Atmospheric Research 246 (2020) 105117

Contents lists available at ScienceDirect

Atmospheric Research

journal homepage: www.elsevier.com/locate/atmosres

Using new neighborhood-based intensity-scale verification metrics to evaluate WRF precipitation forecasts at 4 and 12 km grid spacings

Biyu Yu[a,c], Kefeng Zhu[a,d,*], Ming Xue[a,b], Bowen Zhou[a]

[a] Key Laboratory of Mesoscale Severe Weather, Ministry of Education and School of Atmospheric Sciences, Nanjing University, Nanjing 210093, China
[b] Center for Analysis and Prediction of Storms, School of Meteorology, University of Oklahoma, 73072, United States
[c] East China Regional Air Traffic Management Bureau of Civil Aviation of China, Shanghai, China
[d] Nanjing Joint Institute for Atmospheric Sciences, Chinese Academy of Meteorological Sciences, Nanjing 210041, China

ARTICLE INFO

Keywords:
Verification
Convection-permitting model evaluation
Scale-dependent methods
Meiyu season

ABSTRACT

Wavelet-decomposition-based intensity-scale skill (ISS) score is a verification metric which decomposes the forecast fields into different scales and then calculates verification scores. However, due to the "double-penalty" issues, ISS at small scales are often very low even when the forecasts appear subjectively skillful. The displacement error is an important reason for the low ISS. To address this problem, verification methods based on a combination of neighborhood and scale-separation verification approaches are explored. Instead of calculating ISS, a neighborhood-based fractions skill score at different spatial scales, which we call IS_FSS is proposed. Additionally, to reduce the impact of intensity bias, percentile-based instead of fixed thresholds are used in IS_FSS, leading to ISP_FSS. Those two newly developed verification scores are then used to assess WRF forecasts at 4 km and 12 km grid spacings (WRF-4 and WRF-12, respectively) for a case and the entire Meiyu season of 2016. Compared to ISS scores, both IS_FSS and ISP_FSS show more positive verification scores of both WRF-4 and WRF-12 at small scales. Moreover, IS_FSS and ISP_FSS are able to differentiate WRF-12 and WRF-4 forecasts at smaller scales when ISS cannot. Both scores indicate that WRF-4 outperforms WRF-12 for spatial scales from 12 km through 96 km. The scores of WRF-12 improve more when using ISP_FSS than WRF-4, because the former has higher intensity bias.

1. Introduction

With the advancement of computing power, operational numerical weather prediction (NWP) models are heading towards convection-allowing or convection-resolving resolutions (at grid spacings less than 4–5 km) (Xue et al., 2007; Funatsu et al., 2009; Gowan et al., 2018; Sun et al., 2019). At such resolutions, many convective-scale features, including many aspects of convective storms, can be resolved. Many studies have shown that models at convection-allowing or convection-resolving resolutions produce better precipitation forecasts in terms of spatial distribution, and diurnal variation of precipitation as well as precipitation structures (Burgess et al., 2002; Davies et al., 2005; Clark et al., 2007; Weisman et al., 2008; Love et al., 2011; Zhu et al., 2018). However, when it comes to quantitative precipitation verification, commonly used verification scores do not always reveal better forecast skills for such high-resolution models compared to lower resolution ones. As shown later in this paper, standard intensity-scale skill (ISS) scores of the 4-km Weather Research and Forecast (WRF, Skamarock et al., 2005) model precipitation forecasts show negative ISS scores for small scales. The ISS scores use the Haar-wavelet to decompose the forecast fields into a range of scales for which verification scores are calculated individually.

There are many reasons why useful forecast skill scores are not seen in objective assessments for small scales. Spatial and intensity errors are two of the major error sources. Although high-resolution models can successfully predict precipitation structures that are of value to forecasters and end users, such structures usually cannot exactly match observations in location and timing beyond several hours of forecast. Additionally, convective systems are highly variable in their size, intensity, structure, orientation, timing, etc. (Lorenz, 1969; Brandes et al., 2003; Hintsa et al., 2004), therefore exact predictions in all details and scales are impossible beyond several hours due to predictability limits. Certain aspects of the forecast error should be tolerated while skillful aspects of the forecasts should be properly reflected in the verification scores used.

Various verification methods have been proposed in recent years to

* Corresponding author at: School of Atmospheric Sciences, Nanjing University, Nanjing 210093, China.
E-mail address: kefeng@nju.edu.cn (K. Zhu).

https://doi.org/10.1016/j.atmosres.2020.105117
Received 10 April 2020; Received in revised form 27 June 2020; Accepted 27 June 2020
Available online 01 July 2020

address some of the above issues. The methods can be divided into four categories (Gilleland et al., 2009): feature-based (or object-based) (Ebert and McBride, 2000; Davis et al., 2006a), field deformation (Wernli et al., 2008; Gilleland et al., 2010a), neighborhood (or fuzzy) (Casati et al., 2008; Ebert, 2009), and scale-separation (or decomposition) (Casati et al., 2004; Casati, 2010) techniques. Another class of method is based on distance measures (Dorninger et al., 2018), as discussed in more detail in Gilleland et al. (2020). There have been projects focusing on inter-comparing and understanding a large number of forecast verification methods, including the first spatial verification methods intercomparison project (ICP; Gilleland et al., 2009), and its second phase called the Mesoscale Verification Intercomparison over Complex Terrain (MesoVICT) project (Dorninger et al., 2018). Several review papers exist on forecast verification methods, including Davis et al. (2006b), Ebert (2008), Gilleland et al. (2010b), and Brown et al. (2012). Interested readers are referred to those papers and references therein.

The neighborhood techniques can be described as filtering methods, which evaluate model skill based on filtered fields. Instead of point-to-point verification, neighborhood methods act as a smoothing filter to the binary probabilities. Such methods are useful to understand the contribution of spatial error (Ebert, 2009). However, the neighborhood method cannot measure scale-dependent error. Scale-separation methods are designed to address this issue. Such methods separate or decompose the forecast fields into different scales by applying several single-band-pass spatial filters or performing wavelet or spectral decomposition. Verification scores are then calculated for different scales (Briggs and Levine, 1997; Zepeda-Arce et al., 2000; Jung and Leutbecher, 2008; Weniger et al., 2017; Buschow and Friederichs, 2020).

Among various scale-separation methods, the intensity-scale technique (Casati et al., 2004) is quite popular. By performing wavelet decomposition, forecast skills can be obtained at different intensity and spatial scales (Mittermaier, 2006; Casati, 2010; Stratman et al., 2013). However, because skill scores for the smaller scales are still calculated on the point-to-point basis, the double-penalty issue remains. It is still often difficult for such skill scores to show useful skill for high-resolution models. Mittermaier (2006) showed that both 4- and 12-km resolution models mostly record negative skill scores in terms of ISS for scales from 12 km to 96 km when using the Haar wavelet decomposition. Negative ISS scores indicate that the model forecast is worse than a random forecast, but in reality the forecast may still be useful (because of correct forecast of features, for example). Displacement and intensity errors are often the main causes of negative skill scores for small-scale convection. Besides, the intensity-scale methods are still penalized by intensity and/or position errors. The neighborhood approach, on the other hand, can provide information on the scale at which the forecast becomes skillful, and is thereby less influenced by position and intensity errors. Therefore, it is desirable to combine the scale separation and neighborhood approaches within score metrics for the purpose of measuring different aspects of errors while not getting over-penalized by intensity and/or position errors.

In this study, with the aim of reducing the negative impacts of displacement and intensity errors on scale-separation evaluation metrics so that the calculated verification scores better reflect the value of forecasts, we take the idea of neighborhood-based methods and apply it to the regular ISS, and develop a new intensity-scale-based fraction skill score (IS_FSS). As a variation to IS_FSS, percentile-based thresholds are used to replace fixed-value thresholds (Zhu et al., 2015) to reduce the effect of intensity error, resulting in a percentile-based IS_FSS or ISP_FSS. These metrics are then applied to 4 km and 12 km WRF forecasts over China for an individual Meiyu frontal precipitation case, then to the entire Meiyu season of 2016 and the performances of the metrics are evaluated. The goal is to obtain objective verification scores that can reduce the impact of the "double -penalty" on small scales and better match subjective evaluations.

The Meiyu season is one of the major rainy period of warm season rainfall (May to September) in China. In the Yangtze River-Huaihe Valleys, it usually begins around mid-June and lasts for up to four weeks (Bao et al., 2011). During this period, a west-east-oriented rainband extending thousands of kilometers forms along the quasi-stationary Meiyu front. Along the Meiyu front, mesoscale convective systems are prevalent, which often contain embedded intensive convection including squall lines. As a result, the precipitation fields contain multi-scale structures, including intense localized precipitation cores and wide spread stratiform precipitation regions. In the meantime, afternoon convection is active in southern China, and more organized quasi-linear convective systems such as squall lines and bow echoes are common. For such multiscale precipitation, verification metrics capable of distinguishing forecast skills at different scales are highly desirable.

The rest of the paper is organized as follows. In section 2, we briefly summarize relevant verification metrics, including scale-separation and neighborhood-based methods, and put forward our proposed new metrics. The forecast datasets and observations used for the verifications and their processing are described in section 3. Evaluation results with various verification metrics for a single Meiyu frontal precipitation case are presented in section 4, followed in section 5 by the application of the metrics to the entire Meiyu season (~40 days) of 2016. A summary is given in section 6.

2. Verification metrics

Existing scale-decomposition and neighborhood-based fuzzy verification metrics as well as our proposed ones are discussed in this section.

2.1. Intensity-scale separation methods and intensity-scale skill score

The intensity-scale verification method employed in this study is based on Casati et al. (2004), which we briefly summarize here. The first step is to transform the model forecast (Y) and observation (X) fields into binary fields for each threshold. If the gridded precipitation value is greater than a predetermined threshold, it is assigned 1; otherwise, it is assigned 0. In this study, intensity thresholds u are set to 4, 13, 25 and 60 mm for 6-hour accumulated rainfall. The same thresholds are used by the National Meteorological Center in China.

$$I_Y = \begin{cases} 0 & Y < u \\ 1 & Y \geq u \end{cases}, I_X = \begin{cases} 0 & X < u \\ 1 & X \geq u \end{cases} \quad (1)$$

Once the binary fields are obtained, the wavelet scale components for a given threshold are obtained by applying a two-dimensional discrete Haar wavelet decomposition to the binary fields $(I_*)_u = \sum_{l=1}^{L} I_{u,l}$. Here, "*" represents either forecast or observation. l is the spatial scale and L is the maximum decomposition level. The binary error of a given threshold u is calculated as $Z_u = (I_Y)_u - (I_X)_u = \sum_{l=1}^{L} Z_{u,l}$. The process of multi-level wavelet decomposition is explained in detail in Casati et al. (2004)

For each intensity threshold u and spatial scale l, the mean square error (*MSE*) is calculated as $MSE_{u,l} = \overline{Z_{u,l}^2}$ where the overbar denotes the average. Also, random *MSE* is estimated through the following equation:

$$MSE_u^{random} = FBIAS_u \times BR_u \times (1 - BR_u) + BR_u \times (1 - FBIAS_u \times BR_u) \quad (2)$$

where $FBIAS_u$ is the frequency bias and BR_u is the base rate (which is calculated as the probability of observed precipitation occurrence within the verification domain) (Stratman et al., 2013) for a given threshold u. The ISS score is then calculated as a function of precipitation intensity and the spatial scale of the errors in terms of *MSE*,

$$ISS_{u,l} = 1 - \frac{MSE_{u,l}}{MSE_u^{random}/(L+1)} \quad (3)$$

Positive ISS values are associated with skillful forecasts, whereas negative ISS values are associated with no forecast skill (Casati, 2010).

2.2. Intensity-scale-based fractions skill score

The ISS score is still based on point-to-point comparisons. Displacement error will have a significant impact on ISS, especially for small spatial scales. For the intensity-scale-based fractions skill score we propose in this section, we keep the procedure of ISS, including the steps for obtaining the binary field and wavelet decomposition, but we do not calculate ISS in the final step. Instead, the fractions skill score (FSS, Roberts, 2008; Roberts and Lean, 2008), which is based on the idea of neighborhood verification (Ebert, 2008), is used to assess the forecast skill of each wavelet component. In doing so, the impact of displacement error on the verification score can be considered. To distinguish from the original ISS verification score, we refer to this score as intensity-scale-based fractions skill score or IS_FSS.

During the score calculations, wavelet decomposition is applied to the binary fields of observation I_X and forecast I_Y separately. The squared value of the binary wavelet decomposition term in the forecast/observation field $I_{u,l}{}^2$ in section 2.1 can be regarded as energy for a given threshold and spatial scale (Garg et al., 2011). Once $I_{u,l}$ is obtained, the binary field for a given threshold and spatial scale is calculated as follows. If gridded energy $I_{u,l}{}^2$ is non-zero, it is assigned 1; otherwise, it is assigned 0. The "yes" event occurrence probability of forecast, $P_{u,l}{}^{fcst}$, and of observation, $P_{u,l}{}^{obs}$, are then calculated within the $n \times n$ grid box. IS_FSS is calculated as follows:

$$FSS_{u,l} = 1 - \frac{\frac{1}{N}\sum_N (P_{u,l}^{fcst} - P_{u,l}^{obs})^2}{\frac{1}{N}\left(\sum_N (P_{u,l}^{fcst})^2 + \sum_N (P_{u,l}^{obs})^2\right)} \quad (4)$$

$$IS\_FSS_{u,l} = FSS_{u,l} - FSS_{u,l}^{useful} \quad (5)$$

where N is the number of neighborhood grid points for each neighborhood size; and $FSS_{u,l}{}^{useful} = 0.5 + f_{u,l}{}^o/2$, where $f_{u,l}{}^o$ is the observation hit rate (Roberts and Lean, 2008). It takes into account the random forecast skill, which is also treated as a reasonable "target skill". Following the regular $FSS_{u,l}$, the forecast is considered skillful when $IS\_FSS_{u,l} > 0$, whereas it has no skill when less than zero (Roberts and Lean, 2008). $IS\_FSS_{u,l}$ is applied to each wavelet scale l for a given threshold u, as the regular ISS score does (see Eq. (3)).

2.3. Percentile-based thresholds for IS_FSS

Many studies have pointed out that convection-allowing models tend to over-predict the precipitation intensity (Murata et al., 2016; Karki et al., 2017; Zhu et al., 2018). To be able to better assess a model's ability in predicting precipitation pattern and distribution, forecast bias often needs to be removed before calculating scores (Hamill, 1999). Zhu et al. (2018) used a bias-corrected Gilbert skill score (GSS, Gandin and Murphy, 1992) to assess forecasts from models with different resolutions. To reduce the impact of the intensity bias on the verification score, we replace the fixed threshold with a percentile-based threshold (Roberts and Lean, 2008; Zhu et al., 2015) during the initial determination of the binary fields (Eq. (1)) when calculating IS_FSS. The rest of the calculation is the same as with IS_FSS, and we refer to this verification score as percentile-based IS_FSS or ISP_FSS. In this study, four thresholds at the 75th, 90th, 95th and 99th percentiles are used in place of the 4, 13, 25 and 60 mm thresholds for 6-hour accumulated rainfall. Similar percentile thresholds were used by Roberts and Lean (2008). For calculating the kth percentile, we first sort the data in ascending order. The kth percentile is a value on a scale of 100 that indicates the percent of a distribution that is equal to or below it.

3. Data and methods

3.1. Forecasts and observations

Since 2013, a convection-permitting resolution model that covers the entire Chinese mainland has been run at Nanjing University, China, in real-time, twice a day at 0000 and 1200 UTC, during the June to August months, starting from analyses of the NCEP operational Global Forecasting System (GFS), and forced at the lateral boundaries by GFS forecasts. The system employs WRF as the forecast model and has a horizontal grid resolution of 4 km (hereafter referred to as WRF-4). A single grid covering the entire continental China was used. The configurations of WRF-4 are described in detail in (Zhu et al., 2018). Briefly, the physics schemes used are the Morrison double-moment microphysics scheme (Morrison et al., 2005), the Asymmetrical Convective Model planetary boundary layer scheme, version 2 (Pleim, 2007), the Pleim–Xiu land-surface model (Pleim and Xiu, 1995), the Pleim–Xiu surface layer schemes (Pleim, 2006), and the CAM short- and long-wave radiation schemes (Collins et al., 2004). No cumulus parameterization was employed. Detailed assessments of the 2013–2014 forecasts show that WRF-4 outperforms many global operational forecasts in terms of spatial distribution, intensity, and diurnal variations of precipitation over China. The convection-allowing resolution was believed to be an important contributor (Zhu et al., 2018).

To investigate the sensitivity of precipitation forecasts to grid spacing, WRF forecasts with a 12 km grid spacing (WRF-12) was run for the summer of 2016. In terms of model physics, WRF-12 was configured the same way as WRF-4, except that Grell-3 (Grell, 1993) cumulus parameterization was used in addition. The WRF-12 domain was somewhat larger than that of WRF-4 (Fig. 1). In this study, the performances of WRF-12 and WRF-4 are compared in terms of precipitation forecast skills using various evaluation metrics.

The observations used for precipitation verification are the $0.05^\circ \times 0.05^\circ$ hourly merged precipitation products/analysis (Shen et al., 2014) (over 70°–140°E, 15°–60°N) from the National Meteorological Information Center of the China Meteorological Administration (CMA). This precipitation dataset is a merged product from three gridded observations, including: $0.05^\circ \times 0.05^\circ$ Chinese Precipitation Analyses, which are generated using ~40,000 rain gauge observations (Zheng et al., 2016), $0.01^\circ \times 0.01^\circ$ gridded radar-estimated precipitation (Zheng et al., 2016), and NOAA CMORPH (Climate Precipitation Center Morphing) precipitation product, with an 8-km grid spacing, estimated from multiple satellite sensors (Joyce et al., 2004; Shen et al., 2010). Verifications at independent stations show that this merged precipitation product has smaller RMSEs than any of the three gridded precipitation products (Pan et al., 2015). This high-resolution gridded precipitation product allows for grid-based verifications, especially those employing scale decomposition.

3.2. Data preprocessing and choice of verification grid

For the intensity-scale verification employing the Haar wavelet, a $2^n \times 2^n$ square grid is required. Here, a 128×128 ($n = 7$) grid with a 12 km grid spacing is selected as the verification domain (see Fig. 1, gray area). This area, including south China and the region along the Yangtze River, are two major heavy rain areas during the summer. All the observations and forecasts are then interpolated to the same verification domain using the neighborhood budget interpolation method (Zheng et al., 2016). This method ensures area conservation for total precipitation.

The merged CMA gridded precipitation product has a 0.05° or about 5 km horizontal grid spacing. In principle, it is suitable for verification of our 4-km WRF forecasts. However, because the products were derived from satellite, rain gauge and radar observations of various

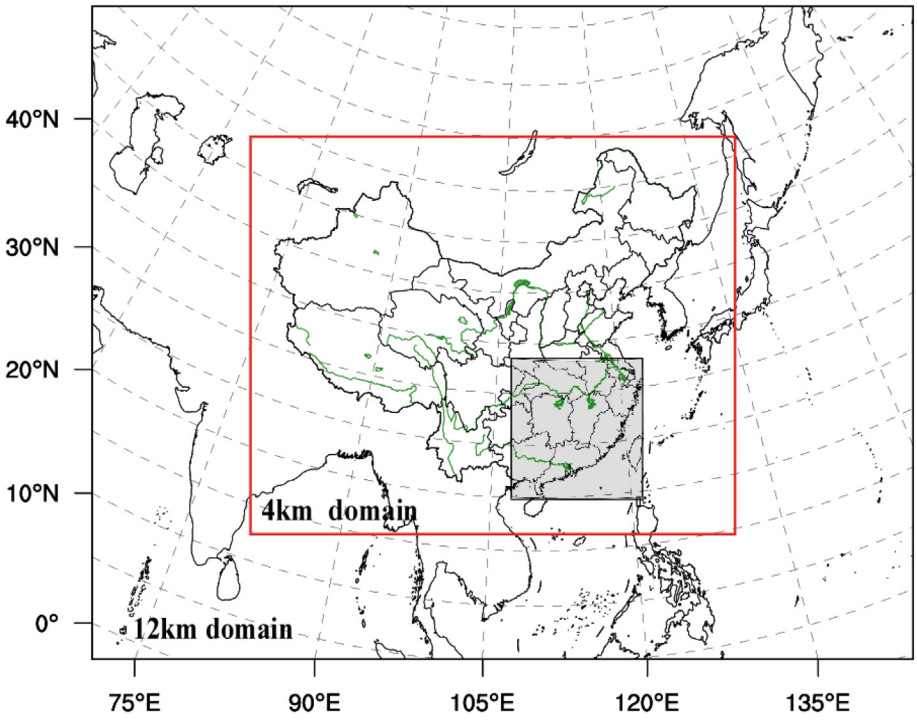

Fig. 1. The outer box is the domain for the 12 km forecast with 600 × 800 grids, and the inner red box is for the 4 km forecast with 1080 × 1408 grids. The gray-shaded square is a domain with 128 × 128 grid points used for verification. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

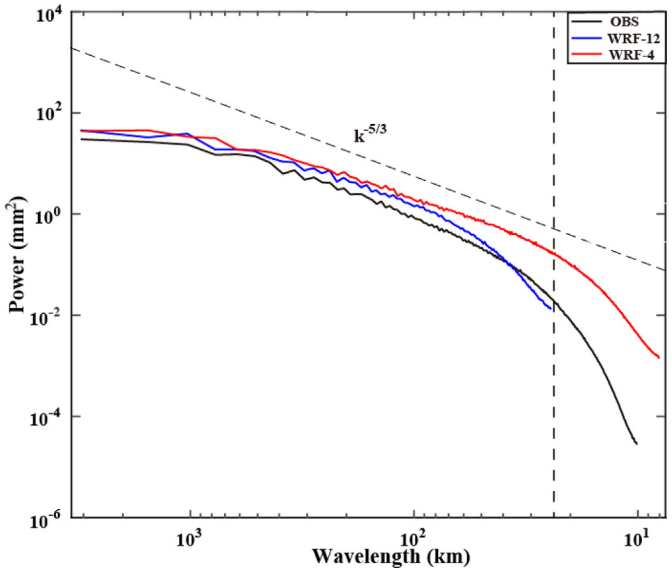

Fig. 2. Power spectra of CMA precipitation product (black), WRF-12 (blue), and WRF-4 (red) forecast precipitation over 24 h accumulation periods, averaged over June–August of 2016. Also plotted as dashed line is a − 5/3 slope expected of the power spectra of mesoscale to convective scale motion (Skamarock, 2004). The vertical dashed black line indicates where the wavelength equals 24 km. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

resolutions, the product may or may not resolve precipitation details up to the maximum resolution allowed by the gridded dataset, i.e., a wavelength of two grid spacings. To see what scales are actually well resolved in the CMA precipitation product, we examine its energy spectra. The average energy spectra of the CMA precipitation fields for summer of 2016, as well as those of WRF-12 and WRF-4 24 h precipitation forecasts, are computed using a two-dimensional discrete cosine transform (deElia et al., 2002). Here, the observation domain for the spectral analysis is slightly larger than the verification domain. This size discrepancy results because spectral analyses are applied to the native grid. We use the minimum latitude-longitude rectangle surrounding the verification domain as the observation analysis domain. The results of the spectral analyses are plotted in Fig. 2. For convective scale to mesoscale motions, with wavelengths in the range of ~1 km to ~100 km, the kinetic energy spectra are expected to have a slope of −5/3 in the log-log space (Lindborg, 1999; Skamarock, 2004). Studies (e.g., Surcel et al., 2014; Snook et al., 2019a) show that high-resolution observed precipitation estimates maintain the −5/3 slope down to kilometer wavelengths. A line with the −5/3 slope is plotted in Fig. 2 for reference.

It can be seen in Fig. 2 that the precipitation energy spectra for WRF-4 maintain a − 5/3 or a somewhat shallower slope from wavelengths of about 200 km to about 20 km (~5Δx) while that of WRF-12 starts to drop off at about 100 km. The curve for the CMA precipitation product (black line) has a slope close to −5/3 between 500 km and 40 km then starts to drop off quickly. This drop-off in slope suggests that the CMA precipitation product, even at a ~ 5 km grid spacing (and being able to resolve up to 10 km wavelength), is lacking energy for wavelengths below ~40 km. Such spectral drop-off is likely due to insufficient resolution of the rainfall measurements and spatial smoothing involved in analyzing the observations on a regular grid (Surcel et al., 2014; Wong and Skamarock, 2016; Snook et al., 2019b). For this reason, it would be unfair to perform verification of WRF-12 and WRF-4 forecasts on the native ~5 km grid of the CMA precipitation product; doing so will actually negatively penalize WRF-4, because it correctly predicts more energy at the shorter wavelengths that is actually missing in the verification data (see differences between the black line and the red line below 24 km wavelength). Therefore, in this study, we interpolate both WRF-4 precipitation forecasts and CMA merged precipitation data to the 12 km WRF-12 grid, which permits wavelengths down to 24 km wavelength. At 12 km grid spacing, the spectra of both CMA products and WRF-4 are relatively well-resolved.

4. Verification results for a Meiyu front case

In this section, we examine the behaviors of various verification metrics when applied to the WRF-12 and WRF-4 precipitation forecasts, using a typical Meiyu front case as an example. Associated with the

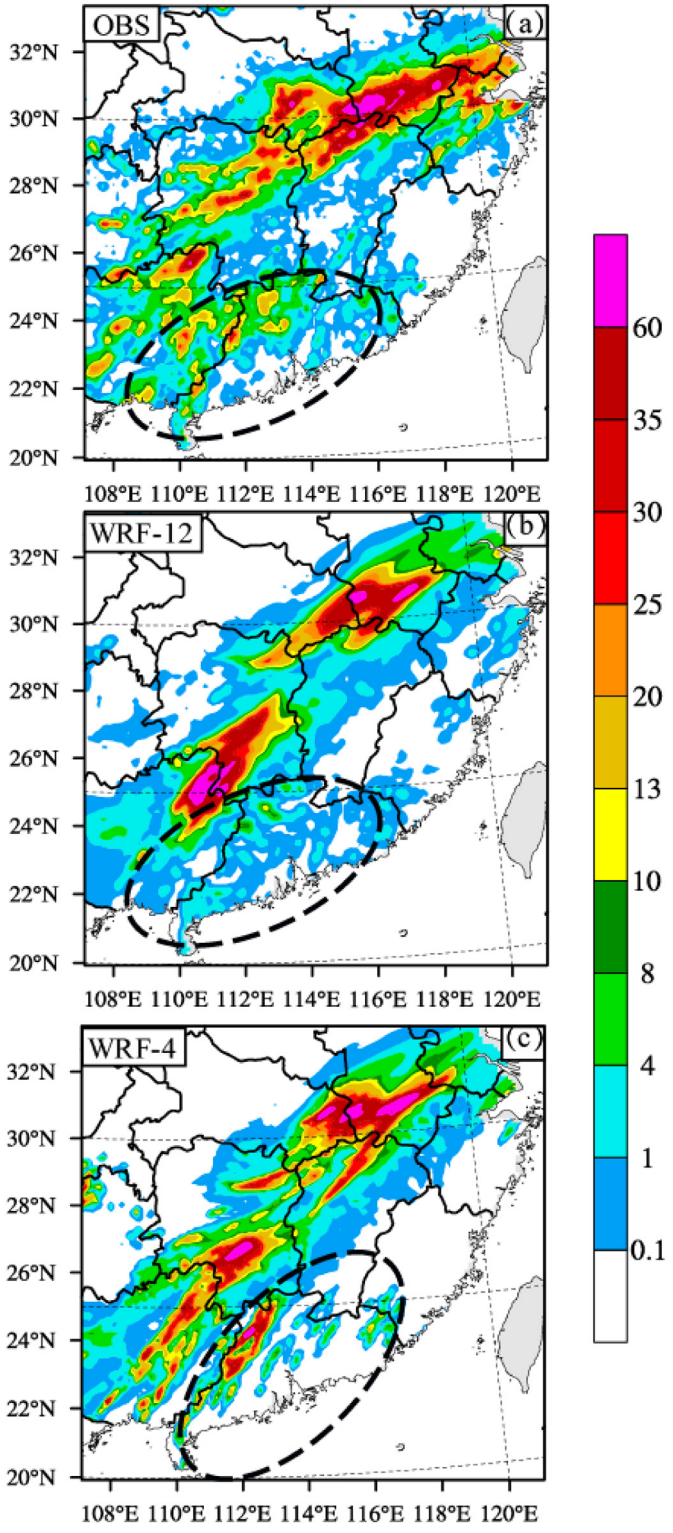

Fig. 3. The 6-h accumulated rainfall from 1400 LST to 2000 LST on 1 July 2016 from (a) observations, (b) WRF-12, and (c) WRF-4 forecasts.

Meiyu front is a loosely organized mesoscale rainband with convective elements embedded within and outside. In other words, the precipitation fields contain multi-scale structures.

4.1. Results of original ISS score

Fig. 3 shows the 6-h accumulated rainfall from July 1, 2016, associated with a typical Meiyu front. It can be seen that in the observations (Fig. 3a) the main rainband has a southwest–northeast orientation with a heavy north precipitation center near (31°N, 116°E) and a smaller area of intense precipitation at the southwestern end of rainband (~26°N, 110°E). At the same time, in south China there are many small areas of more scattered precipitation that are associated with afternoon convection (enclosed by the ellipse), which is relatively hard to predict. In both WRF-4 and WRF-12, the northern portion of the rainband with heavy precipitation is predicted quite well, with about the right orientations and the maximum precipitation locations. The precipitation near the southwestern end is much over-predicted in both coverage and intensity in WRF-12, however, and the center location is also off to the south by about 100 km (Fig. 3b). In comparison, the structure and size of this area of precipitation are predicted quite well in WRF-4, although the location is displaced northeastward somewhat (Fig. 3c). Moreover WRF-4 is able to capture some of the precipitation associated with the after convection over the hilly southern China (in the north part of the ellipse in Fig. 3c) while WRF-12 misses essentially all of the heavy precipitation there (Fig. 3b), presumably due to the relatively poor performance of cumulus parameterization scheme for capturing convective precipitation (Qiao and Liang, 2015).

Using the original ISS, both WRF-12 and WRF-4 show positive scores for spatial scales from 192 km and up (except for a low threshold of WRF-12) but mostly negative scores for spatial scales below 96 km (Fig. 4). For scales at and above 192 km, WRF-4 clearly performs better than WRF-12, having higher or perfect scores. Similar to the findings of Mittermaier (2006) comparing 4 and 12 km forecasts, the negative region of WRF-4 is smaller than that of WRF-12, suggesting that WRF-4 has better forecast guidance. However, both forecasts have negative values for small scales, making it hard to tell which one is better.

To understand the behaviors of the ISS as applied to the current case, we plot in Fig. 5 the wavelet components of different thresholds for observation and forecasts. Here, only spatial scale 12 km is shown. The summed patterns of different thresholds are close to their corresponding observation and forecasts in Fig. 3, i.e., 4 mm mainly corresponds to the outer boundary of the rainband (Fig. 5a-c) while 60 mm corresponds to the heavy-rain center (Fig. 5i-l). In general, the results of WRF-4 appear closer to the observation. For all thresholds, the wavelet components of WRF-4 get a similar number of nonzero points (colored) as observed, suggesting that WRF-4 has smaller bias than WRF-12, consistent with the subjective assessment of Fig. 3. Take the 13 mm threshold for an example (Fig. 5d-f), the number of nonzero points of WRF-4 is closer to the observation while WRF-12 significantly underestimates the points not only for the main rainband but also for the convection south of the main rainband (compare the circled area in Fig. 3). WRF-12 fails to predict the southern convection entirely.

Given that ISS measures agreement between forecasts and observations, it is still a “point-to-point” verification. The observation and forecast pairs usually have serious “double-penalty” issues which will significantly lower the ISS score. The negative ISS of WRF-4 is mainly due to the position error. Even for the smallest threshold of 4 mm per 6 h (Fig. 5a, c), a perfect match between forecast and observation pairs (corresponding mainly to the rain band boundary) is not possible at the 12 km scale. For this case, owing to position errors, WRF-4 produces larger negative ISS scores than WRF-12 because of the “double-penalty” issue. In comparison, WRF-12 also suffers from position biases but at fewer points, the “double-penalty” impact is therefore less, leading to less negative ISS scores than that of WRF-4. The fewer number of points exceeding the threshold is because WRF-12 has significant intensity low biases. The energy for different thresholds at the 12 km scale is clearly underestimated for WRF-12 (see first and second columns in Fig. 5). Scales from 24 km to 96 km have similar behaviors except that the bigger the scale the less the effect of position error (not shown). In all, ISS at small spatial scales is likely to be strongly affected by position error, which is a major cause of negative ISS. In this case, the intensity bias is another major error of WRF-12. These above analyses suggest

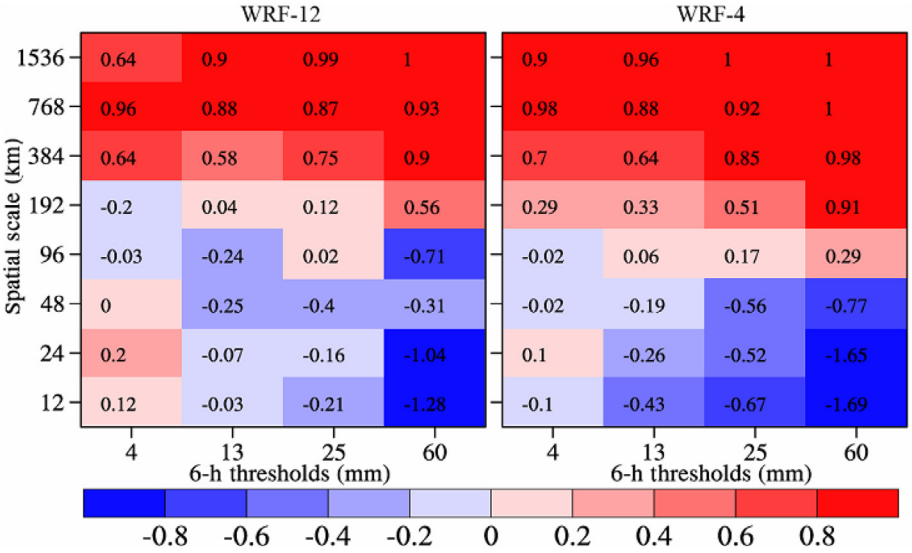

Fig. 4. Original ISS scores of the 6-h accumulated rainfall from 1400 LST to 2000 LST on 1 July 2016 for (a) WRF-12 and (b) WRF-4 forecasts. ISS ≤0 no forecast skill; 1 perfect.

that scores allowing for position and intensity errors are needed when the predicted pattern is a major concern.

4.2. IS_FSS and ISP_FSS results

In this subsection, the alterative fractions skill score FSS-based scores, as described in sections 2.2 and 2.3, are evaluated. With this approach, the binary field determination and wavelet decomposition remain the same as with the original intensity-scale approach, while the ISS is replaced with the widely used probability evaluation score, the FSS. Because FSS considers occurrence of events within a given radius or neighborhood, the displacement can be considered. The IS_FSS and ISP_FSS can be regarded as the probability forecast evaluation at different spatial scales.

Fig. 6 shows the IS_FSSs for 6-h precipitation forecasts with FSS neighborhood widths of 1, 3, and 9 grid intervals. Note that the verification domain is 1536 × 1536 km (128 grids × 12 km grid spacing), therefore the wavelets of spatial scales above 96 km only have a few pixels of data. For example, for the spatial scale of 192 km, there are only 8 × 8 wavelet pixel blocks, it is therefore unsuitable for using a neighborhood width of 9. Therefore, only spatial scales from 12 to 96 km are shown in Fig. 6. For the neighborhood width of 1 grid point, the overall pattern is similar to that of original ISS, with mostly negative IS_FSSs for both WRF-12 and WRF-4 forecasts, suggesting no useful forecasting skill for most spatial scales and thresholds (Fig. 6a, b). WRF-4 has a slightly smaller negative region than WRF-12. The increase in neighborhood width greatly increases the verification score. For the neighborhood width of 3 grid points, more than half of the WRF-4 scores at different spatial scales and thresholds are positive (Fig. 6d). For a neighborhood width of 9 grid points, only two values are negative for WRF-4 (in Fig. 6e), corresponding to large thresholds of 25 mm and 60 mm and small spatial scale of 12 km. In comparison, the IS_FSSs of WRF-12 are also improved, but not as much as for WRF-4. As discussed earlier, WRF-4 better predicts some small-scale precipitation features while WRF-12 misses many of them. WRF-12 suffers from significant intensity and position errors while the errors of WRF-4 are mainly in position. Compared with the original ISS in Fig. 4, which shows mostly negative scores for both WRF-4 and WRF-12 at spatial scales below 96 km, the IS_FSS gives a greater proportion of positive areas as the neighborhood width increases for both WRF-4 and WRF-12. Meanwhile, qualitative differences between WRF-4 and WRF-12 are much better reflected through IS_FSS than through the original ISS.

Intensity error can affect evaluation scores also, and in some cases in undesirable ways. For example, the GSS can be artificially inflated by over-forecasting (Hamill, 1999). To better assess the predicted patterns and structures of, for example, precipitation, it is often desirable to remove the forecast intensity bias first, at least to separate the effects of bias on skill scores (Hamill, 1999). One of the ways to alleviate the effects of intensity bias is to use percentile-based thresholds instead of fixed threshold values (e.g., Zhu et al., 2015). In Fig. 7, percentile-based thresholds are employed to obtain the initial binary observation and forecast fields to obtain ISP_FSSs for the WRF forecasts with neighborhood widths of 1, 3 and 9 grid points. Compared to fixed-threshold IS_FSS, ISP_FSS greatly increases the FSSs of WRF-12 for neighborhood widths of 3 and 9 grid points. It can be seen that most patches in Fig. 7c and e turn positive, especially for low thresholds. For the observation, the 75th, 90th, 95th and 99th percentiles of this case correspond to 8.8, 20.4, 28.9 and 56.8 mm, respectively, while for the WRF-12 forecasts they are 3.2, 14.3, 31.0 and 61.4 mm, respectively. The light rains (75th, 90th) are underestimated, while heavy rains (95th, 99th) are overestimated. The use of a percentile-based threshold reduces the effects of intensity bias for both light and heavy rains. Note that the area coverage of light rain is much larger than that of heavy rain. Therefore, it is unsurprising that the ISP_FSS increases more for low thresholds than high thresholds for WRF-12. For WRF-4, the corresponding thresholds are 8.0, 21.1, 32.9 and 61.2 mm respectively. The light rains are very close to observations while the heavier rains are overestimated. With the use of a percentile-based threshold, the FSS is further improved for most thresholds and scales, especially for WRF-12. Comparing the two resolutions, WRF-4 is still better than WRF-12 in terms of ISP_FSS. However, the latter benefits more from the use of a percentile-based threshold because it has more intensity biases.

Overall, without a neighborhood, i.e., with a width of one grid interval, both IS_FSS and ISP_FSS are mostly negative for most spatial scales and thresholds. Increasing the neighborhood width turns mostly negative FSSs to positive, suggesting that the position error significantly affects the verification score. The use of a percentile-based threshold further reduces the contribution of intensity bias. In this case, WRF-12 forecasts benefit considerably because of its larger intensity bias, while for WRF-4 lower thresholds see a benefit. Based on the above results, IS_FSS is a good choice in distinguishing the quality of WRF forecasts at 4 and 12 km grid spacings, while ISP_FSS better reflects the quality of predicted rainfall pattern and structure without being affected significantly by intensity biases. In the next section, we further examine

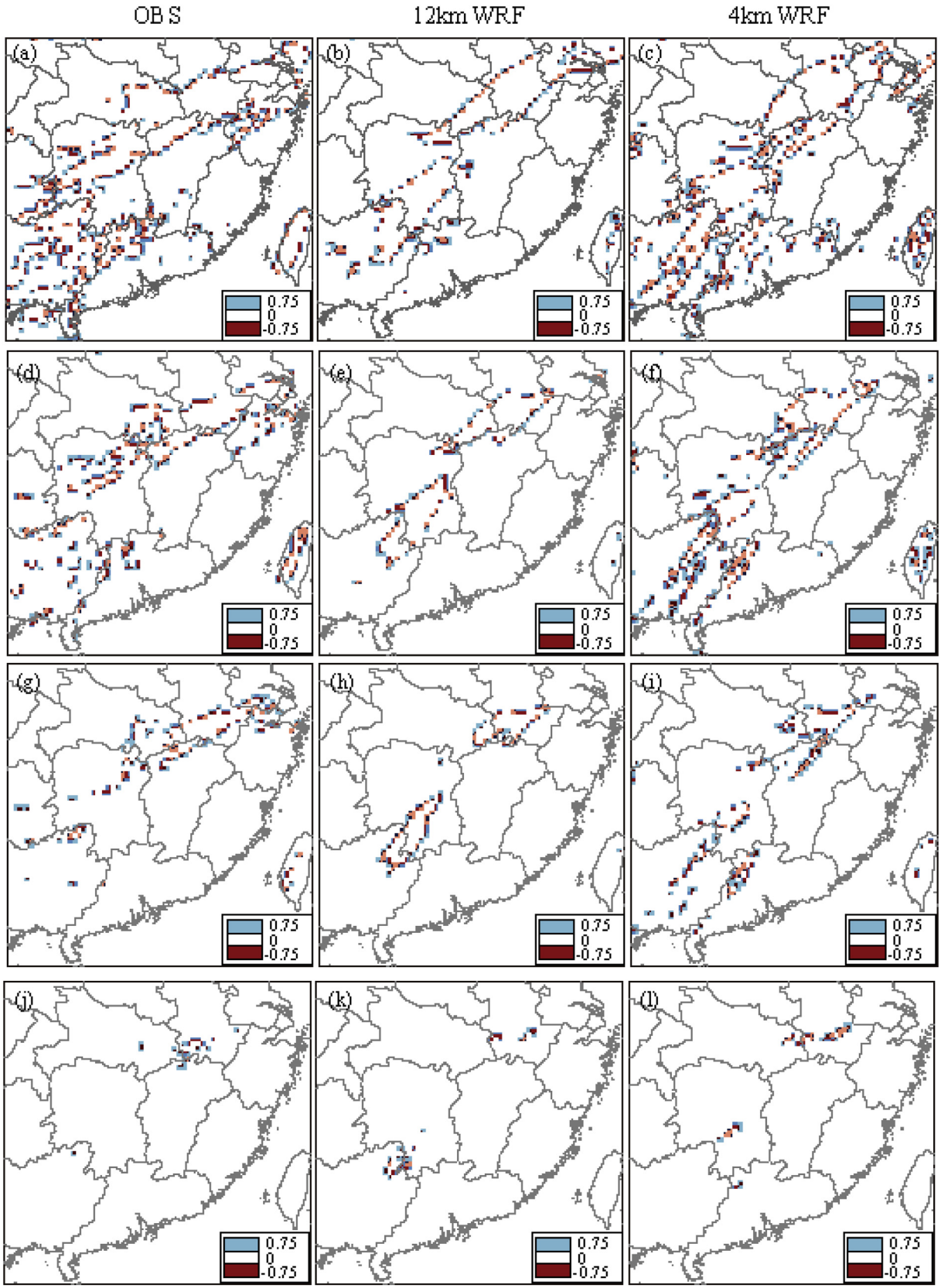

Fig. 5. Wavelet scale component at the 12 km of 6-h accumulated rainfall from 1400 LST to 2000 LST on 1 July 2016 for observations (left column) and forecasts (middle column for WRF-12 and right column for WRF-4) at different thresholds: (a–c) 4 mm, (d–f) 13 mm, (g–i) 25 mm, and (j–l) 60 mm.

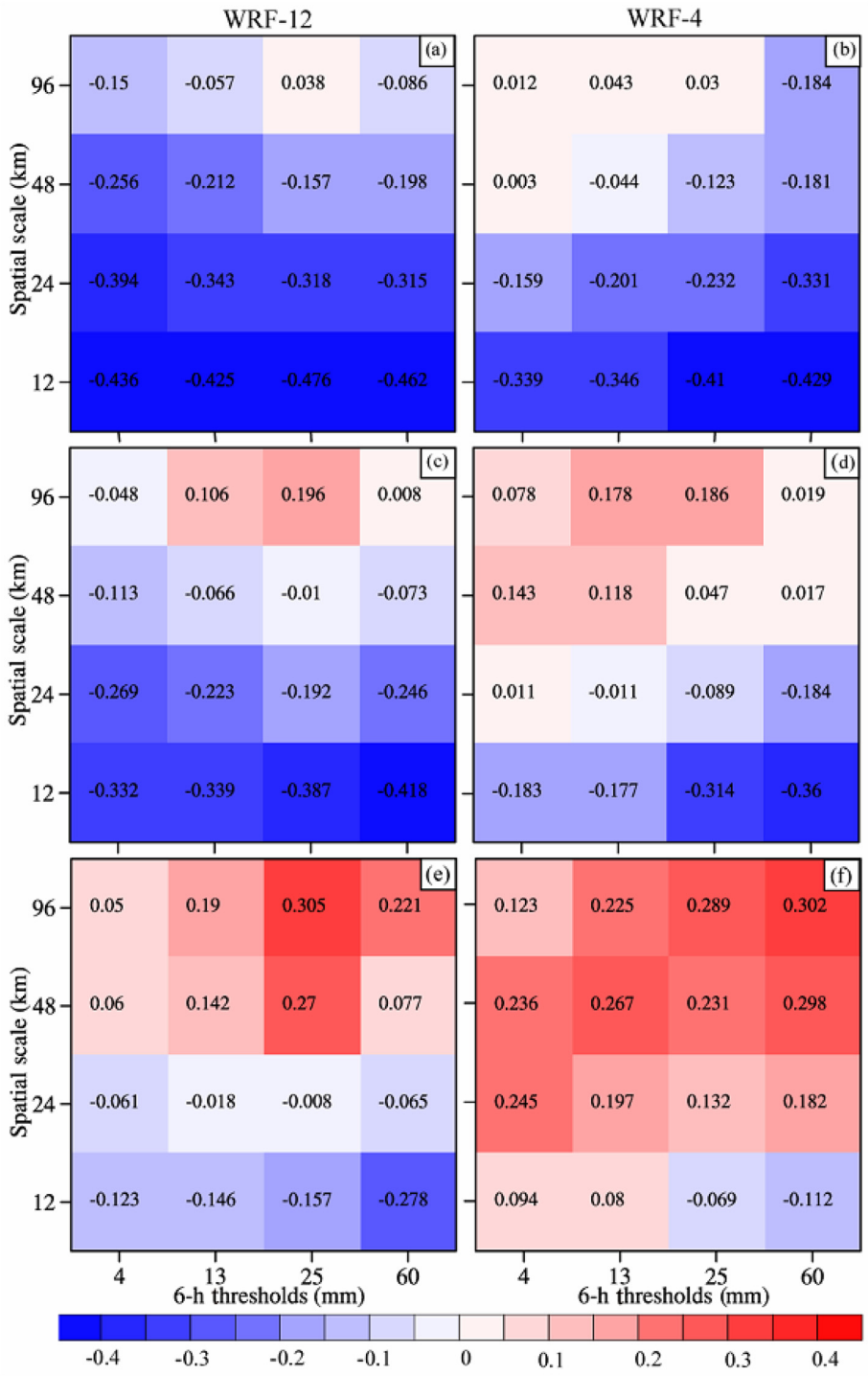

Fig. 6. IS_FSSs of 6-h accumulated rainfall from 1400 LST to 2000 LST on 1 July 2016 for (a) WRF-12 forecast and (b) WRF-4 forecast, for neighborhood widths of 1 (upper row), 3 (middle row) and 9 (bottom row) grid points. IS_FSS ≤ 0 no forecast skill; 1 perfect.

IS_FSS and ISP_FSS by applying them to precipitation forecasts of a full Meiyu season.

5. Verification scores for an entire Meiyu season

In this section, we apply ISS, IS_FSS and ISP_FSS to an entire 2016 Meiyu season in China, from June 11 to July 20. Within this period, 5 days are influenced by typhoon and are excluded.

Fig. 8 displays the time series of observed and simulated hourly rainfall averaged over the entire Meiyu season within the verification domain. All forecasts begin from 1200 UTC on each day. The first 12-h forecasts are not evaluated because of the spin-up of model forecast. The observation shows a clear bimodal pattern, with an afternoon peak at around 1600 LST and a morning peak at 0800 LST. For the WRF-4 forecasts, the diurnal variation is reproduced well, but the afternoon peak is underestimated and the morning peak overestimated. The former corresponds to afternoon thermal convection that is harder to predict. The early morning peak is explained by Xue et al. (2018) as being caused by boundary layer inertial oscillations that produce the strongest low-level convergence forcing along the Meiyu front in early

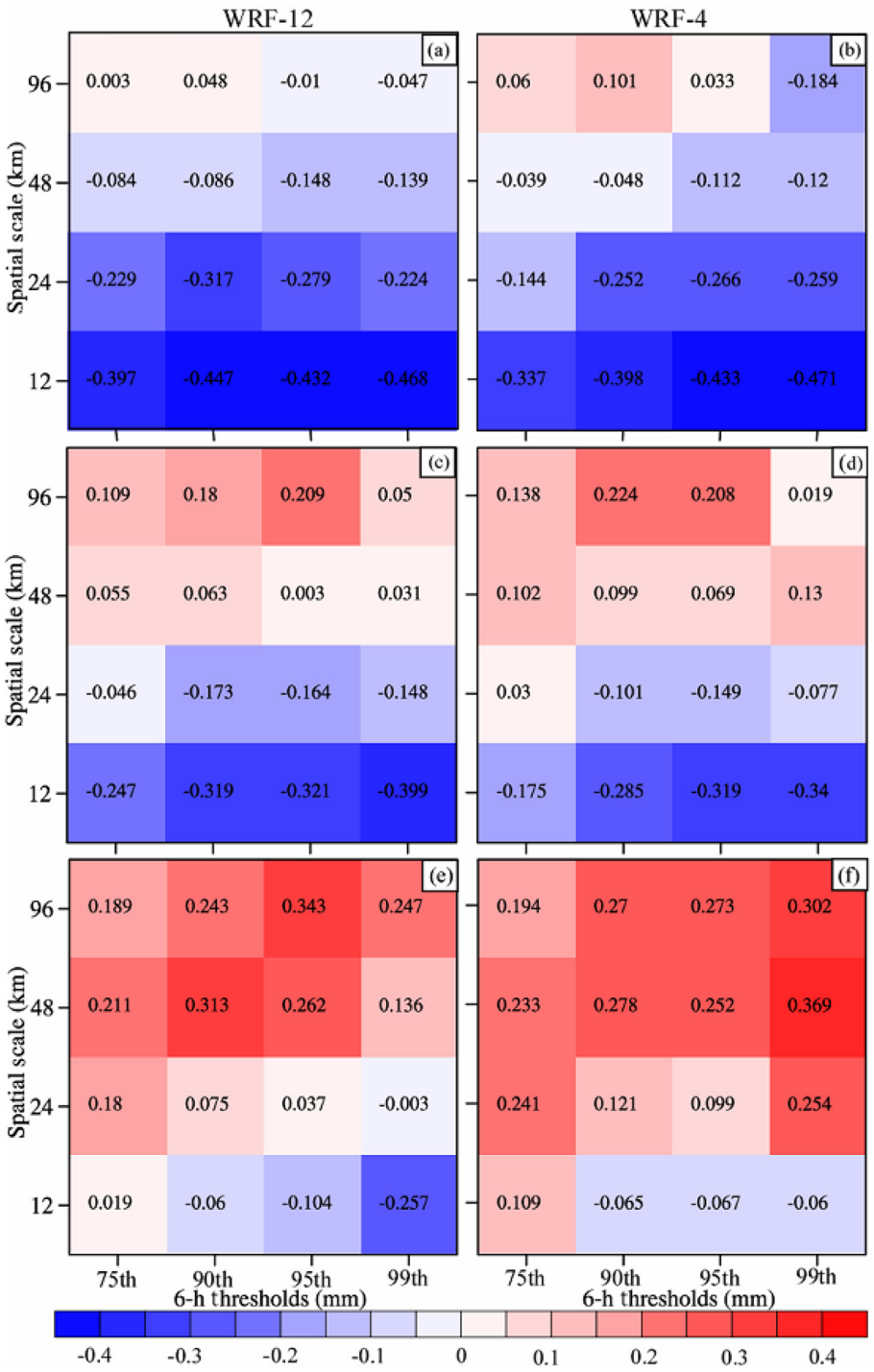

Fig. 7. As in Fig. 6 but for ISP_FSS. The numbers on the *x*-axis represent the percentile thresholds. ISP_FSS ≤0 no forecast skill; 1 perfect.

morning hours. In the WRF-12 forecasts, the afternoon peak is mostly missing. Earlier studies, such as Clark et al. (2009), have also shown that convection-permitting forecasts have much better skill in capturing precipitation diurnal cycles than forecasts that rely on cumulus parameterization. In the following, the afternoon period (1400–2000 LST), when the two forecasts show the greatest difference, is focused on.

Fig. 9 shows the original ISSs for the entire Meiyu period. Here, the ISS score for each spatial scale is calculated using the summed components of all cases (Ebert, 2009). As in the case study, although we can clearly see better forecasts in WRF-4, it is hard to state in terms of the ISS that WRF-4 is better than WRF-12, since most of the patches at the spatial scales from 12 km to 96 km are negative. For larger thresholds and smaller spatial scales (right-bottom corner), similar to the case study, WRF-4 shows larger negative ISS scores than WRF-12, indicative of larger forecast error. Fig. 10 shows the IS_FSS results. Note that we only display the spatial scales from 12 km to 96 km. Most of the afternoon precipitation is from scattered afternoon thermal convection of relatively small spatial scales. Without consideration of a neighborhood width, the IS_FSSs of WRF-4 and WRF-12 are all negative, indicating no useful forecast skill. Similar to the case study, increasing the neighborhood width increases IS_FSS. It can be seen that for a neighborhood width of 9 grid points, most of the WRF-4 patches turn positive. Again, as in the case study, WRF-4 produces clearly higher IS_FSSs than WRF-12, suggesting better prediction skills of the former.

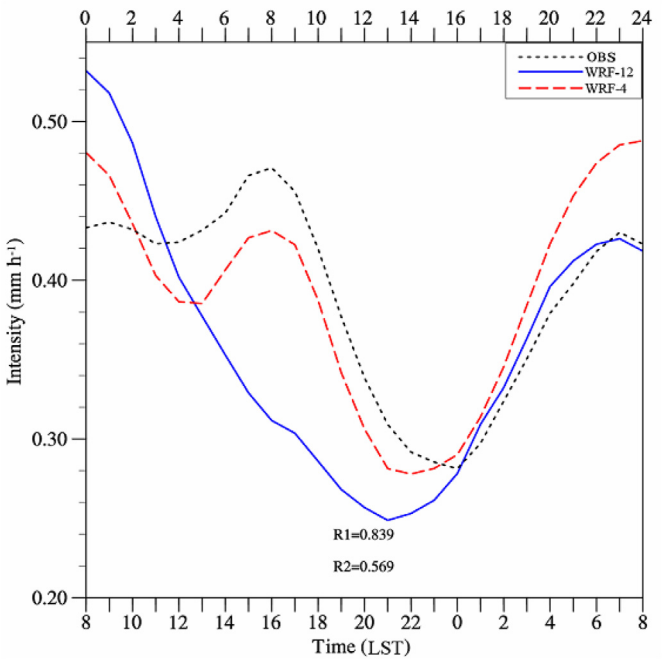

Fig. 8. Diurnal variation of hourly precipitation for the observation and 12-h to 36-h WRF forecasts starting from 12 UTC, averaged over verification domain (gray-shaded area) in Fig. 1 for the entire Meiyu season from June 11 to July 20, 2016. Here, about 5 days under the influence of typhoon are excluded for the calculation. All the verification scores bellow are also calculated without typhoon days.

Fig. 11 displays the ISP_FSS results. For all forecasts and observations, the 75th, 90th, 95th and 99th percentiles are calculated from all the precipitation data points during the Meiyu season. They are: (5.9 mm, 14.2 mm, 21.8 mm, 43.7 mm), (2.6 mm, 10.6 mm, 21.4 mm, 54.4 mm) and (7.4 mm, 19.9 mm, 31.5 mm, 65.3 mm) for the observation and WRF-12 and WRF-4 forecasts, respectively. Compared to IS_FSS, ISP_FSS reduces the effects of intensity bias, and therefore increases the FSSs at most patches, especially for WRF-12. For both WRF-4 and WRF-12, increasing the neighborhood width further increases the FSS. The change of WRF-12 is more obvious because it significantly underestimates the precipitation intensity in the afternoon (also see Fig. 8). Comparing the forecasts at the two different resolutions, using a neighborhood width of 9 grid points, again WRF-4 performs better than WRF-12 for all spatial scales and thresholds.

6. Summary

In 2016, the WRF model was run at Nanjing University, China, at grid spacings of 12 km and 4 km (WRF-12 and WRF-4 respectively) during the summer period from June through August, in realtime for the 4 km grid. Subjective assessment suggests that WRF-4 performs better than WRF-12 in predicting precipitation, especially for smaller-scale afternoon convection. In this paper, a wavelet-based multi-scale intensity-scale skill (ISS) score is first applied to a representative Meiyu precipitation case to check if such an objective verification can give similar conclusions as subjective assessment. It turns out that the ISS scores do not show better performance of WRF-4 compared to WRF-12 at small spatial scales where WRF-4 should have clear advantage based on subjective evaluations. The ISSs at scales from 12 to 96 km are mostly negative for both WRF-4 and WRF-12. For higher precipitation thresholds and smaller spatial scales, WRF-4 even shows larger error than WRF-12. This is because ISS is a point-to-point verification method where position error will lead to a serious “double-penalty” for high resolution forecasts. Because the gridded precipitation data lack full power below 40 km based on power spectra analysis, all objective verifications are performed for scales at and above 12 km in this paper.

For practical purpose, we care about whether a model has the ability to correctly predict detailed structures of precipitation fields, including small-scale features associated with localized convection. At such scales, position biases are almost unavoidable but do not necessarily deem the forecasts useless. In fact, correct prediction of the spatial extent, structure and intensity of localized intense precipitation can be very useful even with position errors. Two methods are proposed for the purpose of reducing the negative impacts of position error on the verification score. Here, the neighborhood fuzzy matching idea is borrowed. The first method involves replacing the ISS with the widely used fractions skill score-based neighborhood verification score, or FSS, while all other steps of the original ISS remain the same. We name this score IS_FSS. The second method is similar to IS_FSS, except that in the step to obtain the initial binary fields, the threshold as a fixed value is replaced by a percentile-based threshold. In doing so, the impact of forecast intensity bias can be reduced. We name this score percentile-based IS_FSS or ISP_FSS. These two new scores can be interpreted as probability forecast skills within a neighborhood range for different spatial scales.

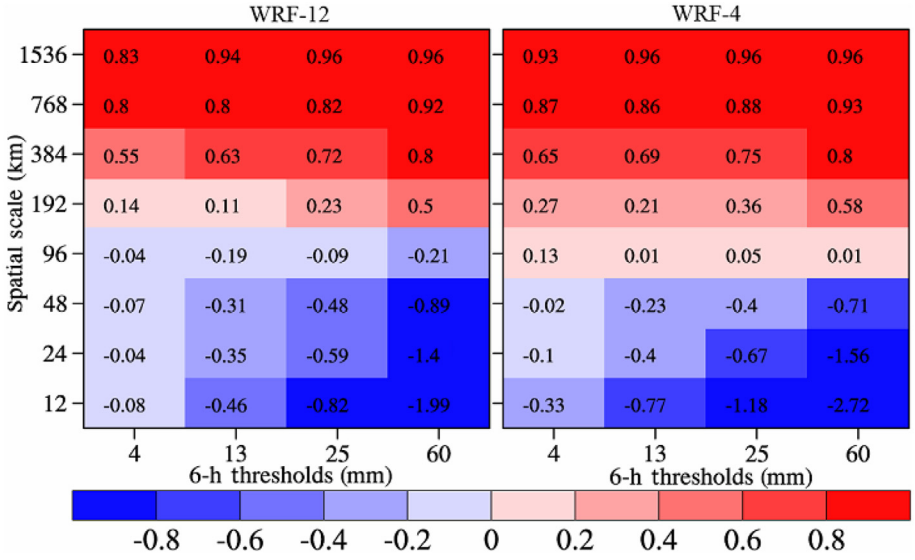

Fig. 9. As in Fig. 4 but for 6-h (1400–2000 LST) accumulated rainfall forecasts of WRF-12 and WRF-4, of the entire Meiyu period of 2016.

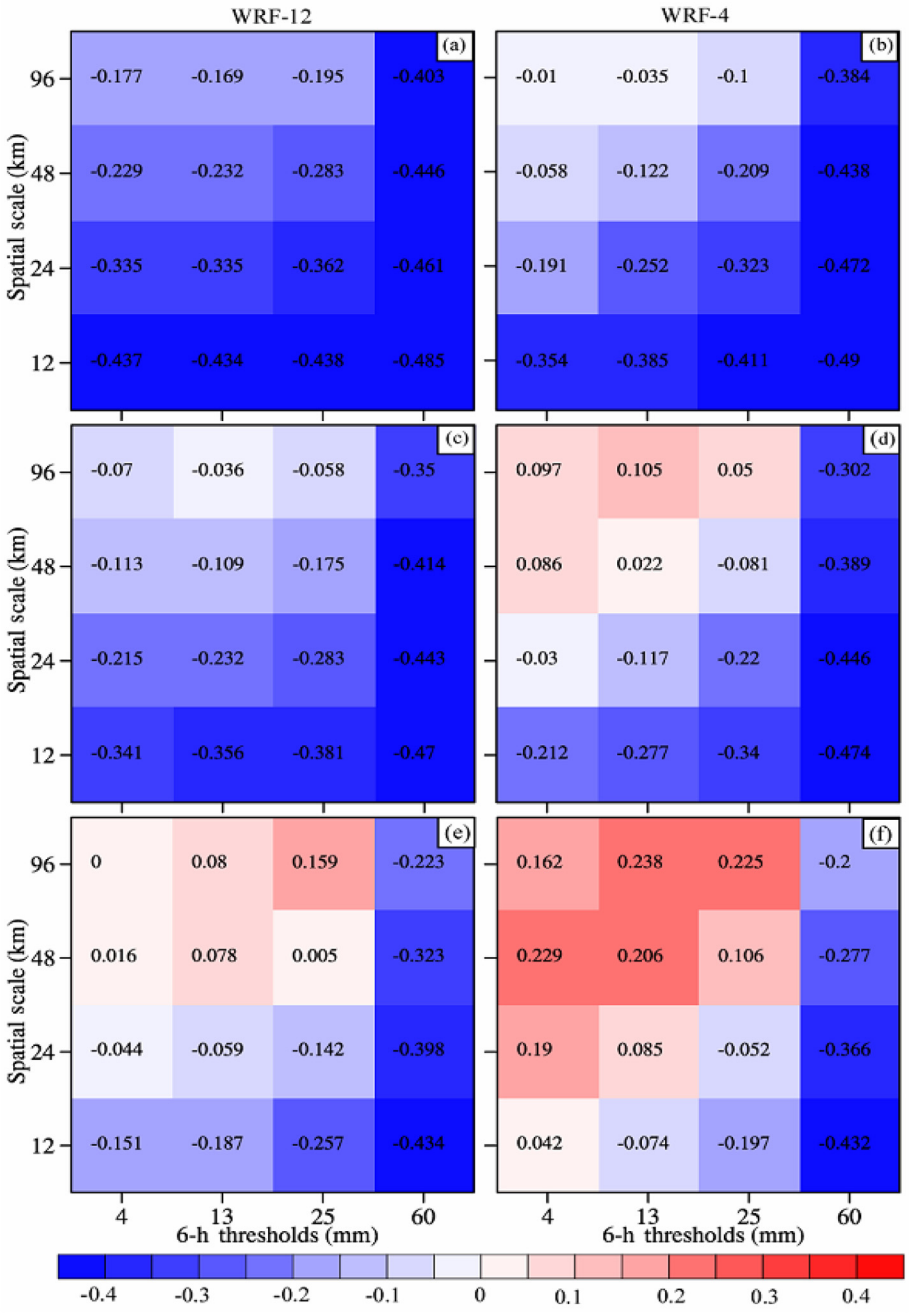

Fig. 10. As in Fig. 9 but for IS_FSS.

A Meiyu front case is first used to test the newly developed verification scores. For IS_FSS, a notable difference is that we apply the neighborhood idea in the skill score calculation (by using FSS instead of ISS). The overall pattern (the larger the spatial scale, the higher the score) is similar to that in the original ISS. With the tolerance of displacement error, WRF-4 shows positive scores at most spatial scales and thresholds. The fraction of useful forecast skill across scales and thresholds also improves for WRF-12, but not as much as that of WRF-4. For ISP_FSS, WRF-12 benefits more because the original forecast has larger intensity biases. Nonetheless, WRF-4 still performs better than WRF-12 for nearly all spatial scales and thresholds.

Finally, we apply IS_FSS and ISP_FSS to the entire Meiyu season of 2016, and similar conclusions are obtained. Compared to the original ISS, both IS_FSS or ISP_FSS show a greater fraction of useful forecast skill, suggesting that verification scores at small scales are sensitive to position and intensity errors. Based on either IS_FSS or ISP_FSS, WRF-4 performs better than WRF-12, agreeing with subjective assessment. Finally, we point out that the Haar wavelet decomposition is used to evaluate scale-dependent precipitation forecast skills in this paper for all metrics used. Other approaches for scale decompositions should also be exploited and compared in future studies.

Acknowledgments

This work was primarily supported by the National Key Research and Development Program of China (Grant No. 2018YFC1507604) and the National Natural Science Foundation of China (Grant No. 41975124 and 4173065).

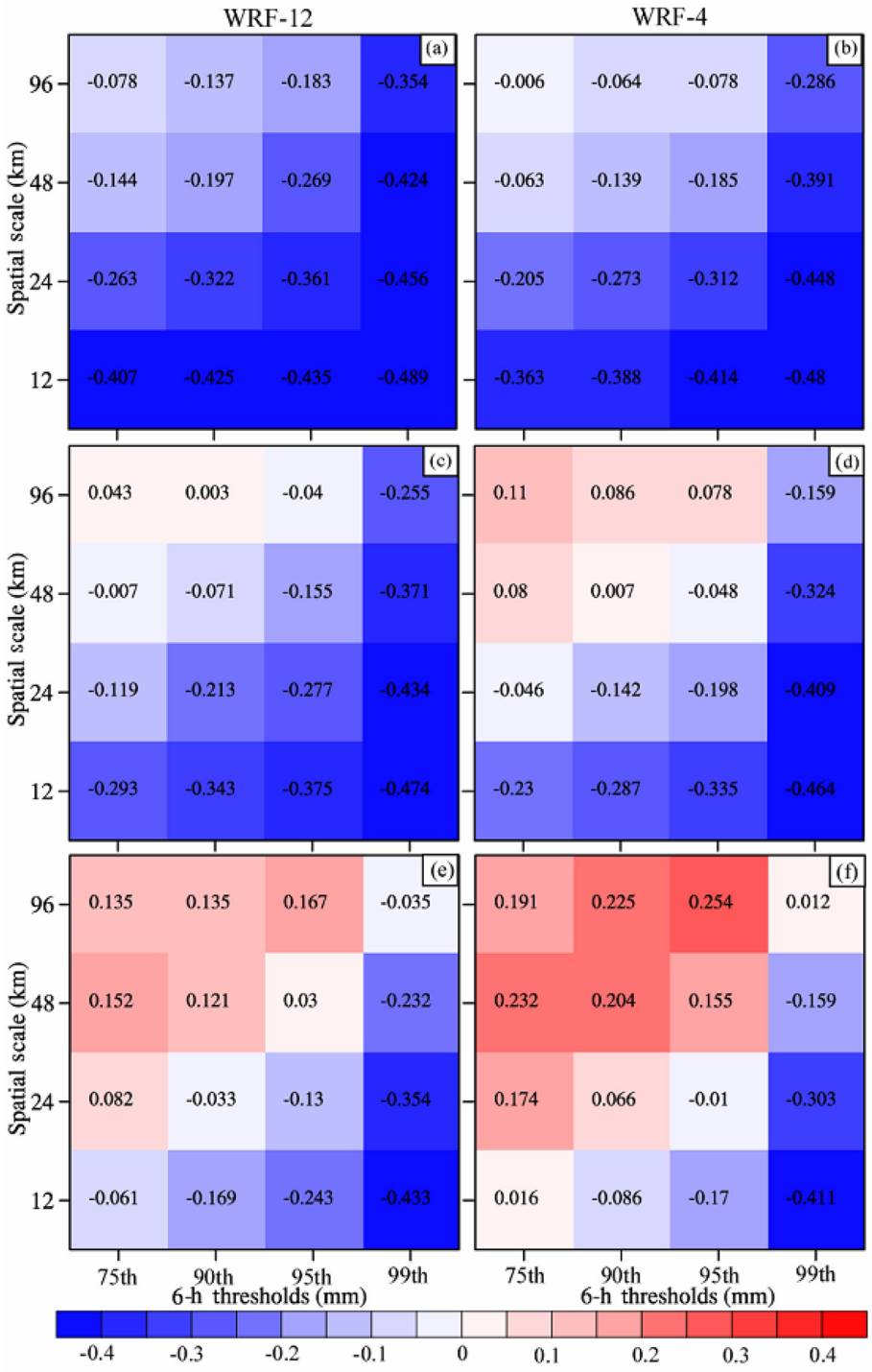

Fig. 11. As in Fig. 9 but for ISP_FSS. The numbers on the *x-axis* represent the percentile thresholds.

References

Bao, X., Zhang, F., Sun, J., 2011. Diurnal variations of warm-season precipitation east of the Tibetan Plateau over China. Mon. Weather Rev. 139, 2790–2810.

Brandes, E.A., Zhang, G., Vivekanandan, J., 2003. An evaluation of a drop distribution-based polarimetric radar rainfall estimator. J. Appl. Meteorol. 42, 652–660.

Briggs, W.M., Levine, R.A., 1997. Wavelets and field forecast verification. Mon. Weather Rev. 125, 1329–1341.

Brown, B.G., Gilleland, E., Ebert, E.E., 2012. Forecasts of Spatial Fields. *Forecast Verification: A Practitioner's Guide in Atmospheric Science (eds I.T. Jolliffe and D.B. Stephenson)*. Wiley, Chichester, West Sussex, UK, pp. 95–117.

Burgess, D.W., Magsig, M.A., Wurman, J., Dowell, D.C., Richardson, Y., 2002. Radar Observations of the 3 May 1999 Oklahoma City Tornado. Weather Forecast. 17, 456–471.

Buschow, S., Friederichs, P., 2020. Using wavelets to verify the scale structure of precipitation forecasts. Adv. Stat. Climatol. Meteorol. Oceanogr. 6, 13.

Casati, B., 2010. New Developments of the Intensity-Scale Technique within the Spatial Verification Methods Intercomparison Project. Weather Forecast. 25, 113–143.

Casati, B., Ross, G., Stephenson, D.B., 2004. A new intensity-scale approach for the verification of spatial precipitation forecasts. Meteorol. Appl. 11, 141–154.

Casati, B., Wilson, L., Stephenson, D., Nurmi, P., Ghelli, A., Pocernich, M., Damrath, U., Ebert, E., Brown, B., Mason, S., 2008. Forecast verification: current status and future directions. Meteorol. Appl. 15, 3–18.

Clark, A.J., Gallus Jr., W.A., Chen, T.-C., 2007. Comparison of the diurnal precipitation cycle in convection-resolving and non-convection-resolving mesoscale models. Mon. Weather Rev. 135, 3456–3473.

Clark, A.J., Gallus Jr., W.A., Xue, M., Kong, F., 2009. A comparison of precipitation forecast skill between small convection-permitting and large convection-

parameterizing ensembles. Weather Forecast. 24, 1121–1140.
Collins, W.D., Rasch, P.J., Boville, B.A., Hack, J.J., Mccaa, J.R., Williamson, D.L., Kiehl, J.T., Briegleb, B., Bitz, C., Lin, S.-J., Zhang, M., Dai, Y., 2004. Description of the NCAR Community Atmosphere Model (CAM 3.0). NCAR Technical Note NCAR/TN-464 + STR. https://doi.org/10.5065/D63N21CH. 226.
Davies, T., Cullen, M.J.P., Malcolm, A.J., Mawson, M.H., Staniforth, A., White, A.A., Wood, N., 2005. A new dynamical core for the Met Office's global and regional modelling of the atmosphere. Q. J. R. Meteorol. Soc. 131, 1759–1782.
Davis, C., Brown, B., Bullock, R., 2006a. Object-based verification of precipitation forecasts. Part II: application to convective rain systems. Mon. Weather Rev. 134, 1785–1795.
Davis, C., Brown, B., Bullock, R., 2006b. Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas. Mon. Weather Rev. 134, 1772–1784.
deElia, R., Laprise, R., Denis, B., 2002. Forecasting skill limits of nested, limited-area models: a perfect-model approach. Mon. Weather Rev. 130, 2006–2023.
Dorninger, M., Gilleland, E., Casati, B., Mittermaier, M.P., Ebert, E.E., Brown, B.G., Wilson, L.J., 2018. The setup of the mesovict project. Bull. Am. Meteorol. Soc. 99, 1887–1906.
Ebert, E., 2008. Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework. Meteorol. Appl. 15, 51–64.
Ebert, E., 2009. Neighborhood verification: a strategy for rewarding close forecasts. Weather Forecast. 24, 1498–1510.
Ebert, E., McBride, J., 2000. Verification of precipitation in weather systems: Determination of systematic errors. J. Hydrol. 239, 179–202.
Funatsu, B.M., Claud, C., Chaboureau, J.P., 2009. Comparison between the large-scale environments of moderate and intense precipitating systems in the Mediterranean Region. Mon. Weather Rev. 137, 3933–3959.
Gandin, L.S., Murphy, A.H., 1992. Equitable skill scores for categorical forecasts. Mon. Weather Rev. 120, 361–370.
Garg, G., Singh, V., Gupta, J.R.P., Mittal, A.P., 2011. Relative wavelet energy as a new feature extractor for sleep classification using EEG signals. Int. J. Biomed. Signal Process. 2, 75–80.
Gilleland, E., Ahijevych, D., Brown, B.G., Casati, B., Ebert, E.E., 2009. Intercomparison of spatial forecast verification methods. Weather Forecast. 24, 1416–1430.
Gilleland, E., Lindström, J., Lindgren, F., 2010a. Analyzing the image warp forecast verification method on precipitation fields from the ICP. Weather Forecast. 25, 1249–1262.
Gilleland, E., Ahijevych, D.A., Brown, B.G., Ebert, E.E., 2010b. Verifying forecasts spatially. Bull. Am. Meteorol. Soc. 91, 1365–1373.
Gilleland, E., Skok, G., Brown, B.G., Casati, B., Dorninger, M., Mittermaier, M.P., Roberts, N., Wilson, L.J., 2020. A novel set of geometric verification test fields with application to distance measures. Mon. Weather Rev. 148, 1653–1673.
Gowan, T.M., Steenburgh, W.J., Schwartz, C.S., 2018. Validation of mountain precipitation forecasts from the convection-permitting NCAR Ensemble and operational forecast systems over the Western United States. Weather Forecast. 33, 739–765.
Grell, G.A., 1993. Prognostic evaluation of assumptions used by cumulus parameterizations. Mon. Weather Rev. 121, 764–787.
Hamill, T.M., 1999. Hypothesis tests for evaluating numerical precipitation forecasts. Weather Forecast. 14, 155–167.
Hintsa, E.J., Allsup, G.P., Eck, C.F., Hosom, D.S., Purcell, M.J., Roberts, A.A., Scott, D.R., Sholkovitz, E.R., Rawlins, W.T., Mulhall, P.A., Lightner, K., McMillan, W.W., Song, J., Newchurch, M.J., 2004. New ozone measurement systems for autonomous operation on Ocean Buoys and Towers{ast}. J. Atmos. Ocean. Technol. 21, 1007–1016.
Joyce, R.J., Janowiak, J.E., Arkin, P.A., Xie, P.P., 2004. CMORPH: a method that produces global precipitation estimates from passive microwave and infrared data at high spatial and temporal resolution. J. Hydrometeorol. 5, 487–503.
Jung, T., Leutbecher, M., 2008. Scale-dependent verification of ensemble forecasts. Q. J. R. Meteorol. Soc. 134, 973–984.
Karki, R., Hasson, S.U., Gerlitz, L., Scholten, T., Schickhoff, U., Böhner, J., 2017. Quantifying the added value of convection-permitting climate simulations in complex terrain: a systematic evaluation of WRF over the Himalayas. Earth Syst. Dyn. 8, 507–528.
Lindborg, E., 1999. Can the atmospheric kinetic energy spectrum be explained by two-dimensional turbulence? J. Fluid Mech. 388, 259–288.
Lorenz, E.N., 1969. The predictability of a flow which possesses many scales of motion. Tellus 21, 289–307.
Love, B.S., Matthews, A.J., Lister, G.M., 2011. The diurnal cycle of precipitation over the Maritime Continent in a high-resolution atmospheric model. Q. J. R. Meteorol. Soc. 137, 934–947.
Mittermaier, M.P., 2006. Using an intensity-scale technique to assess the added benefit of high-resolution model precipitation forecasts. Atmos. Sci. Lett. 7, 36–42.
Morrison, H., Curry, J., Shupe, M., Zuidema, P., 2005. A new double-moment microphysics parameterization for application in cloud and climate models. Part II: Single-column modeling of Arctic clouds. J. Atmos. Sci. 62, 1678–1693.
Murata, A., Sasaki, H., Kawase, H., Nosaka, M., 2016. Evaluation of precipitation over an oceanic region of Japan in convection-permitting regional climate model simulations. Clim. Dyn. 48, 1–14.
Pan, Y., Shen, Y., Yu, J., Xiong, A., 2015. An experiment of high-resolution gauge-radar-satellite combined precipitation retrieval based on the Bayesian merging method. Acta Meteorol. Sin. (in chinese) 73, 177–186.
Pleim, J.E., 2006. A simple, efficient solution of flux–profile relationships in the atmospheric surface layer. J. Appl. Meteorol. Climatol. 45, 341–347.
Pleim, J.E., 2007. A combined local and nonlocal closure model for the atmospheric boundary layer. Part I: model description and testing. J. Appl. Meteorol. Climatol. 46, 1383–1395.
Pleim, J.E., Xiu, A., 1995. Development and testing of a surface flux and planetary boundary layer model for application in mesoscale models. J. Appl. Meteorol. 34, 16–32.
Qiao, F., Liang, X.-Z., 2015. Effects of cumulus parameterizations on predictions of summer flood in the Central United States. Clim. Dyn. 45, 727–744.
Roberts, N., 2008. Assessing the spatial and temporal variation in the skill of precipitation forecasts from an NWP model. Meteorol. Appl. 15, 163–169.
Roberts, N.M., Lean, H.W., 2008. Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. Mon. Weather Rev. 136, 78–97.
Shen, Y., Xiong, A.Y., Wang, Y., Xie, P.P., 2010. Performance of high-resolution satellite precipitation products over China. J. Geophys. Res.-Atmos. 115.
Shen, Y., Zhao, P., Pan, Y., Yu, J.J., 2014. A high spatiotemporal gauge-satellite merged precipitation analysis over China. J. Geophys. Res.-Atmos. 119, 3063–3075.
Skamarock, W.C., 2004. Evaluating mesoscale NWP models using kinetic energy spectra. Mon. Weather Rev. 132, 3019–3032.
Skamarock, W.C., Klemp, J.B., Dudhia, J., Gill, D.O., Barker, D.M., Wang, W., Powers, J.D., 2005. A Description of the Advanced Research WRF Version 2. (88 pp).
Snook, N., Kong, F.Y., Brewster, K.A., Xue, M., Thomas, K.W., Supinie, T.A., Perfater, S., Albright, B., 2019a. Evaluation of convection-permitting precipitation forecast products using WRF, NMMB, and FV3 for the 2016-17 NOAA hydrometeorology testbed flash flood and intense rainfall experiments. Weather Forecast. 34, 781–804.
Snook, N., Kong, F., Brewster, K.A., Xue, M., Thomas, K.W., Supinie, T.A., Perfater, S., Albright, B., 2019b. Evaluation of convection-permitting precipitation forecast products using WRF, NMMB, and FV3 for the 2016–17 NOAA hydrometeorology testbed flash flood and intense rainfall experiments. Weather Forecast. 34, 781–804.
Stratman, D.R., Coniglio, M.C., Koch, S.E., Xue, M., 2013. Use of multiple verification methods to evaluate forecasts of convection from hot- and cold-start convection-allowing models. Weather Forecast. 28, 119–138.
Sun, Z., Xue, M., Zhu, K., Zhou, B., 2019. Prediction of an EF4 supercell tornado in Funing, China: Resolution dependency of simulated tornadoes and their structures. Atmos. Res. 229, 175–189.
Surcel, M., Zawadzki, I., Yau, M.K., 2014. On the filtering properties of ensemble averaging for storm-scale precipitation forecasts. Mon. Weather Rev. 142, 1093–1105.
Weisman, M.L., Davis, C., Wang, W., Manning, K.W., Klemp, J.B., 2008. Experiences with 0–36-h explicit convective forecasts with the WRF-ARW model. Weather Forecast. 23, 407–437.
Weniger, M., Kapp, F., Friederichs, P., 2017. Spatial verification using wavelet transforms: a review. Q. J. R. Meteorol. Soc. 143, 120–136.
Wernli, H., Paulat, M., Hagen, M., Frei, C., 2008. SAL—A novel quality measure for the verification of quantitative precipitation forecasts. Mon. Weather Rev. 136, 4470–4487.
Wong, M., Skamarock, W.C., 2016. Spectral characteristics of convective-scale precipitation observations and forecasts. Mon. Weather Rev. 144, 4183–4196.
Xue, M., Kong, F., Weber, D., Thomas, K.W., Wang, Y., Brewster, K., Droegemeier, K.K., Weiss, J.S.K.S.J., Bright, D.R., Wandishin, M.S., Coniglio, M.C., Du, J., 2007. CAPS realtime storm-scale ensemble and high-resolution forecasts as part of the NOAA hazardous Weather Testbed 2007 spring experiment. In: *22nd Conference on Weather Analysis and Forecasting/18th Conference on Numerical Weather Prediction* American Meteor Society, CDROM 3B.1.
Xue, M., Luo, X., Zhu, K.F., Sun, Z.Q., Fei, J.F., 2018. The controlling role of boundary layer inertial oscillations in Meiyu Frontal precipitation and its diurnal cycles over China. J. Geophys. Res.-Atmos. 123, 5090–5115.
Zepeda-Arce, J., Foufoula-Georgiou, E., Droegemeier, K.K., 2000. Space-time rainfall organization and its role in validating quantitative precipitation forecasts. J. Geophys. Res.-Atmos. 105, 10129–10146.
Zheng, Y., Xue, M., Li, B., Chen, J., Tao, Z., 2016. Spatial characteristics of extreme rainfall over China with hourly through 24-hour accumulation periods based on national-level hourly rain gauge data. Adv. Atmos. Sci. 33, 1218–1232.
Zhu, K.F., Yang, Y., Xue, M., 2015. Percentile-based neighborhood precipitation verification and its application to a landfalling tropical storm case with radar data assimilation. Adv. Atmos. Sci. 32, 1449–1459.
Zhu, K., Xue, M., Zhou, B., Zhao, K., Sun, Z., Fu, P., Zheng, Y., Zhang, X., Meng, Q., 2018. Evaluation of real-time convection-permitting precipitation forecasts in China during the 2013–2014 summer season. J. Geophys. Res.-Atmos. 123, 1037–1064.