

LB/TH/08/2026

TH6137

**INTEGRATING REMOTE SENSING AND
UNSUPERVISED MACHINE LEARNING TECHNIQUES FOR
RUNOFF PREDICTION IN UNGAUGED CATCHMENTS
THROUGH
SPATIOTEMPORAL CATCHMENT SIMILARITY ANALYSIS**

Pilana Vithanage Ravindu Prabhatha Madhumal

248076N

Degree of Master of Science in Civil Engineering

Department of Civil Engineering

University of Moratuwa

Sri Lanka

December 2025

**INTEGRATING REMOTE SENSING AND
UNSUPERVISED MACHINE LEARNING TECHNIQUES FOR
RUNOFF PREDICTION IN UNGAUGED CATCHMENTS
THROUGH
SPATIOTEMPORAL CATCHMENT SIMILARITY ANALYSIS**

Pilana Vithanage Ravindu Prabhatha Madhumal

208076N

Supervised by

Dr T. M. N. Wijayaratna

Thesis submitted in partial fulfillment of the requirements for the degree of
Master of Science in Civil Engineering

UNESCO Madanjeet Singh Centre for South
Asia Water Management (UMCSAWM)

Department of Civil Engineering
University of Moratuwa
Sri Lanka

December 2025

DECLARATION OF THE CANDIDATE AND SUPERVISOR

“I declare that this is my own work, and this thesis/dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text”.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my thesis/dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Pilana Vithanage Ravindu
Prabhatha Madhumal

15/11/2025

Date

The above candidate has carried out research for the Master’s Thesis/Dissertation under my supervision.

Dr T. M. N. Wijayaratna

29/12/2025

Date

ABSTRACT

Integrating Remote Sensing and Unsupervised Machine Learning Techniques for Runoff Prediction in Ungauged Catchments Through Spatiotemporal Catchment Similarity

Predicting streamflow in ungauged basins is a significant challenge in hydrology. This study presents a framework that integrates remote sensing and machine learning to predict runoff for ungauged catchments by analyzing catchment similarity and evaluating whether training on hydrologically similar catchments outperforms training on all available streamflow and precipitation data across catchments. The study analyzed 30 catchments across Sri Lanka's wet, intermediate, and dry zones using two decades of streamflow and precipitation data. An initial suite of 14 hydrological signatures derived from daily river discharge and rainfall time series was refined via correlation analysis ($|r| \geq 0.7$) into six non-redundant signatures (Runoff Ratio, Baseflow Index, Flow Duration Curve Slope, Recession Constant, Precipitation Elasticity, Rising Limb Density) that together represent the complete unit hydrograph. For classification, multiple algorithms (K-Means, K-Medoids, Hierarchical) and distance metrics (Euclidean, Manhattan, Cosine) were evaluated. Hierarchical clustering with cosine distance proved optimal, forming seven distinct clusters with a high Silhouette Score of 0.525. Additional validation using satellite-derived data confirmed strong hydrological coherence (Hydrological Coherence Index = 1.48). For ungauged applications, signatures were predicted from satellite data, achieving a pairwise accuracy of 87.4% in replicating cluster relationships, although exact membership accuracy was lower (60%) due to borderline catchments. The prediction compared three approaches for pseudo-ungauged basins: Monaragala (dry zone) and Ellagawa (wet zone). In both the Dry and Wet zones, the traditional HEC-Hydrologic Modeling System parameter transfer failed to deliver sufficient predictions (Monaragala: Nash-Sutcliffe Efficiency (NSE) = -0.407; Ellagawa: NSE = 0.425). Long Short-Term Memory (LSTM) models were universally superior, but the optimal strategy was regime dependent. In the Wet Zone, the Regionalized LSTM (trained on similar donors) performed best (Ellagawa: NSE = 0.709, Root Mean Square Error (RMSE) = 4.20 mm/day). Conversely, in the Dry Zone, the Global LSTM (trained on all 30 catchments) was more effective (Monaragala: NSE = 0.535, RMSE = 2.88 mm/day), suggesting that data volume aids learning of complex arid hydrology, though low-flow prediction remained a challenge for all models. The findings mandate a regime-specific prediction strategy: hydrological similarity guides optimal donor selection in wet zones, while a global data approach is more effective in dry zones.

Keywords: Clustering, HEC-HMS, Hydrological Signatures, LSTM, Regionalization

DEDICATION

This thesis is dedicated to the people whose presence, encouragement, and quiet persistence shaped my progress more than they may ever realise. Their support often arrived at moments when the work felt heavy, and these small acts of reassurance gradually built the resilience I needed to continue.

To my supervisor, Dr T. M. N. Wijayaratna, I extend my deep appreciation. His guidance did not stop at academic advice but frequently gave me the confidence to attempt ideas I once thought beyond my reach.

My family has remained central to this achievement, and I acknowledge my mother, father, and sister for their quiet strength, steady encouragement, and unwavering confidence in my ability to complete this work. After my father's passing, my mother carried an extraordinary share of responsibility, and her support sustained me through demanding stages of this degree without seeking recognition for her efforts.

My friends played a crucial role by offering motivation, understanding, and companionship during uncertain times, reminding me that perseverance is often a collective effort. This dedication expresses my gratitude to each of them, as their support shaped my thesis and will continue to guide my work.

ACKNOWLEDGEMENT

First, I would like to express my sincere appreciation to my supervisor, Dr. T. M. N. Wijayaratna, whose steady guidance, thoughtful critique, and willingness to share expertise have shaped the direction and quality of this research. His support created an environment in which ideas could develop with clarity, and his patience during demanding phases of the study contributed significantly to my academic progress.

I also acknowledge the academic staff of the UMCSAWM at the University of Moratuwa, including the Director and Centre Chairman, Prof. Lalith Rajapakse, whose commitment to teaching and continuous encouragement strengthened my understanding throughout the course. The academic staff of the Department of Civil Engineering at the University of Moratuwa, including the Research Coordinator, Prof. H. R. Pasindu, also offered valuable guidance and created a learning structure that consistently supported my development.

I am grateful to Eng. A. Gunasekara, Director General of Irrigation, and Eng. L. S. Sooriyabandara, Director of Irrigation for Hydrology and Disaster Management, together with the staff of the Irrigation Department, for providing essential streamflow observations that were central to the analytical procedures of this work. I also thank the Department of Meteorology for their assistance in providing the meteorological datasets required for the hydrological assessment.

The scholarship support offered by the South Asia Foundation contributed directly to my ability to pursue this degree, and I remain appreciative of their commitment to regional education.

I extend my gratitude to the administrative staff of the Department of Civil Engineering and the UMCSAWM, including Ms Vinu Kalanika and Ms Janani Nisansala, whose consistent assistance with academic matters eased the practical aspects of this journey and ensured that the broader research environment remained supportive and well-organized.

Lastly, I appreciate my fellow researchers and colleagues for their camaraderie and constructive discussions throughout this journey.

TABLE OF CONTENT

Declaration of the Candidate and Supervisor	III
Abstract.....	V
Dedication	VII
Acknowledgement	IX
Table of Content.....	XI
List of Figures.....	XV
List of Tables	XIX
List of Abbreviations	XXI
1 Introduction	1
1.1 Background.....	1
1.2 Problem Statement.....	3
1.3 Aims and Objectives.....	4
1.3.1 Main objective.....	4
1.3.2 Specific objectives.....	4
1.4 Scope and Limitations.....	4
1.5 Significance of the Study	5
1.6 Structure of the Thesis	6
2 Literature Review.....	7
2.1 Runoff Prediction.....	7
2.1.1 Traditional regionalization methods	7
2.1.2 Deep learning methods	9
2.2 Catchment Classification	12
2.2.1 Hydrological signatures.....	12

2.2.2	Clustering methods for classification	20
2.3	Classification Validation Framework	27
2.3.1	Remote sensing data for validation.....	28
2.4	Runoff Prediction Method Comparison.....	30
2.4.1	Hydrological model selection.....	30
2.4.2	HEC-HMS methods selection.....	31
2.4.3	Model evaluation	33
2.5	Literature Summery	34
3	Methodology	37
3.1	Initial Stage	38
3.1.1	Study area	38
3.1.2	Data quality control and gap-filling protocol.....	42
3.1.3	Areal rainfall estimation and discharge normalization	42
3.2	Specific Objective 1	43
3.2.1	Catchment classification.....	43
3.2.2	Hydrological validation of classification.....	44
3.3	Specific Objective 2.....	46
3.3.1	Prediction of hydrological signatures from satellite data.....	47
3.3.2	Cluster accuracy assessment and donor catchment identification	48
3.3.3	Streamflow prediction using regionalized and global LSTM models	49
3.4	Specific Objective 3.....	49
3.4.1	Model development and parameter transfer	49
4	Result and Discussion.....	51
4.1	Hydrological Signature Selection and Dimensionality Reduction.....	51
4.2	Catchment Classification Results from Selected Clustering Algorithms	53
4.3	Hydrological Validation of Cluster Solutions Using Satellite-Derived Composite Attributes	54

4.3.1	Composite attribute score calculation.....	54
4.3.2	Within-cluster variance calculations.....	55
4.3.3	Between-cluster variance calculations.....	56
4.3.4	Synthesis of hydrological coherence index.....	57
4.4	Optimum Classification Algorithm and Distance Method Selection.....	58
4.4.1	K-Medoids clustering with cosine distance	59
4.4.2	Hierarchical clustering with cosine distance.....	64
4.5	Prediction of Hydrological Signatures from Satellite Data	69
4.6	Assessment of Cluster Prediction Accuracy Using Pairwise Comparison	72
4.7	Regional LSTM Performance	74
4.7.1	Dry zone: Monaragala case study.....	74
4.7.2	Wet zone: Ellagawa case study.....	78
4.8	Global LSTM Performance.....	81
4.8.1	Dry zone: Monaragala case study.....	81
4.8.2	Wet zone: Ellagawa case study.....	83
4.9	HEC-HMS Model Setup and Parameterization	85
4.9.1	Dry zone case study: Parameter transfer in Monaragala catchment	85
4.9.2	Wet zone case study: Parameter transfer in the Ellagawa catchment	90
4.10	Prediction Results Comparison.....	95
5	Conclusions and Recommendations	98
5.1	Conclusions.....	98
5.2	Recommendations.....	100
Annexure 1.....		115
Annexure 2.....		120
Annexure 3.....		122
Annexure 4.....		128

Table of Content

Annexure 5.....	148
Annexure 6.....	150
Annexure 7.....	156

LIST OF FIGURES

Figure 3-1: Methodological flowchart-----	37
Figure 3-2: Spatial distribution of selected catchments -----	40
Figure 4-1: Pairwise correlation matrix of the fifteen computed hydrological signatures -----	51
Figure 4-2: Correlation matrix of the six filtered hydrological signatures selected for cluster analysis -----	52
Figure 4-3: WCV based on CAS for each hydrological signature across all clustering methods -----	56
Figure 4-4: BCV based on CAS for each hydrological signature across all clustering methods -----	57
Figure 4-5: HCI values for all clustering method configurations -----	57
Figure 4-6: Principal component projections of the seven-cluster K-Medoids solution with cosine distance-----	59
Figure 4-7: Distribution of the six hydrological signatures across the seven K-Medoids clusters-----	60
Figure 4-8: Spatial distribution of the seven hydrological clusters derived from the K-Medoids algorithm with cosine distance across the study catchments -----	61
Figure 4-9: Temporal variation of streamflow and rainfall for Kataragama catchment during 08/2017 -----	63
Figure 4-10: Dendrogram illustrating the hierarchical cluster structure using cosine distance -----	65
Figure 4-11: Principal component projections of the seven-cluster hierarchical solution -----	65
Figure 4-12: Spatial distribution of the seven hydrological clusters derived from the Hierarchical algorithm with cosine distance across the study catchments -----	66
Figure 4-13: Distribution of the six hydrological signatures across the seven Hierarchical clusters --	67
Figure 4-14: Predicted versus observed hydrological signatures from the satellite-based regression models (a) - (f) -----	72
Figure 4-15: Pairwise Comparison Matrix of the original and predicted cluster distributions -----	73
Figure 4-16: Regional LSTM simulated and observed hydrographs for first two years of data: Monaragala dry zone -----	76
Figure 4-17: Observed and simulated hydrograph for pseudo-ungauged Monaragala (2014–2016)--	77
Figure 4-18: Predicted and observed sorted flow duration curve for pseudo-ungauged Monaragala (2014–2024) -----	77
Figure 4-19: Predicted and observed unsorted flow duration curve for pseudo-ungauged Monaragala (2014–2024) -----	78
Figure 4-20: Regional LSTM simulated and observed hydrographs for first two years of data: Ellagawa wet zone -----	79
Figure 4-21: Observed and simulated hydrograph for pseudo-ungauged Ellagawa (2003–2005) -----	80
Figure 4-22: Predicted and observed flow duration curve for pseudo-ungauged Ellagawa (2003–2017) -----	80

Figure 4-23: Predicted and observed flow duration curve for pseudo-ungauged Ellagawa (2003–2017)	81
Figure 4-24: Observed and simulated hydrograph for pseudo-ungauged Monaragala (2014–2016) --	82
Figure 4-25: Predicted and observed sorted flow duration curve for pseudo-ungauged Monaragala (2014–2024)	82
Figure 4-26: Predicted and observed sorted flow duration curve for pseudo-ungauged Monaragala (2014–2024)	83
Figure 4-27: Observed and simulated hydrograph for pseudo-ungauged Ellagawa (2003–2005) ----	84
Figure 4-28: Predicted and observed sorted flow duration curve for pseudo-ungauged Ellagawa (2003–2017)	84
Figure 4-29: Predicted and observed unsorted flow duration curve for pseudo-ungauged Ellagawa (2003–2017)	85
Figure 4-30: First two years of calibration hydrograph for Padiyathalawa catchment (2005–2007) --	86
Figure 4-31: Sorted flow duration curve for Padiyathalawa catchment - calibration period (2005–2015)	86
Figure 4-32: Unsorted flow duration curve for Padiyathalawa catchment - calibration period (2005–2015)	87
Figure 4-33: First two years of validation hydrograph for Padiyathalawa catchment (2015–2017) ---	87
Figure 4-34: Sorted flow duration curve for Padiyathalawa catchment - validation period (2015–2025)	88
Figure 4-35: Unsorted flow duration curve for Padiyathalawa catchment - validation period (2015–2025)	88
Figure 4-36: First two years of simulated vs observed hydrograph for Monaragala catchment using transferred parameters (2014–2016) -----	89
Figure 4-37: Sorted flow duration curve for Monaragala catchment using transferred parameters (2014–2024)	89
Figure 4-38: Sorted flow duration curve for Monaragala catchment using transferred parameters (2014–2024)	90
Figure 4-39: First two years of calibration hydrograph for Ratnapura catchment (2005–2007) -----	91
Figure 4-40: Sorted flow duration curve for Ratnapura catchment - calibration period (2005–2012)	91
Figure 4-41: Sorted flow duration curve for Ratnapura catchment - calibration period (2005–2012)	92
Figure 4-42: First two years of validation hydrograph for Ratnapura catchment (2012–2014) -----	92
Figure 4-43: Sorted flow duration curve for Ratnapura catchment - validation period (2012–2017) -	93
Figure 4-44: Unsorted flow duration curve for Ratnapura catchment - validation period (2012–2017)	93

Figure 4-45: First two years simulated vs observed hydrograph for Ellagawa catchment using transferred parameters (2004–2017)-----	94
Figure 4-46: Sorted flow duration curve for Ellagawa catchment using transferred parameters (2004–2017)-----	94
Figure 4-47: Sorted flow duration curve for Ellagawa catchment using transferred parameters (2004–2017)-----	95

LIST OF TABLES

Table 2-1: Applications of Satellite platforms in Sri Lanka.....	29
Table 3-1: Details of selected catchments.....	38
Table 3-2: Remote sensing data sources used in this study.....	40
Table 4-1: Silhouette score and DBI for all clustering algorithm-distance metric combinations.....	53
Table 4-2: Satellite products and correlation coefficients used for CAS calculation.....	54
Table 4-3: Satellite-derived predictor variables selected for the hydrological signature prediction models	69
Table 4-4: Pairwise accuracy results.....	73
Table 4-5: Performance comparison of streamflow prediction approaches for dry zone and wet zone catchments	96

LIST OF ABBREVIATIONS

ANOVA - Analysis of Variance

BFI - Baseflow Index

BCSI - Between-Cluster Separation Index

BCV - Between-Cluster Variance

BD - Bulk Density

CI - Canopy Interception

CLYPPT - Clay Content

CoV - Coefficient of Variation

CAS - Composite Attribute Score

DBI - Davies-Bouldin Index

DL - Deep Learning

DBSCAN - Density-Based Spatial Clustering of Applications with Noise

DEM - Digital Elevation Model

ET - Evapotranspiration

FI - Flashiness Index

FDSC - Flow Duration Curve Slope

GRU - Gated Recurrent Unit

GLDAS - Global Land Data Assimilation System

GWS - Groundwater Storage

HFF - High Flow Frequency

HSD - Honestly Significant Difference

HEC-HMS - Hydrologic Engineering Center's Hydrologic Modeling System

HCI - Hydrological Coherence Index

LST - Land Surface Temperature

LSTM - Long Short-Term Memory

LFF - Low Flow Frequency

MAR - Mean Annual Runoff

MICE - Multivariate Imputation by Chained Equations

NSE - Nash Sutcliffe Efficiency Coefficient

NSPD - National Soil Property Database

NDVI - Normalized Difference Vegetation Index

PAM - Partitioning Around Medoids

PUB - Prediction in Ungauged Basins

PC - Principal Component

PCA - Principal Component Analysis

PRE - Precipitation Elasticity of Runoff

PSM - Profile Soil Moisture

Q₅ - 5th Streamflow Percentile (Q5)

Q₉₅ - 95th Streamflow Percentile (Q95)

RC - Recession Constant

RNNs - Recurrent Neural Networks

RLD - Rising Limb Density

RMSE - Root Mean Square Error

RSM - Root Zone Soil Moisture

RR - Runoff ratio

SNDPPT - Sand Content

SLTPPT - Silt Content

SWAT - Soil and Water Assessment Tool

SMA - Soil Moisture Accounting

US - United States

VMCDUL - Volumetric Moisture Content Drainage Upper Limit

VMCWP -Volumetric Moisture Content Wilting Point

WCHI - Within-Cluster Homogeneity Index

WCV - Within-Cluster Variance

WMO - World Meteorological Organization

ZFF - Zero Flow Frequency

CHAPTER 1

1 INTRODUCTION

1.1 Background

All growing pressures on global water resources have intensified concerns regarding long-term water security, as climate-driven changes alter precipitation regimes and heighten the likelihood of droughts and floods in many regions (Vörösmarty et al., 2010). The rise in population and economic development places greater stress on river basins, heightening the need for precise evaluations of runoff patterns to support water distribution, disaster preparedness, and the protection of ecosystems (Sivapalan et al., 2003). However, these requirements remain unfulfilled by the extensive number of basins worldwide in which streamflow is either unmeasured or recorded only intermittently, leaving substantial gaps in understanding catchment response dynamics (Hrachowitz et al., 2013). Addressing this challenge is crucial for protecting communities vulnerable to hydrological extremes and for ensuring that development trajectories align with sustainable water use.

Classical approaches to predicting catchment runoff in ungauged basins have centred on regionalisation, a strategy designed to estimate rainfall-runoff model parameters in basins lacking streamflow observations by transferring information from gauged donor sites. Early methods relied on spatial proximity, assuming that catchments located close to one another share similar hydroclimatic conditions and therefore exhibit comparable hydrological responses (Parajka et al., 2005). Subsequent developments introduced physical similarity frameworks, in which catchments were grouped based on attributes such as slope, drainage density, land cover, and climatic indices, enabling model parameters from structurally similar donors to be applied to target basins (Oudin et al., 2008). Regression-based approaches later emerged, linking calibrated model parameters to physiographic descriptors through statistical relationships. However, the strength of these regressions was often weak and difficult to interpret, as many model

parameters lacked uniqueness (He et al., 2011). Despite these refinements, all prominent regionalisation families face notable limitations. They depend heavily on the availability and quality of catchment descriptors, assume relatively simple relationships in systems governed by nonlinear and threshold-driven processes, and frequently exhibit considerable uncertainty relative to full local calibration, especially in heterogeneous landscapes (Oudin et al., 2008; Parajka et al., 2005). These constraints have motivated a growing shift toward data-rich and process-flexible approaches.

Recent developments in hydrology have been shaped by two significant technological advances that broaden the capacity to observe and model water cycle processes. The first is the maturation of satellite-based remote sensing, which now delivers global measurements of precipitation, evapotranspiration, soil moisture, vegetation dynamics, and terrestrial water storage at temporal frequencies suitable for hydrological analysis. These products collectively extend monitoring capacity into regions where ground observations are sparse or absent (Huang et al., 2025; Levizzani & Cattani, 2019; Papa et al., 2023). The second advancement is the rapid expansion of machine learning techniques in hydrology, driven in part by evidence that deep learning (DL) models extract relationships from large multi-catchment datasets that traditional theory-driven models fail to represent (Nearing et al., 2021). Long Short-Term Memory (LSTM) networks are particularly effective because their architecture reproduces sequential dependencies that reflect storage-driven hydrological behaviour (Kratzert et al., 2018).

Although recent developments in remote sensing and DL broaden the opportunities for hydrological prediction, a central limitation persists when these tools are applied to ungauged settings. Current large-sample datasets, such as the Caravan community resource, demonstrate that hydrological behavior varies markedly across climatic, geological, and land-cover gradients, underscoring the difficulty of assuming that models trained on diverse basins can reliably generalize to unmonitored locations (Kratzert et al., 2023). Machine learning models trained in a global configuration often follow a data-maximisation principle, yet this approach may overlook fundamental differences in catchment functioning that arise from thresholds, storage dynamics, and

landscape controls (Addor et al., 2018). Evidence from DL studies indicates that LSTM networks can transfer information across basins, but their performance remains sensitive to the representativeness of the training catchments, particularly when hydrological regimes diverge (Frame et al., 2022; Kratzert et al., 2018). This raises a critical, unresolved question: whether prediction in ungauged basins is best supported by large, heterogeneous datasets or by focused subsets that capture hydrological similarity. The present research addresses this question by integrating remote sensing attributes to define similarity and evaluating contrasting LSTM data strategies against established regionalisation methods.

1.2 Problem Statement

Reliable prediction of streamflow in ungauged basins remains one of the most persistent constraints in hydrology, limiting the ability to understand runoff behaviour in data-scarce regions. This challenge forms a central component of the Prediction in Ungauged Basins (PUB) agenda, which emphasises the difficulty of making sound hydrological inferences where no gauging records exist (Sivapalan et al., 2003). Although recent advances in data-driven modelling have expanded what can be inferred from large observational datasets, their direct application to ungauged basins raises a strategic uncertainty. DL networks have demonstrated strong predictive skill in gauged environments and can learn nonlinear hydrological responses across diverse basins (Kratzert et al., 2018; Nearing et al., 2021). However, without discharge data at the target site, the question of how to select appropriate training catchments remains unresolved.

The dominant approach pools all available donor catchments into a single global model, assuming that broader data coverage enhances generalisation (Frame et al., 2022). Yet this assumption does not account for the pronounced spatial variability in climate, geology, and land cover that shapes catchment functioning, nor for the difficulty of learning coherent hydrological signals from highly heterogeneous datasets (Kratzert et al., 2018). An alternative perspective suggests that training should be limited to hydrologically similar donors; however, there is no consensus on whether

such a similarity-based approach yields more reliable predictions compared to global configurations.

Consequently, the primary issue lies in the lack of a solid framework to determine whether the hydrological context or dataset size provides a stronger basis for DL prediction in ungauged basins, as well as the absence of scalable methods to define similarity using remotely sensed indicators.

1.3 Aims and Objectives

1.3.1 Main objective

To develop a novel methodology for identifying hydrologically similar catchments by analyzing spatiotemporal hydrometeorological data from gauged catchments, utilizing machine learning techniques, and validating the findings through hydrological modeling.

1.3.2 Specific objectives

- I. To analyze the temporal variation of hydrometeorological data and the spatial variation of catchment properties for clustering hydrologically similar catchments.
- II. To develop a framework by combining statistical regionalization techniques with machine learning models for predicting runoff in ungauged catchments.
- III. To validate the developed runoff prediction methodology for ungauged catchments using hydrological modeling.

1.4 Scope and Limitations

This study investigates streamflow prediction in ungauged catchments using hydrological similarity analysis and machine learning, based on 30 gauged catchments distributed across Sri Lanka that represent wet, intermediate, and dry hydroclimatic regimes. The study domain comprises independent river subbasins without major reservoirs, selected to span a broad range of catchment areas, elevations, and climatic settings relevant to regional hydrological analysis. The temporal scope spans 2005 to

2025, using continuous daily streamflow and precipitation records to ensure consistency in the computation of hydrological signatures and in model evaluation. The thematic focus is on characterising catchment hydrological behaviour using streamflow signatures and on examining the relationships among functional similarity, catchment classification, and predictive performance in ungauged streamflow estimation.

Methodologically, this study employs hydrological signature computation and unsupervised clustering for catchment classification, followed by LSTM neural network modelling to evaluate ungauged streamflow prediction under regional and global training strategies. Data inputs include daily streamflow and precipitation observations from 30 gauging stations, supplemented by satellite-derived land-surface, soil, vegetation, and topographic datasets used exclusively for additional validation. The study examines functional hydrological similarity and data-driven prediction performance. Still, it is limited to daily time-step analysis and does not address sub-daily hydrological processes, event-scale flood modelling, or real-time forecasting applications. The framework relies on observed streamflow and precipitation data and does not apply bias correction or uncertainty quantification. Furthermore, the study evaluates established clustering techniques and machine learning models and does not seek to develop new clustering algorithms or neural network architectures. Catchments influenced by major reservoirs or significant flow regulation are excluded, and interpretation of results is restricted to the analysed catchment set and historical observation period, without extending findings to future climate scenarios or asserting transferability beyond the defined study domain.

1.5 Significance of the Study

This study contributes to hydrological science by advancing understanding of how functional hydrological similarity influences streamflow predictability in ungauged catchments, and by addressing ongoing debate over the relative importance of hydrological context versus data volume in data-driven modelling. The findings provide hydrologists and water resources engineers with a structured framework that integrates catchment classification and machine learning to support ungauged

streamflow prediction, enabling more informed model selection and training strategies in data-scarce environments. For water management agencies and planning authorities, the research offers evidence-based insights to support regionalisation decisions, particularly where streamflow observations are limited but catchment-level planning and water allocation assessments are required. At the societal level, the study supports improved evaluation of water availability and hydrological behaviour in poorly monitored basins, contributing to more reliable water resources planning, enhanced resilience to hydrological extremes, and reduced uncertainty in infrastructure and environmental management decisions. Methodologically, the research demonstrates an integrated approach combining hydrological signatures, unsupervised clustering, satellite-derived validation, and LSTM networks, which can be replicated in other data-limited regions to evaluate ungauged prediction strategies systematically. Ultimately, this work supports the broader goal of strengthening sustainable water resources management by improving the scientific basis for streamflow prediction in ungauged catchments, which remains a fundamental challenge in hydrology.

1.6 Structure of the Thesis

This thesis is structured into five chapters focused on hydrological similarity analysis and ungauged streamflow prediction. Chapter One introduces the research background, problem statement, objectives, and thesis structure. Chapter Two reviews literature on runoff prediction, catchment classification, hydrological signatures, clustering methods, and regionalisation techniques. Chapter Three details the study area, data sources, quality control, hydrological signature computation, clustering framework, validation methodology, and the creation of streamflow prediction models. Chapter Four presents and discusses the results, including signature selection, catchment classification, validation using satellite data, clustering performance, and model evaluations. Finally, Chapter Five summarizes key findings and offers recommendations for future research and practical applications in ungauged catchment hydrology.

CHAPTER 2

2 LITERATURE REVIEW

This chapter presents a review of the scientific literature that forms the foundation for this research, structured to narrate the progression from the fundamental challenge of PUB to the innovative methodology employed. It begins with an examination of traditional regionalization methods, highlighting their limitations, and shifts to the paradigm of hydrological signatures as indicators of catchment functionality and their role in catchment classification. The analysis then addresses the geospatial revolution, emphasizing the value of satellite remote sensing in characterizing basins lacking ground measurements. It further investigates the emergence of data-driven modeling, particularly LSTM networks in rainfall-runoff modeling, and evaluates the strategic dilemma surrounding global versus localized data for LSTM regionalization.

2.1 Runoff Prediction

2.1.1 Traditional regionalization methods

Hydrological regionalisation forms the classical foundation for addressing the PUB problem, offering a structured means of transferring information from gauged to ungauged catchments. These regionalisation approaches are typically grouped into three methodological families, each relying on a different interpretation of similarity between catchments (Hrachowitz et al., 2013). Spatial proximity methods, among the earliest applied techniques, assume that neighbouring basins experience broadly comparable climatic and physiographic influences, allowing parameter sets or observed runoff characteristics from a nearby gauged site to serve as proxies for an ungauged one. This rationale remains evident in large-sample assessments, where nearest-neighbour selection has repeatedly been tested as a pragmatic baseline strategy (Oudin et al., 2008). A second group of methods applies physical similarity principles by linking catchments through quantifiable descriptors such as soils, geology, vegetation, and climatic regimes. These approaches, which featured prominently in the PUB decade, treat basin form as a surrogate for basin function, reflecting the long-

standing idea that hydrological response emerges from the integrated influence of landscape properties (He et al., 2011; Wagener et al., 2007). The third category encompasses parameter regression, in which calibrated model parameters from gauged basins are statistically related to catchment attributes, enabling direct parameter estimation for ungauged sites (Parajka et al., 2005). Studies comparing these three families repeatedly highlight their shared objective: to infer hydrological behaviour using transferable information while compensating for the absence of streamflow records (Hrachowitz et al., 2013). Although these approaches provided the initial scientific framework for PUB, the diversity of hydrological environments has revealed varying degrees of success across regions, setting the stage for a closer examination of their known limitations in contemporary practice.

The effectiveness of established regionalization approaches in hydrology has limitations that affect their reliability in various contexts. Spatial proximity methods rely on the principle that neighbouring basins exhibit similar behaviour. However, this assumption often fails due to geographical differences, changes in land use, and variations in climate. These factors can cause sudden changes in runoff response, making proximity a poor indicator of similarity in many cases (Arsenault & Brissette, 2014; Razavi & Coulibaly, 2013). On the other hand, physical similarity methods seem more defensible in theory. Yet, they are hampered by the lack of reliable catchment descriptors, which can be coarse, inconsistent, or even missing in some areas. Research indicates that incomplete or low-resolution attribute sets diminish the effectiveness of similarity measures, undermining the stability of transferred parameter sets and ultimately limiting predictive skill (Arsenault & Brissette, 2014; He et al., 2011). Parameter regression approaches add another layer of complexity. Hydrological models often show equifinality, meaning that different parameter sets can produce similarly acceptable results. This non-uniqueness makes regression relationships challenging to interpret and reduces confidence in any specific transferred parameter set (Beven, 2006). Comparisons of traditional methods reveal that prediction success rates seldom exceed moderate levels, with significant variability across different hydroclimatic regions (Razavi & Coulibaly, 2013). Furthermore, the structural uncertainties in both conceptual and physically based models complicate matters.

Issues such as over-parameterization, scale mismatches, and high data requirements often render these models unsuitable for ungauged areas, particularly where observation networks are sparse (Razavi & Coulibaly, 2013). Collectively, these shortcomings highlight a significant limitation: traditional methods struggle to capture the integrated, dynamic processes that influence catchment responses. Recognizing this has led to a growing interest in data-driven approaches that can learn hydrological behavior directly from observations.

2.1.2 Deep learning methods

DL has brought forward a wide range of modeling strategies for predicting runoff. These methods aim to represent the complex hydrological processes involved accurately. Initially, researchers focused on Artificial Neural Networks (ANNs). These networks could learn nonlinear relationships between rainfall and runoff. However, they struggled with the sequential dependencies that are typical in hydrological systems. This is mainly because their feed-forward structures do not retain any historical information (Patil et al., 2024). Recurrent Neural Networks (RNNs) were developed to overcome the shortcomings of conventional neural networks by concentrating on sequential data. However, standard RNNs continued to struggle with issues related to vanishing gradients, which hindered their ability to model long-term dependencies effectively (Li et al., 2024). As an alternative, Convolutional Neural Networks (CNNs) emerged. These networks excel at capturing spatial patterns within gridded rainfall and physiographic data. Further advancements led to the development of the Gated Recurrent Unit (GRU). This architecture simplifies recurrent modeling by using fewer parameters while maintaining comparable performance for short-term runoff predictions, all with greater computational efficiency (Gao et al., 2020). In recent times, the Transformer architecture has received increasing attention. Its strength lies in establishing long-range relationships within meteorological time series through its unique attention mechanisms. However, hydrological studies indicate that its benefits often depend on the context and typically require substantial datasets to surpass the performance of recurrent models (Liu et al., 2024). Against this landscape of modelling options, the LSTM network has emerged as the most consistently validated architecture for rainfall-runoff prediction. Its formulation was designed to

address the vanishing-gradient limitations that affect basic recurrent structures, introducing a series of gating operations that regulate how information is stored and discarded across time. These mechanisms make LSTM particularly effective at representing hydrological processes governed by delayed or cumulative effects, including soil moisture memory, antecedent wetness, and storage–release dynamics (Kratzert et al., 2018). Evidence from large-sample hydrological studies confirms that LSTM models frequently meet or exceed the accuracy achieved by a wide range of conceptual models, particularly under conditions where rainfall–runoff relationships exhibit strong non-linearity and long-range temporal dependencies (Frame et al., 2022; Nearing et al., 2021). Although GRU models offer computational advantages and have demonstrated similar short-term forecasting accuracy, their reduced parameterisation may limit their ability to represent longer hydrological memory, particularly in catchments where slow-responding stores exert a dominant influence on runoff generation (Gao et al., 2020). Transformers, despite their success in other fields, remain less reliable in hydrology because their performance is sensitive to the size of the training data and often fails to surpass that of recurrent models in daily runoff simulations (Liu et al., 2022). These findings collectively support the selection of LSTM for this study, as its empirical track record, structural ability to model state-dependent processes, and broad validation across hydrological regimes align closely with the requirements of prediction in ungauged basins.

Within the growing portfolio of DL architectures applied to rainfall-runoff modelling, the LSTM network has become the most frequently adopted approach, largely because its internal gating structure enables the representation of long-range dependencies that govern hydrological response. Its formulation was developed to address the vanishing-gradient limitations observed in earlier recurrent architectures, a feature that enables the model to retain information about antecedent conditions, such as soil moisture, cumulative storage, and delayed runoff contributions (Kratzert et al., 2018). Empirical evaluations across large-sample benchmark datasets demonstrate that LSTM models consistently deliver strong performance and, in many instances, achieve accuracy comparable to or exceeding that of conceptual hydrological models, particularly when rainfall-runoff relationships exhibit marked non-linearity (Frame et al., 2022; Nearing

et al., 2021). Studies applying LSTM in a range of hydroclimatic settings further support its versatility and reliability for daily streamflow simulation, reinforcing its position as one of the most thoroughly validated DL architectures in hydrology (Gao et al., 2020). Although no single machine learning model can be considered universally superior, the breadth of evidence available for LSTM provides a defensible foundation for its use in this research, offering a robust baseline from which data-driven prediction in ungauged basins can be explored.

The use of LSTM networks in rainfall–runoff modeling presents a key methodological dilemma: is it better to train the model using a single combined dataset, or do regionalized approaches provide a more accurate representation of hydrological variation? Large-sample studies have shown that pooled LSTM models benefit from extensive cross-catchment information, allowing the network to learn general hydrological behaviours from hundreds of basins simultaneously (Kratzert et al., 2019). Although this establishes the value of broad training domains, several investigations reveal that hydrological processes exhibit pronounced spatial structure. Analyses of national-scale LSTM post-processing demonstrate that predictive skill varies considerably across hydro-regions, suggesting that model performance is influenced by regional hydrological distinctiveness (Frame et al., 2021). Complementary evidence from regime-characteristics research shows that catchments with similar flow regimes cluster consistently, indicating that hydrological similarity carries meaningful predictive information (Sawicz et al., 2011). Furthermore, recent LSTM-based regional modeling experiments demonstrate that learned embedding layers naturally group basins with similar rainfall–runoff behavior, providing mechanistic support for regionalized training strategies (Kratzert et al., 2019). Collectively, these findings suggest that a universal LSTM may not fully capture the nuanced behavior of underrepresented catchments. This reinforces the need for a systematic, data-driven catchment classification framework capable of identifying appropriate donor basins for regionalised LSTM models.

2.2 Catchment Classification

Catchment classification provides the foundational framework for hydrological regionalization by systematically grouping basins based on shared characteristics, thereby enabling the transfer of information. Traditional classification schemes have primarily relied on two types of attributes: static physical descriptors (e.g., soil type, geology, topography) and climatic indices (e.g., mean annual precipitation, aridity index) (McDonnell & Woods, 2004; Wagener et al., 2007). While these static methods offer simplicity, they possess a critical limitation: they classify catchments based on their structure rather than their function. A catchment with specific soils and climate can exhibit vastly different hydrological behaviors due to factors like subsurface heterogeneity and vegetation dynamics. Consequently, the paradigm has progressively shifted towards classification based on hydrological signatures, quantitative metrics derived directly from the streamflow hydrograph that represent integrated catchment response (Olden & Poff, 2003; Sawicz et al., 2011). Signatures such as the Baseflow Index (BFI) and the Flow Duration Curve (FDC) slope directly encapsulate the dynamic outcome of rainfall-runoff processes, including water partitioning, storage dynamics, and event responsiveness. This approach, often termed functional classification, ensures that grouped catchments not only look similar on a map but actually behave similarly in a hydrological context, providing a more process-rich basis for regionalization (Addor et al., 2018).

2.2.1 Hydrological signatures

The efficacy of functional classification using hydrological signatures is well established in the literature. Sawicz et al. (2011) successfully classified over 300 catchments across the U.S. using a suite of signatures, revealing distinct regional patterns in hydrological behavior tied to climate and landscape controls. Similarly, Addor et al. (2018) demonstrated that signature-based groups provided a more robust foundation for regionalization than groupings based solely on climate or physical attributes. These studies, among others, confirm that signatures distil complex hydrological processes into comparable, quantitative metrics. Building upon this proven framework, this analysis initially identified a broad set of fourteen hydrological

signatures, deliberately chosen from the literature to collectively represent the major components of the flow regime. These metrics span over critical aspects, including water balance, flow variability, event responsiveness, and baseflow characteristics. The following section provides a detailed exposition of each selected signature, its formulation, and the specific facet of catchment function it is designed to quantify.

2.2.1.1 *Runoff ratio (RR)*

The *RR* quantifies the catchment's overall water yield efficiency by representing the proportion of total precipitation that is converted into streamflow (as shown in [2-1]). This signature integrates all water loss processes and reflects the catchment's fundamental water partitioning behavior (McMillan, 2021; Sawicz et al., 2011; Yadav et al., 2007).

$$RR = \frac{\sum_{t=1}^T q_t}{\sum_{t=1}^T P_t} \quad [2-1]$$

According to this definition,

q_t = streamflow at time t (mm/day),

P_t = precipitation at time t (mm/day),

T = total timesteps.

2.2.1.2 *Mean annual runoff (MAR)*

MAR represents the catchment's long-term water production potential and serves as a fundamental measure of water availability (as shown in [2-2]). This signature integrates climatic and landscape controls on hydrological response over interannual timescales (McMillan, 2021; Richter et al., 1996; Yadav et al., 2007).

$$MAR = \frac{1}{N} \sum_{y=1}^N \sum_{d=1}^{D_y} q_{y,d} \quad [2-2]$$

According to this definition,

N = number of years,

D_y = days in year,

$q_{y,d}$ = daily streamflow (mm/day).

2.2.1.3 5th streamflow percentile (Q_5)

The Q_5 defines the low-flow magnitude that is exceeded 95% of the time, providing a standardized measure for drought severity and aquatic habitat suitability (as shown in [2-3]). This signature represents a critical lower boundary of the flow regime that is resilient to extreme outliers in the data (Addor et al., 2018; McMillan, 2021).

$$Q_5 = \text{percentile}(Q, 5) \quad [2-3]$$

According to this definition,

$$Q = \text{time series of streamflow (mm/day)}.$$

2.2.1.4 95th streamflow percentile (Q_{95})

The Q_{95} indicates the high-flow magnitude that is exceeded only 5% of the time, characterizing the upper range of common flow events before extreme flooding (as shown in [2-4]). It is a key indicator of the system's energy and its capacity for sediment transport (Addor et al., 2018; McMillan, 2021).

$$Q_{95} = \text{percentile}(Q, 95) \quad [2-4]$$

According to this definition,

$$Q = \text{time series of streamflow (mm/day)}.$$

2.2.1.5 Precipitation elasticity of runoff (PRE)

The PRE measures the sensitivity of streamflow to interannual precipitation variations, indicating how catchment response amplifies or dampens climatic fluctuations (as shown in [2-5]). High elasticity values suggest strong climate sensitivity and potential vulnerability to drought conditions (McMillan, 2021; Sankarasubramanian et al., 2001).

$$PRE = \frac{dQ}{dP} \frac{P}{Q} \quad [2-5]$$

According to this definition,

$$Q = \text{annual streamflow (mm/year)},$$

$$P = \text{annual precipitation (mm/year)}.$$

2.2.1.6 Flashiness index (FI)

The *FI* characterizes the rapidity of streamflow response to rainfall inputs by measuring day-to-day flow variability (as shown in [2-6]). High flashiness indicates quick water transmission through the catchment, typically associated with limited storage capacity (Baker et al., 2004).

$$FI = \frac{\sum_{t=1}^{T-1} |q_{t+1} - q_t|}{\sum_{t=1}^T q_t} \quad [2-6]$$

According to this definition,

q_t = streamflow at time t (mm/day),

T = total timesteps.

2.2.1.7 Flow duration curve slope (FDCS)

The *FDCS* quantifies flow variability by measuring the steepness of the flow duration curve between intermediate exceedance probabilities (as shown in [2-7]). Steeper slopes indicate greater flow variability and flashier response characteristics (McMillan, 2021; Sawicz et al., 2011; Yokoo & Sivapalan, 2011).

$$FDCS = \frac{\ln(Q_{33}) - \ln(Q_{66})}{0.33} \quad [2-7]$$

According to this definition,

Q_{33}, Q_{66} = 33rd and 66th flow percentiles.

2.2.1.8 Rising limb density (RLD)

The *RLD* quantifies the frequency of hydrograph rising limbs, representing how often the catchment generates quick-flow responses to rainfall events (as shown in [2-8]). Higher densities indicate more frequent storm responses and shorter time of concentration (Sawicz et al., 2011; Shamir et al., 2005).

$$RLD = \frac{N_{\text{rising limbs}}}{N_{\text{rising days}}} \quad [2-8]$$

According to this definition,

$N_{\text{rising limbs}}$ = number of rising limbs,

$N_{\text{rising days}}$ = total number of days of the rising.

2.2.1.9 Low flow frequency (LFF)

The *LFF* quantifies the frequency of flows falling below a hydrologically significant low threshold, providing a direct measure of drought occurrence (as shown in [2-9]). Using a dynamic threshold based on the median ensures that this frequency reflects catchment-specific dryness (Addor et al., 2018; McMillan, 2021).

$$LFF = \frac{\sum_{t=1}^T I(q_t < 0.2\tilde{Q})}{T} \quad [2-9]$$

According to this definition,

I = indicator function (1 if true, zero otherwise),

q_t = streamflow at time t (mm/day),

$\tilde{Q}$ = median streamflow (mm/day),

T = total timesteps.

2.2.1.10 High flow frequency (HFF)

The *HFF* measures how often flows exceed a threshold signifying a pronounced high-flow event (as shown in [2-10]). Using a dynamic threshold based on the median ensures that this frequency reflects catchment-specific wetness (Addor et al., 2018; McMillan, 2021).

$$HFF = \frac{\sum_{t=1}^T I(q_t > 2\tilde{Q})}{T} \quad [2-10]$$

According to this definition,

I = indicator function (1 if true, zero otherwise),

q_t = streamflow at time t (mm/day),

$\tilde{Q}$ = median streamflow (mm/day),

T = total timesteps.

2.2.1.11 Zero flow frequency (ZFF)

The *ZFF* calculates the proportion of time the channel experiences no measurable flow, a critical signature for classifying ephemeral and intermittent streams (as shown in [2-11]) (Addor et al., 2018; McMillan, 2021).

$$ZFF = \frac{\sum_{t=1}^T I(q_t = 0)}{T} \quad [2-11]$$

According to this definition,

I = indicator function (1 if true, zero otherwise),

q_t = streamflow at time t (mm/day),

T = total timesteps.

2.2.1.12 Coefficient of variation (CoV)

The *CoV* standardizes the variability of the streamflow time series by its mean, representing the relative spread of flows around the average condition (as shown in [2-12]). A high *CoV* indicates a flashy, unpredictable regime, while a low *CoV* suggests stable, regulated flows (McMahon et al., 2007; Richter et al., 1996).

$$CoV = \frac{\sigma Q}{\mu Q} \quad [2-12]$$

According to this definition,

σQ = standard deviation of streamflow (mm/day),

μQ = mean streamflow (mm/day).

2.2.1.13 Recession constant (RC)

The *RC* quantifies the rate of streamflow decline during rain-free periods, reflecting catchment drainage efficiency and groundwater storage characteristics (as shown in [2-13]). Lower values indicate slower drainage and a greater water retention capacity (Tallaksen, 1995).

$$q_t = q_0 \cdot e^{-k_{rec} \cdot t} \quad [2-13]$$

According to this definition,

q_t = streamflow at time t (mm/day),

q_0 = initial flow streamflow (mm/day),

k_{rec} = recession constant (1/day).

2.2.1.14 Baseflow index (BFI)

The *BFI* represents the long-term proportion of streamflow derived from groundwater and other delayed sources, indicating the catchment's baseflow sustainability capacity (as shown in [2-14]). Higher values suggest strong groundwater connectivity and stable flow regimes (Eckhardt, 2005; Nathan & McMahon, 1990).

$$BFI = \frac{\sum_{t=1}^T b_t}{\sum_{t=1}^T q_t} \quad [2-14]$$

According to this definition,

b_t = baseflow at time t (mm/day),

q_t = streamflow at time t (mm/day),

T = total timesteps.

Following the definition of a comprehensive suite of signatures, a critical subsequent step involves addressing the inherent statistical challenge of multicollinearity. Hydrological signatures are rarely independent. They often quantify overlapping aspects of catchment function, leading to strong intercorrelations. Employing a full set of correlated variables in any subsequent multivariate application is statistically problematic, as it unfairly weights the analysis towards the shared underlying process, thereby distorting the true, multi-faceted nature of hydrological similarity (Olden & Poff, 2003). Several statistical methods are available for this purpose. Hydrological literature presents a spectrum of methods for this purpose. Principal Component Analysis (PCA) is a common technique that transforms the original correlated variables into a new set of uncorrelated principal components, capturing the maximum variance. Although these components can become abstract hybrids of processes, making hydrological interpretation challenging (Jolliffe & Cadima, 2016; Olden & Poff, 2003). As an alternative, feature selection methods like Minimum Redundancy Maximum Relevance (mRMR) aim to identify an optimal subset of original signatures that are maximally relevant to a target variable while being minimally redundant with each other (Peng et al., 2005). Some studies also employ clustering of the signatures themselves to group correlated indices and then select a representative from each cluster (Addor et al., 2018). The choice among these techniques involves a fundamental trade-off between statistical efficiency and the preservation of physical interpretability. Methods like PCA, while statistically powerful, often obfuscate the direct physical meaning of the original signatures, which is a critical requirement for a process-based catchment classification (Wagener et al., 2007).

Given the need for interpretable results, a simple correlation-based filtering approach was selected for this study over more complex transformations. First, all signatures

were standardized to bring them within a common range (as shown in [2-15]). Then, a pairwise Pearson correlation matrix was calculated to identify clusters where the coefficients exceeded a predefined threshold (as shown in [2-16]), a well-established benchmark for defining hydrological redundancy (Olden & Poff, 2003). This straightforward method offers a key advantage: transparency. Unlike PCA, which creates composite axes, this method preserves the original, physically meaningful signature variables. The final output is a "minimum viable signature set" that avoids statistical redundancy while maintaining a direct, intuitive connection to catchment response dynamics, providing a robust and interpretable foundation for subsequent classification.

$$Z_i = \frac{X_i - \mu_X}{\sigma_X} \quad [2-15]$$

According to this definition,

Z_i = standardized value for catchment i ,

X_i = raw signature value for catchment i ,

μ_X = mean of the signature across all catchments,

σ_X = standard deviation of the signature across all catchments.

$$r_{ij} = \frac{\sum_{k=1}^n (Z_{i,k} - \bar{Z}_i)(Z_{j,k} - \bar{Z}_j)}{\sqrt{\sum_{k=1}^n (Z_{i,k} - \bar{Z}_i)^2 \sum_{k=1}^n (Z_{j,k} - \bar{Z}_j)^2}} \quad [2-16]$$

According to this definition,

Z_i, Z_j = two standardized signatures.

After identifying redundant signature pairs, the next critical step is to choose a single representative from each pair. A hierarchy should inform this choice of criteria grounded in hydrological principles, with hydrological interpretability taking precedence. The selected signature must possess clear physical meaning, reflecting the essential features of the unit hydrograph, which is fundamental for comparing catchment response functions (Sawicz et al., 2011). The aim is to compile a final set of signatures that collectively capture the core elements of the unit hydrograph, which is characterized by its distinct rising limb, peak, recession limb, and baseflow

components (Dooge, 1959). While it is recognized that a collection of indices cannot entirely reconstruct a hydrograph, prioritizing signatures that represent these different phases helps create a more meaningful representation of the data. For instance, a signature that indicates the speed of the rise should be chosen for the rising limb, while a signature that measures flow depletion rates is vital for understanding the recession phase. This method ensures that the final selection of signatures transcends mere statistical representation and encapsulates the minimum necessary components of catchment response, establishing a coherent basis for comparison (Yadav et al., 2007).

2.2.2 Clustering methods for classification

Following the selection of a non-redundant signature set, the subsequent task is to delineate hydrologically coherent groups of catchments using an objective and transparent clustering framework. Since there are no predefined catchment labels exist for supervised learning, unsupervised methods remain the primary analytical route for exploratory regionalisation. Foundational texts in data mining describe several major families of clustering algorithms: centroid-based techniques such as K-Means, which partition observations around cluster means; density-based approaches like Density-based spatial clustering of applications with noise (DBSCAN), which identify groups of arbitrary shape; model-based or distributional algorithms that fit probabilistic structures; and hierarchical methods that construct tree-like clustering architectures (Tan et al., 2019). These approaches have been widely adopted in hydrology, where multiple studies demonstrate how methodological choice affects the resulting catchment groups and their hydrological interpretation. For example, Sharghi et al. (2018) applied four distinct clustering methods, including K-Means, Ward's method, and self-organising maps, and found that each produced different optimal cluster configurations, underscoring the sensitivity of regionalisation to algorithmic assumptions. Similarly, Modiri and Bárdossy (2021) demonstrated that hierarchical clustering combined with scaling techniques yielded spatially coherent but method-dependent groups, thereby reinforcing the value of comparative evaluation, concluding that no single clustering method universally outperforms others. Given this evidence, and to prioritise interpretability over algorithmic novelty, the present study employs a

suite of established and transparent clustering approaches to ensure robust, reproducible, and hydrologically meaningful classification outcomes.

2.2.2.1 K-Means clustering

The K-Means algorithm has become a widely used tool in hydrological classification because it offers a computationally efficient and conceptually straightforward method for partitioning catchments into groups with similar characteristics. Its iterative procedure minimises within-cluster variance, producing compact clusters around central centroids, which makes it well-suited for large hydrological datasets. Studies applying unsupervised approaches to catchment similarity frequently include K-Means as a core method. For example, Ghaneei et al. (2024) emphasise that hydrological classification studies routinely consider K-Means, hierarchical clustering, and density-based algorithms, reflecting their established role in regionalisation frameworks. Similarly, Sharghi et al. (2018) applied K-Means, alongside other clustering techniques, to demonstrate that method selection influences the resulting catchment groupings, underscoring its importance in comparative assessments. Although the method assumes roughly spherical clusters and is sensitive to initial centroid placement, its balance of simplicity, interpretability, and performance justifies its inclusion. Consequently, K-Means is adopted as one of the clustering techniques in this study.

2.2.2.2 K-Medoids clustering

K-Medoids clustering offers a robust alternative to centroid-based methods by selecting actual observations as cluster centres, which greatly enhances resilience against outliers and noisy hydrological measurements. This makes it attractive for catchment classification, where signature datasets often contain anomalous values arising from extreme events or measurement artefacts. The Partitioning Around Medoids (PAM) algorithm is the most widely used implementation; it minimises total dissimilarity within clusters and accommodates a range of distance metrics, allowing flexibility in hydrological applications (Madhulatha, 2011). Its use in water-related studies is well documented. Wu et al. (2022) applied the K-Medoids algorithm to Ugandan inflow data and reported that it “showed strong robustness to outliers

compared with K-Means,” demonstrating its value for heterogeneous hydrological datasets. Zhang et al. (2025) included K-Medoids alongside hierarchical methods when classifying over 1400 flood events, emphasising its suitability within partitional clustering frameworks. Given these strengths, K-Medoids is adopted in this study as part of the multi-algorithm classification strategy.

2.2.2.3 Hierarchical clustering

Hierarchical clustering plays a central role in hydrological regionalisation because it provides a multiscale view of similarity that is especially valuable for complex catchment behaviour. Unlike partitional methods, it does not require the analyst to predefine the number of clusters, allowing the full nested structure of hydrological relationships to emerge through the dendrogram. This property has made agglomerative approaches particularly common in hydrology. For example, Brunner et al. (2020) applied hierarchical clustering to 671 US catchments, demonstrating that the resulting regime classes were coherent not only in their hydrograph shapes but also in their drought, flood, and physiographic attributes, highlighting their capacity to reveal meaningful functional groupings. Chiverton et al. (2015) also employed hierarchical clustering to group UK benchmark basins based on temporal dependence patterns, identifying four distinct catchment types with clear geological and soil controls. Similarly, Bajracharya and Jain (2021) employed hierarchical clustering to classify catchments based on width-function characteristics, demonstrating its suitability for diverse hydrological datasets. Due to its interpretability and ability to reveal hydrological structure at multiple scales, hierarchical clustering is included as a key method in this classification framework.

The performance and interpretation of any clustering algorithm are fundamentally governed by the chosen distance metric, which quantitatively defines the concept of "similarity" between catchments. Different metrics capture distinct aspects of hydrological resemblance, leading to potentially divergent groupings from the same signature data.

2.2.2.4 Euclidean distance

As a foundational metric, Euclidean distance measures the straight-line separation between two points in a multidimensional signature space. It calculates the root of squared differences across all signature dimensions (as shown in [2-17]), making it particularly sensitive to magnitude differences between catchments (Tan et al., 2019). This characteristic proves valuable when absolute differences in signature values carry hydrological significance, such as when distinguishing between high-yield and low-yield catchments.

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad [2-17]$$

Its use is well established in hydrology: Oppel and Schumann (2020) applied Euclidean distance in their machine learning based catchment grouping, noting its suitability for representing differences in catchment properties. Similarly, Sharghi et al. (2018) employed Euclidean distance as the core metric in their comparison of K-Means with Self Organizing Map clustering, highlighting that standard clustering algorithms “commonly rely on Euclidean distance” when grouping hydro-climatological time-series. Given its prevalence, transparency, and compatibility with most clustering algorithms, Euclidean distance is adopted in this study as a baseline metric within the multi-distance similarity framework.

2.2.2.5 Cosine distance

Cosine distance provides a complementary perspective by measuring the angular separation between catchment signature vectors, effectively comparing their directional alignment rather than their absolute magnitudes (as shown in [2-18]). The metric calculates similarity based on the pattern of hydrological behavior, making it insensitive to the scale of streamflow (Tan et al., 2019). This characteristic proves particularly valuable when comparing catchments of different sizes but similar functional responses, as it emphasizes the relative importance of different signatures within each catchment's hydrological profile.

$$d(x, y) = 1 - \frac{x \cdot y}{\|x\| \|y\|} \quad [2-18]$$

In hydrological applications, this approach has revealed patterns that magnitude-sensitive metrics might obscure. In hydrological applications, this angle-based comparison highlights patterns that magnitude-sensitive metrics might overlook. For example, Oppel and Schumann (2020) demonstrated that machine-learning analyses of runoff dynamics benefit from metrics that can capture functional similarities rather than absolute values, noting that pattern-based descriptors can better reveal the dominant controls. By integrating cosine distance into this framework, it ensures that the classification system recognises functional hydrological similarities that extend beyond simple magnitude-based comparisons.

2.2.2.6 Manhattan distance

Manhattan distance provides an alternative geometric perspective by summing the absolute differences between signature values across all dimensions (as shown in [2-19]). This grid-based measurement method is particularly useful for handling datasets that display outliers or include noisy signature values, as it reduces the influence of extreme values compared to the squared terms used in Euclidean distance. (Tan et al., 2019).

$$d(x, y) = \sum_{i=1}^n |x_i - y_i| \quad [2-19]$$

In hydrological contexts, this robustness makes Manhattan distance valuable for regionalization studies where certain catchments may exhibit anomalous behavior due to measurement errors or unique local conditions. Rao and Srinivas (2008) explicitly identify the Manhattan (city-block) metric as a standard distance measure in regionalization-oriented clustering, noting that it is widely used due to its simplicity and interpretability. Likewise, Xulu et al. (2019) evaluate the Manhattan distance in their drought-response clustering and demonstrate that it performs competitively, achieving an accuracy of 85.9%, which highlights its practical effectiveness for detecting environmental similarity. By incorporating Manhattan distance alongside Euclidean and cosine measures, this framework captures complementary aspects of catchment dissimilarity while remaining resilient to data anomalies.

Beyond the choice of distance metrics, in hierarchical clustering, the linkage criterion governs how these pairwise distances are aggregated to measure similarity between

emerging groups of catchments. Therefore, the choice of linkage criterion has a significant influence on the structure of the cluster tree. Ward's linkage, closely linked with Euclidean distance, minimizes the total within-cluster sum of squares, resulting in compact groups that reflect Euclidean geometry (Murtagh & Legendre, 2014). Studies in hydrology have shown that this approach yields stable catchment clusters, particularly with hydro-climatic datasets (Jehn et al., 2020). For non-Euclidean measures, such as cosine and Manhattan distances, average linkage calculates the mean dissimilarity between cluster members. This method strikes a balance between the chaining tendency of single linkage and the compactness of complete linkage, making it suitable for hydrological classifications (Bu et al., 2020). By pairing Ward's method with Euclidean distance and average linkage with other measures, the classification enhances the robustness of hydrological clusters.

2.2.2.7 *Quality of the clusters: Silhouette Score (SS)*

Following cluster formation, a critical next step involves quantitatively assessing the quality of the resulting groupings. A primary objective is to achieve a structure where catchments within the same cluster demonstrate high similarity (cohesion), while being distinctly different from those in other clusters (separation). Internal validation indices provide quantitative measures to assess this balance. The Silhouette Score serves as a prominent metric, calculating for each catchment how similar it is to its own cluster compared to other clusters (as shown in [2-20]), ranging from -1 to 1 (Shahapure & Nicholas, 2020).

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad [2-20]$$

According to this definition,

$a(i)$ = mean intra-cluster distance for catchment i ,

$b(i)$ = mean nearest-cluster distance for catchment i .

Its applicability in hydrology is well established, as demonstrated in recent regionalisation studies that use silhouette values to evaluate the compactness and separation of catchment groups (Narbondo et al., 2020). Further support comes from signature-based hydrological clustering, where the silhouette index is routinely applied to assess the internal consistency of generated groups (Chérif & Gargouri-Ellouze,

2023). A similar application of the metric is observed in environmental water-related analyses, where it facilitates the evaluation of clustering outputs (Liu et al., 2024). The broad and consistent adoption of the silhouette score across these hydrological and environmental studies provides strong justification for its implementation in this analytical framework.

2.2.2.8 *Quality of the clusters: Davies-Bouldin Index (DBI)*

The DBI offers a complementary perspective to silhouette analysis by evaluating the ratio between within-cluster dispersion and the separation from the most similar neighbouring cluster (as shown in [2-21]), where lower values signify clearer and more compact groups. Its formulation, grounded in comparing cluster scatter against inter-cluster distances, provides an efficient global indicator of structural quality and has been widely adopted across clustering research (Davies & Bouldin, 1979).

$$DBI = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right) \quad [2-21]$$

According to this definition,

σ_i = average distance of points in cluster i to the centroid c_i ,

$d(c_i, c_j)$ = distance between centroids.

Hydrological applications also extensively utilize this metric. Several studies implementing cluster-based regionalization frameworks have incorporated DBI to assess the adequacy of hydrological grouping, noting its ability to reflect how tightly catchments cluster around representative hydrological behaviours. For example, applications in water quality and streamflow regionalization have utilized DBI to quantify overall cluster compactness and separation (An et al., 2016; Ghaneei et al., 2024). The repeated adoption of this index across hydrology and environmental clustering research provides strong justification for its inclusion in this framework as a complementary measure of cluster validity.

2.2.2.9 *Statistical significance testing*

To quantitatively validate that the identified catchment clusters represent statistically distinct hydrological regimes, the one-way Analysis of Variance (ANOVA) with

Tukey's Honestly Significant Difference (HSD) post-hoc analysis is employed (Fisher, 1992; Tukey, 1949). This approach tests whether significant differences exist in the means of hydrological signatures across clusters, thereby addressing the null hypothesis that all cluster means are equal. The ANOVA F-statistic is computed as shown in [2-22].

$$F = \frac{MSB}{MSW} \quad [2-22]$$

According to this definition,

MSB = between-groups mean squares,

MSW = within-group means squares.

When ANOVA reveals significant differences ($\alpha = 0.05$), Tukey's HSD identifies specific cluster pairs exhibiting significant differences through the test statistic, as follows [2-23].

$$q = \frac{(\bar{y}_i - \bar{y}_j)}{\sqrt{\frac{MSW}{n}}} \quad [2-23]$$

According to this definition,

$\bar{y}_i, \bar{y}_j$ = cluster means,

MSW = within-group mean squares,

n = cluster size.

This approach has been successfully applied in hydrological classification studies (Ahmadireskety et al., 2021; Bakure et al., 2020) to ensure statistically identified groups represent meaningful hydrological distinctions.

2.3 Classification Validation Framework

Internal validation indices, such as the Silhouette Score and DBI, provide vital initial diagnostics for assessing cluster compactness and separation. However, their primary limitation lies in evaluating clustering results based on the same dataset used to form the groups, leading to circular reasoning that does not confirm whether the resulting clusters represent significant hydrological entities. As a result, hydrological science is increasingly calling for additional validation methods that assess cluster robustness

against independent data sources. One prevalent strategy is to analyze whether clusters display distinctive features linked to known physical catchment attributes excluded from the clustering process, such as geological formations or dominant processes of runoff generation. For example, Addor et al. (2018) successfully validated their clusters by demonstrating that they aligned with established regional patterns of streamflow. Additionally, Knoben et al. (2018) emphasized the importance of ensuring that clusters are interpretable in terms of processes, which they achieved by examining the consistency of climate and landscape characteristics within their grouped data. This movement towards process-informed additional validation is vital for transitioning from merely statistical groupings to classifications that are truly meaningful from a hydrological perspective.

2.3.1 Remote sensing data for validation

The emergence of rich remote sensing data products offers a powerful, independent pathway for validating hydrological clusters. This method tests the core hypothesis that catchments grouped as functionally similar should also exhibit homogeneity in their remotely sensed biophysical and geophysical characteristics, which represent the very drivers of hydrological behavior. As noted by Addor et al. (2018), catchment classifications require validation against independent physical data to ensure their hydrological relevance. Satellite-derived products, including soil moisture, land surface temperature, vegetation indices, and topographic metrics, offer continuous spatial datasets that serve as independent proxies for processes influencing catchment behavior. Beck et al. (2017) demonstrated that remote sensing products can effectively characterize spatial patterns in hydrologically relevant variables across different scales. The strength of this approach lies in its ability to provide synoptic, spatially continuous validation data that remain independent of the in-situ streamflow measurements used for initial cluster development. The following table (Table 2-1) presents selected remote sensing datasets, particularly suitable for this validation approach within the Sri Lankan context.

Table 2-1: Applications of Satellite platforms in Sri Lanka

Satellite/Platform	Applications in the Sri Lankan Context
Moderate Resolution Imaging Spectroradiometer (MODIS)	Alahacoon et al. (2021) successfully tracked nationwide agricultural drought patterns using MODIS Normalized Difference Vegetation Index (NDVI) and Land Surface Temperature (LST) data. Selvarajah et al. (2021) applied MODIS surface reflectance information to support Sri Lanka's drought-flood hazard evaluations. Alahacoon et al. (2018) used MODIS eight-day reflectance products to identify long-term flood-affected regions.
Global Land Data Assimilation System (GLDAS)	Janani et al. (2023) utilized the GLDAS soil moisture data to analyze drought variability in the Lower Bhavani Basin. Madhavi et al. (2022) applied GLDAS soil moisture estimates to validate satellite-retrieved soil moisture patterns in southern India. A. Li et al. (2015) integrated GLDAS atmospheric fields with MODIS data to model evapotranspiration across South Asia.
Sentinel-2	J. Li et al. (2021) utilized Sentinel-2 multi-spectral imagery to delineate surface water in northern Sri Lanka, demonstrating its capability to map small and dynamic water bodies with high spatial resolution. Gopalakrishnan and Kumar (2020) utilized Sentinel-2 Multi-Spectral Instrument data to evaluate soil salinity patterns across the Jaffna Peninsula, demonstrating its effectiveness in detecting salinity-related spectral responses in semi-arid landscapes.

Based on the demonstrated success of the above studies, these three platforms are confirmed as suitable for the Sri Lankan context. Consequently, three products are selected for the independent validation of hydrological clusters in this study.

To complement the satellite-derived datasets, this study incorporates two critical geospatial resources specifically relevant to Sri Lanka's hydrological context. The national-scale digital soil property maps by Wimalasiri et al. (2020) provide comprehensive soil characteristics up to a depth of 100 cm at a resolution of 1 km, offering essential data on subsurface properties that influence infiltration and water retention (European Space Agency, 2024). Additionally, the Copernicus Digital Elevation Model (DEM) GLO-30 digital elevation model provides high-resolution (30m) topographic data, which is crucial for analyzing terrain controls on hydrological processes.

2.4 Runoff Prediction Method Comparison

Building upon the catchment classification framework established in Section 2.1, this study implements a comparative analysis of streamflow prediction approaches for ungauged basins. It evaluates the effectiveness of two machine learning approaches: a regionalized LSTM trained on hydrologically similar donor catchments and a global LSTM using all available data. These modern techniques are benchmarked against traditional hydrological modeling with parameter regionalization, where calibrated parameters are transferred from similar gauged catchments. This method has proven useful in Sri Lanka, as demonstrated by its applications in the Deduru and Kelani river basins (Gayashan & Rajapakse, 2024; Rajendran et al., 2020). Selecting an appropriate traditional model requires careful consideration of local validation and compatibility with regionalization techniques relevant to Sri Lankan conditions.

2.4.1 Hydrological model selection

Hydrological modelling efforts in Sri Lanka have largely focused on the Hydrologic Engineering Center's Hydrologic Modeling System (HEC-HMS) and the Soil and Water Assessment Tool (SWAT), both of which are widely applied across the island's diverse climatic zones. SWAT has been used for long-term water-balance and

watershed assessments in several basins; yet HEC-HMS remains the most extensively calibrated, validated, and operationally adopted model in the Sri Lankan context (Iresh et al., 2021; Rajendran et al., 2020). Studies across multiple river basins consistently demonstrate its strong performance. In the Upper Deduru Oya basin, HEC-HMS produced reliable runoff estimates for ungauged catchment conditions (Rajendran et al., 2020). Comparative calibration work has further highlighted the model's sensitivity and adaptability to local rainfall–runoff dynamics (Gunathilake et al., 2021). Applications in the Attanagalu Oya basin confirmed its capability for extreme-related hydrological assessment (Herath & Wijesekera, 2021; Madhushankha & Wijesekera, 2021), while work in the Maha Oya basin reported calibration and validation accuracies exceeding 75% and 80% (Herath & Wijesekera, 2021). Long-term ecological assessments in the Gurugoda Oya sub-catchment also demonstrated robust performance in streamflow simulation (Munasinghe et al., 2021). Collectively, these findings establish HEC-HMS as the most validated, widely applicable, and operationally supported hydrological model in Sri Lanka, providing a strong baseline for comparative analysis in this study.

2.4.2 HEC-HMS methods selection

The HEC-HMS provides flexible modeling approaches, including lumped, semi-distributed, and distributed configurations. For this regionalization study, a lumped model approach was selected for each catchment to align with the parameter transfer methodology between gauged and ungauged basins. The selection of appropriate model components, including canopy, surface, loss, and baseflow methods, proves crucial for developing hydrologically realistic representations of catchment processes.

2.4.2.1 Canopy method

The HEC-HMS provides three options for canopy representation: dynamic, simple, and gridded simple. This study employs the simple canopy method due to its focus on model simplicity, long-term streamflow simulations, and a straightforward representation of the plant canopy.

2.4.2.2 *Surface method*

The surface method in HEC-HMS is optional and generally used in continuous simulation scenarios. Two main surface methods are recommended: Gridded Simple Surface and Simple Surface (U.S. Army Corps of Engineers, 2000). After considering factors such as model simplicity and the need for parameters, the study has selected the simple surface method as the most suitable approach.

2.4.2.3 *Loss method*

The HEC-HMS offers various methods for simulating infiltration losses. For event modeling, options include the initial constant method, the SCS curve number method, exponential, Green-Ampt, and Smith-Parlange methods. For continuous modeling, methods such as the deficit constant loss method and the soil moisture accounting (SMA) method are available. The deficit constant loss method tracks changes in moisture content using a single soil layer, whereas the SMA method employs three layers to represent moisture variations across the soil profile (U.S. Army Corps of Engineers, 2000). Since the study aims to parameterize regionalization, it is essential to understand groundwater dynamics in the soil profile, as they significantly influence river base flow. Therefore, the SMA method has been chosen as more appropriate than the deficit and constant loss method for this research, and it will be applied accordingly.

2.4.2.4 *Transform method*

The transform model in HEC-HMS plays a vital role in translating excess rainfall into direct runoff, offering seven different approaches. Among these options are unit hydrograph (UH) methods, including Clark, Snyder, and SCS techniques. Users have the option to define UH or S-graph ordinates for the transformation of the model. Furthermore, the modified Clark method functions as a linear quasi-distributed UH method, incorporating gridded data for rainfall and evaporation. However, the SCS method is particularly prominent in HEC-HMS applications due to its versatility across different settings, ease of parameter estimation, and dependable results that are comparable to those of more complex methods (U.S. Army Corps of Engineers, 2000).

Consequently, the SCS method will be employed as the transform method for this research.

2.4.2.5 *Baseflow method*

The HEC-HMS offers five distinct techniques for characterizing baseflow contributions to streamflow in basins. The recession method is characterized by an exponential reduction of baseflow following an occurrence or a sequence of events. The constant monthly method, which is suitable for continuous simulation, relies on observed basin data; however, its availability might pose a challenge. The linear reservoir baseflow method maintains mass by directing infiltrated rainfall to the stream, thereby ensuring a consistent flow. Likewise, the nonlinear Boussinesq method yields results that are similar to those achieved with the recession method. For this research, the linear reservoir baseflow method was selected in line with the HEC-HMS user manual. This approach is best suited for use with the SMA loss method, which facilitates lateral water movement between the upper and lower groundwater layers, contributing to baseflow. A portion of this water percolates as aquifer recharge (U.S. Army Corps of Engineers, 2000).

2.4.3 **Model evaluation**

The performance of the competing streamflow prediction approaches is quantitatively assessed using established objective functions and statistical metrics. These metrics serve to impartially quantify predictive accuracy, with each measure providing unique insight into different aspects of the hydrograph's fidelity.

2.4.3.1 *Nash Sutcliffe Efficiency coefficient (NSE)*

The NSE coefficient serves as a primary metric for evaluating hydrological model performance, quantifying the relative magnitude of residual variance compared to the variance in observed data (Melsen et al., 2025). Its formulation is shown as [2-24].

$$NSE = 1 - \frac{\sum_{i=1}^n (Q_{obs,i} - Q_{sim,i})^2}{\sum_{i=1}^n (Q_{obs,i} - Q_{obs})^2} \quad [2-24]$$

According to this definition,

$$Q_{obs,i}, Q_{sim,i} = \text{served and simulated flows at time step } i \text{ (mm/day)},$$

Q_{obs} = mean observed flow (mm/day).

measures how well the simulation predicts observed values, with particular sensitivity to peak flows due to the squared differences. Evaluation metric, NSE, was selected for its widespread acceptance in hydrological studies and its ability to assess overall fit, especially for high-flow events. The coefficient ranges from $-\infty$ to 1, where values ≤ 0 indicate the mean observed value is a better predictor than the model, 0-0.65 represents generally unsatisfactory performance, and values > 0.65 indicate satisfactory to excellent model performance (Moriasi et al., 2007).

2.4.3.2 Root Mean Square Error (RMSE)

The RMSE provides a complementary measure of model accuracy, particularly effective for evaluating mid-range and low-flow conditions due to its sensitivity to moderate deviations (Moriasi et al., 2007). The metric is calculated as [2-25].

$$RMSE = 1 - \frac{\sqrt{\sum_{i=1}^n (Q_{obs,i} - Q_{sim,i})^2}}{n} \quad [2-25]$$

According to this definition,

$Q_{obs,i}, Q_{sim,i}$ = served and simulated flows at time step i (mm/day),

n = number of observations.

Unlike NSE, which focuses on variance reduction, RMSE equally weights errors across all flow ranges, making it particularly valuable for assessing baseflow and moderate flow simulation. Values range from 0 to ∞ , with lower values indicating better performance, where 0 represents perfect prediction (Jain & Singh, 2005).

2.5 Literature Summery

This research addresses the persistent challenge of PUB by developing an integrated framework that synthesizes hydrological signatures, machine learning clustering, and remote sensing validation. The methodology progresses through several distinct phases, each employing specifically selected techniques to ensure robust and interpretable results.

The foundation of this study lies in functional catchment classification using hydrological signatures. Initially, 14 hydrological signatures were computed, covering water balance (RR , MAR), flow variability (FI , $FDCS$, CoV), event responsiveness (RLD , PRE), baseflow characteristics (BFI , RC), and flow extremes (Q_5 , Q_{95} , HFF , LFF , ZFF). These signatures are standardized and refined through correlation filtering to eliminate multicollinearity while preserving signatures that represent essential components of the unit hydrograph.

For catchment classification, implement three clustering algorithms: K-Means, K-Medoids, and Hierarchical clustering, each offering complementary approaches to identifying hydrologically similar groups. The analysis employs three distance metrics: Euclidean, Cosine, and Manhattan distances, capturing different aspects of hydrological similarity. In hierarchical clustering, apply Ward's linkage with Euclidean distance and average linkage with Cosine and Manhattan distances to ensure methodologically consistent groupings. Cluster quality is assessed using Silhouette Score and DBI, with statistical significance of inter-cluster differences validated through one-way ANOVA with Tukey's HSD post-hoc analysis ($\alpha = 0.05$).

Cluster validation utilizes independent remote sensing data, specifically MODIS (evapotranspiration (ET), LST), Sentinel-2 (land cover), and SRTM DEM (topography) products, supplemented by national-scale digital soil maps and Copernicus DEM GLO-30. These datasets provide physical verification of cluster coherence through analysis of within-cluster variance in biophysical characteristics.

The core PUB investigation compares three streamflow prediction approaches: a regionalized LSTM trained on hydrologically similar catchments, a global LSTM using all available data, and traditional hydrological modeling with parameter transfer. For traditional modeling, select HEC-HMS configured with a lumped model, Simple Canopy, Simple Surface, SMA loss, SCS transform, and Linear Reservoir baseflow methods. Model performance is evaluated using NSE and RMSE (mid-low flows) to provide a comprehensive assessment across the entire flow regime.

This structured approach ensures a methodologically rigorous investigation into whether the hydrological context (clustered donors) outperforms the data volume (all

donors) for streamflow prediction in ungauged basins, while maintaining physical interpretability throughout the analytical framework.

CHAPTER 3

3 METHODOLOGY

This chapter delineates the comprehensive research methodology employed to develop and validate a novel framework for streamflow prediction in ungauged basins. The procedural workflow is summarized in Figure 3-1.

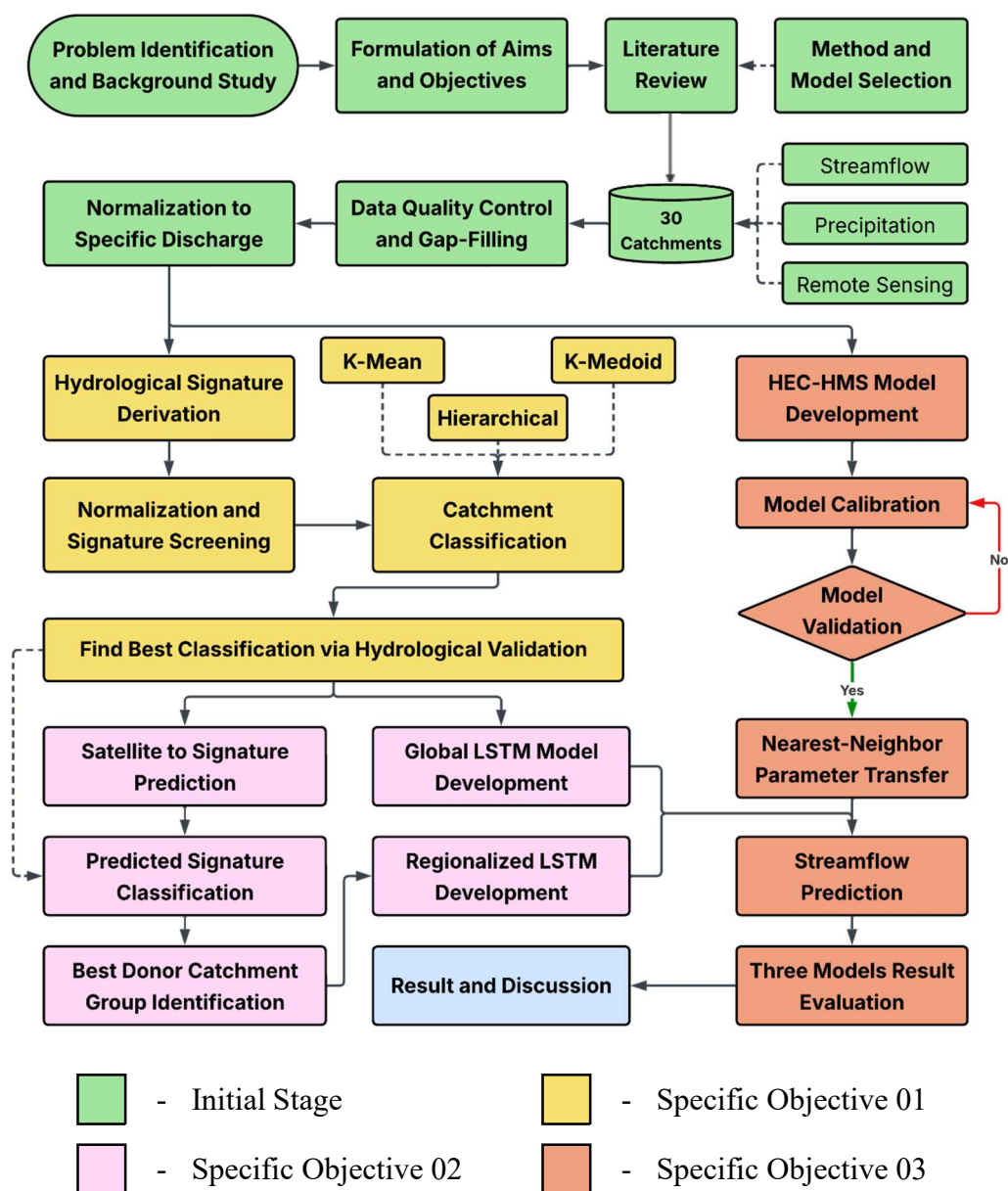


Figure 3-1: Methodological flowchart

3.1 Initial Stage

The issue of predicting streamflow in basins without gauges was addressed using a comparative approach that assessed data-driven methods alongside process-based modeling frameworks. The study aimed to identify the optimal hydrological prediction strategy for Sri Lanka's diverse catchments. Three specific objectives were defined: to classify catchments based on their hydrological function, to validate these classifications using independent satellite data, and to compare the performance of three distinct prediction models rigorously. The methodological selection was informed by a comprehensive review of hydrological literature, which led to the adoption of unsupervised clustering for objective catchment classification, LSTM networks for their capacity to model temporal dependencies, and the HEC-HMS model as a benchmark representative of traditional hydrological frameworks.

3.1.1 Study area

A hydrological observatory has been established, featuring thirty gauged catchments which do not have major reservoirs, deliberately selected from Sri Lanka's Wet, Intermediate, and Dry Zones to ensure a comprehensive representation of the island's hydro-climatic diversity. The key attributes of these catchments are listed in Table 3-1, and their spatial distribution is illustrated in Figure 3-2. Daily streamflow and precipitation records were obtained from the Irrigation Departments and the Departments of Meteorology for a study period of approximately two decades, from 2005 to 2025.

Table 3-1: Details of selected catchments

Sub-basin	Catchment ID	Catchment Area (km ²)	Main River Basin	Climatic Zone
Badalgama	01	1,311	Maha Oya	Wet
Baddegama	02	773	Gin Ganga	Wet
Buttala	03	151	Menik Ganga	Intermediate
Calidoniya	04	184	Mahawali Ganga	Wet
Deraniyagala	05	180	Kelani Ganga	Wet

Sub-basin	Catchment ID	Catchment Area (km ²)	Main River Basin	Climatic Zone
Dunamale	06	160	Attanagalu Oya	Wet
Ellagawa	07	1,415	Kalu Ganga	Wet
Giriulla	08	1,169	Maha Oya	Wet
Glencourse	09	1,528	Kelani Ganga	Wet
Hanwella	10	1,822	Kelani Ganga	Wet
Holombuwa	11	159	Kelani Ganga	Wet
Kataragama	12	800	Menik Ganga	Intermediate / Dry
Kithulgala	13	432	Kelani Ganga	Wet
Kudaoya	14	300	Kirindi Oya	Intermediate / Dry
Mahaoya	15	492	Mundeni Ara	Intermediate / Dry
Millakanda	16	771	Kalu Ganga	Wet
Monaragala	17	230	Mundeni Ara	Intermediate / Dry
Nakkala	18	222	Kumbukkan Oya	Intermediate
Nawalapitiya	19	142	Mahawali Ganga	Wet
Norwood	20	95	Kelani Ganga	Wet
Padiyathalawa	21	168	Maduru Oya	Intermediate
Patupaula	22	2,642	Kalu Ganga	Wet
Pitabeddara	23	293	Nilwala Ganga	Wet
Ratnapura	24	635	Kalu Ganga	Wet
Siyambalanduwa	25	291	Heda Oya	Intermediate / Dry
Thaldena	26	279	Mahawali Ganga	Intermediate
Thanamalwila	27	712	Kirindi Oya	Intermediate / Dry
Thawalama	28	355	Gin Ganga	Wet
Urawa	29	74	Nilwala Ganga	Wet
Wellawaya	30	169	Kirindi Oya	Intermediate

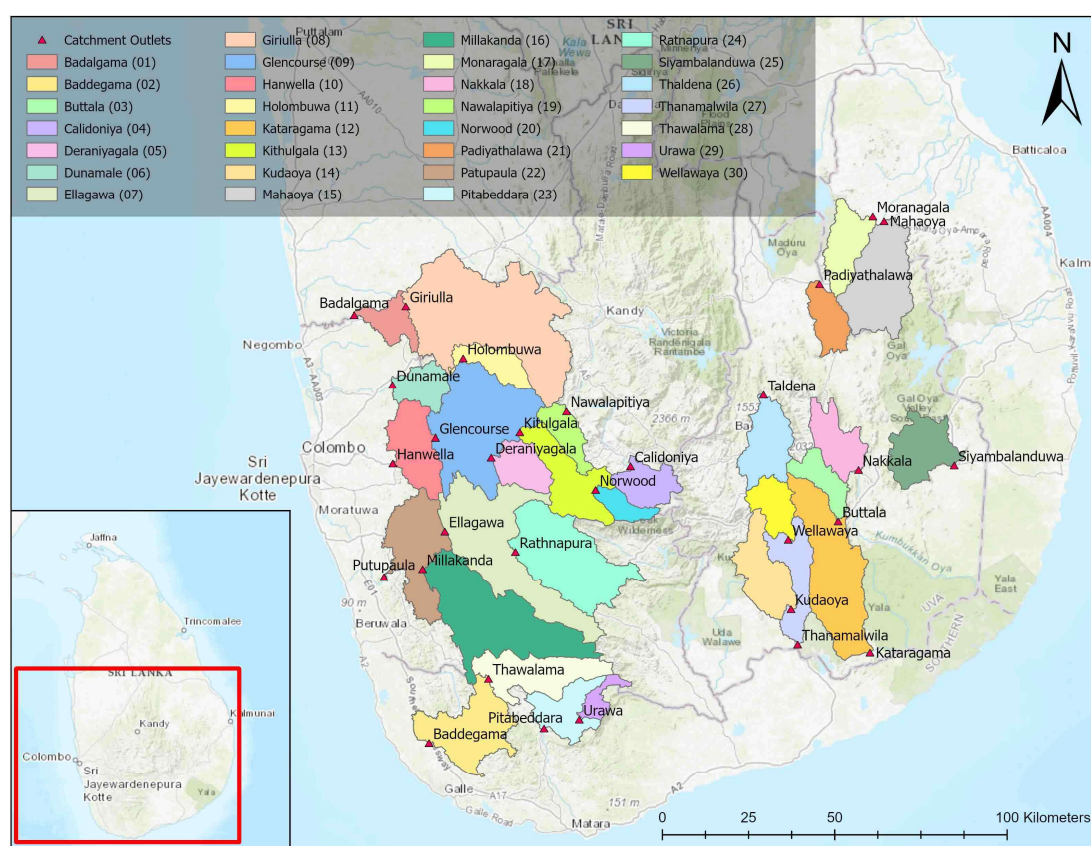


Figure 3-2: Spatial distribution of selected catchments

Additionally, a multi-faceted geospatial database was created by integrating data from various remote sensing platforms. The specific satellites, sensors, and derived data products used are detailed in Table 3-2. The spatio-temporal means were calculated for each remote sensing band to represent distinct attributes within the catchments.

Table 3-2: Remote sensing data sources used in this study

Satellite Platform	Spatial Resolution	Temporal Resolution	Utilized Time Period	Utilized Products
MODIS (MOD16A2GF, MOD13A3, MOD21C2)	500 m, 1 km, 5.6 km	8-day composites, Monthly composites, Daily	2000– 2025	ET, NDVI, LST

Satellite Platform	Spatial Resolution	Temporal Resolution	Utilized Time Period	Utilized Products
Sentinel-2A/B	10 m	Annual composites	2017–2024	Land Cover (Water, Trees, Crops, Built, Rangelands)
GLDAS-2.2	~ 5 km	Daily aggregates	2003–2025	Profile Soil Moisture (PSM), Root Zone Soil Moisture (RSM), Groundwater Storage (GWS), Canopy Interception (CI)
Copernicus DEM	30 m	Static	-	Elevation, Area, Slope
National Soil Property Database (NSPD)	1 km	Static	-	Bulk Density (BD), Sand Content (SNDPPT), Silt Content (SLTPPT), Clay Content (CLYPPT), Volumetric Moisture Content Drainage Upper Limit (VMCDUL), Volumetric Moisture Content Wilting Point (VMCWP)

Here, the NSPD dataset provides characteristics at five discrete depth layers from 0 to 100 cm. A weighted average calculation was applied to aggregate these discrete layers into two hydrologically meaningful depth intervals for analysis. The layer values are calculated as shown in Equation [3-1].

$$\text{Weighted Average} = \frac{\sum(\text{Layer Value} \times \text{Layer Thickness})}{\text{Total Thickness}} \quad [3-1]$$

The surface layer (0–5 cm) was analyzed to represent controls on initial abstraction, infiltration, and surface runoff generation. The full soil profile (0–100 cm) was

characterized to quantify the total soil water storage capacity of each catchment, a critical factor influencing baseflow sustainment and drought resilience.

3.1.2 Data quality control and gap-filling protocol

The integrity of the hydrological time series was ensured through visual inspection and statistical checks of all daily rainfall and streamflow data, which identified missing entries and potential outliers. A critical quality control step involved excluding any time series where the missing data exceeded 10% of the record length, in accordance with World Meteorological Organization (WMO) guidelines for hydrological analysis (WMO, 2025). For stations with gaps below this threshold, the Multivariate Imputation by Chained Equations (MICE) algorithm was implemented. This method was selected over simpler alternatives due to its ability to model conditional dependence between variables, thereby preserving the multivariate structure and temporal autocorrelation of the dataset, a critical requirement for subsequent time-series analysis.

Annexure 1 shows a summary of the data quality used in this study for streamflow and precipitation.

3.1.3 Areal rainfall estimation and discharge normalization

A continuous precipitation field for each catchment was derived by calculating areal rainfall using the Thiessen polygon method (Alfred, 1911). This technique assigned weights to individual rain gauges based on their representative area within the catchment boundary, computed using Equation [3-2].

$$\text{Areal precipitation } (P_A) = \sum_{i=1}^n (w_i \cdot P_i) \quad [3-2]$$

According to this definition,

W_i = weight of the i -th gauge,

P_i = precipitation of the i -th gauge (mm/day).

To enable direct comparison of hydrological response across catchments of different sizes, all volumetric streamflow data (Q , in $\text{m}^3 \text{s}^{-1}$) were converted to specific discharge (q , in mm day^{-1}). This normalization was performed using Equation [3-3].

$$q = k \frac{Q}{A} \quad [3-3]$$

According to this definition,

A = catchment area (km²)

K = unit conversion constant (86.4)

This process removed the influence of basin scale, allowing the intrinsic runoff generation efficiency of each catchment to be expressed and compared directly.

3.2 Specific Objective 1

An initial set of fourteen hydrological signatures was computed to represent key aspects of catchment behaviour, including flow frequency, variability, and water balance. To mitigate multicollinearity and ensure parsimony, a pairwise Pearson correlation analysis was conducted. Signatures exceeding an absolute correlation threshold of $|r| \geq 0.65$ were deemed redundant. From each correlated pair, one signature was systematically excluded based on criteria of hydrological interpretability and prevalence in literature, resulting in a final set of six non-redundant signatures. This final set was intentionally curated to collectively represent the core components of a unit hydrograph: the rising limb, peak flow, recession limb, and baseflow.

3.2.1 Catchment classification

Three distinct clustering algorithms were implemented to ensure methodological robustness: K-Means, K-Medoids, and Hierarchical clustering. Given the sensitivity of results to the distance metric, multiple combinations were evaluated:

- I. K-Means clustering + Euclidean distance (Initialization: K-Means++)
- II. K-Medoids clustering + Euclidean distance (Initialization: K-Medoids++)
- III. K-Medoids clustering + Cosine distance (Initialization: K-Medoids++)
- IV. K-Medoids clustering + Manhattan distance (Initialization: K-Medoids++)
- V. Hierarchical clustering + Euclidean distance (Linkage: Ward)
- VI. Hierarchical clustering + Euclidean distance (Linkage: Average)
- VII. Hierarchical clustering + Cosine distance (Linkage: Average)
- VIII. Hierarchical clustering + Manhattan distance (Linkage: Average)

Unlike hierarchical clustering, which yields deterministic results, K-Means and K-Medoids are stochastic, often producing different cluster groups with each run due to random initialization of centroids or medoids. This variability persists even with intelligent seeding methods, such as K-Means++ and K-Medoids++, as they do not guarantee convergence to the same final result. To rigorously address this stochasticity and consistently identify the most stable and optimal partition, each algorithm was executed with 100 random initializations per k-value.

The determination of the ultimate number of clusters (K-final) was influenced by a dual criterion essential for selecting donor catchments in the regionalization process. The algorithm iterated over a range of K values, terminating at the first instance of a singleton cluster (a cluster containing only a single catchment), as such a group is hydrologically meaningless for providing donor data. Within this viable range of k, the optimal value was determined by identifying the configuration that best balanced two complementary internal validation metrics: the Silhouette Score (also known as the elbow method), which maximizes within-cluster homogeneity, and the DBI, which minimizes between-cluster similarity.

3.2.2 Hydrological validation of classification

To verify that the statistically derived clusters represented genuine hydrological similarity, an additional validation framework was employed using independent satellite-derived data. The core hypothesis was that hydrologically coherent clusters should exhibit minimal internal variance and significant separation in their physical attributes as observed from space.

For each of the six hydrological signatures, a set of satellite-derived variables from GLDAS, MODIS, Sentinel-2, the NSPD, and the Copernicus DEM was assembled. A Pearson correlation analysis was conducted between each signature and all the satellite variables. Only variables with a correlation coefficient $|r| > 0.5$ were retained for the subsequent validation of that specific signature, ensuring hydrological relevance.

For every catchment and signature, a Composite Attribute Score (CAS) was calculated as a weighted integration of the selected, normalized satellite variables, which means

this score quantifies the hydrological signature using satellite data. The calculation is defined by Equation [3-4].

$$CAS = \frac{\sum_{i=1}^n (r_i \cdot Z_i)}{\sum_{i=1}^n |r_i|} \quad [3-4]$$

According to this definition,

r_i = Pearson correlation coefficient between the i -th satellite variable and the hydrological signature,

Z_i = Z-score normalized value of the i -th satellite variable for the catchment,

n = number of satellite variables with $|r| > 0.5$ for the signature.

The clustering quality was then quantified using two variance metrics calculated from the CAS values. The Within-Cluster Variance (WCV) measures the homogeneity, or internal coherence, of each cluster, as defined in Equation [3-5].

$$WCV_{base} = \frac{1}{n} \sum_{j=1}^n (CAS_j - \mu_{CAS})^2 \quad [3-5]$$

According to this definition,

n = number of catchments in the cluster

CAS_j = Composite Attribute Score of the j -th catchment in the cluster

μ_{CAS} = mean CAS of all catchments within the cluster

To penalize hydrologically questionable small clusters that are statistically unreliable for regionalization, an inverse penalty factor was applied. Clusters with fewer than three catchments ($n < n_{min}$) had their variances artificially inflated using Equation [3-6]. The final, aggregate WCV for each signature was then computed as the size-weighted average of the penalized variances across all clusters.

$$WCV_{penalized} = WCV_{base} \times \left(\frac{n_{min}}{n} \right) \quad [3-6]$$

According to this definition,

$$n_{min} = 3.$$

Conversely, the Between-Cluster Variance (BCV) quantifies the degree of separation between the different cluster means, as calculated in Equation [3-7].

$$BCV = \frac{\sum_{k=1}^K n_k (\mu_k - \mu_{global})^2}{N_{total}} \quad [3-7]$$

According to this definition,

- K = total number of clusters,
- n_k = number of catchments in cluster k ,
- μ_k = mean *CAS* of cluster k ,
- μ_{global} = global mean *CAS* across all catchments,
- N_{total} = total number of catchments.

Finally, these two complementary variance metrics were synthesized into a Hydrological Coherence Index (HCI) for a holistic assessment, as defined in Equation [3-8].

$$HCI = \frac{BCSI}{WCHI + \epsilon} \quad [3-8]$$

According to this definition,

- $BCSI$ = Between-Cluster Separation Index, average BCV across all six signatures,
- $WCHI$ = Within-Cluster Homogeneity Index, average penalized WCV across all six signatures,
- ϵ = negligible constant (0.0001) to prevent division by zero.

A high HCI value indicates a clustering solution with strong separation between groups and high internal homogeneity within groups, thereby providing robust, quantitative evidence that the statistically derived clusters are hydrologically meaningful and suitable for subsequent regionalization.

3.3 Specific Objective 2

The optimal catchment classification scheme, validated internally and hydrologically, provides a robust framework for donor selection. However, a fundamental challenge

persists for ungauged basins: the absence of observed streamflow from which to compute the requisite hydrological signatures. To overcome this, a methodology was developed to predict the six key hydrological signatures directly from satellite data.

3.3.1 Prediction of hydrological signatures from satellite data

For each signature, the four most highly correlated satellite-derived variables were selected as predictors from the pool of GLDAS, MODIS, Sentinel-2, soil, and topographic data. Given the limited sample size of 30 catchments, a second-order polynomial regression model was employed to avoid overfitting associated with more complex models. This model structure captures potential nonlinear relationships while maintaining parsimony, and the relationship is defined by Equation [3-9].

$$\hat{S}_j = \beta_0 + \sum_{i=1}^4 \beta_i \cdot B_i + \sum_{i=1}^4 \sum_{k=i}^4 \gamma_{ik} \cdot B_i \cdot B_k + \epsilon \quad [3-9]$$

According to this definition,

$\hat{S}_j$ = predicted value of hydrological signature j ,

B_i = normalized value of the i -th selected satellite variable,

$\beta_0, \beta_i, \gamma_{ik}$ = intercept, linear, and interaction coefficients, respectively,

ϵ = error term.

The predictive performance of the satellite-to-signature models was rigorously evaluated using a leave-8-out cross-validation strategy. For this study, two catchments, Monaragala from the dry zone and Ellagawa from the wet zone, were designated as pseudo-ungauged basins for evaluation purposes. Consequently, in all model training iterations, these two catchments were permanently held out from the training set, ensuring they were never used for model calibration. The model was trained on the remaining 22 catchments and then used to predict the signatures for the 8-catchment validation set, which always included Monaragala and Ellagawa alongside six other randomly selected catchments. This process was repeated across multiple random splits of the data. The evaluation of model performance was conducted using the correlation coefficient (r) and RMSE to compare the observed and predicted signatures.

3.3.2 Cluster accuracy assessment and donor catchment identification

Following the leave-8-out cross-validation, the single iteration that yielded the optimal evaluation metrics (the highest r and the lowest RMSE across the test catchments) was selected. The corresponding set of predicted signatures for all 30 catchments from this best-performing model run was retained for all subsequent analysis. The optimal clustering algorithm identified under the *Specific Objective 1* section was then executed using this single, best-available matrix of predicted signatures. The accuracy of these satellite-driven clusters was quantified through a pairwise comparison with the benchmark clusters derived from observed signatures, as defined by Equation [3-10].

$$\text{Pairwise Accuracy} = \frac{\text{Number of Consistent Pairs}}{\text{Total Number of Pairs}} \quad [3-10]$$

The pairwise accuracy assessment evaluates the consistency of relationships between all possible catchment pairs across the two clustering solutions. For a set of N catchments, the total number of unique pairs is given by $\frac{N(N-1)}{2}$.

For each unique pair of catchments (i, j), their cluster membership was compared between the observed-signature clusters (C_{obs}) and the predicted-signature clusters (C_{pred}). A pair is deemed "consistent" if the clustering relationship between the two catchments is preserved across both solutions. Specifically, consistency is achieved when:

- I. **Agreement in similarity:** Catchments i and j are assigned to the same cluster in C_{obs} and are also assigned to the same cluster in C_{pred} (both clustering solutions identify them as hydrologically similar), or
- II. **Agreement in dissimilarity:** Catchments i and j are assigned to different clusters in C_{obs} and are also assigned to different clusters in C_{pred} (both clustering solutions identify them as hydrologically dissimilar).

Conversely, a pair is considered "inconsistent" when the clustering relationship is contradictory between the two solutions: that is, when catchments are grouped together in one clustering but separated in the other.

A high pairwise accuracy confirms that the satellite-based method reliably replicates the hydrological similarity and dissimilarity relationships between catchments, thereby enabling its confident use for classifying ungauged basins. It can also be used to find the best donor catchment group.

3.3.3 Streamflow prediction using regionalized and global LSTM models

The efficacy of the satellite-based donor selection was evaluated by implementing two distinct LSTM modeling strategies for streamflow prediction in pseudo-ungauged basins.

The first was a Cluster-Based Regional LSTM. For a given pseudo-ungauged catchment, a single LSTM model was trained exclusively on the data from the donor catchments within its assigned hydrologically similar cluster. The second was a Global LSTM, trained on the entire ensemble of all gauged catchments, irrespective of hydrological class. Both models used identical architectures, with daily sequences of precipitation, potential evapotranspiration, and satellite-based signatures as inputs to predict specific discharge. Model performance was quantitatively compared using the NSE and RMSE metrics to determine the optimal prediction strategy for different hydro-climatic regimes.

3.4 Specific Objective 3

To establish a benchmark against the machine learning approaches, a traditional hydrological modeling procedure was implemented using the HEC-HMS. As mentioned in the previous section, two representative pseudo-ungauged catchments were selected for this evaluation: Monaragala, located in the dry zone, and Ellagawa, located in the wet zone. For each of these, considering geographical proximity, Padiyathalawa was selected as the donor for Monaragala, and Ratnapura was selected as the donor for Ellagawa.

3.4.1 Model development and parameter transfer

Two separate HEC-HMS models were developed and rigorously calibrated for these donor catchments. Following the successful calibration and validation of the donor

catchment models, a direct parameter transfer approach was employed for prediction in the pseudo-ungauged basins. The complete set of calibrated parameters from the Padiyathalawa model was directly applied to the Monaragala catchment without modification. Similarly, the full parameter set from the Ratnapura model was transferred to the Ellagawa catchment. This method operates on the fundamental premise of hydrological similarity, assuming that catchments within lower proximity will respond to precipitation inputs in a functionally analogous manner and thus can be described by identical model parameters.

The performance of this parameter transfer regionalization technique was quantitatively assessed by running the transferred models for the validation period in the pseudo-ungauged catchments and comparing the simulated streamflow against the observed records. Performance was evaluated using the NSE and RMSE, providing a direct comparison with the LSTM-based approaches and establishing a traditional modeling benchmark for the study.

CHAPTER 4

4 RESULT AND DISCUSSION

4.1 Hydrological Signature Selection and Dimensionality Reduction

The initial correlation analysis of the fourteen computed hydrological signatures revealed significant multicollinearity, as visually depicted in Figure 4-1. Multiple signature pairs exhibited correlation coefficients exceeding $|r| = 0.7$, indicating substantial redundancy and the potential for over-representing specific hydrological processes in the subsequent multivariate analysis.

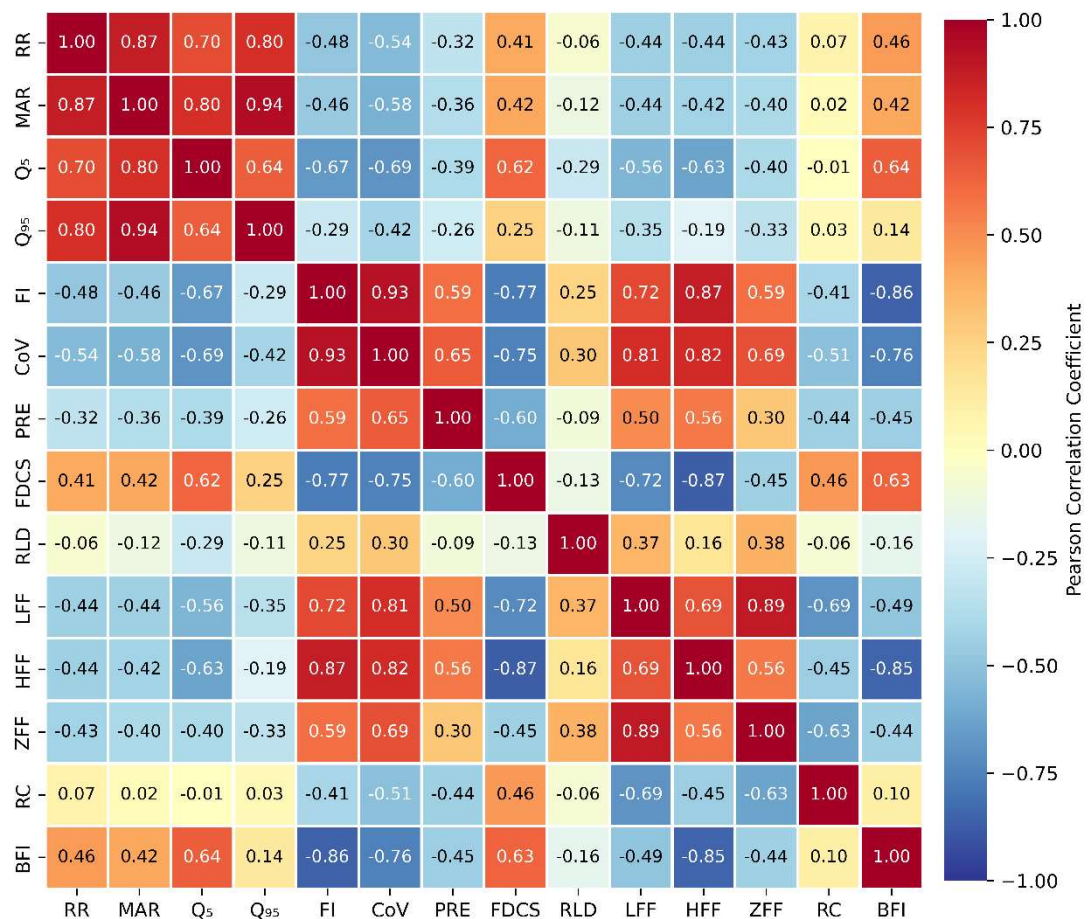


Figure 4-1: Pairwise correlation matrix of the fifteen computed hydrological signatures

Application of the defined correlation filter successfully mitigated this issue, refining the dataset to a suite of six non-redundant signatures: Runoff Ratio (*RR*), Baseflow Index (*BFI*), Flow Duration Curve Slope (*FDCS*), Recession Constant (*RC*), Precipitation-Runoff Elasticity (*PRE*), and Rising Limb Density (*RLD*). This final selection, presented in Figure 4-2.

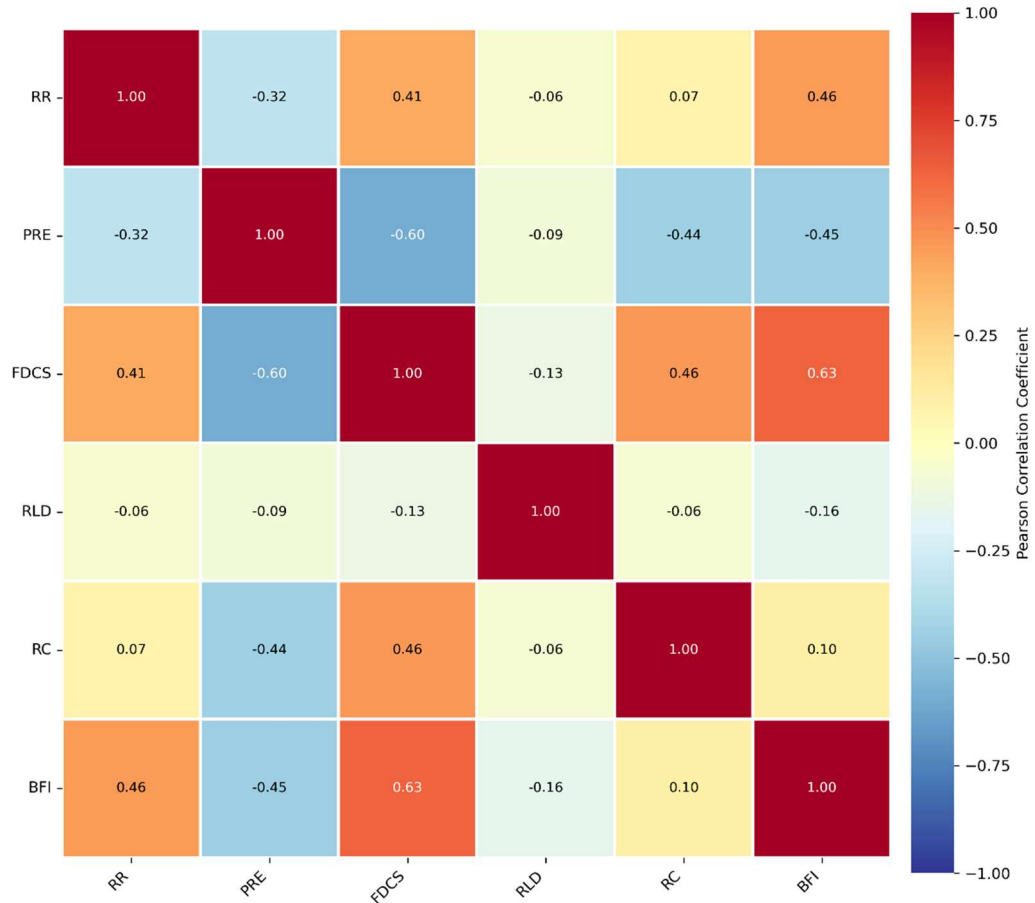


Figure 4-2: Correlation matrix of the six filtered hydrological signatures selected for cluster analysis

Critically, these six signatures collectively represent the core components of a unit hydrograph. The *RLD* captures the rising limb response frequency, while *FDCS* and *PE* describe the peak flow magnitude and sensitivity. The *RC* explicitly quantifies the recession limb dynamics, and the *BFI* represents the baseflow contribution. The *RR* provides the foundational input-output water balance, integrating the overall catchment efficiency. This conceptual alignment confirms that the selected signature set is not merely a statistical artifact but a robust, process-informed basis for catchment classification. Annexure 2 shows all 14 signature values.

4.2 Catchment Classification Results from Selected Clustering Algorithms

The performance of all tested algorithm-distance metric combinations is summarized in Table 4-1, which presents the corresponding Silhouette Score and DBI for each configuration. The process for determining the optimal number of clusters (K) for each of these leading methods, selected based on the highest Silhouette score and lowest DBI.

Table 4-1: Silhouette score and DBI for all clustering algorithm-distance metric combinations

Algorithm	Distance Metric	Initialization / Linkage	Optimum No of Clusters	Silhouette Score	DBI
K-Means	Euclidean	K-Means++	2	0.383	1.027
K-Medoids	Cosine	K-Medoids++	7	0.501	1.054
K-Medoids	Manhattan	K-Medoids++	2	0.406	1.027
K-Medoids	Euclidean	K-Medoids++	2	0.383	1.027
Hierarchical	Euclidean	Ward	4	0.254	1.151
Hierarchical	Cosine	Average	7	0.525	1.009
Hierarchical	Manhattan	Average	2	0.379	0.855
Hierarchical	Euclidean	Average	2	0.363	0.855

The results indicate that clustering performance was highly sensitive to both the algorithm and the distance metric. Based on this initial evaluation, the two top-performing methods were Hierarchical clustering with cosine distance and K-Medoids clustering with cosine distance, achieving the highest silhouette scores and favourable DBI values.

The statistical superiority of the Hierarchical and K-Medoids methods with cosine distance required verification that their clusters represented genuine hydrological similarity.

4.3 Hydrological Validation of Cluster Solutions Using Satellite-Derived Composite Attributes

4.3.1 Composite attribute score calculation

Validation was assessed by calculating HCI derived from CAS as mentioned in the methodology section. The CAS for each signature was computed using the four most highly correlated satellite products, detailed in Table 4-2, which served as weighted inputs to translate satellite observations into a proxy for each catchment's hydrological function, as described in Equation [3-4].

Table 4-2: Satellite products and correlation coefficients used for CAS calculation

Hydrological Signature	Satellite Product	Correlation Coefficient (r)
BFI	MODIS: LST	-0.56
FDCS	GLDAS: Canopy Water	-0.51
	MODIS: LST	-0.60
	GDLAS: GWS	0.56
	GDLAS: PSM	0.56
RC	NSPD: SNDPPT 0-100 cm	-0.57
	NSPD: SNDPPT 0-5 cm	-0.51
	NSPD: BD 0-100 cm	-0.51
	NSPD: SLTPPT 0-100 cm	0.50
	NSPD: VMCWP 0-5 cm	0.54
	NSPD: VMCDUL 0-100 cm	0.57
RLD	NSPD: CLYPPT 0-100 cm	-0.62
	Copernicus DEM: Area	-0.59
RR	MODIS: LST	-0.73
	Sanitel 2: Crops Cover	-0.69
	GDLAS: CI	-0.59
	Sanitel 2: Water Cover	-0.53
	NSPD: BD 0-5 cm	-0.52
	MODIS: NDVI	0.55
	NSPD: CLYPPT 0-5 cm	0.57
	Copernicus DEM: Slope	0.57
	MODIS: ET	0.58
	GDLAS: GWS	0.61
	GDLAS: PSM	0.63

Correlation analysis between hydrological signatures and satellite products generally matched established principles. For the *BFI*, a negative correlation with LST is expected, as higher temperatures increase evaporation and lower groundwater recharge, reducing baseflow. The lack of strong correlations with groundwater or soil attributes ($|r| > 0.5$) suggests satellite proxies may not fully represent baseflow processes.

The *RC* closely relates to soil properties, supporting the theory. Negative correlations with sand content and bulk density show faster drainage in coarse soils, while positive correlations with silt and moisture at field capacity indicate slower recession in finer soils. These findings support using soil grids to explain catchment storage and release. For *RLD*, negative correlations align with hydrological theory. Larger catchments have slower, less frequent hydrograph rises, while more clay increases the frequency of continuous runoff events, as reflected in *RLD* values. Notably, no satellite products strongly correlated with *PRE*, which measures streamflow's sensitivity to precipitation. This suggests that satellite data may not capture the primary factors driving climatic sensitivity in these catchments.

FDCS correlations matched expectations. Higher soil moisture and groundwater storage flatten the flow-duration curve by moderating flows. Lower canopy water is associated with drier, less forested catchments. *RR* showed complex correlations. *RR* decreases with higher LST and increases with soil moisture and groundwater storage. Positive correlations with clay content and slope suggest local context shapes water yield.

4.3.2 Within-cluster variance calculations

Following the calculation of CAS, the hydrological coherence of each clustering method was assessed by computing the WCV and BCV for each signature, as shown in Figure 4-3 and Figure 4-4.

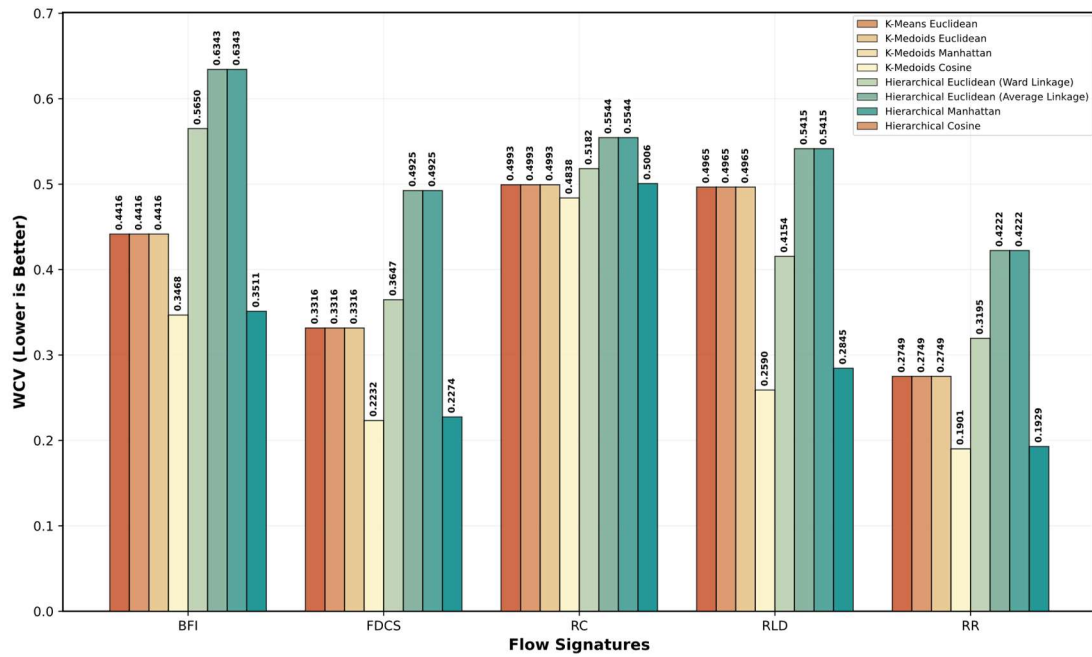


Figure 4-3: WCV based on CAS for each hydrological signature across all clustering methods

The analysis of WCV in Figure 4-3 offers a key hydrological view of cluster performance. K-Medoids (cosine) and Hierarchical (cosine) methods yield the most homogeneous clusters, shown by consistently lowest WCV on most signatures. While internal metrics favored Hierarchical, WCV points to K-Medoids as slightly better for hydrological homogeneity. Exceptionally low WCV for *BFI*, *FDCS*, *RLD*, and *RR* shows catchments within clusters are nearly identical in remote sensing characteristics, supporting a strong physical grouping. Higher WCV for the *RC* suggests remote sensing attributes are less effective at distinguishing cluster differences for this signature, which may be influenced by factors that satellites cannot fully capture.

4.3.3 Between-cluster variance calculations

The BCV analysis, presented in Figure 4-4, reinforces the findings from the WCV assessment, demonstrating a consistent performance hierarchy among the clustering methods. The K-Medoids (cosine) and Hierarchical (cosine) configurations again emerge as superior, achieving the highest degree of separation between cluster means based on the Composite Attribute Scores, with the K-Medoids method holding a slight edge. This superior separation is particularly evident for signatures such as *BFI*, *RR*, *FDCS*, and *RLD*, confirming that the clusters derived using cosine distance represent

distinct, physically identifiable hydrological regimes. Conversely, the *RC* signature again demonstrates relatively low BCV across all methods.

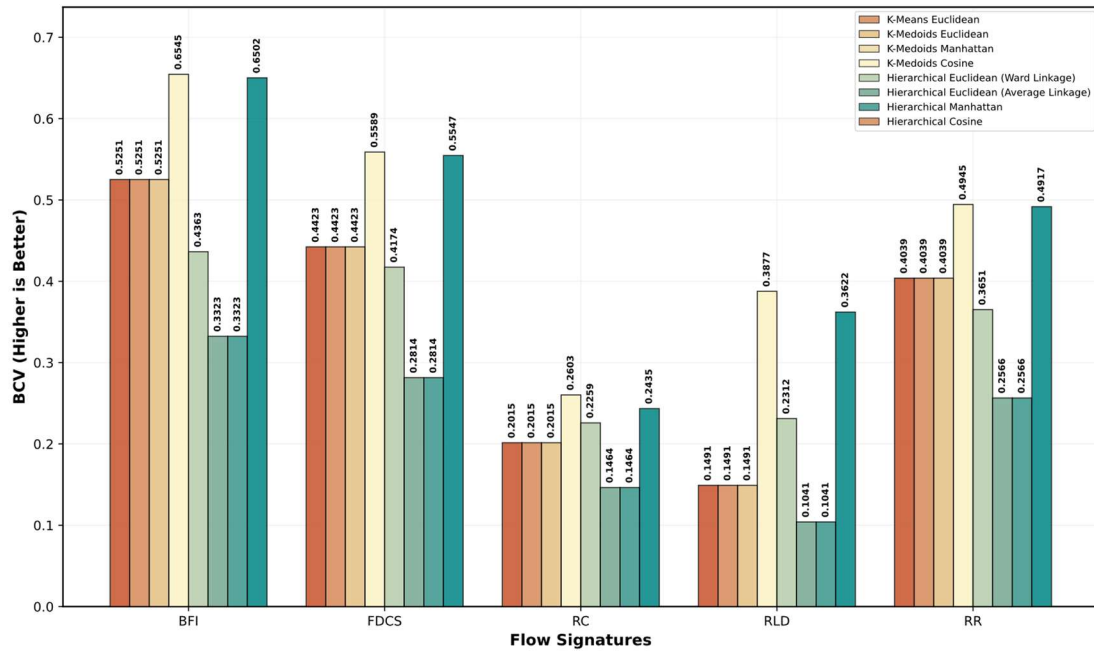


Figure 4-4: BCV based on CAS for each hydrological signature across all clustering methods

4.3.4 Synthesis of hydrological coherence index

The HCI (Figure 4-5) summarizes cluster homogeneity and separation. K-Medoids with cosine distance gives the most hydrologically coherent clusters, with slightly higher HCI than Hierarchical. This indicates a better balance between internal consistency and external differentiation.

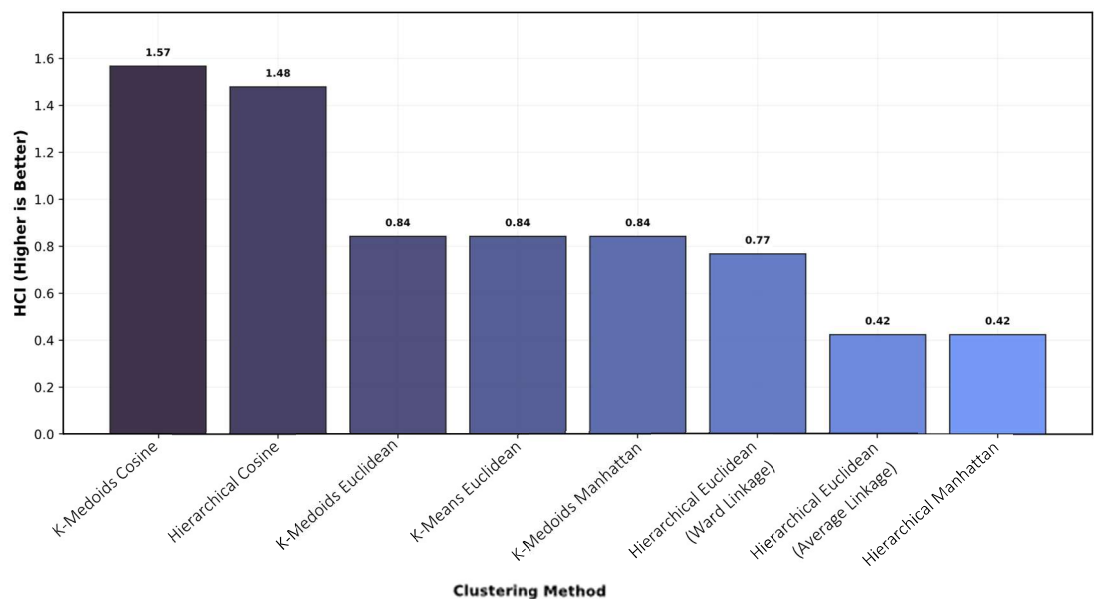


Figure 4-5: HCI values for all clustering method configurations

Comparing the two top methods shows a key distinction: Hierarchical clustering scored highest on internal metrics, but K-Medoids was slightly better in external hydrological validation. While differences are small, this highlights that statistical and physical validation can yield different views on cluster quality. Both methods deserve detailed hydrological interpretation for regionalization.

4.4 Optimum Classification Algorithm and Distance Method Selection

Prior to a detailed comparison of the K-Medoids and Hierarchical clustering algorithms, it is essential to address their common, high-performing component: the cosine distance. Its superior performance, despite ignoring the absolute magnitudes of catchment attributes, indicates that the multi-attribute profile, or the relative balance among characteristics, is a more robust indicator of hydrological similarity than their absolute values.

This finding is supported by the literature. Ali et al. (2012) found that grouping catchments by physical properties alone did not match groupings based on flow indices or storage estimates, indicating that magnitude alone is insufficient for predicting hydrological behavior. Instead, the pattern or configuration of attributes, which the cosine distance captures, is critical. Research by Chérif and Gargouri-Ellouze (2023) also validates cosine distance, showing it consistently forms better groups than magnitude-based metrics. Both correlation and cosine distances measure pattern similarity, not absolute values. Their high-performance shows that the structure and relationships between attributes, not their magnitude, are key for classification.

Cosine distance also excels in high-dimensional problems, where magnitude-based distances like Euclidean become less meaningful. By normalizing vectors and focusing on orientation, cosine distance clusters catchments by structural and functional "signatures".

Therefore, even though cosine distance ignores magnitude, it outperforms other metrics for hydrological classification because it captures the key proportional relationships among catchment attributes. This principle holds across different climatic regions and supports robust hydrological regionalization.

4.4.1 K-Medoids clustering with cosine distance

The K-Medoids algorithm, configured with cosine distance and K-Medoids++ initialization, emerged as a leading method, producing an optimal partition into 7 clusters with a Silhouette Score of 0.501 and a DBI of 1.054. The spatial distribution of these clusters within the reduced-dimensionality space is visualized through a series of principal component (PC) plots, including PC1 vs. PC2, PC1 vs. PC3, PC2 vs. PC3, and a composite three-dimensional plot of PC1, PC2, and PC3 (Figure 4-6 a-d). Visual inspection of these projections indicates that while the majority of clusters are well-separated, a degree of overlap is perceptible between Cluster 3 and Cluster 6, suggesting a less distinct boundary.

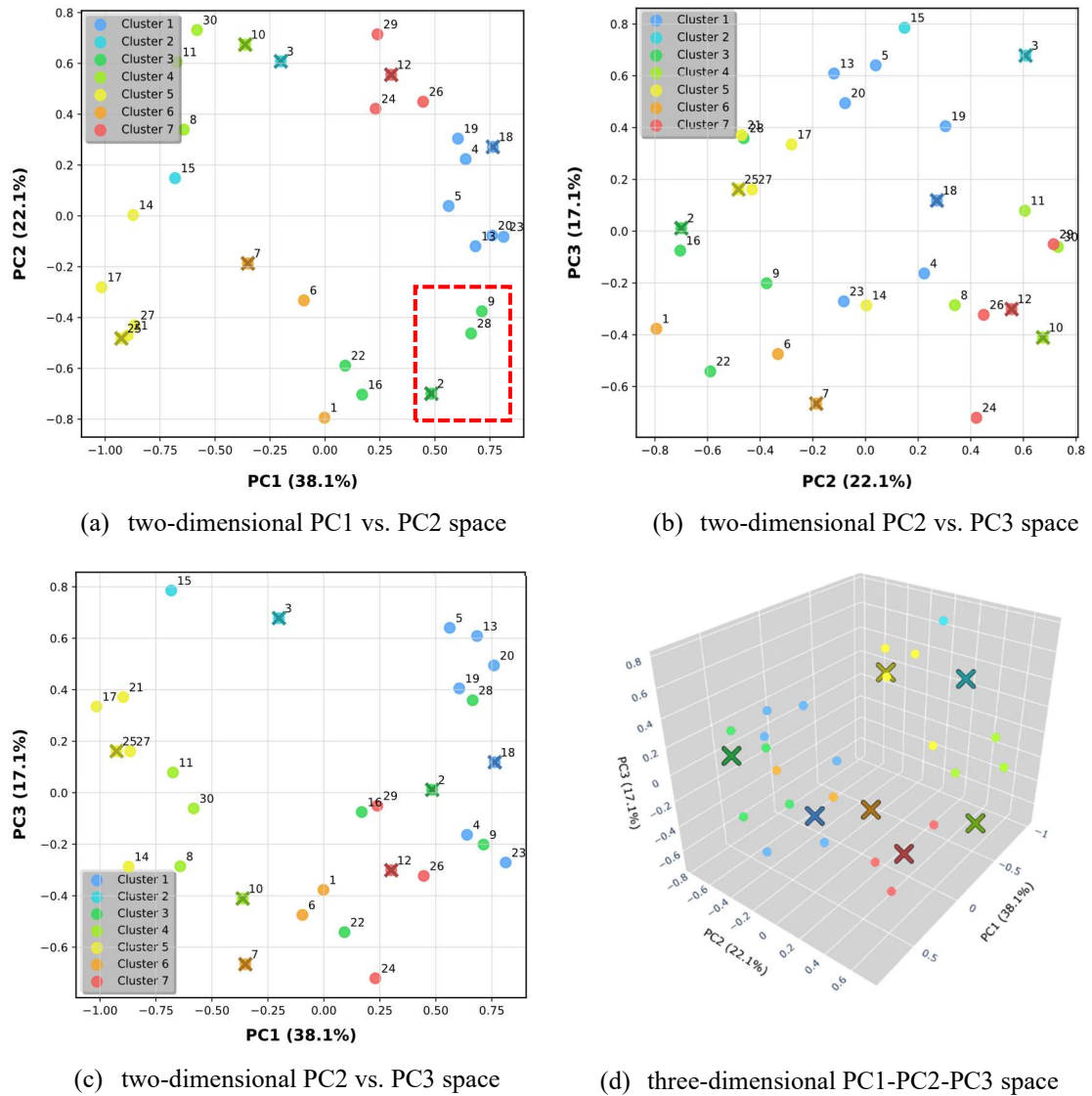


Figure 4-6: Principal component projections of the seven-cluster K-Medoids solution with cosine distance

This ambiguity suggests that some catchments within these groups may exhibit transitional or hybrid hydrological characteristics (e.g., Catchments 1, 16, and 22 may show the same hydrological behaviour), rendering their classification less definitive. To definitively diagnose the nature of this overlap and assess the hydrological distinctness of all clusters, a detailed analysis of the distribution of individual hydrological signatures within each cluster is required. This analysis is presented in the following section via the spatial distribution of clusters and statistical box plots (See Figure 4-7 & Figure 4-8).

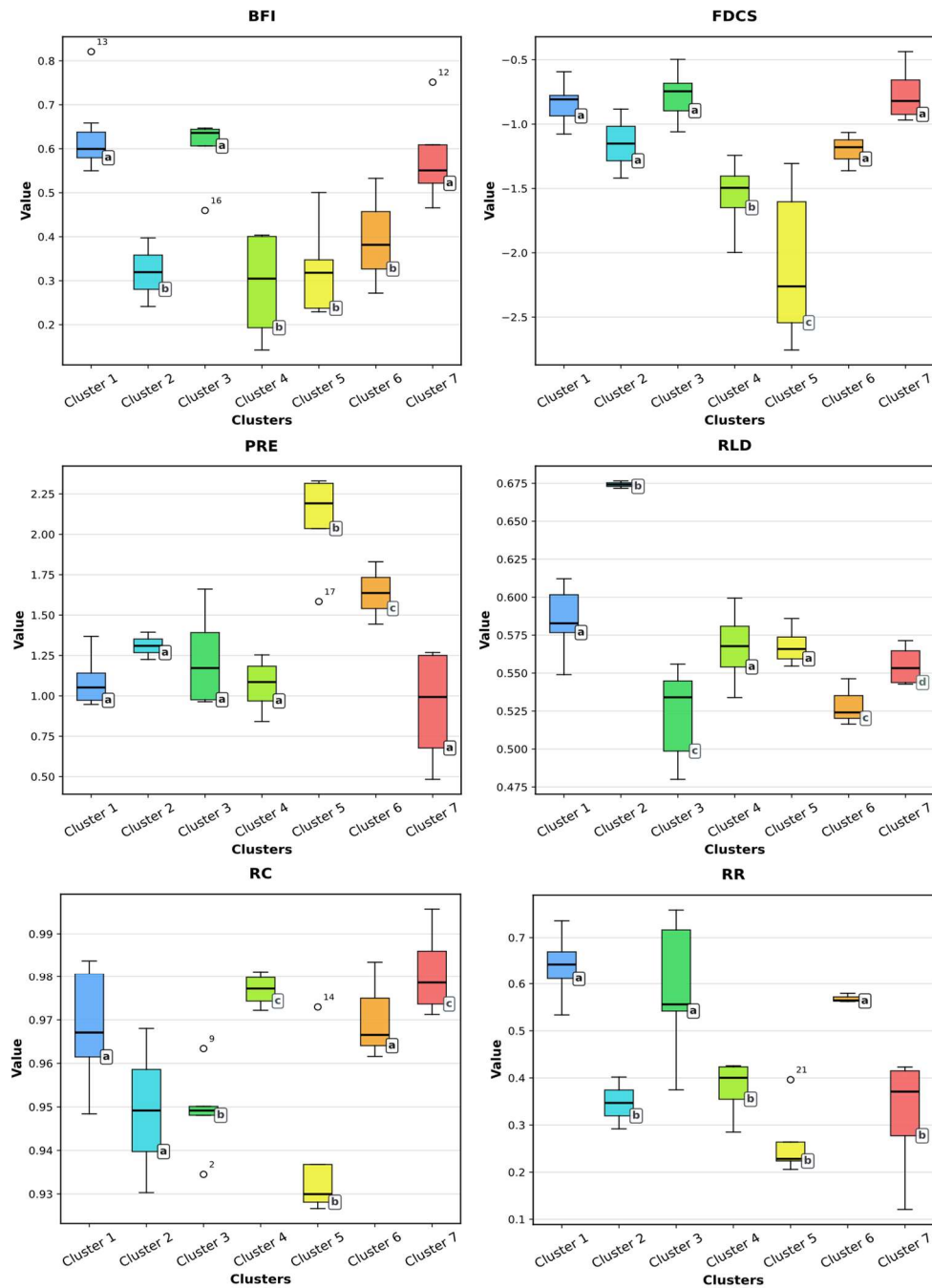


Figure 4-7: Distribution of the six hydrological signatures across the seven K-Medoids clusters

In the box plots, the central line represents the median value, the box indicates the interquartile range (25th–75th percentile), and the whiskers show the data spread excluding outliers. Letters above the box plots (*a*, *b*) denote results of post-hoc statistical comparison, where clusters sharing the same letter are not significantly different, while clusters with different letters indicate statistically significant differences at $p < 0.05$.

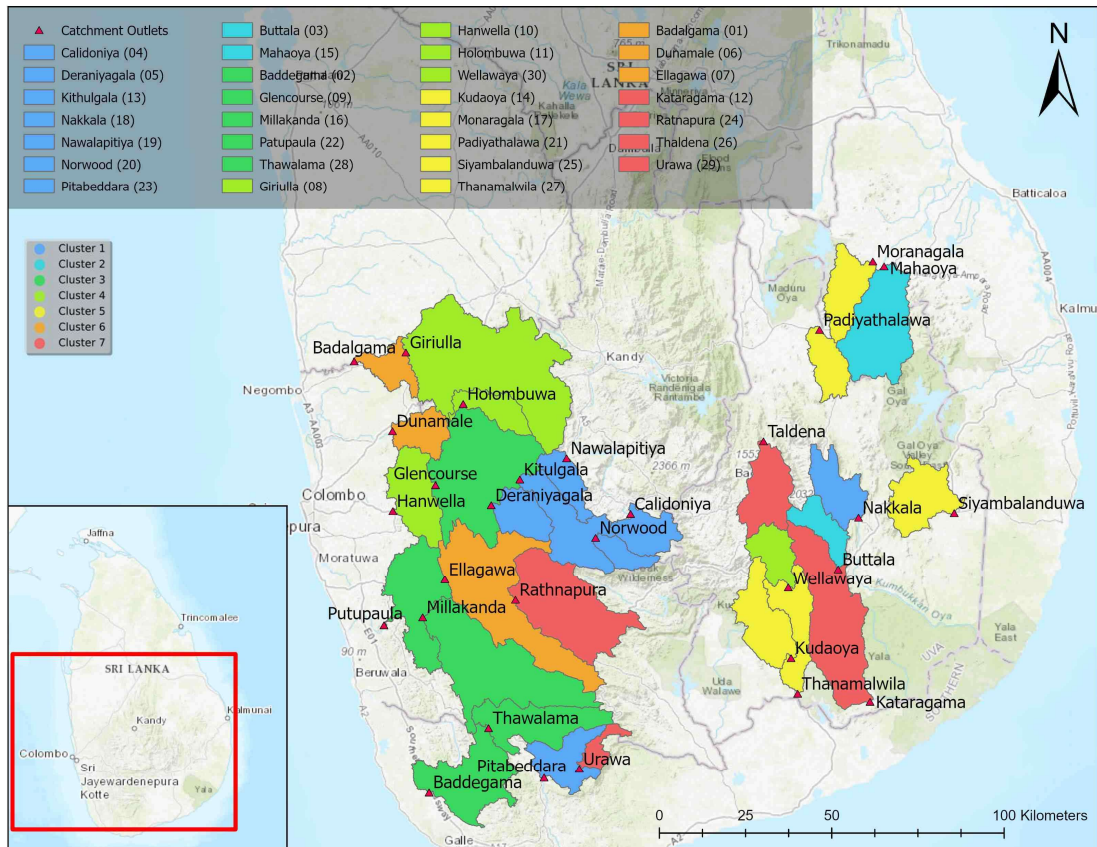


Figure 4-8: Spatial distribution of the seven hydrological clusters derived from the K-Medoids algorithm with cosine distance across the study catchments

The spatial distribution of K-Medoids clusters (Figure 4-8) shows clear geographical patterns, with nearby catchments often grouped together, supporting the principal hydrological assumption that proximity often indicates similar hydrological behavior. However, some hydrologically similar catchments are far apart, revealing a key limitation of relying only on geography. These findings highlight the value of classification-based regionalization, which can identify functionally similar catchments regardless of location.

Cluster 1 (Calidoniya, Deraniyagala, Kithulgala, Nakkala, Nawalapitiya, Norwood, Pitabeddara) shows very high baseflow and a high runoff ratio. The physical data confirm that these are mostly mid to high-elevation catchments with very high forest cover and excellent groundwater storage, providing a clear physical basis for the strong baseflow and high, stable runoff yields. The spatial distribution confirms that all catchments are located in the Sri Lankan hilly wet forest zone. The dense vegetation and deep soils facilitate high infiltration and storage, leading to strong baseflow contribution and high, stable runoff yields with controlled flashiness.

Cluster 2 (Buttala, Mahaoya) is defined by very frequent storm responses and low baseflow. Data on land cover and soil properties validate this, showing that these catchments have shallow groundwater storage and are dominated by rangelands and sandy soils, leading to rapid, frequent surface runoff and limited baseflow. Cluster 3 (Baddegama, Glencourse, Millakanda, Patupaula, Thawalama) exhibits high baseflow, a high runoff ratio, and low flashiness. An assessment of the physical attributes confirms a consistent profile of high groundwater storage, high soil moisture, and significant tree cover across these catchments, validating this stable, high-yield hydrological regime. The spatial distribution of cluster 3 confirms that it includes some of the largest catchments in the wet zone, prone to flooding.

Cluster 4 (Giriulla, Hanwella, Holombuwa, Wellawaya) shows low baseflow and a flashy flow regime. The physical characteristics derived from remote sensing and storage data confirm that these catchments have moderate to low groundwater storage and notable impervious areas, which align with limited infiltration and a quick, flashy response. Cluster 5 (Kudaoya, Monaragala, Padiyathalawa, Siyambalanduwa, Thanamalwila) is the most flashy and sensitive cluster, with a very low runoff ratio. The integration of physical data from soil and land cover sources strongly supports this, indicating a uniform setting of very low groundwater storage, low soil moisture, and sandy soils, creating a "leaky" system that validates the extreme flashiness and low runoff. The spatial distribution of clusters again confirms that all these catchments are located in the dry zone.

Cluster 6 (Badalgama, Dunamale, Ellagawa) displays a moderate baseflow and a high runoff ratio. The physical data provides validation for this mix. Some catchments show

urban influence from land cover maps, driving the high runoff ratio, while others have better underlying storage that supports the moderate baseflow. Cluster 7 (Kataragama, Ratnapura, Thaldena, Urawa) has moderate to high baseflow and the slowest recession. According to the physical groundwater and soil storage data, the behavior is strongly influenced by catchments with good storage capacity, which validates the pattern of a buffered, stable flow regime dominated by sustained groundwater discharge.

However, some catchments in the K-Medoids results appear as outliers or are better suited to other clusters. These cases often reflect transitional or unique hydrological behavior due to specific combinations of physical factors.

- a. Kataragama (12) (Cluster 7) is an outlier. While Cluster 7 usually has moderate to high baseflow, slow recession (high RC value), and the lowest PRE , Kataragama stands out on nearly every signature: it has the highest BFI ($= 0.751$), least $FDCS$ ($= -0.437$), highest RC ($= 0.996$), and lowest RLD and RR in its cluster. Its unique mix of very high BFI , very slow recession, and very low PRE does not fit any other cluster either. Cluster 5 has high PRE , while Clusters 3 and 1 have low BFI .

However, the derived hydrological signatures for Kataragama appear inconsistent with the known physical characteristics of the catchment. Physically, Kataragama is characterized by low groundwater storage capacity and is situated in a dry, low-slope area with high evapotranspiration. Under natural conditions, such characteristics would typically result in low baseflow contribution, rapid recession, and high sensitivity to rainfall variability. As illustrated in Figure 4-9, which present one month of streamflow and rainfall variation, the observed flow record shows minimal recession response to rainfall events with almost constant streamflow.

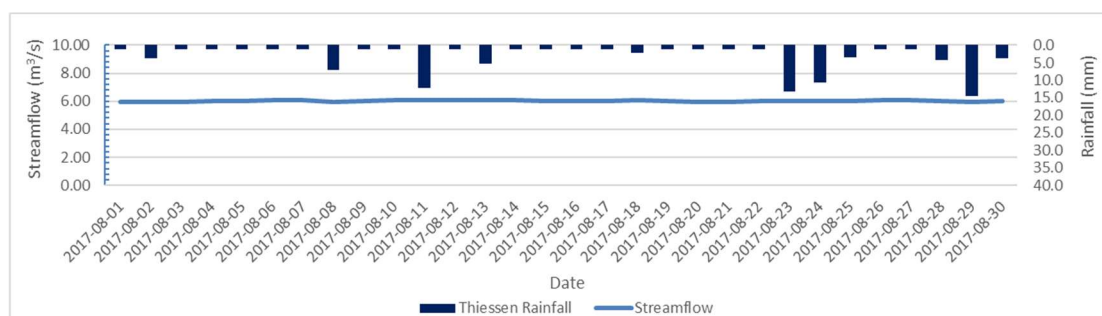


Figure 4-9: Temporal variation of streamflow and rainfall for Kataragama catchment during 08/2017

This contrasts sharply with field observations indicating that the Kataragama River goes completely dry during the dry season, suggesting very low groundwater storage capacity or rapid groundwater recession. The discrepancy between the calculated signatures (particularly the very high *BFI* and slow recession coefficient) and the observed intermittent flow behavior suggests potential data quality concerns. These may include streamflow measurements taken under pooling conditions where flow continuity is absent or instrument limitations during low-flow periods.

Therefore, the clustering algorithm struggled to place this catchment due to the contradiction between its calculated signatures and those of hydrologically similar systems and this case underscores the critical importance of combining signature-based analysis with catchment-specific hydrological knowledge, even when employing DL approaches. Understanding the physical behavior and data quality constraints of individual catchments is essential for correctly interpreting outliers and avoiding misclassification based on potentially unreliable streamflow records.

- b. Badalgama (02) (Cluster 6) is a borderline case. While Cluster 6 has moderate *BFI* and high *RR*, Badalgama has the highest *BFI* ($= 0.533$) and *RR* ($= 0.580$), making it an extreme within Cluster 6. Its high values and large built-up area (12.14%) fit better with Cluster 3, which is defined by high *BFI*, *RR*, and enhanced runoff from urbanization. As noted in Figure 4-6 (a), Badalgama may act more like a Cluster 3 catchment. Similarly, Patupaula (22) (Cluster 3) also shows similar kinds of characteristics.

4.4.2 Hierarchical clustering with cosine distance

The Hierarchical clustering algorithm with cosine distance and Ward linkage was also a top performer, producing 7 clusters (Silhouette Score 0.525, DBI 1.009). The hierarchical classification is visualized in a dendrogram (Figure 4-10), which illustrates the nested relationships and merger distances among catchments. Cluster arrangement in reduced-dimensional space appears in principal component plots (Figure 4-11 a–d), while the spatial distribution of the seven clusters is shown in Figure

4-12. Visual inspection of these projections reveals generally strong separation between the cluster groups, though some intermixing is observable at the boundaries of Clusters 4 and 5, suggesting these groups may share specific transitional hydrological characteristics.

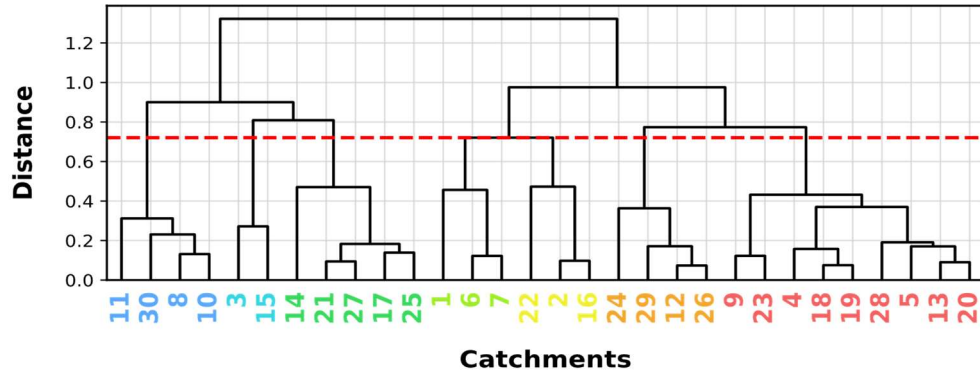
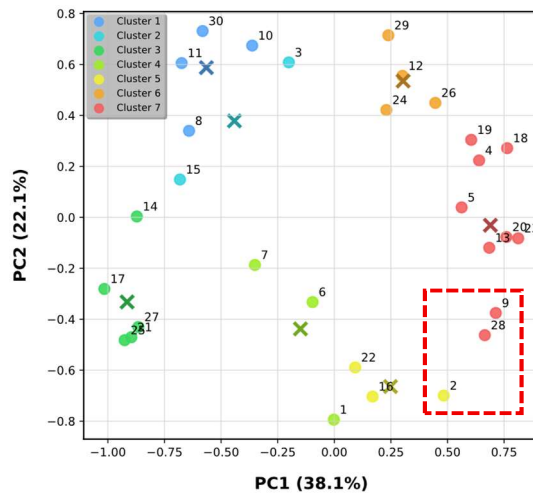
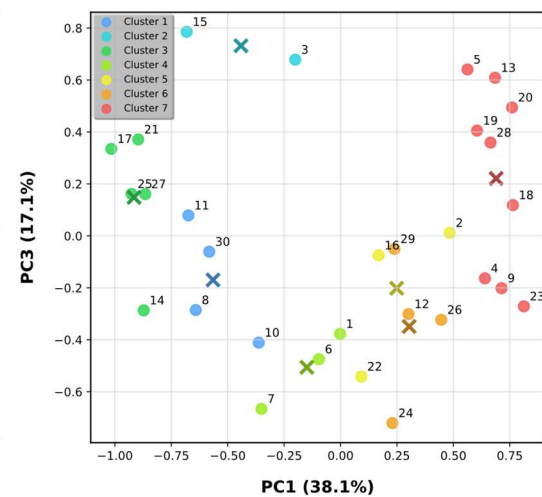


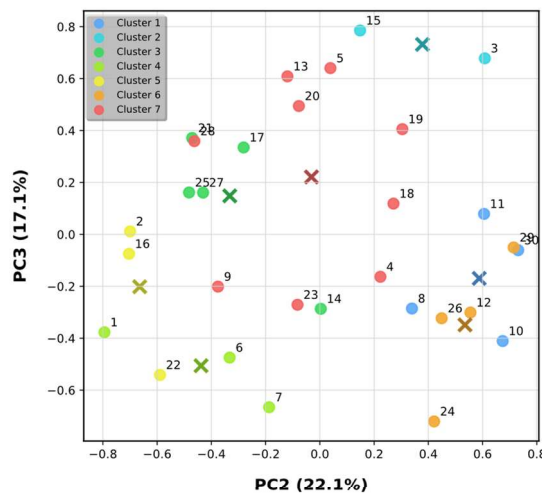
Figure 4-10: Dendrogram illustrating the hierarchical cluster structure using cosine distance



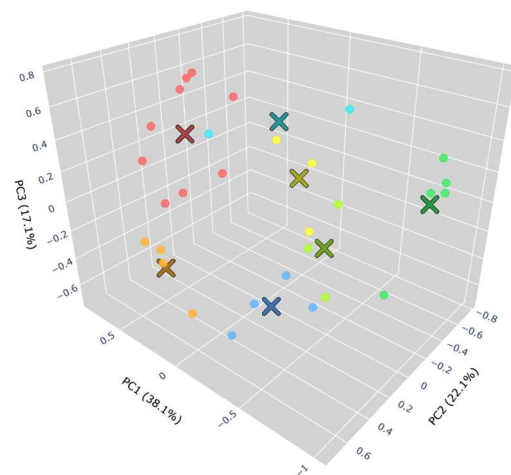
(a) two-dimensional PC1 vs. PC2 space



(b) two-dimensional PC2 vs. PC3 space



(c) two-dimensional PC2 vs. PC3 space



(d) three-dimensional PC1-PC2-PC3 space

Figure 4-11: Principal component projections of the seven-cluster hierarchical solution



algorithm with cosine distance across the study catchments

Similar to K-Medoids, Hierarchical (Figure 4-12) solutions reveal a strong consensus, reinforcing the finding that geographically proximate catchments frequently exhibit hydrological similarity. Specifically, five of seven clusters are functionally equivalent across methods, differing only in labels: Hierarchical Cluster 1 matches K-Medoids Cluster 4; Cluster 2 matches Cluster 2; Cluster 3 matches Cluster 5; Cluster 4 matches Cluster 6; and Cluster 6 matches Cluster 7. This high level of agreement validates the robustness of the identified hydrological regions, while the minor differences highlight catchments with transitional characteristics that are sensitive to algorithmic treatment.

The following cluster variation boxplots in Figure 4-13 show that the main differences between the methods stem from the catchments Glencourse and Thawalama. In the K-Medoids result, Cluster 1 contains Calidoniya, Deraniyagala, Kithulgala, Nakkala, Nawalapitiya, Norwood, and Pitabeddara. The hierarchical clustering formed Cluster 7 by adding Thawalama and Glencourse to this same group. Conversely, the K-Medoids Cluster 3 consists of Baddegama, Glencourse, Millakanda, Patupaula, and

Thawalama. The hierarchical equivalent, Cluster 5, is formed by removing Thawalama and Glencourse from this set. Therefore, only a difference in variation is expected between clusters 5 and 7 in hierarchical clustering, and this reassignment of Glencourse and Thawalama between algorithms highlights that these catchments have transitional characteristics, making their classification sensitive to the chosen clustering method.

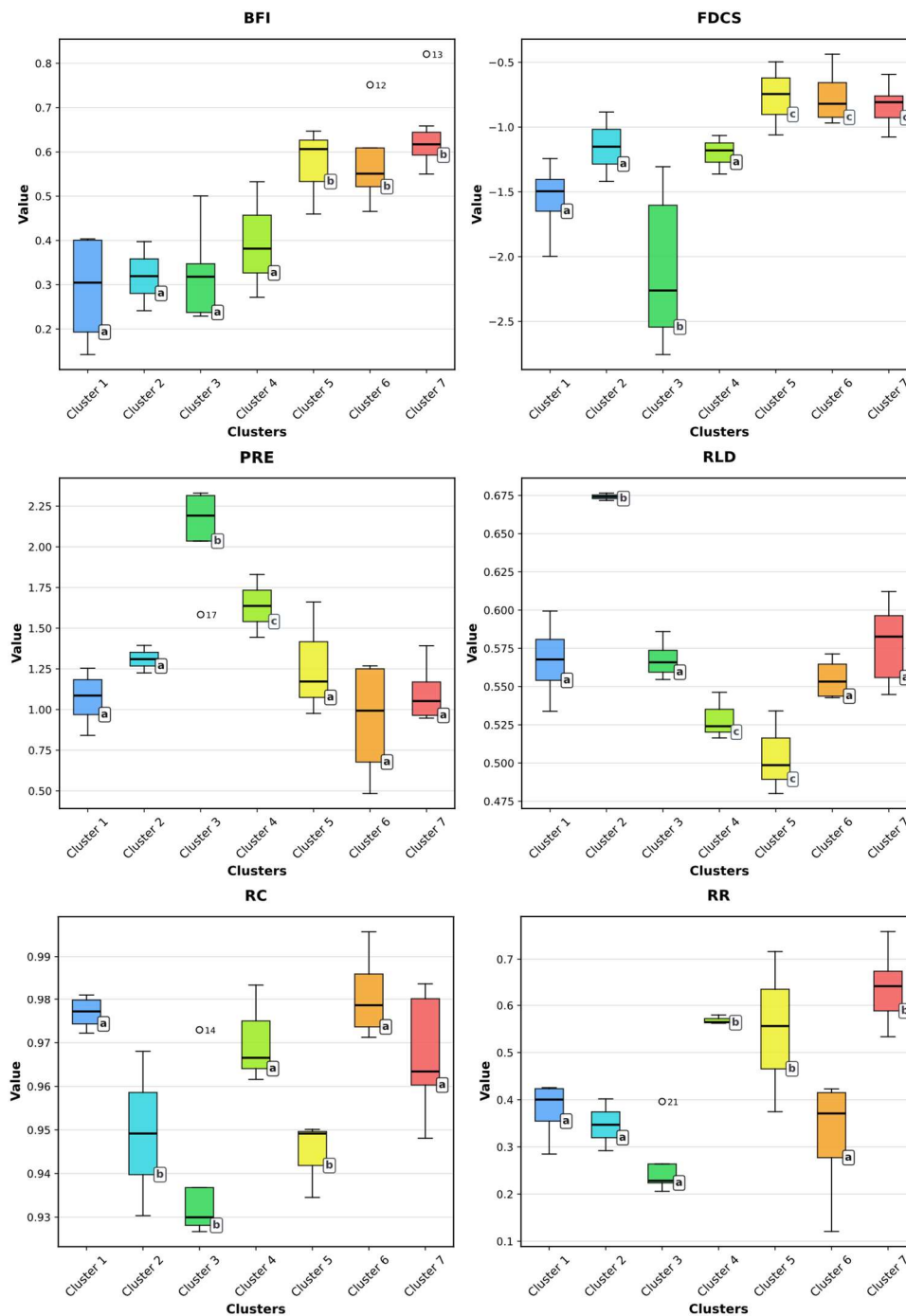


Figure 4-13: Distribution of the six hydrological signatures across the seven Hierarchical clusters

In the box plots as well, the central line represents the median value, the box indicates the interquartile range (25th–75th percentile), and the whiskers show the data spread excluding outliers. Letters above the box plots (*a*, *b*) denote results of post-hoc statistical comparison, where clusters sharing the same letter are not significantly different, while clusters with different letters indicate statistically significant differences at $p < 0.05$.

Cluster 5 is characterized by high baseflow and a moderately high runoff ratio, with less flashy behavior. The physical characteristics derived from remote sensing and soil databases provide clear validation, showing that these catchments have consistently high groundwater storage, high soil moisture, and significant tree cover, which promotes infiltration and sustains baseflow. Cluster 7 represents the most high-yield and stable group, with the highest baseflow and runoff ratio. According to the physical characteristics from terrain and land cover analysis, these are predominantly high-elevation, forested catchments with excellent groundwater storage, which fully explains their efficient and sustained hydrological performance. But there are not many differences in the variation of boxes.

Similarly to the previous K-Medoids result, Kataragama (2) appears as an outlier, but it is a unique case that does not support other clusters. Patupaula (22) and Beddegama (02) exhibit borderline behavior within the hierarchical cluster. This is because a few changes occurred in the cluster compared to the previous K-Medoids result.

Based on hydrological validation using HCI values, which showed minimal performance difference between K-Medoids clustering + cosine (HCI=1.57) and hierarchical clustering + cosine (HCI=1.48) and considering the absence of bias-corrected satellite data in this study, hierarchical clustering with cosine distance was selected as the optimal classification approach. This decision is further supported by its superior statistical metrics, achieving a Silhouette Score of 0.525 and a DBI of 1.009, compared to the K-Medoids results of 0.501 and 1.054, respectively. Consequently, the hierarchical clustering with cosine distance outcome was utilized for the subsequent objective of donor catchment identification. For other cluster combinations, boxplots and spatial variation are shown in Annexure 3.

4.5 Prediction of Hydrological Signatures from Satellite Data

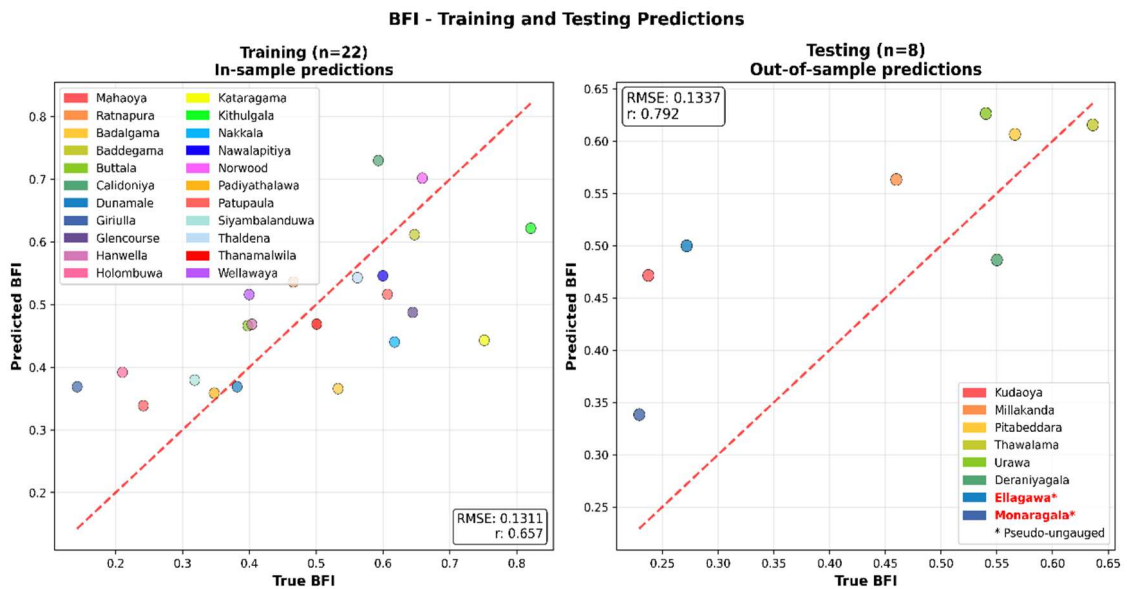
The operationalization of the catchment classification framework for ungauged basins necessitated the development of predictive models to estimate hydrological signatures directly from satellite data. For each of the six signatures, the four satellite-derived products exhibiting the highest magnitude of correlation ($|r|$) were selected as predictor variables. This set of selected features for each signature is detailed in Table 4-3.

Table 4-3: Satellite-derived predictor variables selected for the hydrological signature prediction models

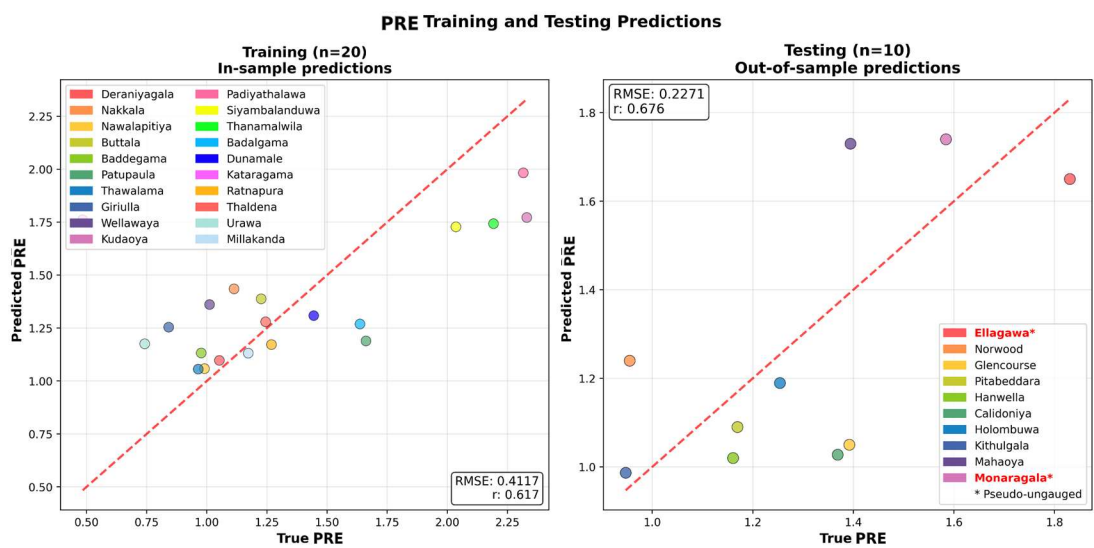
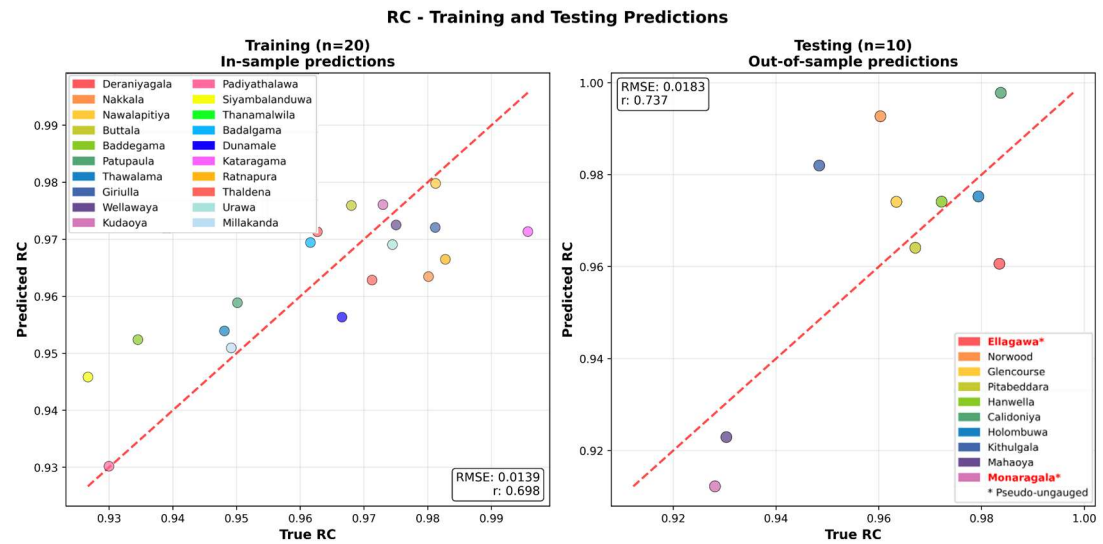
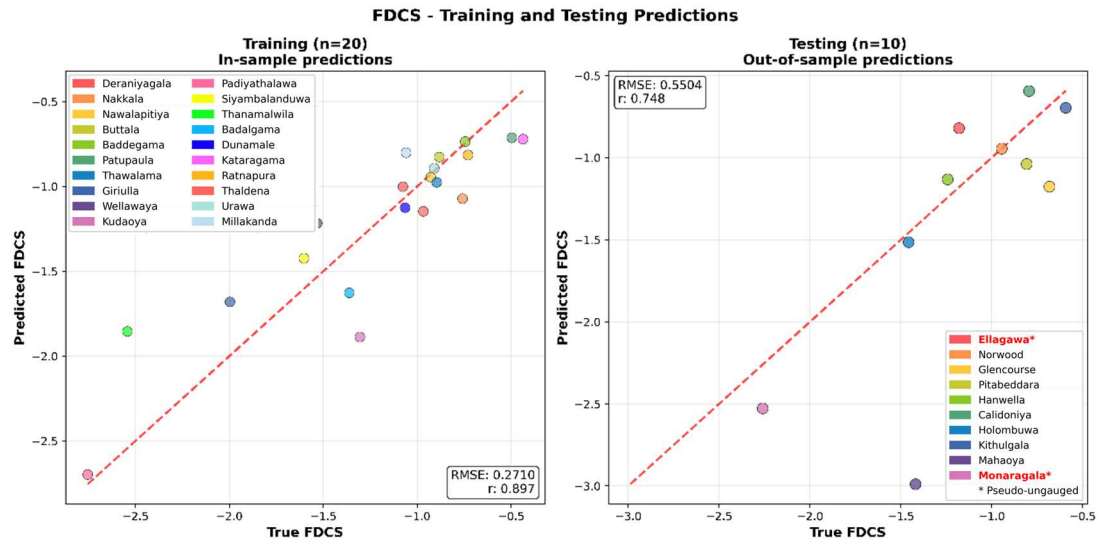
Hydrological Signature	Satellite Product	Correlation Coefficient (r)
BFI	MODIS: LST	-0.56
	BD 0-5 cm	-0.49
	SNDPPT 0-5cm	-0.46
	SLTPPT 0-5cm	0.40
FDCS	MODIS: LST	-0.60
	GLDAS: CI	-0.51
	GDLAS: GWS	0.56
	GDLAS: PSM	0.56
RC	NSPD: SNDPPT 0-100cm	-0.57
	NSPD: BD 0-100cm	-0.51
	NSPD: VMCPW 0-100cm	0.54
	NSPD: VMCDUL 0-100cm	0.57
PRE	GLDAS: CI	0.45
	MODIS: ET	-0.40
	MODIS: LST	0.47
	Slope	-0.40
RLD	NSPD: CLYPPT 0-100 cm	-0.62
	Area	-0.59
	Sanitel 2: Rangelands	0.42
	GLDAS: CI	0.48

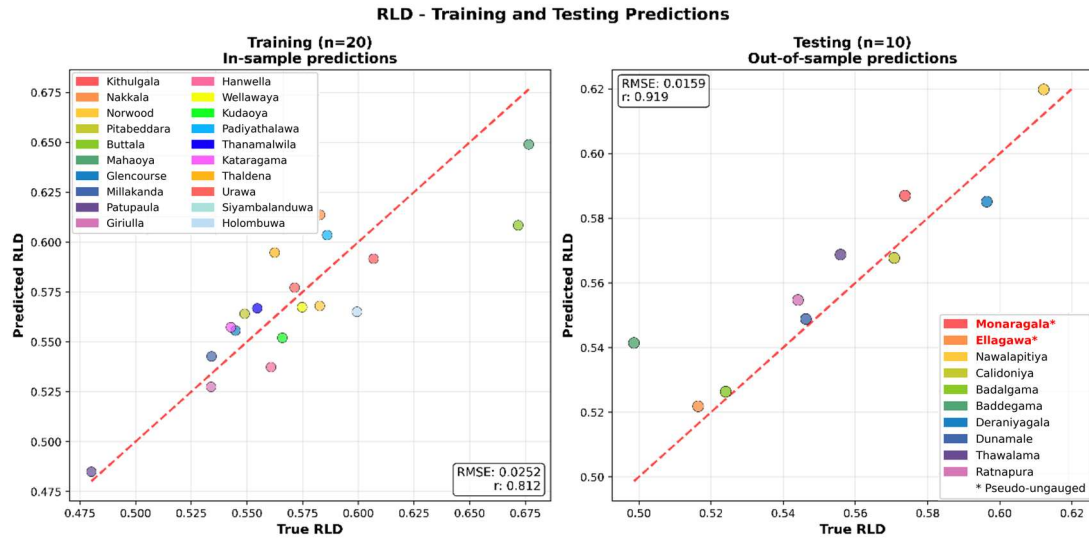
Hydrological Signature	Satellite Product	Correlation Coefficient (r)
RR	MODIS: LST	-0.73
	MODIS: ET	0.58
	GDLAS: GWS	0.61
	GDLAS: PSM	0.63

As evidenced in Table 4-3, the number of satellite-derived predictors meeting the correlation threshold ($|r| > 0.5$) varied significantly across the hydrological signatures. Notably, the *BFI* and *PRE* possessed fewer strongly correlated predictors compared to other signatures. This initial finding suggests that predicting these two specific signatures from the available satellite data would present a greater challenge, as the model has less informative data to leverage. The subsequent scatter plots in Figure 4-14 (a-f), which display the predicted versus actual signature values for both the training set (20 catchments) and the test set (8 catchments, which consistently included the pseudo-ungauged catchments Monaragala and Ellagawa), confirm this expectation and provide a detailed visual assessment of the prediction accuracy for all six signatures.

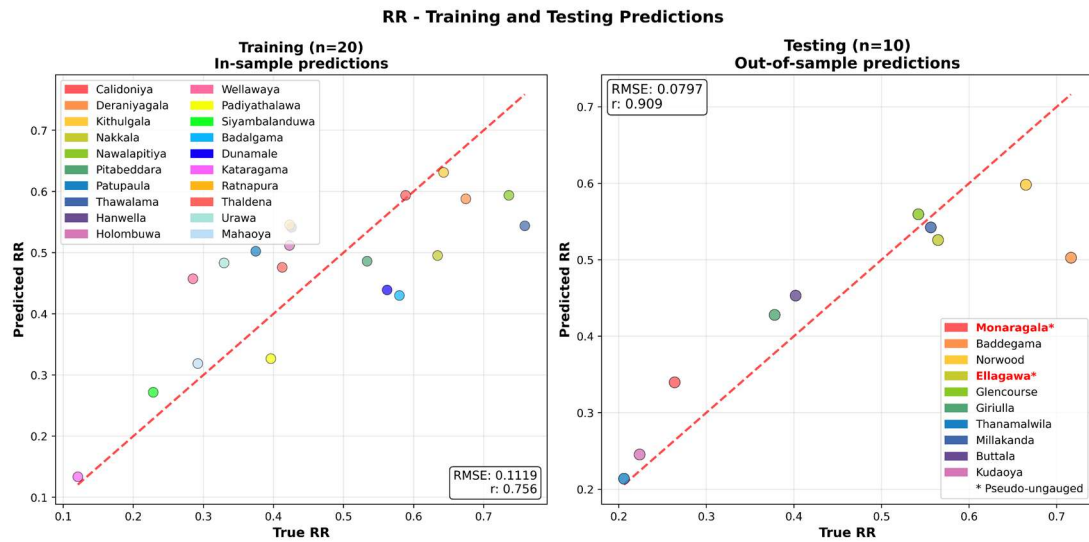


(a) BFI





(e) RLD



(f) RR

Figure 4-14: Predicted versus observed hydrological signatures from the satellite-based regression models (a) - (f)

As expected, *PRE* showed lower correlation values in training (0.61) and testing (0.67) than 0.7. Signature *BFI* had a lower training value (0.66) but a good testing value (0.79). All other predicted signatures showed Pearson correlations of nearly or above 0.7 with the original signature values, which is satisfactory.

4.6 Assessment of Cluster Prediction Accuracy Using Pairwise Comparison

These predicted signatures were then clustered into seven groups using the previously selected best classification algorithm, hierarchical clustering with cosine distance. The

following Figure 4-15 and Table 4-4 show the Pairwise Comparison Matrix of the original and predicted cluster distributions, and the pairwise accuracy results, respectively.

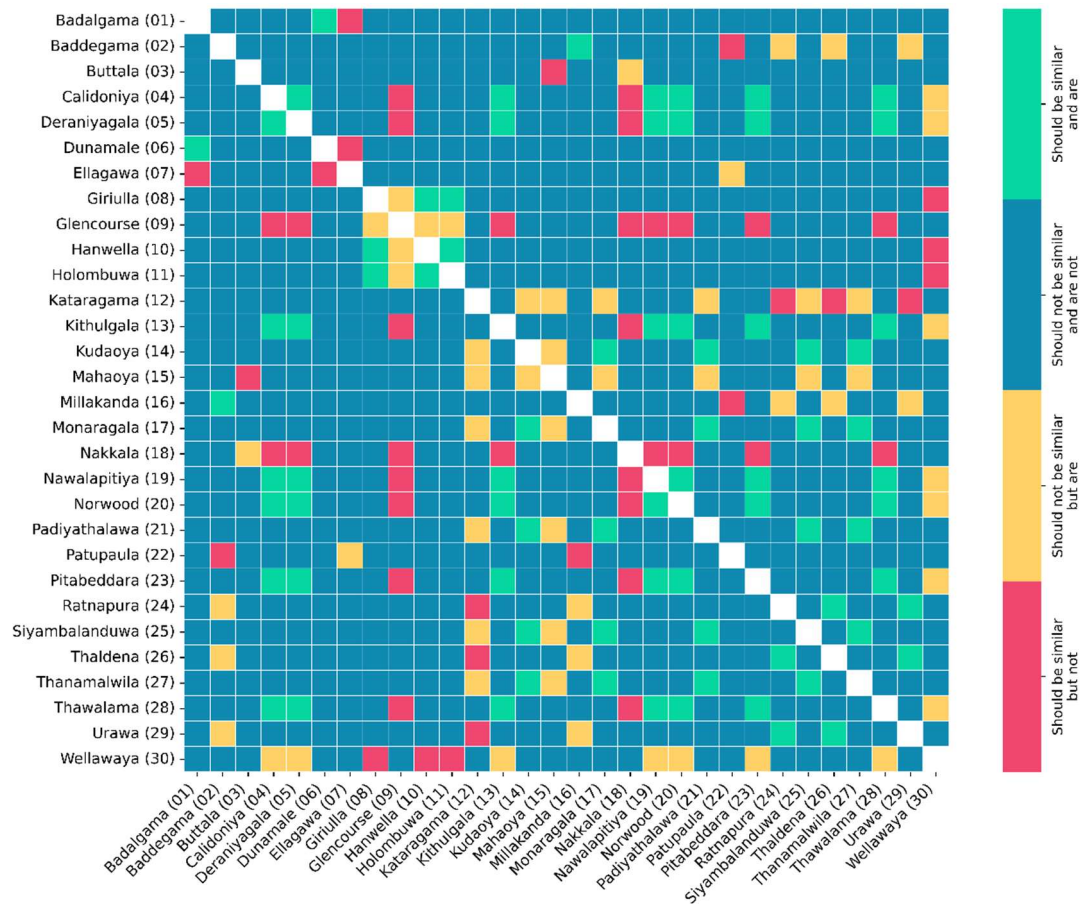


Figure 4-15: Pairwise Comparison Matrix of the original and predicted cluster distributions

Table 4-4: Pairwise accuracy results

Pair Type	Description	No of Pairs	Percentage
Inconsistent pairs (380)	Should be similar and are	39	9.0%
	Should not be similar and are not	341	78.3%
Consistent pairs (55)	Should be similar but not	26	6.0%
	Should not be similar but are	29	6.7%

Pair Type	Description	No of Pairs	Percentage
Total pairs (435)	Pairwise accuracy	$341 + 39$ $= 380$	87.4%
Identical pairs (39)	Similarity accuracy	$39 / (39 + 26)$ $= 39/65$	60%

According to the above table, our predicted signatures achieved 87.4% accuracy. However, similarity accuracy is just 60%. Similarity accuracy measures whether the original cluster alignment is precisely identical to the predicted signature-based cluster alignment. Since it is only 60%, this is just satisfactory with limited training data. Most catchments in the predicted signature-based cluster showed borderline behavior, which is the main reason. With 60% accuracy, donor catchment clusters can still be identified for ungauged catchments using this method.

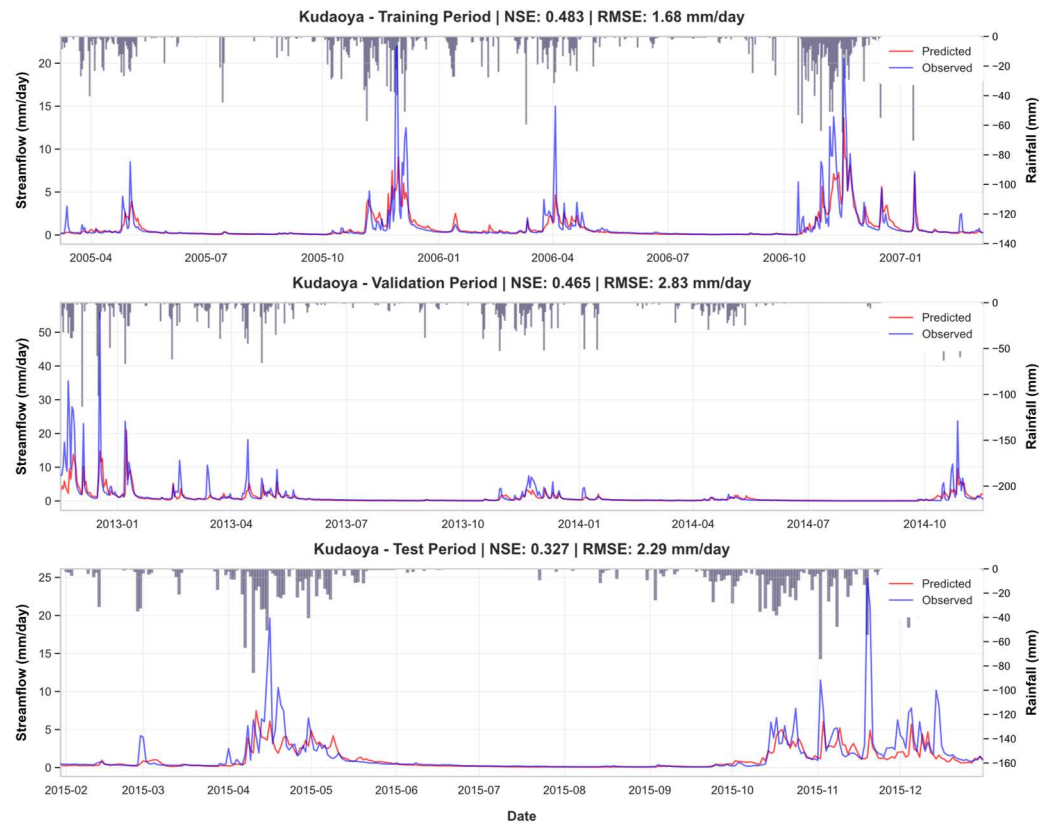
4.7 Regional LSTM Performance

For this study, two gauged catchments in the wet and dry zones were used as pseudo-ungauged catchments for the evaluation: Ellagawa (wet zone) and Monaragala (dry zone). Since satellite-based predicted signature clusters achieved 60% similarity, we used the original signature-based cluster arrangement. One main goal of this study is to validate whether regionalized LSTM can outperform global LSTM, for which accurate regional data is essential.

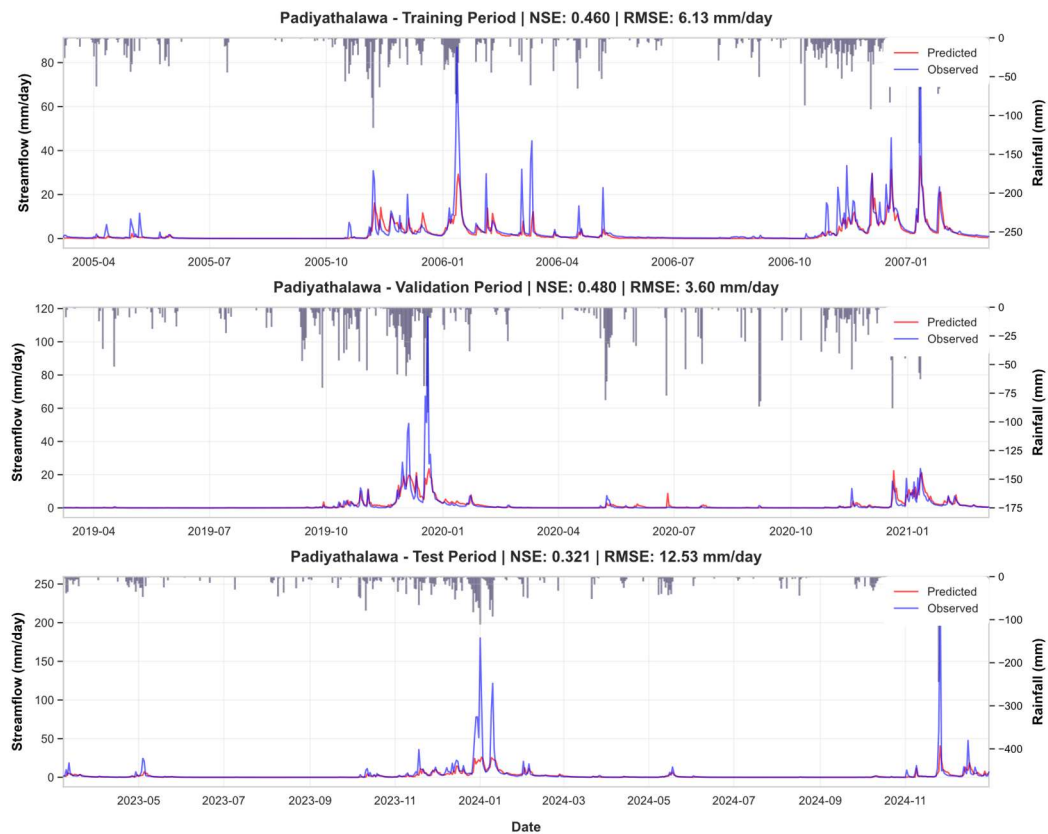
4.7.1 Dry zone: Monaragala case study

Monaragala clustered with four other dry zone catchments according to hierarchical clustering: Kudaoya, Padiyathalawa, Siyambalanduwa, and Thanamalwila. To predict runoff in pseudo-ungauged Monaragala, a regional LSTM was trained by inputting rainfall, satellite-based evapotranspiration, and six signature values. Evaluation metrics, NSE and RMSE, were used to evaluate the model training process. The pooled LSTM showed: Training: NSE = 0.43, RMSE = 4.13 mm/day; Validation: NSE = 0.46, RMSE = 2.86 mm/day; Testing: NSE = 0.33, RMSE = 7.96 mm/day, which is not

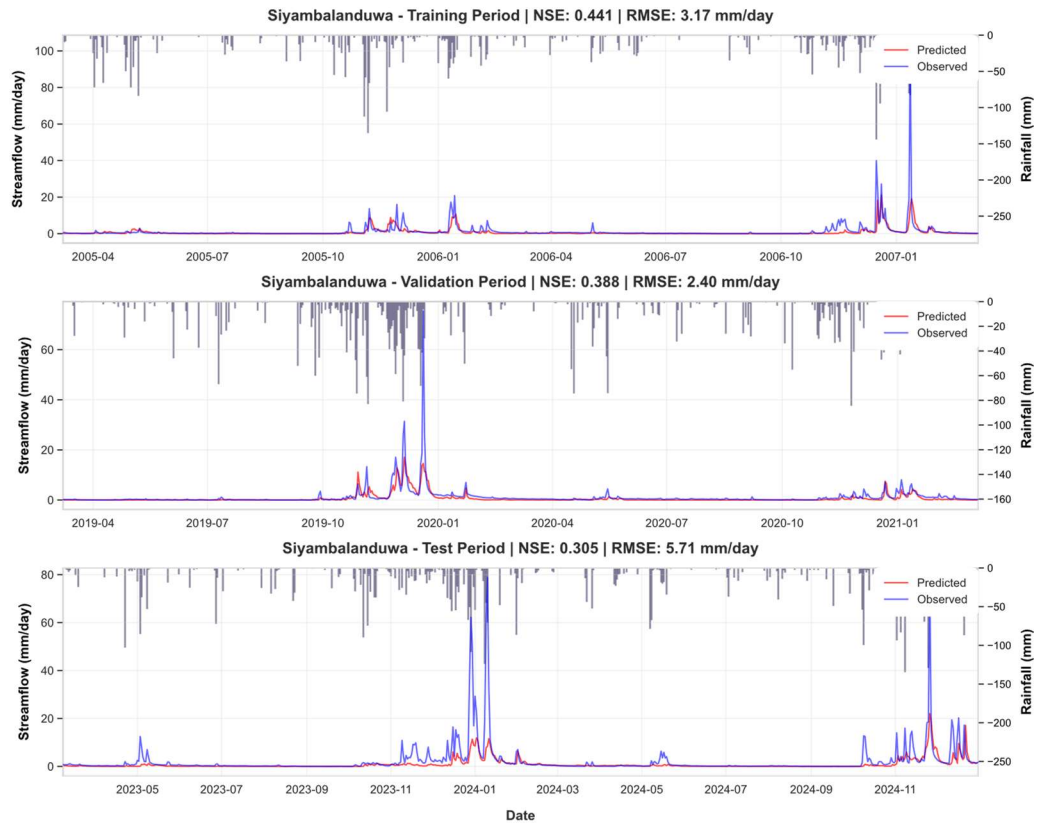
satisfactory. Each catchment's first two year of training, validation, and testing hydrograph is shown in Figure 4-16 (a)-(d).



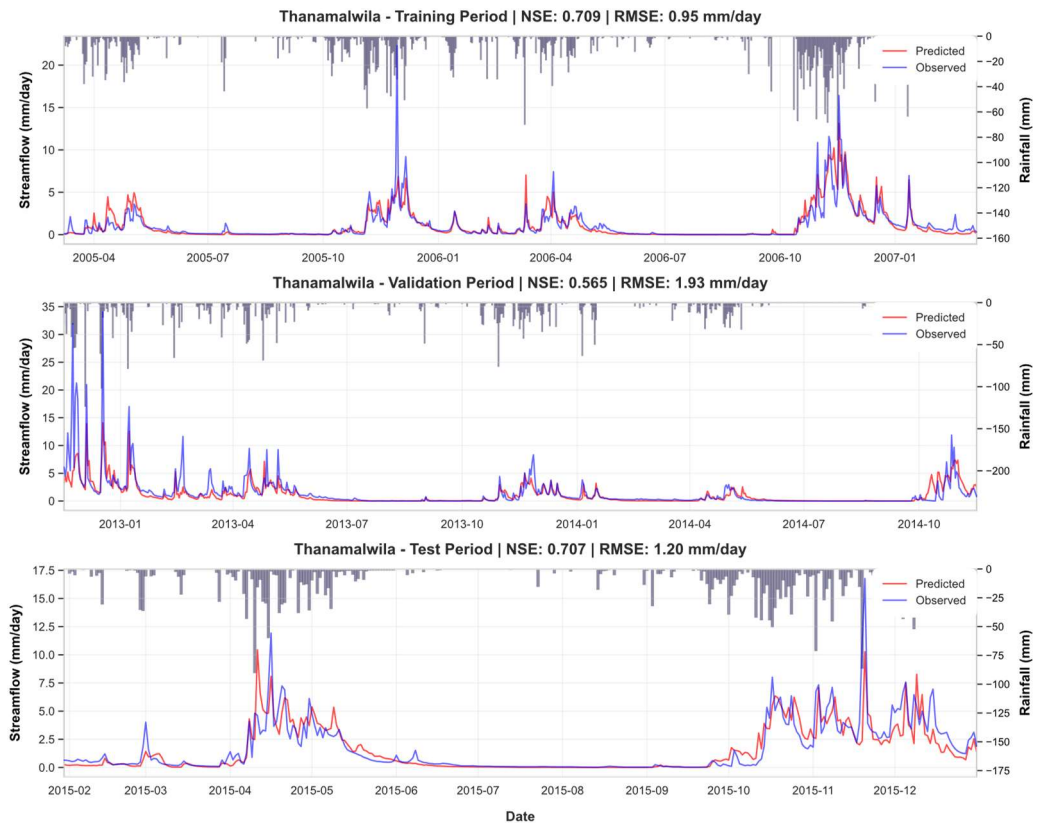
(a) Kudaoya



(b) Padiyathalawa



(c) Siyambalanduwa



(d) Thanamalwila

Figure 4-16: Regional LSTM simulated and observed hydrographs for first two years of data: Monaragala dry zone

According to the above data, the model achieved satisfactory learning only on the Thanamalwila data, and it failed to train effectively for peak prediction most of the time. The following figures show: Figure 4-17, the first two years of hydrograph model predictions for pseudo-ungauged Monaragala; and Figure 4-18, and Figure 4-19: the sorted and unsorted flow duration curve.

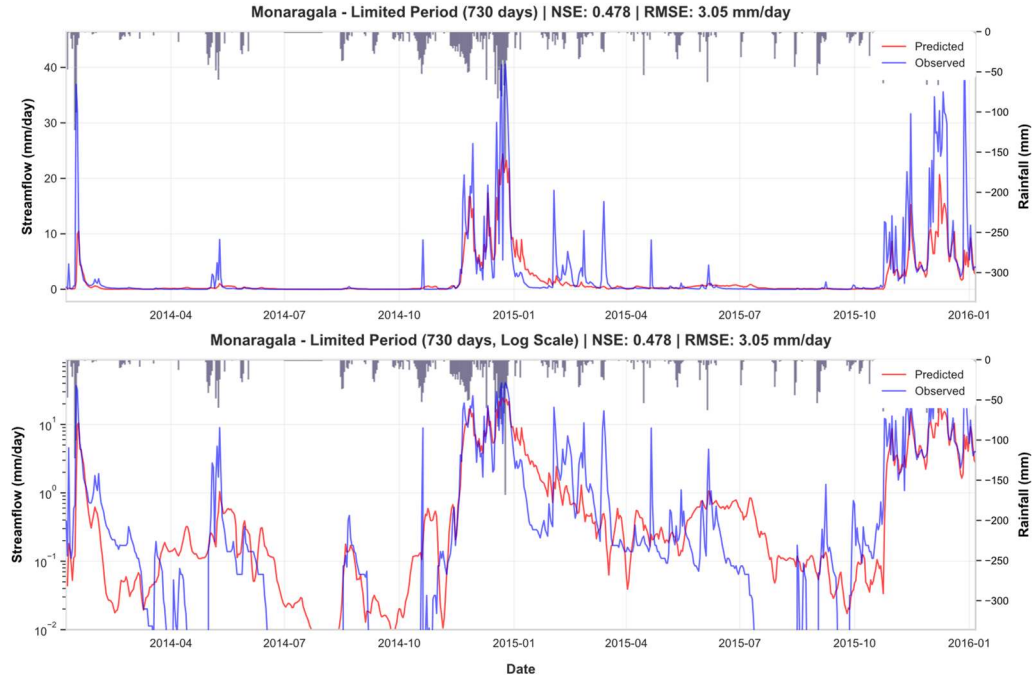


Figure 4-17: Observed and simulated hydrograph for pseudo-ungauged Monaragala (2014–2016)

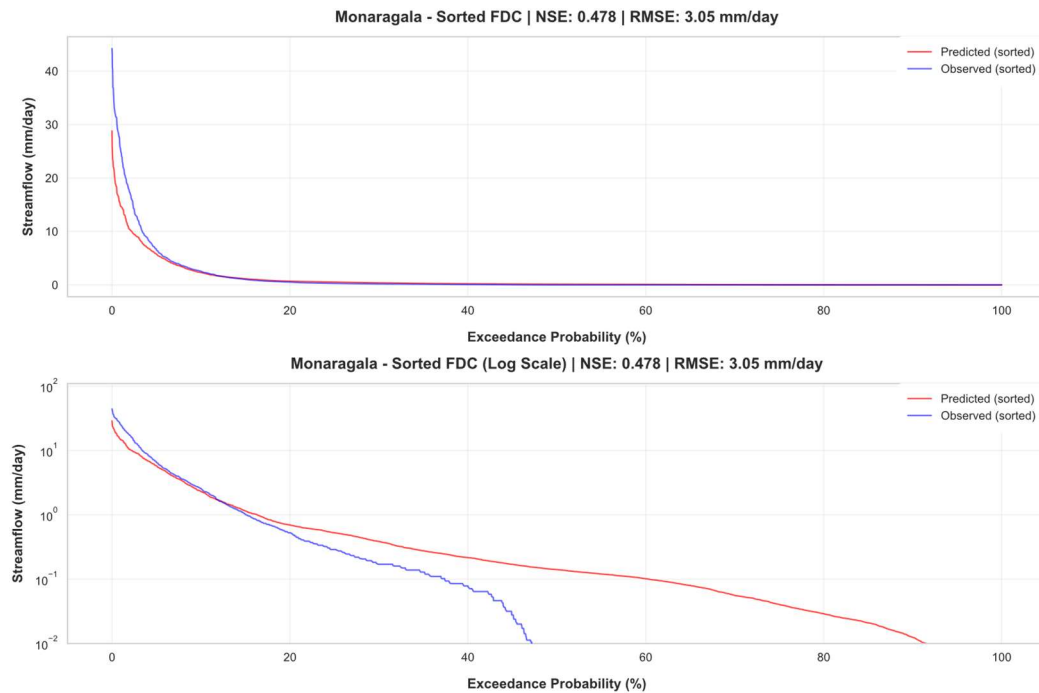


Figure 4-18: Predicted and observed sorted flow duration curve for pseudo-ungauged Monaragala (2014–2024)

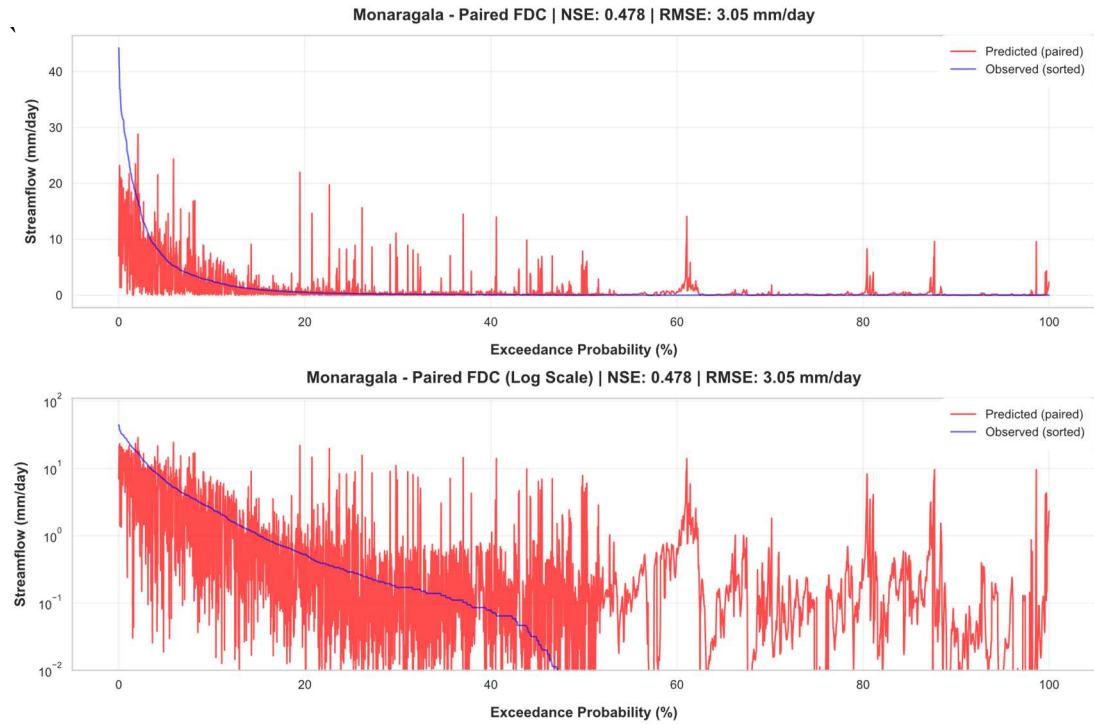
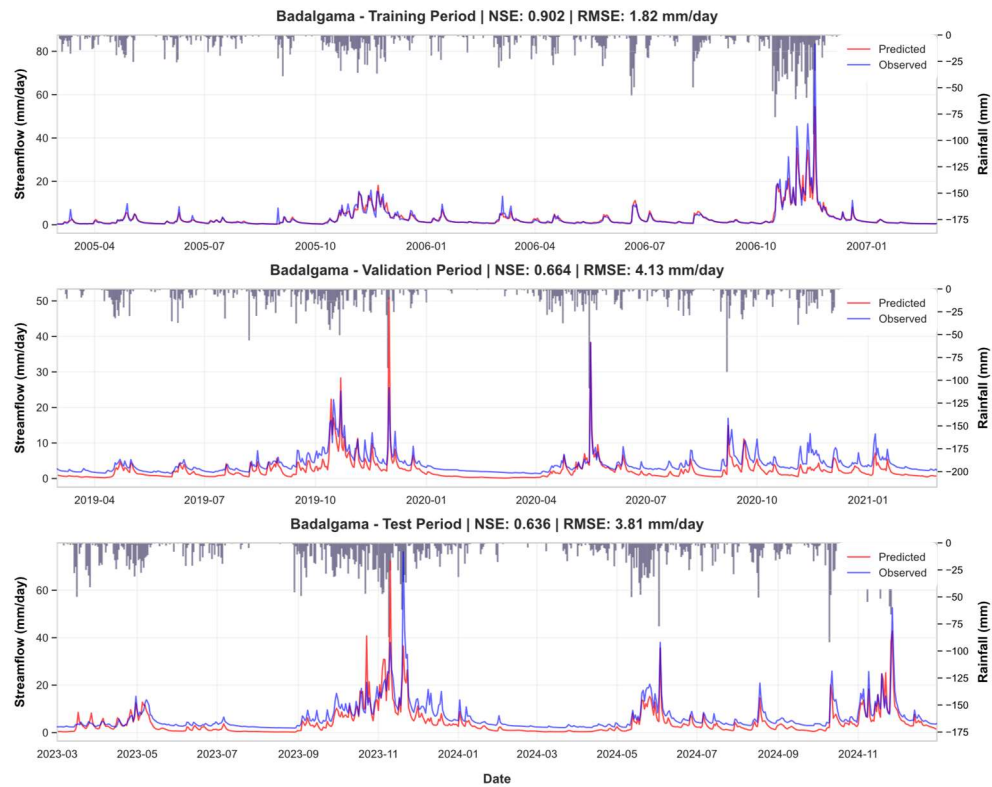


Figure 4-19: Predicted and observed unsorted flow duration curve for pseudo-ungauged Monaragala (2014–2024)

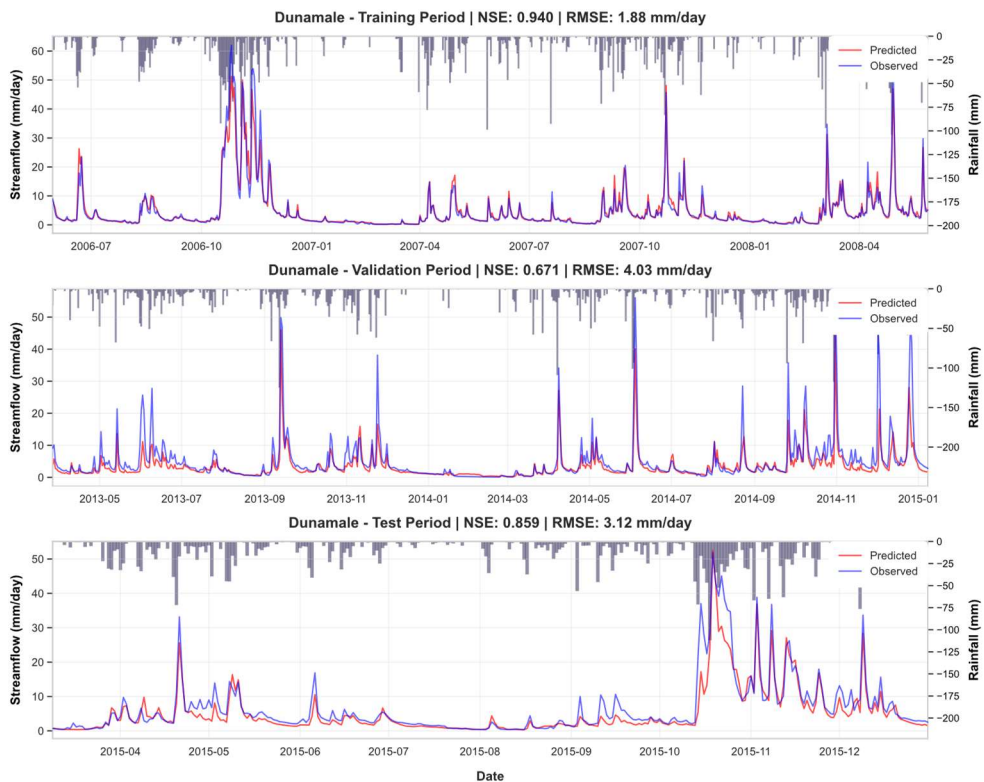
According to hydrographs, the model fails to predict small peaks, while extreme peaks are underpredicted. The flow duration curve confirms that Monaragala is a low-flow dominant catchment, and predictions fail to capture those low flows. Overall, for the dry zone, the regionalized LSTM did not show good performance. Annexure 4 presents the additional flow duration curves and hydrographs corresponding to the training and prediction data series relevant to regional LSTM: dry zone case study.

4.7.2 Wet zone: Ellagawa case study

According to hierarchical clustering, Ellagawa clustered with two other wet zone catchments: Badalgama and Dunamale. The regional LSTM was trained using Dunamale and Badalgama rainfall, evapotranspiration, and signature values to predict streamflow. The pooled LSTM showed satisfactory performance: Training: NSE = 0.92, RMSE = 1.84 mm/day; Validation: NSE = 0.66, RMSE = 4.10 mm/day; Testing: NSE = 0.73, RMSE = 3.16 mm/day. The first two years of training, validation, and testing hydrographs for each catchment are shown in Figure 4-20.



(a) Badalgama



(b) Dunamale

Figure 4-20: Regional LSTM simulated and observed hydrographs for first two years of data: Ellagawa wet zone

According to visual evidence and metric values, the model learned well from the two-catchment data, which is satisfactory. The following figures show: Figure 4-21, the first two years of hydrograph model predictions for pseudo-ungauged Ellagawa; and Figure 4-22 and Figure 4-23, the sorted and unsorted flow duration curves respectively.

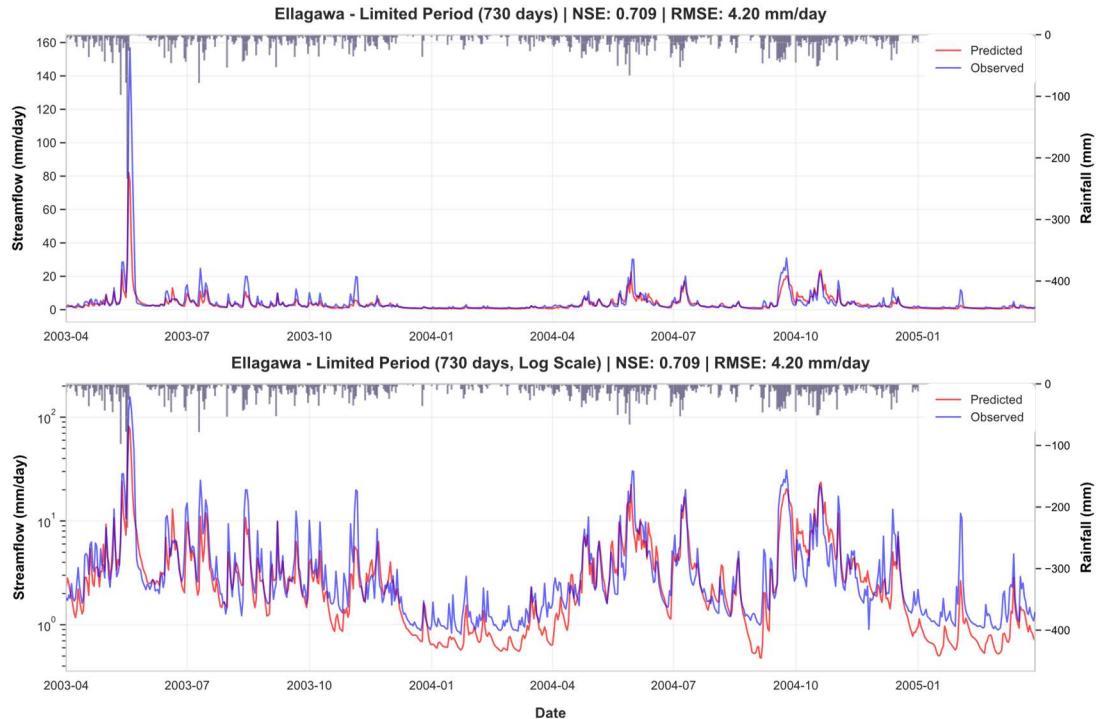


Figure 4-21: Observed and simulated hydrograph for pseudo-ungauged Ellagawa (2003–2005)

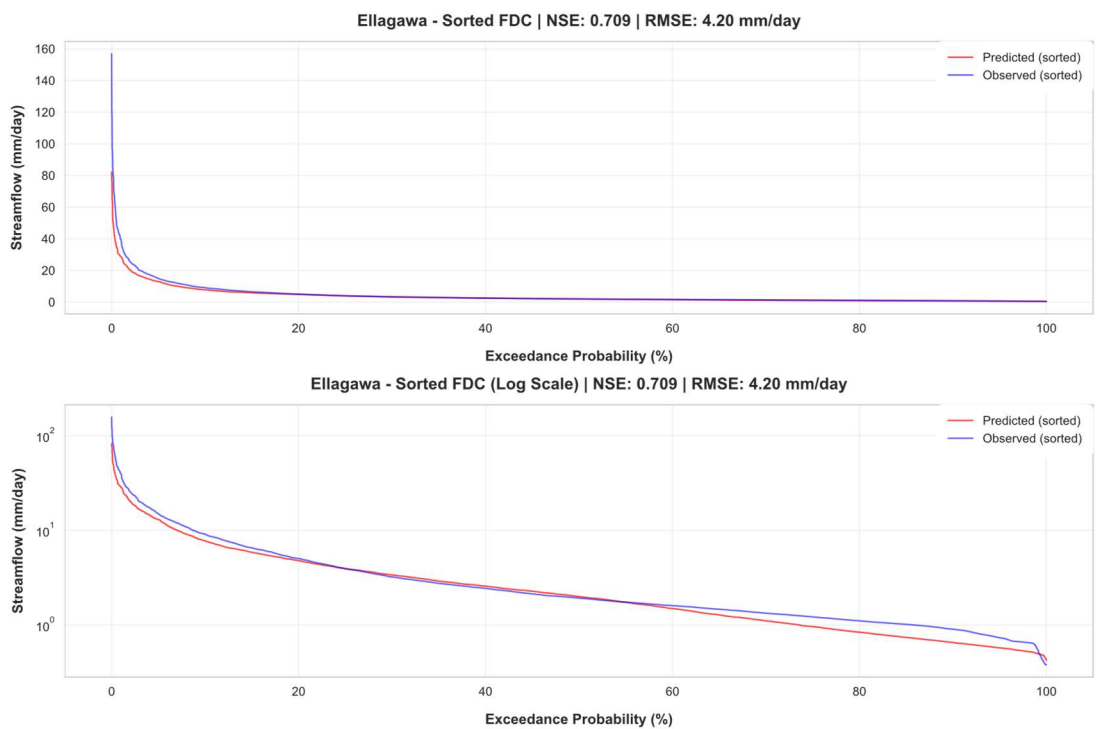


Figure 4-22: Predicted and observed flow duration curve for pseudo-ungauged Ellagawa (2003–2017)

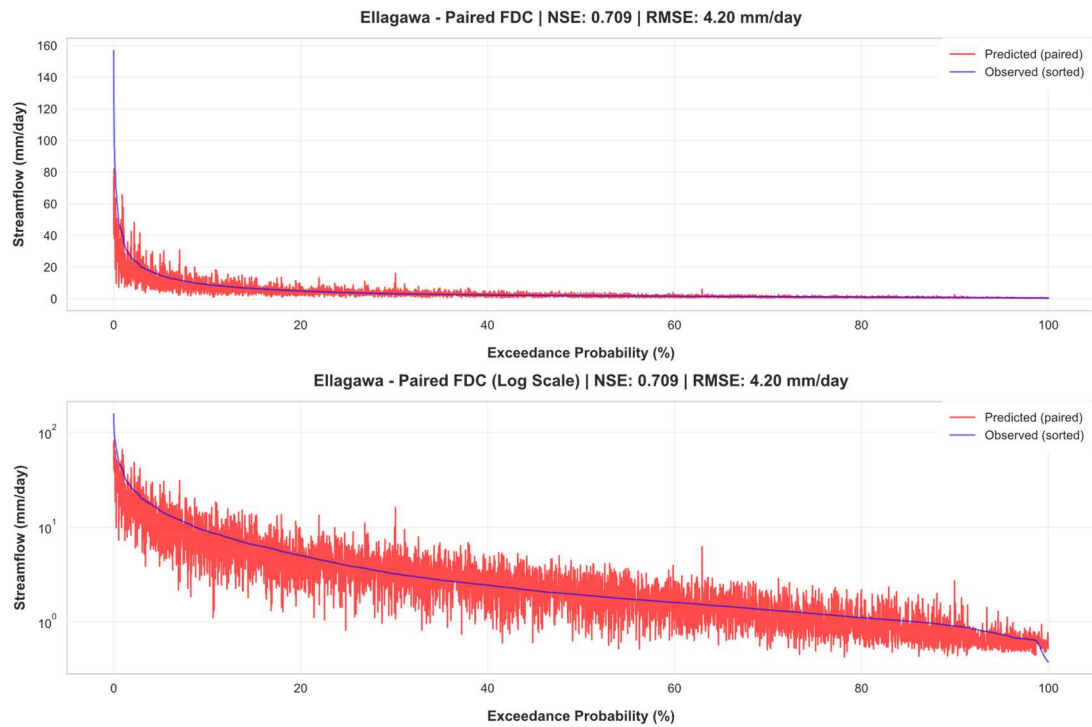


Figure 4-23: Predicted and observed flow duration curve for pseudo-ungauged Ellagawa (2003–2017)

The hydrograph and flow duration curve, along with evaluation metric values, show that the regional LSTM for the wet zone Ellagawa performs well. However, there is still some failure to predict extremes precisely. Additional streamflow hydrographs, along with sorted and unsorted flow duration curves related to the training and prediction phases of the regional LSTM case studies, are presented in Annexure 4.

4.8 Global LSTM Performance

The next step involved predicting two pseudo-ungauged catchments using a global LSTM, by training it with data from all catchments.

4.8.1 Dry zone: Monaragala case study

This time, all 30 catchment data sets were fed to the LSTM model to predict Monaragala pseudo-ungauged streamflow. The global LSTM achieved Training: NSE = 0.62, RMSE = 4.13 mm/day; Validation: NSE = 0.61, RMSE = 4.38 mm/day; Testing: NSE = 0.59, RMSE = 5.22 mm/day, reflecting satisfactory though not superior learning. Figure 4-24 shows the first two years of hydrograph model

predictions for pseudo-ungauged Monaragala, and Figure 4-25, Figure 4-27 show sorted and unsorted flow duration curves respectively.

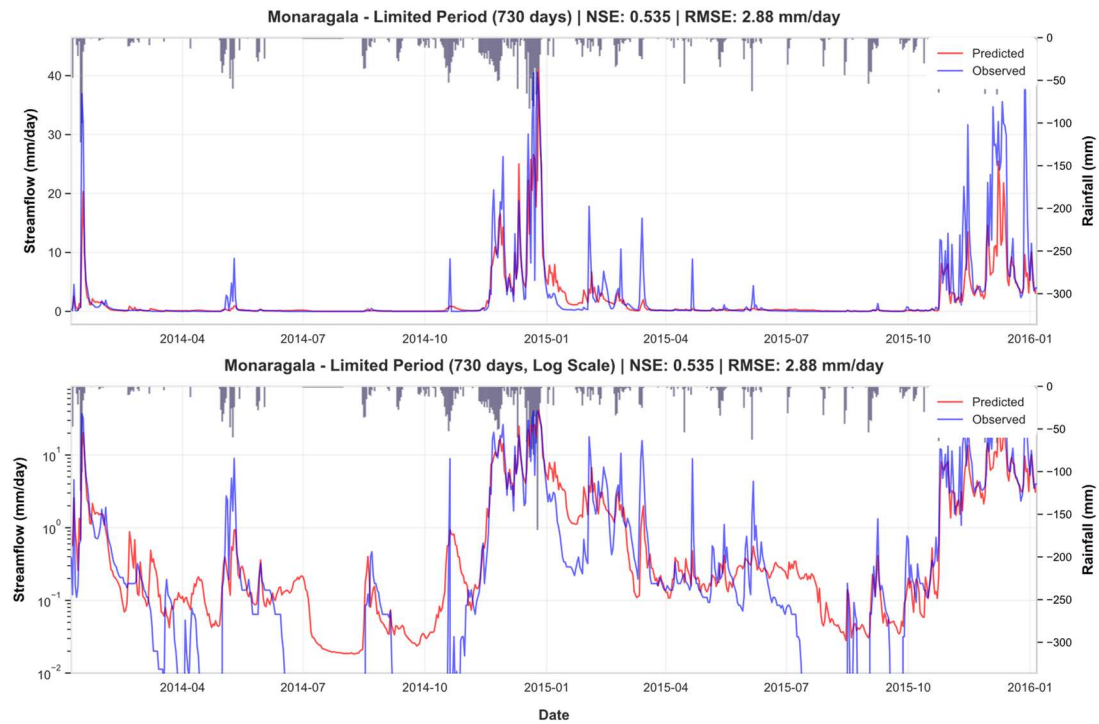


Figure 4-24: Observed and simulated hydrograph for pseudo-ungauged Monaragala (2014–2016)

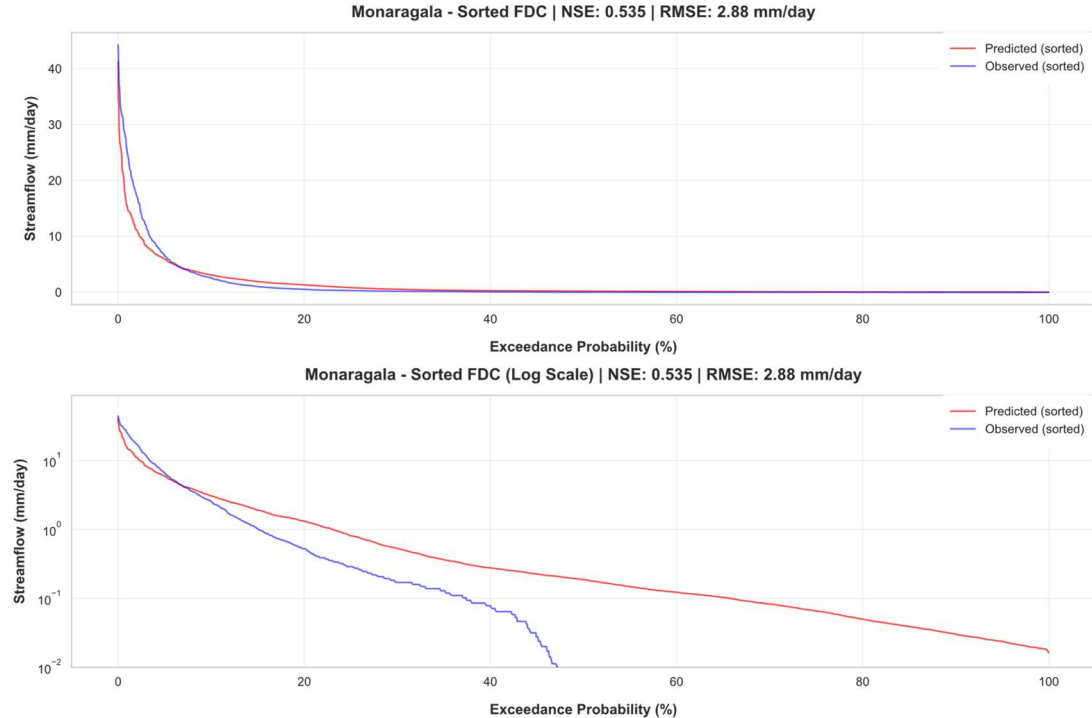


Figure 4-25: Predicted and observed sorted flow duration curve for pseudo-ungauged Monaragala (2014–2024)

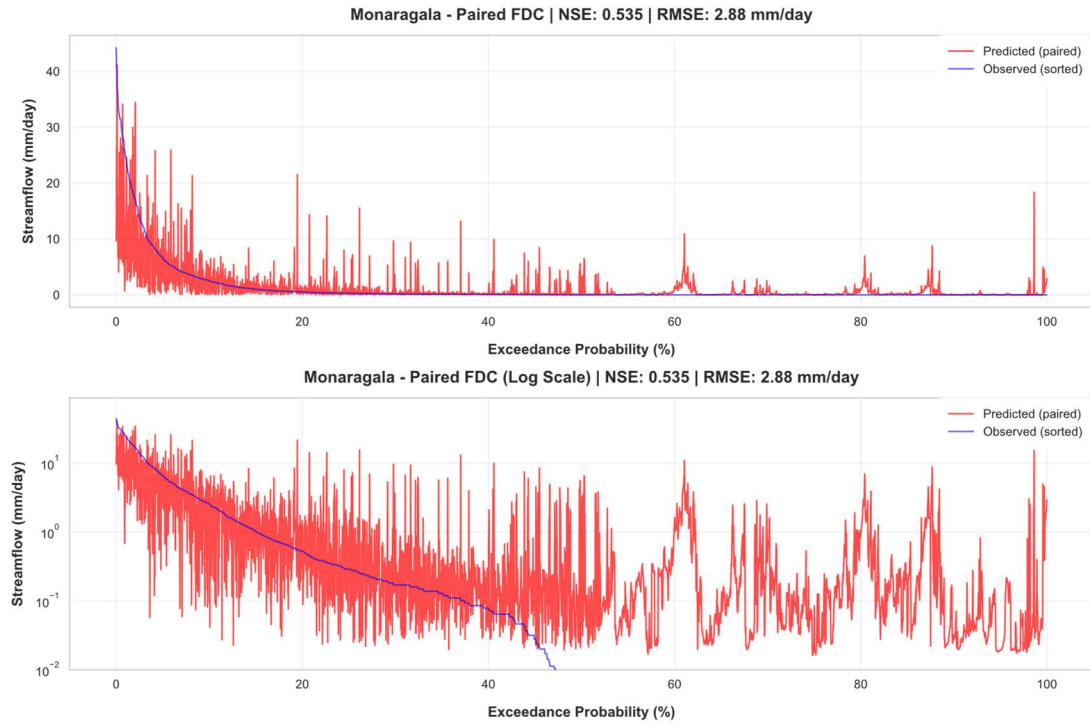


Figure 4-26: Predicted and observed sorted flow duration curve for pseudo-ungauged Monaragala (2014–2024)

It shows better NSE and RMSE values than the regional LSTM, but the results are still not satisfactory. The flow duration curve indicates that the global LSTM model also fails to predict low flows in Monaragala. However, this global LSTM shows some improvement in extreme event prediction. The complete streamflow hydrograph for the full prediction period of the Monaragala catchment is provided in Annexure 5.

4.8.2 Wet zone: Ellagawa case study

The global LSTM for the wet zone case study achieved Training: NSE = 0.66, RMSE = 3.89 mm/day; Validation: NSE = 0.67, RMSE = 4.01 mm/day; Testing: NSE = 0.65, RMSE = 4.76 mm/day, reflecting satisfactory but not superior learning. However, these evaluation metrics are lower than those of the regional LSTM for the wet zone case study. Figure 4-27 shows the first two years of hydrograph model predictions for pseudo-ungauged Monaragala, and Figure 4-28, Figure 4-29 show the sorted and unsorted flow duration curves respectively.

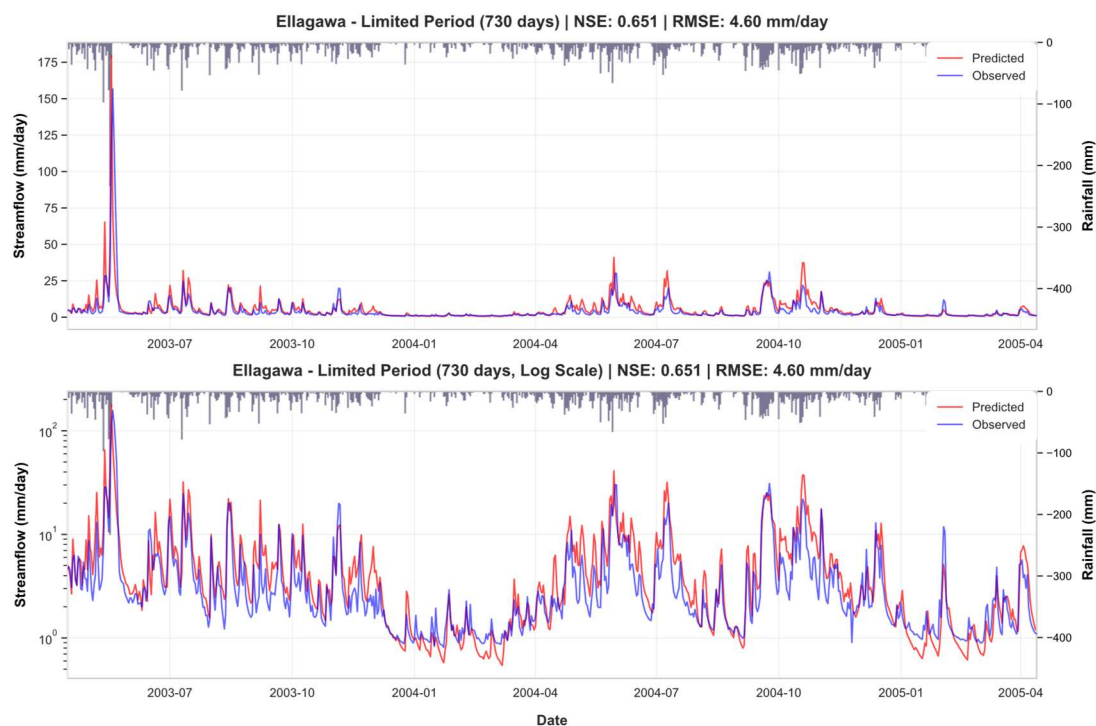


Figure 4-27: Observed and simulated hydrograph for pseudo-ungauged Ellagawa (2003–2005)

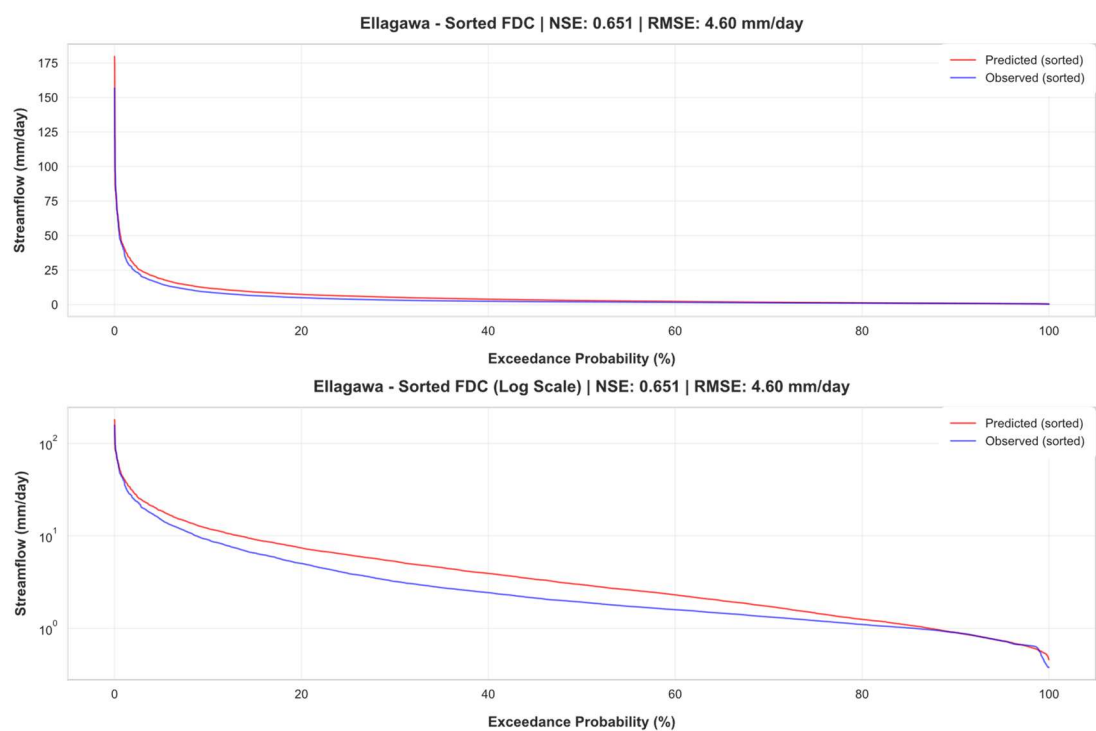


Figure 4-28: Predicted and observed sorted flow duration curve for pseudo-ungauged Ellagawa (2003–2017)

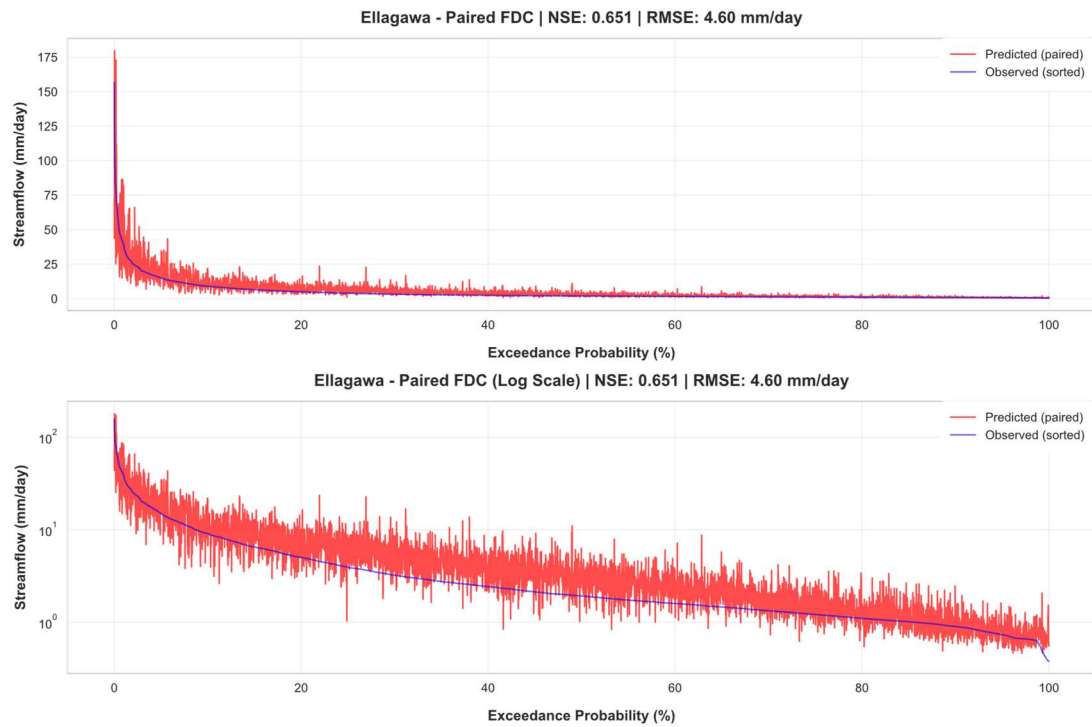


Figure 4-29: Predicted and observed unsorted flow duration curve for pseudo-ungauged Ellagawa (2003–2017)

The model shows moderate performance with an NSE value of 0.65 and an RMSE of 4.60 mm/day. The flow duration curve indicates that the model predicts extremes but fails in most mid-flow ranges, and it overpredicts low flows. The complete streamflow hydrograph for the full prediction period of the Monaragala catchment is provided in Annexure 5.

4.9 HEC-HMS Model Setup and Parameterization

This step involves building two HEC-HMS models for two proximate catchments of the ungauged basins and using calibrated parameters to predict runoff in the ungauged catchments. Padiyathalawa was used as the proximate catchment for Monaragala in the dry zone, while Ratnapura was used for Ellagawa in the wet zone.

4.9.1 Dry zone case study: Parameter transfer in Monaragala catchment

Padiyathalawa's 20 years of data were used, with 10 years for calibration (2005–2015) and 10 years for validation (2015–2025). During calibration, the model showed $NSE = 0.65$ and $RMSE = 4.86$ mm/day. For validation, $NSE = 0.65$ and $RMSE = 4.35$

mm/day. These results are satisfactory for hydrological modeling but not ideal. Figure 4-30 to Figure 4-32 show the hydrograph for the first two years of calibration and the flow duration curves for the calibration period. Figure 4-37 to Figure 4-35 present the same for the validation period.

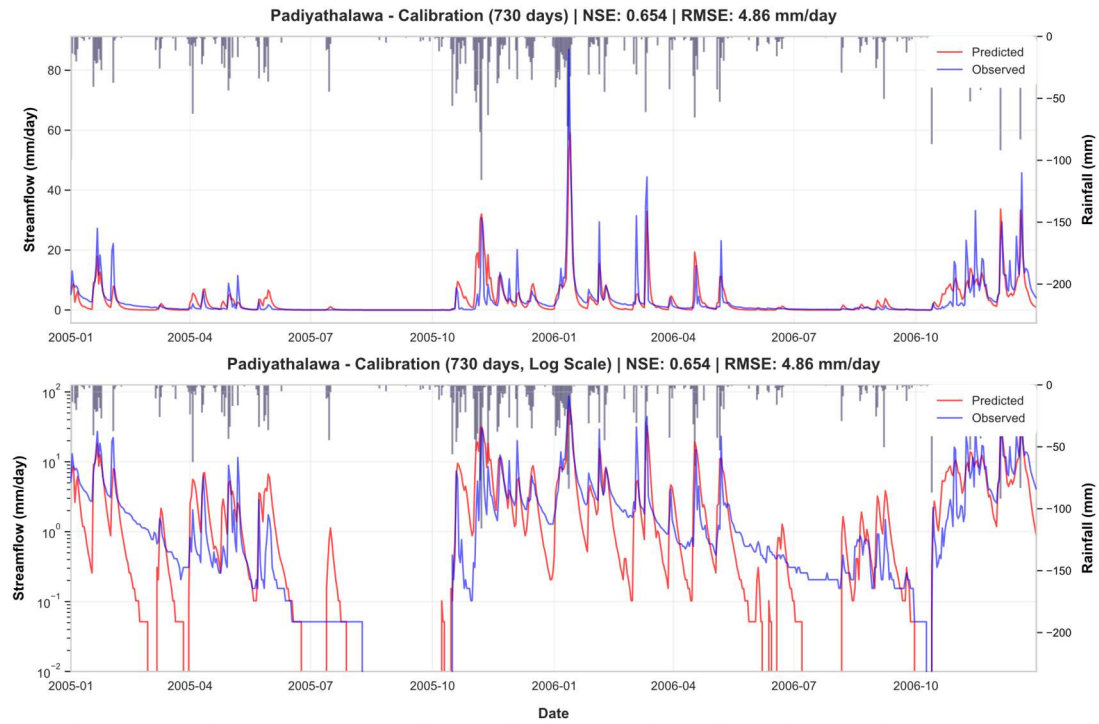


Figure 4-30: First two years of calibration hydrograph for Padiyathalawa catchment (2005–2007)

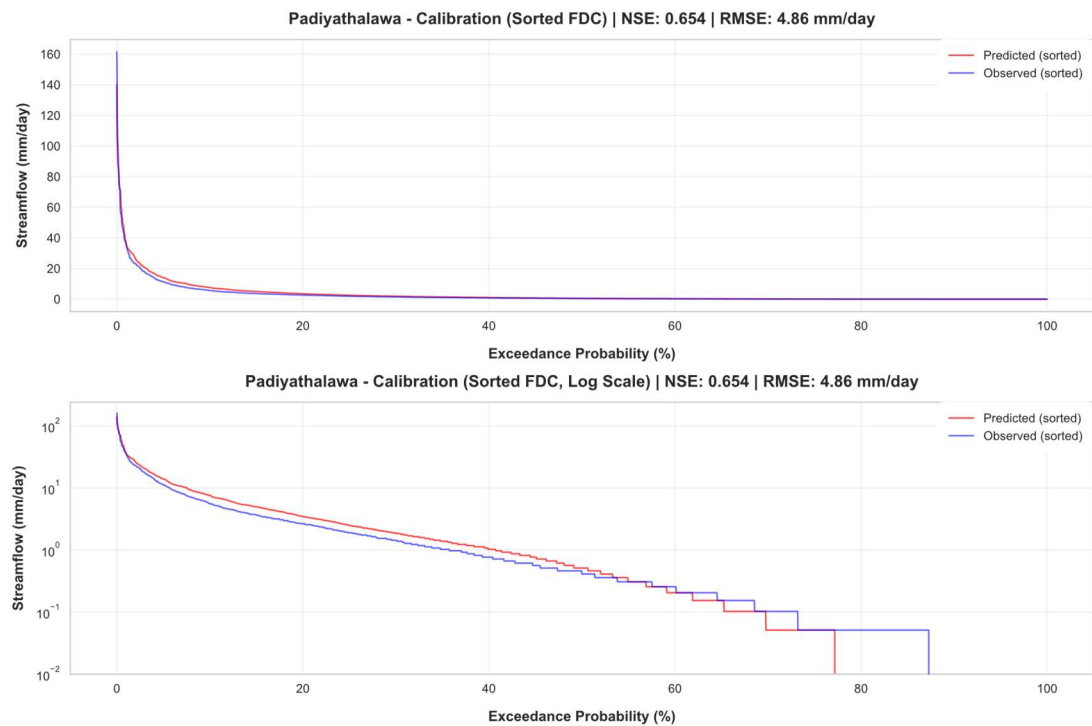


Figure 4-31: Sorted flow duration curve for Padiyathalawa catchment - calibration period (2005–2015)

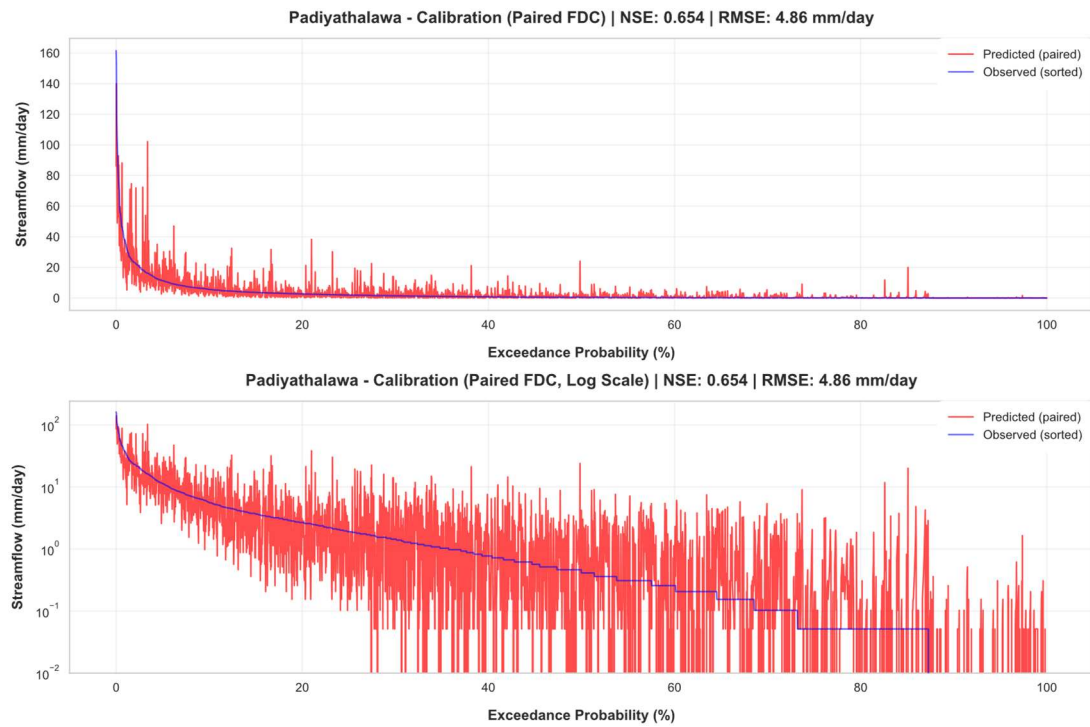


Figure 4-32: Unsorted flow duration curve for Padiyathalawa catchment - calibration period (2005–2015)

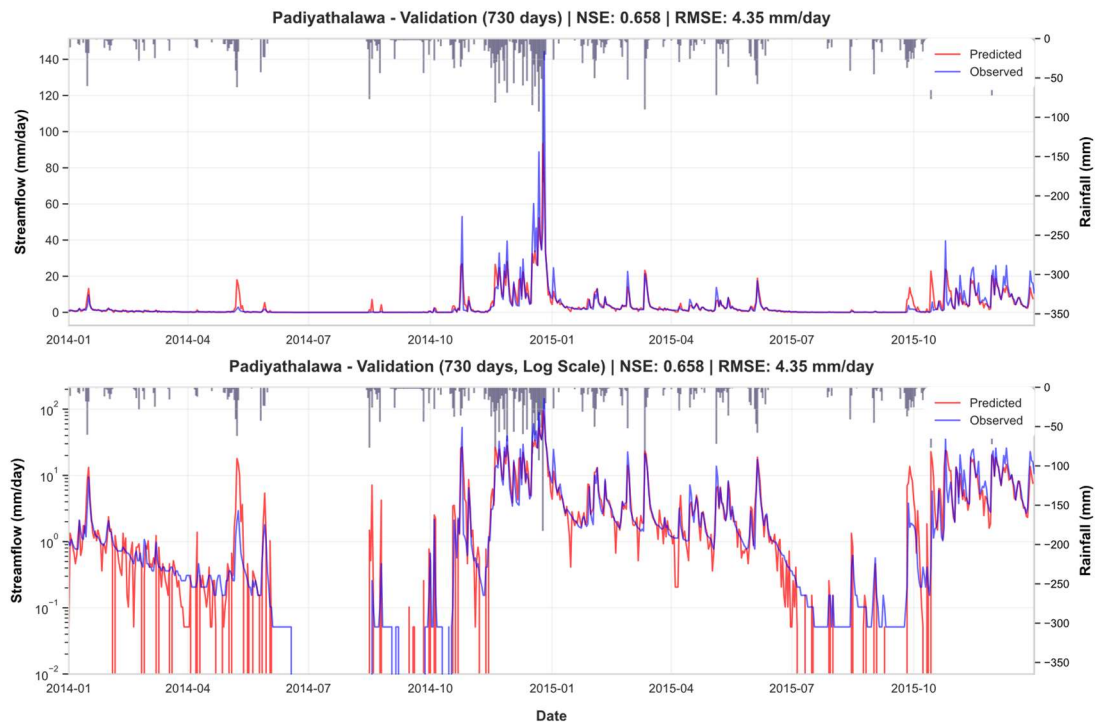


Figure 4-33: First two years of validation hydrograph for Padiyathalawa catchment (2015–2017)

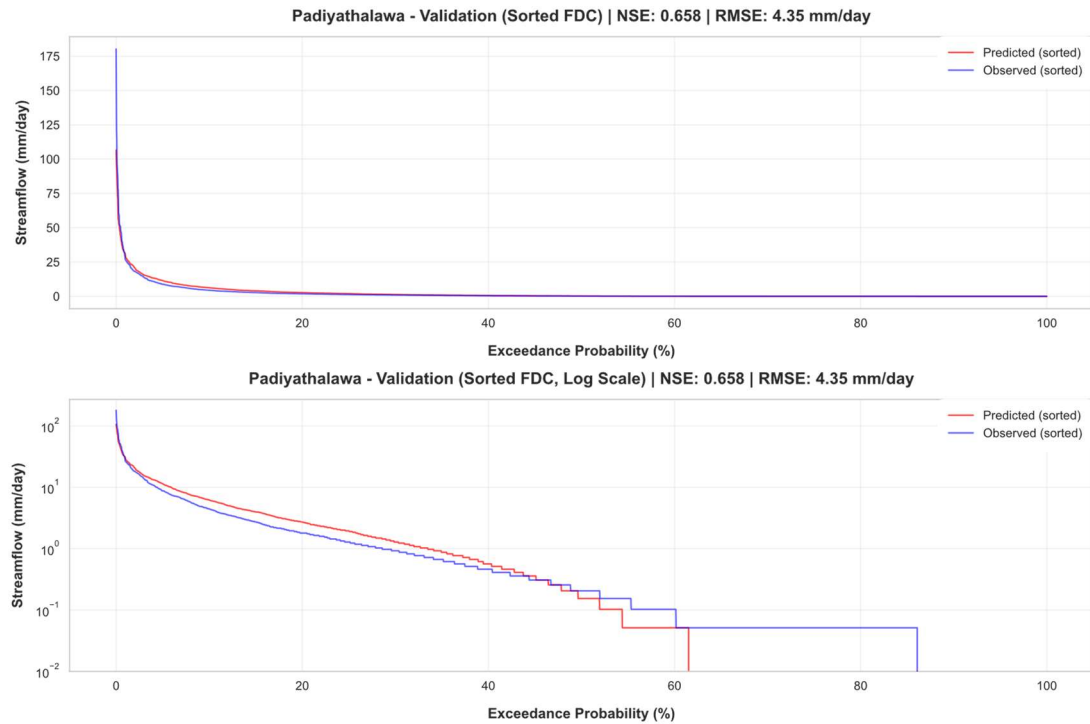


Figure 4-34: Sorted flow duration curve for Padiyathalawa catchment - validation period (2015–2025)

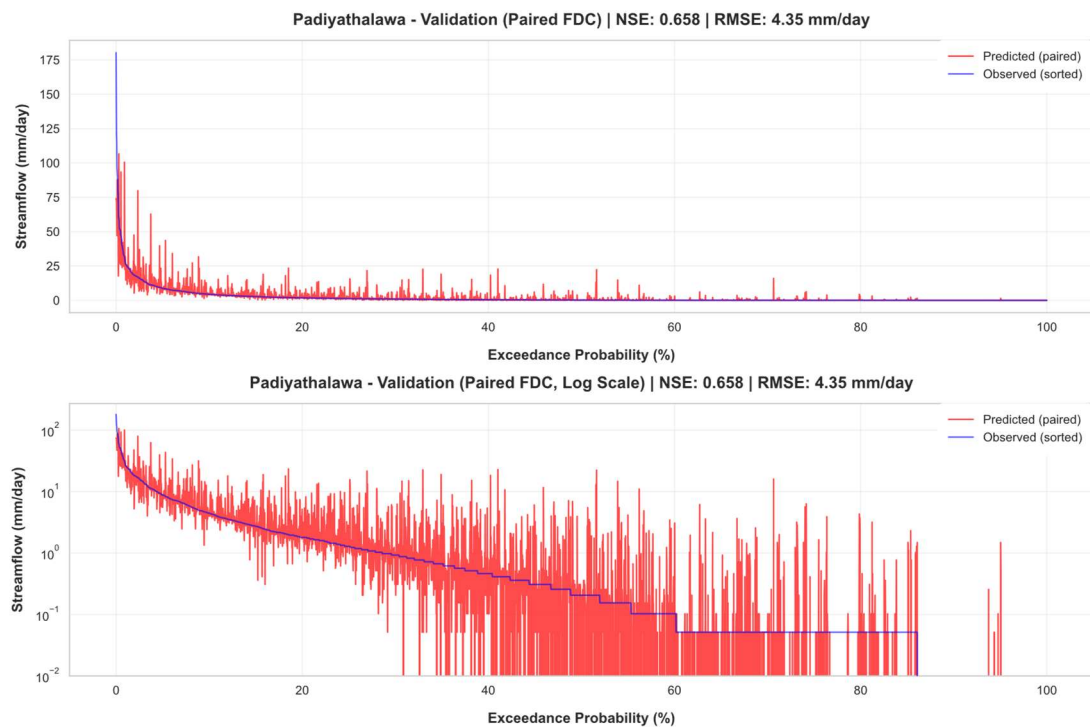


Figure 4-35: Unsorted flow duration curve for Padiyathalawa catchment - validation period (2015–2025)

According to the hydrograph visuals and flow duration curve for calibration, the model predicts high flows and overpredicts mid flows, while low flows are not captured

clearly. The observed data also show some quality issues. In validation, the model again predicts extremes but overpredicts mid flows and clearly fails to predict low flows. Calibrated parameters are shown in Annexure 7. Figure 4-36 to Figure 4-38 display the results of running the HEC-HMS model with the same parameter set for Monaragala.

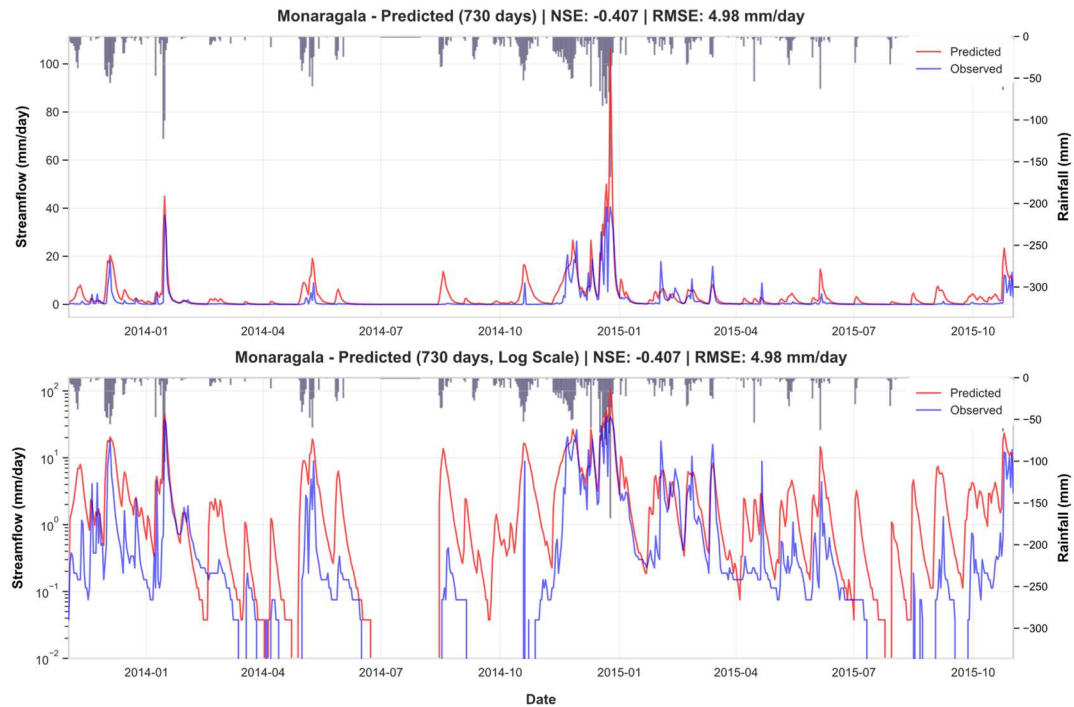


Figure 4-36: First two years of simulated vs observed hydrograph for Monaragala catchment using transferred parameters (2014–2016)

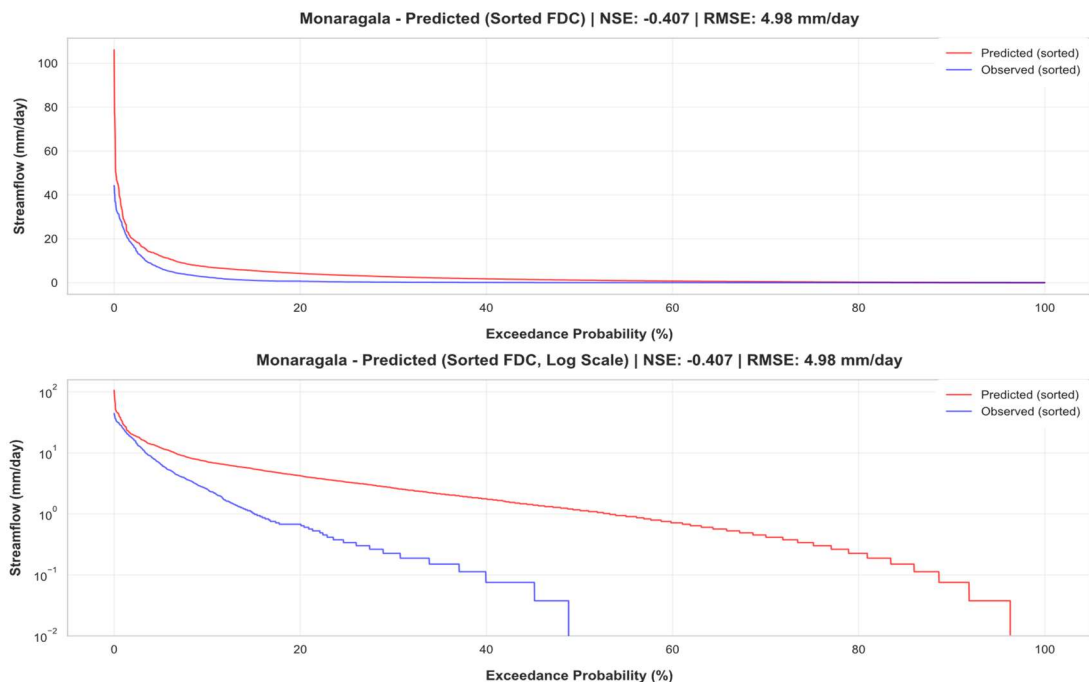


Figure 4-37: Sorted flow duration curve for Monaragala catchment using transferred parameters (2014–2024)

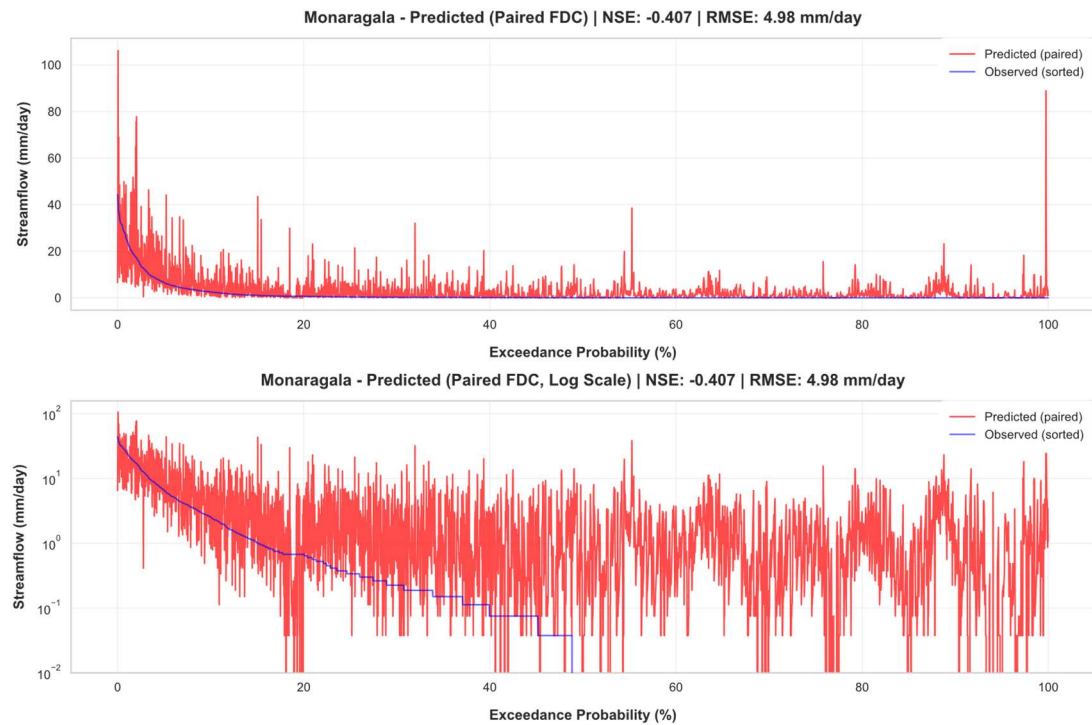


Figure 4-38: Sorted flow duration curve for Monaragala catchment using transferred parameters (2014–2024)

Annexure 6 contains the complete calibration and validation hydrographs for the full study period for donor and ungauged catchment, displayed in both linear and logarithmic scales. According to the NSE and RMSE values, as well as the plots above, the prediction was not sufficient. The hydrograph shows peaks in the predictions that do not correspond to observed values, while the flow duration curve indicates no match between predicted and observed data. This confirms that streamflow prediction for the dry zone Monaragala Catchment using transferred parameters in HEC-HMS was not successful.

4.9.2 Wet zone case study: Parameter transfer in the Ellagawa catchment

For streamflow prediction of the wet zone pseudo-ungauged catchment, Ellagawa, Ratnapura, a neighbouring catchment, was selected. Ratnapura provided 12 years of data (2005–2017), with 7 years used for calibration (2005–2012) and 5 years for validation (2012–2017). Figure 4-39 to Figure 4-41 show the hydrograph for the first two years of calibration and the flow duration curves for the calibration period. Figure 4-42 and Figure 4-44 present the same for the validation period. Calibrated parameters are available in Annexure 7.

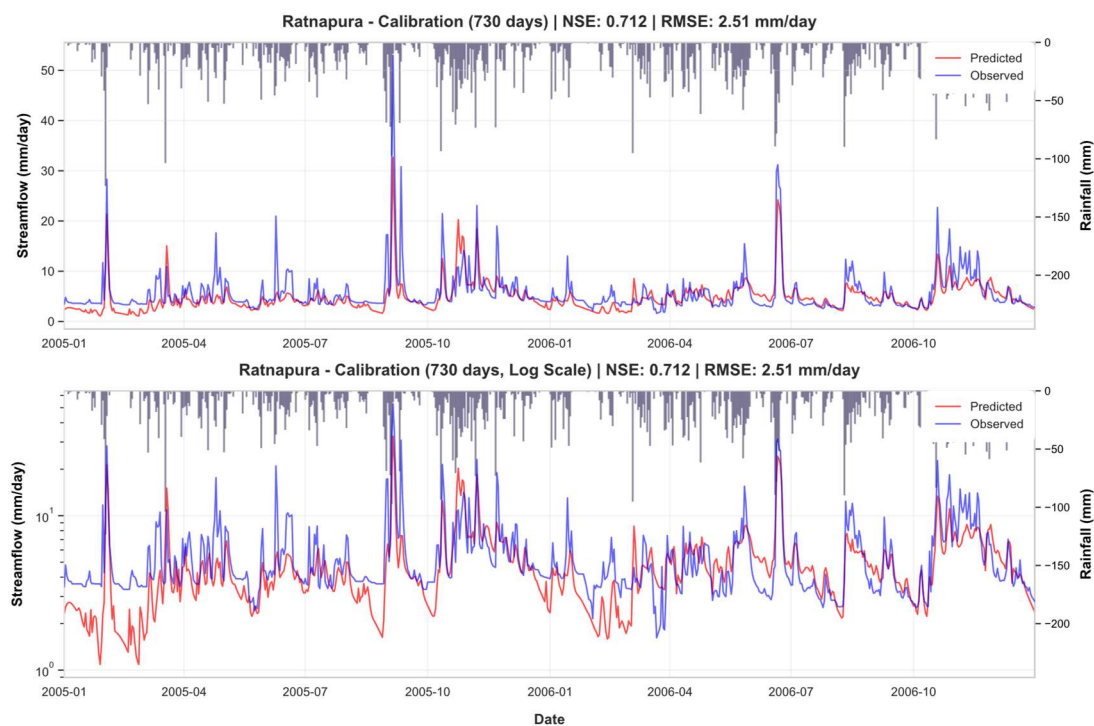


Figure 4-39: First two years of calibration hydrograph for Ratnapura catchment (2005–2007)

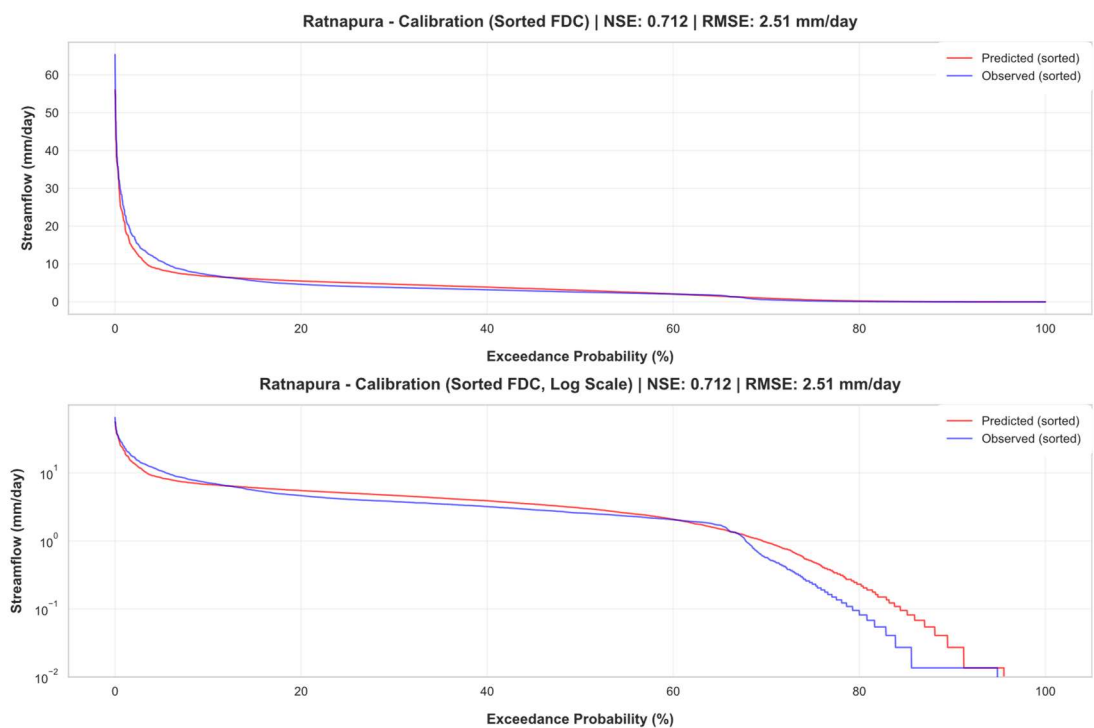


Figure 4-40: Sorted flow duration curve for Ratnapura catchment - calibration period (2005–2012)

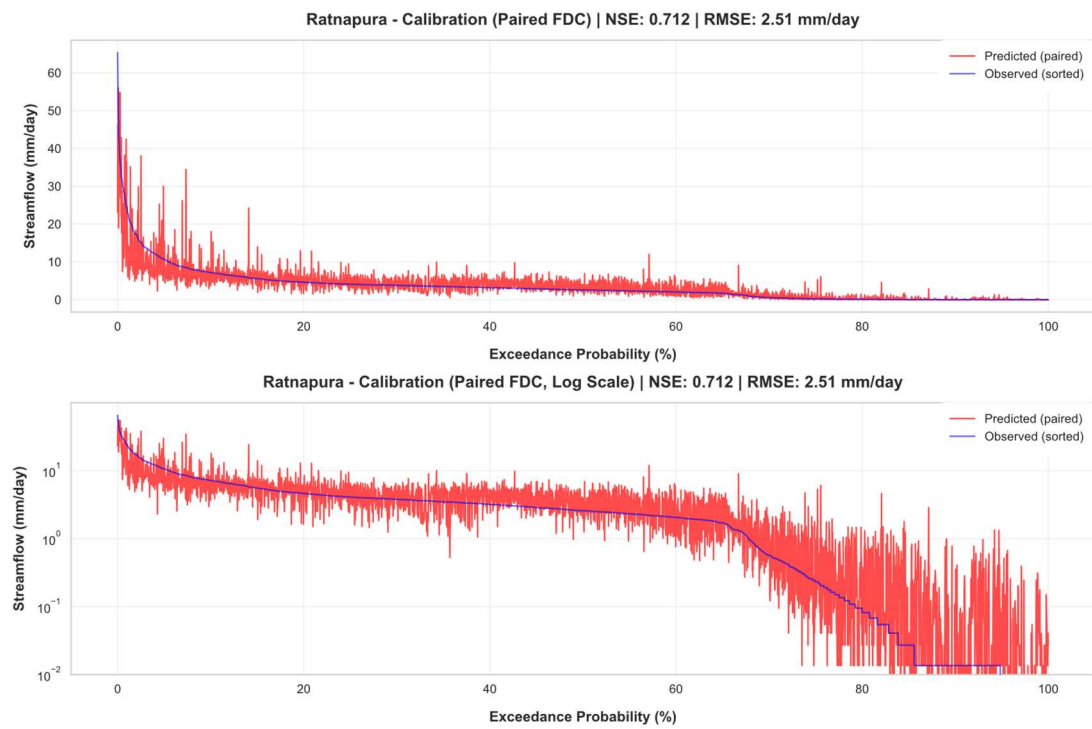


Figure 4-41: Sorted flow duration curve for Ratnapura catchment - calibration period (2005–2012)

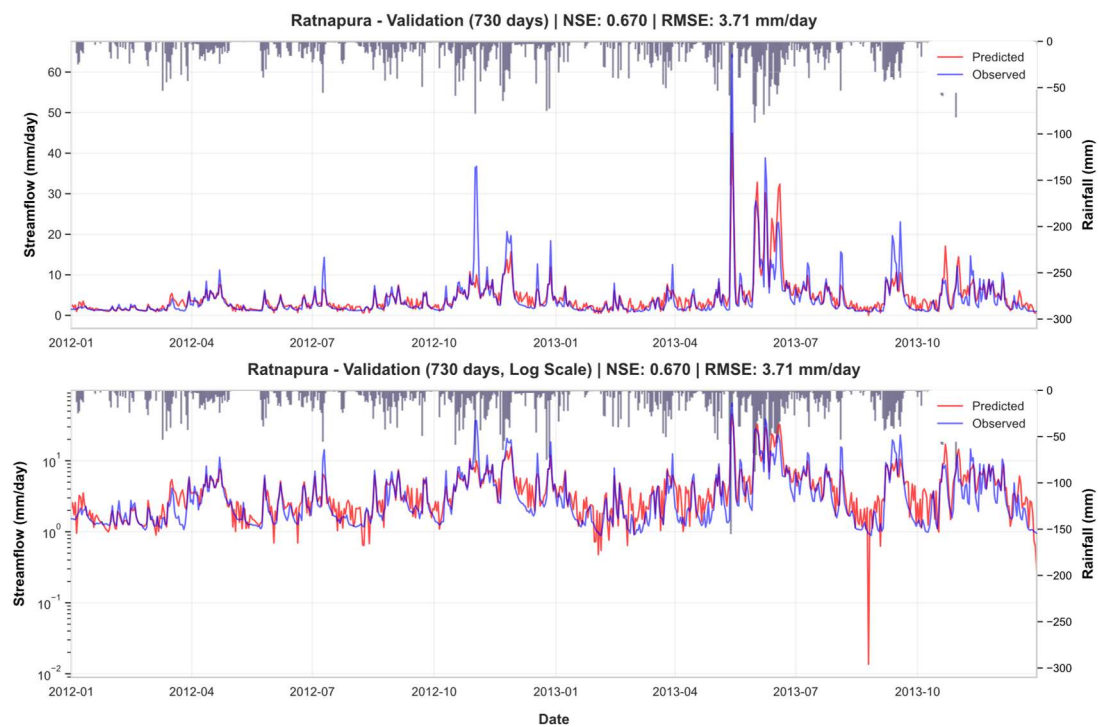


Figure 4-42: First two years of validation hydrograph for Ratnapura catchment (2012–2014)

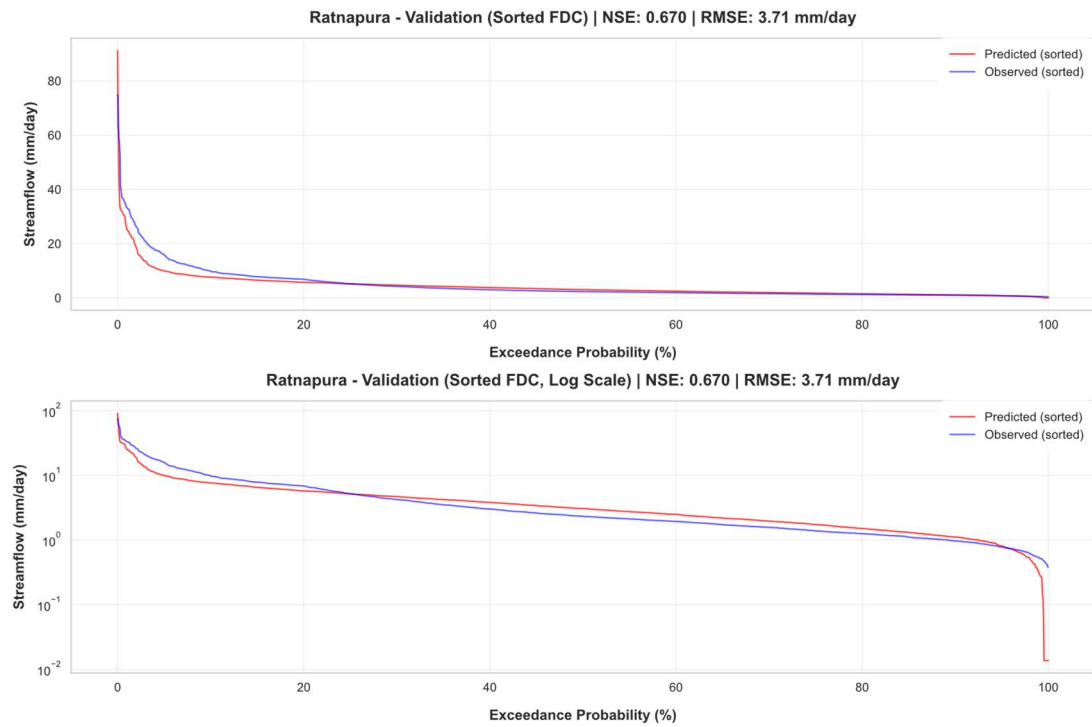


Figure 4-43: Sorted flow duration curve for Ratnapura catchment - validation period (2012–2017)

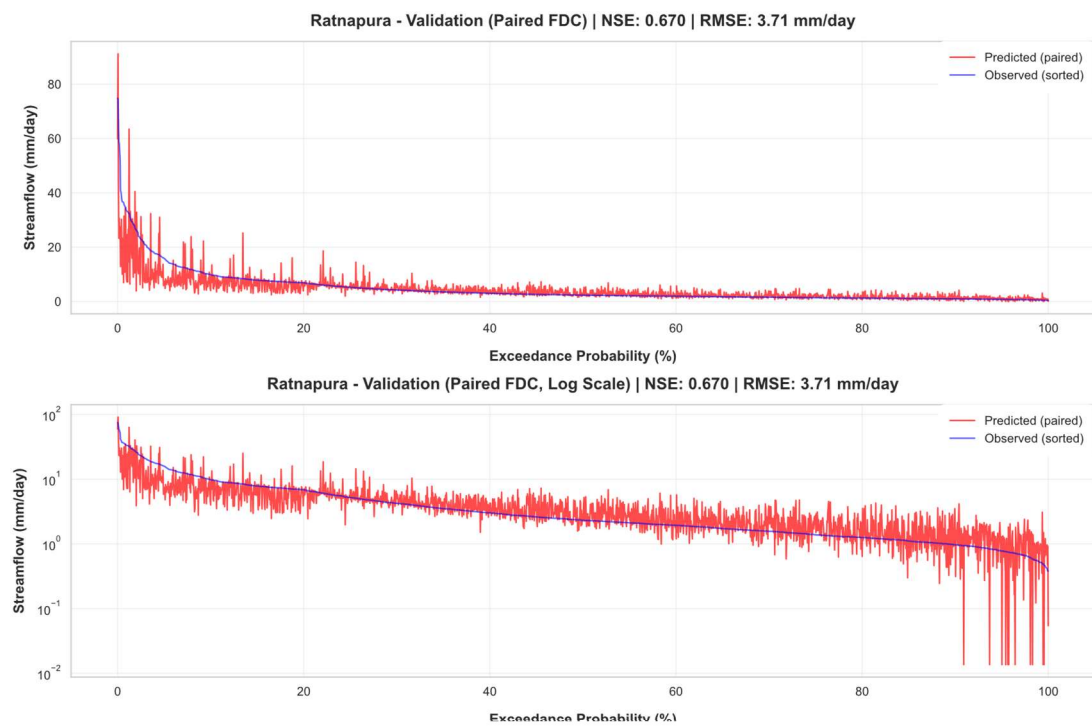


Figure 4-44: Unsorted flow duration curve for Ratnapura catchment - validation period (2012–2017)

During calibration, the model achieved an NSE of 0.71 and an RMSE of 2.51 mm/day. For validation, the model produced an NSE of 0.67 and an RMSE of 3.71 mm/day.

These results are satisfactory for hydrological modeling. According to the calibration flow duration curve, the model's calibrated parameters fail to predict low flows, which are overpredicted. Mid-flow and extreme predictions are satisfactory. The validation period shows similar behaviour: low flows are underpredicted, while mid and extreme flows are predicted satisfactorily. Figure 4-45 to Figure 4-47 display the results of running the HEC-HMS model with the same parameter set for Ellagawa.

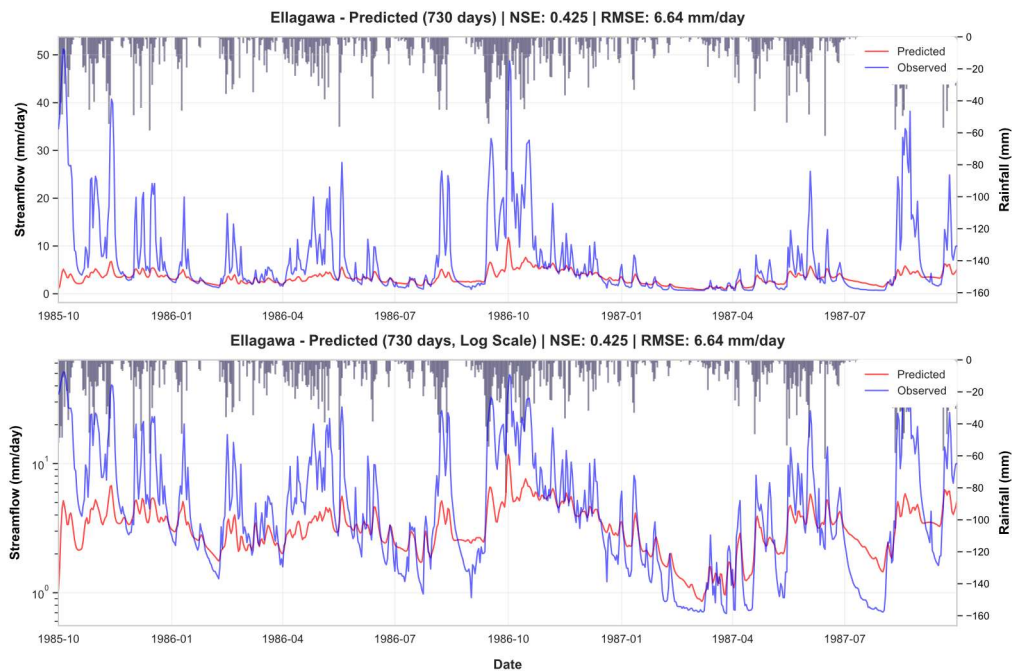


Figure 4-45: First two years simulated vs observed hydrograph for Ellagawa catchment using transferred parameters (2004–2017)

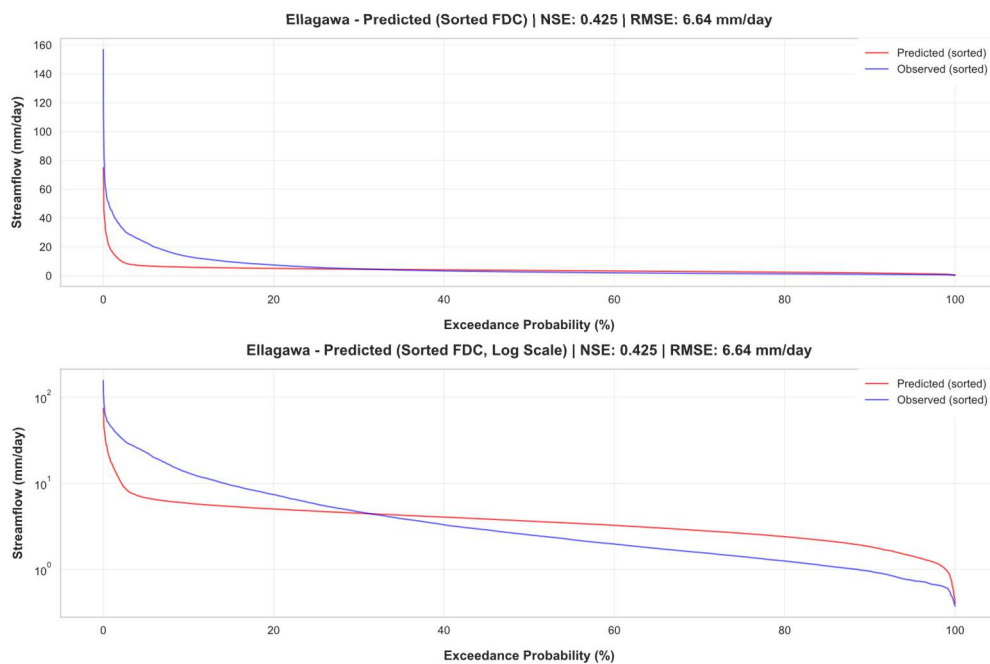


Figure 4-46: Sorted flow duration curve for Ellagawa catchment using transferred parameters (2004–2017)

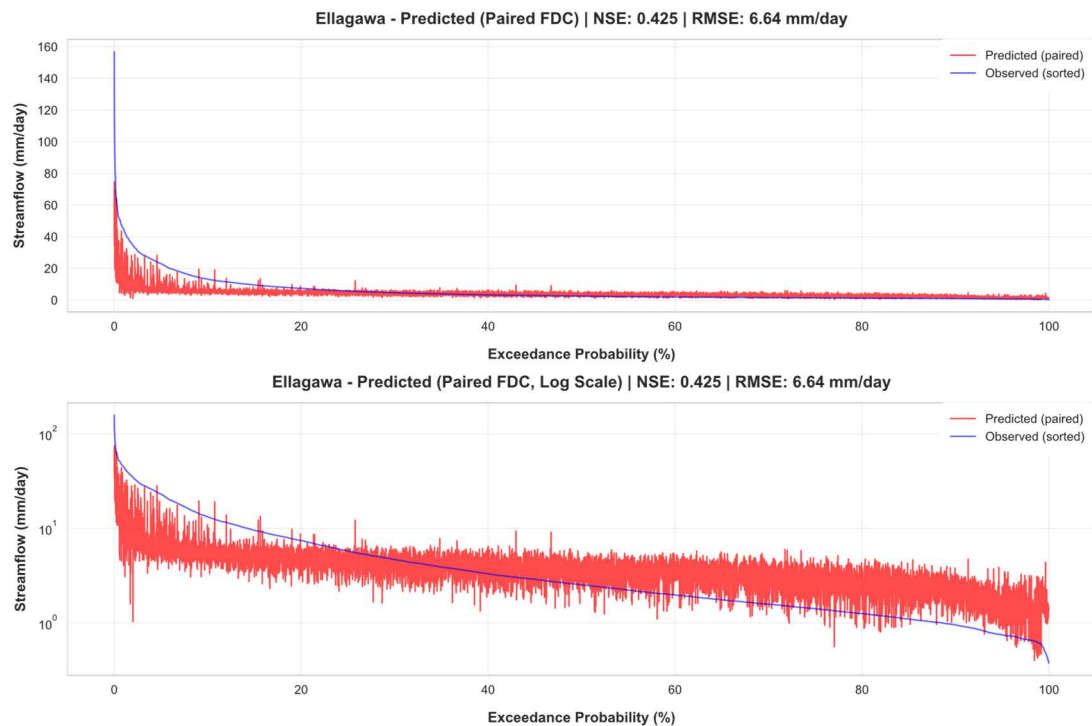


Figure 4-47: Sorted flow duration curve for Ellagawa catchment using transferred parameters (2004–2017)

Similar to the dry zone prediction, the NSE and RMSE values, along with the accompanying plots, indicate that the prediction was insufficient, though it performed slightly better than in the dry zone. The hydrograph shows that the predicted peaks follow the timing of the observed peaks, but not their magnitudes. The flow duration curve shows no alignment between the predicted and observed data across the entire flow range; peaks are underpredicted, and low flows are overpredicted. This confirms that streamflow prediction for the wet zone of the Ellagawa catchment using transferred parameters in HEC-HMS was not successful. Annexure 6 contains the complete calibration and validation hydrographs for the full study period for donor and ungauged catchment, displayed in both linear and logarithmic scales.

4.10 Prediction Results Comparison

The following Table 4-5 shows the evaluation metrics for predictions using three approaches in both the dry and wet zones.

Table 4-5: Performance comparison of streamflow prediction approaches for dry zone and wet zone catchments

Approach	Catchment (Zone)	NSE	RMSE (mm/day)
Regionalized LSTM	Monaragala (Dry)	0.478	3.05
	Ellagawa (Wet)	0.709	4.20
Global LSTM	Monaragala (Dry)	0.535	2.88
	Ellagawa (Wet)	0.651	4.60
HEC-HMS (Parameter Transfer)	Monaragala (Dry)	- 0.407	4.98
	Ellagawa (Wet)	0.425	6.64

According to these results, only Ellagawa produced satisfactory prediction results, both with the global LSTM and the regional LSTM. This outcome confirms our assumption that a regional model trained on similar catchments provides superior predictions compared to a globally trained LSTM.

In any case, dry zone predictions were not satisfactory. However, when comparing the three approaches, the global LSTM produced slightly better results. The regional LSTM showed lower training performance in the dry zone, and in Monaragala, it failed to predict peaks, whereas the global LSTM performed slightly better. This may be because the global dataset contains more extreme values for the model to learn from, while the regional dry zone model has fewer peaks, making extreme events more complex and requiring more data. The global model, which incorporates wet and intermediate zone data, can capture better overall patterns. However, the complexity of extremes, especially in the dry zone after prolonged dry periods, remains a challenge for accurate prediction. Predicting perennial river scenarios and low-flow events remains challenging for both global and regional LSTMs. Overall, LSTM models provided better predictions for ungauged river flow.

The main reason traditional modeling fails is that, when using a proximate catchment for training, the model already contains some error because it is impossible to calibrate a catchment to exactly match observed flow. This error is then carried over to the

ungauged catchment. Additionally, traditional hydrological models trained on data from only one catchment are limited in their predictive capability.

CHAPTER 5

5 CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusions

This research developed an integrated framework for streamflow prediction in ungauged basins using catchment similarity, remote sensing, and machine learning. It shows that the hydrological context, defined by a validated classification system, offers a stronger basis for prediction than relying solely on maximizing data.

Selecting hydrological signatures was a critical first step. By filtering highly correlated metrics ($|r| \geq 0.7$), the initial 14 signatures were reduced to six: *RR*, *BFI*, *FDCS*, *RC*, *PRE*, and *RLD*. This concise set was chosen to capture core unit-hydrograph processes, ensuring a meaningful, process-based representation.

Catchment classification showed that the choice of distance metric was more important than the clustering algorithm for defining hydrological groups. Of eight combinations tested, those using cosine distance performed best. Hierarchical clustering with cosine distance yielded the highest internal validation metrics (Silhouette Score = 0.525, DBI = 1.009) and formed seven clusters. K-Medoids with cosine distance was a close competitor (Silhouette Score = 0.501, DBI = 1.054). The superiority of cosine distance underscores that the proportional pattern of catchment attributes, rather than their absolute values, better reflects hydrological similarity. Principal component projections confirmed that the clusters were generally well separated, though some overlap was observed between Clusters 3 and 6 in the K-Medoids solution and between Clusters 4 and 5 in the hierarchical solution, indicating transitional catchments.

The physical realism of the statistically derived clusters was validated using an independent, satellite-derived CAS. The HCI combined WCV and BCV across all six signatures. Results showed that K-Medoids with cosine distance produced the most hydrologically coherent clusters (HCI = 1.57), slightly outperforming Hierarchical (cosine) (HCI = 1.48). This additional validation confirmed that the clusters represented distinct physio-hydrological regimes.

For operational use in ungauged basins, hydrological signatures were predicted from satellite data. Performance was satisfactory but depended on the strength of the correlation. Signatures with strong satellite proxies, such as *RR*, were well predicted ($r \approx 0.79$), whereas *PRE* with poor satellite proxies had lower accuracy ($r \approx 0.676$). Using predicted signatures for clustering achieved 87.4% pairwise accuracy, but exact cluster membership was matched only 60% of the time. This gap was mainly due to the reassignment of borderline catchments, such as Badalgama and Glencourse, which are sensitive to minor prediction errors. Still, the method was effective in identifying hydrologically similar donor clusters for ungauged catchments.

The ultimate test compared streamflow prediction strategies for pseudo-ungauged catchments. The traditional HEC-HMS parameter transfer approach performed poorly, yielding an NSE of -0.407 for Monaragala (Dry Zone) and 0.425 for Ellagawa (Wet Zone). In contrast, LSTM models consistently provided superior predictions. Critically, the optimal LSTM strategy was zone-dependent. In the Wet Zone (Ellagawa), the Regionalized LSTM trained on similar donors (Badalgama, Dunamale) was superior, achieving an NSE of 0.709 (RMSE = 4.20 mm/day), outperforming the Global LSTM (NSE = 0.651, RMSE = 4.60 mm/day). Conversely, in the Dry Zone (Monaragala), the Global LSTM (NSE = 0.535, RMSE = 2.88 mm/day) slightly surpassed the Regionalized LSTM (NSE = 0.478, RMSE = 3.05 mm/day). This suggests that data-limited dry-zone catchments benefit from broader hydrological patterns learned from a global dataset, including peak flows from wetter regions. However, both models struggled with the complex, non-linear extremes following prolonged dry periods and consistently failed to capture low-flow dynamics, as evidenced by the poor fit in the low-flow sections of the Flow Duration Curves (e.g., Figure 4-22, Figure 4-25).

In conclusion, this thesis provides a quantitative and validated framework that moves beyond the traditional proximity-based paradigm. It establishes that a similarity-based approach, defined by a curated suite of signatures and clustered using pattern-sensitive metrics like cosine distance, is hydrologically meaningful. For prediction, while LSTM models are unequivocally superior to traditional methods, the choice between

regional and global configurations is not universal and must be informed by the hydroclimatic regime of the target ungauged basin.

5.2 Recommendations

To advance the framework developed in this study, the following specific actions are recommended, targeting the core challenges of borderline catchments, dry zone prediction, and model integration:

1. The current study employs hard clustering methods (e.g., K-Means, K-Medoids, hierarchical clustering) to group catchments based on their hydrological similarity. In hard clustering, each catchment is assigned exclusively to a single cluster based on its proximity to the cluster centroid or its position in the dendrogram. While this approach provides clear delineation of hydrologically similar catchment groups, it introduces a significant limitation for catchments located near cluster boundaries. These borderline catchments may exhibit characteristics that are nearly equidistant from multiple cluster centroids, yet the hard assignment forces them into a single group. Consequently, when selecting donor catchments for ungauged basins using regionalization approaches or for training predictive models such as LSTMs, borderline catchments are restricted to donors from only their assigned cluster, even though catchments in neighboring clusters may be equally or more hydrologically similar.

This limitation is particularly problematic in heterogeneous hydrological landscapes where catchment characteristics exist along continua rather than forming distinct, well-separated groups. The artificial boundaries imposed by hard clustering can result in suboptimal donor selection, potentially excluding highly suitable donor catchments that happen to fall on the opposite side of a cluster boundary. This issue is exemplified by catchments like Beddegama (02), which exhibits unique characteristics that do not align well with any single cluster.

To address this limitation, future work should implement soft clustering or fuzzy membership methods (e.g., Fuzzy C-Means, Gaussian Mixture Models) that allow each catchment to have partial membership in multiple clusters simultaneously. In soft clustering, rather than a binary assignment (belong or not belong), each catchment receives a membership coefficient for every cluster, representing its degree of

similarity to that group. For instance, a borderline catchment might have 0.6 membership in Cluster 1, 0.3 in Cluster 2, and 0.1 in Cluster 3.

2. Augment Dry Zone Training with Globally Analogous Data: Since both the global and regional LSTM models performed poorly in the dry zone, data scarcity is a key issue. To improve this, pre-train the LSTM on global datasets from hydrologically similar catchments, such as semi-arid basins in India, Africa, or Australia. Incorporating this data helps the model better learn dryland hydrology before fine-tuning on local data, potentially improving the simulation of low flows and peaks.

3. Implement an Ensemble HEC-HMS Modeling Approach: Replace the single-parameter transfer method with an ensemble approach, using multiple parameter sets from all donor catchments. Running HEC-HMS with this ensemble produces a range of simulated flows, offering more robust predictions and a more precise representation of uncertainty, thus creating a stronger benchmark.

4. Develop a Hybrid Physics-Informed Machine Learning Model: Since LSTMs capture extremes better and HEC-HMS struggles with low flows, a hybrid model is the next step. Build a Physics-Informed Neural Network (PINN) that uses LSTM layers for temporal learning, but constrains predictions with physical laws in the loss function. This merges LSTM predictive power with physical consistency, eliminates parameter transfer, and could improve predictions, especially for low flows.

BIBLIOGRAPHY

- Addor, N., Nearing, G., Prieto, C., Newman, A. J., Le Vine, N., & Clark, M. P. (2018). A Ranking of Hydrological Signatures Based on Their Predictability in Space. *Water Resources Research*, 54(11), 8792–8812. <https://doi.org/10.1029/2018WR022606>
- Ahmadireskety, A., Da Silva, B. F., Awkerman, J. A., Aufmuth, J., Yost, R. A., & Bowden, J. A. (2021). Per- and polyfluoroalkyl substances (PFAS) in sediments collected from the Pensacola Bay System watershed. *Environmental Advances*, 5. <https://doi.org/10.1016/j.envadv.2021.100088>
- Alahacoon, N., Edirisinghe, M., & Ranagalage, M. (2021). Satellite-based meteorological and agricultural drought monitoring for agricultural sustainability in Sri Lanka. *Sustainability (Switzerland)*, 13(6). <https://doi.org/10.3390/su13063427>
- Alahacoon, N., Matheswaran, K., Pani, P., & Amarnath, G. (2018). A decadal historical satellite data and rainfall trend analysis (2001-2016) for flood hazard mapping in Sri Lanka. *Remote Sensing*, 10(3). <https://doi.org/10.3390/rs10030448>
- Alfred, H. T. (1911). District No. 10, Great basin. *Monthly Water Review*, 39, 0–1248. [https://doi.org/https://doi.org/10.1175/1520-0493\(1911\)39%3C1248a:DNGB%3E2.0.CO;2](https://doi.org/https://doi.org/10.1175/1520-0493(1911)39%3C1248a:DNGB%3E2.0.CO;2)
- An, Y., Zou, Z., & Li, R. (2016). Descriptive characteristics of surface water quality in Hong Kong by a self-organising map. *International Journal of Environmental Research and Public Health*, 13(1). <https://doi.org/10.3390/ijerph13010115>
- Arsenault, R., & Brissette, F. P. (2014). Continuous streamflow prediction in ungauged basins: The effects of equifinality and parameter set selection on uncertainty in regionalization approaches. *Water Resources Research*, 50(7), 6135–6153. <https://doi.org/10.1002/2013WR014898>

- Bajracharya, P., & Jain, S. (2021). *Hydrologic Similarity Based on Width Function and Hypsometry: An Unsupervised Learning Approach*. <https://doi.org/10.21203/rs.3.rs-614652/v1>
- Baker, D. B., Richards, R. P., Loftus, T. T., & Kramer, J. W. (2004). A new flashiness index: Characteristics and applications to Midwestern rivers and streams. *JAWRA Journal of the American Water Resources Association*, 40(2), 503–522. <https://doi.org/https://doi.org/10.1111/j.1752-1688.2004.tb01046.x>
- Bakure, B. Z., Fikadu, S., & Malu, A. (2020). Analysis of physicochemical water quality parameters for streams under agricultural, urban and forest land-use types: in the case of gilgel Gibe catchment, Southwest Ethiopia. *Applied Water Science*, 10(11). <https://doi.org/10.1007/s13201-020-01318-9>
- Beven, K. (2006). A manifesto for the equifinality thesis. *Journal of Hydrology*, 320(1–2), 18–36. <https://doi.org/10.1016/j.jhydrol.2005.07.007>
- Brunner, M. I., Melsen, L. A., Newman, A. J., Wood, A. W., & Clark, M. P. (2020). Future streamflow regime changes in the United States: Assessment using functional classification. *Hydrology and Earth System Sciences*, 24(8), 3951–3966. <https://doi.org/10.5194/hess-24-3951-2020>
- Bu, J., Liu, W., Pan, Z., & Ling, K. (2020). Comparative study of hydrochemical classification based on different hierarchical cluster analysis methods. *International Journal of Environmental Research and Public Health*, 17(24), 1–23. <https://doi.org/10.3390/ijerph17249515>
- Chérif, R., & Gargouri-Ellouze, E. (2023). Flow signatures and catchment's attributes for HCA clustering in a hydrologic similarity assessment (Tunisian case). *Scientific Reports*, 13(1). <https://doi.org/10.1038/s41598-023-38608-6>
- Chiverton, A., Hannaford, J., Holman, I., Corstanje, R., Prudhomme, C., Bloomfield, J., & Hess, T. M. (2015). Which catchment characteristics control the temporal dependence structure of daily river flows? *Hydrological Processes*, 29(6), 1353–1369. <https://doi.org/10.1002/hyp.10252>

- Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1*(2), 224–227. <https://doi.org/10.1109/TPAMI.1979.4766909>
- Dooge, J. C. I. (1959). *A General Theory of the Unit Hydrograph* (Vol. 64, Issue 2).
- Eckhardt, K. (2005). How to construct recursive digital filters for baseflow separation. *Hydrological Processes, 19*(2), 507–515. <https://doi.org/10.1002/hyp.5675>
- Fisher, R. A. (1992). Statistical Methods for Research Workers. In N. L. Kotz Samuel and Johnson (Ed.), *Breakthroughs in Statistics: Methodology and Distribution* (pp. 66–70). Springer New York. https://doi.org/10.1007/978-1-4612-4380-9_6
- Frame, J. M., Kratzert, F., Klotz, D., Gauch, M., Shelev, G., Gilon, O., Qualls, L. M., Gupta, H. V., & Nearing, G. S. (2022). Deep learning rainfall-runoff predictions of extreme events. *Hydrology and Earth System Sciences, 26*(13), 3377–3392. <https://doi.org/10.5194/hess-26-3377-2022>
- Frame, J. M., Kratzert, F., Raney, A., Rahman, M., Salas, F. R., & Nearing, G. S. (2021). Post-Processing the National Water Model with Long Short-Term Memory Networks for Streamflow Predictions and Model Diagnostics. *Journal of the American Water Resources Association, 57*(6), 885–905. <https://doi.org/10.1111/1752-1688.12964>
- Gao, S., Huang, Y., Zhang, S., Han, J., Wang, G., Zhang, M., & Lin, Q. (2020). Short-term runoff prediction with GRU and LSTM networks without requiring time step optimization during sample generation. *Journal of Hydrology, 589*. <https://doi.org/10.1016/j.jhydrol.2020.125188>
- Gayashan, K., & Rajapakse, L. (2024). Transferability of Basin Hydrological Parameters for Assessing Hydrological Processes in Sri Lankan Wet Zone River Basins. *2024 Moratuwa Engineering Research Conference (MERCon)*, 302–307. <https://doi.org/10.1109/MERCon63886.2024.10689223>
- Ghaneei, P., Foroumandi, E., & Moradkhani, H. (2024). Enhancing Streamflow Prediction in Ungauged Basins Using a Nonlinear Knowledge-Based Framework

- and Deep Learning. *Water Resources Research*, 60(11). <https://doi.org/10.1029/2024WR037152>
- Gopalakrishnan, T., & Kumar, L. (2020). Modeling and mapping of soil salinity and its impact on paddy lands in Jaffna Peninsula, Sri Lanka. *Sustainability (Switzerland)*, 12(20), 1–15. <https://doi.org/10.3390/su12208317>
- Gunathilake, M. B., Karunanayake, C., Gunathilake, A. S., Marasingha, N., Samarasinghe, J. T., Bandara, I. M., & Rathnayake, U. (2021). Hydrological Models and Artificial Neural Networks (ANNs) to Simulate Streamflow in a Tropical Catchment of Sri Lanka. *Applied Computational Intelligence and Soft Computing*, 2021. <https://doi.org/10.1155/2021/6683389>
- He, Y., Bárdossy, A., & Zehe, E. (2011). A review of regionalisation for continuous streamflow simulation. *Hydrology and Earth System Sciences*, 15(11), 3539–3553. <https://doi.org/10.5194/hess-15-3539-2011>
- Herath, M. H. B. C. W., & Wijesekera, N. T. S. (2021). Evaluation of HEC-HMS Model for Water Resources Management in Maha Oya Basin in Sri Lanka. *Engineer: Journal of the Institution of Engineers, Sri Lanka*, 54(2), 45. <https://doi.org/10.4038/engineer.v54i2.7441>
- Hrachowitz, M., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., Arheimer, B., Blume, T., Clark, M. P., Ehret, U., Fenicia, F., Freer, J. E., Gelfan, A., Gupta, H. V., Hughes, D. A., Hut, R. W., Montanari, A., Pande, S., Tetzlaff, D., ... Cudennec, C. (2013a). A decade of Predictions in Ungauged Basins (PUB)-a review. In *Hydrological Sciences Journal* (Vol. 58, Issue 6, pp. 1198–1255). <https://doi.org/10.1080/02626667.2013.803183>
- Hrachowitz, M., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., Arheimer, B., Blume, T., Clark, M. P., Ehret, U., Fenicia, F., Freer, J. E., Gelfan, A., Gupta, H. V., Hughes, D. A., Hut, R. W., Montanari, A., Pande, S., Tetzlaff, D., ... Cudennec, C. (2013b). A decade of Predictions in Ungauged Basins (PUB)-a review. In *Hydrological Sciences Journal* (Vol. 58, Issue 6, pp. 1198–1255). <https://doi.org/10.1080/02626667.2013.803183>

- Huang, J., Sehgal, V., Alvarez, L. V., Brocca, L., Cai, S., Cheng, R., Cheng, X., Du, J., El Masri, B., Endsley, K. A., Fang, Y., Hu, J., Jampani, M., Kibria, M. G., Koren, G., Li, L., Liu, L., Mao, J., Moreno, H. A., ... Fisher, J. B. (2025). Remotely Sensed High-Resolution Soil Moisture and Evapotranspiration: Bridging the Gap Between Science and Society. In *Water Resources Research* (Vol. 61, Issue 5). John Wiley and Sons Inc. <https://doi.org/10.1029/2024WR037929>
- Iresh, A. D. S., Marasingha, A. G. N. S., Wedanda, A. M. T. S. H., Wickramasekara, G. P., Wickramasooriya, M. D. J. P., & Premathilaka, M. T. C. (2021). Development of a Hydrological Model for Kala Oya Basin Using SWAT Model. *Engineer: Journal of the Institution of Engineers, Sri Lanka*, 54(1), 57. <https://doi.org/10.4038/engineer.v54i1.7435>
- Janani, N., Balaji, K., Nagarajan, K., Thiagarajan, G., & Duraisamy, M. R. (2023). Trend analysis and variability of satellite-based soil moisture data for the Lower Bhavani basin, Tamil Nadu using Google Earth Engine. *Journal of Applied and Natural Science*, 15(2), 555–559. <https://doi.org/10.31018/jans.v15i2.4515>
- Jehn, F. U., Bestian, K., Breuer, L., Kraft, P., & Houska, T. (2020). Using hydrological and climatic catchment clusters to explore drivers of catchment behavior. *Hydrology and Earth System Sciences*, 24(3), 1081–1100. <https://doi.org/10.5194/hess-24-1081-2020>
- Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. In *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* (Vol. 374, Issue 2065). Royal Society of London. <https://doi.org/10.1098/rsta.2015.0202>
- Knoben, W. J. M., Woods, R. A., & Freer, J. E. (2018). A Quantitative Hydrological Climate Classification Evaluated With Independent Streamflow Data. *Water Resources Research*, 54(7), 5088–5109. <https://doi.org/10.1029/2018WR022913>
- Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall-runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrology*

and Earth System Sciences, 22(11), 6005–6022. <https://doi.org/10.5194/hess-22-6005-2018>

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). *Towards Learning Universal, Regional, and Local Hydrological Behaviors via Machine-Learning Applied to Large-Sample Datasets*. <http://arxiv.org/abs/1907.08456>

Kratzert, F., Nearing, G., Addor, N., Erickson, T., Gauch, M., Gilon, O., Gudmundsson, L., Hassidim, A., Klotz, D., Nevo, S., Shalev, G., & Matias, Y. (2023). Caravan - A global community dataset for large-sample hydrology. *Scientific Data*, 10(1). <https://doi.org/10.1038/s41597-023-01975-w>

Levizzani, V., & Cattani, E. (2019). Satellite remote sensing of precipitation and the terrestrial water cycle in a changing climate. In *Remote Sensing* (Vol. 11, Issue 19). MDPI AG. <https://doi.org/10.3390/rs11192301>

Li, A., Zhao, W., & Deng, W. (2015). A quantitative inspection on spatio-temporal variation of remote sensing-based estimates of land surface evapotranspiration in South Asia. *Remote Sensing*, 7(4), 4726–4752. <https://doi.org/10.3390/rs70404726>

Li, J., Peng, B., Wei, Y., & Ye, H. (2021). Accurate extraction of surface water in complex environment based on Google Earth Engine and Sentinel-2. *PLoS ONE*, 16(6 June). <https://doi.org/10.1371/journal.pone.0253209>

Li, W., Liu, C., Hu, C., Niu, C., Li, R., Li, M., Xu, Y., & Tian, L. (2024). Application of a hybrid algorithm of LSTM and Transformer based on random search optimization for improving rainfall-runoff simulation. *Scientific Reports*, 14(1). <https://doi.org/10.1038/s41598-024-62127-7>

Liu, J., Bian, Y., Lawson, K., & Shen, C. (2024). Probing the limit of hydrologic predictability with the Transformer network. *Journal of Hydrology*, 637, 131389. <https://doi.org/https://doi.org/10.1016/j.jhydrol.2024.131389>

- Liu, J., Wu, D., Mohammed, H., & Seidu, R. (2024). A Novel Method for Anomaly Detection and Signal Calibration in Water Quality Monitoring of an Urban Water Supply System. *Water (Switzerland)*, 16(9). <https://doi.org/10.3390/w16091238>
- Madhavi, G. N., Sharath Kumar, P., Chipade, R. A., Bhate, J., & Sarma, T. V. C. (2022). Estimation and Validation Study of Soil Moisture Using GPS-IR Technique Over a Tropical Region: Variability of SM With Rainfall and Energy Fluxes. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15, 42–49. <https://doi.org/10.1109/JSTARS.2021.3127469>
- Madhulatha, T. S. (2011). Comparison between K-Means and K-Medoids Clustering Algorithms. In *CCIS* (Vol. 198).
- Madhushankha, J. M. L., & Wijesekera, N. T. S. (2021). Application of HEC-HMS Model to Estimate Daily Streamflow in Badddegama Watershed of Gin Ganga Basin Sri Lanka. *Engineer: Journal of the Institution of Engineers, Sri Lanka*, 54(1), 89. <https://doi.org/10.4038/engineer.v54i1.7438>
- McDonnell, J. J., & Woods, R. (2004). On the need for catchment classification. In *Journal of Hydrology* (Vol. 299, Issues 1–2, pp. 2–3). Elsevier B.V. <https://doi.org/10.1016/j.jhydrol.2004.09.003>
- McMahon, T. A., Vogel, R. M., Pegram, G. G. S., Peel, M. C., & Etkin, D. (2007). Global streamflows - Part 2: Reservoir storage-yield performance. *Journal of Hydrology*, 347(3–4), 260–271. <https://doi.org/10.1016/j.jhydrol.2007.09.021>
- McMillan, H. K. (2021). A review of hydrologic signatures and their applications. *Wiley Interdisciplinary Reviews: Water*, 8(1). <https://doi.org/10.1002/wat2.1499>
- Melsen, L. A., Puy, A., Torfs, P. J. J. F., & Saltelli, A. (2025). The rise of the Nash-Sutcliffe efficiency in hydrology. *Hydrological Sciences Journal*, 70(8), 1248–1259. <https://doi.org/10.1080/02626667.2025.2475105>
- Modiri, E., & Bárdossy, A. (2021). Clustering simultaneous occurrences of the extreme floods in the neckar catchment. *Water (Switzerland)*, 13(4). <https://doi.org/10.3390/w13040399>

- Moriasi, D. N., Arnold, J. G., Liew, M. W. Van, Bingner, R. L., Harmel, R. D., & Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, 50(3), 885–900.
- Munasinghe, D. S. N., Najim, M. M. M., Quadroni, S., & Musthafa, M. M. (2021). Impacts of streamflow alteration on benthic macroinvertebrates by mini-hydro diversion in Sri Lanka. *Scientific Reports*, 11(1). <https://doi.org/10.1038/s41598-020-79576-5>
- Murtagh, F., & Legendre, P. (2014). Ward’s Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward’s Criterion? *Journal of Classification*, 31(3), 274–295. <https://doi.org/10.1007/s00357-014-9161-z>
- Narbondo, S., Gorgoglione, A., Crisci, M., & Chreties, C. (2020). Enhancing physical similarity approach to predict runoff in ungauged watersheds in sub-tropical regions. *Water (Switzerland)*, 12(2). <https://doi.org/10.3390/w12020528>
- Nathan, R. J., & McMahon, T. A. (1990). Evaluation of Automated Techniques for Base Flow and Recession Analyses. In *WATER RESOURCES RESEARCH* (Vol. 26, Issue 7).
- Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., Prieto, C., & Gupta, H. V. (2021). What Role Does Hydrological Science Play in the Age of Machine Learning? In *Water Resources Research* (Vol. 57, Issue 3). Blackwell Publishing Ltd. <https://doi.org/10.1029/2020WR028091>
- Olden, J. D., & Poff, N. L. (2003). Redundancy and the choice of hydrologic indices for characterizing streamflow regimes. *River Research and Applications*, 19(2), 101–121. <https://doi.org/10.1002/rra.700>
- Oppel, H., & Schumann, A. H. (2020). Machine learning based identification of dominant controls on runoff dynamics. *Hydrological Processes*, 34(11), 2450–2465. <https://doi.org/10.1002/hyp.13740>
- Oudin, L., Andréassian, V., Perrin, C., Michel, C., & Le Moine, N. (2008). Spatial proximity, physical similarity, regression and ungaged catchments: A comparison

- of regionalization approaches based on 913 French catchments. *Water Resources Research*, 44(3). <https://doi.org/10.1029/2007WR006240>
- Papa, F., Crétaux, J. F., Grippa, M., Robert, E., Trigg, M., Tshimanga, R. M., Kitambo, B., Paris, A., Carr, A., Fleischmann, A. S., de Fleury, M., Gbetkom, P. G., Calmettes, B., & Calmant, S. (2023). Water Resources in Africa under Global Change: Monitoring Surface Waters from Space. In *Surveys in Geophysics* (Vol. 44, Issue 1, pp. 43–93). Springer Science and Business Media B.V. <https://doi.org/10.1007/s10712-022-09700-9>
- Parajka, J., Merz, R., & Blöschl, G. (2005). Hydrology and Earth System Sciences A comparison of regionalisation methods for catchment model parameters. In *Hydrology and Earth System Sciences* (Vol. 9). www.copernicus.org/EGU/hess/hess/9/157/SRef-ID:1607-7938/hess/2005-9-157EuropeanGeosciencesUnion
- Patil, S., Wandre, S., Patil, M., Dhekale, J., Bansude, S., Gavit, B., & Ghatge, J. (2024). *A review on artificial neural network and hydrology*. 8, 571–576.
- Peng, H., Long, F., & Ding, C. (2005). *Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy*.
- Rajendran, M., Gunawardena, E. R. N., & Dayawansa, N. D. K. (2020). Runoff Prediction in an Ungauged Catchment of Upper Deduru Oya Basin, Sri Lanka: A Comparison of HEC-HMS and WEAP Models. *International Journal of Progressive Sciences and Technologies (IJPSAT)*, 18(2), 121–129. <http://ijpsat.ijshjournals.org>
- Rao, A. R., & Srinivas, V. V. (2008). *Regionalization of Watersheds: An Approach Based on Cluster Analysis* (Vol. 58). Springer Science & Business Media.
- Razavi, T., & Coulibaly, P. (2013). Streamflow Prediction in Ungauged Basins: Review of Regionalization Methods. *Journal of Hydrologic Engineering*, 18(8), 958–975. [https://doi.org/10.1061/\(asce\)he.1943-5584.0000690](https://doi.org/10.1061/(asce)he.1943-5584.0000690)

- Richter, B. D., Baumgartner, J. V, Powell, J., & Braun, D. P. (1996). A Method for Assessing Hydrologic Alteration within Ecosystems. *Conservation Biology*, 10(4), 1163–1174. [https://doi.org/https://doi.org/10.1046/j.1523-1739.1996.10041163.x](https://doi.org/10.1046/j.1523-1739.1996.10041163.x)
- Sankarasubramanian, A., Vogel, R. M., & Limbrunner, J. F. (2001). Climate elasticity of streamflow in the United States. *Water Resources Research*, 37(6), 1771–1781. <https://doi.org/10.1029/2000WR900330>
- Sawicz, K., Wagener, T., Sivapalan, M., Troch, P. A., & Carrillo, G. (2011). Catchment classification: Empirical analysis of hydrologic similarity based on catchment function in the eastern USA. *Hydrology and Earth System Sciences*, 15(9), 2895–2911. <https://doi.org/10.5194/hess-15-2895-2011>
- Selvarajah, H., Koike, T., Rasmy, M., Tamakawa, K., Yamamoto, A., Kitsuregawa, M., & Zhou, L. (2021). Development of an integrated approach for the assessment of climate change impacts on the hydro-meteorological characteristics of the mahaweli river basin, Sri Lanka. *Water (Switzerland)*, 13(9). <https://doi.org/10.3390/w13091218>
- Shahapure, K. R., & Nicholas, C. (2020). Cluster quality analysis using silhouette score. *Proceedings - 2020 IEEE 7th International Conference on Data Science and Advanced Analytics, DSAA 2020*, 747–748. <https://doi.org/10.1109/DSAA49011.2020.00096>
- Shamir, E., Imam, B., Morin, E., Gupta, H. V., & Sorooshian, S. (2005). The role of hydrograph indices in parameter estimation of rainfall-runoff models. *Hydrological Processes*, 19(11), 2187–2207. <https://doi.org/10.1002/hyp.5676>
- Sharghi, E., Nourani, V., Soleimani, S., & Sadikoglu, F. (2018). Application of different clustering approaches to hydroclimatological catchment regionalization in mountainous regions, a case study in Utah State. *Journal of Mountain Science*, 15(3), 461–484. <https://doi.org/10.1007/s11629-017-4454-4>
- Sivapalan, M., Takeuchi, K., Franks, S. W., Gupta, V. K., Karambiri, H., Lakshmi, V., Liang, X., McDonnell, J. J., Mendiondo, E. M., O’Connell, P. E., Oki, T., Pomeroy, J. W., Schertzer, D., Uhlenbrook, S., & Zehe, E. (2003). IAHS Decade

- on Predictions in Ungauged Basins (PUB), 2003-2012: Shaping an exciting future for the hydrological sciences. *Hydrological Sciences Journal*, 48(6), 857–880. <https://doi.org/10.1623/hysj.48.6.857.51421>
- Tallaksen, L. M. (1995). A review of baseflow recession analysis. In *Journal of Hydrology* (Vol. 165).
- Tan, P.-Ning., Steinbach, Michael., Karpatne, Anuj., & Kumar, Vipin. (2019). *Introduction to data mining*. Pearson Education, Inc.
- Tukey, J. W. (1949). Comparing Individual Means in the Analysis of Variance. In *Source: Biometrics* (Vol. 5, Issue 2).
- U.S. Army Corps of Engineers, H. E. C. (HEC). (2000). *Hydrologic Modeling System HEC-HMS Technical Reference Manual CPD-74B*. [https://www.caee.utexas.edu/prof/maidment/CE365KSpr16/HEC/HEC-HMS_Technical%20Reference%20Manual_\(CPD-74B\).pdf](https://www.caee.utexas.edu/prof/maidment/CE365KSpr16/HEC/HEC-HMS_Technical%20Reference%20Manual_(CPD-74B).pdf)
- Vörösmarty, C. J., McIntyre, P. B., Gessner, M. O., Dudgeon, D., Prusevich, A., Green, P., Glidden, S., Bunn, S. E., Sullivan, C. A., Liermann, C. R., & Davies, P. M. (2010). Erratum: Global threats to human water security and river biodiversity (*Nature* (2010) 467 (555-561)). In *Nature* (Vol. 468, Issue 7321, p. 334). <https://doi.org/10.1038/nature09549>
- Wagener, T., Sivapalan, M., Troch, P., & Woods, R. (2007). Catchment Classification and Hydrologic Similarity. *Geography Compass*, 1(4), 901–931. <https://doi.org/10.1111/j.1749-8198.2007.00039.x>
- Wimalasiri, E. M., Jahanshiri, E., Suhairi, T. A. S. T. M., Mapa, R. B., Karunaratne, A. S., Vidhanarachchi, L. P., Udayangani, H., Nizar, N. M. M., & Azam-Ali, S. N. (2020). The first version of nation-wide open 3D soil database for Sri Lanka. *Data in Brief*, 33. <https://doi.org/10.1016/j.dib.2020.106342>
- World Meteorological Organization (WMO). (2025). *Guide to Hydrological Practice Volume I Hydrology-From Measurement to Hydrological Information*. <https://library.wmo.int/idurl/4/35804>

- Wu, X., Yin, S., Cheng, C., Chen, Z., & Su, H. (2022). *SSDP Model with Inflow Clustering for Hydropower System Operation*. <https://doi.org/10.21203/rs.3.rs-2011605/v1>
- Xulu, S., Peerbhay, K., Gebreslasie, M., & Ismail, R. (2019). Unsupervised clustering of forest response to drought stress in Zululand region, South Africa. *Forests*, 10(7). <https://doi.org/10.3390/f10070531>
- Yadav, M., Wagener, T., & Gupta, H. (2007). Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins. *Advances in Water Resources*, 30(8), 1756–1774. <https://doi.org/10.1016/j.advwatres.2007.01.005>
- Yokoo, Y., & Sivapalan, M. (2011). Towards reconstruction of the flow duration curve: development of a conceptual framework with a physical basis. *Hydrology and Earth System Sciences*, 15(9), 2805–2819. <https://doi.org/10.5194/hess-15-2805-2011>
- Zhang, Y., Zhang, Y., Zhai, X., Xia, J., Tang, Q., Wang, W., Wu, J., Niu, X., & Han, B. (2025). Understanding meteorological and physio-geographical controls of variability of flood event classes in headstream catchments of China. *Hydrology and Earth System Sciences*, 29(14), 3257–3275. <https://doi.org/10.5194/hess-29-3257-2025>

ANNEXURE 1

CATCHMENT DATA INVENTORY

Table: Information on selected streamflow gauges and precipitation gauges used in this study

Catchment	Data Period	Streamflow Gap (%)	Precipitation Gauges	Gap (%)
Badalgama	2005 -2024	0.1	Holombuwa	0.0
			Handessa - Daulagala	4.1
			Polgahawela (Coconut)	2.9
			Kurunegala	0.4
			Aranayake Govt. Hospital	2.9
			Ambepussa Govt. Farm	5.4
Baddegama	2006 -2024	1.5	Deviturai Estate	7.4
			Pallegama Rathnayake GP	7.4
			Beausejour (Lower)	11.9
			Hiniduma	5.9
			Thawalama	0.3
			Anningkanda Estate	6.9
Buttala	2016 -2023	0.2	Canawarella Group	5.8
			Nakkala	0.0

Catchment	Data Period	Streamflow Gap (%)	Precipitation Gauges	Gap (%)
Calidoniya	2004 -2015	0.0	Bopatthalawa	1.51
			Sita Eliya	6.6
Deraniyagala	2005 -2024	0.0	Laxapana	2.1
			Kitulgala	0.0
Dunamale	2006 -2015	0.0	Pasyala	4.2
			Dunamale	0.0
			Vincit Estate	0.0
Ellagawa	1985 - 2017	0.3	Galatura Estate	2.3
			Ratnapura	0.0
			Wellandura Tea Factory	4.4
			Pussella S.P.	0.7
			Eheliyagoda S.P.	2.8
			Hapugastenna Estate	3.6
Giriulla	2005 -2024	1.8	Holombuwa	0.0
			Handessa- Daulagala	4.1
			Polgahawela (Coconut)	2.9
			Kurunegala	0.4
			Aranayake Govt. Hospital	2.9
			Andigama Farm	1.2
Glencourse	2005 -2020	0.0	Pussella S.P.	2.1
			Maussakelle	0.5
			Campion Estate	2.0
			Alupolla Group	3.1
			Kitulgala	0.0
			Holombuwa	0.0

Catchment	Data Period	Streamflow Gap (%)	Precipitation Gauges	Gap (%)
Glencourse	2005 -2020	0.0	Laxapana	2.6
			Undugoda	2.6
Hanwella	2005 -2020	0.0	Maussakelle	0.5
			Alupolla Group	3.1
			Holombuwa	0.0
			Laxapana	2.6
			Kenilworth (Strathellie)	1.0
			Pussella S.P.	2.1
Holombuwa	2005 -2024	0.0	Holombuwa	0.0
			Undugoda	3.8
Kataragama	2014 -2024	4.2	Monaragala	0.8
			Buttala	0.0
			Ranugalla	0.8
Kithulgala	2004 -2024	0.0	Alupolla Group	2.4
			Maussakelle	0.4
			Laxapana	2.1
			Kenilworth (Strathellie)	5.8
Kudaoya	2005 -2015	0.4	Kudaoya	0.7
			Poonagala Group	5.3
Mahaoya	2010 -2024	6.5	Ekiriyankumbura	1.7
			Kolaniyaya	1.7
Millakanda	2005 -2021	0.1	Thawalama	0.4
			Pelawatte	8.1
			Galatura Estate	4.0
			Usk Valley S.P.	6.5

Catchment	Data Period	Streamflow Gap (%)	Precipitation Gauges	Gap (%)
Monaragala	2010 -2024	7.3	Kolaniyaya	1.7
			Padiyathalawa	0.0
			Kongaspanwila	0.6
Nakkala	2016 -2024	0.1	Nakkala	0.0
			Mahadowa Estate	0.0
Nawalapitiya	2004 -2016	0.0	Watawala	1.4
			Canyon	0.7
			Kenilworth (Strathellie)	1.4
Norwood	2005 -2023	0.1	Norwood	0.1
			Campion Estate	11.4
Padiyathalawa	2005 -2024	0.0	Ekiriyankumbura	1.3
			Padiyathalawa	0.1
Patupaula	2005 -2017	0.2	Rayigama	5.8
			Pussella S.P.	1.3
			Geekiyanakanda Estate	5.1
			Usk Valley S.P.	6.4
			Galatura Estate	5.1
			Depedena Group	3.2
			Ratnapura	0.0
			Hapugastenna Estate	8.3
Pitabeddara	2005 -2022	0.1	Anningkanda Estate	5.7
			Pallegama Rathnayake Gp	7.2
			Urawa	0.0
			Dampahala Tea Factory	0.9

Catchment	Data Period	Streamflow Gap (%)	Precipitation Gauges	Gap (%)
Ratnapura	2005 -2017	0.7	Lellopitiya Estate	12.1
			Alupolla Group	2.6
			Ratnapura	0.0
Siyambalanduwa	2005 -2024	0	Siyambalanduwa	0.8
Thaldena	2011 -2024	0.1	Taldena	0.6
			Badulla	1.2
Thanamalwila	2005 -2015	0	Kudaoya	0.8
			Poonagala Group	5.3
Thawalama	2005 -2024	0.1	Anningkanda Estate	6.5
			Pallegama Rathnayake Gp	6.9
			Thawalama	0.4
Urawa	2005 -2022	0.6	Dampahala Tea Factory	0.9
			Urawa	0.0
			Anningkanda Estate	5.7
Wellawaya	2005 -2015	0	Poonagala Group	5.3

ANNEXURE 2

COMPLETE HYDROLOGICAL SIGNATURE DATASET FOR ALL STUDY CATCHMENTS

Table: Computed hydrological signature values for selected catchments

Catchment	RR	MAR	Q ₅	Q ₉₅	FI	CoV	RE	FDCS	RLD	LFF	HFF	ZFF	RC	BFI
Pitabeddara	0.534	1708.6	1.03	13.87	0.37	1.42	1.17	-1.85	0.55	0.01	0.21	0.00	0.97	0.57
Kudaoya	0.224	411.9	0.06	4.77	0.60	2.63	2.33	-2.91	0.57	0.05	0.31	0.01	0.97	0.24
Glencourse	0.542	1990.7	1.07	14.47	0.28	1.18	1.39	-1.53	0.54	0.01	0.21	0.00	0.96	0.64
Buttala	0.402	947.1	0.40	11.23	0.33	1.77	1.23	-2.03	0.67	0.01	0.28	0.00	0.97	0.40
Padiyathalawa	0.385	862.3	0.00	10.07	0.62	3.68	2.42	-6.57	0.59	0.32	0.39	0.06	0.93	0.35
Siyambalanduwa	0.228	445.8	0.00	5.05	0.64	3.56	2.04	-3.70	0.56	0.30	0.35	0.22	0.93	0.32
Urawa	0.329	1059.3	0.55	8.56	0.38	1.24	0.74	-2.08	0.57	0.01	0.20	0.00	0.97	0.54
Holombuwa	0.423	1402.9	0.09	15.68	0.52	1.91	1.25	-3.31	0.60	0.15	0.30	0.00	0.98	0.21
Deraniyagala	0.674	3009.4	0.96	24.70	0.35	1.21	1.05	-2.54	0.60	0.06	0.24	0.00	0.96	0.55
Wellawaya	0.285	858.1	0.11	8.66	0.40	1.68	1.01	-3.50	0.57	0.10	0.32	0.01	0.98	0.40
Nawalapitiya	0.736	3338.9	1.84	27.00	0.27	1.03	0.99	-2.16	0.61	0.03	0.17	0.00	0.98	0.60
Thanamalwila	0.206	432.5	0.00	4.45	0.42	1.81	2.19	-5.82	0.55	0.28	0.38	0.11	0.94	0.50
Giriulla	0.378	860.7	0.14	11.37	0.55	2.25	0.84	-4.37	0.53	0.00	0.40	0.00	0.98	0.14
Mahaoya	0.292	628.3	0.00	7.43	0.59	3.37	1.39	-3.21	0.68	0.47	0.36	0.46	0.93	0.24

Catchment	RR	MAR	Q ₅	Q ₉₅	FI	CoV	RE	FDCS	RLD	LFF	HFF	ZFF	RC	BFI
Nakkala	0.634	1230.3	1.03	10.02	0.28	1.29	1.11	-1.74	0.58	0.02	0.17	0.00	0.98	0.62
Thawalama	0.759	3055.4	1.70	22.22	0.32	0.97	0.96	-2.07	0.56	0.02	0.18	0.00	0.95	0.64
Badalgama	0.580	1304.3	0.27	11.61	0.35	1.75	1.64	-3.17	0.52	0.12	0.23	0.00	0.96	0.53
Monaragala	0.264	590.4	0.00	9.03	0.60	2.93	1.58	-4.99	0.57	0.43	0.44	0.39	0.93	0.23
Norwood	0.665	1777.7	0.37	14.36	0.24	1.11	0.95	-2.18	0.58	0.09	0.21	0.00	0.96	0.66
Ratnapura	0.423	1764.5	1.00	13.88	0.33	1.16	1.27	-1.68	0.54	0.01	0.20	0.00	0.98	0.47
Calidoniya	0.588	1101.8	1.07	8.45	0.22	0.95	1.37	-1.88	0.57	0.00	0.18	0.00	0.98	0.59
Thaldena	0.412	773.9	0.34	6.59	0.32	1.75	1.24	-2.20	0.56	0.01	0.23	0.00	0.97	0.56
Dunamale	0.562	1606.0	0.34	16.64	0.40	1.67	1.44	-2.46	0.55	0.08	0.25	0.00	0.97	0.38
Kataragama	0.121	212.9	0.37	1.23	0.11	0.70	0.48	-1.00	0.54	0.00	0.06	0.00	1.00	0.75
Kithulgala	0.643	2378.8	2.00	14.53	0.17	0.69	0.95	-1.36	0.61	0.01	0.10	0.00	0.95	0.82
Hanwella	0.426	1707.9	0.85	16.23	0.34	1.36	1.16	-2.78	0.56	0.00	0.30	0.00	0.97	0.40
Millakanda	0.556	2506.5	1.39	23.72	0.30	1.28	1.17	-2.42	0.53	0.01	0.26	0.00	0.95	0.46
Ellagawa	0.564	2013.2	0.75	22.97	0.33	1.55	1.83	-2.65	0.52	0.00	0.28	0.00	0.98	0.27
Patupaula	0.375	1499.0	1.72	11.73	0.22	1.06	1.66	-1.10	0.48	0.00	0.17	0.00	0.95	0.61
Baddegama	0.716	2775.9	2.61	22.38	0.24	0.86	0.98	-1.66	0.50	0.00	0.22	0.00	0.93	0.65

ANNEXURE 3

CLUSTER ANALYSIS RESULTS

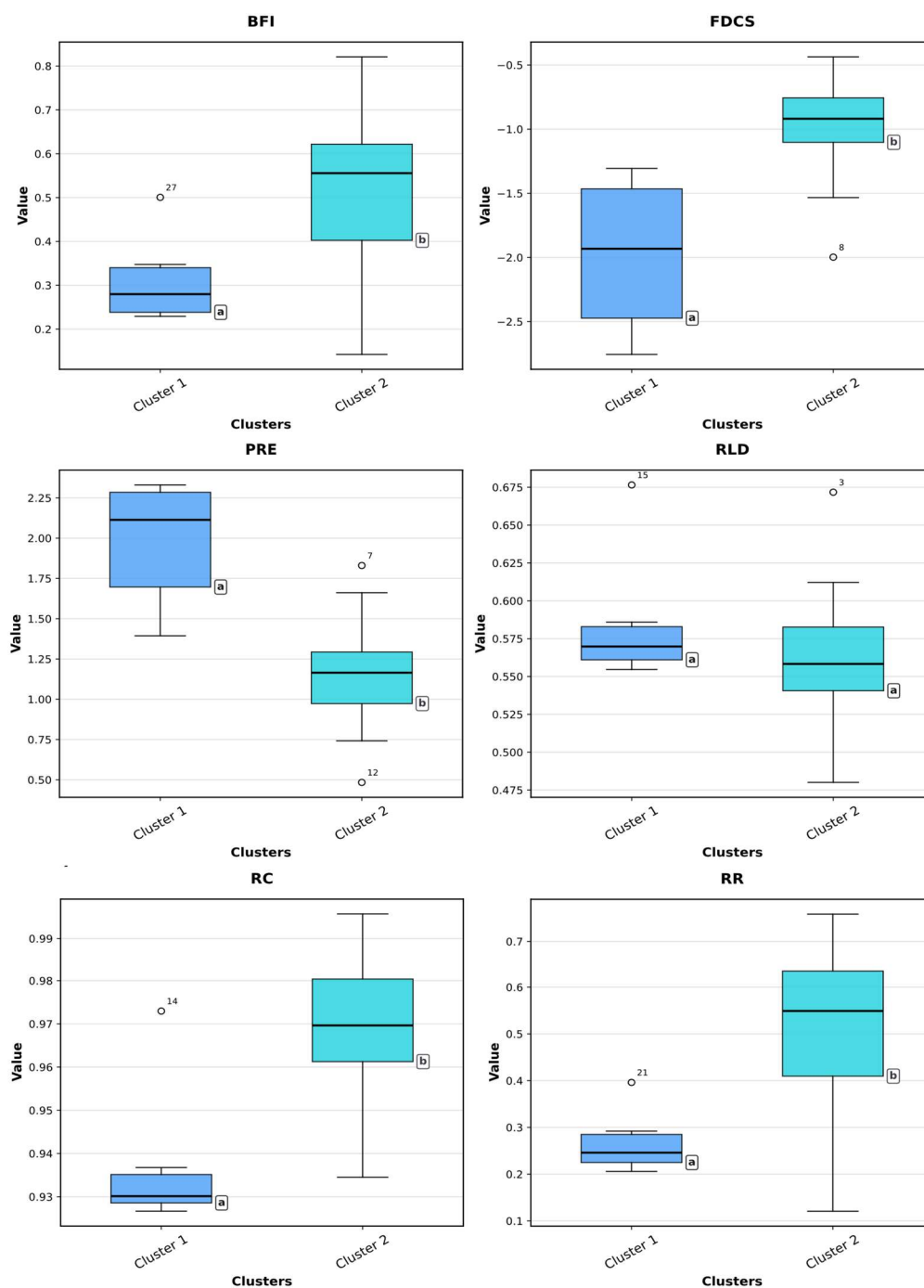


Figure: Signature distribution box plots for K-Means Euclidean, K-Medoids Euclidean and K-Medoids Manhattan clustering

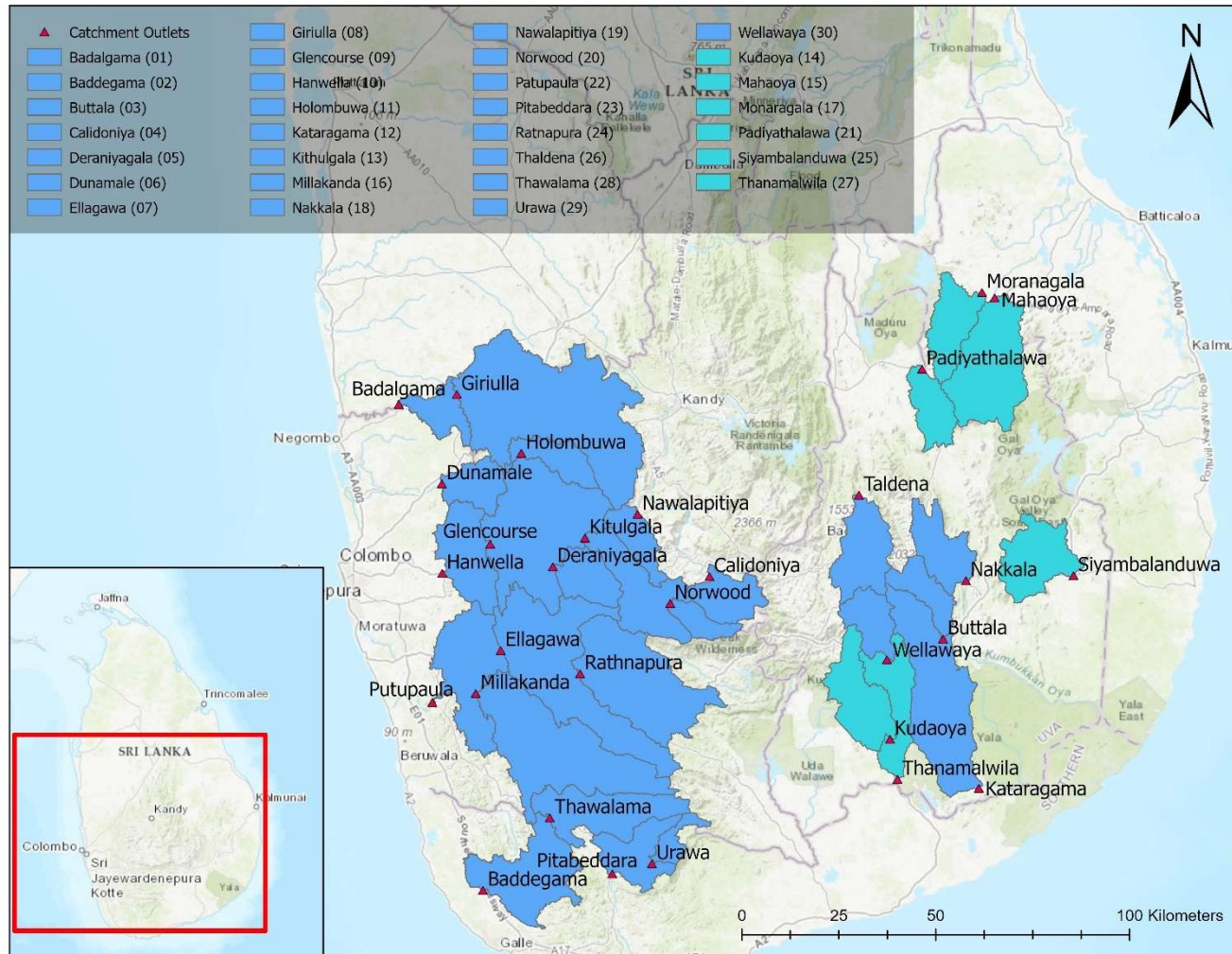


Figure: Spatial distribution of catchment clusters for K-Means Euclidean, K-Medoids Euclidean and K-Medoids Manhattan clustering

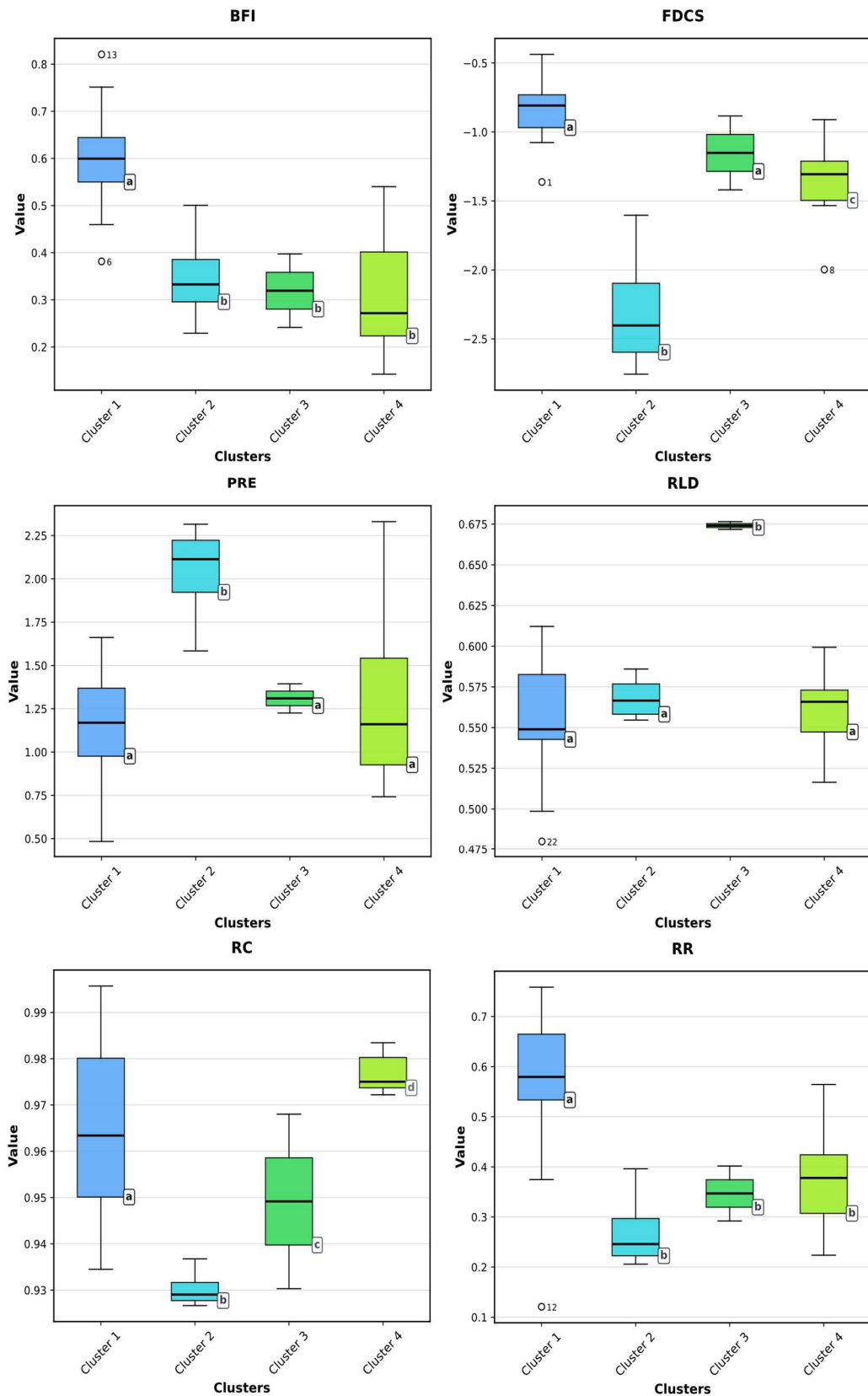


Figure: Signature distribution box plots for Hierarchical Euclidean (Ward Linkage) clustering

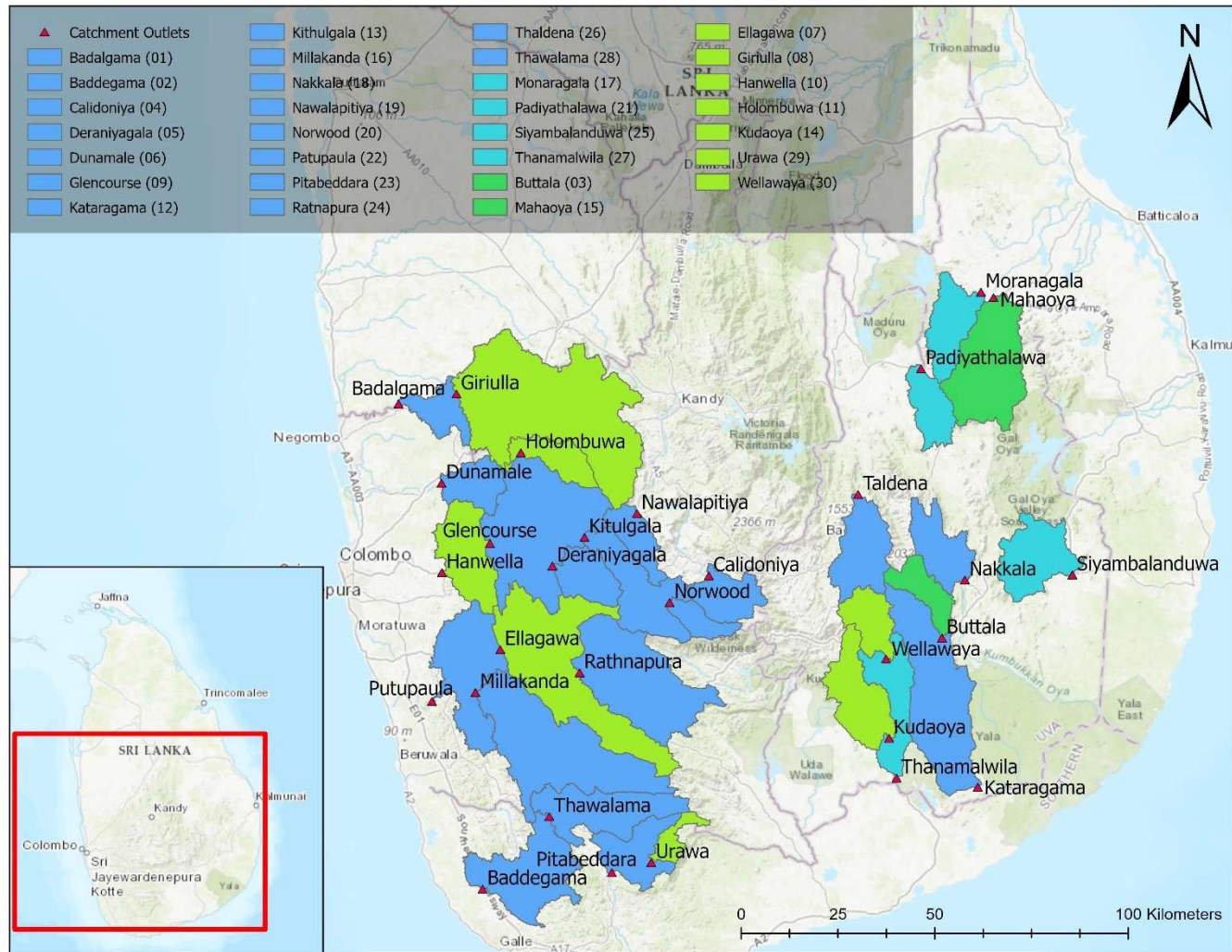


Figure: Spatial distribution of catchment clusters for Hierarchical Euclidean (Ward Linkage) clustering

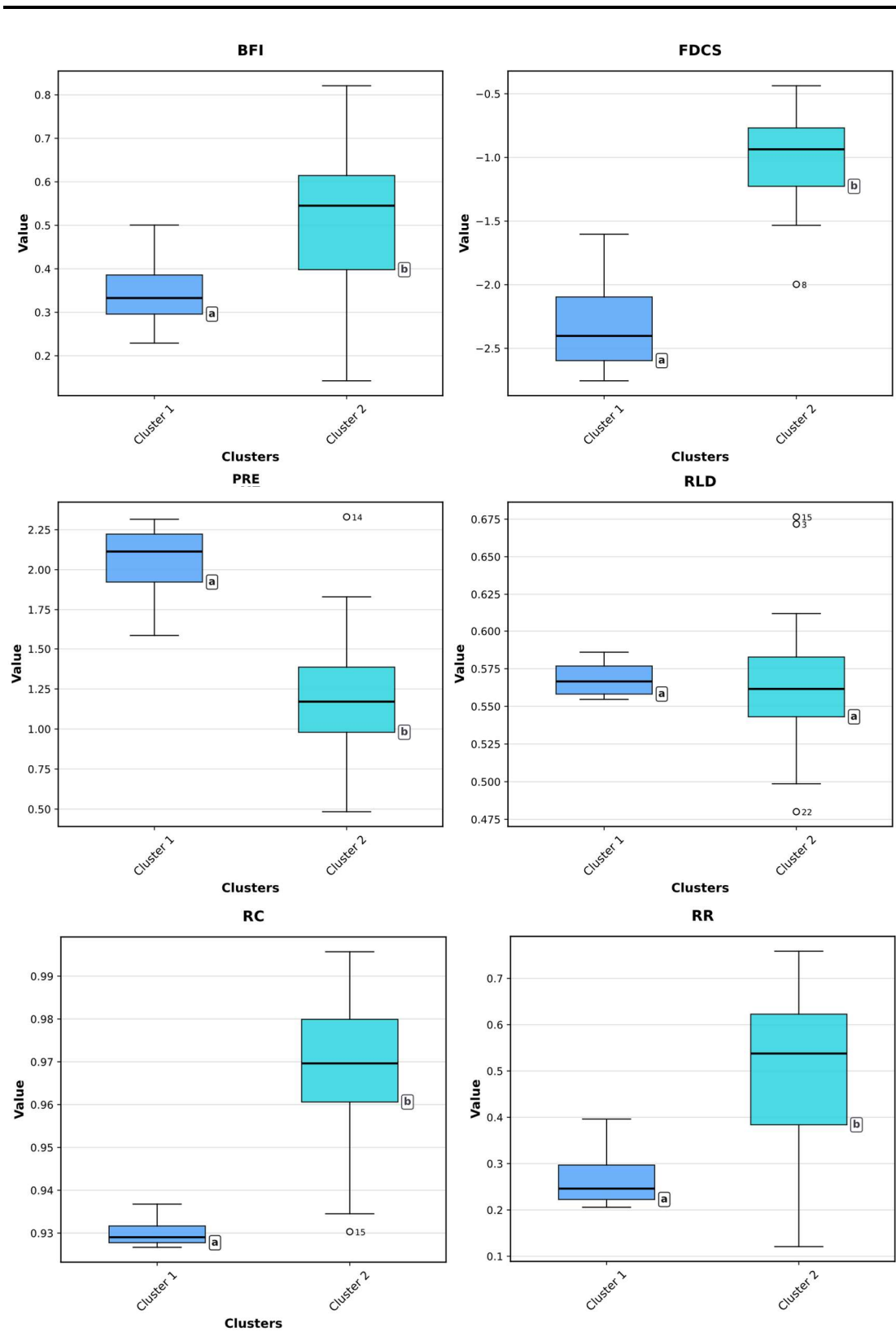


Figure: Signature distribution box plots for Hierarchical Euclidean (Average Linckage)
Hierarchical Manhattan clustering

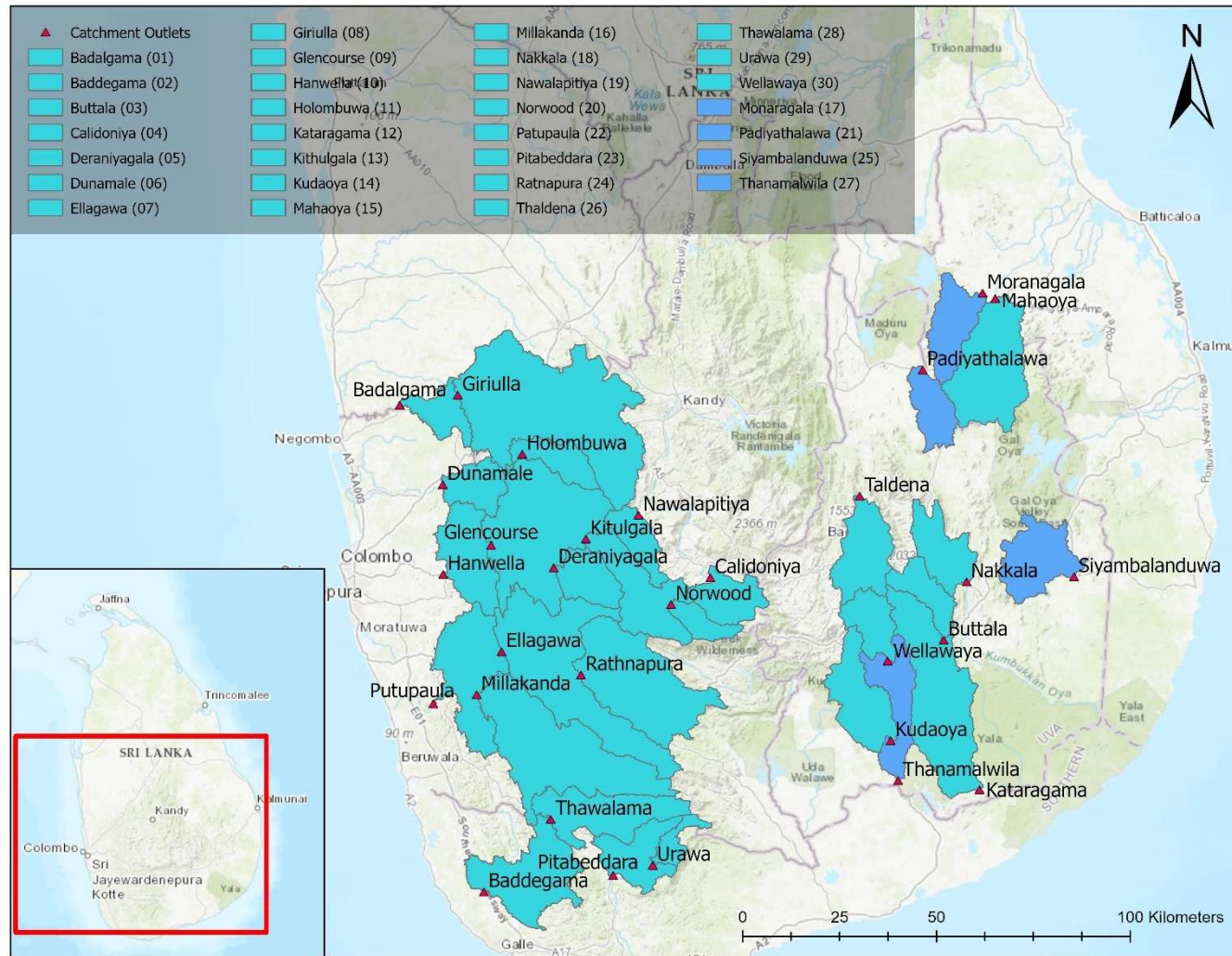


Figure: Spatial distribution of catchment clusters for Hierarchical Euclidean (Average Linkage) Hierarchical Manhattan clustering

ANNEXURE 4

FULL-PERIOD TIME SERIES (TRAINING DATA: REGIONAL LSTM): DRY ZONE CASE STUDY

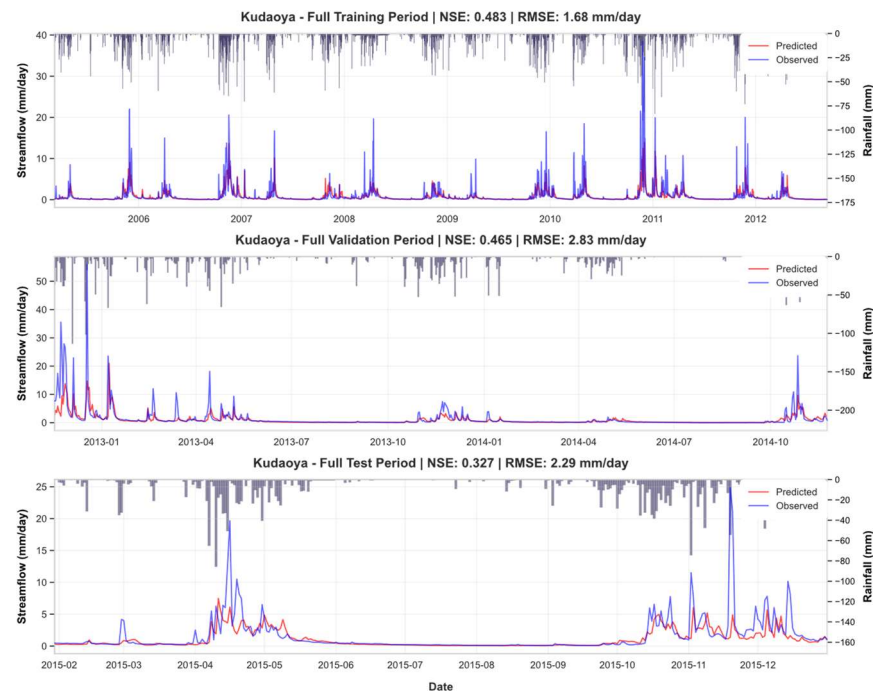


Figure: Complete training, validating, testing period streamflow hydrograph for Kudaoya (linear scale)

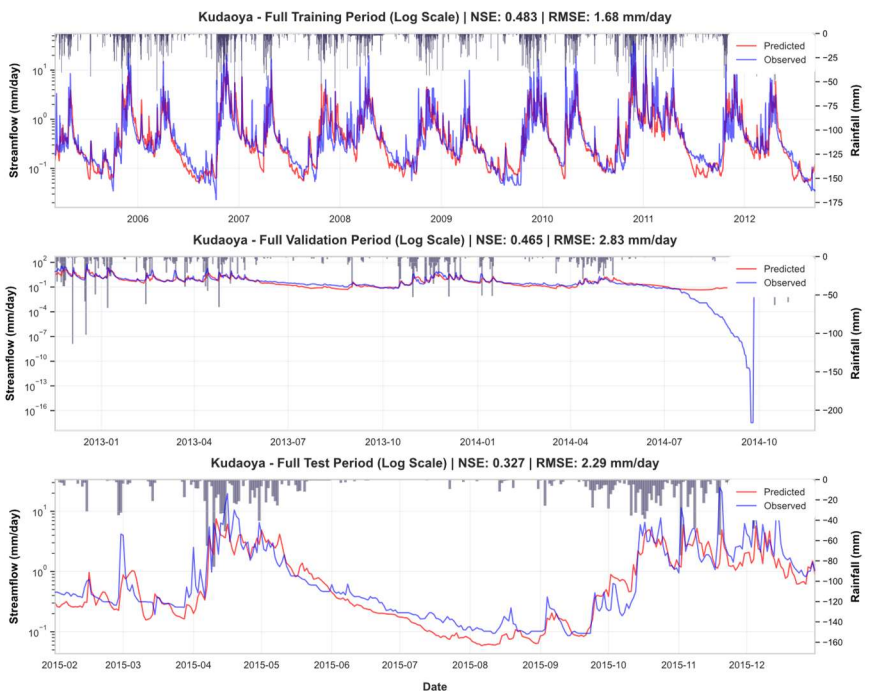


Figure: Complete training, validating, testing period streamflow hydrograph for Kudaoya (log scale)

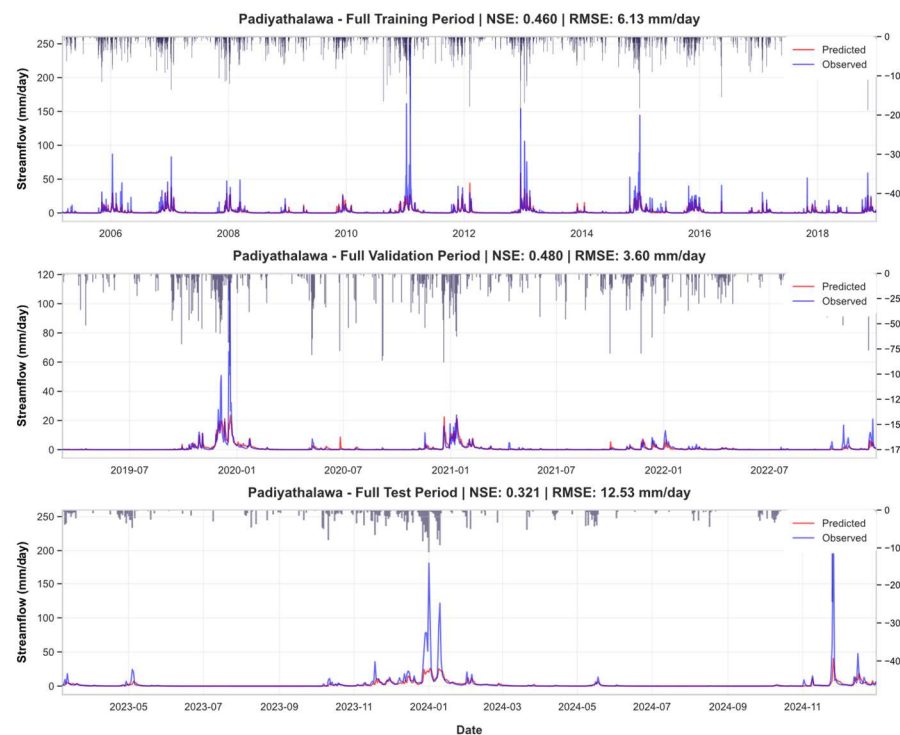


Figure: Complete training, validating, testing period streamflow hydrograph for Padiyathalawa (linear scale)

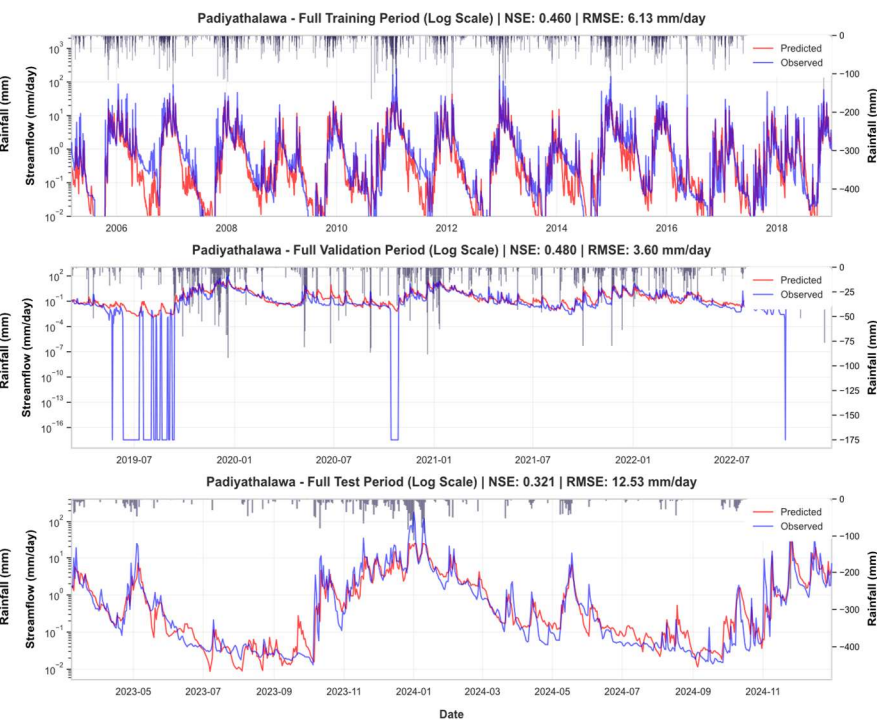


Figure: Complete training, validating, testing period streamflow hydrograph for Padiyathalawa (log scale)

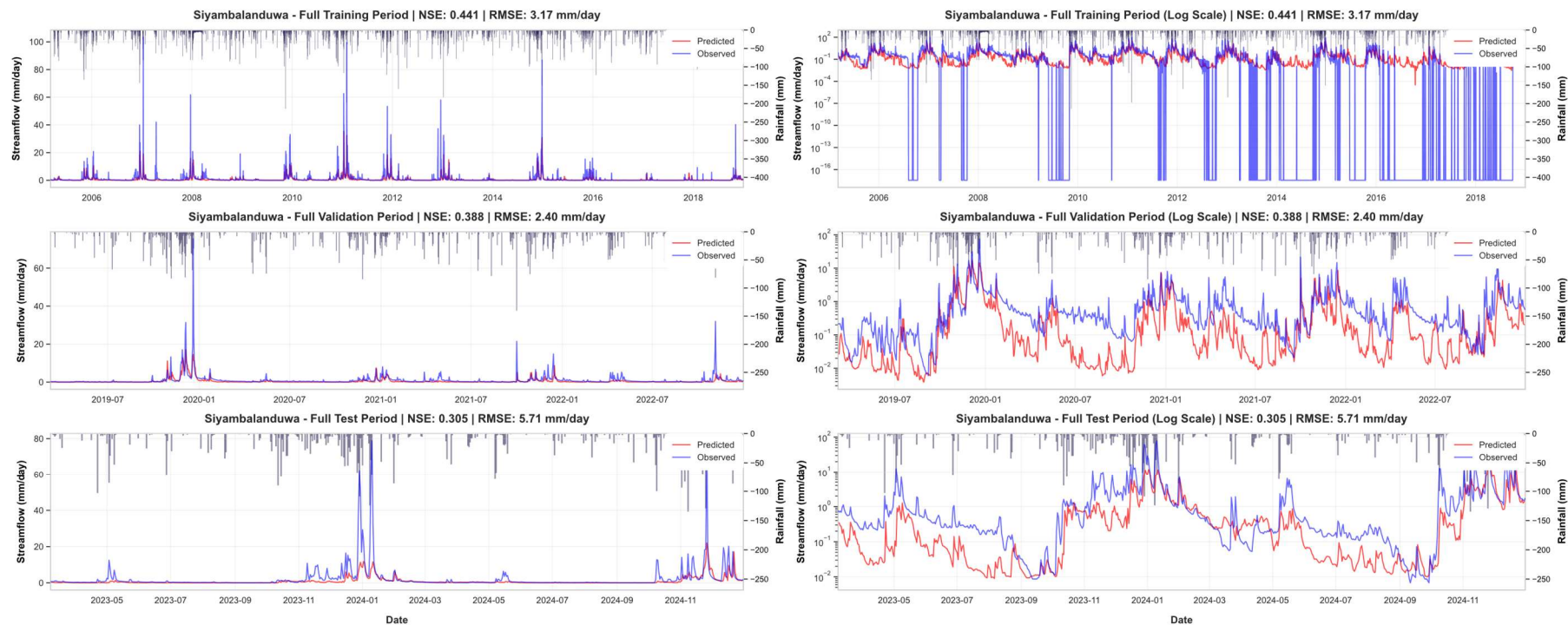


Figure: Complete training, validating, testing period streamflow hydrograph for Siyambalanduwa (linear scale)

Figure: Complete training, validating, testing period streamflow hydrograph for Siyambalanduwa (log scale)

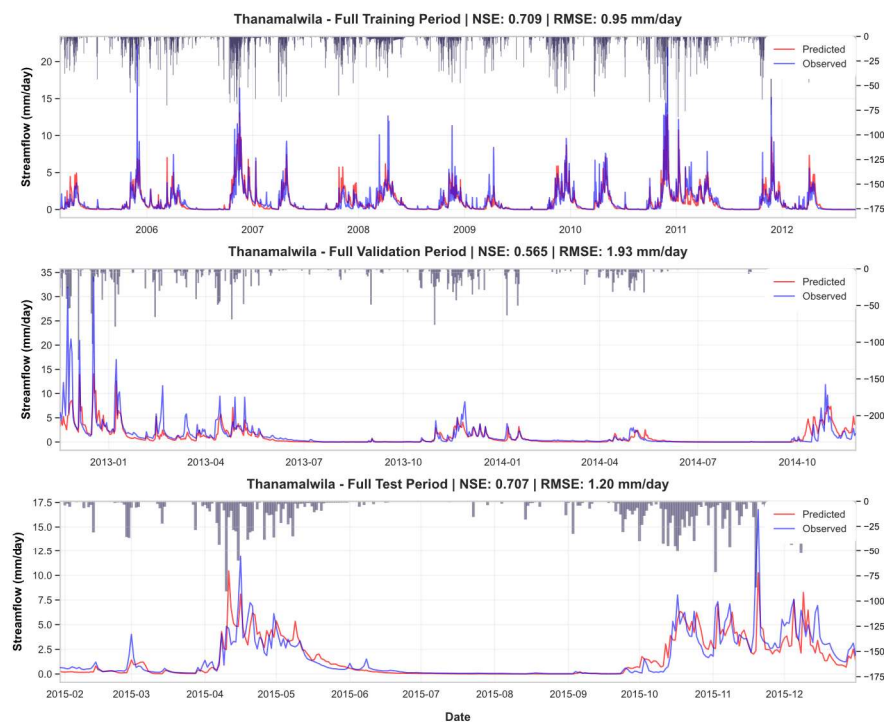


Figure: Complete training, validating, testing period streamflow hydrograph for Thanamalwila (linear scale)

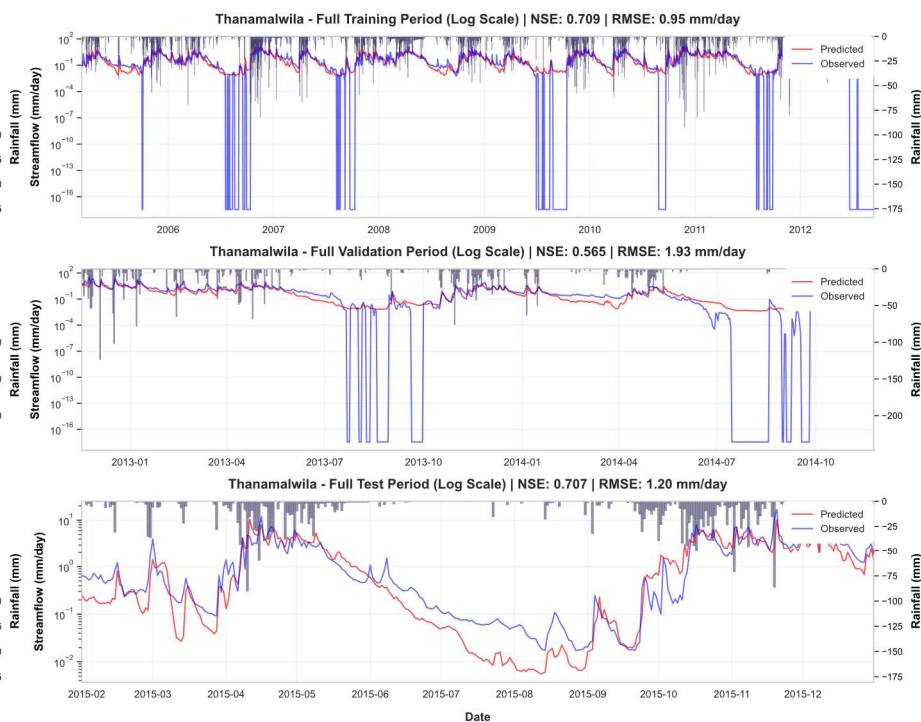


Figure: Complete training, validating, testing period streamflow hydrograph for Thanamalwila (log scale)

FULL-PERIOD FLOW DURATION CURVES (TRAINING DATA: REGIONAL LSTM): DRY ZONE CASE STUDY

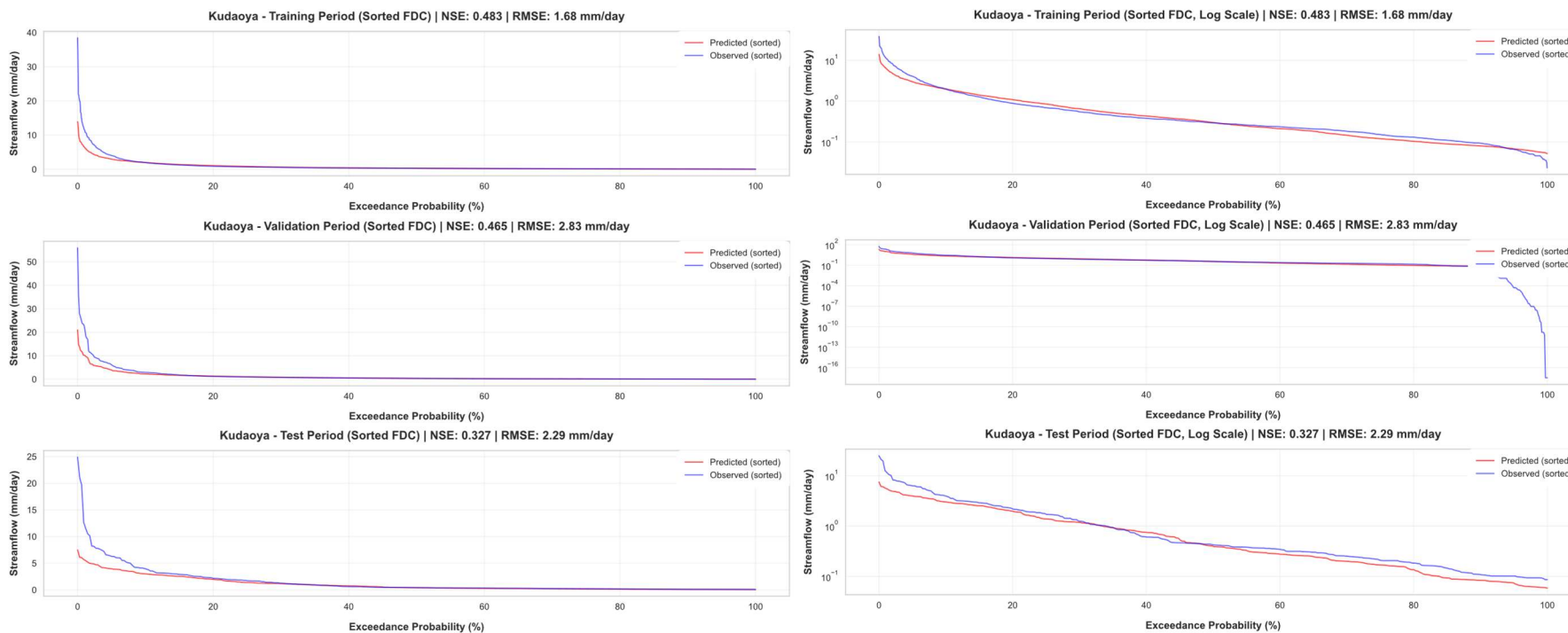


Figure: Complete training, validating, testing period sorted flow duration curves for Kudaoya (linear scale)

Figure: Complete training, validating, testing period sorted flow duration curves for Kudaoya (log scale)

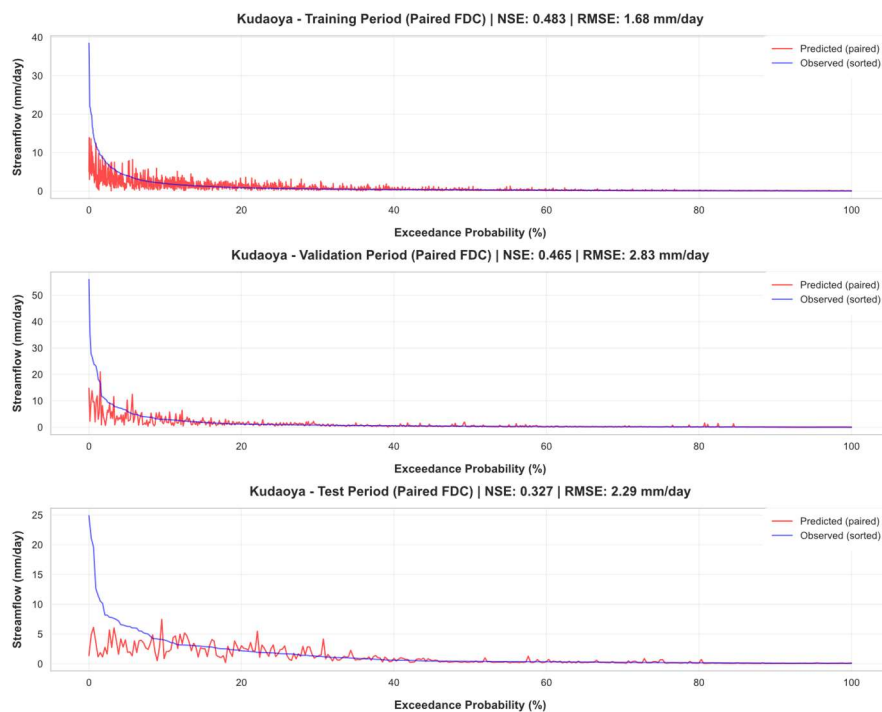


Figure: Complete training, validating, testing period unsorted flow duration curves for Kudaoya (linear scale)

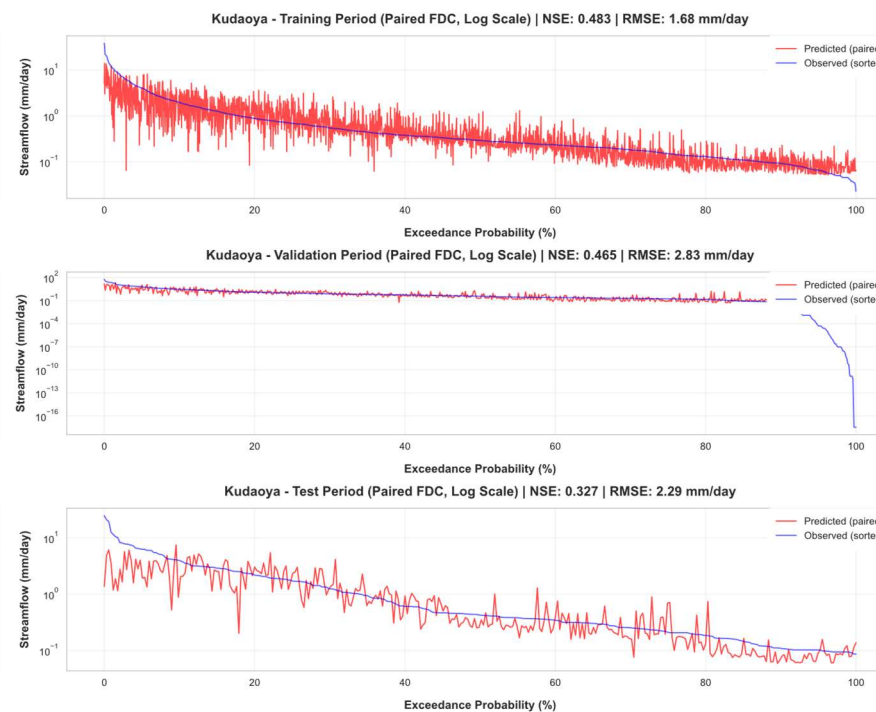


Figure: Complete training, validating, testing period unsorted flow duration curves for Kudaoya (log scale)

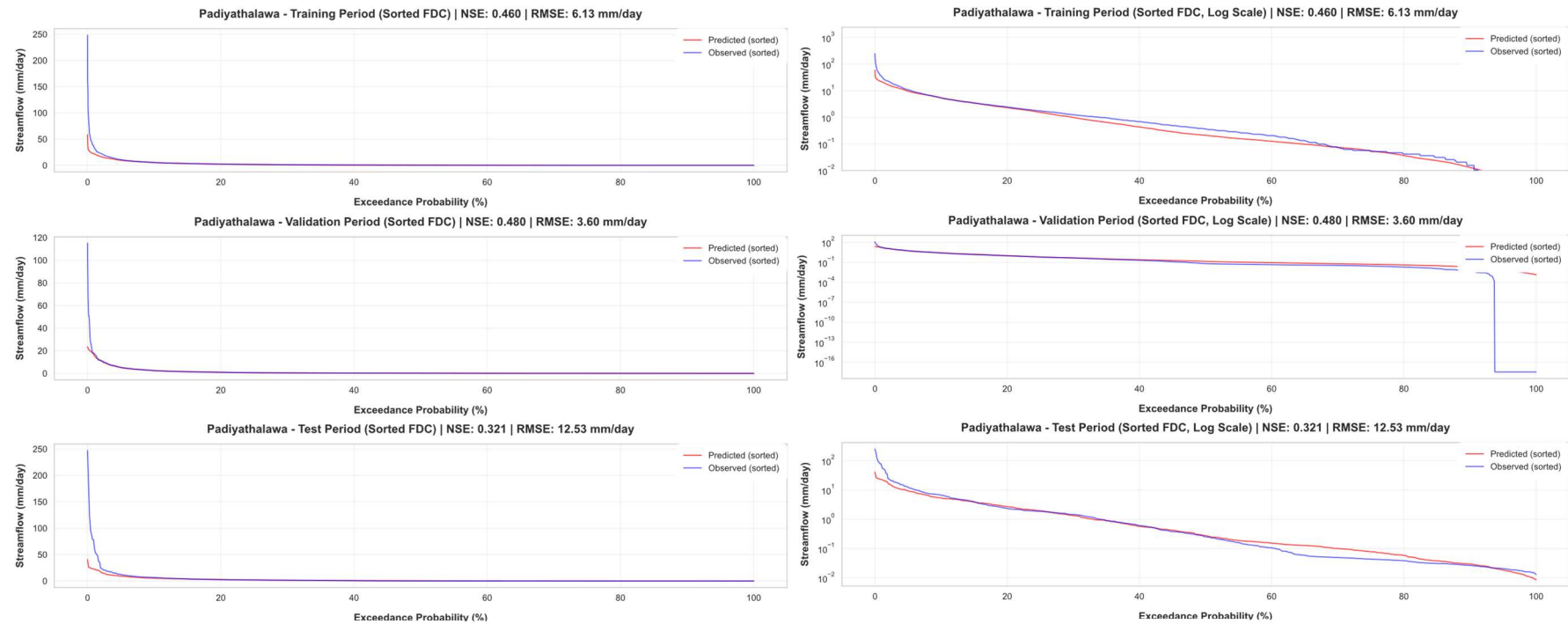


Figure: Complete training, validating, testing period sorted flow duration curves for Padiyathalawa (linear scale)

Figure: Complete training, validating, testing period sorted flow duration curves for Padiyathalawa (log scale)

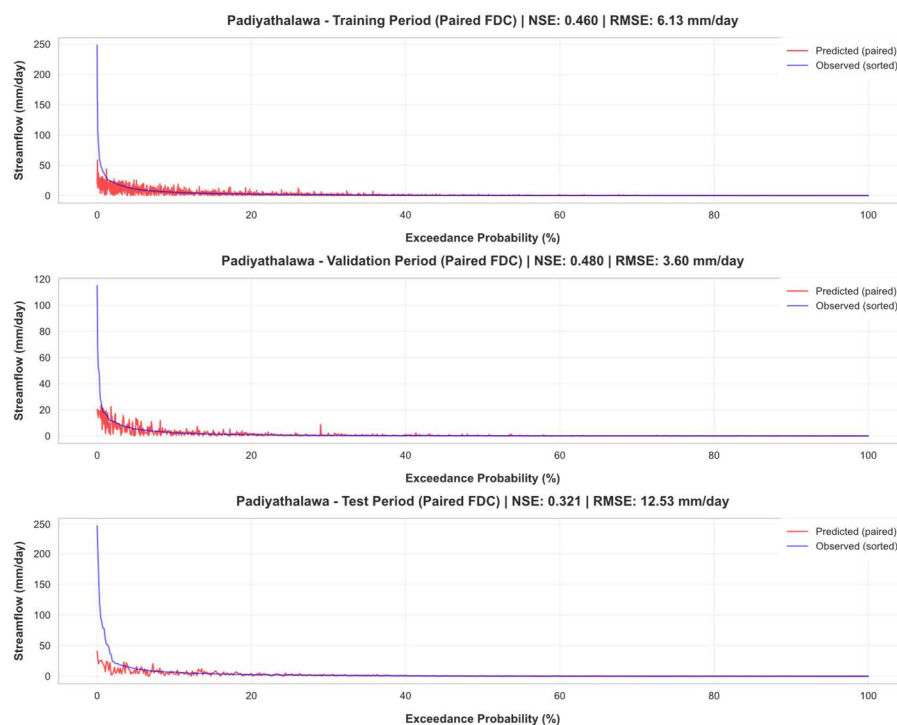


Figure: Complete training, validating, testing period unsorted flow duration curves for Padiyathalawa (linear scale)

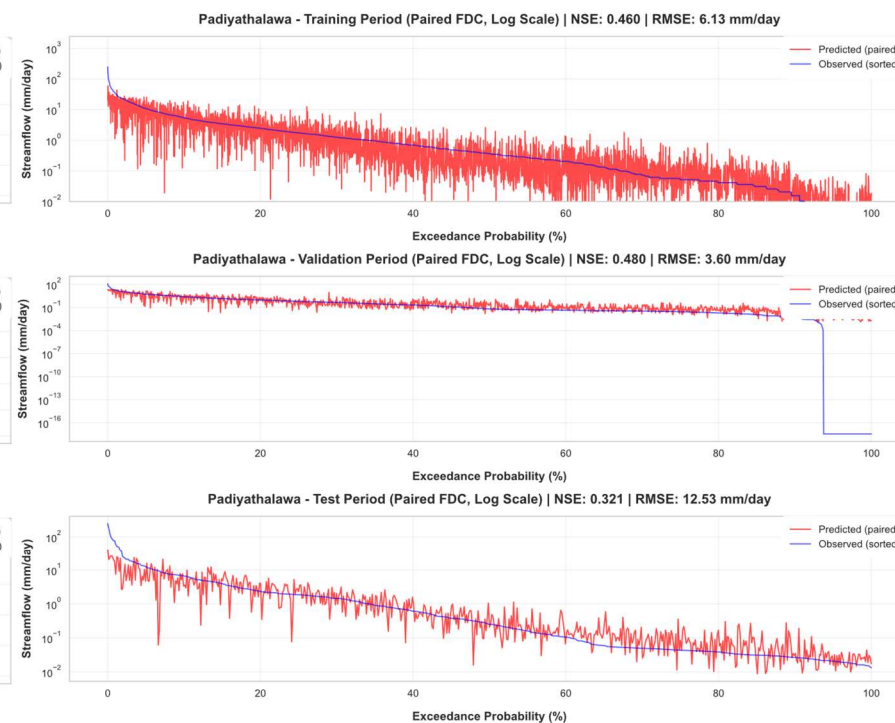


Figure: Complete training, validating, testing period unsorted flow duration curves for Padiyathalawa (log scale)

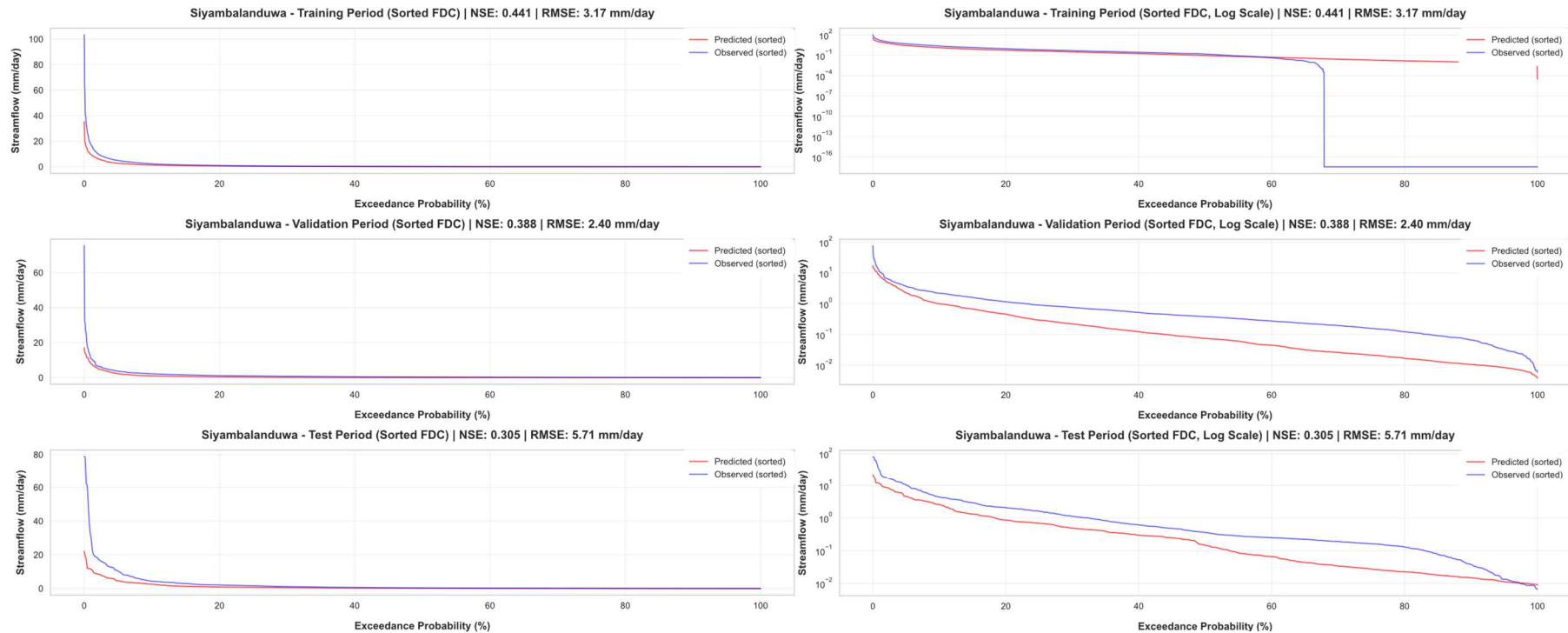


Figure: Complete training, validating, testing period sorted flow duration curves for Siyambalanduwa (linear scale)

Figure: Complete training, validating, testing period sorted flow duration curves for Siyambalanduwa (log scale)

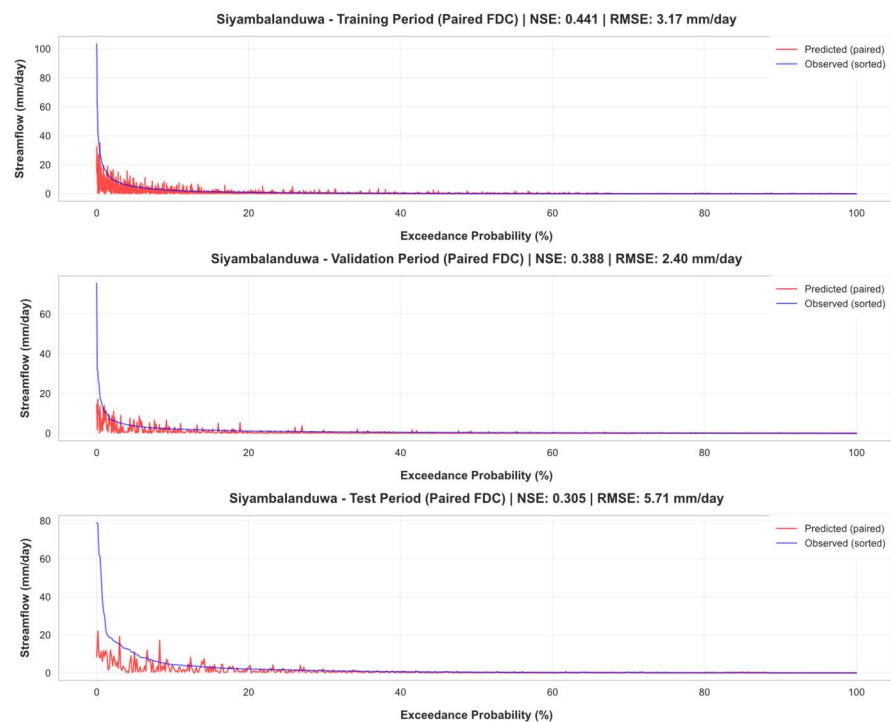


Figure: Complete training, validating, testing period unsorted flow duration curves for Siyambalanduwa (linear scale)

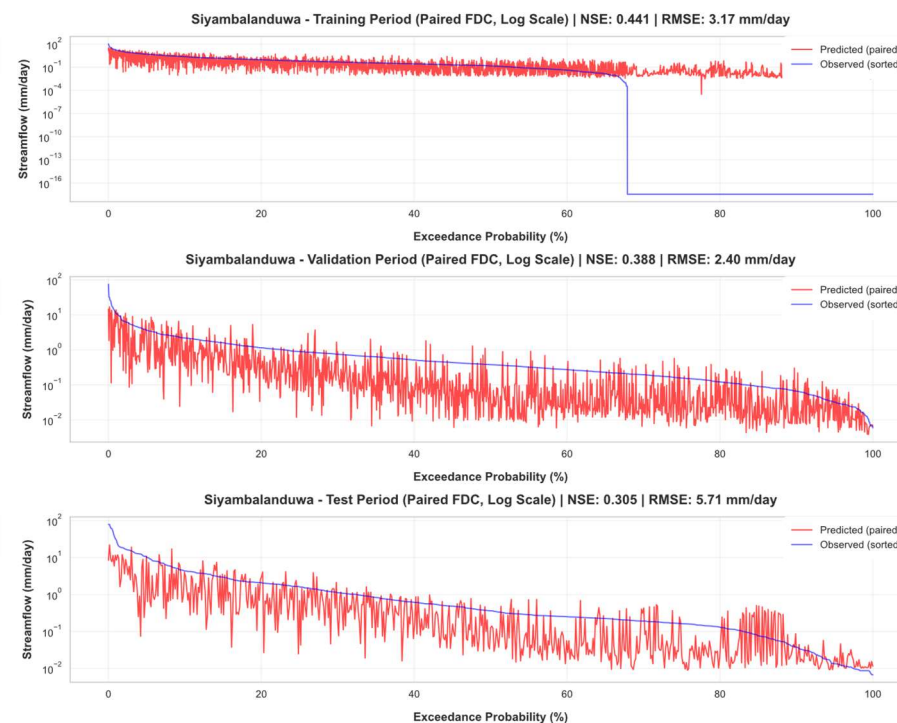


Figure: Complete training, validating, testing period unsorted flow duration curves for Siyambalanduwa (log scale)

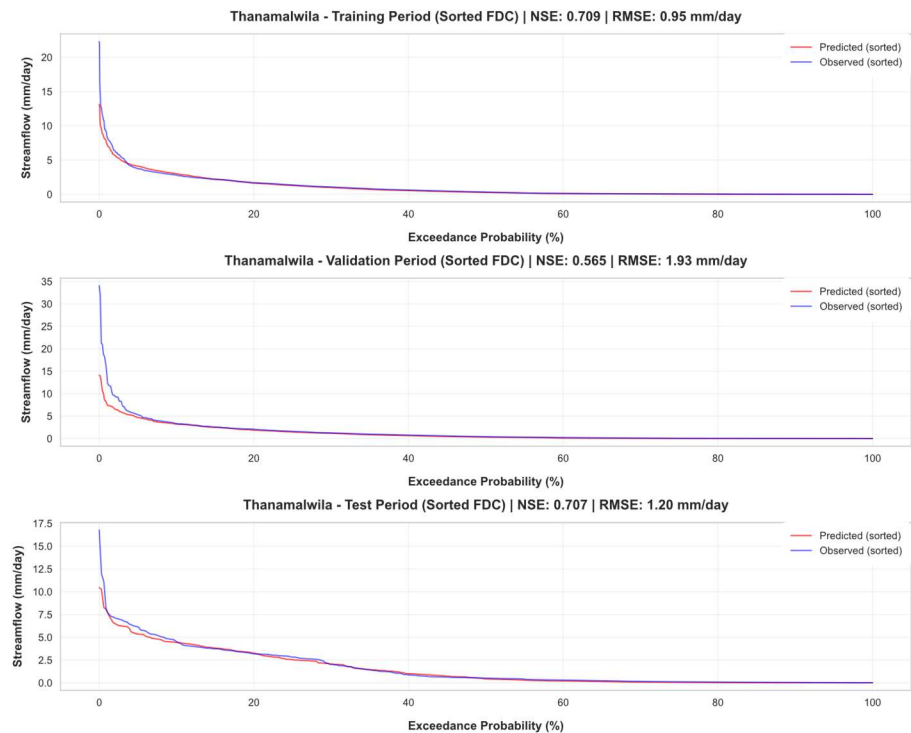


Figure: Complete training, validating, testing period sorted flow duration curves for Thanamalwila (linear scale)

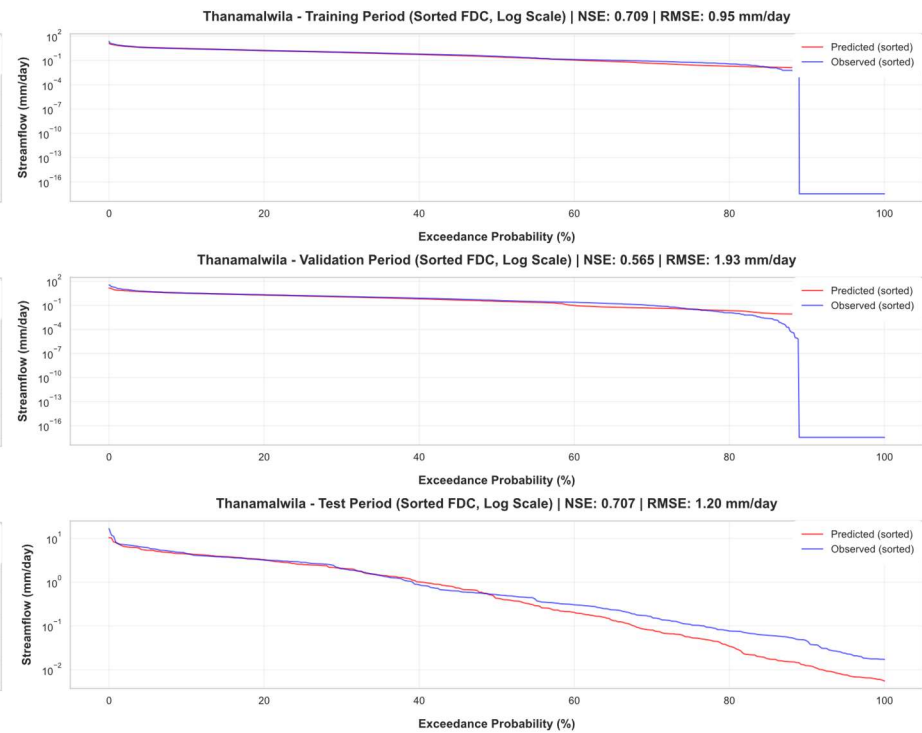


Figure: Complete training, validating, testing period sorted flow duration curves for Thanamalwila (log scale)

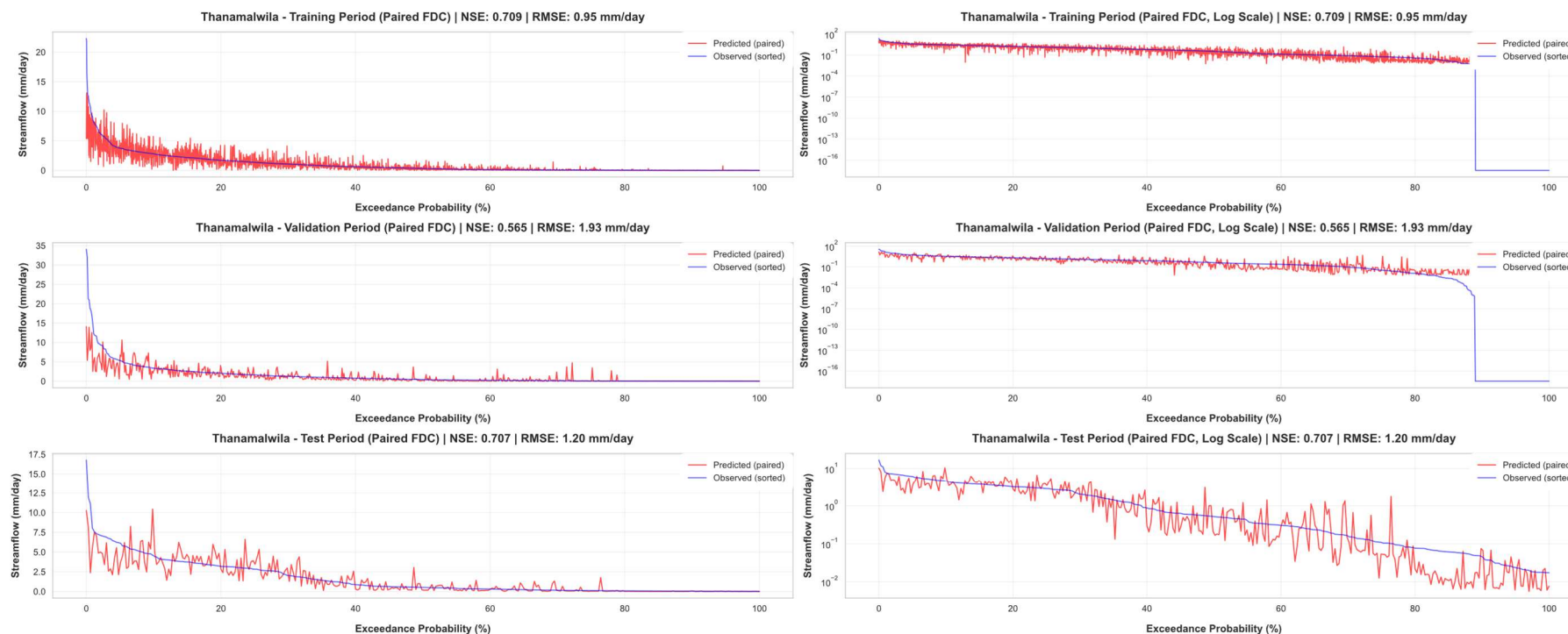


Figure: Complete training, validating, testing period unsorted flow duration curves for Thanamalwila (linear scale)

Figure: Complete training, validating, testing period unsorted flow duration curves for Thanamalwila (log scale)

FULL-PERIOD TIME SERIES (PREDICTION DATA: REGIONAL LSTM): DRY ZONE CASE STUDY

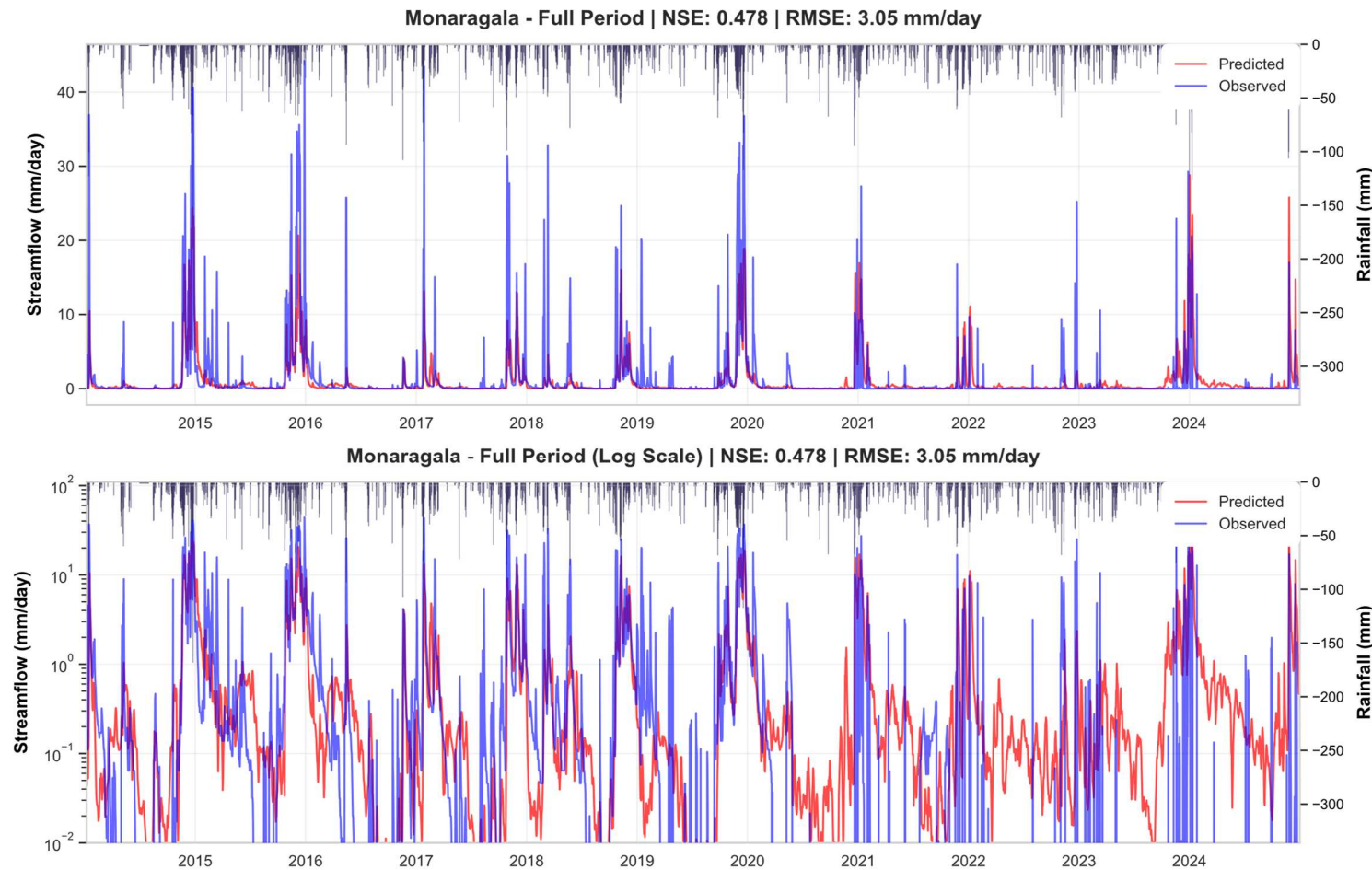


Figure: Complete prediction data period streamflow hydrograph for Monaragala (2014–2025)

FULL-PERIOD TIME SERIES (TRAINING DATA: REGIONAL LSTM): WET ZONE CASE STUDY

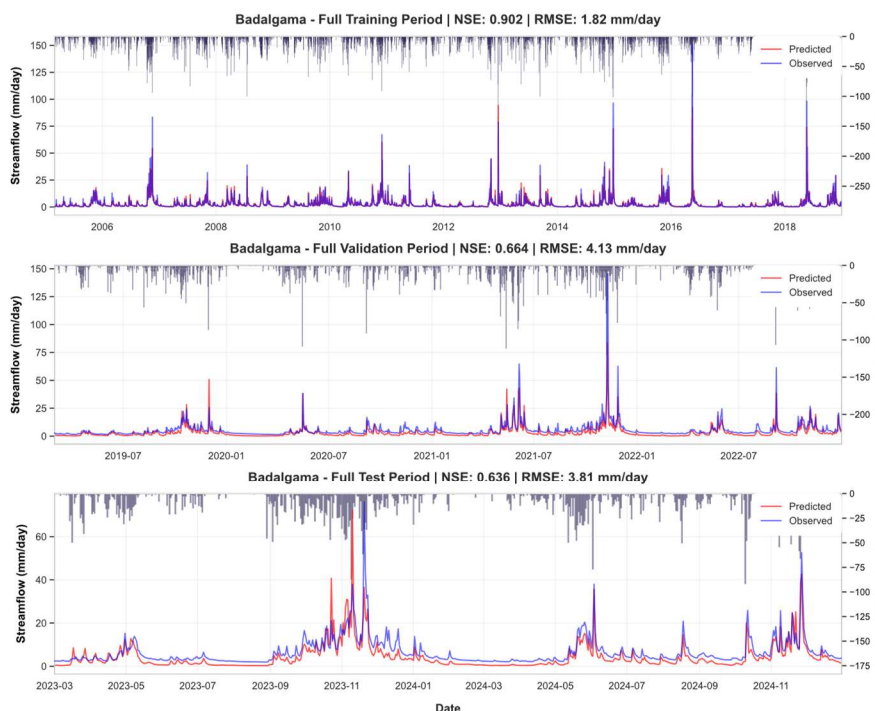


Figure: Complete training, validating, testing period streamflow hydrograph for Badalgama (linear scale)

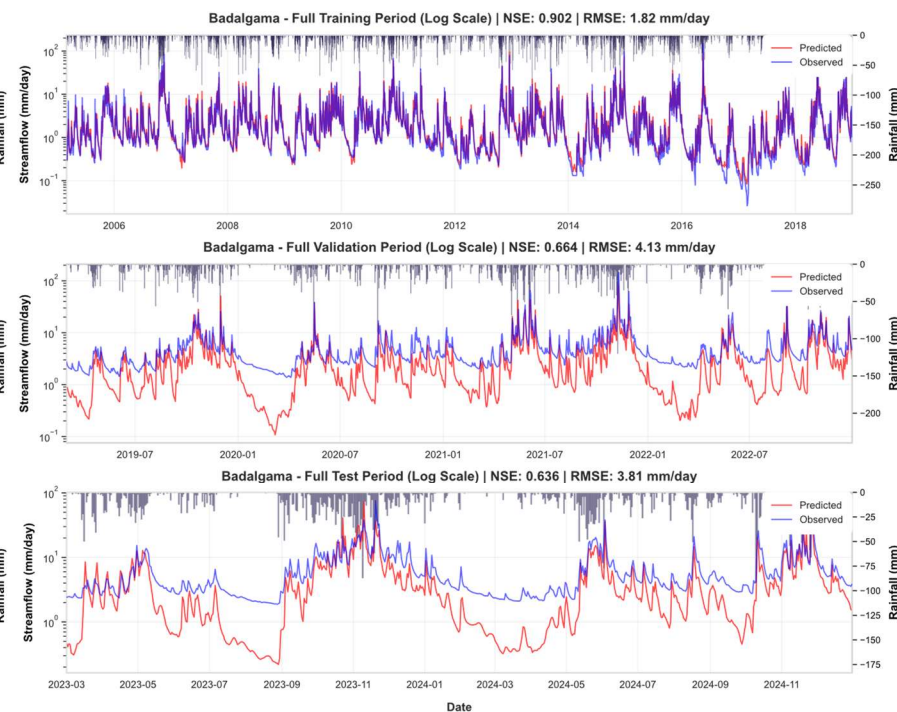


Figure: Complete training, validating, testing period streamflow hydrograph for Badalgama (log scale)

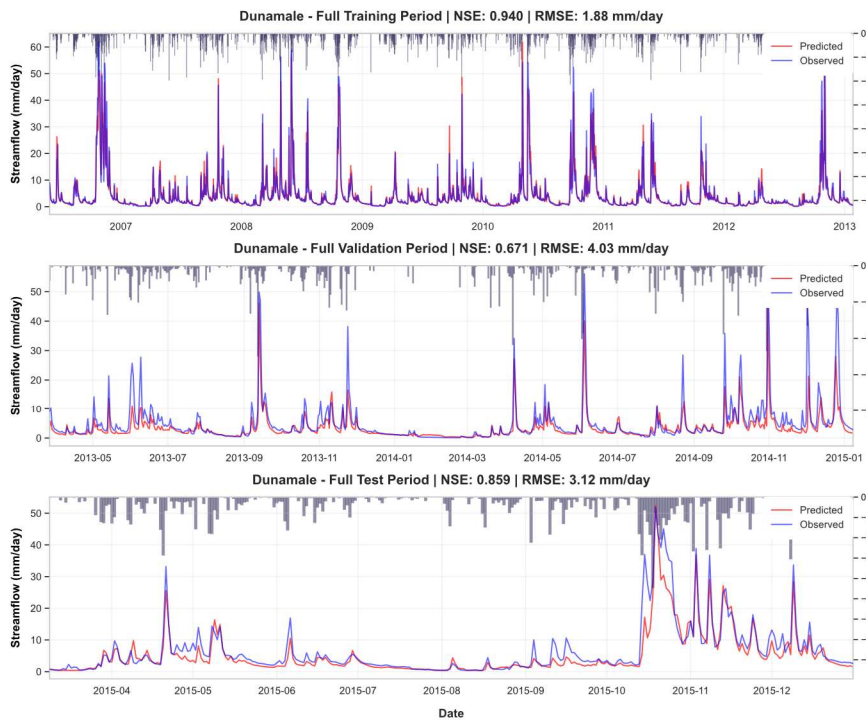


Figure: Complete training, validating, testing period streamflow hydrograph for Dunamale (linear scale)

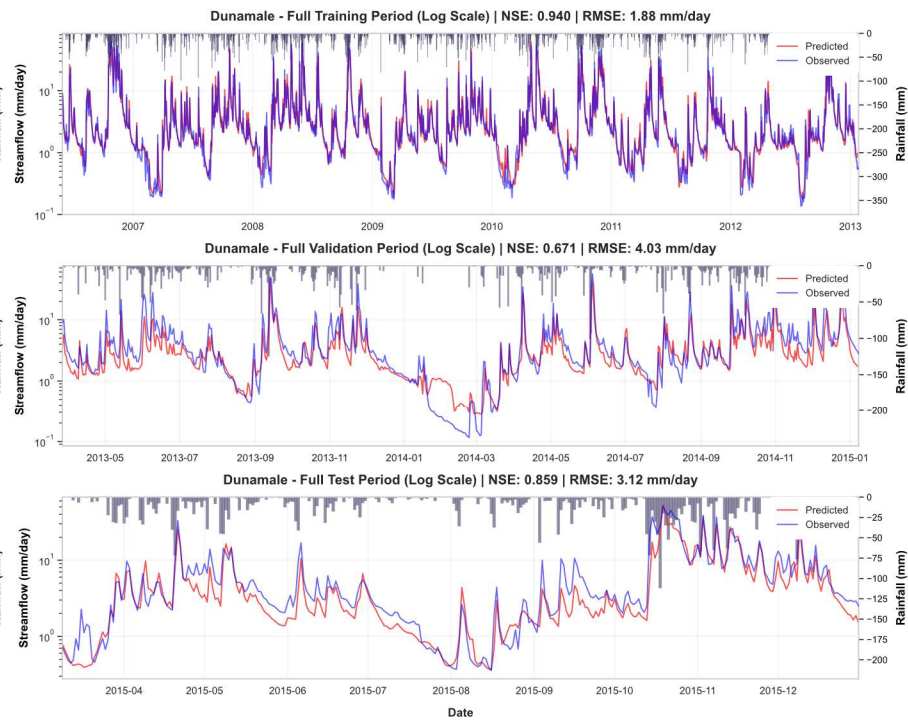


Figure: Complete training, validating, testing period streamflow hydrograph for Dunamale (log scale)

FULL-PERIOD FLOW DURATION CURVES (TRAINING DATA: REGIONAL LSTM): WET ZONE CASE STUDY

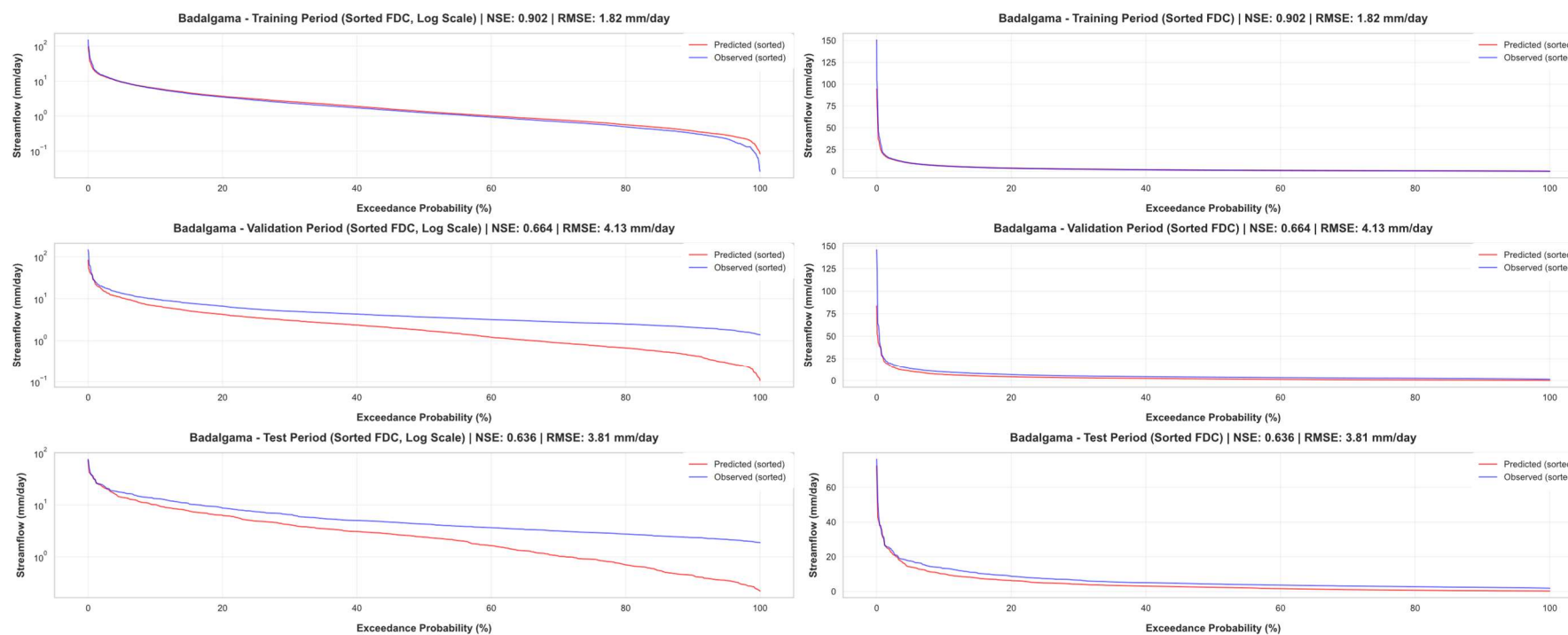


Figure: Complete training, validating, testing period sorted flow duration curves for Badalgama (linear scale)

Figure: Complete training, validating, testing period sorted flow duration curves for Badalgama (log scale)

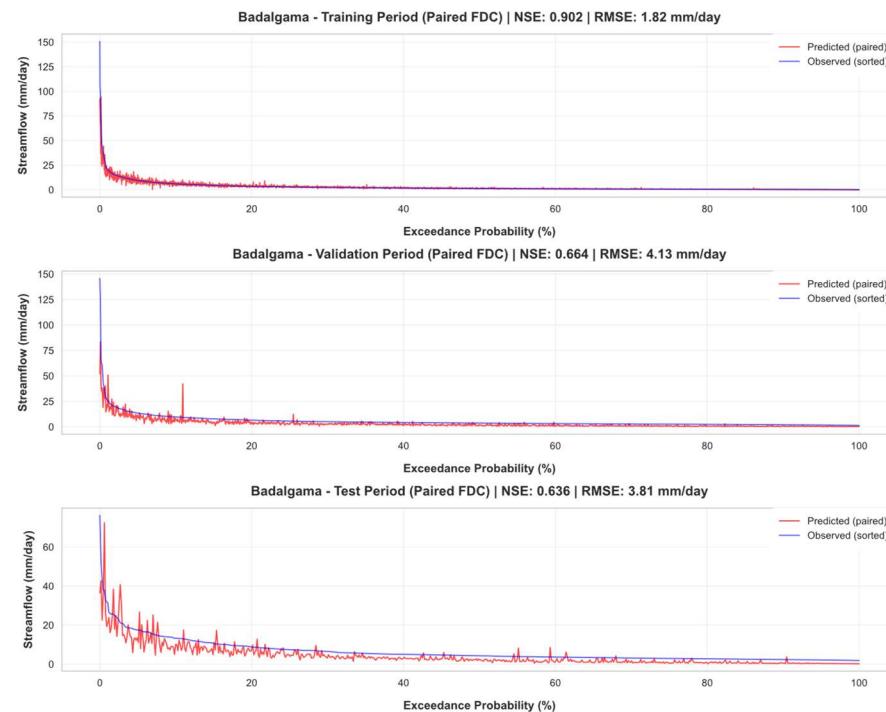


Figure: Complete training, validating, testing period unsorted flow duration curves for Badalgama (linear scale)

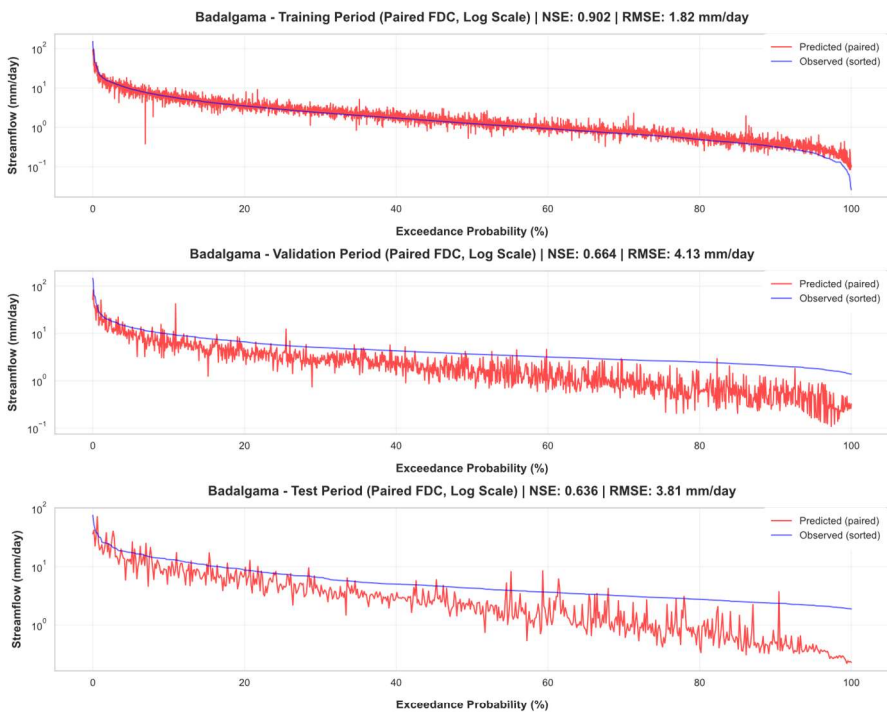


Figure: Complete training, validating, testing period unsorted flow duration curves for Badalgama (log scale)

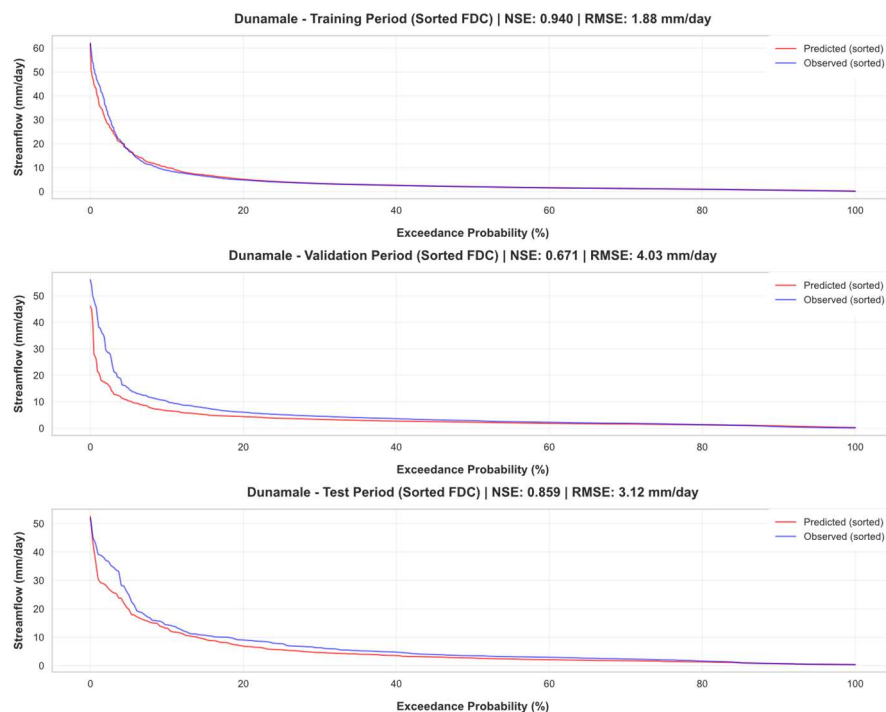


Figure: Complete training, validating, testing period sorted flow duration curves for Dunamale (linear scale)

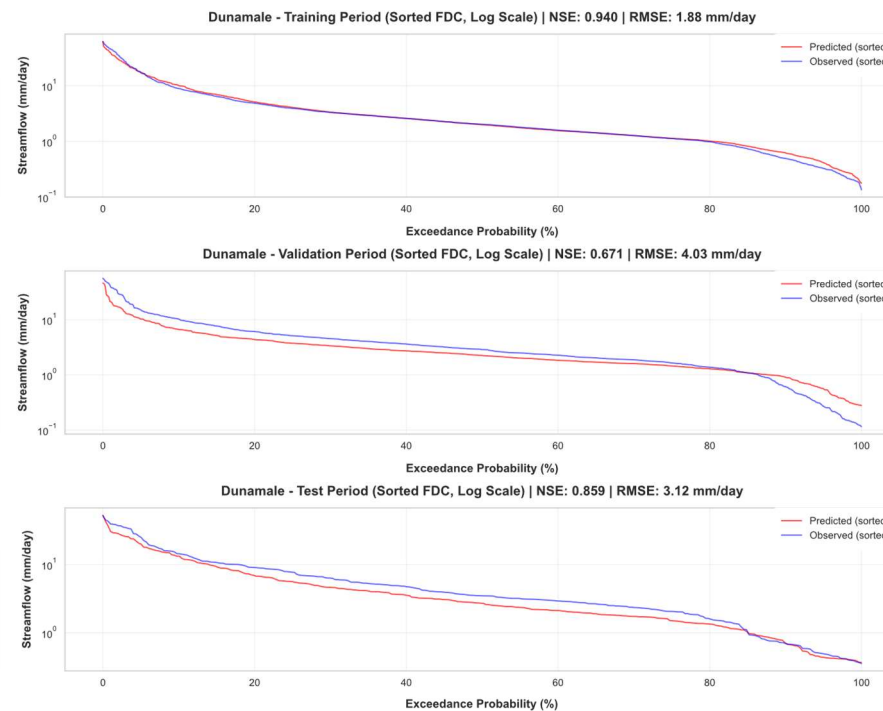


Figure: Complete training, validating, testing period sorted flow duration curves for Dunamale (log scale)

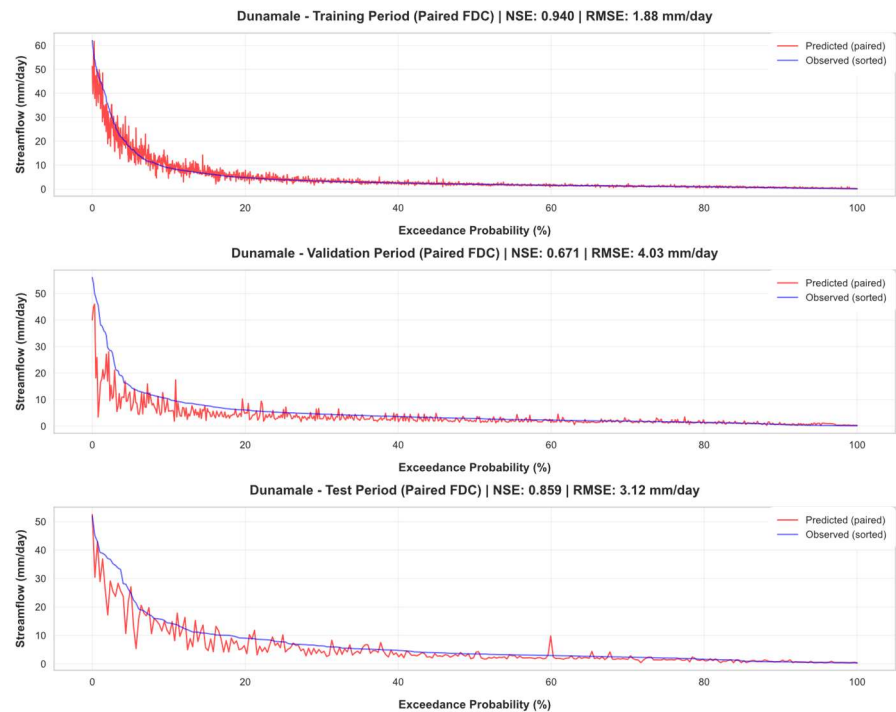


Figure: Complete training, validating, testing period unsorted flow duration curves for Dunamale (linear scale)

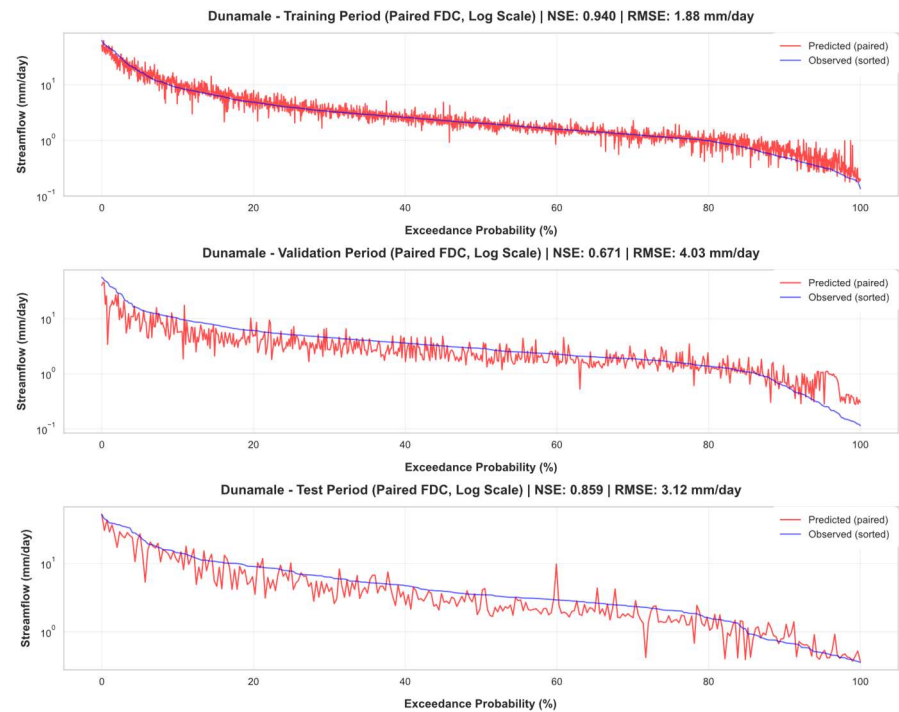


Figure: Complete training, validating, testing period unsorted flow duration curves for Dunamale (log scale)

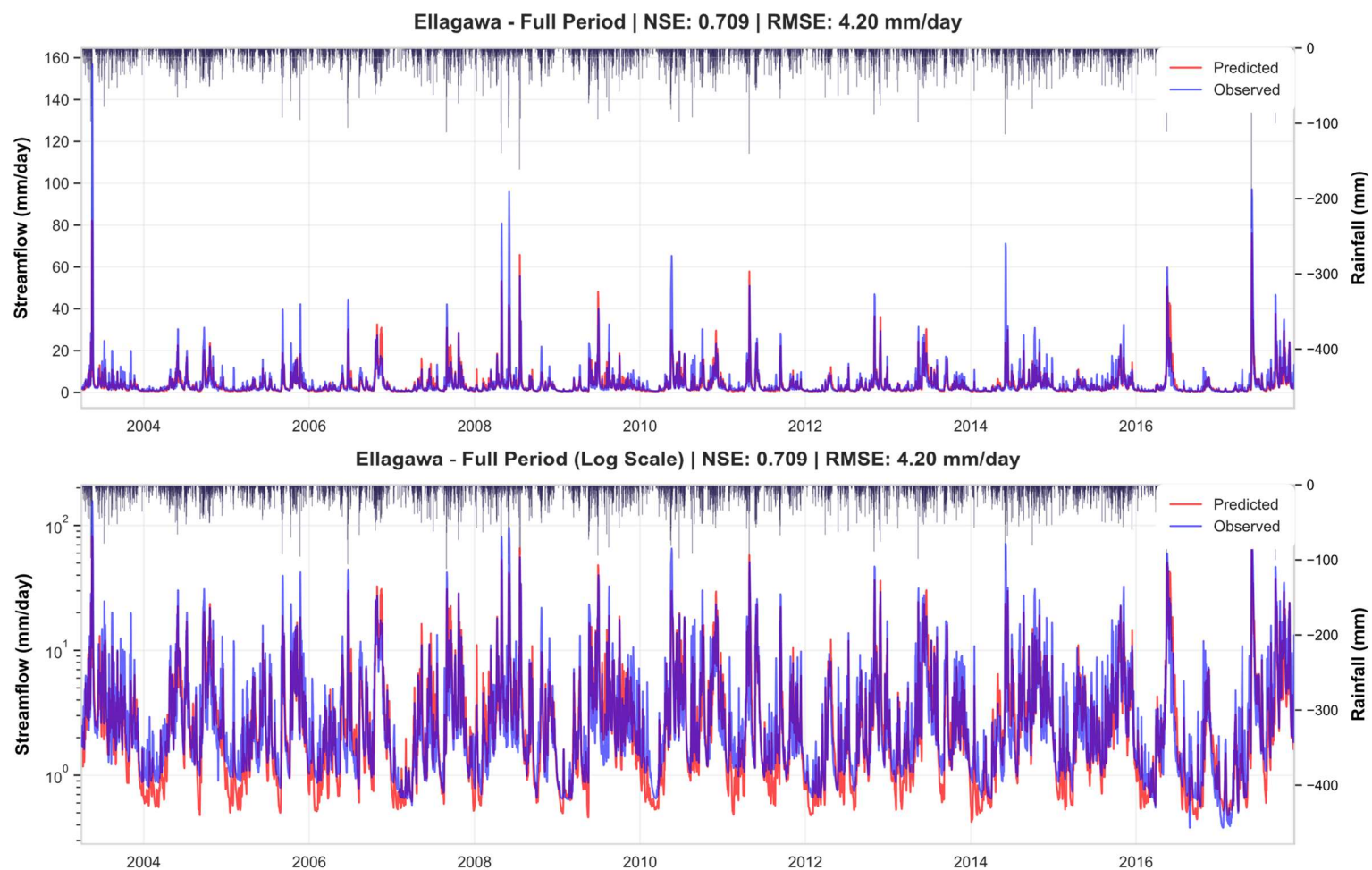
FULL-PERIOD TIME SERIES (PREDICTION DATA: REGIONAL LSTM): WET ZONE CASE STUDY

Figure: Complete prediction data period streamflow hydrograph for Ellagawa (2004–2017)

ANNEXURE 5

FULL-PERIOD TIME SERIES (PREDICTION DATA: GLOBAL LSTM): DRY ZONE CASE STUDY

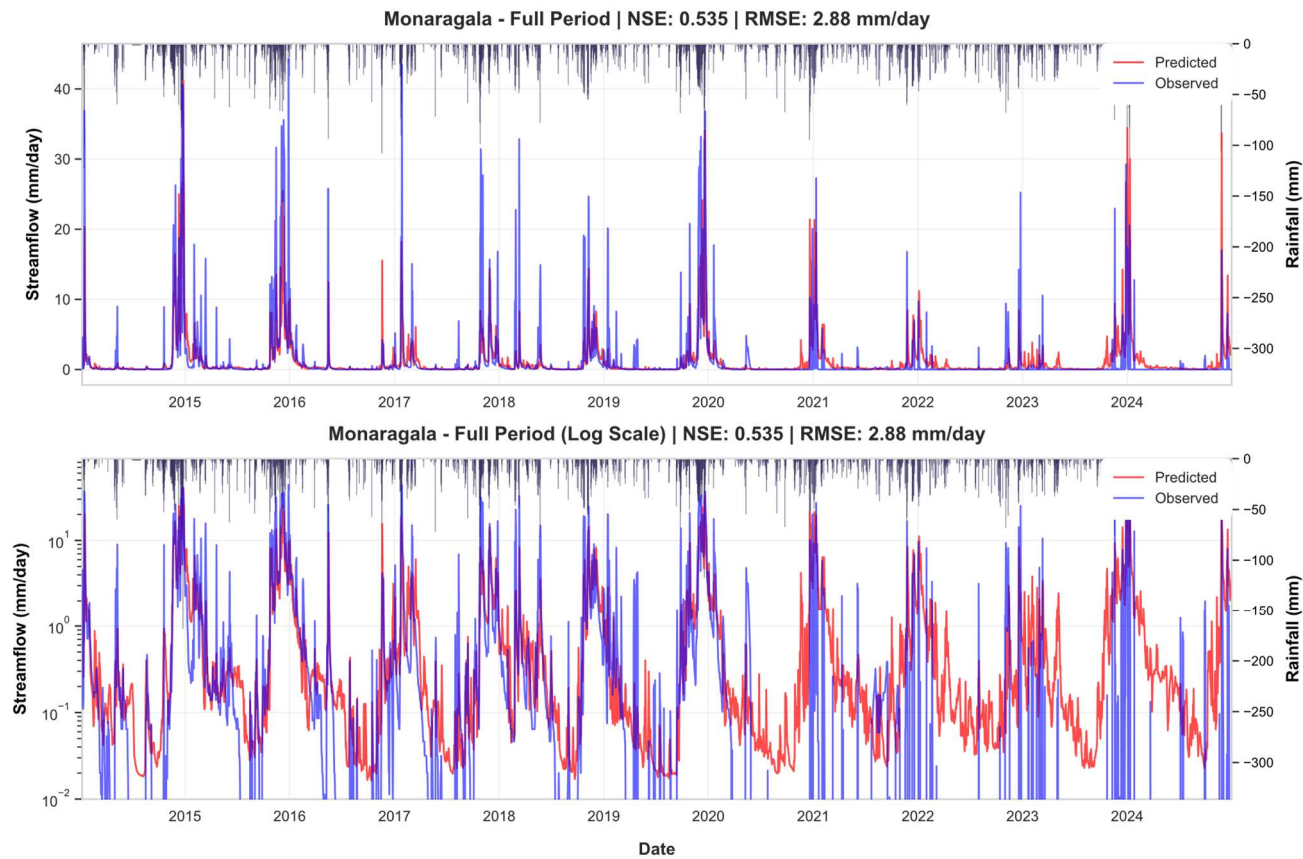


Figure: Complete prediction data period streamflow hydrograph for Monaragala (2014–2025)

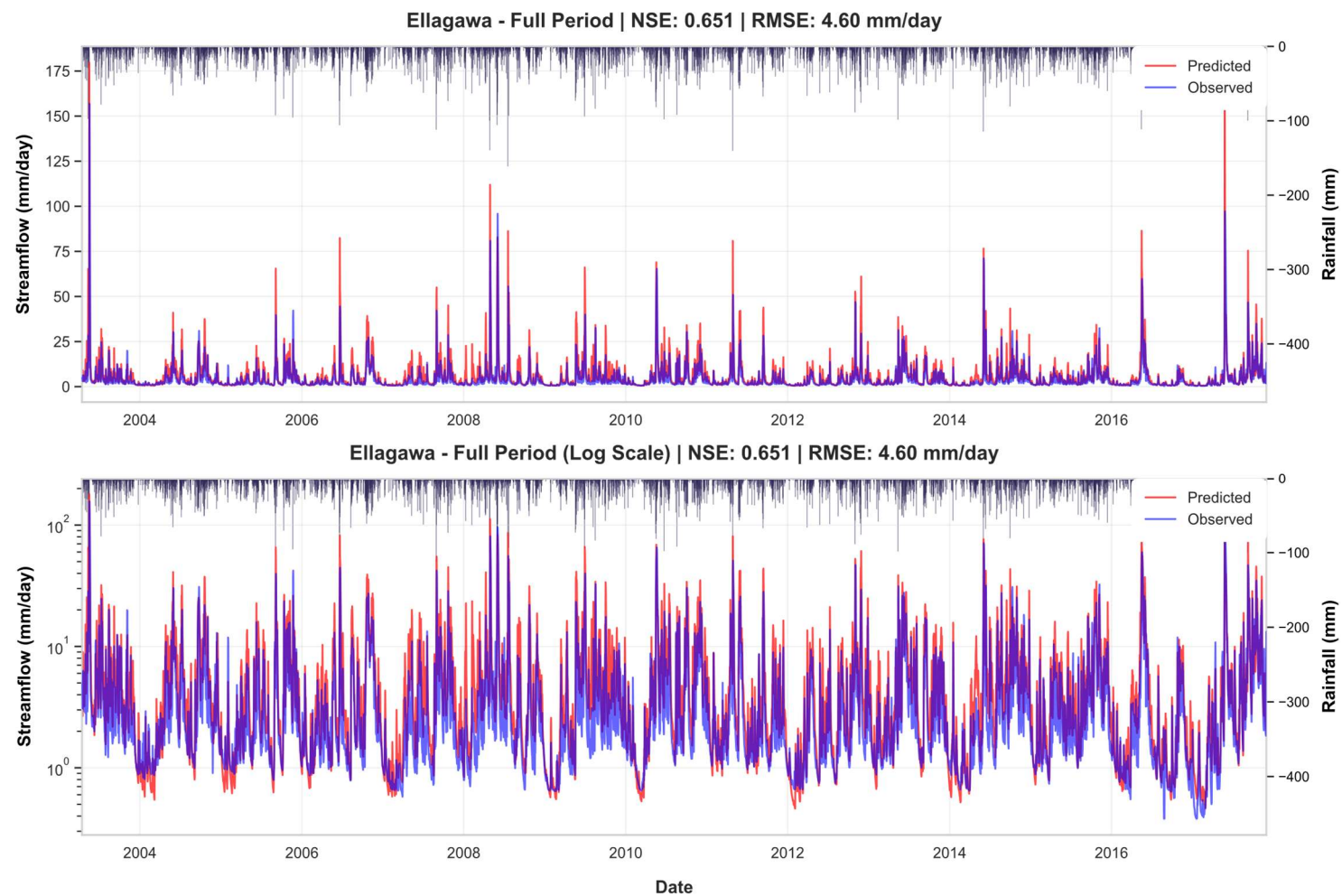
FULL-PERIOD TIME SERIES (PREDICTION DATA: GLOBAL LSTM): WET ZONE CASE STUDY

Figure: Complete prediction data period streamflow hydrograph for Ellagawa (2004–2017)

ANNEXURE 6

FULL-PERIOD TIME SERIES (DONOR DATA: HEC-HMS): DRY ZONE CASE

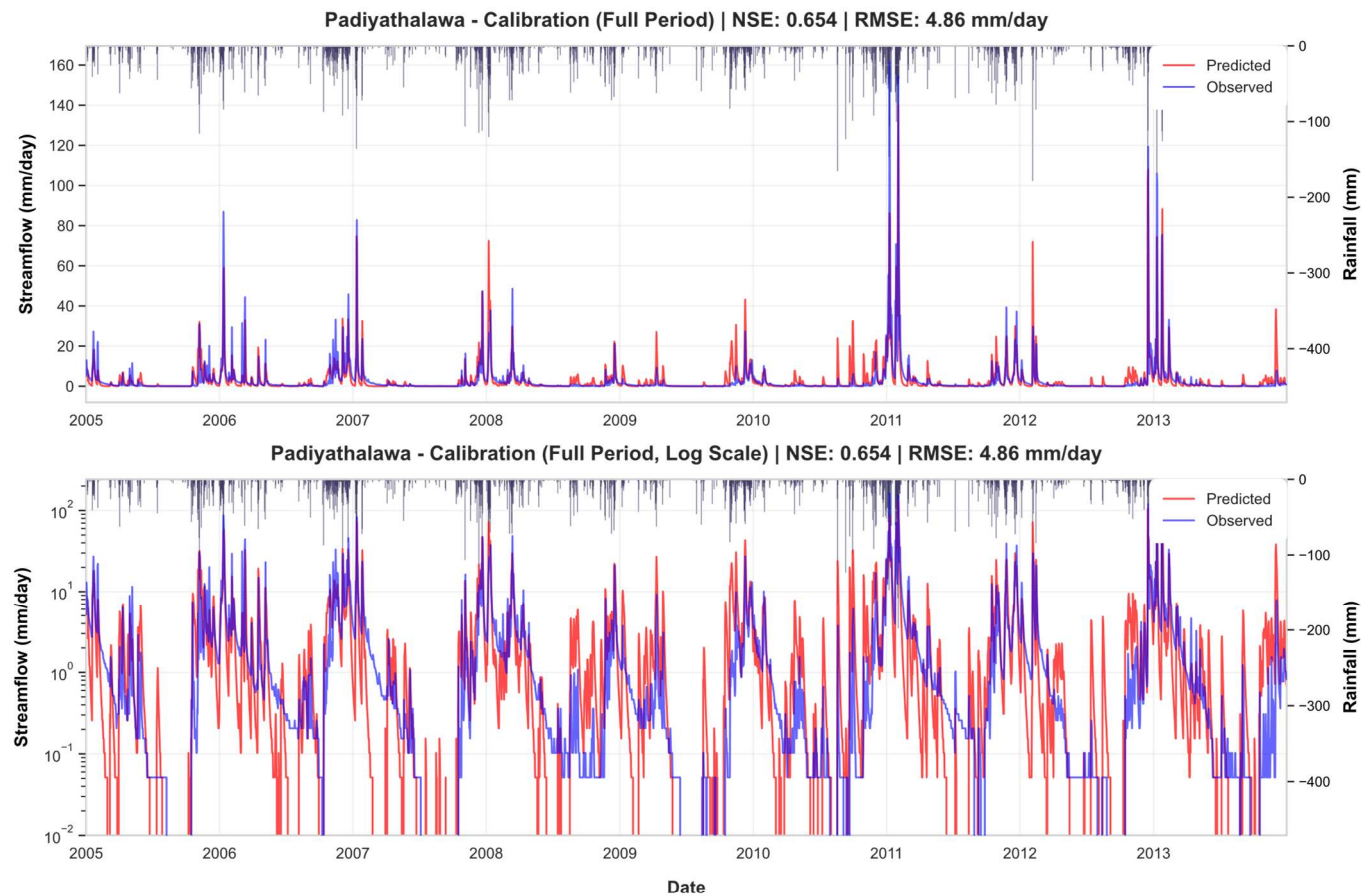


Figure: Calibration hydrograph for Padiyathalawa catchment (2005–2015)

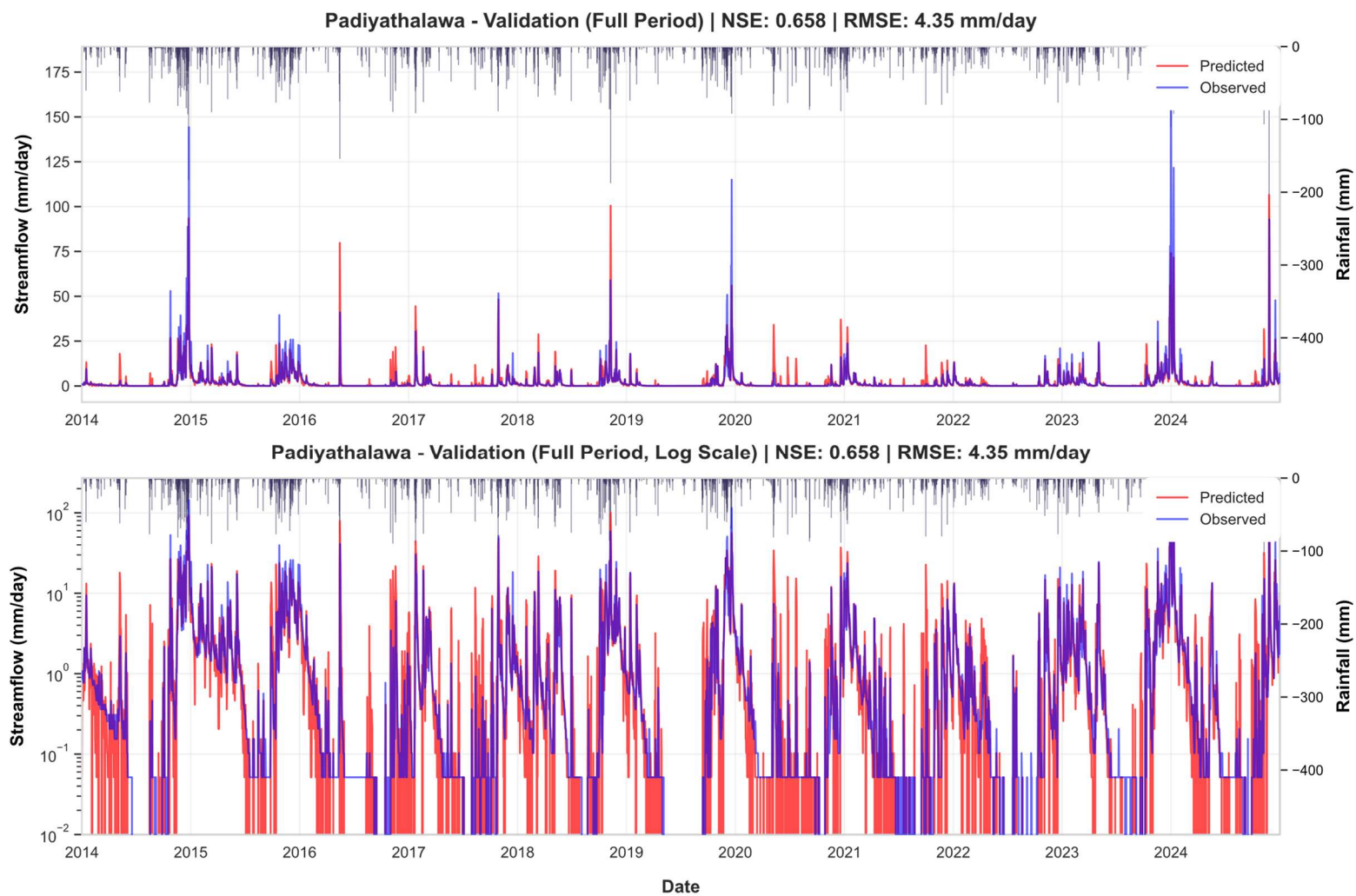


Figure: Calibration hydrograph for Padiyathalawa catchment (2015–2025)

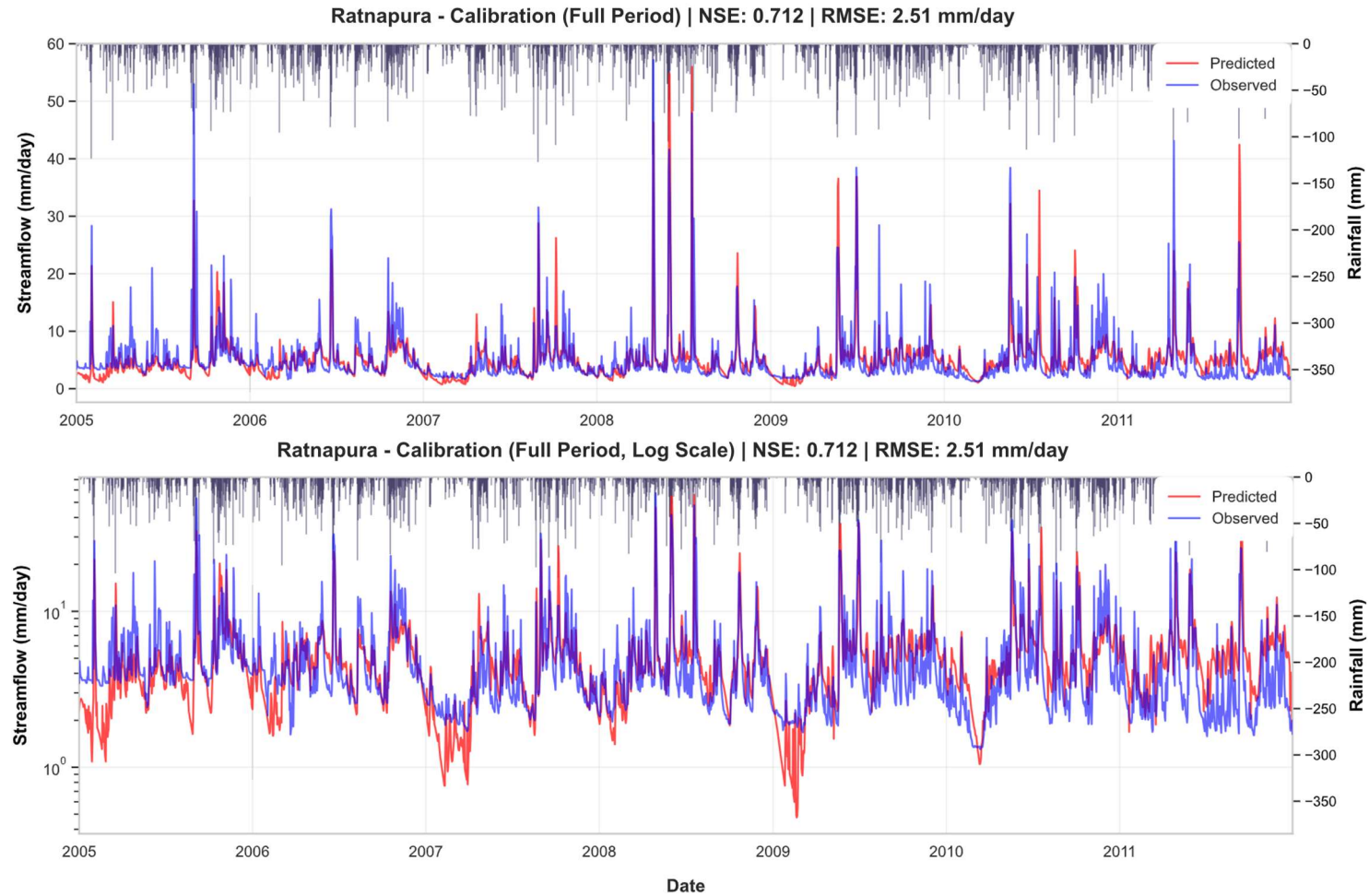
FULL-PERIOD TIME SERIES (DONOR DATA: HEC-HMS): WET ZONE CASE STUDY

Figure: Calibration hydrograph for Ratnapura catchment (2005–2012)

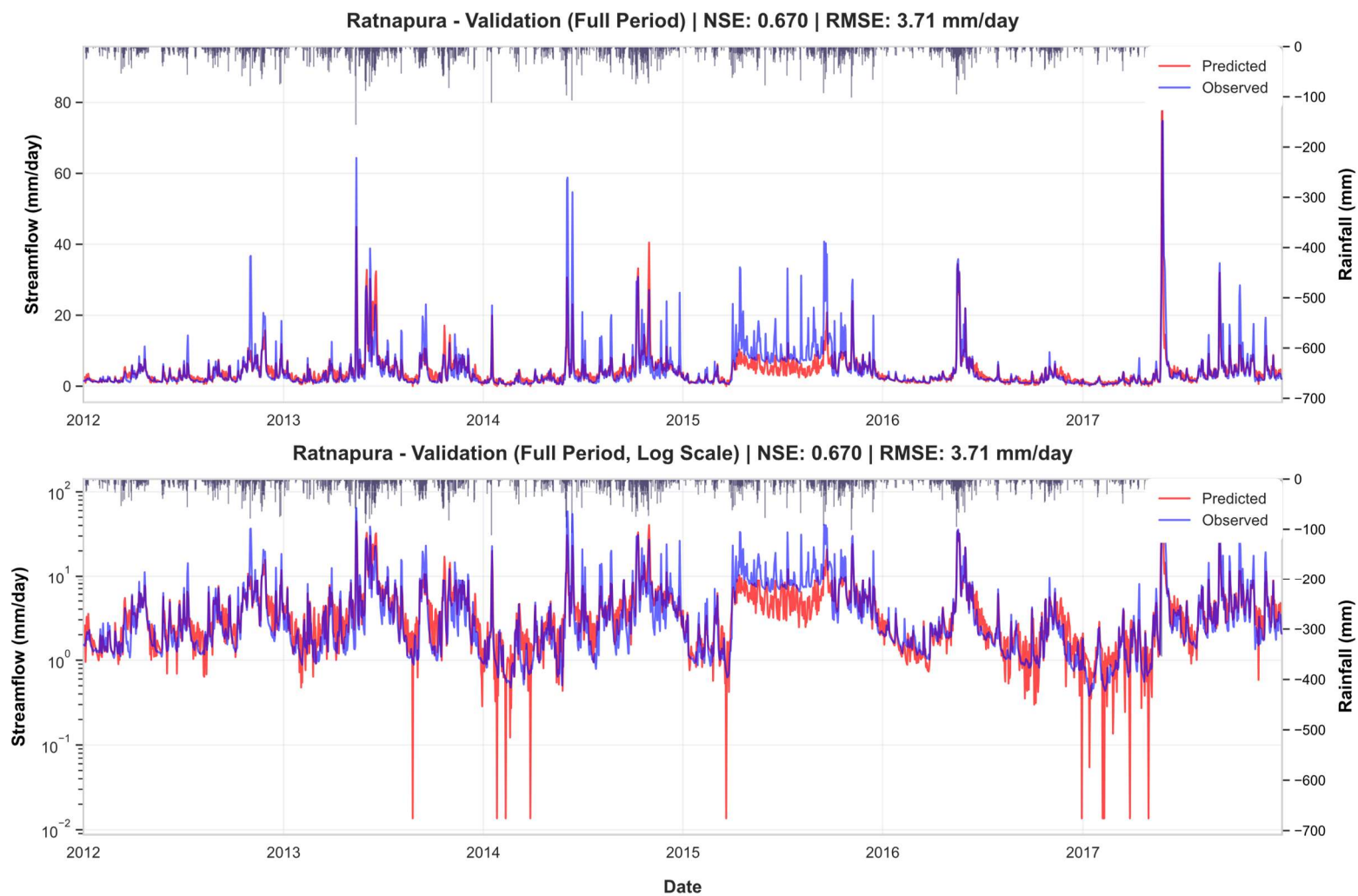


Figure: Validation hydrograph for Ratnapura catchment (2012–2018)

FULL-PERIOD TIME SERIES (PREDICTION DATA: HEC-HMS): DRY ZONE CASE STUDY

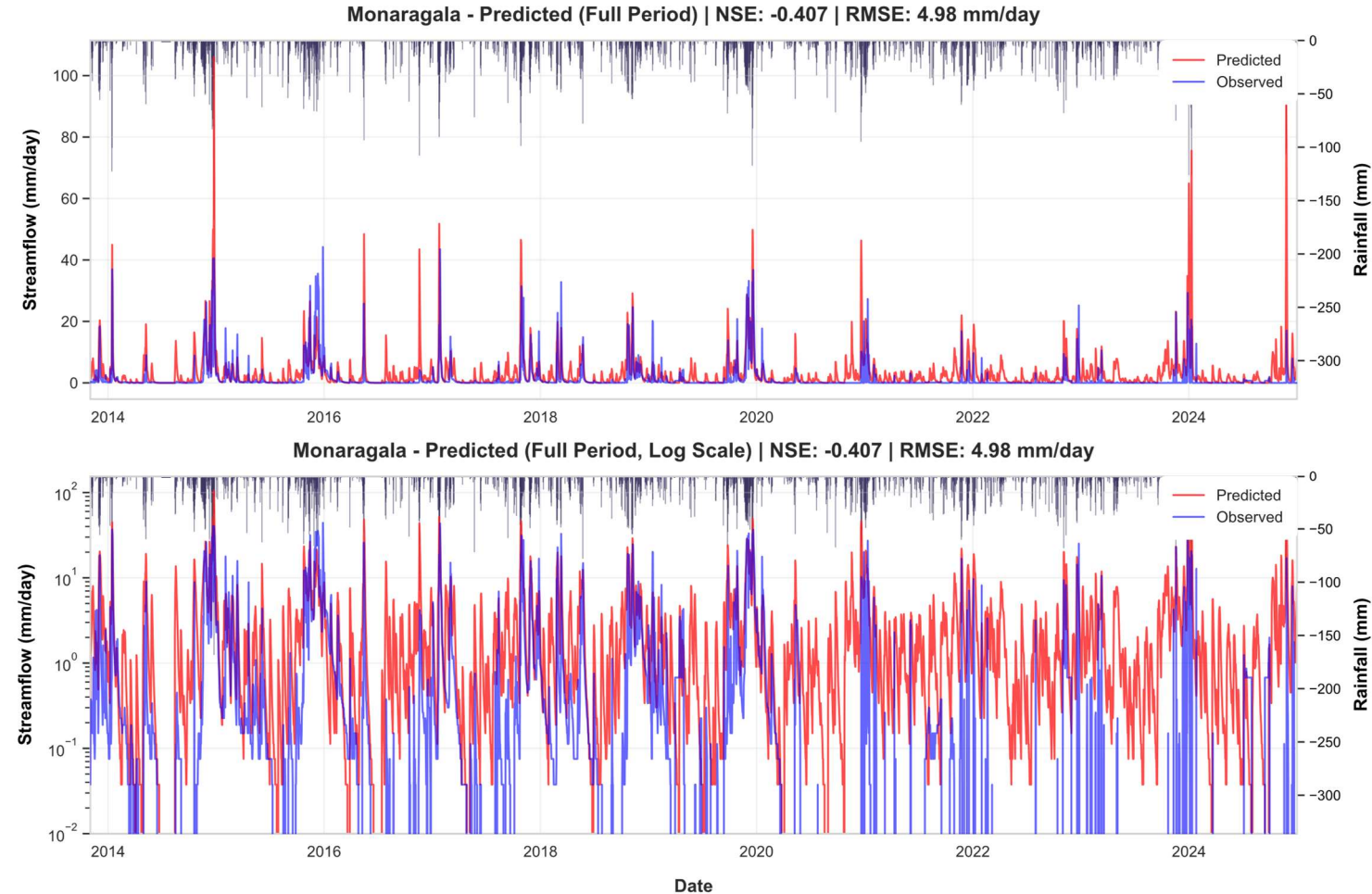


Figure: Complete prediction data period streamflow hydrograph for Monaragala (2014–2025)

FULL-PERIOD TIME SERIES (PREDICTION DATA: HEC-HMS): WET ZONE CASE STUDY

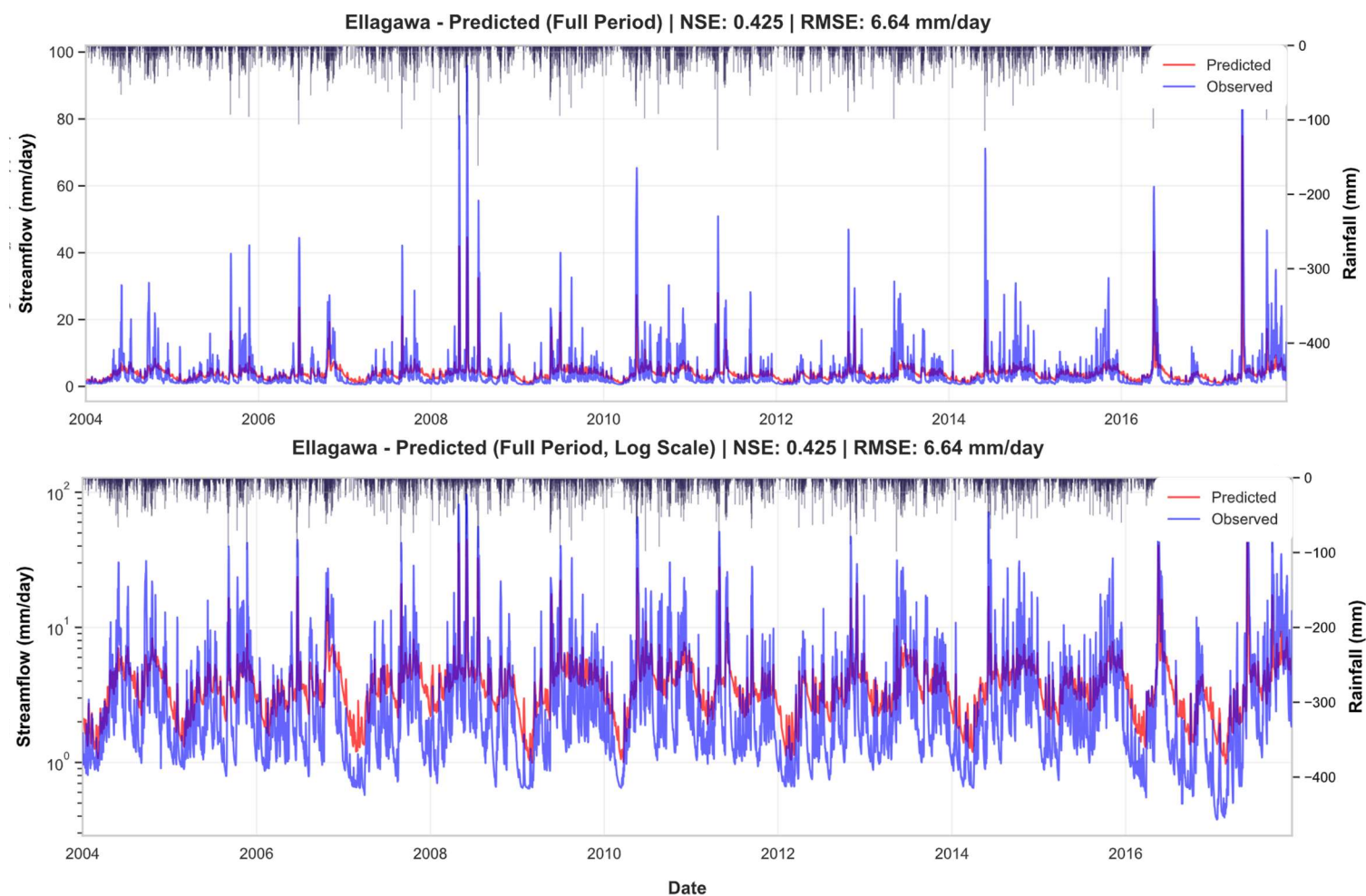


Figure: Complete prediction data period streamflow hydrograph for Ellagawa (1988–2017)

ANNEXURE 7

CALIBRATED HEC-HMS PARAMETERS

Table: Calibrated HEC-HMS parameters

Parameter	Dry Zone: Padiyathalawa Parameter Values	Wet Zone: Ratnapura Parameter Values
Canopy: Simple Canopy		
Initial Storage (%)	12.9	10.02
Crop Coefficient	1	1.1
Max Storage (mm)	7	3.96
Surface: Simple Surface		
Initial Storage (%)	20	20
Max Storage (mm)	10	7
Loss: Soil Moisture Accounting		
Soil (%)	38	52
Groundwater 1 (%)	52	20
Groundwater 2 (%)	29	50
Max Infiltration (mm/hr)	15	7
Impervious (%)	10	6
Soil Storage (mm)	300	25
Tension Storage (mm)	230	20
Soil Percolation (mm/hr)	5	5.6

Parameter	Dry Zone:	Wet Zone:
	Padiyathalawa Parameter Values	Ratnapura Parameter Values
GW 1 Storage (mm)	150	30
GW 1 Percolation (mm/hr)	1.009	2.5
GW 1 Coefficient(hr)	450	30
GW 2 Storage (mm)	400	50
GW 2 Percolation (mm/hr)	8	1.57
GW 2 Coefficient(hr)	550	150
Transform: SCS Unit Hydrograph		
Lag Time (min)	2180	150
Baseflow: Linear Reservoir		
Layers	2	2
GW 1 Flow Type	Baseflow	Baseflow
GW 1 Initial (m ³ /s)	17	10
GW 1 Fraction	0.6	0.5
GW 1 Coefficient (hr)	450	48
GW 1 Reservoirs	1	1
GW 2 Flow Type	Baseflow	Baseflow
GW 2 Initial (m ³ /s)	0	0
GW 2 Fraction	0.4	0.5
GW 2 Coefficient (hr)	500	150
GW 2 Reservoirs	1	1

The findings, interpretations and conclusions expressed in this thesis/dissertation are entirely based on the results of the individual research study and should not be attributed in any manner to or do neither necessarily reflect the views of UNESCO Madanjeet Singh Centre for South Asia Water Management (UMCSAWM), nor of the individual members of the MSc panel, nor of their respective organizations.