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.

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.

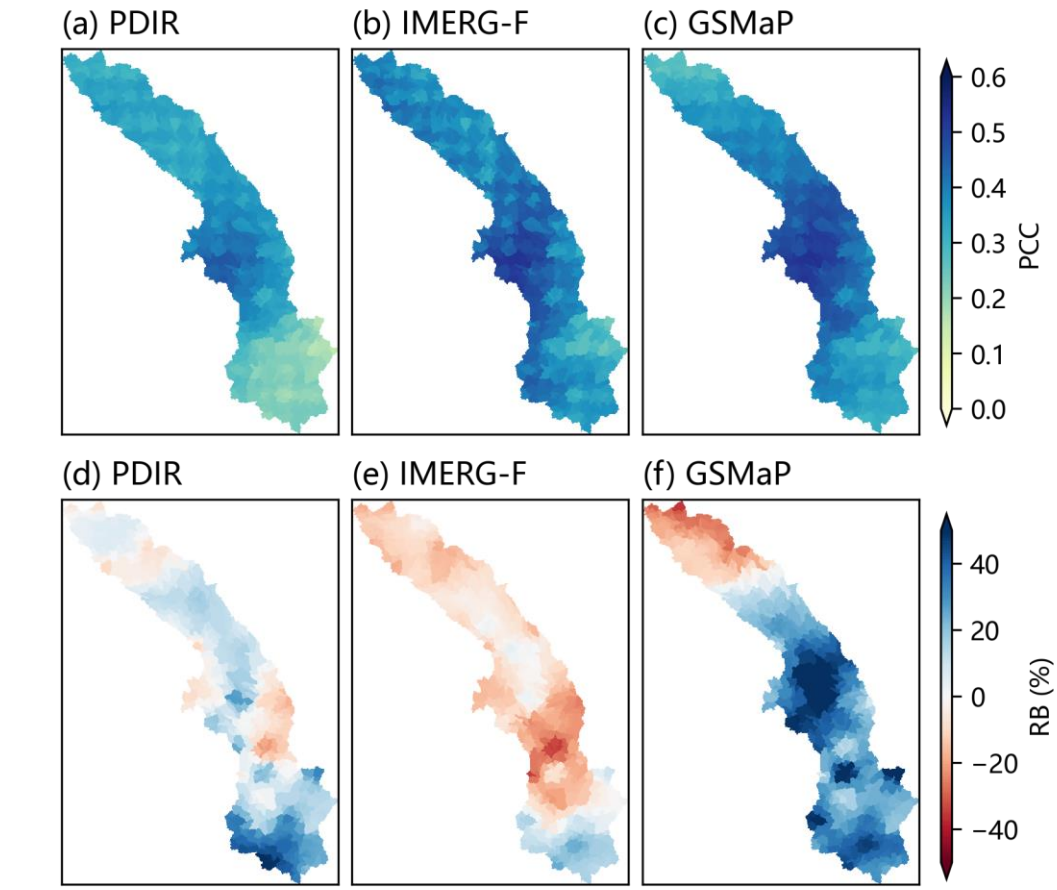


Figure S4. The PCC and RB of three satellite precipitation estimations for 522 sub-basins.

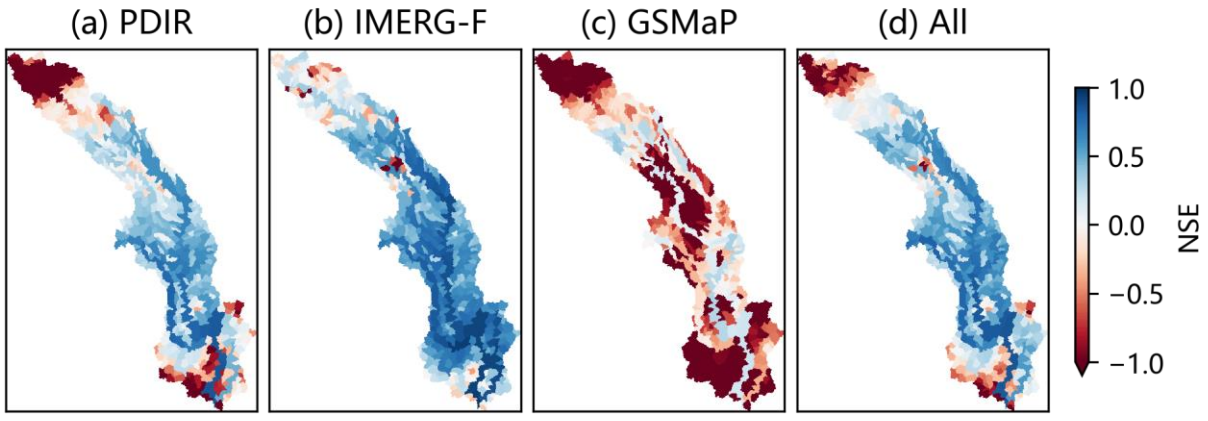


Figure 3. The NSE of uncorrected streamflow simulations for 522 sub-basins.

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 measure 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.

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 will only show the CRPS result in the main text and move the twCRPS results to supplementary material.

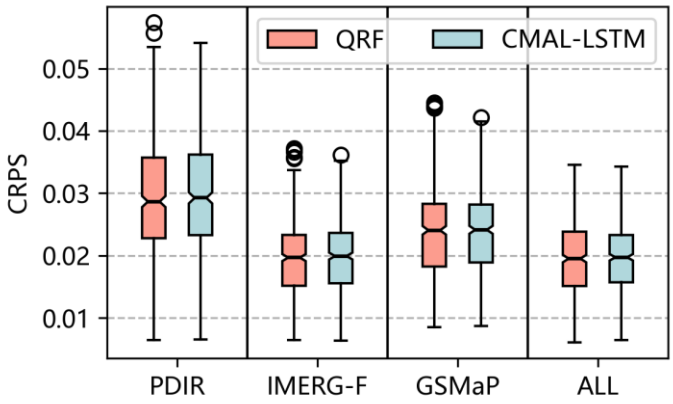


Figure 4. The boxplot of CRPS for different post-processing experiments.

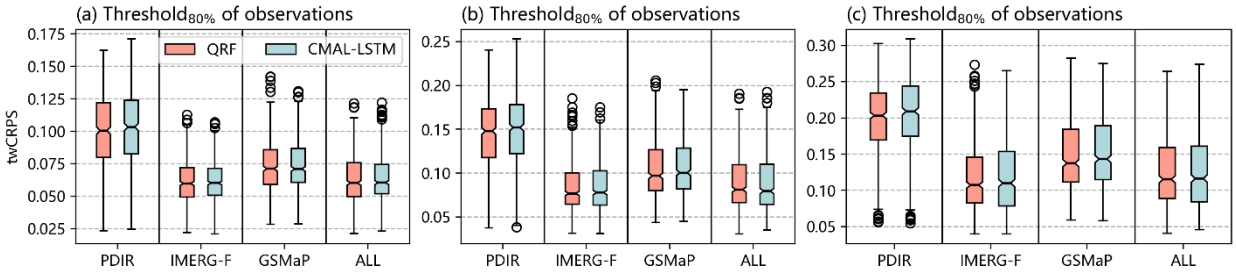


Figure S5. The boxplot of twCRPS for different post-processing experiments.

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 figure below). 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.

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.

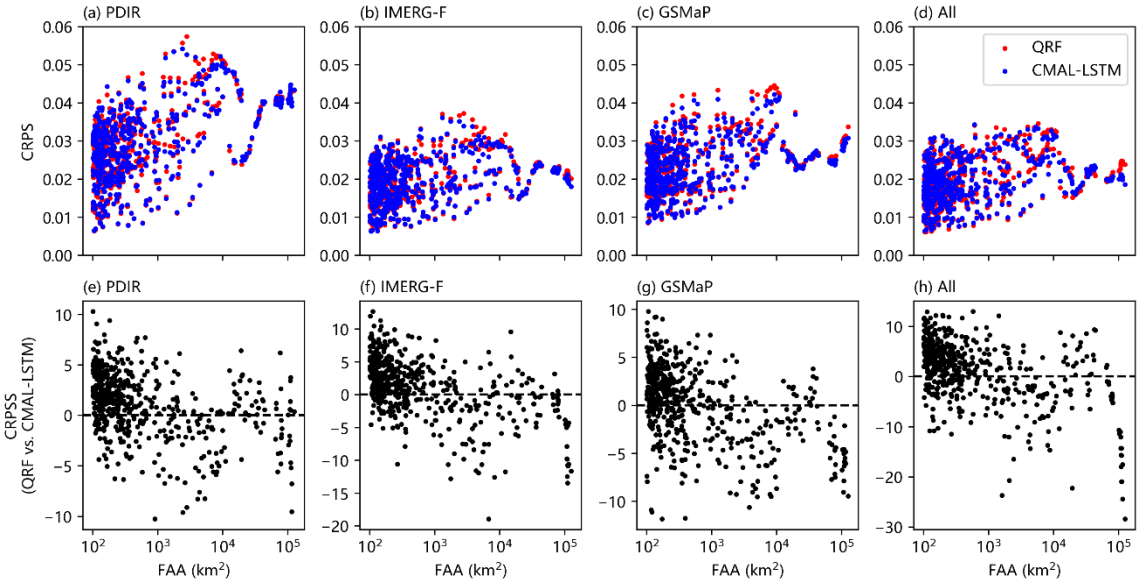


Figure 6. The relationship between (a-d) CRPS, (e-h) CRPSS and FAA.

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

FAA	Number of sub-basins	PDIR		IEMRG		GSMAP		ALL	
		QRF	CMAL-LSTM	QRF	CMAL-LSTM	QRF	CMAL-LSTM	QRF	CMAL-LSTM
0-2	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	14	7	7	0	**14**	0	**14**	0	**14**

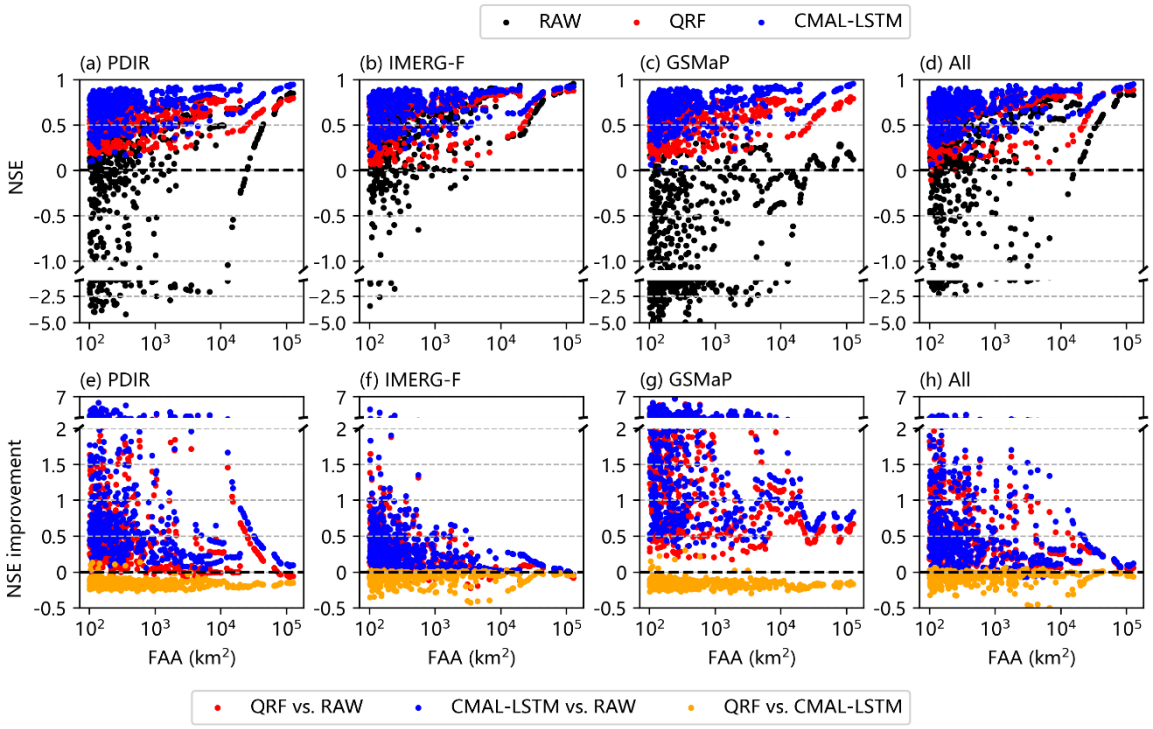


Figure 11. The relationship between (a-d) NSE, (e-h) NSE improvement and FAA.

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

FAA	Number of sub-basins	PDIR		IEMRG		GSMAP		ALL	
		QRF	CMAL-LSTM	QRF	CMAL-LSTM	QRF	CMAL-LSTM	QRF	CMAL-LSTM
0-2	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	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.

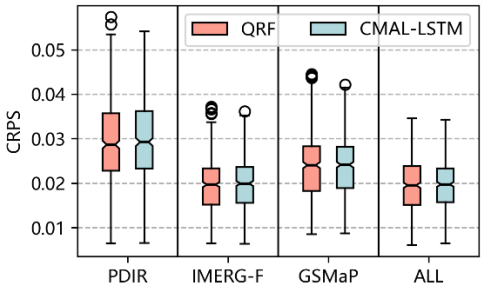
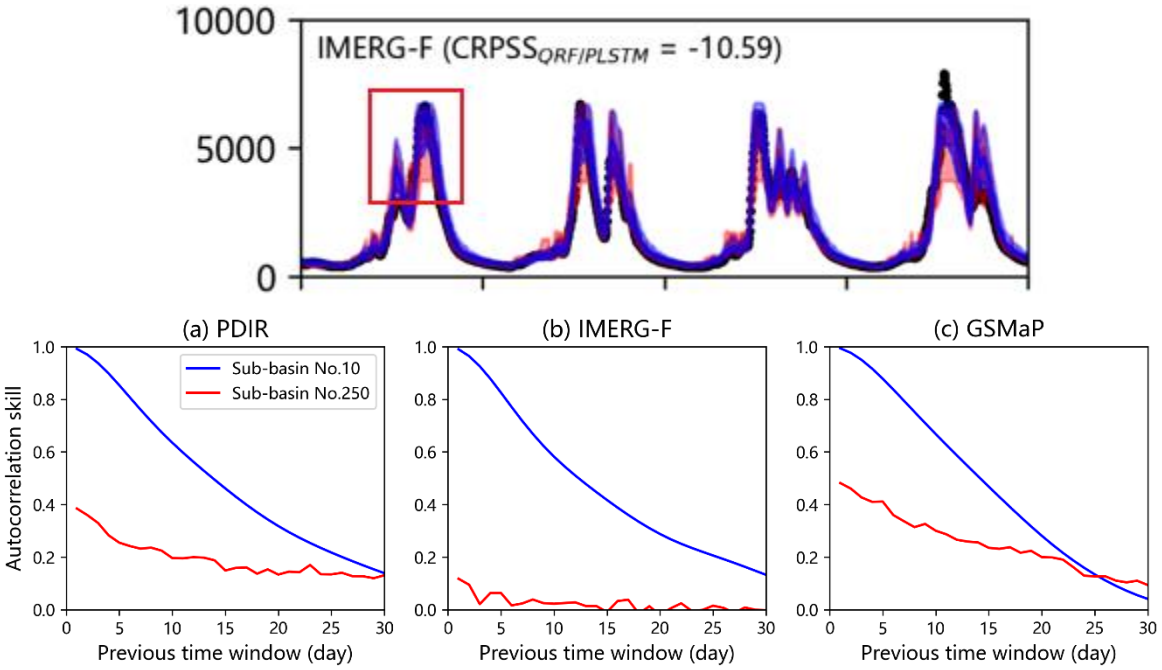


Figure 4. The boxplot of CRPS for different post-processing experiments.

- 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.



- 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.

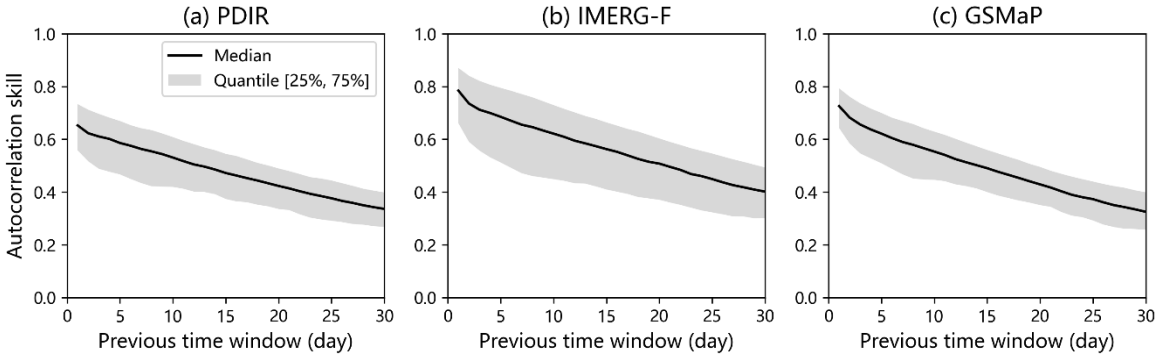


Figure S2. Streamflow autocorrelation skill with different previous time window.

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 will condense the above interpretation to add it to the discussion section. But we have to admit that the above 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.

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 (e.g., CRPSS and NSE).

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.

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 will replace 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.

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. Please see 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. 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 will restructure the discussion section.

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 will restructure the manuscript sections.

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 will carefully check the text and correct errors. And, we will ask 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”.

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.

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 will add 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 will do so.

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 will rewrite this part to shorten the introduction of the two post-processing models.

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 will rewrite this equation.

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.

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 will use “paired simulation-observation data pairs” instead of “standard dataset”.

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 will rewrite it to make it more relevant to this study.

Other technical corrections

Response: Thank you very much for your comments, suggestions and writing tips. We will adapt 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