Dear Editor Prof. Romano and anonymous reviewers:

We are grateful for your consideration of our manuscript entitled “Comparing quantile regression forest and mixture density long short-term memory models for probabilistic post-processing of satellite precipitation-driven streamflow simulations" [HESS-2022-377], and we also very much appreciate your constructive comments and continued 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.

Section 1: Response to Editor

Dear Authors,

The two experts are willing to evaluate even your revised version and I would exploit this opportunity allowing for the concerns raised during the previous step of evaluations.

Response: Thank you for sharing the evaluation, continued guidance and feedback regarding our manuscript. We appreciate the time and effort both you and the reviewers have invested in evaluating our work. We have double-checked our revised manuscript and hope that it addresses your concerns and those of the reviewers.

Dear authors,

Despite the significant improvement in your article with the revised version, it cannot yet be accepted for final publication in HESS. Ref.#1 was quite satisfied with the new version and did not have any further comments, even though is willing to consider another revised version. Ref.#2 raised additional concerns that were not just based on your responses to the previous stage. I am in complete agreement with the additional comments and suggestions of Ref.#2 that require more improvements, and I encourage you to re-arrange your manuscript accordingly. Please upload a new version of your article, together with point-by-point replies to the last Ref.#2's comments. If you disagree with some comments, clearly explain why.

Response: Thank you for sharing the evaluation, continued guidance and feedback regarding our manuscript. We appreciate the time and effort both you and the reviewers have invested in evaluating our work.

Section 2: Response to Reviewer #1

Reviewer 1

The authors have addressed my major concern by narrowing down their scope to the comparison between an ML (QRF) and a DL (MD-LSTM) methods. I don't have further comments.

Response: Thank you for taking the time to review our manuscript again and acknowledging the revisions we made in response to your previous comments. We sincerely appreciate your valuable feedback, which significantly contributed to improving the quality and clarity of our work.

Section 3: Response to Reviewer #2

Reviewer 2

I reviewed the initial version of the manuscript for which I gave a positive recommendation for publication after minor changes. As the reviewing process continues, this review aims to provide some possible improvements to the manuscript.

Response: We appreciate your continued involvement and dedication to improving the quality of our manuscript. We are committed to addressing the subsequent comments to ensure our work meets the standards. We hope these revisions will make our research more accessible to readers.

Although I do not consider myself an expert on machine learning methods, I agree with reviewer #2 that the distinction between "Machine Learning (ML) and Deep Learning (DL) algorithms" is far from clear in the introduction. There is a small sentence at l. 86-87 indicating that the authors "use the term "ML models" to refer to non-DL models, while specifically designating " DL models" to refer to models based on deep learning techniques" but it does not help the reader to understand what are the major differences between these families of algorithms.

Response: Thank you for your insightful comments and for bringing our attention to the distinction between Machine Learning (ML) and Deep Learning (DL) algorithms in the introduction again. Considering your feedback, we revised our introduction part upon the differences between ML and DL.

> *Over the past few years, machine learning (ML) and deep learning (DL) algorithms have emerged as powerful tools in hydrological modelling (Sit et al., 2020; Zounemat-Kermani et al., 2021; Shen and Lawson, 2021; Fang et al., 2022). ML comprises a broad range of algorithms, with commonly used models such as random forest, support vector machines, and clustering methods. DL, a specialized subset of ML, emphasizes algorithms modelled after the architecture of artificial neural networks, including models like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks. In this study, we use the term "ML models" to refer to non-DL models, while specifically designating "DL models" to refer to models based on deep learning techniques.*

Apart from this semantic choice, the very long introduction does not help to clarify the motivation of the study. Lines 64-116 mix up different aspects of the literature and the objectives of the study, which makes them particularly hard to understand. Lines 133-139 are also not very informative in this regard at this stage of the manuscript. I strongly recommend reducing these paragraphs and adopting a more consistent presentation (discussion of the literature and then the objectives/materials of the study).

Response: Thank you for pointing out this problem. We have restructured the introduction part to ensure a logical flow of information.

> ***[Basic knowledge and traditional methods]***
>
> *Satellite precipitation introduces notable uncertainties in hydrological modelling. Various strategies, such as meteorological pre-processing and hydrological post-processing, have emerged to address this challenge (Schaake et al., 2007; Wang et al., 2009; Ye et al., 2014, 2015; Li et al., 2017; Dong et al., 2020; Shen et al., 2021; Zhang et al., 2022a). Meteorological pre-processing predominantly focuses on achieving bias-corrected precipitation estimates. This is often realized by fusing satellite precipitation data with ground observations to mitigate input uncertainty (Xu et al., 2020; Zhang et al., 2022a). Conversely, hydrological post-processing leverages observed streamflow to rectify simulations or predictions, providing an added layer of refinement, especially if the meteorological pre-processing stage falls short. Both these strategies can be employed for deterministic and probabilistic predictions (Ye et al., 2014; Tyralis et al., 2019). Given the inherent autocorrelation observed in streamflow time series, two main methods stand out for hydrological post-processing. The first employs autoregressive models anchored on*

residuals, using these residuals as predictors to adjust forecast errors (Li et al., 2015, 2016; Zhang et al., 2018). The second method employs the Model Output Statistics (MOS) concept, leveraging simulated streamflow as a primary predictor to establish statistical relationships between simulations and observations.

[literature and discussion of ML and DL]

Over the past few years, machine learning (ML) and deep learning (DL) algorithms have emerged as powerful tools in hydrological modelling (Sit et al., 2020; Zounemat-Kermani et al., 2021; Shen and Lawson, 2021; Fang et al., 2022). ML comprises a broad range of algorithms, with commonly used models such as random forest, support vector machines, and clustering methods. DL, a specialized subset of ML, emphasizes algorithms modelled after the architecture of artificial neural networks, including models like convolutional neural networks, recurrent neural networks, and long short-term memory networks. In this study, we use the term "ML models" to refer to non-DL models, while specifically designating "DL models" to refer to models based on deep learning techniques. In the hydrological field, both random forest (RF) and long short-term memory (LSTM) models are widely used and considered state-of-the-art approaches for various tasks and applications. The RF model and its probabilistic variant, the QRF model, have demonstrated capabilities in bias correction and streamflow simulation (Shen et al, 2022; Tyralis et al., 2019; Zhang and Ye, 2021). For example, Shen et al. (2022) used the RF model as a hydrological post-processor to enhance the simulation performance of the large-scale hydrological model PCR-GLOBAL (PCRaster Global Water Balance) model at three hydrological stations in the Rhine basin. Tyralis et al. (2019) compared the usability of the statistical model (e.g., quantile regression) and the machine learning algorithm (e.g., quantile regression forests) as hydrological post-processors on the CAMELS (Catchment Attributes and Meteorology for Large-sample Studies) dataset. And the results showed that the quantile regression forests model outperformed the quantile regression. In the context of bias correction applications, RF models have also exhibited superior performance compared to other machine models (Zhang and Ye, 2021). The LSTM model, on the other hand, has gained widespread recognition as leading choice in hydrological applications (Kratzert et al., 2018, 2019). For example, 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). Frame et al. (2021) utilized LSTM to develop a post-processor that can effectively improve the accuracy of the U.S. National Hydrologic Model. They validated the performance of the proposed post-processor on the CAMELS dataset, which consists of 531 watersheds across North American. By integrating with Gaussian models (Zhu et al., 2020), stochastic deactivation of neurons (Althoff et al., 2021), and Bayesian perspective (Li et al., 2021, 2022), LSTM further solidified its reputation for delivering reliable probabilistic predictions. More recently, Klotz et al. (2022) compared the use of dropout and three Gaussian mixture density models for uncertainty estimation in LSTM rainfall-runoff modelling. They found that the mixture density model outperformed the random dropout model and provided more reliable probabilistic information.

[Research gaps]

While both RF and LSTM models have seen significant advancements and widespread application, a thorough comparative analysis specifically within the context of hydrological probabilistic post-processing is yet to be undertaken. Through their hierarchical feature learning, DL models, especially LSTMs, can autonomously extract insights from raw hydrological data, capturing long-term dependencies and patterns without extensive feature engineering. In contrast, with ML models like RF, effort is often required to select relevant features to adequately represent the data. Additionally, DL models can effectively leverage massive datasets, leading to enhanced generalization and improved accuracy in hydrological prediction tasks. On the other hand, ML

models may face limitations in capturing intricate patterns from large hydrological datasets. Notwithstanding pieces of evidence, it is essential to conduct a direct and focused comparison between RF and LSTM models in the specific context of hydrological probabilistic post-processing to better understand their respective strengths and limitations, such as the scope of application, model performance and computational efficiency.

[Objectives, scopes and significance]

Hydrological probabilistic post-processing represents a big-data task with the involvement of large datasets and a substantial number of ensemble members. The complex relationships between input and output variables in hydrological systems necessitate advanced modelling techniques to achieve accurate and reliable predictions. Therefore, in this study, we attempt to comprehensively compare the performance of the two most widely used ML and DL models for streamflow probabilistic post-processing: quantile regression forests (QRF) and countable mixtures of asymmetric Laplacians LSTM (CMAL-LSTM), at a sub-basin scale daily streamflow, respectively. In particular, a full model comparison is performed in a complex basin with 522 nested sub-basins in southwest China. Three sets of global satellite precipitation products are applied to generate uncorrected streamflow simulations. The three precipitation products represent different algorithms. Also, they have been proven to have relatively good accuracy in our previous study (Zhang et al., 2021b). These satellite precipitation products are compared in two scenarios: a single-product simulation and a multi-product simulation, both serving as input features for streamflow post-processing. A variety of evaluation metrics are used to assess the performance of the proposed models, including probabilistic metrics for multi-point prediction and deterministic metrics for single-point prediction. Additionally, the study also analyze the relationship between model performance and basin size by considering the disparity in the flow accumulation area of the sub-basins. Through a comparative analysis of QRF and CMAL-LSTM models in hydrological probabilistic post-processing, this study aims to provide clarity on their respective merits and drawbacks. The insights garnered will also guide the selection of other ML and DL methodologies with similar model architectures.

I also agree with reviewer #2 that the main message concerning the relationship between the performances of the postprocessing methods and the drainage area is not very convincing. The main basis for this statement is provided in Figs. 6 and 10. The differences of performances between QRF and CMAL-LSTM are far from clear when the merged precipitation product “All” is used.

Response: To clearly delineate the differences between the two models, we have presented the results of Figs 6 and 10 as tables within the article and supplement. Table 2 aligns with Fig. 6, while Table S2 mirrors Fig. 10. A distinct contrast between the two models is evident. In the "All" experiment, CMAL-LSTM excels in larger subbasins, whereas QRF is more effective in smaller subbasins. In the PDIR experiment, while the disparity is subtle, it still reveals nuanced distinctions between the models. Notably, the CMAL-LSTM experiment has a more pronounced CRPSS score, suggesting its superiority over the QRF in larger subbasins.

Table 2. The probabilistic performance of two post-processing models for different FAA intervals. The bold numbers indicate better performance in each group.

FAA (10^4 km^2)	Number of sub-basins	PDIR		IMERG-F		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**

Table S2. The deterministic performance of two post-processing models for different FAA intervals. The bold numbers indicate better performance in each group.

FAA (10^4 km^2)	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**

The better performances of CMAL-MSTM for basins larger than 60,000 km2 concern only 27 catchments for which I understand that there is redundancy (i.e. the largest basin of 127,164 km2 encompasses other large basins) but this is poorly described in Section 2.1.

Response: We recognize the concern raised regarding the larger basins. We acknowledge it's true that some of these basins, particularly those exceeding 60,000 km^2, are nested, we've clarified this aspect in the Section 2.1 and other parts of the manuscript using term “nested”. We also mentioned the “Location, elevation, area, flow accumulation area and flow direction of each sub-basin can be found in Table S1.” in our manuscript.

In addition, although some of these basins are geographically nested, each undergoes independent post-processing in our study. Mathematically speaking, their treatments are distinct. Furthermore, our primary focus remains on the relative performance improvement compared to the uncorrected streamflow in each sub-basin.

Regarding the imbalance in the number of subbasins, we also discussed the shortcomings of this part in detail in the discussion. It is hoped that this can be further explored in future research.

> *Second, there exists data imbalance among the studied sub-basins. Among the selected 522 sub-basins, it can be observed that model performance is related to the catchment size. However, the number of sub-basins corresponding to each of the five intervals (100–20,000 km^2, 20,000–40,000 km^2, 40,000–60,000 km^2, 60,000–100,000 km^2, and greater than 100,000 km^2) are 476, 15, 4, 13 and 14, respectively. Only 5.2% of the sub-basins have a catchment area larger than 60,000 km2. This could potentially affect the generality of conclusions drawn. To address this limitation, more extensive and balanced datasets (such as Caravan, Kratzert et al., 2023) are needed to be utilized to achieve further validation of the research findings and a better understanding of different post-processing models.*

Looking at the first lines of Figs 6 and 10, my understanding is that QRF and CMAL-MSTM perform equally for the precipitation inputs “All”.

Response: Indeed, when observing the first lines of Figures 6 and 10, both QRF and CMAL-MSTM exhibit similar performance for the "All" precipitation inputs. While the first lines provide an absolute measure of their performances, it's the relative scores in the second lines that are central to our analysis. Meanwhile, as we mentioned before, we added Tables 2 and S2 to quantify and highlight their relative differences.

Table 2. The probabilistic performance of two post-processing models for different FAA intervals. The bold numbers indicate better performance in each group.

FAA (10^4 km^2)	Number of sub-basins	PDIR		IMERG-F		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**

Table S2. The deterministic performance of two post-processing models for different FAA intervals. The bold numbers indicate better performance in each group.

FAA (10^4 km^2)	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**

In addition, there are a few minor issues that should be fixed:

- L.289-299: F and x are not defined in Eqs. 2-3. It would be helpful to indicate what x is in this study (the “observations”, i.e. the streamflow reference),

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

> *x represents the observations, i.e., the streamflow reference. F(y) is the CDF obtained from the probabilistic members for the corrected streamflow.*

- L.308-309: a definition of the reliability diagram is missing (e.g. “plots the cdf of the streamflow reference as a function of …”),

Response: Thank you for your comment and suggestion. We have added and revised the description of reliability diagram.

> *The reliability diagram serves as a diagnostic graph to assess the agreement between predicted probabilities and observed frequencies (Jolliffe and Stephenson, 2012). It plots the observed frequencies of events against the predicted probabilities, specifically plotting the cumulative distribution function (CDF) of the streamflow reference as a function of the forecasted probability.*

- L.319: a definition of sharpness is provided afterwards (at l.323-325). At this point, it is just indicated that sharpness is important to assess without knowing what it refers to,

Response: Thank you for your comment and suggestion. We have added and revised the description of Sharpness.

> *Sharpness refers to the precision or tightness of a probabilistic prediction, capturing how closely the predicted probability distributions align with the observations. Essentially, a sharp forecast indicates that the predicted uncertainties are relatively narrow and closely resemble the observed data points, reflecting a more accurate representation of the true uncertainty in the predictions. 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).*

- L.389: Fig.5: missing space,

Response: Thank you for your considerate reminder. We have fixed this issue.

- Table 2: The square is missing for the units of the FAA (km2),

Response: Thank you for your considerate reminder. We have fixed this issue accordingly.

- Table 3: I guess there is a problem with the metric VAR for low flows because it does not correspond to the square of STD,

Response: Thank you for pointing out this error. We appreciate your attention to detail. Indeed, the numbers we calculated for VAR with respect to low flows are quite small. To emphasize the results, we employed scientific notation (e.g., 10^-4), but inadvertently omitted this notation in the table. We have added the appropriate notation to the table.

Table 3. Sharpness statistics in high-flow and low-flow seasons. The bold numbers indicate better performance in each group.

Flow seasons	Metric	PDIR		IMERG-F		GSMaP		All	
		QRF	CMAL-LSTM	QRF	CMAL-LSTM	QRF	CMAL-LSTM	QRF	CMAL-LSTM
High-flow	MAD	**0.046**	0.048	**0.047**	0.052	**0.050**	0.054	**0.045**	0.047
	STD	**0.109**	0.112	**0.133**	0.139	**0.129**	0.133	**0.129**	0.134

(May–Oct.)	VAR	**0.013**	0.014	**0.020**	0.021	**0.018**	0.019	**0.018**	0.020
	DIS_{25-75}	0.0714	**0.0703**	**0.0753**	0.0757	**0.0781**	0.0785	0.0710	**0.0687**
	DIS_{5-95}	**0.184**	0.194	**0.192**	0.215	**0.206**	0.223	**0.184**	0.195
	CO_{25-75} (%)	**51.5**	50.1	**76.9**	76.0	**64.2**	62.8	**73.3**	71.4
	CO_{5-95} (%)	100	100	100	100	100	100	100	100
Low-flow (Nov.–Apr.)	MAD	**0.0085**	0.0100	**0.0073**	0.0094	**0.0088**	0.0104	**0.0064**	0.0069
	STD	**0.0264**	0.0284	**0.0280**	0.0301	**0.0305**	0.0323	**0.0258**	0.0262
	VAR (10^{-4})	**8.32**	9.48	**9.10**	10.47	**10.40**	11.52	**7.71**	7.86
	DIS_{25-75}	**0.0121**	0.0124	**0.0099**	0.0112	**0.0121**	0.0122	0.0086	**0.0086**
	DIS_{5-95}	**0.033**	0.039	**0.029**	0.037	**0.036**	0.042	**0.026**	0.027
	CO_{25-75} (%)	72.2	**75.1**	88.8	**90.2**	69.1	**73.9**	**79.6**	79.2
	CO_{5-95} (%)	100	100	100	100	100	100	100	100

- L.478: Fig.9: missing space,

Response: Thank you for your considerate reminder. We have fixed this issue.

- L.480: the results of Fig. 10 are not described,

Response: Thank you for your comment. We have revised the description of the results of Fig. 10.

> *In Fig. 10, we observe a consistent trend: as the FAA of the sub-basin increases, the performance of the model also increases. Notably, the CMAL-LSTM model consistently surpasses the QRF model across all experiments, which is further supported by the statistics in Table S2. However, as the FAA of sub-basin increase, the performance gap between the CMAL-LSTM model and QRF model begins to diminish, especially in the IMERG-F driven experiment. In contrast, for experiments such as PDIR, GSMaP and multi-product (All), and the increase in FAA has little effect on the performance difference between the CMAL-LSTM and QRF models.*

- Caption of Figure 10: b-h -> e-h,

Response: We appreciate your attention to detail. We have fixed this issue.

- L.497-498: I agree with this statement, but it could be detailed and referenced,

Response: Thank you for your comment. We have expanded our discussion on the factors influencing flood peak and have incorporated pertinent references.

> *Flood peaks have always posed a challenging problem in hydrological simulation due to many factors, such as spatial and temporal variability in rainfall extreme, soil moisture conditions, and catchment characteristics (Brunner et al., 2019; Jiang et al., 2022). Furthermore, slight deviations can lead to significant discrepancies in flood risk assessments (Parodi et al., 2020). Given these challenges, it highlights the necessity of probabilistic post-processing.*

Brunner, M. I., Hingray, B., Zappa, M., & Favre, A. C.: Future trends in the interdependence between flood peaks and volumes: Hydro - climatological drivers and uncertainty. Water Resources Research, 55(6), 4745-4759, https://doi.org/10.1029/2019WR024701, 2019.

Jiang, S., Zheng, Y., Wang, C., and Babovic, V.: Uncovering Flooding Mechanisms Across the Contiguous United States Through Interpretive Deep Learning on Representative Catchments, Water Resources Research, 58(1), e2021WR030185, https://doi.org/10.1029/2021WR030185, 2022.

Parodi, M. U., Giardino, A., Van Dongeren, A., Pearson, S. G., Bricker, J. D., and Reniers, A. J.: Uncertainties in coastal flood risk assessments in small island developing states. Natural Hazards and Earth System Sciences, 20(9), 2397-2414, https://doi.org/10.5194/nhess-20-2397-2020, 2020.

- L.568: Regrading -> regarding,

Response: We appreciate your attention to detail. We have fixed this issue.

- L.52-54: the limitations of satellite data should be discussed (coarse resolution which limits its use for small basins i.e. with an area smaller than 200 km2),

Response: Thank you for drawing our attention to this, enabling a more comprehensive and balanced presentation. Consequently, we have updated our description as follows:

However, uncertainties persist in these products due to various factors, including data sources and algorithms. Additionally, the coarse resolution still limits their use for small basins, i.e., with an area smaller than 200 km2 (Tian et al., 2009; Zhang et al., 2021a).

- There are some references that are cited and not present in the list of references (e.g. Bellier et al., 2018).

Response: Thank you for your meticulous feedback regarding the inconsistencies in our manuscript. We have thoroughly reviewed the entire manuscript and implemented specific revisions.