

## A Comparison between 2D and 3D Rescaling Masks of Initial Condition Perturbation in a 3-km Storm-Scale Ensemble Prediction System

GUO DENG,<sup>a,b,c</sup> JUN DU,<sup>d</sup> YUSHU ZHOU,<sup>e,f</sup> LING YAN,<sup>e,g</sup> JING CHEN,<sup>a,b,c</sup> FAJING CHEN,<sup>a,b,c</sup> HONGQI LI,<sup>a,b,c</sup>  
AND JINGZHOU WANG<sup>a,b,c</sup>

<sup>a</sup> Earth System Modeling and Prediction Center, Chinese Academy of Meteorological Sciences, Beijing, China

<sup>b</sup> National Meteorological Center, Chinese Academy of Meteorological Sciences, Beijing, China

<sup>c</sup> State Key Laboratory of Severe Weather, Chinese Academy of Meteorological Sciences, Beijing, China

<sup>d</sup> Environmental Modeling Center/NCEP/NWS/NOAA, College Park, Maryland

<sup>e</sup> Key Laboratory of Cloud-Precipitation Physics and Severe Storms, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China

<sup>f</sup> University of Chinese Academy of Sciences, Beijing, China

<sup>g</sup> Key Laboratory of Meteorological Disasters by Ministry of Education/Joint International Research Laboratory of Climate and Environment Change/Collaborative Innovation Center on Forecast and Evaluation of Meteorological Disasters, Nanjing University of Information Science and Technology, Nanjing, China

(Manuscript received 16 April 2022, in final form 26 September 2022)

**ABSTRACT:** Using a 3-km regional ensemble prediction system (EPS), this study tested a three-dimensional (3D) rescaling mask for initial condition (IC) perturbation. Whether the 3D mask-based EPS improves ensemble forecasts over current two-dimensional (2D) mask-based EPS has been evaluated in three aspects: ensemble mean, spread, and probability. The forecasts of wind, temperature, geopotential height, sea level pressure, and precipitation were examined for a summer month (1–28 July 2018) and a winter month (1–27 February 2019) over a region in North China. The EPS was run twice per day (initiated at 0000 and 1200 UTC) to 36 h in forecast length, providing 56 warm-season forecast cases and 54 cold-season cases for verification. The warm and cold seasons are verified separately for comparison. The study found the following: 1) The vertical profile of IC perturbation becomes closer to that of analysis uncertainty with the 3D rescaling mask. 2) Ensemble performance is significantly improved in all three aspects. The biggest improvement is in the ensemble spread, followed by the probabilistic forecast, and the least improvement is in the ensemble mean forecast. Larger improvements are seen in the warm season than in the cold season. 3) More improvement is in the shorter time range (<24 h) than in the longer range. 4) Surface and lower-level variables are improved more than upper-level ones. 5) The underlying mechanism for the improvement has been investigated. Convective instability is found to be responsible for the spread increment and, thus, overall ensemble forecast improvement. Therefore, using a 3D rescaling mask is recommended for an EPS to increase its utility especially for shorter time range and surface weather elements.

**SIGNIFICANT STATEMENT:** A weather prediction model is a complex system that consists of nonlinear differential equations. Small errors in either its inputs or model itself will grow with time during model integration, which will contaminate a forecast. To quantify such contamination (“uncertainty”) of a forecast, the ensemble forecasting technique is used. An ensemble of forecasts is a multiple of model runs at the same time but with slightly “perturbed” inputs or model versions. These small perturbations are supposed to represent true “uncertainty” in inputs or model representation. This study proposed a technique that makes a perturbation’s vertical structure more resemble real uncertainty (intrinsic error) in input data and confirmed that it can significantly improve ensemble forecast quality especially for a shorter time range and lower-level weather elements. It is found that convective instability is responsible for the improvement.


**KEYWORDS:** Ensembles; Numerical weather prediction/forecasting; Short-range prediction; Regional models

### 1. Introduction

A numerical weather prediction (NWP) model is only an approximation of real atmosphere and contains unavoidable errors in various components like initial conditions (ICs) and

model physics. These small errors will nonlinearly grow with time during model integration to contaminate a forecast due to chaotic nature of models (Lorenz 1963; Epstein 1969; Leith 1974). Therefore, it is necessary to quantify predictability or uncertainty associated with a model forecast. Since ensemble forecasting is a model-based dynamical approach to quantify forecast uncertainty, an ensemble prediction system (EPS) has now become a standard modeling system at major numerical weather prediction centers in the world (e.g., Buizza et al. 2018).

A key technical component for an EPS is to perturb ICs. There are many existing IC perturbation methods (see the

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

Corresponding authors: Guo Deng, deng719@cma.gov.cn; Jun Du, jun.du@noaa.gov

DOI: 10.1175/WAF-D-22-0073.1

© 2023 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the [AMS Copyright Policy \(www.ametsoc.org/PUBSReuseLicenses\)](#).

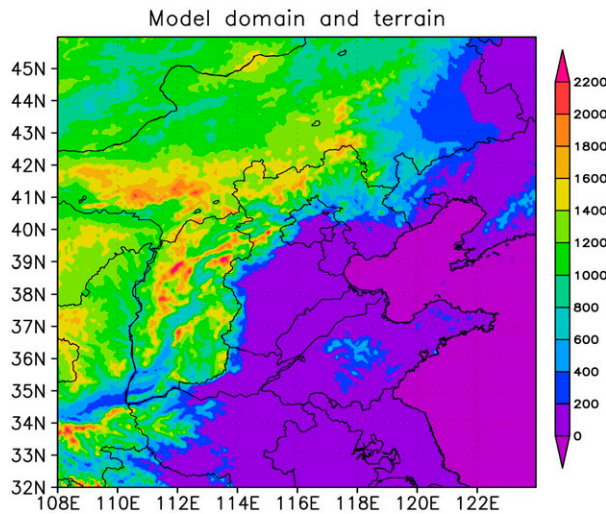


FIG. 1. Model domain ( $32^{\circ}$ – $46.01^{\circ}$ N,  $108^{\circ}$ – $123.99^{\circ}$ E,  $\sim 1600$  km  $\times$  1400 km), about 3-km horizontal resolution with  $468 \times 534$  grid points, plotted together with terrain height (m). There are 2413 auto rain gauge stations within the domain, which are used for precipitation verification.

review of those methods in [Du et al. 2018](#)). Some are random perturbations (e.g., Monte Carlo method), some are just a collection of multiple existing analyses (e.g., multi-analysis approach), some are focusing on error structure in analysis (e.g., breeding method), some are mathematically targeting spread growth at a future time (e.g., singular vector), some are simulating observational errors (e.g., ETKF). [Du et al. \(2018\)](#) has described each of the methods in detail and discussed their advantages and shortages. Based on our own and other's experiences in developing EPSs, three basic scientific principles are behind all those IC-perturbation generation schemes (e.g., [Lacarra and Talagrand 1988](#); [Du 2002](#); [Du et al. 2018](#)). One is the *representativeness* that a perturbation should represent the true uncertainty of an IC by keeping perturbation size similar to analysis uncertainty (intrinsic error in an analysis). Another is the *growing structure* that perturbation's spatial structure should contain atmospheric growing or unstable modes (e.g., baroclinic and convective instabilities) so that a perturbation will grow when a model integrates with time. The third is the *independence* that orthogonality of perturbations among ensemble members needs to be maximized so that perturbations can grow independently into different directions during model integration to fully span forecast uncertainty space. Given a finite ensemble size, the number of ensemble member is normally much smaller than the model's degree of freedom. Therefore, the second and third principles are necessary for a limited-size ensemble to effectively encompass all possible scenarios of a future atmospheric state. Horizontal perturbation structure is taken care of by well-designed perturbation schemes. For example, a blending of ensemble transform Kalman filter (ETKF; [Wang and Bishop 2003](#)) and bred vector (BV; [Toth and Kalnay 1997](#)) is used as a base IC perturbation scheme to provide raw IC perturbation

to be rescaled in this study (see [section 2b](#)). This study will focus on improving the first principle by making an IC perturbation be closer to analysis uncertainty (intrinsic analysis error) in vertical distribution.

To satisfy the first principle, a rescaling mask is normally used to adjust (either amplify or reduce) perturbation magnitude after an initial "raw perturbation" being generated from an IC perturbation scheme. Currently, a rescaling mask is two-dimensional (2D), where a rescaling factor is calculated at a representative model level (reference level) and then applied indistinguishably to all levels such as in the NCEP global ([Toth and Kalnay 1997](#); [Zhou et al. 2016, 2017](#)) and regional EPSs ([Du and Tracton 2001](#)), CMA EPS ([Deng et al. 2010](#); [Liu et al. 2013](#)), and other EPSs (e.g., [Anderson and Anderson 1999](#)). As a result, vertical distribution of IC perturbations with a 2D rescaling mask often does not closely match vertical distribution of analysis uncertainty (e.g., [Zhou et al. 2017](#)). To mitigate the problem, this study extended from 2D to a three-dimensional (3D) rescaling mask in a 3-km storm-scale EPS, then systematically evaluated if it can improve ensemble forecast performance and further investigated how it is achieved.

Extension from 2D to 3D rescaling mask has also gained attention to others. For example, Met Office uses different rescaling factors for planetary boundary layer, stratosphere, and troposphere in a global EPS ([Flowerdew and Bowler 2013](#)). A study has also been done using the NCEP coarse-resolution global EPS (GEFS; [Ma et al. 2014](#)), which showed an improvement in ensemble performance of basic atmospheric state variables in a fall transition season (September–November) with a 3D rescaling mask (derived from a hybrid ensemble data assimilation system). Is this conclusion also valid for a high-resolution regional EPS and other sensible weather elements like precipitation? Thus, [Wang et al. \(2021\)](#) used a 3D rescaling mask to ETKF perturbation in CMA's Global/Regional Assimilation and Prediction Enhanced System-based regional EPS (GRAPES-REPS) and showed an improvement in ensemble performance including precipitation forecast. Following the bred vector's masking strategy, their mask is derived from time-averaged difference between two independent analyses (ECMWF and GRAPES) and varies both horizontally and vertically (only horizontally in the original BV's mask). However, there are three possible deficiencies in their work. One is about the robustness of their result since their work is only based on a very short period of 9 days in spring transition season (7–15 May 2019). The second is that their rescaling factor is not directly related to the analysis uncertainty of its own model's data assimilation system but static differences between two independent analyses. The third is the horizontal variation of their rescaling factor, which will alter the raw perturbation's spatial structure coming out of a perturbation scheme and possibly destroy fast-growing modes. There were other 3D rescaling related works that did not explicitly and systematically compare forecast performance between 2D and 3D masking strategies. For example, [Feng et al. \(2019\)](#) used 3D rescaling in storm-scale ensemble and focused on testing an ensemble-sensitivity analysis-based perturbation method rather than 2D versus 3D comparison for a squall-line case. From the review of the past work, we can see that a systematic comparison between 2D and

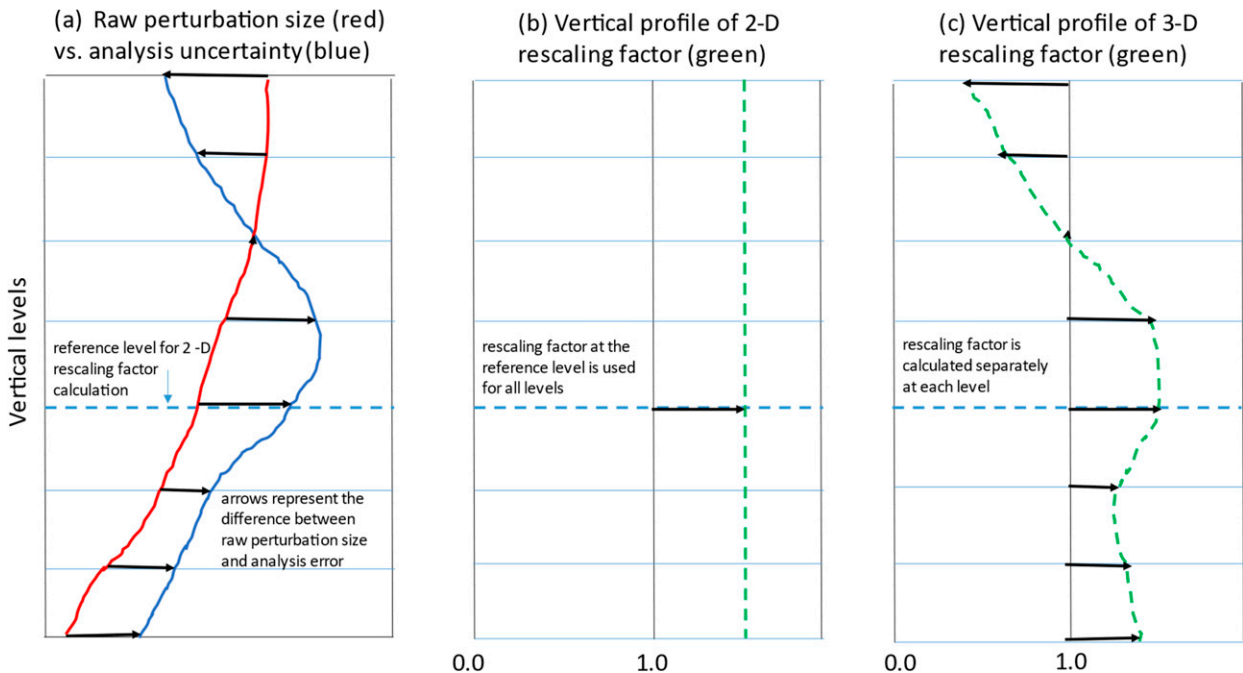


FIG. 2. A schematic illustration how the rescaling factor is calculated in 2D and 3D rescaling masks. (a) Vertical profiles of “raw” IC perturbation from real time (red) and “analysis uncertainty” averaged over a past period (blue), where the arrows indicate their “differences” at each level. (b) 2D rescaling factor (green dashed line) is calculated based on the difference between the raw IC perturbation and analysis uncertainty at a given reference level [e.g., a midlevel shown in (a)] and then applied to all levels uniformly. (c) 3D rescaling factors (green dashed line) are calculated based on the difference between the raw IC perturbation and analysis uncertainty at each individual level [shown in (a)] and then applied to each corresponding level separately. Since the rescaling factor is mathematically defined as a ratio of analysis uncertainty to raw perturbation, it is a non-unit value: 1.0 means “no change” (i.e., raw perturbation will be used as new perturbation), >1.0 means “enlargement” (i.e., raw perturbation will be enlarged as new perturbation), and <1.0 means “reduction” (i.e., raw perturbation will be reduced as new perturbation).

improved 3D rescaling masks in a high-resolution storm-scale EPS over a longer time period including both warm and cold seasons is still needed, which motivates us to perform this study. To have a robust and generalized conclusion, this study will systematically compare the ensemble performances of both atmospheric state variables and precipitation between 2D and 3D rescaling in a 3-km storm-scale regional EPS (GRAPES-EPS) for one summer month and one winter month. The summer and winter months will be verified separately for a comparison. The possible improvement mechanism will also be explored to understand how and why it is achieved. Comparing to Wang et al. (2021), the 3D rescaling mask employed by this study is different in the following two aspects: 1) rescaling factor is derived from a comparison to the estimated analysis uncertainty of its own model’s data assimilation system (rather than a third independent data assimilation system such as ECMWF’s), and 2) rescaling factor varies vertically but not horizontally so that the horizontal structure (fast-growing modes) of raw perturbation remains after rescaling (see section 2b for details).

In the remaining part of this paper, we will describe the ensemble model configuration, the 3D mask design, and data in section 2. Section 3 compares the performances of ensemble forecasts between 3D and 2D masks in terms of ensemble mean, spread and probabilistic forecasts, as well

as an exploration of possible underlying mechanism. A summary and discussion are given in section 4.

## 2. Experiment design

### a. Model and EPS configuration

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

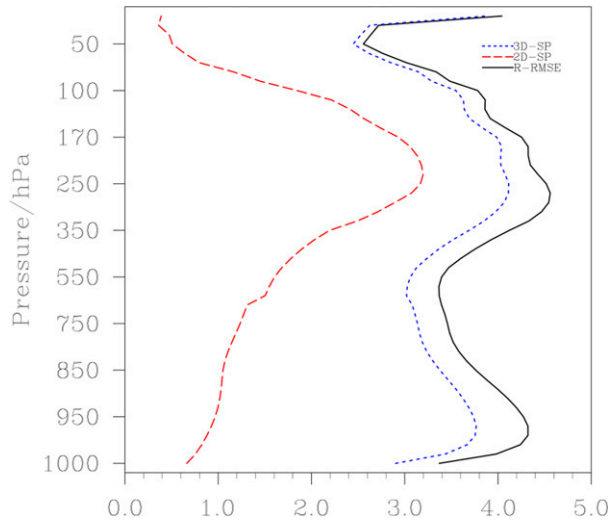


FIG. 3. The vertical profiles of IC perturbation derived from the 2D (red) and 3D (blue) rescaling masks as well as the analysis uncertainty (black) of the full wind speed ( $\text{m s}^{-1}$ ). It is averaged over 1–28 Jul 2018 (including both 0000 and 1200 UTC cycles). The vertical axis is the approximate isobaric level (hPa) corresponding to model level.

latitude–longitude grid with a horizontal resolution of  $0.03^\circ$  (about 3 km) and a vertical resolution of 51 levels. The model domain covers the following area:  $32^\circ\text{--}46.01^\circ\text{N}$  and  $108^\circ\text{--}123.99^\circ\text{E}$  ( $468 \times 534 = 249\,912$  grid points, Fig. 1) with complex terrain from mountain to sea.

The GRAPES-Meso-based regional EPS (GRAPES-REPS) consists of 15 members, including a control run and 14 perturbed ensemble members. The lateral boundary conditions (LBCs) are provided by a T639 (about 28 km) global EPS “T639-GEFS” which is also running operationally at CMA (Ma et al. 2008). In this study, the GRAPES-EPS runs two cycles per day, initiated at 0000 and 1200 UTC, respectively. The forecast length is 36 h. A blended method is used to create IC perturbations by combining the smaller-scale ETKF perturbations (Wang and Bishop 2003) from GRAPES-REPS and the larger-scale breeding perturbations (Toth and Kalnay 1997) from T639-GEFS. The details about this blended method are described in Zhang et al. (2015). The benefit of a blended perturbation has been well recognized in the scientific community by taking advantages of both smaller-scale and larger-scale information (e.g., Caron 2013; Du et al. 2015; Wang et al. 2014; Zhang et al. 2015). It has also proven to be effective in ameliorating the problem of mismatch in the LBCs (Caron 2013; Wang et al. 2014; Zhang et al. 2015). There are five state variables (zonal wind  $u$ , meridional wind  $v$ , potential temperature  $\theta$ , dimensionless pressure  $\pi$ , and specific humidity  $q$ ) in the GRAPES model. All their ICs are perturbed in hoping to have the fullest sampling in IC uncertainty space. A 2D rescaling mask (Fig. 2b) is currently used to adjust IC perturbation magnitude in the operational GRAPES-REPS, which is used as the control experiment to be compared to the new 3D scheme in this study.

Besides the perturbations to IC, physics perturbation is also included in the GRAPES-REPS in the following two ways. One uses multi-physics approach by varying PBL schemes and convective parameterization schemes (see the Table 1 of Xia et al. 2019 or Chen et al. 2020). Another is the stochastically perturbed parameterization tendency approach (SPPT, Buizza et al. 1999; Du et al. 2018) which is also described in detail in Xia et al. (2019). Both the control EPS and experiment EPS use the exact same configurations in raw IC perturbation-generating scheme and physics perturbation method except the rescaling mask.

### b. Three-dimensional (3D) rescaling mask

Rescaling factor is conceptually defined as a ratio of “analysis uncertainty” to “raw IC perturbation.” For the control EPS, a 2D rescaling mask is used, where the rescaling factor or coefficient is calculated within a representative midlayer ( $\sim 550\text{--}500$  hPa) and then applied to all vertical levels constantly, as schematically shown in Fig. 2b. In this study, we replace the 2D with a 3D rescaling mask where the rescaling factor  $\mu(k)$  varies with vertical level  $k$  as shown in Eq. (1) and schematically in Fig. 2c:

$$\mu(k) = \frac{\sqrt{[U_e(k)]^2 + [V_e(k)]^2}}{\sqrt{[\text{sp}_u(k)]^2 + [\text{sp}_v(k)]^2}}, \quad (1)$$

where  $U_e(k)$  and  $V_e(k)$  are domain-averaged *analysis uncertainty* of zonal wind  $u$  and meridional wind  $v$  at level  $k$ ; and  $\text{sp}_u(k)$  and  $\text{sp}_v(k)$  are domain-averaged raw IC perturbation size (*spread*) of  $u$  and  $v$  at level  $k$  at model initialization time before being rescaled. The analysis uncertainty is an intrinsic error in IC estimated by the GRAPES 3-DVAR data assimilation system (Ma et al. 2009). Intrinsic analysis error (“uncertainty”) is estimated by using the NMC method (Parrish and Derber 1992; Wu and Purser 2002), where the horizontal correlation length scale is estimated with a Gauss function linear filtering method (Zhuang et al. 2019). Such analysis uncertainty is estimated daily. An average of those analysis uncertainties over a past period (a year) can be used to represent typical uncertainty in the analysis that is used to initiate GRAPES\_Meso model. From the Eq. (1), we can see that the rescaling factor is the ratio of “analysis uncertainty” to “raw IC perturbation” (i.e., the ensemble spread at initialization time before rescaling) of wind speed or kinetic energy. Therefore, a new perturbation  $X'_{\text{new}}(i, j, k)$  for a variable  $X$  at level  $k$  at a grid point  $(i, j)$  can be obtained by multiplying the raw perturbation  $X'_{\text{raw}}(i, j, k)$  by the rescaling factor  $\mu(k)$  at each grid point  $(i, j)$ :

$$X'_{\text{new}}(i, j, k) = \mu(k) \times X'_{\text{raw}}(i, j, k). \quad (2)$$

Equations (1) and (2) suggest that when raw perturbation of wind speed is larger (smaller) than analyzed wind speed uncertainty, the rescaling factor is less (greater) than 1, suggesting that the raw perturbation will be scaled down (up). The raw perturbation will remain the same only when the rescaling factor is equal to 1 (no rescaling). Since all state variables are internally connected or coupled to each other through

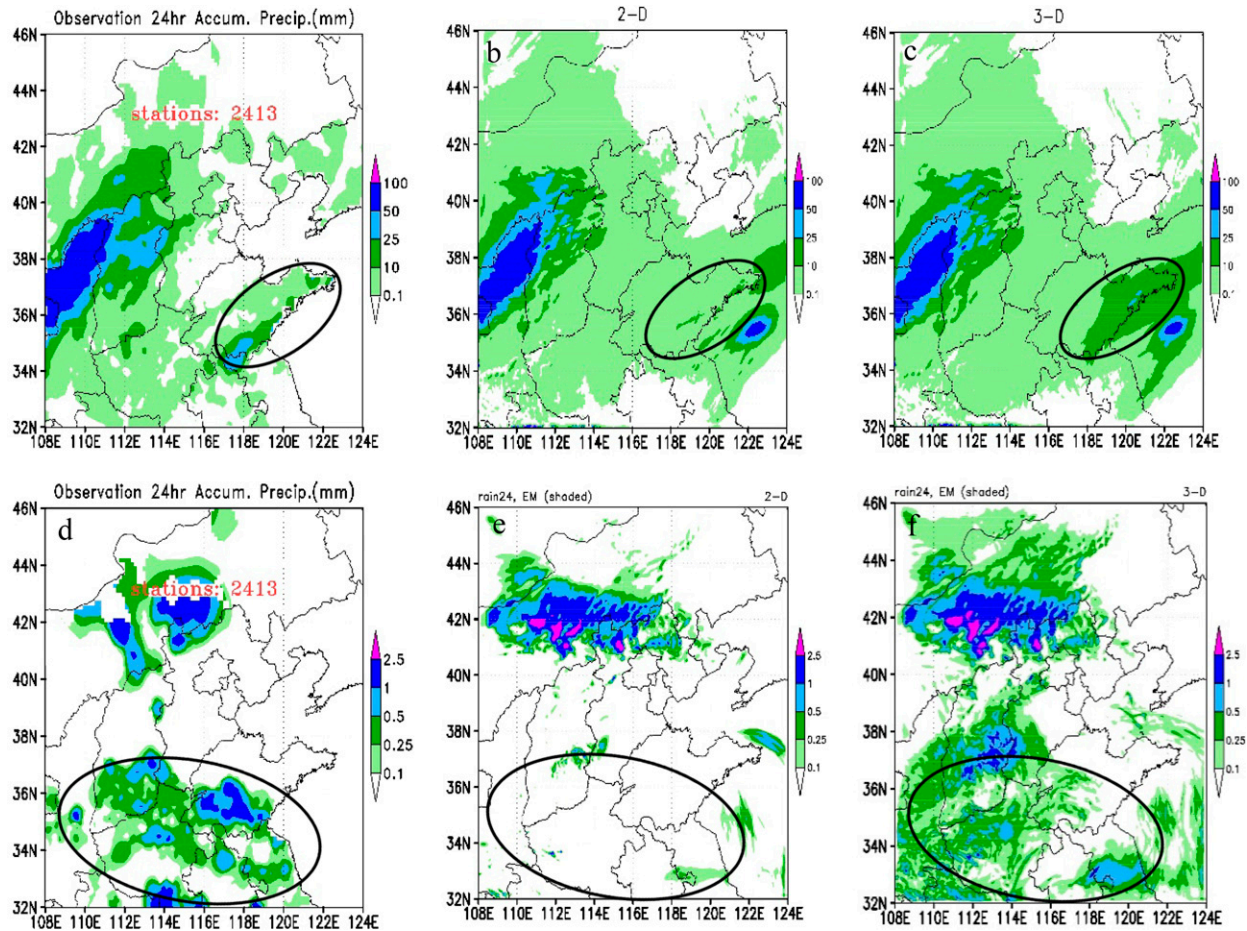


FIG. 4. The 24 h-accumulated precipitation (mm). (top) A warm-season heavy rain event (0000 UTC, 1000–0000 UTC 11 Jul 2008): (a) observation, (b) ensemble mean forecast of the control EPS (2D), and (c) ensemble mean forecast of the experiment EPS (3D). (bottom) A cold-season snow event (0000 UTC, 1000–0000 UTC 11 Feb 2019): (d) observation, (e) ensemble mean forecast of the control EPS (2D), and (f) ensemble mean forecast of the experiment EPS (3D).

dynamical equations, so are their perturbations. For example, if temperature perturbation is too small, perturbations in moisture and wind fields are likely too small too. Therefore, rescaling factor  $\mu(k)$  derived from one or two variables (normally wind and/or temperature) will be uniformly applied to all perturbed variables (wind  $u$  and  $v$ , temperature, moisture, and pressure) in EPS design. For example, in deriving rescaling factor in Eq. (1) were calculated based on gridpoint values at each grid point, the final resulting perturbation field would be the same as the time-averaged analysis uncertainty field, which could destroy the growing-mode structure imbedded in the raw perturbation created by a well-designed perturbation scheme (such as singular vector or bred vector). By the way, if one chooses a rescaling factor of a particular level (reference level  $l$ ),  $\mu = \mu(l)$ , and applies it to Eq. (2) for all levels  $k$ , the Eq. (2) is simplified to a 2D rescaling mask:

one value for each level (i.e., a same value for all grid points). By doing this, spatial structure of a raw perturbation field at each level will be reserved after being rescaled through Eq. (2). This is because the spatial structure of a perturbation field is important for spread growth (e.g., Li et al. 2009), which is the second principle mentioned in the introduction. If the rescaling factor in Eq. (1) were calculated based on gridpoint values at each grid point, the final resulting perturbation field would be the same as the time-averaged analysis uncertainty field, which could destroy the growing-mode structure imbedded in the raw perturbation created by a well-designed perturbation scheme (such as singular vector or bred vector). By the way, if one chooses a rescaling factor of a particular level (reference level  $l$ ),  $\mu = \mu(l)$ , and applies it to Eq. (2) for all levels  $k$ , the Eq. (2) is simplified to a 2D rescaling mask:

$$X'_{\text{new}}(i, j, k) = \mu \times X'_{\text{raw}}(i, j, k). \quad (3)$$

Note that the rescaling factor in Eq. (1) is calculated based on domain averaged values, which means that there is only

The difference between 2D and 3D mask is also illustrated in Fig. 2.

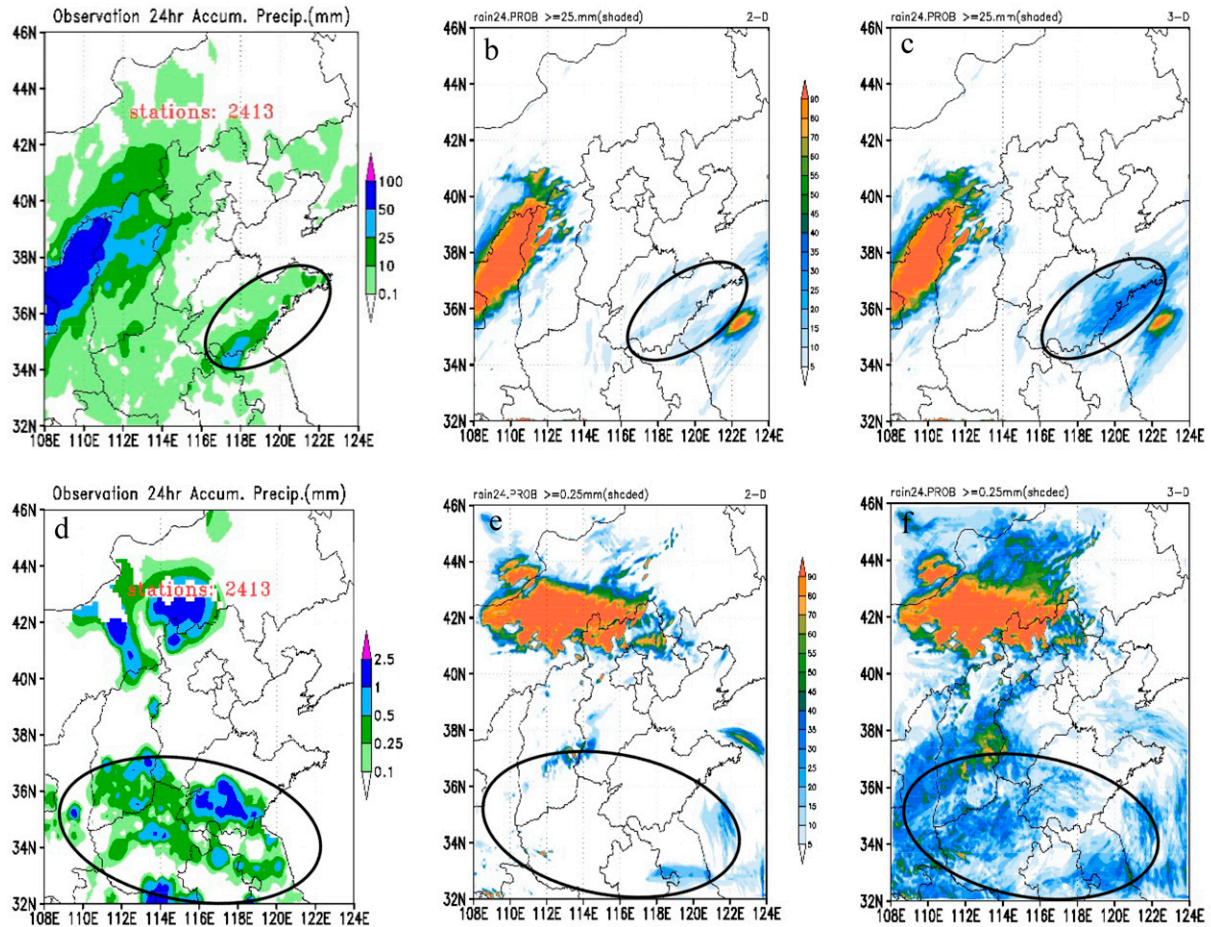


FIG. 5. The 24 h-accumulated precipitation (mm). (top) A warm-season heavy rain event (0000 UTC, 1000–0000 UTC 11 Jul 2008): (a) observation, (b) probability of exceeding 25 mm day<sup>-1</sup> predicted by the control EPS (2D), and (c) probability of exceeding 25 mm day<sup>-1</sup> predicted by the experiment EPS (3D). (bottom) A cold-season snow event (0000 UTC, 1000–0000 UTC 11 Feb 2019): (d) observation, (e) probability of exceeding 0.25 mm day<sup>-1</sup> liquid water equivalent predicted by the control EPS (2D), and (f) probability of exceeding 0.25 mm day<sup>-1</sup> liquid water equivalent by the experiment EPS (3D).

From the above discussion, we can see that the rescaling factor in this study is calculated differently from Wang et al. (2021) in the following two ways. 1) We use the model's own data assimilation system to estimate analysis error as the “analysis uncertainty” used in Eq. (1) in hoping to be more relevant to the IC of the model, while Wang et al. (2021) uses the long-term average of two independent analyses (GRAPES and ECMWF) as “analysis uncertainty” which reflects more of general statistical property of analysis error. 2) Our rescaling factor varies only vertically but not horizontally in hoping to preserve the horizontal structure of raw fast-growing perturbations at each level, while the rescaling factor of Wang et al. (2021) varies both vertically and horizontally and adjusts perturbation magnitude differently at each grid point, which will alter horizontal spatial structure of the raw IC perturbations.

### c. Cases and data

A summer month (28 days from 1 to 28 July 2018) is examined first for warm season, which is a meteorologically active

period with about 11% above normal precipitation (Zhang and Sun 2018). Since the GRAPES-REPS runs twice per day (0000 and 1200 UTC cycles) to 36 h, there are a total of  $28 \times 2 = 56$  36-h forecasts during this period. Therefore, the warm-season verification results presented in section 3 is the average of these 56 forecast cases over the model domain. Then, a weather-active winter month (1–27 February 2019 with a total of 54 36-h forecasts) is repeated for cold season. The warm-season results are compared to the cold-season results for a deeper understanding.

The 3-km GRAPES-Meso analysis is used as truth for verification of all variables except for precipitation. There are  $468 \times 534 = 249912$  data points can be used for a robust verification at each forecast hour. The analysis is produced by a 3DVAR data assimilation system (the method is described by Ma et al. 2009). Station observation, rawinsonde, aircraft data (ACARS), and satellite data are assimilated. No radar data are assimilated. For precipitation verification, auto rain gauge observation is used. There are a total of 2413 auto rain gauges within the model domain (Fig. 1).

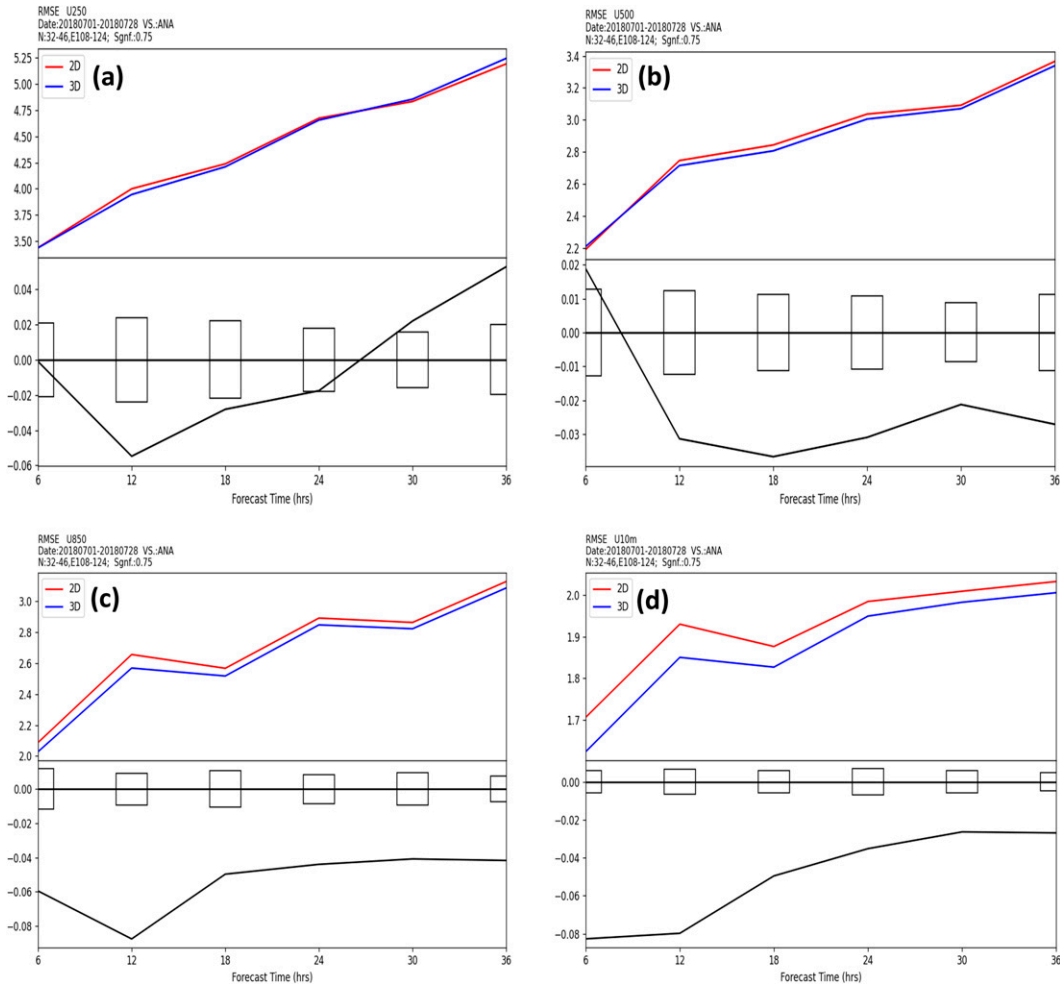


FIG. 6. The RMSE ( $\text{m s}^{-1}$ ) of the ensemble mean forecasts of  $U$  for (a) 250-hPa, (b) 500-hPa, (c) 850-hPa levels, and (d) 10 m above ground level. The horizontal axis is the forecast hour. The blue curve is for the 3D mask, and the red curve is for the 2D mask. It is averaged over 56 forecast cases (0000 and 1200 UTC cycles of 1–28 Jul 2018) and the model domain. The percentage of the improvement at 6, 12, 18, 24, 30, and 36 h: 0.08%, 1.44%, 0.68%, 0.26%,  $-0.5\%$ , and  $-0.85\%$  at 250 hPa (average = 0.2%) in (a);  $-0.93\%$ , 1.1%, 1.2%, 1.0%, 0.7%, and 0.8% at 500 hPa (average = 0.5%) in (b); 2.88%, 3.44%, 1.95%, 1.55%, 1.4%, and 1.4% at 850 hPa (average = 2.1%) in (c); and 4.89%, 4.41%, 2.68%, 1.89%, 1.38%, and 1.44% at 10 m (average = 2.8%) in (d). The statistical significance  $t$ -test results are shown in the bottom panel of each plot (also see Fig. 16), where the box indicates the range of 75% significance level and the curve is RMSE difference (3D – 2D). If the curve is below (above) the box, the 3D mask forecast is significantly improved (degraded) over the 2D mask forecast; if the curve is within the box, the change is neutral (no significant improvement or degradation).

### 3. Results

#### a. Vertical structure of IC perturbation

Figure 3 shows the vertical profiles of the wind perturbation from the 2D (red dash line) and 3D (blue dash line) masks, compared to the analysis uncertainty (black solid line). The perturbations are averaged over the model domain and the experiment period (1–28 July 2018). It shows that the perturbation with the 2D rescaling mask is too small in the entire atmosphere and has incorrect vertical distribution below 700-hPa level, while the perturbation with the 3D rescaling mask is much closer to the analysis uncertainty with correct vertical distribution in the

entire atmosphere. In other words, the IC perturbation can better represent the real analysis uncertainty through the 3D rescaling mask, which is a desired feature (the first principle of “representativeness”) for a good IC perturbation design. Similar result is for the cold season (not shown).

#### b. Forecast improvements

In this section, all aspects of ensemble forecasts (ensemble mean, spread and probability) will be thoroughly examined to see if they can be improved after using a more realistic IC perturbation in vertical distribution through the 3D rescaling mask. Besides showing two cases, seven scoring rules are used

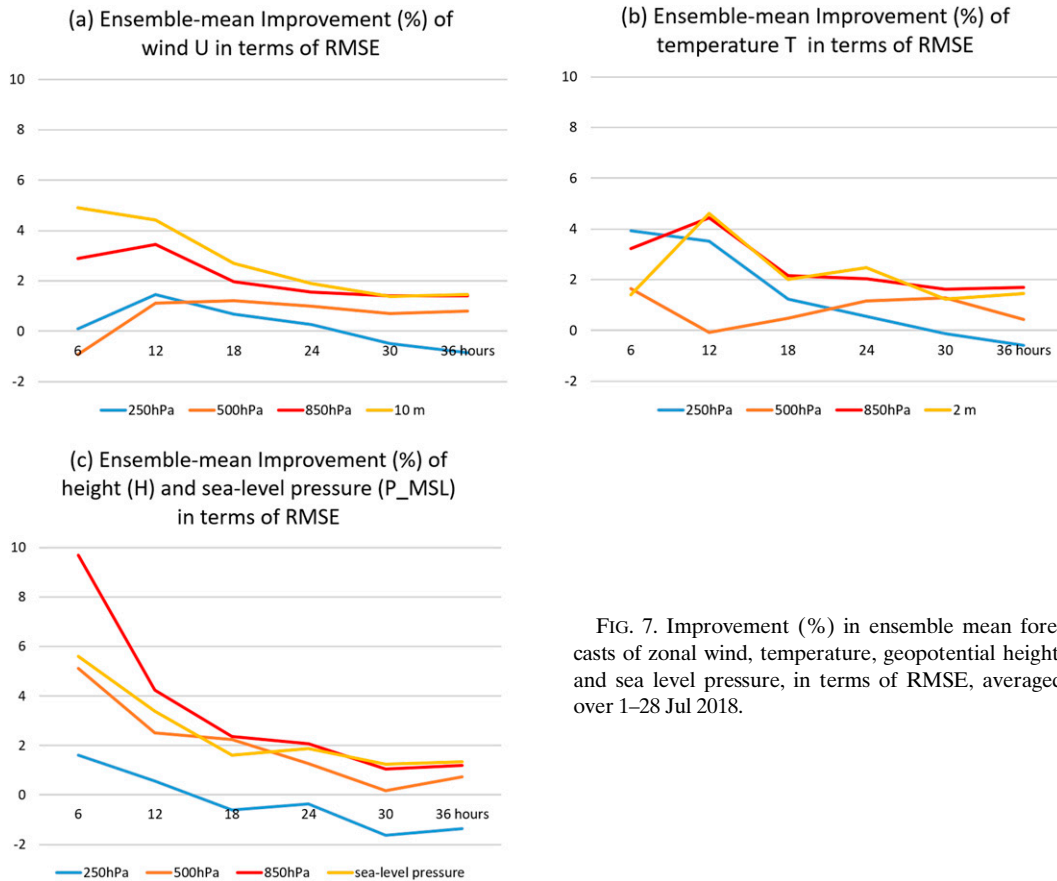


FIG. 7. Improvement (%) in ensemble mean forecasts of zonal wind, temperature, geopotential height, and sea level pressure, in terms of RMSE, averaged over 1–28 Jul 2018.

to measure different aspects of an ensemble performance: one (RMSE) for ensemble mean, three (spread, outlier, and consistency) for ensemble spread, and three (CRPS, BS, and AROC) for probabilistic forecasts. These scores will be briefly described when they are used hereafter. A review of ensemble verification scores can be found in Jolliffe and Stephenson (2003) as well as Du and Zhou (2017). Verification will be carried out at three representative levels: upper (250 hPa), middle (500 hPa), and lower (850 hPa) levels for wind ( $U$  and  $V$ ), temperature ( $T$ ), and geopotential height ( $H$ ). The four surface weather elements were also verified, which includes 2-m temperature  $T$ , 10-m wind ( $U$  and  $V$ ), precipitation, and mean sea level pressure ( $P_{MSL}$ ). Individual results are presented in sections 3b(1)–3b(4), while a summary score card of all results is presented in section 3b(5). Since there are so many scores being used to verify each of those variables for two seasons (warm and cold season), a large number of figures are produced. It is not feasible to show all of them in the article. Therefore, only representative variables of warm season will be demonstrated as examples in individual verification results, but all variables will be shown in summary figures. To show as many variables as possible, different variables are demonstrated for different aspects (mean, spread, and probability) of ensemble performance. In section 3b(5), the cold-season results will be compared to the warm-season results. The scores are robustly calculated over the entire model domain using 249 912 ( $468 \times 534$ ) data points at each forecast hour. The

verification results shown are the average of either all summer cases or all winter cases unless specified otherwise.

#### 1) CASE SHOW

Figure 4 shows the ensemble mean forecasts of 24-h accumulated precipitation for two cases: a warm-season case (the top panel) and a cold-season case (the bottom panel). For the warm-season case (initiated from 0000 UTC 10 July 2018), the forecast of 2D rescaling mask (Fig. 4b) obviously missed the observed northeast–southwest-oriented heavier precipitation band ( $\geq 10$  mm with some areas of exceeding 25 mm) along the coast of Shandong peninsula (the highlighted area of Fig. 4a), while the forecast of 3D rescaling mask provided the information of this event (Fig. 4c). Similarly, for the cold-season case (initiated from 0000 UTC 10 February 2019), the forecast of 2D rescaling mask (Fig. 4e) completely missed a large area of observed snow event (with liquid water equivalent of 0.1–2.5 mm which is a high-impact event given lower-latitude and larger areal coverage) spreading from the southern part of North China to Jiangsu Province (the highlighted south portion of the domain) (Fig. 4d), while the forecast of 3D rescaling mask predicted it (Fig. 4f). The enhanced information in the ensemble mean implies that the observed events were correctly captured by more members in the 3D ensemble than in the 2D ensemble if it is not completely missed by the 2D EPS. To demonstrate this, the



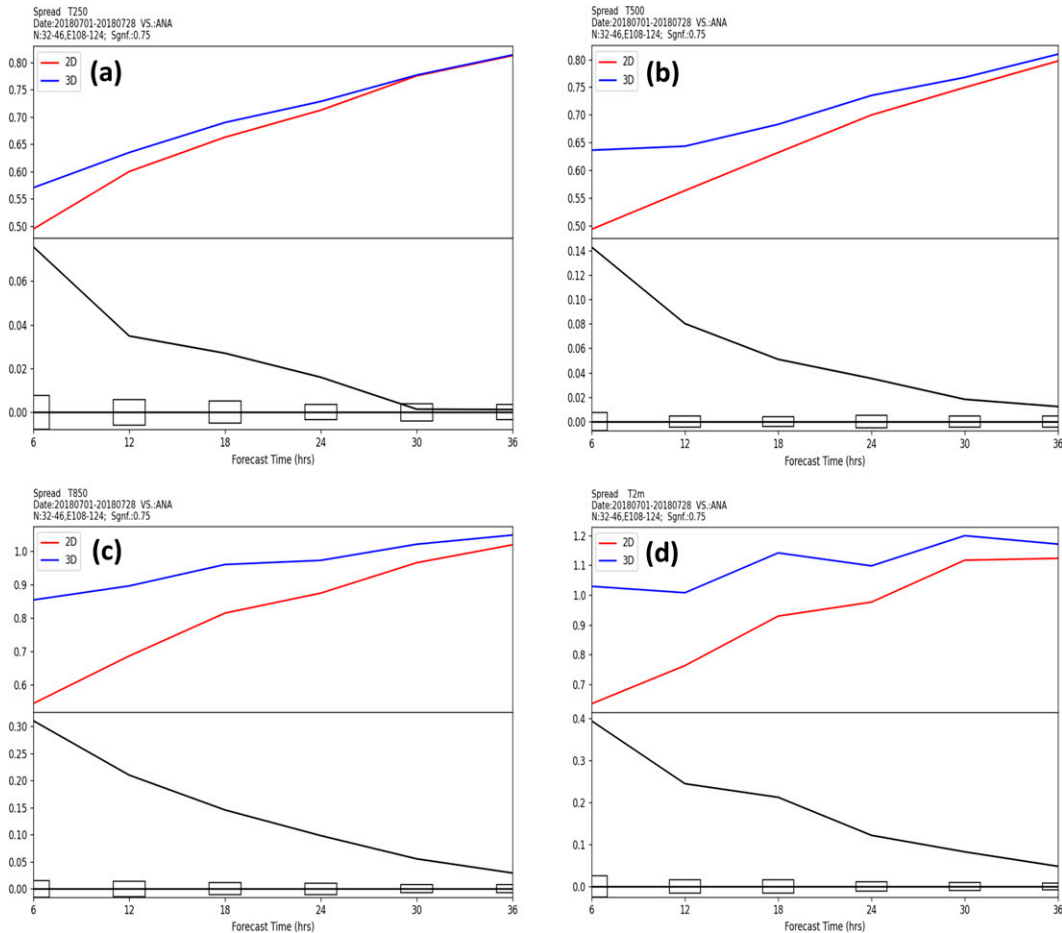


FIG. 8. The ensemble spread (K) of temperature for (a) 250-hPa, (b) 500-hPa, (c) 850-hPa levels, and (d) 2 m above ground level. The horizontal axis is forecast hour. The blue curve is for the 3D mask, and the red curve is for the 2D mask. It is averaged over 56 forecast cases (0000 and 1200 UTC cycles of 1–28 Jul 2018) and the model domain. The percentage of the improvement at 6, 12, 18, 24, 30 and 36 h: 16.56%, 6.46%, 4.19%, 1.93%, 0.16%, and 0.23% at 250 hPa (average = 4.9%) in (a); 29.1%, 14.31%, 8.08%, 5.14%, 2.5%, and 1.65% at 500 hPa (average = 10.1%) in (b); 60.47%, 32.13%, 18.89%, 11.9%, 6.29%, and 3.32% at 850 hPa (average = 22.2%) in (c); and 64.75%, 32.48%, 22.89%, 12.71%, 7.98%, and 4.52% at 2 m (average = 24.2%) in (d). The statistical significance *t*-test results are shown in the bottom panel of each plot (also see Fig. 16), where the box indicates the range of 75% significance level and the curve is spread difference (3D – 2D). If the curve is above (below) the box, the 3D mask forecast has significantly larger (smaller) spread than the 2D mask forecast; if the curve is within the box, the change is neutral (no significant increase or decrease).

corresponding probabilistic forecasts are shown in Fig. 5. Over the highlighted observed 25-mm area of the summer case (Fig. 5a), the probability of exceeding 25 mm in the 3D ensemble is about 10%–15% (Fig. 5c), while it is 0% in the 2D ensemble (Fig. 5b). For the highlighted winter snow case (Fig. 5d), the probability of exceeding 0.25-mm liquid water equivalent in the 3D ensemble is generally 30%–35% with some over 40% (Fig. 5f), while it is largely 0% in the 2D ensemble (Fig. 5e). This useful probability information contained in the 3D ensemble will certainly give forecasters a heads up of these incoming events.

Admittedly, a success comes not free but with an expense. Due to the imperfect model physics and ICs, a larger diversity in an ensemble will also lead to more false alarm cases

although some ensemble members might contain more accurate forecast information. This adverse side effect is typical for an EPS. For example, some precipitation areas in the 3D ensemble forecasts were also falsely expanded in both cases, which is especially obvious in the winter case (e.g., Figs. 4e,f). There is always a trade-off between increasing ensemble diversity and false alarm rate in building an EPS. As a matter of fact, the primary mission of an EPS is not to provide an accurate deterministic forecast but estimate reliable confidence or uncertainty information associated with a forecast (Du et al. 2018). Individual cases give us only a snapshot but not the overall performance of an ensemble. That is why it needs many cases but not just one case in evaluating an EPS. Many cases will be statistically evaluated for ensemble mean, spread

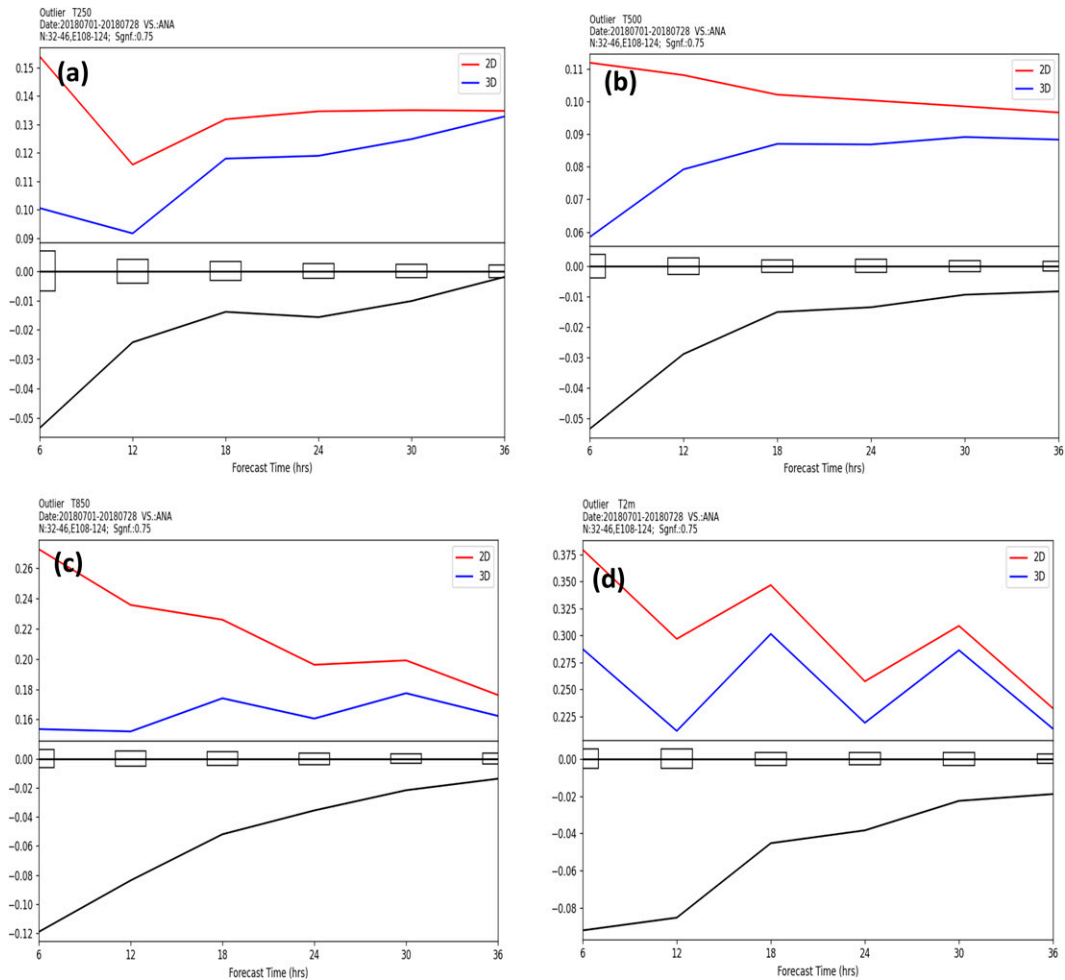


FIG. 9. The outlier score (%) of ensemble spreads of temperature for (a) 250-hPa, (b) 500-hPa, (c) 850-hPa levels, and (d) 2 m above ground level. The horizontal axis is forecast hour. The blue curve is for the 3D mask, and the red curve is for the 2D mask. It is averaged over 56 forecast cases (0000 and 1200 UTC cycles of 1–28 Jul 2018) and the model domain. The percentage of the improvement at 6, 12, 18, 24, 30, and 36 h: 34.7%, 20.97%, 10.47%, 11.59%, 7.56%, and 1.48% at 250 hPa (average = 14.5%) in (a); 47.76%, 26.67%, 14.71%, 13.46%, 9.54%, and 8.59% at 500 hPa (average = 20.1%) in (b); 43.63%, 35.51%, 22.97%, 18.15%, 10.9%, and 7.84% at 850 hPa (average = 23.2%) in (c); and 24.26%, 28.7%, 13.04%, 14.91%, 7.31%, and 8.13% at 2 m (average = 16.1%) in (d). The statistical significance *t*-test results are shown in the bottom panel of each plot (also see Fig. 16), where the box indicates the range of 75% significance level and the curve is the outlier difference (3D – 2D). If the curve is below (above) the box, the outlier of 3D mask forecast is significantly reduced (increased) over that of 2D mask forecast; if the curve is within the box, the change is neutral (no significant reduction or increase).

and probabilistic forecasts hereafter [sections 3b(2)–3b(5)]. For an under-dispersive EPS (which is the case in this study), increasing diversity among ensemble members is generally beneficial [see section 3b(3) about ensemble spread evaluation]. By the way, an ensemble mean forecast will theoretically overestimate areal coverage for lighter precipitation and underestimate it for heavier precipitation during arithmetic averaging for a non-Gaussian distributed variable like precipitation, which was discussed by Du et al. (1997). Therefore, probabilistic information is more preferred than ensemble mean in predicting precipitation.

## 2) ENSEMBLE MEAN FORECASTS

Below we will use root-mean-squared error (RMSE) to quantitatively measure the ensemble mean performance of more variables. Figure 6 is the RMSE of zonal wind *U* at four levels varying with forecast hours. The 3D mask run (blue curve) has less error than the 2D mask run (red curve) at all levels and all forecast hours (except for the 250 hPa at 30 and 36 h). The improvement is larger at surface and lower level than upper level. For example, the averaged improvement of 6–36-h *U* forecasts is about 2.8%, 2.1%, 0.5%, and 0.2% at 10 m above ground level and 850, 500, and 250 hPa,

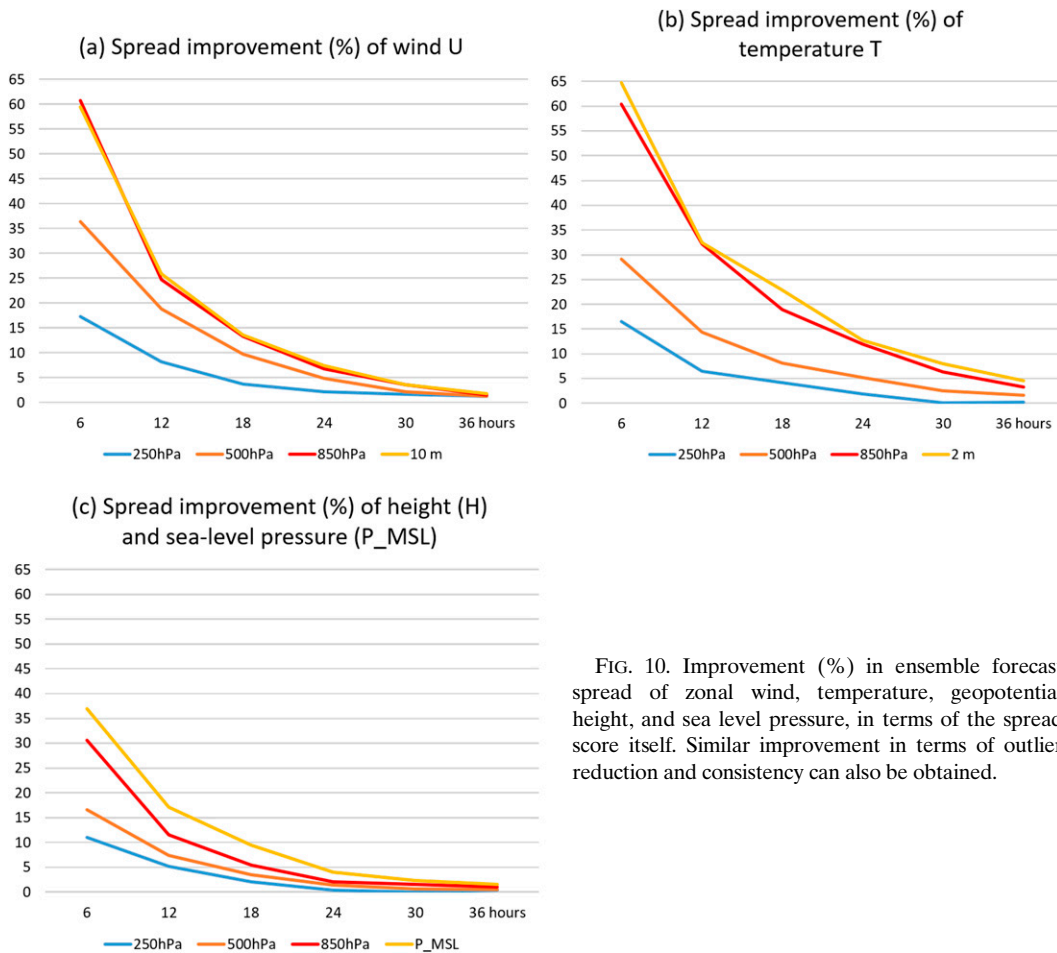


FIG. 10. Improvement (%) in ensemble forecast spread of zonal wind, temperature, geopotential height, and sea level pressure, in terms of the spread score itself. Similar improvement in terms of outlier reduction and consistency can also be obtained.

respectively. The improvement decreases with the increase of forecast length too (see the percentage values of the improvement listed in the figure caption). Results of temperature and geopotential height/pressure are similar to that of wind: the 3D mask runs have generally less error than the 2D mask runs at all levels and all forecast hours. Decreasing of the improvement with the increase of altitude and forecast hour are also observed. For example, the averaged improvement of 6–36-h temperature forecasts is about 2.2%, 2.5%, 0.8%, and 1.4% at 2 m and 850, 500, and 250 hPa, respectively. The averaged improvement of 6–36-h geopotential height or pressure forecasts is about 2.5%, 3.4%, 2.0%, and 0.3% at sea level, 850, 500, and 250 hPa, respectively. All the improvements of ensemble mean forecasts are summarized in Fig. 7, where the relative improvement among levels and the improvement’s decreasing trend with forecast hours can be clearly seen.

By the way, although the improvement in ensemble mean forecasts is noticeable and statistically significant, it is the least compared to the improvements in ensemble spread and probabilistic forecasts (to be discussed in the next two subsections).

### 3) ENSEMBLE SPREAD

Three scores are used to measure the quality of ensemble spread. One is the ensemble spread itself, which is defined as a standard deviation of ensemble members’ forecasts with respect to ensemble mean. For an underdispersive EPS such as this one (the GRAPES EPS used in this study is under-dispersive), ensemble spread is not large enough to match ensemble mean forecast error (not shown). Therefore, increasing ensemble spread is a positive improvement. Figure 8 shows the ensemble spread of temperature at the four levels varying with forecast hours. The ensemble spread is greatly enhanced in the 3D mask run at all levels and forecast hours. These improvements are statistically significant (except for 30 and 36 h at 250 mb). Like the improvement in ensemble mean, the following two features are even more obvious for ensemble spread: the improvement decreases with the increase of forecast length and the increase of altitude. The averaged improvement of 6–36-h ensemble spread is about 24.2%, 22.2%, 10.1%, and 4.9% for 2-m and 850-, 500-, and 250-hPa temperature, respectively. Similar results are observed for the zonal wind *U* and geopotential height/pressure forecasts. The averaged improvement of 6–36-h ensemble spread is about 18.6%, 18.4%, 12.2%, and 5.7% for

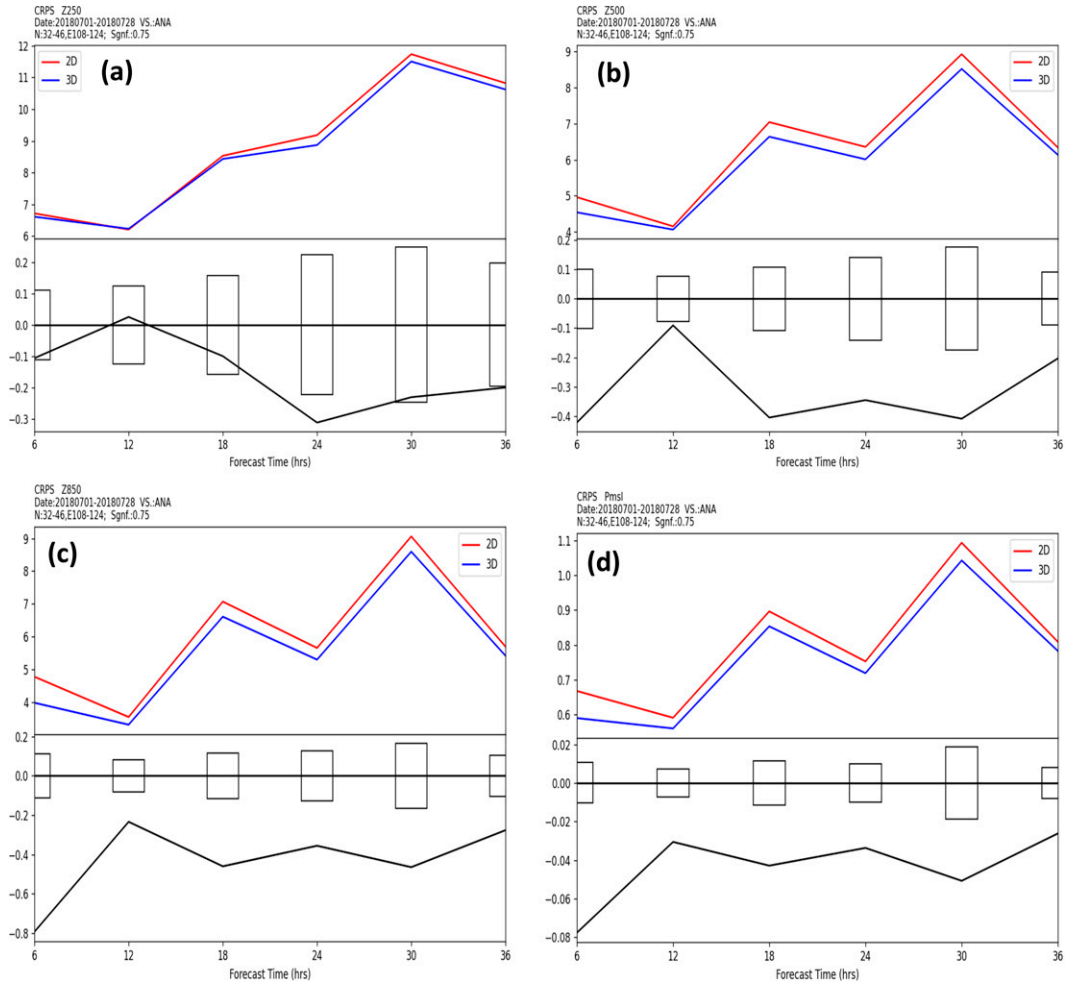


FIG. 11. The CRPS of probabilistic forecasts of geopotential height  $H$  for (a) 250-hPa, (b) 500-hPa, and (c) 850-hPa levels, as well as (d) mean sea level pressure  $P_{MSL}$ . The horizontal axis is forecast hour. The blue curve is for the 3D mask, and the red curve is for the 2D mask. It is averaged over 56 forecast cases (0000 and 1200 UTC cycles of 1–28 Jul 2018) and the model domain. It is averaged over the period of 1–28 Jul 2018 (including both 0000 and 1200 UTC cycles). The percentage of the improvement at 6, 12, 18, 24, 30, and 36 h: 1.58%, 0.41%, 1.17%, 3.39%, 1.97%, and 1.85% at 250 hPa (average = 1.7%) in (a); 8.51%, 2.19%, 5.75%, 5.44%, 4.58%, and 3.21% at 500 hPa (average = 5.0%) in (b); 16.59%, 6.57%, 6.51%, 6.28%, 5.13%, and 4.85% at 850 hPa (average = 7.7%) in (c); and 11.63%, 5.18%, 4.79%, 4.48%, 4.65%, and 3.25% at sea level (average = 6.0%) in (d). The statistical significance  $t$ -test results are shown in the bottom panel of each plot (also see Fig. 16), where the box indicates the range of 75% significance level and the curve is CRPS difference (3D – 2D). If the curve is below (above) the box, the 3D mask forecast is significantly improved (degraded) over the 2D mask forecast; if the curve is within the box, the change is neutral (no significant improvement or degradation).

10-m and 850-, 500-, and 250-hPa wind  $U$ , respectively. The averaged improvement of 6–36-h ensemble spread is about 11.9%, 8.7%, 5.0%, and 3.2% for mean sea level pressure and 850-, 500-, and 250-hPa geopotential height, respectively. Among the three variables, the least improvement in ensemble spread is in the mass field (3.2%–11.9%) compared to the wind (5.7%–18.6%) and temperature fields (4.9%–24.2%).

The second score is the outlier that counts how often (in %) an observation falls out of an ensemble envelope (i.e., the ensemble forecast range from minimum member to maximum member). It is to measure the capability of an ensemble to

encompass an observation. The outlier is derived from the rank histogram (or Talagrand distribution) by adding the two far-end bins together. Forecasters and users at the CMA and its field offices normally prefer the outlier score more than the histogram because the outlier directly shows the missing rate of an EPS forecast. Since the outlier is a missing rate, the smaller the better. As the spread increases (with no worsening in ensemble mean position at the same time), the outlier has also significantly reduced. For example, the average reduction of 6–36-h outlier is about 16.1%, 23.2%, 20.1%, and 14.5% for 2-m and 850-, 500-, and 250-hPa temperature, respectively

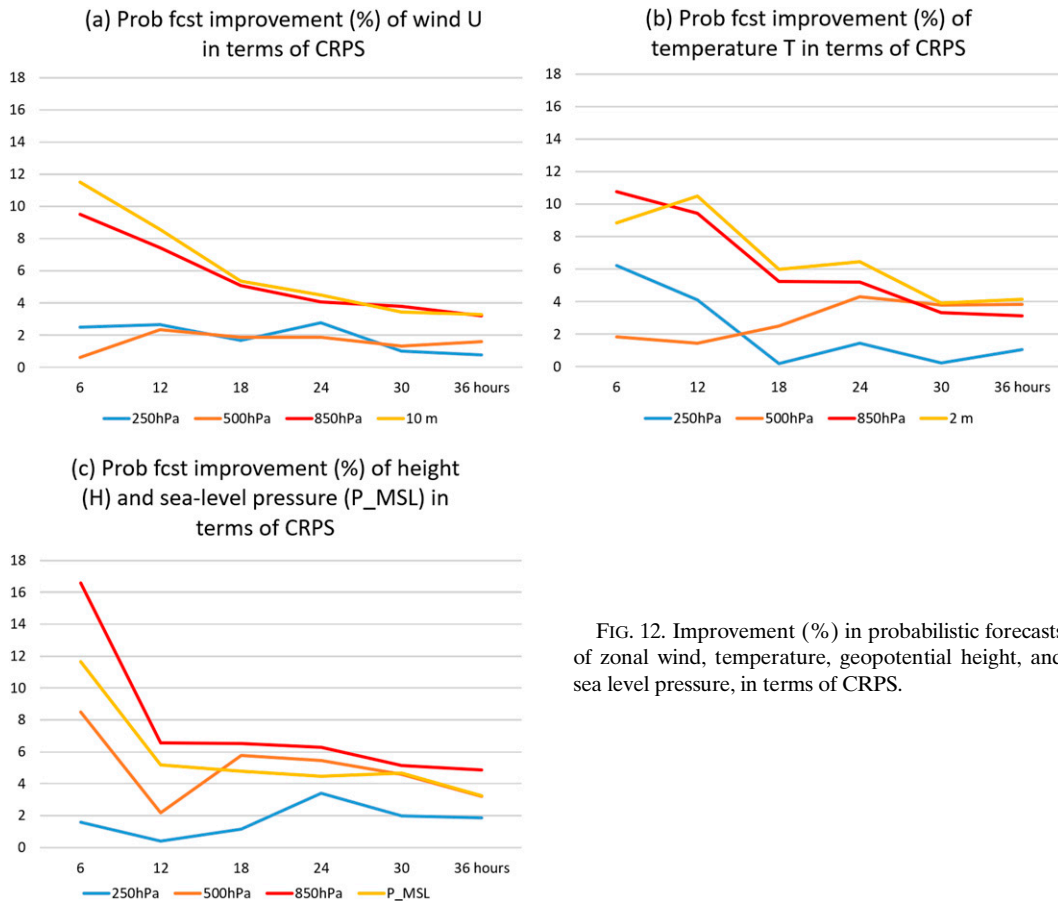


FIG. 12. Improvement (%) in probabilistic forecasts of zonal wind, temperature, geopotential height, and sea level pressure, in terms of CRPS.

(Fig. 9). The average reduction of 6–36-h outlier is about 18.7%, 24.3%, 20.7%, and 12.1% for 10-m and 850-, 500-, and 250-hPa wind  $U$ , respectively. The average reduction of 6–36-h outlier is about 14.4%, 15.0%, 12.3%, and 9.0% for mean sea level pressure and 850-, 500-, and 250-hPa geopotential height, respectively. Again, the improvement in outlier also decreases with the increase of forecast length and the increase of altitude. More improvement is in temperature (14.5%–23.2%) and wind (12.1%–24.3%) than in geopotential height and pressure (9.0%–15.0%).

Since outliers could be improved at the expense of overdispersion in ensemble spread, the third score “consistency” is also evaluated to monitor the spread–skill relation. If both outlier and consistency are improved, it is a desired healthy improvement in ensemble spread. “consistency” is a score to measure spread–skill relationship, which is normally defined as the ratio of ensemble mean forecast’s RMSE to ensemble spread. For an ideal ensemble, consistency score is close to 1.0 since ensemble spread simulates possible error in ensemble mean forecast. If consistency is greater (less) than 1.0, an ensemble is underdispersive (overdispersive), indicating ensemble spread is smaller (greater) than ensemble mean forecast error. In this study, an alternative version of consistency is used as Eq. (4):

$$\text{consistency} = 1 - \frac{\text{RMSE of ens mean fcst}}{\text{ensemble spread}}. \quad (4)$$

With this new definition, consistency is close to 0.0 for a perfect ensemble (spread = RMSE); consistency < 0 for an underdispersive ensemble (spread < RMSE); and consistency > 0 for an overdispersive ensemble (spread > RMSE). The results in terms of consistency score are included in the summary score card (later Fig. 16), which shows similar significant (at 95%–99.7% level) improvements as the spread and outlier. In other words, the increase of ensemble spread is generally in the right direction and does not cause overdispersion (as a matter of fact, the ensemble spread is still too small for lower level and surface variables). Figure 10 summarizes all the improvements of ensemble forecast spread, where we can clearly see the relative improvement among vertical levels and variables as well as the improvement’s decreasing trend with forecast length as we have discussed above.

#### 4) PROBABILISTIC FORECASTS

The continuous ranked probability score (CRPS) is often used to evaluate performance of probabilistic forecasts. Analogous to the mean squared error for deterministic forecasts, CRPS is a mean squared difference between predicted cumulative probability density function (CDF) and observed CDF (either 0 or 1) over continuous mutually exclusive and collectively exhaustive categories (see the appendix of Du et al. 1997). It is a negatively oriented score, i.e., the smaller the

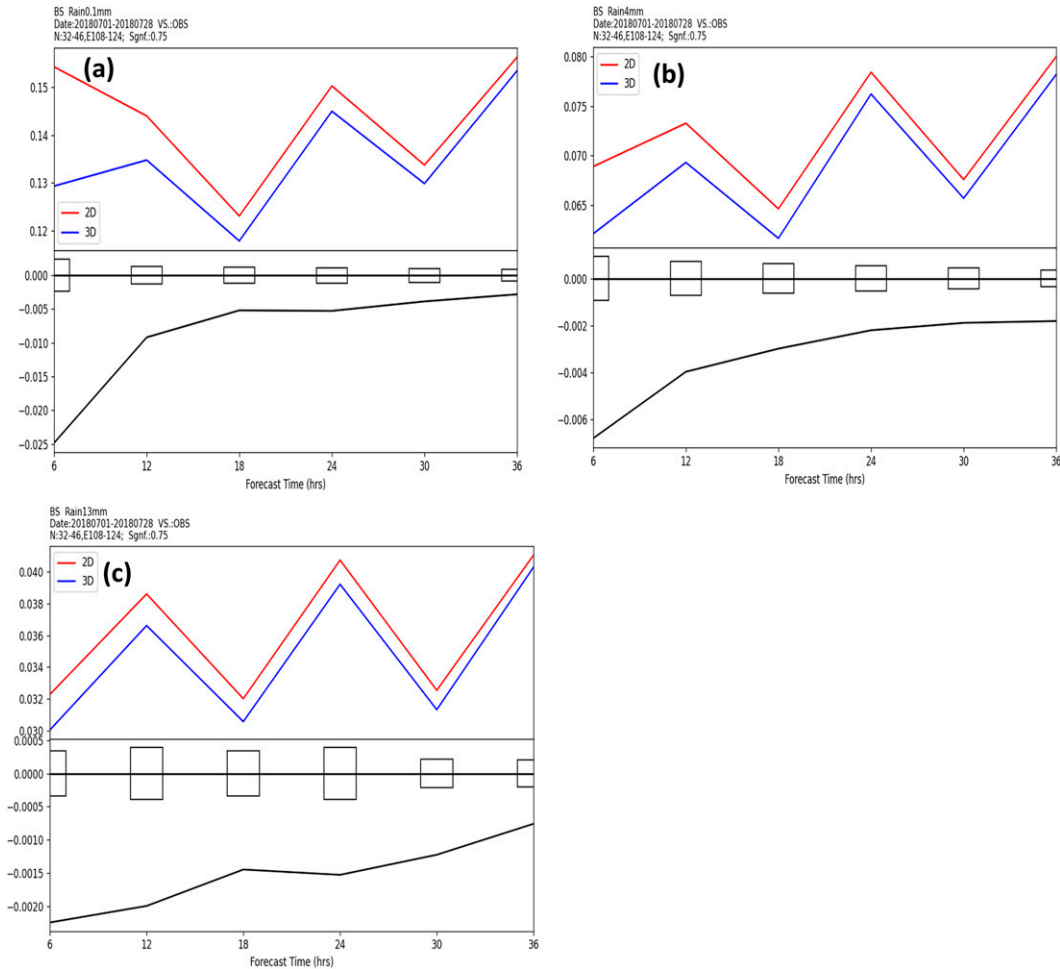


FIG. 13. BS of probabilistic precipitation forecasts for (a) light [ $>0.1 \text{ mm (6 h)}^{-1}$ ], (b) moderate [ $>4 \text{ mm (6 h)}^{-1}$ ], and (c) heavy [ $>13 \text{ mm (6 h)}^{-1}$ ] precipitation events. The horizontal axis is forecast hour. The blue curve is for the 3D mask, and the red curve is for the 2D mask. It is averaged over 56 forecast cases (0000 and 1200 UTC cycles of 1–28 Jul 2018) and the model domain. The percentage of the improvement at 6, 12, 18, 24, 30 and 36 h: 15.23%, 6.05%, 3.72%, 2.79%, 2.24%, and 1.76% for light rain (average = 5.3%) in (a); 8.6%, 4.75%, 3.52%, 2.24%, 2.96%, and 2.3% for moderate rain (average = 4.1%) in (b); and 6.15%, 5.35%, 3.88%, 3.02%, 4.09%, and 1.49% for heavy rain (average = 4.0%) in (c). The statistical significance *t*-test results are shown in the bottom panel of each plot (also see Fig. 16), where the box indicates the range of 75% significance level and the curve is BS difference (3D – 2D). If the curve is below (above) the box, the 3D mask forecast is significantly improved (degraded) over the 2D mask forecast; if the curve is within the box, the change is neutral (no significant improvement or degradation).

better, with zero as a perfect score. CRPS is impacted by both reliability and resolution of a forecast (Jolliffe and Stephenson 2003). Figure 11 is the CRPS of geopotential height at various levels as well as sea level pressure varying with forecast hours.

The 3D mask run has significantly less error (lower CRPS) over the 2D mask run at all levels and all forecast hours. More improvement is at lower levels (850 hPa and sea level) than the higher levels (250 and 500 hPa), which is more obvious in

TABLE 1. Days of observed precipitation categories during the verification period.

1–28 Jul 2018 (30–36-h forecasts cover 28.5 days into 0000–1200 29 Jul)	0000–0600 UTC	0600–1200 UTC	1200–1800 UTC	1800–2400 UTC
Light rain [ $0.1\text{--}4.0 \text{ mm (6 h)}^{-1}$ ]	29	29	28	28
Moderate rain [ $4.0\text{--}13.0 \text{ mm (6 h)}^{-1}$ ]	29	29	28	28
Heavy rain [ $\geq 13.0 \text{ mm (6 h)}^{-1}$ ]	26 (3 days w/o heavy rain)	29	26 (2 days w/o heavy rain)	27 (2 days w/o heavy rain)

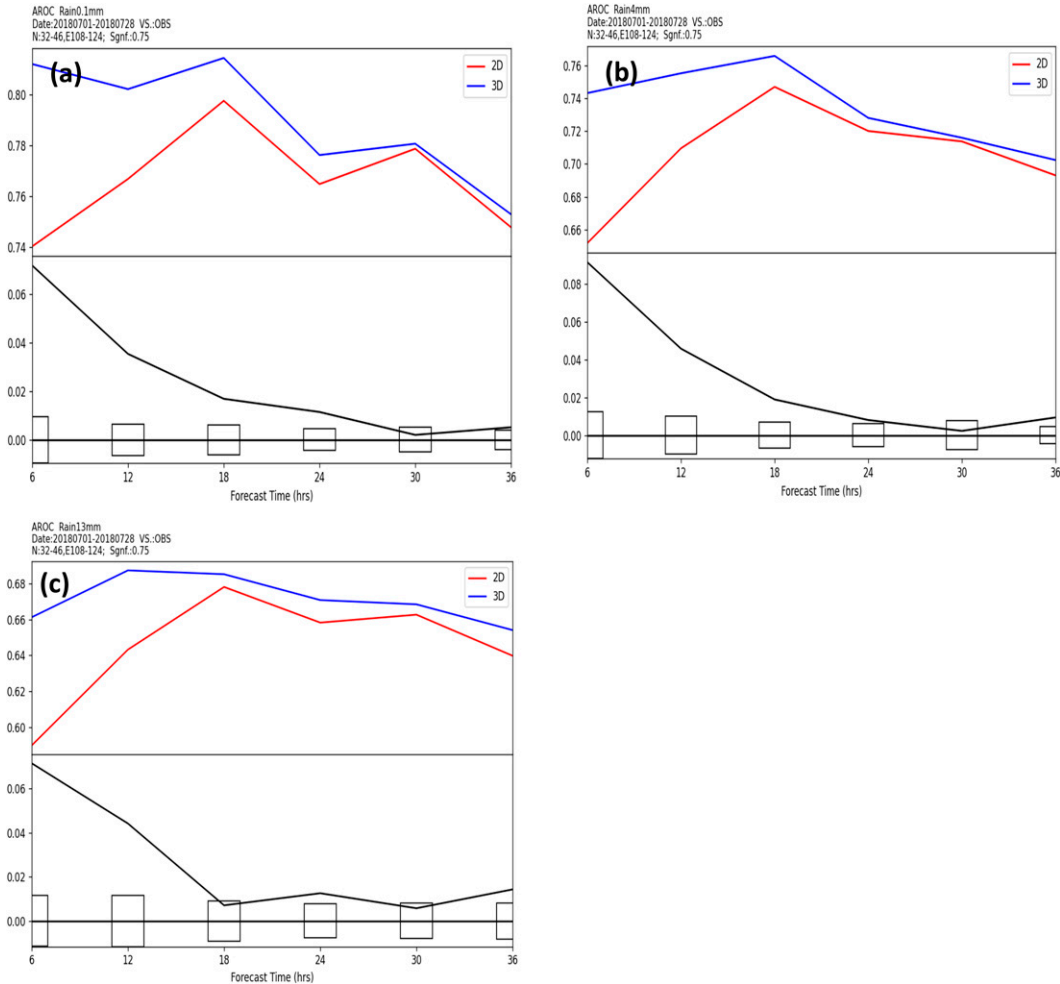


FIG. 14. AROC of probabilistic precipitation forecasts for (a) light [ $>0.1 \text{ mm (6 h)}^{-1}$ ], (b) moderate [ $>4 \text{ mm (6 h)}^{-1}$ ], and (c) heavy [ $>13 \text{ mm (6 h)}^{-1}$ ] precipitation events. The horizontal axis is forecast hour. The blue curve is for the 3D mask, and the red curve is for the 2D mask. It is averaged over 56 forecast cases (0000 and 1200 UTC cycles of 1–28 Jul 2018) and the model domain. The percentage of the improvement at 6, 12, 18, 24, 30 and 36 h: 10.05%, 4.73%, 2.11%, 1.5%, 0.1%, and 0.58% for light rain (average = 3.2%) in (a); 14.75%, 6.87%, 2.72%, 1.07%, 0.27%, and 1.28% for moderate rain (average = 4.5%) in (b); and 12.16%, 6.91%, 0.87%, 1.77%, 0.8%, and 1.97% for heavy rain (average = 4.1%) in (c). The statistical significance *t*-test results are shown in the bottom panel of each plot (also see Fig. 16), where the box indicates the range of 75% significance level and the curve is AROC difference (3D – 2D). If the curve is above (below) the box, the 3D mask forecast is significantly improved (degraded) over the 2D mask forecast; if the curve is within the box, the change is neutral (no significant improvement or degradation).

wind and temperature fields (Fig. 12). The average improvement of 6–36 h CRPS score is about 6.0%, 7.7%, 5.0%, and 1.7% for P\_MSL and 850-, 500-, and 250-hPa geopotential height, respectively. This improvement decreases quickly from 6 to 12 or 18 h and remains generally at a same level for the rest of forecast hours (can be seen more clearly from Fig. 12). Similar results were observed for wind and temperature. The average improvement of 6–36 h CRPS score is about 6.1%, 5.5%, 1.6%, and 1.9% for 10-m and 850-, 500-, and 250-hPa wind *U*, respectively. The average improvement of 6–36-h CRPS score is about 6.6%, 6.2%, 3.0%, and 2.2% for 2-m and 850-, 500-, and 250-hPa temperature, respectively. Figure 12

summarizes all the improvements in probabilistic forecasts, where we can see the relative improvement among levels and the improvement’s decreasing trend with forecast hours clearer.

When the number of forecast categories is reduced to two (e.g., rain or no rain), the CRPS becomes the Brier score (BS). Figure 13 is the BS of probabilistic precipitation forecasts for light [ $\geq 0.1 \text{ mm (6 h)}^{-1}$ ], moderate [ $\geq 4 \text{ mm (6 h)}^{-1}$ ], and heavy rain [ $\geq 13 \text{ mm (6 h)}^{-1}$ ]. The error has been significantly reduced (lower BS) especially prior to 18 h. The improvement quickly decreases with the increase of forecast hours prior to 24 h, and then remains similarly for the rest of forecast hours. On average over 6–36-h forecast range, the

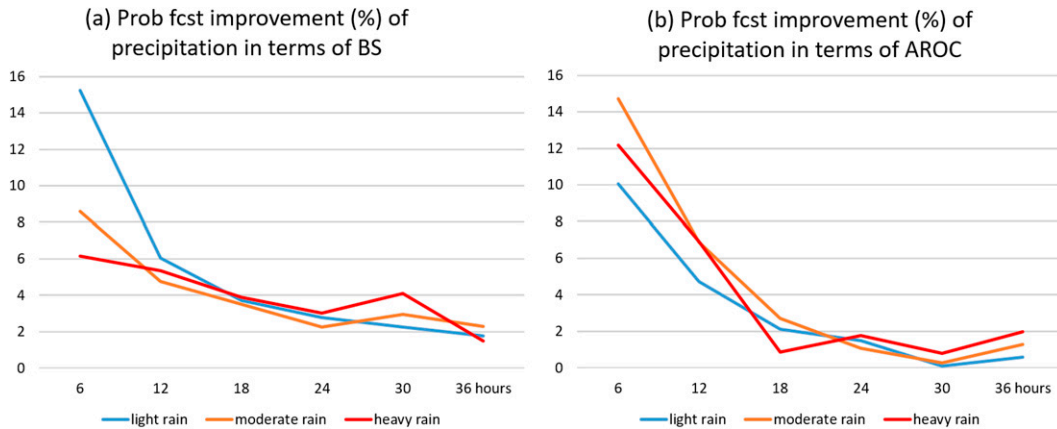


FIG. 15. Improvement (%) in probabilistic forecasts of three precipitation categories (light, moderate, and heavy), in terms of (a) BS and (b) AROC.

improvement is similar (about 4% and 5%) for all three rain categories, which indicates that the 3D rescaling mask cannot only improve light rain but also heavy rain events. However, readers need to keep in mind that although light, moderate and heavy rain events occur almost every day (Table 1) during the verification periods, the areal coverage of heavy rain is smaller than light and moderate categories, caution should be used in interpreting heavy rain category results.

Another measure to verify probabilistic precipitation forecasts is the relative operating characteristic (ROC). An ROC curve is plotted on a plane where the probability of detection (hit rate) is in the vertical axis ( $y$  axis) against the probability of false detection (false alarm rate) in the horizontal axis ( $x$  axis). Therefore, it measures ability of a forecast to discriminate between event (occurrence) and nonevent (nonoccurrence) (i.e., the resolution aspect of a forecast). For a good forecast, the hit rate should be as high as possible, while the false alarm rate should be as low as possible at the same time. The area under the ROC curve (AROC) is often used as a summary score of forecast resolution. The AROC value ranges from 0 (the worst score, hit rate is 0.0 and false alarm rate is 1.0 or ROC curve is the  $x$  axis) to 1 (the perfect score, hit rate is 1.0 and false alarm rate is 0.0 or the ROC curve is alone the  $y$ -axis forming a  $1 \times 1$  square). The diagonal line (AROC is 0.5) indicates that hit rate is equal to false alarm rate (50%), which is a boundary to distinguish a random forecast. In other words, a forecast is better (worse) than a random forecast when AROC is greater (less) than 0.5, indicating that hit rate exceeds (is lower than) false alarm rate. Figure 14 shows the AROC scores for the same three rain categories (light, moderate and heavy). Both 2D and 3D mask runs are skillful (AROC > 0.5). The improvement in terms of AROC is also observed for all three rain categories. On average over 6–36-h forecast range, the improvement is similar (around 4%) for all three rain categories. The improvement is, however, particularly striking prior to 18 h with the statistical confidence level of exceeding at least 75%. This is because the 3D mask run has almost eliminated the model spinup issue of the 2D mask run in precipitation forecasts. For example, the

forecast skill in the 2D mask run unusually increases with forecast hours prior to 18 h and then naturally decreases with the increase of forecast length. This “spinup” phenomenon has been overcome in the 3D mask run. After that the improvement slightly decreases or remains similarly with forecast time. Figure 15 summarizes the improvements in probabilistic quantitative precipitation forecasts in terms of both BS and AROC, where the improvement’s decreasing trend with forecast hours can be clearly seen for all three categories.

##### 5) A SUMMARY OF ALL SCORES

All the verification scores of ensemble mean, spread and probabilistic forecasts are summarized in a scorecard for all variables (Fig. 16). From the scorecard, we can see the following. (i) The benefit of 3D rescaling mask to ensemble performance is obvious: out of 420 verification measurements, 49% (205/420) is significantly improved (at 75%, 95%, and 99.7% levels), 50% (210/420) is comparable or neutral (not statistically significant), and only 1% (5/420) is significantly degraded (mainly upper level geopotential height spread). (ii) Significant improvement (at 75%, 95%, and 99.7% levels) is mainly occurred in the early forecast hours primarily prior to 24 h and lower levels (850 hPa and surface). The decreasing impact of IC perturbation methods with forecast length is similar to what Li et al. (2017) found in their study. (iii) Overall, the biggest improvement occurred in the ensemble spread in terms of outlier and consistency. Due to the increased spread, ensemble envelope can encompass observation more often (i.e., reduced outlier), and ensemble spread is more representative to forecast error of ensemble mean (i.e., improved consistency or spread–skill relationship). The second biggest improvement is probabilistic forecast in terms of CRPS, BS and AROC, which becomes more reliable and sharper. The least improved is ensemble mean forecasts in terms of RMSE. This is expected given the fact that the base model GRAPES\_Meso has a quite large forecast bias and the bias is a big part of ensemble mean forecast error (see Wang et al. 2018). Unless a model-based bias correction scheme (such as Chen et al. 2020) is implemented



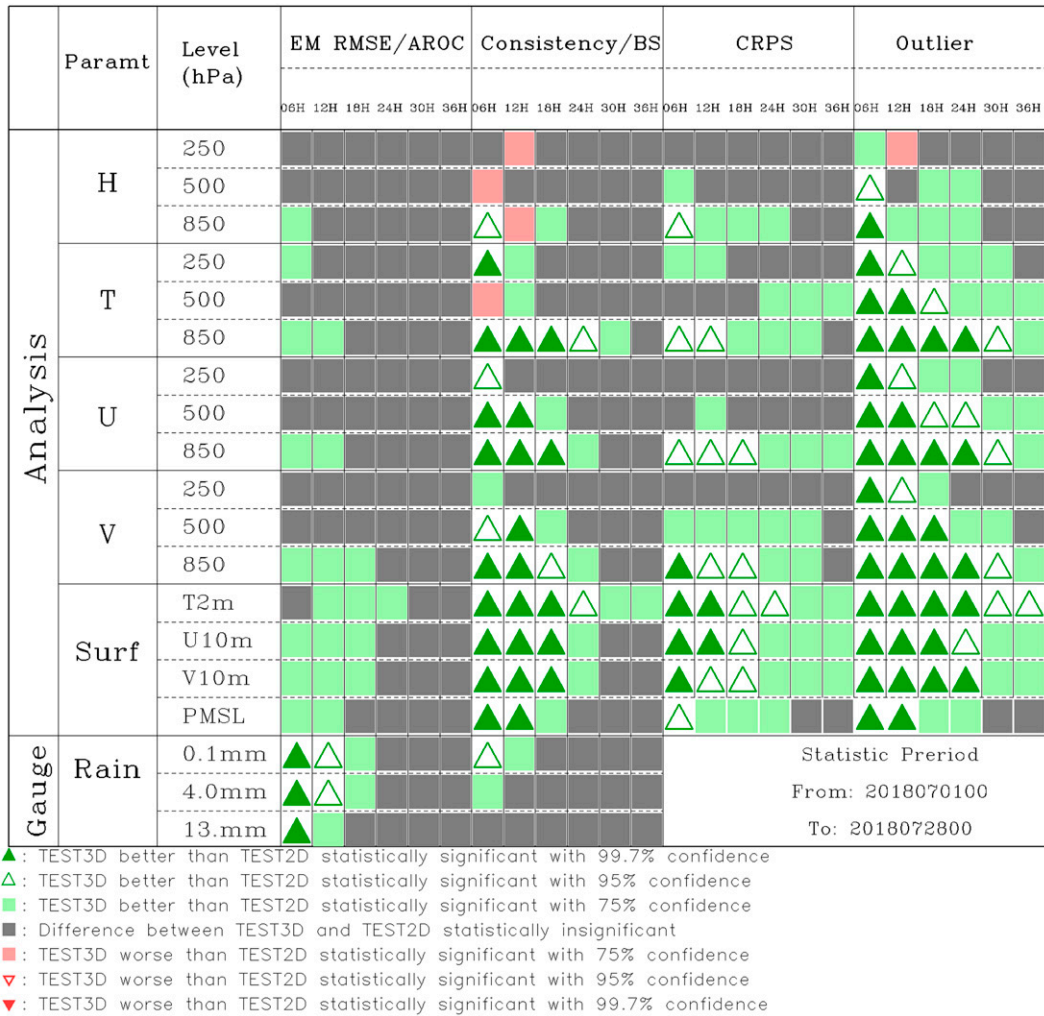


FIG. 16. Scorecards of the 3D mask experiment for the warm season (1–28 Jul 2018). Green indicates an improvement, red indicates a degradation, and gray is neutral (changes not statistically significant) with respect to the control 2D mask run. Different symbols are associated with different level of statistical significance of a *t* test (see the legend for the details). Note that the scores AROC and BS were used for precipitation forecasts only. For score definitions, please see the related text.

together with an ensemble perturbation scheme (such as Xia et al. 2019), pure ensemble perturbation techniques address random error only but not model bias (Du et al. 2018).

To better demonstrate this layered improvement structure among ensemble mean, spread, and probabilistic information, Fig. 17 shows an example of surface variables (10-m *U*, 2-m *T*, *P*\_MSL, and light, moderate, and heavy precipitation), where improvements in ensemble mean, spread and probabilistic forecasts are compared side by side. For the average of 6–36-h forecasts (Fig. 17a), the improvement is about 20% for ensemble spread, 5% for probabilistic forecasts, and 2.5% for ensemble mean forecasts. For the average of first 18-h forecasts (Fig. 17b), the improvement is about 30% for ensemble spread, 9% for probabilistic forecasts, and 4% for ensemble mean forecasts.

We have repeated the same set of experiment and verification for a cold-season period: 1–27 February 2019. Similar results are also seen although the improvement is in a lesser degree compared to the warm-season period. All verification scores of the cold-season cases are summarized in the scorecard in Fig. 18. We can see that 27% (114/420) of all scores have statistically significant improvement, 72% (303/420) of them are comparable or neutral (not statistically significant), and less than 1% (3/420) is significantly degraded (overdispersion). Also note that due to no enough sample for the heavy rain category in the cold season, some statistical significance levels cannot be calculated but just leave blank in the scorecard for heavy rain category (Fig. 18). In the following section, we will try to preliminarily explore how this 3D rescaling method might work and to understand why it works better in

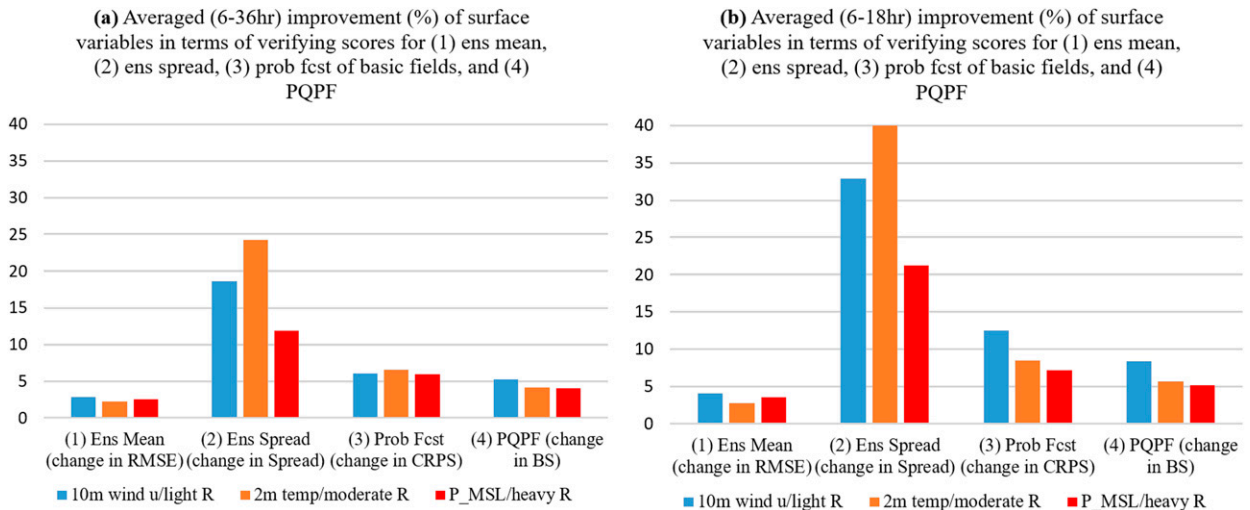


FIG. 17. Improvement (%) of surface variables (10-m wind  $U$ , 2-m temperature, sea level pressure, and light/moderate/heavy precipitations) for 1) ensemble mean forecasts in terms of RMSE, 2) ensemble forecast spread in terms of spread itself, 3) probabilistic forecasts of basic fields in terms of CRPS, and 4) probabilistic quantitative precipitation forecasts (PQPF) of three categories (light, moderate, and heavy) in terms of BS. (a) Averaged over 6–36 h and (b) averaged over 6–18 h.

warm season than in cold season and why the improvement decreases with the increasing of forecast length.

### c. Improvement mechanism

From the above analysis such as Figs. 16 and 17, we see that most obvious improvement is in the boost of ensemble spread, which subsequently results in the improvements in probability distribution and ensemble mean forecasts for an underdispersive EPS. In this section, we will explore possible mechanism leading to the differences in ensemble spread growth as well as the different performances between warm and cold seasons. Since ensemble IC perturbation growth is related to atmospheric instabilities (Toth and Kalnay 1997), the relationship between instability and ensemble spread change (e.g., spread increase from 2D rescaling-based EPS to 3D rescaling-based EPS) is investigated.

There are two dominant instabilities for weather to develop: baroclinic and convective instabilities. Baroclinic instability is a process by which disturbances draw (kinetic) energy from the mean-flow potential energy through warm air rising and cold air sinking in zones with large horizontal temperature gradient. Given the thermal wind balance, it is also reflected in vertical wind shear. Therefore, it can also be defined as the nonuniform vertical distribution of mean zonal flow ( $\bar{u}$ ):

$$\text{baroclinicity} \propto -\frac{\partial}{\partial p} \left( \frac{\partial \bar{u}}{\partial p} \right) \begin{cases} > 0 \text{ unstable} \\ = 0 \text{ neutral} \\ < 0 \text{ stable} \end{cases}, \quad (5)$$

where  $p$  is pressure. For example, zonal wind speed increases with height (decreases with pressure) below a jet stream and decreases with height (increases with pressure) above it, which results in baroclinicity  $> 0$  and causes very strong baroclinic instability near a jet stream. Convective instability is

related to the vertical distribution of atmospheric thermal condition and defined as

$$\text{convective instability} = \frac{\partial \theta_e}{\partial p} \begin{cases} > 0 \text{ unstable} \\ = 0 \text{ neutral} \\ < 0 \text{ stable} \end{cases}, \quad (6)$$

where  $\theta_e$  is equivalent potential temperature. In a convectively unstable (stable) atmosphere, when an air parcel rises it will accelerate (decelerate) because it will be warmer (cooler) than its surrounding environmental air.

We have, therefore, examined if the spread change from the 2D to 3D rescaling-factor based EPS is related to the baroclinic or convective instabilities. The instabilities are calculated from the ensemble mean forecast of the 3D rescaling-factor EPS. The horizontal temperature gradient magnitude is used for the baroclinic instability, and Eq. (6) is used for the convective instability. Figure 19 compares the spread increment (color shaded) of 850-hPa specific humidity to the baroclinic instability (contour) at forecast hours of 18, 24, 30, and 36, initialized at 0000 UTC 10 February 2019 [the same cold-season snow event demonstrated in section 3b(1)]. The baroclinic instability area is not well organized but scattered in smaller scale and does not match to the spread increment area. Figure 20 compares the spread increment to the convective instability for the same cold-season case.

In contrast to the baroclinic instability, the convective instability is more organized at the larger scale and matches the spread increment area better. Figures 21 and 22 are the same as Figs. 19 and 20 but initiated from 0000 UTC 10 July 2018 [the same warm-season heavy rain case of section 3b(1)], where the similar results are observed as in the snow event.

To examine this relationship more quantitatively, spatial correlation between the spread change of 850-hPa specific

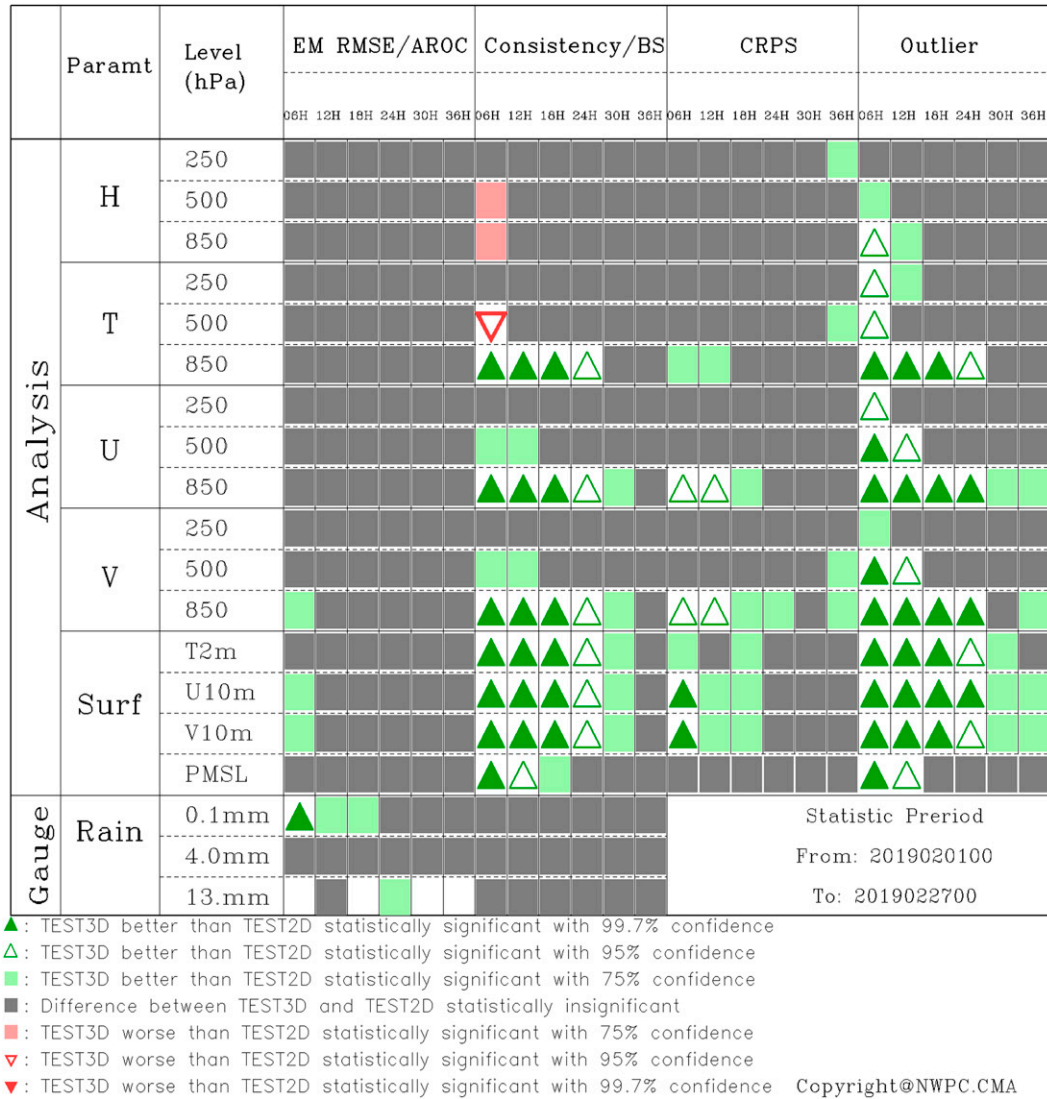


FIG. 18. As in Fig. 16, but for the cold season (1–27 Feb 2019). Note that there is not a large enough sample size to calculate statistical significance level for the heavy precipitation category [ $>13 \text{ mm (6 h)}^{-1}$ ] at 6, 18, 30, and 36 forecast hours for AROC.

humidity and the instabilities have been calculated for the summer and winter experiment periods. The result is shown in Fig. 23. We can see that the spread change is more correlated to convective instability than baroclinic instability for both seasons. The correlation seems to be stronger in summer than winter. This result suggests that the increase of ensemble spread in the experiment EPS is likely through the convective instability. Given that the EPS is in storm-scale (3 km), it is not surprising to see that convective instability plays a more important role than baroclinic instability for the spread growth. Therefore, to have a more effective storm-scale EPS of short-range forecasts over a region, its IC perturbation should be designed to target the structure of convective instability. This might be different for a synoptic-scale and global EPS where IC perturbation targets more on baroclinic instability (Toth and Kalnay 1997).

Comparing the summer case (Fig. 22) with the winter case (Fig. 20), we can see that the convective instability is much stronger in warm season ( $\sim 2.0 \times 0.00001 \text{ K Pa}^{-1}$ ) than in cold season ( $\sim 0.5 \times 0.00001 \text{ K Pa}^{-1}$ ). This explains why the spread increment and forecast improvement is more in the summer month than in the winter month. Since convective instability is a fast-growing mode, ensemble spread associated with it becomes saturated quickly with time. This could also explain why the improvement decreases with the increase of forecast length. This result implies that retaining or further increasing ensemble spread beyond a certain forecast length might be more difficult for a storm-scale EPS than a synoptic-scale global EPS, which EPS developers should pay attention to. Note that since this investigation into the relationship between ensemble spread growth and atmospheric instabilities in a storm-scale EPS is very preliminary (based on only one

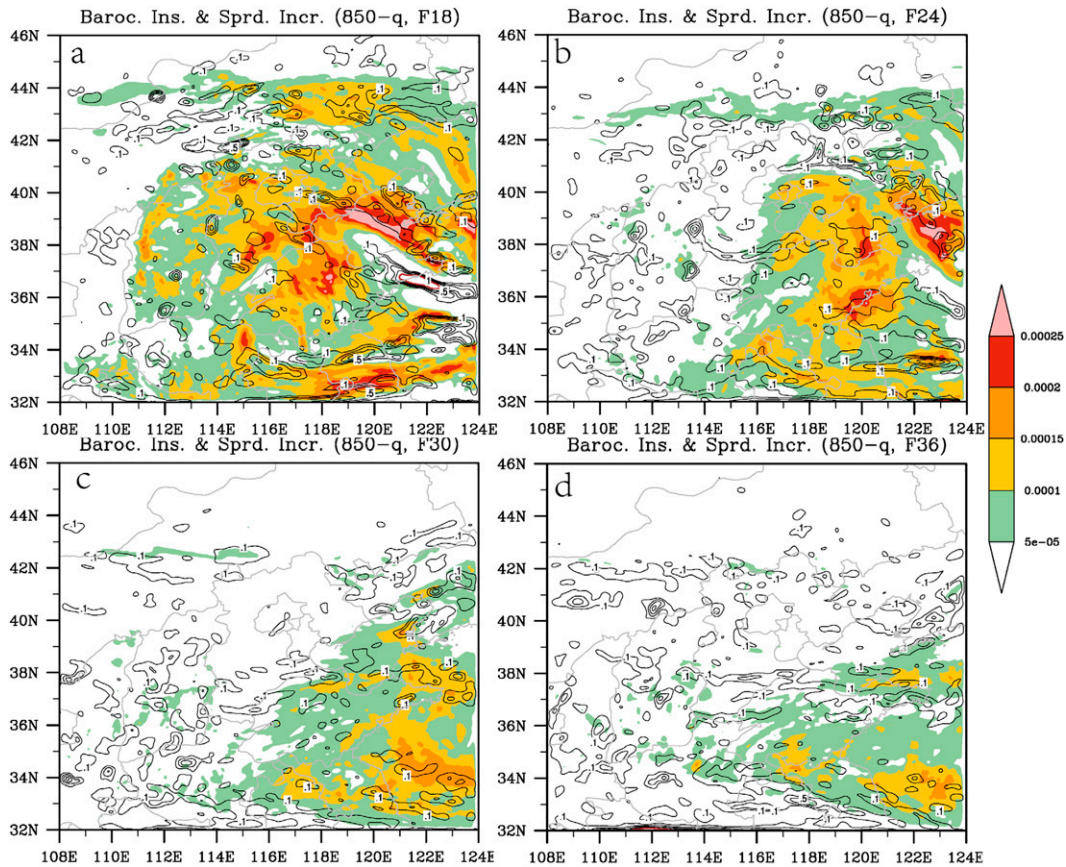


FIG. 19. Baroclinic instability (derived from the 3D EPS's ensemble mean forecast) (contour;  $0.00001 \text{ K Pa}^{-1}$ ) and the ensemble spread increment (3D EPS spread - 2D EPS spread) of specific humidity (color;  $\text{kg kg}^{-1}$ ) at the (a) 18-, (b) 24-, (c) 30-, and (d) 36-h forecast, initiated from 0000 UTC 10 Feb 2019 (a cold-season case).

variable 850-hPa specific humidity), the result here only tends to shed light on this topic but not definitive. More rigorous in-depth study is needed to thoroughly understand storm-scale EPS' behaviors.

#### 4. Summary and discussion

A 3D mask to rescale IC perturbation has been proposed and tested using a 3-km storm-scale EPS. A systematic evaluation has been conducted to examine if it improves ensemble forecasts. The forecasts of wind, temperature, and geopotential height at various levels, sea level pressure and precipitation have been examined in terms of ensemble mean forecast, ensemble spread and probabilistic forecast. The study was carried out in a summer month (1–28 July 2018) and a winter month (1–27 February 2019) over a region in North China. The experiment runs twice per day initiated at 0000 and 1200 UTC into 36 h in forecast length, providing a total of 56 36-h forecast warm-season cases and 54 cold-season cases for verification. To compare the differences in performance, warm and cold seasons were verified separately. Below is a summary of the findings with some discussions.

- (i) As intended, the 3D mask makes the IC perturbation more representative to analysis uncertainty than the 2D mask. The vertical profile of the IC perturbation size is much closer to the estimated intrinsic analysis error.
- (ii) The performance of ensemble forecasts has been significantly improved in all aspects including ensemble mean forecast, ensemble spread and probabilistic forecasts. The most improvement occurred in ensemble spread, followed by probabilistic forecasts, while the least improvement is associated with the ensemble mean forecast. This could be explained by the fact that an improvement to IC perturbation method mainly improves ensemble diversity (spread) but does not reduce model bias, while bias error constitutes a major part of ensemble mean forecast error for the base model. The improvements are larger in warm season than in cold season.
- (iii) The improvement decreases with the increase of forecast length. This result is consistent with the conclusion of Li et al. (2017) who found that the resulting ensemble spread from vastly different perturbation-generating methods became similar to each other when forecast length increased.

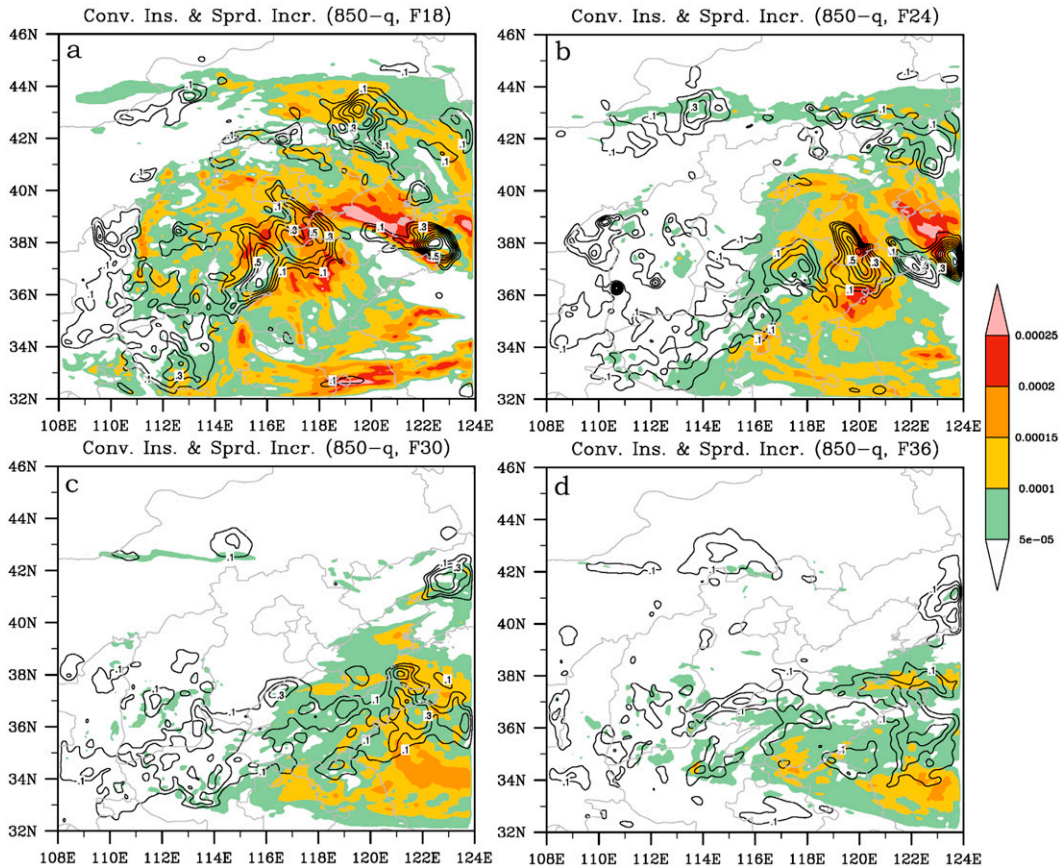


FIG. 20. Convective instability (derived from the 3D EPS’s ensemble mean forecast) (contour;  $0.00001 \text{ K m}^{-1}$ ) and the ensemble spread increment (3D EPS spread – 2D EPS spread) of 850-hPa specific humidity (color;  $\text{kg kg}^{-1}$ ) at the (a) 18-, (b) 24-, (c) 30-, and (d) 36-h forecast, initiated from 0000 UTC 10 Feb 2019 (a cold-season case).

- (iv) The improvement is found to be greater in lower levels and surface than upper levels. It is probably due to the fact that near surface variables are normally more under-dispersive than upper air variables in an ensemble (Du et al. 2018), providing more room to be improved in lower levels than in upper levels.
- (v) The increment of ensemble spread from the 2D to 3D rescaling-factor based IC perturbation is realized likely through convective instability rather than baroclinic instability. This explains why the forecast improvement is more in the summer month (stronger convective instability) than in the winter month. This could also explain why the improvement decreases with the increase of forecast length because convective instability is a fast-growing mode and becomes saturated quickly with time. This result is preliminary and needs to be further studied because it is based on only one variable 850-hPa specific humidity. If ensemble spread growth is indeed more sensitive to convective instability than baroclinic instability in a storm-scale EPS of short-range forecasts, IC perturbation should be designed to target the structure of convective instability to have a more effective ensemble of forecasts. This is different from designing a

synoptic-scale and global EPS which IC perturbation targets more on baroclinic instability. On a negative side of this result is that retaining or further increasing ensemble spread beyond a certain forecast length could be more challenging for a storm-scale EPS than a synoptic-scale global EPS.

This study recommends that a 3D rescaling mask should be used to replace a commonly used 2D one in current operational EPSs. It is important for shorter time range and near surface weather element forecasts, on which are particularly focused by storm-scale ensembles. Storm-scale EPS has been proven to be useful for high impact weather events (Roberts et al. 2020). There is a side benefit of this 3D rescaling mask for data assimilation. Being more representative to true analysis uncertainty, 3D rescaled IC perturbations could also be more useful in ensemble-conventional hybrid data assimilation such as ensemble Kalman filter (EnKF; Zhou et al. 2017). Computationally, the calculation of 3D rescaling factor costs almost nothing and can be done instantaneously in a super-computer. The rescaling factor [Eq. (1)] and new perturbation [Eq. (2)] are calculated only once at model’s initial time for each forecast cycle. The only difference between the 2D and

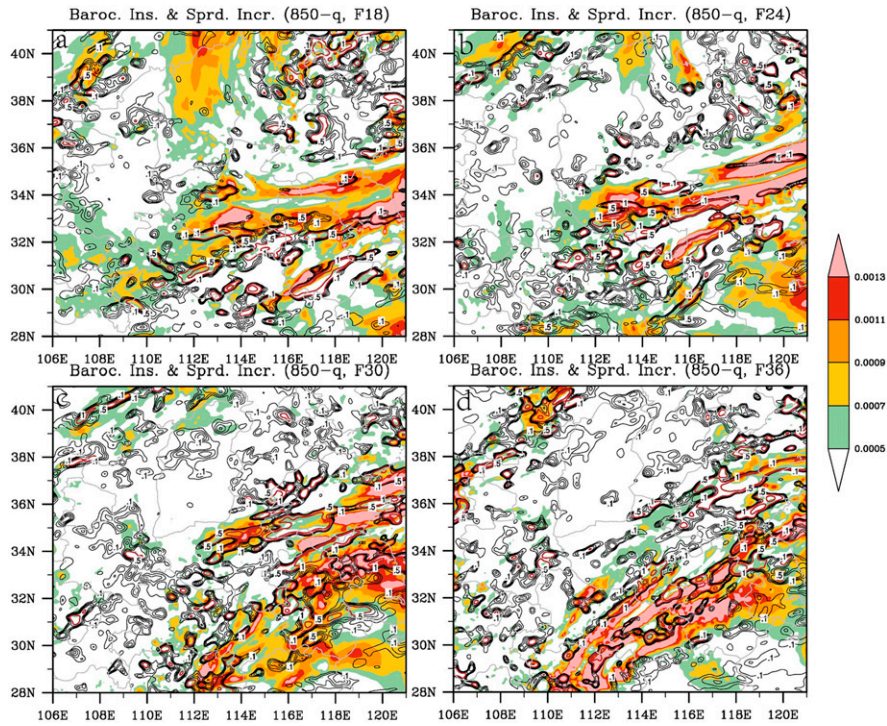


FIG. 21. Same baroclinic instability as in Fig. 19, but for the warm-season case (0000 UTC 10 Jul 2018).

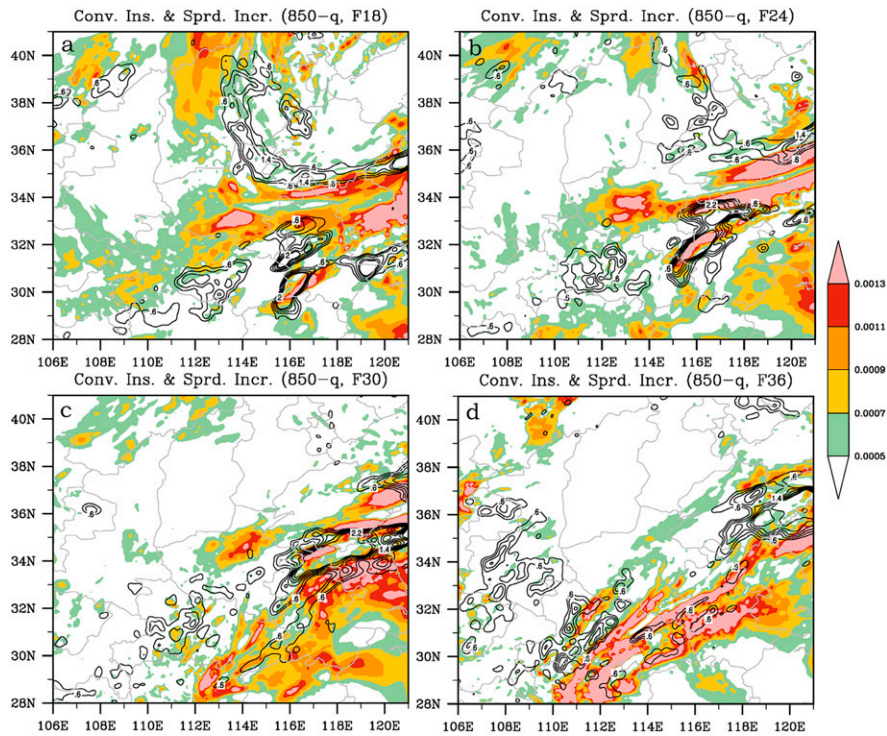


FIG. 22. Same convective instability as in Fig. 20, but for the warm-season case (0000 UTC 10 Jul 2018).

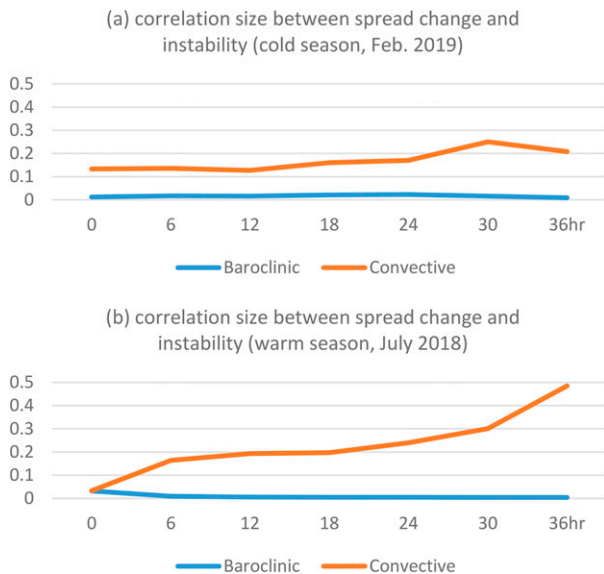


FIG. 23. Correlation between the spread change of 850-hPa specific humidity and instabilities (baroclinic instability in blue and convective instability in brown) for (a) cold season (February 2019) and (b) warm season (July 2018) over the model domain.

3D rescaling factor is that the former is calculated at one level, while the latter is calculated at all model levels. In other words, the extra computing time for the 3D rescaling factor can be practically neglected. Finally, a possible pitfall of this new 3D rescaling mask is that it could lead to over-dispersion of ensemble spread if an original 2D rescaling-based EPS has already an adequate spread–skill relation. However, this situation should occur rarely because almost all current EPSs in operation are underdispersive including the GRAPES EPS of this study. Theoretically, an EPS should not be overdispersive if a model’s uncertainty sources are not fully sampled in a perturbation method. Spurious ensemble spread only occurs when a perturbation method is not well designed or tuned like the ad hoc multimodel method, where a right answer could stem from a wrong reason. Therefore, how to increase diversity among ensemble members is still a main task of EPS design by improving perturbation methods nowadays (Du et al. 2018). The 3D rescaling mask demonstrated in this study is a low-hanging fruit to improve ensemble diversity.

**Acknowledgments.** The work was jointly supported by National Key Research and Development Program (2018YFF0300103), National Natural Science Foundation of China (41975137, 42175012, and 41475097), and Key Scientific and Technology Research and Development Program of Jilin Province (20180201035SF).

**Data availability statement.** The experiments and verification data are accessible at <https://pan.baidu.com/s/1m1jmWuFR9dh-Fucikcpymw>. The password is 1111, and the folder name is REPS\_DATA.

## REFERENCES

- Anderson, J. L., and S. L. Anderson, 1999: A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Mon. Wea. Rev.*, **127**, 2741–2758, [https://doi.org/10.1175/1520-0493\(1999\)127<2741:AMCIOT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1999)127<2741:AMCIOT>2.0.CO;2).
- Buizza, R., M. Miller, and T. N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **125**, 2887–2908, <https://doi.org/10.1002/qj.49712556006>.
- , J. Du, Z. Toth, and D. Hou, 2018: Major operational ensemble prediction systems (EPS) and the future of EPS. *Handbook of Hydrometeorological Ensemble Forecasting*, Q. Duan et al., Eds., Springer, 1–43, [https://doi.org/10.1007/978-3-642-40457-3\\_14-1](https://doi.org/10.1007/978-3-642-40457-3_14-1).
- Caron, J. F., 2013: Mismatching perturbations at the lateral boundaries in limited-area ensemble forecasting: A case study. *Mon. Wea. Rev.*, **141**, 356–374, <https://doi.org/10.1175/MWR-D-12-00051.1>.
- Chen, D. H., and X. S. Shen, 2006: Recent progress on GRAPES research and application. *J. Appl. Meteor. Sci.*, **17**, 773–777, <https://doi.org/10.3969/j.issn.1001-7313.2006.06.014>.
- , J. S. Xue, and X. S. Yang, 2008: New generation of multi-scale NWP system (GRAPES): General scientific design. *Chin. Sci. Bull.*, **53**, 3433–3445, <https://doi.org/10.1007/s11434-008-0494-z>.
- Chen, J., J. Wang, J. Du, Y. Xia, F. Chen, and H. Li, 2020: Forecast bias correction through model integration: A dynamical wholesale approach. *Quart. J. Roy. Meteor. Soc.*, **146**, 1149–1168, <https://doi.org/10.1002/qj.3730>.
- Deng, G., and Coauthors, 2010: Development of mesoscale ensemble prediction system at National Meteorological Center. *Chin. J. Appl. Meteor. Sci.*, **21**, 513–523, <https://doi.org/10.3969/j.issn.1001-7313.2010.05.001>.
- Du, J., 2002: Present situation and prospects of ensemble numerical prediction. *J. Appl. Meteor. Sci.*, **13**, 16–28.
- , and M. S. Tracton, 2001: Implementation of a real-time short-range ensemble forecasting system at NCEP: An update. Preprints, *Ninth Conf. on Mesoscale Processes*, Ft. Lauderdale, FL, Amer. Meteor. Soc., P4.9, 355–356, [https://ams.confex.com/ams/WAF-NWP-MESO/techprogram/paper\\_23074.htm](https://ams.confex.com/ams/WAF-NWP-MESO/techprogram/paper_23074.htm).
- , and B. Zhou, 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://link.springer.com/chapter/10.1007/978-3-319-45229-6\\_10](https://link.springer.com/chapter/10.1007/978-3-319-45229-6_10).
- , S. L. Mullen, and F. Sanders, 1997: Short-range ensemble forecasting of quantitative precipitation. *Mon. Wea. Rev.*, **125**, 2427–2459, [https://doi.org/10.1175/1520-0493\(1997\)125<2427:SREFOQ>2.0.CO;2](https://doi.org/10.1175/1520-0493(1997)125<2427:SREFOQ>2.0.CO;2).
- , G. DiMego, B. Zhou, D. Jovic, B. Ferrier, and B. Yang, 2015: Short Range Ensemble Forecast (SREF) system at NCEP: Recent development and future transition. *23rd Conf. on Numerical Weather Prediction/27th Conf. on Weather Analysis and Forecasting*, Chicago, IL, Amer. Meteor. Soc., 2A.5, <https://ams.confex.com/ams/27WAF23NWP/webprogram/Paper273421.html>.
- , and Coauthors, 2018: Ensemble methods for meteorological predictions. *Handbook of Hydrometeorological Ensemble Forecasting*, Q. Duan et al., Eds., Springer, 1–52, [https://doi.org/10.1007/978-3-642-40457-3\\_13-1](https://doi.org/10.1007/978-3-642-40457-3_13-1).
- Dudhia, J., 1989: Numerical study of convection observed during the Winter Monsoon Experiment using a mesoscale two-dimensional model. *J. Atmos. Sci.*, **46**, 3077–3107, [https://doi.org/10.1175/1520-0469\(1989\)046<3077:NSOCOD>2.0.CO;2](https://doi.org/10.1175/1520-0469(1989)046<3077:NSOCOD>2.0.CO;2).

- Epstein, E. S., 1969: Stochastic dynamic prediction. *Tellus*, **21**, 739–759, <https://doi.org/10.3402/tellusa.v21i6.10143>.
- Feng, Y., J. Min, X. Zhuang, and S. Wang, 2019: Ensemble sensitivity analysis-based ensemble transform with 3D rescaling initialization method for storm-scale ensemble forecast. *Atmosphere*, **10**, 24, <https://doi.org/10.3390/atmos10010024>.
- Flowerdew, J., and N. Bowler, 2013: On-line calibration of the vertical distribution of ensemble spread. *Quart. J. Roy. Meteor. Soc.*, **139**, 1863–1874, <https://doi.org/10.1002/qj.2072>.
- Hong, S. Y., and H. L. Pan, 1996: Nonlocal boundary layer vertical diffusion in a medium-range forecast model. *Mon. Wea. Rev.*, **124**, 2322–2339, [https://doi.org/10.1175/1520-0493\(1996\)124<2322:NBLVDI>2.0.CO;2](https://doi.org/10.1175/1520-0493(1996)124<2322:NBLVDI>2.0.CO;2).
- , and J. O. Lim, 2006: The WRF Single Moment 6-class microphysics scheme (WSM6). *J. Korean Meteor. Soc.*, **42**, 129–151.
- Jolliffe, I. T., and D. B. Stephenson, Eds., 2003: *Forecast Verification*. Wiley Press, 247 pp.
- Lacarra, J. F., and O. Talagrand, 1988: Short range evolution of small perturbation in a barotropic model. *Tellus*, **40A**, 81–95, <https://doi.org/10.1111/j.1600-0870.1988.tb00408.x>.
- Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.*, **102**, 409–418, [https://doi.org/10.1175/1520-0493\(1974\)102<0409:TSOMCF>2.0.CO;2](https://doi.org/10.1175/1520-0493(1974)102<0409:TSOMCF>2.0.CO;2).
- Li, J., J. Du, M. Wang, and C. Cui, 2009: Experiments of perturbing initial conditions in the development of mesoscale ensemble prediction system for heavy rainstorm forecasting. *Plateau Meteor.*, **28**, 1365–1375.
- , —, Y. Liu, and J. Xu, 2017: Similarities and differences in the evolution of ensemble spread using various ensemble perturbation methods including topography perturbation. *Acta Meteor. Sin.*, **75**, 123–146, <https://doi.org/10.11676/qxxb2017.011>.
- Liu, Y. Z., X. S. Shen, and X. L. Li, 2013: Research on the singular vector perturbation of the GRAPES global model based on the total energy norm. *Acta Meteor. Sin.*, **71**, 517–526, <https://doi.org/10.3969/j.issn.1004-4965.2013.03.020>.
- Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.*, **20**, 130–141, [https://doi.org/10.1175/1520-0469\(1963\)020<0130:DNF>2.0.CO;2](https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2).
- Ma, J., Y. Zhu, D. Hou, X. Zhou, and M. Peña, 2014: Ensemble transform with 3D rescaling initialization method. *Mon. Wea. Rev.*, **142**, 4053–4073, <https://doi.org/10.1175/MWR-D-13-00367.1>.
- Ma, X., J. Xue, and W. Lu, 2008: Preliminary study on ensemble transform Kalman filter based initial perturbation scheme in GRAPES global ensemble prediction. *Acta Meteor. Sin.*, **66**, 526–536, <https://doi.org/10.11676/qxxb2008.050>.
- , Z. Zhuang, J. Xue, and W. Lu, 2009: Development of the three-dimensional variational data assimilation system for the nonhydrostatic GRAPES. *Acta Meteor. Sin.*, **23**, 725–737.
- Mahrt, L., and M. Ek, 1984: The influence of atmospheric stability on potential evaporation. *J. Climate Appl. Meteor.*, **23**, 222–234, [https://doi.org/10.1175/1520-0450\(1984\)023<0222:TIOASO>2.0.CO;2](https://doi.org/10.1175/1520-0450(1984)023<0222:TIOASO>2.0.CO;2).
- Mlawer, E. J., S. J. Taubman, P. D. Brown, M. J. Iacono, and S. A. Clough, 1997: Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-*k* model for the longwave. *J. Geophys. Res.*, **102**, 16 663–16 682, <https://doi.org/10.1029/97JD00237>.
- Molteni, F., R. Buizza, T. Parmer, and T. Petroloagis, 1996: The ECMWF ensemble prediction system: Methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122**, 73–119, <https://doi.org/10.1002/qj.49712252905>.
- Noilhan, J., and S. Planton, 1989: A simple parametrization of land surface processes for meteorological models. *Mon. Wea. Rev.*, **117**, 536–549, [https://doi.org/10.1175/1520-0493\(1989\)117<0536:ASPOLS>2.0.CO;2](https://doi.org/10.1175/1520-0493(1989)117<0536:ASPOLS>2.0.CO;2).
- Parrish, D., and J. Derber, 1992: The National Meteorological Center spectral statistical interpolation analysis. *Mon. Wea. Rev.*, **120**, 1747–1763, [https://doi.org/10.1175/1520-0493\(1992\)120<1747:TNMCS>2.0.CO;2](https://doi.org/10.1175/1520-0493(1992)120<1747:TNMCS>2.0.CO;2).
- Roberts, B., T. G. Burkely, I. L. Jirak, A. J. Clark, D. C. Dowell, X. Wang, and Y. Wang, 2020: What does a convection-allowing ensemble of opportunity buy us in forecasting thunderstorms. *Wea. Forecasting*, **35**, 2293–2316, <https://doi.org/10.1175/WAF-D-20-0069.1>.
- Toth, Z., and E. Kalnay, 1997: Ensemble forecasting at NCEP: The breeding method. *Mon. Wea. Rev.*, **125**, 3297–3319, [https://doi.org/10.1175/1520-0493\(1997\)125<3297:EFANAT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1997)125<3297:EFANAT>2.0.CO;2).
- Wang, J., J. Chen, J. Du, Y. Zhang, Y. Xia, and D. Guo, 2018: Sensitivity of ensemble forecast verification to model bias. *Mon. Wea. Rev.*, **146**, 781–796, <https://doi.org/10.1175/MWR-D-17-0223.1>.
- , —, H. Zhang, H. Tian, and Y. Shi, 2021: Initial perturbations based on ensemble transform Kalman filter with rescaling method for ensemble forecasting. *Wea. Forecasting*, **36**, 823–842, <https://doi.org/10.1175/WAF-D-20-0176.1>.
- 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](https://doi.org/10.1175/1520-0469(2003)060<1140:ACOBAE>2.0.CO;2).
- Wang, Y., M. Bellus, J. Geleyn, X. Ma, W. Tian, and F. Weidle, 2014: A new method for generating initial condition perturbations in regional ensemble prediction system: Blending. *Mon. Wea. Rev.*, **142**, 2043–2059, <https://doi.org/10.1175/MWR-D-12-00354.1>.
- Wu, W., and R. J. Purser, 2002: Three-dimensional variational analysis with spatially inhomogeneous covariances. *Mon. Wea. Rev.*, **130**, 2905–2916, [https://doi.org/10.1175/1520-0493\(2002\)130<2905:TDVAWS>2.0.CO;2](https://doi.org/10.1175/1520-0493(2002)130<2905:TDVAWS>2.0.CO;2).
- Xia, Y., J. Chen, J. Du, X. Zhi, J. Wang, and X. Li, 2019: A unified scheme of stochastic physics and bias correction in an ensemble model to reduce both random and systematic errors. *Wea. Forecasting*, **34**, 1675–1691, <https://doi.org/10.1175/WAF-D-19-0032.1>.
- Zhang, H. B., J. Chen, X. F. Zhi, Y. Wang, and Y. N. Wang, 2015: Study on multi-scale blending initial condition perturbations for a regional ensemble prediction system. *Adv. Atmos. Sci.*, **32**, 1143–1155, <https://doi.org/10.1007/s00376-015-4232-6>.
- Zhang, X., and J. Sun, 2018: Analysis of the July 2018 atmospheric circulation and weather. *Meteor. Mon.*, **44**, 1370–1376.
- Zhou, X., Y. Zhu, D. Hou, and D. Kleist, 2016: Comparison of the ensemble transform and the ensemble Kalman filter in the NCEP global ensemble forecast system. *Wea. Forecasting*, **31**, 2057–2074, <https://doi.org/10.1175/WAF-D-16-0109.1>.
- , —, —, Y. Lou, J. Peng, and R. Wobus, 2017: Performances of the new NCEP global ensemble forecast system in a parallel experiment. *Wea. Forecasting*, **32**, 1989–2004, <https://doi.org/10.1175/WAF-D-17-0023.1>.
- Zhuang, Z., J. Xue, and H. Lu, 2014: Experiments of global GRAPES-3DVar analysis based on pressure level and prediction system. *Plateau Meteor.*, **33**, 666–674.
- , R. Wang, J. Wang, and J. Gong, 2019: Characteristics and application of background errors in GRAPES\_Meso. *J. Appl. Meteor. Sci.*, **30**, 316–331, <https://doi.org/10.11898/1001-7313.20190306>.