https://doi.org/10.5194/egusphere-2023-2613
Preprint. Discussion started: 18 December 2023

A General Comprehensive Evaluation Method for Cross-Scale Precipitation Forecasts

Bing Zhang[1,2], Mingjian Zeng[1,2]*, Anning Huang[3]*, Zhengkun Qin[4,5], Couhua Liu[6], Wenru Shi[1,2], Xin Li[1,2], Kefeng Zhu[1,2], Chunlei Gu[3], Jialing Zhou[1,2]

1. Nanjing Joint Institute for Atmospheric Sciences, Nanjing, China
2. Key Laboratory of Transportation Meteorology of China Meteorological Administration, Nanjing, China
3. School of Atmospheric Sciences, Nanjing University, Nanjing, China
4. School of Atmospheric Science, Nanjing University of Information Science & Technology, Nanjing, China
5. Joint Center of Data Assimilation Research and Applications, Nanjing University of Information Science & Technology, Nanjing, China
6. National Meteorological Center, Beijing, China

Correspondence: Mingjian Zeng (swordzmj@qq.com) and Anning Huang (anhuang@nju.edu.cn)

Short summary

By directly analysing the proximity of precipitation forecasts and observations, a precipitation forecast accuracy score (PAS) method was constructed. This method does not utilize the traditional contingency table-based classification verification, can replace the threat score (TS), equitable threat score (ETS) and other skill score methods, and can be used to calculate the accuracy of numerical models or quantitative precipitation forecasts.

https://doi.org/10.5194/egusphere-2023-2613
Preprint. Discussion started: 18 December 2023

Abstract

With the development of refined numerical forecasts, the problems such as the score distortion due to the division of precipitation thresholds in both traditional and improved scoring methods for precipitation forecast and the increasing subjective risk arisen from the scale setting of the neighbourhood spatial verification method have become increasingly prominent. To solve this issue, a general comprehensive evaluation method (GCEM) has been developed for cross-scale precipitation forecasts by directly analysing the proximity of precipitation forecasts and observations in this study. In addition to the core element of the precipitation forecast accuracy score (PAS) index, the GCEM system also includes score indices for insufficient precipitation forecasts, excessive precipitation forecasts, precipitation forecast biases and clear/rainy forecasts. The PAS does not distinguish the magnitude of precipitation and delimit the area of influence, it constitutes a fair scoring formula with objective performance and can be suitable for evaluating the rainfall events such as general and extreme precipitation. The PAS can be used to calculate the accuracy of numerical models or quantitative precipitation forecasts, enabling the quantitative evaluation of the comprehensive capability of various refined precipitation forecasting products. Based on the GCEM, comparative experiments between the PAS and TS are conducted for two typical precipitation weather processes. The results show that relative to TS, the PAS aligns with subjective expectations much more, indicating that the PAS is more reasonable than the TS. In addition, other indices of the GCEM are utilized to analyse the range and extent of both insufficient and excessive forecasts of precipitation, as well as the precipitation forecast ability in two weather processes. These indices not only provide overall scores for individual cases similar to the TS but also offer two-dimensional score distribution plots, which can comprehensively reflect the performance and characteristics of

https://doi.org/10.5194/egusphere-2023-2613
Preprint. Discussion started: 18 December 2023

precipitation forecasts. Both theoretical and practical applications demonstrate that the GCEM exhibits distinct advantages and potential promotion and application value compared to the various mainstream precipitation forecast verification methods.

1. Introduction

Precipitation is one of the most important forecasting elements in weather forecasting (Bi et al., 2016; Han al., 2023). Short duration heavy rainfall often lead to flooding and geological disasters, causing widespread and severe impacts (Zhong et al., 2022; Yang et al., 2023). Precipitation forecasts, as focuses and challenges in meteorological department operations, have drawn widespread attention from governments, societies and the public (Bi et al., 2016; Hao et al., 2023). Scientifically evaluating precipitation forecasts helps people gain a clear understanding of the current precipitation forecast levels and maintain appropriate psychological expectations for such forecasts. Moreover, such evaluations assist forecasters in rationally analysing the quality and characteristics of quantitative precipitation forecast systems and aid researchers in understanding the level, strengths and weaknesses of various types of forecasting systems, which in turn, offers valuable insights to improve these systems (Zhong et al., 2022; Zhang et al., 2023; Liu et al., 2022; Gofa et al., 2018). However, there are several shortcomings in current precipitation verification approaches. For instance, traditional scoring methods often fail to reflect model performance improvements, and small errors in the location or timing of convective features can lead to false alarms and missed events. A model's utility is limited by diagnosing model errors such as a displaced forecast feature or an incorrect mode of convective organization (Ahijevych et al., 2009). For high-resolution precipitation forecasts, even if the spatial distribution and intensity of precipitation are consistent with the observations, slight spatial and temporal deviations between forecasts and observations may still result in a large false alarm ratio and missed alarm ratio, leading to lower forecast scores (Zhao et al., 2018). With the rapid development of seamless fine quantitative precipitation forecasts, the need for objective and rational evaluations of the accuracy and characteristics of precipitation forecasts have become increasingly important and urgent (Chen et al., 2021).

Precipitation forecast verification involves various methods, including traditional contingency table-based classification verification and spatial verification methods. The traditional verification method can be traced back to 1884 when Finley introduced a dichotomous contingency table for tornado forecasts and evaluated these forecasts using the proportion correct scoring method (Finley, 1884). Subsequently, systematic attention was given to the evaluation of forecast classification methods, and Finley's forecast verification method became a classic example of the discussion of forecast scoring methods (Murphy, 1996). Shortly thereafter, Gilbert (1984) proposed two scoring methods, namely, the verification and success ratios in forecasting. The verification ratio later became known as the threat score (TS) (Palmer et al., 1949) or the critical success index (Donaldson et al., 1975; Mason, 1989). In forecasting, the success ratio is referred to as the Gilbert skill score (GSS) (Schaefer, 1990) or the equitable threat score (ETS) (Doswell et al., 1990; Gandin et al., 1992). The TS encourages correct event forecasts (hits) and accounts for the impacts on the false alarm and missed alarm ratios, which can better guide forecasters or research and development personnel in making reasonable subjective and objective predictions compared to relying solely on simple "accuracy". Meanwhile, the ETS eliminates the influence of random forecasts on the score, resulting in a fairer skill score (Liu et al., 2022).

For rare-event forecast verification, in addition to the TS and ETS, the methods such as the

https://doi.org/10.5194/egusphere-2023-2613
Preprint. Discussion started: 18 December 2023

Peirce skill score (PSS) (Peirce, 1884; Hanssen et al., 1965; Murphy et al., 1985; Flueck, 1987) and the Heidke skill score (HSS) (Doolittle, 1885; Doolittle, 1888; Heidke, 1926) can be used. The PSS is a fair score index that is equal to the hit rate minus the false detection probability; the HSS eliminates the influence of random forecasts, and the results can reflect the forecast skill (Liu et al., 2022). Many studies have reviewed and compared these two scoring methods (Doswell et al., 1990; Schaefer, 1990; Marzban, 1998; Mason, 2003). In extreme weather event verification (including severe convective weather such as short duration heavy rainfall), the traditional scoring methods (such as the TS and ETS) for dichotomous events often yield scores of zero when the occurrence probability of the object being verified is very low. Therefore, Stephenson proposed the extreme dependency score (EDS) for evaluating extreme events. The EDS has the advantage that different forecast systems converge to different values and has no explicit dependence on the bias of the prediction system (Stephenson et al., 2008; Casati et al., 2008).

It has been more than a century since Gilbert proposed the two scoring concepts, i.e., the verification and success ratios in forecasting (later known as the TS and ETS). The TS and ETS have been widely used for the performance evaluation of threshold-based event forecasts despite their evident shortcomings (Stephenson et al., 2008). Today, in various forecast verification applications, including high-resolution quantitative precipitation and extreme weather forecast verification, the TS and ETS remain mainstream approaches (Tang et al., 2017; Wei et al., 2019; Chen et al., 2021; Liu et al., 2023). Of course, with the continuous introduction of new scoring methods, some of the problems in traditional verification have been solved. However, the advantageous position of the TS remains unchallenged. The reasons for which, although varied, are worthy of attention, but include its objectivity and practicality.

The traditional TS categorizes precipitation according to thresholds and performs verification using a dichotomous contingency table. The TS can be viewed as a measure of the forecast accuracy that excludes the hit forecasts for “non-occurrence” precipitation events (referred to as no precipitation), and its calculation formula being simple, objective and standardized. However, there are two main limitations of the TS. First, precipitation is categorized by thresholds based on the contingency table, which has limitations in terms of classification. The drawback of artificially dividing precipitation into different threshold ranges is that it may not guarantee that two adjacent precipitation values fall within the same threshold range and slightly different precipitation values may not be within the same threshold, which can lead to precipitation score distortion. The second limitation is related to the so-called "double penalty" issue. With the development of high-resolution numerical weather forecasting and the shortening of the spacing between model grid points, some medium- and small-scale phenomena have been portrayed by models. However, it is difficult for high-resolution numerical forecasts to match the characteristics of the observed medium- and small-scale forecasts. In addition, traditional scoring methods often cannot reflect these improvements in terms of the model performance. Assuming a constant forecast area, when there is a small deviation in the timing and location of events between a forecast and an observation, be both "false alarms" and "missed alarms" will occur and is referred to as the "double penalty" phenomenon. This phenomenon leads to a score lower than the subjective expected result, making it difficult to obtain appropriate verification scores when a forecast that “looks good” is not as good as one that “looks bad” (Ahijevych et al., 2009; Wilks, 2006; Ebert, 2008; Chen et al., 2021). For low-probability events with limited sample size for verification, such as torrential rain and short-term heavy rainfall, the “double penalty” issue becomes more prominent. The TS and ETS for torrential

https://doi.org/10.5194/egusphere-2023-2613
Preprint. Discussion started: 18 December 2023

rain tends to align closely the unskillful portion of the scoring values (Chen et al., 2019). In recent years, new mainstream scoring methods have mainly addressed the abovementioned limitations but still have shortcomings. Such methods include the improved gradient decreasing method, which still results in poor scores for good forecasts, and the neighbourhood spatial verification method, which has too many subjective components and may miss medium and small scale information.

To address the limitations of threshold-based precipitation classification, and improve the verification effect, e.g., the gradient decreasing method (hereinafter referred to as the magnitude improved TS) is used to verify the accuracy of rainstorm forecasts (Yang Dong et al., 2017), and appropriate weights are assigned to close forecast values to avoid scores of zero (Table 1). However, the magnitude-improved TS still has limitations. For example, if the observed 24-hour accumulated precipitation is 50 mm, when forecast A is 48 mm while forecast B is 98 mm, it is evident that forecast A is better than forecast B. According to the original TS, forecast B scores 1 point, while forecast A does not score any points. For the magnitude-improved TS, forecast B scores 1 point, as it still falls within the same magnitude category as the observed precipitation, while forecast A only scores 0.4 points, which still fails to reflect the fact that forecast A is superior to forecast B (Table 2). By employing the new scoring method, i.e., the precipitation forecast accuracy score (PAS), which will be discussed later, forecast A scores 0.998 points, while forecast B scores 0.398 points, confirming the rationality and validity of this new method.

To address the “double penalty” issue, a common approach is to employ the neighbourhood spatial verification method (also known as the fuzzy method), which has two specific processing forms. The first form is simple upscaling, which uses a certain method (including value averaging, maximum, value weighting, etc.) to select values within the scale range, adjusting the high-resolution

https://doi.org/10.5194/egusphere-2023-2613
Preprint. Discussion started: 18 December 2023

forecast and observation information to a larger scale to reduce the accidental information of high-resolution data, and then using the traditional skill score. Although this upscaling method provides traditional skill scores at different scales (Yates et al., 2006; Weygandt et al., 2004), it cannot address the issue of excessive smoothness of the precipitation fields during the upscaling process (Zhao et al., 2018), which may result in the omission of some small- to medium-scale information (Zepeda et al., 2000). The other form is the improved neighbourhood spatial verification method proposed by Roberts and Lean (2008). By referring to the Murphy skill score, this method obtains comprehensive evaluation information by comparing the occurrence frequency (probability) of precipitation within different scale windows. If the forecasted occurrence frequency closely approximates the observed occurrence frequency, the forecast is considered value (Zhao et al., 2018). From the perspective of the precipitation occurrence probability within the analysis region, the precipitation occurrence probability for observations and forecasts is the ratio of the precipitation area to the analysed area of the region, which is referred to as the fraction skill score (FSS). This method also effectively improves the “double penalty” problem.

The neighbourhood spatial verification method considers values that are spatially and temporally adjacent between forecasts and observations during the matching process, thus relaxing the strict requirements for spatiotemporal matching (Ebert, 2008; Casati et al., 2008). However, since the determination of the neighbourhood range is a rather subjective process, it hinders the standardization of verification scores and lacks comparability, which may negatively affect objective quantitative verification. Numerous experiments have shown that there is an obvious improvement in the scoring values after adopting the neighbourhood spatial verification method (Chen et al., 2019), particularly for forecasts of large-magnitude precipitation. Nevertheless, the purpose of scoring is not

https://doi.org/10.5194/egusphere-2023-2613
Preprint. Discussion started: 18 December 2023

to achieve a monotonous increase in scoring values but rather to follow the principle of objectivity as much as possible. Errors are errors and cannot be solved by simply lowering the standard. Instead, reasonable and fair criteria should be utilized to reflect the true extent of errors.

Currently, numerical weather forecasts and intelligent gridded forecasts have been developed to output high-resolution precipitation products, while precipitation observations, whether in the form of gridded or station data, are already high-resolution. Staying at the dichotomous classification level for precipitation verification not only wastes existing data resources but also fails to meet the evaluation requirements of refined forecasts. Therefore, to adapt to the development of refined forecasts, a new scoring method is needed. In light of this, a comprehensive verification index for precipitation forecasts is designed, and the following five aspects are considered. (1) The impact of categorical events on rainstorm forecasts should be reduced. In particular, high-resolution forecasts can refer to continuous variables for scoring methods. (2) The design of the scoring method should aim to minimize subjective factors such as the artificial range division and condition settings, ensuring the scoring objectivity and comparability. (3) The designed scoring performance indices should possess ideal attributes such as fairness, base rate independence, suitability for extreme events, and boundedness as much as possible. (4) The devised scoring method should be easy to promote, concise and efficient, with clear concepts and scientific rationality. (5) Different comprehensive verification indices for precipitation forecasts should reflect the forecast performance and characteristics of high-resolution quantitative precipitation from various perspectives.

In this study, on the basis of analysing the limitations of traditional verification methods as well as improved methods, a new general comprehensive evaluation method (GCEM) for cross-scale precipitation prediction is proposed. This method is applied and verified through practical examples.

https://doi.org/10.5194/egusphere-2023-2613
Preprint. Discussion started: 18 December 2023

The remainder of this paper is organized as follows. Section 2 provides an overview of various scoring indices and their attributes in the GCEM and introduces the optimization processing method for the PAS index in the application. Through ideal experiments, the characteristics of the scoring methods are analysed based on the score curves described in Section 3. Section 4 presents the comparative experiments conducted between the new scoring method and the traditional scoring method based on typical cases. Finally, the summary and discussion are given in Section 5.

2 Cross-scale general comprehensive evaluation method

2.1 General comprehensive evaluation method overview

To address the issues of "distorted scores due to the division of precipitation thresholds and increased subjective risks brought about by the setting of the neighbourhood spatial verification method" in traditional and improved precipitation scoring methods, referring to the verification method for heavy rainfall forecasts based on predictability (Chen et al., 2019) and combining the advantages of the relative and absolute errors in this study, a GCEM is constructed by directly analysing the proximity of forecasted precipitation to observed precipitation. It primarily includes the PAS, and the expression of its core scoring function as follows.

$$\mathrm{PAS} = \begin{cases} \sin\left(\frac{\pi}{2}\cdot\frac{x}{u}\right), & 0 \le x < u \\ e^{-\left(\frac{x-u}{u}\right)^2}, & 0 < u \le x \end{cases} \quad ,(1)$$

where PAS represents the scoring value, x is the forecasted precipitation (mm), and u is the observed precipitation (mm). The PAS falls between 0 and 1, where a higher score indicates a better precipitation forecast effect. When PAS = 1, it signifies a perfect forecast, indicating that the forecasted and observed precipitation match entirely. When the observed precipitation is not forecasted, PAS = 0. When the forecasted precipitation amount is sufficiently large, PAS -> 0 (Fig. 1).

https://doi.org/10.5194/egusphere-2023-2613
Preprint. Discussion started: 18 December 2023

221 Additionally, considering the characteristic large fluctuations in the function curve when the observed precipitation is less than 10 mm, a smoothing optimization is applied to Eq. (1) (see Section 2.2 for details).

224 The GCEM system also includes the following indices:

225 (1) Accuracy score of the insufficient precipitation forecast (IPS), whose core scoring function expression is as follows.

227 $$\mathrm{IPS} = \sin\left(\frac{\pi}{2}\cdot\frac{x}{u}\right) - 1,\ 0 \le x < u \quad , \quad (2)$$

228 where IPS represents the scoring value, reflecting the degree of underestimation in precipitation forecasts when the forecasted value is less than the observed value. The IPS falls between [−1, 0), where a scoring value closer to 0 indicates a lower degree of underestimation.

231 (2) Accuracy score of the excessive precipitation forecast (EPS), whose core scoring function expression is as follows.

233 $$\mathrm{EPS} = 1 - e^{-\left(\frac{x-u}{u}\right)^2}, \quad 0 < u < x \quad , (3)$$

234 where EPS represents the scoring value, reflecting the degree of overestimation in precipitation forecasts when the forecasted value exceeds the observed value. The EPS falls between (0, 1], where a scoring value closer to 0 indicates a lower degree of overestimation.

237 (3) Accuracy score of the insufficient and excessive precipitation forecast (IEPS), whose core scoring function expression is as follows.

239 $$\mathrm{IEPS} = \begin{cases} \sin\left(\frac{\pi}{2}\cdot\frac{x}{u}\right) - 1, & 0 \le x < u \\ 0, & x = u \\ 1 - e^{-\left(\frac{x-u}{u}\right)^2}, & 0 < u < x \end{cases} \quad , (4)$$

240 where IEPS represents the scoring value, reflecting the degree of deviation of the forecasted

https://doi.org/10.5194/egusphere-2023-2613
Preprint. Discussion started: 18 December 2023

precipitation from the observed precipitation. The IEPS falls between [−1, 1], where a scoring value closer to 0 indicates a lower degree of deviation. An IEPS less (greater) than 0 indicates an insufficient (excessive) forecast, and an IEPS equal to 0 represents an unbiased forecast.

(4) The PAS clear/rainy forecast accuracy score (PASC), whose scoring function expression is as follows.

$$\mathrm{PASC} = \frac{1}{m+n}\left(\sum_m \mathrm{PAS}_{|ux0.1} + \sum_n \mathrm{PASN}\right) \quad , (5)$$

where PASN denotes the score of the correctly forecasted non-precipitation event, and its formula is given as:

$\mathrm{PASN} = \begin{cases} 1, & 0 \le u < 0.1 \text{ and } 0 \le x < 0.1 \\ 0, & u \ge 0.1 \text{ or } x \ge 0.1 \end{cases}$. PASC represents the PAS scoring value for clear/rainy forecasts, and $\mathrm{PAS}_{|ux0.1}$ denotes the overall PAS for precipitation forecasts under specific conditions where the observed precipitation $u \ge 0.1$ mm or the forecasted precipitation $x \ge 0.1$ mm. m is the number of stations or grid points for $\mathrm{PAS}_{|ux0.1}$, and n is the number of stations or grid points where non-precipitation events are correctly forecasted.

The discussion below pertains to the characteristics of the PAS scoring method. As an ideal performance indicator, the PAS has the attributes of boundedness, fairness, sensitivity disparity, suitability for extreme events and moderate symmetry.

(1) Boundedness. The PAS scoring values range between 0 and 1. A PAS score of 1 represents an ideal forecast, while a score of 0 indicates that there is observed precipitation but no forecasted precipitation or that the forecasted precipitation is sufficiently large. The scoring range is consistent with that of traditional TS, making it easy to compare and evaluate the scoring methods and suitable for practical forecast verification applications.

(2) Fairness. The PAS scoring method constitutes a scoring formula in an objective form

https://doi.org/10.5194/egusphere-2023-2613
Preprint. Discussion started: 18 December 2023

without a subjective boundary definition. Precipitation forecasts are verified without magnitude or delimitation of the area of influence, and the closer to the observed situation the forecast is, the higher the score, which is fair.

(3) Sensitivity disparity. From the Chinese national standard GB/T 28592—2012 "Grade of precipitation" on the classification of precipitation grades, the public is more sensitive to low-grade precipitation forecasts. As rainfall intensity increases, the public's sensitivity gradually decreases; that is, the public has a higher tolerance for errors in response to heavier rainfall forecasts. In other words, large errors in the forecasts of heavy rainfall events may be considered equivalent to smaller errors in weaker rainfall events in terms of forecast scoring. As shown in Fig. 1, the intersection point on the PAS scoring curves for the observed precipitation amounts of 25 mm and 100 mm corresponds to a forecasted amount of 42.4 mm. That is, the forecast errors are 17.4 mm and 57.6 mm for the observed 24-hour accumulated precipitation of 25 mm and 100 mm, respectively, while the scores are both 0.62. From the perspective of forecast service effectiveness, this aligns with the general public perception.

(4) Suitability for extreme events. From the PAS scoring curves for forecasts corresponding to different observed precipitation amounts (u = 10, 25, 50 and 100 mm) (Fig. 1), it is evident that the PAS scoring method performs well in evaluating precipitation events forecasts at the level of torrential rain and above. For example, when the observed precipitation is 100 mm, with forecasted amounts of 59 mm and 147.2 mm, the PASs are both 0.8, whereas the TSs are 0 and 1, and the improved TSs are 0.8 and 1, respectively. This result indicates that the PAS is suitable for scoring heavy rainfall events, meeting the general applicability requirements as a scoring method that does not degrade due to extreme events.

https://doi.org/10.5194/egusphere-2023-2613
Preprint. Discussion started: 18 December 2023

(5) Moderate symmetry. In Eq. (1), observed precipitation is the independent variable x, and forecasted precipitation is the parameter u. Then, the equation is rewritten as:

$$\mathrm{PAS} = \begin{cases} e^{-\left(\frac{u-x}{x}\right)^2}, & 0 < x \le u \\ \sin\left(\frac{\pi}{2}\cdot\frac{u}{x}\right), & u \le x \end{cases} \quad , (6)$$

Similarly, for different magnitudes of forecasted precipitation (u = 10, 25, 50 and 100 mm) and observed precipitation ranging from 0 to 300 mm, the corresponding forecast scores are shown in Fig. 2. The forecast scores also vary with the degree of proximity between forecasts and observations. Figs. 1 and 2 exhibit similar trends but are not identical, illustrating that the PAS possesses moderate symmetry.

2.2 PAS verification for precipitation forecasts

From the properties of the core verification function of the PAS, it is noted that when the observed precipitation u < 10 mm, there is a large gradient in the PAS curve. A slight change in the forecasted value (x) can result in a large fluctuation in the PAS. To account for this characteristic, based on a comprehensive analysis in combination with the sensitivity of forecasters and the public to small-scale precipitation, a smoothing optimization scheme is applied to the PAS curve for accumulated precipitation below 10 mm. Similarly, the IPS, EPS, IEPS and PASC curves are appropriately smoothed and optimized according to their respective definitions.

Let forecasted precipitation x = 0 mm, 1 > PAS > 0 for observed precipitation, and 0 < u < 10 mm.

Assumption:

(1) $\mathrm{PAS} = 0.6\mathrm{PAS}_{|u\to 0}$ when u = 0 mm, and $x \neq 0$ mm;

$\mathrm{PAS}_{|u\to 0}$ denotes the PAS for the case of observed precipitation 0 < u ≤ 0.1 mm;

https://doi.org/10.5194/egusphere-2023-2613
Preprint. Discussion started: 18 December 2023

(2) $\mathrm{PAS} = 0.6\mathrm{PAS}_{|u\to 0}$ when x = 0 mm, and u < 10 mm;

$\mathrm{PAS}_{|x\to 0}$ denotes the PAS for the case of forecasted precipitation 0 < x ≤ 0.1 mm.

1. When the observed precipitation u = 0 mm (Fig. 3a), let $\mathrm{PAS} = 0.6\mathrm{PAS}_{|u\to 0}$. Then,

$$\mathrm{PAS} = 0.6e^{-\left(\frac{x}{10}\right)^2} \qquad x > 0 \tag{7}$$

2. When the forecasted precipitation x = 0 mm and the observed precipitation u < 10 mm (Fig. 3b), let $\mathrm{PAS} = 0.6\mathrm{PAS}_{|x\to 0}$. Then,

$$\mathrm{PAS} = 0.6\sin\left(\frac{\pi}{2}\cdot\frac{10-u}{10}\right),\ 0 < u < 10 \tag{8}$$

3. When the observed precipitation 0 < u < 10 mm and the forecasted precipitation x ≠ 0 (Fig. 3c). Then,

$$\mathrm{PAS} = \begin{cases} \sin\left(\frac{\pi}{2}\cdot\frac{x-u+10}{10}\right), & 0 \le x < u,\ 0 \le u < 10 \\ e^{-\left(\frac{x-u}{10}\right)^2}, & u \le x, \quad 0 \le u < 10 \end{cases} \tag{9}$$

Note that x and u are not both equal to 0 at the same time.

4. When the observed precipitation u ≥ 10 mm (Fig. 3d), then

$$\mathrm{PAS} = \begin{cases} \sin\left(\frac{\pi}{2}\cdot\frac{x}{u}\right), & 0 \le x < u,\ u \ge 10 \\ e^{-\left(\frac{x-u}{u}\right)^2}, & u \le x, \quad u \ge 10 \end{cases} \tag{10}$$

To compare with the traditional scoring method, the new scoring method for precipitation forecasting adopts the "classification before verification, no classification during verification" approach. Scoring for precipitation processes over different accumulated periods is based on but not limited to the commonly used precipitation classification approaches in practical operations, as shown in Tables 3–5.

3 Ideal experimental validation of the new verification method

https://doi.org/10.5194/egusphere-2023-2613
Preprint. Discussion started: 18 December 2023

3.1 Validation of forecast scoring results for general precipitation

General precipitation refers to precipitation ranging from light rain to heavy rain, i.e., 24-hour accumulated precipitation between [0.1 mm, 50 mm). Figure 4 shows the schematic diagram of PAS scores for general precipitation. The forecasted amounts are compared under conditions when the 24-hour accumulated precipitation is 10 mm, 25 mm and 45 mm and the PAS scores are 0.8, 0.7, 0.5 and 0.3 (Table 6). When the observed precipitation is 10 mm, the forecasted amounts of 5.9 mm and 14.7 mm both have a PAS score of 0.8, with differences from the perfect forecast value (10 mm) of 4.1 mm and 4.7 mm, respectively; the forecasted amounts with a PAS score of 0.3 are 1.9 and 21.0 mm, differing by 8.1 mm and 11.0 mm from the perfect forecast value (10 mm), respectively. When the observed precipitation is 25 mm, the forecasted amounts with a PAS score of 0.8 are 14.7 mm and 36.8 mm, with differences from the perfect forecast value (25 mm) of 10.3 mm and 11.8 mm, respectively; the forecasts with a PAS score of 0.5 are 8.3 mm and 45.8 mm, differing by 16.7 mm and 20.8 mm from the perfect forecast value (25 mm), respectively.

For forecasts with the same observed precipitation and the same scores, the absolute errors of an insufficient forecast and observation are smaller than those of an excessive forecast and observation, and the higher the scores are, the closer the absolute errors of the forecasts. When the observed precipitation is 50 mm, only the insufficient precipitation forecast is scored since a precipitation forecast exceeding 50 mm is not considered within the scope of general precipitation evaluation. The scoring experimental results align with expectations.

3.2 Validation of forecast scoring results for precipitation at the level of torrential rain and above

Figure 5 shows a schematic diagram of the PASs when the amount of precipitation exceeds the

https://doi.org/10.5194/egusphere-2023-2613
Preprint. Discussion started: 18 December 2023

storm magnitude. The predicted precipitation is compared when the 24-hour cumulative observed precipitation is 25 mm, 50 mm, and 100 mm with PAS scores of 0.877, 0.7, 0.5, 0.3, and 0.1 (Table 7). When the observed precipitation is 25 mm, forecasts ≥ 50 mm are only involved in the rating, with PASs of 0.3 and 0.1 for forecasts of 52.4 and 62.9 mm, respectively.

When the PAS is 0.877 and the observed precipitation is 50 mm, the predicted values are 34.1 and 68.1 mm, respectively; when the observed precipitation is 100 mm, the predicted values are 68.1 and 136.2 mm, respectively. When the observed precipitation is 50 or 100 mm, the prediction is 68.1 mm, with a score of 0.877. The absolute error is 18.1 mm for the excessive precipitation forecast and 31.9 mm for the insufficient precipitation forecast. This result indicates that the scoring tolerance increases as the grade of observed precipitation increases and is gradually expanded through continuous changes, avoiding discontinuous increases caused by magnitude changes.

When the observed precipitation is 50 mm and the PAS is 0.3, the insufficient forecast is 9.7 mm and the excessive forecast is 104.9 mm. When the observed precipitation is 100 mm, the predictions for a PAS of 0.3 are 19.4 and 209.7 mm, respectively. When the observed precipitation is 50 mm, the insufficient forecast with a PAS of 0.1 is 3.2 mm, and the excessive forecast is 125.9 mm. When the observed precipitation is 100 mm, the predictions with a PAS of 0.1 are 6.1 and 251.7 mm, respectively.

Under constant observed precipitation conditions, for forecasts with the same score, the absolute error between the insufficient forecast and the observed precipitation is smaller than that between the excessive forecast and the observed precipitation. The higher the score is, the smaller the absolute error between the forecast and the observation. Moreover, the scoring tolerance increases with increasing observed precipitation. The scoring experimental results conform to expectations.

https://doi.org/10.5194/egusphere-2023-2613
Preprint. Discussion started: 18 December 2023

4 Example-based comparative experiments for the new verification method

4.1 Introduction of typical cases

Comparative experiments of traditional TS and PAS are conducted for 12-hour accumulated precipitation of two typical cases. One case pertains to the precipitation weather process occurring at 12:00 UTC on July 16, 2019 (referred to as “Case 1”), which is dominated by a weak weather system. The other case relates to the precipitation weather process occurring at 12:00 UTC on June 13, 2020 (referred to as “Case 2”), which is predominantly associated with a strong weather system.

Both of these precipitation cases are associated with precipitation during the Meiyu period. Case 1, which occurred during the Meiyu period of 2019 and was featured by scattered precipitation under weak synoptic-scale forcing. The low intensity shear line system is located south of the Yangtze River. There are two precipitation concentration areas, one at the intersection of Hunan Province and Jiangxi Province and the other covering the majority of Zhejiang Province. The precipitation process in Case 2 (July 11–13) represents the first round of widespread rainstorms during the Meiyu period of 2020, including heavy precipitation affected by a low-level vortex shear system. The western section of the low-level vortex shear is relatively stable, while the eastern section slightly presses southwards. Southwesterly airflow developed and pushed northwards, and a strong wind speed belt persisted for a long time in the Jianghuai region. Moreover, the Jianghan–Jianghuai region maintained a high-energy and high-moisture state, resulting in persistent heavy rainfall.

A subjective analysis of these two weather processes reveals that for the event on July 16, 2019 (Fig. 6), the forecasted precipitation intensity and rainfall areas are relatively consistent with the observations. There are two distinct heavy rainfall areas in the east and south parts of the Yangtze River, with particularly high accuracy in forecasting scattered rainstorms in Zhejiang Province

https://doi.org/10.5194/egusphere-2023-2613
Preprint. Discussion started: 18 December 2023

located in the eastern section. In contrast, for the precipitation weather process on June 13, 2020 (Fig. 7), it is evident that there is an overestimation of the precipitation forecast.

4.2 Data and methods

The observed precipitation data are provided by the China Meteorological Administration multisource merged precipitation analysis system (CMPAS), developed by the National Meteorological Information Center of China. The CMPAS integrates hourly precipitation data from nearly 40,000 automatic meteorological stations in China and provides radar-based quantitative precipitation estimation and satellite-retrieved precipitation products with a spatial resolution of 0.05° × 0.05°. The predicted precipitation data with 3 km resolution are from the Precision Weather Analysis and Forecasting System (PWAFS) model, a regional refined forecast model, developed by the Jiangsu Provincial Meteorological Bureau. These data are output once per hour.

The specific methods are as follows.

(1) Determine the verification domain and verification points. The verification domain covers the Huang–Huai region of China (28°N–38°N, 111°E–123°E). The verification points are defined based on the grid points of the observed precipitation data, their spatial resolution is 0.05° × 0.05°, and the total number of verification grid points is 48,000 (200 × 240).

(2) Prepare the observed and forecasted precipitation data and interpolate the forecasted precipitation data onto the observed grid points. The observed 12-hour accumulated precipitation data are derived by accumulating the hourly precipitation data from the CMPAS. The forecasted 12-hour accumulated precipitation data are obtained by subtracting the zero-field data from the 12-hour forecast field data. Since the grid points of the observed and forecasted precipitation data do not coincide and the grid spacing is small, the nearest neighbour method is used in this study to

https://doi.org/10.5194/egusphere-2023-2613
Preprint. Discussion started: 18 December 2023

match the forecasted data to the grid points of the observed precipitation. Specifically, the forecasted data on the grid point nearest to the observed grid point are used as the forecasted value for this observed grid point.

(3) Analyse the relationship between forecasted precipitation and observed precipitation. The scores for each verification grid point and the overall scores for each verification area are calculated based on the scoring formula for each index in the GCEM system. Then, the verification result file is generated in NetCDF format. On this basis, the distribution maps for the scores of various indices in the GCEM system are produced. Additionally, the total TS and clear/rainy TS for different precipitation magnitudes within the verification area are calculated based on the TS and clear/rainy TS formulas.

4.3 Analysis of the comparative experiment results

For the precipitation process on July 16, 2019, traditional TSs for different rainfall categories, such as clear/rainy and 12-hour accumulated precipitation ≥ 0.1 mm, ≥ 10 mm, ≥ 25 mm and ≥ 50 mm, are all lower than the traditional scores for the weather process on June 13, 2020. For example, the TS is 0.381 for 12-hour accumulated precipitation ≥ 0.1 mm at 12:00 UTC on July 16, 2019 (Table 8), while this score is 0.625 for that at 12:00 UTC on June 13, 2020 (Table 9), which differs from the subjective judgement.

For the precipitation process at 12:00 UTC on July 16, 2019, the PASs for clear/rainy and 12-hour accumulated precipitation ≥ 0.1 mm, ≥ 10 mm and ≥ 25 mm are all higher than those for the precipitation process on June 13, 2020. For instance, the overall PAS is 0.617 for 12-hour accumulated precipitation ≥ 0.1 mm at 12:00 UTC on July 16, 2019. This PAS is higher than the PAS of 0.457 for the precipitation process at 12:00 UTC on June 13, 2020, which aligns with the

https://doi.org/10.5194/egusphere-2023-2613
Preprint. Discussion started: 18 December 2023

subjective judgement.

For the precipitation process at 12:00 UTC on July 16, 2019, the PAS for each magnitude is higher than the corresponding TS, addressing the issue of TSs being lower. For the precipitation process at 12:00 UTC on June 13, 2020, the PASs for clear/rainy and the magnitudes of ≥ 0.1 mm and ≥ 10 mm are lower than the corresponding TSs, whereas the PASs for the magnitudes of ≥ 25 mm and ≥ 50 mm are higher than the corresponding TSs. This result indicates that the PAS is different from the magnitude-improved TS and the neighbourhood spatial verification method, both of which blindly increase the tolerance, leading to a monotonous increase in scores. This result also demonstrates that the PAS has good discrimination ability for extreme events.The PAS assigns objective scores based on the proximity of the forecast to the observation, making it more reliable for precipitation evaluation than the TS.

4.4 Analysis of the indices in the new verification method

The GCEM includes not only the core element of the PAS index but also the IPS, EPS, IEPS and PASC indices.

Regarding the issue of analysing the sources of errors from the verification results, objectively tracing these errors back from a single score can only determine whether an error was “insufficient (missed alarm)” or “excessive (false alarm)”. However, the advantage of the GCEM lies in its ability to decompose the score to each verification point and examine the forecast performance at each point, which is different from the dichotomous evaluation with only 0 and 1. These indices not only provide overall scores for individual cases similar to the TS but also offer two-dimensional score distribution plots, which can comprehensively reflect the performance and characteristics of precipitation forecasts.

https://doi.org/10.5194/egusphere-2023-2613
Preprint. Discussion started: 18 December 2023

Figure 8 shows the distribution for the PASC scores of 12-hour accumulated precipitation. In these two cases, due to the high accuracy of non-precipitation forecasts, the overall PASC scores are relatively high. However, for Case 1, the scores in Zhejiang are lower and scattered within a small area. In contrast, for Case 2, there is a large area occupying most of the Jianghuai region with low scores. Therefore, the PASC score of Case 1 (0.808) is higher than that of Case 2 (0.734).

Figure 9 shows the PAS distributions of 12-hour accumulated precipitation with magnitudes of ≥ 0.1 mm, ≥ 10 mm and ≥ 25 mm. The blank points in the figure are the points that are excluded in the scoring, following the scoring principle of “classification before verification, no classification during verification” described in Section 2. From the PAS distributions of different magnitudes, for Case 1, the scoring areas in Zhejiang exhibit alternatively distributed high and low scores. In contrast, for Case 2, the scoring areas in the Jianghuai region have a larger area of low scores than high scores. Therefore, Case 1 has higher PASs for the three categories (≥ 0.1 mm, ≥ 10 mm and ≥ 25 mm) than Case 2, and the distributions also allow for distinguishing the areas with better and poorer forecasting performance.

Figure 10 shows the IPS, EPS and IEPS distributions of 12-hour accumulated precipitation. In terms of the IPS, for Case 1, the large-value IPS areas are located at the intersection of Anhui, Zhejiang and Jiangxi in the Hunan–Jiangxi region, as well as in the southern part of Hebei. For Case 2, the large-value IPS areas are situated along the Yangtze River in Anhui and Jiangxi, as well as at the intersection of Henan and Shanxi. The IPSs for Case 1 and Case 2 are −0.376 and −0.400, respectively, indicating that Case 2 shows a slightly higher level of insufficient forecasts (Table 10). In terms of the EPS, for Case 1, the large-value EPS areas are in Zhejiang and Jiangxi. In contrast, for Case 2, the large-value EPS areas are located in most of Hunan, Hubei, Anhui and Jiangsu,

https://doi.org/10.5194/egusphere-2023-2613
Preprint. Discussion started: 18 December 2023

exhibiting a wide southwest–northeast orientation with a large area and depth. The EPS for Case 2 is larger than that for Case 1. The IEPS score is a comprehensive reflection of under- and over-precipitation, and its value reflects the degree of insufficient and excessive precipitation forecasts. From the distributions of insufficient and excessive precipitation forecasts in Case 1, it is evident that the insufficient and excessive forecasts are roughly equivalent, each with an IEPS of 0.057. However, for Case 2, the distribution of the excessive forecast is obviously larger than that of the insufficient forecast, with an IEPS of 0.325. This result indicates that Case 2 has poorer forecasting performance, with larger excessive forecasts being an important factor.

Consequently, analysing the locations of insufficient and excessive precipitation forecasts from the figures in conjunction with the characteristics of the forecasting process can provide useful insights for improving forecasts.

5 Discussion and conclusion

By analysing the advantages and disadvantages of the traditional TS, magnitude-improved TS and neighbourhood spatial verification method, a new precipitation verification method GCEM was designed and constructed from the perspective of the proximity of the forecast to the observation. This method consists of the core element of the PAS, as well as multiple elements such as IPS, EPS, IEPS and PASC.

The PAS index consists of sine and e-exponential functions. Additionally, considering the characteristics of large fluctuations in the function curves when observed precipitation is less than 10 mm, the formula has been smoothed for optimization. The PAS method adopts the principle of “classification before verification, no classification during verification”, which can serve as an alternative to skill scores such as the TS and ETS for verifying quantitative precipitation forecasts.

https://doi.org/10.5194/egusphere-2023-2613
Preprint. Discussion started: 18 December 2023

This method is characterized by objective and transparent rules and easy generalization. Moreover, this possesses attributes of an ideal precipitation scoring method, such as fairness, boundedness and moderate symmetry. Therefore, it can be used to calculate the accuracy of numerical models or quantitative precipitation forecasts, as well as evaluate the comprehensive forecasting capabilities of various refined quantitative precipitation forecast products. The GCEM can also evaluate the performance of numerical forecasts on clear/rain forecasts, as well as insufficient precipitation forecasts, excessive precipitation forecasts and precipitation forecast biases. In addition to the overall score, two-dimensional score distribution maps can be generated for each index in the GCEM system. These maps offer a comprehensive reflection of the precipitation forecasting performance of the numerical models and serve as a reference for improving model forecasts.

This new verification method is validated based on the forecast scoring results for general precipitation and precipitation at the level of torrential rain and above, and the verification results align with expectations. Comparative experiments are also conducted on two cases using the new verification method. For Case 1, the subjective judgement is relatively good, but the TS is lower. Conversely, for Case 2, the subjective judgement is poorer, yet the TS is higher. Verification using the PAS reveals that forecasts with better subjective judgement receive higher scores, and forecasts with poorer subjective judgement receive lower scores. Therefore, the PAS aligns with public expectations.

In addition, the National Meteorological Center of China conducted long-term series large-scale sample testing on this method in 2023. Based on the ECMWF model's 24-hour and 48-hour precipitation forecasts from March 2022 to February 2023, the assessment results show that compared with the TS, the PAS is less affected by the randomness of the sample, and the relative size

https://doi.org/10.5194/egusphere-2023-2613
Preprint. Discussion started: 18 December 2023

relationship of different time forecast scores is more stable.

From the construction of the GCEM to ideal experiments and case analyses, it is evident that this evaluation system, especially the PAS method, is a suitable method for quantitative precipitation evaluation. However, the PAS still has subjective flaws, such as the determination of coefficients in the PAS expression [0.6 in Eqs. (7) and (8)] when the observed or forecasted precipitation is 0 mm. Once these coefficients are determined, they apply to all precipitation scoring, thus becoming an objective component in practice.

Code and data availability. The data are provided at https://www.doi.org/10.5281/zenodo.10251028. The source code of this work can be found at https://www.doi.org/10.5281/zenodo.10251028.

Author contributions. BZ designed the evaluation method, completed the experiments, and wrote the paper. MZ provided advice on the planning and application of the evaluation method. AH provided suggestions for the evaluation method and contributed to paper revisions, ZQ contributed to paper revisions, and CL provided long-term series large-scale sample comparison test results for the evaluation method. All authors discussed the results and commented on the paper.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors.

Acknowledgements. This study was supported by grid precipitation analysis data, forecasted precipitation data and numerical computing capabilities provided by the Jiangsu Provincial

Meteorological Bureau of China, as well as long-term series large-scale sample testing conducted at the National Meteorological Center of China.

Financial support. Financial support. This research has been supported by the National Key Research and Development Program of China (Grant no. 2021YFC3000904).

References

Ahijevych D, Gilleland E, Barbara G B, et al.: Application of spatial verification methods to idealized and NWP-gridded precipitation forecasts, Wea Forecasting, 24:1485-1497, 2009.

Bi B G, Dai K, Wang Y, et al.: Advances in techniques of quantitative precipitation forecast, J Appl Meteor Sci,27(5):534-549 (in Chinese) doi: 10.11898/1001-7313.20160503, 2016.

Casati D B, Wilson L J, Stephenson D B, et al.: Forecast verification: current status and future directions, Meteorological Applications, 15(1):3-18, 2008.

Chen F., Chen J., Wei Q., Li J., Liu C., Yang D., Zhao B., Zhang Z.: A new verification method for heavy rainfall forecast based on predictability II: Verification method and test[J], Acta Meteorologica Sinica, 77(1): 28-42. doi: 10.11676/qxxb2019.003, 2019.

Chen H M, Li P X, and Zhao Y.: A review and outlook of verification and evaluation of precipitation forecast at convection-permitting resolution, Adv Meteor Sci Technol,11(3):155-164 (in Chinese) doi: 10.3969/j.issn.2095-1973.2021.03.018, 2021.

Donaldson R J, Dyer R M, Krauss M J.: An objective evaluator of techniques for predicting severe weather events, 9th Conf Severe Local Storms, Norman, Oklahoma, Amer Meteor Soc, 321-326, 1975.

Doswell CA III, Davies-Jones R, and Keller DL.: On summary measures of skill in rare event forecasting based on contingency tables, Weather and Forecasting 5: 576–585, 1990.

Doolittle M H.: Association ratios, Bull Philos Soc Washington, 10:83-87;94-96, 1888.

Doolittle M H.: The verification of predictions, Amer Meteor J, 2:327-329, 1885.

Ebert E E.: Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework[J], Meteor Appl,15(1):51-64, 2008.

Finley J P. :Tornado predictions, Amer Meteor J., 1: 85-88, 1884.

https://doi.org/10.5194/egusphere-2023-2613
Preprint. Discussion started: 18 December 2023

Flueck J A.: A study of some measures of forecast verification, 10th Conf Probability and Statistics in Atmospheric Science, Edmonton, Alberta, Amer Meteor Soc, 69-73, 1987.

Gandin L S, Murphy A H.: Equitable scores for categorical forecasts, Monthly Weather Review 120: 361–370, 1992.

Gilbert G K.: Finley's tornado predictions, Amer Meteor J, 1: 166-172, 1884.

Gofa F,Boucouvala D, Louka P, et al.: Spatial verification approaches as a tool to evaluate the performance of high resolution precipitation forecasts[J]. Atmos Res, 208:78-87, 2018.

Han F., Tang W., Zhou C., Sheng J.,and Zhang X.: Improving a precipitation nowcasting algorithm based on the SWAN system and related application assessment. Acta Meteorologica Sinica， 81(2):304-315 doi: 10.11676/qxxb2023.20220066, 2023.

Hanssen A W, Kuipers W J A.: On the relationship between the frequency of rain and various meteorological parameters, Mededeelingen en Verhandelingen, 81:2-15, 1965.

Hao C., Yu B., Dai Y., Zhi X., And Zhang Y.: Statistical downscaling research on spatio-temporal distributions of summer precipitation across the beijing region[J], Meteor Mon,49(7):843-854,2023.

Heidke P.: Berechnung des Erfolges und der Güte der Windstärkevorhersagen im Sturmwarnungdienst (Calculation of the success and goodness of strong wind forecasts in the storm warning service), Geogr Ann Stockholm, 8: 301-349, 1926.

Liu C., Lin J., Dai K., Cao Y., Wei Q.: An evaluation method suitable for precipitation forecasts and services[J], Torrential Rain and Disasters, 41(6): 712-719. doi: 10.12406/byzh.2021-203, 2022.

Liu C, Dai K, Lin J, et al.: Design and implementation of whole process evaluation program library of weather forecast[J], Meteor Mon,49(3):351-364(inChinese), 2023.

Liu J., Ren C., Zhao Z., Chen C., Wang Y., And Cai K.: Comparative analysis on verification of heavy rainfall forecasts in different regional models[J].Meteor Mon,48(10):1292-1302, 2022.

Marzban C.: Scalar measures of performance in rare-event situations. Weather and Forecasting 13: 753–763, 1998.

Mason IB.: Binary events. in forecast verification: a practitioner's guide in atmospheric science, Jolliffe IT, Stephenson DB (eds). John Wiley and Sons: Chichester, UK; 37–76, 2003.

Mason IB.: Dependence of the critical success index on sample climate and threshold probability, Australian Meteorological Magazine 37: 75–81, 1989.

Murphy A H.: The finley affair: a signal event in the history of forecast verification, Wea Forecasting, 11:

https://doi.org/10.5194/egusphere-2023-2613
Preprint. Discussion started: 18 December 2023

3-20,1996.

Murphy A H, Daan H.: Forecast evaluation // Murphy A H, Katz R W. Probability, statistics, and decision making in the atmospheric sciences. Westview : Westview Press, 379-437, 1985.

Palmer W C, Allen R A.: Note on the accuracy of forecasts concerning the rain problem, US Weather Bureau,1-4,1949.

Peirce C S.:The numerical measure of the success of prediction, Science, 4:453-454, 1884.

Roberts N M, Lean H W.: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events[J], Mon Wea Rev, 136:78-97, 2008.

Schaefer JT.: The critical success index as an indicator of warning skill, Weather and Forecasting 5: 570–575, 1990.

Stephenson D B, Casati B, Ferro C A T, et al.: The extreme dependency score: A non-vanishing measure for forecasts of rare events. Meteor Appl, 15(1): 41-50, 2008.

Tang W., Zhou Q., Liu X., et al.: Analysis on verification of national severe convective weather categorical forecasts, Meteorological Monthly, 43(1): 67-76, DOI: 10.7519/j.issn.1000-0526.2017.01.007, 2017.

Wei Q., Li W., Peng S., et al.: Development and application of national verification system in CMA, J Appl Meteor Sci, 30(2): 245-256. DOI: 10.11898/1001-7313.20190211, 2019.

Weygandt S S, Loughe A F, Benjamin S G ,et al.: Scale sensitivities in model precipitation skill scores during IHOP[C]//22nd Conf Severe Local Storms, Amer Met Soc, Hyannis, MA, 4-8 October 2004.

Wilks, D. S.: Statistical methods in the atmospheric sciences. 2nd ed. Elsevier, 627 pp, 2006.

Yang Y., Yin J., Wang D., Liu Y., Lu Y., Zhang W., and Xu S.: ABM-based emergency evacuation modelling during urban pluvial floods: A “7.20” pluvial flood event study in Zhengzhou, Henan Province. Science China Earth Sciences, 66(2): 282–291, https://doi.org/10.1007/s11430-022-1015-6, 2023.

Yang D., Gao X., and Zhang W.: Research and improvement of a new rainstorm forecast accuracy verification scheme [C]//Proceedings of the 34th Annual Meeting of the Chinese Meteorological Society S1 Disaster Weather Monitoring, Analysis and Forecast, 549-559,2017.

Yates E., S Anquetin, V Ducrocq, et al.: Point and areal validation of forecast precipitation fields[J]. Meteorol Appl, 13: 1-20, 2006.

Zepeda-Arce J, E Foufoula-Georgiou, Droegemeier K K.: Space-time rainfall organization and its role in validating quantitative precipitation forecasts[J], J Geophys Res, 105(D8), 10 129-10 146,2000.

Zhang, B., Zeng, M., Huang, A., Qin, Z., Liu, C., Shi, W., Li, X., Zhu, K., Gu, C., and Zhou, J.: A general

https://doi.org/10.5194/egusphere-2023-2613
Preprint. Discussion started: 18 December 2023

comprehensive evaluation method for cross-scale precipitation forecasts, Zenodo [code and data set], https://www.doi.org/10.5281/zenodo.10251028, 2023.

Zhang Y., Yu H., Zhang M., Yang Y., and Meng Z.: Uncertainties and error growth in forecasting the record-breaking rainfall in Zhengzhou, Henan on 19–20 July 2021. Science China Earth Sciences, 65(10): 1903–1920, https://doi.org/10.1007/s11430-022-9991-4, 2022.

Zhao B., Zhang B.: Application of neighborhood spatial verification method on precipitation evaluation[J], Torrential Rain and Disasters,37(1):1-7, 2018.

Zhong Q, Sun Z, Chen H, Li J, and Shen L.: Multi model forecast biases of the diurnal variations of intense rainfall in the Beijing-Tianjin-Hebei region. Science China Earth Sciences, 65(8):1490−1509, https://doi.org/10.1007/s11430-021-9905-4, 2022.

Table 1. Gradient decrease scoring table for station-by-station (time) rainstorm forecasts. The values are normalized, i.e., score = original data/100.

Observation (mm)	Forecast (mm)			
	25-49.9	50.0-99.9	100.0-249.9	≥250
<25.0	--	0	0	0
25.0-39.9	--	0.4	0	0
40.0-49.9	--	0.7	0.4	0
50.0-99.9	0.4	1	0.8	0.4
100.0-249.9	0	0.8	1	0.9
≥250.0	0	0.4	0.8	1

https://doi.org/10.5194/egusphere-2023-2613
Preprint. Discussion started: 18 December 2023

Table 2. Examples of station-specific rainstorm precipitation scoring

	Observation	Forecast A	Forecast B	Correct, Reasonable or False
Precipitation	50 mm	48 mm	98 mm	--
Forecast effect	--	Good	Bad	Correct
Classic TS	--	0	1	False
Improved TS	--	0.4	1	False
PAS	--	0.998	0.398	Reasonable

Table 3. Classification of PAS for short-term heavy rainfall.

Scoring name	Notes on the scoring application
$PAS_{\|ux10}$	PAS score for 1 h observed precipitation u ≥ 10 mm or forecasted precipitation x ≥ 10 mm
$PAS_{\|ux20}$	PAS score for 1 h observed precipitation u ≥ 20 mm or forecasted precipitation x ≥ 20 mm

https://doi.org/10.5194/egusphere-2023-2613
Preprint. Discussion started: 18 December 2023

Table 4. Classification of PAS for 12 h accumulated precipitation.

Scoring name	Notes on the scoring application
PASC	12 h PAS clear and precipitation forecast accuracy score
$PAS_{\|ux0.1}$	12 h PAS overall precipitation prediction verification score i.e., PAS score for observed precipitation u ≥ 0.1 mm or forecasted precipitation x ≥ 0.1 mm
$PAS_{\|ux10}$	PAS score for 12 h observed precipitation u ≥ 10 mm or forecasted precipitation x ≥ 10 mm
$PAS_{\|ux25}$	PAS score for 12 h observed precipitation u ≥ 25 mm or forecasted precipitation x ≥ 25 mm
$PAS_{\|ux50}$	PAS score for 12 h observed precipitation u ≥ 50 mm or forecasted precipitation x ≥ 50 mm
$PAS_{\|ux100}$	PAS score for 12 h observed precipitation u ≥ 100 mm or forecasted precipitation x ≥ 100 mm

https://doi.org/10.5194/egusphere-2023-2613
Preprint. Discussion started: 18 December 2023

Table 5. Classification of PAS for 24 h accumulated precipitation.

Scoring name	Notes on the scoring application
PASC	24 h PAS clear and precipitation forecast accuracy score
$PAS_{\|ux0.1}$	24 h PAS overall precipitation prediction verification score i.e., PAS score for observed precipitation u ≥ 0.1 mm or forecasted precipitation x ≥ 0.1 mm
$PAS_{\|ux10}$	PAS score for 24 h observed precipitation u ≥ 10 mm or forecasted precipitation x ≥ 10 mm
$PAS_{\|ux25}$	PAS score for 24 h observed precipitation u ≥ 25 mm or forecasted precipitation x ≥ 25 mm
$PAS_{\|ux50}$	PAS score for 24 h observed precipitation u ≥ 50 mm or forecasted precipitation x ≥ 50 mm
$PAS_{\|ux100}$	PAS score for 24 h observed precipitation u ≥ 100 mm or forecasted precipitation x ≥ 100 mm

Table 6. Examples of forecast verification scores for general precipitation (u = 25, 50 and 100 mm).

PAS value	Observation u=10 mm		Observation u=25 mm		Observation u=45 mm	Observation u=50 mm (No comparison)
	Insufficient forecast x	Excessive forecast x	Insufficient forecast x	Excessive forecast x	Insufficient forecast x	Insufficient forecast x
PAS=0.8	5.9	14.7	14.7	36.8	26.6	29.5
PAS=0.7	4.9	16.0	12.3	39.9	22.2	24.7
PAS=0.5	3.3	18.3	8.3	45.8	15.0	16.7
PAS=0.3	1.9	21.0	4.8	--	8.7	9.7

https://doi.org/10.5194/egusphere-2023-2613
Preprint. Discussion started: 18 December 2023

Table 7. Same as Table 6, but for precipitation at the level of torrential rain and above (u = 25, 50 and 100 mm).

PAS value	Observation u=25 mm	Observation u=50 mm		Observation u=100 mm	
	Excessive forecast x	Insufficient forecast x	Excessive forecast x	Insufficient forecast x	Excessive forecast x
PAS=0.877	--	34.1	68.1	68.1	136.2
PAS=0.7	--	24.7	79.9	49.4	159.7
PAS=0.5	--	16.7	91.6	33.3	183.3
PAS=0.3	52.4	9.7	104.9	19.4	209.7
PAS=0.1	62.9	3.2	125.9	6.4	251.7

Table 8. PAS and TS of 12 h accumulated precipitation at 12:00 UTC on July 16, 2019.

	Clear/rainy	≥0.1 mm	≥10 mm	≥25 mm	≥50 mm
PAS	0.808	0.617	0.256	0.200	0.104
TS	0.690	0.381	0.194	0.076	0.006

Table 9. Same as Table 8, but for 12:00 UTC on June 13, 2020.

	Clear/rainy	≥0.1 mm	≥10 mm	≥25 mm	≥50 mm
PAS	0.734	0.457	0.228	0.185	0.116
TS	0.816	0.625	0.338	0.149	0.036

Table 10. Accuracy scores of insufficient precipitation forecast (IPS), excessive precipitation forecast (EPS) and insufficient and excessive precipitation forecast (IEPS) of 12 h accumulated precipitation for two precipitation processes.

	IPS	EPS	IEPS
Case 1	-0.376	0.389	0.057
Case 2	-0.400	0.597	0.325

https://doi.org/10.5194/egusphere-2023-2613
Preprint. Discussion started: 18 December 2023

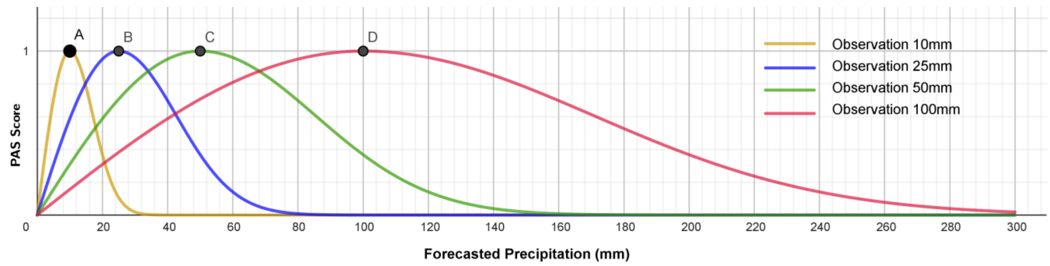

Figure 1. Schematic diagram of the precipitation forecast accuracy score (PAS) curves when the observed precipitation amounts are 10, 25, 50 and 100 mm.

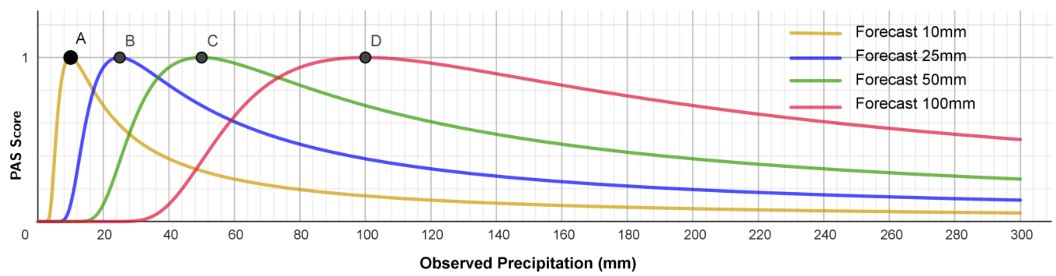

Figure 2. PAS curves corresponding to different forecasted precipitation amounts (u = 10, 25, 50 and 100 mm).

https://doi.org/10.5194/egusphere-2023-2613
Preprint. Discussion started: 18 December 2023

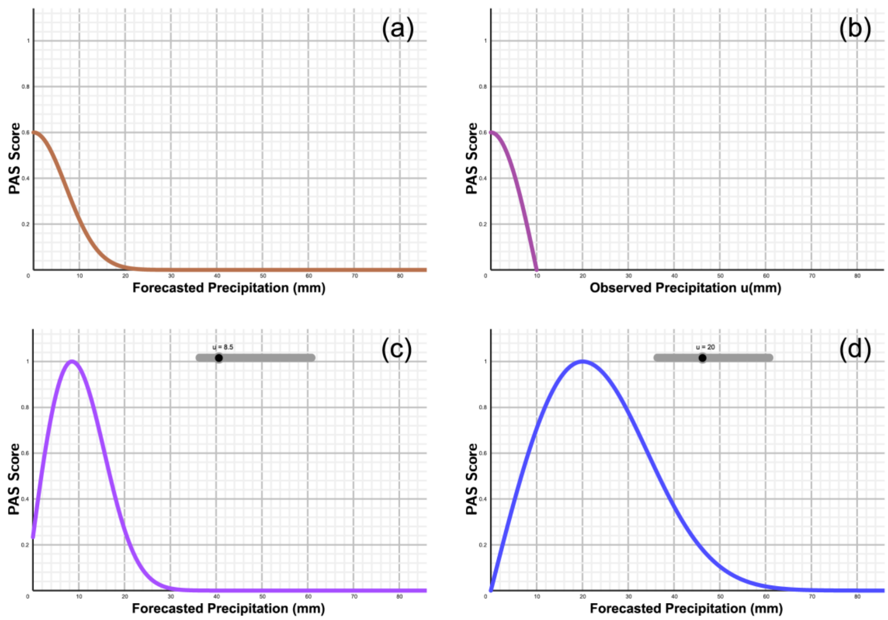

Figure 3. PAS curves of precipitation forecasts when (a) the observed precipitation u = 0 mm, (b) the observed precipitation u < 10 mm and the forecasted precipitation x = 0 mm (the horizontal coordinate denotes the observed precipitation u), (c) the observed precipitation 0 ≤ u < 10 mm, and (d) the observed precipitation u ≥ 10 mm.

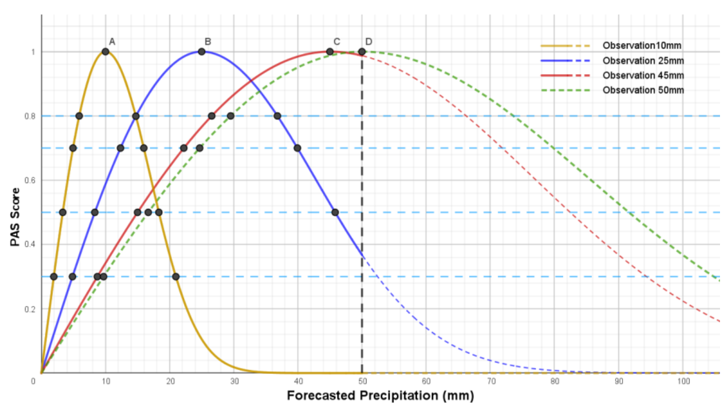

Figure 4. PAS curves of the forecasts under general precipitation conditions (*u* = 10, 25 and 45 mm).

https://doi.org/10.5194/egusphere-2023-2613
Preprint. Discussion started: 18 December 2023

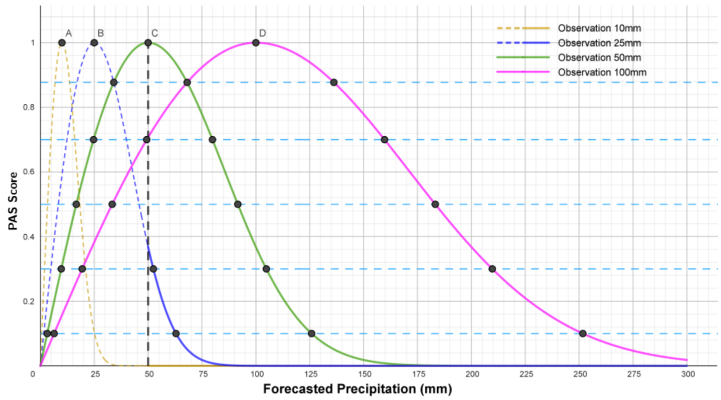

Figure 5. Same as Fig. 4, but for precipitation at the level of torrential rain and above (u = 25, 50 and 100 mm).

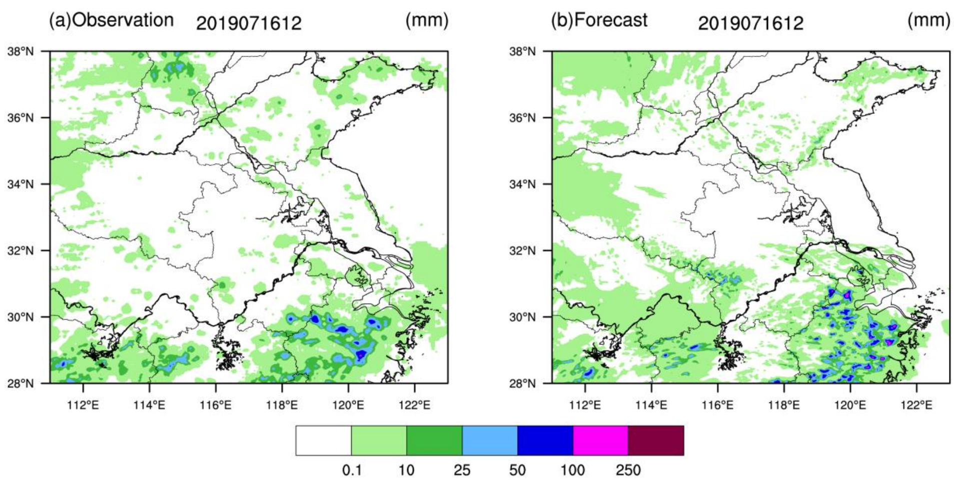

Figure 6. Accumulated precipitation (a) observed and (b) forecasted from 00:00 to 12:00 UTC on July 16, 2019.

https://doi.org/10.5194/egusphere-2023-2613
Preprint. Discussion started: 18 December 2023

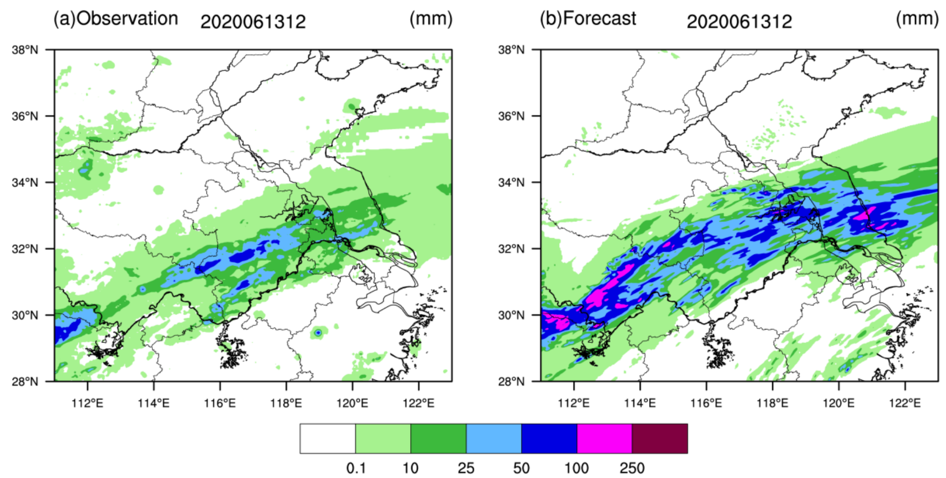

Figure 7. Accumulated precipitation (a) observed and (b) forecasted from 00:00 to 12:00 UTC on June 13, 2020.

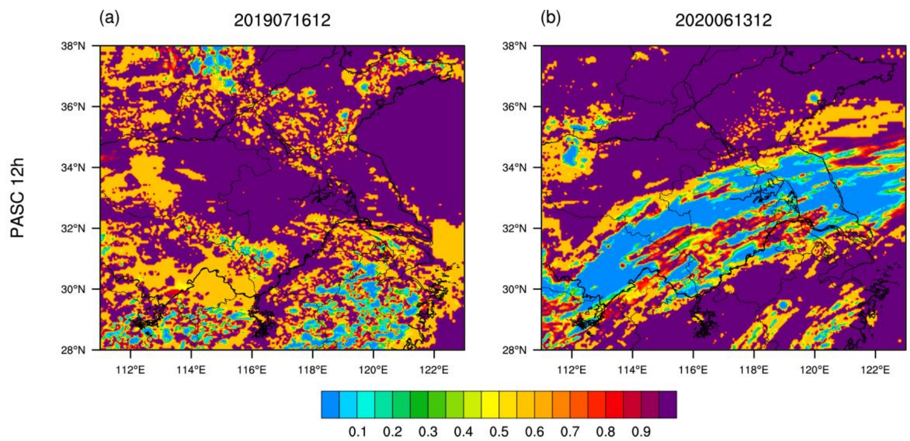

Figure 8. Distributions of the PAS clear/rainy forecast accuracy score (PASC) of 12 h accumulated precipitation for (a) case 1 from 00:00–12:00 UTC on July 16, 2019, and (b) case 2 from 00:00–12:00 UTC on June 13, 2020.

https://doi.org/10.5194/egusphere-2023-2613
Preprint. Discussion started: 18 December 2023

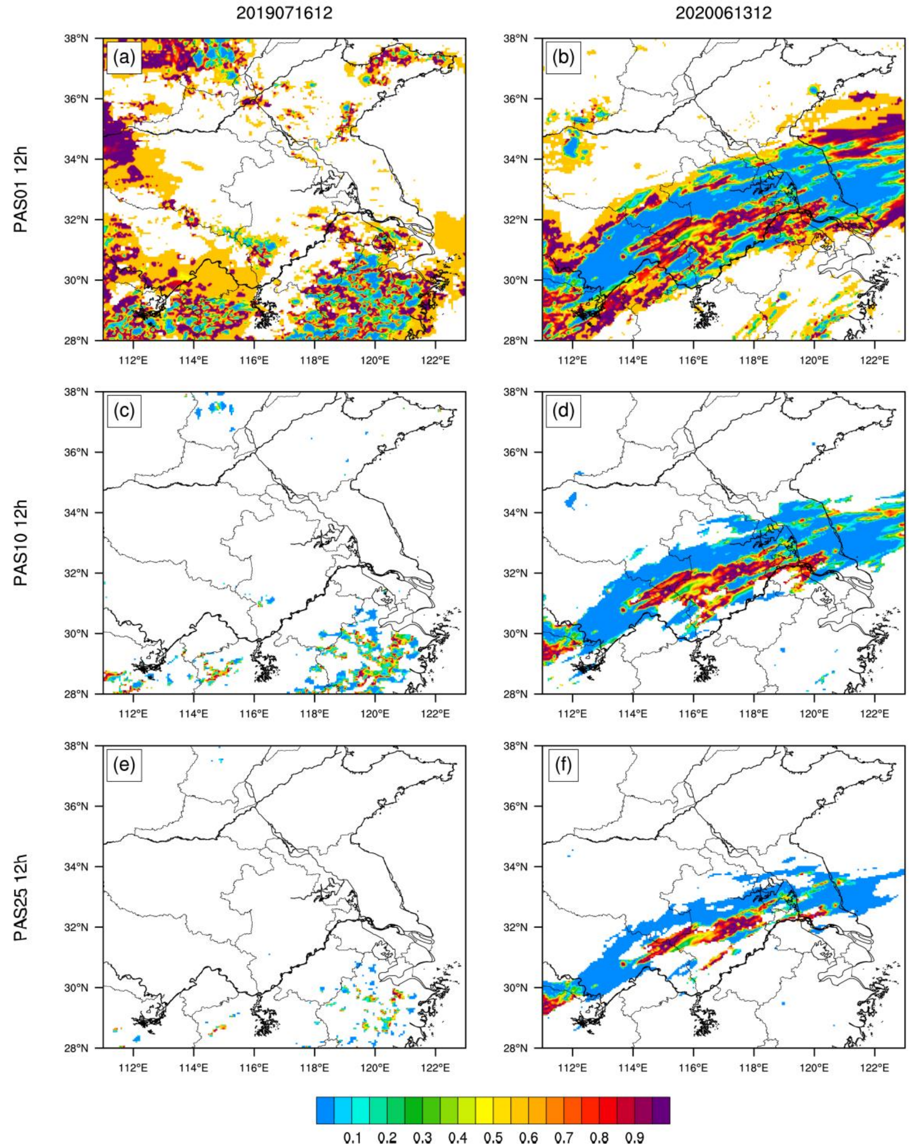

Figure 9. Distributions of PAS of 12 h accumulated precipitation ≥ 0.1 mm for (a) case 1 from 00:00–12:00 UTC on July 16, 2019 and (b) case 2 from 00:00–12:00 UTC on June 13, 2020, ≥10 mm for (c) case 1 and (d) case 2, and ≥ 25 mm for (e) case 1 and (f) case 2.

https://doi.org/10.5194/egusphere-2023-2613
Preprint. Discussion started: 18 December 2023

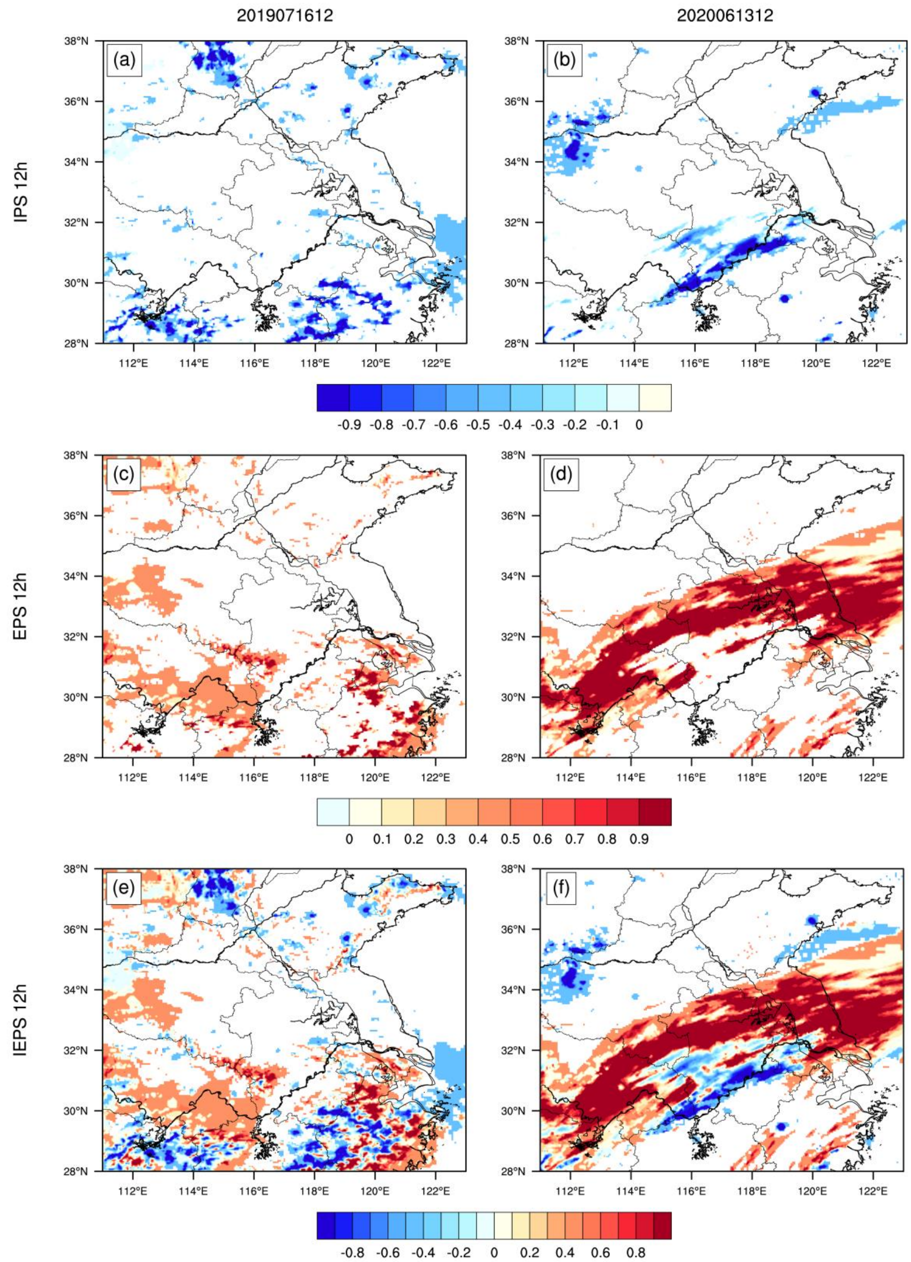

Figure 10. Distributions of IPS of 12 h accumulated precipitation for (a) case 1 from 00:00–12:00 UTC on July 16, 2019, and (b) case 2 from 00:00–12:00 UTC on June 13, 2020, EPS for (c) case 1 and (d) case 2, and IEPS for (e) case 1 and (f) case 2.