A General Comprehensive Evaluation Method for Cross-Scale Precipitation Forecasts

Bing Zhang[1], Mingjian Zeng[1]*, Anning Huang[2]*, Zhengkun Qin[3,4], Couhua Liu[5], Wenru Shi[1],

Xin Li[1], Kefeng Zhu[1], Chunlei Gu[2], Jialing Zhou[1]

1. Key Laboratory of Transportation Meteorology of China Meteorological Administration, Nanjing Joint Institute for Atmospheric Sciences, Nanjing 210041, China

2. School of Atmospheric Sciences, Nanjing University, Nanjing 210023, China

3. School of Atmospheric Sciences, Nanjing University of Information Science & Technology, Nanjing 210044, China

4. Joint Center of Data Assimilation Research and Applications, Nanjing University of Information Science & Technology, Nanjing 210044, China

5. National Meteorological Centre, Beijing 100081, China

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

Short summary

By directly analyzing the proximity of precipitation forecasts and observations, a precipitation forecast accuracy score (PAS) method was constructed. This method does not utilize 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.

Abstract

With the development of refined numerical forecasts, problems such as score distortion due to the division of precipitation thresholds in both traditional and improved scoring methods for precipitation forecasts and the increasing subjective risk arising from the scale setting of the neighbourhood spatial verification method have become increasingly prominent. To address these issues, a general comprehensive evaluation method (GCEM) is developed for cross-scale precipitation forecasts by directly analyzing the proximity of precipitation forecasts and observations in this study. In addition to the core indicator of the precipitation accuracy score (PAS), 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 does not delimit the area of influence; it constitutes a fair scoring formula with objective performance and can be suitable for evaluating 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 threat score (TS) are conducted for two typical precipitation weather processes. The results show that relative to the TS, the PAS better aligns with subjective expectations, indicating that the PAS is more reasonable than the TS. In the case of an extreme precipitation event in Henan, China, two high-resolution models were evaluated using the PAS, TS, and fraction skill score (FSS), verifying the evaluation ability of PAS scoring for predicting extreme precipitation events. In addition, other indices of the GCEM are utilized to analyze the range and extent of both insufficient and excessive forecasts of precipitation, as well as the precipitation

forecasting ability for different weather processes. These indices not only provide overall scores similar to those of the TS for individual cases but also support two-dimensional score distribution plots, which can comprehensively reflect the performance and characteristics of 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 et al., 2023). Short duration heavy rainfall often leads 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 analyzing 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., 2022; Liu et al., 2022b; Gofa et al., 2018). However, there are several shortcomings in current precipitation verification approaches. For instance, with traditional scoring methods, small errors in the location or timing of convective features can lead to false alarms and missed events, and their utility is limited regarding diagnosing

67 model errors such as a displaced forecast feature or an incorrect mode of convective organization;

68 thus, traditional scoring methods often fail to reflect model performance improvements (Ahijevych et

69 al., 2009). For high-resolution precipitation forecasts, even if the spatial distribution and intensity of

70 precipitation are consistent with the observations, slight spatial and temporal deviations between

71 forecasts and observations may still result in a large false alarm ratio and missed alarm ratio, leading

72 to lower forecast scores (Zhao and Zhang, 2018). With the rapid development of seamless fine

73 quantitative precipitation forecasts, the need for objective and rational evaluations of the accuracy

74 and characteristics of precipitation forecasts has become increasingly important and urgent (Chen et

75 al., 2021).

76   Precipitation forecast verification involves various methods, including traditional contingency

77 table-based classification verification and spatial verification methods. The traditional verification

78 method can be traced back to 1884, when Finley introduced a dichotomous contingency table for

79 tornado forecasts and evaluated these forecasts using the proportion correct scoring method (Finley,

80 1884). Subsequently, systematic attention was given to the evaluation of forecast classification

81 methods, and Finley's forecast verification method became a classic example of the discussion of

82 forecast scoring methods (Murphy, 1996). Gilbert (1984) proposed two scoring methods, namely, the

83 ratio of verification and the ratio of success in forecasting. The ratio of verification later became

84 known as the threat score (TS) (Palmer and Allen, 1949) or the critical success index (Donaldson et

85 al., 1975; Mason, 1989). The ratio of success is referred to as the Gilbert skill score (GSS) (Schaefer,

86 1990) or the equitable threat score (ETS) (Doswell et al., 1990; Gandin and Murphy, 1992). The TS

87 encourages correct event forecasts (hits) and accounts for the impacts on the false alarm and missed

88 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., 2022a).

In addition to the TS and ETS, the methods of traditional contingency table-based classification verification include Peirce skill score (PSS) (Peirce, 1884; Hanssen and Kuipers, 1965; Murphy and Daan, 1985; Flueck, 1987), Heidke skill score (HSS) (Doolittle, 1885; Doolittle, 1888; Heidke, 1926), probability of detection (POD), frequency bias (BIAS), accuracy (ACC), false alarm ratio (FAR), missing ratio (MR), probability of false detection (POFD), etc. 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., 2022a). 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 two scoring concepts, i.e., the ratio of verification and the ratio of success 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). With the continuous introduction of new scoring methods, several problems in traditional verification have been solved. However, the advantageous position of the TS remains unchallenged. Although the reasons for this are varied, its objectivity and practicality merit attention.

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 forecast accuracy that excludes hit forecasts for "non-occurrence" precipitation events (referred to as no precipitation), and its calculation formula is 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 cannot guarantee that two adjacent precipitation values will always fall within the same threshold range. Slightly different precipitation values are not 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 meso- 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 meso- and small-scale forecasts, resulting in traditional scoring methods often cannot reflect these improvements in terms of 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, both "false alarms" and "missed alarms" will occur,

which is referred to as the "double penalty" phenomenon. This phenomenon leads to a score lower than the subjectively 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 rain are often at the unskilled end of the scoring values (Chen et al., 2019). In recent years, new mainstream scoring methods have addressed most of 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 meso- 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 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 and forecast B is 98 mm, forecast A is evidently 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 scores only 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 accuracy score (PAS), which will be discussed later,

155 forecast A scores 0.998 points, while forecast B scores 0.398 points, confirming the rationality and

156 validity of this new method.

157 To address the "double penalty" issue, a common approach is to employ the neighbourhood

158 spatial verification method (also known as the fuzzy method), which has two specific processing

159 forms. The first approach is simple upscaling, which uses a certain method (such as value averaging,

160 maximum, value weighting) to select values within the scale range(Chen et al., 2019), adjusting the

161 high-resolution forecast and observation information to a larger scale to reduce the accidental

162 information of high-resolution data, and then using the traditional skill score (Yates et al., 2006;

163 Weygandt et al., 2004). The other form is the improved neighbourhood spatial verification method

164 proposed by Roberts and Lean (2008). By referring to the Murphy skill score, this method obtains

165 comprehensive evaluation information by comparing the occurrence frequency (probability) of

166 precipitation within different scale windows. If the forecasted occurrence frequency closely

167 approximates the observed occurrence frequency, the forecast is considered valuable (Zhao and

168 Zhang, 2018). From the perspective of the precipitation occurrence probability within the analysis

169 region, the precipitation occurrence probability for observations and forecasts is the ratio of the

170 precipitation area to the analyzed area of the region, which is referred to as the fraction skill score

171 (FSS). These two processing methods effectively solve the "double penalty" problem, but neither can

172 address the issue of excessive smoothness of the precipitation fields during the upscaling process

173 (Zhao and Zhang, 2018), which may result in the omission of some small- to meso-scale information

174 (Zepeda-Arce et al., 2000).

175 The neighbourhood spatial verification method considers values that are spatially and

176 temporally adjacent between forecasts and observations during the matching process, thus relaxing

177 the strict requirements for spatiotemporal matching (Ebert, 2008; Casati et al., 2008). However, since

178 the determination of the neighbourhood range is a rather subjective process, it hinders the

179 standardization of verification scores and lacks comparability, which may negatively affect objective

180 quantitative verification. Numerous experiments have shown that there is an obvious improvement in

181 the scoring values after adopting the neighbourhood spatial verification method (Chen et al., 2019),

182 particularly for forecasts of large-magnitude precipitation. Nevertheless, the purpose of scoring is not

183 to achieve a monotonous increase in scoring values but rather to follow the principle of objectivity as

184 much as possible. Errors are errors and cannot be solved by simply lowering the standard. Instead,

185 reasonable and fair criteria should be utilized to reflect the true extent of errors.

186   Currently, numerical weather forecasts and intelligent gridded forecasts have been developed to

187 output high-resolution precipitation products, while precipitation observations, whether in the form

188 of gridded or station data, are already high-resolution. Staying at the dichotomous classification level

189 for precipitation verification not only wastes existing data resources but also fails to meet the

190 evaluation requirements of refined forecasts. Therefore, to adapt to the development of refined

191 forecasts, a new scoring method is needed. In light of this, a comprehensive verification index for

192 precipitation forecasts is designed, and the following five aspects are considered. (1) The impact of

193 categorical events on rainstorm forecasts should be reduced. In particular, high-resolution forecasts

194 can refer to continuous variables for scoring methods. Especially for the evaluation of

195 high-resolution precipitation forecasts, the scoring method of continuous variables can be borrowed

196 for reference. (2) The design of the scoring method should aim to minimize subjective factors such as

197 the artificial range division and condition settings, ensuring scoring objectivity and comparability. (3)

198 The designed scoring performance indices should possess ideal attributes such as fairness,

199 independent of climatological probability, suitability for extreme events, and boundedness as much

200 as possible. (4) The devised scoring method should be easy to promote, concise and efficient, with

201 clear concepts and scientific rationality. (5) Different comprehensive verification indices for

202 precipitation forecasts should reflect the forecasting performance and characteristics of

203 high-resolution quantitative precipitation products from various perspectives.

204   In this study, on the basis of analyzing the limitations of traditional verification methods as well

205 as improved methods, a new general comprehensive evaluation method (GCEM) for cross-scale

206 precipitation prediction is proposed. This method is applied and verified through practical examples.

207 The remainder of this paper is organized as follows. Section 2 provides an overview of various

208 scoring indices and their attributes in the GCEM and introduces the optimization processing method

209 for the PAS index in the application. Through ideal experiments, the characteristics of the scoring

210 methods are analyzed based on the score curves described in Section 3. Section 4 presents

211 comparative experiments, including the new scoring method, the traditional scoring method and the

212 neighbourhood spatial verification method based on typical cases. Finally, a summary and discussion

213 are presented in Section 5.

214 **2 Cross-scale general comprehensive evaluation method**

215 **2.1 Overview of the general comprehensive evaluation method**

216   To address the issues of "distorted scores due to the division of precipitation thresholds and

217 increased subjective risks brought about by the setting of the neighbourhood spatial verification

218 method" in traditional and improved precipitation scoring methods, this study refers to the

219 verification method for heavy rainfall forecasts based on predictability (Chen et al., 2019) and

220 combines the advantages of relative and absolute errors. A GCEM is constructed by directly

221   analyzing the proximity of forecasted precipitation to observed precipitation. It primarily includes

222   the PAS, and the expression of its core scoring function is as follows:

223
$$PAS = \begin{cases} \sin\left(\frac{\pi}{2} \cdot \frac{x}{u}\right), & 0 \le x < u \\ e^{-\left(\frac{x-u}{u}\right)^2}, & 0 < u \le x \end{cases} \tag{1}$$

224   where PAS represents the scoring value, x is the forecasted precipitation (mm), and u is the observed

225   precipitation (mm). The PAS falls between 0 and 1, where a higher score indicates a better

226   precipitation forecast effect. When PAS = 1, it signifies a perfect forecast, indicating that the

227   forecasted and observed precipitation match entirely. For Eq. (1), given the observation value $u>0$

228   mm, when the forecasted precipitation is 0 mm, then PAS=0, indicating that the model has no

229   forecast skill. When the forecasted precipitation amount is sufficiently large, PAS $\rightarrow$ 0, indicating no

230   forecast skill as well (Fig. 1). Additionally, considering the large fluctuation characteristics of the

231   function curve when the observed precipitation is less than 10 mm, Eq. (1) was smoothed and

232   optimized (see Section 2.2 for details).

233   The GCEM system also includes the following indices:

234   (1) Insufficient precipitation index (IPI), whose core scoring function expression is as follows:

235
$$IPI = \sin\left(\frac{\pi}{2} \cdot \frac{x}{u}\right) - 1, \ 0 \le x < u \tag{2}$$

236   where IPI represents the scoring value, reflecting the degree of underestimation in precipitation

237   forecasts when the forecasted value is less than the observed value. The IPI falls within $[-1, 0)$,

238   where a value closer to 0 indicates a lower degree of underestimation.

239   (2) Excessive precipitation index (EPI), whose core scoring function expression is as follows:

240
$$EPI = 1 - e^{-\left(\frac{x-u}{u}\right)^2}, \quad 0 < u < x \tag{3}$$

241   where EPI represents the scoring value, reflecting the degree of overestimation in precipitation

242   forecasts when the forecasted value exceeds the observed value. The EPI falls within (0, 1), where a

243   value closer to 0 indicates a lower degree of overestimation.

244   (3) Insufficient and excessive precipitation index (IEPI), whose core scoring function

245   expression is as follows:

246
$$IEPI = \begin{cases} \sin\left(\frac{\pi}{2} \cdot \frac{x}{u}\right) - 1, & 0 \le x < u \\ 1 - e^{-\left(\frac{x-u}{u}\right)^2}, & 0 < u \le x \end{cases} \quad (4)$$

247   where IEPI represents the scoring value, reflecting the degree of deviation of the forecasted

248   precipitation from the observed precipitation. The IEPI falls within $[-1, 1)$, where a value closer to 0

249   indicates a lower degree of forecast deviation. An IEPI less (more) than 0 indicates an insufficient

250   (excessive) forecast, and an IEPI equal to 0 represents an unbiased forecast.

251   Additional explanation: Eqs. (2-4) are a series of theoretical indicator formulas derived from Eq.

252   (1), therefore Eqs. (2-4) are referred as the core calculation formulas for the IPI, EPI, and IEPI,

253   respectively. In practical applications, the optimized solution will be used (see Section 2.2) to

254   calculate the IPI, EPI, and IEPI for the situations of $u \ge 0.1$ mm or $x \ge 0.1$ mm.

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

256   as follows.

257
$$PASC = \begin{cases} 1 & 0 \le u < 0.1 \text{ and } 0 \le x < 0.1 \\ PAS_{|ux0.1} & u \ge 0.1 \text{ or } x \ge 0.1 \end{cases} \quad (5)$$

258   where PASC represents the PAS scoring value for clear/rainy forecasts. "$0 \le u < 0.1$ and $0 \le x <$

259   0.1" denotes the correctly forecasted non-precipitation event with PASC=1. $PAS_{|ux0.1}$ denotes the

260   overall PAS for precipitation forecasts under specific conditions where the observed precipitation $u \ge$

261   0.1 mm or the forecasted precipitation $x \ge 0.1$ mm.

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 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. According to 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 amounts of 25 mm and 100 mm, respectively, while the scores are both 0.62. From the perspective of forecast service effectiveness,

284   this aligns with general public perception.

285   (4) Suitability for extreme events. From the PAS scoring curves for forecasts corresponding to

286   different observed precipitation amounts (u = 10, 25, 50 and 100 mm) (Fig. 1), it is evident that the

287   PAS scoring method performs well in evaluating precipitation event forecasts at the level of

288   torrential rain and above. For example, when the observed precipitation is 100 mm, with forecasted

289   amounts of 59 mm and 147.2 mm, the PASs are both 0.8, whereas the TSs are 0 and 1, and the

290   improved TSs are 0.8 and 1, respectively. This result indicates that the PAS is suitable for scoring

291   heavy rainfall events, meeting the general applicability requirements as a scoring method that does

292   not degrade due to extreme events.

293   (5) Moderate symmetry. In Eq. (1), let the observed precipitation is the independent variable u,

294   and the forecasted precipitation is the parameter x. Similarly, for different magnitudes of forecasted

295   precipitation (parameter x = 10, 25, 50 and 100 mm) and observed precipitation (variable u) ranging

296   from 0 to 300 mm, the corresponding scores are shown in Fig. 2. The scores also vary with the

297   degree of proximity between forecasts and observations. Figures 1 and 2 exhibit similar trends but

298   are not identical, illustrating that the PAS possesses moderate symmetry.

299   2.2 PAS verification for precipitation forecasts

300   From the properties of the core verification function of the PAS, it is noted that when the

301   observed precipitation u < 10 mm, there is a large gradient in the PAS curve. A slight change in the

302   forecasted value (x) can result in a large fluctuation in the PAS. To account for this characteristic,

303   based on a comprehensive analysis in combination with the sensitivity of forecasters and the public

304   to small-scale precipitation, a smoothing optimization scheme is applied to the PAS curve for

305   accumulated precipitation below 10 mm. Similarly, the IPI, EPI, IEPI and PASC curves are

306   appropriately smoothed and optimized according to their respective definitions.

307   Assumptions:

308   (1) PAS $= 0.6\text{PAS}_{|u\to0}$ when $u = 0$ mm, and $x \neq 0$ mm;

309     $\text{PAS}_{|u\to0}$ denotes the PAS for the case of observed precipitation $0 < u \leq 0.1$ mm;

310   (2) PAS $= 0.6\text{PAS}_{|x\to0}$ when $x = 0$ mm, and $0 < u < 10$ mm;

311     $\text{PAS}_{|x\to0}$ denotes the PAS for the case of forecasted precipitation $0 < x \leq 0.1$ mm.

312   1. When the observed precipitation $u = 0$ mm and the forecasted precipitation $x > 0$ mm (Fig.

313   3a), let PAS $= 0.6\text{PAS}_{|u\to0}$, then,

314     $$\text{PAS} = 0.6e^{-\left(\frac{x}{10}\right)^2} \qquad x > 0 \tag{6}$$

315   2. When the forecasted precipitation $x = 0$ mm and the observed precipitation $0 < u < 10$ mm

316   (Fig. 3b), let PAS $= 0.6\text{PAS}_{|x\to0}$, then,

317     $$\text{PAS} = 0.6\sin\left(\frac{\pi}{2}\cdot\frac{10-u}{10}\right), \quad 0 < u < 10 \tag{7}$$

318   The coefficient was set to 0.6. According to Eqs. (6-7), when the situation is the observation

319   u=0 mm and forecast x=0.1 mm or the observation u=0.1 mm and forecast x=0 mm, PAS=0.6,

320   suggesting that the forecast effect has just reached the standard, like when the ACC reaches 0.6,

321   which indicates that the model forecast effect is available (Zhao and Zhang, 2018).

322   3. When the observed precipitation $0 < u < 10$ mm and the forecasted precipitation $x \neq 0$ (Fig.

323   3c), then,

324     $$\text{PAS} = \begin{cases} \sin\left(\frac{\pi}{2}\cdot\frac{x-u+10}{10}\right), & 0 < x < u, \ 0 < u < 10 \\ e^{-\left(\frac{x-u}{10}\right)^2}, & u \leq x, \quad 0 < u < 10 \end{cases} \tag{8}$$

325   4. When the observed precipitation $u \geq 10$ mm (Fig. 3d), then

$$326 \qquad PAS = \begin{cases} \sin\left(\dfrac{\pi}{2} \cdot \dfrac{x}{u}\right), & 0 \leq x < u, \ u \geq 10 \\[2ex] e^{-\left(\frac{x-u}{u}\right)^2}, & u \leq x, \quad u \geq 10 \end{cases} \qquad (9)$$

327   To compare with the traditional scoring method, the new scoring method for precipitation

328   forecasting adopts the "classification before verification, no classification during verification"

329   approach. Scoring for precipitation processes over different accumulation periods is referenced but

330   not limited to the commonly used precipitation classification approaches in practical operations, as

331   shown in Tables 3-5.

332   3 Ideal experimental validation of the new verification method

333   3.1 Validation of forecast scoring results for general precipitation

334   General precipitation refers to precipitation ranging from light rain to heavy rain, i.e., 24-hour

335   accumulated precipitation within [0.1 mm, 50 mm). Figure 4 shows the schematic diagram of PAS

336   scores for general precipitation. The forecasted amounts are compared under conditions when the

337   24-hour accumulated precipitation is 10 mm, 25 mm and 45 mm and the PAS scores are 0.8, 0.7, 0.5

338   and 0.3 (Table 6). When the observed precipitation is 10 mm, the forecasted amounts of 5.9 mm and

339   14.7 mm both have a PAS score of 0.8, with differences from the perfect forecast value (10 mm) of

340   4.1 mm and 4.7 mm, respectively; the forecasted amounts with a PAS score of 0.3 are 1.9 and 21.0

341   mm, differing by 8.1 mm and 11.0 mm from the perfect forecast value (10 mm), respectively. When

342   the observed precipitation is 25 mm, the forecasted amounts with a PAS score of 0.8 are 14.7 mm

343   and 36.8 mm, with differences from the perfect forecast value (25 mm) of 10.3 mm and 11.8 mm,

344   respectively; the forecasts with a PAS score of 0.5 are 8.3 mm and 45.8 mm, differing by 16.7 mm

345   and 20.8 mm from the perfect forecast value (25 mm), respectively.

346   For forecasts with the same observed precipitation and the same scores, the absolute errors of an

347 insufficient forecast and observation are smaller than those of an excessive forecast and observation,

348 and the higher the scores are, the closer the absolute errors of the forecasts. When the observed

349 precipitation is 50 mm, only the insufficient precipitation forecast is scored since a precipitation

350 forecast exceeding 50 mm is not considered within the scope of general precipitation evaluation. The

351 scoring experimental results align with expectations.

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

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

355 storm magnitude. The predicted precipitation is compared when the 24-hour cumulative observed

356 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

357 7). When the observed precipitation is 25 mm, only forecasts ≥ 50 mm are involved in the rating,

358 with PASs of 0.3 and 0.1 for forecasts of 52.4 and 62.9 mm, respectively.

359 When the PAS is 0.877 and the observed precipitation is 50 mm, the predicted values are 34.1

360 and 68.1 mm, respectively; when the observed precipitation is 100 mm, the predicted values are 68.1

361 and 136.2 mm, respectively. When the observed precipitation is 50 or 100 mm, the prediction is 68.1

362 mm, with a score of 0.877. The absolute error is 18.1 mm for the excessive precipitation forecast and

363 31.9 mm for the insufficient precipitation forecast. This result indicates that the scoring tolerance

364 increases as the grade of observed precipitation increases and gradually expands through continuous

365 changes, avoiding discontinuous increases caused by changes in magnitude.

366 When the observed precipitation is 50 mm and the PAS is 0.3, the insufficient forecast is 9.7

367 mm and the excessive forecast is 104.9 mm. When the observed precipitation is 100 mm, the

368 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.

4 Example-based comparative experiments for the new verification method

Different examples are selected for the new precipitation verification method, and its multifaceted characteristics are demonstrated through comparative experiments. In Section 4.1, two typical cases are selected, the performance characteristics of the PAS and TS are compared, and the indicators of insufficient and excessive forecasts and spatial verification in the GCEM are analyzed. In Section 4.2, typical case of extreme precipitation event is selected, and the forecast results of different high-resolution models using the PAS, TS, and FSS methods are evaluated to verify the advantages and characteristics of the new precipitation verification method for extreme precipitation events.

4.1 Comparative experiments of two typical processes

4.1.1 Introduction of typical cases

Comparative experiments of PAS and traditional TS are conducted for 12-hour accumulated precipitation for two typical cases. One case pertains to the precipitation weather process occurring during 00:00 to 12:00 UTC on 16 July 2019 (referred to as "Case 1"), which is dominated by a weak

391 weather system. The other case relates to the precipitation weather process occurring during 00:00 to

392 12:00 UTC on 13 June 2020 (referred to as "Case 2"), which is predominantly associated with a

393 strong weather system.

394   Both precipitation cases are associated with precipitation during the Meiyu period. Case 1,

395 which occurred during the Meiyu period of 2019 and was characterized by scattered precipitation

396 under weak synoptic-scale forcing. The low-intensity shear line system is located south of the

397 Yangtze River. There are two precipitation concentration areas, one at the intersection of Hunan

398 Province and Jiangxi Province and the other covering the majority of Zhejiang Province. The

399 precipitation process in Case 2 (12–13 June) was the first round of widespread rainstorms during the

400 Meiyu period of 2020, including heavy precipitation affected by a low-level vortex shear system.

401 The western section of the low-level vortex shear is relatively stable, while the eastern section

402 slightly presses southwards. Southwesterly airflow developed and pushed northwards, and a strong

403 wind speed belt persisted for a long time in the Jianghuai region. Moreover, the Jianghan–Jianghuai

404 region maintained a high-energy and high-moisture state, resulting in persistent heavy rainfall.

405   A subjective analysis of these two weather processes reveals that for the event on 16 July 2019

406 (Fig. 6), the forecasted precipitation intensity and rainfall areas are relatively consistent with the

407 observations. There are two distinct heavy rainfall areas in the eastern and southern parts of the

408 Yangtze River, with particularly high accuracy in forecasting scattered rainstorms in Zhejiang

409 Province located in the eastern section. In contrast, for the precipitation weather process on 13 June

410 2020 (Fig. 7), it is evident that there is an overestimation of the precipitation forecast.

411 **4.1.2 Data and methods**

412   The observed precipitation data are provided by the China Meteorological Administration

multisource merged precipitation analysis system (CMPAS), developed by the National

Meteorological Information Centre 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

$0.05° \times 0.05°$. 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 $0.05° \times 0.05°$,

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

match the forecasted data to the grid points of the observed precipitation. Specifically, the forecasted

data on the model grid nearest to the observed grid are used as the forecasted value at this observed

grid.

(3) Analyze the relationship between the 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, 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.

4.1.3 Analysis of the comparative experiment results

For the precipitation process on 16 July 2019, the 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 TSs for the weather process on 13 June 2020. For example, the TS is 0.381 for 12-hour accumulated precipitation ≥ 0.1 mm during 00:00 to 12:00 UTC on 16 July 2019 (Table 8), while this score is 0.625 for that during 00:00 to 12:00 UTC on 13 June 2020 (Table 9), which differs from the subjective judgement.

For the precipitation process during 00:00 to 12:00 UTC on 16 July 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 during 00:00 to 12:00 UTC on 13 June 2020. For instance, the overall PAS is 0.617 for 12-hour accumulated precipitation ≥ 0.1 mm during 00:00 to 12:00 UTC on 16 July 2019. This PAS is higher than the PAS of 0.457 for the precipitation process during 00:00 to 12:00 UTC on 13 June 2020, which aligns with subjective judgement.

For the precipitation process during 00:00 to 12:00 UTC on 16 July 2019, the PAS for each magnitude is higher than the corresponding TS, addressing the issue of TSs being lower. For the precipitation process during 00:00 to 12:00 UTC on 13 June 2020, the PASs for clear/rainy and the

457 magnitudes of ≥ 0.1 mm and ≥ 10 mm are lower than the corresponding TSs, whereas the PASs for

458 the magnitudes of ≥ 25 mm and ≥ 50 mm are higher than the corresponding TSs. This result indicates

459 that the PAS is different from the magnitude-improved TS and the neighbourhood spatial verification

460 method. Both the magnitude-improved TS and the neighbourhood spatial verification method

461 increase the tolerance, leading to a monotonous increase in scores. This result also demonstrates that

462 the PAS has good discrimination ability for extreme events. The PAS assigns scores based on the

463 proximity of the forecast to the observation, making it more reliable for precipitation evaluation than

464 the TS.

465 **4.1.4 Analysis of the indices in the new verification method**

466   Modern forecast verification is based mainly on spatial verification methods to compensate for

467 the shortcomings of traditional methods. The literature review of Gilleland et al. (2009) defines four

468 main categories of methods: neighbourhood, scale separation, features based, and field deformation

469 (Ahijevych et al., 2009). These methods can analyze more comprehensively in specific individual

470 cases, but seem to be less able to provide direct overall scoring results than traditional scoring

471 methods in the statistics of long time series. GCEM is based on point-to-point scoring statistics,

472 without a radius of influence, no isolation of features at each scale, and no definition of objects in the

473 forecast and observation to analyze the similarity of the objects or to fit the forecast objects through

474 deformation operations. However, the GCEM still has spatial attributes that can discriminate spatial

475 forecast characteristics (e.g., insufficient or excessive forecasting scenarios) for different categories

476 of precipitation, and the GCEM can carry out statistical verification of long time series and produce

477 overall scoring results.

478   Regarding the issue of analyzing the sources of errors from the verification results, objectively

479 tracing these errors back from a single score can only determine whether an error was "insufficient

480 (missed alarm)" or "excessive (false alarm)". However, the advantage of the GCEM lies in its ability

481 to decompose the score for each verification point and examine the forecasting performance at each

482 point, which is different from the dichotomous evaluation approach with only 0 and 1 outputs. These

483 indices not only provide overall scores for individual cases similar to the TS but also offer

484 two-dimensional score distribution plots, which can comprehensively reflect the performance and

485 characteristics of precipitation forecasts.

486   Figure 8 shows the distributions of the 12-hour accumulated precipitation PASC scores. In these

487 two cases, due to the high accuracy of non-precipitation forecasts, the overall PASC scores are

488 relatively high. However, for Case 1, the scores in Zhejiang are lower and scattered within a small

489 area. In contrast, for Case 2, there is a large area occupying most of the Jianghuai region with low

490 scores. Therefore, the PASC score of Case 1 (0.808) is higher than that of Case 2 (0.734).

491   Figure 9 shows the PAS distributions of 12-hour accumulated precipitation with magnitudes of

492 ≥ 0.1 mm, ≥ 10 mm and ≥ 25 mm. The blank points in the figure are the points that are excluded in

493 the scoring, following the scoring principle of "classification before verification, no classification

494 during verification" described in Section 2. From the PAS distributions of different magnitudes, for

495 Case 1, the high and low scores in the Zhejiang region are scattered among them. In contrast, for

496 Case 2, the scoring areas in the Jianghuai region have a larger area of low scores than high scores.

497 Therefore, Case 1 has higher PASs for the three categories (≥ 0.1 mm, ≥ 10 mm and ≥ 25 mm) than

498 Case 2, and the distributions also allow for distinguishing the areas with better and worse forecasting

499 performance.

500   Figure 10 shows the IPI, EPI and IEPI distributions of 12-hour accumulated precipitation. In

terms of the IPI, for Case 1, the large-value IPI 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 IPI areas are situated along the Yangtze River in Anhui and Jiangxi, as well as at the intersection of Henan and Shanxi. The IPIs 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 EPI, for Case 1, the large-value EPI areas are in Zhejiang and Jiangxi. In contrast, for Case 2, the large-value EPI areas are located in most of Hunan, Hubei, Anhui and Jiangsu, exhibiting a wide southwest–northeast orientation with a large area and degree. The EPI for Case 2 is larger than that for Case 1. The IEPI 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, with an IEPI of 0.057. However, for Case 2, the distribution of the excessive forecasts is obviously larger than that of the insufficient forecasts, with an IEPI of 0.325. This result indicates that Case 2 has poorer forecasting performance, with larger excessive forecasts being an important factor.

Consequently, analyzing 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.

4.2 Comparison experiment of extreme rainfall events

4.2.1 Introduction of the "21.7" extreme rainstorm event in Henan, China

From 17 to 23 July 2021, a rare extreme rainstorm event occurred in Henan Province, China. The extremely heavy rainstorm started in the southeastern of Henan Province in the morning of 17

523 July, then extended to the northern region, and ended in the morning of 23 July, lasting more than 6

524 days. The rainstorm occurred against the background of typhoon, Huang-Huai vortex, shear line and

525 convergence line, and was caused by the coupling of the low-level jet and boundary layer jet,

526 combined with the uplift of terrain (Wang et al., 2022; Su et al., 2021; Shi et al., 2021).

527   The period from 00:00 UTC on 18 July to 00:00 UTC on 22 July 2021 is the concentrated

528 period of heavy precipitation. To facilitate the study, the heavy rainstorm process is divided into three

529 periods: (1) 00:00 UTC on 19 July - 00:00 UTC on 20 July 2021. (2) 00:00 UTC on 20 July - 00:00

530 UTC on 21 July 2021. (3) 00:00 UTC on 21 July - 00:00 UTC on 22 July 2021. (Figs. 11a-c).

531 **4.2.2 Data and methods**

532   The observed precipitation data are provided by the CMPAS, with a spatial resolution of 0.05° ×

533 0.05°, similar to the case in Section 4.1. The forecast data come from two models. One is the PWAFS

534 model, which has a horizontal resolution of 3 km, similar to the case in Section 4.1. The other is the

535 global-regional assessment and prediction system (GRAPES) model independently developed by the

536 China Meteorological Administration, which has a horizontal resolution of 3 km.

537   (1) Determine the verification domain and verification points. The verification domain covers

538 the region of (30°N-40°N, 107.5°E-117.5°E). The verification points are defined based on the grid

539 points of the observed precipitation data, their spatial resolution is 0.05° × 0.05°, and the total

540 number of verification grid points is 40,401 (201 × 201).

541   (2) Prepare the observed and forecasted precipitation data and interpolate the forecasted

542 precipitation data onto the observed grid points. The 24-hour cumulative precipitation observation

543 data of the three periods were obtained from the 24-hour precipitation data of the CMPAS. The

544 forecast precipitation data in the three periods are the cumulative precipitation with a forecast time of

545   12 to 36 hours. For the case described in Section 4.1, the nearest neighbour method is used to match

546   the forecast data to the grid points of the observed precipitation.

547   (3) Analyze the relationship between the forecasted precipitation and observed precipitation.

548   PAS, TS and FSS were compared for the extreme rainstorm event in Henan, China.

549   As mentioned earlier, the FSS belongs to the neighbourhood category of spatial verification

550   methods and is an advanced evaluation method widely used in recent years. It can still yield valuable

551   scores when the model prediction intensity is spatially biased and can also represent the scale

552   information of forecasting skills. Therefore, in this case, the FSS scoring method was added for

553   comparative experiments. For FSS verification, 15 km, 25 km, 45 km, 75 km and 120 km are used as

554   the neighbourhood distances.

555   The brief steps of FSS calculation are as follows: 1. Determine the domain scope. Set the

556   neighbourhood point n, such as when n=3 (n is odd), the neighbourhood range is 15 km × 15 km, 2.

557   Calculate the spatial density in the observed binary observation fields (Eq.10), 3. Calculate the

558   spatial density in the binary forecast fields(Eq.11), 4. Calculate $FSS_{(n)}$ (Eq.12). (Please refer to the

559   article of Roberts and Lean (2008) for details.)

560
$$O(n)(i,j) = \frac{1}{n^2}\sum_{k=1}^{n}\sum_{i=1}^{n} I_o\left[i+k-1-\frac{n-1}{2}, j+l-1-\frac{n-1}{2}\right] \quad (10)$$

561
$$M(n)(i,j) = \frac{1}{n^2}\sum_{k=1}^{n}\sum_{i=1}^{n} I_M\left[i+k-1-\frac{n-1}{2}, j+l-1-\frac{n-1}{2}\right] \quad (11)$$

562
$$FSS_{(n)} = 1 - \frac{\frac{1}{N_xN_y}\sum_{i=1}^{N_x}\sum_{j=1}^{N_y}\left[O_{(n)i,j}-M_{(n)i,j}\right]^2}{\frac{1}{N_xN_y}\left[\sum_{i=1}^{N_x}\sum_{j=1}^{N_y}O_{(n)i,j}^2+\sum_{i=1}^{N_x}\sum_{j=1}^{N_y}M_{(n)i,j}^2\right]} \quad (12)$$

563   where i ranges from 1 to N_x, N_x is the number of columns in the domain and j ranges from 1 to N_y,

564   N_y is the number of rows. I_o and I_M are binary fields. $O_{(n)}(i, j)$ is the resultant field of observed

565   fractions for a square of length n. $M_{(n)}(i, j)$ is the resultant field of model forecast fractions obtained.

566   **4.2.3 Analysis of the comparative experiment results**

567   (1) Questionnaire survey of the effectiveness of model forecasting

568   Fifty-two questionnaires were completed by 32 researchers and 20 forecasters. The names of the

569   PWAFS and GRAPES models used for comparison were omitted and replaced with Model 1 and

570   Model 2, respectively.

571   The survey results show that 52 people believe that the forecasting effect of periods A and C of

572   Model 2 (GRAPES) is good, 19 people believe that the forecasting effect of period B of Model 1

573   (PWAFS) is good, and 33 people believe that the forecasting effect of period B of Model 2

574   (GRAPES) is good. Fifty-two people think that Model 2 (GRAPES) is good in general.

575   (2) Indices analysis and comparison between the two models

576   The high-resolution regional models used for evaluation are (1) PWAFS 3km and (2) GRAPES

577   3km, and the modelled precipitation is the accumulated precipitation of 24 hours in the forecast

578   12-36 hours. The evaluation results are as follows (Tables 11-16):

579   The results show that in this process, the evaluation results of different methods on the forecast

580   skill of the PWAFS and GRAPES models are basically consistent and in line with the subjective

581   evaluation statistical results. However, PAS scores have obvious advantages in the evaluation of

582   rainstorms and above, especially extreme rainstorms. It can be seen from the six rating scales that the

583   TS and FSS have almost no ability to evaluate precipitation above 250 mm, and the scores are

584   generally at the unskilled end of 0 and no more than 0.2 (Chen et al., 2019). The PAS scores can also

585   distinguish differences and provide different scores for situations where the forecasting effect is

586   good.

587   For example, when evaluating precipitation above 250 mm, the scores of TS for PWAFS in all

588   three periods are 0.000, and the scores of GRAPES in the three periods are 0.000, 0.045 and 0.044.

589 The scores of FSS (45 km) for the PWAFS in all three periods are 0.000, and the scores of GRAPES

590 in the three periods are 0.000, 0.218 and 0.137, respectively. This indicates that the TS and FSS (45

591 km) have little ability to assess the heavy rainfall of this process.

592 The PAS scores for PWAFS in the three periods are 0.229, 0.302 and 0.153, and those for

593 GRAPES in the three periods are 0.338, 0.637 and 0.528, indicating that PAS has the ability to

594 evaluate heavy rainstorms (above 250 mm) in this process. The evaluation results show that

595 GRAPES is superior to PWAFS in predicting heavy rainfall.

596 The evaluation capabilities of PAS, TS, and FSS for precipitation above 100 mm are further

597 analyzed. The scores of TS for the PWAFS (GRAPES) are 0.035, 0.257, and 0.042 (0.178, 0.451,

598 and 0.284) in the three periods, respectively. The scores of FSS (45 km) are 0.129, 0.550, and 0.103

599 (0.432, 0.767, and 0.613) for the PWAFS (GRAPES) in the three periods, respectively. The

600 evaluation effect of FSS (45 km) is better than that of TS. The evaluation feature of FSS is to

601 examine the predictability scale of the model to reflect its predictive ability; however, due to the

602 subjectivity of selecting neighbourhood scales, its score lacks comparability. While the PAS scores

603 are 0.246, 0.492 and 0.253 (0.573, 0.581 and 0.492) for the PWAFS (GRAPES) in the three periods,

604 it can be seen that the PAS also has a good ability to assess heavy rainstorms in this process.

605 In small-magnitude precipitation (above light and moderate rain) verification, the FSS scores

606 tend to approach 1 as the neighbourhood distance expands, making it difficult to compare forecast

607 differences between models. The PAS scores can also distinguish the differences in forecast

608 effectiveness for small-magnitude precipitation.

609 In conclusion, different scoring methods were used to evaluate the skill of different models to

610 predict extreme precipitation events in July 2021 in Henan, China, and the evaluation characteristics

611 of different scoring methods were indicated. The results show that the PAS scoring method has

612 obvious advantages in the evaluation of extreme precipitation events and can also reflect the

613 differences in the small magnitude precipitation forecasting effects of the models well compared to

614 those of the TS and FSS methods.

5 Discussion and conclusion

616 By analyzing the advantages and disadvantages of the traditional TS, magnitude-improved TS

617 and neighbourhood spatial verification methods, a new precipitation verification method, GCEM,

618 was designed and constructed from the perspective of the proximity of the forecast to the observation.

619 This method consists of the core indicator of the PAS, as well as multiple indicators such as IPI, EPI,

620 IEPI and PASC.

621 The PAS index consists of sine and e-exponential functions. Additionally, considering the

622 characteristics of large fluctuations in the function curves when observed precipitation is less than 10

623 mm, the formula has been smoothed for optimization. The PAS method adopts the principle of

624 "classification before verification, no classification during verification", which can serve as an

625 alternative to skill scores such as the TS and ETS for verifying quantitative precipitation forecasts.

626 This method is characterized by objective and transparent rules and easy generalization. Moreover,

627 this approach possesses attributes of an ideal precipitation scoring method, such as fairness,

628 boundedness and moderate symmetry. Therefore, it can be used to calculate the accuracy of

629 numerical models or quantitative precipitation forecasts, as well as evaluate the comprehensive

630 forecasting capabilities of various refined quantitative precipitation forecast products. The GCEM

631 can also evaluate the performance of numerical forecasts on clear/rain forecasts, as well as

632 insufficient precipitation forecasts, excessive precipitation forecasts and precipitation forecast biases.

633   In addition to the overall score, two-dimensional score distribution maps can be generated for each

634   index in the GCEM system. These maps offer a comprehensive reflection of the precipitation

635   forecasting performance of the numerical models and serve as a reference for improving model

636   forecasts.

637   This new verification method is validated based on the forecast scoring results for general

638   precipitation and precipitation at the level of torrential rain and above, and the verification results

639   align with expectations. Comparative experiments are also conducted on two typical processes using

640   the new verification method. For Case 1, the subjective judgement is relatively good, but the TS is

641   lower. Conversely, for Case 2, the subjective judgement is poorer, yet the TS is higher. Verification

642   using the PAS reveals that forecasts with better subjective judgement receive higher scores, and

643   forecasts with poorer subjective judgement receive lower scores. Therefore, PAS aligns with public

644   expectations.

645   The PAS, TS and FSS methods were used to compare and verify the "21.7" extreme

646   precipitation event in Henan, China, to reflect the evaluation characteristics of different scoring

647   methods. The results show that the PAS scoring method can not only reflect the difference in the

648   small-magnitude precipitation forecast effect of models, but also has obvious advantages in the

649   evaluation of extreme precipitation events.

650   In addition, the National Meteorological Centre of China conducted long-term series large-scale

651   sample testing on this method in 2023. Based on the ECMWF model's 24-hour and 48-hour

652   precipitation forecasts from March 2022 to February 2023, the assessment results show that

653   compared to the TS, the PAS is less affected by the randomness of the sample, and the relative size

654   relationship of different time forecast scores is more stable.

From the construction of the GCEM to ideal experiments and case analysis, 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. (6) and (7)] 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 source code and data of this work can be found at https://doi.org/10.5281/zenodo.10951799 (Zhang et al., 2024). The readme file can be found at https://doi.org/10.5281/zenodo.10951799 (Zhang et al., 2024), which includes the compiling environment and steps to repeat this work, as well as other relevant content descriptions (code, data, output files, module code main interfaces, etc.).

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.

677 *Acknowledgements.* This study was supported by grid precipitation analysis data, forecasted

678 precipitation data and numerical computing capabilities provided by the Jiangsu Provincial

679 Meteorological Bureau of China, as well as long-term series large-scale sample testing conducted at

680 the National Meteorological Centre of China.

681 *Financial support.* This research is supported by the National Key Research and Development

682 Program of China under Grant 2021YFC3000904, the Beijige Open Research Fund under Grant

683 BJG202403 and the Jiangsu Collaborative Innovation Center for Climate Change.

684

References

686 Ahijevych, D., Gilleland, E., Brown, B. G., and Ebert, E. E.: Application of spatial verification methods to

687   idealized   and   NWP-gridded   precipitation   forecasts,   Weather   Forecast.,   24,   1485–1497,

688   https://doi.org/10.1175/2009WAF2222298.1, 2009.

689 Bi, B., Dai, K., Wang, Y., Fu, J., Cao, Y., and Liu, C.: Advances in techniques of quantitative precipitation forecast,

690   Journal of Applied Meteorological Science, 27, 534–549, http://doi.org/10.11898/1001-7313.20160503, 2016.

691 Casati, B., Wilson, L. J., Stephenson, D. B., Nurmi, P., Ghelli, A., Pocernich, M., Damrath, U., Ebert, E. E., Brown,

692   B. G., and Mason, S.: Forecast verification: current status and future directions, Meteorol. Appl., 15, 3–18,

693   https://doi.org/10.1002/met.52, 2008.

694 Chen, F., Chen, J., Wei, Q., Li, J., Liu, C., Yang, D., Zhao, B., and Zhang, Z.: A new verification method for heavy

695   rainfall forecast based on predictability II: Verification method and test, Acta. Meteorol. Sin., 77, 28–42.

696   https://doi.org/10.11676/qxxb2019.003, 2019.

697 Chen, H., Li, P., and Zhao, Y.: A review and outlook of verification and evaluation of precipitation forecast at

698   convection-permitting resolution, Advances in Meteorological Science and Technology, 11, 155–164,

699   https://doi.org/10.3969/j.issn.2095-1973.2021.03.018, 2021.

700 Donaldson, R. J., Dyer, R. M., and Krauss, M. J.: An objective evaluator of techniques for predicting severe

701   weather events, Ninth Conference on Severe Local Storms, Norman, Oklahoma, United States of America,

702   American Meteorological Society [preprint], 321–326, 1975.

Doswell, C. A. III, Davies-Jones, R., and Keller, D. L.: On summary measures of skill in rare event forecasting based on contingency tables, Weather Forecast., 5, 576–585, https://doi.org/10.1175/1520-0434(1990)005<0576:OSMOSI>2.0.CO;2, 1990.

Doolittle, M. H.: The verification of predictions, Bulletin of the Philosophical Society of Washington, 7, 122–127, 1885.

Doolittle, M. H.: Association ratios, Bulletin of the Philosophical Society of Washington, 10, 83–87, 94–96, 1888.

Ebert, E. E.: Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework, Meteorol. Appl., 15, 51–64, https://doi.org/10.1002/met.25, 2008.

Finley, J. P. :Tornado predictions, American Meteorological Journal, 1, 85–88, 1884.

Flueck, J. A.: A study of some measures of forecast verification, in: 10th Conf. Probability and Statistics in Atmospheric Science, Edmonton, Alberta, Canada, American Meteorological Society [preprint], 69–73, 1987.

Gandin, L. S., and Murphy, A. H.: Equitable scores for categorical forecasts, Mon. Weather Rev., 120, 361–370, https://doi.org/10.1175/1520-0493(1992)120<0361:ESSFCF>2.0.CO;2, 1992.

Gilbert, G. K.: Finley's tornado predictions, American Meteorological Journal, 1, 166–172, 1884.

Gilleland, E., Ahijevych, D., Brown, B. G., Casati, B., and Ebert, E. E.: Intercomparison of Spatial Forecast Verification Methods. Weather Forecast., 24, 1416–1430. https://doi.org/10.1175/2009WAF2222269.1, 2009.

Gofa, F., Boucouvala, D., Louka, P., and Flocas, H.A.: Spatial verification approaches as a tool to evaluate the performance of high resolution precipitation forecasts, Atmos. Res., 208, 78–87, https://doi.org/10.1016/j.atmosres.2017.09.021, 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. Meteorol. Sin., 81, 304–315, https://doi.org/10.11676/qxxb2023.20220066, 2023.

Hanssen, A. W, and 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, Meteorological Monthly, 49, 843–854, http://doi.org/10.7519/j.issn.1000-0526.2023.050201, 2023.

Heidke, P.: Berechnung des erfolges und der güte der windstärkevorhersagen im sturmwarnungdienst, Geografika Annaler, 8, 301–349, 1926.

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

Liu, C., Dai, K., Lin, J., Wei, Q., Li N., Wang, B., Tang B., Guo Y.,Zhu W., Tang J., and Zeng X.: Design and implementation of whole process evaluation program library of weather forecast, Meteorological Monthly, 49, 351–364, http://doi.org/10.7519/j.issn.1000-0526.2022.050902, 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, Meteorological Monthly, 48, 1292-1302, http://doi.org/10.7519/j.issn.1000-0526.2022.050502, 2022b.

Marzban, C.: Scalar measures of performance in rare-event situations, Weather Forecast., 13, 753–763, https://doi.org/10.1175/1520-0434(1998)013<0753:SMOPIR>2.0.CO;2, 1998.

Mason, I. B.: Dependence of the critical success index on sample climate and threshold probability, Aust. Meteorol. Mag., 37, 75–81, 1989.

Mason, I. B.: Binary events. In forecast verification: a practitioner's guide in atmospheric science, edited by: Jolliffe IT and Stephenson DB, John Wiley and Sons, Chichester, United Kingdom, 37–76, 2003.

Murphy, A. H.: The Finley affair: a signal event in the history of forecast verification, Weather Forecast., 11, 3–20, https://doi.org/10.1175/1520-0434(1996)011<0003:TFAASE>2.0.CO;2, 1996.

Murphy, A. H., and Daan, H.: Forecast evaluation, in: Probability, statistics, and decision making in the atmospheric sciences, edited by: Murphy, A. H. and Katz, R. W., Westview Press, Boulder, Colorado, United States of America, 379–437, 1985.

Palmer, W. C., and Allen, R. A.: Note on the accuracy of forecasts concerning the rain problem, U.S. Weather Bureau, 1–4, 1949.

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

Roberts, N. M., and Lean, H. W.: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events, Mon. Weather Rev., 136, 78–97, https://doi.org/10.1175/2007MWR2123.1, 2008.

Schaefer, J. T.: The critical success index as an indicator of warning skill, Weather Forecast., 5, 570–575, https://doi.org/10.1175/1520-0434(1990)005<0570:TCSIAA>2.0.CO;2, 1990.

Shi, W., Li, X., Zeng, M., Zhang, B., Wang, H., Zhu, K., and Zhuge, X.: Multi-model comparison and high-resolution regional model forecast analysis for the "7 · 20" Zhengzhou severe heavy rain, Transactions of

Atmospheric Sciences, 44, 688–702, https://doi.org/10.13878/j.cnki.dqkxxb.20210823001, 2021.

Stephenson, D. B., Casati, B., Ferro, C. A. T., and Wilson, C. A.: The extreme dependency score: A non-vanishing measure for forecasts of rare events, Meteorol. Appl., 15, 41–50, https://doi.org/10.1002/met.53, 2008.

Su, A., Lü, X., Cui, L., Li, Z., Xi, L., and Li H.: The Basic Observational Analysis of "7.20" Extreme Rainstorm in Zhengzhou, Torrential Rain and Disasters, 40, 445–454, https://doi.org/10.3969/j.issn.1004-9045.2021.05.001, 2021.

Tang, W., Zhou, Q., Liu, X., Zhu, W., and Mao, X.: Analysis on verification of national severe convective weather categorical forecasts, Meteorological Monthly, 43, 67–76, https://doi.org/10.7519/j.issn.1000-0526.2017.01.007, 2017.

Wang, X., Yang, H., Cui, C., Li, C., Qi, H., Du, M., Wang, J., and Wang, X.: Analysis of unusual climatic characteristics of precipitation and four typical extreme weather processes in China in 2021, Torrential Rain and Disasters, 41, 489–500, https://doi.org/10.12406/byzh.2022-045, 2022.

Wei, Q., Li, W., Peng, S., Xue, F., Zhao, S., Zhang, J., and Qi, D.: Development and application of national verification system in CMA, Journal of Applied Meteorological Science, 30, 245–256. https://doi.org/10.11898/1001-7313.20190211, 2019.

Weygandt, S. S., Loughe, A. F., Benjamin, S. G., and Mahoney, J. L.: Scale sensitivities in model precipitation skill scores during IHOP, in: 22nd Conference Severe Local Storms, Hyannis, Massachusetts, United States of America, American Meteorological Society, 4–8 October 2004.

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences (2nd ed), Academic Press, 630 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, 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, in: Proceedings of the 34th Annual Meeting of the Chinese Meteorological Society, S1 Disaster Weather Monitoring, Analysis and Forecast, 549–559, 2017.

Yates, E., Anquetin, S., Ducrocq, V., Creutin, J.-D., Ricard, D., and Chancibault, K.: Point and areal validation of forecast precipitation fields, Meteorol. Appl., 13: 1–20, https://doi.org/10.1017/S1350482705001921, 2006.

Zepeda-Arce. J., Foufoula-Georgiou, E., and Droegemeier, K. K.: Space-time rainfall organization and its role in validating quantitative precipitation forecasts, J. Geophys. Res., 105, 10129–10146,

790   https://doi.org/10.1029/1999JD901087, 2000.

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

792   for "A general comprehensive evaluation method for cross-scale precipitation forecasts", Zenodo [code and

793   data set], https://doi.org/10.5281/zenodo.10951799, 2024.

794 Zhang, Y., Yu, H., Zhang, M., Yang, Y., and Meng, Z.: Uncertainties and error growth in forecasting the

795   record-breaking rainfall in Zhengzhou, Henan on 19–20 July 2021, Science China Earth Sciences, 65,

796   1903–1920, https://doi.org/10.1007/s11430-022-9991-4, 2022.

797 Zhao, B., and Zhang, B.: Application of neighborhood spatial verification method on precipitation evaluation,

798   Torrential Rain and Disasters, 37, 1–7, https://doi.org/10.3969/j.issn.1004-9045.2018.01.001, 2018.

799 Zhong, Q., Sun, Z., Chen, H., Li, J., and Shen, L.: Multi model forecast biases of the diurnal variations of intense

800   rainfall in the Beijing-Tianjin-Hebei region, Science China Earth Sciences, 65, 1490−1509,

801   https://doi.org/10.1007/s11430-021-9905-4, 2022.

802

803

804

805

806 **Table 1.** Gradient decrease scoring table for station-by-station (time) rainstorm forecasts. The values are

807 normalized, i.e., score = original data/100.

Observation	Forecast (mm)			
(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

808

809

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

810

811

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

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

812

813

814

815

816

817

818

819

820

821

822

823

Table 4. Classification of PAS for 12-hour accumulated precipitation.

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

824

825

Table 5. Classification of PAS for 24-hour accumulated precipitation.

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

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-hour accumulated precipitation from 00:00 to 12:00 UTC on 16 July 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 from 00:00 to 12:00 UTC on 13 June 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

835

836 **Table 10.** Accuracy indices of insufficient precipitation forecast (IPI), excessive precipitation forecast (EPI) and

837 insufficient and excessive precipitation forecast (IEPI) of 12-hour accumulated precipitation for two precipitation

838 processes.

	IPI	EPI	IEPI
Case 1	-0.376	0.389	0.057
Case 2	-0.400	0.597	0.325

839

840 **Table 11.** PAS, TS and FSS scores of PWAFS 24-hour accumulated precipitation from 00:00 UTC on 19 July to

841 00:00 UTC on 20 July 2021.

	clear/rainy	≥0.1 mm	≥10 mm	≥25 mm	≥50 mm	≥100 mm	≥250 mm
PAS	0.598	0.487	0.301	0.256	0.254	0.246	0.229
TS	0.823	0.774	0.377	0.229	0.115	0.035	0.000
FSS(15 km)	---	0.909	0.637	0.452	0.259	0.090	0.000
FSS(25 km)	---	0.923	0.680	0.486	0.281	0.102	0.000
FSS(45 km)	---	0.939	0.732	0.526	0.307	0.129	0.000
FSS(75 km)	---	0.953	0.778	0.559	0.335	0.180	0.003
FSS(120 km)	---	0.964	0.820	0.592	0.365	0.226	0.007

842

843 **Table 12.** PAS, TS and FSS scores of PWAFS 24-hour accumulated precipitation from 00:00 UTC on 20 July to

844 00:00 UTC on 21 July 2021.

	clear/rainy	≥0.1 mm	≥10 mm	≥25 mm	≥50 mm	≥100 mm	≥250 mm
PAS	0.653	0.578	0.408	0.398	0.427	0.492	0.302
TS	0.789	0.743	0.500	0.429	0.434	0.257	0.000
FSS(15 km)	---	0.891	0.750	0.690	0.687	0.475	0.000
FSS(25 km)	---	0.908	0.789	0.731	0.725	0.507	0.000
FSS(45 km)	---	0.928	0.837	0.782	0.771	0.550	0.000
FSS(75 km)	---	0.945	0.878	0.831	0.815	0.598	0.003
FSS(120 km)	---	0.958	0.912	0.877	0.858	0.654	0.042

845

846

847 **Table 13.** PAS, TS and FSS scores of PWAFS 24-hour accumulated precipitation from 00:00 UTC on 21 July to

848 00:00 UTC on 22 July 2021.

	clear/rainy	≥0.1 mm	≥10 mm	≥25 mm	≥50 mm	≥100 mm	≥250 mm
PAS	0.656	0.533	0.346	0.322	0.296	0.253	0.153
TS	0.802	0.731	0.469	0.318	0.169	0.042	0.000
FSS(15 km)	---	0.887	0.714	0.563	0.352	0.093	0.000
FSS(25 km)	---	0.905	0.747	0.599	0.381	0.096	0.000
FSS(45 km)	---	0.924	0.784	0.644	0.414	0.103	0.000
FSS(75 km)	---	0.940	0.813	0.685	0.443	0.111	0.000
FSS(120 km)	---	0.952	0.840	0.723	0.474	0.120	0.000

849

850 **Table 14.** PAS, TS and FSS scores of GRAPES 24-hour accumulated precipitation from 00:00 UTC on 19 July to

851 00:00 UTC on 20 July 2021.

	clear/rainy	≥0.1 mm	≥10 mm	≥25 mm	≥50 mm	≥100 mm	≥250 mm
PAS	0.665	0.549	0.396	0.414	0.494	0.573	0.338
TS	0.804	0.735	0.422	0.358	0.312	0.178	0.000
FSS(15 km)	---	0.884	0.689	0.629	0.576	0.365	0.000
FSS(25 km)	---	0.901	0.742	0.688	0.633	0.400	0.000
FSS(45 km)	---	0.922	0.809	0.759	0.704	0.432	0.000
FSS(75 km)	---	0.939	0.865	0.817	0.758	0.457	0.000
FSS(120 km)	---	0.950	0.907	0.862	0.786	0.494	0.000

852

853 **Table 15.** PAS, TS and FSS scores of GRAPES 24-hour accumulated precipitation from 00:00 UTC on 20 July to

854 00:00 UTC on 21 July 2021.

	clear/rainy	≥0.1 mm	≥10 mm	≥25 mm	≥50 mm	≥100 mm	≥250 mm
PAS	0.669	0.580	0.438	0.451	0.504	0.581	0.637
TS	0.801	0.746	0.544	0.438	0.431	0.451	0.045
FSS(15 km)	---	0.891	0.774	0.693	0.683	0.687	0.127
FSS(25 km)	---	0.909	0.808	0.737	0.727	0.721	0.167
FSS(45 km)	---	0.930	0.850	0.793	0.787	0.767	0.218
FSS(75 km)	---	0.947	0.884	0.843	0.847	0.818	0.233
FSS(120 km)	---	0.960	0.913	0.885	0.897	0.864	0.238

855

856

857

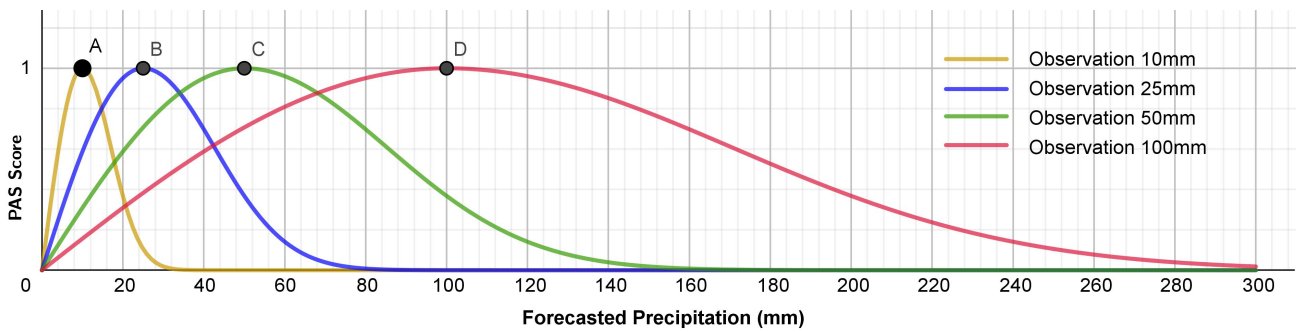
858 **Table 16.** PAS, TS and FSS scores of GRAPES 24-hour accumulated precipitation from 00:00 UTC on 21 July to

859 00:00 UTC on 22 July 2021.

	clear/rainy	≥0.1 mm	≥10 mm	≥25 mm	≥50 mm	≥100 mm	≥250 mm
PAS	0.694	0.566	0.407	0.425	0.462	0.492	0.528
TS	0.796	0.710	0.559	0.501	0.410	0.284	0.044
FSS(15 km)	---	0.875	0.799	0.752	0.667	0.508	0.092
FSS(25 km)	---	0.897	0.842	0.793	0.713	0.548	0.102
FSS(45 km)	---	0.924	0.889	0.842	0.772	0.613	0.137
FSS(75 km)	---	0.945	0.925	0.883	0.823	0.690	0.192
FSS(120 km)	---	0.960	0.949	0.911	0.858	0.757	0.257

860

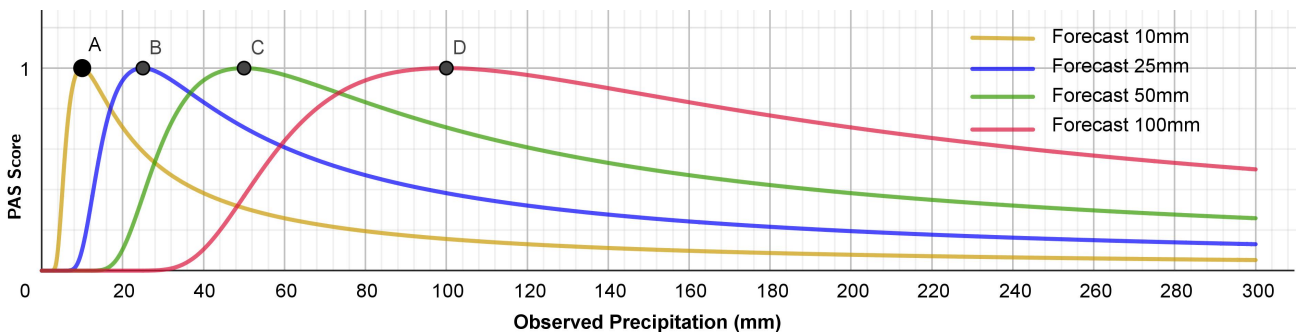
861

862

863

864 **Figure 1.** Schematic diagram of the precipitation forecast accuracy score (PAS) curves when the observed

865 precipitation amounts are 10, 25, 50 and 100 mm.

866

867

868

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

870

871

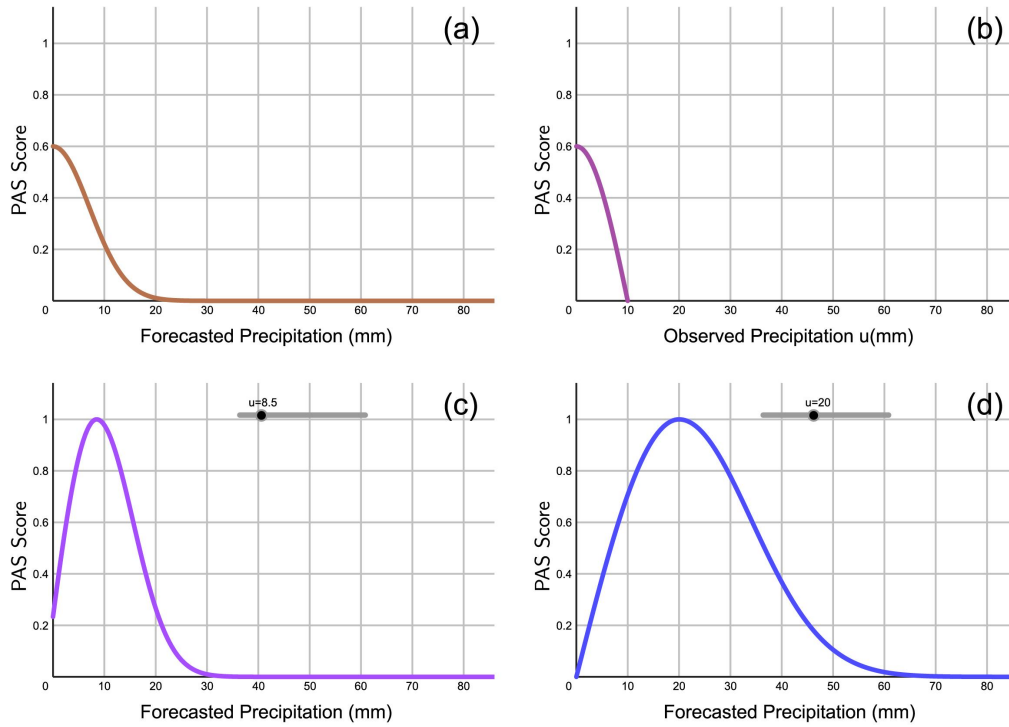



Figure 3. PAS curves of precipitation forecasts when (a) the observed precipitation u = 0 mm and the forecasted

precipitation x > 0 mm, (b) the observed precipitation 0 < 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

the forecasted precipitation x > 0 mm, and (d) the observed precipitation u ≥ 10 mm.

878

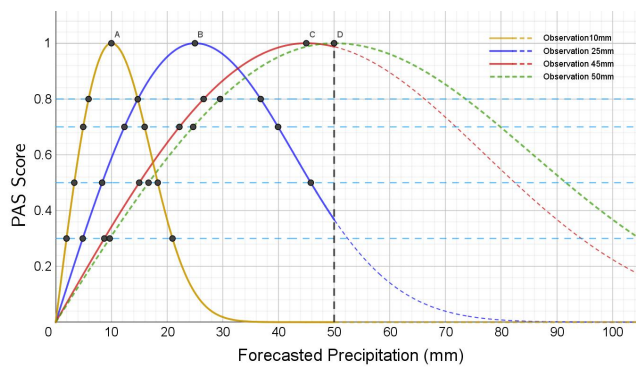



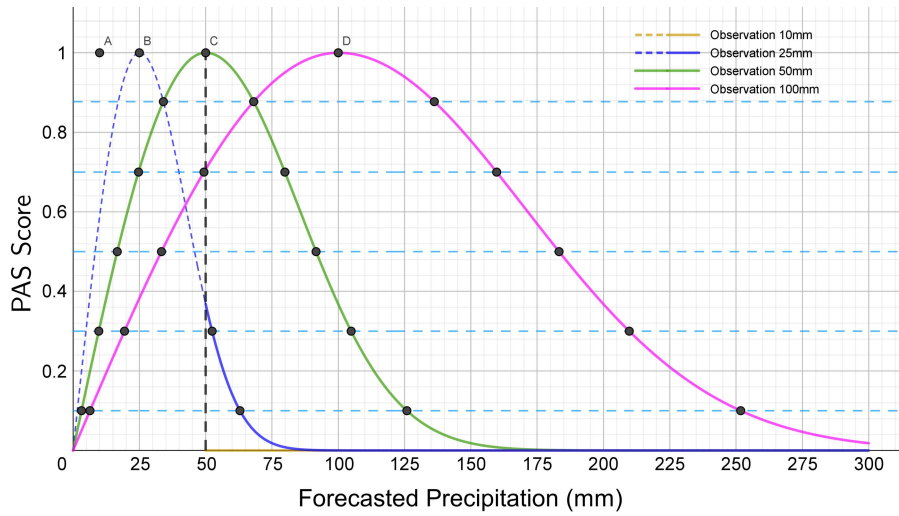
Figure 4. PAS curves of the forecasts under general precipitation conditions (u = 10, 25 and 45 mm). The solid line

part of the curve in the figure is involved in the comparison, the dashed line part is not involved in the comparison,

10 mm observed precipitation is represented by the orange line, 25 mm observed precipitation is represented by the

blue line, 45 mm observed precipitation is represented by the red line, and 50 mm observed precipitation is

represented by the green line.

884

885

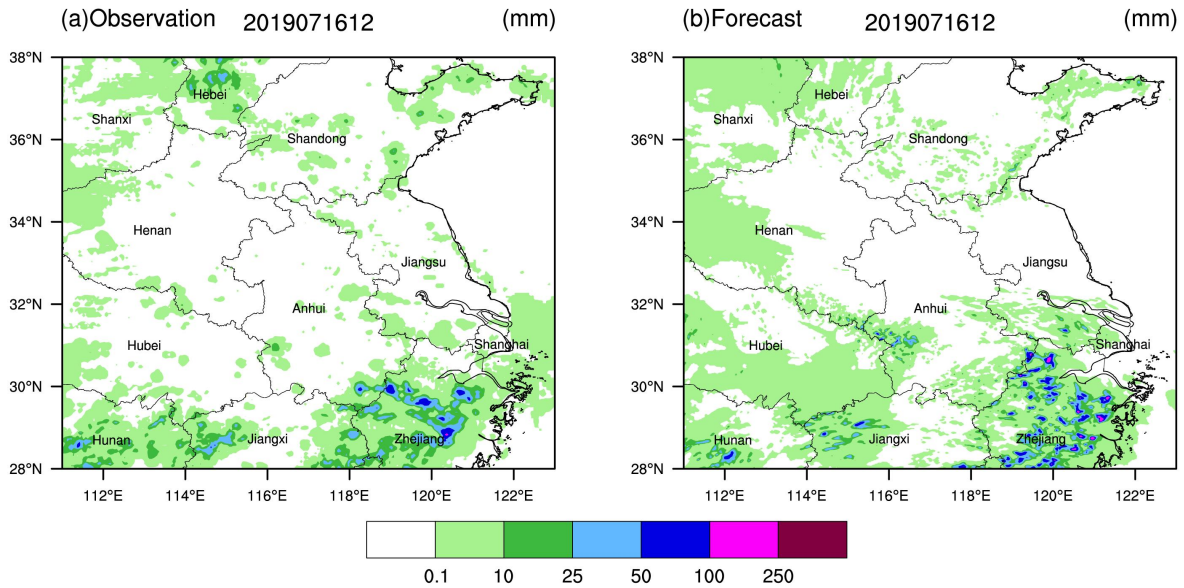
886 **Figure 5.** Same as Fig. 4, but for precipitation at the level of torrential rain and above (u = 25, 50 and 100 mm).

887 The solid line part of the curve in the figure is involved in the comparison, the dashed line part is not involved in

888 the comparison, 10 mm observed precipitation is represented by the orange line, 25 mm observed precipitation is

889 represented by the blue line, 50 mm observed precipitation is represented by the green line, and 100 mm observed

890 precipitation is represented by the red line.

891

892

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

894

895

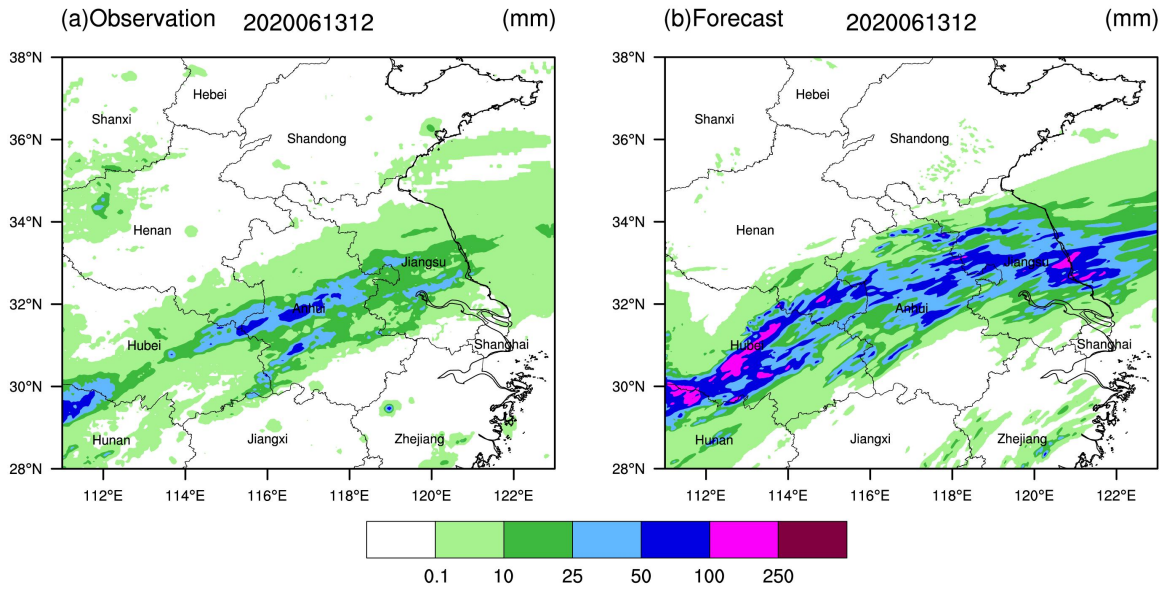



Figure 7. Accumulated precipitation (a) observed and (b) forecasted from 00:00 to 12:00 UTC on 13 June 2020.

897

898

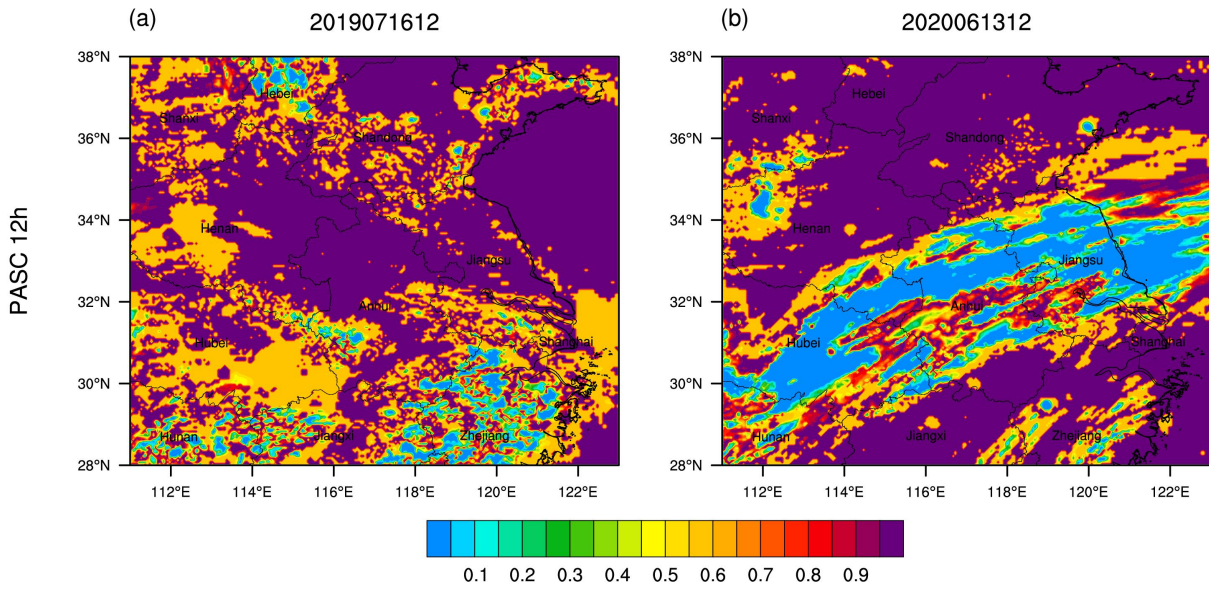



Figure 8. Distributions of the PAS clear/rainy forecast accuracy score (PASC) of 12-hour accumulated precipitation

for (a) Case 1 from 00:00 to 12:00 UTC on 16 July 2019 and (b) Case 2 from 00:00 to 12:00 UTC on 13 June 2020.

901

902

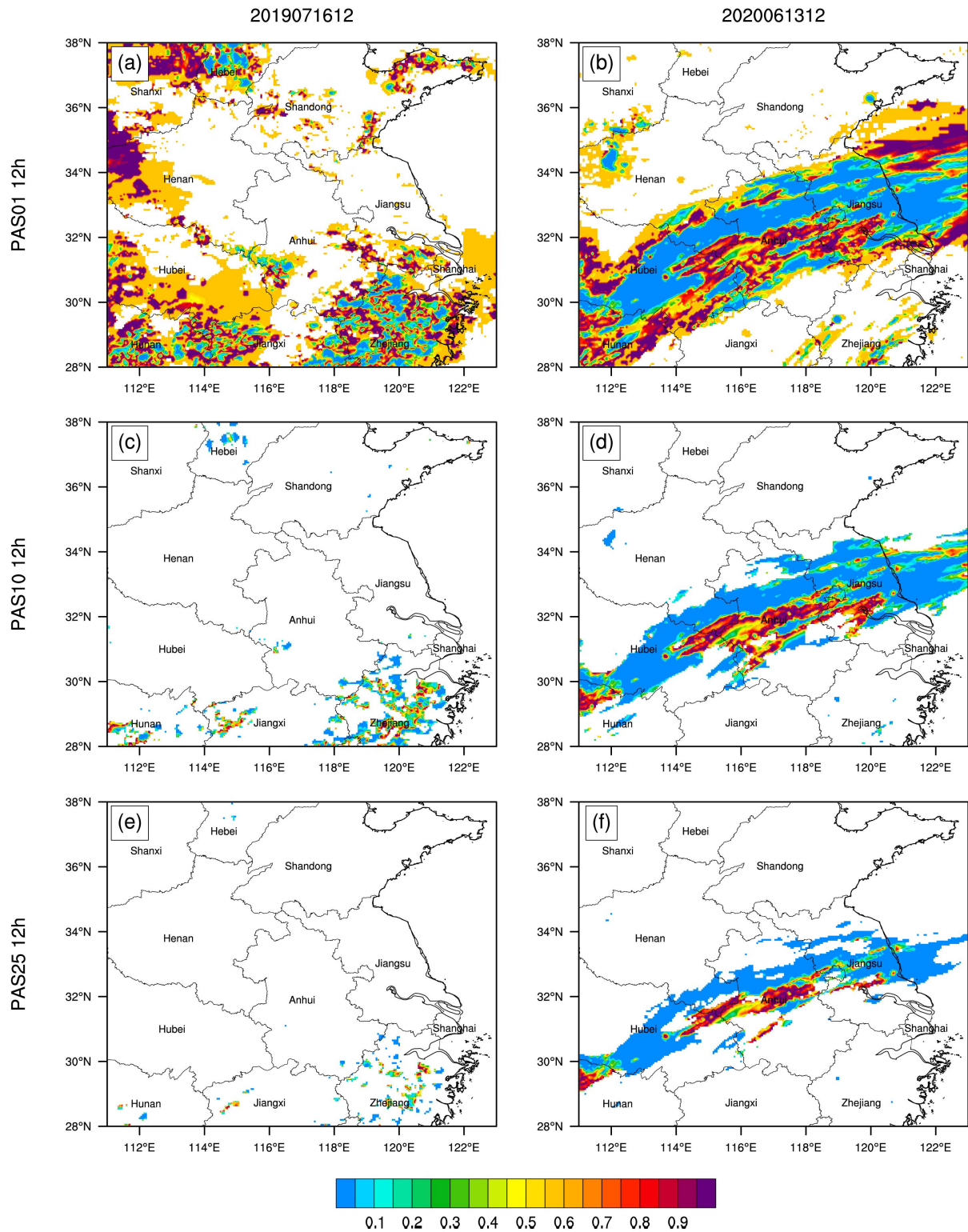
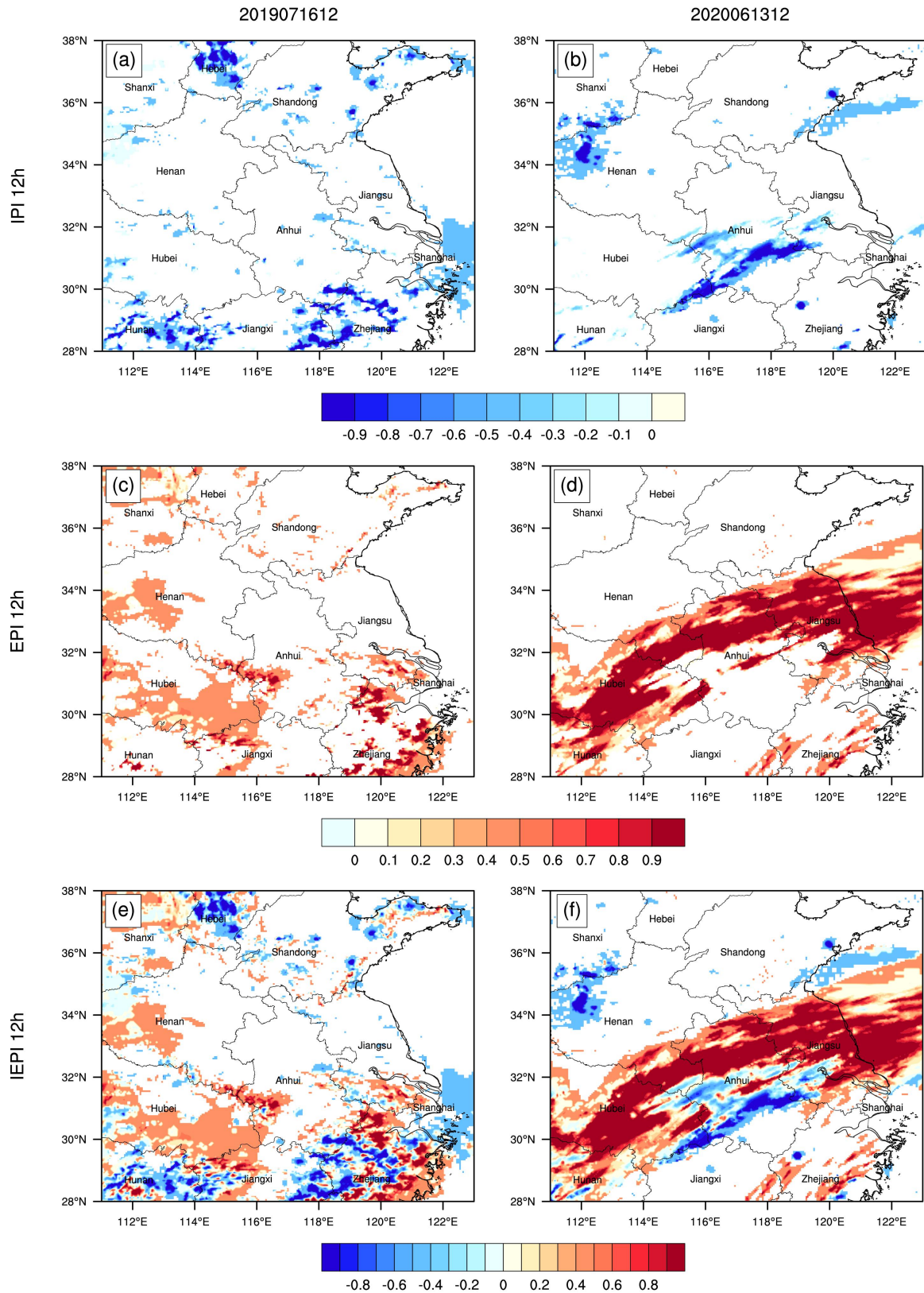



Figure 9. Distributions of PAS of 12-hour accumulated precipitation, ≥ 0.1 mm for (a) Case 1 from 00:00 to 12:00

UTC on 16 July 2019 and (b) Case 2 from 00:00 to 12:00 UTC on 13 June 2020, ≥10 mm for (c) Case 1 and (d)

Case 2, and ≥ 25 mm for (e) Case 1 and (f) Case 2.

906

907

908 **Figure 10.** Distributions of IPI of 12-hour accumulated precipitation for (a) Case 1 from 00:00 to 12:00 UTC on

909 16 July 2019, and (b) Case 2 from 00:00 to 12:00 UTC on 13 June 2020, EPI for (c) Case 1 and (d) Case 2, and

910 IEPI for (e) Case 1 and (f) Case 2.

911

912

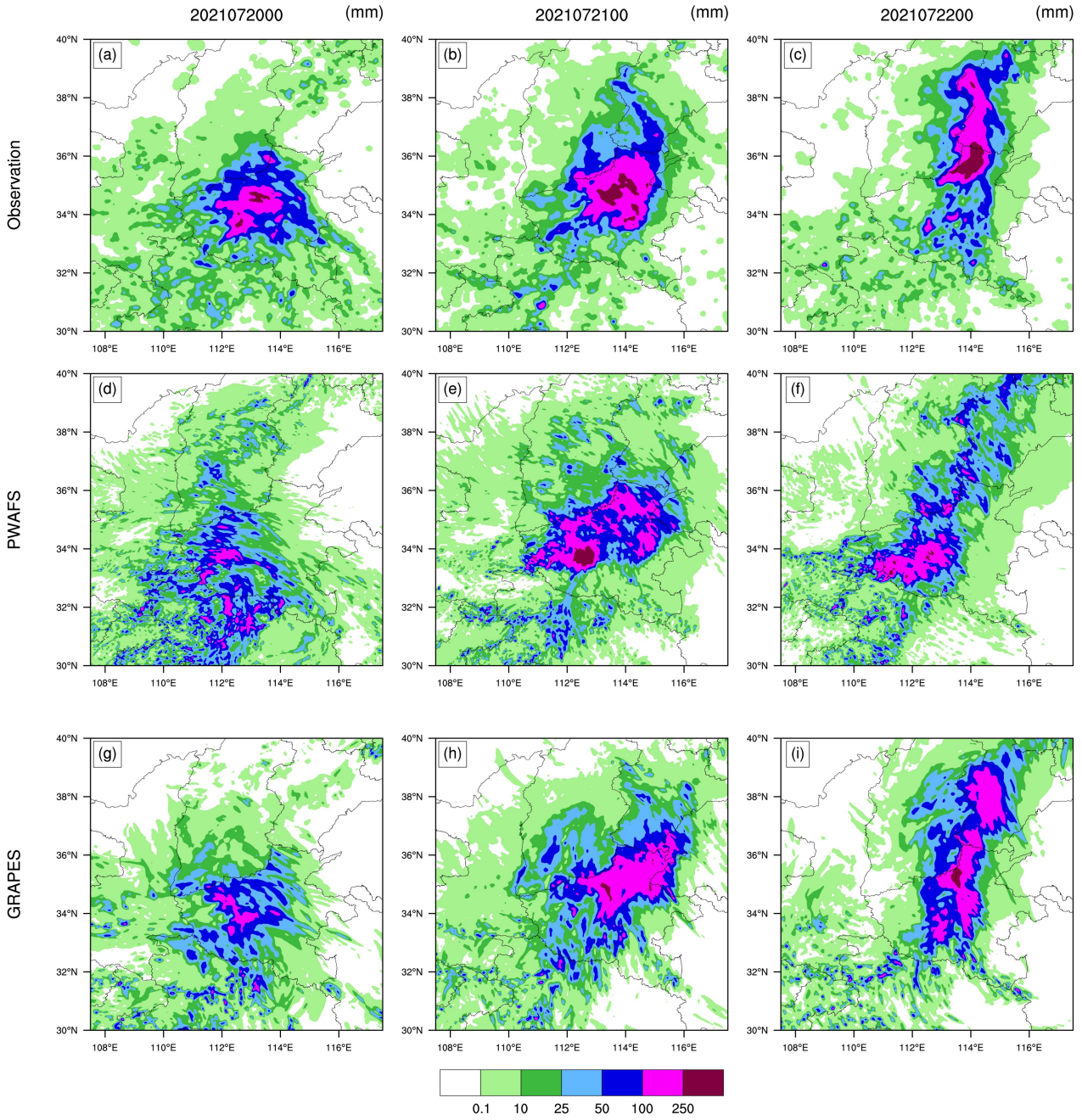



Figure 11. Distribution of observed and forecasted 24-hour accumulated precipitation. (a) Observation, (d) PWAFS,

(g) GRAPES from 00:00 UTC on 19 July to 00:00 UTC on 20 July 2021; (b) Observation, (e) PWAFS, (h)

GRAPES from 00:00 UTC on 20 July to 00:00 UTC on 21 July 2021; (c) Observation, (f) PWAFS, (i) GRAPES

from 00:00 UTC on 21 July to 00:00 UTC on 22 July 2021.

917