The authors have applied multiple hydrological models using multiple precipitation datasets at various temporal and spatial resolutions to predict flood peaks. Clearly, a lot of hard work has been put into this research. The results are interesting. However, certain parts of their methodology are quite unclear to me, which is why I am unable to provide a fair opinion. I am also a bit disappointed by the lack of effort in providing insights, which would make their study complete. My specific comments are given below.

Point 1: 3.2.2. It is not clear whether you have considered peak discharge only or all the data points of the flood hydrographs. If the former is true, the number of data points is very small for any meaningful calibration. The term 'flood event' has not been explicitly defined, which gives rise to additional confusion.

Response1: Thank you very much for your comment. When we trained the models, we use all the data points of the flood hydrographs instead of just the peak discharges. For the term "flood event", ***In 2.2 Data description***, we have explained how we choose flood events: "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 with flood peak exceeding the threshold of 8,600 m^3/s in this study."

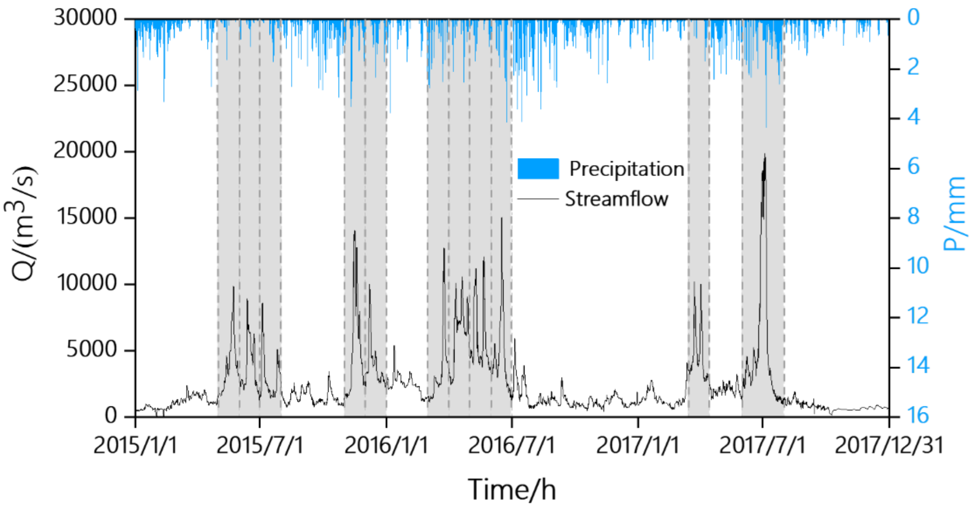

Fig. 2. Time series of observed hourly streamflow in Xiangtan station and basin-average precipitation from CMA, with eleven selected flood events covered by shaded areas.

Point 2: 3.3. It is not clear if Eq. (1) uses only flood peaks or all the data points in the time series for computing NSE. If the former is the case, the metric is not reliable since there are not many data points considered by the author.

Response2: Thank you very much for your comment. To quantitatively evaluate the performance of flood events simulation, three evaluation indices are selected in this study, namely NSE, BIAS-P and KGE. In Eq. (1) and Eq. (3), we used all the streamflow data points of 11 flood events and calculated the evaluation indices for each flood event separately.

We have modified the relevant description of Eq. (1) and Eq. (3) in Page 12 Lines 284-287: 'Where Q_o^t and Q_s^t are the values of the observed and simulated flood events at time t; Q_o^p and Q_s^p are the observed and simulated peaks of the flood events; r is the linear correlation between observations and simulations, α a measure of the flow variability error, and β a bias term.'

Point 3: Eq. (2): Bias is not typically presented in this way. Again, how reliable is the equation when there are so few data points?

Response3: Thank you very much for your comment. To quantitatively evaluate the performance of the flood peaks simulation, BIAS-P is selected in this study instead of the common BIAS. BIAS-P provides a more comprehensive reflection of the errors between observed and simulated flood peaks.

Point 4: 4.1. The results are not very surprising since you have calibrated the model for flood peaks only. I would be surprised if you also show an improvement in overall NSE (i.e., NSE considering all the data points).

Response4: As we respond above, we use all the data points of the flood hydrographs to calibrate the models, and the best model is selected by maximizing the mean NSE of the flood events simulation. All trained models show an improvement in overall NSE, but the selected models are those that performs the best on flood events simulation.

Point 5: 4.2. The effects of precipitation data type on model performance are quite informative. However, no proper explanation is provided in the discussion section, which makes the analysis incomplete. Line 460 is unclear. What do you mean by error propagation? Please explain instead of merely citing another paper.

Response5: Sorry for the misunderstanding. The section 4.2 is about the impact of spatial resolutions of precipitation on flood events simulation, rather than the effects of precipitation data type on model performance. To investigate the impact of spatial resolutions of precipitation on flood events simulation, the IMERG-E, IMERG-L, IMERG-F, and CMA are adopted to force the SWAT model, the DHSVM model and the LSTM model under 0.1 °, 0.25 ° and 0.5 °.

The relevant discussion about the effects of precipitation data type on model performance is presented in ***5.2 "Comparison of the performance of precipitation products on flood events simulation at different spatio-temporal resolutions".***

We are very sorry for the difficulty in reading. The sentence in Line 460 has been re-edited:

Page 20 Lines 464-466: 'Furthermore, when driven by IMERG, HBV outperforms SWAT and DHSVM, especially by IMERG-E and IMERG-L. It is because the hydrological model with a simpler structure can reduce the impact of errors in radar rainfall estimation, which is better constrained during its propagation in the hydrological process (Zhu et al. 2013).'

References:

Zhu, D., Peng, D. Z., and Cluckie, I. D.: Statistical analysis of error propagation from radar rainfall to hydrological models, Hydrology and Earth System Sciences, 17, 1445-1453, 10.5194/hess-17-1445-2013, 2013.

Point 6: 4.3. The results look interesting. Again, no explanation is provided. For example, why does 0.25-degree data give the best 75th NSE and the worst 25th NSE for HBV (Figure 7a)?

Response6: Thank you very much for your comment. Fig. 7 show the performance of flood events simulation based on SWAT, DHSVM, and LSTM forced by precipitation with different spatial resolutions. But the HBV is not included in this part. Since our target is to explore the impacts of different resolutions on flood events simulation, we focus on the overall performance of the model, therefore, the mean NSE is used in our study for it is more suitable for flood events as many previous studies proved (Yu et al. 2018, Kao et al. 2020).

References:

Kao, I. F., Zhou, Y., Chang, L.-C., and Chang, F.-J.: Exploring a Long Short-Term Memory based Encoder-Decoder framework for multi-step-ahead flood forecasting, Journal of Hydrology, 583, 10.1016/j.jhydrol.2020.124631, 2020.

Yu, D., Xie, P., Dong, X., Hu, X., Liu, J., Li, Y., Peng, T., Ma, H., Wang, K., and Xu, S.: Improvement of the SWAT model for event-based flood simulation on a sub-daily timescale, Hydrology and Earth System Sciences, 22, 5001-5019, 10.5194/hess-22-5001-2018, 2018

Point 7: The abstract says LSTM is outperforming other models. Figure 6 says HBV is better than LSTM.

Response7: Thank you very much for your comment. Fig. 6 shows the distributions of NSE and BIAS-P values to illustrate the impact of calibration strategies on flood events simulation. It can be seen that flood events simulation with LSTM shows better performance than HBV, for the mean NSE values of CMA, IMERG-E, IMERG-F increase from 0.79, 0.62, 0.75 based on HBV to 0.82, 0.76, 0.77 based on LSTM. For IMERG-L, the mean NSE of LSTM is slightly lower than HBV, but the BIAS-P of LSTM show better performance than HBV. As a whole, we can summarize that LSTM is better than HBV. And there is a section about the comparison of different models on flood events simulation in the discussion part. Please refer to "***5.3 Comparison of different models on flood events simulation***" for details.

Point 8: Line 90: The term 'physically based' is typically used for hydrological models based on hydrodynamic equations. The models you are referring to are typically called conceptual models. This is just a semantic issue though.

Response8: Thank you for pointing it out. Yes, HBV is a conceptual model, while SWAT and DHSVM are physically based models. And we revise the corresponding sentences as you suggested.

Page 2 Lines 38: 'Numerous models are applied to simulate the flood events, most of which are conceptual/physically based models.'

Page 4 Lines 91-94: 'Therefore, three widely used and typical conceptual/physically based models (lumped HBV model, semi-distributed SWAT model, and distributed DHSVM model), and one data-driven model (LSTM) which shows good performance in hydrological simulation, are employed to probe the impacts of spatio-temporal resolutions of precipitation on flood events simulation.'

Page 7 Lines 167-171: 'As mentioned above, three widely used and typical conceptual/physically based models (lumped HBV model, semi-distributed SWAT model, and distributed DHSVM model), and one data-driven model (LSTM), are employed to probe the impacts of spatio-temporal resolutions of precipitation on flood events simulation.'