Exploring the Applicability and Reliability of Machine Learning tools in Streamflow Forecasting

by

Kavindra Lakmal Lewkebandara

A thesis submitted to the Faculty of Graduate Studies of
The University of Manitoba
in partial fulfilment of the requirements of the degree of

MASTER OF SCIENCE

Department of Civil Engineering
University of Manitoba
Winnipeg, Canada

Abstract

Recent advancements in machine learning, particularly Long Short-Term Memory (LSTM) networks, have demonstrated remarkable success in hydrological modelling, often outperforming traditional hydrological models. This thesis provides a comprehensive analysis of LSTM networks for streamflow forecasting under various conditions. First, the impact of incorporating historical streamflow data as an input was evaluated, demonstrating significant improvements in prediction accuracy across diverse catchments. While LSTM outperformed the persistence for one-day-ahead forecasts, accuracy decreased for longer lead times for both models. The effect of noisy precipitation inputs was subsequently investigated, revealing that while noise generally reduces performance, LSTMs trained with noisy data exhibit resilience. Basin sensitivity to precipitation noise varied and correlated with catchment attributes. Lastly, interpolation and extrapolation under stationary and non-stationary climate scenarios were examined. LSTM performed remarkably well under stationary conditions but showed biases when predicting under changing precipitation regimes, highlighting challenges in extrapolation. In conclusion, the thesis summarizes the key findings, addresses its limitations, and suggests avenues for future research, such as incorporating forecasted forcing data and hybrid models development to improve the robustness of LSTM-based streamflow forecasting.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Professor Ricardo Mantilla, for his unwavering guidance, encouragement, and support throughout this research journey. His expertise, patience, and insightful feedback have been invaluable in shaping my work and helping me overcome challenges along the way. I am also sincerely grateful to Professor Chandra Rajulapati and Professor Chris Henry for serving on my thesis examination committee. Their thoughtful comments and suggestions were instrumental in refining my work and elevating it to a higher standard. I would also like to extend my heartfelt thanks to Professor Shawn Clark and Professor Donghoon Lee for sharing their knowledge with me. Their teachings have not only enriched my understanding but have also motivated me to strive for excellence in my academic pursuits.

Additionally, I would like to extend my sincere appreciation to my colleagues for taking the time to show interest in my work and for sharing their constructive ideas and suggestions that contributed to the success of my research. I truly appreciate their willingness to help and inspire me. Lastly, but certainly not least, I want to thank my family and friends for their endless encouragement, understanding, and patience throughout this journey. Their love and support have been my greatest strength, and I am deeply thankful for their belief in me, especially during the most challenging times.

Dedications

To my beloved family

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Streamflow forecasting is a critical component of water resource management, essential for applications such as flood prediction, drought management, and water supply planning (Hunt et al., 2022; Kratzert et al., 2018; Lafon et al., 2013). Traditionally, hydrological models, including physically based, conceptual, and data-driven approaches, have been employed for streamflow simulation and forecasting. However, these traditional models often face challenges in capturing the non-linear and non-stationary characteristics of streamflow, particularly when dealing with large datasets, parameter complexity, computational expenses, and inherent limitations (Kratzert et al., 2018; Arsenault et al., 2023).

In past decade, machine learning (ML) techniques, particularly Long Short-Term Memory (LSTM) algorithm, have emerged as a promising alternative (Arsenault et al., 2023). LSTMs, a type of recurrent neural network (RNN), are designed to model long-term dependencies in sequential data and have shown a strong performance in streamflow simulation and

forecasting against traditional hydrological models and other ML approaches like Artificial Neural Networks (ANNs) (Le et al., 2021; Kratzert et al., 2018; Demiray et al., 2024). Their ability to learn complex hydrological relationships directly from data, without relying on predefined physical equations, makes them well-suited for capturing the complex dynamics of the hydrological cycle (Gauch et al., 2021). The integration of LSTMs with other modeling techniques, forming hybrid models, has also shown potential for achieving more accurate and robust streamflow predictions (J. Liu et al., 2024, Yifru et al., 2024).

1.2 Problem Definition

Despite the advancements in LSTM-based streamflow modeling, several research gaps and challenges remain. One key area involves understanding the impact of the inclusion of historical streamflow data as an input feature on LSTM performance.

Another crucial challenge lies in the forecasting capability of LSTM models over different prediction horizons. While LSTMs can perform well for short-term forecasts, their accuracy typically degrades as the lead time increases (Lin et al., 2024). Understanding how LSTM performance compares to simpler benchmark models, such as persistence methods, for multi-day ahead forecasts is important. Furthermore, traditional performance metrics might not adequately capture model behavior during critical hydrological periods, such as high-flow and flood events, necessitating a more focused evaluation of LSTM performance under such conditions.

The reliability of LSTM models also significantly depends the quality of input data, particularly precipitation, which is the primary forcing of the hydrological cycle (Lafon et al., 2013). Uncertainties and noise inherent in precipitation data from various sources can impact

the accuracy of streamflow predictions (Frame et al., 2021). Understanding the sensitivity of LSTM models to noisy precipitation inputs, identifying regions where this sensitivity is more pronounced, and assessing the model's ability to filter out such errors are critical for real-world applications.

Finally, with the anticipation of changing climate patterns leading to hydrological events that have not been observed before, the ability of data-driven models like LSTMs to generalize beyond the historical data they were trained on becomes a significant concern (Beven, 2024). Distinguishing between interpolation (predictions within the training data range) and extrapolation (predictions outside the training data range) is crucial for assessing the reliability of LSTM forecasts under stationary and non-stationary conditions.

1.3 Research Objectives

This thesis aims to address these challenges and enhance the understanding of LSTM networks for streamflow forecasting through the following key objectives:

- to evaluate the effectiveness of incorporating historical streamflow data as an input feature in LSTM models for improving streamflow prediction accuracy across a large number of diverse catchments,
- to compare the capability of LSTM models for multiple days ahead streamflow forecasting against the persistence,
- to investigate the impact of noisy precipitation input data on the accuracy of LSTM models for streamflow prediction, and to analyze the sensitivity of different basins to varying levels of noise in precipitation,

- to analyze the performance of LSTM models in streamflow forecasting under stationary and non-stationary scenarios, focusing on their ability to interpolate and extrapolate to hydrological conditions that may lie outside the historical training data range.

1.4 Outline of the Thesis

This thesis is structured as a **Sandwich Style Thesis** and it includes the following chapters:

- Chapter 2: **The Effectiveness of LSTMs in Streamflow Forecasting Using Historical Data** investigates the impact of incorporating historical streamflow data as an input for LSTM models and evaluates their performance for multi-day ahead forecasting, comparing them against a persistence model. It also examines the model performance during non-baseflow periods.

- Chapter 3: **The Effect of Noisy Precipitation Inputs on Streamflow Predictability of LSTMs** explores the sensitivity of LSTM models to uncertainties in precipitation input data. It analyzes the impact of varying levels of noise on streamflow prediction accuracy across different basins and examines the relationship between model sensitivity and catchment characteristics.

- Chapter 4: **Interpolation & Extrapolation in Streamflow Forecasting in Stationary & Non-Stationary Scenarios using LSTMs** examines the reliability of LSTM models in predicting streamflow under stationary and non-stationary climate scenarios simulated using a physically based hydrological model. It focuses on the ability of the model to generalize to hydrological conditions that involve interpolation and extrapolation beyond the historical training data.

- Chapter 5: **Conclusion of the Thesis** summarizes the key findings & the conclusion of the thesis, limitations, and suggests potential directions for future research.

Through these chapters, this thesis aims to provide a comprehensive analysis of the capabilities and limitations of LSTM networks for streamflow forecasting under various conditions, contributing to the ongoing advancements in hydrological modeling and prediction.

Chapter 2

The Effectiveness of LSTMs in Streamflow Forecasting Using Historical Data

2.1 Introduction & Background

Streamflow forecasting is one of the most crucial elements of water resource management (Hunt et al., 2022). As traditional hydrological models, physically based, conceptual and data-driven approaches have been used for decades to simulate streamflow. However, capturing non-linear and non-stationary streamflow characteristics using traditional models has been challenging due to difficulties related to large data, parameter complexity, computational expense, and limitations (Kratzert et al., 2018; Arsenault et al., 2023). Recently, machine learning (ML) techniques such as Long Term Sort Term Memory (LSTM) networks have emerged as a novel approach that could address those difficulties (Arsenault et al., 2023; Gauch et al., 2021).

2.1.1 Performance of LSTMs against Traditional Models

Many studies have exhibited the superior performance of LSTM networks in streamflow simulation and forecasting compared to traditional hydrological models (Frame et al., 2021; Kratzert et al., 2021; Lees et al., 2021). For example, Kratzert et al. (2018) showed that LSTM models could slightly outperformed the Sacramento Soil Moisture Accounting Model (SAC-SMA), a well-established hydrological model. Frame et al. (2021) also found that LSTM models significantly improved streamflow predictions compared to the U.S. National Water Model (NWM). In a study across 669 catchments in Great Britain, Lees et al. (2021) showed that their LSTM models outperformed a set of benchmark conceptual models. Gauch et al. (2021) also demonstrated that LSTM models performed better than the NWM at both time scales (Daily and Hourly) they considered. These findings suggest that LSTM models has ability to trace complex hydrological relationships directly from data, without relying on predefined physical equations. Furthermore, unlike traditional Artificial Neural Networks (ANN), LSTM models, along with other Recurrent Neural Network (RNN) types such as Gated Recurrent Units (GRU), demonstrate greater stability and performance in streamflow forecasting (Le et al., 2021).

LSTM's ability to discover long-term dependencies in sequential data is crucial for hydrological modelling. Catchments often exhibit memory effects, where past conditions influence future streamflow. Traditional RNNs have limitations in learning such long-term dependencies, whereas LSTMs are designed to overcome this weakness (Kratzert et al., 2018).

2.1.2 Historical Streamflow Data as Input for LSTM

The use of previous streamflow data as an input variable significantly could enhance the performance of LSTM models, because streamflow series exhibit strong temporal dependencies, with current streamflow values highly correlated with past values. Analysis Lin et al. (2024) revealed that lagged streamflow contributed more significantly to the forecast results than lagged precipitation Pokharel and Roy (2024) also used lagged streamflow values as input in their CNN-LSTM model and showed improved performance for streamflow prediction. Frame et al. (2021) demonstrated that LSTM models effectively learn rainfall-runoff dynamics, gaining minimal additional insight from the conceptualizations encoded in the NWM.

2.1.3 Forecasting Capability of LSTM

While LSTM models are powerful tools for streamflow forecasting, their performance could degrade as the prediction horizon increases. The accuracy of predictions tends to decrease with longer lead times (Ghimire & Krajewski, 2019). For instance, Lin et al. (2024) showed that while their naive LSTM model forecasted the 1-day-ahead streamflow with an higher accuracy, the performance decreased as the lead time increased. This highlights a limitation of purely data-driven approaches that rely on patterns in the training data, but those patterns may not fully capture future dynamics.

2.1.4 Performance of the Persistence

The naive method, the Persistence model, serves as a simple yet reliable benchmark for short term streamflow forecasting. This method assumes that future streamflow would be same as the current streamflow, relying on the inherent characteristics of streamflow series. Lin et

al. (2024) found that for 1 day, 2 days, and 3 days ahead streamflow forecasts, their naive method achieved Nash–Sutcliffe Efficiency (NSE) higher than 0.5 in 88%, 65%, and 52% of basins respectively. This highlights that very simple assumptions of persistence can capture a significant amount of streamflow predictability at short time scales. The naive method provides a baseline for comparing and assessing the performance of more complex models such as LSTM.

2.1.5 Baseflow Separation and its Influence on Model Evaluation

Accurate evaluation of streamflow prediction models might require the separation of baseflow from total streamflow (Szilagyi, 2004). Baseflow, representing the sustained groundwater contribution to streamflow, exhibits distinct characteristics from quick flow or direct runoff, which is primarily driven by rainfall events. Various methods exist for baseflow separation, ranging from simple digital filters (Eckhardt, 2012) to more complex physically based models (Szilagyi, 2004). The choice of baseflow separation method can significantly influence the assessment of model performance, particularly in the context of comparing different forecasting techniques. Selecting an appropriate baseflow separation technique is therefore crucial for ensuring accurate and reliable analyses of streamflow data (Szilagyi, 2004; Xie et al., 2020). This study will employ the Eckhardt digital filter method. Xie et al. (2020) compared 9 different methods for more than 1800 catchments across the contiguous United States and finally showed that Eckhardt digital filter method has the best performance.

Despite the progress in LSTM-based streamflow modelling, several research gaps remain. While Gauch et al. (2021) utilized the CAMELS-US dataset and demonstrated superior performance of LSTMs over traditional hydrological models incorporating multiple input variables (including convective fraction, longwave radiation, potential energy, potential

evaporation, pressure, shortwave radiation, specific humidity, temperature, total precipitation and wind) and catchment attributes, they did not explore the impact of including previous streamflow data as an input. It remains unclear whether incorporating previous streamflow would lead to further improvements in LSTM performance.

Gauch et al. (2021) and many others, focused on same time-step streamflow predictions (using input data up to the current time step to predict the streamflow at current time step). However, the literature shows that the performance of LSTM models degrades with longer lead times. It is not clear how much the LSTM performance degrades for multiple days ahead streamflow forecasting, and whether its performance remains better than the simplest models like Persistence. Hence, it's important to explore the performance of LSTM models in forecasting streamflow at different lead times and compare them with simpler approaches like the naive method (Lin et al., 2024). This will help determine the potential applicability of LSTM models for real-time forecasting.

Traditional performance metrics such as NSE, Kling–Gupta Efficiency (KGE), and Root Mean Square Error (RMSE), are calculated using all streamflow values (Gupta et al., 2009). However, these metrics may not adequately capture model performance during critical periods, such as high-flow and flood events. There is a need to evaluate LSTM performance specifically during medium to high streamflow periods to understand how well models predict those events. By focusing on the non-baseflow periods, we can more accurately assess a model's capability to mitigate flood risks, which is a primary concern for water resource management.

In conclusion, the literature suggests that LSTM networks and their hybrid variations are promising tools for streamflow prediction, offering improvements over traditional models. However, their performance can be affected by the prediction horizon and the complexity of the hydrological processes. The current research gaps involve the potential of previous

streamflow data, multi-day forecasting capability, and performance during high-flow events. Further research should focus on addressing these challenges, particularly in the context of long-term forecasting and in poorly gauged basins.

2.2 Methodology

2.2.1 Data

The same data was incorporated in this study to compare it with a previous benchmark study (Gauch et al., 2021). Therefore, this study was also done on the CAMELS-US data set. CAMELS-US data set includes 671 catchments in the continental United States ranging in size from 4 to 25,000 km^2. These catchments were selected from the existing gauged catchments in the United States since they are mostly natural and have long period of gauge records (1980–2018) available from the United States Geological Survey National Water Information System (Newman et al., 2015; Kratzert, Klotz, Herrnegger, et al., 2019). Gauch et al. (2021) selected only 516 basins out of 671 basins in CAMELS-US data which are not larger than 2000 km^2 and have hourly streamflow data available from the USGS Water Information System. All observed streamflow data, meteorological forcing data and catchment attributes used for this study were downloaded from the data repository for the benchmark study by Gauch et al. (2021) which is publicly available. Hourly observed streamflow data that were averaged to daily scale for this study were originally from the USGS Water Information System. Hourly meteorological forcing data was originally from NLDAS-2 product, which provides hourly meteorological data since 1979 (Xia et al., 2012). Gauch et al. (2021) spatially averaged forcing variables for each basin and temporally averaged them on the daily scale. Catchment attributes data were originally from CAMELS-US data set. The dataset was divided into three distinct periods, as outlined in Table 2.1, to simulate the training, validation, and testing of the LSTM models. This study incorporated 11 forcing variables (see Table 2.2) and 27 catchment attributes (see Table 2.3) as inputs for LSTM models respectively.

Table 2.1: Simulation Periods.

Simulation	**Data Period**
Train	1990 - 2003
Validation	2004 - 2008
Test	2009 - 2018

2.2.2 Persistence Model

The Persistence method assumes that streamflow remains constant over time, using today's observed value as the prediction for upcoming days (Ghimire & Krajewski, 2019). This method uses the strong temporal autocorrelation typically found in short-term streamflow data, making it a fair and simple method for short term streamflow forecasting. The Persistence approach was used to serve as a baseline for comparing LSTM forecasting models.

For a streamflow observation X_t measured at time t, the persistence forecast at lead time Δt can be expressed as (Ghimire & Krajewski, 2019):

$$X_{(t+\Delta t)} = X_t \ . \tag{2.1}$$

2.2.3 Baseflow separation

Eckhardt method (Eckhardt, 2004) was used to calculate baseflow in this study as per the recommendation provided by Xie et al. (2020). The two-parameter (recession constant & maximum baseflow index) digital filter equation for the Eckhardt method can be expressed as:

$$b_k = \frac{(1 - BFI_{max})\alpha b_{k-1} + (1 - \alpha)BFI_{max} y_k}{(1 - BFI_{max})} \ , \tag{2.2}$$

where b_t is the baseflow and f_t is the total streamflow at t^{th} time step. α is the recession constant that can be found by recession analysis (Eckhardt, 2008). The maximum baseflow index (BFI_{max}) is determined based on field studies of hydrogeological characteristics. According to Eckhardt (2008), BFI_{max} values vary depending on stream and aquifer types: it is 0.8 for perennial streams over porous aquifers, 0.5 for ephemeral streams over porous aquifers, and 0.25 for perennial streams over hard rock aquifers.

To estimate the BFI_{max} without conducting field investigations of hydrogeological conditions, Collischonn and Fan (2012) proposed a backward filter method that relies on the recession constant, written as:

$$b_{k-1} = \frac{b_k}{\alpha} \ ; b_k \leq y_k \ . \tag{2.3}$$

The daily streamflow is processed using the backward filter method derived from Eq. 2.3, and BFI_{max} is determined by the ratio of the maximum potential total baseflow to the total streamflow. This backward filter method was used to compute distinct BFI_{max} values for each basin, partially capturing the variability in soil properties and spatial differences in hydroclimatic factors.

This base flow separation method was used during the second experiment of this study (which is explained in Subsection 2.2.4) to identify the non-baseflow periods for each basin and evaluate the streamflow predictability of LSTM and Persistence in non-baseflow periods.

2.2.4 Long Short-Term Memory (LSTM) network

Long-Short-Term Memory (LSTM) networks (Hochreiter & Schmidhuber, 1997) are a special type of recurrent neural networks (RNNs) developed to capture long term dependencies in sequential data. They utilize an internal memory cell that is dynamically updated through gating mechanisms, including an input gate (regulating new information flow), a forget gate (controlling memory retention), and an output gate (managing state-to-output transmission). The structure of an LSTM unit is depicted in Figure 2.1.

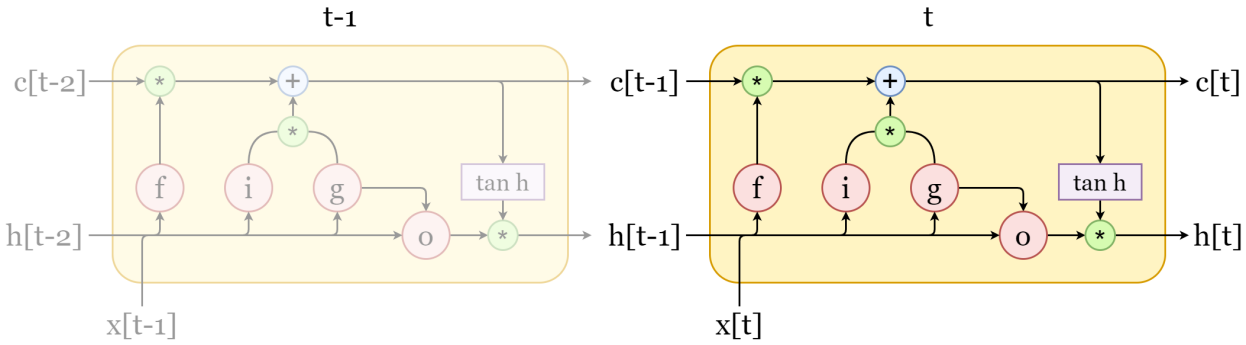

Figure 2.1: Schematic architecture of a standard LSTM cell.

$i[t]$, $f[t]$, and $o[t]$ in Figure 2.1 are the input gate, the forget gate, and the output gate, respectively; $g[t]$ is the cell input; and $x[t]$ is the network input at the time step t. $h[t-1]$ is the recurrent input; $c[t-1]$ is the cell state from the previous time step. tanh is the hyperbolic tangent function, and $(*)$ indicates multiplication by elements. For a more detailed description of LSTMs, especially in the context of rainfall-runoff modelling, refer to Kratzert, Klotz, Shalev, et al. (2019).

In all Chapter 2, Chapter 3 and Chapter 4, I utilized NeuralHydrology, a Python library designed for implementing deep learning models, such as LSTMs, in hydrological

applications. Developed primarily for students and researchers, NeuralHydrology simplifies the process of applying deep learning to rainfall-runoff modeling and other other problems related to hydrology (Kratzert et al., 2022). Users can specify various configurations of the model including input and target data, architecture, evaluation metrics, and training/validation/testing periods through a straightforward YAML configuration file.

Table 2.2 and Table 2.3 show the meteorological forcing variables and catchment attributes used as input data for the LSTM model.

Table 2.2: Dynamic inputs to the LSTM model.

Variable	**Unit**
Total precipitation	kg m^{-2}
Air temperature	K
Surface pressure	Pa
Surface downward-longwave radiation	W m^{-2}
Surface downward-shortwave radiation	W m^{-2}
Specific humidity	kg kg^{-1}
Potential energy	J kg^{-1}
Potential evaporation	J kg^{-1}
Convective fraction	-
u wind component	m s^{-1}
v wind component	m s^{-1}

Table 2.3: Static inputs to the LSTM model.

Attribute	**Unit**
Precipitation mean	mm/day
Potential Evapotranspiration mean	mm/day
Aridity index	-
Precipitation seasonality	-
Snow fraction	-
High precipitation frequency	days/year
High precipitation duration	days
Low precipitation frequency	days/year
Low precipitation duration	days
Elevation	meter above sea level
Slope	m/km
Area	km^2
Forest fraction	-
LAI (Leaf Area Index) max	-
LAI (Leaf Area Index) difference	-
GVF (Green Vegetation Fraction) max	-
GVF (Green Vegetation Fraction) difference	-
Soil depth (Pelletier)	m
Soil depth (STATSGO)	m
Soil Porosity	-
Soil conductivity	cm/hr
Max water content	m
Sand fraction	%
Silt fraction	%
Clay fraction	%
Carbonate rocks fraction	-
Geological permeability	m^2

Refer to Appendix A for descriptions of catchment attributes. These meteorological forcing variables and catchment attributes are used in the benchmark LSTM model. The model used all dynamic input variables within the look-back period up to t^{th} day and made the streamflow prediction of the t^{th} day, as shown in Figure 2.2. For our first experiment, past streamflow was incorporated as an additional dynamic input to the benchmark model. For our second experiment, the model was set up to predict 1, 2 and 3 days ahead streamflow separately.

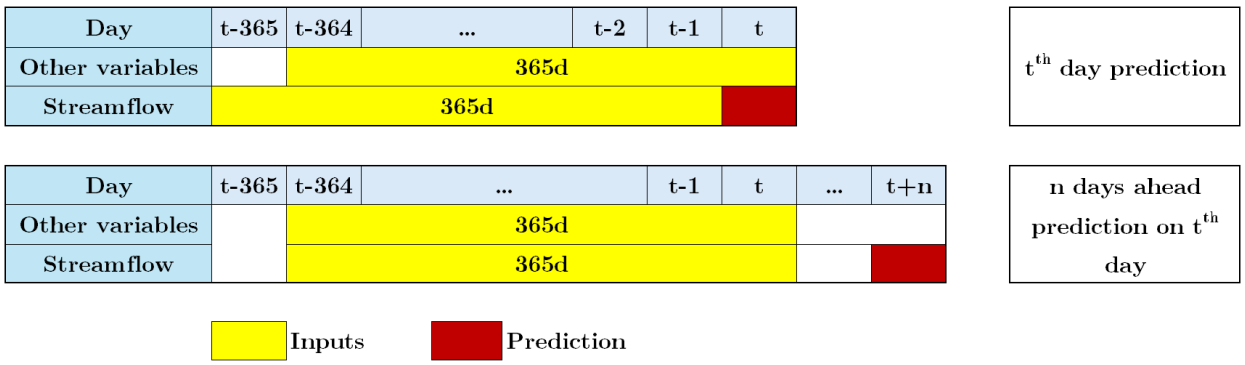

Figure 2.2: Time windows of input variable sequences and the prediction (top panel) first experiment (bottom panel) second experiment.

The benchmark LSTM model aligns with the Naive LSTM model investigated by Gauch et al. (2021), the same hyperparameters they used were adopted for the benchmark model. When we incorporated past streamflow as an additional dynamic input to the benchmark model, the hyperparameters were tuned for the new model, focusing on only four hyperparameters (Hidden Size, Output Dropout, Batch Size, Sequence Length). This decision was made because extremely long input sequences significantly increase training time. All other parameters and configuration settings remained consistent with those in Gauch et al. (2021)'s study. For further details on hyperparameter tuning, please refer to Appendix B. The

selected hyperparameter combination also remained the same for the second experiment. NeuralHydrology (Kratzert et al., 2022), a python library which allows users to run deep learning models was used in this study to set up LSTM models.

2.2.5 Evaluation metrics

Since single evaluation metric can not fully demonstrate the reliability and accuracy of streamflow predictions, multiple evaluation metrics were considered to comprehensively analyze and represent the results. The four evaluation metrics used in this Chapter are

$$\text{Pearson Correlation Coefficient } (r) = \frac{\sum_{i=1}^{n}(\hat{y}_i - \mu_{\hat{y}})\sum_{i=1}^{n}(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \mu_{\hat{y}})^2 \sum_{i=1}^{n}(y_i - \mu_y)^2}} \ , \tag{2.4}$$

$$\text{Nash Sutcliffe Efficiency } (NSE) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \mu_y)^2} \ , \tag{2.5}$$

$$\text{Kling Gupta Efficiency } (KGE) = 1 - \sqrt{(r-1)^2 + \left(\frac{\sigma_{\hat{y}}}{\sigma_y} - 1\right)^2 + \left(\frac{\mu_{\hat{y}}}{\mu_y} - 1\right)^2} \ , \tag{2.6}$$

$$\text{Mean Absolute Error } (MAE) = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \ , \tag{2.7}$$

where y_i and $\hat{y}_i$ represent observed and predicted values respectively, while μ and σ represent mean and standard deviation.

2.3 Results & Discussion

2.3.1 Incorporating historical streamflow data as an input

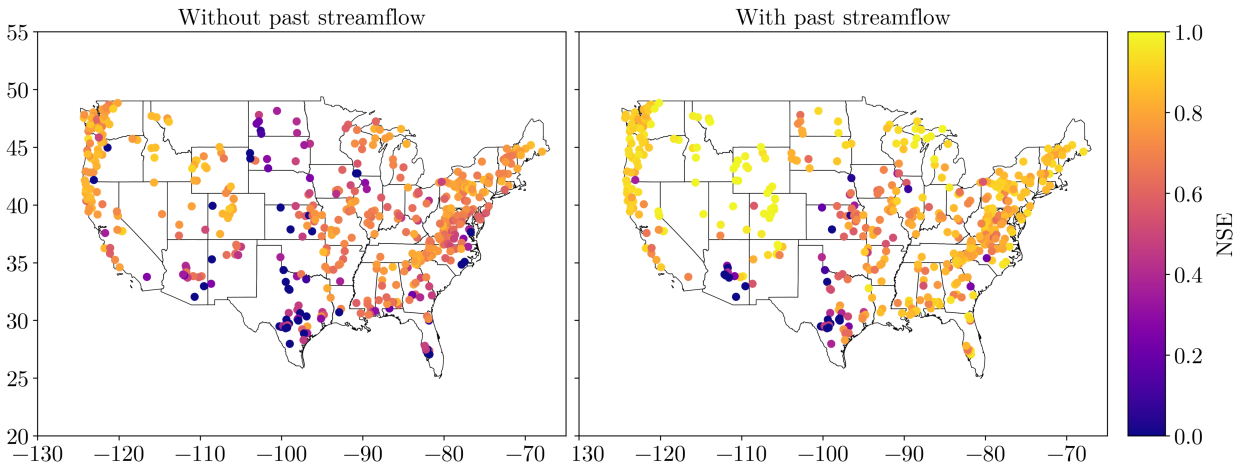

Figure 2.3: NSE for LSTM model (left) without (right) with previous streamflow data as an input.

The impact of including historical streamflow data as an input variable in LSTM models is significant for streamflow prediction, leading to a notable improvement in model performance. As seen in Figure 2.3, the integration of past streamflow data results in a notable increase in NSE values across a large number of basins. Specifically, most basins located in the western and eastern regions of the United States exhibit NSE values exceeding 0.8, indicating a high degree of accuracy when incorporating this historical streamflow data. However, it's worth noting that not every basin demonstrates the same level of improvement, with certain areas in the central, southern, and southwestern U.S. not showing enhanced NSE scores with this additional data. This variability suggests that the effectiveness of historical streamflow data as an input is not uniform across all regions, and other factors might be at play in those

basins where the improvement is less significant.

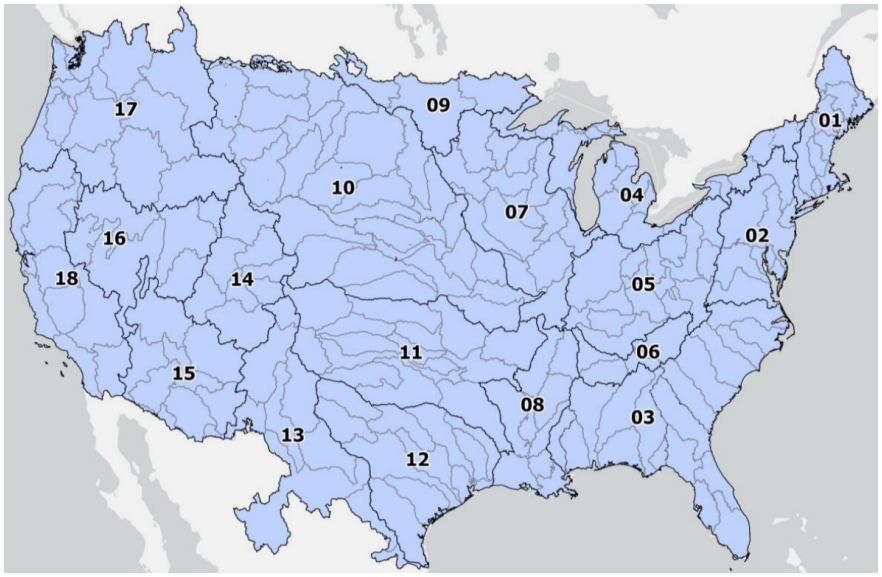

Figure 2.4: USGS Hydrologic Unit Code (HUC) map for the Contiguous United States (Incorporated & the Spatial Sciences Laboratory at Texas AM, 2023).

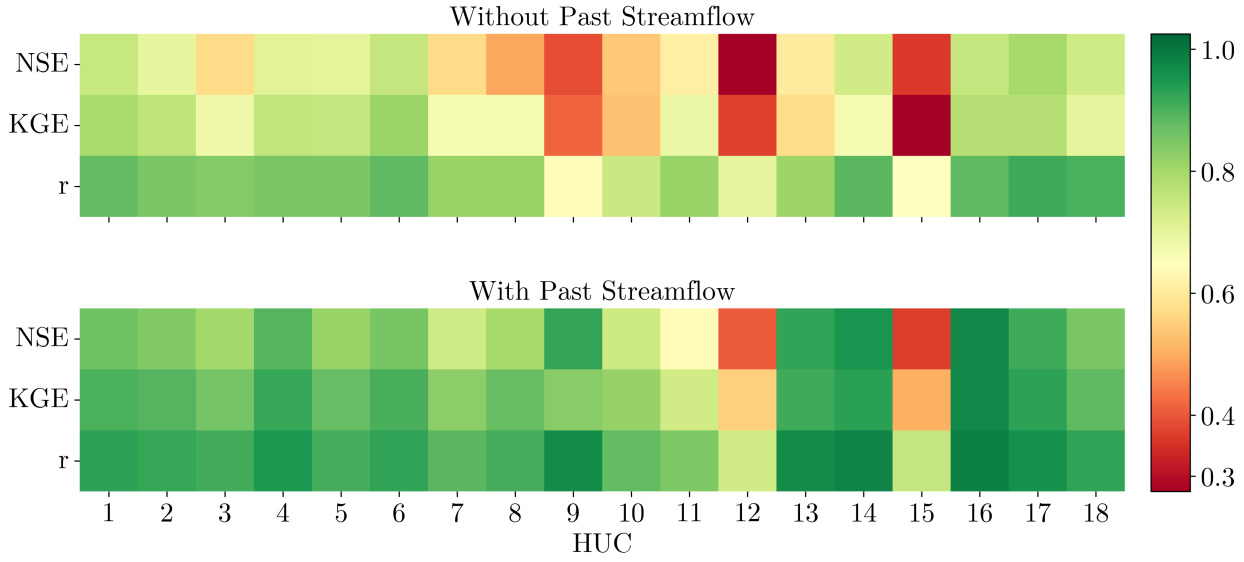

Figure 2.5: Spatially (HUC wise) averaged metrics (top) without (bottom) with past streamflow data as an input.

Further examination of spatially averaged metrics across 18 hydrological regions (Figure 2.4) within the Contiguous United States (CONUS), as shown in Figure 2.5, provides additional support for the overall improvement in model performance. The three metrics used (NSE, KGE, and r) show increases in all 18 regions when historical streamflow is considered in the LSTM model. Regions such as HUC 7 through 10, and HUC 13, which initially presented lower metric values, exhibit considerable gains in mean NSE and KGE. Conversely, areas like HUC 12 and 15 continue to display relatively lower NSE and KGE scores, which implies that the performance of the LSTM model in these regions might be limited by other factors that require further investigation. These results emphasize the time-dependent nature of streamflow, as evidenced by strong correlations between present and past streamflow values. This is supported by the findings of Lin et al. (2024), who showed that lagged streamflow is more influential on forecasting accuracy compared to lagged precipitation, further highlighting the importance of using historical streamflow data in models.

The relationship between catchment characteristics and LSTM model performance metrics, as analyzed in Figure 2.6, reveals some important insights. The heatmap highlights significant correlation patterns within a specific subset of basins. Figure 2.7 identifies this subset of 364 basins out of 516 basins (more than 70% of the total) where LSTM performance shows a strong correlation with certain catchment attributes.

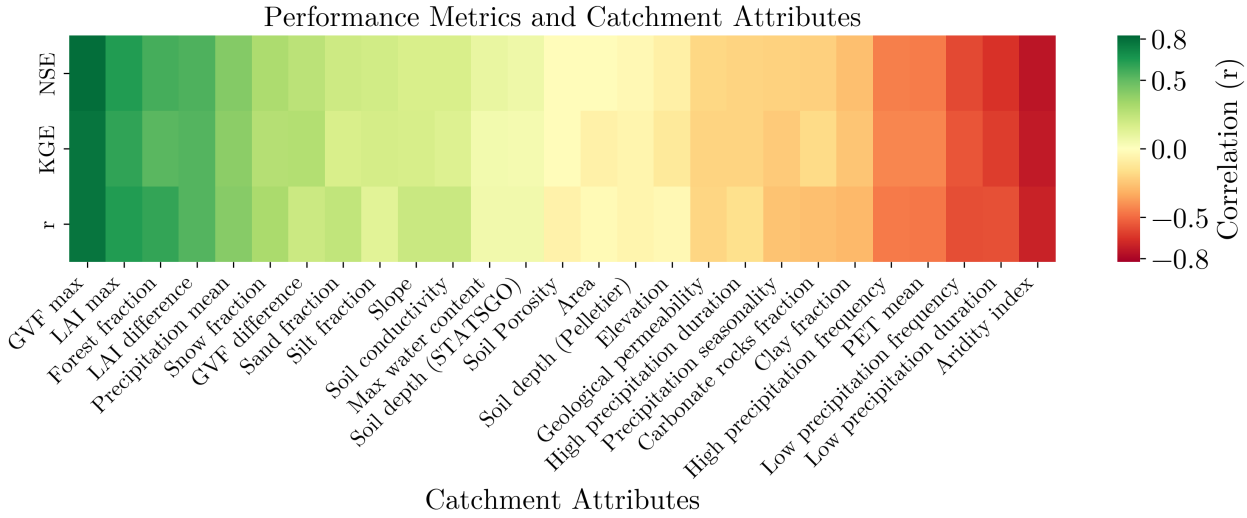

Figure 2.6: Correlation between catchment attributes and performance metrics.

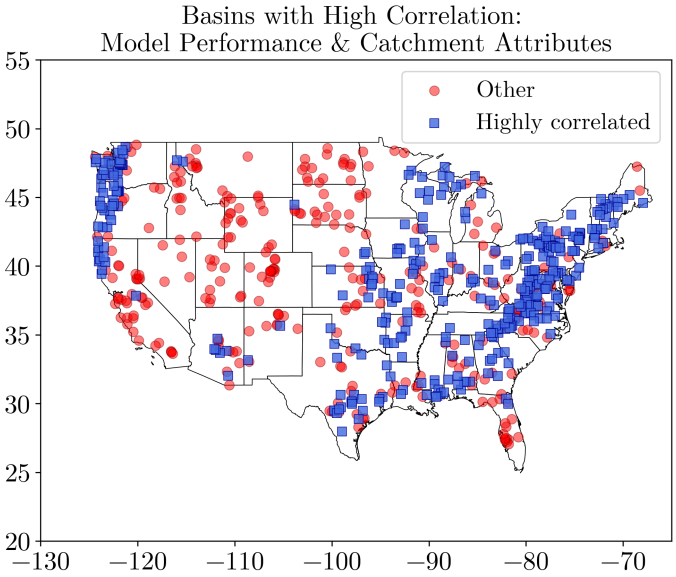

Figure 2.7: Basins with high correlation between LSTM performance and certain catchment attributes.

Within this selected subset of basins, the performance metrics NSE, KGE, and r are

highly positively correlated with the maximum monthly mean of green vegetation fraction (GVF Max), the maximum monthly mean of leaf area index (LAI Max), and the fraction of the catchment covered by forest. This suggests that LSTM models tend to perform better in basins with higher vegetation density and forest cover. On the other hand, NSE, KGE, and r are highly negatively correlated with the aridity index, low precipitation duration, and low precipitation frequency. This indicates that LSTM models may struggle to accurately predict streamflow in regions with arid conditions or where precipitation events are infrequent and short in duration.

In simpler terms, within the selected subset of basins, the accuracy of LSTM streamflow predictions is higher in basins with lower aridity, shorter and less frequent low precipitation events, and greater vegetation and forest cover. Conversely, basins with arid conditions, long and frequent low precipitation, and less vegetation tend to pose greater challenges for LSTM models. These findings align with the understanding that catchment characteristics significantly influence hydrological modelling (Addor et al., 2017; Kratzert, Klotz, Shalev, et al., 2019) and highlight the importance of adapting models to the dominant characteristics of a region. Gauch et al. (2021) demonstrated that LSTM models can handle a wide range of input variables and catchment attributes, but these complex interactions can still impact the accuracy of streamflow predictions.

Figure 2.8 illustrates the improved performance of the LSTM model when using past streamflow as input. Most of the data points in the scatter plots, each representing an individual basin, are positioned above the 1-to-1 line for NSE, KGE, and Pearson Correlation, and below the 1-to-1 line for RMSE. This confirms that including historical streamflow data leads to increases in NSE, KGE, and Pearson r values, as well as lower RMSE values.

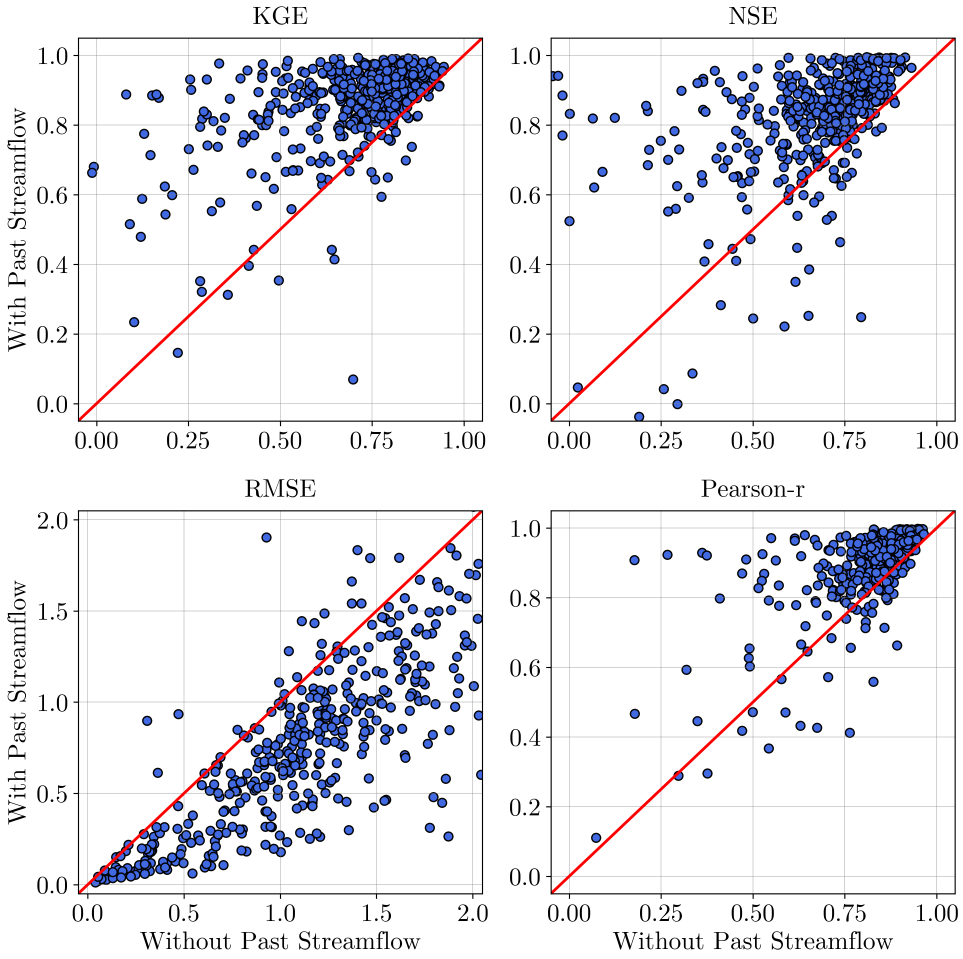

Figure 2.8: Comparison of performance metrics of LSTM models with (y axis) and without (x axis) past streamflow as an input.

The cumulative distribution function (CDF) curves for NSE values, shown in Figure 2.9, further illustrate that models incorporating past streamflow data show a higher frequency of higher NSE values. For instance, the data show that when historical streamflow data were included, over 60% of the basins achieved NSE values greater than 0.8, compared to only about 20% when this input was not considered. These findings collectively underscore the considerable positive influence that incorporating historical streamflow data has on the

accuracy of streamflow predictions using LSTM models.

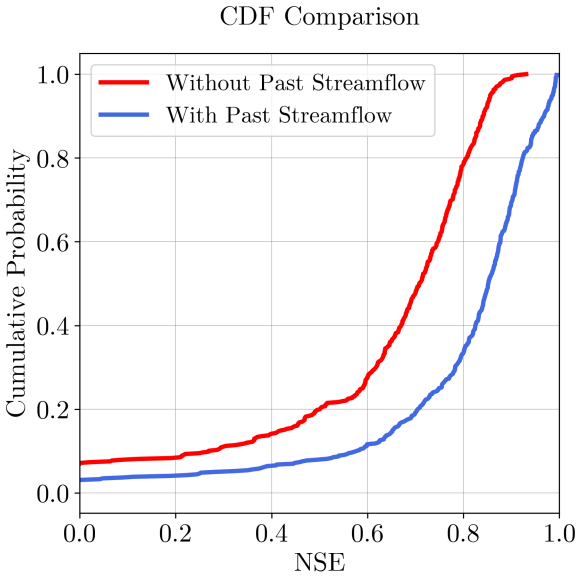

Figure 2.9: CDF for NSE of basins.

2.3.2 Multiple days ahead streamflow forecasting using LSTM

Despite the overall strong performance of LSTM models when past streamflow is used as an input, it's also been observed that predictive accuracy diminishes with increasing forecast lead times.

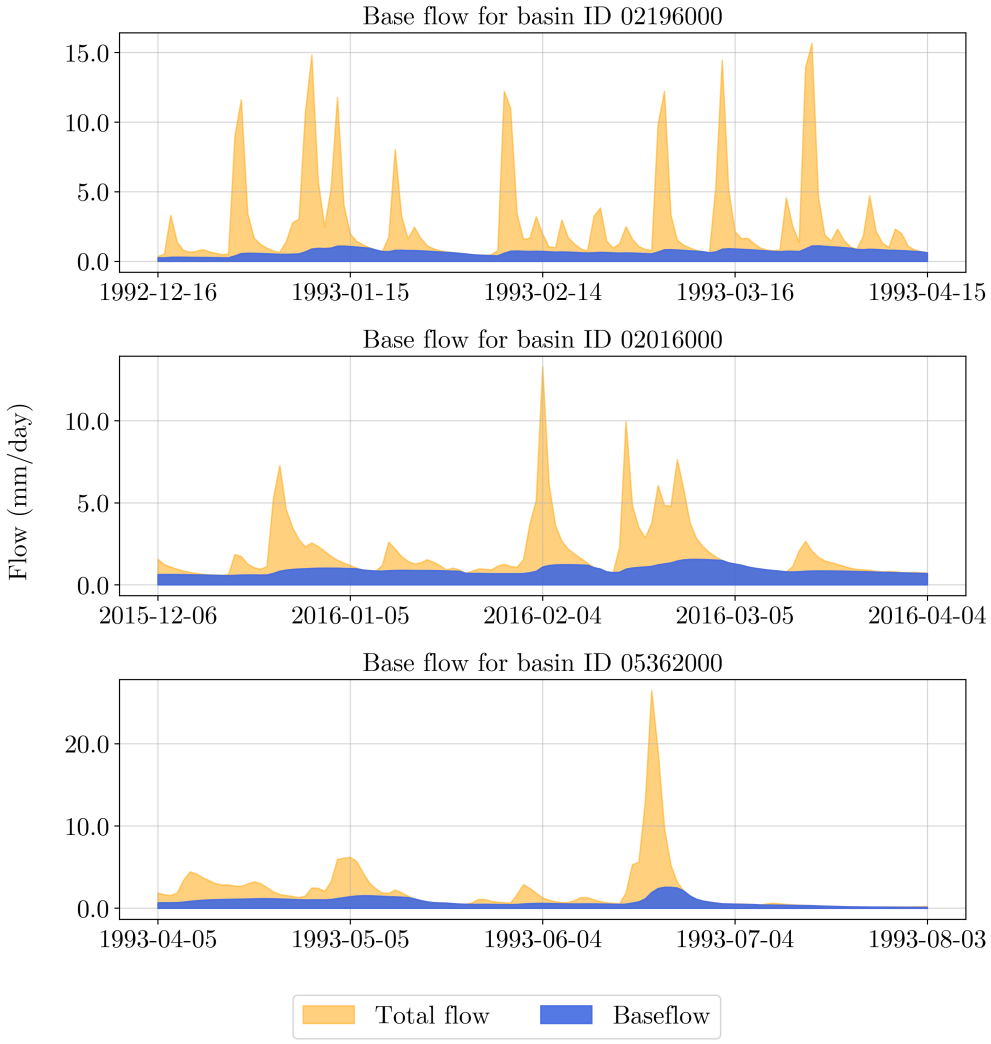

Figure 2.10: Baseflow change in different basins.

To evaluate the streamflow forecasting accuracy of LSTMs and Persistence, baseflow

periods for each basin were identified and only the non-baseflow periods were considered when calculating evaluation metrics. Figure 2.10 depicts how baseflow changes with time in a few different basins.

A comparison of the LSTM model performance against the Persistence model offers additional context, particularly for multiple-day-ahead forecasts. The Persistence model makes predictions by assuming the future streamflow would be same as the current streamflow. As seen in Figure 2.11, the LSTM model, compared to the Persistence model, demonstrates superior performance in terms of r, MAE and RMSE across all multiple-day-ahead forecasts. However, the NSE does not show a clear advantage for the LSTM model except in one-day-ahead forecasting. In fact, for two- and three-day-ahead streamflow forecasts, both models yield negative NSE values in most basins, indicating unreliable predictions at longer lead times with only a few exceptions among the 516 basins.

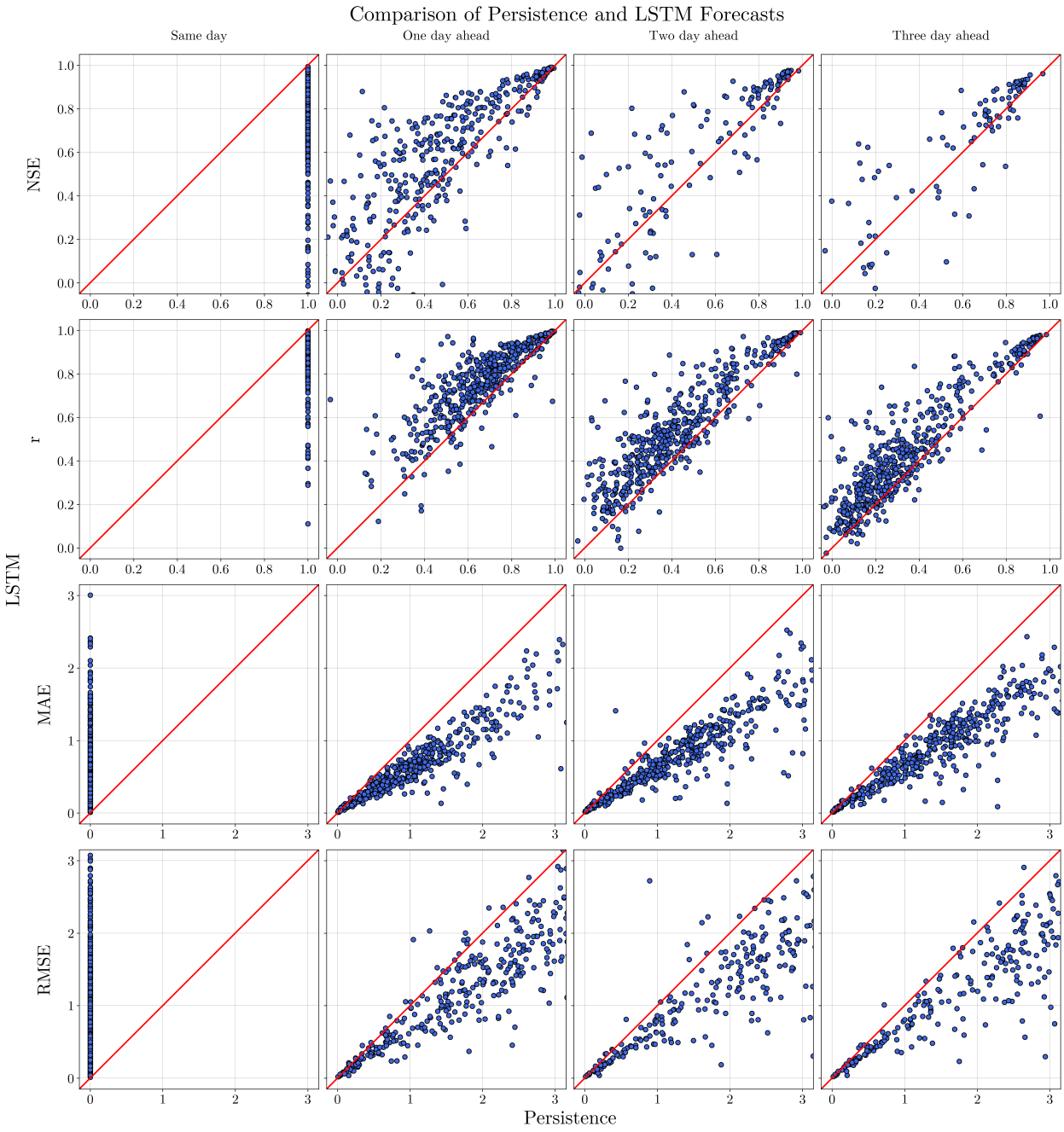

Figure 2.11: Performance metric comparison between LSTM and Persistence (1st row) NSE (2nd row) Pearson Correlation (3rd row) MAE (4th row) RMSE (1st column) one day ahead (2nd column) 2 days ahead (3rd column) 3 days ahead forecast.

Figure 2.12 and Figure 2.13 illustrate again the substantial decline in NSE values of both Persistence and LSTM models as the forecast lead time increases. For instance, when predicting current-day streamflow, NSE value of LSTM model remain above 0.7 for more than 76% of the basins, but that decreases to around 31% for one-day-ahead forecasts, and further down to about 14% & 10% for two and three-days ahead forecasts respectively. On the other hand, 21% of basins have NSE of Persistence higher than 0.7 for one-day-ahead forecasts, but that percentage decreases to around 12% & 9% for two and three-days ahead forecasts respectively. As shown in both Figure 2.11 and Figure 2.12, the LSTM outperforms the Persistence model in one-day-ahead forecasts. However, both models performed poorly in most basins for two-day and three-day-ahead forecasts.

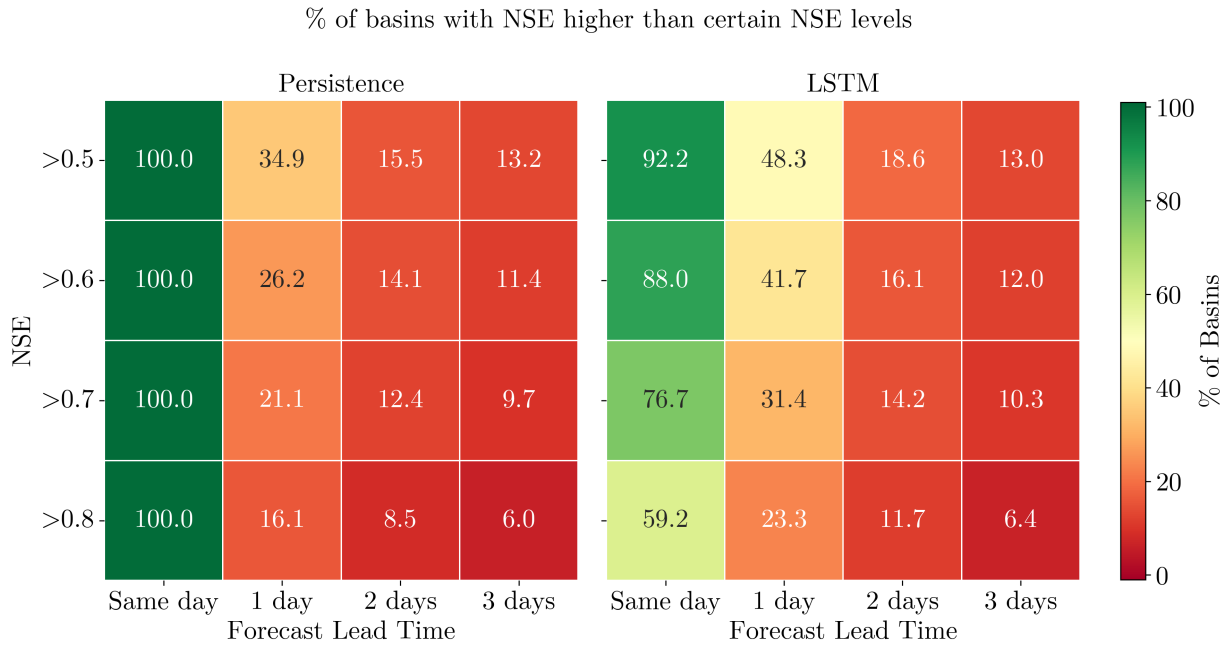

Figure 2.12: NSE drop in LSTM and Persistence models as the lead time increases. Rows represent percentage of basins that exhibits NSE higher than certain NSE levels.

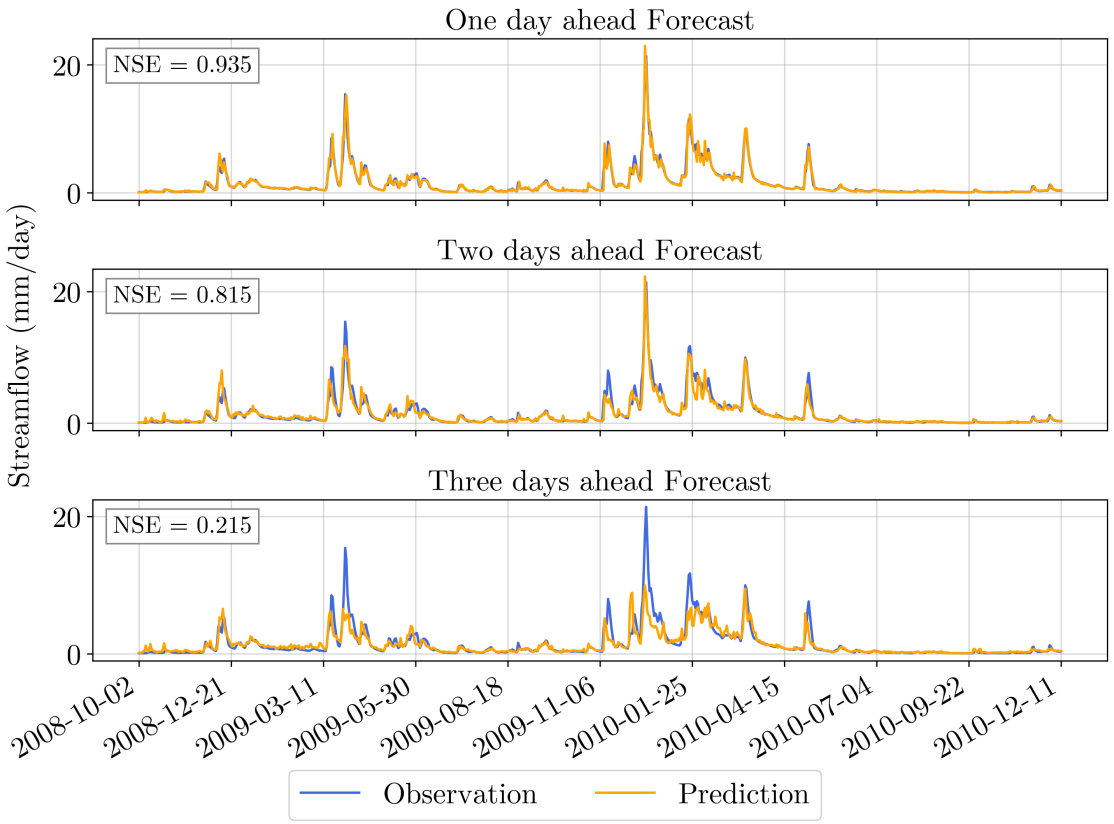

Figure 2.13: Observed & forecasted hydrograph comparison with different lead times (Basin ID : 02371500 , Area : 1292.79 km^2).

2.4 Conclusion

The findings of this study offer valuable insights into the use of LSTM networks for streamflow prediction. One of the most significant conclusions is the critical role of incorporating historical streamflow data as an input feature in LSTM models. The study demonstrates that including past streamflow information leads to a substantial enhancement in the performance of these models. This improvement is particularly noticeable in river basins situated in the western and eastern regions of the United States, where the inclusion of historical data results in significant gains in NSE values. This highlights the inherent temporal dependencies within streamflow series, where past values strongly influence present and future streamflow.

However, the effectiveness of incorporating historical streamflow data is not uniform across all geographical regions. Certain basins in the central, southern, and southwestern U.S. did not exhibit the same level of improvement in NSE scores with the addition of the historical streamflow data. This suggests that while past streamflow is generally a crucial predictor, other catchment-specific factors might play a more dominant role in these regions.

Furthermore, the study reveals the influence of catchment characteristics on the performance of LSTM models. Specifically, the aridity index, as well as low precipitation duration and frequency, show a negative correlation with LSTM performance metrics. This indicates that LSTM models may face challenges in accurately predicting streamflow in basins characterized by arid conditions and infrequent or short precipitation events. Conversely, the fraction of forest and vegetation cover exhibits a positive correlation with LSTM performance. This suggests that LSTM models tend to perform better in basins with higher vegetation density and more extensive forest cover. These findings underscore the importance of considering the hydrological setting and dominant characteristics of a region when developing

and implementing streamflow prediction models.

Another key finding pertains to the forecasting capability of LSTM models over different lead times. While LSTM models demonstrate strong performance in predicting same-day streamflow, their accuracy tends to decrease as the forecast lead time increases. This decline in performance with longer lead times suggests a limitation inherent in data-driven approaches that rely solely on historical data.

In comparison to the Persistence model, which assumes future streamflow would be same as the current streamflow, the LSTM model shows better performance for one-day-ahead forecasting in terms of NSE. Notably, for two- and three-day-ahead streamflow forecasts, both the LSTM and the Persistence models struggle, often resulting negative NSE values for most basins, indicating unreliable predictions at these longer lead times. This highlights the significant challenge of using data-driven methods solely relying on historical forcing data, alone for streamflow forecasting beyond very short time frames.

Overall, the study reinforces the potential of LSTM models as a powerful tool for streamflow prediction, particularly when historical streamflow data is incorporated. However, the findings also emphasize that relying solely on historical forcing data in LSTM models has inherent limitations, especially when forecasting streamflow at longer lead times.

To overcome these limitations, future research could explore incorporating forecasted forcing data (such as future precipitation and temperature) into LSTM models, which has the potential to enhance their performance in longer-term streamflow forecasting. Additionally, the development and investigation of hybrid models that combine the strengths of data-driven approaches like LSTM with physically based hydrological models could offer a promising avenue for achieving more accurate and reliable streamflow predictions across various time scales. Addressing these challenges and exploring these potential improvements are essential

for advancing the field of streamflow prediction and supporting effective water resource management in diverse hydrological settings.

Chapter 3

The Effect of Noisy Precipitation Inputs on Streamflow Predictability of LSTMs

3.1 Introduction & Background

An accurate prediction of streamflow is critical for effective water resource management, flood forecasting, and a variety of other applications (Hunt et al., 2022; Nifa et al., 2023). Hydrological models are essential tools, simulating the movement of water through the hydrological cycle and predicting streamflow based on various inputs (Kratzert et al., 2018; Lafon et al., 2013). These models range from simple conceptual models to complex physically based models, each with its own set of assumptions and limitations (Arsenault et al., 2023; Hong et al., 2006). However, a common challenge across all types of hydrological models is the inherent uncertainty associated with their inputs, parameters, and structure (Huard &

Mailhot, 2006; Pathiraja et al., 2018).

Precipitation, as the primary driver of the hydrological cycle, plays a crucial role in streamflow generation (Lafon et al., 2013). Accurate and reliable precipitation data are therefore essential for hydrological modelling and streamflow prediction. Precipitation data can be obtained from various sources, including ground-based rain gauges, weather radar, and satellite remote sensing (Hong et al., 2006). However, each of these sources has its own limitations and uncertainties, such as spatial distribution issues, measurement errors, and resolution constraints (Vergara et al., 2014). Rain gauges, while providing direct measurements of rainfall at a point, can be sparse and unevenly distributed, particularly in mountainous or remote regions (Feng et al., 2021; Lafon et al., 2013). Weather radar, on the other hand, can provide spatially continuous estimates of rainfall but is subject to errors due to beam broadening, attenuation, and ground clutter (Vergara et al., 2014). Satellite-based precipitation estimates offer global coverage but often have coarse spatial and temporal resolutions. Additionally, they may be affected by errors due to cloud cover, sensor limitations, and retrieval algorithms (Hong et al., 2006; Vergara et al., 2014).

Errors in precipitation data could significantly impact the accuracy of streamflow predictions. Overestimation or underestimation of rainfall can lead to corresponding errors in simulated streamflow, affecting flood forecasts, water supply estimates, and other critical decisions. Therefore, it is essential to account for the uncertainty in precipitation data when using hydrological models for streamflow prediction (Hong et al., 2006; Pathiraja et al., 2018; Vergara et al., 2014).

In recent years, machine learning (ML) techniques, particularly Long Short-Term Memory (LSTM) networks, have emerged as promising tools for hydrological modelling and streamflow prediction (Arsenault et al., 2023). LSTMs are a type of recurrent neural network (RNN) that

can effectively capture the temporal dependencies in sequential data, making them well-suited for modeling the complex dynamics of the hydrological cycle (Hunt et al., 2022). Several studies have demonstrated the ability of LSTMs to achieve state-of-the-art performance in streamflow prediction, often outperforming traditional hydrological models (Feng et al., 2021; Kratzert, Klotz, Herrnegger, et al., 2019; Kratzert, Klotz, Shalev, et al., 2019; Nifa et al., 2023).

LSTMs can learn complex relationships between precipitation inputs and streamflow outputs (Feng et al., 2021). They can also incorporate other relevant inputs, such as temperature, solar radiation, and catchment characteristics, to improve their predictive accuracy (Kratzert, Klotz, Herrnegger, et al., 2019; Kratzert, Klotz, Shalev, et al., 2019). Furthermore, LSTMs can be trained on large datasets of historical streamflow and meteorological data, allowing them to capture a wide range of hydrological conditions and adapt to changing climate patterns (Feng et al., 2021; Nifa et al., 2023).

Despite their advantages, LSTMs still face challenges because of uncertainty in precipitation data (Frame et al., 2021). While LSTMs can learn to filter out some of the noise in the inputs, their performance can still be degraded by significant errors in precipitation estimates. Therefore, it is important to investigate the sensitivity of LSTMs to precipitation errors and to develop strategies for mitigating their impact on streamflow predictions (S. Liu et al., 2023).

Many studies have explored the impact of precipitation errors on hydrological modelling and streamflow prediction. Hong et al. (2006) evaluated the influence of precipitation error on streamflow forecasting uncertainty using a conceptual rainfall-runoff model and a Monte Carlo simulation approach. They found that the error in precipitation data could significantly affect the accuracy of streamflow forecasts, particularly during extreme events.

Kratzert et al. (2021) investigated the effect of different meteorological forcing products on the performance of LSTMs for streamflow prediction. They found that the forcing product choice could significantly affect the accuracy of streamflow predictions, highlighting the importance of using accurate precipitation data. Pathiraja et al. (2018) proposed a data-driven framework for estimating model uncertainty in hydrologic data assimilation. They used kernel conditional density estimation to estimate the probability density of model errors and found that their approach improved the accuracy of streamflow forecasts. S. Liu et al. (2023) stated that noises/errors observations consequently affect streamflow predictions, underscoring their sensitivity to input variability. This finding is supported by Frame et al. (2021), who observed that errors in atmospheric forcings, primarily precipitation, significantly affect the accuracy of LSTM streamflow predictions using data from the National Water Model (NWM).

Despite the advances in LSTM hydrological modelling, the challenge of streamflow prediction using noisy precipitation inputs remains an active area of research. The goal of this study is to evaluate the performance of LSTM models for streamflow prediction when the model is fed with noisy precipitation data. This study aims to address the following key questions:

- How does the introduction of noise in precipitation input data affect the accuracy of LSTM models for streamflow prediction?
- Are there specific regions or basins where LSTM models are more sensitive to precipitation noise?
- Can LSTM models effectively identify and filter out precipitation errors?

By addressing these questions, this study aims to provide valuable insights into the

capabilities and limitations of LSTM models for streamflow prediction under real-world conditions, where precipitation data is often subject to various sources of error and uncertainty. The findings of this study might help guiding the development of more robust and reliable hydrological models for water resource management and flood risk assessment.

3.2 Methodology

3.2.1 Data

In this study, the same dataset utilized in Chapter 2 was used. Specifically, the dataset includes observed streamflow data, meteorological forcing data, and catchment attributes, all of which were sourced from the publicly available data repository associated with the study by Gauch et al. (2021). The observed streamflow data were originally from the United States Geological Survey (USGS) National Water Information System (NWIS), a reliable and widely used source for hydrological data in the United States. The meteorological forcing data were originally from the NLDAS-2 (North American Land Data Assimilation System) product (Xia et al., 2012). Additionally, the catchment attributes were originally sourced from the CAMELS-US dataset (Newman et al., 2015), which offers a comprehensive set of physical and climatic characteristics for 671 catchments across the United States.

3.2.2 LSTM model Set-Up

Dynamic inputs (except the precipitation) and static inputs and the temporal splitting of data for model training, validation, and testing of the LSTM model remained unchanged from the LSTM model which was simulated with historical streamflow data in Chapter 2 to maintain consistency in the experimental setup. The LSTM network architecture used in Chapter 2 was also retained in this study, with minor modifications. Specifically, noise was introduced to the last three days of precipitation (P) inputs to investigate the model's sensitivity to uncertainties in the input data. The noise (ε) at time step t was characterized as:

$$P_{t-i}(\text{modified}) = P_{t-i}(\text{observed}) \left(1 \pm \frac{\varepsilon}{100}\right) \ , \tag{3.1}$$

where $i \in \{0, 1, 2\}$, $\varepsilon \sim uniform(0, r)$, and $r \in \{10, 30, 50\}$.

Three distinct scenarios were considered, each corresponding to different maximum noise levels: 10%, 30%, and 50%. In each scenario, precipitation data for three days was modified using Eq. 3.1. Additionally, within each scenario, the modification was applied in two different ways:

① Testing Period Only: The modification was applied exclusively to the precipitation inputs during the testing period,

② Training and Testing Periods: The modification was applied to the precipitation inputs in both the training and testing periods.

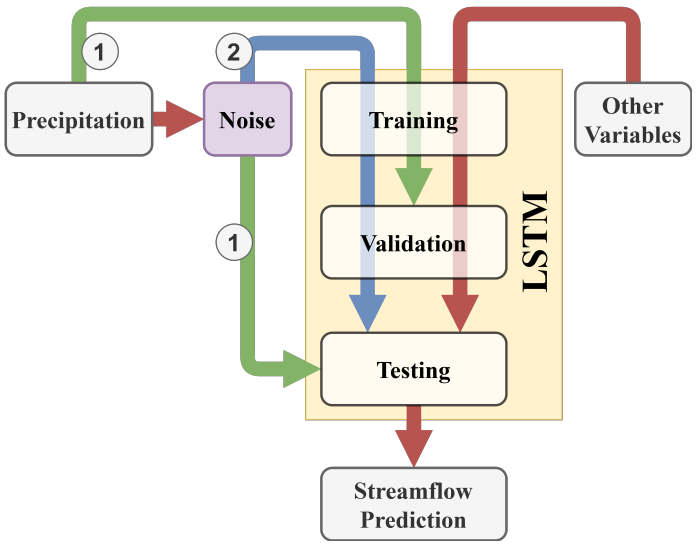

Figure 3.1: Flowchart of the methodology.

This approach resulted in a total of six different scenarios being analyzed. Due to time constraints, the hyperparameters of the LSTM network, which were previously tuned in Chapter 2, were reused across all scenarios to maintain consistency and comparability. The LSTM structure used in this study is designed only to predict 1 day of streamflow at a time, following the previous 365 days of the dynamic inputs as shown in Figure 3.2.

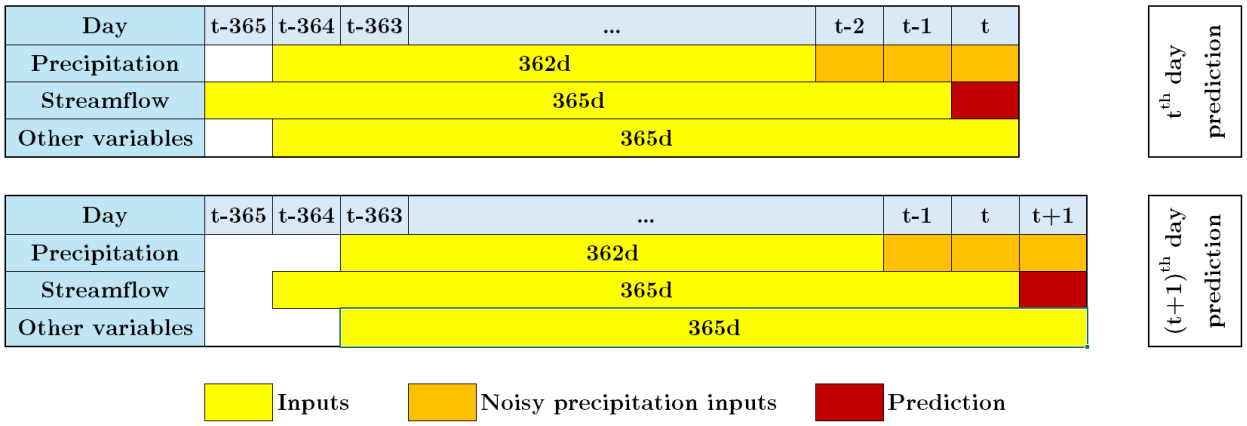

Figure 3.2: Time windows of look-back period and the prediction.

3.2.3 Evaluation Metrics

The performance of LSTM streamflow predictions was evaluated using Nash-Sutcliffe Efficiency (NSE) and Mean Absolute Percentage Error (MAPE). The Figure 3.13 highlights this subset of 384 basins out of 516 basins (more than 74% of the total) where LSTM's sensitivity strongly correlates with certain catchment attributes. MAPE which is defined as the average absolute percentage difference between predicted values and observed values, can be expressed as:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \ , \tag{3.2}$$

where y_i and $\hat{y}_i$ represent the observed and predicted values, respectively, while $\bar{y}_i$ is the mean

of y_i.

To analyze how the MAPE (y) increases with precipitation noise (x) across different basins, the MAPE values for each basin were fitted to a curve of the form $y = a + bx^c$, where a, b, and c are parameters specific to each basin. To provide a clearer comparison of how MAPE changes with precipitation noise in different basins, the curve $y = bx^c$ was plotted for each basin, excluding the parameter a (interception). To quantify the sensitivity of LSTM model performance to precipitation noise, a new metric k was introduced. This metric k is defined as a function of two components:

i. Cumulative MAPE Increment (A): The total increase in MAPE across all noise levels,

ii. Rate of MAPE Increment at Maximum Noise Level (S_{50}): The rate at which MAPE increases at the highest noise level (50%).

The definition of k can be written as:

$$k = normalize\left(A^{\alpha} S_{50}^{\beta}\right) \quad , \quad k \in [0, 1] \ , \tag{3.3}$$

$$A = \int_0^{50} bx^c \, dx = \left(\frac{b}{c+1}\right) \times 50^{c+1} \ , \tag{3.4}$$

$$S_{50} = \left.\frac{dy}{dx}\right|_{x=50} = 50^{c+1} bc \ . \tag{3.5}$$

To identify optimal values for α & β that suit well across all basins, a basin exhibiting a 25% MAPE increase (or decrease) at 50% noisy precipitation, with a $\pm$0.5% MAPE increment rate, was selected as a moderately sensitive basin. The parameters α & β were then iteratively adjusted until the sensitivity k for that basin reach 0.5. The metric k thus captures both the overall sensitivity of the LSTM model to noise and how rapidly the error grows as noise

reaches its maximum level. This provides a comprehensive measure of the model's sensitivity to precipitation noise.

3.3 Results & Discussion

3.3.1 Training & Testing LSTMs with noisy precipitation

The main question addressed whether LSTMs could identify precipitation errors and predict streamflow without affecting their performance when trained with erroneous precipitation inputs. Figure 3.3 presents the NSE of streamflow predictions across all basins for varying noise levels in precipitation inputs, comparing the model's performance with and without noisy precipitation inputs. The right column of Figure 3.3 illustrates this comparison.

The results indicate that the introduction of noise in the precipitation input data generally leads to a decline in the LSTMs' performance for streamflow predictions across most basins. Interestingly, the extent of this performance drop does not vary significantly across different noise levels in precipitation. Surprisingly, there are a few basins where LSTMs perform better with noisy precipitation inputs. However, most basins still exhibit relatively high NSE values, even with 50% noise in the precipitation inputs. In general, basins with poorer performance are mainly located in the central and eastern regions of the United States. This suggests that the geographical location and regional climate characteristics may influence the model's sensitivity to noisy data.

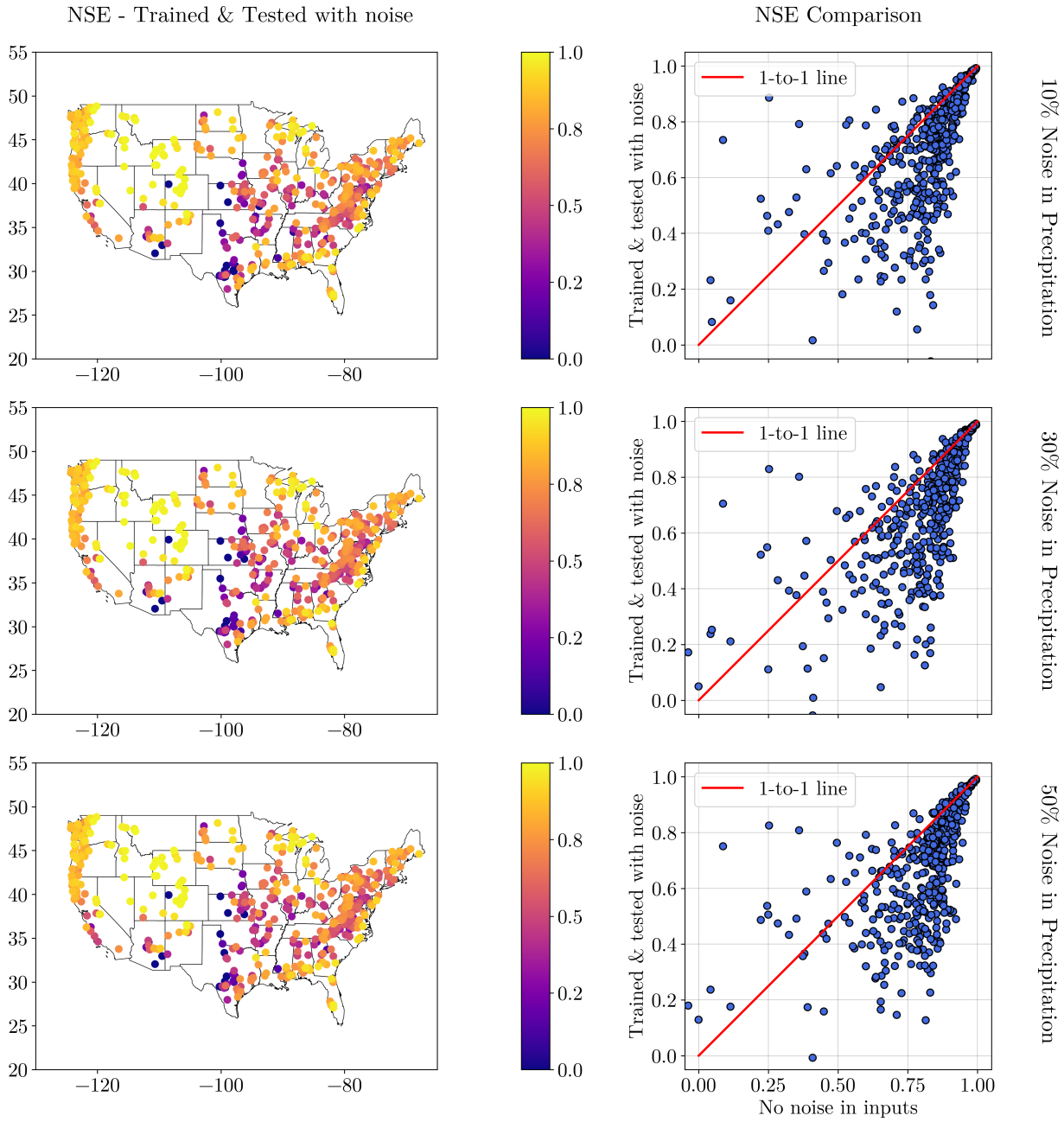

Figure 3.3: (Left column) NSE for streamflow predictions from LSTM trained & tested with noisy precipitation, (Right column) NSE comparison: "no noise in inputs" (x axis) vs "trained & tested with noise" (y axis), (Rows) represent different noise levels.

Figure 3.4 presents the NSE results when the model is tested (but not trained) with erroneous precipitation inputs. The subplots in the right column of Figure 3.4 compare the LSTM performance between scenarios where the model is trained and not trained with noisy precipitation inputs. The results indicate a significant decline in LSTM performance across most basins compared to when the model was both trained and tested with noisy precipitation inputs. Examining the spatial distribution of LSTM performance reveals that the eastern half of the US is particularly affected by the introduction of noise in the testing inputs. This demonstrates that training and testing with noisy precipitation inputs yields better results than just testing LSTMs with noisy precipitation inputs alone.

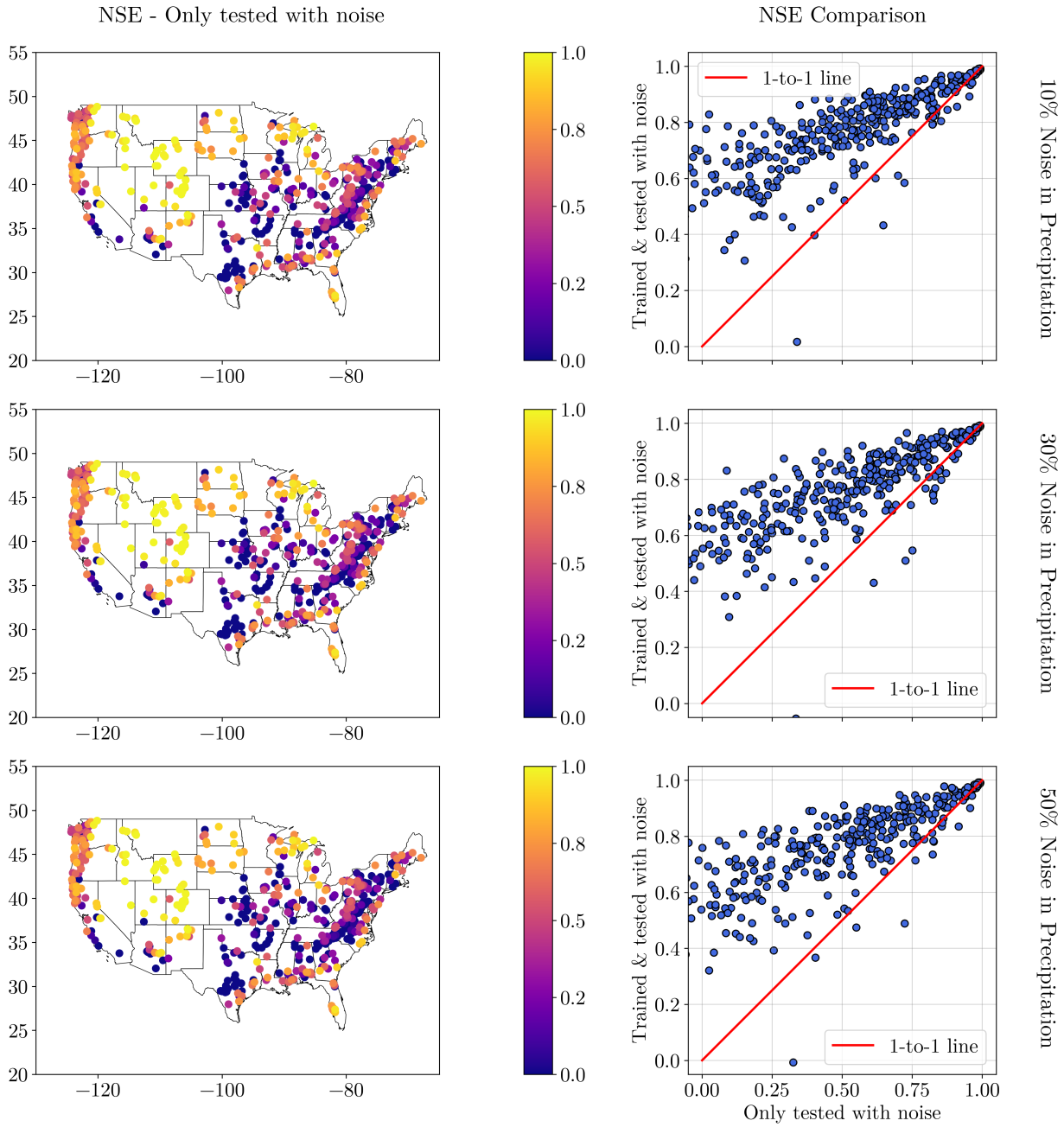

Figure 3.4: (Left column) NSE for streamflow predictions from LSTM only tested with noisy precipitation, (Right column) NSE comparison: "only tested with noise" (x axis) vs "trained & tested with noise" (y axis), (Rows) represent different noise levels.

Figure 4.7 illustrates the percentage of basins, out of the 516 considered in this study, that exceed specific higher NSE values when the LSTM model is trained and tested with noisy precipitation inputs. For instance, more than 55% of the 516 basins achieve an NSE higher than 0.7, even with a 50% noise level in the input precipitation data. This can be regarded as a notably strong performance in streamflow predictions. This underscores the resilience of LSTMs in capturing streamflow dynamics even when trained with significant levels of noise in precipitation.

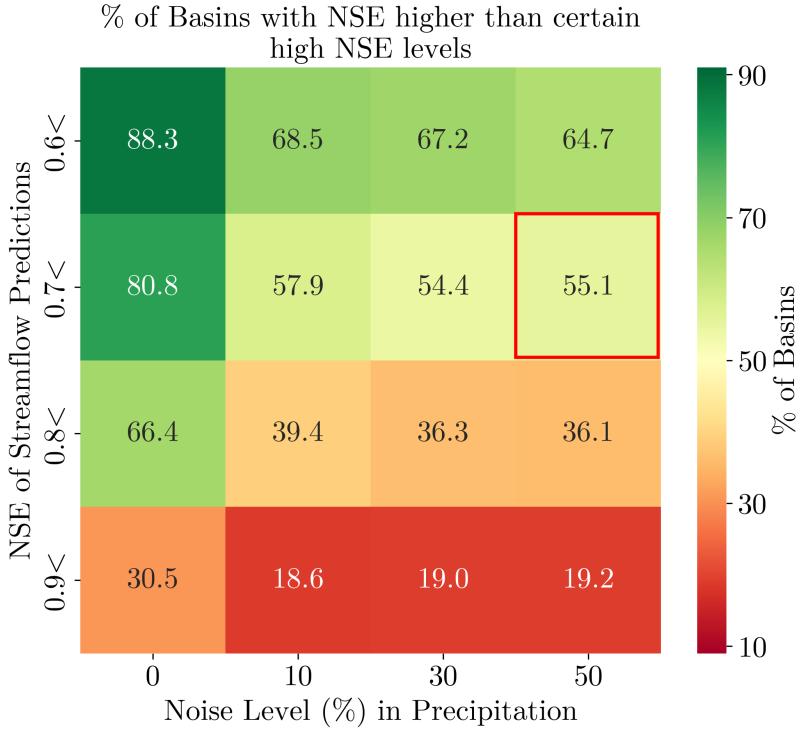

Figure 3.5: Percentage of total basins that exhibit higher NSE for different noise levels in precipitation.

3.3.2 Mean Absolute Percentage Error (MAPE) in Peak Flows

Figure 3.6 presents the MAPE of LSTM peak flow predictions when the model was trained and tested using noisy precipitation inputs. A similar performance pattern to that of the NSE can be observed. The peak flow MAPE does not significantly vary as the noise level in precipitation increases. However, most basins exhibit a MAPE exceeding 10%, with some reaching up to 100%, except for a few basins in the mid-western half of the United States, which show very low peak flow MAPE.

The right column of Figure 3.6 compares the peak flow MAPE between LSTM models with and without noisy precipitation inputs. For most basins, the peak flow MAPE falls within the 1-to-1 (red) and 1-to-2.5 (green) lines, indicating that the MAPE can increase by 0 to 150% when noisy precipitation data is introduced into the model. Interestingly, there are a few basins where the addition of noise to the precipitation input data results in a decrease in peak flow MAPE.

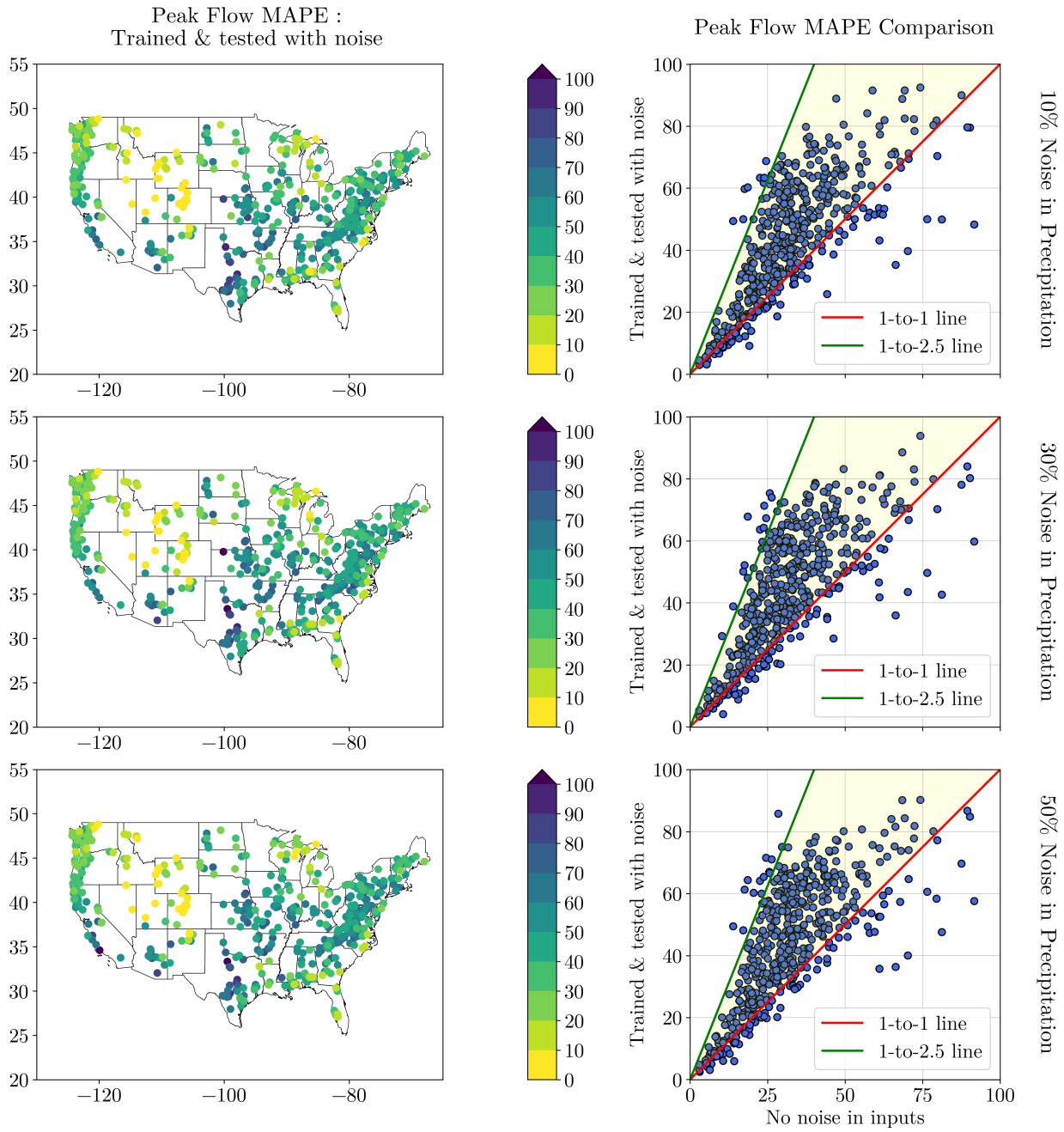

Figure 3.6: (Left column) MAPE for peak flow predictions from LSTM trained & tested with noisy precipitation, (Right column) Peak flow MAPE comparison: "no noise in inputs" (x axis) vs "trained & tested with noise" (y axis), (Rows) represent different noise levels.

3.3.3 Mean Absolute Percentage Error (MAPE) in Total Flows

Figure 3.7 illustrates the MAPE of LSTM streamflow predictions. Compared to the peak flow MAPE, a greater number of dark blue scatters are observed, indicating that more basins exhibit MAPE values exceeding 100% when considering the entire streamflow time series. In contrast, certain basins in the eastern United States, represented by light green scatters, demonstrate lower MAPE values than their peak flow MAPE counterparts.

The right column of Figure 3.7 compares the peak flow MAPE between LSTM models with and without noisy precipitation inputs. A notable observation is that most of the basins fall within the 1-to-1 and 1-to-2 lines, suggesting that the increase in MAPE due to the introduction of noise in precipitation is generally less pronounced than the increase in peak flow MAPE. Additionally, the scatters are more densely concentrated within the 0 to 40% MAPE range on the (y) axis, indicating that the increment in MAPE is relatively low compared to the increment in peak flow MAPE. Therefore, the performance of the LSTM model can be interpreted slightly differently depending on the specific metric being evaluated.

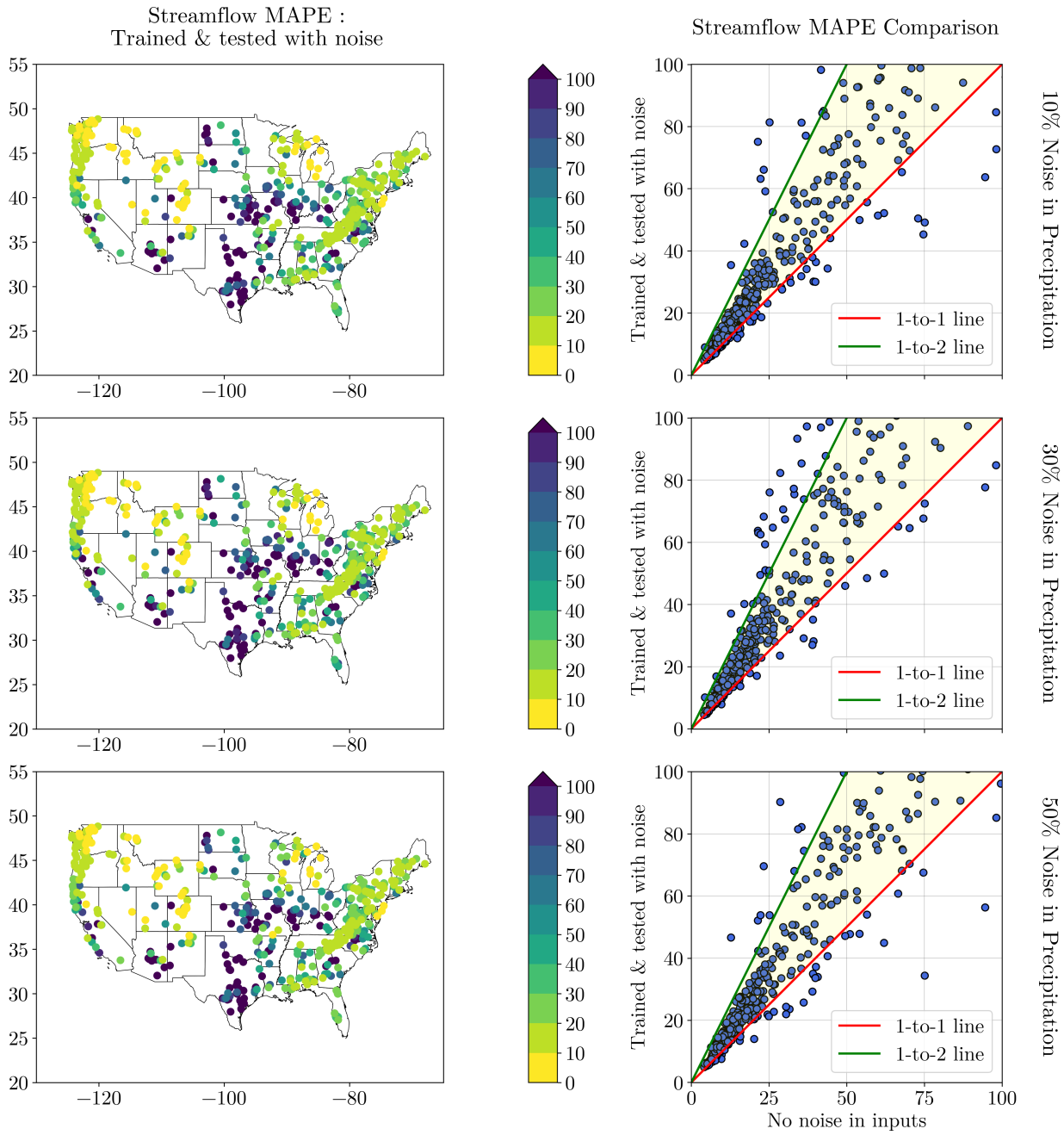

Figure 3.7: (Left column) MAPE for streamflow predictions from LSTM trained & tested with noisy precipitation, (Right column) Streamflow MAPE comparison: "no noise in inputs" (x axis) vs "trained & tested with noise" (y axis), (Rows) represent different noise levels.

3.3.4 Sensitivity of LSTM to the Noise in Precipitation in Different Basins

Figure 3.8 illustrates how the MAPE of LSTM streamflow predictions across different basins varies with the tested noise range in precipitation input data. Each best-fit curve ($y = bx^c$ excluding a (*interception*) for clearer comparison) corresponds to a specific basin. As depicted in the Figure 3.8, the sensitivity of LSTM predictions to noise in precipitation inputs varies differently as the noise level increases. Most basins exhibit high resistance to noise in precipitation, maintaining their MAPE increment at a very low level. As previously observed, these basins limit their MAPE increment to around 40%. However, a significant number of basins show a substantial increase in MAPE due to higher b and c parameters in their best-fit curves.

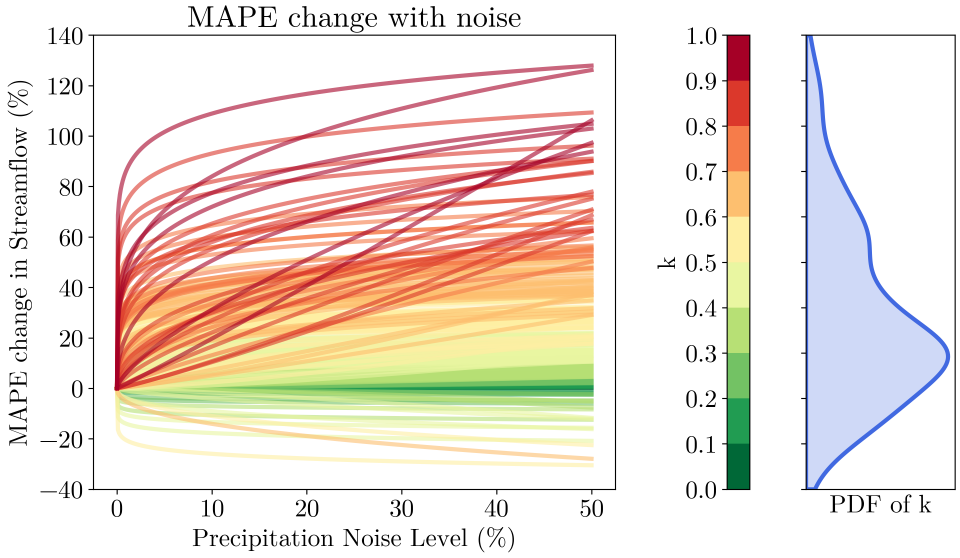

Figure 3.8: (Left) MAPE change in streamflow with noise level in precipitation & (Right) PDF of model sensitivity (k).

Figure 3.9 presents the Probability Density Function (PDF) of the R^2 values of the best-fit curves. While most basins have an R^2 value of around 0.94, approximately 30% of the basins have an R^2 value less than 0.7, indicating that the sensitivity of these basins to noise in precipitation input data is not well captured by the best-fit curves in Figure 3.8.

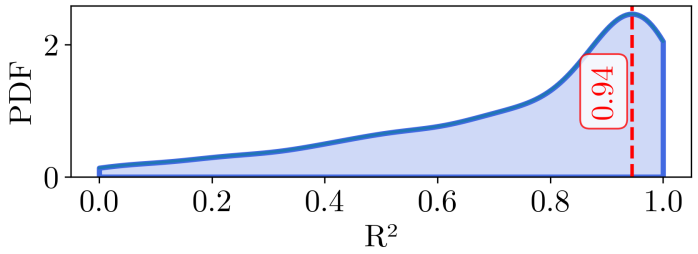

Figure 3.9: PDF for R^2 values of best fit curves in Figure 3.8.

Additionally, Figure 3.8 includes the PDF of parameter k, which is a function of both the accumulated MAPE increase (area under the curve) and the slope of the curve at 50%. Optimal values for α & β in Eq. 3.3 were determined to be 0.2 and 0.1, respectively, by assuming that $y = 0.5x$ curve (or line) represents a moderately sensitive ($k = 0.5$) basin.

Figure 3.10 demonstrates that LSTM performance in most basins exhibits limited sensitivity to the tested noise levels in precipitation input data. Specifically, 73% of the basins show less than moderate sensitivity, indicating that the model performance in those basins remains resilient to varying noise levels of precipitation input data. Figure 3.11 displays the sensitivity k for each basin, providing further insight into the spatial variability of LSTM sensitivity across different basins.

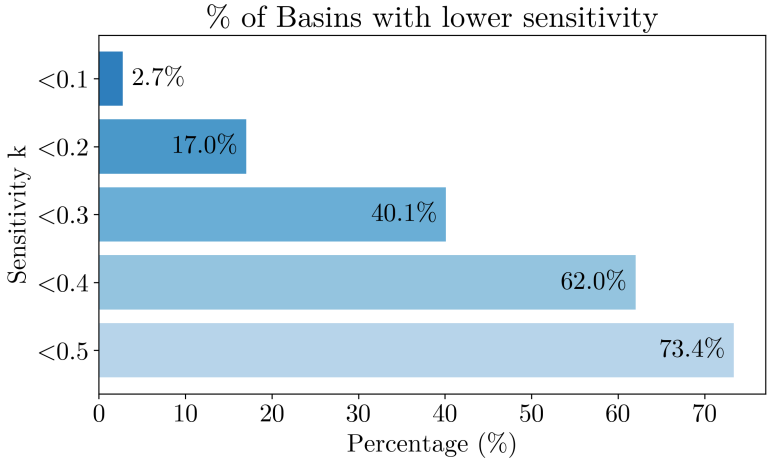

Figure 3.10: Percentage of basins with k lower than certain low k values.

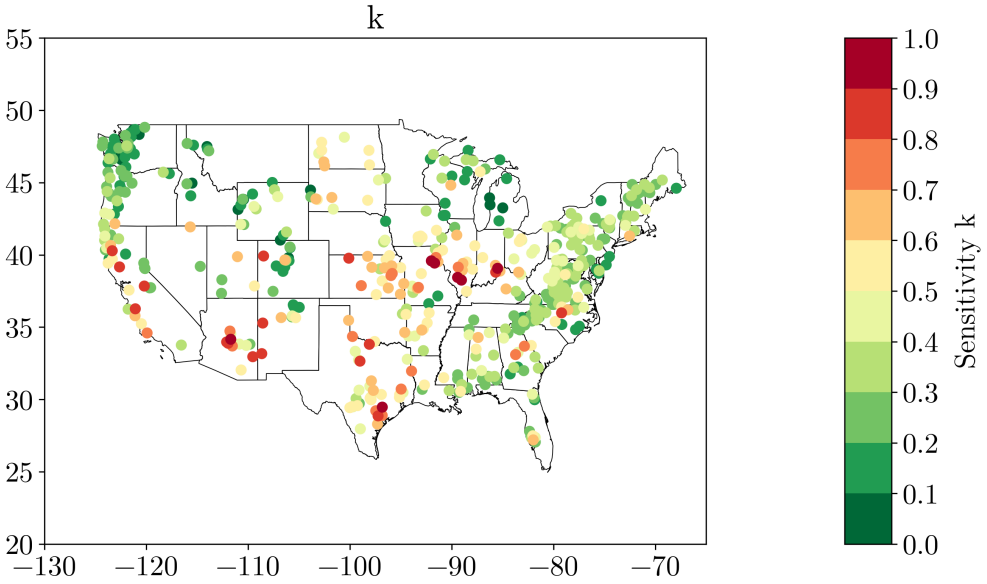

Figure 3.11: Spatial variability of LSTM's performance sensitivity (k) to noise in precipitation.

The observation that LSTM models exhibit varying degrees of sensitivity to noisy precipitation data across different basins underscores the importance of considering regional and catchment-specific characteristics in hydrological modelling. Figure 3.12 presents how the sensitivity of LSTM to the noise in precipitation is correlated with catchment attributes within a specific subset of basins.

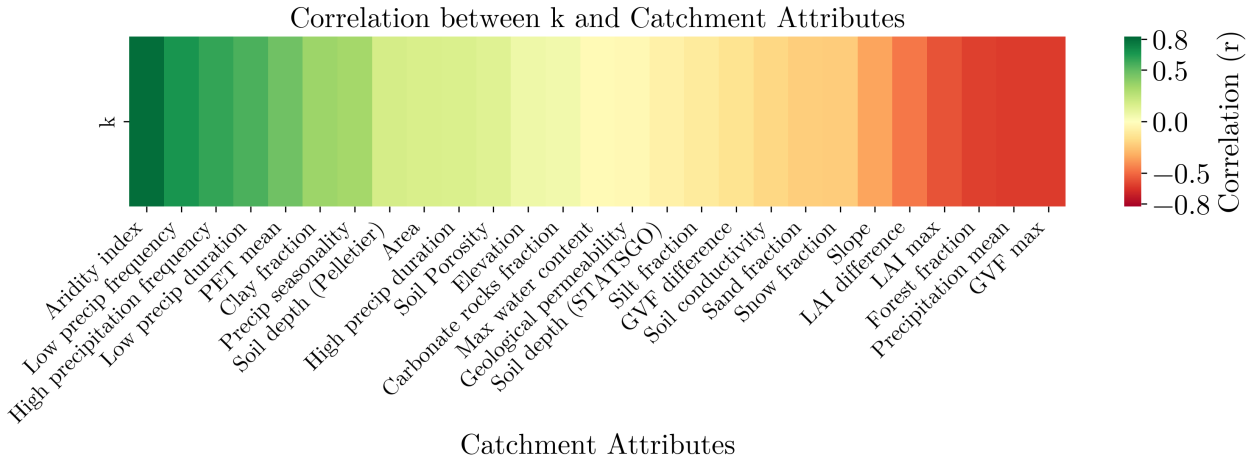

Figure 3.12: Correlation between LSTM sensitivity (k) and catchment attributes.

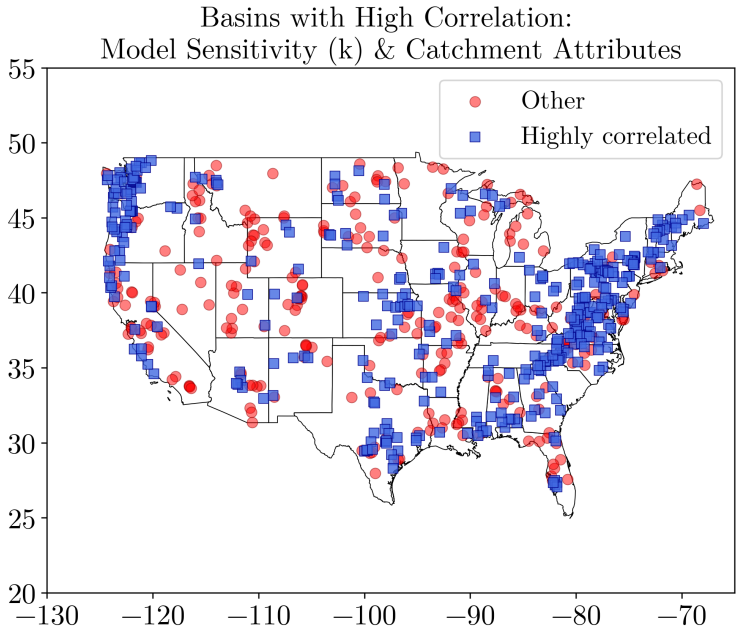

Figure 3.13: Subset of basins that exhibit a higher correlation between models' sensitivity (k) and catchment attributes.

Figure 3.13 highlights this subset of 384 basins out of 516 basins (more than 74% of the total) where LSTM's sensitivity strongly correlates with certain catchment attributes. Within the selected subset of basins, the LSTM's sensitivity is higher in basins with higher aridity, more frequent high and low precipitation events, lower precipitation mean, and less vegetation and forest cover. Conversely, basins with humid conditions, less frequent high and low precipitation events, higher precipitation mean, and greater vegetation and forest cover tend to be less sensitive to the noise in precipitation input data. This makes sense because basins with greater vegetation and forest cover experience higher canopy interception, which likely mitigates the impact of noise in precipitation input data on runoff response. On the other hand, basins with humid conditions, less frequent extreme precipitation events, and higher average precipitation tend to experience more consistent medium precipitation events,

leading to a stable long-term rainfall-runoff response. This stability likely allows the LSTM to effectively filter out short-term noise in precipitation data, as the model can rely on the overall consistency of the hydrological system.

Furthermore, the structure and hyperparameters of the LSTM model itself can play a role in determining its sensitivity to noisy data. Overall, results have shown that while LSTMs are generally robust, their performance can be affected by noise in precipitation data, particularly when the models are not trained with such data. The sensitivity varies by region and catchment characteristics, highlighting the need to tailor the model configuration carefully.

3.4 Conclusion

The findings of this study contribute important insights into the robustness of LSTM networks in handling noisy precipitation inputs for streamflow prediction. While traditional hydrological models are often highly sensitive to errors in precipitation data (Hong et al., 2006), LSTMs demonstrate a notable ability to maintain reliable performance even under significant noise levels. However, the degree of sensitivity varies across basins, emphasizing the role of regional characteristics and the importance of tailored model training.

A key finding is that training LSTMs with noisy precipitation data enhances their ability to predict streamflow during testing. When the model was exposed to noise during both training and testing, performance declines in metrics like NSE were less severe compared to scenarios where noise was introduced only during testing. This suggests that LSTMs can adapt to input uncertainties when trained under realistic, imperfect conditions. This is a critical advantage for regions with error-prone precipitation measurements.

The study also reveals slight differences in performance metrics. While MAPE change for total streamflow ranges mostly up to 100%, MAPE for peak flows showed greater variability than 100%. This difference implies that LSTMs prioritize overall hydrological trends over precise peak flow predictions when faced with noisy inputs. Such behaviour could be advantageous for applications like water resource planning, where long-term trends matter more than instantaneous peaks. However, it also signals a limitation for flood forecasting, where accurate peak predictions are critical. These outcomes match with Kratzert et al. (2021), who emphasized that model performance depends heavily on the chosen evaluation metric.

Additionally, the finding that over 55% of basins maintained NSE values above 0.7 even at

50% noise suggests that LSTMs could outperform traditional models in basins with imperfect data. This is particularly relevant for ungauged basins, where data quality is often a limiting factor (Arsenault et al., 2023). Another notable insight is the variability in sensitivity across basins, as quantified by the parameter k. While most basins exhibited low sensitivity, a subset experienced a significant performance decline with increasing noise. This variability underscores the influence of catchment-specific factors on model performance. The sensitivity parameter k appears to be strongly influenced by catchment characteristics, such as the frequency of high and low precipitation events, mean precipitation, aridity, and vegetation or forest cover. These factors may serve as physical drivers of LSTM resilience to noisy precipitation inputs. Additionally, the resilience observed in most basins suggests that LSTMs can effectively filter out noise when catchment dynamics align well with the model's learned patterns. To further enhance robustness, these insights could be integrated into tailored training strategies or hybrid approaches that combine LSTMs with physical models.

Another point to highlight is why certain basins perform better with noisy inputs. It can be concluded that in these basins, LSTMs generalize better from noisy data when exposed to such data during training. This agrees with the findings of Arsenault et al. (2023) that noise regularization during training leads to smoother density estimates and improved generalization. This indicates that LSTMs can learn underlying patterns even from imperfect data, provided they are trained with similar imperfections. Additionally, ensemble methods, as proposed by Feng et al. (2021), could improve LSTMs by quantifying prediction uncertainty and improving reliability.

However, the study's limitations must be mentioned. The use of uniformly distributed noise may not fully replicate errors in real-world precipitation estimates, which often exhibit spatial correlation or systematic biases (Lafon et al., 2013). Furthermore, noise was applied

only to the last three days of precipitation data, whereas longer-term errors could have compounding effects on model predictions. Future work could explore diverse noise structures, such as temporally correlated errors or region-specific bias patterns, to better mimic real-world input data uncertainties.

In conclusion, this study reinforces the potential of LSTMs as a competent tool for streamflow prediction under noisy data conditions. Their ability to generalize from imperfect inputs, along with their adaptability to regional variations, presents them as a viable alternative to traditional hydrological models. However, successful deployment requires careful consideration of tailored model configurations, noise characteristics, and the selection of performance evaluation metrics. By addressing these challenges in future research, LSTMs could be an important approach to improving hydrological forecasts and global water resource management.

Chapter 4

Interpolation & Extrapolation in Streamflow Forecasting in Stationary & Non-Stationary Scenarios using LSTMs

4.1 Introduction & Background

Recent studies have shown that machine learning (ML) models, such as LSTMs, GRUs, and Transformers, perform more reliably in streamflow prediction tests (both hindcasts and forecasts) compared to complex high-resolution hydrological models (Demiray et al., 2024). This challenges the long-held assumption that a deeper understanding of hydrological processes would automatically lead to better streamflow predictions, including floods and droughts. However, while ML models excel in many cases, their reliability and performance in

situations beyond their training data, such as extreme or unseen conditions, remain uncertain (Beven, 2024).

In a recent study, Gurbuz et al. (2024) proposed and tested a methodology to benchmark data-driven algorithms for streamflow prediction creating artificially generated streamflow data using physically based hydrological models under very controlled conditions. Their approach considered the implementation of the Hillslope-Link distributed hydrological Model (HLM) for a 4,385 km^2 basin forced by storms generated using the stochastic storm transposition (SST) framework. Gurbuz Gurbuz et al. (2024) demonstrated that ML algorithms can effectively identify the input-output relations between the average rainfall over a basin and the streamflow (as time series) at multiple sub-basin outlets under very general conditions of space-time variability of flood-generating storm systems.

In this study, the work by Gurbuz et al. (2024) is extended to ask a new question: How reliable are trained ML algorithms at predicting streamflow fluctuations that have never been observed in the "historical" record? This question goes to the heart of what these black/grey-box tools represent mathematically: a deterministic estimate for the input-output relationship between meteorological forcings and streamflow. Consequently, when any of these grey/black-box models predict a hydrograph, there are two possible cases for the prediction, 1) interpolation, which means that the hydrograph and peak flow being predicted are within the range of flows and meteorological conditions observed in the past, and 2) extrapolation, the case where the event being predicted is significantly larger or smaller than anything observed in the past. This investigation is crucial in the context of climate change, where increasing atmospheric water-holding capacity leads to unprecedented record-breaking storms in intensity, duration, and spatial coverage, and where the anticipated response of watersheds remains unknown. ML is already being used to fill this knowledge gap under the implicit

premise that "if it served us well in the past, it shall serve us well in the future."

Several studies have directly tackled the ability of ML models to extrapolate to extreme hydrological events. Frame et al. (2021) conducted a study specifically to test the hypothesis that data-driven models become less reliable in extreme events compared to process-based models. They trained LSTM and mass-conserving LSTM (MC-LSTM) models, along with conceptual and process-based models, and evaluated their performance on high-return-period streamflow events. Their findings contradicted the initial hypothesis, indicating that the data-driven models, including both pure ML and physics-informed ML models, performed better at predicting peak flows across almost all conditions, even for extreme events that were not included in the training data. This suggests a strong generalization capability of modern deep learning models for the prediction of extreme events. The study partitioned data according to frequency of occurrence, using the discharge of the 5-year return period as a threshold for training and testing.

Acuña Espinoza et al. (2025) also investigated the generalization capabilities of data-driven, hybrid, and conceptual models for predicting extreme hydrological events. They followed a methodology similar to Frame et al. (2021). Their experiments involved evaluating model performance for peak flows. While their study did not offer a definitive conclusion on one architecture being significantly better than another in all scenarios, they systematically tested the models' ability to handle increasingly extreme events.

Song et al. (2025) focused on generalizing hybrid models combining deep learning with process-based equations (differentiable models) to extreme floods outside the training data. They conducted temporal extrapolation tests by training models on water years with lower return period flows and holding out years with higher return periods. Their results indicated that these hybrid models could generalize reasonably well to unseen extreme events and that

incorporating interpretable structural priors could further improve this generalizability. This highlights the potential of integrating physical knowledge into deep learning architectures to enhance their reliability in extrapolating to unseen conditions.

These studies suggest that modern deep learning models, and increasingly hybrid approaches, demonstrate a promising ability to predict streamflow fluctuations outside of their training range, including extreme events. For further understanding the capabilities and limitations of ML algorithms in predicting truly unseen hydrological events, the question of interpolation vs. extrapolation is addressed by creating three scenarios of data generation where 1845 generated storms were applied to the HLM model to create an artificial record of streamflow data for 101 consecutive years. The first case is the Stationary scenario, where the generated storms were applied randomly in time; the second case is a Non-Stationary scenario with increasing precipitation (NS-Increasing), where the smallest generated storms were applied first and the largest were applied last; and the third case is a Non-Stationary scenario with decreasing precipitation (NS-Decreasing) where the storms were applied from largest to smallest.

In this Chapter, Section 4.2 describes the study area, watershed characteristics, the setup of the SST framework, the high-resolution HLM model, and the ML model used to predict the artificially generated data. Moreover, it describe the experimental design to create the Stationary and Non-Stationary scenarios and details about the resampling of data. Section 4.3 presents the numerical experiment results and the discussion of results. Finally, Section 4.4 presents the conclusions of this study.

4.2 Methodology

4.2.1 Study area

The Turkey River Basin (Figure 4.1) is a significant hydrological region spanning 4,385 km^2. It hosts multiple streamflow gauging stations, including the most downstream station at Garber and four sub-catchments in Littleport, Elkader, Eldorado, and Spillville. These stations provide crucial hydrological modelling, flood forecasting, and water resource management data. Several studies have implemented hydrological models of different degrees of complexity for this region (Politano et al., 2023) that provide a high level of accuracy in hindcast streamflow evaluations. In particular, the performance of the HLM model has been tested in several studies. The HLM model decomposes the basin into 237,000 individual hillslope-scale control volumes ($\sim$0.02 km^2), providing a high-resolution characterization of the network of natural and engineered features that regulate water flow, soil moisture dynamics, and erosion patterns. This watershed was used by Perez et al. (2019) to study various processes affecting peak flow response; it has been used to assess the impact of storm spatial and temporal details (Zhu et al., 2018); and to determine climate-driven changes in snowmelt and soil moisture seasonality (Yu et al., 2019) on peak flow patterns. The basin is typically further subdivided into sub-watersheds determined by the location of the USGS streamflow gauges in the basin. This partitioning was preserved in this study because it represents a typical scenario where grey/black box models are implemented and tested. The description of each sub-watershed is provided in Table 4.1.

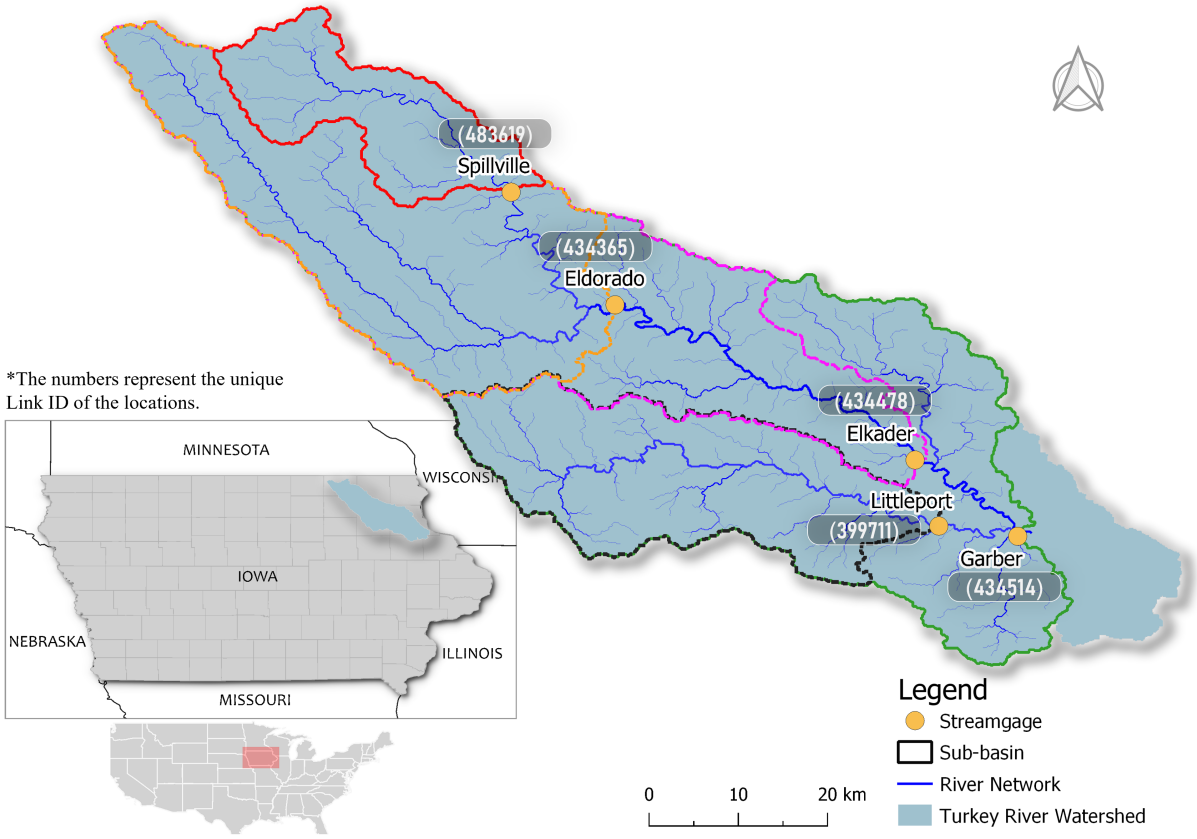

Figure 4.1: Map of the Turkey River Watershed in Iowa and its sub-catchments (Gurbuz et al., 2024).

Table 4.1: Description of Turkey River Watershed along with its sub-catchment.

Catchment	Description	Link ID	Area (km²)
Spillville	Turkey River at Spillville	483619	458.8
Littleport	Volga River at Littleport	399711	909.2
Eldorado	Turkey River near Eldorado	434365	1667.3
Elkader	Turkey River above French Hollow Creek at Elkader	434478	2359.5
Garber	Turkey River at Garber	434514	4031.8

4.2.2 Rainfall Data

The SST framework, implemented through the RainyDay software (Wright et al., 2017), was used to generate a catalog of synthetic, yet realistic, storm events for the Turkey River Basin. The SST method works by resampling and spatially transposing observed rainfall events from a surrounding domain (A'). In brief, the generation of synthetic-realistic storms through RainyDay consists of five key steps: (1) defining the transposition domain A', (2) creating a storm catalogue from the largest observed storms in the domain, (3) randomly determining the number of storms per year using a Poisson distribution (with a rate of 20 storms per year), (4) randomly selecting storms from the catalogue, and (5) spatially transposing the selected storms in both east-west and north-south directions to generate new rainfall fields. Steps 4 and 5 can be repeated T_{max} times to then generate T_{max} number of synthetically realistic storm events. A transposition region defined by latitudes from 40.2°N to 45°N and longitudes from 90.2°W to 96.7°W along with the April–November Stage IV rainfall data from 2002 to 2018 (Du, 2011) to select the largest 300, non-overlapping storm events based on rainfall accumulation over durations of up to 72 hours were used for Turkey River basin. Then, (T_{max} =) 10,000 storm events were generated. By transposing storms across a broad region, this method leverages the space-for-time trade-off, allowing rare and extreme rainfall events that may not have been observed locally to be represented. This approach enables a more comprehensive analysis of extreme storm impacts and provides a robust foundation for flood frequency analysis and risk assessment. For more information on the SST methodology, please refer to Wright et al. (2017) and Gurbuz et al. (2024).

4.2.3 Hydrological Digital-Twin

The Hillslope-Link Model (HLM) was used as the best representation of the hydrological system for the Turkey River Basin, serving as a digital twin to conduct simulations under both Stationary and Non-Stationary scenarios. This model decomposes the landscape into hillslope-channel-link components, consisting of 237,000 hillslope-link units for the Turkey River Basin, with an average area of 0.018 km^2, derived from a 1-meter resolution LiDAR digital elevation model. This hydrological model incorporates non-linear channel routing, evaporation, infiltration, surface ponding, and effective water depths in both the upper soil and subsurface layers (Quintero et al., 2020). Changes in storage components are modeled using a system of nonlinear ordinary differential equations, solved through a parallel Runge-Kutta method with asynchronous integration (Small et al., 2013). The HLM has been extensively tested in various studies and has consistently delivered accurate depictions of streamflow dynamics and flood patterns (Krajewski et al., 2017; Mantilla, 2007; Mantilla et al., 2006; Quintero et al., 2020;Gurbuz et al., 2024).

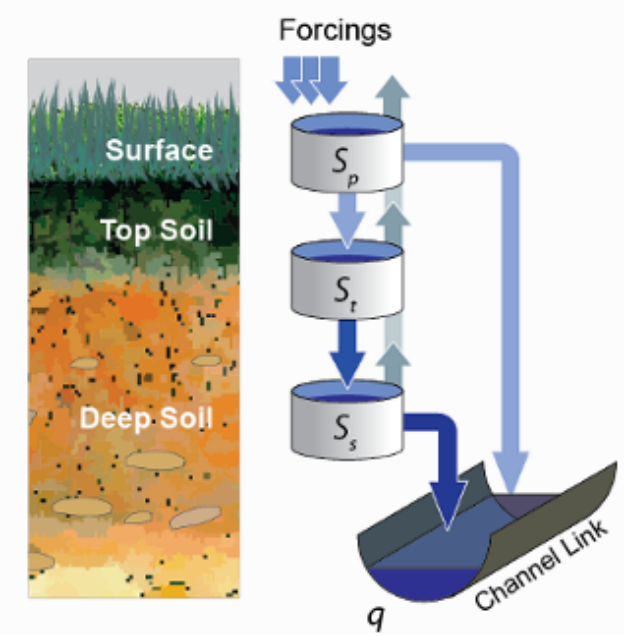

Figure 4.2: Top Layer Hydrological (254) Model in HLM (Asynch, 2023).

HLM: Top layer Hydrological model (Figure 4.2) which was used in this study produces continuous records of streamflow for all links in the river network, that are sampled every hour. Each storm in the SST catalogue lasts for 3 days and they are applied over the entire catchment domain, followed by a period of no rain of 17 days. Therefore, an event inside the digital twin is relatively simple, with a new storm occurring every 20 days. Only the streamflow hydrographs calculated at the 5 selected watersheds are used for this study. The hourly data was averaged to obtain daily records. Figure 4.3 shows the daily data over the full period at Garber station for all 3 scenarios (Stationary, NS-Increasing and NS-Decreasing worlds). The HLM implementation follows the same configuration as presented in Perez et al. (2019) and Gurbuz et al. (2024).

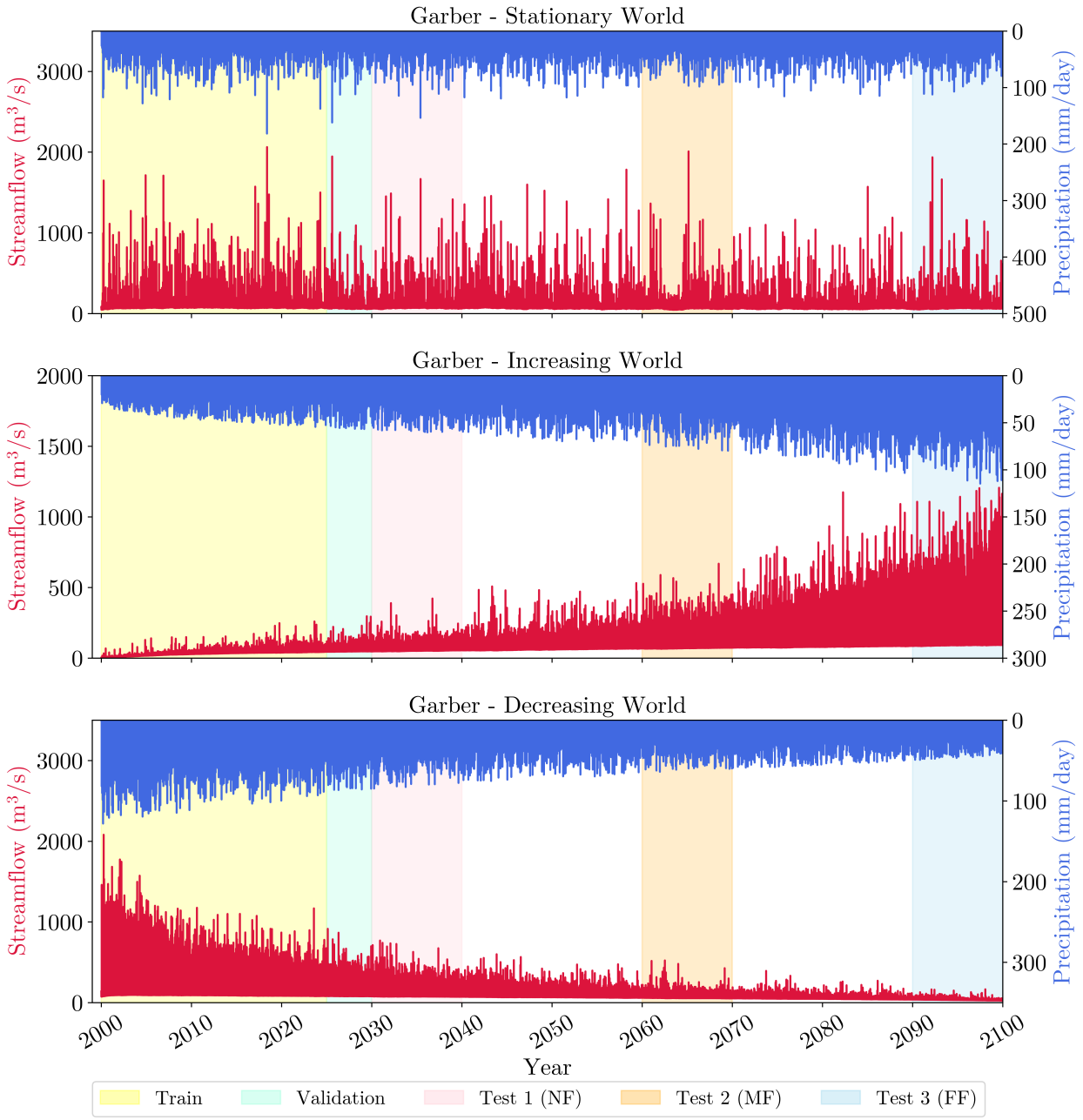

Figure 4.3: Precipitation and streamflow data with simulation data splits (top panel) for the Stationary world, (middle panel) for NS-Increasing world, (bottom panel) for NS-Decreasing world at Garber.

4.2.4 Scenarios and Test Periods

Three data generation scenarios are created where 1844 generated storms were applied to the HLM model to create a streamflow record for 101 consecutive years. The first case is the Stationary scenario, where the generated storms were applied randomly in time. The second case is a Non-Stationary scenario with increasing precipitation (NS-Increasing world), where the smallest generated storms were applied first and the largest were applied last. The third case is a Non-Stationary scenario with decreasing precipitation (NS-Decreasing world) where the storms were applied from largest to smallest. Four data periods are selected in the 101-year data record. The first 30 years of artificial data, which represent historical records, are selected for training and validation, then three 10-year periods are selected to represent the Near Future (NF; 2031-2040), the Mid Future (MF; 2061-2070), and the Far Future (FF; 2091-2100). The timeline for the experimental design is shown in Table 4.2.

Table 4.2: Data split for different simulation periods.

World	2001-2025	2026-2030	2031-2040	2061-2070	2091-2100
Stationary NS-Increasing NS-Decreasing	Train	Validation	Test 1: Near Future (NF)	Test 2: Mid Future (MF)	Test 3:Far Future (FF)

4.2.5 Machine Learning Model Set-Up

In this study, a single LSTM was trained, evaluated, and tested separately for each basin. Streamflow predictions were generated for only one daily step at a time. Rainfall was the only dynamic input while no static attribute was considered. During model evaluation, any negative predictions in the original value space were set to zero, meaning no negative discharges were

allowed. NSE was used as the loss. As the lengthy input sequences significantly increases the time required for training, a relatively small parameter grid was taken in to the consideration for hyperparameter tuning for the model. All possible hyperparameter combinations (243 different combinations) within the selected grid were considered and the model was employed for the outlet (Garber) station of the basin in all 3 scenarios and calculated the mean NSE for future test periods (NF, MF & FF). The combination with the highest mean NSE was selected for the outlet in all 3 scenarios. The tested hyperparameter grid and selected combination are given in Table 4.3. Using tuned hyperparameters, LSTM models were trained and tested in all 5 stations and NSE metric was calculated for the streamflow predictions in different worlds and different future periods.

Table 4.3: Tested hyperparameter grid and selected combination for LSTM model.

Hyperparameter	**Tested Grid**	**Selected**
Hidden size	128, 256, 512	512
Output Dropout	0.2, 0.4, 0.6	0.6
Batch size	128, 256, 512	128
Sequence Length	20, 40, 60	20
Epochs	30, 40, 50	50

4.3 Results & Discussion

4.3.1 Prediction of Peak Flows

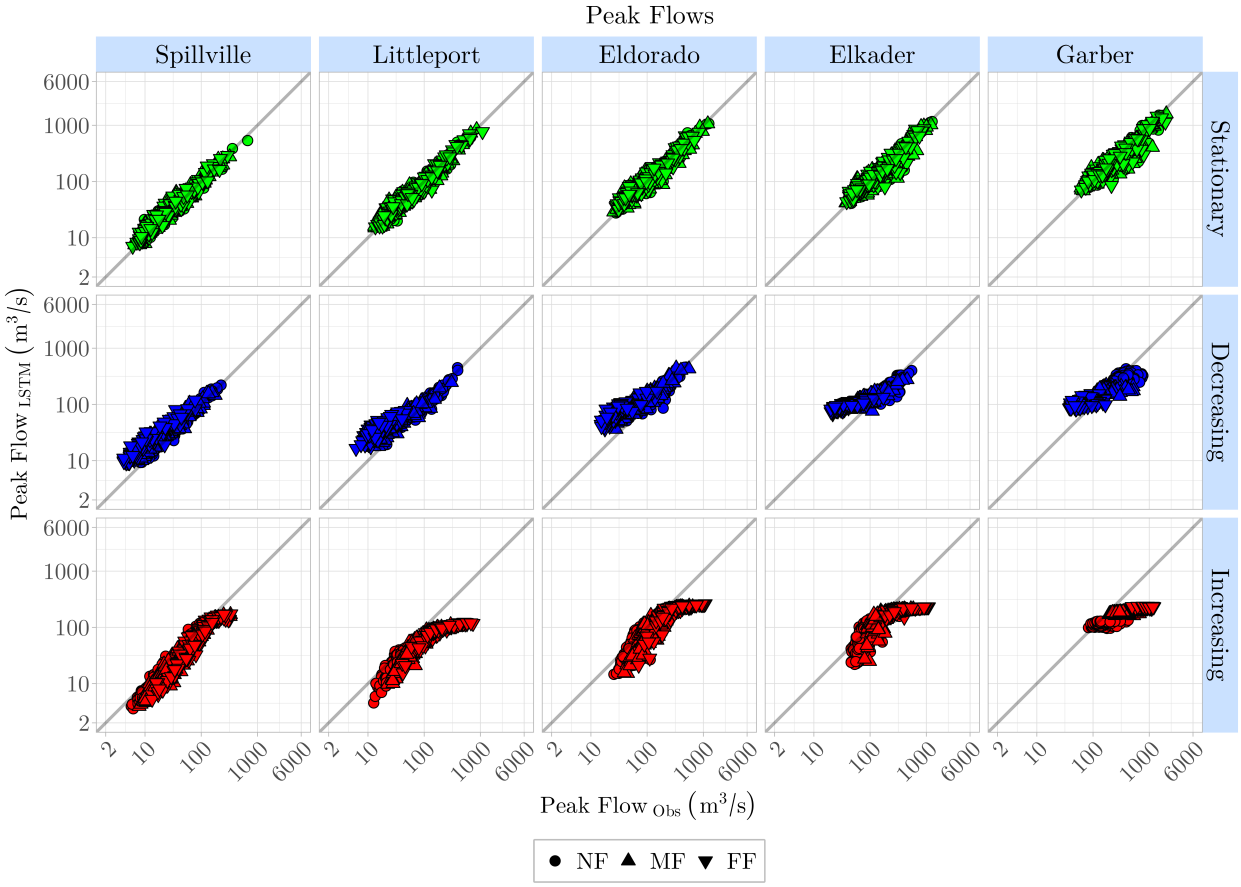

Figure 4.4: Observed and LSTM peak flows during NF, MF & FF periods for each catchment (Color & shape of the scatters represents the world & the future test periods respectively, Catchments are ordered by size in ascending order).

Figure 4.4 illustrates the ability of the LSTM model to predict peak flows for each of the five catchments: Spillville, Littleport, Eldorado, Elkader, and Garber. Figure 4.4 presents scatter plots comparing observed peak flows with the peak flows predicted by the LSTM model under Stationary, NS-Decreasing, and NS-Increasing worlds. Each column represents a different catchment, and each row represents a different world. Different markers are used to represent the three future test periods (NF, MF and FF). A gray diagonal line is present in each plot, representing a perfect match between observed and predicted values. The closer the scatter points are to this line, the better the model's prediction for that specific catchment, world, and future period.

Under the Stationary world, the close alignment between observed and simulated peak flows across all five catchments suggests that the model successfully predicted peak flows across various future test periods for each catchment. For Elkader and Garber, two catchments with the largest areas, the spread of the scatter points around the 1 to 1 line is slightly larger compared to Spillville, Littleport, and Eldorado, meaning the model performance in predicting peak flows is narrowly less accurate than that for other three catchments. For the smallest catchment, Spillville in Non-Stationary worlds, the scatter points still seem to follow the general trend of the diagonal line, indicating the best model performance compared to other catchments in these Non-Stationary worlds. For other catchments, the LSTM model showed similar performance in Non-Stationary worlds. Still, the peak flow predictions deviated from observations more significantly at higher peak flows in the NS-Increasing world and at lower peak flows in the NS-Decreasing world. Overall, the model tended to underestimate the peak flows in the NS-Increasing world and overestimate the peak flows in the NS-Decreasing world.

4.3.2 Prediction of Streamflow Event Volumes

Figure 4.5 displays the performance of the LSTM model in predicting the volume of streamflow events for each of the five catchments in Stationary and Non-Stationary worlds and three future test periods (NF, MF, FF). The markers again correspond to the future test periods.

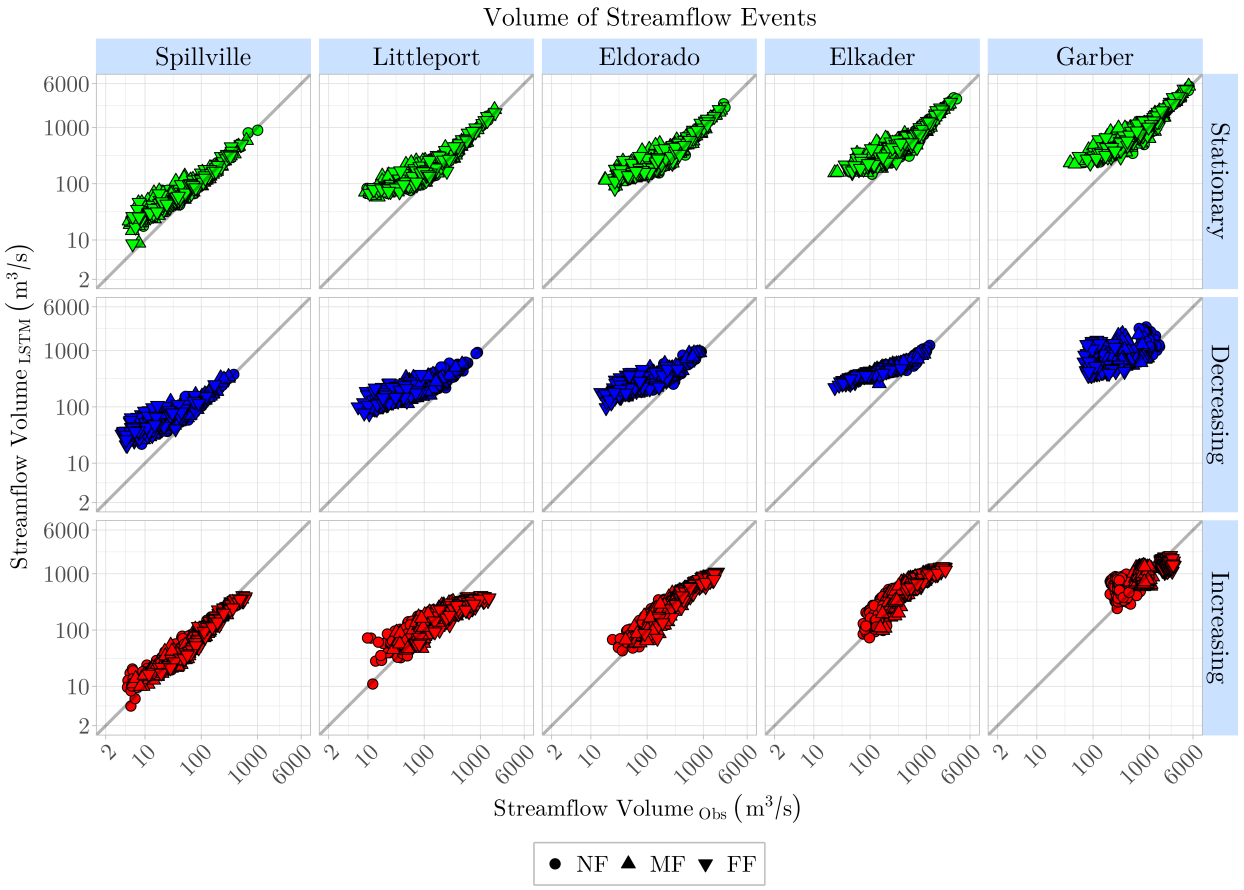

Figure 4.5: Observed & LSTM volumes of streamflow events during NF, MF & FF periods for each catchment (Color & shape of the scatters represents the world & the future test periods respectively, Catchments are ordered by size in ascending order).

In all three worlds, the observed and simulated streamflow volumes showed better alignment for Spillville compared to the other catchments. This suggests that the LSTM model was generally more accurate in predicting the total volume of streamflow events for the smallest catchment, regardless of the future climate scenario. It can be observed that in general, predicted volumes show similar variation in each world separately. This implies that for a given future climate trend and future test period, the model's performance in predicting streamflow event volumes was relatively consistent across the different catchments, although there might be differences in the overall accuracy as mentioned for Spillville.

In the Stationary world, the model predicted higher streamflow event volumes well, however, many points at lower observed volumes appear above the diagonal line, indicating the model struggled to predict lower streamflow volumes accurately. In the NS-Decreasing world, the predicted volumes are relatively overestimated, and at the same time, the deviation increases as the observed flow volume decreases. This is due to the fact that many events exceeded the model's training limits, and its predictive ability diminished after a certain threshold of rainfall events. This suggests that the model might lead to less accurate predictions when exposed to rainfall events outside of the trained data range.

Interestingly, under the NS-Increasing world, the scatter points for all catchments generally appear to be more tightly clustered around the diagonal line compared to the Stationary and NS-Decreasing worlds, showing the best predictability across different worlds. This implies that even though LSTM was unable to predict peak flow magnitudes accurately in NS-Increasing world, the LSTM streamflow hydrograph might align reasonably well with the observed streamflow hydrograph if the baseflow from both hydrographs is excluded. While the overall performance in the NS-Increasing world seems better, a closer inspection of the plots for Littleport and Garber might reveal a slightly larger spread of points compared to

Spillville, Eldorado, and Elkader.

4.3.3 Prediction of Rainfall-Runoff Relationship

Figure 4.6 depicts a scatter plot of observed and predicted rainfall-runoff relationships for each world. The colour and shape of the scatters represent the catchment and the future test period respectively. In all three worlds, there is a noticeable spread in the observed runoff values for a given rainfall amount, particularly at lower to medium rainfall intensities. This indicates that factors beyond just the amount of rainfall influence the resulting runoff, meaning the complex behaviour of the observed rainfall-runoff relationship.

Comparing the left and right columns of Figure 4.6, it appears that the predicted rainfall-runoff relationships show a smoother, more defined curve for all catchments compared to the more scattered points in the left column (observed). This suggests that the LSTM model captured a general trend but missed some of the complex variability present in the observed relationship. Examining the middle row of 4.6, it seems that for most catchments, the predicted runoff values tend to be higher than the observed runoff values for a given rainfall amount, indicating an overestimation of runoff by the model in the NS-Decreasing world. On the other hand, looking at the bottom row of 4.6, the predicted runoff values appear to be generally lower than the observed runoff values for similar rainfall amounts, suggesting an underestimation of runoff by the model in the NS-Increasing world. Overall, the model performed best at replicating the observed rainfall-runoff relationship under the Stationary world.

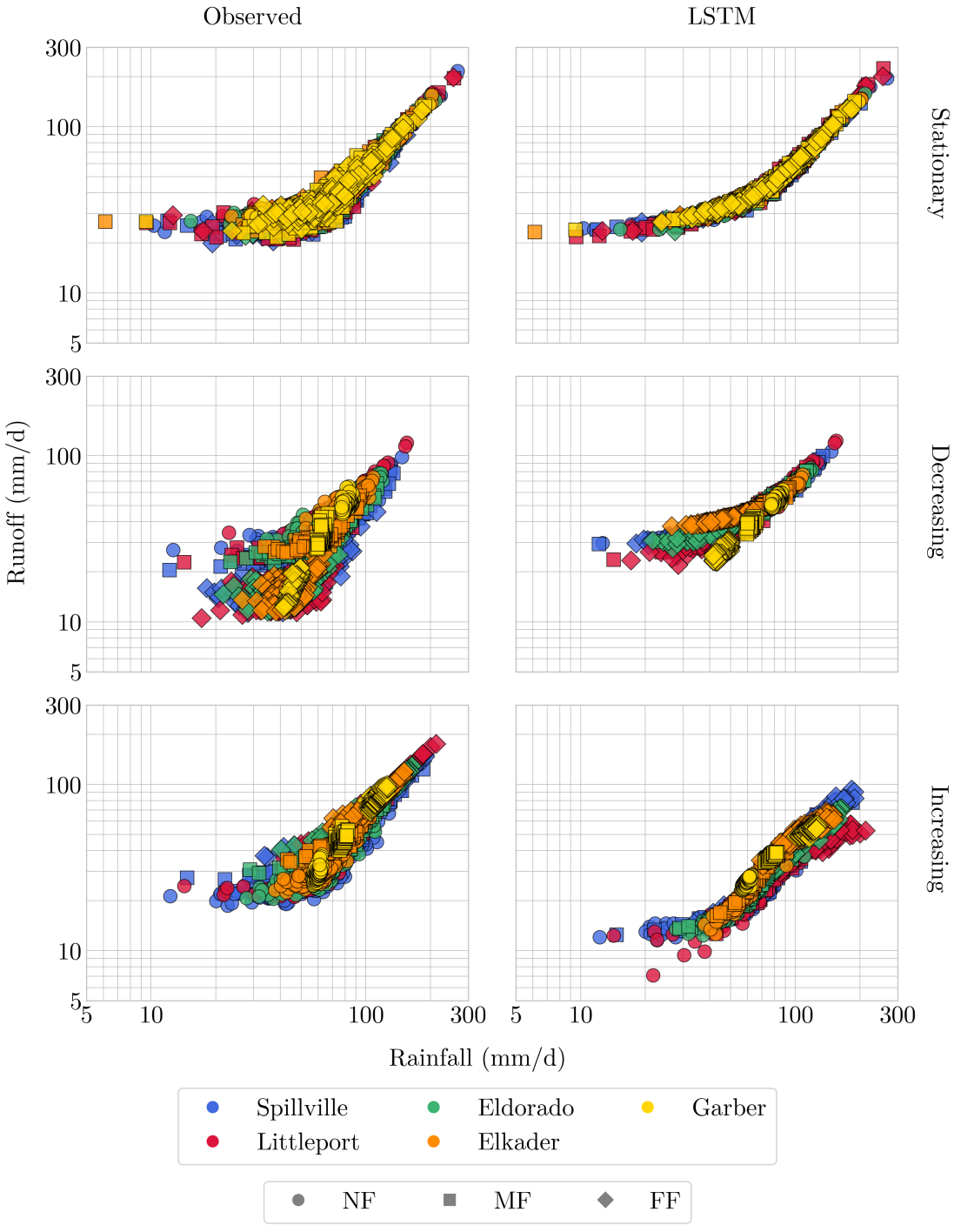

Figure 4.6: Observed and Predicted rainfall-runoff relationship (Color & shape of the scatters represents the Catchment & the future test periods respectively).

4.3.4 Overall Performance

Figure 4.7 presents the NSE values across different worlds, catchments, and model periods. The rows in the Figure 4.7 represent the five catchments ordered by ascending drainage area, and the columns are grouped by the world (Stationary, NS-Increasing, NS-Decreasing) and then further divided by the model periods (Train, Validation, NF, MF, FF).

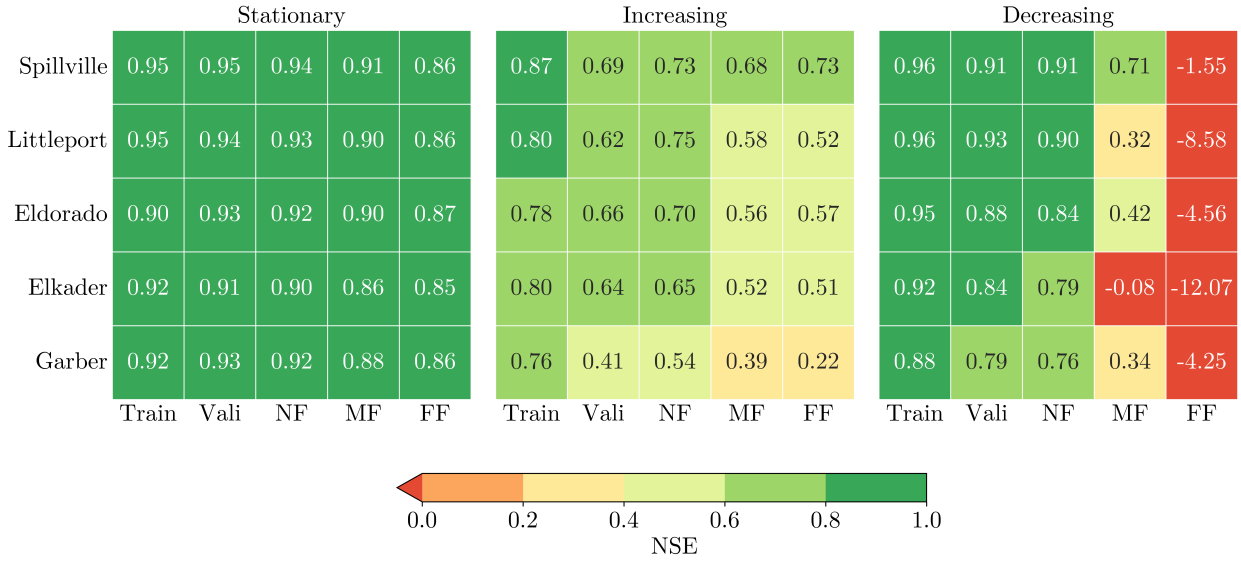

Figure 4.7: NSE across each catchment, world and simulation periods.

In the Stationary world and across all the catchments, the model achieved remarkably higher NSE values, showing the strong capability of LSTMs to predict streamflow within the data range it has seen before. Looking at the first group of columns in Figure 4.7, the NSE values for training and validation are generally high (mostly above 0.90) for all five catchments. During testing periods (NF, MF, FF) in the Stationary world, the NSE values are also relatively high, although there is some variability across catchments and future testing periods. During these testing periods, even the lowest NSE achieved by the model

was above 0.85, which is generally identified as a strong model performance based on NSE. This supports the idea that LSTM, with its deep learning approach, can better capture the complexities and non-linearities in hydrological processes under stationary conditions.

Examining the third group of columns in Figure 4.7, the higher NSE values for train and validation in the NS-Decreasing world show that the model performed significantly well during train, and validation, similar to the Stationary world. Notably, during NF, the model was able to achieve NSE higher than 0.75 across all catchments. However, for the MF and FF periods under the NS-Decreasing world, the NSE values tend to be lower, and even negative in FF, indicating a decline in model performance as the future prediction period moves further away from the training data. Except during MF in Spillville, across all catchments during MF and FF, the model performance was significantly poor.

There is an interesting observation under the NS-Increasing condition where the model performances during train, validation, and NF are not as good as in the NS-Decreasing world, but in MF and FF they are better than in the NS-Decreasing world. In the second group of columns in Figure 4.7, the NSE values for train and validation are generally lower compared to the Stationary and NS-Decreasing worlds. However, for the MF and FF periods in the NS-Increasing world, the NSE values show an improvement compared to the corresponding MF and FF values in the NS-Decreasing world. Except for the outlet station Garber, across all other catchments, the model was able to achieve NSE higher than 0.51 even during FF.

In general, the catchment area seems to play a crucial role in LSTM's performance accuracy. By looking at the trends in NSE values across the columns for the NF, MF, and FF periods within the NS-Increasing and NS-Decreasing worlds, it can be observed that the model performance decreases gradually as the catchment area increases during all test periods. Overall, the smallest catchment, Spillville, has the best model performance across

all testing periods and worlds. Conversely, the two largest catchments have the worst model performance across all testing periods and worlds.

Figure 4.8 further illustrates how the model performance changes in future testing periods across different worlds and catchments. The plots show the NSE values for each catchment under Stationary, NS-Increasing, and NS-Decreasing worlds. Markers connected by lines represent different future test periods. As noticed before, the predictive accuracy of the model decreases as the testing window moves away from the train and validation periods. In Figure 4.8 as for all catchments and worlds, the NSE values tend to be highest for NF, followed by MF, and lowest for FF. There is also a noticeable dramatic drop in NSE in the NS-Decreasing world during FF.

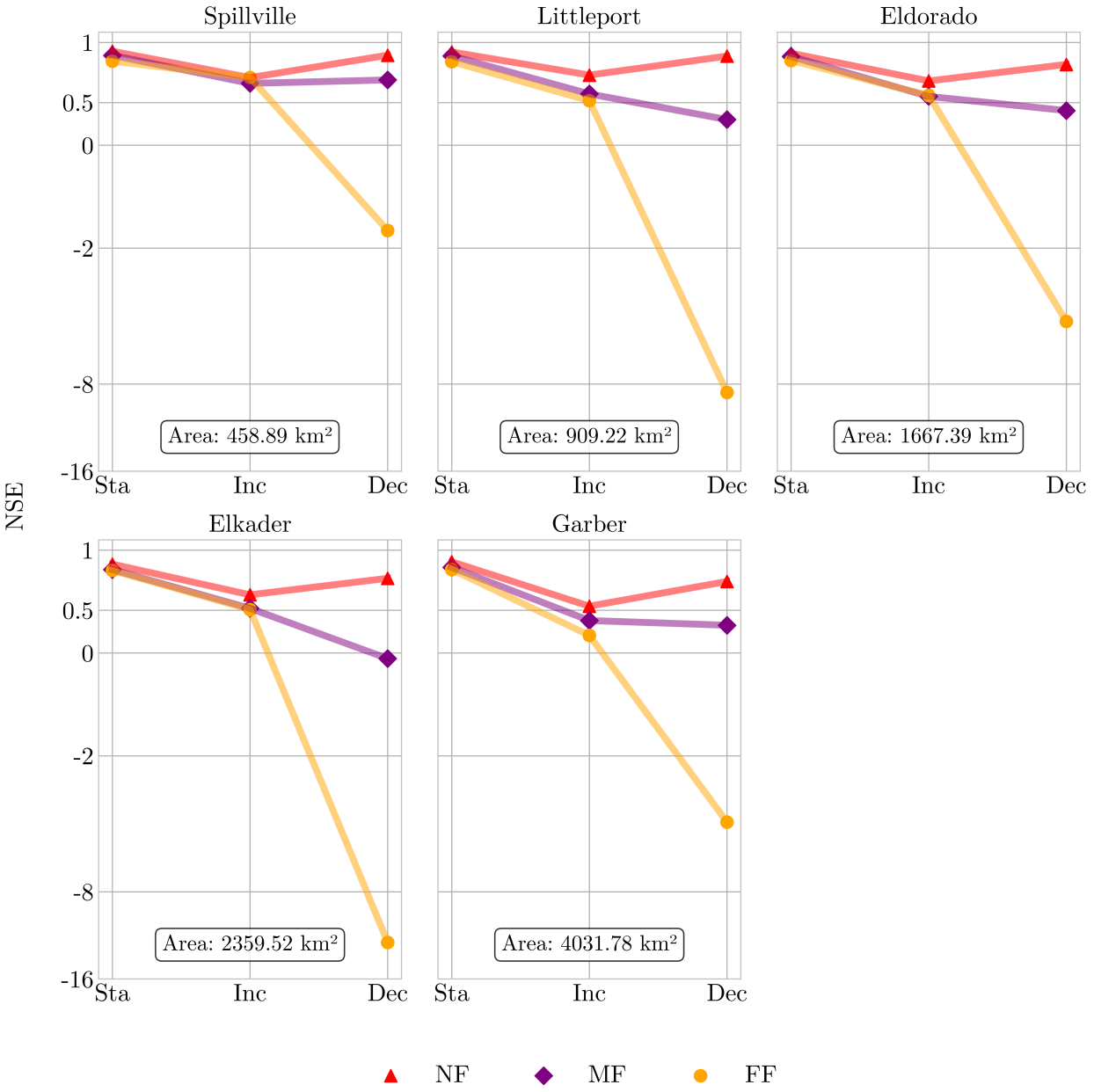

Figure 4.8: NSE variation across different catchments during future testing periods.

Figure 4.9 shows zoomed-in hydrographs for the Garber station, displaying observed precipitation, streamflow, and LSTM streamflow in one of the NF years (Year 2065) for the Stationary (top), NS-Increasing (middle), and NS-Decreasing (bottom) worlds. These hydrographs further confirm some observations made earlier in the context of peak flows, streamflow event volumes, and overall performance. Under the Stationary world, the LSTM hydrograph shows a near-perfect alignment with the observed hydrograph, demonstrating the model's strong streamflow predictability under stationary conditions. However, under Non-Stationary worlds, a common observation is that the model tended to make peak flows with similar magnitudes consistently once it is exposed to rainfall storms that is outside the range of training data. Moreover, the hydrographs show that LSTMs overestimate and underestimate streamflow under NS-Decreasing and NS-Increasing worlds, respectively, specifically during baseflow periods. This tendency aligns with the earlier findings regarding the model's behavior in Non-Stationary climate conditions.

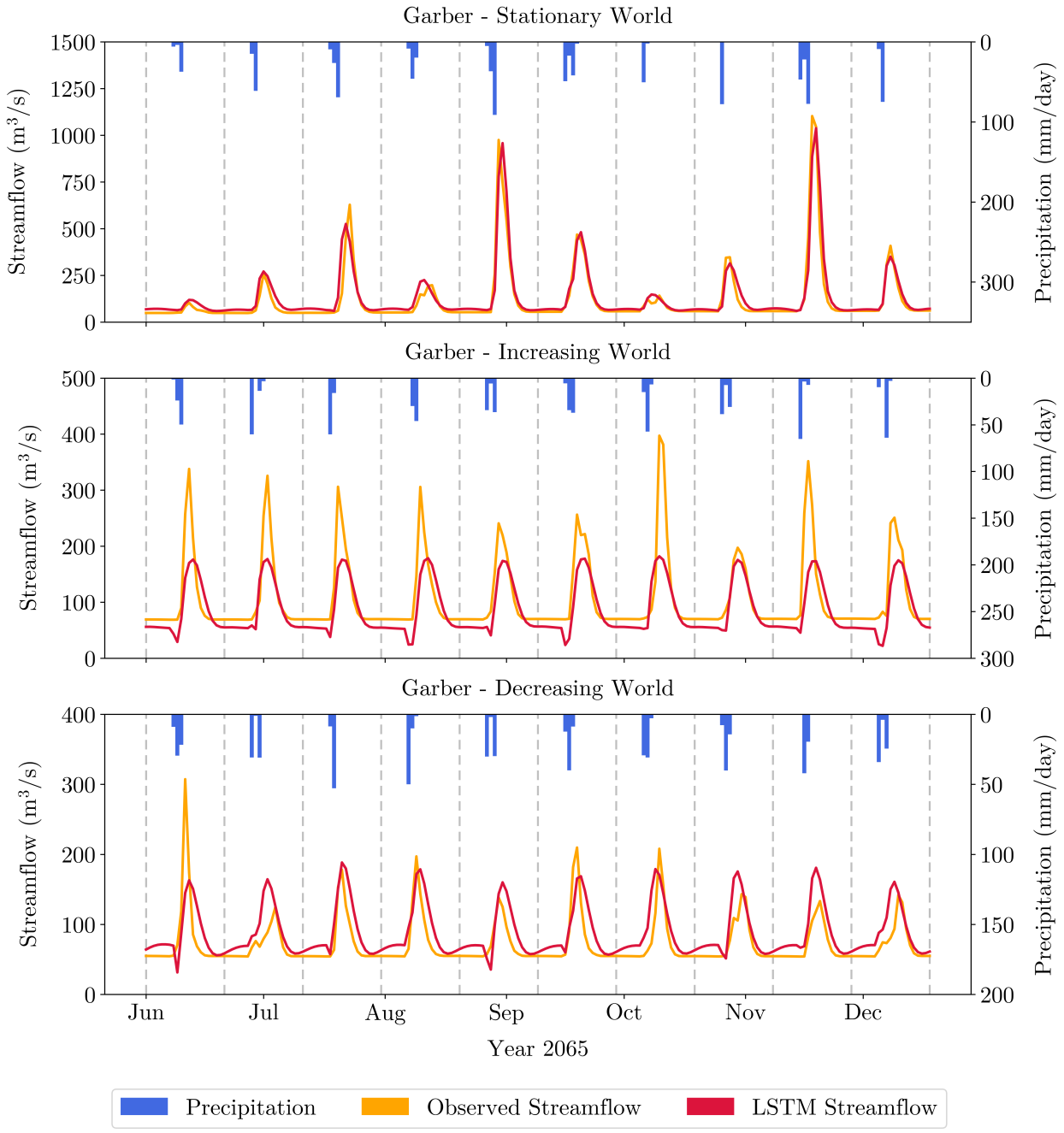

Figure 4.9: Zoomed-in hydrographs for Garber station, showing observed precipitation, observed streamflow and LSTM streamflow in one of MF years for (top) Stationary (middle) NS-Increasing (bottom) NS-Decreasing worlds.

4.4 Conclusion

This study investigated the reliability of machine learning (ML) algorithms, specifically LSTM networks, in predicting streamflow fluctuations, particularly in scenarios that extend beyond the historical record used for training. The central question addressed was how well these grey/black-box models perform when faced with hydrological events that involve either interpolation (conditions within the training range) or extrapolation (conditions outside the training range). This is especially relevant given the anticipated impacts of climate change, which are expected to lead to unprecedented extreme weather events.

The findings reinforce the observation from existing literature that ML algorithms can indeed perform reliably for streamflow prediction under conditions similar to those they have been trained on. In the Stationary world, where the statistical characteristics of the precipitation events remained consistent over time, the LSTM model demonstrated a strong ability to predict peak flows, streamflow event volumes, and the overall rainfall-runoff relationship across all five sub-catchments of the Turkey River Basin. The high NSE values achieved during the training, validation, and future testing periods in the Stationary world underscore the capability of deep learning models to capture the complex, non-linear relationships inherent in hydrological processes when the conditions remain stable.

However, the study also highlights important limitations when LSTM models are applied to predict streamflow under Non-Stationary conditions, where the characteristics of precipitation events change over time, mimicking potential climate change scenarios. In the Non-Stationary scenario with decreasing precipitation, the LSTM model tended to overestimate streamflow event volumes, particularly at lower observed flows, and the overall performance, as measured by NSE, declined in the mid and far future periods. This suggests that when exposed to

rainfall events smaller than those predominantly seen during training, the model's predictive ability diminishes.

Conversely, in the Non-Stationary scenario with increasing precipitation, the LSTM model generally underestimated peak flows and runoff volumes, especially for higher peak flows. While the predictability of streamflow event volumes appeared better in this increasing precipitation scenario, the discrepancies in peak flow prediction indicate that the model struggled to accurately capture the magnitude of extreme events exceeding the training data range. Interestingly, the overall performance (NSE) in the increasing precipitation world, while initially lower than in the decreasing scenario, showed better results in the mid and far-future periods. This might suggest a greater capacity to adapt to increasingly extreme events compared to decreasing ones, although accurate peak flow prediction remains a challenge.

A crucial observation across all scenarios is the influence of catchment size on the model's performance. The smallest catchment, Spillville, consistently exhibited the best predictive accuracy, while the two largest catchments, Elkader and Garber, generally showed the poorest performance. This suggests that the spatial scale and complexity of the watershed may play a significant role in the ability of data-driven models to generalize, particularly when faced with changing conditions.

These findings have significant implications for the use of ML in streamflow forecasting, especially in the context of climate change. While ML models have demonstrated superior performance in hindcast and forecast tests under historical conditions, their reliability when extrapolating to unprecedented hydrological events remains a key concern. This study indicates that while these models show promise in adapting to some forms of non-stationarity, they can exhibit significant biases in predicting peak flows, streamflow event volumes and

rainfall-runoff relationships when faced with events substantially outside their training experience.

Therefore, the implicit assumption that "if it served us well in the past it shall serve us well in the future" needs careful consideration when applying ML models for future hydrological predictions. While studies have shown that data-driven models can sometimes outperform process-based models even for extreme events, the analysis of this study suggests that the nature of the Non-Stationarity (increasing vs. decreasing precipitation) and the characteristics of the watershed can significantly influence the reliability of these predictions.

In conclusion, while ML algorithms offer a powerful tool for streamflow prediction, especially under stationary conditions, their application in a Non-Stationary future requires caution. Understanding their limitations in extrapolation scenarios and actively working towards enhancing their robustness and reliability for unseen hydrological events is essential for effective water resource management and risk assessment in a changing climate. The findings of this study contribute to this understanding by highlighting the differential performance of LSTM models under increasing and decreasing precipitation regimes and across different catchment sizes.

Chapter 5

Conclusion of the Thesis

5.1 Summary

Addressing the research gaps identified in the Section 1.2: Problem Definition, this thesis specifically evaluated the effectiveness of incorporating historical streamflow data, the impact of noisy precipitation inputs, and the performance of LSTMs in interpolating and extrapolating under stationary and non-stationary hydrological scenarios.

Chapter 2 investigated the effectiveness of including historical streamflow data as an input feature in LSTM models. The results demonstrated a significant enhancement in streamflow prediction accuracy across many diverse catchments when past streamflow information was integrated. This finding underscores the inherent temporal dependencies within the streamflow. However, the study also revealed that the benefit of incorporating historical streamflow data was not uniform across all regions and showed that certain catchment characteristics might be more influential in certain areas. Furthermore, while LSTM models showed strong performance for same-day streamflow prediction, their accuracy declined

with increasing forecast lead times. When compared against the Persistence model for multiple-day-ahead forecasts, LSTM demonstrated superior performance in terms of Pearson correlation, MAE, and RMSE, but did not show a clear advantage in NSE, particularly for longer lead times where both models exhibited poor performance.

Chapter 3 explored the impact of noisy precipitation input data on the accuracy of LSTM models. The study found that the introduction of noise generally led to a decline in the performance of the LSTM model, although models trained with noisy precipitation showed greater resilience when tested with similar noise. Notably, a substantial percentage of basins maintained relatively high NSE values even under significant noise levels, indicating a degree of robustness in capturing streamflow dynamics from imperfect data. The sensitivity of LSTM performance to precipitation noise varied considerably across different basins and was found to correlate with certain catchment attributes.

Chapter 4 examined the reliability of LSTM models in predicting streamflow under stationary and non-stationary climate scenarios simulated using a physically based hydrological model. Under stationary conditions, the LSTM model demonstrated remarkably high NSE values, indicating a strong ability to predict streamflow within the range of the training data. However, under non-stationary scenarios with increasing or decreasing precipitation, the model exhibited biases, tending to underestimate peak flows and runoff volumes in the increasing precipitation scenario and overestimate them in the decreasing scenario. The study also highlighted the influence of catchment size, with smaller catchments generally showing better predictive accuracy across all scenarios. These findings suggest that while LSTMs can learn complex hydrological relationships, their ability to extrapolate reliably to unseen hydrological conditions remains a significant challenge, particularly under evolving climate patterns.

5.2 Limitations

Chapter 2 & Chapter 3 primarily focused on the CAMELS-US dataset to investigate the impact of historical streamflow and noisy precipitation. While this dataset offers a large and diverse set of catchments, the findings might not be universally applicable to other geographical regions with different hydrological regimes or data availability. Furthermore, the investigation of noisy precipitation inputs involved the introduction of uniformly distributed random noise to a limited number of recent precipitation days. Real-world precipitation errors can exhibit more complex spatial and temporal patterns, and future research could explore the impact of more realistic noise structures. Chapter 4 relied on synthetic streamflow data generated by a physically based model. While this approach allows for controlled experiments under specific climate scenarios, the results depend on the accuracy and assumptions of the underlying hydrological model. Finally, the hyperparameter tuning for the LSTM models was conducted with certain constraints, and more extensive tuning might potentially yield more positive results.

5.3 Directions for Future Research

Based on findings of this thesis, several directions for future research can be identified:

- Incorporating forecasted meteorological data: To overcome the limitations in longer-term streamflow forecasting, future studies should explore the integration of forecasted precipitation, temperature, and/or other input data into LSTM models,
- Development of hybrid models: Combining the strengths of data-driven approaches like LSTM with physically based hydrological models could lead to more robust and reliable

streamflow predictions across various time scales and under changing conditions,

- Investigating different noise characteristics and uncertainty quantification: Future research could explore the impact of more realistic and complex precipitation error structures on LSTM performance and incorporate methods for quantifying prediction uncertainty,

- Evaluating advanced machine learning architectures: Exploring the potential of more recent machine learning architectures, such as Transformers or Informers, for streamflow forecasting under stationary and non-stationary conditions.

5.4 Final Conclusion

This thesis provides a comprehensive analysis of the effectiveness and limitations of LSTM networks for streamflow forecasting. The findings highlight the significant potential of LSTMs, particularly when leveraging historical streamflow data. However, the challenges encountered with longer lead times, noisy inputs, and especially under non-stationary conditions underscore the need for future research and development to improve the reliability and robustness of these models for hydrological forecasting and water resource management in a changing climate. By addressing the limitations identified and investigating the suggested directions for future research, machine learning hydrological modelling can contribute to more advanced and effective water resource management practices.

References

Acuña Espinoza, E., Loritz, R., Kratzert, F., Klotz, D., Gauch, M., Álvarez Chaves, M., & Ehret, U. (2025, March). Analyzing the generalization capabilities of a hybrid hydrological model for extrapolation to extreme events. *Hydrology and Earth System Sciences*, *29*(5), 1277–1294. Retrieved from http://dx.doi.org/10.5194/hess-29-1277-2025 doi: 10.5194/hess-29-1277-2025

Addor, N., Newman, A. J., Mizukami, N., & Clark, M. P. (2017, October). The camels data set: catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences*, *21*(10), 5293–5313. Retrieved from http://dx.doi.org/10.5194/hess-21-5293-2017 doi: 10.5194/hess-21-5293-2017

Arsenault, R., Martel, J. L., Brunet, F., Brissette, F., & Mai, J. (2023). Continuous streamflow prediction in ungauged basins: Long short-Term memory neural networks clearly outperform traditional hydrological models. *Hydrology and Earth System Sciences*, *27*(1), 139–157. doi: 10.5194/hess-27-139-2023

Asynch. (2023). *Built-in models — asynch documentation.* https://asynch.readthedocs.io/en/latest/builtin_models.html.

Beven, K. J. (2024, October). A short history of philosophies of hydrological model evaluation and hypothesis testing. *WIREs Water*, *12*(1). Retrieved from http://dx.doi.org/10.1002/wat2.1761 doi: 10.1002/wat2.1761

Collischonn, W., & Fan, F. M. (2012, June). Defining parameters for eckhardt's digital baseflow filter. *Hydrological Processes*, *27*(18), 2614–2622. Retrieved from http://dx.doi.org/10.1002/hyp.9391 doi: 10.1002/hyp.9391

Demiray, B. Z., Sit, M., Mermer, O., & Demir, I. (2024, April). Enhancing hydrological modeling with transformers: a case study for 24-h streamflow prediction. *Water Science amp; Technology*, *89*(9), 2326–2341. Retrieved from http://dx.doi.org/10.2166/wst.2024.110 doi: 10.2166/wst.2024.110

Du, J. (2011). *Ncep/emc 4km gridded data (grib) stage iv data.* UCAR/NCAR - Earth Observing Laboratory. Retrieved from https://doi.org/10.5065/D6PG1QDD doi: 10.5065/D6PG1QDD

Eckhardt, K. (2004, December). How to construct recursive digital filters for baseflow separation. *Hydrological Processes*, *19*(2), 507–515. Retrieved from http://dx.doi

.org/10.1002/hyp.5675 doi: 10.1002/hyp.5675

Eckhardt, K. (2008, April). A comparison of baseflow indices, which were calculated with seven different baseflow separation methods. *Journal of Hydrology*, *352*(1–2), 168–173. Retrieved from http://dx.doi.org/10.1016/j.jhydrol.2008.01.005 doi: 10.1016/j.jhydrol.2008.01.005

Eckhardt, K. (2012, February). Technical note: Analytical sensitivity analysis of a two parameter recursive digital baseflow separation filter. *Hydrology and Earth System Sciences*, *16*(2), 451–455. Retrieved from http://dx.doi.org/10.5194/hess-16-451-2012 doi: 10.5194/hess-16-451-2012

Feng, D., Lawson, K., & Shen, C. (2021). Mitigating Prediction Error of Deep Learning Streamflow Models in Large Data-Sparse Regions With Ensemble Modeling and Soft Data. *Geophysical Research Letters*, *48*(14), 1–12. doi: 10.1029/2021GL092999

Frame, J. M., Kratzert, F., Raney, A., Rahman, M., Salas, F. R., & Nearing, G. S. (2021, November). Post-processing the national water model with long short-term memory networks for streamflow predictions and model diagnostics. *JAWRA Journal of the American Water Resources Association*, *57*(6), 885–905. Retrieved from http://dx.doi.org/10.1111/1752-1688.12964 doi: 10.1111/1752-1688.12964

Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Lin, J., & Hochreiter, S. (2021, April). Rainfall–runoff prediction at multiple timescales with a single long short-term memory network. *Hydrology and Earth System Sciences*, *25*(4), 2045–2062. Retrieved from http://dx.doi.org/10.5194/hess-25-2045-2021 doi: 10.5194/hess-25-2045-2021

Ghimire, G. R., & Krajewski, W. F. (2019, December). Exploring persistence in streamflow forecasting. *JAWRA Journal of the American Water Resources Association*, *56*(3), 542–550. Retrieved from http://dx.doi.org/10.1111/1752-1688.12821 doi: 10.1111/1752-1688.12821

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009, October). Decomposition of the mean squared error and nse performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, *377*(1–2), 80–91. Retrieved from http://dx.doi.org/10.1016/j.jhydrol.2009.08.003 doi: 10.1016/j.jhydrol.2009.08.003

Gurbuz, F., Mudireddy, A., Mantilla, R., & Xiao, S. (2024, January). Using a physics-based hydrological model and storm transposition to investigate machine-learning algorithms for streamflow prediction. *Journal of Hydrology*, *628*, 130504. Retrieved from http://dx.doi.org/10.1016/j.jhydrol.2023.130504 doi: 10.1016/j.jhydrol.2023.130504

Hochreiter, S., & Schmidhuber, J. (1997, November). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. Retrieved from http://dx.doi.org/10.1162/neco.1997.9.8.1735 doi: 10.1162/neco.1997.9.8.1735

Hong, Y., Hsu, K. L., Moradkhani, H., & Sorooshian, S. (2006). Uncertainty quantification of satellite precipitation estimation and Monte Carlo assessment of the error propagation into hydrologic response. *Water Resources Research*, *42*(8), 1–15. doi: 10.1029/2005WR004398

Huard, D., & Mailhot, A. (2006). A Bayesian perspective on input uncertainly in model calibration: Application to hydrological model "abc". *Water Resources Research*, *42*(7), 1–14. doi: 10.1029/2005WR004661

Hunt, K. M., Matthews, G. R., Pappenberger, F., & Prudhomme, C. (2022). Using a long short-term memory (LSTM) neural network to boost river streamflow forecasts over the western United States. *Hydrology and Earth System Sciences*, *26*(21), 5449–5472. doi: 10.5194/hess-26-5449-2022

Incorporated, I., & the Spatial Sciences Laboratory at Texas AM. (2023, May). *Hawqs 2.0: Hydrologic and water quality system* (Tech. Rep.). U.S. Environmental Protection Agency.

Krajewski, W. F., Ceynar, D., Demir, I., Goska, R., Kruger, A., Langel, C., ... Young, N. C. (2017, March). Real-time flood forecasting and information system for the state of iowa. *Bulletin of the American Meteorological Society*, *98*(3), 539–554. Retrieved from http://dx.doi.org/10.1175/BAMS-D-15-00243.1 doi: 10.1175/bams-d-15-00243.1

Kratzert, F., Gauch, M., Nearing, G., & Klotz, D. (2022, March). Neuralhydrology — a python library for deep learning research in hydrology. *Journal of Open Source Software*, *7*(71), 4050. Retrieved from http://dx.doi.org/10.21105/joss.04050 doi: 10.21105/joss.04050

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018, November). Rainfall–runoff modelling using long short-term memory (lstm) networks. *Hydrology and Earth System Sciences*, *22*(11), 6005–6022. Retrieved from http://dx.doi.org/10.5194/hess-22-6005-2018 doi: 10.5194/hess-22-6005-2018

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. S. (2019, December). Toward improved predictions in ungauged basins: Exploiting the power of machine learning. *Water Resources Research*, *55*(12), 11344–11354. Retrieved from http://dx.doi.org/10.1029/2019WR026065 doi: 10.1029/2019wr026065

Kratzert, F., Klotz, D., Hochreiter, S., & Nearing, G. S. (2021, May). A note on leveraging synergy in multiple meteorological data sets with deep learning for rainfall–runoff modeling. *Hydrology and Earth System Sciences*, *25*(5), 2685–2703. Retrieved from http://dx.doi.org/10.5194/hess-25-2685-2021 doi: 10.5194/hess-25-2685-2021

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019, December). 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. Retrieved from http://dx.doi.org/10.5194/hess-23-5089-2019 doi: 10.5194/hess-23-5089-2019

Lafon, T., Dadson, S., Buys, G., & Prudhomme, C. (2013). Bias correction of daily precipitation simulated by a regional climate model: A comparison of methods. *International Journal of Climatology*, *33*(6), 1367–1381. doi: 10.1002/joc.3518

Le, X.-H., Nguyen, D.-H., Jung, S., Yeon, M., & Lee, G. (2021). Comparison of deep learning techniques for river streamflow forecasting. *IEEE Access*, *9*, 71805–71820.

Retrieved from http://dx.doi.org/10.1109/ACCESS.2021.3077703 doi: 10.1109/access.2021.3077703

Lees, T., Buechel, M., Anderson, B., Slater, L., Reece, S., Coxon, G., & Dadson, S. J. (2021, October). Benchmarking data-driven rainfall–runoff models in great britain: a comparison of long short-term memory (lstm)-based models with four lumped conceptual models. *Hydrology and Earth System Sciences*, *25*(10), 5517–5534. Retrieved from http://dx.doi.org/10.5194/hess-25-5517-2021 doi: 10.5194/hess-25-5517-2021

Lin, Y., Wang, D., Jiang, T., & Kang, A. (2024, March). Assessing objective functions in streamflow prediction model training based on the naïve method. *Water*, *16*(5), 777. Retrieved from http://dx.doi.org/10.3390/w16050777 doi: 10.3390/w16050777

Liu, J., Koch, J., Stisen, S., Troldborg, L., & Schneider, R. J. M. (2024, July). A national-scale hybrid model for enhanced streamflow estimation – consolidating a physically based hydrological model with long short-term memory (lstm) networks. *Hydrology and Earth System Sciences*, *28*(13), 2871–2893. Retrieved from http://dx.doi.org/10.5194/hess-28-2871-2024 doi: 10.5194/hess-28-2871-2024

Liu, S., Lu, D., Painter, S. L., Griffiths, N. A., & Pierce, E. M. (2023). Uncertainty quantification of machine learning models to improve streamflow prediction under changing climate and environmental conditions. *Frontiers in Water*, *5*(Ml). doi: 10.3389/frwa.2023.1150126

Mantilla, R. (2007). *Physical basis of statistical scaling in peak flows and stream flow hydrographs for topologic and spatially embedded random self-similar channel networks* (PhD Thesis). University of Colorado.

Mantilla, R., Gupta, V. K., & J. Mesa, O. (2006, May). Role of coupled flow dynamics and real network structures on hortonian scaling of peak flows. *Journal of Hydrology*, *322*(1–4), 155–167. Retrieved from http://dx.doi.org/10.1016/j.jhydrol.2005.03.022 doi: 10.1016/j.jhydrol.2005.03.022

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., ... Duan, Q. (2015, January). Development of a large-sample watershed-scale hydrometeorological data set for the contiguous usa: data set characteristics and assessment of regional variability in hydrologic model performance. *Hydrology and Earth System Sciences*, *19*(1), 209–223. Retrieved from http://dx.doi.org/10.5194/hess-19-209-2015 doi: 10.5194/hess-19-209-2015

Nifa, K., Boudhar, A., Ouatiki, H., Elyoussfi, H., Bargam, B., & Chehbouni, A. (2023). Deep Learning Approach with LSTM for Daily Streamflow Prediction in a Semi-Arid Area: A Case Study of Oum Er-Rbia River Basin, Morocco. *Water (Switzerland)*, *15*(2). doi: 10.3390/w15020262

Pathiraja, S., Moradkhani, H., Marshall, L., Sharma, A., & Geenens, G. (2018). Data-Driven Model Uncertainty Estimation in Hydrologic Data Assimilation. *Water Resources Research*, *54*(2), 1252–1280. doi: 10.1002/2018WR022627

Perez, G., Mantilla, R., Krajewski, W. F., & Wright, D. B. (2019, November). Using

physically based synthetic peak flows to assess local and regional flood frequency analysis methods. *Water Resources Research*, *55*(11), 8384–8403. Retrieved from http://dx.doi.org/10.1029/2019WR024827 doi: 10.1029/2019wr024827

Pokharel, S., & Roy, T. (2024, October). A parsimonious setup for streamflow forecasting using cnn-lstm. *Journal of Hydroinformatics*, *26*(11), 2751–2761. Retrieved from http://dx.doi.org/10.2166/hydro.2024.114 doi: 10.2166/hydro.2024.114

Politano, M., Arenas, A., & Weber, L. (2023, June). A process-based hydrological model for continuous multi-year simulations of large-scale watersheds. *International Journal of River Basin Management*, *23*(1), 15–28. Retrieved from http://dx.doi.org/10.1080/15715124.2023.2216937 doi: 10.1080/15715124.2023.2216937

Quintero, F., Krajewski, W. F., Seo, B.-C., & Mantilla, R. (2020, May). Improvement and evaluation of the iowa flood center hillslope link model (hlm) by calibration-free approach. *Journal of Hydrology*, *584*, 124686. Retrieved from http://dx.doi.org/10.1016/j.jhydrol.2020.124686 doi: 10.1016/j.jhydrol.2020.124686

Small, S. J., Jay, L. O., Mantilla, R., Curtu, R., Cunha, L. K., Fonley, M., & Krajewski, W. F. (2013, March). An asynchronous solver for systems of odes linked by a directed tree structure. *Advances in Water Resources*, *53*, 23–32. Retrieved from http://dx.doi.org/10.1016/j.advwatres.2012.10.011 doi: 10.1016/j.advwatres.2012.10.011

Song, Y., Sawadekar, K., Frame, J. M., Pan, M., Clark, M., Knoben, W. J. M., ... Shen, C. (2025, March). Physics-informed, differentiable hydrologic models for capturing unseen extreme events. *ESS Open Archive*. Retrieved from http://dx.doi.org/10.22541/essoar.172304428.82707157/v2 (Preprint) doi: 10.22541/essoar.172304428.82707157/v2

Szilagyi, J. (2004, July). Heuristic continuous base flow separation. *Journal of Hydrologic Engineering*, *9*(4), 311–318. Retrieved from http://dx.doi.org/10.1061/(ASCE)1084-0699(2004)9:4(311) doi: 10.1061/(asce)1084-0699(2004)9:4(311)

Vergara, H., Hong, Y., Gourley, J. J., Anagnostou, E. N., Maggioni, V., Stampoulis, D., & Kirstetter, P. E. (2014). Effects of resolution of satellite-based rainfall estimates on hydrologic modeling: Skill at different scales. *Journal of Hydrometeorology*, *15*(2), 593–613. doi: 10.1175/JHM-D-12-0113.1

Wright, D. B., Mantilla, R., & Peters-Lidard, C. D. (2017, April). A remote sensing-based tool for assessing rainfall-driven hazards. *Environmental Modelling amp; Software*, *90*, 34–54. Retrieved from http://dx.doi.org/10.1016/j.envsoft.2016.12.006 doi: 10.1016/j.envsoft.2016.12.006

Xia, Y., Mitchell, K., Ek, M., Sheffield, J., Cosgrove, B., Wood, E., ... Mocko, D. (2012, February). Continental-scale water and energy flux analysis and validation for the north american land data assimilation system project phase 2 (nldas-2): 1. intercomparison and application of model products. *Journal of Geophysical Research: Atmospheres*, *117*(D3). Retrieved from http://dx.doi.org/10.1029/2011JD016048 doi: 10.1029/2011jd016048

Xie, J., Liu, X., Wang, K., Yang, T., Liang, K., & Liu, C. (2020, April). Evaluation of typical methods for baseflow separation in the contiguous united states. *Journal of Hydrology*, *583*, 124628. Retrieved from http://dx.doi.org/10.1016/j.jhydrol.2020.124628 doi: 10.1016/j.jhydrol.2020.124628

Yifru, B. A., Lim, K. J., Bae, J. H., Park, W., & Lee, S. (2024, March). A hybrid deep learning approach for streamflow prediction utilizing watershed memory and process-based modeling. *Hydrology Research*, *55*(4), 498–518. Retrieved from http://dx.doi.org/10.2166/nh.2024.016 doi: 10.2166/nh.2024.016

Yu, G., Wright, D. B., Zhu, Z., Smith, C., & Holman, K. D. (2019, May). Process-based flood frequency analysis in an agricultural watershed exhibiting nonstationary flood seasonality. *Hydrology and Earth System Sciences*, *23*(5), 2225–2243. Retrieved from http://dx.doi.org/10.5194/hess-23-2225-2019 doi: 10.5194/hess-23-2225-2019

Zhu, Z., Wright, D. B., & Yu, G. (2018, November). The impact of rainfall space-time structure in flood frequency analysis. *Water Resources Research*, *54*(11), 8983–8998. Retrieved from http://dx.doi.org/10.1029/2018WR023550 doi: 10.1029/2018wr023550

Appendix A

A.1 Description of Catchment Attributes of CAMELS-US Basins

Table A.1: Description of Catchment Attributes

Attribute	Description
Precipitation mean	Mean daily precipitation
PET mean	Mean daily potential evapotranspiration
Aridity index	Ratio of Mean PET to Mean Precipitation
Precipitation seasonality	Seasonality and timing of precipitation (estimated using sine curves to represent the annual temperature and precipitation cycles, positive [negative] values indicate that precipitation peaks in summer [winter], values close to 0 indicate uniform precipitation throughout the year)
Snow fraction	Fraction of precipitation falling as snow (i.e., on days colder than 0°C)
High precipitation frequency	Frequency of high precipitation days ($\geq$ 5 times mean daily precipitation)
High precipitation duration	Average duration of high precipitation events (number of consecutive days $\geq$ 5 times mean daily precipitation)
Low precipitation frequency	Frequency of dry days (1 mm/day)

Continued on next page

Attribute	**Description**
Low precip duration	Average duration of dry periods (number of consecutive days < 1 mm/day)
Elevation	Catchment mean elevation
Slope	Catchment mean slope
Area	Catchment area (Geospatial Fabric estimate)
Forest fraction	Fraction of catchment covered by forest
LAI max	Maximum monthly mean of leaf area index
LAI difference	Difference between the max. and min. mean of the leaf area index
GVF max	Maximum monthly mean of green vegetation fraction
GVF difference	Difference between the maximum and minimum monthly mean of the green vegetation fraction
Soil depth (Pelletier)	Depth to bedrock (maximum 50 m)
Soil depth (STATSGO)	Soil depth (maximum 1.5m, layers marked as water and bedrock were excluded)
Soil Porosity	Volumetric porosity
Soil conductivity	Saturated hydraulic conductivity
Max water content	Maximum water content of the soil
Sand fraction	Fraction of sand in the soil
Silt fraction	Fraction of silt in the soil
Clay fraction	Fraction of clay in the soil
Carbonate rocks fraction	Fraction of the catchment area characterized as “carbonate sedimentary rocks”
Geological permeability	Surface permeability (log10)

Appendix B

B.1 Hyperparameter Tuning of LSTM model

For tuning the hyperparameters of the LSTM model where past streamflow data were incorporated as an additional dynamic input variable in the benchmark LSTM model, we changed four hyperparameters (Hidden Size, Batch Size, Dropout, Sequence Length) in a small grid of values. Among the 81 different hyperparameter combinations, the combination with the highest median NSE in the validation period was selected. Note that the model was trained only for one epoch for hyperparameter tuning due to time constraints and high computational requirements. Table B.1 lists the tested parameter grid and selected values.

Table B.1: Tested parameter grid and selected values

Hyperparameter	**Tested Grid**	**Selected**
Hidden Size	128, 256, 512	256
Output Dropout	0.2, 0.4, 0.6	0.4
Batch Size	128, 256, 512	256
Sequence Length (months)	9, 12, 15	12

Figure B.1 presents the median NSE for each hyperparameter combination tested. Table B.2 lists the other parameters and configurations of LSTM models which are the same as in the Naive LSTM model by Gauch et al. (2021).

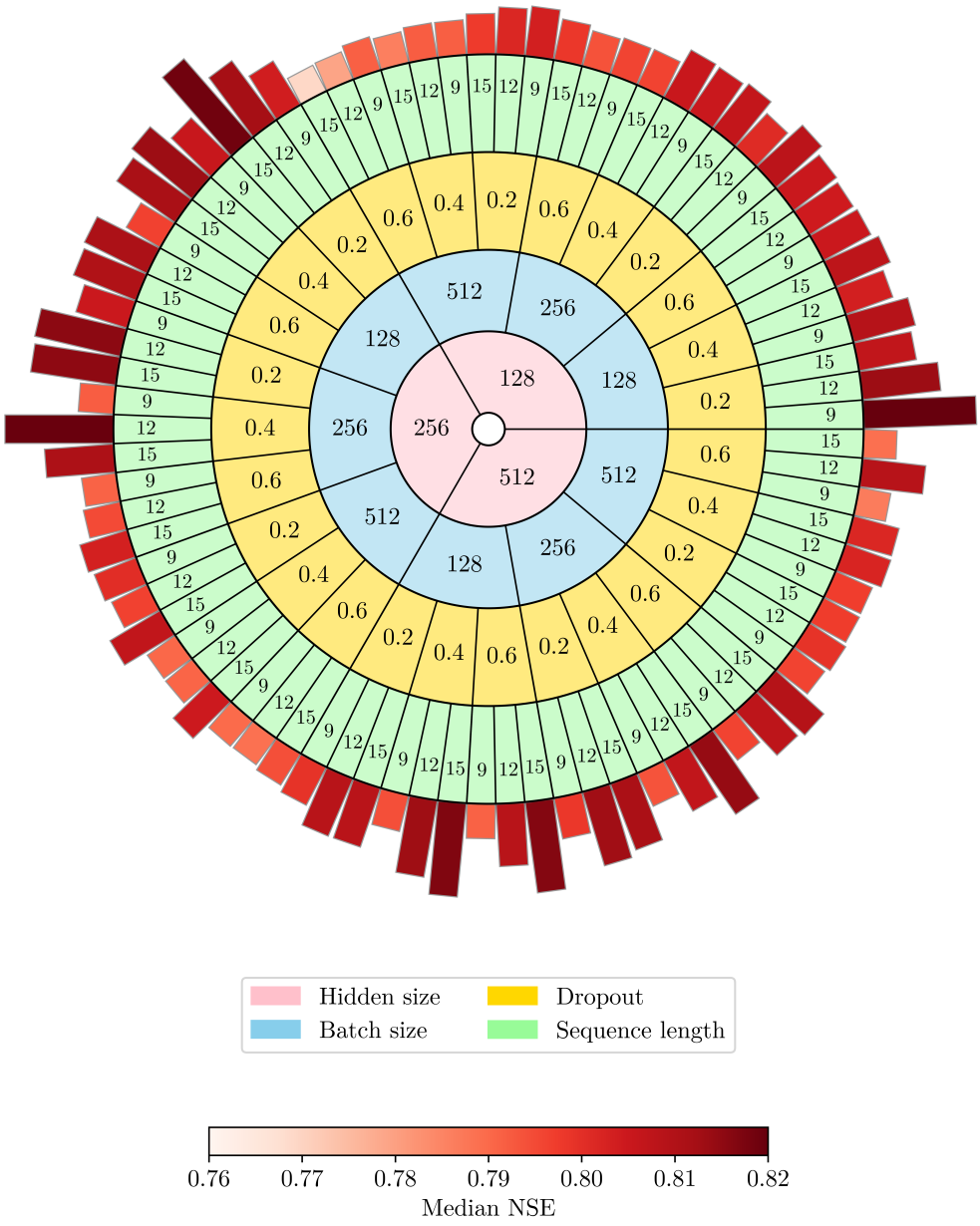

Figure B.1: The median NSE for each hyperparameter combination tested

Table B.2: Other parameters and configurations used in LSTM models

Configuration	**Selection**	**Parameter**	**Selection**
Head	Regression	Number of Epochs	30
Output Activation	Linear	Learning Rate:	0: 0.001
Optimizer	Adam		10: 0.0005
Loss	NSE		25: 0.0001