We thank the review for the 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 below.

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 reviewer for pointing out that interpretability and model complexity are not sufficiently explained in the manuscript. In the new version of the manuscript, we will add about it, as follows "The interpretability of N-HiTS and N-BEATS is achieved through their designs. N-HiTS approach aims to enhance the accuracy of long-term time-series forecasts. This is achieved through the implementation of techniques that combine hierarchical interpolation with multi-scale data sampling methods., allowing it to focus on different data aspects, which prioritizes features essential for flood trends, i.e., larger scale. N-BEATS leverages interpretable configurations with trend and seasonality projections, enabling it to decompose time series data into intuitive components. N-BEATS interpretable architecture is recommended for scarce data settings (such as flooding event), as it regularizes its predictions through projections unto harmonic and trend basis. These approaches improve model transparency by allowing understanding of how each part of the model contributes to the final prediction, particularly when applied to complex flood patterns." In addition to that, we will include additional visualization within the models' architecture diagrams to illustrate the interpretability of their outputs.

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 will clarify 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 will be included in the results section of the paper.

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) will strengthen model evaluations. After analyzing the general prediction performance with NSE, RMSE and MAE, we concluded that KGE will provide insights into peak flow timing and the alignment between predicted and observed values—a key aspect of flood forecasting. This addition will be included in the result section of the revised manuscript.

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. We will add more discussion about this when we revise the manuscript.

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 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. 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. 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 will include in the revised manuscript the 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."

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' generalization 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. We will clarify the language in the paper when we revise the paper.

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. Additionally, I will include in the "Open Research" section of the manuscript that to support reproducibility and facilitate further research,

the source code for this study will be uploaded on the Zenodo, providing open access to all codes and model configurations used in our experiments.

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 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 flood generating rainfall as well as shallow aquifer/groundwater contribution. This point will be added to the conclusion section of the revised manuscript.