We thank the review for the positive and constructive feedback. Please find below the answers to your questions. All answers will be implemented in the revised version of the manuscript. Our reply to each of your questions/suggestions can be found in blue below.

Remark: All references to lines in the revised manuscript refer to the version with track-changes,

Answers to Anonymous Referee #1' s comments:

Comment 1: *Are precipitation, temperature, and humidity enough as input variables for your neural networks?*

Authors' answer: In this study, precipitation, temperature, and humidity are the only variables available to develop the models, hence they were selected as primary input variables based on their direct impact on flood generation processes. The sensitivity analysis that we have performed and presented in Section 3.4 of the discussion manuscript, showed that N-HiTS and N-BEATS models maintained high performance even when individual meteorological variables were excluded, indicating the models' robustness to input variations. This robustness likely stems from the models' ability to learn essential patterns from historical discharge data alone, particularly the rainfall-runoff relationships in headwater streams. We have added additional performance metrics to Section 3.4 of the updated manuscript, and clarified that both models exhibited strong performance, even without the inclusion of precipitation, temperature, or humidity data (lines 771-772, of the manuscript with track-changes version).

Comment 2: *The forcing station is a single point in the watershed while the runoff generation should be attributed to the water convergence involving a large area of the watershed, do you think a single station can represent these complex processes at large areas?*

Authors' answer: Thank you for this important remark. Indeed, a single station may not fully represent the spatial heterogeneity of larger watersheds. We acknowledge that using a single station can provide localized information and data for small watershed, such as headwater streams where runoff generation is more responsive to immediate meteorological conditions. However, for the particular case we have applied the methods to, capture significant runoff and flood dynamics in these streams, with a single station. The new version of the manuscript incorporated these important aspects, especially that for broader areas, more spatially distributed data could improve model accuracy. Moreover, additional discussion has been incorporated into the conclusion section, specifically in lines 973–978 of the reviewed version of the manuscript, the track-changes version of it.

Comment 3: *You mentioned, your models predicted one hour ahead? Is this meaningful for flood prediction? In other words, is this enough time to escape once people know the flood will arrive one hour later.*

Authors' answer: You raised a valid point. Although a one-hour lead time may seem brief, it is meaningful in the context of accurate flash flood prediction particularly in an area with a proximity to large metropolitan cities, where rapid response is critical. Although, we tested the model to forecast up to 72 hours in advance based on National Weather Service near real time river forecast data. We found that one-hour lead time provides the best near real time performance. We included this in the conclusion part of the revised manuscript, as this is an important remark for the readers. Additional discussion has been incorporated into the conclusion section, specifically in lines 905–907, of the reviewed version of the manuscript.

Comment 4: *Did you train each NN model for each watershed? Trained based on one watershed and then transferred to the other one? Or trained both watersheds together?*

Authors' answer: Thank you for your comment. Each neural network model was trained separately for each watershed to account for local hydrological characteristics. The independent training approach enabled the models to capture watershed-specific runoff generation and flood dynamics, optimizing prediction accuracy within each unique environment. Transfer learning was not implemented. We included this remark in the new version of the manuscript in the result section (lines 590–594, of the revised manuscript, please see track-changes version).

We thank the review for the constructive feedback. Please find below the answers to your questions. All answers are implemented in the revised version of the manuscript. Our reply to each of your questions/suggestions can be found in blue below.

Remark: All references to lines in the revised manuscript refer to the version with track-changes,

Answers to Anonymous Referee #2' s comments:

Comment 1: *Interpretability and Model Complexity: The paper claims that N-HiTS and N-BEATS models offer interpretability. However, further elaboration on how these models achieve interpretability would strengthen the paper. Including visual examples or providing a more explicit breakdown of how interpretability manifests in model outputs could clarify this for readers who may be less familiar with these architectures.*

Authors' answer: Thank you for pointing out that interpretability and model complexity are not sufficiently explained in the manuscript. In the new version of the manuscript, we enhanced this portion. The discussion has been added throughout the paper, specifically in lines 920–931. Additionally, the architecture figures for the N-HiTS and N-BEATS models have been modified/updated to better represent the interpretability of these models in lines 320 and 364.

Comment 2: *Hyperparameter Selection: The selection process for critical hyperparameters like the lookback window size is not fully justified. Lookback windows are crucial in sequence-based forecasting, and this choice should either be explored as a hyperparameter or explained in greater detail, particularly given the model's dependency on residuals for subsequent window predictions. Additionally, since a 24-hour lookback window is used, further elaboration on how this length captures relevant hydrological features, like seasonality or trends, would enhance clarity.*

Authors' answer: We agree with your assessment, and we clarified this in the revision. The selection of a 24-hour lookback window was guided by the average Time of Concentration (TC) values for the particular watersheds under study, where the average TC were close to 19 hours in the Lower Dog River watershed, and 22 hours in the Upper Dutchmans Creek watershed. By setting the lookback window to 24 hours, the model could capture essential meteorological data preceding flood events, reflecting both short-term variances and any potential longer trends relevant to hydrological processes. We also evaluated different lookback window sizes (input sizes) from 1 hour to 24 hours to analyze the impact of this hyperparameter on the results. All these mentioned important aspects are included in the results section of the paper, specifically in lines 582–584.

Comment 3: *Metrics Selection: While NSE, RMSE, and MAE are utilized, the omission of the Kling-Gupta Efficiency (KGE) index is notable. KGE is especially relevant for flood forecasting* *as it provides insights into peak flow timing, magnitude, and correlation. Including KGE would add robustness to the evaluation by capturing aspects critical to hydrological modeling.*

Authors' answer: Thank you for this suggestion, which we find it very valuable. The inclusion of Kling-Gupta Efficiency (KGE) can certainly strengthen the model evaluations. The KGE metric has been added as one of the performance metrics in the methodology section in lines 390-396. All tables containing performance metrics in the results section have been updated accordingly.

Comment 4: *Interpretability in Model Outputs: Although the paper claims interpretability for both N-HiTS and N-BEATS, the explanation is somewhat abstract. Providing visual aids or case studies that illustrate interpretability in flood prediction contexts would be beneficial. Specifically, the paper mentions that projections onto harmonic and trend bases improve prediction accuracy, but further clarification on the physical interpretability of these projections would help. Given the use of a 24-hour window, it would be helpful to explain whether trends, network depth, or some other feature captures seasonality and why this choice is appropriate for flood prediction.*

Authors' answer: You raised a valid point. Both N-HiTS and N-BEATS capture trends and seasonality through basis functions in the interpretable configuration. For flood prediction, these components allow models to project periodic and steady trends, enhancing physical interpretability. Additional discussion has been incorporated into the conclusion section, specifically in lines 920–931.

Comment 5: *Uncertainty Analysis: The application of Maximum Likelihood Estimation (MLE) for uncertainty quantification is intriguing. However, more details on how MLE is applied in this context would improve reproducibility. A clearer formulation of MLE within the training process or its integration with multi-quantile loss could better inform readers about the strengths and limitations of this approach. Additionally, bootstrapping methods could help quantify uncertainty and assess whether observed performance differences between models are statistically significant, providing a more robust comparison.*

Authors' answer: The MLE was implemented by optimizing the likelihood function to capture prediction distribution characteristics. We acknowledge that a clearer presentation of how MLE integrates with the multi-quantile loss in training can enhance reproducibility, while bootstrapping could provide additional quantification of model prediction variability. Given the paper scope and length incorporating multiple uncertainty quantification methods would lose the focus of the presented content. The authors are already conducting a separate study on different uncertainty quantification methods for these models, with the aim to publish the findings in a subsequent paper. Additional clarification about MQL has been added into the methodology section, specifically in lines 440-465. We appreciate the comment.

Comment 6: *Separate Model Training for Each Catchment: Each model was trained separately for each catchment, rather than training a single model on both catchments. This approach limits the assessment of the models' generalizability across different hydrological conditions. Training a unified model on data from both catchments would provide insights into the model's adaptability and robustness across diverse environments, which is crucial for broader flood prediction applications. I recommend including an analysis of a single model trained across both catchments to evaluate cross-catchment performance.*

Authors' answer: Thank you for the comment. We included in the revised manuscript the below explanation:

"The decision to train separate models for each catchment was made to account for the unique hydrological characteristics and local features specific to each watershed. By training models individually, we aimed to optimize performance by tailoring each model to the distinct rainfall-runoff relationship inherent in each catchment." Additional clarification has been added into the result section, specifically in lines 590–594.

Comment 7: *Data Splits for Training, Validation, and Testing: It appears the observational data up to October 1, 2022, was used for training, and data from October 1, 2022, to March 28, 2023, was used for validation. However, the absence of an unseen test set to demonstrate generalization capabilities raises concerns. Dividing the dataset into three splits (training, validation, and testing) would allow for hyperparameter optimization on the validation set and final results on an unseen test set, demonstrating the model's generalization. Including metrics like loss curves for the training and validation sets or evaluation metrics on a test set would help assess model performance and detect overfitting thereby enhancing reliability.*

Authors' answer: Thank you for this direction. In our study, the models were trained and validated on data up to October 1, 2022. We then used data from October 1, 2022, onward as an unseen test set to evaluate the models' forecasting capabilities. This approach allowed us to optimize hyperparameters during training and validation while ensuring the final performance metrics reflected the models' ability to predict on unseen data. The clarification has been added into the result section, specifically in lines 588–590.

Comment 8: *Model Reproducibility: Simplifying the explanation of the Multi-Quantile Loss (MQL) function could make the methodology more accessible. Additionally, code availability or pseudocode in an appendix would enhance reproducibility and facilitate further exploration by other researchers.*

Authors' answer: We agree that simplifying the explanation of the Multi-Quantile Loss (MQL) function would improve accessibility for readers. The clarification about MQL has been added into the methodology section in lines 440-465. For the open research section, we mentioned that the source code for this study will be available after publication of the paper results in a repository, specifically Zenodo.

Comment 9: *Input Sensitivity Inconsistency (Line 568-569): The statement here suggests that the models are indeed sensitive to input conditions, especially during extreme events. However, in the following section, the paper concludes that the models are not sensitive to input data, which presents an inconsistency. This contradiction should be addressed.*

Authors' answer: Thanks for your comments. In these sections of the manuscript, the model results indicate challenges in capturing peak rates during flashy floods, which represent anomalies in discharge and deviate from typical rainfall-response patterns in the time series data. In addition, Lines 568-569 discussed the deficiency of both N-BEATS and N-HiTS models in capturing the dynamics of the recession curve which is directly related to groundwater contribution to flood hydrograph. However, both models are technically insensitive to rainfall data as an input variable, suggesting they can learn from discharge patterns (which inherently include precipitation effects) without requiring meteorological data. Since the models are trained on regular discharge patterns, they encounter difficulties to capture the peak rates and the recession curve due to short duration, intense runoff as well as a shallow aquifer/groundwater contribution. This discussion and clarification have been incorporated into the conclusion section, specifically in lines 936–941.