of Eta/RSM short-range ensemble forecasts. *Mon. Wea. Rev.*, **125**, 1312–1327, https://doi.org/10.1175/1520-0493(1997)125<1312:VOERSR>2.0.CO;2.

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

Ho, C. K., E. Hawkins, L. Shaffrey, J. Bröcker, L. Hermanson, J. M. Murphy, D. M. Smith, and R. Eade, 2013: Examining reliability of seasonal to decadal sea surface temperature forecasts: The role of ensemble dispersion. *Geophys. Res. Lett.*, **40**, 5770–5775, https://doi.org/10.1002/2013GL057630.

Hopson, T. M., 2014: Assessing the ensemble spread–error relationship. *Mon. Wea. Rev.*, **142**, 1125–1142, https://doi.org/10.1175/MWR-D-12-00111.1.

Jolliffe, I., and D. Stephenson, 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Wiley and Sons, 240 pp.

Jones, M. S., B. A. Colle, and J. S. Tongue, 2007: Evaluation of a mesoscale short-range ensemble forecast system over the northeast United States. *Wea. Forecasting*, **22**, 36–55, https://doi.org/10.1175/WAF973.1.

Kay, J. K., and H. M. Kim, 2014: Characteristics of initial perturbations in the ensemble prediction system of the Korea meteorological administration. *Wea. Forecasting*, **29**, 563–581, https://doi.org/10.1175/WAF-D-13-00097.1.

Kolczynski, W. C., D. R. Stauffer, S. E. Haupt, N. S. Altman, and A. Deng, 2011: Investigation of ensemble variance as a measure of true forecast variance. *Mon. Wea. Rev.*, **139**, 3954–3963, https://doi.org/10.1175/MWR-D-10-05081.1.

Li, L., Y. Li, H. Tian, and B. Cui, 2011: Study of bias-correction in T213 Global ensemble forecast. *Meteor. Mon.*, **37**, 31–38.

McCollor, D., and R. Stull, 2009: Evaluation of probabilistic medium-range temperature forecasts from the North American ensemble forecast system. *Wea. Forecasting*, **24**, 3–17, https://doi.org/10.1175/2008WAF2222130.1.

Reynolds, C. A., J. A. Ridout, and J. G. McLay, 2011: Examination of parameter variations in the U.S. Navy Global Ensemble. *Tellus*, **63A**, 841–857, https://doi.org/10.1111/j.1600-0870.2011.00532.x.

Rodwell, M. J., S. T. K. Lang, N. B. Ingleby, N. Bormann, E. Holm, F. Rabier, D. S. Richardson, and M. Yamaguchi, 2016: Reliability in ensemble data assimilation. *Quart. J. Roy. Meteor. Soc.*, **142**, 443–454, https://doi.org/10.1002/qj.2663.

Saetra, O., H. Hersbach, J.-R. Bidlot, and D. S. Richardson, 2004: Effects of observation errors on the statistics for ensemble spread and reliability. *Mon. Wea. Rev.*, **132**, 1487–1501, https://doi.org/10.1175/1520-0493(2004)132<1487:EOOEOT>2.0.CO;2.

Stensrud, D. J., and N. Yussouf, 2005: Bias-corrected short-range ensemble forecasts of near surface variables. *Meteor. Appl.*, **12**, 217–230, https://doi.org/10.1017/S135048270500174X.

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

——, 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.

Talagrand, O., R. Vautard, and B. Strauss, 1997: Evaluation of probabilistic prediction systems. *Proc. ECMWF Workshop on Predictability*, Reading, United Kingdom, ECMWF, 1–25. [Available from ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, United Kingdom.]

Toth, Z., O. Talagrand, G. Candille, and Y. Zhu, 2003: Probability and ensemble forecasts. *Forecast Verification: A Practitioner's Guide in Atmospheric Science,* I. T. Jolliffe and D. B. Stephenson, Eds., Wiley, 137–163.

Wang, X., and C. H. Bishop, 2003: A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. *J. Atmos. Sci.*, **60**, 1140–1158, https://doi.org/10.1175/1520-0469(2003)060<1140:ACOBAE>2.0.CO;2.

——, ——, and S. J. Julier, 2004: Which is better, an ensemble of positive–negative pairs or a centered spherical simplex ensemble? *Mon. Wea. Rev.*, **132**, 1590–1605, https://doi.org/10.1175/1520-0493(2004)132<1590:WIBAEO>2.0.CO;2.

Wei, M., Z. Toth, R. Wobus, and Y. Zhu, 2008: Initial perturbations based on the ensemble transform (ET) technique in the NCEP global operational forecast system. *Tellus*, **60A**, 62–79, https://doi.org/10.1111/j.1600-0870.2007.00273.x.

Weisheimer, A., S. Corti, T. Palmer, and F. Vitart, 2014: Addressing model error through atmospheric stochastic physical parametrizations: Impact on the coupled ECMWF seasonal forecasting system. *Philos. Trans. Roy. Soc. London*, **372A**, 2018, https://doi.org/10.1098/rsta.2013.0290.

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

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

Zhu, Y., and Z. Toth, 2008: Ensemble based probabilistic forecast verification. *19th Conf. on Probability and Statistics,* New Orleans, LA, Amer. Meteor. Soc., 2.2, https://ams.confex.com/ams/88Annual/techprogram/paper_131645.htm.