Overall comments:

The article titled "A two-step merging strategy for incorporating multi-source precipitation products and gauge observations using machine learning classification and regression over China" assesses the effectiveness of three machine learning-based algorithms to merge precipitation products over China. The article is very well written, concise, relevant, and I enjoyed reading it. I believe that it fulfills the requirements to be published in HESS. In the following points, I describe some points and suggestions that I think can be considered to increase the quality of the manuscript.

Response: Thanks very much for your valuable and meaningful suggestions on our manuscript. These comments significantly improve the quality of this manuscript. We have tried our best to carefully study the suggestions you raised and made corresponding modifications to the manuscript. The grammar of the manuscript has been polished by native English speakers. The responses to the reviewer's comments are as follows:

1- ERA5 has already superseded ERA-Interim; therefore, I recommend using ERA5 instead. Similarly, the authors used TMPA 3B42 (with IMERG), which is no longer in production. For reproducibility purposes, it is essential to use products that can still be acquired.

Response: This is a nice suggestion considering future product updates. As you said, ERA5 has already superseded ERA-Interim and IMERG superseded TMPA. IMERG inherits the advantages of TRMM and makes many improvements. Therefore, we remove TRMM and keep IMERG. The previous study (Xu et al., 2022) demonstrated that the overall performance of ERA5-Land is superior to ERA5 at the daily scale, the ERA5-Land is used to replace ERA-Interim. Hence, six precipitation products are used in this study, including IMERG, GSMaP, CHIRPS, CMORPH, PERSIANN-CDR, and ERA5-Land.

Xu, J., Ma, Z., Yan, S., Peng, J., 2022: Do ERA5 and ERA5-land precipitation estimates outperform satellite-based precipitation products? A comprehensive comparison between state-of-the-art model-based and satellite-based precipitation products over mainland China. Journal of Hydrology, 605, 127353.

2- I suggest including a figure showing the number and location of stations used in the GPCC product in the appendix or supplement material. This information will be beneficial for the readers.

Response: The figure showing the number and location of stations used in the GPCC product have

been added in the revised manuscript in the appendix, and shown as follows:

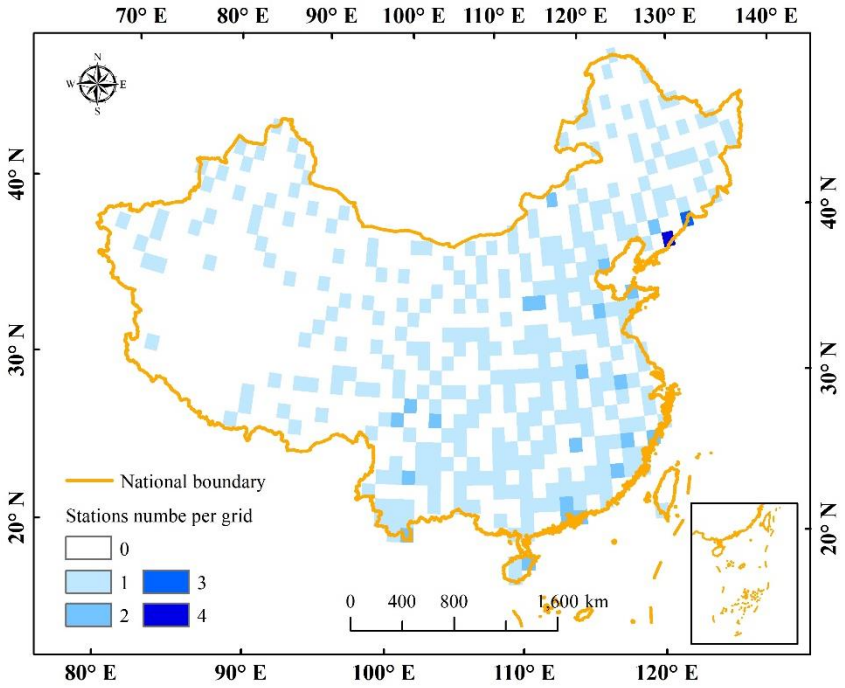

Fig. S1 The number and location of stations used in GPCC over China

From the latest GPCC dataset, the number of China's International Exchange Stations used in GPCC has fluctuated between 360-370 (In Fig. S1, the number is 362 July 2015), which has increased in recent years. Before 2017, only about 200 China's stations are used in GPCC (Tang et al., 2016). Despite the use of these stations, satellite precipitation products are corrected based on monthly GPCC, making it insufficient to improve daily performance.

Tang, G., Ma, Y., Long, D., Zhong, L., Hong, Y., 2016: Evaluation of GPM Day-1 IMERG and TMPA Version-7 legacy products over Mainland China at multiple spatiotemporal scales, Journal of hydrology, 533, 152-167.

3- It would be helpful to describe whether the ground-based data were quality controlled.

Response: The ground-based data were quality controlled and the corresponding description are added in the revised manuscript.

4- The readers would benefit substantially from an improved description of the methodology. I liked Section 3.1 as it is very clear and informative, as well as the description of the machine learning algorithms and their respective figures. However, I believe that the description of the two-step merging strategy (Section 3.2) can be further improved. The authors mention that first, the gauge

observations are classified into wet and dry days to be used as the benchmark for classification models. Can the authors explain what the reason behind this is? It is not clear how the dry days were separated from wet days. How is a day classified as dry? Is the classification performed by grid-cell and by day? What is the final result of this classification? How can data scarcity affect this classification? Solving these questions is crucial for the understanding of the manuscript. Additionally, the authors mention that a regression process was applied to the wet days for the cold and warm periods. How were the models trained independently for the warm and cold periods? Perhaps it would be helpful to do a diagram for the first and second steps where the process is clearly explained. This will increase the manuscript's impact and is essential as the article presents these novel merging procedures.

Response: Thank you for your encouragement and suggestions. Your suggestions and questions are answered point by point as follows:

(1) Q: The authors mention that first, the gauge observations are classified into wet and dry days to be used as the benchmark for classification models. Can the authors explain what the reason behind this is?

A: The biases of precipitation products mainly come from overestimating/underestimating the amounts of hit events, and failing to correctly distinguish precipitation occurrence, including false alarm and missed events. It is difficult to reduce all biases by directly correcting the precipitation amount of all samples. Correctly judging whether precipitation occurs is an important way to improve precipitation detection efficiency. Therefore, the purpose of the first step is to classify precipitation to reduce the missed and false alarmed bias.

(2) Q: It is not clear how the dry days were separated from wet days. (a) How is a day classified as dry? (b) Is the classification performed by grid-cell and by day? (c) What is the final result of this classification?

A: **(a)** The gauge observations are distinguished to wet/dry days according to the 0.1mm/d threshold value (Lei, et al., 2020; Yu et al., 22020; Jiang et al., 2021) and used as the benchmark for classification, the wet day is set as 1, dry day is set as 0. The feature values of MSPs and covariables corresponding to each grid are applied to construct XGBoost, GBDT, and RF classification models, respectively. The model determines whether a day in the grid is a wet day or a dry day according to the classification probability, i.e., if the probability of a day >0.5, it is

judged as a wet day and the output value is 1. Otherwise, it is judged as a dry day and the return value is 0. **(b)** The classification is performed by grid-cell in time series. The orange dotted line in Fig. R1 shows the process of model applying grid by gird. **(c)** Hence, the classification result contains only wet and dry days (0,1) of each grid and does not involve precipitation intensity.

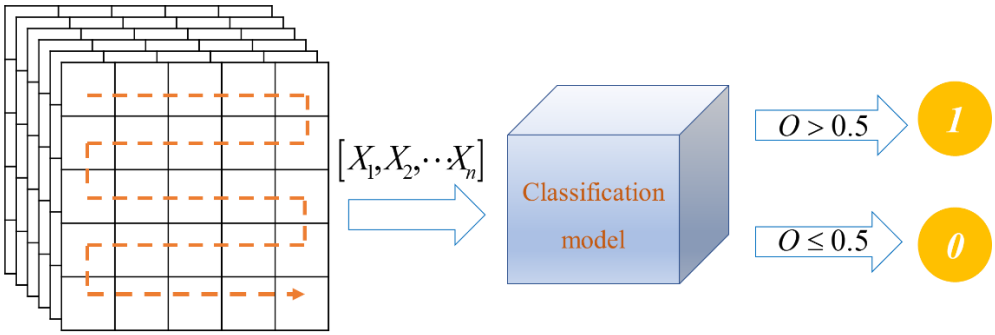

Fig. R1 the flowchart of classification

(3) How can data scarcity affect this classification?

The influence of data density (10%, 30%, 50%,70% of the total gauges) on classification result has been discussed in discussion part.

(4) Additionally, the authors mention that a regression process was applied to the wet days for the cold and warm periods. How were the models trained independently for the warm and cold periods? Perhaps it would be helpful to do a diagram for the first and second steps where the process is clearly explained.

According to the annual distribution characteristics of precipitation, we group all input datasets into two seasons: warm season (May and October) and cold season (November to April). The model is constructed and trained in warm and cold seasons using divided independent datasets, which leads to six classification and six regression models (i.e., two seasons with three models for classification and regression). Fig. 2 in the manuscript has illustrated the overall flowchart of merging strategy including the first and second steps, I hope that my detailed description can make reviewer and readers clearly understand the method of this study. The description of the methodology has also improved in the revised manuscript.

5- I liked the idea of using the semivariogram as a spatial autocorrelation variable :). Can the authors discuss the influence that the selection of a particular semivariogram model can have in applying

the method?

Response: Thank you for your suggestions. The widely used semivariogram models include: spherical, exponential, Gaussian, power, and linear. We have discussed the different of the Kriging_based prediction (KP) based on five semivariogram models. The expresses of five models as follows:

(1) Spherical model:

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ C_0 + C\left(\frac{3}{2}\cdot\frac{b}{a} - \frac{1}{2}\cdot\frac{b^3}{a^3}\right) & 0 < h \le a \\ C_0 + C & h > a \end{cases} \quad \text{(S1)}$$

(2) Exponential model:

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ C_0 + C\left(1 - \exp\left(\frac{-h}{r}\right)\right) & h > 0 \end{cases} \quad \text{(S2)}$$

where $\gamma(h)$ is semivariogram, h is the distance, C_0, C, and a is the nugget, sill, and range, respectively.

(3) Gaussian model:

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ C_0 + C\left(1 - \exp\left(\frac{h^2}{r^2}\right)\right) & h > 0 \end{cases} \quad \text{(S3)}$$

where the range is $\sqrt[2]{3}a$

(4) Power model:

$$\gamma(h) = h^a \qquad 0 < a \le 2 \quad \text{(S4)}$$

(5) Linear model:

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ C_0 + C\left(\frac{h}{a}\right) & 0 < h \le a \\ C_0 + C & h > a \end{cases} \quad \text{(S5)}$$

In order to compared the performance of the five semivariogram models, the Kriging_based predictions (KP) of total 2372 gauges are estimated and validated. The accuracy of KP will directly influence the model training and merging results. The evaluated results of different model are show in Table R1.

Table R1 The performance of KPs estimated from five models

Metrics	Spherical	Exponential	Gaussian	Power	Linear
CC	0.806	**0.810**	0.782	0.799	0.803
RMSE	4.530	**4.486**	4.862	4.625	4.582

RB	0.028	0.032	0.044	0.040	**0.006**
FAR	0.276	0.284	**0.269**	0.302	0.282
POD	0.931	**0.943**	0.895	0.942	0.937
CSI	**0.688**	0.687	0.674	0.670	0.685
KGE	**0.692**	0.685	0.684	0.661	0.675
β	1.028	1.032	1.044	1.040	**1.006**
γ	0.830	0.816	**0.876**	0.798	0.814
precision	**0.724**	0.716	0.731	0.698	0.718
HSS	**0.708**	0.706	0.696	0.686	0.705

Note: the values in bold represent the best performing values.

It can be seen from Table R1 that the overall performance of five models is good. The performance of spherical model shows the best CC, RMSE, and RB. The exponential model shows the best CSI, KGE, *precision*, HSS. Therefore, there is slightly different between spherical and exponential models. While the latter three models (Gussian, Power, and Linear) are inferior to the former two. KP is an important variable in precipitation merging, its performance will directly affect the training accuracy of the ML model. However, the difference of semivariogram models is relatively small and the spherical model with slight better performance is adopted in this study. Therefore, the influence of different models on KP can reflect its influence in the merging method application.

6- For reproducibility purposes, please mention the parameters of the RF, GBDT, and XGBoost that were used while training the models during the warm and cold period.

Response: The optimal parameters of the RF, GBDT, and XGBoost classification and regression models during the warm and cold periods are displayed in the appendix.

7- The authors evaluated the performance of two categorical indices (CSI and HSS) over different precipitation intensities. This is a very good idea as the detection of no-precipitation events mainly masks the categorical performance. In this sense, as the objective of this article is to assess and compare the effectiveness of merging precipitation products using different ML techniques, I

suggest evaluating all categorical metrics over these precipitation intensities. This separation into rain intensities will provide additional insights regarding the performance of these merged products.

Response: Thank you for your suggestion. We have added all categorical metrics (POD, FAR, CSI, precision, FB, and HSS) over these precipitation intensities in the revised manuscript.

8- The first two sections of the Results are a bit puzzling for me. In Section 4.1, "Performance assessment for classification results", the results obtained during the first step of the merging procedure are shown. However, in this section, the authors evaluate precipitation intensities (see Fig 6), which I believe, according to the methodology, are the result of applying the regression models over the wet days. Later on, in Section 4.2, "Performance assessment for regression results", the authors mention that regression models predict the final results presented in this section. By improving the explanation in the methodology, these two sections will be much clearer, and this issue can be solved.

Response: Thank you for your nice suggest. This is indeed a problem that cannot be ignored, we have adjusted the presentation of the results. The results show the performance of the final three merged precipitation (MSMPs: PXGB2, PGBDT2, and PRF2) from different aspects rather than exhibit the performance of classification and regression results separately. We evaluate the MSMPs from the evaluation of precipitation detection ability and precipitation intensity. The title of the Section 4.1 and 4.2 is revised to 4.1 "**Evaluation the precipitation detection ability of MSMPs**" and 4.2 "**Evaluation the precipitation amounts of MSMPs**". In this way, the advantage of classification and regression models could be well explored and the structure of this section is clearer than before. Meanwhile, we also improving the explanation in the methodology in the revised manuscript.

11- L355: significantly is a statistical loaded term, which must be accompanied by its respective p-value. If the p-value is not provided, I suggest using the word "substantially" instead. Please apply this throughout the manuscript.

Response: Thanks. We have applied "substantially" to replaced "significantly" throughout the manuscript.

12- L395: The authors mention the following "...although Kriging exhibits better performance than original MSPs, its accuracy is strongly dependent on gauge density. This only gauge-based interpolation method would have limited in complex mountainous areas with few gauges." I would be cautious with this statement. Although I agree with it, the ML algorithms cannot predict values outside of their training range, which could be translated into plausible underestimating precipitation over high elevations. Additionally, these techniques are also affected by the size of the training sample. Therefore, the ML techniques, in a sense, have limited performance in complex mountainous areas with few gauges as well.

Response: I very agree with your suggestion. In order to avoid misunderstanding and inappropriate statement. We have reorganized the sentences as follows and revised in manuscript:

although Kriging exhibits better performance than original MSPs, it is only based on gauge observations and does not combine other climate variables associated with precipitation. When MSPs, gauge, and multiple covariates are considered, the MSMPs are more accurate than Kriging.

13- L400: CC or r is also a component of the KGE. Also, see L412. I suggest including it inside of the KGE.

Response: Thank you for your suggestion. We have put CC in the component of the KGE and revised the corresponding statement in manuscript.

14- Figure7: Nice Figure! I think that it should be KGE instead of KEG.

Response: We have revised the KEG to KGE in Figure 7 in the revised manuscript.