An Objective Approach to Generating Multi-Physics Ensemble Precipitation Forecasts Based on the WRF Model

Chenwei SHEN[1], Qingyun DUAN[2*], Wei GONG[1], Yanjun GAN[3], Zhenhua DI[1], Chen WANG[4], and Shiguang MIAO[5]

1 *State Key Laboratory of Earth Surface Processes and Resource Ecology*, *Faculty of Geographical Science*, *Beijing Normal University*, *Beijing* 100875
2 *State Key Laboratory of Hydrology–Water Resources and Hydraulic Engineering and College of Hydrology & Water Resources*, *Hohai University*, *Nanjing* 210098
3 *State Key Laboratory of Severe Weather*, *Chinese Academy of Meteorological Sciences*, *China Meteorological Administration*, *Beijing* 100081
4 *South China Botanical Garden*, *Chinese Academy of Sciences*, *Guangzhou* 510650
5 *Institute of Urban Meteorology*, *China Meteorological Administration*, *Beijing* 100081

(Received November 14, 2019; in final form February 19, 2020)

ABSTRACT

Selecting proper parameterization scheme combinations for a particular application is of great interest to the Weather Research and Forecasting (WRF) model users. This study aims to develop an objective method for identifying a set of scheme combinations to form a multi-physics ensemble suitable for short-range precipitation forecasting in the Greater Beijing area. The ensemble is created by using statistical techniques and some heuristics. An initial sample of 90 scheme combinations was first generated by using Latin hypercube sampling (LHS). Then, after several rounds of screening, a final ensemble of 40 combinations were chosen. The ensemble forecasts generated for both the training and verification cases using these combinations were evaluated based on several verification metrics, including threat score (TS), Brier score (BS), relative operating characteristics (ROC), and ranked probability score (RPS). The results show that TS of the final ensemble improved by 9%–33% over that of the initial ensemble. The reliability was improved for rain ⩽ 10 mm day^{-1}, but decreased slightly for rain > 10 mm day^{-1} due to insufficient samples. The resolution remained about the same. The final ensemble forecasts were better than that generated from randomly sampled scheme combinations. These results suggest that the proposed approach is an effective way to select a multi-physics ensemble for generating accurate and reliable forecasts.

Key words: ensemble precipitation forecast, Weather Research and Forecasting (WRF) model, multi-physics, verification, bootstrapping

Citation: Shen, C. W., Q. Y. Duan, W. Gong, et al., 2020: An objective approach to generating multi-physics ensemble precipitation forecasts based on the WRF model. *J. Meteor. Res.*, **34**(3), 601–620, doi: 10.1007/s13351-020-9198-3.

1. Introduction

Much progress has been made in the last half century in developing sophisticated, physically realistic numerical weather prediction (NWP) models. Those models have been widely used today for weather forecasting with lead times up to two weeks into the future and have played a critical role in emergency management to alleviate the impact of severe weather events (Skamarock et al., 2008; Du and Qian, 2014). Traditional weather forecasts, known as deterministic forecasts, have been issued in the form of a single space–time series for variables of interest. There are inherent limitations to predict the future state of the atmosphere by single deterministic forecasts due to the chaotic nature of the fluid dynamic equations involved (Lorenz, 1965). The predictive skill of regional NWP models is affected by three major sources of error: (1) the errors in the initial conditions needed by a model

Supported by the Chinese Academy of Sciences Strategic Pioneering Program (XDA20060401), China Meteorological Administration Special Public Welfare Research Fund (GYHY201506002), National Basic Research Program of China (2015CB953703), and Intergovernment Key International S&T Innovation Cooperation Program (2016YFE0102400).

*Corresponding author: qyduan@hhu.edu.cn.

to generate a forecast, which are amplified by the chaos effects as the lead time increases; (2) the errors due to idealistic abstraction by parameterization schemes in representing highly heterogeneous and nonlinear physical processes within a model grid and approximate numerical solutions to the dynamic equations; and (3) the errors from the boundary conditions due to inaccuracy in the global model predictions and inaccurate data specification at the boundary (Tribbia and Baumhefner, 1988; Miao et al., 2019). To counter these limitations, ensemble forecasting methods have emerged in the last 40 years as a promising way to account for the uncertainties due to these errors. Ensemble forecasts are generated by perturbing uncertain factors such as initial and boundary conditions, representations of model physics, or both of them.

Leith (1974) made the first attempt at ensemble forecasting by using Monte Carlo simulations with different initial states to estimate the means and variances of future atmospheric states. However, it was noted that there is not enough dispersion for the ensembles constructed in this way unless a huge number of ensemble members are used and the ensemble probability distribution is a representative sample of the distribution of the actual atmospheric states. To overcome the deficiency of the crude Monte Carlo approach, strategies for increasing ensemble dispersion by identifying the growing modes in the atmospheric states have been developed. Two such strategies have gained prominence: the breeding of growing modes method developed at the NCEP and the singular vector method developed at the ECMWF (Toth and Kalnay, 1993, 1997; Molteni et al., 1996). Ensemble data assimilation methods, such as the ensemble Kalman filter, have also been used to generate ensemble forecasts by perturbing the initial conditions (Burgers et al., 1998; Houtekamer and Mitchell, 1998; Houtekamer and Zhang, 2016). Numerous researchers found that perturbing model physics can also improve the performance of the ensembles (Stensrud et al., 2000; Zheng et al., 2019; Gou et al., 2020).

There are different ways to perturb model physics. One way is to construct stochastic parameterization schemes by adding random perturbations to the physical components in the parameterization schemes (Du and Li, 2014; Berner et al., 2017). Buizza et al. (1999) and Palmer et al. (2005) used random fields as amplification factors to physics tendency terms such as temperature, specific humidity, and wind components to generate stochastically perturbed parameterization tendencies (SPPT). Another scheme, known as the stochastic kinetic energy backscatter (SKEB), attempts to account for the uncertainties arising from scale interactions that exist in real atmosphere, but ignored in grid-scale parameterization schemes (Berner et al., 2009, 2017). There are other schemes available that follow similar principles as the SPPT and SKEB schemes, such as stochastic convective backscatter scheme (Shutts, 2015), physically based stochastic perturbations (Kober and Craig, 2016), and improved stochastic kinetic energy backscatter version 2 (Sanchez et al., 2016).

Another way to perturb model physics is to randomly sample the adjustable model parameters, which are the constants and exponents contained in the equations of parameterization schemes (Bowler et al., 2008; McCabe et al., 2016). Even though some model parameters are well constrained by observations, many parameters are empirical in nature and are subject to large uncertainty. The key to perturbing model parameters is to identify the most sensitive parameters whose variations cause large perturbations in model response (Di et al., 2015, 2018; Quan et al., 2016). A number of meteorological forecasting centers have tried ensemble forecasting with perturbed parameters (Irvine et al., 2013; Murphy et al., 2014; Christensen et al., 2015).

A third way to perturb model physics is to take a multi-model approach in which forecasts from different models are combined to form a forecast ensemble. There are different ways to combine multi-model forecasts (Ebert, 2001). The simplest way, known as a "poor man's ensemble," is to combine forecasts from all models to form an ensemble that can provide a larger spread of forecast, compared to the ensemble based on a single model. More sophisticated multi-model ensemble approaches include the super-ensemble approach developed by Krishnamurti et al. (1999) and the Bayesian model averaging (BMA) method by Raftery et al. (2005), which assign weights to different forecasts based on their agreement with past observations. The Observing System Research and Predictability Experiment (THORPEX) Interactive Grand Global Ensemble (TIGGE) launched in 2005 to enhance medium-range forecast of high impact weather extremes by operational centers has stimulated extensive research on making use of forecasts from truly independent models to generate multi-model ensembles (Bougeault et al., 2010; Sun et al., 2020).

Previous efforts in generating multi-physics ensemble forecasts have been either based on perturbation of a specific model by adding stochastic components to parameterization schemes or based on a combination of independent models. Many Weather Research and Forecasting (WRF) model users create multi-physics ensemble by focusing on selecting the potential schemes of a particular process or by randomly selecting them. For example,

Efstathiou et al. (2013) examined the performance of two commonly used boundary layer schemes (Yonsei University and Mellor–Yamada–Janjic boundary layer schemes) on rainfall simulation over Chaldidiki Peninsula region, and their results show that the Yonsei University scheme performed better. Crétat et al. (2012) checked the performance of 27 WRF parameterization schemes combination on summer rainfall based on three cumulus, planetary boundary layer (PBL), and microphysics schemes, and their results show the necessity of conducting multi-physics ensemble testing. The arrival of the WRF model symbolizes a new era in numerical weather modeling as the WRF model can be regarded as millions of different models contained in a single framework. Trying to form a multi-physics ensemble using the WRF model poses a special challenge because it is impossible to try out all the possible combinations available in WRF, which exceeds two million based on WRF version 3.7.1. (Hereafter known as WRF3.7.1) .

How does one find the suitable combinations for an application over a particular area out of millions of potential combinations? Several studies have attempted to address this issue (Lee et al., 2011; Weusthoff et al., 2011; Lee, 2012). The heuristic approaches used in these studies start with an initial ensemble that contains a high number of ensemble members to ensure large uncertainties contained in the ensemble. Then, redundant ensemble members (i.e., the ensemble members that have high correlations with other ensemble members) are removed step by step while the uncertainties are retained as much as possible (Lee et al., 2011; Lee, 2012). Classification and performance criteria are commonly used in these approaches to identifying which parameterization schemes are suitable for a particular area. Generally, the following steps are taken to identify the desired ensemble members: (1) performance criteria are established to assess the performance of an individual ensemble member; (2) ensemble members are classified into several grade categories according to their performance indices; (3) the ensemble members with bad performance indices are removed, and the good ones are retained. These steps/approaches can only handle a limited number of potential combinations and are not designed to handle all of the potential combinations available in the WRF model.

Our work employs a systematic approach that incorporates several statistical techniques in addition to some heuristics to analyze all the plausible combinations of parameterization schemes available in the WRF model. The aim is to identify a set of parameterization scheme combinations that can be used to form a multi-physics ensemble based on several skill metrics for short-range precipitation forecasts. The paper is organized as follows. Section 2 presents a brief description of methodology. Section 3 describes model set up and verification datasets. Section 4 details the screening results. Section 5 evaluates the ensemble forecasts using screening results and provides further discussions. Section 6 presents conclusions.

2. Methodology

2.1 *The parameterization scheme combination selection procedure*

Our overall goal is to select a set of parameterization scheme combinations from millions of potential ones available from WRF3.7.1 to produce a short-range (3-day) ensemble precipitation forecast with satisfactory skill over the summer monsoon season in the Greater Beijing area. According to Du (2002), a good ensemble forecast should possess three features: (1) equal-likelihood for all ensemble members; (2) the ensemble mean having a good agreement with the observed value as measured by chosen performance metrics; (3) a good ensemble spread property as marked by a proper balance between reliability and resolution (i.e., ensemble distribution being sharp subject to calibration). To find a set of good ensemble members with those features, we have designed the following procedure:

(1) Remove any unsuitable schemes for a specific application from all physical processes (i.e., microphysics, longwave and shortwave radiation, PBL, surface layer, land surface, and cumulus cloud) in WRF3.7.1 based on the WRF3.7.1 User's Guide (2016) and on expert knowledge for given applications.

(2) Select randomly a large initial set of parameterization scheme combinations under allowable computational resources, $\boldsymbol{M} = \{M_i, i = 1, 2, \ldots, N\}$, from all feasible ones using a design of experiment (DOE) approach (to be described later), where M_i denotes a particular parameterization scheme combination formed by choosing one scheme from each of the physical processes and N is the initial sample size. Compute the performance metrics, $\boldsymbol{F}$, over the training period (which consists of a pre-specified number of multi-day cases, 3-day in this study) for each combination M_i in $\boldsymbol{M}$, where $\boldsymbol{F} = \{f_i, i = 1, 2, \ldots, N\}$, and f_i is the performance metrics for M_i.

(3) For each physical process $j = 1, \ldots L$, compute the average performance metrics for scheme k in process j, $\mu_{k,j}$, $k = 1, \ldots, K_j$, where K_j is the number of available schemes in physical process j and L is the total number of physical processes. Then, compute the variance of the performance metrics for process j:

$$\sigma_j^2 = \sum_{k=1}^{K_j} \frac{\left(\mu_{k,j} - \overline{\mu_j}\right)^2}{K_j}, \quad (1)$$

where $\mu_{k,j} = \sum_{i=1}^{L_{k,j}} \frac{\mu_{k,j}^i}{L_{k,j}}$ is the average performance metrics of all parameterization scheme combinations in which scheme k of physical process j appears, $\mu^i_{k,j}$ corresponds to the performance metrics of an individual scheme combination in which scheme k of physical process j appears, $L_{k,j}$ is the number of times scheme k of physical process j appears in the initial set of parameterization scheme combinations, and $\overline{\mu_j} = \sum_{k=1}^{K_j} \frac{\mu_{k,j}}{K_j}$ is the mean performance metrics of all schemes.

(4) Starting from the physical process with the highest variance, $\sigma_{\mathrm{m}}^2 = \max\left\{\sigma_j^2, j = 1, \ldots, L\right\}$, screen the physical parameterization schemes in physical process j by keeping the schemes that are significantly better than the average performance of all parameterization schemes and removing the ones significantly worse than the average performance by using a two-sided test. For the schemes with no significant difference with average performance, remove the schemes whose case-to-case variances are significantly smaller than the average of all schemes.

(5) Start a new round of screening by sampling randomly a new set of combinations from the remaining schemes from the last round using DOE approach, and then repeat Steps (2)–(4) until the final remaining schemes are statistically not different from the mean or when the number of remaining combinations are within a pre-specified number that meets the users' requirements.

Step (1) from the above procedure ensures that only the suitable schemes from the WRF model are considered. Step (2) employs a DOE approach to sample different schemes randomly so that all feasible schemes from each physical process would have an equal chance of being chosen. The variance computed in Step (3) is an indicator of sensitivity of performance metrics to the choice of different schemes in a physical process. If the variance is high, it means that the choice of the schemes has a high sensitivity and thus a big impact on performance metrics. In Step (4), the actual screening is executed by keeping the good performing schemes as candidates to be included in a multi-physics ensemble and removing the bad ones from further consideration. For the rest of the schemes, the ones which display small discrepancies in performance metrics during different events are removed to enhance the diversity of the ensemble members. The final remaining parameterization scheme combinations after the above screening process completes would form the basis for the ensemble forecast experiments to be shown later.

2.2 *Statistical methods for screening parameterization scheme combinations*

In completing the above screening process, three statistical methods are used to objectively choose the multi-physics ensemble members. Two statistical techniques key to the screening methodology are used for screening: DOE approach to sampling parameterization schemes randomly and the statistical significance tests to distinguish the performance metrics of different schemes. The DOE method used in this study is the Latin hypercube sampling (LHS) design, which is a uniform sampling method (Loh, 1996; McKay et al., 2000; Helton and Davis, 2003). This method gives all possible schemes within a physical process the same chance to appear in the ensemble. The statistical significance tests used in this study are the one-sided or two-sided *t*-test. These tests are used to judge which schemes are significantly better or worse than the average performance. The detailed descriptions of the LHS design and *t*-tests are provided in the Appendixes A and B.

2.3 *Performance criteria and verification metrics*

In selecting good performing scheme combinations and in evaluating the ensemble forecasts from the resulting perturbed-physics ensemble members, several performance metrics are used to assess the forecast accuracy and reliability: threat score (TS; Zhao and Carr, 1997), root mean square error (RMSE), Brier score (BS; Brier, 1950), relative operating characteristics (ROC; Stanski et al., 1989), and ranked probability score (RPS; Murphy, 1969, 1971). These metrics were computed based on different 24-h rainfall intensities. When evaluating precipitation events of different intensities (V), performance metrics are usually computed separately for five storm categories: (1) no rain ($V < 0.1$ mm day^{-1}); (2) light rain (0.1 mm day^{-1} $\leqslant V \leqslant$ 10 mm day^{-1}); (3) moderate rain (10 mm day^{-1} $\leqslant V \leqslant$ 25 mm day^{-1}); (4) heavy rain (25 mm day^{-1} $\leqslant V \leqslant$50 mm day^{-1})); and (5) severe storm (V > 50 mm day^{-1}). In this study, due to the limited storm sample size, we use only two categories of rain to compute the performance metrics: $V \leqslant$ 10 mm day^{-1} and $V >$ 10 mm day^{-1}, and they are weighted equally to form a combined performance metric for each forecast. The descriptions of the aforementioned performance metrics are presented in Appendix C.

3. Model setup and verification datasets

3.1 *Model setup*

WRF3.7.1 was set up to run over a two-grid nested domain in this study (Fig. 1), with the outer domain includ-

ing most of northern China (i.e., d01 in Fig. 1) and the inner domain being the Greater Beijing area (i.e., d02 in Fig. 1). The outer domain is composed of 78 × 45 grid cells with a spatial resolution of 27 km, and the inner domain is composed of 85 × 49 grid cells with a spatial resolution of 9 km. The vertical profile for both domains is represented by 38 sigma vertical levels from the land surface to 50-hPa level in the atmosphere. The integration time step is 60 s. The NCEP Final (FNL) operational global analysis data from its Global Data Assimilation System (GDAS), available at the 1° × 1° horizontal resolution and 6-h intervals, were used to generate the initial and lateral boundary conditions. In the study, the PBL and surface layer schemes are used in tandem because the two schemes must be used in combination according to the WRF3.7.1 User's Guide (2016).

3.2 *Verification datasets*

The verification data used to evaluate the model forecasting performance is the China Meteorological Precipitation Analysis (CMPA)-hourly dataset (Shen et al., 2014) from the China Meteorological Administration (CMA), which was generated by merging hourly precipitation data from over 30,000 stations of the automatic weather station network with the Climate Prediction Center Morphing Technique gauge–satellite data (CMORPH; Joyce et al., 2004). The CMPA-hourly dataset has a spatial resolution of 0.1° × 0.1°. When computing the forecasting performance metrics such as TS, RMSE, BS, and ROC, the model outputs over the d02 domain were interpolated spatially to match the observation grids by using bilinear interpolation method.

3.3 *Selection of the training and verification events*

In northern China, rain is usually concentrated from June to August and is mostly caused by convective systems (Jiang et al., 2014). To improve short-range summer precipitation forecasts, 30 typical three-day rainfall cases (see Table 1) were chosen over the June–August period from 2014 to 2017 in the Greater Beijing area for our multi-physics ensemble selection study. Among the 30 cases, 15 of them were randomly chosen for training purpose and the remaining 15 were used for verification purposes. The model integral time spans 78 h, starting from 1800 UTC the day before, with the first 6-h integration for spinning up.

4. The parameterization scheme combination screening process and results

4.1 *Pre-screening and construction of the initial set of parameterization schemes*

Before we construct the initial set of parameterization scheme combinations, some unsuitable schemes for the Greater Beijing area were removed from the pool of potential schemes. For example, the Kessler microphysics scheme is a warm-rain (i.e. no ice) scheme used commonly in idealized cloud modeling studies, and the Held–Suarez temperature relaxation shortwave radiation scheme is also for idealized testing only, and the two have thus been removed from consideration. Table 2 lists all available parameterization schemes after those unsuitable schemes were removed. Some parameterization schemes in certain physical processes are developed with specific coupled physics codes and must be used in tan-

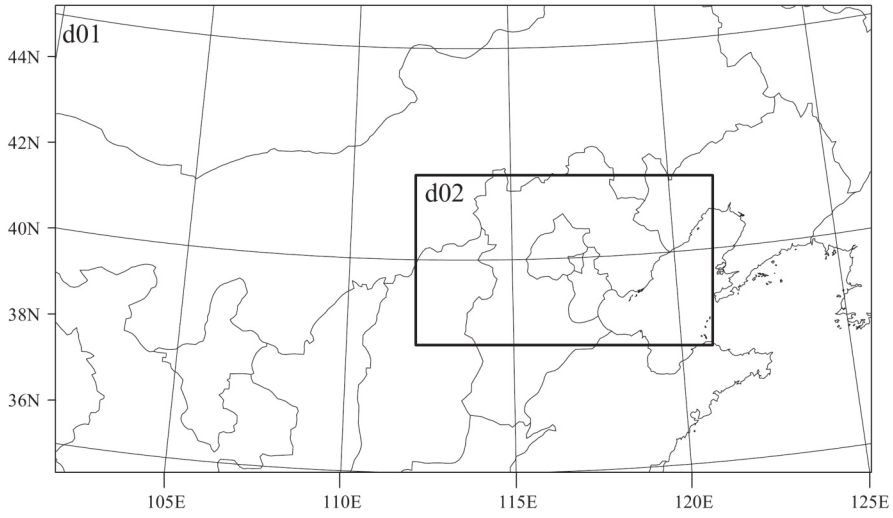

Fig. 1. The horizontal two-level nested grid domain with d01 being the outer grid domain and d02 the inner grid domain encompassing the Greater Beijing area.

Table 1. Starting dates for the 3-day forecasts during June–August of 2014–2017 in the Greater Beijing area, with 15 cases as training set and the other 15 cases as verification set. The integration time spans 78 h including the first 6 h for spinning up

Training set	Verification set
2014-06-30	2014-07-14
2014-07-18	2014-08-08
2014-07-28	2015-07-11
2014-08-02	2015-08-29
2014-08-11	2016-07-11
2014-08-20	2016-07-19
2014-08-26	2016-07-23
2014-08-29	2016-07-29
2015-07-14	2016-08-06
2015-07-16	2016-08-11
2015-07-20	2016-08-17
2015-07-26	2017-07-03
2015-07-31	2017-07-05
2015-08-04	2017-07-20
2015-08-06	2017-07-24

dem. For example, some surface layer and PBL schemes must be chosen together (denoted as PBL + surface or pbl + sfclay in Table 2). There were 15 schemes remaining for the microphysics (mp), 15 schemes for PBL and surface layer tandem (pbl + sfclay), 9 schemes for the cumulus convection (cu), 5 schemes for the shortwave radiation (ra_sw), 6 schemes for the longwave radiation (ra_lw), and 4 schemes for the land surface (sf_surface).

We used the LHS design to uniformly sample parameterization scheme combinations from Table 2. Ninety parameterization scheme combinations were sampled (this number can be enlarged if computational resources permit). For example, each scheme in microphysics appears 6 times, 18 times for longwave radiation, 15 times for shortwave radiation, 6 for PBL + surface, and 10 for cumulus convection schemes. For land surface, two schemes appear 22 times and the other two 23 times.

4.2 *The screening process and results*

After the initial set of 90 scheme combinations was created, each of those schemes was used to generate 3-day forecasts for the 15 training cases. The NCEP FNL data were used to set the initial and lateral boundary conditions for these forecasts. After the forecasts were generated, the performance metric, TS, for all training cases was computed according to Eq. (C1). We then computed the variance σ_j^2 of the performance metrics for each

Table 2. The schemes retained after pre-screening. In left columns of each category of schemes, the numbers in brackets represent the corresponding scheme options in the following screening process. Note that planetary boundary layer (PBL) and surface layer schemes are considered together (denoted as pbl + sfclay) in the screening process. Refer to WRF3.7.1 User's Guide (2016) for complete information

mp		pbl + sfclay		cu		ra_lw		ra_sw		sf_surface	
Scheme	Reference	Scheme	Reference	Scheme	Reference	Scheme	Reference	Scheme	Reference	Scheme	Reference
Lin (2)	Lin et al. (1983, JCAM)	YSU + MM5 (1)	Hong et al. (2006, MWR)	KF (1)	Kain (2004, JAM)	RRTM (1)	Mlawer et al. (1997, JGR)	Dudhia (1)	Dudhia (1989, JAS)	5-layer (1)	Jin et al. (2010, AW)
WSM3 (3)	Hong et al. (2004, MWR)	MYJ + Monin–Obukhov (2)	Janjic (1994, MWR)	BMJ (2)	Janjic (1994, MWR; 2000, JAS)	CAM (3)	Collins et al. (2004, NCAR Tech Note)	Goddard (old) (2)	Chou and Suarez (1994, NASA Tech Memo)	Noah (2)	Mitchell et al. (2005, NCAR Tech Note)
WSM5 (4)	Hong et al. (2004, MWR)	QNSE + QNSE (3)	Sukoriansky et al. (2005, BLM)	GF (3)	Grell et al. (2013, ACP)	RRTMG (4)	Iacono et al. (2008, JGR)	CAM (3)	Collins et al. (2004, NCAR Tech Note)	RUC (3)	Smirnova et al. (2000, JGR)
Ferrier (95)	Rogers et al. (2001, web doc)	MYNN2 + MM5 (4)	Nakanishi and Niino (2006, BLM)	SAS (4)	Pan and Wu (1995, NMC Office Note 409)	New Goddard (5)	Chou and Suarez (1999, NASA Tech Memo)	RRTMG (4)	Iacono et al. (2008, JGR)	Pleim–Xiu (7)	Pleim and Xiu (1995, 2001, JAM)
WSM6 (6)	Hong and Lim (2006, JKMS)	MYNN2 + Monin–Obukhov (5)	Nakanishi and Niino (2006, BLM)	Grell 3 (5)	Grell and Devenyi (2002, GRL)	FLG (7)	Gu et al. (2011, JGR), Fu and Liou (1992, JAS)	Goddard (new) (5)	Chou and Suarez (1999, NASA Tech Memo)		
Goddard (7)	Tao et al. (1989, MWR)	MYNN2 + MYNN (6)	Nakanishi and Niino (2006, BLM)	Tiedkte (6)	Tiedtke (1989, MWR)			FLG (7)	Gu et al. (2011, JGR), Fu and Liou (1992, JAS)		

Continued

mp		pbl + sfclay		cu		ra_lw		ra_sw		sf_surface	
Scheme	Reference	Scheme	Reference	Scheme	Reference	Scheme	Reference	Scheme	Reference	Scheme	Reference
Thompson (8)	Thompson et al. (2008, MWR)	MYNN3 + MYNN (7)	Nakanishi and Niino (2006, BLM)	New SAS (14)	Han and Pan (2011, Wea Forecasting)						
Milbrandt 2-mom (9)	Milbrandt and Yau (2005, JAS)	ACM2 + MM5 (8)	Pleim (2007, JAMC)	GD (93)	Grell and Devenyi (2002, GRL)						
Morrison 2-mom (10)	Morrison et al. (2009, MWR)	BouLac + MM5 (9)	Bougeault and Lacarrere (1989, MWR)	New Tiedkte (16)	Zhang et al. (2011, MWR)						
CAM5.1 (11)	Neale et al. (2012, NCAR Tech Note)	BouLac + Monin–Obukhov (10)	Bougeault and Lacarrere (1989, MWR)								
SBU_YLin (13)	Lin and Colle (2011, MWR)	UW + MM5 (11)	Bretherton and Park (2009, JC)								
WDM5 (14)	Lim and Hong (2010, MWR)	UW + Monin−Obukhov (12)	Bretherton and Park (2009, JC)								
WDM6 (16)	Lim and Hong (2010, MWR)	TEMF + TEMF (13)	Angevine et al. (2010, MWR)								
HUJI fast (30)	Khain et al. (2010, JAS)	GBM + MM5 (14)	Grenier and Bretherton (2001, MWR)								
Thompson aerosol-aware (28)	Thompson and Eidhammer (2014, JAS)	Shin−Hong + MM5 (15)	Shin and Hong (2015, MWR)								
Total 15		15		9		5		6		4	

Abbreviations of the journal titles used above. ACP: *Atmos. Chem. Phys.*; AW: *Adv. Meteor.*; BLM: *Bound.-Layer Meteor.*; JAM: *J. Appl. Meteor.*; JAMC: *J. Appl. Meteor. Climatol.*; JAS: *J. Atmos. Sci.*; JC: *J. Climate*; JCAM: *J. Clim. Appl. Meteor.*; JGR: *J. Geophy. Res.*; JKMS: *J. Korean Meteor. Soc.*; GRL: *Geophy. Res. Lett.*; MWR: *Mon. Wea. Rev.*

physical process to find the sensitivity of performance metrics to the selection of different schemes in physical process j. Note that the greater the variance is, the more sensitive the performance metrics to the selection of different schemes. Figure 2 gives the results of sensitivity of each physical process. Starting with the physical process with the largest σ_j^2, which is microphysics in this case, and then moving to physical process with next highest σ_j^2, longwave radiation, and so on, we performed the two-sided t-test for each physical process according to the procedure described in Section 2.1.

Figure 3 presents the details of the two-sided t-test results for the TS of different schemes. From Fig. 3a, we notice that TS of Goddard scheme (number 7) in microphysics is significantly better than the average for microphysics schemes, and CAM 5.1 scheme (number 11) has a TS that is significantly worse than the average, while the TSs for the rest of the schemes are not significantly different from the average. To further check the results of TS, the observation and error distributions (simulation minus observation) of the simulated rainfall of the Goddard and CAM5.1 schemes averaged for the training period (June–August of 2014–2017) over the domain d02 (see Fig. 1) are shown in Fig. 4. From Fig. 4, we notice that both schemes tend to overestimate the observed rainfall. But for the 48-h simulation, both schemes underestimate the rainfall for the west part of the domain. The RMSE results show that the Goddard scheme shows re-

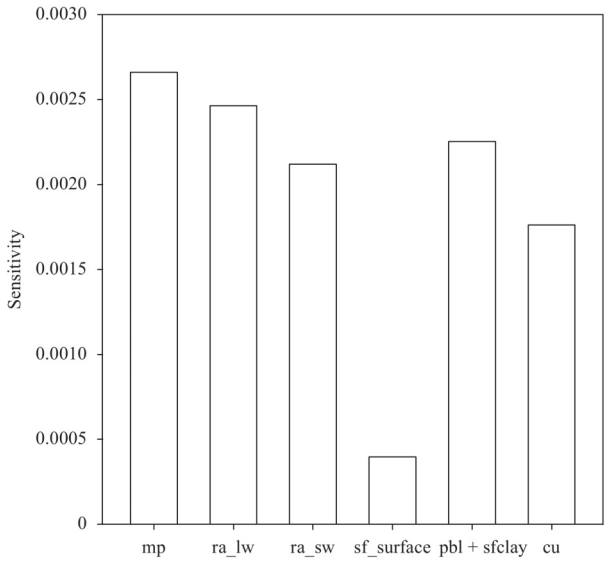

Fig. 2. The between scheme variances (sensitivity) of different physical processes in the first round of screening.

markably smaller simulation error compared to the CAM5.1 scheme, especially for 72-h simulation. Therefore, according to Step (4) in Section 2.1, we retain Goddard scheme and remove CAM 5.1 scheme from further consideration.

For the schemes with TSs that are not significantly different from the average TS, we then performed one-sided *t*-test to remove the schemes that have significantly smaller variances over different cases. Figure 5 gives the one-sided *t*-test results of variances over the 15 training cases at the 97.5% confident level. From Fig. 5a, we notice that the WSM6 (number 6) and SBU_YLin schemes (number 13) have significantly smaller variances than the average variance over the 15 cases and they were removed from further consideration. For microphysics, we retained Lin, WSM3, WSM5, Ferrier, Goddard, Thompson, Milbrandt 2-mom, Morrison 2-mom, WDM5, WDM6, and Thompson aerosol-aware schemes and removed WSM6, CAM 5.1, SBU_YLin, and HUJI SBM "fast" schemes in Table 2 (Note that scheme HUJI SBM "fast" in microphysics is not shown in the figure because the WRF model failed to complete the simulation due to segmentation fault, the same for scheme QNSE + QNSE (number 3) in PBL + surface layer combination). The screening process used for microphysics was repeated for the physical process with the next highest variance, the longwave radiation, whose results are shown in Figs 3d, 5d. The screening for longwave radiation resulted in the Goddard scheme (number 5) as the only scheme with a TS significantly better than the average TS. The TSs of the other schemes are shown to be not significantly different from the average TS. Figure 5d shows that none of the variances for different schemes over the 15 cases are significantly different from each other and no schemes were removed consequently. Thus, we retained all schemes for longwave radiation. We repeated the screening process for the remaining physical processes in the order of decreasing variance. At the end of the screening process, we retained 11 microphysics schemes, 5 longwave schemes, 4 shortwave schemes, 4 land surface schemes, 10 PBL + surface scheme tandems, and 7 cumulus convection schemes, respectively. Table 3 lists all the remaining schemes for each physical process at the end of the first round screening.

We started the second round of screening by uniformly sampling 70 scheme combinations from all of the schemes listed in Table 3. The same screening process as in the first round was carried out to retain and remove schemes according to the procedure described in Section 2.1. At the end of the second round of screening, we retained eight microphysics schemes, four longwave schemes, four shortwave schemes, four land surface schemes, seven PBL + surface tandems, and six cumulus convection scheme, respectively. Table 3 shows the remaining schemes for each physical process at the end of the second round of screening (the scheme names marked with "*" in Table 3 represent the schemes that were removed after the second round). For the sake of brevity, we do not show the detailed results of the second round of screening in the main text, but they are included in the supplemental material (see Figs. S1–S3).

The third round of screening was carried out like the second round. We sampled 48 combinations from the remaining schemes at the end of the second round. At the end of the third round of screening, three microphysics schemes, three PBL + surface schemes, and one cumulus convection scheme (schemes in italics in Table 3) were removed, while five microphysics schemes, four each of longwave radiation, shortwave radiation, land surface, PBL + surface scheme tandems, and five cumulus convection schemes were retained (see Table 3, schemes in boldface were retained). Figures S4–S6 record the results from the third round of screening.

Based on the screening results from the third round, we randomly sampled 40 scheme combinations from the remaining schemes (marked in boldface in Table 3) using LHS method and generated 40-member ensemble forecasts for the 15 training cases (each scheme combination is an ensemble member). We then computed the variances of the TSs of the different schemes. We pro-

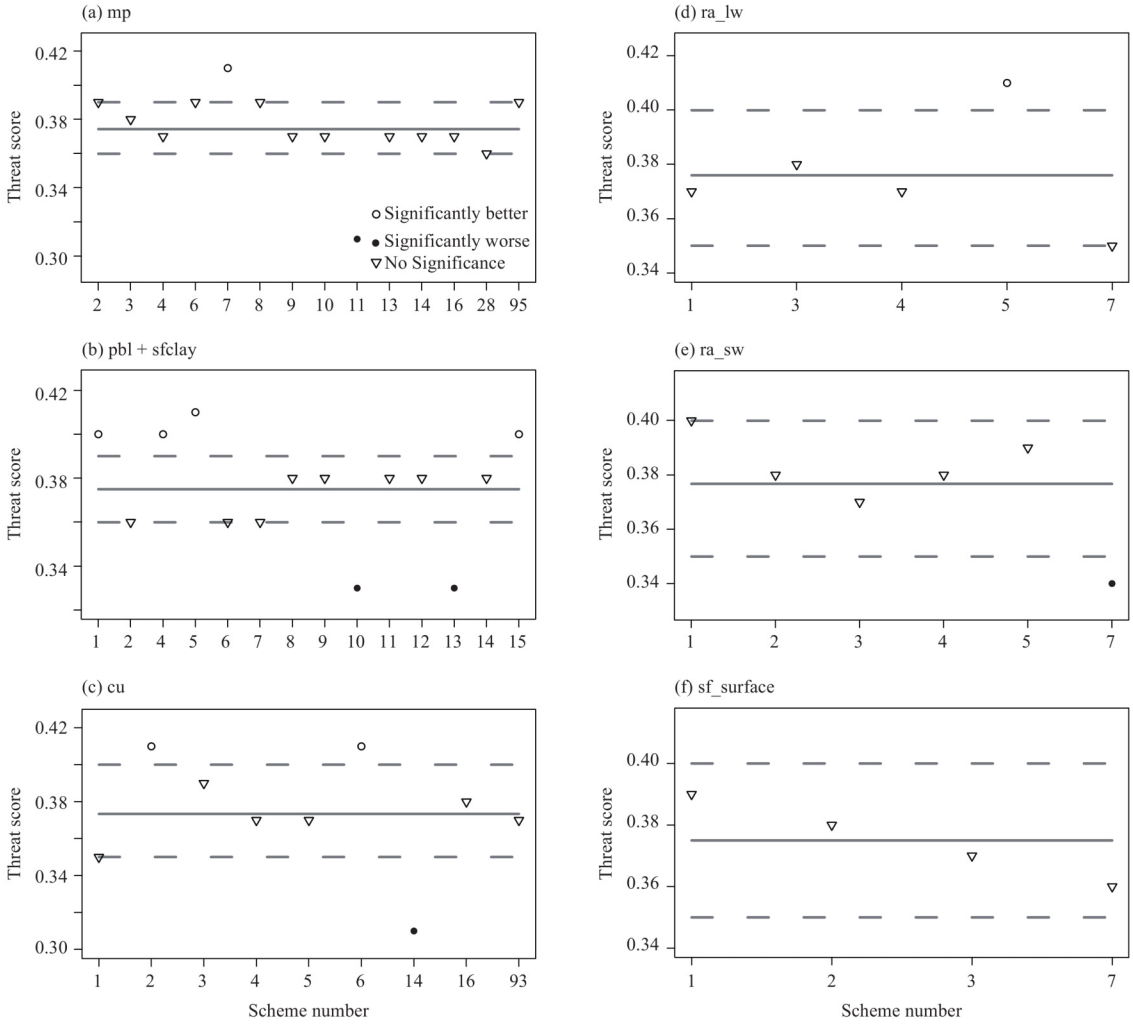

Fig. 3. Threat scores (TSs) of different schemes for (a) mp, (b) pbl + sfclay, (c) cu, (d) ra_lw, (e) ra_sw, and (f) sf_surface with a two-sided *t*-test in the first round of screening. The scheme number (*x*-axis) of corresponding schemes is indicated in Table 2 (numbers in brackets). The solid line represents the average TS over different schemes, and the dashed lines represent the upper and lower bounds of 95% confidence interval for the average TS. Symbols "◦" and "•" denote schemes corresponding to the TS significantly better or worse than the average TS, respectively; while "▽" denotes the TS not significantly different from the average TS.

ceeded to perform the two-sided *t*-tests for the TSs to identify the significantly better or worse schemes (see Fig. 6). Based on the results from the fourth round of screening, we found that the TSs of all schemes were not significantly different from the average TS, and therefore no scheme was removed. The screening process is thus completed, and there is no need for further screening at this point. This also means that any scheme combinations formed by selecting from the remaining schemes in Table 3 (in boldface) are deemed as not statistically different from other schemes.

5. Verification of the multi-physics ensemble forecasts

The 40-member ensemble forecasts of the 15 training cases from the final round of screening represent ensemble forecasts from a multi-physics ensemble because they are formed by using different scheme combinations. We also used the same scheme combinations to generate ensemble forecasts for the 15 verification cases that have not been used in the training. In this section, we evaluate the ensemble forecasts in terms of the ensemble mean and ensemble spread against observations using a number of different verification metrics. The verification results of the ensemble mean and ensemble spread are presented below. We also compare the final ensemble forecasts against the ensemble forecasts generated from scheme combinations randomly drawn (using bootstrapping method) from the initial 90 combinations to ensure that the results are robust statistically.

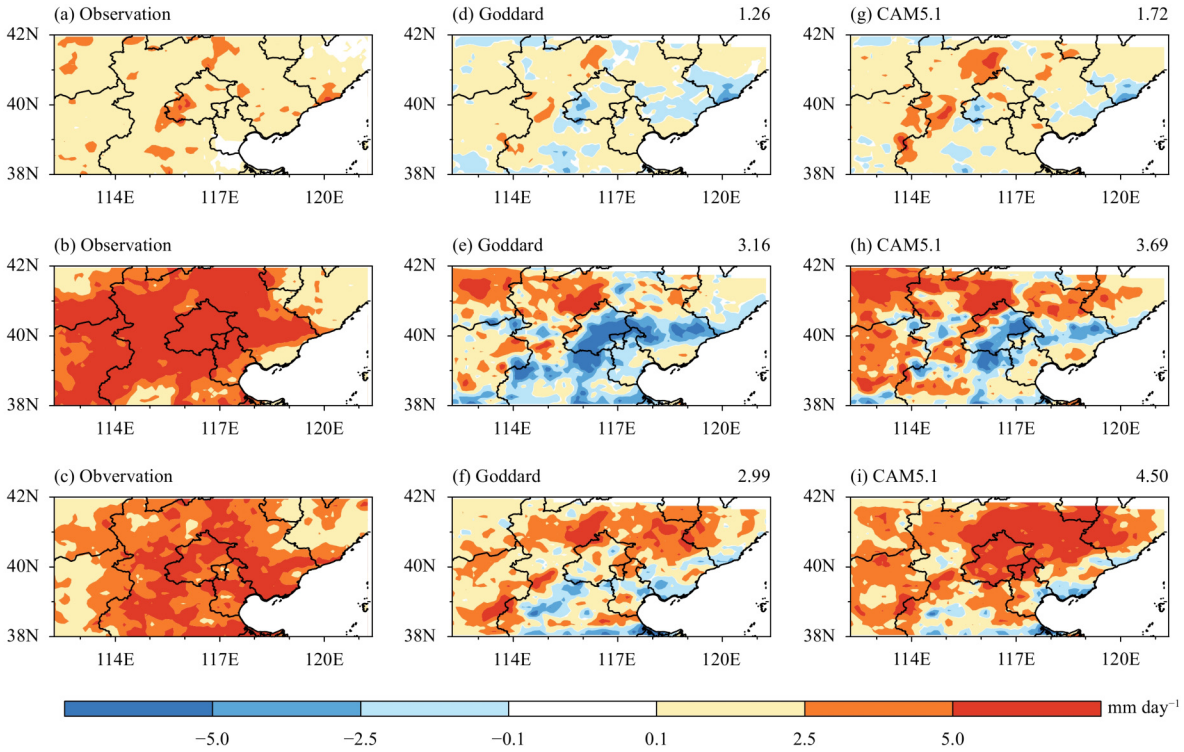

Fig. 4. Spatial distributions of (a–c) observed rainfall (mm day^{-1}) and (d–i) rainfall simulation error (mm day^{-1}) for the training period (June–August of 2014–2017) of (d–f) the Goddard scheme (number 7) and (g–i) the CAM 5.1 scheme (number 11) of microphysics at (d, g) 24-, (e, h) 48-, and (f, i) 72-h forecast lead times. The mean RMSE value (mm day^{-1}) of the Goddard and CAM 5.1 schemes is indicated on the top right of the corresponding panel.

5.1 *Deterministic verification of the ensemble forecast means*

Figure 7 shows the inter-comparison results of the TSs of the final 40 ensemble forecasts (denoted as final ensemble) against the initial 90 ensemble forecasts (denoted as initial ensemble) for the 15 training cases and 15 verification cases for lead times at 24, 48, and 72 h, respectively. The results indicate that the performance of the final ensemble is superior to that of the initial ensemble in all but one training case for forecast lead time at 72 h. In only one case (i.e., case 4 for lead time at 72 h), the TS of the final ensemble is inferior to that of the initial ensemble. The average improvement in the TSs is remarkable for the training cases, at 29%, 22%, and 17% for lead times of 24, 48, and 72 h, respectively, and for verification cases at 33%, 15%, and 9%, respectively.

Figure 8 is similar to Fig. 7, but for the RMSE of the initial and final ensemble forecasts for the training and verification cases. The results of RMSE are similar to those of TS. For most cases, the final ensemble is superior to the initial ensemble, especially for the verification cases that have more rainfall amount than that of the training cases (e.g., case 14 for lead time at 24 h). The average improvement in RMSE for the training cases is 1%, 0.4%, and 2% for lead times of 24, 48, and 72 h, respectively, and for verification cases is 12%, 7%, and 14%, respectively. For a better understanding of the precipitation simulation difference between the two ensembles, distribution patterns of simulation errors are given in Fig. 9, which shows that for both cases, the final ensemble has reduced simulation errors compared to that of the initial ensemble for all lead times. Especially for lead times of 24 and 72 h in the verification cases, the errors in the initial ensemble have been reduced in the final ensemble across the entire spatial domain. The spatial areas corresponding to the maximum error for the initial ensemble have been reduced in the final ensemble. Note that the final ensemble performed better in the verification cases than in the training cases, possibly due to the fact that there is more rainfall in the verification cases as compared to the training cases.

5.2 *Probabilistic verification of the ensemble forecast spreads*

For ensemble forecasts, probabilistic verification metric is a more comprehensive measure of ensemble forecast performance. Here, we use a number of probabilistic verification metrics, including RPS, BS, and ROC. RPS is a measure of how well forecasts expressed as

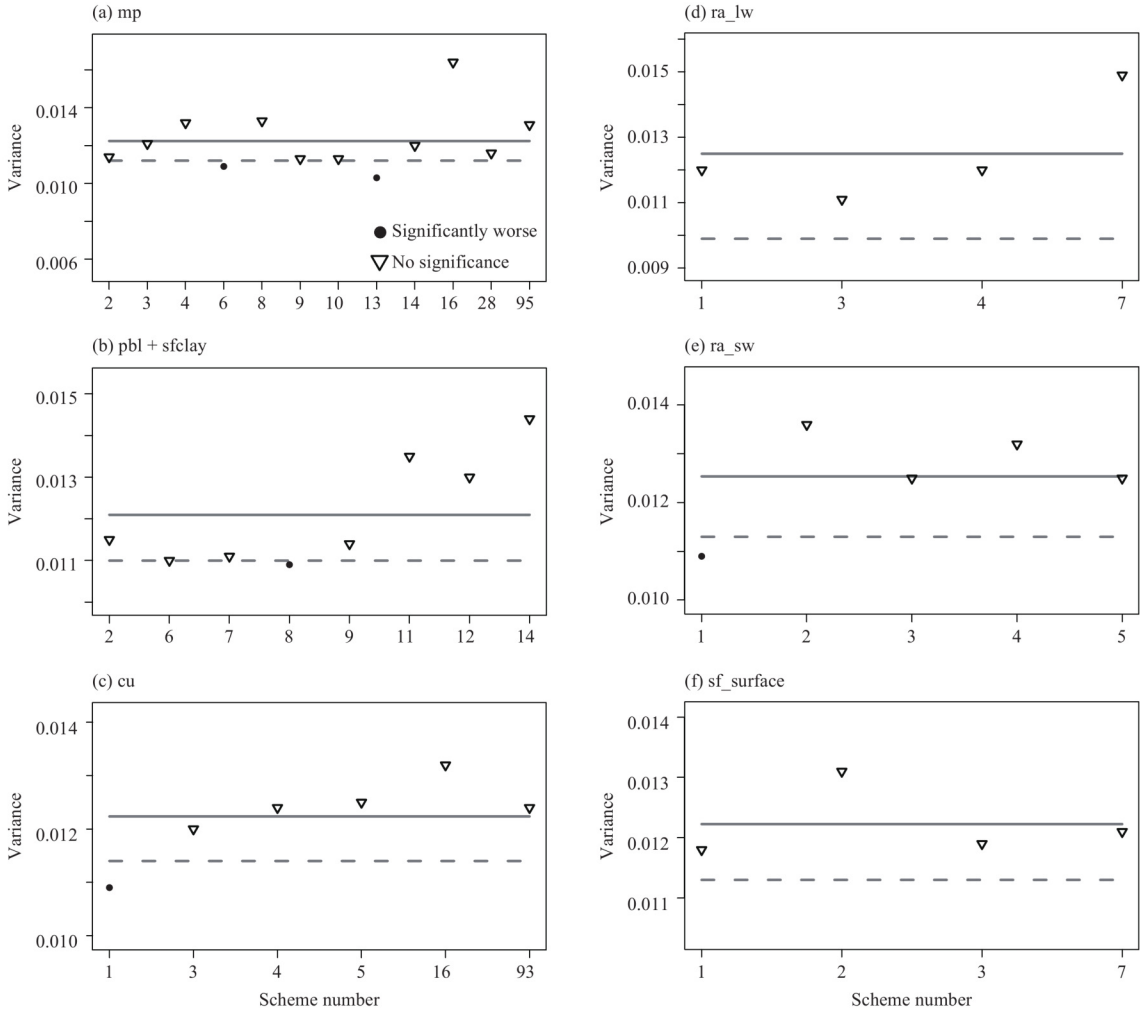

Fig. 5. As in Fig. 3, but for the variances of threat scores of different schemes over the training cases with a one-sided *t*-test.

Table 3. The scheme screening results. The schemes marked by boldface are the remaining schemes after three rounds of screening. The schemes marked with "*" (in italics) are the schemes that are removed in the second (third) round of screening

	Microphysics	Longwave	Shortwave	Land surface	PBL + surface	Cumulus
	Lin	**RRTM**	**Goddard (old)**	**5-layer**	**YSU + MM5**	**BMJ**
	WSM3*	**CAM**	**CAM**	**Noah**	**MYJ + Monin–Obukhov**	*GD*
	WSM5	**RRTMG**	**RRTMG**	**RUC**	*MYNN2.5 + MM5*	SAS*
	Ferrier (95)	**New Goddard**	**New Goddard**	**Pleim–Xiu**	*MYNN2.5 + Monin–Obukhov*	**Grell-3**
	Goddard	FLG*			**MYNN3 + MYNN**	**Tiedkte**
	Thompson*				BouLac + MM5*	**GF**
	Mibrandt				UW + MM5*	**New Tiedkte**
	Morrison				**UW + Monin–Obukhov**	
	WDM5				*GBM + MM5*	
	WDM6				Shin–Hong + MM5*	
	Thompson aerosol-aware*					
Number of schemes	11	5	4	4	10	7

probability distributions match with observations. The RPS values of the initial and final ensembles for all training and verification cases were compared and the results are shown in Fig. 10. We note a remarkable improvement in RPS values of the final ensemble over that of the initial ensemble for all lead times. This improvement is striking in that the RPS value is generally negatively biased for ensemble forecasts with small ensemble sizes (Buizza and Palmer, 1998). In this case, the RPS values for the final ensemble with an ensemble size of 40 are much smaller than that of the initial ensemble with an ensemble size of 90. Note that Figs. 7–10 show the same degree of improvement for both the training cases and the verification cases, indicating that the final ensemble

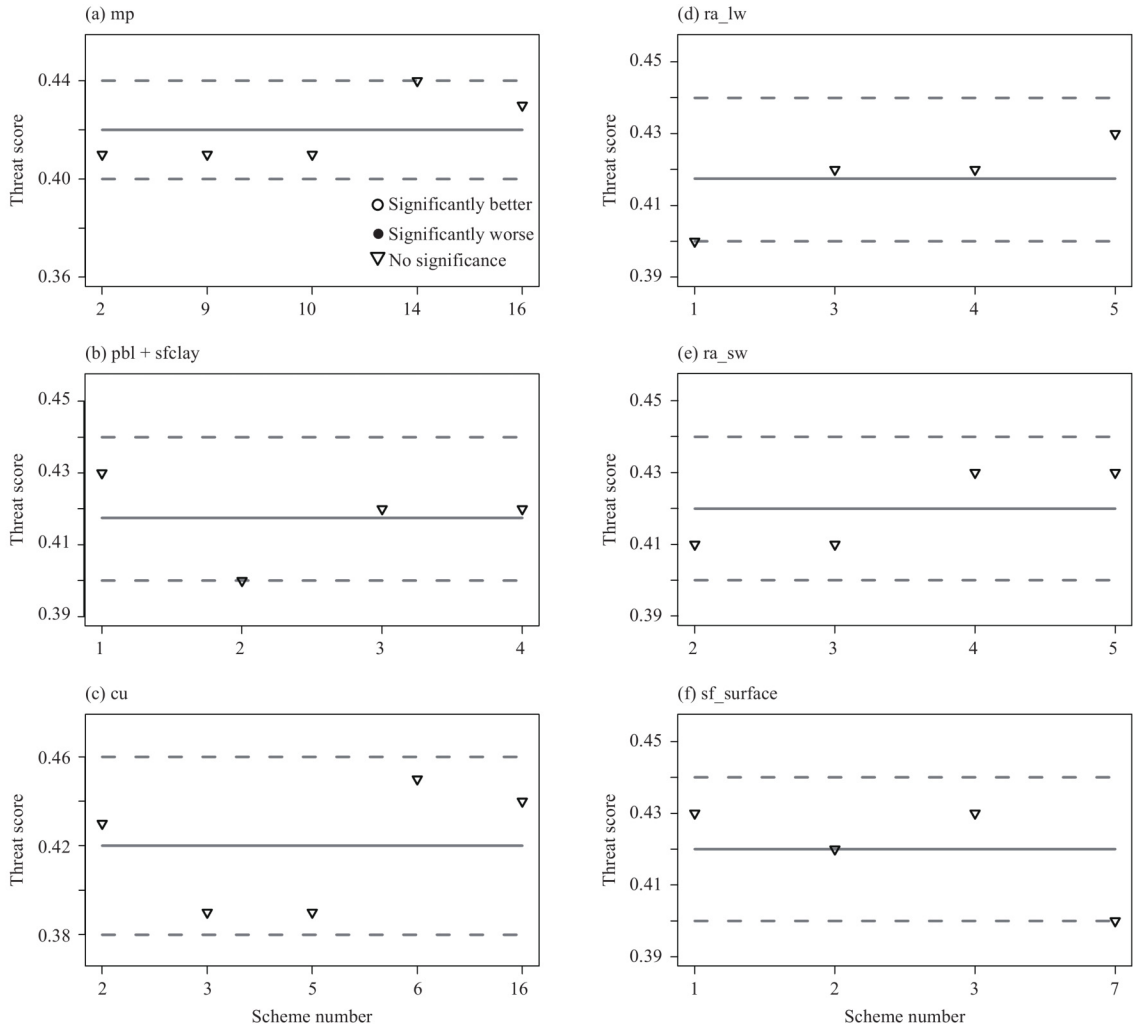

Fig. 6. As in Fig. 3, but for threat scores of different schemes over the training cases with a two-sided *t*-test in the fouth round of screening.

obtained through the screening procedure is effective and the results are transferable to other cases.

Table 4 exhibits the inter-comparison results based on the BSs and the BS decompositions (i.e., reliability, resolution, and uncertainty) for ensemble forecasts generated by the initial and final ensembles at different lead times for two categories of rainfall cases. Because probabilistic forecast verification usually requires a large sample size and the rainfall intensity > 10 mm day^{-1} only accounts for 13% (23,282 grid points) of the total observation grid points in the training and verification cases, we used only two categories of rain for probabilistic verification, with the threshold set at 10 mm day^{-1}. If the 24-h cumulative rain is greater than 10 mm, it is marked as "heavy rain;" otherwise, it is marked as "regular rain." Compared to the initial ensemble, the BSs of the final ensemble have smaller values for "regular rain." The difference is mainly due to the fact that the final ensemble has better reliability (i.e., the forecast probability of the final ensemble has a better agreement with the observed frequency). However, for "heavy rain" forecasts, the BSs of the final ensemble are not as good as those of the initial ensemble. This may be because "heavy rain" cases have fewer samples to obtain reliable statistics and are more difficult to predict (Zhang et al., 2006). The resolutions of the two ensembles are similar to each other, with the initial ensemble having a slight advantage. Therefore, both ensembles have similar ability to separate rain intensity from one category to another. The uncertainty of the two ensembles based on the BSs are not analyzed here because this metric is independent of the forecast quality and needs climatological information to compute (Ferro and Fricker, 2012). Note that the BSs for "heavy rain" forecasts are generally smaller than those for "regular rain" forecasts. We think this may be related to how the statistical method is formulated and does not necessarily mean that the probabilistic forecasts of "heavy rain" cases have a better performance than that for "regu-

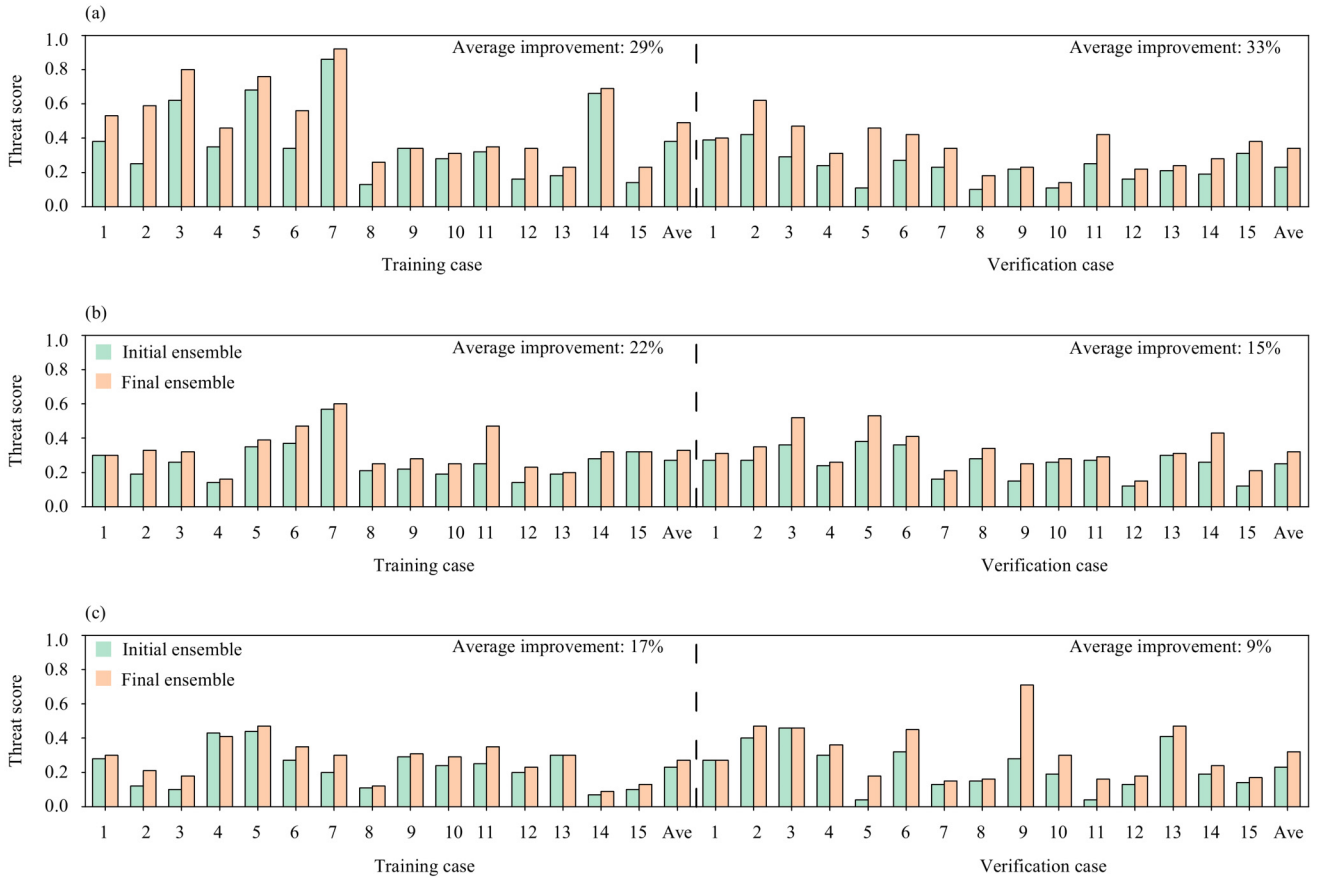

Fig. 7. Comparison of the threat scores of the initial (green bar) and final (orange bar) ensemble forcasts and their average (Ave) values for the training and verification cases (divided by the vertical dashed line) at (a) 24-, (b) 48-, and (c) 72-h forecast lead times.

lar rain" cases. In calculating the BS, the grid is fixed, but the range of the two rain categories varies. Because the range of "regular rain" is larger than that of "heavier rain," the BS value is larger for the rain with a larger range (Atger, 2004; Wang, 2005). This implies that the BS values for forecasts of different rain intensities are not fully comparable.

Figure 11 displays the ROC curves for the two categories of rain cases at different forecast lead times. The ROC curve provides information on the hit rates and false alarm rates expected from use of different probability thresholds and is very useful to discriminate the performance of two sets of ensemble forecasts. As Fig. 11 shows, both sets of ensemble forecasts have excellent skills for all cases, as the ROC curves are all above the diagonal line with ROC area (ROCA) above 0.5, a threshold delineating whether ensemble forecasts having skills or not. The forecasts generated by the initial ensemble have better skills than those by the final ensemble. The advantage of the initial ensemble over the final ensemble is expected, as ROCA is also dependent on ensemble size. Forecasts with a small ensemble size are at a disadvantage compared to forecasts with a large ensemble size. This is consistent with the conclusion from Pellerin et al. (2003) and Marsigli et al. (2005), where in their study, the bigger ensemble takes advantage when compared to smaller ensemble (Mason and Graham, 1999).

5.3 *Comparison of the final ensemble against randomly sampled ensembles*

We have shown that the ensemble forecasts from the final set of ensemble members have better forecast performance over the forecasts generated from the initial ensemble according to a number of verification metrics. One may argue that this advantage does not pass the statistical significance test because the final set of ensemble members represent only a particular set of scheme combinations that is better than the initial ensemble by chance. To validate the effectiveness of the final set of ensemble members, we compare the TS and RPS values of the forecasts from the final ensemble to the TS and RPS distributions of the randomly sampled 40 scheme combinations using bootstrapping method, which generates 1000 sets of 40-member ensemble from the initial ensemble members. The comparison results are shown in

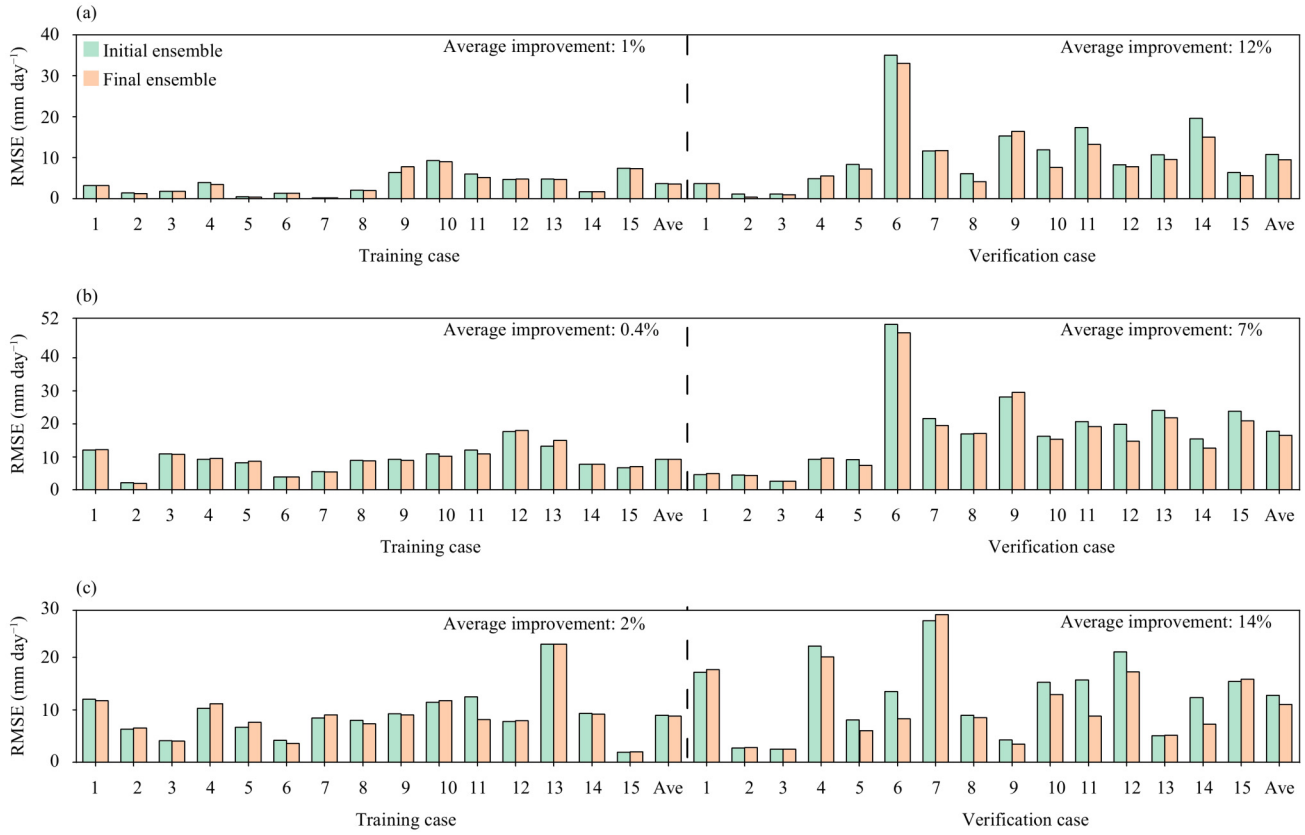

Fig. 8. As in Fig. 7, but for the RMSE of the initial and final ensemble forecasts for the training and verification cases.

violin plots in Fig. 12. Violin plots, which have a kernel density plot on each side, are similar to bar plots, and display the uncertainty due to sampling errors (Hintze and Nelson, 1998). The evaluation results show that there is little chance that the randomly sampled ensemble forecasts have a better performance than the final ensemble forecasts. The average TSs of the final ensemble are all higher than the TS ranges for randomly sampled ensembles. The results for RPS also show clearly that the final ensemble is more likely to generate better forecasts than the randomly sampled ensembles, with the RPS values of the final ensemble well below the median RPS values and only a slight chance of smaller RPS values for some randomly sampled ensembles. The fact that average TS and RPS values of the final ensemble are superior to those of the randomly sampled ensembles supports the argument that the screening process is effective in improving the performance of ensemble forecasts.

6. Conclusions

In this study, we proposed an objective statistical method to select which parameterization schemes should be included to form a multi-physics ensemble for short-range (3-day) summer precipitation forecast over the Greater Beijing area. The screening methodology centers on using statistical variances and significance *t*-tests to determine which schemes are significantly better or worse than the average performance and which schemes should be retained to ensure a large ensemble dispersion. After several rounds of screening, we obtained the final ensemble with 40 ensemble members, each representing a particular scheme combination from the WRF model. Then, we used a series of verification metrics, including TS, RMSE, BS, ROC, and RPS, to evaluate the ensemble forecasts.

The evaluation results of the TS and RMSE suggest that the forecasts from the final ensemble are superior to those of the initial ensemble for almost all cases and forecast lead times. The improvement of the average TS is 33%, 15%, and 9% for lead times of 24, 48, and 72 h, respectively. The average improvement in RMSE is significant for the verification cases at 12%, 7%, and 14%, respectively. The BS evaluation results indicate that the screening process has improved the BS and reliability for "regular rain." For the "heavy rain" cases, performance of the initial ensemble is better than that of the final ensemble, due to the number of "heavy rain" cases being smaller than "regular rain" cases. Hence, the results for "heavy rain" is not as reliable as for "regular rain." The

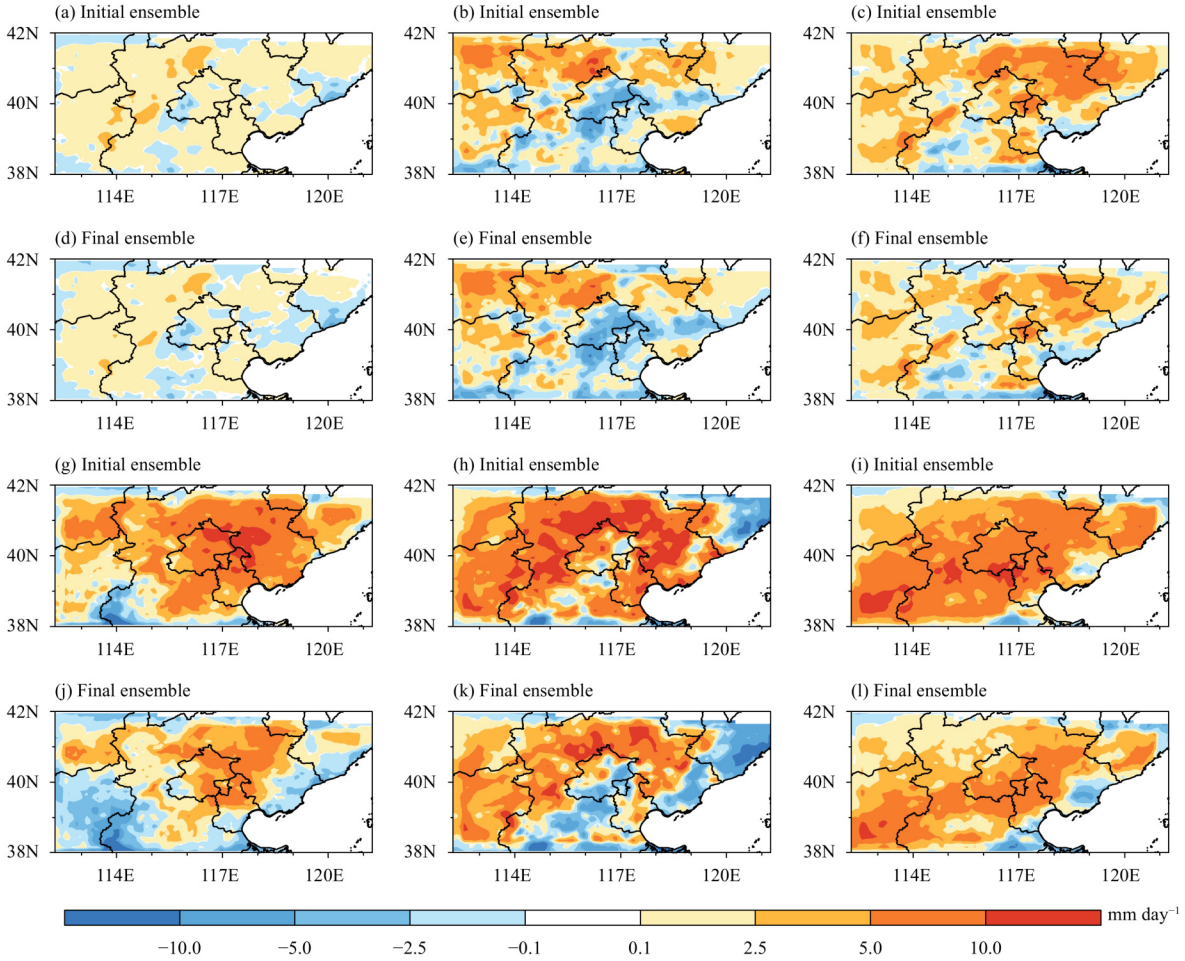

Fig. 9. The error distributions of the (a–c, g–i) initial and (d–f, j–l) final ensemble forcasts for the average of (a–f) training and (g–l) verification cases at (left panels) 24-, (middle panels) 48-, and (right panels) 72-h lead times.

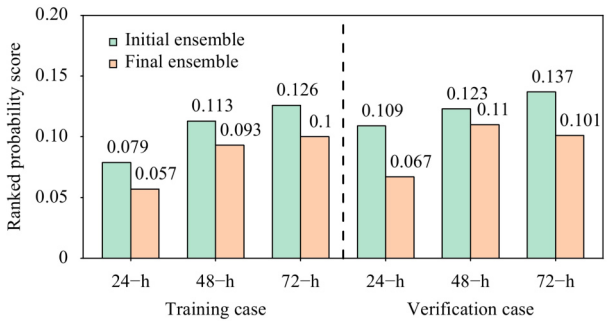

Fig. 10. Comparison of the average ranked probability scores of the initial (green bar) and final (orange bar) ensemble forcasts for the training and verification cases (divided by the vertical dashed line) at 24-, 48-, and 72-h lead times.

ROC evaluation results show that the two ensembles have similar resolutions. Finally, we compared the final ensemble against 1000 randomly sampled ensembles drawn from the initial ensemble to ensure statistical significance of the results. The violin plots show that the RPS and TS values of the final ensemble are statistically better than those of the random sampled ensembles in all cases. These results illustrate that the screening procedure proposed in this study is effective in generating ensemble forecasts with good performance from the WRF model.

In evaluating the ensemble forecasts, we performed the evaluation on both the training cases and the verification cases, so our method is generalizable to cases outside the training data. It has the potential for operational ensemble forecast applications (e.g., related to the data assimilation method or the stochastic perturbations schemes). Although our study only focused on short-range summer precipitation forecasting in the Greater Beijing area, the general procedure can be used for medium- to long-range forecasts, for other areas, and other forecast variables.

Appendix A: LHS design

The LHS is a method for sampling model input space. LHS design uses a stratified sampling scheme to im-

Table 4. Comparison of the Brier score (BS) values for different lead times and rain intensities for the training and verification cases. Reliability, resolution, and uncertainty are the decomposed values from the BS value. The numbers before and after "/" are the BS values for the initial and final ensembles, respectively

	Lead time	24 h		48 h		72 h	
	Rain category	⩽ 10 mm day^{-1}	> 10 mm day^{-1}	⩽ 10 mm day^{-1}	> 10 mm day^{-1}	⩽ 10 mm day^{-1}	> 10 mm day^{-1}
Training case	BS	0.055/0.034	0.031/0.033	0.165/0.160	0.145/0.158	0.153/0.126	0.114/0.123
	Reliability	0.027/0.005	0.003/0.004	0.027/0.021	0.006/0.019	0.047/0.019	0.008/0.016
	Resolution	0.006/0.005	0.006/0.005	0.019/0.018	0.018/0.017	0.010/0.010	0.010/0.009
	Uncertainty	0.034/0.034	0.034/0.034	0.156/0.156	0.156/0.156	0.116/0.116	0.116/0.116
Verification case	BS	0.128/0.122	0.110/0.121	0.165/0.168	0.155/0.167	0.160/0.134	0.113/0.131
	Reliability	0.022/0.010	0.003/0.009	0.018/0.010	0.005/0.009	0.050/0.023	0.002/0.019
	Resolution	0.034/0.028	0.034/0.028	0.068/0.057	0.065/0.057	0.024/0.023	0.023/0.022
	Uncertainty	0.141	0.141	0.215	0.215	0.134	0.134

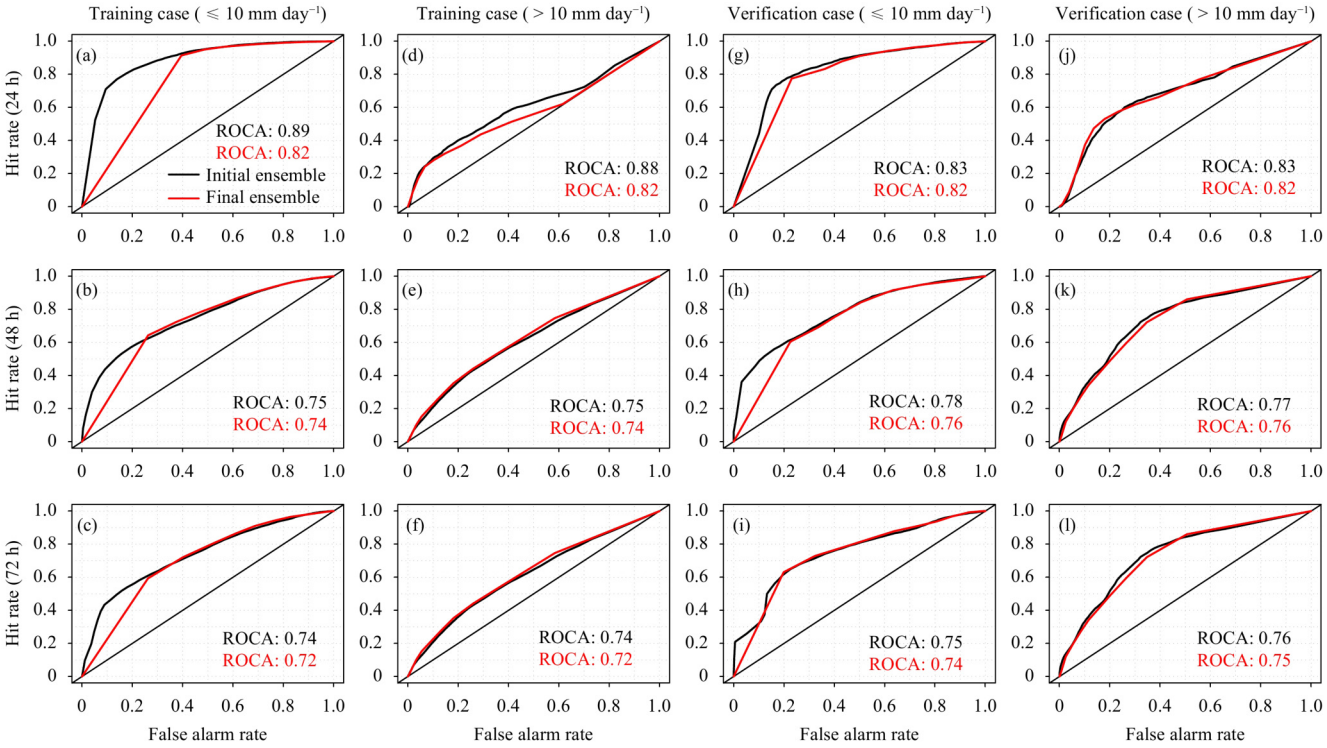

Fig. 11. Comparison of the relative operating characteristics (ROC) curve and ROC area (ROCA) between the initial (black line) and final (red line) ensemble forcasts for the (a–f) training and (g–l) verification cases of two rain intensities at (a, d, g, j) 24-, (b, e, h, k) 48-, and (c, f, i, l) 72-h lead times.

prove the coverage of input space (Loh, 1996; McKay et al., 2000). Compared with the Monte Carlo sampling method, the LHS samples can more evenly across all possible values. Assumed that n_0 samples are needed and n is the dimension of the sample space, $\boldsymbol{x_i}$ is the i-th dimension in the space. Dividing dimension $\boldsymbol{x_i}$ into n_0 intervals and each interval has the same probability. Then, the sample space is divided into $n_0{}^n$ small hypercube, which can be expressed as an array $\boldsymbol{U}$ with n_0 rows and n columns ($n_0 \times n$). Each row of $\boldsymbol{U}$-array corresponds to a small hypercube. Randomly selecting one sample from each small hypercube will get final n_0 samples (Helton and Davis, 2003). Figure A1 gives an illustration of LHS design for a two-dimensional variable.

Appendix B: Description of the one-sided and two-sided *t*-tests

The *t*-test, also known as the Student's test, is one of the most commonly used tests to determine whether the means of two groups of samples are different significantly. In this study, we used both the one-sided and two-sided *t*-tests. The Student's *t* distribution is defined as:

$$t = \frac{\mu_{i,j} - \overline{\mu_j}}{\sigma_j / K_j}, \tag{B1}$$

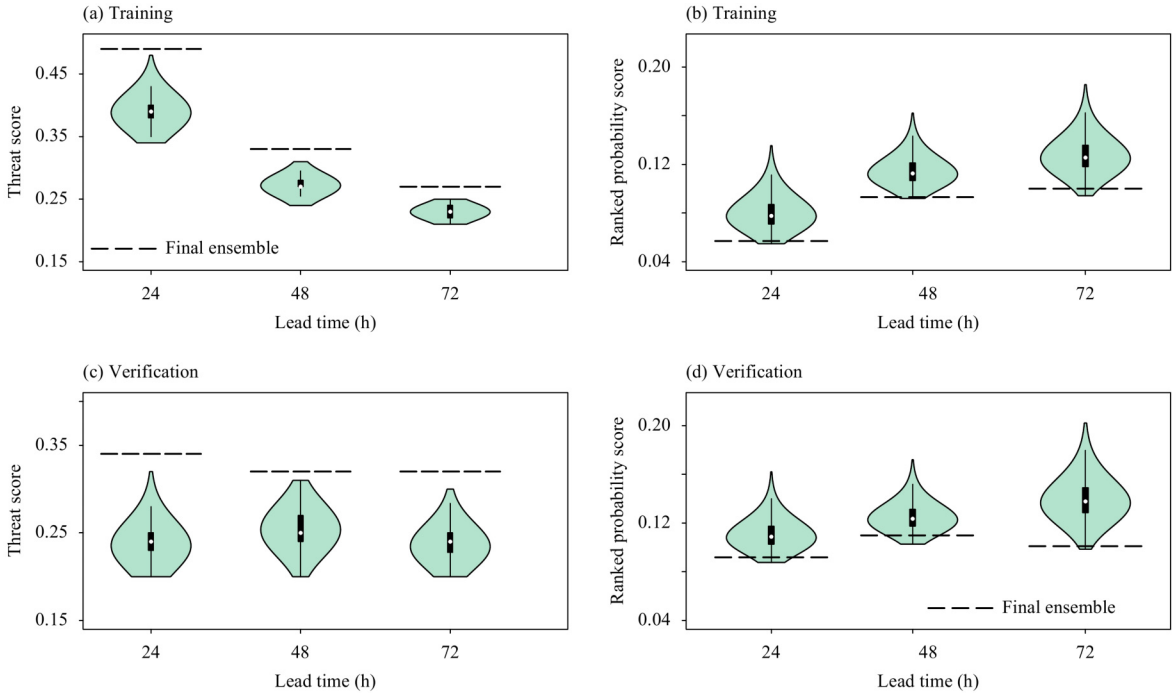

Fig. 12. Violin plots of (a, c) threat score (TS) and (b, d) ranked probability score (RPS) distributions from the ensembles randomly sampled from the initial 90 scheme combinations and the average TS and RPS values from the final ensemble (dashed lines) for the (a, b) training and (c, d) verification cases at 24-, 48-, 72-h lead times.

where $\mu_{i,j}$ is the average performance metrics of all parameterization scheme combinations in which scheme i of the physical process j appears, $\overline{\mu_j} = \sum_{i=1}^{K_j} \frac{\mu_{i,j}}{K_j}$ is the mean performance metrics of all schemes, K_j corresponds to the number of feasible schemes in physical process j, and $\sigma_j = \sqrt{\sum_{i=1}^{K_j} \frac{\left(\mu_{i,j} - \overline{\mu_j}\right)^2}{K_j}}$ is the standard deviation. For a two-sided test, the null hypothesis of our experiments is $\mu_{i,j} = \overline{\mu_j}$ and the alternative hypothesis is $\mu_{i,j} \neq \overline{\mu_j}$. To evaluate the statistical significance of the two-sided t-test, we need to calculate the confidence interval at a specific level, say 95%, and then determine whether the $\mu_{i,j}$ is significantly greater or less than $\overline{\mu_j}$. For the one-sided t-test, the null hypothesis of our experiments is $\mu_{i,j} < \overline{\mu_j}$ and the alternative hypothesis is $\mu_{i,j} \geqslant \overline{\mu_j}$. Here, we need to compute the confidence limit at a specific level, and determine whether the $\mu_{i,j}$ is significantly less than $\overline{\mu_j}$.

Appendix C: Verification metrics

C1. TS

TS measures the fraction of forecasts corresponding to the observations correctly. It is defined as,

$$\mathrm{TS} = a/(a+b+c), \tag{C1}$$

where a, b, and c represent the hits, false alarms, and misses in the contingency table, respectively (Table C1).

The TS value averaged over different thresholds is used to evaluate the performance of precipitation forecast:

$$\mathrm{TS_w} = \sum_{i=1}^{m} w_i(\mathrm{TS})_i, \tag{C2}$$

where m is the number of thresholds used for categorized precipitation events and weight $w_i = g_i/G$, g_i is the number of grids for threshold i, and G is the total number of grids for all thresholds (Zhao and Carr, 1997).

C2. BS

BS measures the mean squared probability error for a binary event (Brier, 1950). It can be decomposed into three terms:

$$\mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)^2 = \frac{1}{N}\sum_{k=1}^{K} n_k\left(p_k - \bar{O}_k\right)^2 - \frac{1}{N}\sum_{k=1}^{K} n_k\left(\bar{O}_k - \bar{O}\right)^2 + \bar{O}\left(1 - \bar{O}\right). \tag{C3}$$

The three terms on the right-hand side of the above equation denote reliability, resolution, and uncertainty of the forecasts, respectively. In Eq. (C3), P_i is the occurrence probability of rain exceeding a certain threshold for a particular event: if the event occurs, $O_i = 1$; otherwise, $O_i = 0$. Moreover, K is the number of forecast probability categories; p_k is the mean forecast probability in kth category; $\bar{O}_k$ is the mean observation frequency in kth

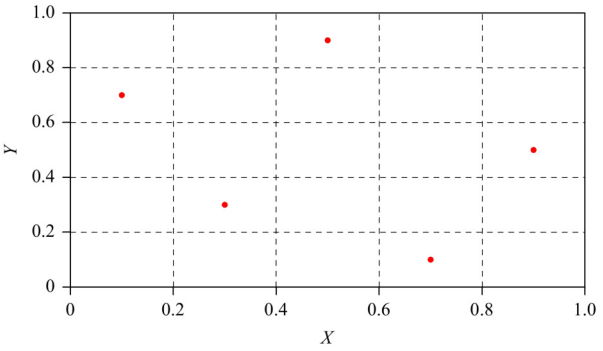

Fig. A1. Illustration of the LHS process for a two-dimensional variable.

category; and $\bar{O}$ is the mean of all observation frequencies. The range for BS is 0–1. It is a negatively oriented score, with the perfect score being 0.

C3. ROC

ROC displays the hit rate versus the false alarm rate, based on a set of increasing probability thresholds to make the "yes/no" decision (Stanski et al., 1989). The specific measure associated with ROC is the area (ROCA) under the curve, which is a measure of resolution. The range of the area is 0–1, where 1 means a perfect ensemble system. A value of 0.5 (diagonal line) means a useless forecast system because it cannot discriminate between occurrence and non-occurrence of an event.

C4. RPS

RPS measures the sum of squared differences in cumulative probability space for a multi-category precipitation probabilistic forecast, which is given by

$$\mathrm{RPS} = \left(\sum_{m=1}^{J} \left\langle \sum_{i=1}^{m} P_i - \sum_{i=1}^{m} O_i \right\rangle^2 \right) / (J-1), \qquad \text{(C4)}$$

where J is the number of forecast categories; P_i is the predicted probability in forecast category i; and O_i is a binary indicator (i.e., 0 = no, 1 = yes) for the observation in category i. RPS penalizes forecasts more severely when their probabilities are farther away from the actual outcome (Murphy, 1969, 1971).

Acknowledgments. Special thanks go to the group of Prof. Shiguang Miao at the Institute of Urban Meteorology, China Meteorological Administration for offering help with the WRF simulations and data collection.

Table C1. Contingency table for computing TS. Symbols *a, b, c,* and *d* represent hits, false alarms, misses, and correct negatives, respectively; and *n* is the number of total events

Event forecast	Event observed		
	Yes	No	Marginal total
Yes	*a*	*b*	*a*+*b*
No	*c*	*d*	*c*+*d*
Marginal total	*a*+*c*	*b*+*d*	*a*+*b*+*c*+*d* = *n*

REFERENCES

Atger, F., 2004: Relative impact of model quality and ensemble deficiencies on the performance of ensemble based probabilistic forecasts evaluated through the Brier score. *Nonlin. Processes Geophys.*, **11**, 399–409, doi: 10.5194/npg-11-399-2004.

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

Berner, J., U. Achatz, L. Batté, et al., 2017: Stochastic parameterization: Toward a new view of weather and climate models. *Bull. Amer. Meteor. Soc.*, **98**, 565–588, doi: 10.1175/BAMS-D-15-00268.1.

Bougeault, P., Z. Toth, C. Bishop, et al., 2010: The THORPEX interactive grand global ensemble. *Bull. Amer. Meteor. Soc.*, **91**, 1059–1072, doi: 10.1175/2010BAMS2853.1.

Bowler, N. E., A. Arribas, K. R. Mylne, et al., 2008: The MOGREPS short-range ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **134**, 703–722, doi: 10.1002/qj.234.

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

Buizza, R., and T. N. Palmer, 1998: Impact of ensemble size on ensemble prediction. *Mon. Wea. Rev.*, **126**, 2503–2518, doi: 10.1175/1520-0493(1998)126<2503:IOESOE>2.0.CO;2.

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

Burgers, G., P. J. van Leeuwen, and G. Evensen, 1998: Analysis scheme in the ensemble Kalman filter. *Mon. Wea. Rev.*, **126**, 1719–1724, doi: 10.1175/1520-0493(1998)126<1719:ASITEK>2.0.CO;2.

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

Crétat, J., B. Pohl, Y, Richard, et al., 2012: Uncertainties in simulating regional climate of Southern Africa: Sensitivity to physical parameterizations using WRF. *Climate Dyn.*, **38**, 613–634, doi: 10.1007/s00382-011-1055-8.

Di, Z. H., Q. Y. Duan, W. Gong, et al., 2015: Assessing WRF model parameter sensitivity: A case study with 5-day summer precipitation forecasting in the Greater Beijing area. *Geophys. Res. Lett.*, **42**, 579–587, doi: 10.1002/2014GL061623.

Di, Z. H., Q. Y. Duan, C. Wang, et al., 2018: Assessing the applicability of WRF optimal parameters under the different precipitation simulations in the Greater Beijing area. *Climate Dyn.*, **50**, 1927–1948, doi: 10.1007/s00382-017-3729-3.

Du, J., 2002: Present situation and prospects of ensemble numerical prediction. *J. Appl. Meteor. Sci.*, **13**, 16–28, doi: 10.3969/j.issn.1001-7313.2002.01.002. (in Chinese)

Du, J., and W. H. Qian, 2014: Three revolutions in weather forecasting. *Adv. Meteor. Sci. Technol.*, **4**, 13–26. (in Chinese)

Du, J., and J. Li, 2014: Application of ensemble methodology to heavy-rain research and prediction. *Adv. Meteor. Sci. Technol.*, **4**, 6–20. (in Chinese)

Ebert, E. E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.*, **129**, 2461–2480, doi: 10.1175/1520-0493(2001)129<2461:AOAPMS>2.0.CO;2.

Efstathiou, G. A., N. M. Zoumakis, D. Melas, et al., 2013: Sensitivity of WRF to boundary layer parameterizations in simulating a heavy rainfall event using different microphysical schemes. Effect on large-scale processes. *Atmos. Res.*, **132–133**, 125–143, doi: 10.1016/j.atmosres.2013.05.004.

Ferro, C. A. T., and T. E. Fricker, 2012: A bias-corrected decomposition of the Brier score. *Quart. J. Roy. Meteor. Soc.*, **138**, 1954–1960, doi: 10.1002/qj.1924.

Gou, J. J., C. Y. Miao, Q. Y. Duan, et al., 2020: Sensitivity analysis-based automatic parameter calibration of the VIC model for streamflow simulations over China. *Water Resour. Res.*, **56**, e2019WR025968, doi: 10.1029/2019WR025968.

Helton, J. C., and F. J. Davis, 2003: Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. *Reliab. Eng. Syst. Saf.*, **81**, 23–69, doi: 10.1016/S0951-8320(03)00058-9.

Hintze, J. L., and R. D. Nelson, 1998: Violin plots: A box plot-density trace synergism. *Am. Stat.*, **52**, 181–184, doi: 10.2307/2685478.

Houtekamer, P. L., and H. L. Mitchell, 1998: Data assimilation using an ensemble Kalman filter technique. *Mon. Wea. Rev.*, **126**, 796–811, doi: 10.1175/1520-0493(1998)126<0796:DAUAEK>2.0.CO;2.

Houtekamer, P. L., and F. Q. Zhang, 2016: Review of the ensemble Kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.*, **144**, 4489–4532, doi: 10.1175/MWR-D-15-0440.1.

Irvine, P. J., L. J. Gregoire, D. J. Lunt, et al., 2013: An efficient method to generate a perturbed parameter ensemble of a fully coupled AOGCM without flux-adjustment. *Geosci. Model Dev.*, **6**, 1447–1462, doi: 10.5194/gmd-6-1447-2013.

Jiang, X. M., H. L. Yuan, M. Xue, et al., 2014: Analysis of a heavy rainfall event over Beijing during 21–22 July 2012 based on high resolution model analyses and forecasts. *J. Meteor. Res.*, **28**, 199–212, doi: 10.1007/s13351-014-3139-y.

Joyce, R. J., J. E. Janowiak, P. A. Arkin, et al., 2004: CMORPH: A method that produces global precipitation estimates from passive microwave and infrared data at high spatial and temporal resolution. *J. Hydrometeor.*, **5**, 487–503, doi: 10.1175/1525-7541(2004)005<0487:CAMTPG>2.0.CO;2.

Kober, K., and G. C. Craig, 2016: Physically based stochastic perturbations (PSP) in the boundary layer to represent uncertainty in convective initiation. *J. Atmos. Sci.*, **73**, 2893–2911, doi: 10.1175/JAS-D-15-0144.1.

Krishnamurti, T. N., C. M. Kishtawal, T. E. LaRow, et al., 1999: Improved weather and seasonal climate forecasts from multimodel superensemble. *Science*, **285**, 1548–1550, doi: 10.1126/science.285.5433.1548.

Lee, J. A., 2012: Techniques for down-selecting numerical weather prediction ensembles. Ph.D. dissertation, Dept. of Meteorology, Pennsylvania State Unversity, USA, 126 pp.

Lee, J. A., W. C. Kolczynski, T. C. McCandless, et al., 2011: An objective methodology for configuring and down-selecting an NWP ensemble for low-level wind prediction. *Mon. Wea. Rev.*, **140**, 2270–2286, doi: 10.1175/MWR-D-11-00065.1.

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.*, **102**, 409–418, doi: 10.1175/1520-0493(1974)102<0409:TSOMCF>2.0.CO;2.

Loh, W. -L., 1996: On Latin hypercube sampling. *Ann. Stat.*, **24**, 2058–2080, doi: 10.1214/aos/1069362310.

Lorenz, E. N., 1965: A study of the predictability of a 28-variable atmospheric model. *Tellus*, **17**, 321–333, doi: 10.3402/tellusa.v17i3.9076.

Marsigli, C., F. Boccanera, A. Montani, et al., 2005: The COSMO-LEPS mesoscale ensemble system: Validation of the methodology and verification. *Nonlin. Processes Geophys.*, **12**, 527–536, doi: 10.5194/npg-12-527-2005.

Mason, S. J., and N. E. Graham, 1999: Conditional probabilities, relative operating characteristics, and relative operating levels. *Wea. Forecasting*, **14**, 713–725, doi: 10.1175/1520-0434(1999)014<0713:CPROCA>2.0.CO;2.

McCabe, A., R. Swinbank, W. Tennant, et al., 2016: Representing model uncertainty in the Met Office convection-permitting ensemble prediction system and its impact on fog forecasting. *Quart. J. Roy. Meteor. Soc.*, **142**, 2897–2910, doi: 10.1002/qj.2876.

McKay, M. D., R. J. Beckman, and W. J. Conover, 2000: A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, **42**, 55–61, doi: 10.2307/1271432.

Miao, C. Y., Q. Y. Duan, Q. H. Sun, et al., 2019: Non-uniform changes in different categories of precipitation intensity across China and the associated large-scale circulations. *Environ. Res. Lett.*, **14**, 025004, doi: 10.1088/1748-9326/aaf306.

Molteni, F., R. Buizza, T. N. Palmer, et al., 1996: The ECMWF ensemble prediction system: Methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122**, 73–119, doi: 10.1002/qj.49712252905.

Murphy, A. H., 1969: On the "ranked probability score". *J. Appl. Meteor.*, **8**, 988–989, doi: 10.1175/1520-0450(1969)008<0988:OTPS>2.0.CO;2.

Murphy, A. H., 1971: A note on the ranked probability score. *J. Appl. Meteor.*, **10**, 155–156, doi: 10.1175/1520-0450(1971)010<0155:ANOTRP>2.0.CO;2.

Murphy, J. M., B. B. B. Booth, C. A. Boulton, et al., 2014: Transient climate changes in a perturbed parameter ensemble of emissions-driven earth system model simulations. *Climate Dyn.*, **43**, 2855–2885, doi: 10.1007/s00382-014-2097-5.

Palmer, T. N., G. J. Shutts, R. Hagedorn, et al., 2005: Representing model uncertainty in weather and climate prediction. *Annu. Rev. Earth Planet. Sci.*, **33**, 163–193, doi: 10.1146/annurev.earth.33.092203.122552.

Pellerin, G., L. Lefaivre, P. Houtekamer, et al., 2003: Increasing the horizontal resolution of ensemble forecasts at CMC. *Nonlin. Processes Geophys.*, **10**, 463–468, doi: 10.5194/npg-10-463-2003.

Quan, J. P., Z. H. Di, Q. Y. Duan, et al., 2016: An evaluation of parametric sensitivities of different meteorological variables simulated by the WRF model. *Quart. J. Roy. Meteor. Soc.*, **142**, 2925–2934, doi: 10.1002/qj.2885.

Raftery, A. E., T. Gneiting, F. Balabdaoui, et al., 2005: Using Bayesian model averaging to calibrate forecast ensembles.

Mon. Wea. Rev., **133**, 1155–1174, doi: 10.1175/MWR2906.1.

Sanchez, C., K. D. Williams, and M. Collins, 2016: Improved stochastic physics schemes for global weather and climate models. *Quart. J. Roy. Meteor. Soc.*, **142**, 147–159, doi: 10.1002/qj.2640.

Shen, Y., P. Zhao, Y. Pan, et al., 2014: A high spatiotemporal gauge–satellite merged precipitation analysis over China. *J. Geophys. Res. Atmos.*, **119**, 3063–3075, doi: 10.1002/2013JD020686.

Shutts, G., 2015: A stochastic convective backscatter scheme for use in ensemble prediction systems. *Quart. J. Roy. Meteor. Soc.*, **141**, 2602–2616, doi: 10.1002/qj.2547.

Skamarock, W. C., J. B. Klemp, J. Dudhia, et al., 2008: A Description of the Advanced Research WRF Version 3. No. NCAR/TN-475+STR, University Corporation for Atmospheric Research, Boulder, Colorado, USA, 113 pp, doi: 10.5065/D68S4MVH.

Stanski, H. R., L. J. Wilson, and W. R. Burrows, 1989: Survey of Common Verification Methods in Meteorology. World Weather Watch Technical Report No. 8, TD No. 358, World Meteorological Organization, Geneva, 114 pp.

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

Sun, Q. H., C. Y. Miao, A. AghaKouchak, et al., 2020: Possible increased frequency of ENSO-related dry and wet conditions over some major watersheds in a warming climate. *Bull. Amer. Meteor. Soc.* doi: 10.1175/BAMS-D-18-0258.1.

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

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

Tribbia, J. J., and D. P. Baumhefner, 1988: The reliability of improvements in deterministic short-range forecasts in the presence of initial state and modeling deficiencies. *Mon. Wea. Rev.*, **116**, 2276–2288, doi: 10.1175/1520-0493(1988)116<2276:TROIID>2.0.CO;2.

Wang, C. X., 2005: Experiments of short-range ensemble precipitation probability forecasts. *J. Appl. Meteor. Sci.*, **16**, 78–88, doi: 10.3969/j.issn.1001-7313.2005.01.009. (in Chinese)

Weusthoff, T., D. Leuenberger, C. Keil, et al., 2011: Best member selection for convective-scale ensembles. *Meteorol. Z.*, **20**, 153–164, doi: 10.1127/0941-2948/2011/0211.

WRF3.7.1, 2016: ARW Version 3 Modeling System User's Guide, 408 pp. Available at https://www2.mmm.ucar.edu/wrf/users/docs/user_guide_V3.7/ARWUsersGuideV3.7.pdf. Accessed on 19 June 2020.

Zhang, F. Q., A. M. Odins, and J. W. Nielsen-Gammon, 2006: Mesoscale predictability of an extreme warm-season precipitation event. *Wea. Forecasting*, **21**, 149–166, doi: 10.1175/WAF909.1.

Zhao, Q. Y., and F. H. Carr, 1997: A prognostic cloud scheme for operational NWP models. *Mon. Wea. Rev.*, **125**, 1931–1953, doi: 10.1175/1520-0493(1997)125<1931:APCSFO>2.0.CO;2.

Zheng, H. Y., C. Y. Miao, J. W. Wu, et al., 2019: Temporal and spatial variations in water discharge and sediment load on the Loess Plateau, China: A high-density study. *Sci. Total Environ.*, **666**, 875–886, doi: 10.1016/j.scitotenv.2019.02.246.

Tech & Copy Editor: Hongqun ZHANG