Dear reviewer:

Thanks a lot for your great efforts to read through this paper again and give very valuable comments. Here we have addressed the comments from you and the detailed description is attached to this document.

Best regards,

Qian Zhu, Xiaodong Qin, Dongyang Zhou, Tiantian Yang, Xinyi Song

Point 1: Earlier I wrote "on. The term 'flood event' has not been explicitly defined." The concern remains unaddressed. The authors write "… 17, where eleven historical flood events are selected with flood peak exceeding the threshold of 8,600 m3/s in this study." Again, what is an event? Where does it start and where it ends? It appears the authors follow some subjective criteria to select the events, which are not elaborated.

Response1: Thank you very much for your comment. The sentence has been re-edited: Page 5 Line 154-157 'Fig. 2 shows the time series of the hourly streamflow and corresponding gauge-based precipitation between 2015 and 2017, where eleven historical flood events are selected in this study. The flood events are the streamflow time series with one-month span whose peak flow exceeded 8600m^3/s, corresponding to 97th approximately the quantile level (Zhu et al., 2020a).'

References:

Zhu, Q., Zhou, D., Luo, Y., Xu, Y.-P., Wang, G., and Gao, X.: Suitability of high-temporal satellite-based precipitation products in flood simulation over a humid region of China, Hydrological Sciences Journal, 66, 104-117, 10.1080/02626667.2020.1844206, 2020a.

Point 2: I am not comfortable with the overly simplistic conclusion that LSTM is better than HBV. The results provide a much more nuanced picture. Figure 6 NSE plots: HBV is more consistent across the data products compare to LSTM. For

instance, LSTM's 25th percentile is much lower compared to HBV's for IMERG-L. A statement like "LSTM has a higher likelihood of success" would be much more acceptable.

Response2:

Thank you very much for your suggestion and we agree with it. We add some discussion to compare the performance of LSTM and HBV in Figure 10 besides Figure 6, and the corresponding sentence has been re-edited:

Page 18-19 Line 535-538: 'The comparisons of SWAT, DHSVM and LSTM at different spatial resolutions are also illustrated. As a data-driven approach, LSTM shows better performance than SWAT and DHSVM in terms of flood events simulation and shows reduced uncertainty and a higher likelihood of success than HBV, which is considered an appropriate model in this case.'

Point 3: I reiterate my earlier statement: there is no surprise that model performance improved for flood events after calibrating exclusively for flood events. It is widely acknowledged that calibration with respect to a specific objective function leads to its improvement. As I said earlier, I would be surprised to see overall NSE (NSE for the whole time series) improving after calibration with respect to flood events.

Response3:

Thank you for this comment, and we calculate the NSE to see how the overall NSE performs for the whole time series. The results are presented in the following table. According to the Table 1, it can be seen that generally the calibration strategy II shows better performance than the strategy I in overall NSE, for the mean NSE values of HBV, SWAT, DHSVM, LSTM increase from 0.79, 0.77, 0.80, and 0.88 with calibration strategy I to 0.80, 0.80, 0.86, 0.89 with calibration strategy II. This is probably due to the fact that the NSE is more sensitive to changes in flood peaks (Huang et al., 2019) and calibration strategy II can better capture the flood peaks compared to calibration strategy I. Based on your comment, we have added some discussion in ***"5.1 Comparison of two different***

calibration strategies", and the details are as follows:

Page 14 Line 425-431: 'Although we targeted in difference between the two strategies in flood events simulation, their performances in the whole streamflow simulation time series are also compared, which is presented in Table 1 (The mean value is the average NSE of the four precipitation products with the same calibration strategy). According to the mean NSE values, calibration strategy II outperforms calibration strategy I. To be specific, for HBV, SWAT, DHSVM and LSTM models, among the four precipitation products, there are two, three, three and three NSE values larger with calibration strategy II than that with calibration strategy I.'

Table. 1 The NSE values of the whole streamflow simulation time series forced by CMA, IMERG-E, IMERG-L, IMERG-F

Model	Strategies	CMA	IMERG-E	IMERG-L	IMERG-F	Mean
HBV	Strategies I	0.77	0.77	0.72	0.88	0.79
	Strategies II	0.73	0.81	0.82	0.86	0.80
SWAT	Strategies I	0.83	0.75	0.76	0.73	0.77
	Strategies II	0.83	0.84	0.82	0.70	0.80
DHSVM	Strategies I	0.86	0.75	0.75	0.85	0.80
	Strategies II	0.82	0.87	0.86	0.87	0.86
LSTM	Strategies I	0.92	0.89	0.87	0.85	0.88
	Strategies II	0.93	0.91	0.86	0.85	0.89

References:

Huang, Y., Bárdossy, A., and Zhang, K.: Sensitivity of hydrological models to temporal and spatial resolutions of rainfall data, Hydrology and Earth System Sciences, 23, 2647-2663, 10.5194/hess-23-2647-2019, 2019.

Point 4: As I mentioned earlier, there are numerous results but proportionally less discussion. For instance, no explanation is provided for why a 0.25-degree resolution appears to perform well for NSE-CMA-SWAT (apologies for the earlier typo) but not for NSE-CMA-LSTM. Similarly, a 0.5-degree resolution seems to work for KGE-CMA-SWAT but not for KGE-CMA-LSTM. The authors have predominantly presented the results without a thorough critical analysis, which is the point I am emphasizing. Once again, I am not suggesting that the study is

irrelevant, but I believe that additional effort is needed to enhance the paper's overall appeal.

Response4: Thank you for your comment, and we have added some results and discussion about this issue in ***'5.2 Comparison of the performance of precipitation products on flood events simulation at different spatio-temporal resolutions'***:

Page 15-16 Line 478-491: 'In order to compare the performance of different models on flood events simulation in the same spatial resolutions, some results presented in Fig.7 are illustrated in Fig.9. Overall, the LSTM shows better performance in most cases, for instance, in Fig. 9 (a) and Fig. 9 (c), LSTM is better than other models with the largest mean NSE and the smallest range between 25th and 75th percentile. There is also exception, for example, in Fig. 9 (b), the range of NSE between 25th and 75th percentile of SWAT with CMA is smaller than that of LSTM, but its mean and medium values of NSE are lower. Therefore, it can be summarized that the performance of LSTM has a higher likelihood of success than the other models. For KGE at 0.1° (Fig.9 (d)), LSTM also show better performance than the other models expect that simulated with CMA, with which DHSVM is better than LSTM, and they show similar results with 0.5° (Fig. 9 (e)).'

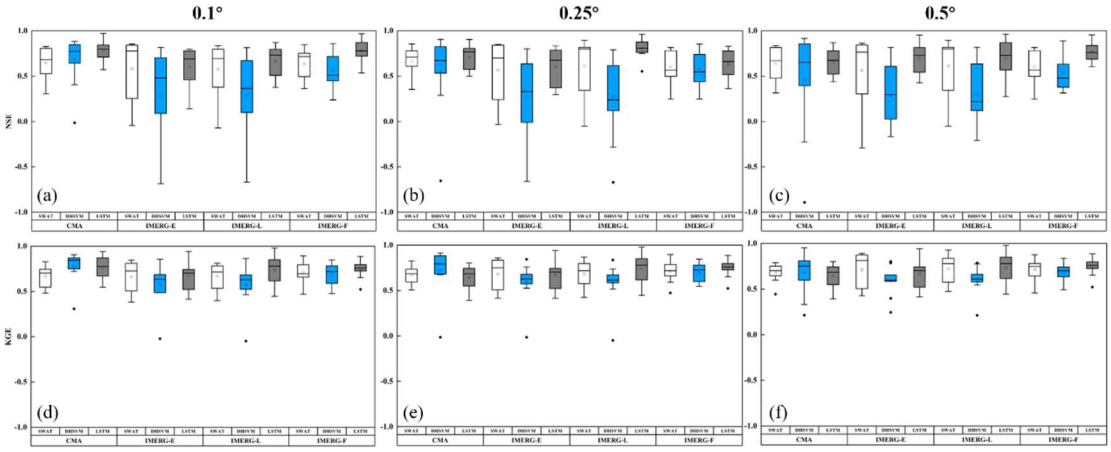

Fig. 9. The NSE and KGE of flood events simulation forced by CMA, IMERG-E, IMERG-L and IMERG-F with different spatial resolutions. The box plots show the 25th, 50th, and 75th percentiles, and the mean value is given and shown by a square.

As we stated in the manuscript, for the 0.25-degree resolution, it performs well for NSE-

CMA-SWAT and also for NSE-CMA-LSTM. Based on the mean and medium NSE values, NSE-CMA-LSTM is even better, with corresponding values increasing from 0.64 to 0.66.

To better explain the effect of spatial resolution on different models, we have added additional details in ***'5.2 Comparison of the performance of precipitation products on flood events simulation at different spatio-temporal resolutions'***:

Page 16 Line 492-513: 'The influence of spatio-temporal resolution on flood events simulation is affected by model structure. For instance, based on NSE, the SWAT shows the best performance at 0.25° with CMA forcing, but the LSTM shows the best performance at 0.1°. Similarly, based on KGE, the SWAT performs the best at 0.5° with CMA forcing, but the LSTM has the best performance at 0.1°. On one hand, the difference in performance between NSE and KGE is due to their different statistical focus, with NSE giving larger weights to high values, especially flood peaks, which leads to different performance with different statistical metrics. On the other hand, the difference between SWAT and LSTM is due to their model structure. The SWAT operates as a physically driven model, where the impact of the spatial resolution of the precipitation dataset will propagate during hydrological process, which makes finer spatial resolution does not necessarily lead to the improved performance, as indicated by studies such as Huang et al. (2019). This is exemplified by the SWAT performs better at 0.25° with CMA forcing based on NSE, while it performs better at 0.5° based on KGE. Regarding LSTM, as a deep learning model, some studies have highlighted significant performance enhancements when applied to larger, reliable datasets (Sun et al., 2017). Consequently, when forced by CMA and IMERG-F, LSTM shows the best performance across all statistical metrics at 0.1°, rather than at 0.25° or 0.5°. The deviations observed in IMERG-E and IMERG-L from this pattern are likely attributable to inherent errors within the precipitation product itself. We previously evaluated the applicability of the IMERG dataset in the Xiangjiang River Basin, and found that IMERG-E and IMERG-L have larger uncertainties and errors than IMERG-F (Zhu et al., 2020a). The CMA has been confirmed by several studies to be a more reliable

precipitation product in the Xiangjiang River Basin and always used as a reference precipitation product (Wang et al., 2017; Tang et al., 2017; Su et al., 2020). This probably makes IMERG-E and IMERG-L do not bring enough performance improvement to LSTM when the spatial resolution is finer.'

References:

Huang, Y., Bárdossy, A., and Zhang, K.: Sensitivity of hydrological models to temporal and spatial resolutions of rainfall data, Hydrology and Earth System Sciences, 23, 2647-2663, 10.5194/hess-23-2647-2019, 2019.

Su, J., Lü, H., Crow, W. T., Zhu, Y., and Cui, Y.: The Effect of Spatiotemporal Resolution Degradation on the Accuracy of IMERG Products over the Huai River Basin, Journal of Hydrometeorology, 21, 1073-1088, 10.1175/jhm-d-19-0158.1, 2020.

Sun, C., Shrivastava, A., Singh, S., and Gupta, A.: Revisiting Unreasonable Effectiveness of Data in Deep Learning Era, Ieee I Conf Comp Vis, 843-852, 10.1109/Iccv.2017.97, 2017.

Tang, G., Zeng, Z., Ma, M., Liu, R., Wen, Y., and Hong, Y.: Can Near-Real-Time Satellite Precipitation Products Capture Rainstorms and Guide Flood Warning for the 2016 Summer in South China?, IEEE Geoscience and Remote Sensing Letters, 14, 1208-1212, 10.1109/lgrs.2017.2702137, 2017.

Zhu, Q., Zhou, D., Luo, Y., Xu, Y.-P., Wang, G., and Gao, X.: Suitability of high-temporal satellite-based precipitation products in flood simulation over a humid region of China, Hydrological Sciences Journal, 66, 104-117, 10.1080/02626667.2020.1844206, 2020a.

Point 5: Mean NSE does not make a lot of sense if the distribution is skewed (which is very likely).

Response 5: Thank you very much for your comment. In most cases in our study, the medium NSE performs the same pattern as the mean NSE. For instance, in Fig. 6, the mean NSE values of HBV, SWAT, LSTM driven by IMERG-F increase from 0.54, 0.44, 0.56 with calibration strategy I to 0.67, 0.57, 0.63 with calibration strategy II while the medium NSE values increase from 0.68, 0.53, 0.98 to 0.79, 0.68, 0.99. BIAS-P also shows the same pattern between the medium BIAS-P and the mean BIAS-P. For the

mean BIAS-P values of HBV, SWAT, LSTM driven by IMERG-E decrease from 27.0%, 29.8%, 22.4% with calibration strategy I to 21.2%, 23.8%, 18.3% with calibration strategy II, while the medium BIAS-P values decrease from 34.3%, 37.6%, 27.2% to 27.8%, 32.5%, 23.1%. For the mean BIAS-P values of HBV, SWAT, LSTM driven by IMERG-F decrease from 14.5%, 26.0%, 16.0% with calibration strategy I to 13.1%, 14.0%, 15.5% with calibration strategy II while the medium BIAS-P values decrease from 15.0%, 24.7%, 17.5% to 11.3%, 11.3%, 13.1%.

Similar to Fig. 6, Fig. 7 and Fig. 8 show the same pattern between the medium NSE and the mean NSE in most case. For instance, in Fig. 7, the SWAT driven by CMA shows the best performance at 0.25° with the mean NSE of 0.66 and the medium NSE of 0.71. The SWAT driven by IMERG-E and IMERG-L show the best performance at 0.5° with the mean NSE of 0.57, 0.61 and the medium NSE of 0.76, 0.80. The SWAT driven by IMERG- F shows the best performance at 0.1°with the mean NSE of 0.63 and the medium NSE of 0.72. The LSTM driven by CMA and IMERG-F show the best performance at 0.1° with the mean NSE of 0.78, 0.78 and the medium NSE of 0.80, 0.78. The LSTM driven by IMERG-E shows the best performance at 0.5°with the mean NSE of 0.7 and the medium NSE of 0.73. The LSTM driven by IMERG-L shows the best performance at 0.25°with the mean NSE of 0.81 and the medium NSE of 0.81. In Fig. 8, the HBV driven by CMA and IMERG-F show the best performance at the hourly scale with the mean NSE of 0.81, 0.77 and the medium NSE of 0.82, 0.80. The DHSVM driven by IMERG-E and IMERG-L also show the best performance at the hourly scale with the mean NSE of 0.36, 0.37 and the medium NSE of 0.59, 0.54. So, in this study the medium NSE and the mean NSE are considered to be representative of the overall performance.

In some cases, as you said, mean NSE does not make a lot of sense if the distribution is skewed. For instance, in Fig. 6, the LSTM driven by CMA shows better medium NSE with calibration strategy I while better mean NSE with calibration strategy II. In Fig. 7, the DHSVM driven by IMERG-L shows the best medium NSE of 0.48 at 0.1° while the best mean NSE of 0.30 at 0.5° . In order to better describe the results, we use the 25^{th} NSE and 75^{th} NSE to discuss the uncertainty of the results when the distribution is

skewed.

These sentences have been re-edited:

Page 10 Line 323-331: 'For the LSTM, the NSE values of flood events simulation also show higher mean values and smaller uncertainty based on the strategy II for all precipitation products. The flood events simulation based on IMERG-L shows the most significant improvement with the mean NSE value increasing from 0.62 with the strategy I to 0.77 with the strategy II. The flood events simulation based on CMA and IMERG-E show lightly lower medium NSE values of 0.94, 0.88 with the strategy II than 0.95, 0.99 with strategy I. But they show higher 25th NSE with strategy II, especially LSTM driven by IMERG-E, which increases from 0.58 with strategy I to 0.66 with strategy II. Therefore, although strategy II has a lower median performance than strategy I in individual cases, it still significantly improves the performance of LSTM, particularly in terms of uncertainty.'

Page 10 Line 341-350: 'For instance, CMA performs the best at 0.25° with the mean BIAS-P of 26.5%, while IMERG-E, IMERG-L and IMERG-F display the best performance at 0.5° with the mean BIAS-P of 23.7%, 22.9% and 13.8%, respectively. Similar to its performance in BIAS-P, in terms of mean NSE, CMA also performs the best under 0.25° with the mean NSE of 0.66. IMERG-E presents little difference at different spatial resolutions, while IMERG-L performs slightly better at 0.5° with the mean NSE of 0.61and the medium NSE of 0.76.'

Page 11-12 Line 367-383: 'Similar to DHSVM, LSTM shows different performance forced by precipitation with different spatial resolutions. CMA and IMERG-F performs the best at 0.1° with the mean BIAS-P of 18.64% ,15.55% and mean NSE of 0.78. The 25th NSE of flood events simulated with CMA increases from 0.52 to 0.72, the 75th NSE increases from 0.78 to 0.83 while the spatial resolution is finer. By contrast, IMERG-E performs the best at 0.5° with the mean NSE of 0.69 and medium NSE of 0.68 while IMERG-L performs the best at 0.25° with the mean NSE of 0.80 and the medium NSE of 0.81. In the light of BIAS, IMERG-E and IMERG-L achieve the best performance

on flood events simulation at 0.5°, the mean values of which are 24.55%, and 18.27%, 0.77. In contrast of BIAS-P, LSTM driven by IMERG-L shows the best KGE at 0.25° with the mean KGE of 0.76 and the smallest uncertainty, which is the same as NSE. Compared with the SWAT and DHSVM, the LSTM shows better performance on flood events simulation. The mean NSEs of LSTM are higher than 0.7 in most cases, while the mean NSEs of SWAT is around 0.6, and the largest mean NSE of DHSVM is 0.68. The 25^{th} NSE of LSTM are higher than 0.5 in most cases, while the 25^{th} NSE of DHSVM is around 0.15. The smallest 75^{th} NSE of LSTM is 0.78, while the 75^{th} NSE of DHSVM are around 0.6. The mean KGEs of SWAT and LSTM are similarly around 0.7, which are around 0.6 for DHSVM. In addition, LSTM also shows a relatively lower BIAS-P (the mean values less than 25%).'

Page 14-15 Line 445-459: 'As illustrated in Fig.7 and Fig.8, the performance of precipitation products on flood events simulation is affected by both the spatial and temporal resolutions. Impacts of spatial resolution on flood events simulation behave differently among different models and precipitation sources. For the study area, under 0.25° spatial resolution, the CMA obtains the best flood events simulation based on SWAT. The impact of spatial resolution on the capture of precipitation variability during flood event periods can propagate to the flood events simulation. The best results are obtained under 0.25° spatial resolution, the possible reason can be that finer spatial resolution (0.1°) increases the uncertainty of precipitation sets, nevertheless coarser spatial resolution (0.5°) decreases the sufficiency of datasets. For SWAT driven by CMA, it shows the best 75^{th} NSE and the worst 25^{th} NSE at 0.5° while the DHSVM driven by CMA shows the same pattern at 0.5°, which proved that coarser spatial resolution decreases the sufficiency of datasets. But the DHSVM driven by CMA shows the best performance at 0.1°, which proves that the effects of increasing and decreasing spatial resolution are simultaneous and affect different models differently. It indicates that the choice of the dataset is influenced by the resolution range, which must be adapted to the model definition, for the proper spatial resolution is essential to both minimize the uncertainty and assure the sufficiency (Grusson et al., 2017).'

References:

Grusson, Y., Anctil, F., Sauvage, S., and Sánchez Pérez, J.: Testing the SWAT Model with Gridded Weather Data of Different Spatial Resolutions, Water, 9, 10.3390/w9010054, 2017.