atmosphere

MDPI

Article

Comparative Evaluation of Rainfall Forecasts during the Summer of 2020 over Central East China

Yakai Guo [1,2], Changliang Shao [3,*] and Aifang Su [1,2,*]

[1] China Meteorological Administration Henan Meteorological Bureau, Zhengzhou 450003, China; guoykhmb@126.com
[2] China Meteorological Administration Key Laboratory of Agro-Meteorological Support and Application Technology of Henan Province, Zhengzhou 450003, China
[3] China Meteorological Administration Meteorological Observation Centre, Beijing 100080, China
* Correspondence: shchl1@163.com (C.S.); afsu011@sohu.com (A.S.); Tel.: +86-182-1098-6639 (C.S.); +86-186-3839-8288 (A.S.)

Abstract: By using various skill scores and spatial characteristics of spatial verification methods and traditional techniques of the model evaluation tool, the gridded precipitation observation, known as Climate Prediction Center Morphing Technique, gauge observation and three datasets that were derived from local, Shanghai, and Grapes models, respectively, were conducted to assess the 3 lead day rainfall forecast with 0.5 day intervals during the summer of 2020 over Central East China. Results have shown that the local model generally outperforms the other two for the most skill scores but usually with relatively larger uncertainties than the Shanghai model, and it has the least displacement errors for moderate rainfall among the three datasets. However, the rainfall of the Grapes model has been heavily underestimated and is accompanied with a large displacement error. Both the local and Shanghai model can effectively forecast the large-scale convection and rainstorms but over forecast the local convection, while the local model likely over forecasts the local rainstorms. In addition, the Shanghai model slightly favors over forecasting on a broad scale range and a broad threshold range, and the local model slightly misses the rainfall exceeding 100 mm. Generally, for a broadly comparative evaluation on rainfall, the popular dichotomous methods should be recommended when considering reasonable classification of thresholds if the accuracy is highly demanding. In addition, most spatial methods are suggested to conduct with proper pre-handling of non-rainfall event cases. Especially, the verification metrics including spatial characteristic difference information should be recommended to emphasize rewarding the severe events forecast under a global warming background.

Keywords: rainfall verification; rainstorm; skill scores; spatial characteristics; model evaluation

Citation: Guo, Y.; Shao, C.; Su, A. Comparative Evaluation of Rainfall Forecasts during the Summer of 2020 over Central East China. *Atmosphere* **2023**, *14*, 992. https://doi.org/10.3390/atmos14060992

Academic Editor: Tomeu Rigo

Received: 15 May 2023
Revised: 29 May 2023
Accepted: 5 June 2023
Published: 7 June 2023

1. Background

Burning of fossil fuels causes extreme weather events and has damaged the production systems [1–5]. Therefore, it is required to accurately predict the extreme weather events. In particular, rainfall is a highly distinguished aspect of numerical prediction, and its evaluation is not only an essential part of the numerical weather prediction system development, but also serves as an objective basis for practical decision-making [6–10]. Due to the quite complex model physics and forecasting methods [11–13], rainfall prediction capabilities are of great uncertainty [14–17]. As the complex forecasts with finer spatial scales have been developed in recent years [18–20], it has become an essential means for both research and application to obtain their specific attributes and performances through evaluations and comparisons [21,22].

The contingency table based on event occurrence ("yes or no"; dichotomous) count, and a general framework based on joint distributions (called the distributions-oriented approach) have been established successively [23], and further intends to satisfy broader

forecast verification. While for the multi-threshold precipitation verification, no classification of thresholds is perfect and there are many overlaps that cannot be excluded between different thresholds [24]. The equitable scores for categorical forecasts [25] have been furtherly proposed that embody almost all the desirable attributes various works have highlighted [26–28]. Following, several quite popular methods based on paired forecast–observation point comparisons for rainfall verification have been proposed under a proper hypothesis on threshold, occurrence rate, confidence, and others [29], and the complexity and dimensionality of dichotomous verification have been discussed [30,31].

Spatial verification has drawn attention accordingly in recent years. The well-known spatial verification methods intercomparison project (ICP) stage I [32] and stage II [33], which are mostly based on a meta-analysis of ideal precipitation events with several spatial classifications, have been carried out successively to develop verification methods that are directly against various spatial characteristic differences. In addition, the Model Evaluation Tool (MET) has been developed with the addition of the multiple ongoing mentioned verification methods over a decade [34]. MET has integrated broad spatial verification techniques, such as the neighborhood [35], gradient [36], distance-map [37–40], wavelet [41], and model object diagnosis and evaluation (MODE) methods [42,43], and intends to diagnose various spatial measurements for broader datasets.

The high dimensionality problem (defined as the number of probabilities that must be specified to reconstruct the basic distribution of forecasts and observations) [21] is one of the key factors during comparative evaluations in dichotomous methods. This can be reduced by using the threshold or categorical value to divide the rainfall values into binary bins (its value is 1 or 0) [44], which can simplify the complexes in skill comparison among datasets to some extent. Moreover, the sampling uncertainty is another key factor during comparative evaluation of dichotomous methods, because the sample number is always limited in real-world applications. Confidence should be estimated to ensure that apparent differences in skill are real, and not just due to random fluctuations. A measurement without some indication of precision has little meaning [45]. Usually, under the assumptions of stationarity and independence, the confidence interval that indicates lower bounds on the uncertainty in skill is taken as a basic measure on reliability of skill [46]. Furthermore, a nonparametric method, such as resampling (also known as bootstrap), is proposed to be appropriate for estimating the confidence interval of skill scores [47,48].

The dimensionality of rainfall verification is too great when compared to the size of the data set available, e.g., the observed rainfall is usually local and intermittent. Especially, spatial verification demands spatially regular to ensure equitable evaluation on the basis that connections between points are straight. Meanwhile, the characteristics of most spatial methods and the events of dichotomous methods can be of great variation between cases or at different lead times, which can make the resampling strategy too complex to conduct. Therefore, comparative verification studies on the spatial characteristics or events of rainfall forecast peremptorily demand the joint analysis on various measurements under identical spatial conditions or equitable occurrence bases to hopefully address these ongoing mentioned unaccountable issues [35–43].

The limitations of regional rainfall products in an application that should be attributed to either blindness skill scores or biased data can be quite an open problem, and this can be investigated by using the comparative evaluation between skills or datasets [32–34]. Especially, most spatial verification metrics lack a comparison to identify their abilities in verifying regional rainfall under the background of increasingly severe weather events. To fill this gap, this study has evaluated the local precipitation products using various verification methods of MET (Version 10.0.0), and further analyzed the advantages and weaknesses of the methods and products by comparing the uncertainties of skill scores and related characteristics of spatial measurements, which aims at providing better ideas for the inspection and evaluation of the local rainfall products.

2. Datasets and Methods

2.1. Datasets

The Central East China area covers from 30° N to 38° N, and from 109° E to 118° E, and it is located on the south side of the Southern Taihang Mountains (Figure 1a), where frequent rainstorm occurs in local summer. This study uses the ISO-meridional coordinates with an interval of 0.1° to re-grid both the raw forecast and observation product over our study area, which intends to generate paired forecast–observation fields with identical grids for hopefully equitable verification.

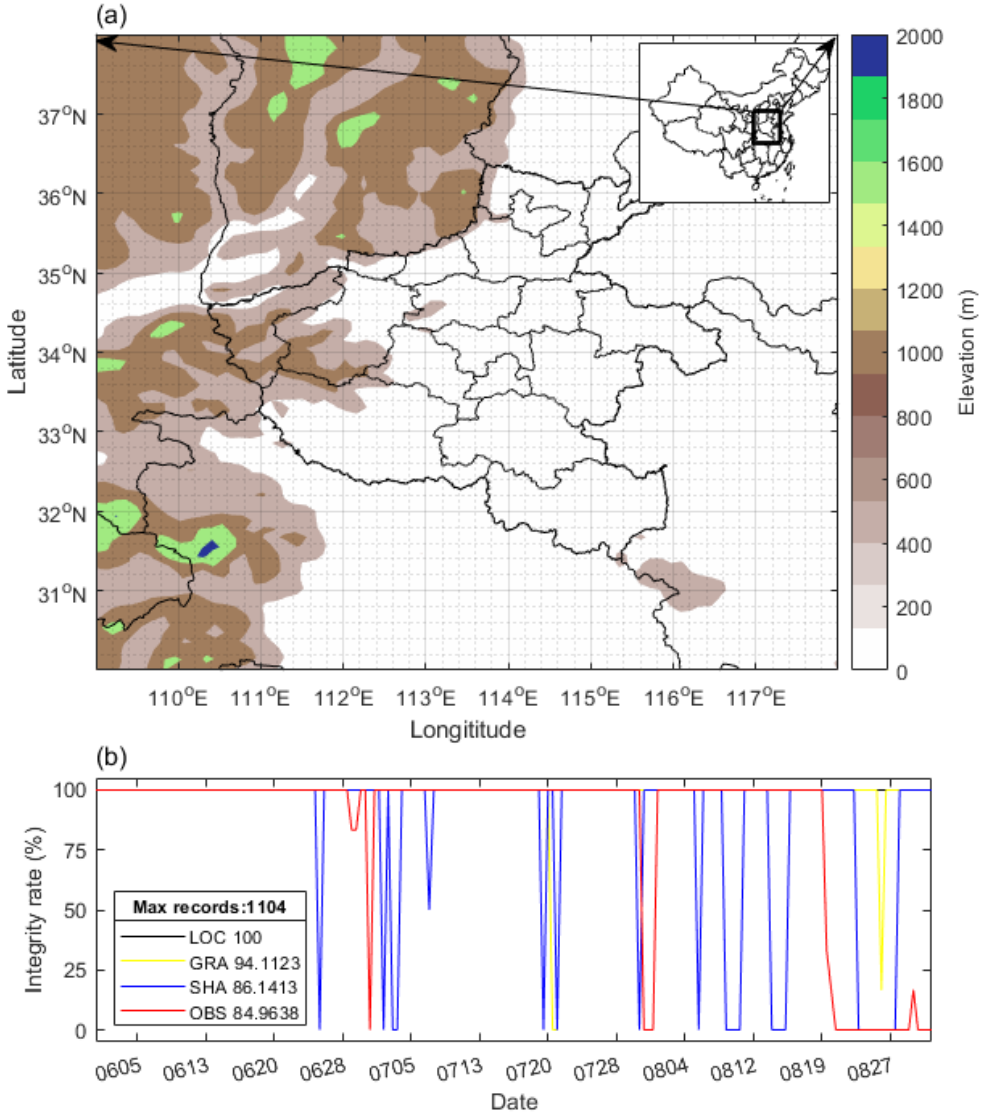

Figure 1. The study area and datasets. (**a**) Topography (shaded) and location of study area (black box; inner figure), and (**b**) integrity of different datasets.

Moreover, the 3 lead day accumulation precipitation (APCP) forecast products with a 9 km resolution and a 3 h interval over our study area are assembled into three datasets with a 0.5 day interval (Table 1). Note that the APCP products are derived from the local model of Henan province (LOC) [25], the CMA Grapes model (GRA) [23], and the model of CMA Shanghai meteorological bureau (SHA) [24]. The three datasets are further re-gridded into three forecast fields with identical grids. Meanwhile, the merged precipitation observations, known as Climate Prediction Center Morphing Technique and gauge observations (CMPA, Version 2.0) [49], have a high spatial and temporal resolution (5 km and 1 h interval), and

are collected and further re-gridded into observation fields (OBS hereafter) with the grids identical to the forecast field.

Table 1. Description of datasets.

Datasets	Fields (Resolution; Period Range)	Description	Raw Resolution
LOC	(0.1°/0.5 day; 3 days)	The APCP products derived from the local model of Henan province	9 km/3 h
GRA	(0.1°/0.5 day; 3 days)	The APCP products derived from CMA Grapes model forecasts	9 km/3 h
SHA	(0.1°/0.5 day; 3 days)	The APCP products derived from the local model of CMA Shanghai meteorological bureau	9 km/3 h
OBS	(0.1°/0.5 day; 3 days)	The gridded APCP observational product known as CMPA (V2.0)	5 km/1 h

APCP = accumulative precipitation, h = hour.

These APCP products are referenced by forecasters for the local weather forecast service, and the products at local 08:00 and 20:00 time (00:00 and 12:00 UTC) are most frequently used for rainfall event decisions. Therefore, in this study, the daily 00 and 12 (UTC) forecast and observation fields during summertime in 2020 are collected for the expected synoptic insights. As seen from Figure 1b, the records of the overall forecast and observation fields are 1104, and the integrity rate of LOC, GRA, SHA, and OBS fields are 100%, 94%, 86%, and 85%, respectively. The integrity rate can be defined as:

$$\frac{1}{n_t}\sum_{i}^{n_t}\frac{l_i}{L} \tag{1}$$

In Equation (1), L is total number of all 3 lead day forecast fields, which represents the expected length of forecast fields or the forecast length, l is number of the available forecast fields in the forecast length, i represents for the i verification at the same initial time (00 or 12), and n_t is the total number of verifications during the whole period of this comparative evaluation. Therefore, integrity here represents for one basis that the number of datasets used for calculations during comparative evaluation, which is usually 100% to make sure every forecast can be verified.

2.2. Methods

This study here has taken two popular verification schemes, labeled dichotomous and neighborhood, and three spatial verification schemes, labeled displaced, decomposed, and featured, to emphasize on assessment of APCP forecast quality and possible insights into verification method differences (Table 2).

The dichotomous scheme includes three widely used skill scores, CSI, GSS, and FBIAS, which labeled the 2 × 2 contingency table, and four skill scores, ACC, HK, HSS, and GER, which labeled the N × 2 contingency table, which each usually measured at one point in the paired forecast–observation field [25–28]. The 2 × 2 contingency table methods are measured with one-threshold-based categories, while the N × 2 contingency table methods are measured with multi-threshold-based categories [29]. The neighborhood scheme includes popular skills, such as FSS and FBS, and two special cases of FSS, AFSS and UFSS, respectively [35]. The neighborhood window of FSS and AFSS is set to 1 in the entire domain, respectively.

Table 2. Description of skill scores.

Short Name	Full Name	Reference Formula *	Perfect Limit No Skill Limit	Description	Type
ACC	Accuracy rate	$ACC = \sum_{i=1}^{K} p_{ii}$	=1 =0	The N × 2 contingency table	Dichotomous
HK	Hanssen–Kuipers discriminant	$HK = \frac{\left(\sum_{i=1}^{K} p_{ii} - \sum_{i=1}^{K} p_i \hat{p}_i\right)}{\left(1 - \sum_{i=1}^{K} p_i \hat{p}_i\right)}$	=1 =0		
HSS	Heidegger skill score	$HSS = \frac{\left(\sum_{i=1}^{K} p_{ii} - \sum_{i=1}^{K} p_i \hat{p}_i\right)}{\left(1 - \sum_{i=1}^{K} p_i p_i\right)}$	=1 ~−∞		
GER	Gerrity score	$GER = \sum_{i=1}^{K} \sum_{j=1}^{K} p_{ij} s_{ij}$	=1 =0		
CSI	Critical success index	$CSI = \frac{a}{a+b+c}$	=1 =0	The 2 × 2 contingency table	
GSS	Gilbert skill score	$GSS = \frac{a - C_1}{a+b+c-C_1}, C_1 = \frac{(a+b)\cdot(a+c)}{t}$	=1 =0		
FBIAS	Frequency bias score	$FBIAS = \frac{a+b}{a+c}$	=1 ~		
FBS	Fractions brier score	$FBS = \frac{1}{n}\sum_{n}\left[\langle P_f \rangle_s - \langle P_o \rangle_s\right]^2$	=0 =1	The neighborhood method	Neighborhood
FSS	Fractions skill score	$FSS = 1 - \frac{FBS}{\frac{1}{n}\left[\sum_n \langle P_f \rangle_s^2 + \sum_n \langle P_o \rangle_s^2\right]}$	=1 =0		
AFSS	Asymptotic fractions skill score	$AFSS = FSS(n = 1)$	=1 =0		
UFSS	Uniform fractions skill score	$UFSS = \frac{1+f_o}{2}$	~ ~		
S1	S1 score	$S1 = 100\frac{\sum_{i=1}^{n}(w_i(e_g))}{\sum_{i=1}^{n}(w_i(G_L))_i}$, $e_g = \left\|\frac{\delta}{\delta x}(f-o)\right\| + \left\|\frac{\delta}{\delta y}(f-o)\right\|, G_L = \max\left(\left\|\frac{\delta f}{\delta x}\right\|, \left\|\frac{\delta o}{\delta x}\right\|\right) + max\left(\left\|\frac{\delta f}{\delta y}\right\|, \left\|\frac{\delta o}{\delta y}\right\|\right)$, $w_i = 1$	=0 ~+∞	The gradient method	Displaced
BM	Baddeley's Δ Metric	$BM = \Delta_{p,\omega}(A,B) = \left[\frac{1}{n}\sum_{s\in D}\left\|\omega(d(s,A)) - \omega(d(s,B))\right\|\right]^{\frac{1}{p}}$	=0 ~+∞	The distance map method	
HD	Hausdorff Distance	$HD = BM(p = \infty)$	=0 ~+∞		
MZM	Mean of Zhu's Measure	$MZM = \lambda\sqrt{\frac{1}{n}\sum_{s\in D}(I_F(s) - I_O(s))^2} + (1-\lambda)\cdot MED(A,B)$	=0 ~+∞		
ISC	Intensity scale skill score	$ISC = SS(t,j) = 1 - MSE(t,j)\cdot\frac{n+1}{MSE(t)_{random}}$	≥0 <0	The wavelet analysis method	Decomposed
TIN	Total of total interest	$TIN = median(T(\alpha)_k \geq 0.7, k \in (1,\ldots,m))$	=1 NULL	MODE	Featured

* Dichotomous. For the N × 2 contingency table, i and j represent for the forecast and observation category respectively, K is the total category number, p and $\hat{p}$ are the relative frequency and the estimated probability function, respectively, and s represents for the corrected score matric. For the 2 × 2 contingency table, the count a,b,c, and d represent for Hit, False alarm, Miss, and Anti hit, respectively; $t = a + b + c + d$. Neighborhood. n is the number of neighborhoods; $\langle P_f \rangle_s$ and $\langle P_o \rangle_s$ represent for the proportion of grid boxes that have forecast and observed events, respectively; f_o is the observation rate. Displaced. For S1, f and o represent for forecast and observations respectively; δx and δy are set to 1; n is the domain size; w is a weight. For HD, BM, and MZM, n is the total number of events, d is the distance map for the respective event ($\frac{A}{B}$) area, and D is its vector; ω is the concave function; and s is the event set. p is a corresponding parameter; $p = 2$ is for BM, and $p = \infty$ is for HD; $MED(A,B)$ is as in the mean error distance. $I_F(s)(I_O(s))$ is the binary field derived from the forecast (observation); λ is a weight. Decomposed. t and j represent for threshold and scale component, respectively. n is the wavelet component index; $MSE(t)_{random}$ is the MSE for the random binary forecast and observation fields. Featured. T represents the total interest; k represent the object index, and m is the total number of objects; α is the entire attribute vector.

The displaced scheme includes S1 scores labeled with the gradient method [36], and three skills HD, MZM [37,38], and BM [39] scores labeled with the distance map method [40]; their perfect score is 0. The decomposed scheme includes the ISC skill labeled with the wavelet analysis (Haar wavelet) [41]; its score varies between −1 and 1. The featured scheme includes TIN skill labeled as a method for object-based diagnostic evaluation (MODE) [42,43]; its score varies between 0.7 and 1. The convolution radius in MODE is set to 6 km in this study, which is slightly larger than the raw observation resolution (Table 1).

For the skill score over the one paired forecast–observation field of the dichotomous methods, MET takes the predefined significance level-based parameter (p; the value of which is usually set to 5%) of the resampling strategy (or bootstrap) to estimate the sampling uncertainty. Since the overall paired forecast–observation fields for different lead times or valid times are of a large quantity, the sampling uncertainties in skill of those overall fields that owns the significance level (usually smaller than 0.05) or the confidence levels (usually larger than 0.95) have been noted along with the skill, and this intends to indicate the overall evaluation on the reliability of one dichotomous skill.

Moreover, the uncertainty difference among verification methods for one dataset is compared to account for the skill sharpness difference, while the skill performance uncertainty among datasets is compared to evaluate the reliability of data quality indicated by the skills.

In addition, since most verification schemes are conducted on the threshold-based categories, the thresholds as 0.1, 1, 5, 10, 25, 50, and 100 mm have been used to define the category of rainfall events. Usually, rainfall between 0.1 and 1 mm, between 1 and 5 mm, between 10 and 25 mm, between 25 and 50 mm, and that between 50 and 100 mm have been related to the drizzle (rainfall or not), light rain, moderate rain, heavy rain, and rainstorm, respectively.

3. Experiments

The experiment of this study is illustrated in Figure 2. The observation and different kinds of APCP forecast products are assembled into forecast–observation pairs with an identical lead day range (3) and interval (0.5), and were further interpolated into the identical grids by using the bi-linear interpolation method. The following calculations distinguished by dichotomous, neighborhood, displaced, decomposed, and featured schemes that are based on MET are conducted to obtain the verification information about skill scores and spatial characteristics of these paired forecast–observation fields. Finally, the overall comparative evaluation among datasets and verification methods are finally conducted to achieve the insights of datasets and methods.

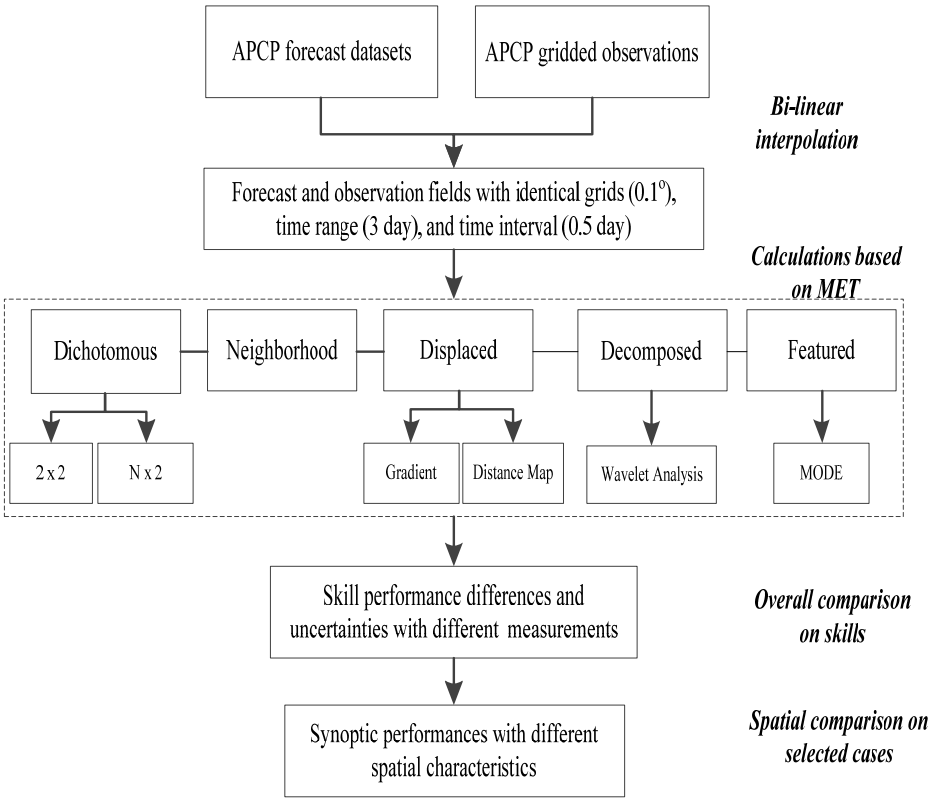

Figure 2. The flowchart of this study.

Moreover, the main eleven rainfall cases of summer 2020 over Middle East China are collected (Table 3) to fulfill the additional synoptic insights of datasets. Except in case 6, which is mainly strong convection, other processes are mixed precipitation of continuity systems and organized convection. Furthermore, except in cases 4 and 7, the rainfall events all occur at local night. It should be noted that each case (usually observation based) is verified with forecasts that has the same valid time but different initial times. This is another

basis that every case can be sufficiently verified during comparative verification, which is usually desired by forecasters.

Table 3. The main rainfall events during summer of 2020 over Central East China.

Index	Periods (mmdd hh)	Falling Area; Convection Location	Description
1	0609 12–0610 00	large; central southern	process and strong convection
2	0611 12–0612 00	large; central eastern	process and strong convection
3	0616 12–0617 00	large; local central eastern	process and strong convection
4	0622 00–0622 12	local; local southern	process edge
5	0627 12–0628 00	large; southern	process and strong convection
6	0704 12–0705 00	local; local northern	strong convection
7	0711 00–0711 12	large; central southern	process and strong convection
8	0718 12–0719 00	large; local southern	process edge
9	0721 12–0722 00	large; central southern	process and strong convection
10	0803 12–0804 00	local; local central eastern	process edge and strong convection
11	0806 12–0807 00	large; local central northern	process and strong convection

The cases are derived from local Precipitation Log Table of Henan meteorological observatory.

4. Results

4.1. Skill Scores

To assess the skill of different forecasts and the skill difference among different methods, the representative skills of the ongoing described methods and their uncertainties are further compared, and this is conducted on the basis that every forecast can be verified.

4.1.1. Dichotomous

As seen in Figure 3a, the categorical samples of LOC and SHA datasets have shown comparatively similar frequencies for almost all threshold-based categories, while for GRA, much more samples than the others can be observed for the threshold less than 0.1 mm. GRA has gained the highest ACC scores on averaged statistics at a confidence level of 95%, but with much more uncertainties, followed by LOC, and SHA is the worst (Figure 3b). Obviously, LOC has shown more skill in the mean statistics than SHA for HK, HSS, and GER, but with more uncertainties, while GRA has no skill (Figure 3c–e).

As GRA has much more weak rainfall samples than the other two and less strong rainfall samples, and ACC has clearly been affected by the large amounts of overlapped anti hits between different thresholds, which could be misleading. This indicates the careful usage of those non-equitable skill scores with multi-categorical values because they are quite sensitive to the forecast frequency [24]. Meanwhile, the no confidence levels of the other three measurements (Figure 3c–e) indicate unreliable skill; if the confidence is not considered, more outliers accompanied with less uncertainties in HK and HSS than GER indicate that GER skill is relatively sharper than the other two during this work. Meanwhile, the measurements, such as HK, HSS, and GER, could favor samples that are evenly distributed at given threshold so that they give a strict penalty on the overlapped anti hits to GRA when compared to ACC.

Moreover, as seen in Figure 4, LOC showed more skill on averaged values (CSI, GSS, and FBIAS) than SHA for different thresholds (0.1, 10, and 50 mm) at a confidence level of 95%, but with more uncertainties, while GRA had no skill. The increased outliers in FBIAS than in CSI and GSS indicate that the immeasurable information in skill uncertainties of the latter could not be ignored (Figure 4c,f,i). In addition, much more uncertainties in all skills for the 50 mm threshold than those for the small thresholds indicate the significant sensitivities of skills to threshold difference (Figure 4g–i). Furthermore, the 2×2 contingency table skills could be generally in favor of a field that has small threshold with large sample numbers.

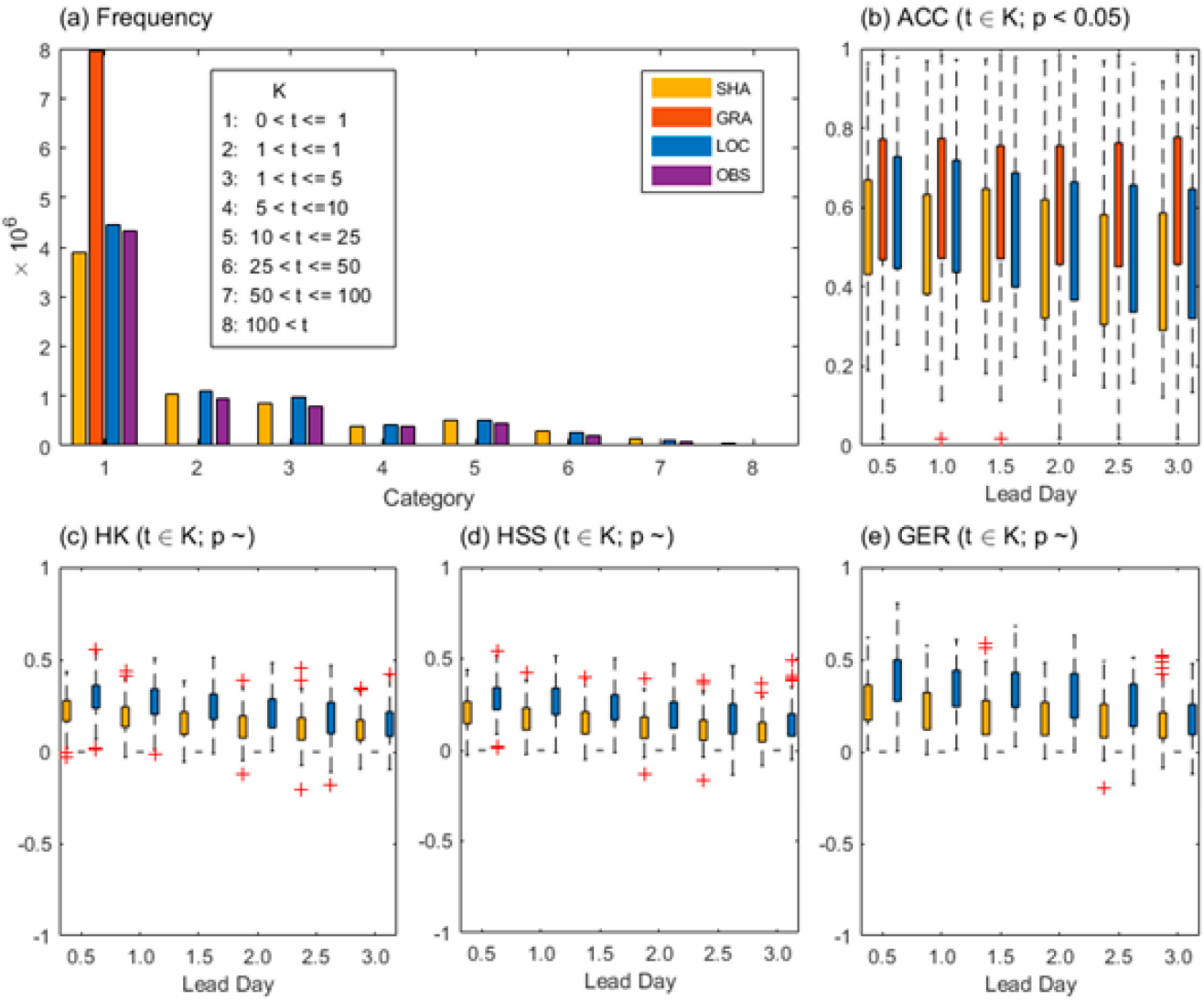

Figure 3. The categorical frequency and multi-category contingency table skills. (**a**) Frequency of different categories for all kinds of data fields; (**b**–**e**) represent ACC, HK, HSS, and GER skills, respectively, as a function of the lead day for different forecast fields. The t represents for the corresponding threshold (units: mm), K represents for the multi-categories, p represents for the significance level that can be obtained from all the available fields, and the red cross symbol (+) of box plot represents for outlier.

The strict penalty on GRA, and the complete confidence on different datasets for those single categorical skills (CSI, GSS, and FBIAS), indicate they are threshold sensitive but reliable. Meanwhile, for all thresholds, the large number of outliers that are far from 1 in FBIAS indicates the immeasurable information that the frequently forecast events are heavily biased. In addition, the significant counts of outliers in both CSI and GSS for the threshold 50 mm indicate that their scores are possibly less informative for the field that has a large threshold with limited samples during the comparative evaluation.

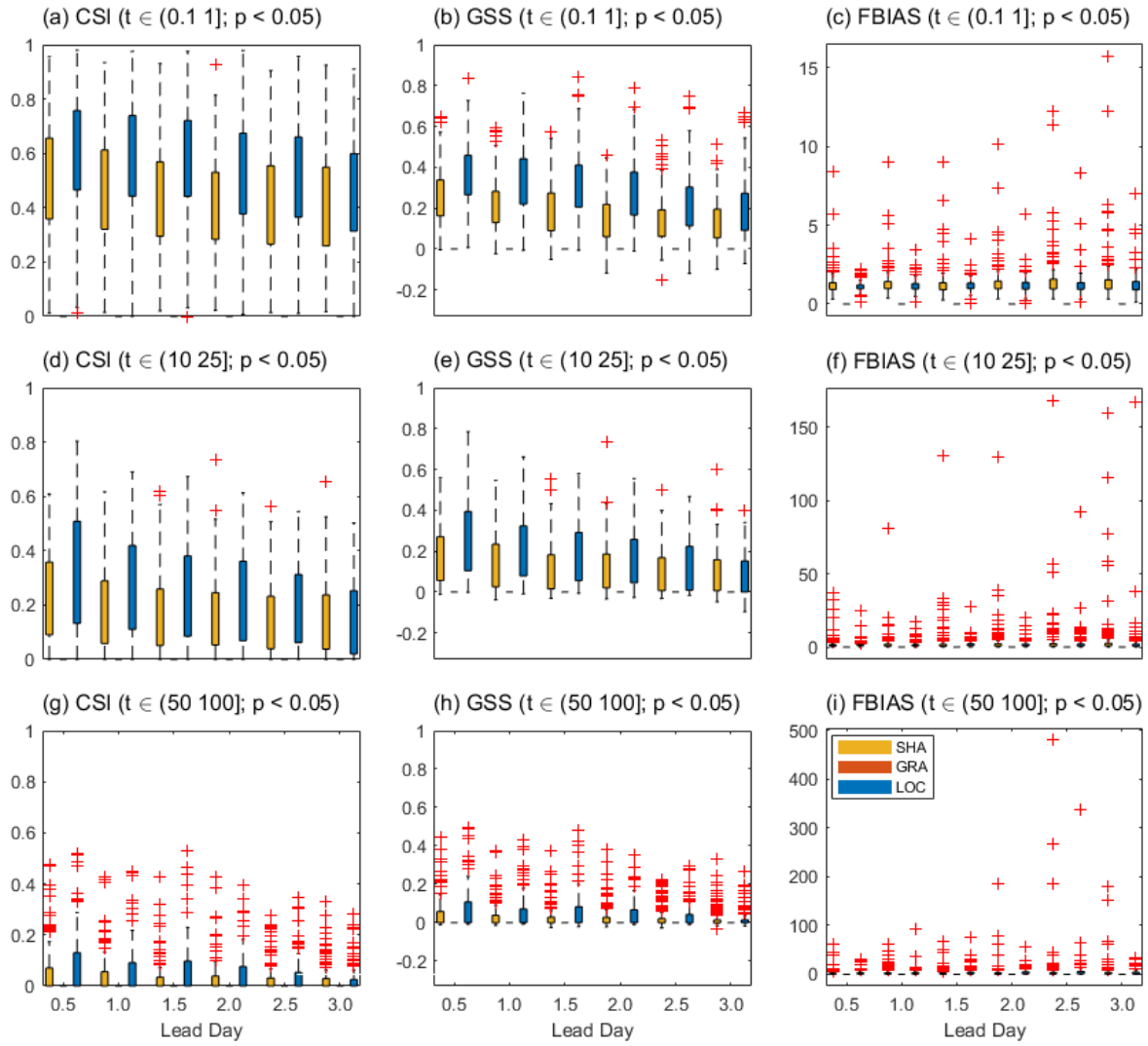

Figure 4. The 2 × 2 contingency table skills. (**a**–**c**), (**d**–**f**), and (**g**–**i**) represent thresholds as 0.1, 10, and 50 (mm), respectively, as a function of the lead day for different forecast fields. The t represents the threshold (units: mm), p represents the significance level that can be obtained from all the available fields, and the red cross symbol (+) of box plot represents for outlier.

Overall, the LOC is more skillful than the other two, and the SHA has the least uncertainties in skills, while GRA has possibly captured the best signal for rainfall or not. Especially, the sharpness of different contingency table skills can be affected by the categorical value, and it also favors one threshold with a large sample number (or optimal threshold). However, the optimal threshold can be small because heavy rainfall is usually a rare event in the real-world. Additionally, these inherent sampling and categorical deficits, called the "double penalties problem" [32,33], have resulted in the dichotomous measurements not to be sharply self-explained for broader application during the comparative evaluation.

4.1.2. Neighborhood

As seen from Figure 5, the LOC has more FSS and AFSS skills on the mean statistics than SHA for both the 0.1 mm and the 10 mm thresholds, while the GRA has no skill. The LOC has no obvious UFSS skill advantage on the mean statistics among the three datasets, while the GRA has the largest FBS mean values at the 0.1 mm threshold but the smallest at the 10 mm threshold. Moreover, more outliers accompanied with less uncertainties of FSS and AFSS for the 0.1 mm threshold than those for the 10 mm threshold can be observed

(Figure 5a,b), which indicates that they are threshold sensitive. In addition, this means that they are sharp for verification of heavier rainfall events. Usually, the discrete small rainfall of the forecast has little chance to be overlapped by the discrete small rainfall of the observation when compared to the continuously neighbored organized rainfall events.

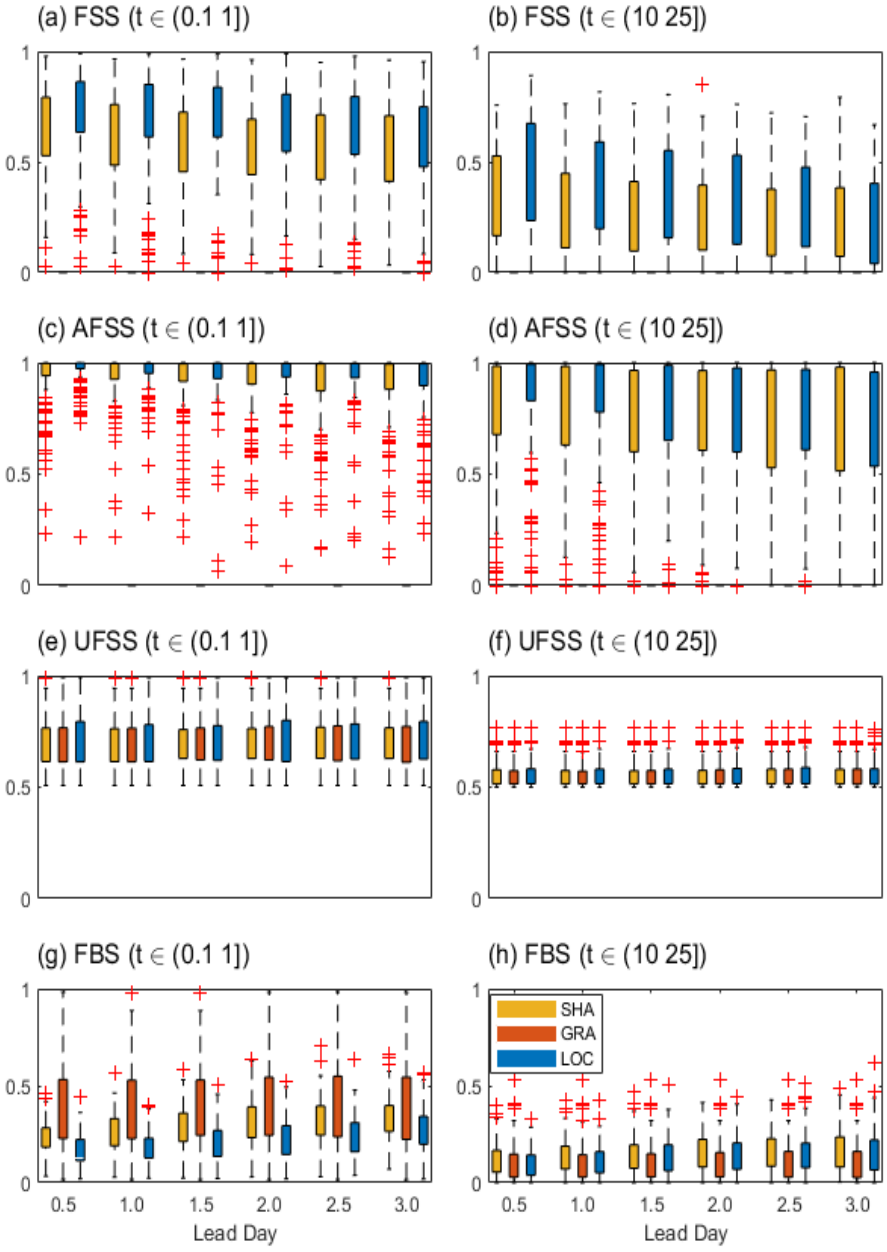

Figure 5. The neighborhood continuous statistics. (**a**–**d**) and (**e**–**h**) represent for FSS, AFSS, UFSS, and FBS at the threshold 0.1 and 10 (mm), respectively, as a function of the lead day for different forecast fields. The red cross symbol (+) of box plot represents for outlier.

The neighborhood window used in FSS and AFSS is one grid in length 0.1° (narrow) and the entire domain (broad), respectively. Better skill performance with less uncertainties can be found in FSS for LOC than that for SHA at the 0.1 mm threshold, and this also can be found in AFSS. This is quite distinguished from the generality of the dichotomous skill, which could give a better performance accompanied with larger uncertainties. This indicates that FSS skill could be less sharp than the dichotomous method to some extent, and the spatial advantages in LOC indicated by AFSS are not robust. Therefore, the choice

of a narrow or broad window size could never be determined for the spatially discrete rainfall during the comparative evaluation.

The zero FSS and AFSS values for GRA usually indicate zero overlaps between forecast and observation on a quite broad neighborhood domain (Figure 5a,b,e,f); however, this is accompanied with no zero FBS values and an equitable observation rate (UFSS). According to the definitions in Table 2, FBS is taken as a correspondence factor of FSS and AFSS. This indicates more possibly unrecognized events in one neighborhood domain, which could be prevalent for FSS. Recall the fact that the GRA has produced extremely weak rainfall but comparatively rainfall events with when compared to the other two datasets (Figure 3a); therefore, the zero FSS and AFSS values could be meaningless.

Overall, LOC outperforms the other two (GRA and SHA) in FSS and AFSS on the averaged skill. Furthermore, FSS favors more the small-value rainfall with non-discrete distributions, and it is possibly less sharp than the dichotomous method. However, a zero FSS value could be possibly meaningless and misleading during the comparative evaluation among distinguished datasets, because no events of both the forecast and observation (or anti hit) in the same verification space defined by the fuzzy neighborhood window is taken as zero skill. This is also mentioned in a nearby study [50]. Thus, FSS is likely unsuitable for the verification of the datasets where highly discrete rainfall frequently occurs during the comparative evaluation.

4.1.3. Displaced

As seen from Figure 6, SHA has larger S1 values on mean statistics than LOC across all the lead days. Furthermore, the S1 values in LOC show more uncertainties than SHA. This indicates that the overall rainfall in LOC is slightly less displaced from the observation when compared to SHA, but mostly during the lead time from 0.5 to 2.5 days. Nevertheless, GRA has no skill.

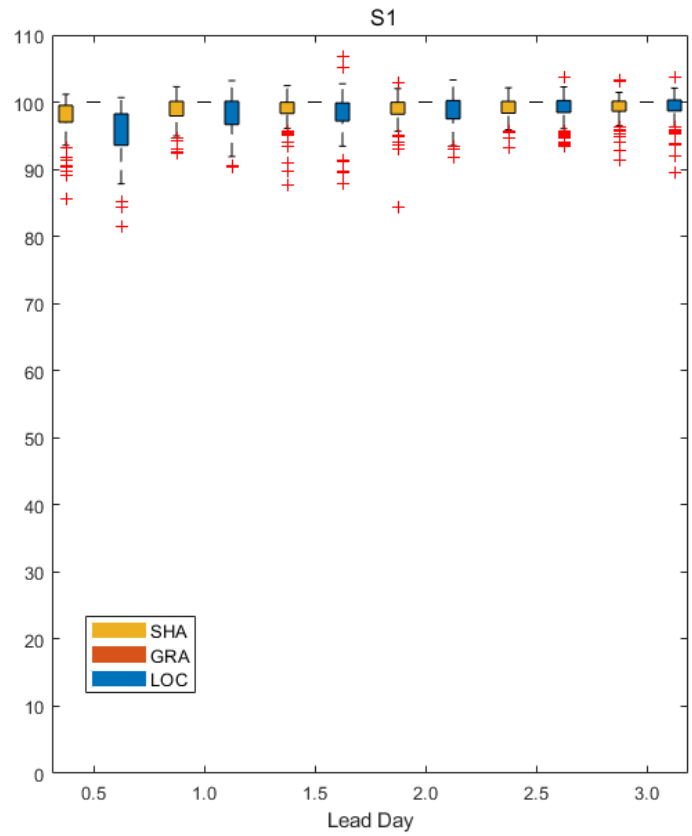

Figure 6. The gradient statistics. S1 score as a function of the lead day for different forecast fields. The red cross symbol (+) of box plot represents for outlier.

Note that S1 only measures the difference in two fields; those higher values (close to 100) than that reported in the previous work [36] indicates that the overall deviation degree from forecast to observation is quite large. Especially, the 100 S1 skill of GRA indicates that the event gradients of the neighbor grid in the forecast or observation field are possibly immeasurable (see Table 2 notes). This should attribute to the large amount of great discrete rainfall event in datasets. Obviously, S1 skill is significantly affected by the high frequency of null rainfall events.

As seen in Figure 7, the whole distance map skills have shown lots of uncertainties. SHA has produced smaller mean HD values than LOC for the 0.1 mm threshold (Figure 7a), while a larger mean HD value of SHA than that of LOC for the 10 mm threshold can be observed (Figure 7b). Meanwhile, LOC has produced smaller mean MZM values than SHA at both 0.1 and 10 mm thresholds (Figure 7c,d). While GRA has no/null HD and MZM skill. The BM skill value of LOC and SHA behave similarly to HD, and GRA has large BM values, which are far from the other two (Figure 7e,f). Obviously, both HD and BM give an opposite skill score estimation for SHA and LOC at the two given thresholds, while MZM insists that LOC has a better performance than SHA at both given thresholds.

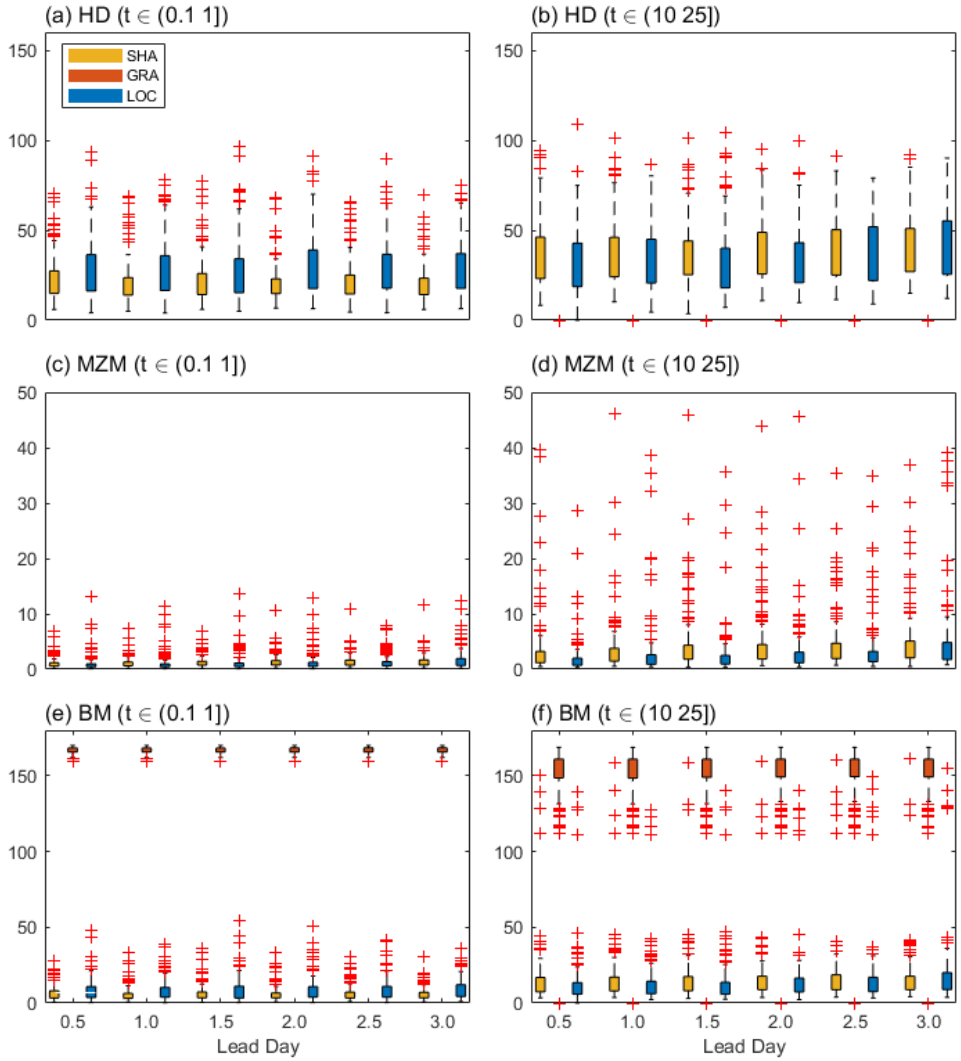

Figure 7. The distance map statistics skills. (**a**,**b**), (**c**,**d**), and (**e**,**f**) represent the measurement for HD, MZM, and BM, respectively, as a function of the lead day for different forecast fields. The red cross symbol (+) of box plot represents for outlier.

Moreover, for the 0.1 mm threshold, HD uncertainties of LOC are larger than those of SHA, but for the 10 mm threshold, those of LOC are likely equitable with those of SHA. Clearly, HD uncertainty performance seems to have nothing to do with HD skill performance in this study. This is similar to BM's behavior; for the 0.1 mm threshold, the least BM uncertainties in GRA is accompanied with the worst skill performance (Figure 7e), which indicates that GRA is the largest displacement deviated with the most confidence among all datasets, while for the 10 mm threshold, this collaborative change phenomenon between uncertainties and skill performance shows the opposite. However, these multipolar differentiation performances related to multi-thresholds in the comparative evaluation can be non-robust because that one data set cannot surely win other datasets in one single skill when the reliability is considered. In contrast, a better MZM performance with less uncertainties indicates that the skill advantage of LOC is more robust than the other two during this study.

Overall, rainfall in LOC is likely to be less displaced from the observation when compared to SHA. This is generally pronounced at 0.5 lead days, indicated by S1 skill (Figure 6), and at the 10 mm threshold, indicated by the distance map skill (Figure 7b,d,f). The prevalent large number outliers of every skill in displaced schemes indicates that they are more easily affected by case differences when compared to FSS or dichotomous skill. This should be attributed to the overstrict distance metrics [34]. However, careful usage should be promoted because unlike MZM and S1 skill, the HD and BM skill uncertainties are quite sensitive to the threshold.

4.1.4. Decomposed

As seen in Figure 8, except for the 50 mm threshold at scale 0.4° (Figure 8f), positive ISC values can be observed. For the 0.1 mm threshold, GRA outperforms the other two at the scale 0.4° and 1.6°, but with largest uncertainties, followed by LOC, and then SHA (Figure 8b,c). Nevertheless, GRA has no/null skill at scale 0.1° and 6.4° (Figure 8a,j). Meanwhile, for the 10 mm threshold, LOC has shown the best ISC skill, followed by SHA, and then GRA. Clearly, GRA has larger uncertainties than the other two (Figure 8b,e,h,k). Moreover, for the 50 mm threshold, great outliers can be observed at scales 0.1° and 6.4° (Figure 8c,i), and the ISC skill advantages in LOC is slight, and only pronounced at the scale 6.4° (Figure 8i).

In the 10 mm threshold at almost all scales, LOC has larger positive ISC values than the other two datasets, which indicates that the convective rainfall events in LOC is the best forecast at almost all scales among the three datasets. GRA can be properly evaluated by using ISC; it shows notable skill advantages for the drizzle events at a broad scale range (from 0.4° to 1.6°), but it could be greatly displaced when compared to the others for the threshold exceeding 10 mm. For the 50 mm threshold, the totally negative skill of all the three datasets (Figure 8f) indicates that the scale of errors between forecast and observation could be possibly larger than 0.4°, while at large scales, such as 6.4°, the general exhibition of quite large positive skills indicates that the rainstorm events of large scales are well forecast easily.

The notable outliers of ISC in both rainstorm and drizzle events indicate that their errors of scales can be heavily changed by case differences. In addition, the uncertainties of ISC are relatively larger for convective events than those for drizzle and rainstorm events. This indicates that ISC rewards the moderate rain most but with sharp scale and threshold discrimination. Generally, it should be noted that ISC can provide the errors of scale that depended on intensity (or threshold), and it is not an accuracy measurement of displacement when compared to the displaced scheme.

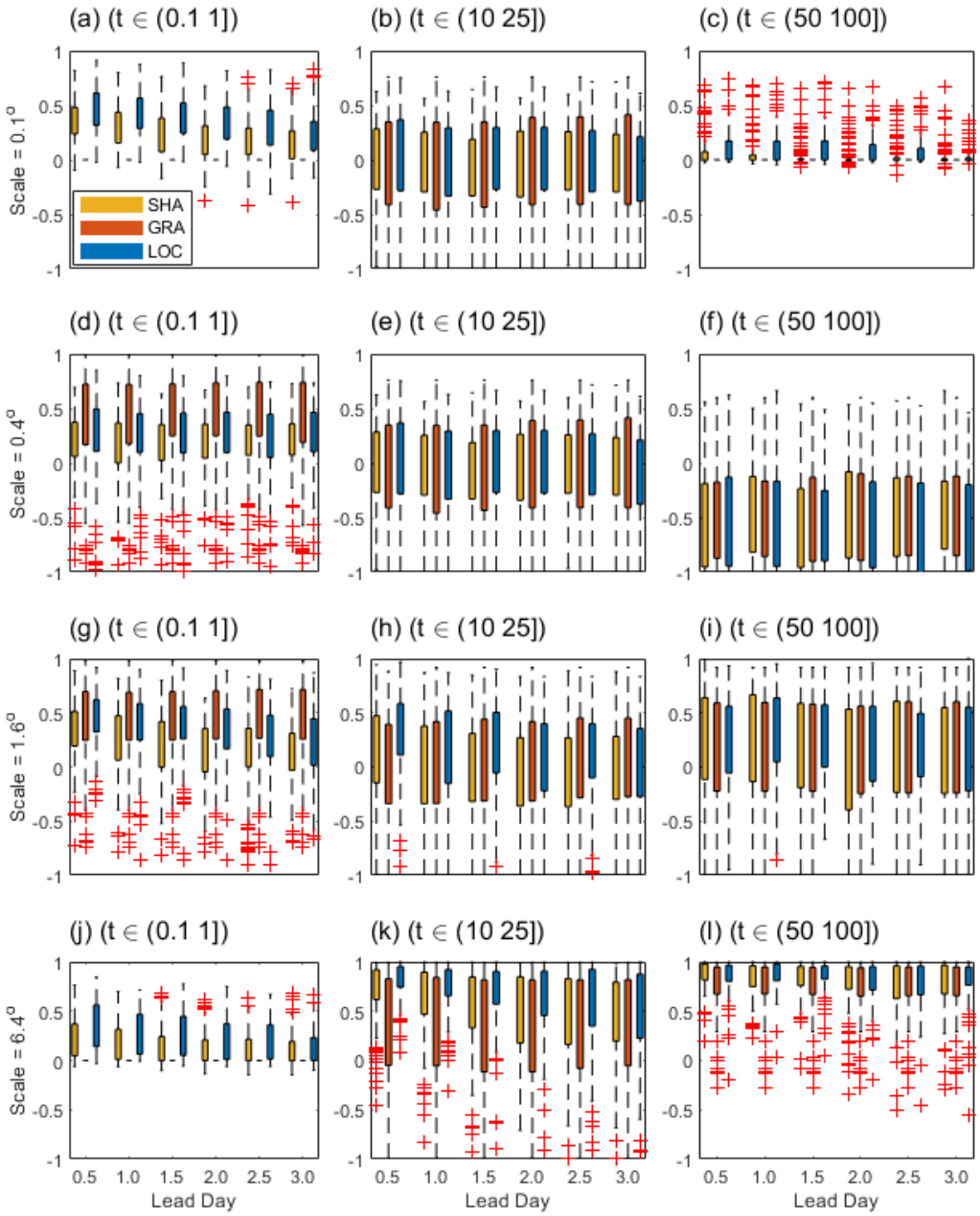

Figure 8. The wavelet intensity scale skill scores. (**a**–**c**), (**d**–**f**), (**g**–**i**), and (**j**–**l**) represent the scales as 0.1, 0.4, 1.6, and 6.4 (degrees), respectively, as a function of the lead day for different forecast fields. The red cross symbol (+) of box plot represents for outlier.

Overall, LOC has predicted the best convective events at any scale, followed by SHA, then GRA. Additionally, the notable ISC skill advantages of GRA for drizzle and rainstorm events indicate that it can avoid the overstrict penalty regular in dichotomous methods, the unrecognized events in neighborhood method, and the multipolar issue in displaced methods. The scale of errors indicated by ISC can be easily related with synoptic systems, and ISC is super suitable for rainfall verification if the accuracy of errors is the secondary needs. However, ISC could be relatively expensive when it is applied to the comparative evaluation because it is not only event-scale sharpened but also event-threshold sharpened.

4.1.5. Featured

As seen in Figure 9, GRA has no TIN skill. Recall the fact that the events in GRA greater than 10 mm are rare and totally displaced, which has resulted in the object clustering

between the forecast and observation having too large a distance difference, which can further result in null total interest. This could be attributed to the small convolution radius around 6 Km (about 0.06°) to some extent. However, since too large convolution radius can cause meaningless objects of MODE, TIN is likely not suitable for the highly discrete event verification.

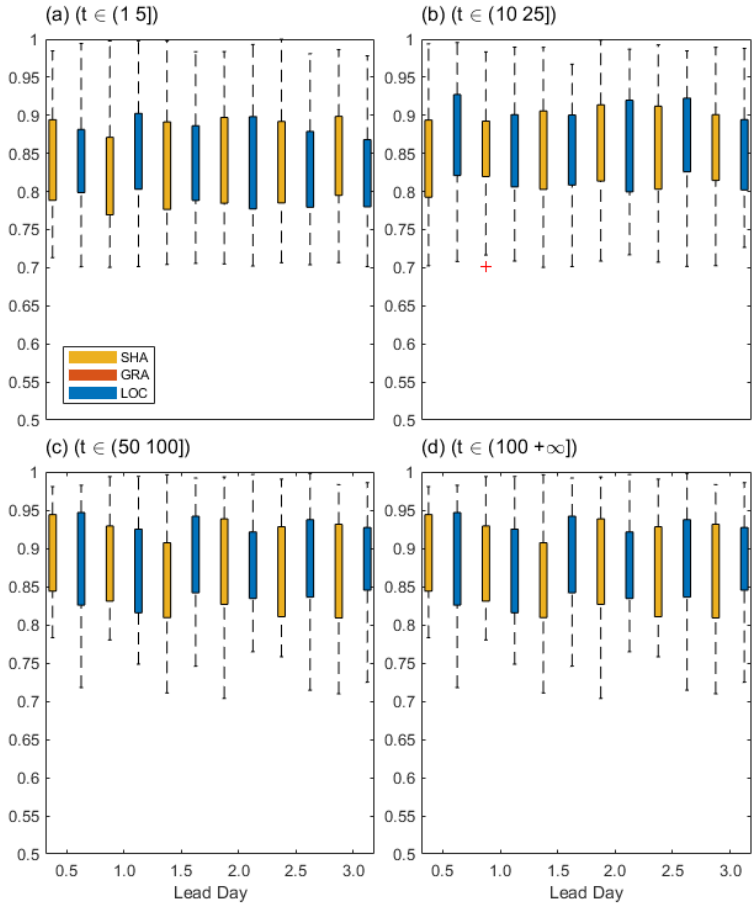

Figure 9. The TIN skill score. (**a**–**d**) represent for thresholds as 1, 10, 50, and 100 (mm) respectively as a function of the lead day for different forecast fields. The red cross symbol (+) of box plot represents for outlier.

For the 1 mm threshold, the mean TIN value of SHA is slightly larger than that of LOC (Figure 9a) at all lead days except the lead day 1. However, for the 10 mm threshold, the mean TIN value of LOC is slightly larger than that of SHA at all lead days except the lead day 3 (Figure 9b). This indicates that the generally relative skill advantage in LOC or SHA is slightly different between the drizzle and convection events, and to the point where this can negligible.

Moreover, for both the 50 and 100 mm thresholds, LOC outperforms SHA at all lead days except the lead day 1 and 2 (Figure 9c,d). This indicates that for the TIN skill of rainstorm or larger threshold events, LOC and SHA have a winner or loser for each other. While in fact, the two kinds of rainfall events possibly occur simultaneously.

The broad TIN uncertainties for both LOC and SHA indicate TIN is highly sensitive to the total interest of different object clusters. For the 1 and 10 mm thresholds, the mean TIN varies around 0.85 (Figure 9a,b), while for the 50 and 100 mm thresholds, it varies around 0.9 (Figure 9c,d). This indicates that TIN favors to reward large threshold events, which can be related to rare rainstorms. Generally, the prevalent broad uncertainty in both LOC and SHA for all thresholds indicate that TIN uncertainties are not sensitive to threshold and event difference during comparative evaluation. Furthermore, the almost unchanged mean

value of TIN in both LOC and SHA indicates that the two datasets have almost equitable spatial characteristics.

Overall, LOC has shown a slight advantage of spatial similarity at most lead days for the threshold exceeding 10 mm. Meanwhile, the lead 1~2 day rainstorm forecasts of SHA are likely more similar with observations than LOC. TIN can provide the estimation of spatial difference indicated by composited object attributes, and it is hopefully accurate because the size of object could be reduced to one point in the limit condition. It can be promoted to evaluate the datasets with seriate events, such as convection or rainstorms, which can be easily related to the synoptic systems. Nevertheless, it is computationally expensive.

4.2. Spatial Characteristics

GRA is totally displacement deviated from the spatially continuous rainfall at a large threshold. Thus, the main rainfall events during the summer of 2020 in both LOC and SHA are further selected for comparison in a forecaster desired way or an observational preferable way. Spatial characteristics as the object clusters and energy squared relative difference (En2RD) that are derived from MODE and Wavelet Analysis, respectively, are further compared to identify their synoptic insights.

4.2.1. The Object Clusters Comparison

It should be noted that clustering is conducted as a two-step technique (merging and matching) in the fuzzy logic method of MODE. Merging refers to grouping together objects in a single field, and matching refers to grouping together objects in different fields, typically the forecast and observed fields [42,43]. In this work, since the initial fields derived from different datasets have been interpolated into identical grids, and the objects clusters of the observation field in one specific event case are mostly equal for different forecast datasets at different leading days; therefore, here we take all matched or unmatched objects clusters in the observation as one overall cluster for comparison convenience.

As seen in Figure 10, three kinds of object clusters including matched and unmatched for the 10 mm threshold are compared between LOC and SHA. Matched clusters for both LOC and SHA can be observed for all events. However, the observed arcuate cluster in case 1 has been clearly under forecast (Figure 10(Aa–Af)), while SHA is relatively less biased than LOC at the 2 lead day (Figure 10(Ad)).

Meanwhile, for case 5, 6, and 10, isolated clusters in the middle north area of Henan province can be observed, which should be related to convection. For case 5, LOC has produced a similar northwest-biased convection when compared to the observation during the lead day from 1 to 2.5, while SHA has totally missed (Figure 10(Ea–Ef)). Moreover, for cases 6, 8, and 10, unmatched clusters mostly in LOC can be observed, while the much larger convection area for both LOC and SHA indicate an over forecast.

As seen from Figure 11, except for case 4, both LOC and SHA can capture the observed clusters for the 50 mm threshold, which should relate to rainstorms. Especially for case 5, 8, and 9, LOC and SHA can well forecast the large area of rainstorms at a lead time from 2.5 up to 3 days. While for case 2, 3, 10, and 11, lots of unmatched cluster pairs in LOC and SHA indicate that forecasts of the small area rainstorms have heavy displacement error and are over forecasted.

In general, both LOC and SHA have shown almost equitable abilities in the 10 mm threshold rainfall forecast, while unmatched pairs of isolated object clusters for both LOC and SHA indicate an over forecast of local convection. Moreover, the well forecasted large area of observed clusters indicates good abilities of both LOC and SHA in large-scale rainstorms, while the unmatched pairs of small areas in LOC indicate an over forecast of local rainstorms.

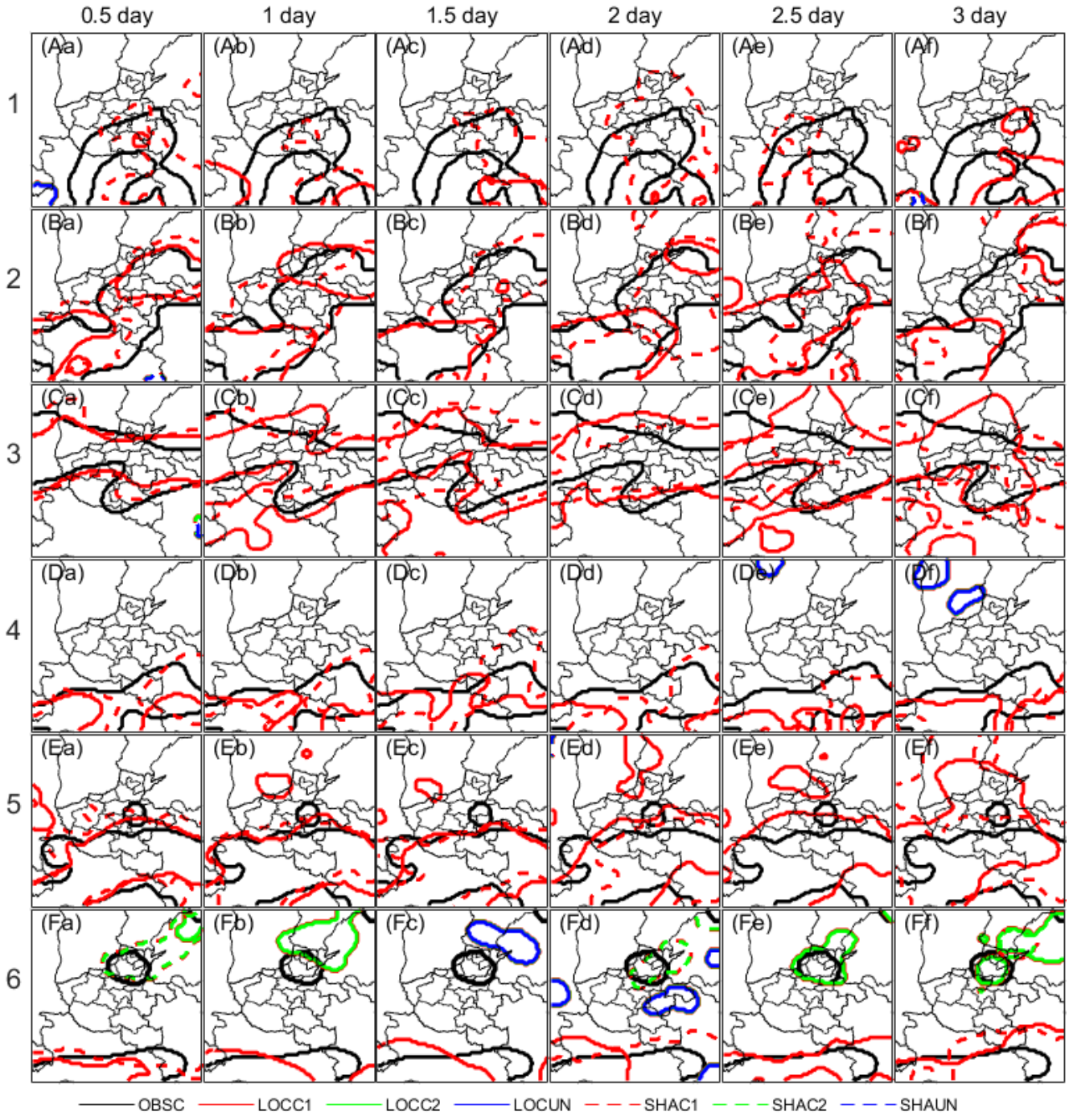

Figure 10. *Cont.*

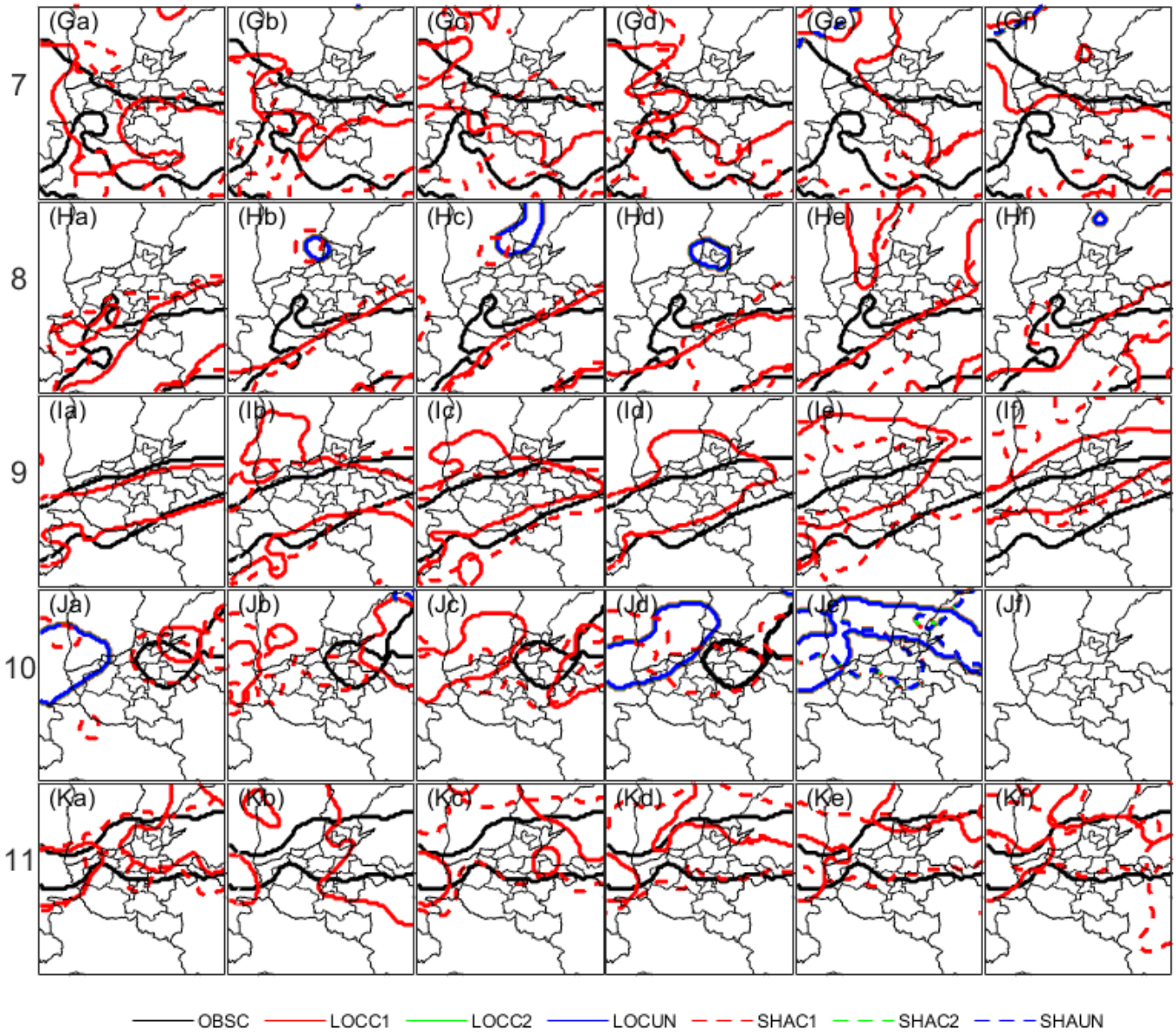

Figure 10. The 10 mm object clusters in different cases. The eleven events are shown in line (**A***–**K***), while the lead day forecasts are shown in row (***a**–***f**). OBSSC = Clusters of OBS; LOCC1 = Matched cluster pair 1 of LOC, LOCC2 = Matched cluster pair 2 of LOC, and LOCUN = Unmatched cluster pairs of LOC; SHAC1 = Matched cluster pair 1 of SHA, SHAC2 = Matched cluster pair 2 of SHA, and SHAUN = Unmatched cluster pairs of SHA.

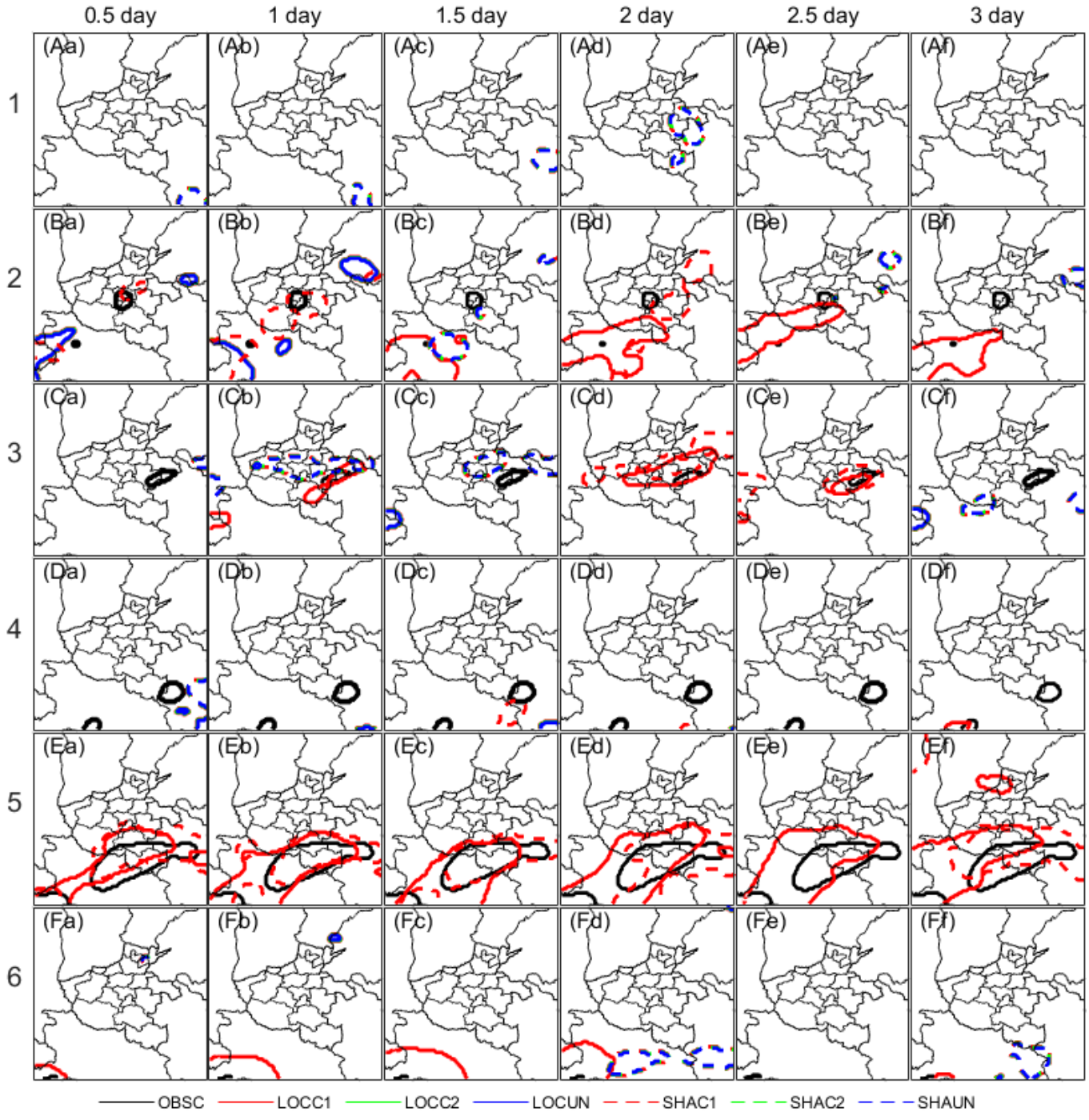

Figure 11. *Cont.*

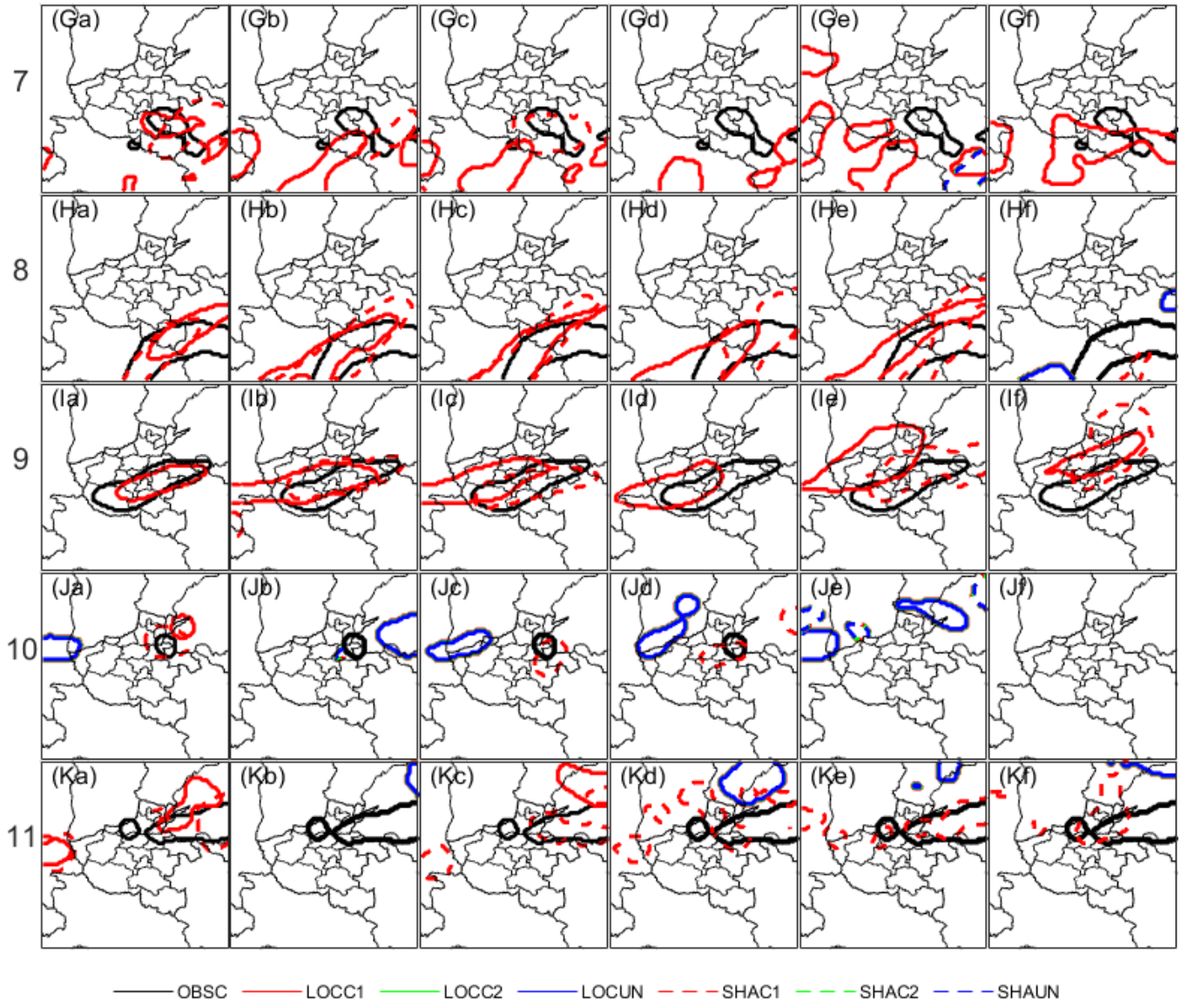

Figure 11. Same as Figure 10 but for the 50 mm object clusters.

However, these objects clusters are hopefully to recognize the similarity or small difference of the overall geometry shape of events in two different fields, because the smaller or isolated objects with large distance errors usually cause null interest. Therefore, it could be not suitable for the largely displaced or non-event cases. In the contrast, these object clusters can be easily related to the specific rainfall events if they are properly predefined, and they can provide an intuitionistic comparison on spatial differences between the forecast and observation. These differences in datasets should be directly attributed to the meteorological systems related with different physical process [7,8,42,43]. This means they are quite suitable for the physically systematic events verification, especially in convection and rainstorm systems used in rainfall verification.

4.2.2. The En2 Relative Difference

It should be noted that the scale components are derived from the decomposition of the wavelet analysis method for each individual threshold field, and are jointly displayed, and so is En2RD [41]. The null scale components of the wavelet analysis could be negligible if the samples for any given threshold are sufficient, e.g., the large area multi-scale precipitation cases during this study. The mean En2RD of multi-scale rainfall cases can account

for the overall difference between forecast and observation squared energies relative to their magnitude.

As seen in Figure 12, the mean En2RD value has shown a distinguished scale and threshold dependence between LOC and SHA. For LOC, the positive values vary between the 1 and 25 mm threshold, and this is pronounced at the scale between 0.1° and 0.8° for all the lead days (Figure 12a–f). This indicates that the LOC has shown an over forecast of events when their thresholds ranged from 1 to 25 mm and scales ranged from 0.1° to 0.8°, and an under forecast of events when their thresholds are larger than 50 mm. Meanwhile, for SHA, the values cover the whole scale and threshold axis. This indicates the SHA forecasts exhibit over forecasts for all thresholds on all scales, pronounced during the lead days among 1~3. Both the LOC and SHA have shown nearly perfect forecasts at the 0.5 lead day. Moreover, it is noted that compared to a previous case, where the En2RD can have a large variation range (−1~1) [41], the small range (−0.1~0.1) for both LOC and SHA indicate that the under forecast and over forecast magnitude is relatively small.

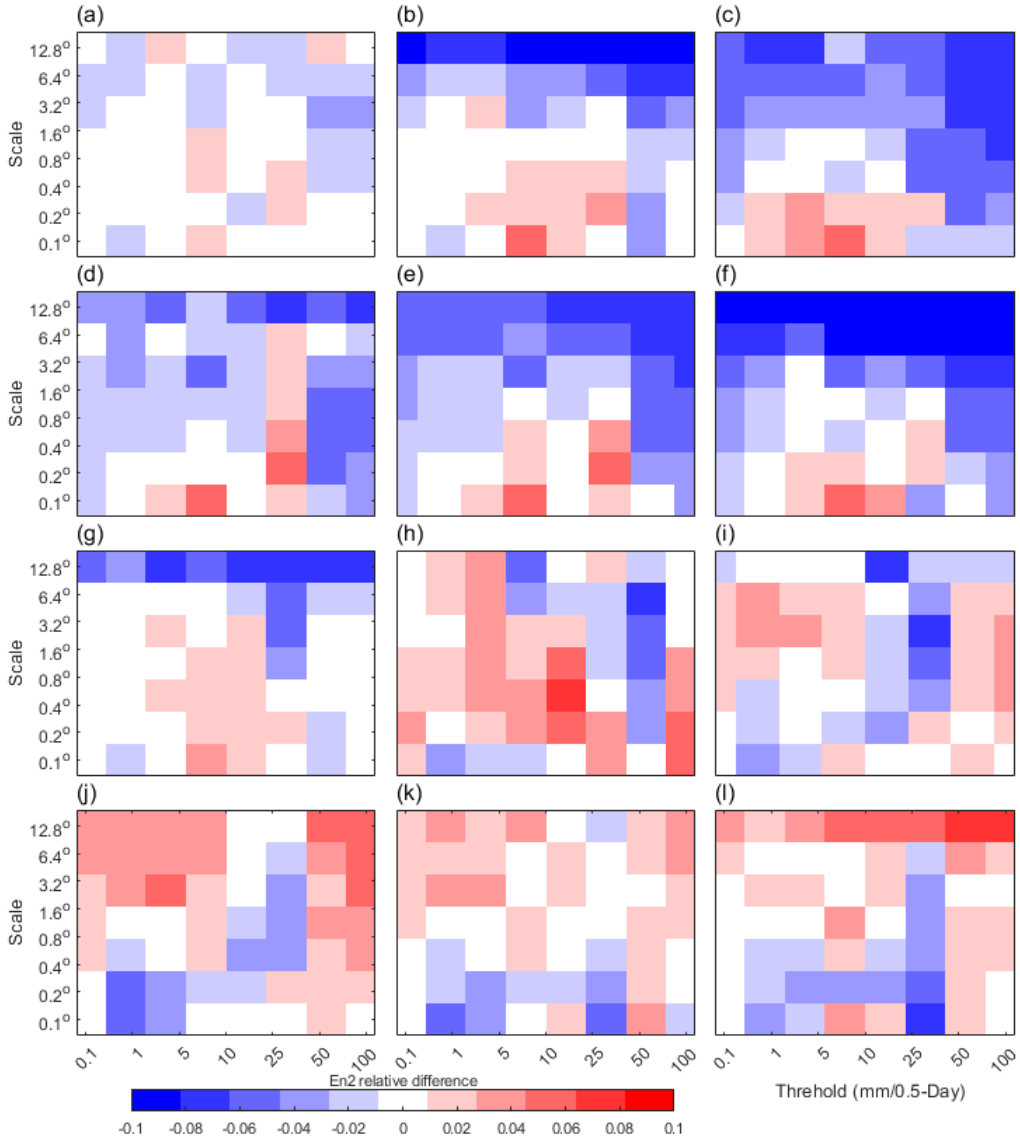

Figure 12. The mean En2 relative differences (En2RD) of all selected cases. (**a**–**f**) are the men En2RD of 0.5, 1, 1.5, 2, 2.5, and 3 lead days, respectively, for LOC, and (**g**–**l**) are the same as (**a**–**f**), but for SHA.

The abovementioned null scale components can be caused by the improper threshold or non-event fields; therefore, En2RD may not be suitable for the cases with too little

rainfall. Meanwhile, the relative magnitude indicated by En2RD could be sensitive to cases, which could make the ranges of the mean En2RD value to be not so meaningful during the multi-cases applications.

Overall, SHA favors over forecasting on a quite a broad scale range, and LOC likely misses the rainfall exceeding 100 mm. The En2RD could provide very intuitive spatial differences distinguished by scales or thresholds, but could be more informative on if applied to the multi-scale rainfall events.

5. Summary

By using various skill scores and spatial characteristics, the gridded precipitation observation CMPAV2.0 and three datasets derived from local (LOC), Shanghai (SHA), and Grapes (GRA) models, respectively, are conducted to assess the 3 lead day and 0.5 day interval rainfall forecasts during the summer of 2020 over Central East China. The results are concluded as follows.

1. For dichotomous measurements, LOC is more skillful than the other two, and the SHA has the least uncertainties in skills, while GRA has captured the best signal for rainfall or not. For neighborhood measurements, LOC slightly outperforms SHA in FSS, AFSS, and FBS skills, but relatively large uncertainties of FSS in LOC can be identified. This indicates that both LOC and SHA forecasts can overlap the observation at a broad neighborhood window, but LOC has more uncertainties.
2. LOC is generally less displaced than SHA for S1, and more pronounced on the lead 0.5 day. Less displacement errors of LOC than that of SHA also can be found for MZM. This advantage of LOC can only be found at the 10 mm threshold for both HD and BM. Moreover, LOC has more intensity scale skills than the other two for the 10 mm threshold at almost all scales. GRA likely has large displacement errors when compared to the other two datasets. In addition, LOC shows slight advantages in spatial similarity with observations when compared to SHA.
3. Both LOC and SHA have shown almost equitable abilities in convection and rainstorms forecast of the large areas but slightly over forecasts in the local convection, while LOC likely over forecasts the local rainstorms. Moreover, the 1~2 lead day rainstorm forecasts of SHA are more similar with observations than LOC. SHA slightly favors over forecasting on a broad scale range and a broad threshold range, and LOC slightly misses the rainfall exceeding 100 mm.

The popularly dichotomous and neighborhood skill advantages of LOC can be identified by using a collection of measurements, and that GRA has few popular skills should be attributed to sampling errors related to its very little heavy rain and particularly heavy drizzle. Moreover, the largely deviated rainfall forecast of GRA can be identified by using different displace measurements, while LOC has slightly little advantage in displacement when compared to SHA, which is pronounced on the lead 0.5 day and/or at the 10 mm threshold. Furthermore, LOC is more spatially similar with the observations than SHA. In addition, both LOC and SHA have shown almost equitable abilities in convection and rainstorms forecast of the large area but with a slight over forecast.

The dichotomous methods are sharpened on the quality of datasets but could be blind to model developers and datasets users because the overstrict penalty makes frequently immeasurable zero or null values, but these are possibly meaningful for the comparative evaluation. The spatial skills derived from the neighborhood, displaced, decomposed, and featured schemes have clearly broadened the dimensionality of rainfall verification; however, during this comparative study, the neighborhood and decomposed skills are likely fuzzy, while the displaced skills behaviors multipolar differentiation performance, and the featured skills are likely too sensitive to the spatial geometry of rainfall event distribution. Especially, the abundant spatial characteristics derived from the decomposed and featured schemes could be powerful assistance of subjective decisions for forecasters.

Nevertheless, except for the ISC and TIN skill, almost all other measurements give too little reward on rainstorm forecast or even heavier rainstorm forecast of all models when

compared to the lighter rainfall forecast during this work. This might indicate that the rare events are too difficult to forecast, or more likely their poor abilities in verifying forecasts with rainstorm or heavier rainstorm events. Since the severe rainfall events accompanied with severe social influence frequently occur recently under a global warming background, the rewards on model abilities in severe weather forecasting including rainfall should be emphasized.

6. Discussion

Although comparative evaluation can map the overall performance of datasets and methods, there are still limitations during this study that the basis of verification can be varied, and/or the verification can be too broad. The event climatic occurrence background could be quite varied in different study areas and time periods, e.g., the dry lands or tropics, the wet or dry years, and so on. Therefore, the categorical values in this study should be cautious in different regions. In addition, the application purpose should be quite different, e.g., decision-making on public service of one event or a fundamental metric of post-proceed training methods usually as a reference to narrow metric candidates and take the possibly chanceful advantages of nearby forecasts to get their hopeful rewards, but not broader skill advantages or differences, as in this study. Therefore, besides the dichotomous metrics, other spatial metrics with their desired accuracy (e.g., displaced and/or featured) could be taken as additional candidates for rewarding regular rainfall decisions, and other metrics should be further studied.

For a fair or equitable comparative evaluation of distinguished rainfall datasets to tell model performance over a specific area, one single skill of limited datasets can be far from enough. Long time series (e.g., one year or more) datasets are usually desired in the verification to identify robust model performance under a changed climatic background. Furthermore, the model physics related to rainfall characteristics are usually temporal–spatial scale issued and parameterized with hopefully solved empirical assumptions, and the decomposed scheme can give a clear view of a decomposed spatial–scale dependent estimate; this can simplify this spatial scale estimation issue of model physics to some extent. More cases should be studied in future work to enhance the application of ISC metrics.

The generally notable skill uncertainties of datasets during this study, which should be likely attributed to the model or method theory difference, indicate that the model or method uncertainties in precipitation forecast can be great. This could be hopefully addressed by the work considering the uncertainties of models or methods [51]. In general, for a broadly comparative evaluation on rainfall, the popular dichotomous methods should be recommended under considering reasonable classification of thresholds if the accuracy is highly demanded. Most spatial methods are suggested to be conducted with proper pre-handling of non-rainfall event cases. Especially, the spatial characteristic difference information could be recommended in a computationally sufficient environment.

Author Contributions: Conceptualization, methodology, validation, formal analysis, Y.G.; investigation, resources, and data curation, C.S.; writing—original draft preparation, C.S.; writing—review and editing, C.S.; visualization, C.S.; supervision, C.S.; project administration, A.S.; funding acquisition, C.S. and A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Science and Technology Project on Innovation Ecosystem Construction at Zhengzhou Supercomputing Center in Henan province (grant number: 201400210800), China Environmental Protection Foundation Blue Mountain Project (grant number: CEPFQS202169-28), and China Meteorological Administration Meteorological Observation Centre "Chip Project" (Xiaoman I).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request.

Acknowledgments: We would like to give our many thanks to those who have made their efforts in advancing this work, and the fellow travelers along the way.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Elahi, E.; Khalid, Z.; Tauni, M.Z.; Zhang, H.; Xing, L.R. Extreme weather events risk to crop-production and the adaptation of innovative management strategies to mitigate the risk: A retrospective survey of rural Punjab, Pakistan. *Technovation* **2021**, *117*, 102255. [CrossRef]
2. Elahi, E.; Khalid, Z.; Zhang, Z.X. Understanding farmers' intention and willingness to install renewable energy technology: A solution to reduce the environmental emissions of agriculture. *Appl. Energy* **2022**, *309*, 118459. [CrossRef]
3. Elahi, E.; Zhang, Z.X.; Khalid, Z.; Xu, H. Application of an artificial neural network to optimise energy inputs: An energy-and cost-saving strategy for commercial poultry farms. *Energy* **2022**, *244*, 123169. [CrossRef]
4. Abbas, A.; Waseem, M.; Ahmad, R.; Khan, K.A.; Zhao, C.; Zhu, J. Sensitivity analysis of greenhouse gas emissions at farm level: Case study of grain and cash crops. *Environ. Sci. Pollut. Res.* **2022**, *29*, 82559–82573. [CrossRef] [PubMed]
5. Abbas, A.; Zhao, C.Y.; Waseem, M.; Khan, K.A.; Ahmad, R. Analysis of Energy Input–Output of Farms and Assessment of Greenhouse Gas Emissions: A Case Study of Cotton Growers. *Front. Env. Sci.* **2022**, *9*, 826838. [CrossRef]
6. Rodwell, M.J.; Richardson, D.S.; Hewson, T.D.; Haiden, T. A new equitable score suitable for verifying precipitation in numerical weather prediction. *Q. J. R. Meteorol. Soc.* **2010**, *136*, 1344–1363. [CrossRef]
7. Pan, L.J.; Zhang, H.F.; Wang, J.P. Progress on verification methods of numerical weather prediction. *Adv. Earth Sci.* **2014**, *29*, 327–335. (In Chinese)
8. Li, J.; Hsu, K.L.; AghaKouchak, A.; Sorooshian, S. Object-Based Assessment of Satellite Precipitation Products. *Remote Sens.* **2016**, *8*, 547. [CrossRef]
9. Shen, F.; Song, L.X.; Li, H.; He, Z.; Xu, D. Effects of different momentum control variables in radar data assimilation on the analysis and forecast of strong convective systems under the background of northeast cold vortex. *Atmos. Res.* **2022**, *230*, 106415. [CrossRef]
10. Xu, D.; Yang, G.; Wu, Z.; Shen, F.; Li, H.; Zhai, D. Evaluate Radar Data Assimilation in Two Momentum Control Variables and the Effect on the Forecast of Southwest China Vortex Precipitation. *Remote Sens.* **2022**, *14*, 3460. [CrossRef]
11. Song, L.; Shen, F.; Shao, C.; Shu, A.; Zhu, L. Impacts of 3DEnVar-Based FY-3D MWHS-2 Radiance Assimilation on Numerical Simulations of Landfalling Typhoon Ampil (2018). *Remote Sens.* **2022**, *14*, 6037. [CrossRef]
12. Zhang, X.; Xu, D.; Liu, R.; Shen, F. Impacts of FY-4A AGRI Radiance Data Assimilation on the Forecast of the Super Typhoon "In-Fa" (2021). *Remote Sens.* **2022**, *14*, 4718. [CrossRef]
13. Shu, A.; Shen, F.; Jiang, L.P.; Zhang, T.; Xu, D. Assimilation of Clear-sky FY-4A AGRI radiances within the WRFDA system for the prediction of a landfalling Typhoon Hagupit (2020). *Atmos. Res.* **2022**, *283*, 106556. [CrossRef]
14. Shen, F.; Min, J. Assimilating AMSU-A radiance data with the WRF hybrid En3DVAR system for track predictions of Typhoon Megi (2010). *Adv. Atmos. Sci.* **2015**, *32*, 1231–1243. [CrossRef]
15. Shen, F.; Min, J.; Xu, D. Assimilation of radar radial velocity data with the WRF Hybrid ETKF—3DVAR system for the prediction of Hurricane Ike (2008). *Atmos. Res.* **2016**, *169*, 127–138. [CrossRef]
16. Shen, F.; Xu, D.; Xue, M.; Min, J. A comparison between EDA-EnVar and ETKF-EnVar data assimilation techniques using radar observations at convective scales through a case study of Hurricane Ike (2008). *Meteorol. Atmos. Phys.* **2017**, *130*, 649–666. [CrossRef]
17. Shen, F.; Xu, D.; Min, J.; Chu, Z.; Li, X. Assimilation of radar radial velocity data with the WRF Hybrid 4DEnVar system for the prediction of Hurricane Ike (2008). *Atmos. Res.* **2020**, *230*, 104622. [CrossRef]
18. Ma, X.L.; Zhuang, Z.R.; Xue, J.S.; Lu, W.S. Development of the 3DVar system for the non hydrostatic numerical prediction model of GRAPES. *Acta Meteorol. Sin.* **2009**, *67*, 11. (In Chinese)
19. Chen, B.D.; Wang, X.F.; Li, H.; Zhang, L. An Overview of the Key Techniques in Rapid Refresh Assimilation and Forecast. *Adv. Meteorol. Sci. Tech.* **2013**, *3*, 29–35. (In Chinese)
20. Guo, Y.K.; Su, A.F. A Meteorological Data Acquisition Method, Device, Computer Equipment, and Storage Medium. CN115392533A, 2022. Available online: https://patents.google.com/patent/CN109819044A/en (accessed on 17 December 2022). (In Chinese).
21. Du, L.M.; Ke, Z.J. A Verification Approach for the Assessment of Extend-range Process Event Prediction. *J. Appli. Meteoro. Sci.* **2013**, *24*, 686–694. (In Chinese)
22. Zhang, H.F.; Pan, L.J.; Yang, X. Compararive Analysis of Precipitation Forecasting Capabilities of ECMWF and Japan High-Resolution Models. *Meteorol. Mon.* **2014**, *40*, 424–432. (In Chinese)
23. Murphy, A.H.; Winkler, R.L. A general framework for forecast verification. *Mon. Weather. Rev.* **1987**, *115*, 1330–1338. [CrossRef]
24. Jolliffe, I.T.; Stephenson, D.B. *Forecast Verification. A Practitioner's Guide in Atmospheric Science*; Wiley and Sons Ltd.: Hoboken, NJ, USA, 2012; p. 240.
25. Gandin, L.S.; Murphy, A.H. Equitable scores for categorical forecasts. *Mon. Weather. Rev.* **1992**, *120*, 361–370. [CrossRef]
26. Heidke, P. Calculation of success and good of strong wind forecasts in storm warning service (Berechnung der erfolges und der gute der windstarkevorhersagen im sturmwarnungdienst). *Geogr. Ann.* **1926**, *8*, 301–349. (In German)
27. Gerrity, J.P., Jr. A note on Gandin and Murphy's equitable skill score. *Mon. Weather. Rev.* **1992**, *120*, 2707–2712. [CrossRef]

28. Hanssen, A.W.; Kuipers, W.J.A. On the Relationship between the Frequency of Rain and Various Meteorological Parameters. Meded. En Verh.; KNMI: Utrechtseweg, The Netherlands, 1965; p. 65.
29. Doswell, C.A.; Davies-Jones, R.; Keller, D. On summary measures of skill in rare event forecasting based on contingency tables. *Wea. Forecast.* **1990**, *5*, 576–585. [CrossRef]
30. Murphy, A.H. Forecast verification, Its complexity and dimensionality. *Mon. Weather. Rev.* **1991**, *119*, 1590–1601. [CrossRef]
31. Murphy, A.H. What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecast.* **1993**, *8*, 281–293. [CrossRef]
32. Gilleland, E.; Ahijevych, D.A.; Brown, B.G.; Ebert, E. Verifying Forecasts Spatially. *Bull. Amer. Meteor. Soc.* **2010**, *91*, 1365–1373. [CrossRef]
33. Dorninger, M.; Gilleland, E.; Casati, B.; Mittermaier, M.P.; Ebert, E.E.; Brown, B.G.; Wilson, L.J. The setup of the mesovict project. *Bull. Amer. Meteor. Soc.* **2018**, *99*, 1887–1906. [CrossRef]
34. Brown, B.; Jensen, T.; Halley-Gotway, J.; Bullock, R.; Gilleland, E.; Fowler, T.; Newman, K.; Adriaansen, D.; Blank, L.; Burek, T.; et al. The Model Evaluation Tools (MET), More than a Decade of Community-Supported Forecast Verification. *Bull. Amer. Meteor. Soc.* **2021**, *102*, E782–E807. [CrossRef]
35. Ebert, E.E. Fuzzy verification of high-resolution gridded forecasts, a review and proposed framework. *Meteorol. Appli.* **2008**, *15*, 51–64. [CrossRef]
36. Teweles, S.; Wobus, H.B. Verification of prognostic charts. *Bull. Amer. Met. Soc.* **1954**, *35*, 455–463. [CrossRef]
37. Ahijevych, D.; Gilleland, E.; Brown, B.G. Application of spatial verification methods to idealized and nwp-gridded precipitation forecasts. *Wea. Forecast.* **2009**, *29*, 1485–1497. [CrossRef]
38. Zhu, M.; Lakshmanan, V.; Zhang, P.; Hong, Y.; Cheng, K.; Chen, S. Spatial verification using a true metric. *Atmos. Res.* **2011**, *102*, 408–419. [CrossRef]
39. Gilleland, E. Novel measures for summarizing high-resolution forecast performance. *Adv. Statist. Climatolo. Meteorol. Oceanogr.* **2021**, *7*, 13–34. [CrossRef]
40. Gilleland, E.; Lee, T.C.M.; Halley-Gotway, J.; Bullock, R.G.; Brown, B.G. Computationally efficient spatial forecast verification using Baddeley's delta image metric. *Mon. Weather. Rev.* **2008**, *136*, 1747–1757. [CrossRef]
41. Casati, B.; Ross, G.; Stephenson, D. A new intensity-scale approach for the verification of spatial precipitation forecasts. *Meteorol. Appli.* **2004**, *11*, 141–154. [CrossRef]
42. Davis, C.A.; Brown, B.G.; Bullock, R.G. Object-based verification of precipitation forecasts, Part I, Methodology and application to mesoscale rain areas. *Mon. Weather. Rev.* **2006**, *134*, 1772–1784. [CrossRef]
43. Davis, C.A.; Brown, B.G.; Bullock, R.G. Object-based verification of precipitation forecasts, Part II, Application to convective rain systems. *Mon. Weather. Rev.* **2006**, *134*, 1785–1795. [CrossRef]
44. Brooks, H.E.; Doswell, C.A. A comparison of measures-oriented and distributions-oriented approaches to forecast verification. *Wea. Forecast.* **1996**, *11*, 288–303. [CrossRef]
45. Stephenson, D.B.; Doblas-Reyes, F.J. Statistical methods for interpreting Monte Carlo forecasts. *Tellus* **2000**, *52A*, 300–322. [CrossRef]
46. Seaman, R.; Mason, I.; Woodcock, F. Confidence intervals for some performance measures of yes/no forecasts. *Austral. Met. Mag.* **1996**, *45*, 49–53.
47. Wilks, D.S. *Statistical Methods in the Atmospheric Sciences: An Introduction*; Academic Press: San Diego, CA, USA, 1995.
48. Gilleland, E. *Confidence Intervals for Forecast Verification. NCAR Technical Note NCAR/TN-479+STR*; UCAR: Boulder, CO, USA, 2010; p. 71.
49. Shen, Y.; Zhao, P.; Pan, Y.; Yu, J. A high spatiotemporal gauge-satellite merged precipitation analysis over China. *J. Geophys. Res. Atmos.* **2014**, *119*, 3063–3075. [CrossRef]
50. Mittermaier, M.P. A "meta" analysis of the fractions skill score: The limiting case and implications for aggregation. *Mon. Weather. Rev.* **2018**, *149*, 3491–3504. [CrossRef]
51. Zhi, X.F.; Peng, T.; Wang, Y.H. Extended range probabilistic forecast of surface air temperature using Bayesian model averaging. *Trans. Atmos. Sci.* **2018**, *41*, 627–636. (In Chinese)