Multi-Model Ensemble Forecast of Precipitation Based on an Object-Based Diagnostic Evaluation

Luying Ji [1,2], Xiefei Zhi *[1,3], Clemens Simmer[2], Shoupeng Zhu [1], Yan Ji [1,]

[1] *Key Laboratory of Meteorological Disasters, Ministry of Education (KLME) / Collaborative Innovation Center on Forecast and Evaluation of Meteorological Disasters (CIC-FEMD), Nanjing University of Information Science & Technology, Nanjing, China*

[2] *Institute for Geosciences – Section Meteorology, University of Bonn, Bonn, Germany*

[3]*Nanjing Joint Center for Atmospheric Research (NJCAR), Nanjing, China*

*Corresponding author: Xiefei ZHI

Email: zhi@nuist.edu.cn

Early Online Release: This preliminary version has been accepted for publication in *Monthly Weather Review*, may be fully cited, and has been assigned DOI 10.1175/MWR-D-19-0266.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

ABSTRACT

We analyzed 24-hour accumulated precipitation forecasts over the four-months period from 1 May to 31 August 2013 over an area located in East Asia covering the region 70.15°E–139.95°E, 15.05°N–58.95°N generated with the Ensemble Prediction Systems (EPSs) from ECMWF, NCEP, UKMO, JMA and CMA contained in the TIGGE dataset. The forecasts are first evaluated with the Method for Object-based Diagnostic Evaluation (MODE). Then a multi-model ensemble (MME) forecast technique based on weights derived from object-based scores is investigated and compared with the equally-weighted MME and the traditional gridpoint-based MME forecast using weights derived from the point-to-point metric, mean absolute error (MAE).

The object-based evaluation revealed that attributes of objects derived from the ensemble members of the five individual EPS forecasts and the observations differ consistently. For instance, their predicted centroid location is more southwestward, their shape is more circular, and their orientation is more meridional than in the observations. The sensitivity of the number of objects and their attributes to methodological parameters is also investigated.

A MME prediction technique based on weights computed from the object-based scores, Median of Maximum Interest (MMI) and Object-based Threat Score (OTS), is explored and the results compared to the ensemble forecasts of the individual EPS, the equally-weighted MME forecast, and the traditional super-ensemble forecast. When

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

using MODE statistics for the forecast evaluation, the object-based MME prediction outperforms all other predictions. This is mainly because of a better prediction of the objects' centroid locations. When using the precipitation-based fractions skill score (FSS), which is not used in either of the weighted MME forecasts, the object-based MME forecasts are slightly better than the equally-weighted MME forecasts but inferior to the traditional super-ensemble forecast based on weights derived from the point-to-point metric, MAE.

Key words: The Method for Object-based Diagnostic Evaluation (MODE), 24-h accumulated precipitation, multi-model ensemble forecasts

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

1. Introduction

In the past two decades, numerical weather forecasting rapidly developed and - besides model improvements - evolved from traditional single deterministic forecasts to ensemble forecasting (Gneiting and Raftery 2005; Bauer et al. 2015). Different forecast systems differ in their overall architecture, spatial resolution, choice of initial conditions, data assimilation technology, and physical parameterization schemes used in the numerical models. Multi-model ensemble (MME) forecasting is an effective way to make use of the forecasts from different Ensemble Prediction Systems (EPS) with the goal being to reduce systematic deviations from observations and thus improve the overall prediction skill. Based on The Observing System Research and Predictability Experiment (THORPEX) program, which provides forecasts from different operational numerical weather prediction (NWP) centers, MME forecasting is currently already widely used. Many studies have shown that the MME forecast performance is superior to the forecast of an individual (one-model-based) EPS (Krishnamurti et al. 1999; Fraley et al. 2010; Zhi et al. 2012; Zhang et al. 2015; He et al. 2015; Ji et al. 2019).

Besides the equally-weighted MME, more complex MME methods, such as linear regression (Krishnamurti et al. 1999; 2000), Bayesian model averaging (BMA; Raftery et al. 2005; Vrugt et al. 2006), ensemble MOS (EMOS; Scheuerer 2014; Scheuerer and Hamill 2015) and artificial neural networks (Yuan et al. 2007; Bakhshaii and Stull 2009) have been proposed and are already widely used for precipitation forecasting (Tebaldi et al. 2004; Ke et al. 2008), typhoon forecasts (Kumar et al. 2003; Jordan et

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

al. 2008), and regional climate predictions (Kharin and Zwiers 2002; Yun et al. 2005). Many studies suggest that unequally-weighted MME forecasts can achieve better skill than equally-weighted ones (Chen et al. 2010; Zhang and Zhi 2015; Kim and Chan 2018). Peng et al. (2002) and Ke et al. (2009) show, however, that they are not always better and may even be worse than the best individual EPS forecast.

Unequally-weighted MME methods often determine the weight of each contributing EPS by their relative performance during a training period, which assumes a certain temporal stability of their forecast performance. Most methods use scores derived from point-to-point comparisons between forecasts and observations, e.g. the weighted ensemble mean (WEMN, Nohara et al. 2006), the bias-removed ensemble mean (BREM, Kharin and Zwiers 2002), and the super-ensemble (SUP, Krishnamurti et al. 2000), which uses the mean absolute error (MAE) during a training period.

However, point-to-point verification scores (e.g. MAE or Equitable Threat Score (ETS)) provide only limited information about the quality of a precipitation forecast because they only compare the observations and predictions point by point without taking, for example, the resemblance of spatial patterns into account (Mass et al. 2002; Baldwin and Kain 2006; Gilleland et al. 2009). Precipitation is highly discontinuous in space and time. Thus, even almost perfect forecasts of e.g. the shapes and sizes of precipitation systems may lead to poor point-to-point scores because of many false alarms and misses known as “double penalty’’ already at small spatial deviations. However, the correct prediction of spatial features like shape, size and approximate

Accepted for publication in *Monthly Weather Review*. DOI10.1175/MWR-D-19-0266.1.

location of extended precipitation fields are important because they can be used as a valuable guidance to improve forecasts, especially of extreme weather.

Several methods have been used to get around the limitations of point by point verification, which categorize into filtering and displacement methods. Filtering methods generally apply smoothing or scale separation to evaluate the forecast for different spatial scales (Marsigli et al. 2006; Roberts and Lean 2008; Ebert 2008; 2009; Casati et al. 2004; 2009; Zepeda-Arce et al. 2000; Harris et al. 2001; Mittermaier 2006; Marzban and Sandgathe 2009), while displacement methods identify discrete features or objects in the forecast and the observations and quantify their respective displacements in terms of location or other attributes (Ebert and McBride 2000; Baldwin and Lakshmivarahan 2003; Keil and Craig 2007; Marzban and Sandgathe 2008; Gilleland et al. 2010).

The Method for Objected-based Diagnostic Evaluation (MODE) developed by Davis et al. (2006a) is adopted for calculating verification scores in this study. MODE is a typical feature-based displacement approach and an example for a spatial diagnostic technique. MODE attempts to mimic the way a human would subjectively evaluate a forecast via setting a precipitation threshold and spatially convoluting (scale-dependent averaging) the precipitation field. The median of maximum interest (MMI; Davis et al. 2009) and the object-based threat score (OTS; Johnson et al. 2011a) are two scores calculated from the attributes of the detected objects in the forecast and in the

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

observations (for more detail see Section 2.3), which are sensitive to different aspects of forecast accuracy (Johnson and Wang 2013).

Although MODE has been most commonly used for the verification of high-resolution model forecasts of convective storms, it can also be applied to lower-resolution numerical model weather forecasts, regional climate simulations, or chemistry model simulations (e.g. Brown et al. 2007; Wolff et al. 2014; Li et al. 2015). Twenty-four hour accumulated precipitation over areas of hundreds of kilometers exhibits characteristic spatial patterns that should be reproduced by model forecasts. Object-based methods allow us to evaluate whether this is indeed the case. In this study, we focus on daily precipitation forecasts with lead times of 1–7 days and venture to improve their prediction skills of shape, size and/or location by a new MME approach, which employs weights derived from object-based scores. We will compare its quality with the predictions achieved by the individual EPS, by an equally-weighted MME forecast, and by an MME forecast using weights based on point-to-point metric, MAE, derived from the precipitation forecasts and the observations during a training period.

First, we evaluate ensemble forecasts of 24-h accumulated precipitation produced by the five ensemble prediction systems (EPSs, i.e. the European Centre for Medium-Range Weather Forecasts (ECMWF), the National Center for Environment Prediction (NCEP), the UK Met Office (UKMO), the Japan Meteorological Agency (JMA), and the China Meteorological Administration (CMA)). For each ensemble member forecast of each individual EPS, MODE is used to obtain the attributes of every identified object.

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

Various attributes of one identified object are compared to the attributes of the best corresponding object in the observations; then the performance of each EPS is represented by the object attribute differences averaged over all objects identified for each ensemble member of an individual contributing EPS.

Second, three MME predictions are computed and their forecast accuracy evaluated using the spatial object-based measures MMI and OTS, but also using the FSS as an independent skill score that was not used for calculating weights in any of the three MME forecasts. We compare the three MME techniques in order to investigate if the MME forecast with object-based weights provides more accurate spatial information in the precipitation forecast.

The remainder of this paper is structured as follows. Section 2 briefly describes the datasets that were used and introduces MODE. In section 3, we present the performance evaluation of the five individual EPSs and of the three MME precipitation forecasting methods. A discussion and major conclusions are provided in section 4.

2. Data and methods

2.1 Data

We used 24-h accumulated precipitation ensemble forecasts produced by ECMWF, NCEP, UKMO, JMA and CMA at 0.5° x 0.5° resolution initialized daily at 1200 UTC for lead times of 1–7 days (Table 1). The data is available from the TIGGE-ECMWF portal (http://apps.ecmwf.int/datasets/data/tigge). TIGGE (The THORPEX Interactive

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

Grand Global Ensemble) is a key component of the THORPEX program; it contains ensemble forecast data from 10 global model prediction centers and has been widely used for scientific research on ensemble forecasting, predictability and the development of products to improve the prediction of severe weather (Breivik et al. 2014; Loeser et al. 2017; Parsons et al. 2017). We analyzed the data for a four-months period from 1 May to 31 August 2013 and over an area located in East Asia covering the region 70.15°E–139.95°E, 15.05°N–58.95°N.

For forecast validation we selected a high-resolution gridded dataset of hourly precipitation, which merged precipitation analyses of the U.S. National Oceanic and Atmospheric Administration Climate Prediction Center morphing technique (CMORPH) given at a spatial resolution of 8 km, with the Chinese gauge-based precipitation analysis based on about 30,000 automatic weather stations. This merged gauge–satellite precipitation product (available at http://data.cma.cn/data/detail/dataCode/SEVP_CLI_CHN_MERGE_CMP_PRE_HOUR_GRID_0.10/) with a resolution of 0.1° x 0.1° used optimal interpolation and the probability density functions of both products, and has been proved to be superior to other similar international products over China (Xie and Xiong 2011; Pan et al. 2012). The verification data were interpolated to 0.5° x 0.5° resolution by bilinear interpolation (Rauscher et al. 2010; Kopparla et al. 2013; Ahmed et al. 2019).

2.2 Method for Object-Based Diagnostic Evaluation (MODE)

Accepted for publication in *Monthly Weather Review*. DOI10.1175/MWR-D-19-0266.1.

MODE sets weight and confidence coefficients for predefined precipitation object attributes and calculates a total interest function based on a fuzzy logic approach, which quantifies the similarity between any two objects (Davis et al. 2006a; Johnson et al. 2013). The predefined attributes are chosen by a particular user for a particular application. In general, MODE consists of four steps: identifying objects, calculating object attributes, finding matching objects between observations and predictions, and assessing the similarity of their attributes.

2.2.1 Identifying Objects and object attributes

In order to extract the spatial boundary of an object, the original precipitation field is spatially smoothed with a convolution radius R (unit: grid points). Then an intensity threshold T (unit: mm (24 h)$^{-1}$) is used to define the boundaries of precipitation objects (Davis et al. 2006a). The original precipitation field within these boundaries then defines the precipitation objects, which are solely determined by the selection of the convolution radius R, which is related to the precipitation scale, and the threshold T, which is related to the precipitation intensity. These two paremeters can be chosen based on the scales of interest. The result of each step is demonstrated in Fig. 1.

We usually pay attention to the overall location of a precipitating system, its size and its shape, especially when dealing with more extreme weather (Johnson and Wang 2013). Therefore, the specific attributes used in our study are the area coverage of precipitation objects, their aspect ratio (the ratio of minor axis to major axis; i.e. 1.0 for a circular object and <1 otherwise) and orientation angle (the orientation of the major

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

axis in degrees counterclockwise starting at zonal orientation), and their centroid location. For matched object pairs (introduced in Section 2.2.2), attribute differences in the four mentioned object attributes (Table 2) are calculated.

2.2.2 Object matching

Object matching creates a pair consisting of one object in the forecasted field and one object in the observed field. Here, we followed Davis et al. (2006a), who determined paired objects solely based on their centroid distance *D* and their areas. If $D < (Area_o^{1/2} + Area_f^{1/2})/2$ with $Area_o$ and $Area_f$ the areas of the observed object and the forecasted object, respectively, both objects create a matched pair. Thus, a matching object pair requires the centroid distance between both to be less than their average size.

2.3 Quantification of similarity of matched object pairs

For a matched pair, its total interest I is computed via

$$I = \frac{\sum_{i=1}^{n} \omega_i c_i G_i}{\sum_{i=1}^{n} \omega_i c_i} \qquad (1)$$

c_i and ω_i are the confidence value and the weight of the attribute *i*, respectively, and *n* is the number of attributes used. While the weight depends only on the particular attribute, the confidence value varies with the sizes and distances of the paired objects (Table 2). G_i is the interest value of the matched objects in terms of attribute *i*; it quantifies the degree of similarity between the objects for that attribute as a monotonic function decreasing from 1 to 0 as the attribute dissimilarity increases (Fig. 2).

2.4 Quantification of object-based forecast accuracy

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

The Median of Maximum Interest (MMI; Davis et al. 2009) and the fuzzy Object-based Threat Score (OTS; Johnson et al. 2011a) are two metrics used to quantify the similarity of the objects in the forecasted and observed fields. The MMI proposed by Davis et al. (2009), which is called the standard MMI in the following, is the median of the maximum total interests in the forecasted and observed fields to which all objects contribute equally regardless of size. The MMI calculated in our study will be slightly larger than the standard MMI, because we first determine the matched objects by their centroid distance and areas, and then the total interest I is only calculated for the matched paris. Thus, unmatched objects are not considered.

The OTS is the fraction of the area of all objects that is contained in matched objects, multiplied by their total interests:

$$OTS = \frac{\sum_{p=1}^{P} I^p (a_f^p + a_o^p)}{A_f + A_o} \tag{2}$$

with P the total number of objects pairs, A_f and A_o the total area of all objects in the forecasted and the observed field, respectively, and a_f^p and a_o^p (p=1,2,…,P) the areas of the p-th paired objects in the forecasted and observed field, respectively. I^p is the total interest value for p-th matched pair. According to Eq. (2) the OTS takes the object area and the number of matched objects into account. Thus, larger objects will contribute more to the OTS than smaller objects, while over-forecasting or under-forecasting the number of objects will decrease the OTS due to more unmatched objects. Both indices range between 0 and 1 and have a value of 1.0 for perfect forecasts. Both

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

scores are used in the present study to quantify two complementary aspects of forecast accuracy.

The MMI of each EPS is the median computed from the 51 members of ECMWF, 21 members of NCEP, 24 members of UKMO, 51 members of JMA, and 15 members of CMA, respectively. The OTS of each EPS is the average OTS of its ensemble members. Both EPS-scores computed over a training period are used as weights to construct the object-based MME prediction as an alternative to point-to-point metrics as used in the classical approach (see next section).

2.5 Different multi-model ensemble types

2.5.1 Traditional grid point-based multi-model ensemble

Super-ensembles have the potential to improve weather and climate forecast skills above individual ensemble forecasts (Kim et al. 2010; Johnson et al. 2014; Krishnamurti et al. 2016). They automatically remove the bias between the observations and model forecasts estimated during a training period, which contributes to the improved prediction skill of multi-model forecasting. In this study, the point-to-point weighted multi-ensemble forecast is defined as:

$$SUP_j = \bar{O} + \sum_{i=1}^{N} \delta_i (Y_{ij} - \bar{Y_i}) \quad (3)$$

$$\delta_i = \left(\frac{1}{T}\sum_{t=1}^{T}|Y_{it} - O_t|\right)^{-1} / \sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T}|Y_{it} - O_t|\right)^{-1} \quad (4)$$

with $\bar{O}$ and $\bar{Y_i}$ respectively the average observed and forecasted value by EPS *i* (*i*=1,2,…,5) computed over a training period T in days, and Y_{ij} the forecast of EPS *i*

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

on day *j* of the forecast period. δ_i is the individual contributing EPS weight, with Y_{it} the forecast value of the *i*-th EPS on day *t* (t=1,2,…,*T*), O_t the respective observation, and *N* the total number of used EPS (in our case *N*=5).

2.5.2 Object-based multi-model ensemble

In this study, the weights for MME forecasts are also calculated by object-scores (i.e. MMI and OTS). As described before, first, the precipitation object with its several object attributes is identified by MODE. Second, one object in the observed field will be matched to one object in the forecast field by satisfying the matching criteria. Third, the similarity between the two matched objects is determined on the basis of the differences in their attributes. Fourth, the similarity values are used to calculate the object-based metrics MMI and/or OTS from which the object-based scores for each EPS are obtained. The performance of each EPS during a training period determines its weight. Tests identified a sliding window of 30 days before the forecast period as the optimal training period. During the training period, for a certain EPS, each ensemble member is evaluated by MODE, and then MMI and/or OTS of this EPS is calculated by the median and/or mean of all ensemble members. Finally, the multi-model ensemble forecasts MME_{MMI} or MME_{OTS} are determined by multiplying the ensemble mean of each contributing EPS by the weight calculated for the training period as follows:

$$MME_{MMI/OTS} = \sum_{i=1}^{N} \delta_i^{MMI/OTS} Y_i \tag{5}$$

$$\delta_i^{MMI} = \frac{1}{T}\sum_{t=1}^{T} MMI_{i,t} \Big/ \sum_{i=1}^{N} \frac{1}{T}\sum_{t=1}^{T} MMI_{i,t} \tag{6}$$

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

$$\delta_i^{OTS} = \frac{1}{T}\sum_{t=1}^{T} OTS_{i,t} / \sum_{i=1}^{N} \frac{1}{T}\sum_{t=1}^{T} OTS_{i,t} \quad (7)$$

with N the number of EPS, Y_i the ensemble mean for the i-th EPS, and $\delta_i^{MMI/OTS}$ the weight of each contributing EPS calculated by MMI or OTS. T is the length of the training period in days, and $MMI_{i,t}$ or $OTS_{i,t}$ is the MMI or OTS value for i-th EPS on day t during the training period.

2.6 Fractions skill score

Besides the MMI and OTS, the Fractions Skill Score, FSS (Roberts and Lean 2007; Roberts 2008), which is not used to generate the weights in the tested MMEs, is applied to evaluate the forecast skill for individual EPS and the MME forecasts. This spatial verification score quantifies forecast skill over different spatial scales. FSS is calculated based on the fractional coverage within a square neighborhood centered on each grid point. FSS requires for a given spatial scale, s, the forecasted and the observed areal fractions M_i and O_i at each grid point, respectively, with precipitation above a given threshold, and is calculated for an area divided into N sub-areas of size s x s as follows:

$$FSS = 1 - \frac{FBS}{FBS_{worst}} \quad (7)$$

$$FBS = \frac{1}{N}\sum_{i=1}^{N}(O_i - M_i)^2 \quad (8)$$

$$FBS_{worst} = \frac{1}{N}\left(\sum_{i=1}^{N} O_i^2 + \sum_{i=1}^{N} M_i^2\right) \quad (9)$$

FBS_{worst} is the largest Fractions Brier Score (FBS), which indicates the case when there are no common non-zero fractions between predictions and observations. The FSS ranges between 0 and 1; 0 stands for a totally mismatched forecast and 1 for a perfect forecast.

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

3. Results

3.1 Individual objects

Convolution radius *R* and threshold value *T* are the only two parameter that influence object recognition and thus affect the values of object attributes such as object number, area, and centroid location. In this section, we analyze the effect of the choice of *R* and *T* on the object attributes. Since the effective resolution of the model dynamics is about seven grid points (Skamarock 2004) and precipitation is generated grid point-wise in the model by the action of parameterizations, we have chosen 3 grid points for the minimum convolution radius *R* as compromise. Since larger *R* values will smooth out especially the interesting heavy precipitation areas, we analyze the impact of different *R* in the intermediate range between 3 and 6 grid points. Only the results for 24-h forecasts are shown; results for other lead times are qualitatively similar.

The variation of the number of objects and their areas with precipitation thresholds and convolution radii in the observations and the forecasts of the ECMWF EPS is displayed in Fig. 3. Observations and forcast exhibit similar behavior; e.g. the number of objects first increases with the precipitation threshold *T* until 5 mm is reached, and then gradually decreases until 25 mm is reached from where the number of objects strongly decrease. Generally, the forecast produces a lower number of objects than the observations, suggesting that the model is more inclined to predict larger continuous precipitation areas. This bias also leads to a large number of false alarms in point-to-point statistics (not shown). The number of objects understandably decreases with

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

increasing convolution radius *R* in both observations and forecasts. Also the average object areas have similar dependencies on *T* and *R* for observations and forecasts, but there are also differences. The average object areas are smaller for the forecasts than for the observations for precipitation less than 5mm. The forecast areas are slightly larger than or equal to the observed areas for higher precipitation thresholds. The average precipitation area is – different from the number of objects - relatively insensitive to the choice of the convolution radius. This is maybe because averaging makes them bigger – but also flatter, so the chosen threshold more or less compensates for that. For a given precipitation threshold, *T*, the number of objects decreases with increasing object area for observations and the ECMWF EPS (Fig. 4). The decrease in object number with increasing area gets larger for higher precipitation thresholds. For larger precipitation thresholds the forecasts produce significantly less objects with larger areas than the observations (be aware, that only the forecast may produce object numbers below 1, because these values are averages over the ensemble members). The effect of *R* and *T* on object number and area is qualitatively similar for the other four models (not shown).

In Figure 5 we compare the distributions of several object attributes between observations and 24-hour predictions by all ensemble members of all five EPSs for a convolution radius *R*=4 grid points (~220km) and a precipitation threshold *T*=10mm as an example. This qualitative analysis of the observed and forecasted daily precipitation distributions is performed in order to investigate if the different numerical models

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

behind the different EPS do capture the observed spatial features. The number of objects decrease rapidly with increasing area (Fig. 5a) with the models having a lower number of very small areas. The object aspect ratio distributions are broad and peak around 0.6 for the observations and at 0.7 for the model forecasts (Fig. 5b). Most objects have an orientation angle between -30 and 30 degrees with the largest number of objects found around 15 degrees especially for the forecasts, which also have secondary peaks at 90 degrees (Fig. 5c). More objects are found in the southern part of the domain, which is also more pronounced in the forecasted field, while the east-west distribution is more even for both observations and model forecasts (Fig. 5d,e). In general, the forecasted and observed distributions are qualitatively similar, which demonstrates that the spatial features of 24-h accumulated precipitation are captured reasonably well by the numerical models.

3.2 Comparison of matched object attributes

We compare now the object attributes of centroid location, aspect ratio, and orientation angle in the matched object pairs. Figures 6 and 7, respectively, show the mean objects' zonal (i.e. east-west) and meridional (i.e. north-south) centroid differences for the five EPSs compared to the observations for different convolution radii and precipitation thresholds. The mean zonal or meridional centroids of the forecasted objects are generally within 0 to 2 grid points of the observed ones for all EPSs. The forecasted objects from all EPSs are located west to the observed objects for thresholds less than 10mm. But for larger thresholds, the objects of NCEP, UKMO and

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

JMA EPSs are eastward from the observed ones. The predicted objects are consistently southward compared to the observed ones for all thresholds and convolution radii. The aspect ratio deviations between model predictions and observations are always positive but small (Fig. 8) indicating that the shape of the forecasted objects is more circular than for the observed objects. The orientation angle differences are on average within 0 to 10 degrees except for the large convolution radii at a threshold of 50mm (Fig. 9). The average positive deviations between forecasted and observed objects indicate a more meridional orientation of the former. In summary, the forecasted objects are more circular, more southwest and have a more meridional orientation than the observed objects. A most likely hypothesis is that these characteristics of forecasted objects are attributed to model dynamics and physics (Johnson et al. 2011b; Johnson and Wang 2013).

Since the four main attributes are not very sensitive to the choice of R (Fig. 6-9) and the difference of object numbers between the forecast and the observation becomes small when R is larger than 3 grid points (Fig. 3), we choose R=4 grid points to investigate the performance of object-based MME forecasting. We have chosen a threshold value of 10 mm for 24-h accumulated precipitation in order to focus on moderate to strong precipitation and exclude light precipitation, which is usually overpredicted in frequency especially by non-convection-permitting model simulations (Giorgi et al. 1992; Golding 2000; Dravitzki and McGregor 2011).

3.3 Multi-model ensemble forecasting

Accepted for publication in *Monthly Weather Review*. DOI10.1175/MWR-D-19-0266.1.

Many studies have shown advantages of MME prediction over predictions from a single EPS (Candille 2009; Beck et al. 2016; Wanders and Wood 2016; Samuels et al. 2018). We first calculate the weights based on the MAE by point-to-point statistics and the MMI or OTS based on MODE, hereafter abbreviated as SUP, MME_{MMI} and MME_{OTS}, respectively. The results of these three MME predictions are weighted ensemble mean forecasts. Thus, they are deterministic forecasts, which we evaluated via MODE taking a threshold of 10mm and a convolution radius of 4 grid points.

The object-based scores for both the individual EPSs and the three MME predictions (i.e. MME_{MMI}, MME_{OTS} and SUP) are compared in Fig. 10. As expected, the forecast skill generally decreases with lead time for all predictions. The ECMWF EPS is more skillful than the other EPSs in terms of MMI and OTS and thus it contributes relatively more to the results of MME_{MMI} and MME_{OTS} (Fig. 11). UKMO EPS performs good for lead times of 1–4 days, and NCEP EPS is better for longer lead times. The relative performance of each EPS is shown by their respective weights. The CMA EPS has the lowest scores and thus contributes the least (Fig. 11). The MME predictions weighted by the MMI and OTS metrics perform similarly well, and perform better than both the individual EPSs and the traditional grid point-based MME prediction based on the point-to-point MAE metric for almost all lead times.

In order to understand why the MME forecasts based on MMI and OTS are better than the single-model ensemble forecasts and the traditional point-to-point MME, we analyze the four main attribute differences (aspect ratio, orientation angle, zonal and

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

meridional centroid location) between the observed and forecasted fields for all lead times (Fig. 12). For all lead times the forecasted objects are on average more circular than the observed ones (Fig. 12a). The object orientation angles resulting from the traditional super-ensemble forecast are somewhat closer to the observations. The orientation angle of the forecasted objects is on average larger than for the observed objects; thus, the forecasted objects have on average a more meridionally oriented orientation than the observed objects (Fig. 12b). For aspect ratio and orientation angle, the MME_{MMI} and MME_{OTS} forecasts on average are not better than the individual model and traditional point-to-point super-ensemble forecasts, while the centroid locations – both latitude and longitude – are better reproduced by both the MME_{MMI} and MME_{OTS} forecasts and are thus the main reason for their overall better performances given the higher weight the centroid locations get in Eq. (1). The traditional point-to-point super-ensemble forecast is unable to predict the location well in our case, especially for the meridional centroid location. But it still beats some individual EPSs for lead times of 3 days and longer. (Figs. 12c,d). The average bias for these four attributes in the MME forecasts is qualitatively similar to the bias of the individual models because all models exhibit similar error characteristics. Accordingly, a MME forecast will suffer from the same errors.

We evaluate the equally-weighted MME mean forecasts (EMME) and the two unequally-weighted MME forecasts MME_{OTS} and SUP with the FSS, which is not used for weight determination in the training periods (Fig. 13). The results for MME_{MMI} are

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

similar to those of MME_{OTS} and thus not displayed in Fig. 13. The FSS always increases with scale; accordingly it is easier to predict precipitation probabilities for larger areas. For all spatial scales, the object-based MME forecasting MME_{OTS} is slightly better than the equally-weighted one (EMME) for all lead times. The grid-point-based MME (SUP) provides the best predictions when evaluated with the FSS. There may be two main reasons for this result. First, precipitation objects often have complicated shapes that are not sufficiently represented by the MODE attributes. In this study, only orientation angle and aspect ratio are used to describe the shape of the precipitation object; thus other meaningful precipitation information may be missed. Second, the grid point-based super-ensemble removes the bias of precipitation intensity between the observations and model forecasts, while the object-based MME in our study removes the spatial bias (e.g. centroid location) but not the precipitation intensity bias.

4. Summary and Discussion

Traditional point-to-point verification methods neglect important spatial information, and are usually insensitive to differences in precipitation location and shape errors. Precipitation is regarded as an object by MODE and several object attributes such as number, area, shapes and centroid locations are identified. The differences in object attributes between the model forecasts and the observations could provide important diagnostic information about prediction biases and help forecasters to better use model forecast products.

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

In this study, the ensemble forecasts from five EPSs (ECMWF, NCEP, UKMO, JMA, and CMA EPS) available from the TIGGE datasets are evaluated via object attributes based on MODE. In addition, we investigate a MME technique based on object-based scores and compare it with the equally-weighted multi-model ensemble mean and super-ensemble forecasts based on the point-to-point metric MAE.

We first analyze the impact of the convolution radius *R* and precipitation threshold *T* on the attributes of the derived precipitation objects. The number of detected objects decreases with increasing convolution radius and precipitation threshold. For all precipitation fields the number of detected objects decreases with increasing object area.

In general, the numerical models could capture the distribution of attributes of the observed precipitation objects, and their forecast skill decreases as expected with lead time. The objects aspect ratio varies between 0.3 and 0.9 and the orientation angles are within ± 30 degrees. More objects are found in the eastern/central and southern portion of the domain than in other parts of the domain. In addition, for matched objects - compared to the observed one - the forecasted object centroid positions by all individual model ensembles are more southward and westward. Forecasted objects tend to be more circular and more southwest-northeast orientated compared to the observed ones. Causes for these features of forecast objects are probably related to dynamical errors and model physics.

For the five EPSs used in this study, the ECMWF EPS performs best. The MME weighted by the spatial metrics outperforms both all single model EPSs and the

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

traditional point-to-point super-ensemble forecast mainly because of the better forecasted object centroid locations when evaluated using the ensemble mean of the object-based metrics. When all EPSs have similar error characteristics, MMEs will not help much. Thus, the causes for such biases – most probably related to model dynamics and parameterization physics - must be found and the models improved accordingly.

When evaluated with the grid point-based (i.e., non-object) metric, FSS, the object-based MME still performs somewhat better than the equally-weighted ensemble mean, but is not as good as the grid point-based MME predictions. This is probably attributed to the use of too few attributes used in our MODE realization and to the inherent bias removal built in the traditional sup-ensemble. MME performance strongly depends on how it is generated, and additional metrics may be used to determine the weights for MME forecasting. Possibly, forecast skill may be further improved by combining different post-processing methods.

The rather small differences between object-based and equally-weighted MME forecasts, in terms of MMI and OTS (not shown), are probably due to similar model biases of the five EPSs in our study domain and suggest an extension of such studies to other domains.

The precipitation objects are identified in our study from the raw ensemble forecasts without any bias correction. Thus, the object-based scores may be improved by appropriate bias correction. Alternatively, appropriate measures of the objects's precipitation intensity could be developed and added as object attributes both for object

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

pair identification and EPS weight determination and potentially improve the forecast skill of object-based MME above pure grid point-based MME even when evaluated by grid point-based metrics. The object-based MME prediction results may also be further improved by excluding the EPSs performing worst in the training period. In addition, the FSS metric can also be employed to determine the weights of the contributing EPSs. Because precipitation structures become increasingly complex as resolution increases, features such as shape and orientation are hard to define at high resolution, thus the FSS might be an alternative to MODE.

Acknowledgements

This study was supported by the National Natural Science Foundation of China (Grant No. 41575104), the NJCAR key project (Grant No. 2016ZD04), the Postgraduate Research & Practice Innovation Program of Jiangsu Province (Grant No. SJKY19_0934) and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

References

Bakhshaii, A., and R. Stull, 2009: Deterministic Ensemble Forecasts Using Gene-Expression Programming, *Wea. Forecasting*, **24,** 1431–1451.

Baldwin, M. E and J. S. Kain, 2006: Sensitivity of several performance measures to displacement error, bias, and event frequency. *Wea. Forecasting*, **21,** 636–648.

Bauer P., A. Thorpe, G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525,** 47-55.

Breivik, Ø., O. J. Aarnes, S. Abdalla, J.-R. Bidlot, and P. A. E. M. Janssen, 2014: Wind and wave extremes over the world oceans from very large ensembles, *Geophys. Res. Lett.*, **41,** 5122–5131.

Brown, B. G., R. Bullock, J. H. Gotway, D. Ahijevych, C. Davis, E. Gilleland, and L. Holland., 2007: Application of the MODE object-based verification tool for the evaluation of model precipitation fields. *22nd Conf. on Weather Analysis and Forecasting and 18th Conf. on Numerical Weather Prediction*, Park City, Utah.

Chen, C. H., C. Y. Li, Y. K. Tan, and T. Wang, 2010: Research of the multi-model super-ensemble prediction based on crossvalidation. *J. Meteor. Res.*, **68,** 464–476.

Davis C. A., B. G. Brown, and R. Bullock, 2006a: Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas. *Mon. Wea. Rev.*, **134,** 1772–1784.

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

Davis, C. A., B. G. Brown, R. Bullock, and J. H. Gotway, 2009: The Method for Object-Based Diagnostic Evaluation (MODE) Applied to Numerical Forecasts from the 2005 NSSL/SPC Spring Program. *Wea Forecasting*, **24,** 1252–1267.

Dravitzki, S., and J. McGregor, 2011: Predictability of heavy precipitation in the Waikato River basin of New Zealand. *Mon. Wea. Rev.*, **139,** 2184–2197.

Fraley, C., A. E. Raftery, and T. Gneiting, 2010: Calibrating multi-model forecast ensembles with exchangeable and missing members using Bayesian model averaging. *Mon. Wea. Rev.*, **138,** 190–202.

Gilleland, E., D. Ahijevych, B. G. Brown, B. Casati, and E. E. Ebert, 2009: Intercomparison of Spatial Forecast Verification Methods. *Wea Forecasting*, **24,** 1416–1430.

Gilleland, E., D. Ahijevych, B. G. Brown, and E. E. Ebert, 2010: Verifying Forecasts Spatially. *Bull. Amer. Meteor. Soc.*, **91,** 1365–1373.

Giorgi, F., G. T. Bates, and S. J. Nieman, 1992: Simulation of the arid climate of the southern Great Basin using a regional climate model. *Bull. Amer. Meteor. Soc.*, **73,** 1807–1823.

Gneiting, T., A. E. Raftery, A. H.Westveld III, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, **133,** 1098–1118,

Golding, B. W., 2000: Quantitative precipitation forecasting in the UK. *J. Hydrol.*, **239,** 286–305.

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

He, C. F., X. F. Zhi, Q. L. You, B. Song, and K. Fraedrich, 2015: Multi-model ensemble forecasts of tropical cyclones in 2010 and 2011 based on the Kalman Filter method. *Meteor. Atmos. Phys.* **127,** 467–479.

Ji, L. Y., X. F. Zhi, S. P. Zhu and K. Fraedrich, 2019: Probabilistic Precipitation Forecasting over East Asia Using Bayesian Model Averaging. *Wea. Forecasting*, **34,** 377–392.

Johnson, A., and X. Wang, 2013: Object-based evaluation of a storm-scale ensemble during the 2009 NOAA Hazardous Weather Testbed Spring Experiment. *Mon. Wea. Rev.*, **141,** 1079–1098.

Johnson, A., X. Wang, F. Kong, and M. Xue, 2011a: Hierarchical Cluster Analysis of a Convection-Allowing Ensemble during the Hazardous Weather Testbed 2009 Spring Experiment. Part I: Development of the Object-Oriented Cluster Analysis Method for Precipitation Fields. *Mon. Wea. Rev.*, **139,** 3673–3693.

Johnson, A., X. Wang, F. Kong, and M. Xue, 2011b: Hierarchical Cluster Analysis of a Convection-Allowing Ensemble during the Hazardous Weather Testbed 2009 Spring Experiment. Part Part II:Season-long ensemble clustering and implication for optimal ensemble design. *Mon. Wea. Rev.*, **139,** 694–3710.

Johnson, A., X. Wang, F. Kong, and M. Xue, 2013: Object-based evaluation of the impact of horizontal grid points on convection-allowing forecasts. *Mon. Wea. Rev.*, **141,** 3413–3425.

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

Johnson, B., V. Kumar, T. N. Krishnamurti, 2014: Rainfall anomaly prediction using statistical downscaling in a multimodel superensemble over tropical South America. *Climate Dyn.*, **43,** 1731–1752.

Jordan, M. R., T. N. Krishnamurti, and C. A. Clayson, 2008: Investigating the utility of using cross-oceanic training sets for superensemble forecasting of eastern Pacific tropical cyclone track and intensity, *Wea. Forecasting*, **23,** 516–522.

Ke, Z. J., W. J. Dong, P. Q. Zhang, J. Wang, and T. B. Zhao, 2009: An Analysis of the Difference between the Multiple Linear Regression Approach and the Multimodel Ensemble Mean, *Adv. Atmos. Sci.*, **26,** 1157–1168.

Kharin, V. V., and F. W. Zwiers, 2002: Climate predictions with multimodel ensembles. *J. Climate*, **15,** 793–799.

Kim, M. K., I. S. Kang, C. K. Park, and Coauthors, 2010: Superensemble prediction of regional precipitation over Korea. *Int. J. Climatol.*, **24,** 777–790.

Kim, O. Y., and J. C. L. Chan, 2018: Cyclone-track based seasonal prediction for South Pacific tropical cyclone activity using APCC multi-model ensemble prediction, *Climate Dyn.*, **51,** 3209–3229.

Krishnamurti, T. N., C. M. Kishtawal, T. E. Larow, and Coauthors, 1999: Improved Weather and Seasonal Climate Forecasts from Multimodel Superensemble, *Science*, **285,** 1548–1550.

Krishnamurti, T. N., C. M Kishtawal, Z. Zhang, and Coauthors, 2000: Multimodel ensemble forecasts for weather and seasonal climate. *J. Climate*, **13,** 4196–4216.

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

Krishnamurti, T. N., V. Kumar, A. Simon, and Coauthors, 2016: A review of multimodel superensemble forecasting for weather, seasonal climate, and hurricanes. *Rev. Geophys*. **54,** 336–377.

Kumar, T. S. V. V., T. N. Krishnamurti, M. Fiorino, and M. Nagata, 2003: Multimodel Superensemble Forecasting of Tropical Cyclones in the Pacific. *Mon. Wea. Rev.*, **131,** 574–583.

Li J, K. Hsu, A. AghaKouchak and S. Sorooshian, 2015: An object-based approach for verification of precipitation estimation. *Int. J. Remote Sens.*, **36,** 513–529.

Loeser, C. F., M. A. Herrera, and I. Szunyogh, 2017: An assessment of the performance of the operational global ensemble forecast systems in predicting the forecast uncertainty. *Wea. Forecasting*, **32,** 149-164.

Mass, C. F.，D. Ovens，K. Westrick and B. A. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts? *Bull. Amer. Meteor. Soc.*, **83,** 407–430.

Nohara, D., A. Kitoh, M. Hosaka, et al., 2006: Impact of climate change on river discharge projected by multimodel ensemble. *J. Hydro.*, **7,** 1076–1089.

Pan, Y., Y. Shen, J. J. Xu, P. Zhao,2012: Analysis of the combined gauge-satellite hourly precipitation over China based on the OI technique, *Acta. Meteorologica Sinica.*, **70,** 1381-1389. (in Chinese)

Parsons, D. B., and Coauthors, 2017: THORPEX Research and the Science of Prediction. *Bull. Amer. Meteor. Soc.*, **98,** 807–830.

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

Peng, P., A. Kumar, and H. Dool, A. G. Barnston, 2002: An analysis of multimodel ensemble predictions for seasonal climate anomalies, *J. Geophys. Res.*, **107(D23)**, ACL-18.

Roberts, N., 2008: Assessing the spatial and temporal variation in the skill of precipitation forecasts from an NWP model. *Meteor. Appl.*, **15,** 163–169.

Roberts, N. M.,, and H. W. Lean, 2007: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. Mon. Wea. Rev., 136, 78–97.

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

Scheuerer, M., 2014: Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quart J. Roy. Meteor. Soc.*, **140,** 1086–1096.

——, and T. M. Hamill, 2015: Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Mon. Wea. Rev.*, **143,** 4578–4596.

Skamarock, W. C., 2004: Evaluating mesoscale NWP models using kinetic energy spectra. *Mon. Wea. Rev*., **132,** 3019–3032.

Tebaldi, C., L. O. Mearns, D. Nychka, and R. L. Smith, 2004: Regional probabilities of precipitation change: A Bayesian analysis of multimodel simulations. *Geophys. Res. Lett.*, **31,** L24213.

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

Vrugt, J. A., M. P. Clark, C. G. H. Diks, Q. Duan, and B. A. Robinson, 2006: Multi-objective calibration of forecast ensembles using Bayesian model averaging, *Geophys. Res. Lett.*, **33,** L19817.

Wolff, J. K., M. Harrold, T. Fowler, J. H. Gotway, L. Nance, and B. G. Brown, 2014: Beyond the basics: Evaluating modelbased precipitation forecasts using traditional, spatial, and object-based methods. *Wea Forecasting*, **29,** 1451–1472.

Xie, P. P., and A. Y. Xiong, 2011: A conceptual model for constructing high-resolution gauge-satellite merged precipitation analyses. *J. Geophys. Res.*, **116,** D21106.

Yuan, H., X. Gao, S. L. Mullen, S. Sorooshian, J. Du, and H. H. Juang, 2007: Calibration of probabilistic quantitative precipitation forecasts with an artificial neural network, *Wea. Forecasting*, **22,** 1287-1303.

Yun, W.T., L. Stefanova, A. K. Mitra, T. S. V. Vijaya Kumar, W. Dewar and T. N. Krishnamurti, 2005: A multi-model superensemble algorithm for seasonal climate prediction using DEMETER forecasts, *Tellus*, **57A,** 280–289.

Zhang, H. B., X. F. Zhi, J. Chen, and Coauthors, 2015: Study of the modification of multi-model ensemble schemes for tropical cyclone forecasts. *J. Trop. Meteor.*,**21,** 389–399.

Zhang, L., and X. F. Zhi, 2015: Multi-model consensus forecasting of low temperature and icy weather over central and southern China in early 2008. *J. Trop. Meteor.*, **21,** 67–75.

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

Zhi, X. F., H. X. Qi, Y. Q. Bai, and C. Z. Lin, 2012: A comparison of three kinds of multi-model ensemble forecast techniques based on the TIGGE data. *J. Meteor. Res.*, **26,** 41–51.

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

Captions

Table 1. Ensemble forecast systems used in this study.

Table 2. Weights and confidence values for pair attributes of matched objects used in this study. CD and CDI denotes the centroid distance and centroid distance interest, respectively. AR is the area ratio (AR= min($Area_o$, $Area_f$)/max($Area_o$, $Area_f$)) and K is the aspect ratio. This table is adopted from Johnson and Wang (2013).

Fig. 1 Observed objects for 24-h accumulated precipitation on 2 Jun 2013. (a) original precipitation field before smoothing; (b) convoluted precipitation field after smoothing with a 4-gridpoint avaraging radius; (c) filtered precipitation field with the precipitation intensity greater or equal to 10mm.

Fig. 2 Interest function G_i for (a) area ratio; (b) centroid distance; (c) aspect ratio difference and (d) angle difference. This figure is adopted from Johnson and Wang (2013).

Fig. 3. The total number of objects and the average objects area for the observation (solid line) and the ECMWF EPS (dashed line) for different convolution radii (colors) and precipitation thresholds (abscissa).

Fig. 4. Number of average precipitation objects vs. average object area for different precipitation thresholds T (line color and type) for the observations (left) and the forecasts of the 51 members of the ECMWF EPS (right) and for increasing convolution radii R (top-to-bottom).

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

Fig. 5. Distribution of objects with specific attribute values as a fraction of the total number of objects for convolution radius R=4 grid points and precipitation threshold T=10mm for observations (black bar) and 24-h lead time predictions from all members of all EPS (white bar). (a) object area, (b) object aspect ratio, (c) object orientation angle, (d) zonal grid point of object centroid, (e) meridional grid point of object centroid.

Fig. 6. The objects mean zonal centroid location of the individual members of the five EPS compared to the observation. (a) ECMWF, (b) NCEP, (c) UKMO, (d) JMA, and (e) CMA.

Fig. 7. The same as Fig. 6, but for meridional centroid location.

Fig. 8. The same as Fig. 6, but for aspect ratio.

Fig. 9. The same as Fig. 6, but for orientation angle.

Fig. 10. (a) MMI and (b) OTS for five individual EPSs and the three multi-model forecasting with R=4 grid points and T=10mm for the lead time of 1-7 days.

Fig. 11. Weights of the five EPSs with lead times of 1-7 days calculated by MMI (right) and OTS (left), respectively.

Fig. 12. The average difference between the forecasted (individual EPSs and multi-model ensemble forecasts) and observed object attribute distributions with R=4 grid points and T=10 mm as a function of lead time for (a) aspect ratio, (b) orientation angle, (c) zonal grid point of centroid and (d) meridional grid point of centroid.

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

Fig. 13. Fractions skill scores against forecast lead days for spatial scales s of 1 grid point (dots), 2 grid points (asterisks), and 3 grid points (triangles) for the multi-model ensemble predictions based on MAE (SUP), equally-weighted multi-model ensemble mean (EMME), and multi-model ensemble predictions based on object-based scores (MME_{OTS}).

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

Tables

Table 1. Ensemble forecast systems used in this study.

Prediction Center	Model spectral resolution	Initial perturbation scheme	Representation of model error and uncertainty	Ensemble members	Max forecast Lead time (day)
ECMWF	T399l62/T255L62	Singular vectors and EDA	SKEB/SPPT	51	15
NCEP	T126L28	Ensemble transform and rescaling	STTP	21	15
UKMO	90km	ETKF	SKEB	24	15
JMA	T106	Singular vectors	SPPT	51	11
CMA	T106	Bred vectors	None	15	10

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

Table 2. Weights and confidence values for pair attributes of matched objects used in this study. CD and CDI denotes the centroid distance and centroid distance interest, respectively. AR is the area ratio (AR= min(area(obs), area(mod))/max(area(obs), area(mod))) and K is the aspect ratio. This table is adopted from Johnson and Wang (2013).

Pair attributes of matched objects	Weight	Confidence
Centroid distance (CD)	2.0	AR
Area ratio (AR)	2.0	1.0 if CD≤160km 1-[(CD-160)/640] if 160<CD<800km 0.0 if CD≥800km
Aspect ratio difference	1.0	CDI×AR
Orientation angle difference	1.0	CDI×AR× $\sqrt{a^2+b^2}$ a,b=$[(K-1)^2/(K^2-1)]^{0.3}$ for the two matched objects

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

Figures

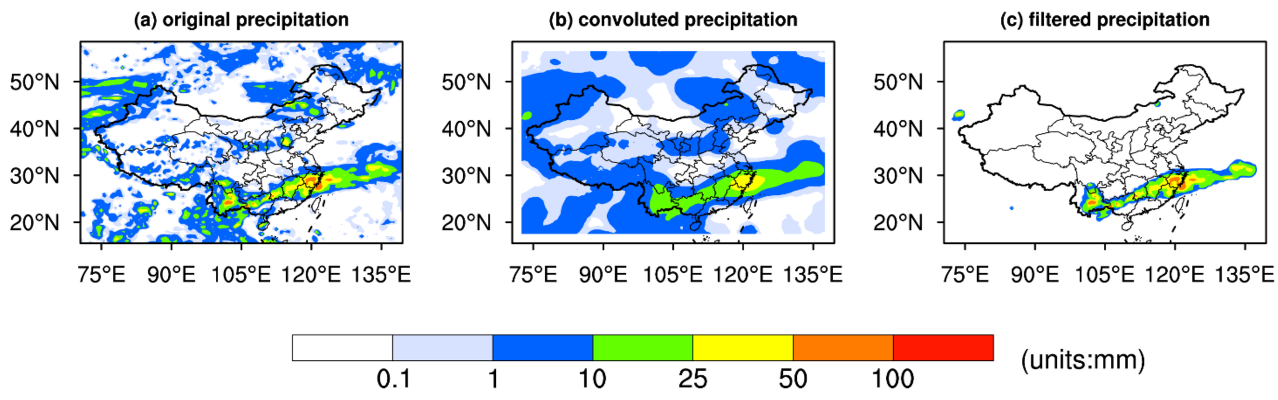


Fig. 1. Observed objects for 24-h accumulated precipitation on 2 Jun 2013. (a) original precipitation field before smoothing; (b) convoluted precipitation field after smoothing with a 4-gridpoint avaraging radius; (c) filtered precipitation field with the precipitation intensity greater or equal to 10mm.

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

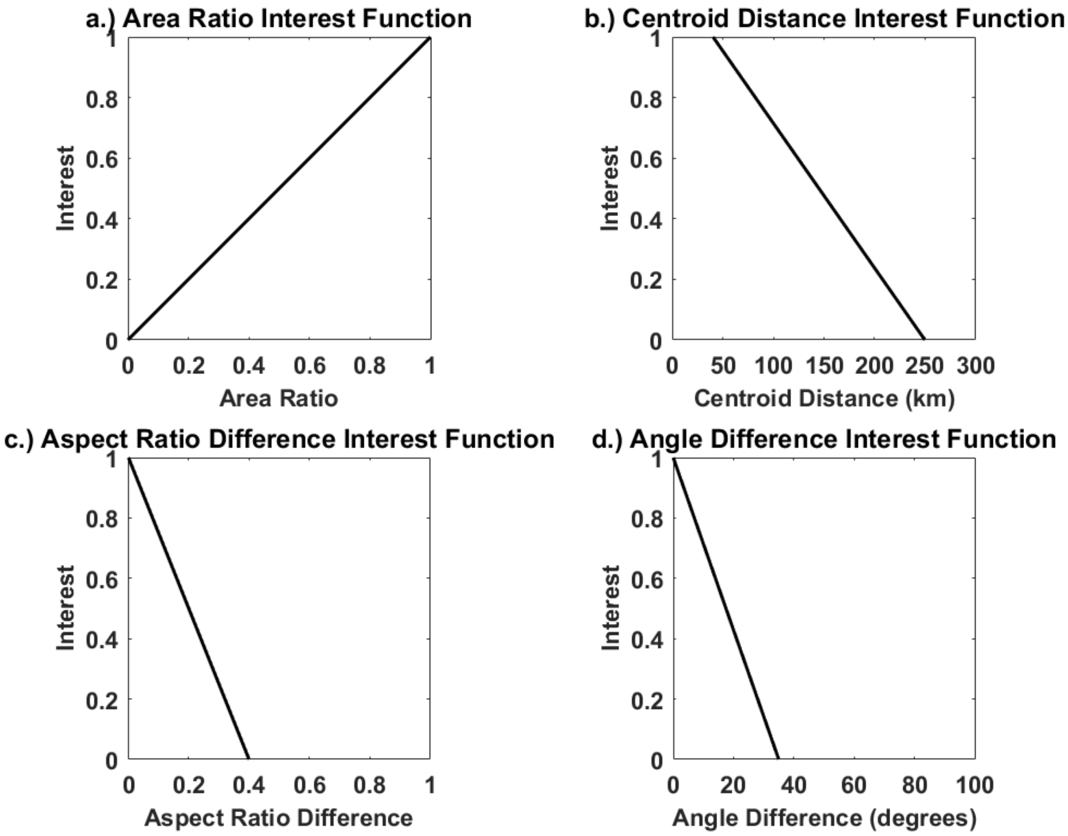


Fig. 2. Interest function G_i for (a) area ratio; (b) centroid distance; (c) aspect ratio difference and (d) angle difference. This figure is adopted from Johnson and Wang (2013).

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

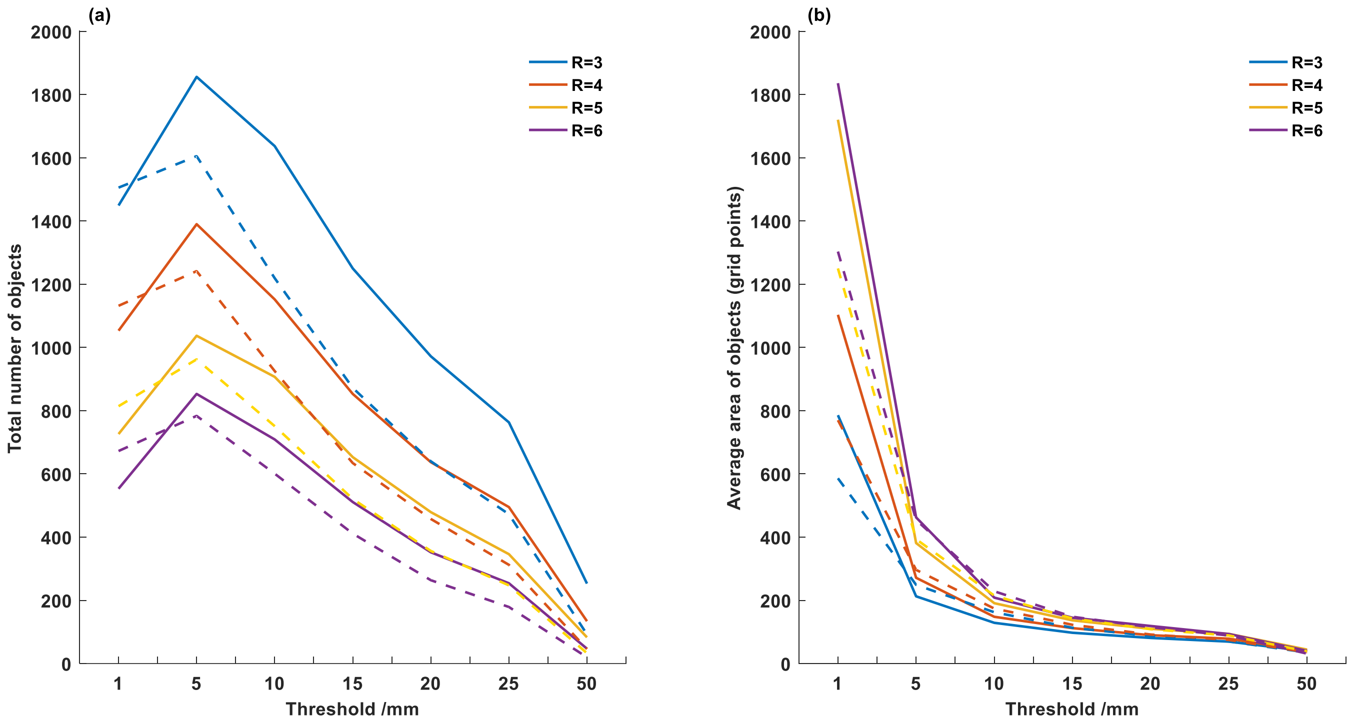


Fig. 3. The total number of objects and the average objects area for the observation (solid line) and the ECMWF EPS (dashed line) for different convolution radii (colors) and precipitation thresholds (abscissa).

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

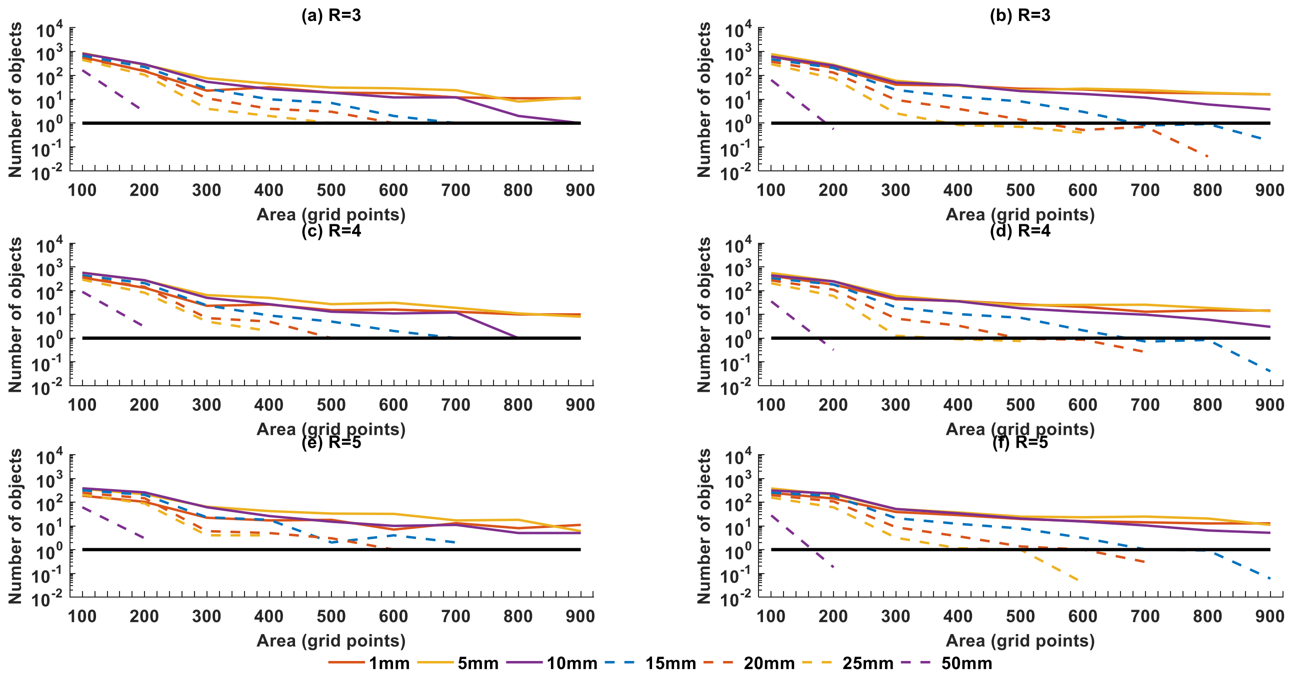


Fig. 4. Number of average precipitation objects vs. average object area for different precipitation thresholds T (line color and type) for the observations (left) and the forecasts of the 51 members of the ECMWF EPS (right) and for increasing convolution radii R (top-to-bottom).

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

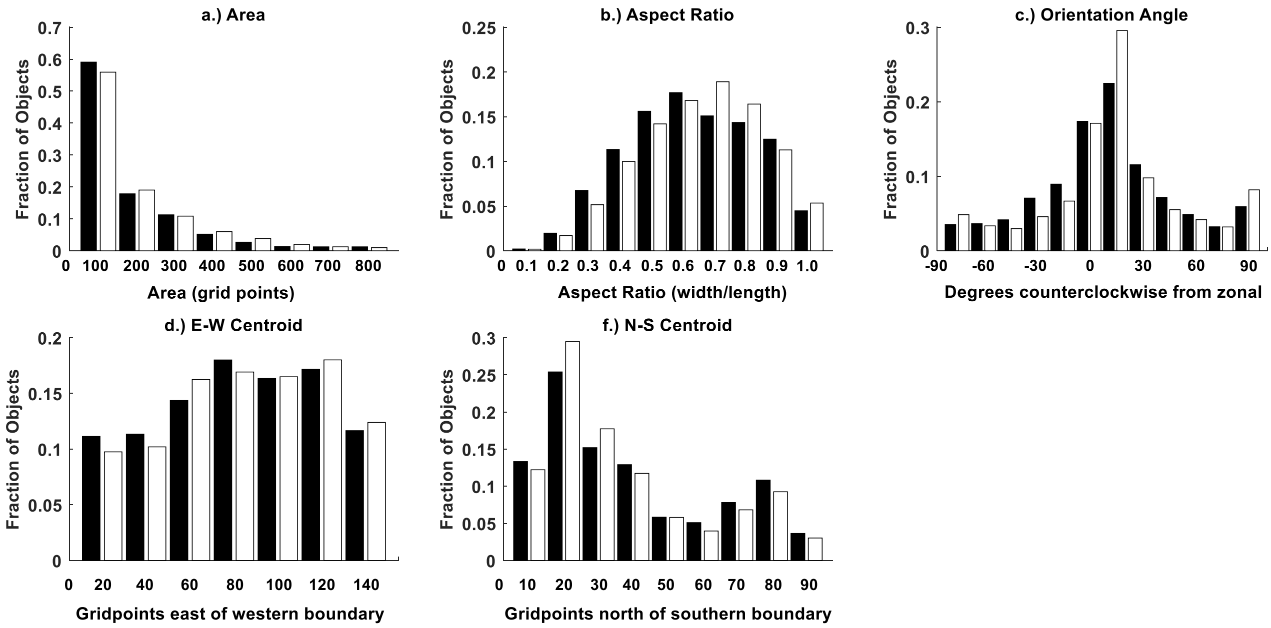


Fig. 5. Distribution of objects with specific attribute values as a fraction of the total number of objects for convolution radius R=4 grid points and precipitation threshold T=10mm for observations (black bar) and 24-h lead time predictions from all members of all EPS (white bar). (a) object area, (b) object aspect ratio, (c) object orientation angle, (d) zonal grid point of object centroid, (e) meridional grid point of object centroid.

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

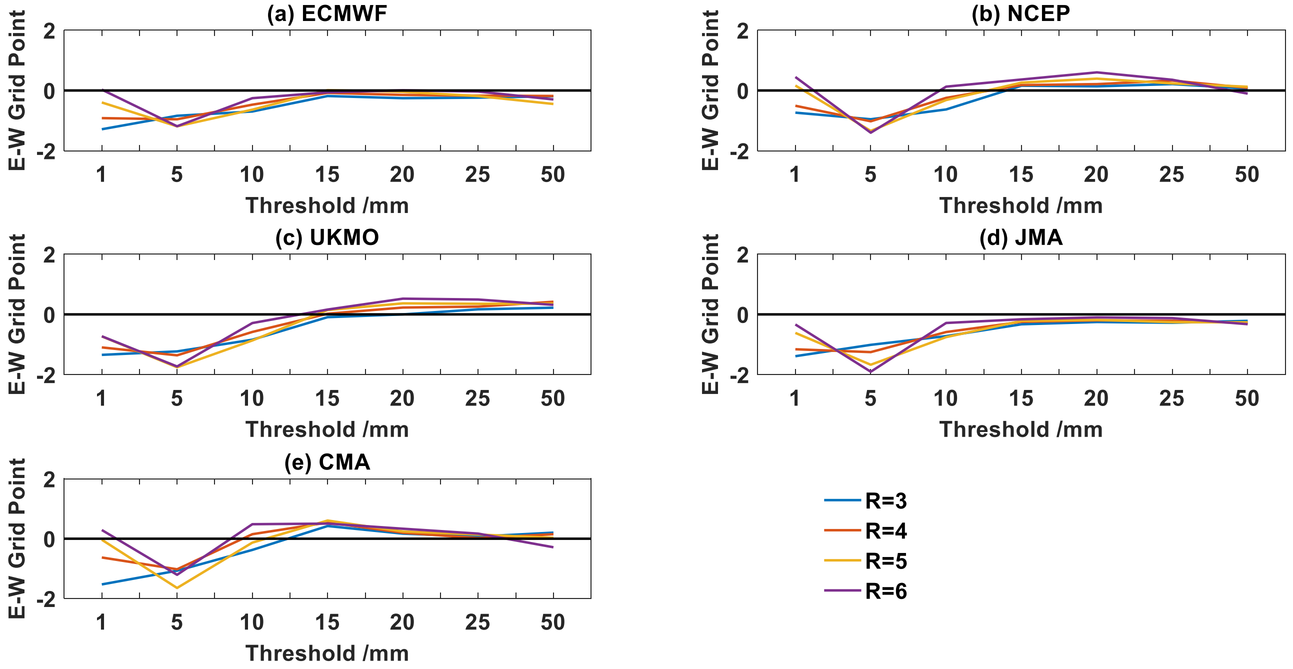


Fig. 6. The objects mean zonal centroid location of the individual members of the five EPS compared to the observation. (a) ECMWF, (b) NCEP, (c) UKMO, (d) JMA, and (e) CMA.

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

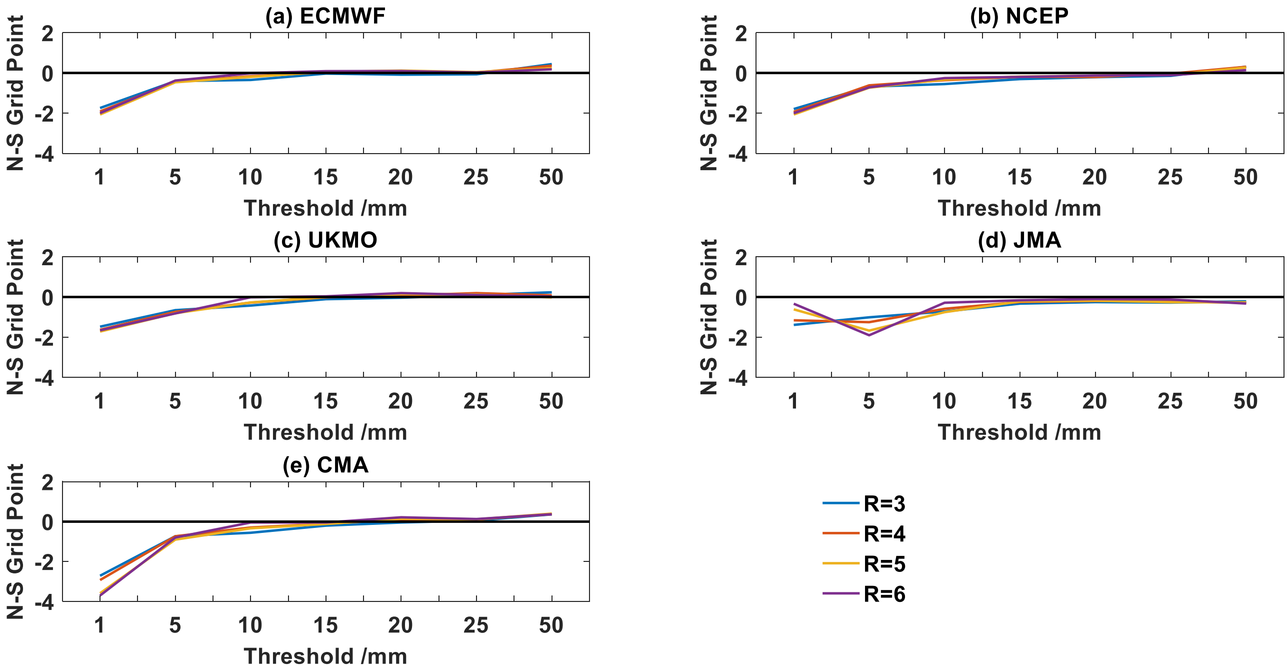
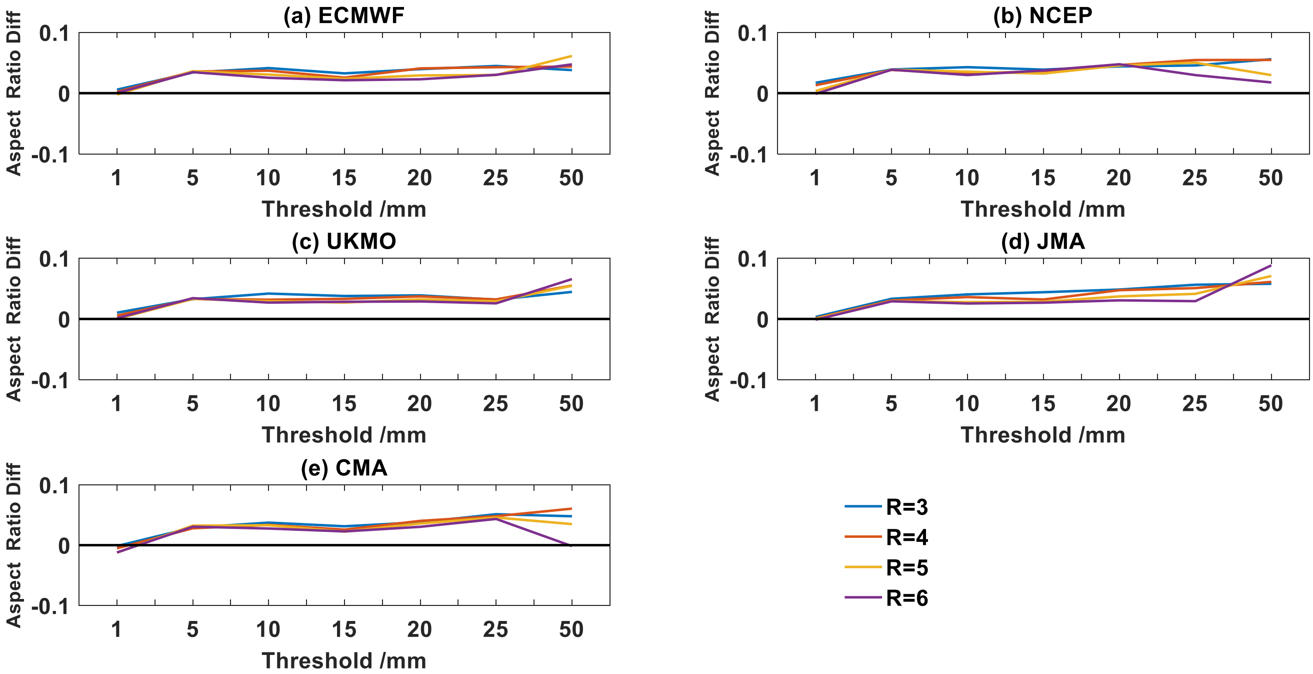


Fig. 7. The same as Fig. 5, but for meridional centroid location.

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.



718

719 **Fig. 8.** The same as Fig. 5, but for aspect ratio.

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

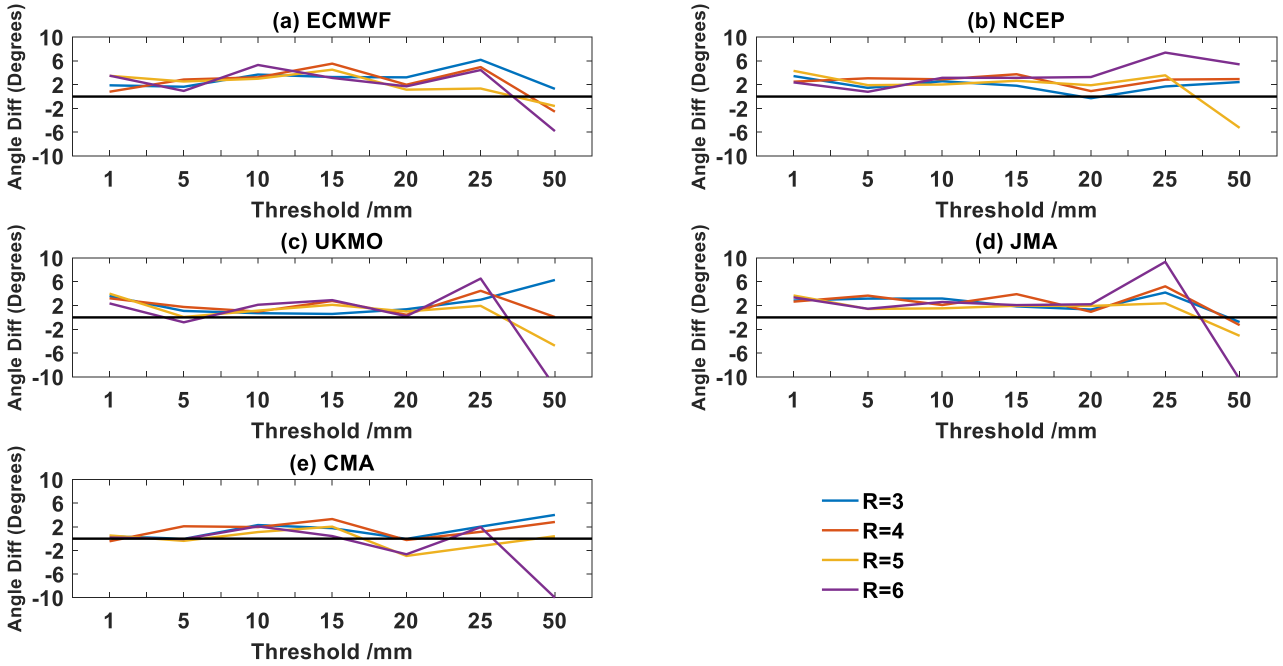


Fig. 9. The same as Fig. 5, but for orientation angle.

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

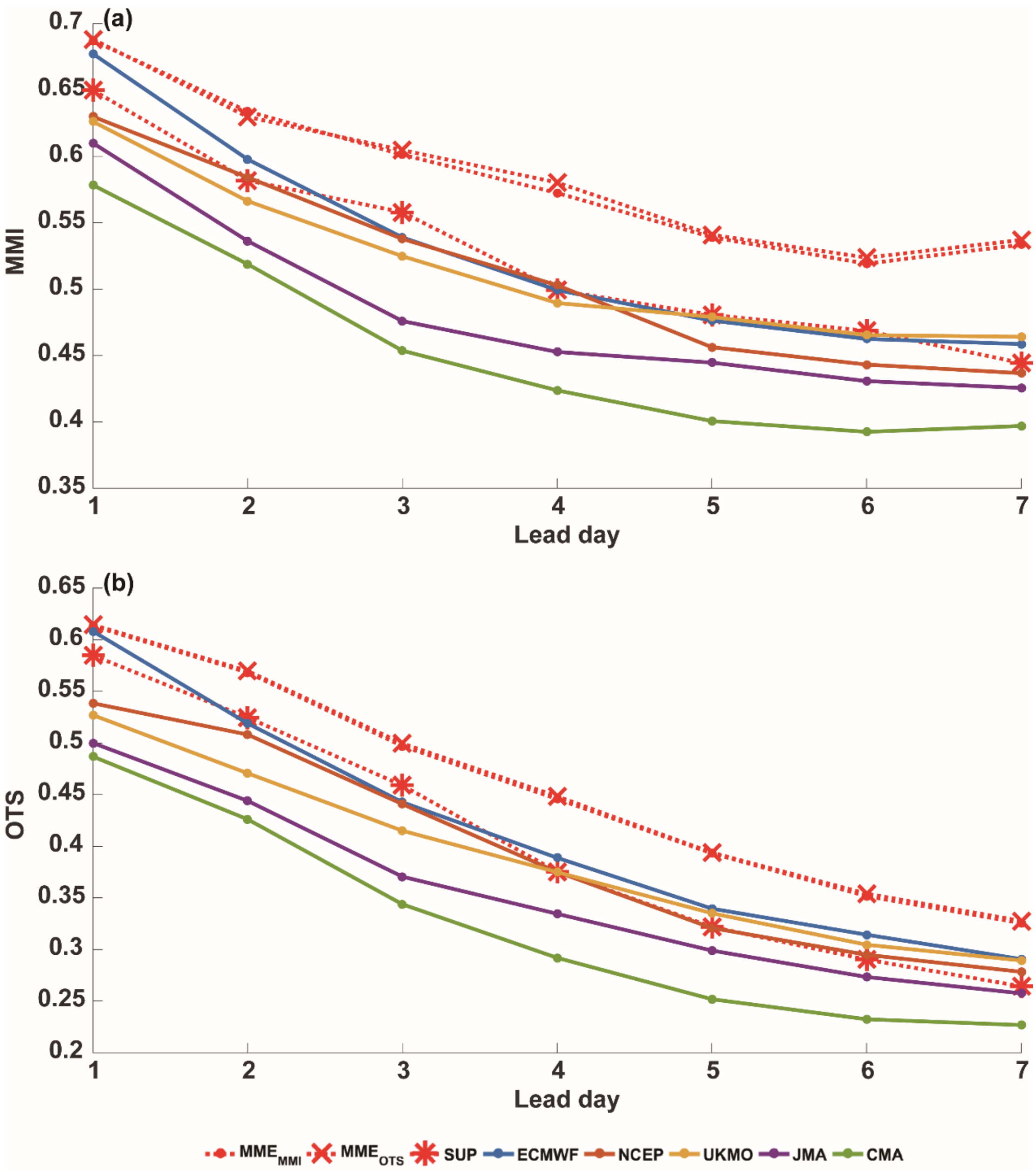


Fig. 10. (a) MMI and (b) OTS for five individual EPSs and the three multi-model forecasting with *R*=4 grid points and *T*=10mm for the lead time of 1-7 days.

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

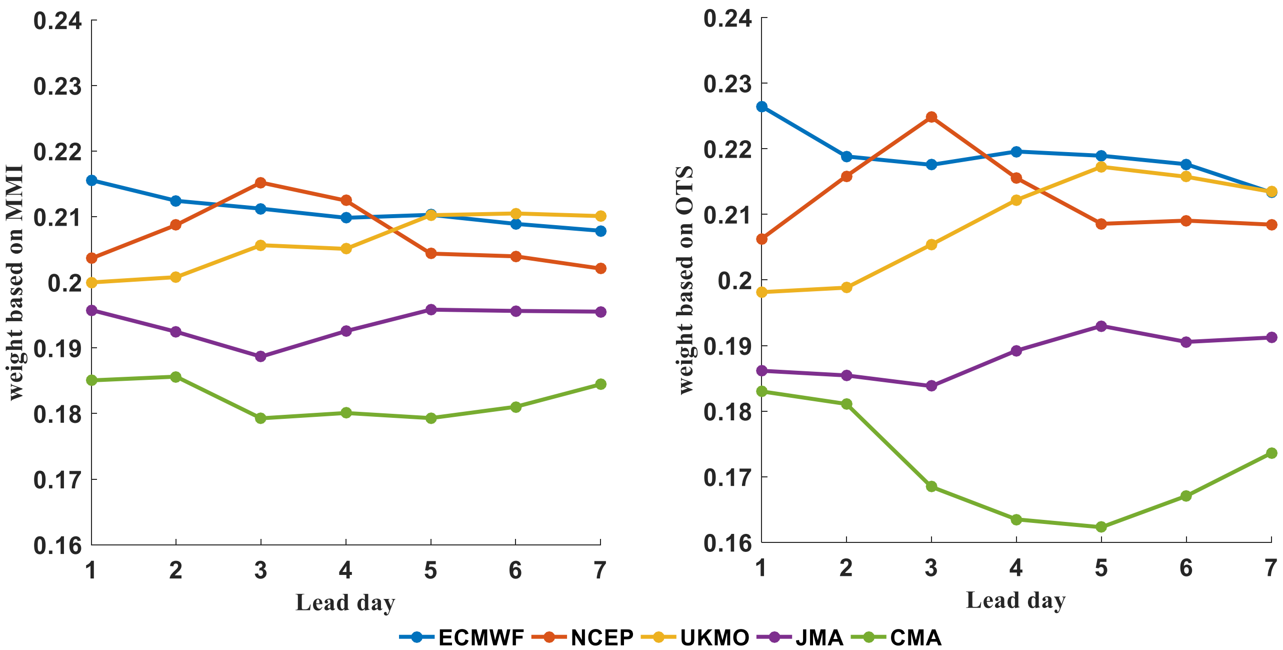


Fig. 11. Weights of the five EPSs with lead times of 1-7 days calculated by MMI (right) and OTS (left), respectively.

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

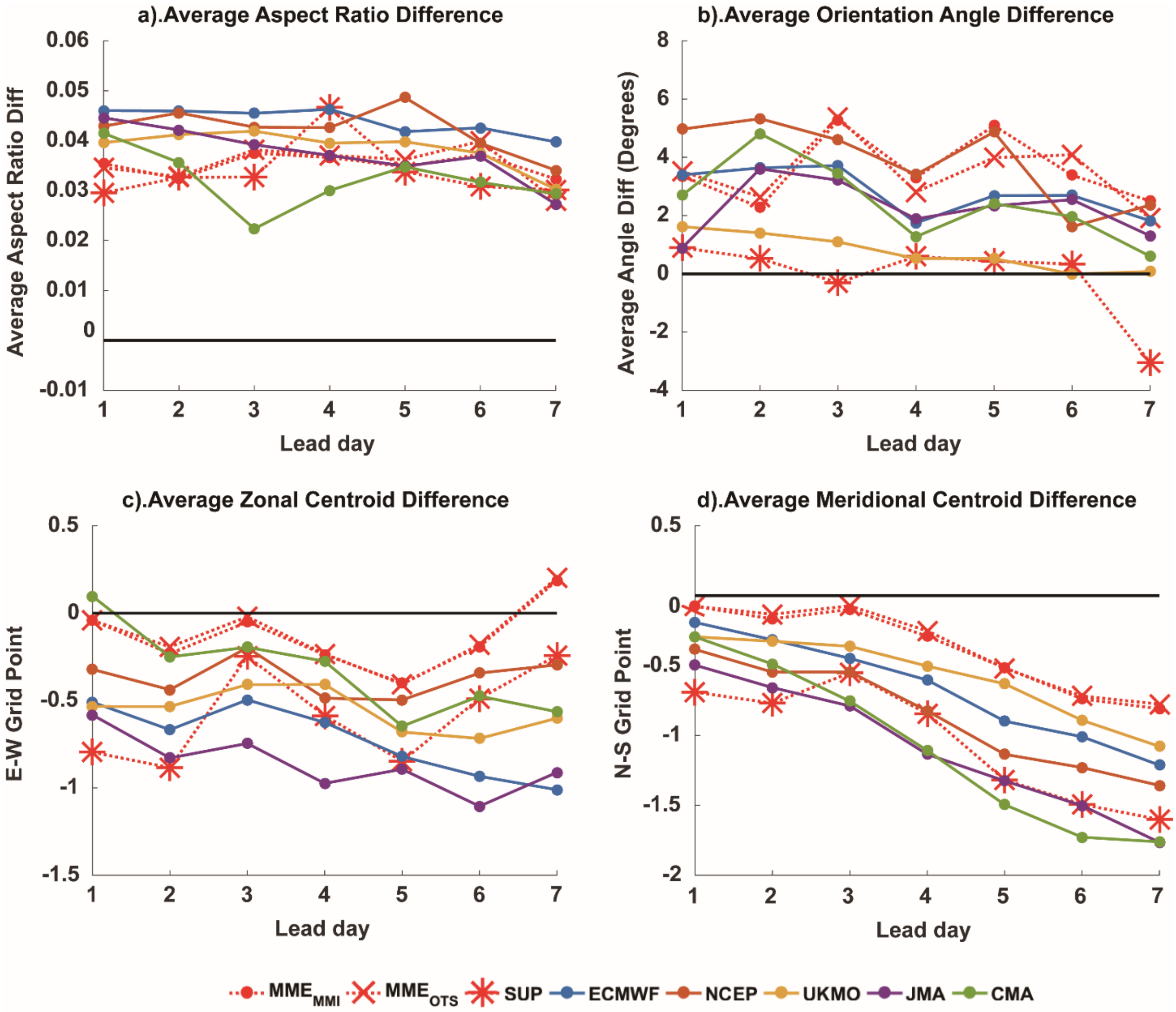


Fig. 12. The average difference between the forecasted (individual EPSs and multi-model ensemble forecasts) and observed object attribute distributions with R=4 grid points and T=10 mm as a function of lead time for (a) aspect ratio, (b) orientation angle, (c) zonal grid point of centroid and (d) meridional grid point of centroid.

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.

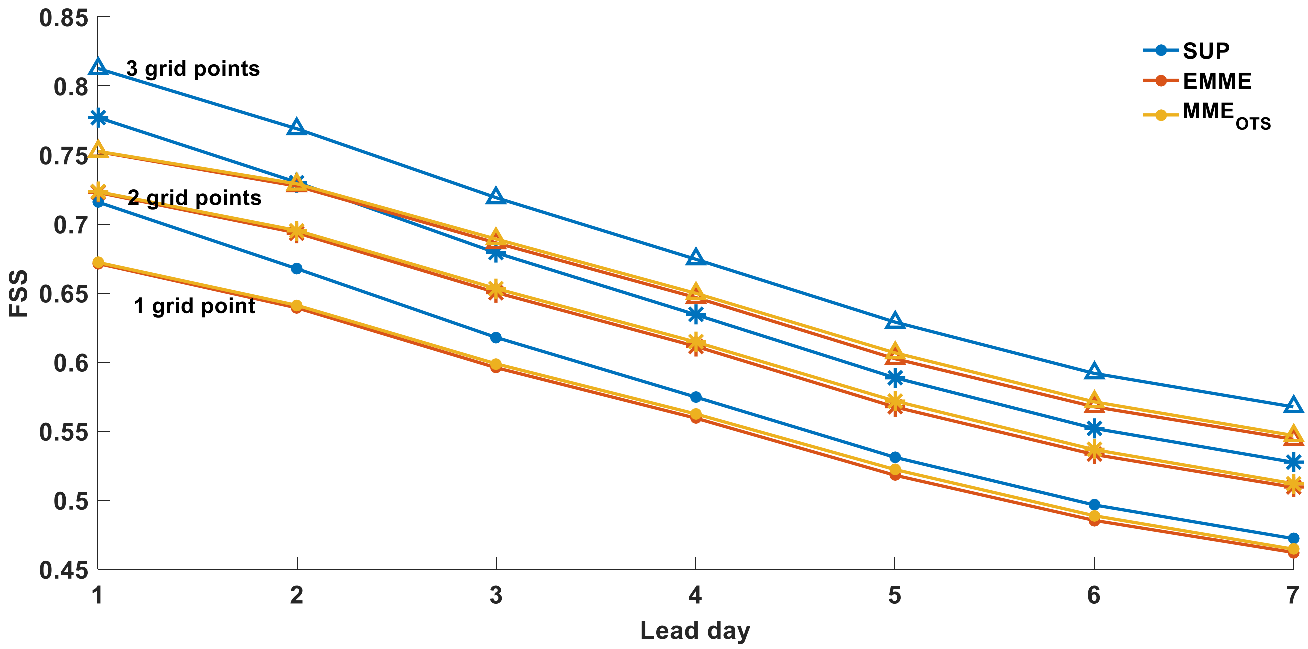


Fig. 13. Fractions skill scores against forecast lead days for spatial scales s of 1 grid point (dots), 2 grid points (asterisks), and 3 grid points (triangles) for the multi-model ensemble predictions based on MAE (SUP), equally-weighted multi-model ensemble mean (EMME), and multi-model ensemble predictions based on object-based scores (MME_{OTS}).

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-19-0266.1.