Dear Editor and Reviewers:

We are grateful for your consideration of our manuscript entitled “Comparing machine learning and deep learning models for probabilistic post-processing of satellite precipitation-driven streamflow simulation" [HESS-2022-377], and we also very much appreciate your constructive comments and useful suggestions, which have enabled us to improve the manuscript. All the comments we received on this study have been considered, and we present our reply to each of them separately. We hope the revised manuscript would satisfy you.

Response to Editor

Your submission was evaluated by three reviewers who provided helpful feedback during the discussion step but have also raised some concerns about parts of your study. Overall, reviewers’ ratings ranged from quite good, or even excellent, to rather poor mainly based on their respective backgrounds and experiences. However, among the other things that I see in your replies that you should revise, I suggest that the scientific significance of your study should be much improved.

I release your paper under major revisions. When submitting the revised version, please upload also detailed point-by-point replies to the comments and concerns received so far, together with possible additional comments that can help evaluate your changes.

Response: Thank you very much for your effort and suggestion. we have rewritten our abstract and most part of main text to improve the scientific significance of our study.

***Abstract.*** *Deep learning (DL) and machine learning (ML) are widely used in hydrological post-processing, which plays a critical role in improving the accuracy of hydrological predictions. However, the trade-off between model performance and computational cost has always been a challenge for hydrologists when selecting a suitable model, particularly for probabilistic post-processing with large ensemble members. Moreover, it is unclear whether the performance differences between DL and ML models is significant in hydrological probabilistic post-processing. Therefore, this study aims to systematically compare the quantile regression forest (QRF) model and countable mixtures of asymmetric Laplacians long short-term memory (CMAL-LSTM) model as hydrological probabilistic post-processors. Specifically, we evaluate their ability in dealing with biased streamflow simulation driven by three satellite precipitation products across 522 sub-basins of Yalong River basin in China. Model performance is comprehensively assessed using a series of scoring metrics from both probabilistic and deterministic perspectives. Our results show that the QRF model and the CMAL-LSTM model are comparable in terms of probabilistic prediction, and their performance is closely related to the flow accumulation area (FAA) of the sub-basin. The QRF model outperforms the CMAL-LSTM model in most of the sub-basins with smaller FAA, while the CMAL-LSTM model has an undebatable advantage in sub-basins with FAA larger than 60,000 km2 in Yalong River basin. In terms of deterministic predictions, the CMAL-LSTM model is preferred, especially when the raw streamflow is poorly simulated and used as an input. However, if we put aside the differences in model performance, the QRF model is more efficient than the CMAL-LSTM model in computation time in all experiments. As a result, this study provides insights into model selection in hydrological post-processing and the trade-offs between model performance and computational efficiency. The findings highlight the importance of considering the specific application scenario, such as the catchment size and the required accuracy level, when selecting a suitable model for hydrological post-processing.*

Response to Referee #1

This study compares two post-processing methods of streamflow simulation obtained using different precipitation products based on satellite data. A comprehensive evaluation is performed on 522 sub-catchments located in China to assess the performances in terms of reliability, sharpness, and various hydrological skills. The paper is well-written and complete, the figures are clear and the interpretations of the results are convincing. My recommendation is that the paper can be accepted for publication after minor corrections which are listed below.

Response: Thank you for your positive comments. Each of your suggestions is very valuable to us as they have greatly improved the quality and readability of the manuscript. The following are point-by-point responses to these comments.

l.44-46: I strongly disagree with this statement. There is no evidence that satellite precipitation estimation is the most promising hydrological model input. As an example, ERA5 is mostly driven by satellite data and is not able to reproduce most of the precipitation features at a high spatial resolution (Bandhauer et al., 2022; Reder et al., 2022), does not reproduce the strong relationships between precipitation characteristics and the topography in mountainous areas, underestimate hourly and daily extreme values and overestimate the number of wet days (Bandhauer et al., 2022). At high spatial and temporal resolutions, the assimilation of ground measurements and/or radar data is needed to reproduce extreme events (Reder et al., 2022). However, I agree that satellite precipitation estimation is valuable in regions where ground measurements are scarce.

Bandhauer, Moritz, Francesco Isotta, Mónika Lakatos, Cristian Lussana, Line Båserud, Beatrix Izsák, Olivér Szentes, Ole Einar Tveito, and Christoph Frei. 2022. “Evaluation of Daily Precipitation Analyses in E-OBS (V19.0e) and ERA5 by Comparison to Regional High-Resolution Datasets in European Regions.” International Journal of Climatology 42 (2): 727–47. https://doi.org/10.1002/joc.7269.

Bellier, Joseph, Isabella Zin, and Guillaume Bontron. 2018. “Generating Coherent Ensemble Forecasts After Hydrological Postprocessing: Adaptations of ECC-Based Methods.” Water Resources Research 54 (8): 5741–62. https://doi.org/10.1029/2018WR022601.

Reder, A., M. Raffa, R. Padulano, G. Rianna, and P. Mercogliano. 2022. “Characterizing Extreme Values of Precipitation at Very High Resolution: An Experiment over Twenty European Cities.” Weather and Climate Extremes 35 (March): 100407. https://doi.org/10.1016/j.wace.2022.100407.

Response: We agree with your opinion on satellite precipitation products. We have weakened the statement here and highlight the significance of satellite precipitation estimation for remote areas.

> *Precipitation information is mainly derived from gauge observations, radar precipitation estimates, satellite precipitation retrievals and reanalysis products (Sun et al., 2018). Gauge stations and radar are limited by the density of the station network and the topography, especially in remote areas such as mountainous regions and high altitudes (Sun et al., 2018; Chen et al., 2020). Reanalysis requires the assimilation of observations from multiple sources and therefore cannot be obtained in real time. Satellite precipitation estimates are available in near-real-time and have shown valuable potentials for applications in regions where ground measurements are scarce. (Jiang and Bauer-Gottwein, 2019; Dembélé et al., 2020).*

l.75: A more recent application of MOS method is provided by Bellier et al. (2018).

Response: Thank you for sharing this more recent application of MOS method, we have added it to our reference.

> *Another way is to use the idea of model output statistics (MOS) (Wang et al., 2009; Bogner and Pappenberger, 2011; Bellier et al., 2018).*

l.80: short memory: I guess that ‘term’ is missing between ‘short‘ and ‘memory’.

Response: Thank you for pointing it out. We have fixed it.

> *For example, long short-term memory (LSTM) models have been used to simulate streamflow in a number of gauged and ungauged basins in North America (Kratzert et al., 2018, 2019), the United Kingdom (Lees et al., 2021), and Europe (Nasreen et al., 2022).*

l.123: serval -> several.

Response: Thank you for correcting this. It was a typo and we have rephrased this sentence.

> *Following the watershed division method of Du et al. (2017), Yalong River basin is divided into 522 sub-basins with catchment area ranging from 100 km$^2$ to 127,164 km$^2$ (Fig. 1b).*

l.195: “so the model is reliable”. Is it possible to rephrase the sentence to indicate that this is an assumption and not your personal judgement? As the authors do not provide evidence that the model is able to reproduce the natural runoff process (I understand that it is not possible), it would be fairer.

Response: Thank you for your suggestion. Here, we would like to state that the calibrated hydrological model meets the needs of the subsequent study. With regard to reviewer 2's suggestion, we agree that deleting this could be a wiser choice and we have deleted this sentence.

l.247: Klotze -> Klotz.

Response: Thank you for correcting this. It was a typo and we have checked our text.

l.255-256: The terms “single-model” and “multi-model” are a bit misleading, as I understand that the authors refer to precipitation products here. I suggest replacing them by “single-precipitation” product and “multi-precipitation” or something similar.

Response: Thank you for your suggestion. We have replaced “multi-model” by “multi-product” in our text.

l.348: Missing dot after “threshold”.

Response: Thank you for pointing out this minor error. We have rephrased this sentence.

> *As the threshold conditions increase, the performance of the multi-product approach is slightly worse than that of IMERG-F (Fig. S5).*

l.448: “Little precipitation events”: I was not sure if the authors refer to localized precipitation events here, or with moderate intensities. Is it possible to be more specific?

Response: Sorry for the confusing information. We want to refer to localized precipitation events. We have rephrased this sentence.

> *For sub-basin No.250 with a smaller catchment area, its rainfall-runoff response is faster, and the fluctuation of streamflow is greater. Localized precipitation events can also cause large pulse flow, which is the main feature of flash floods.*

Response to Referee #2

I reviewed the manuscript entitled “Comparing machine learning and deep learning models for probabilistic post-processing of satellite precipitation-driven streamflow simulation” by Zhang et al. The manuscript compares the uses of a machine learning method (QRF) and a deep learning method (PLSTM) for bias-correction of streamflow simulations. The study uses the reference precipitation-driven streamflow as the reference for the bias-correction instead of the observed streamflow due to the data availability of the region. Overall, I have five major concerns.

Response: Thank you very much for your time. And we are very grateful for the valuable comments and suggestions on our manuscript. Based on the concerns you mentioned, we have made thorough changes to our manuscript accordingly. We hope that our responses will satisfy you.

Here, I would like to start by outlining the changes we have made in the revised manuscript.

- We modified the abstract section to emphasize the scientific significance of the article.
- We rewrote the methods section to shorten the introduction of post-processing models and to present the evaluation metrics and their implications in more detail.
- We reorganized the structure of the Methods, Results, and Discussion sections to increase readability.
- We rewrote the discussion section and focused on the interpretations of results.
- We adjusted some of the figures and tables to highlight the results and reduce redundant information.

Comment 1: Lack of interpretations on results

This study used several statistics for model performance evaluation, namely the continuous rank probability score (CRPS), the weighted CRPS, the reliability diagram, and the sharpness. The figures/tables were used to demonstrate those statistics. My first and biggest concern is the lack of interpretation on the appearance of the figures/tables. For example, I am less familiar with the concept of a reliable diagram; after reading section 4.2.4, I was still not able to understand what Figure 7 and 8 were showing. It seems that the optimum is to have lines following the diagonal line. But how to quantitatively define “close to the diagonal line”? If it is close then it is a reliable prediction. But what exactly is meant for “reliable prediction”? If a line is mostly located above (below) the diagonal line, it is an underestimation (overestimation) of what? Another example is the concept of sharpness. I was not able to understand this concept after reading lines 312-315 where the concept was introduced. After reading section 4.2.5, the section dedicated to the sharpness-related results, I was even more puzzled. The section compared the variability of the different streamflow estimations and it seems that if those statistics show smaller values (lower variability), then the model is better. Again, what is it better for and why? It is hard to interpret the meaning probably due to the lack of descriptions on those two methods (reliable diagram and sharpness). Rather than those, I also found the use of CRPS and twCRPS redundant (see the same pattern between panel a and c and b and d in Figure 4). The patterns of Figure 3 also need to be interpreted properly.

Response: Thank you for your comments and suggestions. We provide point-by-point responses to the above comments according to the order of the articles.

Comment 1.1

The patterns of Figure 3 also need to be interpreted properly.

Response: When using the same calibrated hydrological model, the quality of the precipitation product determines the performance of the streamflow simulation. Using observed precipitation as a reference, we calculated spatial metrics (Pearson correlation coefficient, PCC and Relative bias, RB) of three satellite precipitation products (see Fig. S4 below).

Compared to PDIR and GSMaP, IMERG has both higher PCC and lower RB values. This explains its higher NSE values. Compared to PDIR, GSMaP suffers from larger biases (RB), resulting in its worse performance.

Although GSMAP is a bias-corrected product, it is not guaranteed that it is superior to the near-real-time PDIR. This is because PDIR uses more advanced precipitation retrieval algorithms (Nguyen et al., 2020, 2021). The Precipitation Estimations from Remotely Sensed Information using Artificial Neural Networks (PERSIANN) Dynamic Infrared-Rain rate model (PDIR) utilizes climatological data to construct a dynamic cloud-top brightness temperature ($T_b$)——rain rate relationship. The algorithm is a machine learning method and uses historical observations to calibrate the model parameters during training process. No additional observations are required in the prediction period, so it is a near real-time product.

> *The precipitation product quality plays a crucial role in streamflow performance with the same hydrological model configuration. The high precipitation bias in GSMaP (Fig. S4f in the supplement) leads to high biases in streamflow simulations (Fig. 8b), resulting in the lowest NSE values (Fig. 3c and Fig. 8c) of the three products. The performance of PDIR-driven streamflow is mainly influenced by the poor temporal variability (PCC) against observations (Fig. S4a in the supplement and Fig. 8a).*

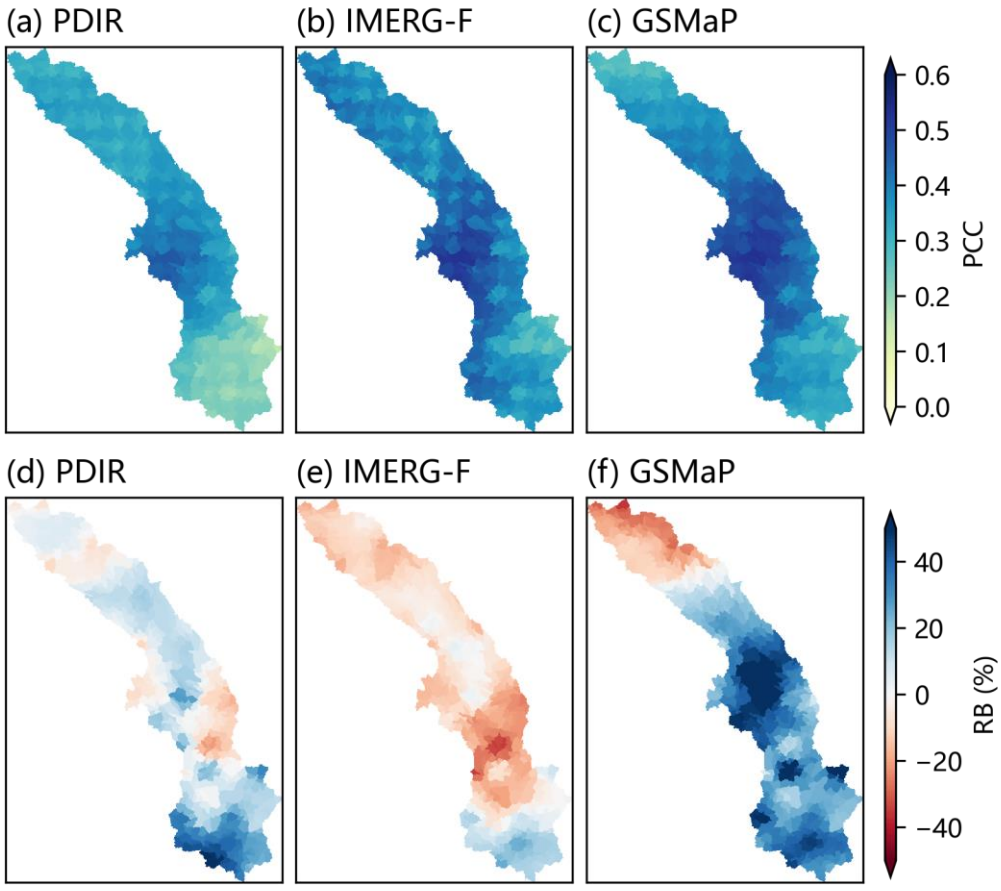



Figure S4. The PCC and RB of three satellite precipitation estimations for 522 sub-basins.

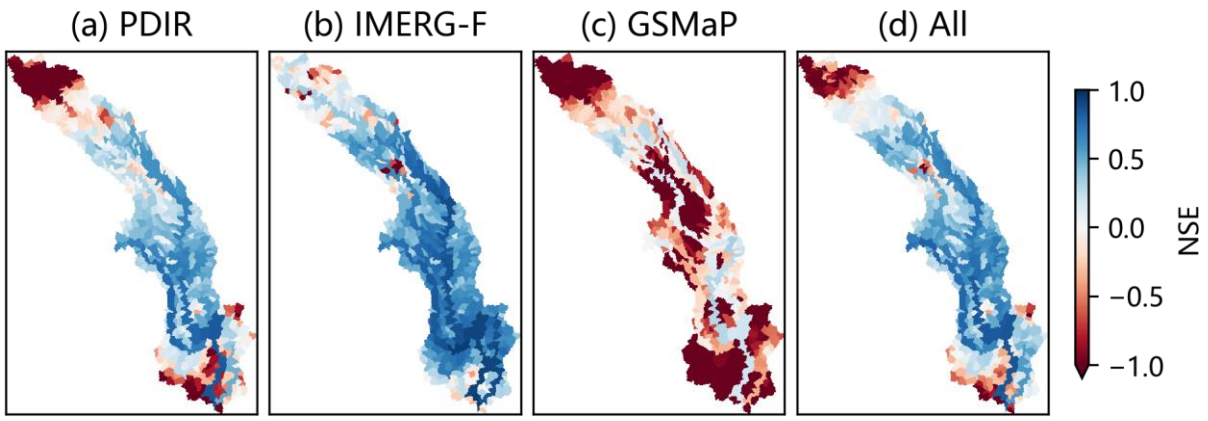



Figure 3. The NSE of uncorrected streamflow simulations for 522 sub-basins.

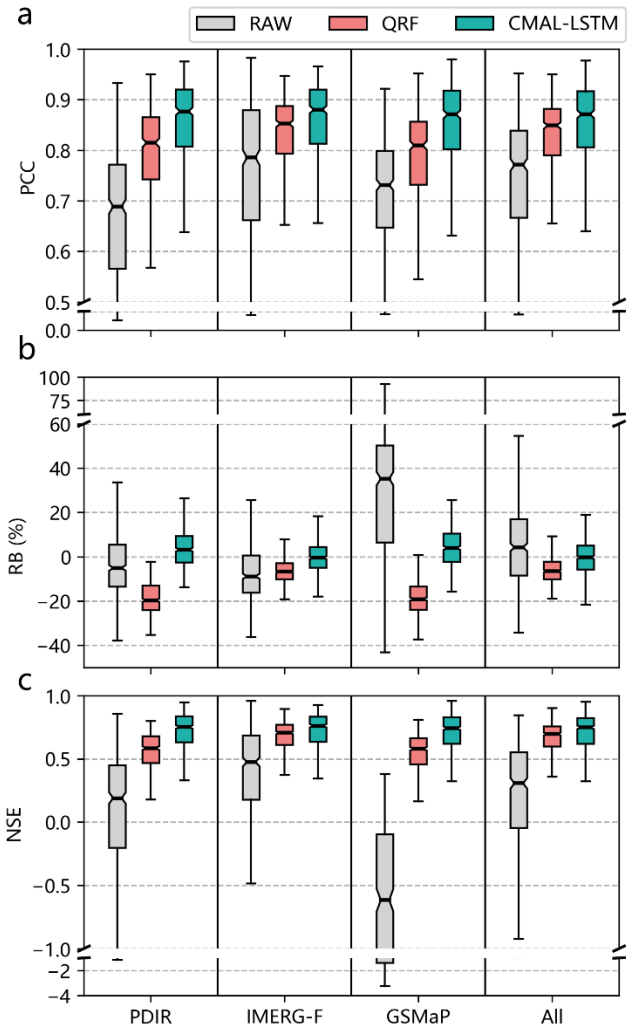



Figure 8. Boxplots of different model performance in 522 sub-basins. (a) PCC; (b) RB; and (c) NSE.

Comment 1.2

This study used several statistics for model performance evaluation, namely the continuous rank probability score (CRPS), the weighted CRPS, the reliability diagram, and the sharpness. The figures/tables were used to demonstrate those statistics. My first and biggest concern is the lack of interpretation on the appearance of the figures/tables. For example, I am less familiar with the concept of a reliable diagram; after reading section 4.2.4, I was still not able to understand what Figure 7 and 8 were showing. It seems that the optimum is to have lines following the diagonal line. But how to quantitatively define "close to the diagonal line"? If it is close then it is a reliable prediction. But what exactly is meant for "reliable prediction"? If a line is mostly located above (below) the diagonal line, it is an underestimation (overestimation) of what? Another example is the concept of sharpness. I was not able to understand this concept after reading lines 312-315 where the concept was introduced. After reading section 4.2.5, the section dedicated to the sharpness-related results, I was even more puzzled. The section compared the variability of the different streamflow estimations and it seems that if those statistics show smaller values (lower variability), then the model is better. Again, what is it better for and why? It is hard to interpret the meaning probably due to the lack of descriptions on those two methods (reliable diagram and sharpness).

Response: Thank you for your very useful comments. We are very sorry for the deficient descriptions of the probabilistic metrics.

In contrast to deterministic (single-point) predictions, probabilistic predictions (multi-point) of continuous variables take the form of predictive cumulative distribution functions. Therefore, the evaluation principle of probabilistic prediction is to compare the relationship between the probability distribution function and the observation (Gneiting et al., 2007). Developed from Murphy (1993), there are nine key attributes to assess forecast quality: bias, correlation, accuracy, skill, reliability, sharpness, resolution, discrimination, and uncertainty (Troin et al., 2021; Huang and Zhao, 2022).

Table. Description of the nine key attributes to assess forecast quality (Murphy, 1993)

| Attributes | Definition |
|---|---|
| Bias | Correspondence between mean forecast and mean observation |
| Correlation | Overall strength of the linear relationship between individual pairs of forecasts and observations |
| Accuracy | Average correspondence between individual pairs of forecasts and observations |
| Skill | Accuracy of forecasts of interest relative to accuracy of forecasts produced by standard of reference |
| **Reliability** | Correspondence between conditional mean observation and conditioning forecast, averaged over all forecast |
| **Sharpness** | Variability of forecasts as described by distribution of forecasts |
| Resolution | Difference between conditional mean observation and unconditional mean observation, averaged over all forecasts |
| Discrimination 1 | Correspondence between conditional mean forecast and conditioning observation, averaged over all observations |
| Discrimination 2 | Difference between conditional mean forecast and unconditional mean forecast, averaged overall observations |
| Uncertainty | Variability of observations as described by distribution of observations |

A common conclusion of the forecast verification literature is that there is no best verification approach combining all attributes sought in assessing a forecast. Gneiting et al. (2007) propose to evaluate predictive performance based on the paradigm of **maximizing the sharpness of the prediction distributions subject to calibration (the same as reliability,** Jolliffe and Stephenson, 2012**)**. In other words, make sure the probabilistic forecasts are reliable, and then make them as sharp as possible. In addition to reliability and sharpness, scoring rules assign numerical scores to probabilistic forecasts and form attractive summary measures of predictive performance, in that they address reliability and sharpness simultaneously. Therefore, in this study, we followed these principles and selected CRPS, reliability diagram and sharpness as our probabilistic (multi-point) metrics.

- CRPS is a widely used proper scoring rule that assesses reliability and sharpness simultaneously (Gneiting et al., 2007). For given probabilistic members, the CRPS calculates the difference between the cumulative distribution function (CDF) of the probabilistic members and the observations. The CRPS is a composite indicator, similar to the NSE. It can give us comprehensive evaluation results, and we use it as the main probabilistic metric. However, the decomposition of CRPS can provide additional information (Candille and Talagrand, 2005). For example, in our study, the QRF model outperforms the CMAL-LSTM model with a lower CRPS value. Perhaps because QRF is both reliable and sharp. Or is it just that QRF is more reliable than CMAL-LSTM. So, we further used reliability and sharpness metrics.
- Reliability measures how closely the forecast probabilities of an event correspond to the actual chance of observing the event. The reliability diagram is a common **graphical tool** to evaluate and summarize this relationship. It consists of plotting observed frequencies against forecast probabilities. The reliability diagram groups the predictions into bins according to the probability (Forecast probability, horizontal axis). The frequency with which the event was observed to occur for this sub-group of predictions is then plotted against the vertical axis (Observed relative frequency). For perfect reliability the forecast probability and the observed relative frequency should be equal, and the plotted points should lie on the diagonal. For example, when the forecast states an event will occur with a probability of 25% then for perfect reliability, the observed relative frequency should occur on 25%. If a line is mostly located above the diagonal line, it is underestimation of probability (underprediction). For example, for a specific event, the forecasted probability is 0.4, but the observed relative frequency is 0.6. The forecast underestimates the actual probability of occurrence.
- Sharpness is the **variability** of forecasts as described by distribution of forecasts. For a set of probabilistic members, sharpness describes the dispersion of the probabilistic quantiles. If the prediction interval is smaller, the probabilistic prediction tends to be deterministic with smaller uncertainty. Therefore, if the statistic is smaller, and the dispersion is smaller, then the model is better. The 50% and 90% quantile intervals are the most common choices in the literature. In Murphy's definition, sharpness is only relevant for predictions, not observations. Focusing only on predictions makes sense, but is one-sided. For example, a sharp prediction interval but misses almost any observation is meaningless. Therefore, some studies also use the coverage of observations by prediction intervals to supplement the evaluation. For example, Ajami et al. (2007) count the number of observations within the 95% prediction interval. Last but not least, the 50% and 90 prediction intervals can only calculate partial probabilistic members, not all quantile members. Therefore, consistent with previous study (Klotz et al., 2022), we finally selected three additional metrics for all probabilistic quantiles, namely Mean absolute deviation (MAD), Standard deviation (STD) and Variance (VAR).

In summary, we use numerical scores and diagnostic plots to explore specific properties of probabilistic predictions and make holistic evaluations. We have rewritten the evaluation metric part.

*We followed the criterion for probabilistic predictions proposed by Gneiting et al. (2007): to maximize the sharpness of the prediction distributions subject to reliability. We both use scoring rules and diagnostic graphs to assess reliability and sharpness holistically.*

*The continuous rank probability score (CRPS) is a widely used proper scoring rule that assesses reliability and sharpness simultaneously (Gneiting et al., 2007). For given probabilistic members, the CRPS calculates the difference between the cumulative distribution function (CDF) of the probabilistic members and the observations. We also used a weighted version of CRPS (threshold weighted CRPS, twCPRS), which is commonly used to give more weight to extreme cases (Gneiting and Ranjan, 2011). These two metrics can be expressed as follows:*

$$CRPS(F,x) = \int_{-\infty}^{\infty}\{F(y) - \mathbf{1}(y \geq x)\}^2 dy \quad (2)$$

$$twCRPS(F,x) = \int_{-\infty}^{\infty}\{F(y) - \mathbf{1}(y \geq x)\}^2 \omega(y) dy \quad (3)$$

*where $\omega(y)$ is a threshold weighted function and is calculated based on the threshold q (80%, 90% and 95% percentiles of observations in this study). When $y \geq q$ ($y < q$), $\omega(y)$ equals 1 (0). $F(y)$ is the CDF obtained from the probabilistic members for the corrected streamflow; $\mathbf{1}(y \geq x)$ is the Heaviside function. The better performing model has both metrics (CRPS and twCRPS) closer to 0.*

*The CRPS skill score (CRPSS) is also used to define the relative differences between the two post-processing models. For QRF and CMAL-LSTM, the CRPSS can be calculated as:*

$$CRPSS_{QRF/PLSTM} = \left(1 - \frac{CRPS_{QRF}}{CRPS_{PLSTM}}\right) \times 100\% \quad (4)$$

*A CRPSS greater than 0 indicates that the QRF model is better than the CMAL-LSTM model, and vice versa.*

*The reliability diagram is used as diagnostic graph to assess the agreement between the predicted probability and the observed frequency (Jolliffe and Stephenson, 2003). Namely, if the predicted probability of a particular event is 30%, then the observed relative frequency should also be 30%. Ultimately, perfectly reliable predictions at multiple levels of probability result in a distribution along the diagonal line corresponding to the same levels of observed frequency. A point above (below) the diagonal line in the reliability diagram indicates that the observed relative frequency is higher (lower) than the predicted probability and that there is an underprediction (overprediction) phenomenon. Here again, three thresholds (80%, 90% and 95%) are chosen to better evaluate the reliability of extreme cases (Yang et al., 2021).*

*Sharpness is a fundamental characteristic of predictive distributions. A sharp probabilistic output corresponds to a low degree of variability in the predictive distribution. To evaluate the sharpness of probabilistic predictions, prediction intervals are commonly employed (Gneiting et al., 2007). For this study, the 50% and 90% percentile intervals were chosen. Furthermore, to establish the relationships between predictive distributions and observations, we assessed the coverage of the prediction intervals over the observations. The average Euclidean distance of the 25% and 75% probabilistic members is adopted as the sharpness metric (DIS25-75) for the 50% prediction interval, and the 5% and 95% probabilistic members were used to compute the sharpness metric (DIS5-95) for the 90% prediction intervals. The ratio of the number of observations in the prediction intervals to the total number of observations was used as the coverage of observations (CO25-75 and CO5-95). In addition, three additional metrics used in a previous study (Klotz et al., 2022) are also employed to calculate the sharpness metric for the full probabilistic members, including mean absolute deviation (MAD), standard deviation (STD) and variance (VAR).*

Comment 1.3

Rather than those, I also found the use of CRPS and twCRPS redundant (see the same pattern between panel a and c and b and d in Figure 4).

Response: We understand your concern regarding possible redundancy. We agree that the similar patterns between CRPS and twCRPS results. CRPS is an integral over the **whole** range of values, while twCRPS is an integral over a **partial** range of values, which is a weighted version of the CRPS and gives more weight to the extreme cases. For this reason, even though the patterns are consistent, they guarantee more convincing results for different cases.

We show the CRPS result in the main text (Fig. 4) and have moved the twCRPS results to supplementary material (Fig. S5).

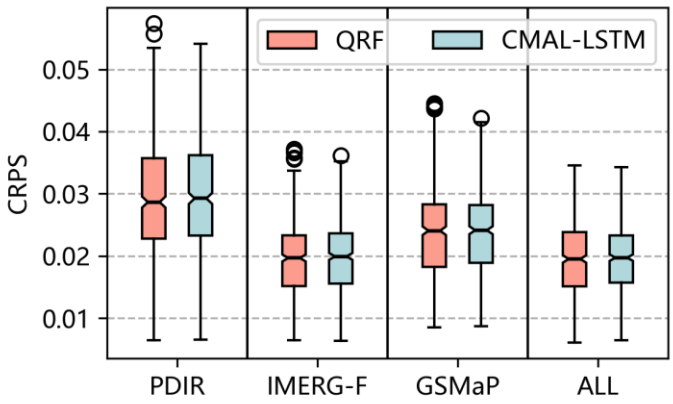



Figure 4. The boxplot of CRPS for different post-processing experiments.

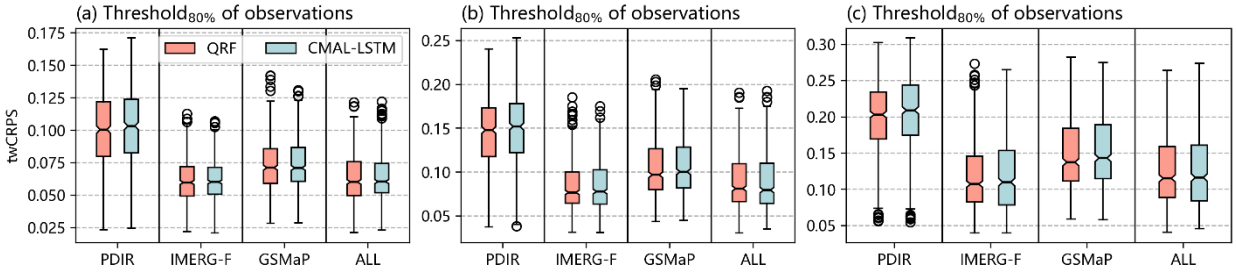



Figure S5. The boxplot of twCRPS for different post-processing experiments.

Comment 2: Drainage area thresholds

Comment 2.1

The authors provided scatter plots between drainage areas and CRPS (CRPSS) in Figure 6. Two different drainage area thresholds (20,000 and 60,000 km$^2$) were used to split the space of the plots for CRPS and CRPSS, respectively. I was not sure how those thresholds were selected. It seems that they are arbitrarily selected by the authors. Moreover, in the latter Figures 7 and 8, only 60,000 km$^2$ was used as the threshold, while in Figure 12, 20,000 km$^2$ was used again. I can't see a clear reason for switching between thresholds.

Response: Thank you very much for your comments and questions. This is a very important and critical question, and one that we struggled with in our analysis.

The analysis of the relationship between model performance and drainage areas and the thresholds were intended to better compare the QRF model with the CMAL-LSTM model, as well as to provide insights for their application in streamflow post-processing. Regrading model performance and drainage areas, it would be very exciting if critical thresholds exited. A number of studies have given different thresholds for various basins. For example, Nijssen (2004) concluded that streamflow errors are large for small drainage area but decreased rapidly for drainage areas larger than about 50000 km$^2$ the Ohio River basin. Mandapaka et al. (2009) showed that the radar-rainfall errors are spatially correlated with a correlation distance of about 20 km in the central Oklahoma region. Nikolopoulos et al. (2010) showed the propagation of the rainfall error depends on the basin size and small watershed (< 400 km$^2$) exhibited a higher ability in dampening the error than larger-sized watersheds in the Posina and Bacchiglione basins.

Based on your reminder, we have rethought this issue carefully. We also used logarithmic axes to show the results (see Fig.6 and Fig.10). We acknowledge that the threshold is indistinguishable and that an explicit threshold may change with the choice of basins, which may not be very informative for readers. Therefore,

we deleted the current threshold lines and added tables to list the number of sub-basins in different FAA intervals (Table 2 and Table S2).

We express these patterns as more general conclusions. For example, the model performance is a function of the drainage area. In the probabilistic post-processing scenario, QRF model outperforms CMAL-LSTM model in most sub-basins with small drainage areas; as the drainage area increases, the CMAL-LSTM model slightly performs better compared to QRF model.

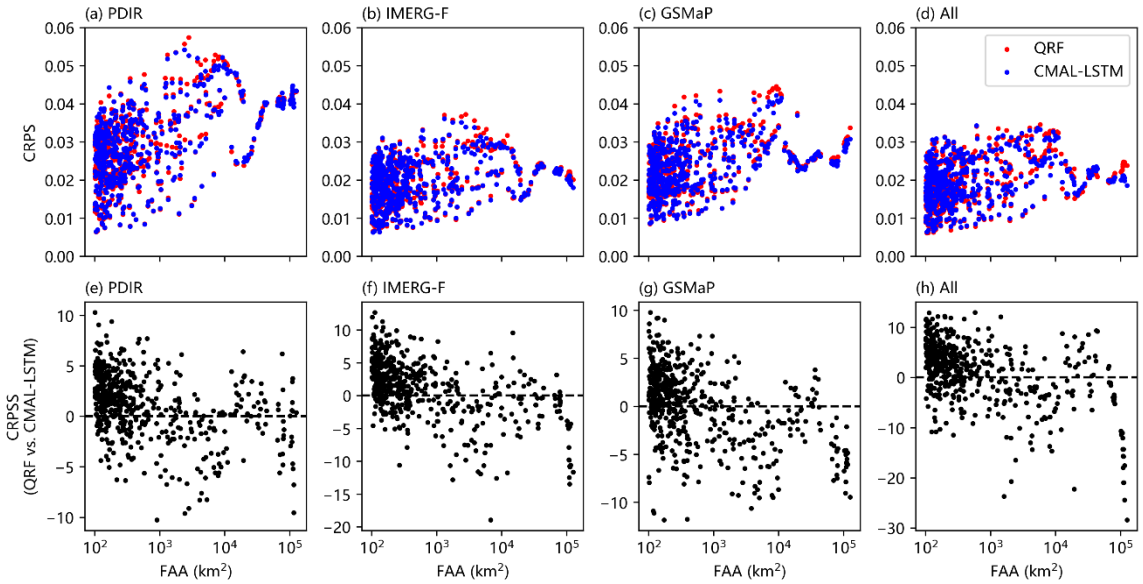



Figure 6. The relationship between (a-d) CRPS, (e-h) CRPSS and FAA.

Table 2. The probabilistic performance of two post-processing models for different FAA intervals.

| FAA ($10^4$ km) | Number of sub-basins | PDIR | | IEMRG | | GSMAP | | ALL | |
|---|---|---|---|---|---|---|---|---|---|
| | | QRF | CMAL-LSTM | QRF | CMAL-LSTM | QRF | CMAL-LSTM | QRF | CMAL-LSTM |
| < 2 | 476 | **331** | 145 | **332** | 144 | **273** | 203 | **320** | 156 |
| 2–4 | 15 | **11** | 4 | 6 | **9** | **9** | 6 | **11** | 4 |
| 4–6 | 4 | **3** | 1 | 1 | **3** | 1 | **3** | **4** | 0 |
| 6–10 | 13 | 4 | **9** | 3 | **10** | 0 | **13** | 2 | **11** |
| > 10 | 14 | 7 | 7 | 0 | **14** | 0 | **14** | 0 | **14** |

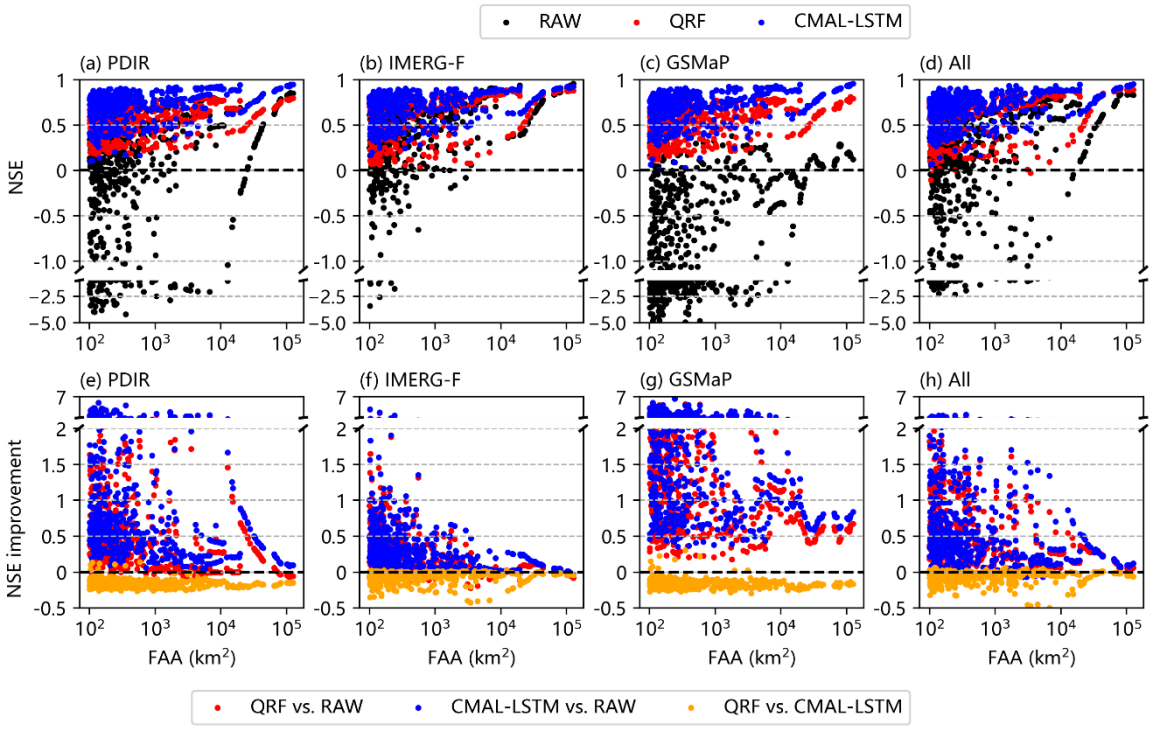



Figure 10. The relationship between (a-d) NSE, (e-h) NSE improvement and FAA.

Table S2. The deterministic performance of two post-processing models for different FAA intervals.

| FAA ($10^4$ km) | Number of sub-basins | PDIR | | IEMRG | | GSMAP | | ALL | |
|---|---|---|---|---|---|---|---|---|---|
| | | QRF | CMAL-LSTM | QRF | CMAL-LSTM | QRF | CMAL-LSTM | QRF | CMAL-LSTM |
| < 2 | 476 | 8 | **468** | 37 | **439** | 10 | **466** | 40 | **436** |
| 2–4 | 15 | 0 | **15** | 2 | **13** | 0 | **15** | 2 | **13** |
| 4–6 | 4 | 0 | **4** | 0 | **4** | 0 | **4** | **4** | 0 |
| 6–10 | 13 | 0 | **13** | 0 | **13** | 0 | **13** | 0 | **13** |
| > 10 | 14 | 0 | **14** | 0 | **14** | 0 | **14** | 0 | **14** |

Comment 2.2

Rather than that, the authors mentioned in several locations that the statistics show dependencies on the drainage area. I don't disagree that the patterns are not random (see Figure 6 and 12), but how do the authors explain those patterns? The current descriptions are merely on the appearance of the plots without convincing explanation.

Response: This question is about the interpretation of the results. Probabilistic and deterministic post-processing are two different tasks, so we explain them separately.

The first scenario is the probabilistic post-processing task. Before interpreting the results, it is worth noting that CRPSS is a relative indicator. In fact, the difference between the QRF model and the CMAL-LSTM model is not very significant. This can also be found in the CRPS boxplot (see Fig. 4). The difference between the QRF model and the CMAL-LSTM model is mainly in the handing of extreme samples and temporal features.

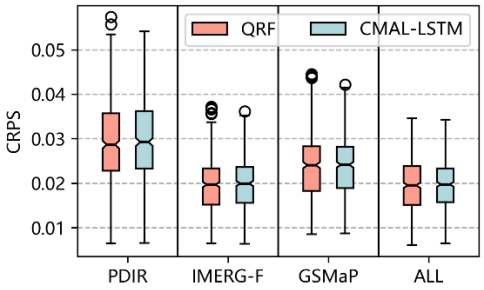
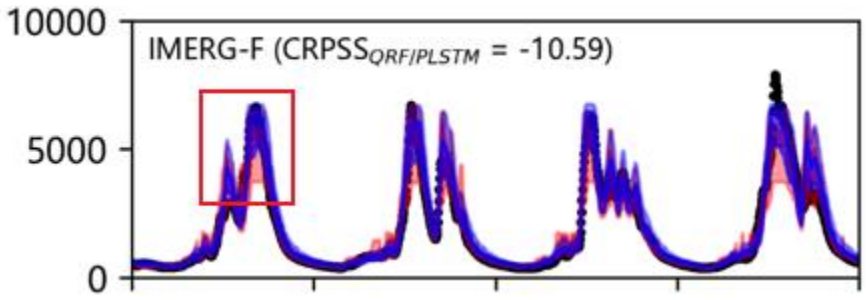



Figure 4. The boxplot of CRPS for different post-processing experiments.

- When using the same input data (e.g., IEMRGF-driven uncorrected streamflow simulation), the differences between model performance are largely attributable to the models themselves. The QRF, a variant of RF, essentially is a decision tree-based classification algorithm. For a given quantile bin, it calculates the MSE between the predicted and actual samples based on a search of historical samples, resulting in outputs for each corresponding quantile. Therefore, for cases where the drainage area is large. The sample of extreme events is small, the prediction interval obtained by the QRF model is narrow (For sub-basin No.10, QRF: $DIS_{25\text{-}75}$=596.8 $m^3/s$; CMAL-LSTM: $DIS_{25\text{-}75}$=676.4 $m^3/s$) and underestimates the flow peak (For sub-basin No.10, QRF: $CO_{25\text{-}75}$=28.8%; CMAL-LSTM: 33.0%). The QRF model only set 100 fixed quantiles. The CMAL-LSTM model firstly samples from the mixed distribution function and then set 100 quantiles. The CMAL-LSTM model infers much wider prediction interval. Last, we used a 10-day time series as features. The QRF model uses time embedding to stack the data and therefore cannot learn more data dynamics. The CMAL-LSTM model is able to learn more data dynamics because of its "gate" functions. In sub-basins with larger drainage area with high autocorrelation skill (e.g., sub-basin No.10), the CMAL-LSTM model architecture can perform better.



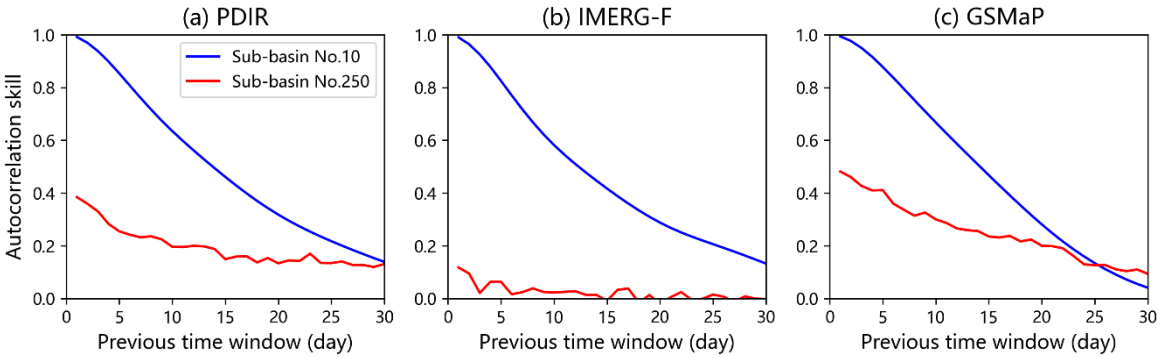



Figure S6. Autocorrelation skill for two randomly selected sub-basins (No.10 and No.250) of different satellite precipitation driven simulation. (a) PDIR, (b) IMERG-F, and (c) GSMaP.

- When using different input data (e.g., PDIR-driven vs. GSMaP-driven uncorrected streamflow simulations), the quality of input data determines the model performance. Compared to GSMaP, the PDIR-driven uncorrected streamflow simulations are less autocorrelated, although it has a higher NSE value.

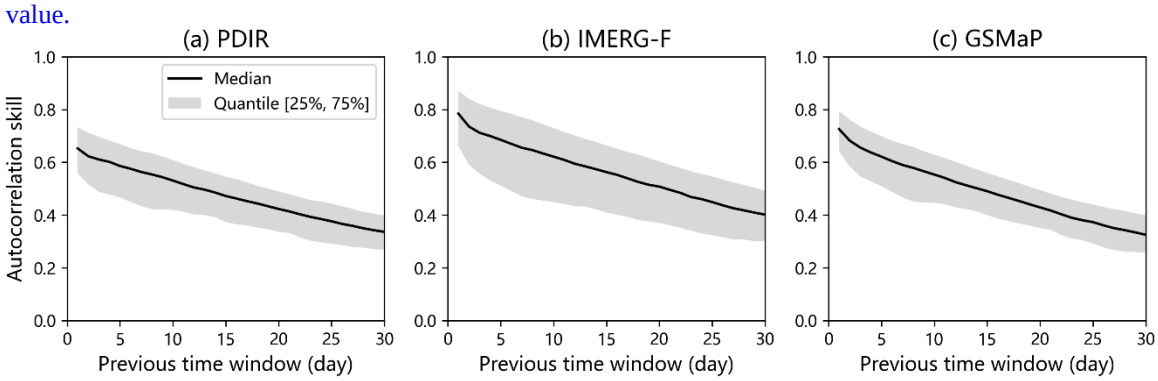



Figure S2. Streamflow autocorrelation skill with different previous time window.

The second scenario is the deterministic post-processing task.

- The main metric we use is the NSE. Due to the presence of the squared term in the NSE formula, the simulation error in the flood peak leads to worse NSE values. The model differences between QRF and CMAL-LSTM are manifested in the treatment of flow peaks and the data dynamics. The limitations of the QRF model for temporal features also result in its worse PCC performance ($PCC_{QRF} < PCC_{CMAL\text{-}LSTM}$).
- For sub-basins with different drainage areas, since the NSE improvement we used is an absolute value, the difference in model performance correlates with the performance of uncorrected streamflow simulations (raw NSE). If the raw NSE is poor, there is a lot of room for improvement in NSE; if the raw NSE is high, there is little room for improvement in NSE.

We have condensed the above interpretation and added it to the discussion section (Sect. 5.1 Model comparison). But we have to admit that the current analysis is still rather superficial. Difficulties in explaining more general patterns may also result from basin imbalances. We will continue to explore related issues by using large-sample hydrology dataset (e.g., Caravan in Kratzert et al., 2023) in future research.

*5.1 Model comparison*

*Previous studies have demonstrated that the quantile regression forests (QRF) approach outperforms other quantile-based models, such as quantile regression and quantile neural networks (Taillardat et al., 2016; Tyralis et al., 2019; Tyralis and Papacharalampous, 2021). Additionally, recent research has indicated the effectiveness of mixture density networks based on the countable mixtures of asymmetric Laplacians models and long short-term memory networks (CMAL-LSTM) for hydrological probabilistic modelling (Klotz et al., 2022). In terms of reliability and sharpness*

*evaluation for probabilistic prediction, CMAL-LSTM has been proven to achieve the best results compared to other models such as LSTM coupled with Gaussian mixture models, uncountable mixtures of asymmetric Laplacians models, and Monte Carlo dropout. These findings suggest that currently, QRF and CMAL-LSTM may be the most effective machine learning and deep learning model for hydrological probabilistic modelling. In this study, we conducted a comprehensive evaluation of the performance of these two advanced data-driven models in the context of streamflow probabilistic post-processing.*

*Our findings suggest that the QRF model outperformed the CMAL-LSTM model in terms of probability prediction in most sub-basins. And the performance difference between the two models was found to be associated with the catchment area of the sub-basins. The QRF model was superior in sub-basins with smaller catchment area, while the CMAL-LSTM model demonstrated better performance in larger sub-basins. However, when evaluated from a deterministic standpoint, the CMAL-LSTM model achieved higher NSE scores than the QRF model across nearly all sub-basins. The authors believe that the primary reason for the disparity in model performance is due to the differences in their respective model structure. As illustrated in Fig 2, the QRF model and the CMAL-LSTM model have dissimilar probabilistic procedure.*

*First, the QRF model and the CMAL-LSTM model differ in their treatment of input features. Specifically, the QRF model utilizes time embedding to flatten time-series features as input for the model. In contrast, the CMAL-LSTM model is capable of better learning the temporal autocorrelation of input features due to the inherent time-series learning capabilities of LSTM. As a result, the CMAL-LSTM model is more responsive to the autocorrelation of uncorrected streamflow features compared to the QRF model. The results depicted in Fig. S6 in the supplement provide evidence to support the interpretation that the performance difference between the QRF model and the CMAL-LSTM model is related to the autocorrelation of input features. The CMAL-LSTM model performs better in the sub-basin No. 250, where streamflow feature autocorrelations are more skillful, than in sub-basin No. 10, where streamflow feature autocorrelation skills are lacking.*

*Second, the QRF model and CMAL-LSTM differ in how they generate probabilistic members. The QRF model calculates the final probabilistic members by grouping them based on a predetermined number of quantiles (100 in this study). In contrast, the CMAL-LSTM model first specifies the form of the probabilistic distribution, then learns the parameters of the distribution using neural networks, and finally obtains the final probabilistic members by sampling. The QRF model produces an approximate and implicit probabilistic distribution, while the CMAL-LSTM model produces an accurate and explicit probabilistic distribution. Moreover, the predicted distribution from the CMAL-LSTM model using the mixture density function is more flexible. As a result, the QRF model produces narrower prediction intervals compared to the CMAL-LSTM model as is reported in Table 3. This is especially true when the sub-basin catchment area is smaller, and the streamflow amplitude is lower. This also explains the reason that the QRF model has higher sharpness in these cases compared to the CMAL-LSTM model. Figure. S7 presents the hydrograph and prediction intervals in two randomly selected sub-basins as an example. In sub-basin No.10, the CMAL-LSTM model achieves a balance between the width of the prediction interval and the observation coverage, which is more important for high-flow predictions and also explains why the CMAL-LSTM model has a higher CRPS value in the sub-basin with larger catchment area. In contrast, although the prediction interval of the QRF model is narrower, it is affected by systematic bias. For example, IMERG-F-QRF underestimates the peak flow in the high-flow season, leading to its smaller CRPS value compared to the CMAL-LSTM model. For sub-basin No.250 with a smaller catchment area, its rainfall-runoff response is faster, and the fluctuation of streamflow is greater. Localized precipitation events can also cause large pulse flow, which is the main feature of flash floods. Therefore, there are relatively more extreme samples. In this case, the QRF model learns and captures more observations with narrower prediction intervals, resulting in a better CRPS value.*

*Third, the QRF model and CMAL-LSTM model differ in their inference process. The QRF model utilizes a decision tree model as its base learner, which is a classification algorithm based on historical searches. Whereas, the CMAL-LSTM model uses a neural network with LSTM layer as its base learner, which is a more powerful fitting model. Due to the differences in model structure, the*

*two models have different abilities to handle extreme events. When extreme event samples are limited, the QRF model tends to underestimate predictions due to its historical search-based approach. On the other hand, the CMAL-LSTM uses the mixture density function for extrapolation. However, both post-processing models still underestimates streamflow extreme events. The QRF model exhibits a higher degree of underestimation in sub-basins with larger catchment areas, resulting in unsatisfactory performance compared to the CMAL-LSTM model in these regions. These discrepancies also lead to lower NSE scores for the QRF model across all sub-basins, as the squared term in the NSE metric increases the sensitivity to high-flow processes which is reported in Fig. S8 in the supplement.*

*Furthermore, besides examining the differences in model performance, we investigated the effects of different input features on the post-processing model by using three different satellite precipitation products in this study. We observed a cascading impact on model performance in the rainfall-runoff and post-processing process. Given a fixed hydrological model, in areas with a small catchment area, the response of streamflow to precipitation is quicker, and the quality of satellite precipitation products directly influences the quality of streamflow prediction through the rainfall-runoff process. The temporal correlation of satellite precipitation determines the temporal correlation of streamflow prediction. Deviations in satellite precipitation led to the biased streamflow prediction, which have a more significant effect on the NSE score of streamflow prediction. This explains the reason that IMERG-F is optimal and PDIR is superior to GSMaP. During the transfer process from raw streamflow to post-processed streamflow, the autocorrelation skill of the raw runoff dictates the performance of the streamflow post-processing model. This clarifies why IMERG-F is still optimal, but GSMaP is superior to PDIR. Based on the results of the multi-product experiment, we observed that the post-processing model can learn better features to a larger extent, however, it cannot completely filter out the information that affects the model accuracy. Regrading information filtering, the CMAL-LSTM model surpasses the QRF model. These findings suggest that although streamflow post-processing can enhance model performance, opting for the best quality product is still a prudent decision when multiple precipitation products are available, and it can also save more computing resources. Another strategy is to execute precipitation post-processing before the hydrological model, which can assist the model to better learn the features and ultimately improve model performance.*

Comment 3: Selection of typical sub-basins

The manuscript dedicated two sections (4.2.6 and 4.3.5) for pilot analysis of two sub-basins. However, I can't see a clear reason for having those pilot analyses. Nor do I see any convincing reason supporting that the two sub-basins chosen are "typical". I don't even know the definition of "typical" here. I think the authors need to address what had been shown by rendering such pilot analysis (the necessity of emphasizing analysis of the two sub-basins)? How do those analyses help to tell the story? Besides, please add in the methodology section the criteria of choosing the "typical" sub-basins.

Response: We are very sorry for the misinformation caused by our terminology. Here, we want to show two "random" sub-basins for "typical" results. The selection criteria are based on the value of CRPSS (QRF vs. CMAL-LSTM). For a sub-basin with a larger drainage area, the CMAL-LSTM model outperforms the QRF model; for a sub-basin with a smaller drainage area, the QRF model outperforms the CMAL-LSTM model. Based on the above criteria, we randomly selected sub-basin No.10 and sub-basin No.250. Hydrographs and predict intervals are more familiar to hydrologists and tend to be chosen in most streamflow simulation studies. Hydrographs and predict intervals can also be used as a diagnostic plot to discover patterns in a long-term streamflow time series as an additional perspective to complement the interpretation of the integrated indicators.

Considering the completeness of the main text's results, we have moved these two sections to the supplement.

Comment 4: Composition of the discussion section

I found the current discussion section superficial, lacking the in-depth explanations on some critical observations from the result section. For example, in section 5.2, the authors mentioned that "In their study, the CMAL-LSTM model achieved the best model performance, which is why we chose it.". But what was used in this study is PLSTM, not CMAL-LSTM by Klotz et al. (2022). Another example is the use of global vs. local models in section 5.3. The authors explain that they chose to train local models because they have limited computational resources. I don't against either training one global model or several local models; I think it is just the choice of the users. But I found this content irrelevant to the science of this study. In my opinion, there are several observations from the result section that are worthy to explain. First, on the hydrological model performance, why are the headwater and the downstream catchments show worse performance than the other catchments (Figure 3)? Why is the gauge-adjusted GSMaP worse than the near-real-time PDIR? Second, on the relative performance between QRF and PLSTM, why is QRF better than PLSTM in the probabilistic evaluation but the reversed situation shown in the deterministic evaluation? Third, on the dependency between metrics and drainage area. As it was mentioned in the previous comment, the patterns are not random. How do we interpret those patterns? Besides, I think it is too much to dedicate two sub-section (section 5.4 and 5.5) on future research directions. Consider merging them and making the writing concise. Discuss something valuable from your results rather than some general facts from literature.

Response: We appreciate your suggestions and tips. We will respond to these points in specific comments below. And we have restructured our discussion section.

Comment 4.1

For example, in section 5.2, the authors mentioned that "In their study, the CMAL-LSTM model achieved the best model performance, which is why we chose it.". But what was used in this study is PLSTM, not CMAL-LSTM by Klotz et al. (2022).

Response: Sorry for this misleading information, we used the CMAL-LSTM model. We have replaced the PLSTM with the CMAL-LSTM throughout the text.

Comment 4.2

Another example is the use of global vs. local models in section 5.3. The authors explain that they chose to train local models because they have limited computational resources. I don't against either training one global model or several local models; I think it is just the choice of the users. But I found this content irrelevant to the science of this study.

Response: Thank you for your comment. This is an important issue and many studies have studied the difference between global and local models (e.g., Kratzert et al., 2019; Fang et al., 2022). For this reason, we will shorten it and include it in the methods section to inform the reader of our choice.

> *Our computing platform is a workstation configured with an Intel(R) Xeon(R) Gold 6226R CPU @ 2.9GHz and an RTX3090 GPU with 24G video memory. It is worth noting that we modelled each sub-basin separately due to the random sampling process of the CMAL-LSTM model exceeding the GPU's video memory. For consistency, the QRF model was also modelled locally.*

Comment 4.3

In my opinion, there are several observations from the result section that are worthy to explain. First, on the hydrological model performance, why are the headwater and the downstream catchments show worse performance than the other catchments (Figure 3)? Why is the gauge-adjusted GSMaP worse than the near-real-time PDIR?

Response: Thank you for your question. When using the same calibrated hydrological model, the quality of the precipitation product determines the performance of the streamflow simulation. Using observed precipitation as a reference, we calculated spatial metrics (Pearson correlation coefficient, PCC and Relative bias, RB) of three satellite precipitation products (see Fig. S4 below).

Compared to PDIR and GSMaP, IMERG has both higher PCC and lower RB values. This explains its higher NSE values. Compared to PDIR, GSMaP suffers from larger biases (RB), resulting in its worse performance.

Although GSMAP is a bias-corrected product, it is not guaranteed that it is superior to the near-real-time PDIR. This is because PDIR uses more advanced precipitation retrieval algorithms (Nguyen et al., 2020, 2021). The Precipitation Estimations from Remotely Sensed Information using Artificial Neural Networks (PERSIANN) Dynamic Infrared-Rain rate model (PDIR) utilizes climatological data to construct a dynamic cloud-top brightness temperature (Tb)——rain rate relationship. The algorithm is a machine learning method and uses historical observations to calibrate the model parameters during training process. No additional observations are required in the prediction period, so it is a near real-time product.

Please see detailed response to comment 1.1.

Second, on the relative performance between QRF and PLSTM, why is QRF better than PLSTM in the probabilistic evaluation but the reversed situation shown in the deterministic evaluation? Third, on the dependency between metrics and drainage area. As it was mentioned in the previous comment, the patterns are not random. How do we interpret those patterns?

Response: Thank you for your question. The prediction interval obtained by the QRF model is narrow, resulting its better probabilistic performance. But the QRF model underestimates the flood peak, resulting its worse deterministic performance. For sub-basin with larger drainage area, high-flow process plays high contribution in performance evaluation, which the CMAL-LSTM model performed better than the QRF model.

Please see response to comment 2.2

Comment 4.4

Besides, I think it is too much to dedicate two sub-section (section 5.4 and 5.5) on future research directions. Consider merging them and making the writing concise. Discuss something valuable from your results rather than some general facts from literature.

Response: Thank you for your suggestions. We have restructured the discussion section.

*5.1 Model comparison*

*5.2 Limitations and future work*

Comment 5: Presentation of materials and writing of the manuscript

Comment 5.1

I think the structure of the sections needs to be improved. Both the methodology and the result sections reach three hierarchical levels. Some sub-sections just have one paragraph. I can see a clear room to make the structure more concise by limiting it to only two hierarchical levels (see my detailed writing tips in the annotated manuscript).

Response: Thank you for your suggestions and tips for our manuscript. Based on your suggestions, we have restructured the manuscript sections.

*3 Methodology*

*3.1 Streamflow reference and uncorrected streamflow simulations*

*3.2 Post-processing model and experimental design*

*3.3 Performance evaluation*

*3.3.1 Probabilistic (multi-point) metrics*

*3.4.2 Deterministic (single-point) metrics*

*4 Results*

*4.1 Uncorrected streamflow simulations*

*4.2 Probabilistic (multi-point) assessment*

*4.2.1 CRPS overall performance*

*4.2.2 The relationship between model performance and flow accumulation area*

*4.2.3 Reliability and sharpness*

*4.3 Deterministic (single-point) assessment*

*4.3.1 Overall model performance*

*4.3.2 The relationship between model performance and flow accumulation area*

*4.3.3 High-flow, low-flow, and peak timing*

Comment 5.2

In addition to that, writing of the manuscript needs to be improved significantly. I can identify grammatical issues and sentences with bad structure. Please pay specific attention to the tense (past vs. present), the use of articles, the use of plural vs. singular form, and the rules of using acronyms.

Response: Thank you very much for your correction and careful review. We have carefully checked the text and correct errors. And, we have asked a native speaker to double-check our manuscript thoroughly.

Comments in the annotated manuscript (selected)

L125: “time scale and step size” Not sure what is the difference? Do you mean time resolution?

Response: Sorry for the misleading information, here we mean computational step size. “time scale” and “step size” were repeated, we will delete “time scale”.

> *Following the watershed division method of Du et al. (2017), Yalong River basin is divided into 522 sub-basins with catchment area ranging from 100 km2 to 127,164 km2 (Fig. 1b). The key to sub-basin delineation is the minimum catchment area threshold (100 km2 in this study), which is related to the total area of the basin, the model architecture complexity, the step size and the spatial resolution of the input data.*

L175: “Figure 2. The framework of this study.” Why do we need the different background colors?

Response: Thank you for your question. Different background colors are used to distinguish different sections. Due to the presence of the boxes, it is ok to delete them.

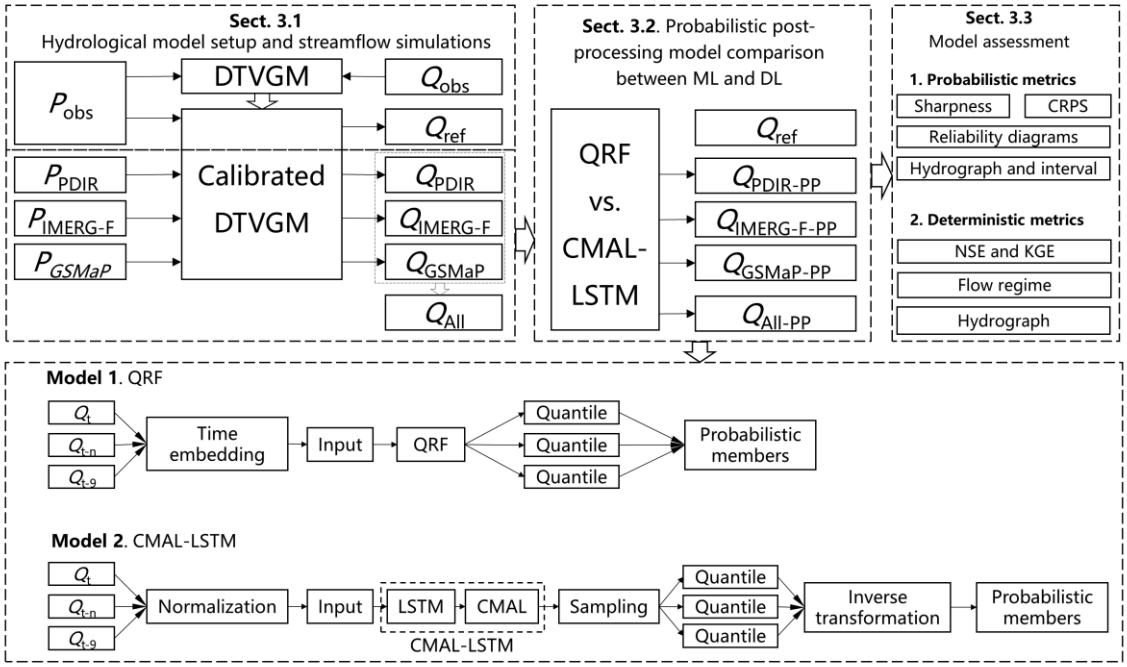



*Figure 2. Framework of this study.*

L184: "runoff is calculated according to water balance." Could this be more specific? I think nearly all hydrological model calculate runoff based on water balance.

Response: Runoff is calculated according to below equation. We have added it in the text.

$$P_t + AW_t = AW_{t+1} + g_1\left(\frac{AW_{u,t}}{C \cdot WM_u}\right)^{g2} \cdot P_t + K_r \cdot AW_{u,t} + K_e \cdot EP_t + K_g \cdot AW_{g,t}$$

*where t is the time step; P is the precipitation (mm); EP is the potential evapotranspiration (mm); AW and WM are soil moisture (mm) and field soil moisture (mm), respectively; u and g are the upper and lower soil layers, respectively; Ke, Kr and Kg are evapotranspiration, interflow and groundwater runoff coefficients, respectively; g1 and g2 are factors describing the non-linear rainfall-runoff relationship; and C is the land cover parameter.*

L194: "The us e hydrological model does …. Model structure and model parameters are neglected." I would suggest to delete this part for two reasons. First, I think the NSE values are acceptable. Second, the writing here is not convincing, especially the first sentence "... but we believe..., so...". There is no "we believe" in science. Every statement needs support. I think a 0.59 of NSE is sufficient to prove that the model is acceptable. Deleting this part could be a wiser choice in my opinion.

Response: Thank you for your suggestion. We agree with you that deleting this part could be a wiser choice, and we have deleted them.

L218: "Three steps are required to implement … get the final prediction." I suggest to delete this part. It is too detailed. This is an application of the well-known RF method, not a modification of the algorithm. So, there is no need to go to those well-known details of RF.

Response: Thank you for your suggestion. We have rewritten this part to shorten the introduction of the two post-processing models.

*The two post-processing models selected are the QRF model (Meinshausen and Ridgeway, 2006) and the CMAL-LSTM model (Klotz et al., 2022). The QRF model was chosen because it enables us to analyse the distribution of the entire data based on different quantiles, and it has been previously used*

*in several studies (Taillardat et al., 2016; Evin et al., 2021; Kasraei et al., 2021; Tyralis et al., 2019; Tyralis and Papacharalampous, 2021). The CMAL-LSTM model is a combination of an LSTM model and a CMAL mixture density function, which allows it to estimate prediction uncertainty. To the best of our knowledge, these two models currently considered state-of-the-art in ML and DL for hydrological probabilistic modelling. Detailed information about each model can be found in their original papers and will not be restated in this study.*

L293: Eq. 1. I found that the form here is not the same as the one in Bröcker (2012). Why?

Response: Thank you for pointing it out. We have fixed this issue.

L613: "We believe … more predictors" I don't think this can help. The key is to have some extreme observations.

Response: We agree with you that having more extreme observations could be helpful. However, learning the rainfall-runoff process through predictors to achieve more accurate inference is also a direction to be explored. We have rephrased this sentence.

*We believe that collecting more data samples and introducing additional predictors and distribution functions for extreme events can lead to further improvements.*

L619: "standard dataset" How do you define "standard dataset"? What criteria need to be satisfied?

Response: Sorry for the misleading terminology. What we want to show here is that our dataset is reusable for others. We have rephrased this sentence.

*Using observed precipitation and three different satellite precipitation products to drive the calibrated hydrological model, we generated a large-sample dataset of 522 sub-basins with paired streamflow reference and biased streamflow simulations.*

L622: "In conclusion, decision-tree …" This is a general fact of ML model, not the conclusion of this study.

Response: Thank you for the comment. We have rewritten our conclusion part.

*In this study, a series of well-designed experiments to compare the performance of two state-of-the-art models for streamflow probabilistic post-processing were conducted: a machine learning model (quantile regression forests) and a deep learning model (countable mixtures of asymmetric Laplacians long short-term memory network). Using observed precipitation and three different satellite precipitation products to drive the calibrated hydrological model, we generated a large-sample dataset of 522 sub-basins with paired streamflow reference and biased streamflow simulations. We evaluated the model performance from both probabilistic and deterministic perspectives, including reliability, sharpness, accuracy, and flow regime, through intuitive case studies. These experiments established a path for understanding the model differences in probabilistic modelling and post-processing, provided practical experience for model selection, and extracted insights for model improvement. It also serves as a reference for establishing benchmark tests for model evaluation, including dataset construction and metrics selection. Furthermore, streamflow post-processing provides dependable data support for a range of downstream tasks, such as flood risk analysis, reservoir scheduling, and water resource management. The empirical findings of this study for the two post-processing models are summarized below.*

*(1) Based on the probabilistic assessment, the QRF and CMAL-LSTM models exhibit comparable performance. However, their model differences are correlated with the flow accumulation area (FAA) of sub-basins. In cases where the catchment area of a sub-basin is small, the QRF model generates a narrower prediction interval, resulting in better CRPS scores compared to the CMAL-LSTM model in most sub-basins. Conversely, for larger sub-basins (over 60,000 km2 in this study), the CMAL-LSTM model outperforms the QRF model due to its ability to learn autocorrelation skills of features and capture extreme values.*

*(2) Based on the deterministic assessment, it can be concluded that the CMAL-LSTM model performs better than the QRF model in capturing high-flow process and flow duration curve. On the other hand, the QRF model tends to underestimate the high-flow process, resulting in worse NSE score across all sub-basins. Both models, however, have the issue of underestimating flood peaks due to sparse samples of extreme events.*

*(3) For the input uncertainties introduced by the different satellite precipitation products, both models are able to reduce their impact on the streamflow simulation. However, the performance of the post-processing models does not improve further in the multi-product experiments. Instead, the inclusion of heavily biased inputs leads to a deterioration in model performance. Opting for a single precipitation product that is best suited to the task at hand is a more prudent approach to safeguard model performance and minimize computational cost, rather than using multiple precipitation products with varying degrees of quality.*

*(4) Given the performance of post-processing models, the author believes they have the potential to be applied to other sources of uncertainty that affect hydrological modelling, such as model structure and parameter uncertainty.*

Other technical corrections

Response: Thank you very much for your comments, suggestions and writing tips. We have adapted all proposed issues.

Reference


Ajami, N. K., Duan, Q., & Sorooshian, S. (2007). An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction. Water Resources Research, 43(1). https://doi.org/10.1029/2005WR004745

Candille, G., & Talagrand, O. (2005). Evaluation of probabilistic prediction systems for a scalar variable. Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography, 131(609), 2131-2150.

Fang, K., Kifer, D., Lawson, K., Feng, D., & Shen, C. (2022). The data synergy effects of time-series deep learning models in hydrology. Water Resources Research. https://doi.org/10.1029/2021WR029583

Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69(2), 243-268. https://doi.org/10.1111/j.1467-9868.2007.00587.x

Huang, Z., & Zhao, T. (2022). Predictive performance of ensemble hydroclimatic forecasts: Verification metrics, diagnostic plots and forecast attributes. Wiley Interdisciplinary Reviews-Water, 9(2). https://doi.org/10.1002/wat2.1580

Jolliffe, I. T., & Stephenson, D. B. (Eds.). (2012). Forecast verification: a practitioner's guide in atmospheric science. John Wiley & Sons.p210

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. Hydrology and Earth System Sciences, 23(12), 5089-5110. https://doi.org/10.5194/hess-23-5089-2019

Kratzert, F., Nearing, G., Addor, N., Erickson, T., Gauch, M., Gilon, O., Gudmundsson, L., Hassidim, A., Klotz, D., Nevo, S., Shalev, G., & Matias, Y. (2023). Caravan - A global community dataset for large-sample hydrology. Scientific Data, 10(1). https://doi.org/10.1038/s41597-023-01975-w

Murphy, A. H. (1993). What is a good forecast? An essay on the nature of goodness in weather forecasting. Weather and Forecasting, 8(2), 281-293. https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2

Nguyen, P., Shearer, E. J., Ombadi, M., Gorooh, V. A., Hsu, K., Sorooshian, S., Logan, W. S., & Ralph, M. (2020). PERSIANN Dynamic Infrared–Rain Rate Model (PDIR) for High-Resolution, Real-Time Satellite Precipitation Estimation. Bulletin of the American Meteorological Society, 101(3), E286-E302. https://doi.org/10.1175/BAMS-D-19-0118.1

Nguyen, P., Ombadi, M., Gorooh, V. A., Shearer, E. J., Sadeghi, M., Sorooshian, S., Hsu, K., Bolvin, D., & Ralph, M. F. (2020). PERSIANN Dynamic Infrared–Rain Rate (PDIR-Now): A Near-Real-Time, Quasi-Global Satellite Precipitation Dataset. Journal of Hydrometeorology, 21(12), 2893-2906. https://doi.org/10.1175/JHM-D-20-0177.1

Nikolopoulos, E. I., Anagnostou, E. N., Hossain, F., Gebremichael, M., & Borga, M. (2010). Understanding the Scale Relationships of Uncertainty Propagation of Satellite Rainfall through a Distributed Hydrologic Model. Journal of Hydrometeorology, 11(2), 520-532. https://doi.org/10.1175/2009JHM1169.1

Troin, M., Arsenault, R., Wood, A. W., Brissette, F., & Martel, J. L. (2021). Generating Ensemble Streamflow Forecasts: A Review of Methods and Approaches Over the Past 40 Years. Water Resources Research, 57(7). https://doi.org/10.1029/2020WR028392

Response to Referee #3

This is a well-written manuscript. This reviewer only has minor technical comments. See attached for details.

Response: Thank you so much for your kind words and positive feedback. We are committed to continuously improving the quality of our manuscript, and your suggestions are invaluable. Thank you again for taking the time to share your comments and suggestions.

L130: the CMA precipitation gauge stations should be included as well on the map.

Response: Thank you for your suggestion. We have added it to Fig.1.

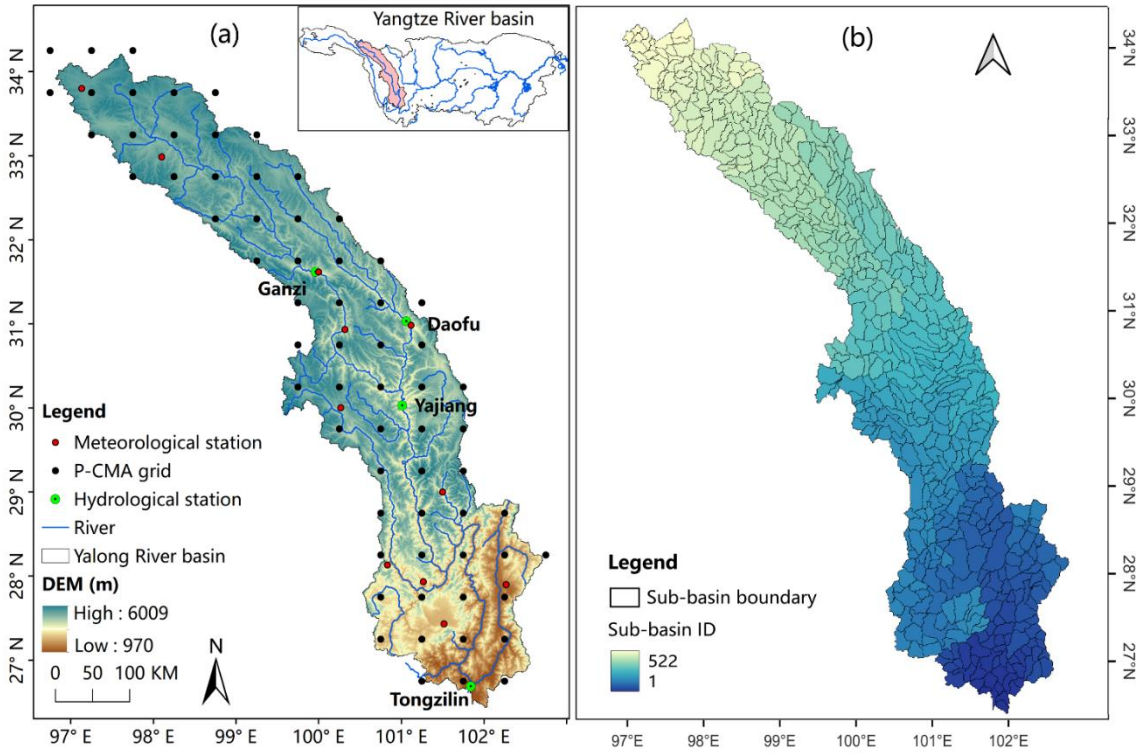



Figure 1. (a) Study area and (b) 522 sub-basins (Zhang et al., 2022a).

L150: “interpolated” is it resampling?

Response: Yes, it is. We have replaced “resample” with “interpolate”.

L160: There are uncertainties among different DEM data. For example, below paper shows that SRTM deviates from the GPS-RTK measurement with min. error of 22m and max. error of 44m over Maqu region, Tibetan Plateau. This may impact the simulated stream flow. Please the author help clarify on this point.

Li, M., Zeng, Y., Lubczynski, M. W., Roy, J., Yu, L., Qian, H., Li, Z., Chen, J., Han, L., Zheng, H., Veldkamp, T., Schoorl, J. M., Hendricks Franssen, H.-J., Hou, K., Zhang, Q., Xu, P., Li, F., Lu, K., Li, Y., and Su, Z.: A first investigation of hydrogeology and hydrogeophysics of the Maqu catchment in the

Yellow River source region, Earth Syst. Sci. Data, 13, 4727–4757, https://doi.org/10.5194/essd-13-4727-2021, 2021

Response: Thank you very much for your suggestion. We have clarified the uncertainty from the DEM data.

> *In the remaining part of this study, the hydrological model is fixed and we mainly post-process the streamflow bias introduced by satellite precipitation, disregarding other sources of uncertainty such as model structure, DEM and other forcing data.*

L577: this is the same as 5.5

Response: Thank you for pointing this out. It was a typo and we have rewritten the discussion section.

L578: are you sure streamflow is the predictor? And not target variable?

Response: Thank you for your question. This study used satellite precipitation-driven uncorrected streamflow simulations as the predictor and ground precipitation-driven streamflow simulations as the target variable. We have specified it.

> *Third, the selection of input features and hydrological models could be extended. In order to maintain model complexity and keep computational costs low, this study only used one variable, uncorrected streamflow, as the predictor.*