Dear Editor Prof. Nunzio Romano and anonymous 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.

Section 1: Response to Editor

Dear Authors,

The revised version of your submission was evaluated by two reviewers who provided contrasting comments. Ref.#1 judged it publishable as it is. Ref.#2, instead, suggested that this revised version should be rejected altogether. The latter reviewer supported this recommendation with a number of concerns and negative comments.

Apart from the fact that Ref.#1 was also partially satisfied with the way this paper discusses the scientific relevance of your study, the concerns and arguments raised by Ref.#2 should now be adequately rebutted.

Please, send a new point-by-point response to the new Ref.#2's appraisal and, if modified, a revised version. Should you disagree with any of Ref.#2 comments and arguments, explain why clearly.

Response: Thank you for sharing the evaluation of our revised manuscript. We appreciate the feedback provided by both Reviewer #1 and Reviewer #2. We have carefully reviewed the comments and have prepared a point-by-point response addressing each concern raised by Reviewer #2. We have also made appropriate revisions to the manuscript based on their feedback.

Section 2: Response to Reviewer #2

I reviewed a revised version of the manuscript entitled “Comparing machine learning and deep learning models for probabilistic post-processing of satellite precipitation-driven streamflow simulation” by Zhang et al. I am still not convinced by this version of the manuscript. I have three major concerns.

Response: Thank you for taking the time to review the revised version of our manuscript. We appreciate your valuable feedback and would like to address your concerns.

My first concern is the necessity of conducting this research. The manuscript mentioned the research gap in line 109-110, quoted here: “to our knowledge there has not been a comparison between ML and DL models for hydrological probabilistic post-processing in the literature.”. Later in line 113-114, it mentioned: “their differences in the field of hydrological probabilistic post-processing, such as the scope of application, model performance and computational efficiency is not well studied.”. Although an explicit objective was not provided, it seems that the authors want to compare the bias-correction performance between Machine Learning (ML) and Deep Learning (DL) algorithms. I doubt this objective can be finished and any general conclusion related to the “general features” of ML and DL can be drawn.

Response: We appreciate your feedback and the opportunity to clarify the necessity and objective of our research.

We have narrowed down our focus to specifically compare QRF models and CMAL-LSTM models. The research gap identified revised manuscript highlights the lack of a comprehensive comparison between RF and LSTM and their probabilistic variants for hydrological probabilistic post-processing in the existing literature. While previous studies have individually examined the performance and characteristics of RF and LSTM models, a direct comparison of their scope of application, model performance, and computational efficiency in the specific context of hydrological probabilistic post-processing is still lacking. So, it is essential to conduct a such study. For this reason, we have added a few sentences to clarify this research gap and the necessity of our study.

> *DL models, like LSTMs, through their hierarchical feature learning, possess the capability to autonomously glean insights from raw hydrological data without the need for extensive feature engineering. Moreover, LSTM models, specifically designed for sequential data, are well-suited for time-series hydrological data, capturing long-term dependencies and patterns. 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 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.*
>
> *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. Given the challenges posed by complex interactions between data, scenarios, and model representations, exploring and selecting the application of both ML and DL models becomes essential.*

As per your suggestion, we have also narrowed down our focus to specifically compare QRF models and CMAL-LSTM models in our revised manuscript. We have also revised the title to accurately reflect the scope of our research.

> *The revised title: "Comparing quantile regression forest and mixture density long short-term memory models for probabilistic post-processing of satellite precipitation-driven streamflow simulations"*

We hope this clarification addresses your concerns regarding the necessity and objectives of our research.

We acknowledge that our analysis cannot cover all ML and DL models. Our objective in this study is to provide insights into the differences and similarities between QRF model and CMAL-LSTM model in the field of hydrological probabilistic post-processing. We also acknowledge that drawing general conclusions about the "general features" of ML and DL may be ambitious. However,

our goal is to contribute to the existing knowledge by providing a comparative analysis of the bias-correction performance and exploring the strengths and limitations of QRF and CMAL-LSTM models in the context of hydrological probabilistic post-processing. Basically, the findings of this study will help guide researchers and practitioners in selecting appropriate models based on their specific requirements and objectives. We will continue to supplement our contributions in the response below.

First, DL is a subset of ML, so a DL algorithm is also an ML algorithm. Set this aside, say the authors want to compare DL and non-DL algorithms in bias-correction, there are a lot of algorithms belonging to the two types. Simply picking a couple of examples (one for each in this case) can't represent the whole DL/non-DL family. Besides, these are data-driven algorithms, its performance in bias-correction of streamflow depends greatly on how one sets the models and how the datasets are like. So, I really don't know how we can get to the conclusion that ML is absolutely better/worse than DL. The current study only compared QRF (representing ML) and CMAL-LSTM (representing DL). I don't see a clear way (nor did the manuscript present such a way) of generalizing the results for other ML and DL algorithms. If we view this study as a case study, comparing QRF and CMAL-LSTM only, we still face the necessity issue. Why do we want to know these two particular algorithms' performance on bias-correction of streamflow? Can we learn from this study if we would need to use algorithms other than QRF and CMAL-LSTM to construct models? Even if we use the same two algorithms, will all the results of this study be consistent if we move to another catchment? I double so because data-driven models depend on the data.
Response: We appreciate your insightful comments and would like to address your concerns regarding the necessity and objective of our research again.

Firstly, we acknowledge that DL is a subset of ML, and DL algorithms fall within the broader category of ML algorithms. To address this concern, we have stated this in the text.

> *"In this paper, 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."*

Secondly, in our study, we aim to compare the performance of QRF (representing ML) and CMAL-LSTM (representing DL) specifically in the context of hydrological post-processing. We agree that there are numerous algorithms within both DL and non-DL categories, and our study focuses on a limited selection as representative examples. We chose to specifically focus on QRF models and CMAL-LSTM models because they are considered to be state-of-the-art machine learning and deep learning models, respectively, in the field of hydrological applications. These models have been widely recognized and extensively used in previous studies, making them suitable candidates for our comparative analysis. To address this concern, we have reorganized our literature in the text.

> *"Two state-of-the-art approaches in ML and DL are the random forest (RF) and Long Short-Term Memory (LSTM) models. 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, 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). 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 distribution 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."

Thirdly, we understand the concern about generalizing the results to the entire DL/non-DL algorithm family. While our study may not encompass all possible algorithms, it provides valuable insights into the performance and characteristics of QRF and CMAL-LSTM algorithms in hydrological post-processing. The basic intention is to contribute to the existing knowledge by providing a comparative analysis of these two algorithms, highlighting their strengths and limitations in the specific context of streamflow bias-correction. While our focus in this study is on the QRF and CMAL-LSTM models, we believe that the insights gained from the comparison study in hydrological probabilistic post-processing will have broader implications beyond just understanding the performance of these two specific methods. The findings will serve as a foundation for making more informed decisions when selecting other ML and DL methods that share similar model families or architectural ideas. For example, the random forest model represents a decision tree-based approach and is based on a historical search algorithm. The QRF model can be seen as an implicit probability distribution modeling approach, where the probabilistic members are determined by dividing the data into quartiles. On the other hand, the LSTM model represents a sequence regression-based modeling method that utilizes neural networks to fit sequential data. Similarly, CMAL-LSTM can be viewed as an explicit probability distribution modeling method, which obtains the final probability distribution function by combining multiple probability distributions. It obtains the final probabilistic members by sampling and then taking the quantiles. Therefore, we believe that while our study focuses on specific models, the underlying ideas and principles are applicable to a broader range of models. By examining and comparing the performance of these models, we aim to provide insights and understanding that can be useful in various other studies and applications.

This study represents a significant opportunity to deepen our comprehension of the strengths and limitations of both QRF and CMAL-LSTM models in hydrological probabilistic

post-processing. By undertaking this exploration, researchers can identify the most appropriate approaches for various aspects of the task, ultimately elevating the reliability and efficiency of hydrological forecasts and predictions. The insights gained from the study will also pave the way for more informed decision-making when selecting other ML and DL methods with similar model families or architectural ideas, tailored precisely to the unique characteristics of hydrological data and specific prediction requirements. As a result, this study has the potential to contribute significantly to the advancement of hydrology as well as its crucial applications in water resource management and flood forecasting.

Fourthly, we appreciate your concerns about the limitations of data-driven models and the dependence on specific catchments and datasets. The sample size of 522 sub-basins selected for our study is considered relatively adequate for obtaining valuable insights and drawing meaningful conclusions. It is important to note that this sample size is comparable to that of the CAMELS (Catchment Attributes and Meteorology for Large-sample Studies) dataset, which has been extensively used in numerous hydrological studies. While we are working with a different dataset, it is still homogeneous in nature, allowing for valid comparisons and inferences. Therefore, our study provides a comparable basis for analysis and contributes to the existing hydrological research.

Another concern of mine is related to one of the conclusions, the dependency of error metrics with drainage area. The authors used drainage area as a key factor to demonstrate the error metrics. They concluded in several locations that the model performance is related to the drainage area. For example, in line 608 they wrote: “their model differences are correlated with the flow accumulation area (FAA) of sub-basins.”. Yet, based on my visual inspections on the related figures (e.g., Figure 6, Figure 10), the drainage area dependency is unclear. I would suggest the authors remove this track of analysis and probably find other factors for a similar investigation.

Response: We appreciate your suggestion to explore alternative factors for investigating the model performance dependency. While we understand the importance of considering multiple factors, we believe that the analysis of drainage area can still provide valuable insights into the relationship between this specific factor and model performance.

Upon careful analysis, we believe that catchment area remains an important factor in our study. Specifically, when we examined the results presented in Figure 6 and compiled them into Table 2, a clear pattern emerged. The QRF model consistently outperformed the CMAL-LSTM model in most sub-basins with smaller catchment areas, as indicated by higher CRPSS values. However, as the catchment area increased, particularly beyond 60,000 km^2 in this study, the CMAL-LSTM model demonstrated a distinct advantage over the QRF model. These findings highlight the influence of catchment area on the performance of the two models and emphasize the need to consider this factor in hydrological probabilistic post-processing.

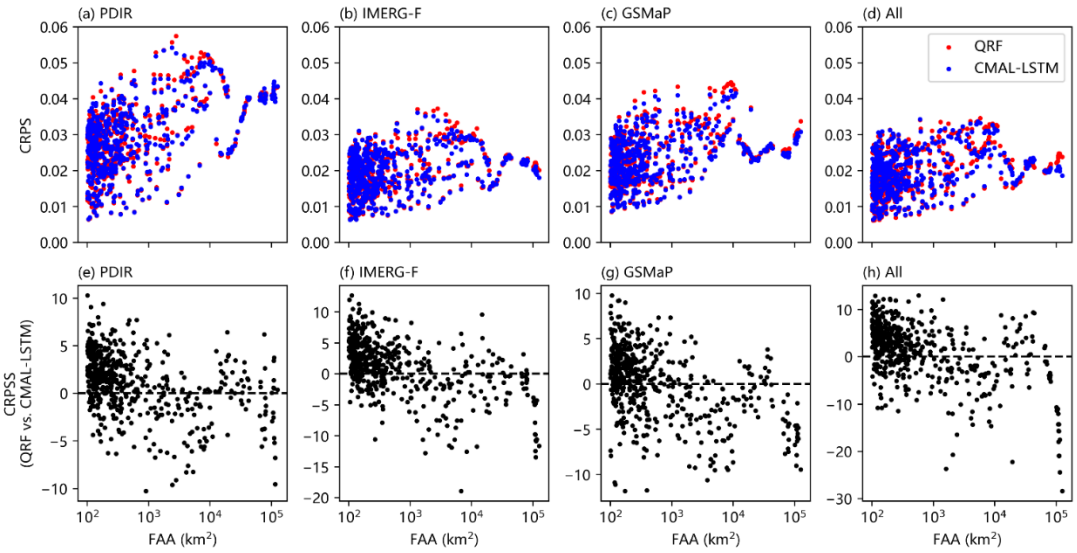

Figure 1. The relationships between (a–d) CRPS, (e–h) CRPSS and FAA.

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)	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**

In fact, we conducted this analysis also based on some priori knowledge. For instance, Nijssen and Lettenmaier (2004) conducted a study using Monte Carlo and Variable Infiltration Capacity (VIC) models, where they investigated simulated precipitation and its performance. Their results indicated a significant decrease in runoff simulation errors as catchment size increased, particularly for catchments larger than 50,000 km^2. Additionally, in another study, researchers found that errors were more pronounced in smaller catchments (less than 400 km^2) compared to larger catchments (Nikolopoulos et al., 2010). Furthermore, we would like to mention a study by Cunha et al. (2012), who focused on watersheds in Central Iowa with catchment areas ranging from 20 to 1600 km^2. Their research highlighted the strong dependence of peak flow simulation uncertainty on the size of the catchment. These previous findings provide supporting evidence for our analysis, and our study contributes to the existing body of knowledge by further investigating the relationship between catchment area and the performance of the QRF and CMAL-LSTM models.

Reference:

Cunha, L.K.; Mandapaka, P.V.; Krajewski, W.F.; Mantilla, R.; Bradley, A.A. Impact of radar-rainfall error structure on estimated flood magnitude across scales: An investigation based on a

parsimonious distributed hydrological model. Water Resour. Res. 2012, 48, W10515.
Nijssen, B.; Lettenmaier, D. Effect of precipitation sampling error on simulated hydrological fluxes and states: Anticipating the Global Precipitation Measurement satellites. J. Geophys. Res. Atmos. 2004, 109, D02103.
Nikolopoulos, E.I.; Anagnostou, E.N.; Hossain, F.; Gebremichael, M.; Borga, M. Understanding the Scale Relationships of Uncertainty Propagation of Satellite Rainfall through a Distributed Hydrologic Model. J. Hydrometeorol. 2010, 11, 520–532.

Last but not the least, a persisting one from the previous round. Presentation of the manuscript is bad. I pointed out last time that the methodology was not well described, especially those different evaluation metrics. I went through the new version and found a similar level of writing, confusing and hard to read. I am still not able to understand the concept of sharpness (Table 3) and the meaning of the reliability diagrams (Figure 7). I didn't point out all the writing issues as reading this manuscript made me tired (see some specific comments in the annotated manuscript). I wish the authors can put more effort into improving the readability of the manuscript.

Response: We appreciate your feedback and acknowledge your concerns regarding the presentation and readability of the manuscript. We apologize for any confusion caused by the methodology description and the interpretation of evaluation metrics. We understand the importance of clear and concise writing in effectively communicating our research findings. We have carefully reviewed the manuscript and made significant revisions to improve its readability and clarity. Specifically, we have carefully reviewed the annotated manuscript and have incorporated your specific comments and suggestions to ensure a more coherent and cohesive presentation.

In the previous round of revision, we made significant efforts to address this issue and improve the overall structure and clarity of the paper. Firstly, we reorganized the methods section to provide a clear and concise description of the research methodology. By merging smaller sections and streamlining the content, we aimed to enhance the overall readability and make it easier for readers to understand the approach. Furthermore, we restructured the results section to focus on the core analytical content, highlighting the comparison between the two methods of interest. We have also included some case studies in the supplementary material, ensuring that the main text remains focused on the key findings and their interpretation. In the discussion section, we have rewritten the section to provide more in-depth analysis and interpretation of the results. We have added two subheadings to address the characteristics of the two methods, highlighting their respective strengths and weaknesses. Additionally, we have included a section on future directions for improvement, allowing for a more comprehensive discussion of the implications and potential advancements in the field. Overall, we have made significant revisions to the organization of sections, paragraphs, and sentences to ensure a logical flow of information and improve the readability of the text.

In this round of revision, we took your suggestions into account and made specific changes to further enhance the explanation of different evaluation metrics. We provided more detailed descriptions and examples to help readers understand the concepts of sharpness (Table 3) and the meaning of reliability diagrams (Figure 7).

Sharpness is a fundamental characteristic of predictive distributions, crucial for assessing assess the precision or tightness of a probabilistic predictions. A sharp probabilistic output corresponds to a low degree of variability in the predictive distribution. In the context of probabilistic hydrological post-processing, sharpness measures how closely the predicted probability distributions align with the observations. A sharp forecast means that the predicted uncertainties are relatively narrow and closer to the observed data points, indicating a more accurate representation of the true uncertainty in the predictions. 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).

The reliability diagram serves as a diagnostic graph to assess the agreement between predicted probabilities and observed frequencies (Jolliffe and Stephenson, 2003). The diagram helps to evaluate the reliability of probabilistic forecasts by comparing the predicted probabilities of events with their corresponding observed relative frequencies. Ideally, in a perfectly reliable forecast, if the predicted probability of a specific event is, for example, 30%, then the observed relative frequency of that event should also be around 30%. Consequently, the reliability diagram would show a distribution of points lying along the diagonal line, indicating a consistent alignment between predicted probabilities and observed frequencies across various probability levels. However, in practice, there may be deviations from perfect reliability. Points on the reliability diagram above the diagonal line suggest that the observed relative frequency is higher than the predicted probability, indicating an underprediction phenomenon. On the other hand, points below the diagonal line indicate that the observed relative frequency is lower than the predicted probability, indicating an overprediction phenomenon.

In this round of revision, we also made our best effort to address the specific writing issues pointed out by you, aiming to make the manuscript more engaging and easier to comprehend. While we understand that there might still be areas for improvement, we made every effort to enhance the readability of the manuscript based on your feedback.