UNIVERSIDAD DE CANTABRIA

ESCUELA DE DOCTORADO DE LA UNIVERSIDAD DE CANTABRIA

DOCTORADO EN INGENIERÍA DE COSTAS, HIDROBIOLOGÍA Y GESTIÓN DE SISTEMAS ACUÁTICOS (IH2O)

Ph.D. THESIS

THEORY-GUIDED DEEP LEARNING FOR IMPROVING PREDICTION AND UNDERSTANDING OF FLASH FLOODS

TESIS DOCTORAL

APRENDIZAJE PROFUNDO GUIADO POR LA TEORÍA PARA MEJORAR LA PREDICCIÓN Y LA COMPRENSIÓN DE LAS INUNDACIONES REPENTINAS

Presentada por: **FARZAD HOSSEINI HOSSEIN ABADI**

Dirigida por: **Dra. CRISTINA PRIETO SIERRA**

Prof./Dr. CÉSAR ALVAREZ DÍAZ

Santander, January 2025

To my wife, ***Masi****,*

whose unwavering support and love made this journey possible.

"I think the brain is essentially a computer and consciousness is like a computer program. It will cease to run when the computer is turned off. Theoretically, it could be re-created on a neural network, but that would be very difficult, as it would require all one's memories."

Stephen Hawking in Time magazine (15 Nov 2010) responded to Elliot Giberson's question "What do you believe happens to our consciousness after death?"

Acknowledgments / Agradecimientos

First and foremost, I want to express my heartfelt gratitude to my wife, Masi, for her unwavering patience and support throughout these challenging and stressful years of my Ph.D. journey. She encouraged me to pursue my dreams and return to academia and research. Masi stood by me every step of the way, from Iran to Spain, during the difficult times of the COVID-19 pandemic and our migration to this new country, where neither of us spoke Spanish at the beginning. I feel incredibly fortunate to have such a supportive family, and I often reflect on how giving birth and raising a child under these circumstances was far more challenging to her than me completing this thesis.

Then, I would like to thank all the individuals who contributed to the completion of this research. I am especially thankful to my supervisors, Prof. César Álvarez Díaz and Dr. Cristina Prieto Sierra, for their invaluable guidance and support throughout my Ph.D. journey. César's expertise, patience, and understanding greatly alleviated the challenges I faced during my research. He consistently went above and beyond to provide assistance and played a pivotal role in managing the administrative aspects of my thesis. I recall one particularly memorable meeting when, as I was deeply engrossed in the AI components of my work, César reminded me not to overlook the hydrological outcomes of my research. This insightful observation illuminated a new perspective for me and ultimately inspired the creation of Chapter 7, where I interpret the results of the deep neural networks within a hydrological context.

On the other hand, Cristina's innovative vision for this research in 2019 and her ongoing guidance were pivotal in shaping this work. Our numerous meetings over the years equipped me with essential tools for my research, including academic writing skills and strategies for effectively addressing reviewers' comments on my publications. I will always be thankful for her support and the valuable connections she facilitated throughout this process with key persons in the field. The contributions of my supervisors had a profound impact on the final versions of my publications and this manuscript.

I extend my gratitude to Dr. Grey Nearing from Google Research for his invaluable guidance on deep learning models and for helping refine the direction of my research. He dedicated several hours to online meetings, teaching me how to work with the NeuralHydrology library for the LSTM model and addressing my challenges with kindness and expertise. Our constructive discussions and his insightful feedback played a significant role in guiding my findings. If he did not ask me to train near 400 optimized LSTMs, I may never come to the idea of ensembles and the finding that works accurate in all locations. I also thank all team members of the NeuralHydrology Python library for their collaboration and open-source contributions.

Furthermore, this work was supported by the Instituto de Hidráulica Ambiental de la Universidad de Cantabria (IHCantabria), which funded my Ph.D. and provided essential computational resources. Specifically, I am thankful to David Del Prado Secadas and all IT team members who helped me overcome extensive computational costs of this deep learning research. Moreover, I am so thankful to Sheila Abad Herrero for her kindness in assisting me with GIS that was a key step in preparing catchments attributes for this machine learning research. I should, also, express my thanks to all IHCantabria members that gathered the Basque Country catchments attributes in different projects. Without these previous works, I would never be able to proceed with data mining ideas of my research.

I wish to acknowledge the Basque Country Water Agency (URA) for granting access to their valuable dataset, which was crucial for my study. Their VALERIA project provided financial support throughout my Ph.D., and the well-documented dataset served as a strong foundation for applying deep learning models.

Additionally, I deeply value the friendships I developed at IHCantabria during my Ph.D. journey. One of the most meaningful relationships I formed was with Carlos Gutiérrez Abascal, who offered me a listening ear from my very first week in Spain. We shared countless moments discussing the ups and downs of my Ph.D. experience, and he was always there to help me navigate the challenges I faced. When I expressed my needs, he made all his efforts in providing logistical support, ensuring access to powerful GPU hardware resources for deep learning training at IHCantabria. Furthermore, he opened his home to my family, and we shared many wonderful moments together. I will forever be grateful for his kindness, which profoundly impacted both my personal and professional life.

I also want to express my appreciation for the wonderful coffee team on the first floor, whose camaraderie transformed my workspace and even positively impacted my family life. Sheila, Bárbara, Cris, Bea, Ana, Xabi, Camino and the others made my research time enjoyable. The coffee breaks were a significant motivation for me to come to the office and continue my work.

Finally, I am specifically grateful to Dr. Bárbara Ondiviela Eizaguirre for her friendly, unofficial support during the last two years of my Ph.D. Her guidance helped me sum up my messy thoughts during moments of confusion under floods of research findings and new ideas on the desk, ultimately leading to the writing of a paper that I initially hesitated to pursue. She told me stop, do not think more and write it down since it matters. This paper, later published in the Journal of Hydrology, became one of the essential components of my research to be proud of. Furthermore, I want to thank Bárbara and Dr. Cristina Galván Arbeiza for reviewing my publication drafts with friendly yet expert feedback, providing constructive insights that significantly improved the final manuscripts. I will always remember their kindness, especially considering their demanding workloads.

Table of contents

Summary 12
Summary: Brief English version 14
Resumen: Versión extendida por capítulos en español 18
Chapter I: Introduction to intersection of Hydrology and Artificial Intelligence 25
1.1. Hydrology and Perceptual Hydrological Models 26
1.1.1. Rainfall-Runoff Modeling 26
1.1.2. Hydrological Modeling as a System 27
1.2. Challenges in Hydrology and Conceptual Hydrological Models 28
1.2.1. Uncertainties and Complexities 28
1.2.2. Hydrological Variability and Model Scaling Challenges 29
1.2.3. Parametrization (Calibration/Validation) 30
1.3. Artificial Intelligence in Hydrology: Development of Intelligent Agents 30
1.3.1. How AIs/MLs/DLs Work? 31
1.3.2. Deep Neural Networks (DNNs) 32
1.3.2.1. Fundamental Concepts of ANN/DNNs 32
1.3.3. Fundamentals of the Learning Process in Deep Learning 33
1.3.3.1. Gradient Descent and Backpropagation Algorithms 34
1.3.3.2. Hyperparameters in DL/DNNs (the Configuration settings) 35
1.3.4. Advanced DNN Architectures in Hydrology Domain 36
1.3.5. Long-Short-Term-Memory Networks (LSTMs) 37
1.3.5.1. Multi-Timescale Prediction (MTS-LSTM) 39
1.3.5.2. NeuralHydrology Library: LSTM-based Hydrological Modeling 41
1.3.6. Transformers 41
1.3.7. Convolutional Neural Networks (CNNs) 42
1.4. Hydroinformatics in Water Science and rainfall-runoff modelling 43
1.4.1. Advanced DL/DNNs in hydrology and rainfall-runoff modeling 44
1.4.3. General Challenges in employing DNNs in Hydrology 51
Chapter II: Research Gaps and Objectives 55
2.1. Unaddressed questions in rainfall-runoff modeling by DLs 56
2.2. Hyperparameter optimization 57
2.3. Uniqueness of the Place Paradigm in Regional hydrological DLs 58
2.4. Some more key challenges in the domain 59
2.5. Research Goal and Objectives of this thesis 60
Chapter III: Case Study and Dataset, General Model setups, Evaluation approaches 63
3.1. Case Study: Basque Country Hydrological System 64
3.2. Dataset: Hydro-Meteorological Time Series 67
3.3. Data Splitting for Model Training and Evaluation 67
3.4. General Model Architecture and Setups 68
3.4.1. Inputs and Targets 69
3.5. Post-Random Search Validation DATASET 69
3.6. Performance Evaluation Methods 70
3.6.1. Evaluating Accuracy 70
3.6.2. Benchmarking 71
3.6.3. Evaluating Computational Costs 71

3.6.4. Statistical Analyses to Study Significant Differences 71
Chapter IV: Hyperparameter Optimization of Regional LSTMs by Random Search 75
4.1. Introduction 76
4.2. Method 78
4.2.1. Definitions and Designing the Hyperparameter Search Space 78
4.2.1.1. (MTS)LSTMs' Hyperparameters and their Definitions 78
4.2.1.2. The employed hyperparameters values in the literature 82
4.2.1.3. Selection of Hyperparameters to be tuned 83
4.2.2. Random search 84
4.2.3. Post-random search 86
4.2.4. Performance evaluation 87
4.3. Results 87
4.3.1. Overall test accuracy of optimized networks 87
4.3.2. Evaluating Catchment-Scale Performance of Optimized Networks 89
4.3.3. Significant Disparities in simulations of the optimized networks 91
4.3.4. Relation between number of random searches and accuracy 93
4.4. Discussion 94
4.4.1. Efficiency and efficacy of random search in optimization of LSTMs 94
4.4.2. Performance metrics values interpretation 94
4.4.3. Complexity of post-random search configuration selection 95
4.4.4. Learning maturity of different optimized regional configurations 95
Chapter V: Ensemble Learning of Optimized Regional LSTMs 99
5.1. Introduction 101
5.2. Method 103
5.2.1. Hyperparameter optimization and ensemble development technique 103
5.2.2. Re-training, test, and ensembles predictions 107
5.2.3. Evaluation and benchmarking the results 107
5.3. Results 108
5.3.1. Benchmarking ensemble learning versus RO and ERO networks 108
5.3.2. Significant disparities in simulations of catchment-wise method 117
5.3.3. Hydrographs confirm outperformance of ensemble learning 117
5.4. Discussion 121
5.4.1. Regional hydrological artificial intelligent agents 121
5.4.2. Catchment-scale performance evaluation of regional models 122
5.4.3. Catchment-wise hyperparameter optimization 125
5.4.4. pros and cons of Top 10 Configs and K-means Configs ensembles 125
5.4.5. Disparities in the learning skills of different configuration ensembles .. 126
5.4.6. Significance of ensemble deep learning in real-life practice 127
5.4.7. Limitations and challenges in training ensemble configurations 128
Chapter VI: Significance of Different Hyperparameters during Optimization 131
6.1. Introduction 132
6.2. Method 133
6.2.1. Generating the Post-Random Search Validation DATASET 134
6.2.2. Random Forest Model for Hyperparameter Impact Analysis 135
6.2.3. Principal Component Analysis (PCA) for Dimensionality Reduction 136
6.3. Results 137
6.3.1. Random Forest models Results 137

6.3.1.1. Catchment-wise Random Forest 138
6.3.1.2. Attribute-wise Random Forest 140
6.3.1.3. Regional Random Forest Trained on Regional val_DATASET 143
6.3.2. Principal Component Analysis 145
6.4. Discussion 148
6.4.1. Interpretation of the outcomes 149
6.4.2. Role of catchments attributes 151
6.4.3. Possible hydrological meaning of input sequence length 151
Chapter VII: Performance analysis of optimized rainfall-runoff modeling LSTMs 155
7.1. Introduction 156
7.2. Method 158
7.2.1. Test_DATASET Setup and Compilation 159
7.2.2. Exploration of Catchment Attribute-Performance Relationships 161
7.3. Results 163
7.3.1. Pearson Correlation Analysis 163
7.3.2. Random Forest Analysis 167
7.3.3. Principal Component Analysis (PCA) 171
7.4. Discussion 175
7.4.1. Intersection of physical hydrology and AI/DL models 175
Conclusions: Overall Findings and Future Works 179
Key Findings 180
Future Research Directions 182
References 184
Appendixes 202
Appendix 01: General Hydrological Definitions in hydrological modeling 204
Appendix 01.01. Hydrological Definitions 204
Appendix 01.02. Fundamental hydrological perspectives 208
Appendix 02: Loss Functions in Hydrology for Performance Evaluation 211
Appendix 02.01. Nash-Sutcliffe Efficiency (NSE) 211
Appendix 02.02. Kling-Gupta Efficiency (KGE) 211
Appendix 02.03. Mean Squared Error (MSE) 213
Appendix 02.04. Root Mean Squared Error (RMSE) 213
Appendix 02.05. Alpha-Nash-Sutcliffe Efficiency (Alpha-NSE) 214
Appendix 02.06. Beta-Nash-Sutcliffe Efficiency (Beta-NSE) 214
Appendix 02.07. Beta-Kling-Gupta Efficiency (Beta-KGE) 215
Appendix 02.08. Pearson's Correlation Coefficient (Pearson-r) 215
Appendix 02.09. High-segment volume (%BiasFHV) 216
Appendix 02.10. Low-segment volume (%Bias FLV) 217
Appendix 02.11. Mid-segment slope (%Bias FMS) 217
Appendix 02.12. Mean difference in Peak Flow Timing (Peak-Timing) 218
Appendix 02.13. Mean Absolute Percentage Error for peaks (MAPE_peak) 218
Appendix 02.14. Fraction of Missed Peaks (missed_peaks) 219
Appendix 03: Codes, Data and reproducibility 221

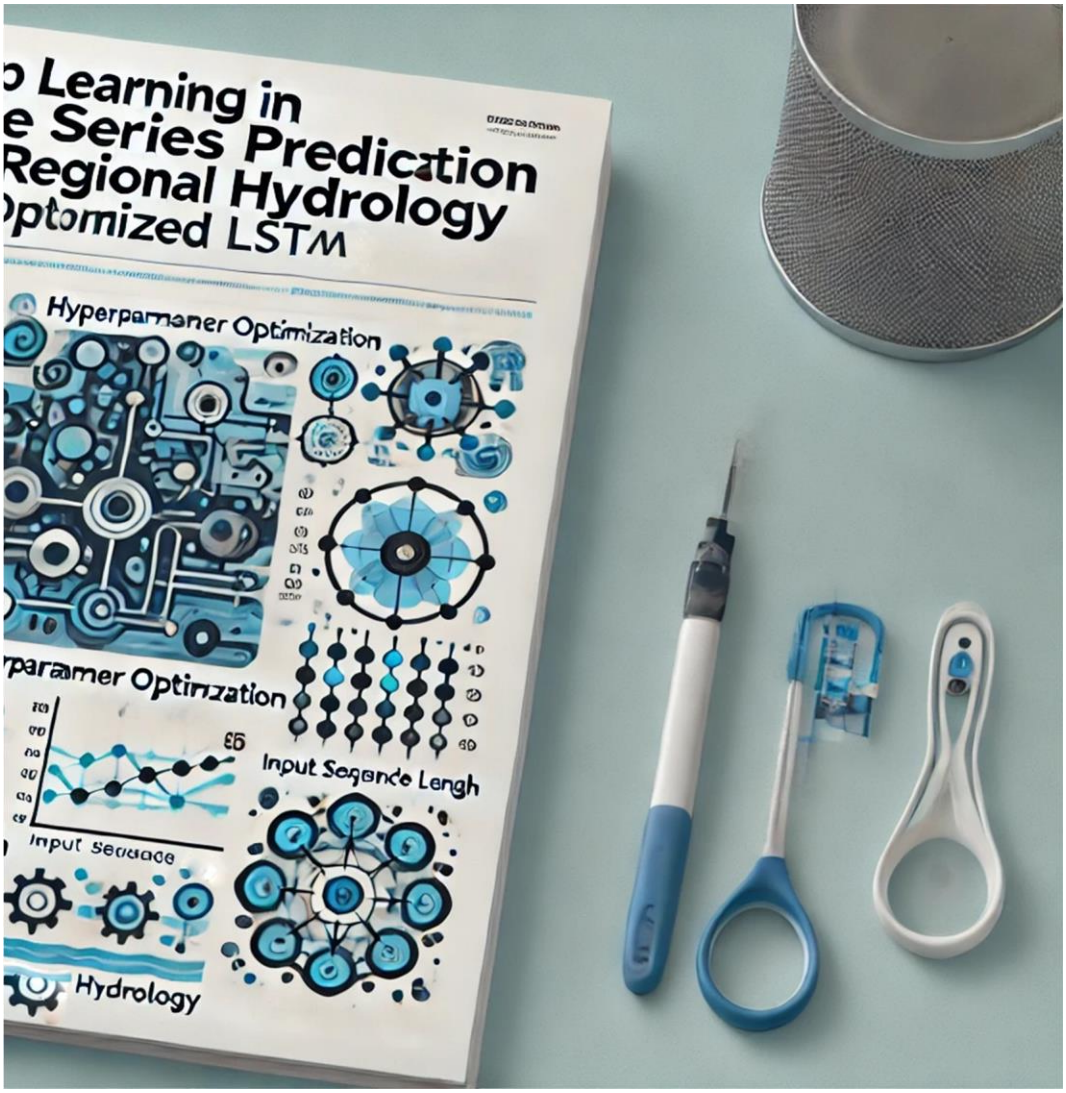
List of Tables

Table 1. A brief summary of 40 URA catchments' attributes. 66
Table 2. Ranges of the input sequence lengths 79
Table 3. Learning rates ranges and schedules for random search 82
Table 4. The defined hyperparameters space designed for random search 85
Table 5. Overall regional performance metrics of RO and ERO. 88
Table 6. Different hyperparameter configurations for the 3 ensemble methods 106
Table 7. Comparison of different configuration ensembles and the ERO network. 123
Table 8. PCA Analysis 147
Table 9. Definitions of the catchments' attributes. 160
Table 10. High Values (-0.3< or >0.3) in the Correlation heatmap 167
Table 11. PCA components' loads and the explained variance ratios 173
Table 12. Basic infiltration rates for various soil types 206
Table 13. The qualification metrics of KGE components 212

List of Figures

Figure 1. General Thesis Outline. 15
Figure 2. Basic Structure of a Neuron in an Artificial Neural Network (ANN) 33
Figure 3. Schematic of an LSTM cell and its difference from a traditional RNN 37
Figure 4. Schematic of an LSTM design for hydrological rainfall-runoff modeling 39
Figure 5. The architecture of the MTS-LSTM model 40
Figure 6. Study area of Basque Country in north of Spain 64
Figure 7. Overall Methodological Strategy developed for this research 68
Figure 8. Methodology designed and employed for hyperparameter optimization 86
Figure 9. Illustration of the frequency distribution of RO and ERO performances 89
Figure 10. Illustration of NSE and KGE distributions of RO and ERO 90
Figure 11. CDF of NSE and KGE metrics for the simulations 91
Figure 12. Statistical analyses comparing the performances of ERO and RO 92
Figure 13. Illustration of the frequency distribution of validation performance metrics 93
Figure 14. Methodology designed and employed for ensemble deep learning 105
Figure 15. Benchmarking performances of the 3 ensembles versus ERO network 108
Figure 16. NSE and KGE test metrics 110
Figure 17. MSE and RMSE test metrics 111
Figure 18. Alpha-NSE and Beta-KGE test metrics 112
Figure 19. Beta-NSE and FMS test metrics 113
Figure 20. FHV and FLV test metrics 114
Figure 21. Peak-timing and Missed-Peaks test metrics 115
Figure 22. Peak-MAPE and Pearson-r test metrics 116
Figure 23. Results of three statistical tests 118
Figure 24. Sample hydrographs showcasing observed streamflow versus predictions 119
Figure 25. Cumulative distribution function plots for simulations 124
Figure 26. Methodology for Assessing Hyperparameter Importance 134
Figure 27. Catchment-wise Random Forest 139
Figure 28. Gini gains show feature importance for the local Random Forest 140
Figure 29. Attribute-wise Random Forest 141
Figure 30. Gini gains show feature importance for the Random Forest model 142
Figure 31. Regional Random Forest 144
Figure 32. Gini gains show feature importance for the regional Random Forest 145
Figure 33. Scree Plots of the PCA applied on regional val_DATASET 146
Figure 34. Scree Plots of the PCA applied on local val_DATASET 146
Figure 35. Biplots of PCA applied on the regional val_DATASET 147
Figure 36. Biplots of PCA applied on the local val_DATASET 148
Figure 37. Correlation heatmap: relationships between attributes & model's metrics 164
Figure 38. Random Forest prediction accuracy 169
Figure 39. Feature importance ranking derived from the Random Forest model 170
Figure 40. PCA results applied to the test_DATASET 172
Figure 41. Biplot of the PCA analysis 172
Figure 42. A schematic of the Natural Water Cycle provided by © USGS 204
Figure 43. Areta catchment in Basque Country, Spain 205

Summary



Summary:

Brief version in English and Extended version in Spanish

Resumen:

Versión breve en inglés y versión extendida en español

Summary: Brief English version

Hydrology, the science of water movement, distribution, and quality, plays a crucial role in managing water resources, particularly in predicting streamflow and water levels—a process referred to as rainfall-runoff modeling. Accurate modeling is essential for various water-dependent sectors and for addressing challenges such as flood risk mitigation and sustainable water resources management. While traditional hydrological models have long been employed to simulate these processes, they often struggle with the inherent complexity and non-linear dynamics of natural systems, especially in humid, flashy catchments like those in the Basque Country, Spain. These limitations arise due to the intricate nature of hydrological processes, which involve numerous variables, extensive data requirements, and varying spatial and temporal scales.

In recent years, the rise of Artificial Intelligence (AI) has led to a paradigm shift in many scientific fields, including hydrology. The advent of Deep Learning (DL) models—such as Long Short-Term Memory (LSTM) networks, Convolutional Neural Networks (CNNs), and Transformers—has offered new ways to tackle some of the challenges faced by traditional models. These Deep Neural Networks (DNNs), inspired by the brain's hierarchical structure, can capture complex temporal patterns and non-linear relationships in large datasets. In particular, LSTMs, with their ability to retain information over time, have emerged as powerful tools for capturing the intricate dynamics of rainfall-runoff processes, offering unprecedented accuracy and flexibility in streamflow and water level predictions.

Despite these advances, still remain some challenges on the use of DL models in hydrology; one of them is the systematic optimization of their hyperparameters. Hyperparameters— the architecture of DL models and their settings that control the training process—play a critical role in determining model performance. However, given the computational costs and the multitude of hyperparameters involved, optimizing these settings for regional hydrological applications has been a challenge. Additionally, there is an ongoing need to understand whether DL models can provide new insights into the hydrological processes they model, moving beyond their role as black-box predictors to becoming tools for scientific discovery (i.e., from predictability to understanding).

This thesis seeks to address these gaps by focusing on the precise hyperparameter optimization of LSTM networks for regional rainfall-runoff modeling. Specifically, the research focuses on two aims:

1) systematically optimizing hyperparameters to improve hourly prediction accuracy across multiple catchments.

2) exploring whether the optimized LSTMs can enhance our understanding of hydrological processes and support decision-making in water resources management.

Figure 1 illustrates the logical structure and flow of this thesis, guiding readers through the overall process and key ideas presented in each chapter.

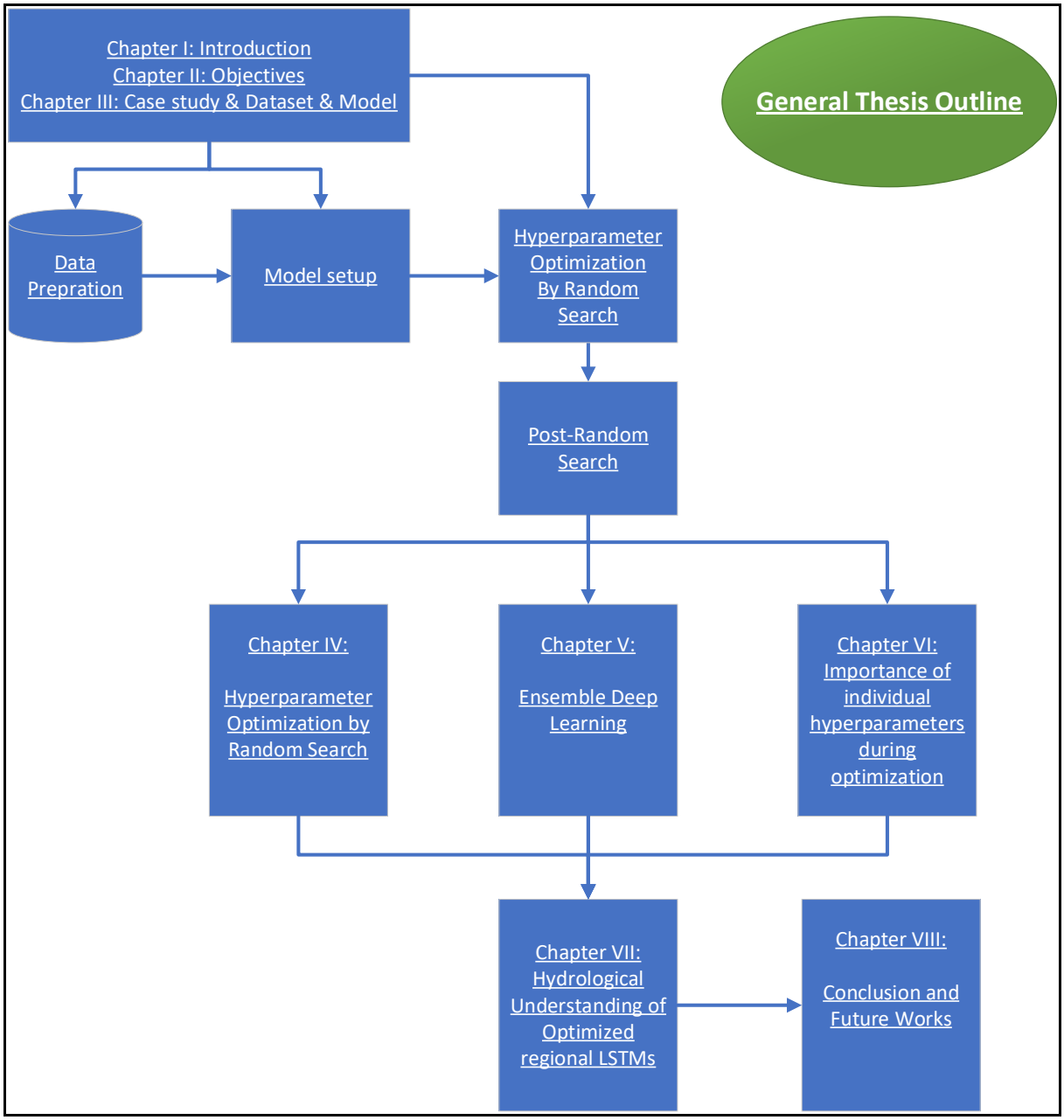


Figure 1. General Thesis Outline: The figure shows the flow of the research and the main idea behind every chapter of the whole research.

Chapter I provides a comprehensive introduction to the interdisciplinary nature of this work, reviewing key concepts in hydrology and AI/DL. The chapter outlines the motivation behind this thesis, particularly the need for more robust and systematic approaches to hyperparameter optimization in DL models for regional rainfall-runoff modeling.

Chapter II identifies the current research gaps in the domain that can be addressed by applying DL models to regional hydrology. And we will define all objectives of this research briefly.

Chapter III sets the theoretical foundation, detailing the general methodologies and materials used in this research, including data collection, initial deep learning model selection and design, and evaluation approaches. A focus is placed on the humid and flashy nature of the Basque Country's catchments, highlighting the challenges these conditions pose for hydrological modeling and the opportunities they offer for testing AI-based methods.

In **Chapter IV**, the thesis explores the hypothesis that random search, a commonly used hyperparameter optimization method, can effectively optimize LSTM networks for regional hydrological predictions. This chapter demonstrates how systematic hyperparameter optimization across multiple catchments can result in highly accurate predictions in different catchments, yielding high Nash-Sutcliffe Efficiency (NSE) and Kling-Gupta Efficiency (KGE) scores for both streamflow and water levels. The analysis also examines the trade-off between search iterations and computational cost, providing insights into the efficiency of random search for regional hydrology. At the end of this chapter, and based on lessons learned from 1000 randomly-tuned regional LSTM networks, we will find some new opportunities that can improve prediction accuracy and propose them for being examined in Chapter 5.

Chapter V investigates the potential of ensemble learning to further enhance regional model performance. It hypothesizes that combining several regionally optimized LSTM configurations, tailored to the unique characteristics of individual catchments, can yield more accurate and robust predictions compared to a single optimized configuration. The findings confirm the effectiveness of ensemble learning, with the "Catchment-wise Configs" ensemble proving especially successful in improving prediction accuracy in all locations by considering the unique hydrological behavior of each catchment.

Chapter VI investigates deeper into the importance of different tuned hyperparameters in shaping final DL model performance, using advanced machine learning techniques such as Random Forest Regression (RF) and Principal Component Analysis (PCA). This chapter hypothesizes that certain hyperparameters have a greater influence on model performance, with their impact varying depending on the hydrological characteristics of individual catchments. The findings provide valuable insights into the significance of particular hyperparameters, offering guidance for more targeted optimization efforts in future studies.

Chapter VII addresses the question of whether regional LSTMs, trained solely on hydro-meteorological data without direct access to catchment attributes, can implicitly learn latent features of catchments that influence their performance. The results suggest that while these models can capture certain catchment-specific characteristics through their input-output data, the inclusion of more explicit catchment information could further improve their predictive accuracy. This chapter underscores the potential for DL models to reveal new scientific insights, contributing to our understanding of hydrological processes in different regions.

Chapter VIII synthesizes the findings from the preceding chapters, drawing conclusions about the effectiveness of LSTM networks and hyperparameter optimization for regional hydrology. The research confirms that systematic hyperparameter optimization and catchment-wise ensemble learning are crucial for maximizing the predictive capabilities of DL models in regional hydrological applications. Additionally, the study highlights the importance of balancing computational cost with model accuracy and robustness, providing practical recommendations for researchers and practitioners in the field.

This thesis offers a robust framework for optimizing DL models in regional rainfall-runoff modeling, with significant implications for real-world applications such as flood risk management and sustainable water resources planning. The results not only improve our ability to predict streamflow and water levels in complex, humid catchments but also open up new avenues for scientific exploration in hydrology through the use of AI. Future research could focus on integrating more diverse datasets, refining ensemble learning techniques, and exploring the interpretability of DL models to further enhance their application in water resources management.

In summary, this work aims to bridge the gap between AI-driven regional hydrological DL modeling and practical applications, offering a systematic approach to hyperparameter optimization and highlighting the potential for LSTMs to contribute to both scientific discovery and effective environmental management.

Resumen: Versión extendida por capítulos en español

Introducción

La hidrología, como ciencia que estudia el movimiento, distribución y calidad del agua, desempeña un papel crucial en la gestión sostenible de los recursos hídricos. Dentro de este campo, la modelización de la relación entre las precipitaciones y el caudal de los ríos (conocido como modelado de lluvia a escorrentía) es esencial para predecir los flujos de agua en ríos y cuencas. Estas predicciones son clave para mitigar los riesgos de inundaciones, planificar infraestructuras y gestionar los recursos hídricos en diferentes entornos. Este aspecto adquiere especial relevancia en cuencas con características húmedas y de respuesta rápida, como las del País Vasco, España.

Históricamente, los modelos hidrológicos basados en procesos físicos han sido la principal herramienta para simular estos fenómenos. Estos modelos, como el SAC-SMA (Sacramento Soil Moisture Accounting Model), VIC (Variable Infiltration Capacity), SWAT (Soil and Water Assessment Tool) y HEC-HMS (Hydrologic Engineering Center-Hydrologic Modeling System), se basan en ecuaciones diferenciales que describen el movimiento del agua a través de las cuencas. Sin embargo, presentan importantes limitaciones cuando se trata de capturar la complejidad no lineal y la variabilidad espacial y temporal inherente a los sistemas hidrológicos.

Para superar estas limitaciones, la inteligencia artificial (IA), y más específicamente los modelos de Deep Learning (DL), han irrumpido en el campo de la hidrología, ofreciendo un enfoque innovador para la modelización de sistemas complejos. A diferencia de los modelos tradicionales, los DL no requieren una comprensión explícita de los procesos físicos subyacentes, sino que aprenden directamente de los datos. Esto resulta particularmente ventajoso en la hidrología, donde las relaciones entre las variables climáticas y las respuestas hidrológicas son altamente no lineales y difíciles de modelar con precisión mediante enfoques físicos convencionales.

Entre las técnicas de Deep Learning, las redes neuronales profundas (Deep Neural Networks o DNNs) y, en particular, las redes de memoria a corto y largo plazo (Long Short-Term Memory networks o LSTMs) han demostrado ser herramientas poderosas para la modelización de sistemas hidrológicos. Estas redes son particularmente efectivas en la captura de dinámicas complejas que los métodos tradicionales no logran representar de manera adecuada.

Las DNN han sido aplicadas con éxito a una amplia variedad de problemas hidrológicos, como la predicción de caudales, la estimación de la recarga de acuíferos y la gestión de riesgos de inundaciones. Diferentes estudios han mostrado que estas redes pueden predecir con precisión los flujos de los ríos en cuencas complejas, mejorando significativamente la capacidad predictiva en comparación con los modelos hidrológicos convencionales. Sin embargo, también presentan desafíos, como la dificultad para entrenar redes profundas debido al problema de generalización y la alta sensibilidad a los hiperparámetros.

Para abordar estos problemas, las redes LSTM han emergido como una variante especializada dentro de las DNN, ideal para modelar fenómenos temporales y secuenciales, como las respuestas hidrológicas. Las LSTM fueron diseñadas específicamente para superar los inconvenientes de las redes neuronales recurrentes, permitiendo procesar secuencias largas de datos y capturar relaciones temporales complejas. Esto las convierte en una herramienta particularmente valiosa para predecir caudales y niveles de agua en sistemas hidrológicos donde las dinámicas de los flujos pueden ser influenciadas por fenómenos de larga duración, como sequías prolongadas o la acumulación de nieve.

Un área de investigación que ha crecido rápidamente es la aplicación de las LSTM para la predicción hidrológica a nivel regional. En este tipo de estudios, las LSTM no solo se entrenan para predecir el caudal en un único punto de salida, sino que también proporcionan predicciones para múltiples cuencas dentro de una región geográfica. Esto permite evaluar la capacidad de las LSTM para generalizar su aprendizaje a cuencas con características hidrológicas y geomorfológicas diversas, lo que es esencial para abordar la variabilidad espacial en la hidrología regional.

Los hiperparámetros en redes neuronales profundas son parámetros clave que determinan el comportamiento y el rendimiento del modelo durante el proceso de entrenamiento. A diferencia de los parámetros internos del modelo, como los pesos, que se ajustan automáticamente a través del aprendizaje, los hiperparámetros deben ser definidos antes del entrenamiento y tienen un impacto significativo en la precisión y eficiencia del modelo. Entre los más importantes se encuentran la tasa de aprendizaje, que controla la rapidez con la que el modelo adapta sus pesos, el tamaño de la red (número de capas y neuronas), y las técnicas de regularización, que ayudan a evitar el sobreajuste. La elección de estos hiperparámetros es fundamental para alcanzar un equilibrio entre la capacidad de generalización del modelo y su exactitud en datos específicos, lo que se convierte en un desafío complejo y esencial para mejorar la calidad de las predicciones en modelos hidrológicos.

La optimización de los hiperparámetros de las redes LSTM, especialmente para aplicaciones regionales, es una de las tareas más críticas y aún no completamente resueltas en este campo. El uso de LSTM en hidrología regional puede mejorarse mediante técnicas de optimización de hiperparámetros, como la búsqueda aleatoria, que permiten maximizar la exactitud de las predicciones. Este fue uno de los principales objetivos de esta tesis.

No obstante, a pesar de los resultados prometedores, el uso de LSTM en hidrología todavía se enfrenta a ciertos desafíos. Por un lado, la necesidad de grandes cantidades de datos hidrometeorológicos de buena calidad para entrenar eficazmente los modelos limita su aplicabilidad en regiones con escasez de datos. Además, las LSTMs son consideradas modelos de "caja negra", lo que significa que es difícil interpretar cómo toman decisiones o qué características específicas emplean para realizar sus predicciones. Esta falta de interpretabilidad ha llevado a investigaciones que buscan desarrollar enfoques híbridos, imponiendo a las LSTMs cumplir con procesos físicos para mejorar tanto el entendimiento como la exactitud de las predicciones.

Esta tesis doctoral se sitúa en la intersección entre la hidrología y la inteligencia artificial, con el objetivo principal de optimizar el uso de redes LSTM en la predicción de caudales y niveles de agua en la hidrología regional. Además de explorar las capacidades predictivas de estas redes, se abordarán los desafíos de interpretabilidad y aplicabilidad en diferentes contextos hidrológicos.

Casos de estudio

El País Vasco, presenta cuencas húmedas y un régimen torrencial (inundaciones repentinas – flash floods), ofrece un caso de estudio ideal para probar la capacidad de las LSTMs y evaluar cómo la optimización sistemática de los hiperparámetros puede mejorar las predicciones hidrológicas. A través de esta tesis, no solo se busca aumentar la exactitud de los modelos, sino también proporcionar un marco metodológico robusto que pueda aplicarse en otras regiones.

El área de estudio seleccionada para esta investigación comprende 40 cuencas ubicadas en el País Vasco, en el norte de España. Estas cuencas se caracterizan por su clima lluvioso y respuestas rápidas a eventos de precipitación intensa, lo que las convierte en un entorno ideal para evaluar los modelos de predicción basados en IA. Se utilizó información hidrometeorológica de alta resolución horaria para entrenar y optimizar las redes LSTM, con el objetivo de mejorar la precisión de las predicciones del caudal y los niveles de agua a nivel regional en diferentes localizaciones.

Estas cuencas también tienen una relevancia práctica, ya que las mejoras en las predicciones hidrológicas tienen un impacto directo en la gestión del riesgo de inundaciones y la asignación de recursos hídricos en una región con importantes actividades económicas y ecológicas dependientes del agua.

Estado del arte, Método y Materiales

Como se ilustra en la Figura 1, El **Capítulo I** proporciona una introducción completa a la naturaleza interdisciplinaria de este trabajo, revisando conceptos clave en hidrología, inteligencia artificial (IA) y Deep Learning (DL). Expone la motivación detrás de esta tesis, específicamente la necesidad de enfoques más sistemáticos para optimizar los hiperparámetros en modelos de DL aplicados al modelado regional de lluvia-caudal.

El **Capítulo II** identifica las brechas en la investigación actual sobre la aplicación de modelos de DL en hidrología regional, estableciendo los objetivos clave de esta investigación.

El **Capítulo III** presenta la base teórica y describe las metodologías y materiales empleados, que incluyen la recolección de datos, la selección y el diseño inicial del modelo de aprendizaje profundo, así como los métodos de evaluación. Este capítulo destaca la naturaleza húmeda y de respuesta rápida de las cuencas del País Vasco, subrayando los desafíos que estas características presentan para el modelado hidrológico y las oportunidades que ofrecen para evaluar métodos basados en IA.

Hallazgos clave

En el **Capítulo IV** ("Optimización de hiperparámetros en redes LSTM para la hidrología regional"), se plantea que la optimización sistemática de hiperparámetros mediante búsqueda aleatoria puede conducir a predicciones hidrológicas más exactas. Aquí se introduce un enfoque innovador, que combina el diseño del espacio de hiperparámetros con búsqueda aleatoria, resultando en una red LSTM regionalmente optimizada para toda la comunidad autónoma del País Vasco. Los altos niveles de exactitud alcanzados en la predicción de caudales y niveles de agua (NSE y KGE de hasta 0,97 en algunas cuencas) validan la eficacia de este método. Un hallazgo clave fue que diferentes configuraciones de hiperparámetros optimizadas producen predicciones significativamente diferentes en distintas cuencas, destacando la capacidad de cada red LSTM para adaptarse a características específicas de cada cuenca (ver Figuras 9-11).

El **Capítulo V** ("Aprendizaje en conjunto de redes LSTM regionalmente optimizadas") explora técnicas de aprendizaje en conjunto para mejorar la precisión y la robustez de predicciones mediante LSTM optimizadas. Se desarrollaron tres estrategias de conjuntos de modelos, y el Catchment-wise Configs ensemble, que ajusta hiperparámetros a características específicas de cada cuenca, fue la más eficaz, superando tanto a las estrategias individuales como a otros enfoques de conjunto (ver Figuras 15-25).

El **Capítulo VI** ("Importancia de los hiperparámetros durante la optimización") examina el impacto de diferentes hiperparámetros en el rendimiento de las redes LSTM en la modelación regional. A través de análisis de componentes principales (PCA) y regresión de Random Forest, se identificó que algunos hiperparámetros, como la longitud de la secuencia de entrada, tienen una influencia estadísticamente significativa en la precisión, especialmente en cuencas de respuesta rápida (ver Figuras 27-36).

El **Capítulo VII** ("Comprensión hidrológica de las LSTMs optimizadas") analiza si las LSTM optimizadas regionalmente pueden captar características latentes específicas de cuenca utilizando solo datos hidrometeorológicos. Los resultados sugieren que estas redes pueden aprender patrones implícitos propios de cada cuenca, esenciales para predicciones precisas (ver Figuras 37-41).

Conclusiones

Esta tesis presenta un estudio exhaustivo sobre la optimización de redes de memoria a largo y corto plazo (LSTM) para la modelización de la relación lluvia-caudal a nivel regional en múltiples cuencas del País Vasco, España. A través de una investigación sistemática, exploramos la hipótesis de que la optimización de hiperparámetros mediante la búsqueda aleatoria, combinada con el aprendizaje en conjunto de redes LSTM optimizadas y teniendo en cuenta la singularidad de cada cuenca, podría mejorar significativamente la precisión predictiva de los modelos de aprendizaje profundo para predicciones de caudal y niveles de agua. Los hallazgos no solo examinan las hipótesis, sino que también abren un camino para investigaciones futuras, ofreciendo valiosos conocimientos para aplicaciones hidrológicas prácticas, especialmente en regiones propensas a inundaciones repentinas, como el País Vasco.

Capítulo IV se planteó la hipótesis de que la optimización sistemática de hiperparámetros mediante búsqueda aleatoria podría lograr una alta precisión en la modelización de la relación lluvia-caudal en todas las cuencas. Los resultados confirmaron que un espacio de búsqueda de hiperparámetros bien diseñado, combinado con la búsqueda aleatoria, conducía a un rendimiento LSTM altamente exacto en diversas ubicaciones. Con modelos entrenados en 40 cuencas, logramos altos puntajes de Eficiencia de Nash-Sutcliffe (NSE) y Eficiencia de Kling-Gupta (KGE), lo cual indica una que el modelo es robusto en términos de exactitud. Incluso con solo 100 iteraciones de búsqueda aleatoria, el modelo Regional Óptimo (RO) mostró una alta exactitud, que fue aún más refinada en el modelo Óptimo Regional Mejorado (ERO) después de 1000 iteraciones. Además, la alta exactitud del RO valida tanto la eficiencia como la eficacia de la búsqueda aleatoria para la optimización de hiperparámetros en redes LSTM regionales.

Un aspecto clave identificado en este capítulo es la importancia de ajustar simultáneamente múltiples hiperparámetros para lograr predicciones confiables. Aunque el incremento en el número de búsquedas mejoró los resultados, se observó que existe un equilibrio entre el costo computacional y la exactitud. La evidencia sugiere que, aunque modelos como el RO no exhiban el mismo grado de madurez de aprendizaje que el ERO, aún pueden capturar efectivamente anomalías hidrológicas y matices que redes más maduras podrían pasar por alto. Este balance entre eficiencia computacional y madurez del modelo es un factor crítico a considerar en futuras aplicaciones de redes LSTM en la hidrología regional.

Además, los diferentes rendimientos estadísticamente significativos de las redes RO y ERO optimizadas respaldan la afirmación de que dos redes LSTM regionales optimizadas con configuraciones de hiperparámetros distintas “piensan” y funcionan de manera diferente, incluso si se someten al mismo enfoque de entrenamiento y utilizan los mismos datos de entrenamiento.

Capítulo V introdujo el aprendizaje en conjunto como un medio para mejorar aún más la exactitud y la robustez de los modelos de predicción basados en LSTM a nivel regional. La hipótesis de que un conjunto de configuraciones optimizadas regionalmente podría superar una configuración única se confirmó mediante pruebas rigurosas. Además, el conjunto de configuraciones específicas por cuenca (Catchment-wise Configs), que ajustaba los hiperparámetros a las características únicas de cada cuenca, superó tanto a la mejor configuración única como a otras estrategias de conjunto.

Este hallazgo subraya la importancia de considerar las características específicas de cada cuenca al seleccionar los hiperparámetros óptimos para los modelos hidrológicos regionales. La importancia de la longitud de la secuencia de entrada, un hiperparámetro destacado en el enfoque Catchment-wise Configs, refuerza aún más la hipótesis de que los modelos hidrológicos se benefician de adaptar los datos de entrada a la dinámica temporal única del flujo de agua en cada cuenca. Este enfoque adaptado resultó ser particularmente efectivo en las cuencas húmedas y torrenciales del País Vasco, donde los procesos de lluvia-escorrentía rápidos exigen modelos que puedan capturar la retención de agua a corto plazo y los tiempos de viaje.

Capítulo VI amplió la investigación sobre el papel de los hiperparámetros individuales en la configuración del rendimiento del modelo. Mediante técnicas avanzadas de aprendizaje automático como la regresión empleando bosques aleatorios (RF) y el análisis de componentes principales (PCA), exploramos la hipótesis de que ciertos hiperparámetros ejercen una mayor influencia que otros, y que su importancia varía según las características hidrológicas de cada cuenca.

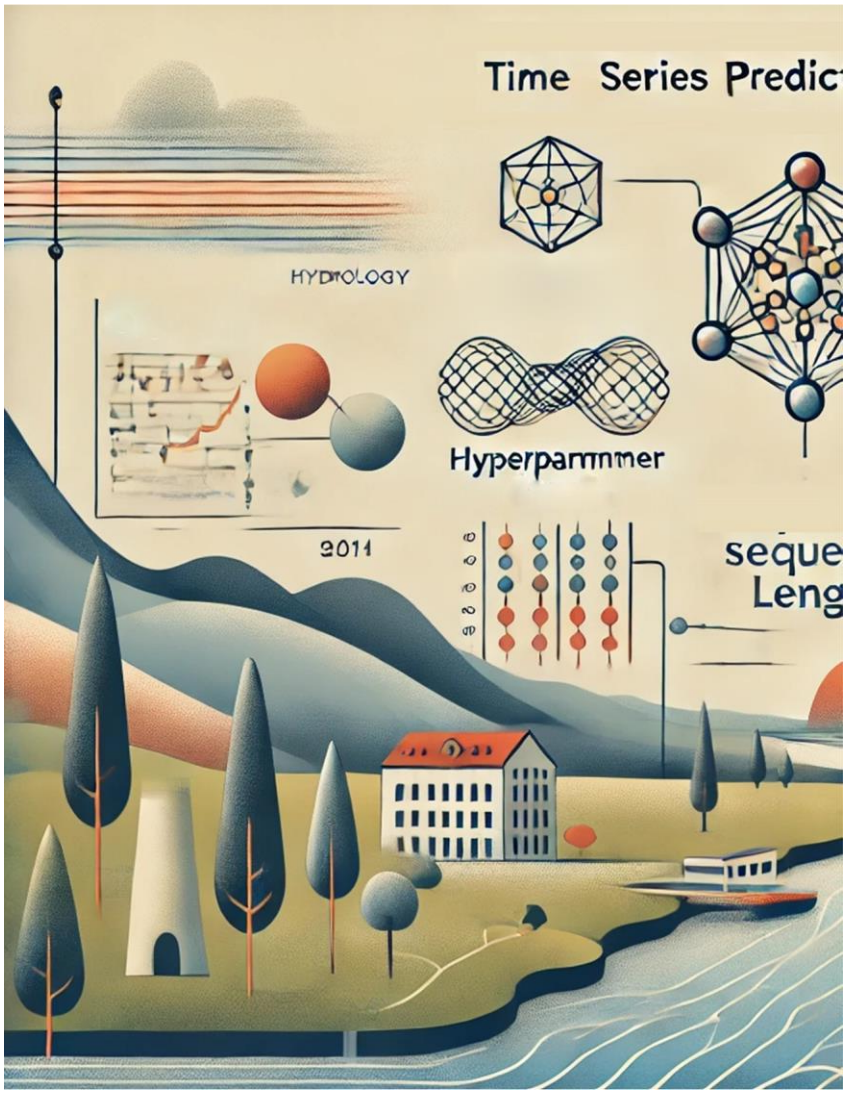
Los resultados confirmaron que la importancia de los hiperparámetros no es la misma en todas las cuencas. Por ejemplo, la longitud de la secuencia de entrada resultó ser particularmente influyente en las cuencas con respuestas hidrológicas rápidas, como las del País Vasco. Estos hallazgos destacan la necesidad de estrategias de optimización de hiperparámetros que tomen en cuenta la diversidad de condiciones hidrológicas dentro de una región. La variabilidad en la importancia de los hiperparámetros entre diferentes cuencas también refuerza la idea de que un enfoque único no es suficiente para maximizar la precisión de las predicciones.

Capítulo VII se planteó la hipótesis de que las redes LSTM, incluso sin acceso directo a atributos específicos de cada cuenca, podrían aprender características hidrológicas latentes a partir de los datos. Los hallazgos confirmaron que el rendimiento de las LSTM optimizadas regionalmente variaba según las características de la cuenca, a pesar de que los modelos se entrenaban exclusivamente con datos hidrometeorológicos.

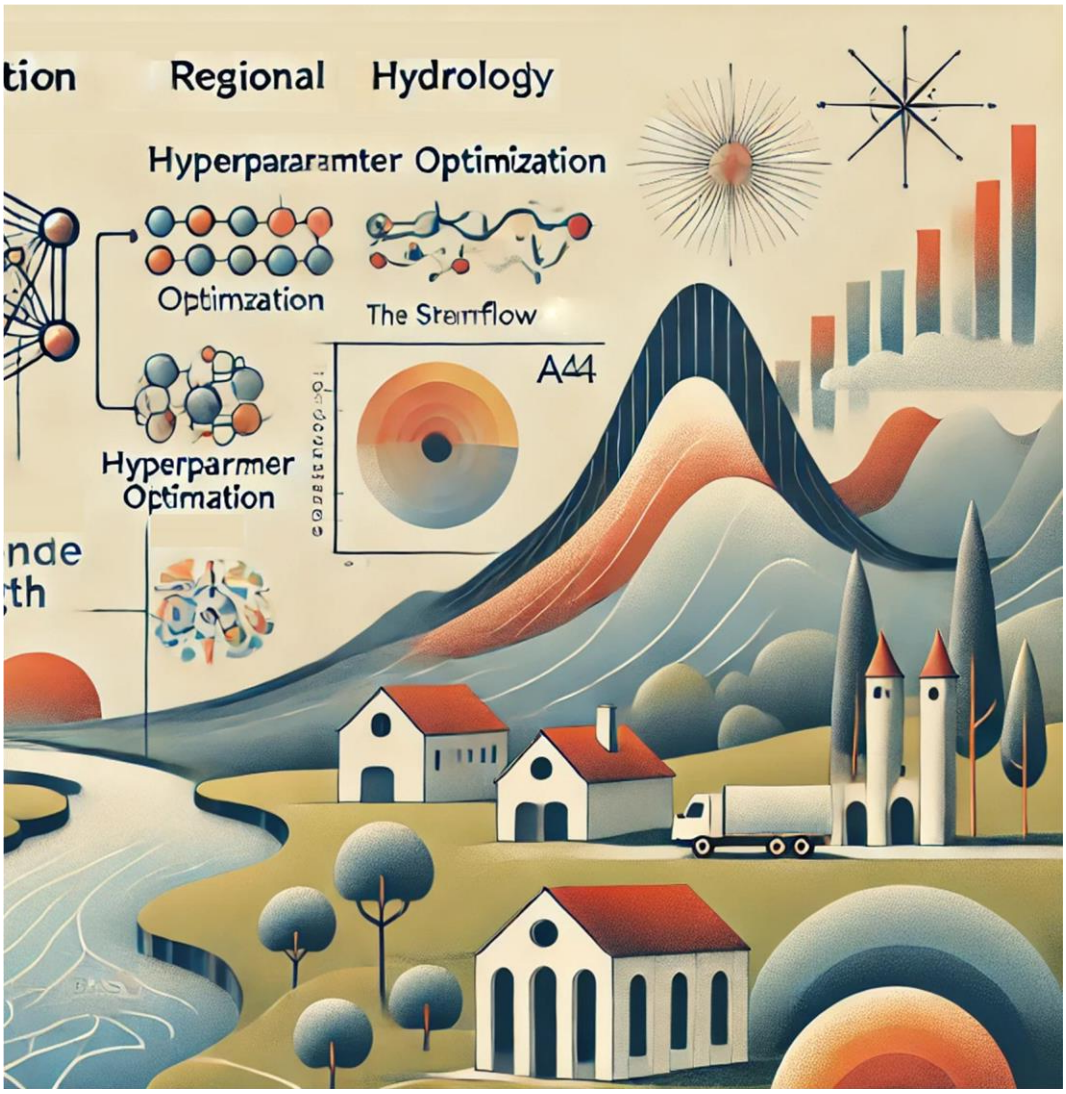
Esta capacidad implícita de aprendizaje es una característica poderosa de los modelos de DL, especialmente en regiones donde los datos detallados de la cuenca pueden ser limitados o inconsistentes. Al predecir tanto el caudal como el nivel de agua simultáneamente, las LSTM pudieron capturar dinámicas únicas de la cuenca, como las reflejadas en las curvas de gasto. Esta habilidad para aprender de los procesos hidrológicos sin entradas explícitas sugiere que las redes LSTM pueden generalizarse bien en cuencas diversas, haciéndolas altamente valiosas para la modelización hidrológica regional.

Los conocimientos obtenidos de esta investigación tienen importantes implicaciones para aplicaciones prácticas en hidrología, particularmente en regiones como el País Vasco, donde las inundaciones repentinas y los rápidos procesos de lluvia-escorrentía representan un riesgo constante. Las predicciones de caudales y niveles de agua lo más exactas y robustas posibles como las proporcionadas por los modelos LSTM regionales optimizados, pueden mejorar la evaluación del riesgo de inundación, los sistemas de alerta temprana y las estrategias de gestión de recursos hídricos.

En conjunto, esta tesis ha demostrado el potencial de las redes LSTM optimizadas para predicciones hidrológicas regionales, particularmente en entornos complejos y propensos a inundaciones rápidas usando como caso de estudio principalmente las cuencas del País Vasco, España. A través de la optimización sistemática de los hiperparámetros, el aprendizaje en conjunto y el aprendizaje implícito de atributos específicos de cada cuenca, hemos desarrollado modelos que logran una alta exactitud en la predicción de caudales y niveles de agua en múltiples cuencas.


Time Series Predic
HYDROLOGY
Hyperparameter
2011
seque
Leng

Chapter I



Introduction to intersection of

Hydrology and Artificial Intelligence

A Comprehensive Literature Review

1.1. Hydrology and Perceptual Hydrological Models

The history of Earth Systems Science is full of several developed physico-mathematical models reproducing and simplifying the Earth's Natural Processes. These models tried to simplify our knowledge and understanding and define the complex processes in the form of some definable concepts for a variety of purposes (Refsgaard et al., 2022).

Hydrology, as one of the known Earth systems sciences, "is the science that encompasses the study of water on and beneath the Earth's surface, the occurrence and movement of water, the physical and chemical properties of water, and its relationship with the living and material components of the environment (Bales, 2015)" either from a natural or an anthropogenic perspective. Hydrology has never extinguished the perceptual modeling approaches to resemble the complex hydrological processes on and under the ground utilizing understandable systems for humans having known input(s) and output(s) with the aid of some scientifically developed hypotheses (Refsgaard et al., 2022; Beven, 2012; Chow et al., 1988). Appendix 01 offers a comprehensive overview of essential hydrological definitions and key concepts to bridge the interdisciplinary gap between hydrologists and AI/DL scientists, ensuring a shared understanding.

1.1.1. Rainfall-Runoff Modeling

Rainfall-runoff modeling, is central to hydrological science and practice. The models are designed to estimate how much runoff a rain event will generate, compute the routing of water downstream, and predict flow at specific points over time. Understanding and predicting streamflow is crucial for managing water resources, preparing for floods and droughts, and ensuring water supply needs are met (Beven, 2012; Chow et al., 1988).

Historically, rainfall-runoff models are categorized based on their structure and approach. Lumped models utilize aggregated catchment inputs and simulate streamflow at a single point, typically the catchment outlet. These models are simpler and computationally less demanding but overlook spatial variability. Distributed models, on the other hand, divide the catchment into smaller grid cells, capturing spatial variability by simulating streamflow at multiple points in every catchment. While these models provide higher resolution, they are often more complex and prone to greater uncertainty (Refsgaard et al., 2022; Beven, 2012; Chow et al., 1988).

The effectiveness of these models varies with catchment characteristics, data availability, and calibration processes. Some models perform exceptionally well in specific catchments but struggle in others due to regional differences or data limitations; the so-called "uniqueness of the place" paradigm (Beven, 2020; 2000). The complexity of distributed models does not always guarantee better performance, and simpler lumped models can sometimes achieve comparable results (Refsgaard et al., 2022; Perrin et al., 2001; Reed et al., 2004). This resulted

in a general believe in traditional hydrology that there is no “one size fits all” model (See Fenicia et al., 2008).

A key challenge in hydrology is regionalization of rainfall-runoff modeling—applying a model effectively across different catchments. Traditional regional hydrological models are often highly parameterized, with defined parameters that are specific to uniqueness of different catchment types, making generalization so difficult. The hypothesis that more detailed models with extensive process descriptions improve accuracy has been questioned, especially when focused on predicting river discharge (Refsgaard et al., 2022; Refsgaard & Knudsen, 1996). Although latest approaches are very promising as instead of transfer mode parameters, what is transfer and using to predict/infer the streamflow in the ungauged catchments are the hydrological indices (see, e.g., Prieto et al, 2019; 2022).

As Albert Einstein famously said, “Everything should be made as simple as possible, but no simpler” (Refsgaard et al., 2022). This principle has inspired a trend towards developing simpler, more parsimonious models that use fewer parameters while still delivering effective performance (Bergström, 1991; Beven, 1989; Jakeman & Hornberger, 1993; Perrin et al., 2003). However, this contrasts with the complexity of advanced DLs, which involve numerous key hyperparameters, a vast number of internal parameters (such as weights and biases), and require sophisticated, time-consuming training processes. The paradox lies in the fact that complex DL models can appear both simple and intricate, depending on the user’s expertise and familiarity with the model’s inner workings.

1.1.2. Hydrological Modeling as a System

Hydrological modeling, when approached as a system, views catchments and other water bodies as interconnected and interdependent components. This system-based perspective offers numerous benefits, enhancing both the interpretation and application of traditional hydrological models.

A system-based approach represents a catchment as a network of interconnected components and processes, adhering to the principles of systems theory. Components may include Inputs (e.g., precipitation, springs), Storages (e.g., reservoirs), and Outputs (e.g., river flow, water transport, consumption). Processes involve Interconnections between the components (e.g., evaporation, evapotranspiration, condensation, infiltration, runoff, underground flows, etc. (Prieto et al., 2021; Clark et al., 2008)).

This approach emphasizes the importance of understanding how each process interacts with the other processes and as well within the broader system. In hydrology, while each catchment can be considered an individual system, it often has known or unknown connections with other catchments (systems); particularly from a regional perspective. This interconnectedness makes regional hydrology more challenging, especially when considering multiple catchments as part of an integrated system and the aim is to achieve accurate regional predictions across various locations simultaneously.

1.2. Challenges in Hydrology and Conceptual Hydrological Models

Hilbert's (1900) perspective on the challenges of solving mathematical problems resonates strongly in hydrology today: "A mathematical problem should be difficult so as to pose a challenge for us, and yet not completely inaccessible so that it does not mock our effort." In hydrology, unresolved challenges persist, particularly in the realm of rainfall-runoff modeling, especially at regional scales. These challenges hinder our understanding of the complex interactions between water, climate, and human activities. Addressing these challenges requires a holistic approach that moves beyond isolated model comparisons to achieve a deeper comprehension of hydrological processes (Blöschl et al., 2019).

In traditional hydrology, accurate rainfall-runoff modeling at regional scales in gauged catchments remains a daunting task due to several persistent sticking points. These include dealing with spatial heterogeneity in catchment properties, uncertainties in data, model structures (identification and selection), and model parameters. These issues not only affect the performance of streamflow predictions but also highlight significant gaps in our current understanding of hydrological processes at both catchment and regional levels. Overcoming these obstacles is critical to move forward towards more reliable, precise and accurate streamflow predictions by hydrological models.

1.2.1. Uncertainties and Complexities

Uncertainty in hydrological modeling is an inherent challenge that must be addressed to improve the reliability of predictions (Prieto et al., 2022; 2021). These uncertainties arise from various sources, including natural variability, data inaccuracies, and model structural limitations.

Uncertainty in Natural Processes: The inherent variabilities in natural processes such as rainfall, soil moisture, and evapotranspiration complicate modeling efforts. These processes are influenced by numerous interacting factors, making the modeling chain (identification, selection and parametrization) difficult.

Data and Measurement Errors: Inaccuracies in streamflow records (or the observation data used to evaluate the performance of the predictions), meteorological data, and other measurements contribute significantly to uncertainty in hydrological predictions.

Model Structure Uncertainty: The simplifications and assumptions made in modeling real-world hydrological processes introduce structural uncertainties. These uncertainties can lead to significant prediction errors, particularly when the model is not able to capture the main processes and parameter values are incorrect or suboptimal. Of course, it is well known the adage that Box posed in 1976 "All models are wrong but some are useful".

Maybe, it is worth to mention here that there is still no method to disentangle input uncertainty from model uncertainty (structure and parameters).

Spatial and Temporal Heterogeneity: The spatial and temporal heterogeneity of catchment characteristics, such as land use, soil properties, and vegetation cover, adds complexity to modeling efforts. Capturing these variations accurately across different scales remains a significant challenge.

Future Climate and Land Use Uncertainty: The unpredictability of future climate conditions and land use changes introduces additional layers of uncertainty, making long-term predictions particularly difficult.

Addressing these uncertainties is critical for enhancing the performance (reliability, accuracy and precision) of hydrological models. Techniques such as Bayesian methods, Monte Carlo Dropout, and Mixture Density Networks have been introduced to quantify uncertainty and improve model interpretability and reliability (Donnelly et al., 2024b; Klotz et al., 2022; Prieto et al., 2022; 2021). Reducing model uncertainty is key to improve the accuracy and practical utility of hydrological models, particularly in the context of regional hydrology, where diverse catchment conditions must be considered.

1.2.2. Hydrological Variability and Model Scaling Challenges

One of the fundamental challenges in hydrological modeling is representing the spatial heterogeneity of catchment properties. Catchments vary significantly in terms of land use, soil properties, geology, vegetation cover, and climate conditions, all of which influence hydrological responses (Beven, 2000; 2020). Addressing this variability is crucial for improving the accuracy and generalizability of rainfall-runoff models.

Spatial heterogeneity affects key hydrological processes, including infiltration, runoff generation, and evapotranspiration. Traditional conceptual models often rely on lumped or semi-distributed parameterizations that struggle to capture local variations in these processes (Beven, 2012). Distributed models attempt to address this issue but require extensive data inputs and computational resources (Blöschl & Sivapalan, 1995).

A major challenge is identifying the dominant controls of spatial variability at different scales. Studies have shown that factors such as topography, soil moisture distribution, and land cover exert strong influence on hydrological responses (Western et al., 2002; Tetzlaff et al., 2009). However, these relationships are often non-linear and context-dependent, making it difficult to generalize findings across diverse catchments.

Remote sensing and geospatial analysis have provided new opportunities to quantify spatial variability at finer resolutions. However, integrating these high-resolution datasets into hydrological models remains an ongoing research challenge. Machine learning and deep learning approaches have emerged as potential solutions for handling complex spatial interactions, but their interpretability and transferability remain key issues in hydrology.

1.2.3. Parametrization (Calibration/Validation)

Parametrization, particularly the calibration and validation of hydrological models, presents another significant challenge. Reliable simulations require accurate calibration, yet this process is fraught with difficulties due to several factors:

Model Performance Dependence: The accuracy of these models is highly dependent on both the model structure and the calibration techniques used, as well as the availability of accurate data. This is particularly challenging for process-based models, which require detailed data that may not be available in all locations.

Calibration Techniques: Calibration in hydrological modeling is the process of determining model parameters, a crucial step in ensuring accurate simulations (Beven, 2012; Kavetski, 2018). Automatic calibration involves using optimization algorithms alongside an objective function designed to minimize errors. Quantitative metrics are typically employed during calibration and validation to evaluate model performance. The choice of calibration method, whether it involves single-objective functions, multi-objective approaches, or hybrid methods integrating machine learning, can significantly impact model accuracy and effectiveness (Kavetski, 2018; Zheng & Wang, 2021). For example, recent techniques like differentiable parameter learning (dPL) leverage deep learning to improve the calibration of perceptual hydrological models (Tsai et al., 2021). Each calibration method has its own advantages and limitations, and its suitability depends on the specific application of the model.

Parameter Regionalization: To overcome the limitations of conceptual models, parameter regionalization techniques have been developed. These techniques use transfer functions to link distributed catchment features to model parameters, thereby improving the representation of spatial heterogeneity. However, these methods may still fall short in accounting for all variations, leading to potential inaccuracies (Hansen et al., 2007; Andersen et al., 2001).

1.3. Artificial Intelligence in Hydrology: Towards the Development of Intelligent Agents

According to Russell and Norvig, 2021, Artificial Intelligence (AI) refers to the development of computer systems capable of performing tasks that typically require human intelligence, such as pattern recognition and decision-making. A key subset of AI is Machine Learning (ML), which enables models to learn from data and improve their performance without being explicitly programmed. Deep Learning (DL) is an advanced form of ML that relies on Deep Neural Networks (DNNs), which consist of multiple processing layers designed to extract complex latent features from large datasets (Goodfellow et al., 2016). Throughout this text, we will frequently use these terms, as they form the foundation of modern AI-driven hydrological modeling.

AI has evolved significantly since its inception in 1950s, and its application in hydrology has the potential to revolutionize how we model and predict complex water systems. At the heart of AI in this domain is the development of intelligent agents (See: Russell and Norvig, 2021)—systems designed to perceive environment data, learn from it, and autonomously act to achieve specific objectives (e.g., predicting rainfall-runoff responses in real-time.)

Understanding the core principles of AI is crucial. Machine Learning (ML), a key subset of AI, along with advanced Deep Learning (DL) models, have significantly contributed to hydrology. By enabling models to adapt to evolving data patterns and environmental changes, ML/DL help address the dynamic and unpredictable nature of hydrological conditions. While AI provides a strong foundation, it still falls short in fully capturing the complexities and uncertainties inherent in hydrological systems. The ultimate goal should not merely be to apply AI to hydrological datasets but to develop intelligent systems with hydrological insights—systems that can autonomously learn and adapt to evolving environmental conditions, thereby improving predictive accuracy over time. Achieving this requires carefully designed architectures and domain-specific optimizations.

Rationality in AI is another critical concept, particularly for hydrological applications. A rational agent aims to take the best possible action based on available data, even under uncertainty (Russell and Norvig, 2021). This is vital in hydrology, where predictive models must determine the most effective actions to take, given the current conditions. Unlike human, AI systems lack intrinsic desires or preferences— at least for now, so these must be clearly defined by hydrologists to ensure the desired predictive outcomes—such as different attitudes for minimizing error in runoff forecasts (e.g., which metrics to choose, or aggregating metrics strategies in regional hydrology, or focusing on high-flows, etc.) or detecting climate change impacts on water systems.

To develop effective intelligent agents, hydrologists must understand the nature of the task's environment—whether it's fully or partially observable, deterministic or stochastic. These characteristics influence how an agent is designed and the strategies it employs to maximize predictive accuracy. For example, an agent in rainfall-runoff modeling should map environment inputs to accurate predictions of surface runoff and streamflow, handling uncertainty and partial observability effectively.

Overall, as AI continues to advance, the challenge is to ensure that these intelligent agents can effectively model the complexities of hydrological processes, ultimately improving our understanding and management of water resources.

1.3.1. How AIs/MLs/DLs Work?

In AI and ML/DL, the ability to transform data into higher-dimensional spaces is crucial for identifying complex patterns and extracting meaningful latent features from massive information buried in large-scale datasets. This transformation is achieved through techniques such as kernel methods and activation functions.

Kernel methods (Hofmann et al., 2008) transform input data into higher-dimensional spaces, such as *Hilbert space*, where linear algorithms can more effectively discern patterns that are not apparent in the original, lower-dimensional space. This transformation allows models to separate and classify data points that are otherwise inseparable, enhancing the model's ability to extract latent features and information for making accurate outcomes in tasks like image recognition, text classification, and time-series analysis.

Similarly, *activation functions*, such as Rectified Linear Unit (ReLU) and Sigmoid, introduce non-linearity into neural networks, enabling them to capture and process intricate dependencies within the data. These functions are crucial for enabling neural networks to learn from and adapt to data that exhibit complex, multi-dimensional relationships.

In hydrology, where data is often non-linear and dynamic, these techniques are particularly valuable. By mapping data into higher-dimensional spaces and applying non-linear activation functions, AI models can more effectively extract relevant features and improve their ability to understand and predict complex hydrological phenomena.

1.3.2. Deep Neural Networks (DNNs)

Artificial Neural Networks (ANNs) are a fundamental component of AI, especially in ML domain. Modeled after the biological neural networks found in animal brains, ANNs consist of interconnected nodes, or artificial "neurons," organized in layers. These networks are designed to recognize patterns, process data, and make decisions similarly to the human brain, albeit in a more structured and simplified manner through a recursive training process involving rewards and penalties (See: Russell & Norvig, 2020).

The development of Deep Neural Networks (DNNs), which are essentially ANNs with multiple hidden layers, has significantly expanded the capability of these models. DNNs are particularly effective at capturing complex, latent features and nonlinear relationships within data, enabling them to tackle more sophisticated tasks than traditional ANNs.

1.3.2.1. Fundamental Concepts of ANN/DNNs

At the core of an ANN/DNNs is the neuron, which receives inputs, processes them through a mathematical function, and produces an/some output(s) (Figure 2). Each connection between neurons is associated with a weight and bias, which determine the influence of the input information on the neurons. During the learning process, these weights are adjusted to minimize errors in the network's predictions or classifications. This adjustment process is typically carried out using the backpropagation algorithm, which plays a critical role in reducing the discrepancy between predicted and actual outcomes, a measure often referred to as the loss function.

The history of neural networks dates back to the 1950s, when Marvin Minsky and Dean Edmonds, undergraduate students at Harvard, developed the first neural network model. The evolution of neural networks was further propelled by the development of the backpropagation algorithm from the early 1960s to the mid-1980s, which facilitated the application of these models to a wide range of learning problems across different domains. Despite the initial enthusiasm, it is important to note that AI models, including ANNs, should be guided by well-founded theories of intelligence. As Russell and Norvig (2021) highlight, the mere ability of a program to find a solution in principle does not imply that it possesses the mechanisms required to find it in practice.

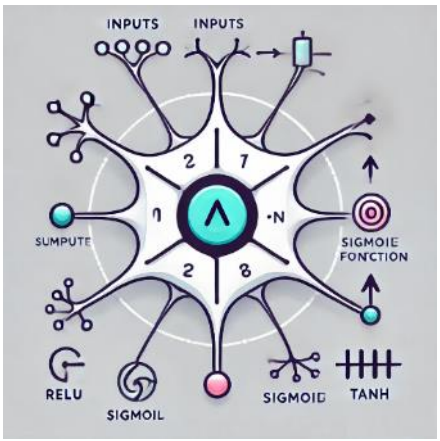


Figure 2. Basic Structure of a Neuron in an Artificial Neural Network (ANN)/Deep Neural Network (DNN). The neuron receives inputs, each associated with a weight. These inputs are summed and passed through an activation function, producing the neuron's output. This output can then be passed to neurons in subsequent layers for further processing, allowing the network to learn and make predictions (The figure was generated by ChatGPT.4).

Neurons of ANN/DNNs are typically organized into three types of layers:

Input Layer: Neurons in this layer receive the initial data, with each neuron may preprocessing a different feature of the input data passing it to next layers.

Hidden Layer(s): Neurons in these layers perform complex computations and feature transformations. The presence of multiple hidden layers allows the network to model intricate patterns and relationships within the data, leading to the emergence of Deep Neural Networks (DNNs).

Output Layer: Neurons in this layer transform the understandings and outcomes of the previous layers to produce the final output of the network, whether it be a classification, regression, decision making or any other form of outcome.

1.3.3. Fundamentals of the Learning Process in Deep Learning

Deep Learning relies on a complex iterative learning processes to build models capable of generalizing from data. This process involves the optimization of model architecture (The configurations) by hyperparameter optimization methods (Bergstra & Bengio, 2010) and numerous internal parameters (weights and biases) through the backpropagation algorithm (Rumelhart et al., 1986) to finally achieve effective learning. Understanding the fundamentals

of this process, including key algorithms and strategies for avoiding common pitfalls like under/overfitting and overparameterization, is crucial for developing robust DL models.

Tom Mitchell (1980) stated that "biases and initial knowledge are at the heart of the ability to generalize beyond observed data". The number of internal parameters (weights and biases) as well as the structural design of the AIs (hyperparameters) are considered to determine models' complexity. Generally, as the complexity of AIs rises, the training set error approaches zero (Russell and Norvig, 2021). Therefore, choosing an appropriate architecture (the configuration settings), inductive biases, and initial weights for DL/DNNs is key to generalization (Hoedt et al., 2021).

A DL model can contain thousands to millions of internal parameters, such as weights and biases. For instance, in a hydrological prediction study by Mai et al. (2022), the developed LSTM networks had approximately 300,000 parameters. However, this number should never be compared to parameters in traditional hydrological models, as DL/DNN parameters do not explicitly represent individual physical properties. Moreover, the effectiveness of a DNN is not solely determined by its size but also by its architecture and connectivity. This is akin to the complexity of a biological brain, where intelligence arises not just from the number of neurons but from their intricate connections and how they interact (the flow of information among different neurons). Thus, a more complex DNN, with numerous connections and fine-tuned hyperparameters, may capture more sophisticated patterns and relationships in data after being well-trained.

1.3.3.1. Gradient Descent and Backpropagation Algorithms

Backpropagation is a fundamental algorithm used for training artificial neural networks (ANN) and DNNs by minimizing the error between predicted and actual outcomes. Introduced by Rumelhart et al. in 1986, backpropagation operates by propagating errors backward through the network to update weights and biases. This iterative process involves calculating the gradient of the loss function with respect to each weight through the chain rule, allowing the model to learn and refine its parameters. The efficiency and effectiveness of backpropagation in training deep neural networks have made it a cornerstone technique in this domain.

The *Gradient Descent* (GD) algorithm is, also, fundamental to training DL/DNNs. It is an optimization technique used to minimize the cost function by iteratively updating the network's internal parameters—weights and biases—based on the gradient of the cost function. The core principle of Gradient Descent is to adjust the parameters in the direction that reduces the error, thereby improving the model's performance.

Several variants of gradient descent enhance the basic algorithm's efficiency. Stochastic Gradient Descent (SGD) introduces randomness by updating the parameters based on individual data points, which can accelerate convergence and help the algorithm escape local minima. Mini-Batch Gradient Descent updates the parameters using subsets of the data, balancing the efficiency of batch processing with the speed of stochastic updates.

Another important variant is the Adam optimizer, which stands for Adaptive Moment estimation. Adam improves on other methods by combining features of Adaptive Gradient algorithm (AdaGrad) with momentum and Root Mean Square Propagation (RMSProp). It adjusts the learning rates for each parameter individually, using information about the average and variance of the gradients. This makes it better at speeding up the training process and handling noisy or sparse data.

Last but not least, the concept of the *Vanishing Gradient* is a common challenge in training DL/DNNs, especially those with many layers. In simple terms, when training deep networks, the gradient—the value used to update the model's parameters during backpropagation—can become very small as it is passed back through the layers. This means that the earlier layers (closer to the input) learn very slowly or stop learning altogether because the updates to their weights are so tiny. As a result, the network may struggle to learn and improve its performance, particularly in capturing complex patterns in the data. This issue, known as the vanishing gradient problem, can hinder the effectiveness of DL/DNNs, making them harder to get trained. Researchers have developed various techniques, such as using different activation functions (like ReLU) and advanced initialization methods, to mitigate this problem and ensure that all layers in the network learn effectively.

1.3.3.2. Hyperparameters in DL/DNNs (the Configuration settings)

Hidden size, learning rate, batch size, length of the input sequence, initial forget gate bias, loss function, dropout rate, standard target noise, regularization terms, optimizer type, LSTM head, and output activation function are some of the used hyperparameters we need to configure in DNNs. They play a pivotal role in training DL/DNN models, impacting their learning and generalization capabilities (Russell & Norvig, 2020; Goodfellow et al., 2016; Shalev-Shwartz & Ben-David, 2014). These settings, denoted by the prefix "hyper-," oversee both the training dynamics and the ultimate structure of the neural network. Therefore, careful attention to network architecture and the hyperparameter configuration is essential for optimizing a DNN's ability to generalize from training data to new, unseen data.

In developing DNNs, hyperparameters determine not only the model's size but also its architectural layout, analogous to how the human brain's various regions specialize in different functions and evolve from birth through childhood. Hyperparameters govern the connections between layers and their neurons, influencing the flow and processing of information within the network. For example, while one layer may be tasked with learning specific data features, the roles and interconnections of these layers are defined by hyperparameters, even though the fine-tuned network gets adjusted during training through weights and biases in backpropagation.

It is important to recognize that simply increasing the size and complexity of a network does not guarantee improved accuracy. Similar to how an elephant, despite its larger brain, does not match human cognitive abilities and intelligence, a larger network does not automatically enhance performance. The efficacy of a DNN depends on its architectural design, which is profoundly influenced by the hyperparameter settings. Proper tuning

(hyperparameter optimization) ensures that the network's structure is optimized for specific tasks, balancing both design and size for optimal performance.

Therefore, effective hyperparameter optimization is essential, particularly in complex domains such as hydrology, where it remains an unresolved challenge (Arsenault et al., 2023). In AI, techniques for hyperparameter optimization include manual tuning, systematic grid search, and random search, each contributing to enhanced model performance, with random search often outperforming other methods (Bergstra & Bengio, 2012).

1.3.4. Advanced DNN Architectures in Hydrology Domain

In the context of hydrology, DNNs provide powerful tools for modeling and predicting the complex, nonlinear relationships inherent in hydrological systems and datasets (Kratzert et al., 2024; Shen & Lawson, 2021). Hydrological processes are influenced by numerous factors such as climate, land uses, topography, soil types and geomorphology, which interact in ways that are often too complex to model using traditional deterministic approaches (Donnelly et al., 2024a). DNNs are particularly well-suited for scenarios where the relationships between variables are not fully understood or are too intricate to be captured by simple mathematical models (Russell & Norvig, 2020; Goodfellow et al., 2016).

Among the different types of DNNs, Long Short-Term Memory (LSTM) networks (Hochreiter & Schmidhuber, 1997) have proven to be particularly effective for time series analyses, which is highly relevant in hydrological rainfall-runoff modeling (Kratzert et al., 2018). LSTMs are a type of Recurrent Neural Network (RNN) designed to capture temporal dependencies in sequential data, making them ideal for tasks such as predicting rainfall patterns or streamflow, where understanding the sequence of past events is crucial for accurate forecasting.

In addition to LSTMs, the new generation of Transformer models (Vaswani et al., 2017) and generative AI techniques have shown significant promise across various applications, including timeseries analyses. Originally developed for natural language processing, Transformers excel at capturing long-range dependencies in data through their attention-based mechanism, eliminating the need for sequential processing. This makes them powerful tools for hydrological modeling (Liu et al., 2024), where they can be leveraged to analyze complex temporal relationships in large datasets. Moreover, Generative AI, with its ability to create new data instances and explore potential scenarios, further enhances the capability of these models in hydrological forecasting and simulation.

Regardless of the type of DNN used, whether LSTMs, Transformers, Convolutional Neural Networks (CNNs), or traditional feedforward neural networks, it is crucial to apply these models precisely and optimize (Hosseini et al., 2024a) them properly for the specific datasets and problems at hand. The success of DNNs in hydrology depends not only on the choice of architecture but also on careful tuning and validation of the DL models to ensure that they generalize well to new, unseen data and hydrological conditions.

1.3.5. Long-Short-Term-Memory Networks (LSTMs)

Long-Short-Term Memory Networks (LSTMs) are a specialized type of Deep Neural Network (DNN) that evolved from Recurrent Neural Networks (RNNs) (Figure 3). LSTMs are designed as "self-trained memory systems with storage units that can mimic system storage and fluxes" (Shen & Lawson, 2021). Originally developed for sequence modeling in AI, LSTMs have gained extensive applications in machine translation and handwriting recognition (Hochreiter & Schmidhuber, 1997). The LSTM architecture is adept at capturing long-term dependencies in timeseries data, making it particularly suitable for modeling dynamic systems such as hydrological processes in watersheds.

LSTMs are particularly effective at capturing intricate temporal and spatial dynamics in hydrological data, significantly enhancing modeling capabilities (Kratzert et al., 2018a; Shen & Lawson, 2021; Refsgaard et al., 2022). Applications of LSTMs span a diverse range, from rainfall-runoff modeling (Kratzert et al., 2024; Arsenault et al., 2023; Hashemi et al., 2022; Frame et al., 2022; Refsgaard et al., 2022; Gauch et al., 2021; Shen & Lawson, 2021; Nearing et al., 2020a, b; Kratzert et al., 2019a; Kratzert et al., 2018a) to research on climate change impacts (Mahdian et al., 2024), flood resilience (Ahmadi et al., 2024), soil moisture prediction (Feng et al., 2020), soil water erosion (Donnelly et al., 2024a; Khosravi et al., 2023), and water quality and temperature prediction (Rahmani et al., 2021).

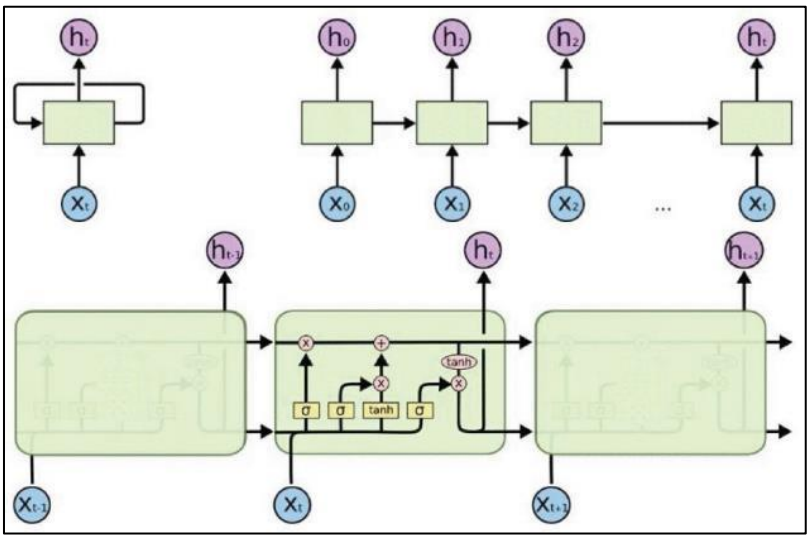


Figure 3. Schematic of an LSTM cell and its difference from a traditional RNN (Fu et al., 2019). The upper left shows a traditional Recurrent Neural Network (RNN), where only some information is retained over time and passed forward through the network. In contrast, the Long Short-Term Memory (LSTM) architecture, shown in the lower part, includes additional gates—specifically the input, forget, and output gates—that regulate the flow of information. These gates allow the LSTM to preserve and update relevant memory over longer sequences, addressing the vanishing gradient problem that commonly occurs in traditional RNNs. As a result, the LSTM is better equipped to capture long-term dependencies and is more effective for complex sequence learning tasks, such as those in the hydrology domain.

Unlike traditional RNNs, LSTMs utilize memory cells that store information over extended time periods, which is crucial for learning relationships between input and output features across longer time scales. This capability allows LSTMs to handle complex temporal patterns and dependencies, addressing issues like vanishing and exploding gradients that commonly affect RNNs. In hydrological modeling, this resilience is essential for capturing phenomena

with extended durations, such as snow accumulation and snowmelt, which occur over longer time spans compared to precipitation events.

LSTMs utilize two states: the cell state and the hidden state, along with three gates: input, forget, and output. The cell state enables long-term memory, while the gates are trained to decide which information should be retained over multiple time steps and which should be discarded. This design addresses the vanishing gradient problem, a common issue in deep learning where gradients diminish exponentially over time, preventing effective learning.

The LSTM architecture incorporates memory cells (c[t]) that act as the system's memory. The memory cells can be modified by the forget gate (f[t]), which removes certain states, and the input gate (i[t]) and cell update (g[t]), which introduce new information. The cell update represents the added information, and the input gate controls which cells receive this new information. Finally, the output gate (o[t]) determines which information stored in the cell states is outputted.

The selection of LSTM networks as the model architecture for streamflow prediction was motivated by their ability to capture long-term dependencies and overcome challenges in hydrological modeling. The LSTM architecture, with its memory cells and gate mechanisms, enables the storage and retrieval of information over extended time periods, making it well-suited for modeling dynamic systems like watersheds.

To optimize the performance of the LSTM networks in streamflow prediction, specific modifications and enhancements can be implemented. These enhancements aim to further capture the complexities and patterns of hydrological processes, leading to improved predictive capabilities. Kratzert et al. (2019a) introduced modifications and enhancements to the standard LSTM architecture to improve its performance in rainfall-runoff modeling (Figure 4). The modified LSTM network operates as follows: Given an input sequence x = [x[1], ..., x[T]] with T time steps, where each element x[t] represents a vector of input features at time step t ($1 \leq t \leq T$), the network goes through a series of calculations involving gates that control the information flow within the LSTM network. These gates include the input gate (i[t]), forget gate (f[t]), and output gate (o[t]), which determine how information is processed and propagated throughout the network. The LSTM equations for the forward pass are as follows (Equations set 1):

$$i[t] = \sigma(W_i x[t] + U_i h[t-1] + b_i)$$

$$f[t] = \sigma(W_f x[t] + U_f h[t-1] + b_f)$$

$$g[t] = tanh(W_g x[t] + U_g h[t-1] + b_g)$$

$$o[t] = \sigma(W_o x[t] + U_o h[t-1] + b_o)$$

$$c[t] = f[t] * c[t-1] + i[t] * g[t]$$

$$h[t] = o[t] * tanh(c[t])$$

Equations 1

In the equations, i[t], f[t], and o[t] represent the input gate, forget gate, and output gate, respectively. The cell input is denoted as g[t], while x[t] represents the input at time step t. The recurrent input from the previous time step is h[t−1], and c[t−1] represents the cell state from the previous time step. The parameters W, U, and b are the learnable weights and biases for each gate. The sigmoid function σ(·) and hyperbolic tangent function tanh(·) are used for activation, and * denotes element-wise multiplication.

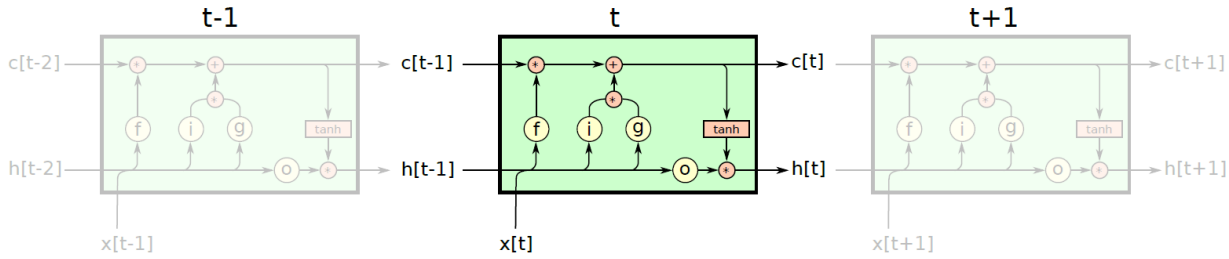


Figure 4. Schematic of an LSTM design for hydrological rainfall-runoff modeling, (Kratzert et al., 2019a).

By utilizing LSTM networks and integrating these enhancements, Kratzert et al. (2019) demonstrated significant improvements in streamflow prediction compared to traditional hydrological models; the LSTM-based approach outperformed not only regionally calibrated hydrological models but also models calibrated individually for each catchment. This highlights the effectiveness of LSTM networks and their potential to address the challenges associated with hydrological modelling.

1.3.5.1. Multi-Timescale Prediction (MTS-LSTM)

To enable the generation of discharge predictions at different time steps while ensuring consistency across timescales, the Multi-Timescale LSTM (MTS-LSTM) architecture was introduced (Gauch et al., 2021). The traditional approach of training separate LSTMs for fine-resolution data (e.g., sub-daily, hourly, sub-hourly) forcings poses significant computational challenges due to the extensive processing required. For example, with an annual dataset, the hourly LSTM networks would need to process 8,760 (365*24) time steps to predict a single hour, resulting in lengthy training times and potential instability in the learning process. Moreover, inconsistencies between the different time steps predictions may arise as the two LSTM networks operate independently (Gauch et al., 2021).

The MTS-LSTM overcomes these challenges by adopting a novel strategy. It incorporates the concept of processing past time steps at lower temporal resolutions, leveraging the fact that fine-grained details from distant past time steps may not significantly impact the accuracy of current predictions. For instance, to predict the discharge on September 10th at 9:00 am, fine-grained data from the preceding few days or weeks may be crucial, but details such as rain occurrence at specific hours several months ago may be less relevant (Gauch et al., 2021).

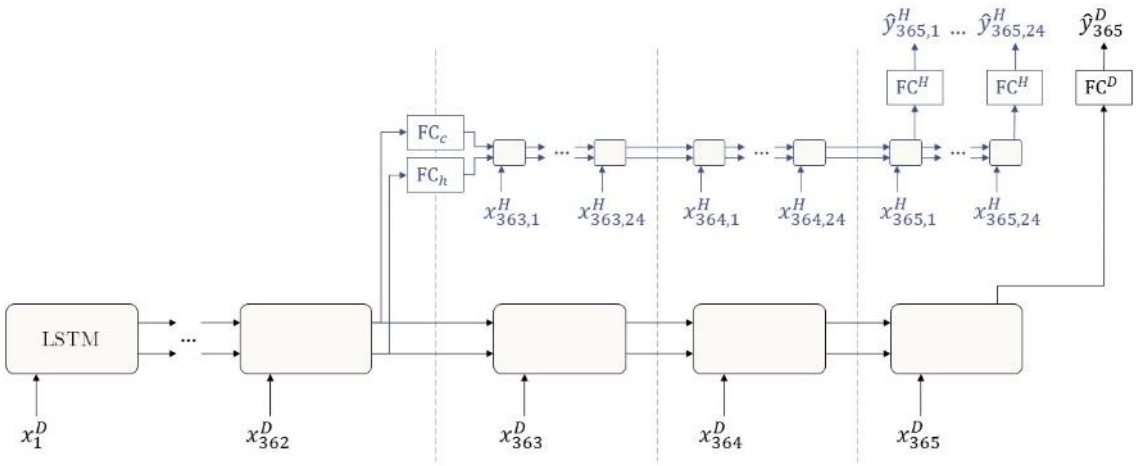


Figure 5. The architecture of the MTS-LSTM model, showing how it processes data at different timescales. Initially, the model handles the first 362 days of input at a daily frequency. From day 363 onwards, the daily LSTM continues processing daily inputs, while the hidden and cell states from day 362 are transformed linearly and used to initialize the hourly LSTM. This hourly LSTM then processes the next 14 days of hourly data to produce 24-hourly forecasts for the current day (from Gauch et al., 2021).

The MTS-LSTM architecture (Figure 5) implements the following approach by parallelizing two daily and hourly LSTMs connecting to each other; Daily meteorological information from a longer time in the past (e.g., 1 year) is fed into the LSTM to predict both daily and hourly discharge. At a specific time point (some days to some months: e.g., 14 days) before the present, the processing diverges into two branches. The first branch continues with daily inputs until it produces the daily prediction for the current day, similar to conventional daily-only prediction models. The second branch, however, introduces a significant innovation. It takes the LSTM state from the recent time (e.g., 14 days before the present), applies a linear transformation to it, and uses the resulting states as the initial states for another LSTM in parallel. This second LSTM processes the recent time (e.g., 14 days) hourly data to generate the 24-hourly predictions for the current day. Consequently, a single forward pass through the MTS-LSTM simultaneously generates both daily and hourly predictions (Gauch et al., 2021).

By employing individual branches for each timescale, the MTS-LSTM model can accommodate different forcing products or multiple sets of forcing data at each timescale. This flexibility allows for the integration of diverse data sources, leading to potential improvements in prediction accuracy. During training, the daily and hourly predictions are linked to promote consistency across timescales. A regularization technique can be applied to the loss function, penalizing the model if the average daily prediction aggregated from the hourly predictions deviates from the daily prediction. This regularization encourages the model to produce coherent hourly predictions that align with the overall daily pattern (Gauch et al., 2021).

In other word, an MTS-LSTM is made up of lots of LSTMs arranged in a branching configuration. At a specific temporal resolution, each LSTM branch processes a portion of the input timeseries. Then sends its states to the other parallel LSTM branch that allows information to be shared between branches (e.g., daily and hourly). Through an additional branch, this pattern was adapted to handovers across data products (rather than just

timescales). This would allow them to combine past and future data in a single prediction pathway, resulting in more precise predictions.

In summary, the MTS-LSTM architecture presents a powerful and innovative solution for multi-timescale prediction, facilitating the generation of both daily and hourly discharge predictions within a single model. It addresses challenges related to computational efficiency and prediction consistency, and its flexible design allows for the integration of diverse data sources, ultimately leading to improved prediction accuracy.

1.3.5.2. NeuralHydrology Library: LSTM-based Hydrological Modeling

There are currently various libraries that facilitate setting and training DLs (e.g., LSTMs), including NeuralHydrology Library that we employed in our research. This is a specialized Python library designed for hydrological modeling using DL techniques, specifically LSTM-based models. Built on PyTorch, NeuralHydrology provides a robust framework for hydrologists and researchers, offering a suite of models, tools, and functionalities tailored to the unique demands of hydrological analysis (Kratzert et al., 2022). The library is notable for its user-friendly interface, comprehensive documentation, and a variety of pre-implemented models, which streamline the integration of LSTM-based DLs into hydrological research and enhance streamflow prediction accuracy (Kratzert et al., 2022).

The developed library is designed to leverage (Graphics Processing Units) GPUs' computational power for accelerating the training and inference processes of Long Short-Term Memory (LSTM) networks (Kratzert et al., 2022). By taking advantage of GPUs' parallel processing capabilities, these models significantly reduce computational time, facilitating the analysis of extensive hydrological datasets, specifically for hourly prediction. Employing this library for rainfall-runoff modeling is suggested regarding it being user-friendly.

The process of selecting optimal hyperparameters is intricate and critical for AI/DL success. Appendix 3 serves a comprehensive definition of MTS-LSTM hyperparameters with detailed strategies we approached to choose them for being optimized in this research.

1.3.6. Transformers

Transformers (Vaswani et al., 2017) are a class of Deep Learning (DL) models, initially designed for natural language processing (NLP) tasks but now widely applied across various domains, including computer vision and time series analysis. Unlike traditional sequence models like RNNs or LSTMs, Transformers utilize a self-attention mechanism that allows them to dynamically weigh the importance of different input elements, independent of their position in the sequence.

Key Features of Transformers:

Self-Attention Mechanism: This mechanism enables the model to focus on different parts of the input sequence, effectively capturing relationships between distant elements, such as words in a sentence or points in a time series.

Positional Encoding: Since Transformers do not process data in a fixed order, positional encodings are used to maintain the sequence information, ensuring that the model can distinguish the order of the input elements.

In machine translation, a Transformer model can accurately translate a sentence from one language to another by understanding the context of each word relative to others in the sequence, rather than relying solely on the proximity of words. This makes Transformers particularly effective for tasks where context is key to understanding the data. Transformers, originally designed for natural language processing, are adept at identifying long-range dependencies within data thanks to their attention-based mechanism, which removes the need for sequential processing. Consequently, they serve as powerful tools in hydrological modeling (Liu et al., 2024), enabling the analysis of complex temporal relationships in large datasets.

1.3.7. Convolutional Neural Networks (CNNs)

CNNs are also a class of deep learning models particularly effective for image processing tasks that have been used for hydrological tasks. They are designed to automatically and adaptively learn spatial hierarchies of features from inputs. CNNs consist of multiple layers, including convolutional layers, pooling layers, and fully connected layers, each performing specific operations that contribute to the model's ability to recognize patterns and objects in data.

Convolutional Layer: Applies a set of filters to the inputs, detecting complex patterns.

Pooling Layer: Reduces the dimensionality of the data by down-sampling, which helps to make the network invariant to small translations of the input.

Fully Connected Layer: Connects every neuron in one layer to every neuron in another, helping the model make final decisions based on the extracted features.

To simplify, if we consider a CNN designed to identify cats in images. The early layers might detect edges and simple textures, while deeper layers could recognize parts of a cat's face, such as eyes or ears, eventually leading to a classification decision. In hydrology, CNNs have been used for several applications, such as:

Satellite Image Analysis: CNNs have been employed to analyze satellite imagery for various hydrological applications, such as mapping flood extents, monitoring land use changes, and assessing water bodies. By detecting patterns and features in high-resolution images, CNNs can classify different types of terrain, vegetation, and water bodies, aiding in the management and prediction of water resources.

Precipitation Estimation: CNNs have been used to estimate precipitation from satellite-based radar and infrared images. By learning the relationship between cloud patterns in the images and rainfall intensity on the ground, CNNs can improve the accuracy of precipitation estimates, which is crucial for flood forecasting and water resource management.

River Ice Monitoring: CNNs can also be applied to monitor river ice formation and break-up from remote sensing images. This is important for predicting ice-related flooding events and managing river flow during winter and spring seasons.

Soil Moisture Prediction: Using CNNs to analyze microwave remote sensing data allows for the estimation of soil moisture content, which is vital for agricultural management, drought prediction, and hydrological modeling. CNNs can effectively capture spatial patterns and variations in soil moisture across large areas.

In each of these examples, CNNs demonstrate their ability to process and analyze complex visual data, making them valuable tools in the field of hydrology for tasks that involve spatial patterns and image-based information.

1.4. Hydroinformatics in Water Science and rainfall-runoff modelling

Several advancements in employing AIs in Earth System Sciences have been presented, and still tremendous opportunities left ahead. As an example of knowledge discovery using AI, a study in the field of climate science could be referred. A team of scientists won the 2018 Gordon bell prize for a DL model that discovers detailed information about extreme weather events that were previously buried in climate data far from human eyes. They employed a supercomputer with specialized GPU to exceed the exaop level (10^{18} operations per second) as the first ML program to do so (Kurth et al., 2018). Later, Rolnick et al. (2019) presented a 60-page catalog of ways in which AIs can be used to tackle climate change.

AI models found their way under the category of data-driven models into hydrology more than three decades ago (Refsgaard et al., 2022). Hydroinformatics is the use of numerical modeling and data-driven models to effectively and sustainably take advantage of water resources. Hydrologists employed different Hydroinformatics techniques to benefit from data repositories and timeseries. The adoption of those techniques originating in the AI community, such as ANN/DNNs, support vector machines (SVMs), genetic algorithms (GAs), and many other ML techniques have been of great interest in Hydroinformatics. However, during the first decades, not only these models were new for hydrologists, but also, there was not adequate hardware to properly address the relevant huge computational costs. In recent years, we are witnessing several achievements in computer and AI sciences, which have opened new opportunities for lots of those impediments (See: Refsgaard et al., 2022; Shen & Lawson 2021; Abrahart et al. 2012).

ANN/DNNs were among the most interesting tools for hydrologists in this story for a variety of applications, including neural networks for rainfall-runoff modeling. According to Abrahart et al. (2012), the race started by “extending the basic models”. Earlier ANNs concentrated on addressing restrictions, such as limited soft/hardware, learning, and

inflexible critics. In a manner similar to multiple linear regression, traditional hydrological input variables were employed to estimate traditional hydrological output variables. The fundamental model was the Feedforward Neural Networks (FFNN): a multi-layer perceptron (MLP) trained using one of the multiple popular techniques (e.g., backpropagation of error (BPE): Rumelhart et al., 1986).

Then "neuro-fuzzy" explorations emerged, which were a hybrid of ANNs and Fuzzy Logic (FL). To develop rules and optimize parameters, neuro-fuzzy systems (NFS) combined the reasoning style of fuzzy systems with the automated learning capabilities of ANNs. In theory, neuro-fuzzy systems can produce transparent explanations of the fundamental processes in a catchment. Maier and Dandy (2000) strongly recommended neuro-fuzzy systems for hydrological modeling; also, the Adaptive Neuro-Fuzzy Inference System (ANFIS, Jang, 1993) has gained popularity owing to its ease of use via the MATLAB Fuzzy Logic Toolbox (Abrahart et al., 2012).

Abrahart et al. (2012) refer to the next phases in the use of ANNs in hydrology as "Neuro-genetic" and "Neuro-wavelet" investigations. Neuro-genetic models were created by incorporating ANNs and evolution-based methods known as neuro-genetic systems. The GA and Evolutionary Algorithm (EA) techniques were later widely used in hydrology. Neuro-wavelet systems were hybrids of ANNs with wavelet decompositions. The wavelet transformation detects frequencies encoded in timeseries as well as the time and position of their occurrence, making it a strong approach for studying and modeling hydrological non-stationary timeseries. The critical thing was not the time variation of the signals, but the underlying mechanisms that generate the data evolution.

Another applicable use of ANNs in the literature was their employment as surrogate models. An emulator (surrogate or "meta-modeler") mimics the behavior of a more complicated model (e.g., a process-based physical model). As a result, hydrological neuro-emulators are ANNs designed to imitate the full functionality of perceptual hydrological models. Neuro-emulation provides higher processing speeds, more efficient and effective hybrid model coupling, and better understanding of the internal dynamics of ANNs; also, they can be trained to represent uncertainty in traditional hydrological models (Abrahart et al., 2012).

1.4.1. Advanced DL/DNNs in hydrology and rainfall-runoff modeling: A more in-depth review

The expanding utilization of DLs/DNNs as stand-alone techniques and in conjunction with other model types has been a notable development in recent years (Nearing et al., 2020a, b; Shen, 2018). This advancement has been made possible by increased access to huge amounts of data and enhancement of computer computations by GPUs. Furthermore, the full potential of information buried in hydrological big data, which is multivariate and distributed across spatial and temporal scales, cannot get unearthed by traditional perceptual hydrological models in terms of input data and parametrization (Nearing et al., 2020a, b). This necessitated

the development of more adaptable models capable of making the best use of the abundance of accessible data sources (Refsgaard et al., 2022).

Latest evidence has shown that DLs can be used to consolidate hydrological process understanding. This results in the development of a novel class of process- or knowledge-guided hybrid models, which, for example, employs perceptual hydrological models' simulations as input to the DLs (Konapala et al., 2020), incorporates first-order conservation laws into DLs (Hoedt et al., 2021; Read et al., 2019), or directly integrate perceptual hydrological models into DLs' architectures (Kraft et al., 2020). Additionally, research work focuses on explainable DLs to unearth their so-called "black-box" concept, with the goal of eventually applying DLs to develop new hydrological process knowledge (Kratzert et al., 2019a,b; Nearing et al., 2020a, b; Refsgaard et al., 2022); the idea of learning from big data.

Moreover, there are no scale-relevant hypotheses in hydrology, although DLs imply their existence. Large-scale hydrological datasets include far more information than hydrologists have been able to put into theory or models. While there is increased interest in DLs among the community, we continue to have fundamentally subjective and non-evidence-based inclinations for approaches that rely on novel forms of "process understanding". In a modeling discipline increasingly dominated by DLs, the hydrology community must work on creating a quantitative knowledge of where and when hydrological process understanding is beneficial (Nearing et al., 2020a, b).

Long Short-Term Memory networks (LSTMs) are one of the most accurate and extrapolatable DLs currently available in hydrological science (Kratzert et al., 2024; 2019a, b; Nearing et al., 2020a; Gauch et al., 2021; Frame et al., 2021a). Because of its memory capacity, it is useful for hydrological modeling (Shen & Lawson, 2021; Kratzert et al., 2018a). Kratzert et al. (2021c) claim that LSTMs can learn relevant hydrologic processes (such as snow accumulation and melting) without being explicitly trained.

Mai et al., 2022 conducted a model inter-comparison investigation to examine and evaluate the simulations of several model configurations over the same research domain. This article evaluates models not only in terms of their ability to generate streamflow, but also simulations of actual evapotranspiration, surface soil moisture, and snow water equivalent. The excellence of the employed LSTMs was noticeable in all experiments; even in the most rigorous spatio-temporal validations, DLs outperformed perceptual hydrological models. The regional LSTM model used in this research was trained for 141 calibration locations simultaneously, resulting in a single trained model for the entire domain that can be executed for any catchment immediately if the appropriate input data is accessible.

Because of the empirical successes of DLs, as well as the fact that process understanding is necessary for modeling the so-called "out-of-sample" catchments (e.g., under accelerating anthropogenic-driven global change), Knowledge-Guided ML is essential for the Earth sciences (Nearing et al., 2020a). Inserting inductive bias by theory-guided DLs is another field of study to incorporate recognized constraints and information into DLs; This paradigm is often referred to as physics- or theory-guided MLs/DLs for simulations of physical processes (Gauch et al., 2021). Several studies have been done so far on physics-informed DLs. For example, graph models stemming from the spatial structure of the river network, or Mass-

Conserving LSTMs (MC-LSTMs) that conserve mass inputs by the architecture of the DLs (e.g., Frame et al., 2022; Nearing et al., 2020a; Hoedt et al., 2021; Tsai et al., 2020).

By pushing the inductive bias of LSTM networks toward accumulating information over time to simulate the redistribution of those accumulated quantities, a Mass-Conserving LSTM (MC-LSTM) was introduced which adheres to the conservation laws; it establishes a new benchmark for predicting peak flows (Hoedt et al., 2021). The article demonstrates that MC-LSTM states correlate with real-world processes in hydrology and are hence interpretable. However, later, Frame et al. (2022; 2021a), suggested that conservation principles may not be advantageous for accurate hydrological modeling, due to input (precipitation) and output (streamflow) flaws. This hypothesis was investigated using physics-informed DLs, and the results showed that imposing closure in the rainfall-runoff mass balance appears to reduce hydrological models' overall accuracy (Hoedt et al., 2021), especially during extreme events. Moreover, Kratzert et al. (2019a) introduced an Entity-Aware-LSTM (EA-LSTM) architectural adaptation to the conventional LSTM architecture that enables learning catchment similarities by embedding as a feature layer in DLs. They demonstrated that this well-acquired catchment similarities with what is expected based on prior hydrological understanding.

A DL structure known as HydroNets exploits prior knowledge of a hydrologic region's sub-basin river structure (Moshe et al., 2020). According to the study, importing prior knowledge of river structure decreases sample intricacy and enables scalable and more precise hydrologic modeling even with only a few years of data. Unlike perceptual hydrological models, DLs can utilize various precipitation data products at the same time. A sensitivity analysis revealed that the LSTMs combine precipitation products differently depending on location, as well as different for simulation of different parts of the hydrograph; whereas temperature estimates between different data products are frequently similar, precipitation estimates are frequently subject to large disagreements (Kratzert et al., 2021b).

In another research, Li et al., (2021), demonstrated that the LSTM networks statistically prioritize the physically important gages above the irrelevant gages. The study says rainfall-runoff predictions are improved when redundant gages are removed. The physical consistency not only indicates that LSTMs are paying more attention to the more relevant gages, but it also gives a method for selecting rainfall gages for the model. Moreover, it is pretty tricky to support prior information in a manner that advantages DLs; for instance, when predictions are the goal, there is evidence demonstrated that conceptual hydrological models' outputs do not meaningfully enhance the predictions of an LSTM (Frame et al., 2022; Gauch et al., 2021).

Tsai et al. (2020) employed DLs to develop articulable hypotheses regarding whether physical elements involving soil texture, soil thickness and slope drove water storage and streamflow to be correlated in a given way in a catchment. A framework for learning a global mapping between inputs (and optionally responses) and parameters known as differentiable parameter learning (dPL) was introduced later by Tsai et al., (2021). Importantly, dPL features favorable scaling curves that previously were unknown to geoscientists. Without requiring reimplementation, the generic architecture encourages the integration of DLs and conceptual hydrological models.

In addition, Jiang et al. (2022) proposed a hybrid knowledge-informed DL that can efficiently perform parameter calibration of a fully integrated process-based traditional hydrological model. They first began by reducing the number of model parameters based on the mutual information between model responses and each parameter. Later, they executed numerous ensemble-runs to construct training sets for the inverse mapping, which picks informative model responses for estimating parameters regarding MI-based parameter sensitivity. The study emphasizes the significance of exploiting data-driven knowledge in DL-assisted model calibration, especially for computationally expansive conceptual hydrological models.

In another study, Rahmani et al. (2021) stated that not only does streamflow influence temperature fluctuations directly, but it also demonstrates multi-faceted hydrologic catchment dynamics, including baseflow contributions and surface runoff residence times, which could improve streamflow temperature modeling. Under the same subject, Rasheed et al. (2022) claimed since various sources of streamflow components (e.g., surface runoff and groundwater discharge) have distinctive flow temperature signatures, water temperature measurements might be able to distinguish their contributions. To calculate both proportions of streamflow origins and temperature, they connected a DL to a process-based model of stream temperature. This method produced an acceptable approximation of stream temperature and could confidently estimate proportions of streamflow origins.

De La Fuente et al. (2022) developed and evaluated Hydro-LSTM, an LSTM modification based on isomorphic linkages between the LSTM structure and the water budget principles for updating the state variables in a dynamic environmental system. They used data from ten distinct catchments with varying hydro-climatic variables to compare the Hydro-LSTM and LSTM architectures. Results reveal that Hydro-LSTM requires fewer nodes (cell states or neurons) to achieve comparable performance to a regular LSTM network. Furthermore, when adopting the new structure, the weights corresponding with the input variables have a more straightforward interpretation. This study demonstrates that it is possible to improve the interpretability of DLs and obtain useful knowledge in the domain.

Regarding interpretable AI, many generic valuable methods exist using DLs; though, they cannot be applied directly in hydrology in terms of the needed expectations. Therefore, “the domain scientists are responsible for customizing DLs for knowledge discovery” (Shen & Lawson, 2021). While there is a lot of research going towards knowledge-guided DLs (Read et al., 2019), there are lots of approaches to get to the objective of combining physics with DLs; for instance, methods for parameter learning (Tsai et al., 2020) and physics-informed-DLs (Shen et al., 2021).

In an adversarial approach, conceptual hydrological models could be utilized to analyze causal controls and distinguish between competing components (Fang et al., 2020a). Experiments on small-scale catchments have shown that the internal states of LSTMs can be interpreted. Lees et al. (2021) tried to determine what information the LSTMs capture about the hydrological system by extracting the tensors that represent the learned translation from inputs (precipitation, temperature) to outputs (discharge). The study demonstrated that

LSTMs contain information corresponding to established hydrological processes, which is related to the concept of variable-capacity soil moisture storage.

Kratzert et al. (2021a) developed two distinct model components to estimate streamflow: one for runoff generation and one for routing. The former was LSTM-based forecasting the discharge contribution of each sub-catchment through the river network. The latter was a Graph Neural Network (GNN) that routes water in a hierarchical order along the river network (Kratzert et al., 2021a). In another research, Gauch et al. (2021a,b) developed a Multi-Timescale LSTM (MTS-LSTM) that can more efficiently provide rainfall-runoff projections at multiple timescales; highly efficient for sub-daily predictions to decrease the expensive computational costs. For instance, training LSTMs on hourly data requires extremely long input sequences making the learning difficult and costly. The multi-timescale architectures are computationally more affordable, accurate and robust. Aside from prediction performance, MTS-LSTM can analyze many input variables at various timeframes (Gauch et al., 2021).

Furthermore, Xie et al. (2021) developed a physics-guided DL known as (PHY-LSTM). The study states employing three physical mechanisms to train DLs: (1) extreme heavy rainfalls when the soil water is saturated, (2) long-duration rainless occurrences when the soil water is drained, as well as (3) the monotonic rainfall-runoff relations. Synthetic samples, different from the actual samples, were developed to provide additional hydrological theories that were not present among data records; the made-up year of the synthetic samples was introduced to the original observations to train DLs to cope with the extreme events. The study states that synthetic samples can successfully improve flood peak simulation and decrease the number of negative streamflow.

In addition, the process of transferring trained model components from one task to another known as Transfer Learning is a helpful approach for such purposes since it facilitates the borrowing of key learned features on much broader datasets (Ma et al. 2021). The literature demonstrates conceptual hydrological models outperform if calibrated for specific catchments and poorly when calibrated regionally. On contrary, as expected, DLs improve with more and diverse training (good quality) data, with respect to their capacity of extracting latent features in massive data. According to Kratzert et al. (2020a), maybe it is because conceptual hydrological models do (must) not learn hydrological processes from data. For example, a trained single LSTM over hundreds of water basins in the US, dramatically outperformed a set of locally-calibrated conceptual hydrological models. The latest is expected and there have been a large body of the community working on model adequacy and model structural error (See, e.g., Prieto et al., 2019; 2021; 2022; Clark et al., 2008; Fenicia et al., 2011; Kavetski et al., 2011a, b). When compared to conceptual hydrological models; a single LSTM outperforms even in out-of-sample ungauged catchments (Kratzert et al. 2019).

Moreover, Ma et al. (2021) stated that LSTM-based streamflow models that were pre-trained over the US could be moved to catchments on other continents using Transfer Learning in the sense of weight initialization and weight freezing. Compared to locally trained models using all water basins, the Transfer Learning models outperformed. Also, Kratzert et al. (2018) provided an example of which a pre-trained LSTM model based on regional data

could be developed by more limited data for an individual catchment; this outperforms LSTMs only-trained on the limited data for the catchment, despite the fact that the regional model may include catchments with a variety of behaviors.

In another study, Khoshkalam et al. (2021) examined the LSTM-based modeling approach to the performance of a recently constructed hydrological forecasting system based on an Ensemble Kalman Filtering data assimilation scheme. To transfer pre-trained information to chosen catchments in Canada, pre-trained LSTM networks with CAMELS dataset were used over the US. The pre-trained LSTM networks were then retrained locally employing data from the specified catchments. Hence, the findings could help construct forecasting systems in areas with insufficient hydro-meteorological records.

In a relevant subject in surface water modeling, reservoir computing has demonstrated phenomenal progress in modeling and should be given more consideration by hydrologists. An LSTM network can simulate reservoir operation; LSTM has been shown to successfully mimic reservoir operation at large scales for a restricted number of applications, particularly for those smaller reservoirs in a catchment (Ouyang et al. 2021). Investigations on whether types of dammed water basins may be well-represented by LSTMs, demonstrated that a consistent modeling technique in which smaller dams (storing about a month of average streamflow or less) are implicitly modeled as part of catchment rainfall-runoff processes, and large reservoirs of particular sorts are explicitly represented. Dammed water basins, also, must be present in the training dataset. Ouyang et al. (2021) claimed that compared to other approaches, LSTM networks outperform in modeling reservoirs.

LSTM networks are not the only DL algorithm suitable for timeseries; Convolutional Neural Networks (CNNs) are, also, applicable to timeseries modeling in the same way that they have been utilized in machine translation. For example, monthly data with 10 years of records is potentially insufficient for an LSTM model to learn. In such scenarios, CNNs can forecast the mismatch and reduce the error with the simulated water storage significantly (Shen & Lawson, 2021). Moreover, the performance of employing lagged data directly as inputs was comparable to applying a CNN unit. Such an approach could discover valuable insights by connecting LSTMs and Data Integration techniques (Feng et al., 2020). According to Feng et al. (2020), the introduced CNN-LSTMs could not give statistically competitive advantages than just incorporating 1-day-lag observations as inputs. They insist that granting the same sequential input information might be hard to construct an architecture that outperforms normal LSTMs substantially. This conclusion is consistent with the literature, where adjustments to LSTM structure have not resulted in meaningful gains.

Another introduced hydrologic Neural ODE model to perform MLs (conventional ANN architectures) for streamflow predictions (Höge et al., 2022), seems as easily interpretable as perceptual hydrological models and performs as well as complex DLs. Neural networks partially or completely substitute model internal processes in the Neural ODE. As a result, such models provide a solution to combine DLs with mechanistic modeling to produce time-continuous outcomes.

Lately, new research started to test functionality of attention-based DLs such as Transformers in hydrology. According to Li et al. (2022a, b), attention mechanisms in Natural

Language Processing (NLP) tasks outperform recurrent structures which could lead to better hydrological applications of MLs. Developing an ML model based on attention mechanisms and discarding recurrent architectures, their model learned long-term dependencies competitively to LSTMs, while decreasing training time. The results reveal that the attention-based model achieves a high NSE and captures peak flows more accurately, resulting in more interpretable models for hydrological modeling applications.

The introduction of new DLs has led to a surge in efforts to compare their performance on identical datasets. For instance, the advent of ChatGPT has directed significant attention towards Transformer models (Vaswani et al., 2017). However, Liu et al., 2024 have shown that their performance, though being competitive, does not statistically significantly surpass that of LSTMs in rainfall-runoff modeling when compared to Kratzert et al., 2024. It should be noted that both studies did not perform a systematic comprehensive hyperparameter search. The former relied on manual tuning, while the latter tuned a limited set of hyperparameters using the grid search method.

Additionally, researchers have explored other models such as Encoder-Decoder architectures, Convolutional Neural Networks (CNNs), and hybrid approaches combining different structures to enhance performance. There is a tendency to rapidly transition from one model to another without thoroughly evaluating whether the previous models were fully and effectively optimized and implemented on the datasets. This raises a critical question: have we mastered the application of these sophisticated AI and deep learning techniques within the hydrological domain?

Finally, referring to Fang et al. (2020) and Klotz et al., (2022), uncertainty estimation is an essential task for hydrology, but it is new time to apply it for hydrologic DLs. Daily projections could be advanced by including moving-average discharge (e.g., discharge over the preceding few days or even mean previous monthly discharge). The research states that having the appropriate hyperparameters and testing dataset, Monte Carlo Dropout with a data noise term may accurately quantify estimation error. Nearing et al. (2020a) demonstrated with adequate training data that DLs can be trained to reflect not just deterministic responses but also the parameters of the uncertainty in outputs (variances or quantiles). In other words, given a large enough sample, a stochastic DL should be able to reflect the fluctuations seen in the training dataset for fairly similar gauged catchments, even if all of those gauged catchments have their own “uniqueness of the place” that is poorly expressed in characteristic indices (See: Beven, 2020).

1.4.2. Ensemble Approaches for Multi-Objective Problems

In developing a robust regional model, the primary aim should be to achieve high-performance metrics across a wide range of catchments rather than relying solely on aggregated regional statistics like median metrics. Regional averages can mask critical shortcomings in individual catchments, where local performance is crucial for accurate prediction deficiencies. To address this, a focus on each catchment’s unique performance and specific requirements is necessary. Approaches such as “Multi-Objective Recommendations” (Zheng & Wang, 2021) and “Model soups” (Wortsman et al., 2022) highlight the importance

of balancing performance across diverse objectives in different domains. In regional hydrological predictions using deep learning, even outlier catchments with poorer performance must not be ignored.

Ensemble methods, which are widely used in fields including hydrology (Opitz & Maclin, 1999; Polikar, 2006; Prieto et al., 2021, 2022; Höge et al., 2018; Carneiro et al., 2022), offer promising solutions for addressing variability in model performance across different catchments. Potentially, ensemble approaches—where multiple models are combined, could yield higher predictive accuracy and adaptability than any single model. For hydrological applications, such ensemble methods can be strategically adapted to meet multi-objective challenges, such as optimizing for both accuracy across varying catchment characteristics. By tailoring ensembles to include catchment-wise models, it becomes possible to achieve a balance between generalization and accuracy, accommodating the specific hydrological dynamics of different basin.

Machine learning offers various ensemble methods with distinct advantages and trade-offs, making them well-suited to multi-objective problems in regional modeling. Bagging (Bootstrap Aggregating) and Voting are two of the most widely implemented techniques in ensemble learning (Breiman, 1996; Dietterich, 2000). Bagging involves training multiple models on different subsets of training data and averaging their predictions, which reduces variance and helps prevent overfitting (Breiman, 1996). In contrast, Voting, usually applied for classification objectives, combines the predictions of multiple models through majority voting or averaging, providing a straightforward yet powerful method to boost model performance (Polikar, 2006; Opitz & Maclin, 1999). These ensemble methods help address the diverse requirements of each catchment in a region by effectively capturing individual performance nuances that a single model may overlook.

1.4.3. General Challenges and Considerations in employing DNNs in Hydrology

While DNNs offer a flexible and powerful approach to modeling hydrological systems, they come with several challenges:

Data Requirements: DNNs typically require large datasets for training to achieve high accuracy. In hydrology, where data might be sparse or incomplete, this can be a significant limitation. But when using conceptual models, the limitation comes from the hypotheses we formulate when we develop the model. So, as expected, each approach has pros and cons.

Structural design and hyperparameter optimization: The architecture of a DNN—such as the number of layers, the number of neurons per layer, and the types of activation functions—plays a critical role in its performance. These hyperparameters need to be carefully tuned to optimize the model's accuracy and generalization capabilities. Improper design and suboptimal hyperparameter settings can lead to poor performance and bad learning habits, including overfitting or underfitting the data or skewness to some specific trends in the data

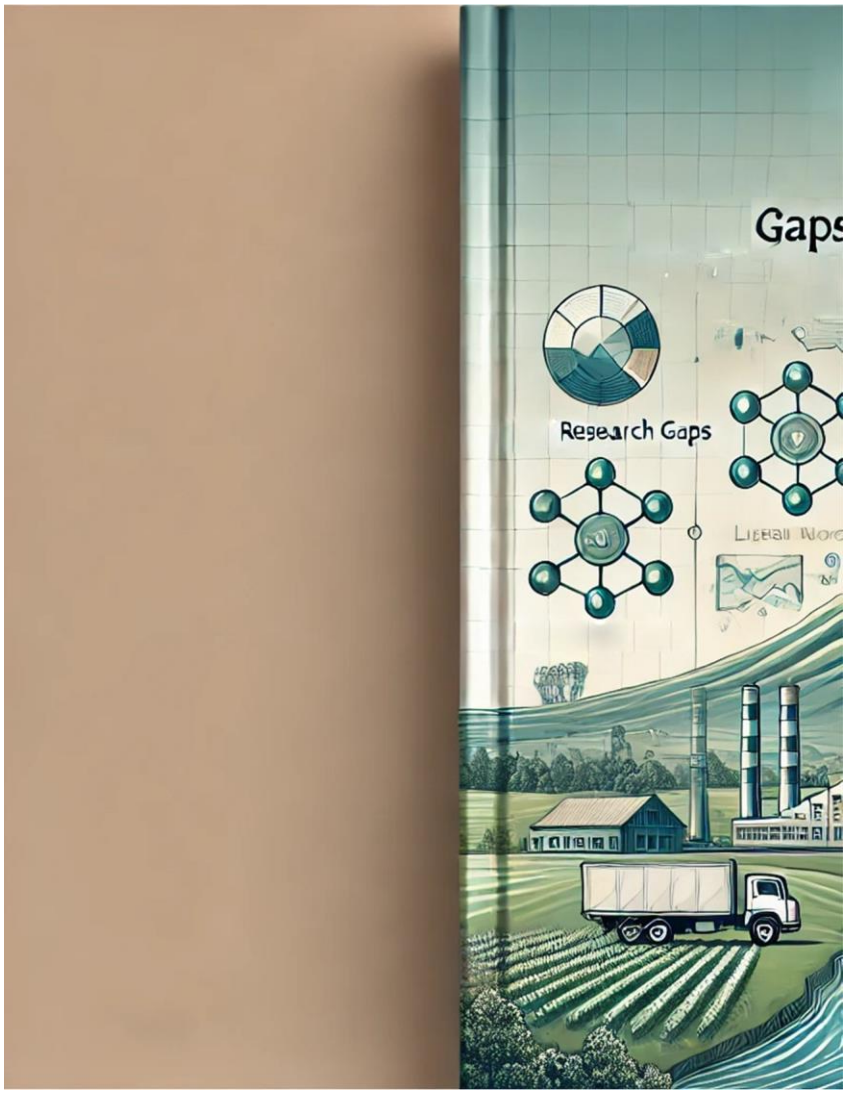
and specific learning behaviors by the trained networks. In hydrology, as mentioned, where the relationships between variables can be highly complex and nonlinear, finding the right architecture and hyperparameter configurations is crucial for the success of the DLs.

Overfitting/Underfitting: DNNs with a large number of internal parameters, such as excessive hidden layers or neurons, can be prone to overfitting, where the model performs well on training data but poorly on unseen data. This issue is particularly problematic in hydrological applications, where the ability to generalize to new conditions is crucial. Conversely, underfitting occurs when the model is too simple or not well-trained, with insufficiently modified parameters to capture the underlying patterns in the data for the specific task, leading to poor performance on both training and test data. Balancing these risks is essential for developing robust and accurate hydrological DL models. Which can remind to the problem between model complexity and the number of parameters.

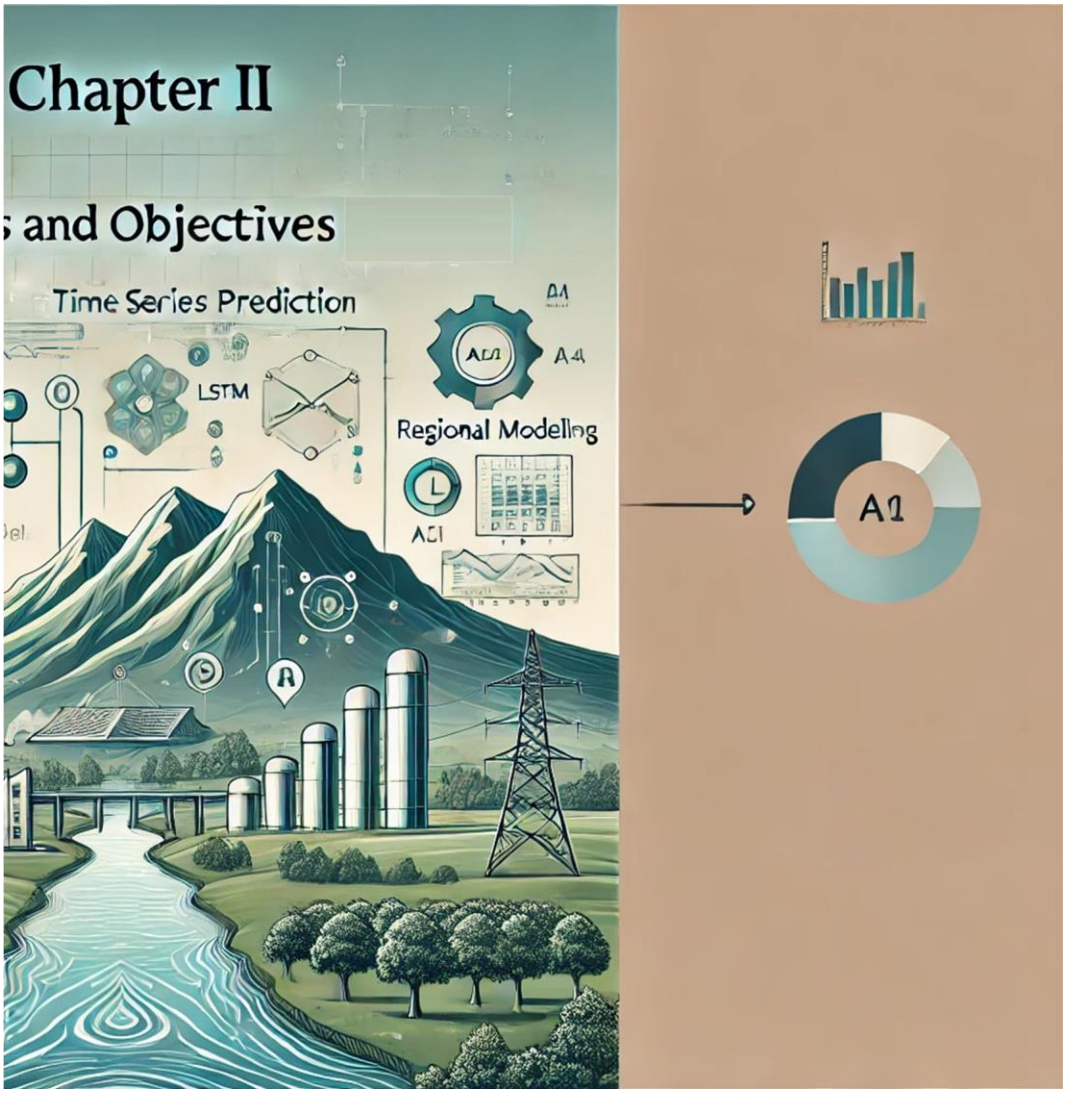
Interpretability: Unlike conceptual hydrological models, which are often based on "physical" principles, DNNs are of the so-called generation of "black-box" data-driven models. This lack of interpretability can be a drawback when understanding the underlying processes is as important as making accurate predictions. Although years of experience have proven that prediction and understanding are two different things, which in an ideal world without upscaling, uncertainty, heterogeneity problems, etc. they should converge.

Computational Cost: Training DNNs, can be computationally intensive, requiring significant resources in terms of processing power, time, and memory. This may limit their applicability in real-time forecasting or in resource-constrained environments.

As computational power increases and more hydrological data becomes available, the application of DNNs in hydrology is expected to expand. Recent research has focused on integrating physical models with DNNs to improve interpretability and robustness, developing more efficient training algorithms to handle large datasets, and applying DNNs to new areas of hydrology such as climate change impact assessment. Additionally, the continued development and application of advanced architectures like LSTMs and Transformers hold promise for further enhancing the accuracy and applicability of DNNs in hydrological modeling. However, careful consideration of their limitations and challenges is essential to ensure their successful application in the domain of hydrology.



Chapter II



Research Gaps and Objectives

2.1. Unaddressed questions in rainfall-runoff modeling by DLs

Over decades, hydrologists have debated the most critical factors limiting model performance, focusing on model structure, parameterization, and data quality (Beven, 2002; Hrachowitz & Clark, 2017; Refsgaard et al., 2022). Shen and Lawson (2021) envision DLs as integral to hydrology, offering high precision and efficiency. Despite advancements, as whatever model, DL models face similar challenges.

Recent advances in rainfall-runoff modeling have brought to light ongoing challenges, particularly in balancing model complexity, accuracy, and generalizability. Conceptual hydrological models are valued for their interpretability, which aids in understanding predictions. However, they often struggle with generalization and calibration issues. In contrast, deep learning (DL) models have shown significant potential in improving predictive accuracy and extracting complex latent features from Earth science datasets. Despite these advances two key gaps remain in hyperparameter optimization and regional applicability:

1. Shen & Lawson (2021) suggest integrating DLs with perceptual hydrological models for more comprehensive multi-physics modeling. This approach could enhance DLs' interpretability and applicability in hydrology. The need for DLs that incorporate physical processes and provide interpretable, non-lossy outcomes is crucial for scientific progress (Gharari et al., 2021). Karpatne et al. (2017) note that DL and process-based models represent extremes in knowledge discovery, and the integration of physical principles into DLs remains an open question (Jiang et al., 2020).
2. Beven (2020) raises critical questions about DLs, such as their ability to simulate poorly understood catchments and the extent to which DL outputs can inform process-based understandings. Integrating process information could enhance DL predictions, but challenges remain in leveraging local hydrological knowledge and addressing scale-dependent processes.

In general, advancing DLs in hydrology by integrating physics, enhancing interpretability, developing hybrid models, and improving uncertainty quantification is still questioned; as well as the importance of distributed DL models, transfer learning, and addressing non-stationarity in the face of climate change. However, DLs have yet to fully overcome challenges like ungauged catchments and regional hydrological predictions, where traditional models still provide crucial insights (Beven, 2020). Additionally, concerns about over-parameterization and the true intelligence of DL models persist, emphasizing the need for approaches that achieve accurate results with minimal complexity.

2.2. Hyperparameter optimization

All the gaps highlighted in 2.1, require of careful tuning of hyperparameters—such as network architecture, learning rates, and so on—to achieve optimal performance. Hyperparameter optimization is critical because it directly influences the model's ability to learn complex patterns, generalize to new data, and avoid issues like overfitting and underfitting (Russell & Norvig, 2020; Goodfellow et al., 2016).

There is a significant research challenge due to the lack of standardized guidelines for configuring and optimizing DL models for hydrological applications. As Arsenault et al. (2023) highlight, hyperparameter optimization remains a major challenge, with no clear protocols established for deep neural networks architecture in hydrology. This challenge can lead to issues like over-parameterization, which affects model performance and interpretability. As always, over-parameterized model structures or architectures may show high accuracy on training data but fail to generalize to new data, while overfitting reduces the model's ability to handle unseen events.

Arsenault et al. (2023) highlight the persistent challenge of hyperparameter optimization in neural network architecture design for hydrological applications. The lack of clear guidelines for configuring AIs has led to a diversity of approaches, fostering innovation but also criticism due to the absence of standardized protocols. Common methods, such as manual tuning through trial and error, focusing on limited number of hyperparameters by grid search, or adopting configurations from previous studies, are used but often lead to suboptimal models due to the limited exploration of the hyperparameter space. Moreover, these approaches can introduce cognitive biases, as human-defined structural inputs by un-tuned hyperparameters may not fully capture the complexity of the underlying processes (Russell & Norvig, 2020; Goodfellow et al., 2016).

Research in AI has shown that random search method, as proposed by Bergstra & Bengio (2012), can offer a more efficient and thorough exploration of the hyperparameter space compared to grid search. Random search allocates resources more effectively by focusing on more influential hyperparameters, leading to better-performing models with fewer computational resources. However, in the field of hydrology, particularly in rainfall-runoff modeling using LSTMs, systematic hyperparameter optimization, especially using random search, has been largely overlooked, probably, due to computational constraints.

In regional studies using the CAMELS-US dataset, seminal works often manually tuned hyperparameters like the number of LSTM layers, and hidden size based on prior experience or employed grid search focusing on limited number of hyperparameters (Kratzert et al., 2024). Although these configurations provide acceptable results, they are not optimized for each catchment, leaving room for a catchment-wise hyperparameter optimization improvement. Moreover, some studies have adopted these configurations without further tuning, potentially missing opportunities for enhancing model accuracy (e.g., Liu et al., 2024).

As an example, the length of the input sequence in LSTM networks, often set to 365 days to capture a "full annual water cycle", has not been systematically tuned; however, evidence

exist on suggesting its hydrological significance (Kratzert et al., 2019; Hashemi et al., 2022). This reliance on default settings underscores the need for more comprehensive systematic hyperparameter optimization strategies in hydrological DL modeling.

Overall, there is a critical need for more systematic and thorough exploration of the hyperparameters space, possibly through random search method, to achieve higher accuracy and better generalization in regional hydrology. Advances in computational power, particularly with modern GPUs and parallel computing, now make it feasible to undertake this kind of extensive optimization methods, which could lead to significant improvements in model performance.

2.3. Uniqueness of the Place Paradigm in Regional hydrological DLs

In regional rainfall-runoff modeling, DLs like LSTMs have demonstrated strong generalization capabilities; however, their performance can vary significantly across different catchments. This variability presents a challenge in applied hydrology, where poor performance in specific locations can undermine the model's overall reliability and trustworthiness for critical applications such as flood resilience and mitigation (Beven, 2020; Prieto et al., 2020). To address this, it is essential to evaluate regional DL models not only on aggregate regional metrics but also on their performance at the catchment level, embracing the "uniqueness of the place" paradigm.

Many studies in regional hydrological modeling focus on overall performance, often reporting median or average performance metrics across all catchments (Liu et al., 2024; Kratzert et al., 2024; Gauch et al., 2021). While this approach provides a general view, it can mask significant variability at the catchment level, potentially leading to misrepresentations of a regional model's true performance. Valiela (2000) points out a drawback of regional comparative studies: "the conclusion being valid only for the dataset on aggregate." In comparative hydrology, poor performance of a regional DL model in specific catchments should not be dismissed as outliers, but rather investigated to understand the unique hydro-geological characteristics that may be influencing these results (Beven, 2020).

In traditional hydrology it might be accepted that a regional model does not perform well in certain locations due to several logical factors like snow presence, reservoirs, or underground flows; traditional hydrology generally accepts that there is no "one size fits all" model (See: Fenicia et al., 2008). However, the advanced capabilities of modern DL models, which can uncover latent features from large datasets, challenges this idea for modern AI/DLs. The persistence of poor performance in certain places raises the question of why intelligently trained regional DL models, such as LSTMs, still struggle in some specific catchments (Beven, 2020).

The goal in developing regional DL models should be to optimize performance across all catchments as much as possible, rather than relying solely on overall aggregated regional metrics. This is a multi-objective problem and needs a broader perspective, specifically when different optimized configurations of regional LSTM networks can exhibit unique strengths

and weaknesses specific to each catchment, influenced by how they handle input data and learn during training (For more in-depth understanding refer to the ideas of “Multi-Objective Recommendations” by Zheng & Wang, 2021, or “Model soups” for deep learning by Wortsman et al., 2022). For instance, one optimized configuration might excel in detecting anomalies like data deficiencies or anthropogenic influences in certain catchments, while another optimized network might perform better across the broader perspective in more catchments.

Therefore, effective regional hydrological DL modeling requires careful consideration of individual catchment performances to ensure their reliability. The selection of the best-performing configurations for regional DLs should be based on their ability to perform well across diverse catchments as much as possible having a greedy appetite, rather than solely on showing a highly aggregated regional metric for all catchments. This approach aligns with the “uniqueness of the place” paradigm, ensuring that the model is accurate and trustworthy in practical hydrological applications.

2.4. Some more key challenges in the domain

Data processing is another complication, with mixed results such as preprocessing techniques, noise and trend removal (Jain and Kumar, 2007). Effective input feature selection and data division are crucial in AI/MLs, yet more systematic research is needed in these areas (Abrahart et al., 2012). The preprocessing of input data and the selection of relevant features are critical to model performance, as per definition, data-driven models learn from data. The research presents mixed results regarding data preprocessing methods, indicating the need for a systematic investigation into these aspects. Effective DL model implementation should consider input preprocessing, feature selection, and data splitting, all of which can significantly impact the overall effectiveness of DL models in hydrological rainfall-runoff modeling.

The physical interpretability of DLs remains contentious. While some research shows that hidden units can map to hydrological processes, skepticism persists until stronger evidence of physical rationality is presented (Lees et al., 2021). Sensitivity analysis and constraints based on physical plausibility offer potential avenues for enhancing model credibility.

Finally, Abrahart et al. (2012) suggested three key research directions for advancing ANNs that can be generalized to new generations of DNNs: (i) discovering optimal structures and improved training regimes, (ii) developing systematic methods for preprocessing and data division, and (iii) hybridizing DLs with physical models to improve overall performance.

2.5. Research Goal and Objectives of this thesis

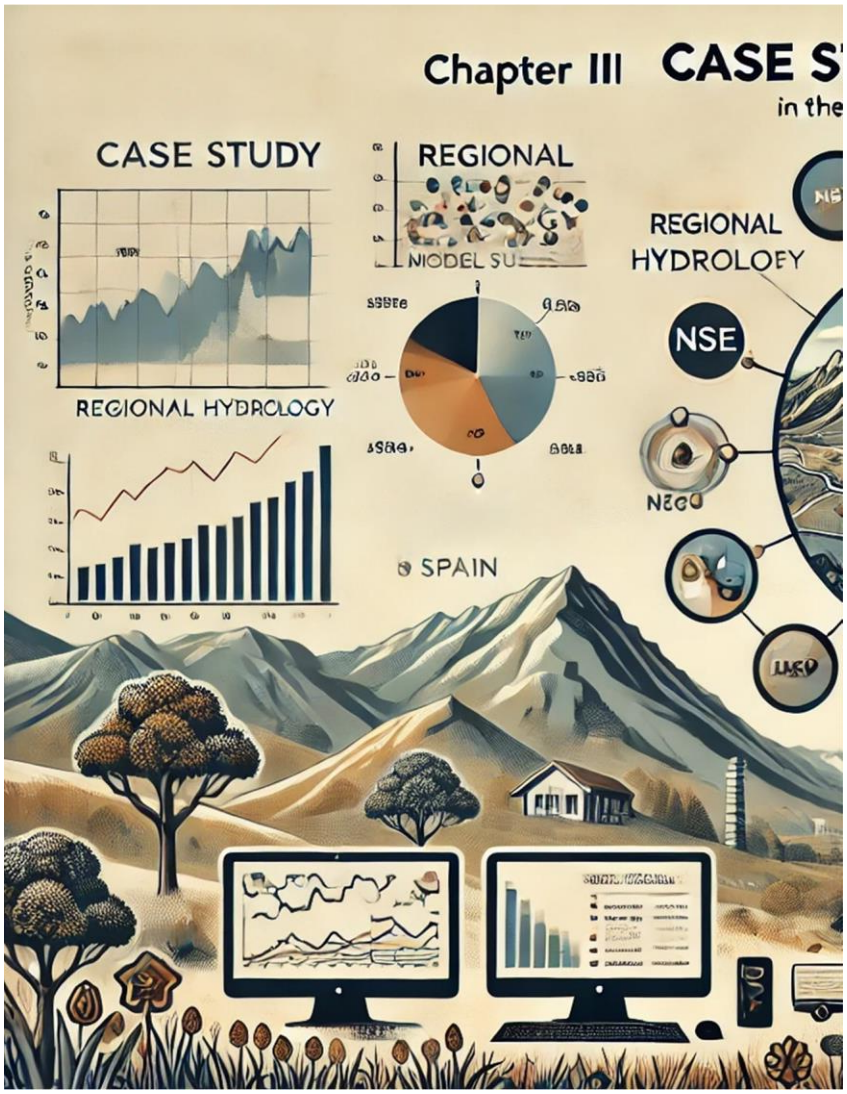
Given the challenges in hydrological modeling, this PhD research hypothesizes that a systematic, regional approach to optimizing hyperparameters for LSTM networks—leveraging random search—can enhance the accuracy and reliability of hydrological predictions. The overarching goal is to establish a systematic protocol for optimizing regional LSTM networks, achieving highly accurate predictions across diverse catchments.

To achieve this goal, this research addresses the following aims:

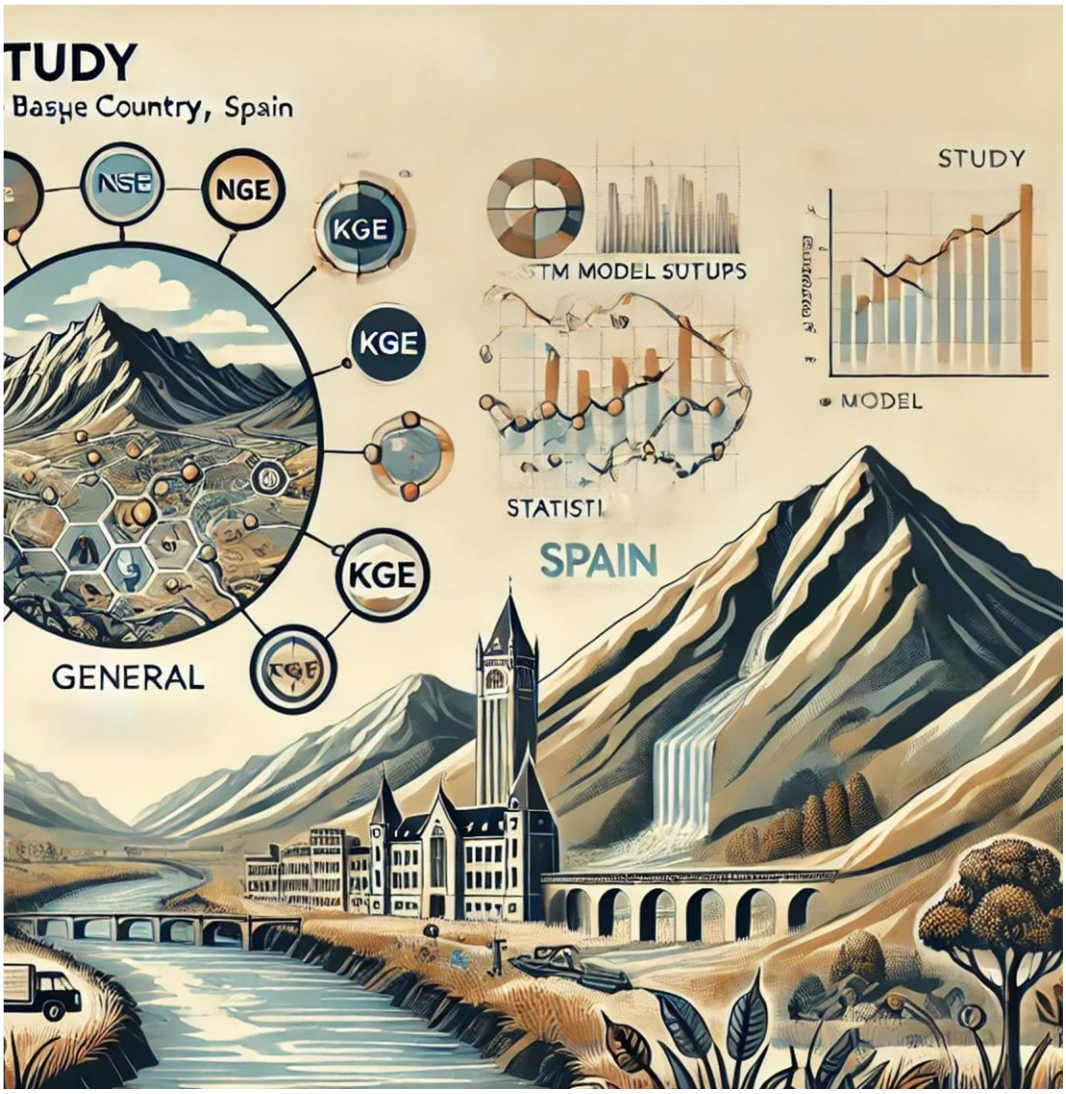
1. **Comprehensive Hyperparameter Optimization through Random Search (Chapter 4)**
   - **Evaluation of Random Search Effectiveness**: Analyze the capacity of systematic random search in identifying optimal configurations for improved model accuracy and generalization across catchments.
   - **Impact of Iteration Count on Model Accuracy**: Examine how increased search iterations influence final model accuracy, balancing computational cost with the accuracy gains achieved.
2. **Performance Assessment of Regionally Optimized Configurations**
   - **To Optimize LSTM Network Performance in Basque Catchments:** Conduct regional hyperparameter optimization across 40 catchments in the Basque Country, aiming to improve hourly streamflow and water level predictions.
   - **Effectiveness of Regional Networks' Configurations:** Assess the need for precise hyperparameter tuning by determining whether variations across configurations significantly affect prediction accuracy across different catchments.
3. **Mitigating Cognitive Bias in configuration selection with Ensemble Learning (Chapter 5)**
   - **Ensemble Method Implementation:** Introduce and evaluate multiple ensembles learning methods to diversify hyperparameter selection and reduce subjective biases as much as possible in configuration choice.
   - **Benchmarking Ensemble vs. Single Configuration Performance:** Compare ensemble methods with the highest-performing single configurations to identify approaches that offer high-accuracy predictions, especially for catchments with unique behaviors.
   - **Catchment-Specific Ensemble Approaches:** Evaluate a catchment-wise ensemble approach that respects the "uniqueness of place" paradigm and contrasts its performance with other ensemble techniques.
4. **Analysis of Hyperparameter Influence on Model Performance (Chapter 6)**
   - **Quantifying Hyperparameter Importance:** Employ techniques like Random Forest Regression and Principal component analysis to rank the impact of different tuned unique hyperparameters, assessing their influence on regional LSTM performance.
   - **Variation in Hyperparameter Impact by Catchment Characteristics:** Investigate how hyperparameters influence model accuracy across different catchment types, exploring possible hydrological significance that emerges.

5. **Understanding the Role of Catchment Attributes in Model Performance (Chapter 7)**
    - **Correlation of Catchment Attributes and LSTM Accuracy:** Assess how catchment-specific physical and hydrological characteristics influence the performance of hyperparameter-optimized LSTMs, identifying attribute-based performance patterns.
    - **Implicit Learning of Catchment-Specific Features:** Explore the extent to which regional LSTMs trained solely on hydrometeorological data can learn catchment-specific patterns without direct access to catchment attributes.
    - **Hydrological Insights from Optimized LSTMs:** Investigate how optimized regional models might reflect underlying hydrological processes, enhancing their interpretability and potential application in water management.

By addressing these aims, this research intentions to elevate the accuracy of hourly streamflow and water level predictions across the Basque Country's catchments - a flashy, humid region in north of Spain along the European Atlantic Ocean, integrating deep learning rigor with hydrological insights. This study contributes to advancing AI applications in hydrology, bridging predictive modeling with domain-specific understanding for enhanced water management practices.


Chapter III
CASE S
in the
CASE STUDY
REGIONAL
REGIONAL HYDROLOGY
NSE
SPAIN

Chapter III



Case Study and Dataset

General Model setups

Evaluation approaches

3.1. Case Study: Basque Country Hydrological System

This study focuses on the hydrological systems of the Basque Country as its case study; a region located in the north of Spain along the European Atlantic coast. This area is characterized by its humid climatology and abundant water resources, making it a critical zone for water resources management and flood prediction. The Basque Country case studies span an area of approximately 4,494 km² and encompasses a wide range of catchment sizes, from as small as 4 km² to as large as 1,000 km² (Figure 6).

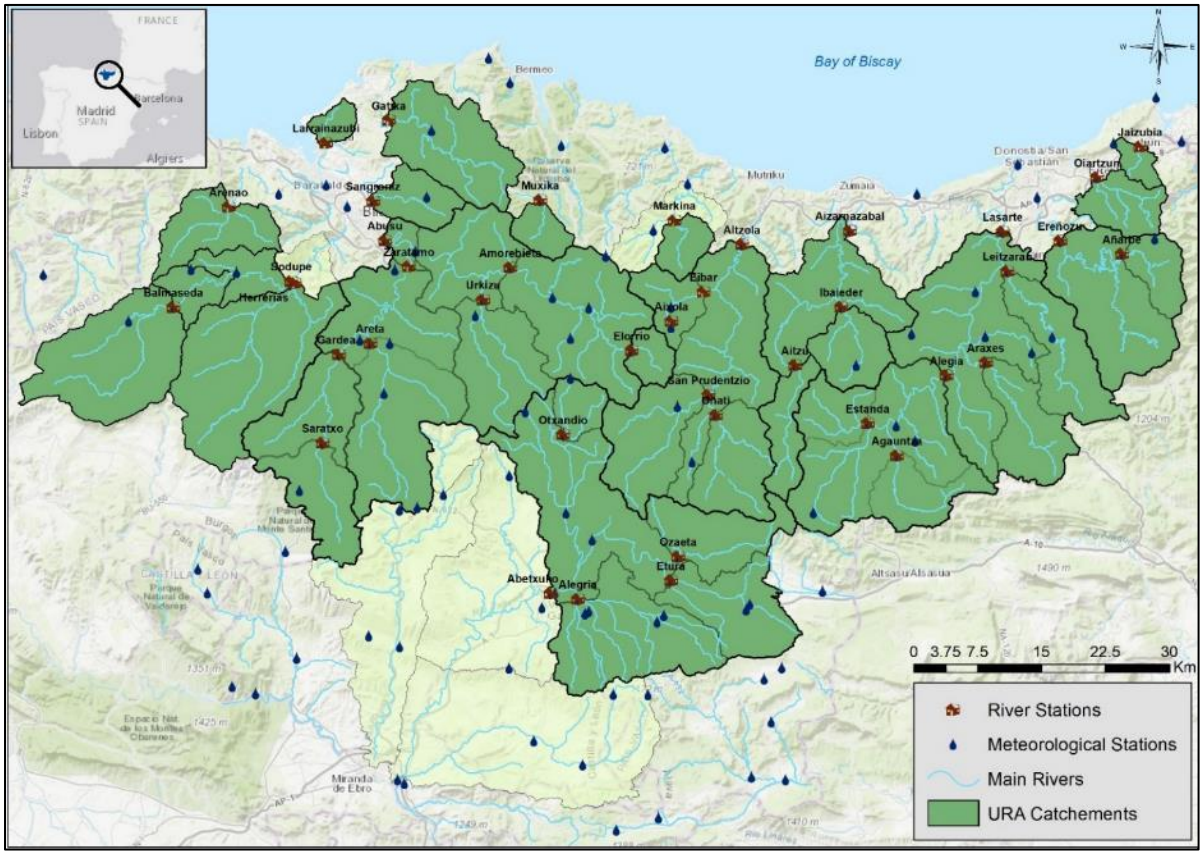


Figure 6. Study area: including 40 catchments of the Basque Country in north of Spain

The Basque Country's catchments (URA) are situated between the Cantabrian Mountains in the northwest, which reach elevations up to 1,300 meters, and the Atlantic Ocean to the north. The region's landscape is predominantly covered by grasslands and evergreen forests, benefiting from the warming influence of the Gulf Stream. This results in a humid and temperate climate, with mean annual temperatures varying between 9°C in the mountainous areas and 15°C in the lower regions. Annual precipitation in the region ranges from 1,200 mm to 1,600 mm, largely driven by the advection of North Atlantic fronts. Table 01 demonstrates a concise review of the most crucial geo-hydrological characters of 40 URA catchments participated in our research.

Given these climatic and geographic conditions, the Basque Country is prone to flashy and intense rainfall events, leading to rapid runoff and a heightened risk of flash floods. The

hydrological characteristics of the region, coupled with its susceptibility to flooding, underscore the importance of reliable, accurate and precise hydrological models for effective water resources management and planning, as well as flood prediction and mitigation.

The Basque Water Agency (URA), a regional government entity, plays a crucial role in managing water policies and resources in this territory. URA has compiled a comprehensive and high-quality dataset of hydro-meteorological time series, recorded at approximately 100 monitoring stations distributed across the tiny region. These stations collect data on various parameters, including rainfall and water levels, at a high temporal resolution of 10 minutes. This dataset serves as the foundation for developing advanced hydrological models, particularly those based on deep learning techniques, which learn from data, to enhance flood forecasting and water management strategies in the region.

Table 1. A brief summary of 40 URA catchments' attributes in the case study.

Row	Station Name	Data Start	Data End	Area Km2	Annual PREC mm	Min Temp ºC	Max Temp ºC	Days with Negative Temp	RC	Aridity Index (AI)	Flow mm	Annual PET mm	GRAD %	Elevation		Land Use									Soil					COND	PERM	HARD
														MIN Hight m	MAX Hight m	UHD %	AGR %	PAS %	BLF %	CNF %	PLT %	SSH %	WAE %	DEN %	CALC %	CONG %	SDIM %	VLC %	WATR %	(1-5) class		
1	Abetxuko	9/3/2010	7/14/2019	679.7	1175	-9.91	39.29	225	0.33	1.31	241.31	968.09	25%	503	1549	3%	18%	13%	36%	7%	8%	14%	1%	1%	88%	0%	10%	0%	2%	4.6	3.0	2.8
2	Abusu	10/1/2000	9/30/2021	1003.1	1176	-4.26	38.14	41	0.61	0.82	718.96	956.83	40%	13	1377	2%	0%	18%	18%	24%	13%	23%	0%	3%	97%	2%	0%	0%	1%	4.9	3.1	3.0
3	Agauntza	10/1/2000	9/30/2021	69.5	1537	-5.80	38.78	47	0.72	0.77	874.63	937.01	47%	184	1412	0%	0%	25%	50%	4%	13%	8%	0%	1%	100%	0%	0%	0%	0%	5.0	3.1	3.0
4	Aitzu	10/1/2000	9/30/2021	56.8	1486	-5.90	37.57	51	0.63	0.67	886.67	951.77	45%	312	1431	1%	0%	21%	22%	31%	17%	5%	0%	3%	100%	0%	0%	0%	0%	5.0	3.1	3.0
5	Aixola	7/1/1987	9/30/2019	4.8	1421	-17.30	38.13	40	0.44	0.65	619.60	929.90	44%	340	750	0%	0%	5%	5%	80%	10%	0%	0%	0%	100%	0%	0%	0%	0%	4.6	3.0	3.0
6	Aizarnazabal	10/1/2000	9/30/2021	273.5	1653	-5.90	37.57	51	0.63	0.67	895.56	950.02	47%	20	1074	1%	1%	32%	14%	43%	4%	5%	0%	0%	97%	3%	0%	0%	0%	5.0	3.1	3.0
7	Alegia	10/1/2000	10/2/2020	329.6	1416	-6.03	39.24	61	0.61	0.76	770.05	959.35	45%	92	1549	1%	0%	27%	30%	14%	18%	10%	0%	1%	100%	0%	0%	0%	0%	5.0	3.1	3.0
8	Alegria	4/30/2010	2/7/2021	185.1	1074	-10.74	39.45	238	0.31	1.18	231.54	868.46	22%	508	1099	4%	36%	3%	42%	0%	5%	9%	1%	0%	81%	0%	19%	0%	0%	4.3	3.0	2.6
9	Altzola	10/1/2000	9/30/2021	460.3	1380	-5.34	39.28	44	0.60	0.74	775.15	955.84	44%	12	1363	2%	0%	26%	20%	25%	15%	11%	0%	1%	99%	0%	0%	0%	0%	5.0	3.1	3.0
10	Amorebieta	10/1/2000	9/30/2021	233.4	1302	-5.72	40.12	51	0.71	0.73	912.70	950.78	43%	65	1330	2%	1%	24%	10%	27%	11%	17%	0%	8%	90%	10%	0%	0%	0%	4.7	2.9	2.9
11	Anarbe	10/1/2000	9/30/2021	47.1	2031	-7.87	39.63	74	0.78	0.42	1744.02	951.18	54%	182	1052	0%	0%	18%	62%	5%	12%	2%	0%	0%	0%	87%	0%	12%	0%	1.9	2.9	4.0
12	Araxes	1/5/2011	9/30/2021	93.1	1616	-7.00	38.69	69	0.75	0.50	1428.12	950.78	49%	119	1429	1%	0%	30%	41%	13%	11%	3%	0%	1%	100%	0%	0%	0%	0%	5.0	3.3	3.0
13	Arenao	5/26/2005	9/1/2020	85.7	1200	-3.74	39.31	14	0.58	0.87	616.01	929.71	38%	45	821	1%	0%	28%	16%	20%	12%	20%	0%	4%	100%	0%	0%	0%	0%	5.0	3.1	3.0
14	Areta	4/1/2013	9/30/2021	190.1	1149	-5.14	37.88	48	0.58	0.89	625.02	953.53	36%	118	1305	1%	0%	14%	37%	15%	9%	23%	0%	1%	100%	0%	0%	0%	0%	5.0	3.0	3.0
15	Balmaseda	10/1/2000	9/30/2021	194.9	1064	-4.07	40.60	18	0.67	0.92	696.91	953.53	34%	172	1331	1%	3%	33%	36%	10%	7%	8%	0%	1%	90%	10%	0%	0%	0%	4.9	2.9	2.9
16	Eibar	12/1/2013	9/30/2021	50.0	1451	-5.04	38.35	25	0.54	0.66	738.08	903.95	44%	94	812	5%	0%	34%	5%	41%	9%	5%	0%	0%	100%	0%	0%	0%	0%	4.6	3.0	3.0
17	Elorrio	1/1/2001	9/30/2021	29.6	1368	-5.02	42.89	18	0.50	0.69	656.77	909.78	40%	167	1116	2%	1%	21%	4%	31%	21%	20%	0%	1%	100%	0%	0%	0%	0%	5.0	3.0	3.0
18	Erenozu	10/1/2000	9/30/2021	215.8	1902	-6.43	38.56	36	0.72	0.45	1444.70	909.78	55%	23	1142	0%	0%	20%	41%	15%	15%	9%	0%	0%	2%	95%	0%	2%	1%	2.3	2.8	3.7
19	Estanda	10/1/2000	9/30/2021	54.6	1352	-5.80	38.78	47	0.48	0.75	587.45	909.78	43%	164	966	2%	0%	26%	15%	32%	13%	11%	0%	0%	100%	0%	0%	0%	0%	5.0	3.0	3.0
20	Etura	12/1/2012	9/30/2021	113.9	1201	-9.47	35.91	135	0.63	1.02	571.04	924.84	20%	548	1150	3%	37%	14%	39%	0%	1%	4%	0%	0%	90%	3%	7%	0%	0%	4.5	3.0	2.8
21	Gardea	10/1/2000	9/30/2021	192.2	1025	-5.55	41.73	61	0.42	0.94	418.20	939.40	34%	134	1183	1%	1%	26%	27%	15%	8%	21%	0%	1%	98%	2%	0%	0%	0%	5.0	3.0	3.0
22	Gatika	12/1/2013	9/30/2021	143.4	1308	-2.88	37.71	11	0.61	0.77	747.34	940.40	28%	5	687	6%	1%	38%	16%	20%	5%	13%	0%	1%	96%	4%	0%	0%	0%	4.8	3.0	3.0
23	Herrerias	10/1/2000	9/30/2021	254.0	1103	-4.07	40.60	18	0.41	0.84	450.38	926.70	39%	50	1184	0%	0%	26%	14%	27%	13%	19%	0%	1%	100%	0%	0%	0%	0%	5.0	3.0	3.0
24	Ibai Eder	10/1/2000	9/30/2021	65.4	1644	-5.90	37.57	51	0.51	0.62	738.76	895.34	49%	87	971	0%	0%	26%	27%	36%	3%	7%	0%	0%	100%	0%	0%	0%	0%	5.0	3.0	3.0
25	Jaizubia	4/1/2013	9/30/2021	18.3	1988	-3.72	41.03	7	0.82	0.60	1244.23	902.12	36%	3	544	10%	0%	20%	33%	1%	9%	26%	0%	0%	52%	32%	0%	16%	0%	3.3	2.8	3.4
26	Larrainazubi	4/1/2013	9/30/2021	19.1	1363	-3.55	39.57	4	0.74	1.01	702.34	953.67	23%	5	254	20%	0%	35%	29%	0%	0%	15%	0%	0%	12%	86%	2%	0%	0%	3.3	2.1	2.1
27	Lasarte	10/1/2000	9/30/2021	791.3	1601	-5.77	38.84	50	0.69	0.62	971.40	874.21	45%	18	1549	1%	0%	28%	28%	21%	11%	9%	0%	1%	83%	17%	0%	0%	0%	4.7	3.1	3.0
28	Leitzaran	10/1/2000	9/30/2021	114.2	1943	-7.34	38.69	69	0.75	0.47	1490.37	938.93	49%	49	1200	1%	0%	21%	22%	42%	4%	11%	0%	0%	17%	83%	0%	0%	0%	3.2	2.7	3.1
29	Markina	12/1/2013	9/30/2021	34.0	1469	-5.46	37.80	28	0.61	0.71	823.09	955.75	45%	76	791	1%	0%	35%	3%	48%	5%	8%	0%	0%	100%	0%	0%	0%	0%	4.8	3.0	3.0
30	Muxika	10/1/2000	9/30/2021	31.4	1351	-3.56	39.10	37	0.47	0.69	650.36	945.24	36%	11	625	0%	1%	25%	5%	43%	11%	14%	0%	0%	97%	3%	0%	0%	0%	4.2	3.0	3.0
31	Oiartzun	10/1/2000	9/30/2020	55.9	2073	-5.49	38.57	27	0.81	0.46	1517.96	860.62	49%	6	831	1%	0%	19%	25%	9%	15%	26%	0%	4%	18%	38%	0%	44%	0%	2.1	2.5	3.8
32	Onati	10/1/2000	9/30/2021	99.1	1442	-5.34	39.28	44	0.63	0.63	997.41	1005.28	47%	193	1362	1%	0%	21%	30%	19%	15%	12%	0%	2%	100%	0%	0%	0%	0%	5.0	3.0	3.0
33	Otxandio	1/1/2003	9/30/2021	35.5	1361	-9.23	38.06	121	0.70	0.71	956.55	962.09	34%	549	1330	1%	0%	11%	40%	29%	7%	8%	0%	5%	100%	0%	0%	0%	0%	5.0	3.0	3.0
34	Ozaeta	4/1/2014	9/30/2021	97.5	1323	-8.58	36.86	90	0.58	1.04	540.12	961.37	29%	549	1547	3%	4%	24%	32%	0%	10%	25%	0%	1%	86%	0%	14%	0%	0%	4.6	3.0	2.7
35	San Prudentzio	10/1/2000	9/30/2021	122.2	1206	-5.34	39.28	44	0.55	0.77	712.99	992.86	42%	171	1146	2%	0%	19%	23%	12%	26%	15%	0%	3%	99%	0%	0%	0%	1%	5.0	3.2	3.0
36	Sangroniz	6/6/2005	9/30/2021	50.8	1281	-4.56	39.87	27	0.50	0.76	598.78	922.30	28%	5	475	11%	2%	26%	18%	16%	6%	19%	0%	1%	100%	0%	0%	0%	0%	4.6	3.0	3.0
37	Saratxo	10/1/2000	9/30/2021	91.2	993	-5.07	41.45	37	0.46	0.94	440.65	910.58	35%	225	1140	1%	1%	27%	40%	5%	1%	23%	0%	2%	94%	6%	0%	0%	0%	4.9	2.9	2.9
38	Sodupe	2/22/2001	8/29/2020	275.8	1118	-4.07	40.60	18	0.68	0.87	710.17	916.49	38%	49	717	1%	0%	27%	10%	28%	15%	19%	0%	0%	99%	1%	0%	0%	0%	5.0	3.0	3.0
39	Urkizu	10/1/2000	9/30/2021	127.1	1178	-5.16	41.29	39	1.02	0.71	1311.44	912.66	42%	68	1377	1%	0%	16%	13%	31%	12%	23%	0%	4%	100%	0%	0%	0%	0%	5.0	3.2	3.0
40	Zaratamo	1/1/2003	9/30/2021	512.3	1107	-4.78	39.28	47	0.57	0.88	607.34	940.69	38%	41	1305	1%	0%	17%	26%	19%	11%	24%	0%	1%	98%	1%	0%	0%	1%	5.0	3.1	3.0

*** PREC: precipitation; Temp: Temperature; RC: Runoff Coefficient; PET: Potential Evapotranspiration; GRAD: Gradiant (Slope)**

*** Land Use Distribution: Urban (UHD), Agriculture (AGR), Pasture (PAS), Broadleaf Forest (BLF), Coniferous Forest (CNF), Plantation (PLT), Shrublands (SSH), Water bodies (WAE)**

*** Soil Composition: CALC: calcareous soils; CONG: conglomerate soils; SDIM: sedimentary soils; VLC: volcanic soils; WATR: wetlands and water associated ecosystems**

*** Soil Composition class: COND (soil conductivity), PERM (permeability), and HARD (hardness)**

3.2. Dataset: Hydro-Meteorological Time Series

The dataset used in this study comprises 21 years of hourly hydro-meteorological time series data from 40 catchments across the Basque Country. The data collection period spans from October 1, 2000, to September 30, 2021. This extensive temporal coverage, combined with the high spatial resolution of data collected from diverse catchment sizes, provides a robust foundation for developing and testing hydrological DL models.

The dataset includes hourly observations of key hydrological variables such as precipitation, temperature, potential evapotranspiration, streamflow, and water level, that will be used for modeling the region's hydrological dynamics. In particular, streamflow data is available for all 40 catchments, while water level data is accessible for 27 catchments in our dataset. Given the region's susceptibility to flash floods, the accurate prediction of streamflow and water levels is critical, making this dataset invaluable for developing predictive regional DL models.

3.3. Data Splitting for Model Training and Evaluation

To ensure rigorous model development and evaluation, the dataset was partitioned into distinct subsets: a training-and-validation set and a test set. The training-and-validation set covers the period from October 1, 2000, to September 30, 2015, with the initial five years (October 1, 2000, to September 30, 2005) reserved for validation purposes during hyperparameter optimization step. The remaining period (October 1, 2015, to September 30, 2021) serves as the test set, which was withheld from all optimized networks during both hyperparameter optimization and the training phases to provide an unbiased evaluation of finally optimized networks' performance.

All Long Short-Term Memory (LSTM) networks developed in this study were regional models, trained using aggregated data from all 40 catchments. This regional approach ensures that the models capture the diverse hydrological behaviors across the catchments, leading to a more comprehensive understanding of the regional hydrological dynamics and aims to develop a one-size-fits-all regional LSTM model.

Despite efforts to maintain uniform training-and-validation periods across all catchments, data availability issues necessitated adjustments. Some catchments had incomplete data for the full training-and-validation period, leading to the inclusion of only 25 catchments in the validation phase during hyperparameter optimization. Nevertheless, the final optimized networks were tested on the entire region, encompassing all 40 catchments, providing a systematic evaluation of model performance.

3.4. General Model Architecture and Setups

In this study, we employed the Multi-Timescale Long Short-Term Memory (MTS-LSTM) architecture, a state-of-the-art neural network designed for hydrological predictions, particularly at hourly intervals. The MTS-LSTM model, as developed by Gauch et al. (2021) and accessible via the NeuralHydrology Python library (Kratzert et al., 2022), represents a significant advancement in the field of DL for hydrology. This architecture addresses the challenge of high computational costs associated with fine temporal resolution by parallelizing two LSTM networks, one for hourly predictions and the other for daily predictions. By doing so, MTS-LSTM efficiently captures both short-term dynamics and longer-term trends in hydrological data, making it particularly well-suited for regions with complex and variable hydrological patterns, such as the Basque Country.

The MTS-LSTM network was meticulously trained and validated using comprehensive hydro-meteorological data collected from 40 catchments within the Basque Country. The training process involved hyperparameter optimization, followed by the selection of the best-performing configurations, which were then retrained on 10 different random seeds and rigorously evaluated. The overall flow of this Ph.D. research—from data preprocessing to model evaluation—is outlined in the flowchart in Figure 7. This figure presents the overarching approach used throughout the thesis. Each chapter, however, delves into specific aspects of this general approach, developing its own detailed methods. While Chapters 4 through 7 follow the same core approach, they each implement unique, chapter-specific methodologies tailored to their respective objectives and analyses.

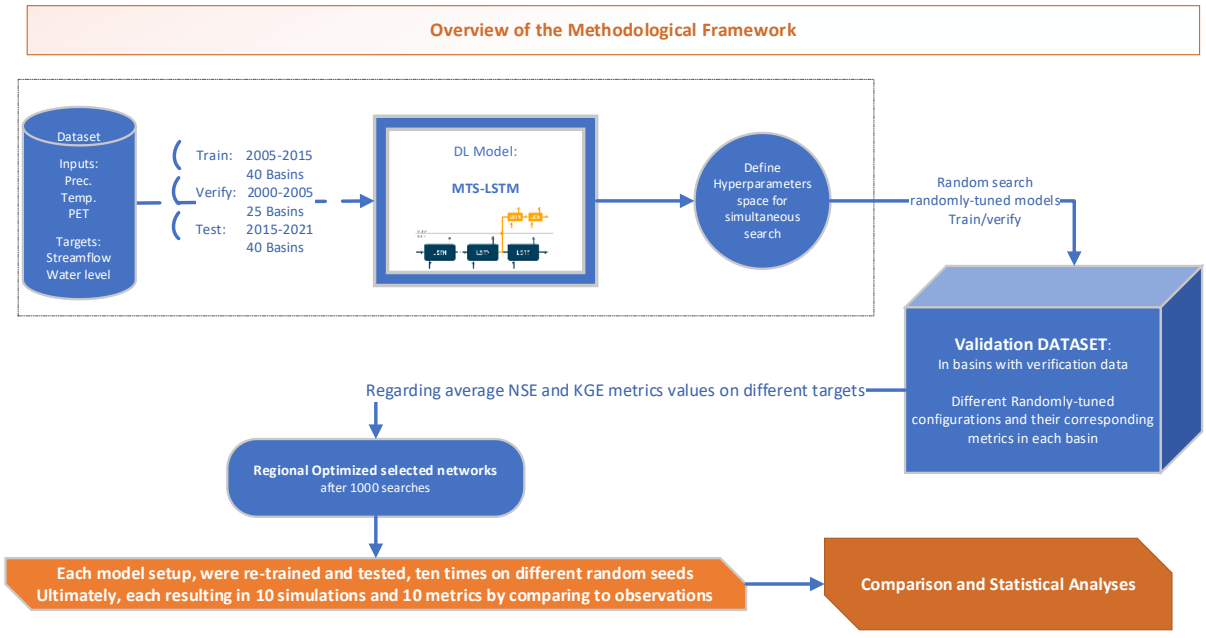


Figure 7. Overall Methodological Strategy developed for this research

This model architecture allows for a detailed and nuanced understanding of the hydrological processes at play, providing critical insights into both short-term and long-term water resource dynamics. The successful application of the MTS-LSTM model in this study underscores its potential as a systematic tool for enhancing flood prediction and water resource management in regions characterized by complex hydrological behavior.

3.4.1. Inputs and Targets

For the MTS-LSTM model implementation, we utilized a set of carefully selected meteorological inputs and hydrological targets to accurately capture the dynamics of the Basque Country's hydrological systems. The input data comprised hourly average precipitation, temperature, and potential evapotranspiration (PET) values, the latter calculated using the Hargreaves and Allen's (2003) equation. These three variables were chosen due to their critical role in driving the region's hydrological processes, particularly under the humid and temperate climate conditions prevalent in the study area.

The choice of inputs reflects the need to model not only the immediate impacts of precipitation but also the underlying energy dynamics, as represented by temperature and PET, which influence evapotranspiration rates and, consequently, water availability in the catchments, regarding generally accepted hydrological definitions. This comprehensive input set allows the MTS-LSTM model to effectively simulate the complex interactions between meteorological conditions and hydrological responses, which are essential for accurate streamflow and water level predictions.

The targets for the model included hourly streamflow and water level measurements at the outlets of all 40 catchments. These targets were chosen for their direct relevance to flood forecasting and water resource management. While streamflow data was available for all 40 catchments, water level data was available for only 27 catchments. Despite this limitation, the inclusion of water level as a target provided valuable additional information during the training phase, offering the model insights into dynamic water storage and the behavior of the hydrological system under varying conditions.

The selected inputs and targets, combined with the advanced capabilities of the MTS-LSTM architecture, enabled the development of a highly accurate and reliable model for hydrological predictions.

3.5. Post-Random Search Validation DATASET

In this research, we applied 1000 random search in an extensive hyperparameter space to optimize regional LSTM networks. The "Post-Random Search Validation DATASET" (with capital letters) includes validation metrics for 25 out of the 40 catchments, derived from 1,000 randomly-tuned configurations after an exhaustive random search in the hyperparameters space. The final dataset comprises 594 successful experiments out of the 1,000 configurations that completed both training and validation, along with their corresponding performance

metrics in the 25 catchments with validation data. This approach enabled a reliable comparison of regionalization strategies and the accuracy of the final optimized LSTM networks, ensuring that the models developed are both robust and reliable for practical applications in the Basque Country's hydrological system.

3.6. Performance Evaluation Methods

3.6.1. Evaluating Accuracy

The performance of the optimized LSTM networks in this research was rigorously evaluated against observed data from the test set. To assess the accuracy of the model predictions, we employed several performance metrics, including Nash-Sutcliffe Efficiency (NSE) (Nash & Sutcliffe, 1970), Kling-Gupta Efficiency (KGE) (Gupta et al., 2009), Mean Squared Error (MSE) (Legates and McCabe, 1999; Makridakis et al., 1993), Root Mean Squared Error (RMSE) (Willmott and Matsuura, 2006), Alpha-Nash-Sutcliffe Efficiency (Alpha-NSE), Beta-Nash-Sutcliffe Efficiency (Beta-NSE), Beta-Kling-Gupta Efficiency (Beta-KGE) (Gupta et al., 2012), Pearson's Correlation Coefficient (Pearson-r), three common biases: High-segment volume (%BiasFHV), Low-segment volume (%BiasFLV), Mid-segment slope (%BiasFMS) (Yilmaz et al., 2008), and three other metrics presented by Kratzert et al. (2020): Mean difference in peak flow timing (Peak-Timing), Mean Absolute Percentage Error for peaks (MAPE_peak), and Fraction of Missed Peaks (missed_peaks). Appendix 02 - Common Hydrological Loss Functions, provides a detailed explanation of each of these evaluation metrics, highlighting its interpretation and significance from a hydrological perspective.

Each of these metrics is designed to capture different aspects of model performance, contributing to a robust assessment of the models' ability to accurately predict hydrological variables. This approach ensures a clear understanding of the metrics' implications and establishes a solid foundation for assessing and discussing the models' performance in subsequent chapters. Specifically, the accuracy of the most employed NSE and KGE metrics in assessing hydrological predictions was affirmed by Gauch et al. (2023). They established that these metrics are robust indicators of overall and high-flow hydrograph quality, although the efficacy in assessing low-flow quality may be limited. Additionally, the stated research highlights the alignment of the quantitative metrics with human preferences, as hundreds of participants tended to favor AI models based on both visual judgments and quantitative assessments.

3.6.2. Benchmarking

Benchmarking was conducted to evaluate and compare the performance of different ensemble methods from Chapter 5 against the best-performing regionally optimized networks presented in Chapter 4, with assessments made at both regional and catchment levels. This comprehensive evaluation primarily focused on various accuracy metrics to ensure a robust and reliable assessment of the proposed methods and optimized networks. By examining these metrics, we aimed to capture the multifaceted aspects of prediction accuracy, providing a thorough understanding of each method's strengths and limitations across diverse hydrological contexts.

3.6.3. Evaluating Computational Costs

In addition to evaluating conventional accuracy metrics, we conducted an analysis of the models' convergence speed and computational efficiency, which are critical factors for the practical deployment of hydrological DL models. Convergence speed refers to the rate at which a model reaches its optimal performance during training, a key determinant of the time and computational resources required for model development. Computational efficiency, on the other hand, encompasses the overall resource utilization, including memory and processing power, necessary to execute the model, particularly in real-time or large-scale applications. These factors are especially pertinent in the context of regional hydrological modeling, where the ability to rapidly and efficiently generate reliable predictions can significantly influence decision-making processes in water resources management and flood forecasting. Our evaluation provides a more holistic understanding of the model's applicability and readiness for operational use, ensuring that the selected configurations are not only accurate but also viable in terms of computational demands.

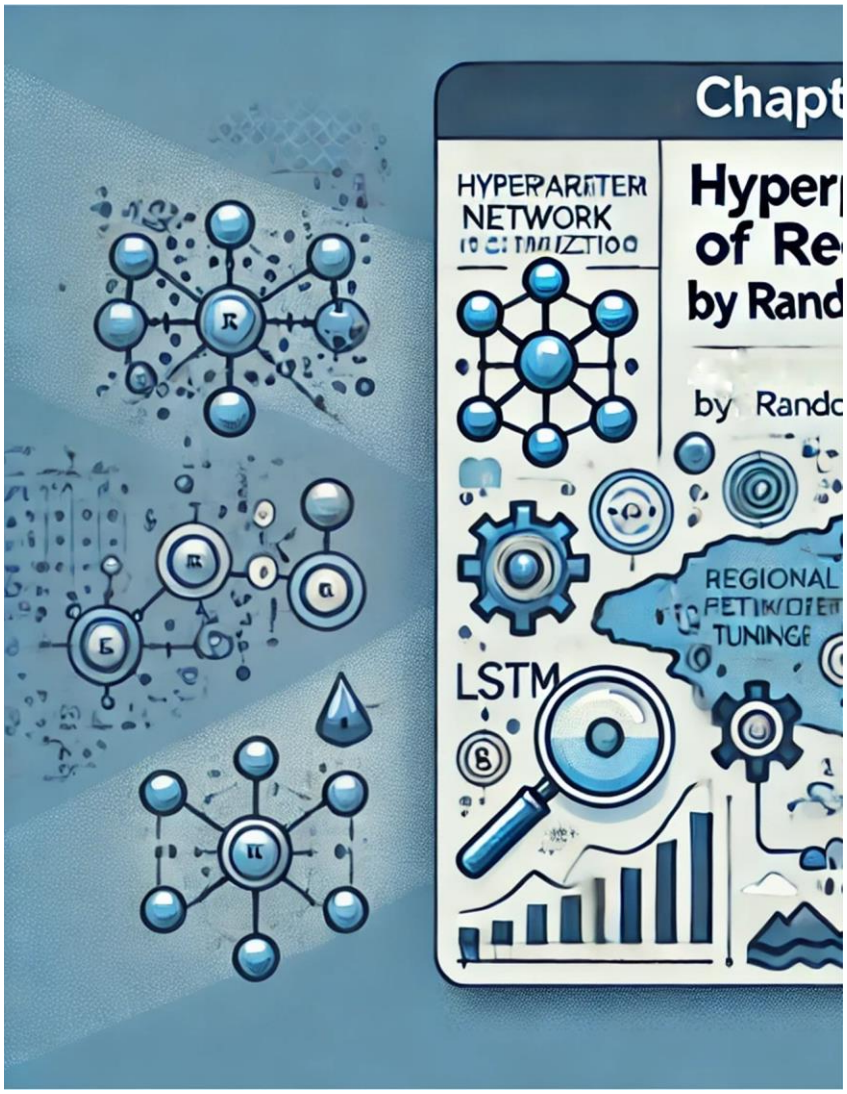
Regarding our research aims in Chapter 03, evaluating the impact of increasing the number of random searches on model performance was another crucial issue. We conducted a detailed statistical analysis on this subject in the aforementioned chapter. This analysis aimed to determine whether and how increasing the search iterations conclude in finding more accurate configurations?

3.6.4. Statistical Analyses to Study Significant Differences

We employed a classical criterion (P-value < 0.05) to reject the null hypothesis, which was posed as H0: there is no statistically significant difference in performance metrics between the different optimized configurations across different catchments. To rigorously compare different optimized models, we focused on the distribution and variability of ten individual performance metrics by each method across all catchments.

Three statistical tests were utilized to conduct this analysis: the Wilcoxon Signed-Rank Test (Wilcoxon, 1945), the Analysis of Variance (ANOVA) test (Moore, 2006), and the Mann–Whitney U test (Mann & Whitney, 1947). Each of these tests serves a specific purpose in the evaluation process. The Wilcoxon Signed-Rank Test compares paired performance metrics in each catchment to assess whether their differences are significantly different from zero, making it ideal for non-normally distributed data or small sample sizes. The ANOVA test analyzes differences among group means in a sample, helping to determine if significant differences exist in prediction outcomes among different hyperparameter configurations. The Mann-Whitney U test, a nonparametric method, evaluates differences between two independent groups, useful when normality and homogeneity of variance cannot be assumed.

By employing a range of statistical tests, we conducted a thorough evaluation of the learning capabilities of different regionally optimized LSTM networks and ensemble learning methods. This systematic approach allowed us to discern the significance of variations in hyperparameter configurations and their impact on prediction outcomes, offering deeper insights into the effectiveness of these regional models across diverse catchments. Moreover, the multi-faceted evaluation framework highlighted the true potential of ensemble learning methods, revealing their value in enhancing model robustness and accuracy, especially in regions with complex hydrological dynamics. The results underscore the importance of the proposed methods in this study for improving regional LSTM predictions, emphasizing their role in addressing the challenges of regional hydrological modeling.



Chapter IV

This chapter is an edited version of the Research paper "Hosseini, Farzad, Prieto, Cristina, & Álvarez, Cesar. (2024). Ensemble learning of catchment-wise optimized LSTMs enhances regional rainfall-runoff modeling: A case study from Basque Country, Spain. Journal of Hydrology, 132269. doi: 10.1016/j.jhydrol.2024.132269"

Hyperparameter Optimization

of Regional LSTMs by Random Search

4.1. Introduction

In hydrology, accurate hourly rainfall-runoff modeling is critical, especially for managing flashy catchments prone to rapid flooding, which pose significant risks to life and property. These predictions are essential for effective water resources management, flood risk mitigation, and supporting economic activities reliant on water availability (Refsgaard et al., 2022; Prieto et al., 2020; Hrachowitz & Clark, 2017). Long Short-Term Memory (LSTM) networks have shown considerable promise in enhancing the accuracy of regional rainfall-runoff modeling, particularly in handling temporal dependencies and capturing complex hydrological processes (Kratzert et al., 2024, 2018). However, optimizing their hyperparameters remains significant a challenge.

The rapid adoption of new DLs, including LSTMs, Transformers, Encoder-Decoder architectures, CNNs, and hybrid approaches, has led to a tendency to move quickly from one model to another. This often happens without fully optimizing and implementing the DL models on the datasets, raising a critical question: have we mastered the application of these sophisticated AI techniques in hydrology?

DLs involve numerous hyperparameters that significantly impact their capacity, learning dynamics, and performance. Effective implementation requires a nuanced understanding of these hyperparameters (Russell & Norvig, 2020; Goodfellow et al., 2016; Bergstra & Bengio, 2012). Precise tuning (hyperparameter optimization) is essential to avoid issues like overfitting, underfitting, and ensuring generalization. Without careful optimization, networks may retain traces of human bias, as structural inputs and certain hyperparameters are often set empirically (Kratzert et al., 2019, 2018; Shen, 2018).

Despite the success of LSTMs, research in hydrological modeling has not fully explored systematic hyperparameter optimization. Arsenault et al. (2023) identify neural network architecture design as an “unresolved problem” in hydrological DLs. Proper hyperparameter optimization, which involves adjusting parameters that define the architecture and learning dynamics of deep networks, is crucial for achieving optimal performance (Russell & Norvig, 2020; Goodfellow et al., 2016).

As stated earlier, systematic approaches to hyperparameter optimization, such as grid search, random search (Bergstra & Bengio, 2012), and Bayesian optimization (Snoek et al., 2012), have been developed to address these challenges. Bergstra & Bengio (2012) argue that not all hyperparameters carry equal significance in the optimization process. Grid search often allocates disproportionate resources to less influential dimensions, leading to inadequate coverage. In contrast, random search explores the hyperparameter space more thoroughly, achieving more accurate results with fewer computational resources.

In hydrology, many studies have relied on manual tuning (e.g., Donnelly et al., 2024a,b; Frame et al., 2022; Mai et al., 2022; Hoedt et al., 2021; Ma et al., 2021; Tsai et al., 2021; Rahmani et al., 2021; Feng et al., 2020; Kratzert et al., 2018; Fang et al., 2017). Limited studies have employed grid search on a small number of hyperparameters chosen by cognitive bias

(e.g., Ahmadi et al., 2024; Klotz et al., 2022; Gauch et al., 2021; Nearing et al., 2021; Kratzert et al., 2019), and even fewer have used Bayesian optimization (Mahdian et al., 2024). Furthermore, there is a tendency to adopt hyperparameter settings from previous studies without thorough validation or exploration of alternative configurations, particularly for similar datasets (e.g., Liu et al., 2024; Xie et al., 2021; Ouyang et al., 2021; Feng et al., 2020; Kratzert et al., 2019).

For example, in rainfall-runoff modeling using LSTMs, seminal comparative studies have focused predominantly on the CAMELS-US dataset (Addor et al., 2017), employing specific hyperparameter configurations proposed by pioneering researchers (Kratzert et al., 2024, 2019, 2018). Kratzert et al. (2018) manually tuned hyperparameters such as the number of LSTM layers, hidden size, dropout rate, and input sequence length. While their architecture proved effective, they acknowledged that more effective LSTM networks could be configured through comprehensive catchment-wise hyperparameter optimization. Additionally, a systematic sensitivity analysis of different hyperparameter combinations was not performed, leaving room for future research.

There are traces of cognitive bias in manually tuned networks in the literature. For instance, the length of the input sequence of LSTMs applied in hydrology is often set to 365 days (Kratzert et al., 2024; 2018), a value chosen to capture the dynamics of a full annual cycle. Although Kratzert et al. (2019) reduced this to 270 days following a limited systematic grid search tuning, they reverted to 365 days in later work without further tuning (Kratzert et al., 2024). However, systematic tuning of this hyperparameter has revealed its hydrological significance (Hashemi et al., 2022; Kratzert et al., 2019).

Overall, hyperparameter optimization for regional hydrological LSTM modeling has received limited attention. This raises concerns about whether the full capabilities of LSTMs have been realized, especially in the context of regional hydrological applications where variability across catchments demands precise model tuning. Existing approaches often overlook the complexity of optimizing multiple hyperparameters simultaneously, leading to suboptimal configurations and limited exploration of the hyperparameters space. As far as we know, no study has conducted a comprehensive, simultaneous systematic hyperparameter optimization exploring various combinations of hyperparameters to achieve high accuracy in regional hydrology. Moreover, the efficiency and efficacy of the random search method in this context remain unexplored, likely due to the historically high computational costs.

Fortunately, advancements in computational resources, particularly new generations of Graphics processing units (GPUs), now make it feasible to conduct a comprehensive systematic hyperparameter optimizations. This chapter hypothesizes that systematic hyperparameter optimization of regional LSTM networks using random search can achieve high accuracy of hydrological predictions. The chapter aims to achieve the following objectives:

Optimization of Regional LSTM Networks: We aim to optimize hyperparameters to achieve high accuracy and reliable hourly streamflow and water level predictions in 40 catchments located in Basque Country, Spain.

Assessment of Random Search Method: We will evaluate the effectiveness of simultaneous systematic hyperparameter optimization using the random search method, addressing the need for comprehensive hyperparameter tuning of regional hydrological LSTM networks.

Analysis of Search Iterations: We will analyze the impact of increasing the number of search iterations on the final accuracy of tuned networks to determine the computational costs of the method and evaluate whether variations in different optimized hyperparameter configurations, regarding the number of searches, result in meaningful disparities among prediction outcomes. This objective ultimately helps us address the key question: "Should the precise optimization of hyperparameters in hydrological DL models be considered a significant task, or can some hyperparameters be ignored based on cognitive bias?"

4.2. Method

4.2.1. Definitions and Designing the Hyperparameter Search Space

4.2.1.1. (MTS)LSTMs' Hyperparameters and their Definitions

In our research, following initial trial and error and considering all the aforementioned factors, we carefully selected 10 key hyperparameters and two schedules for the learning rate for the MTS-LSTM model structure. In total, we optimized 12 hyperparameters, as we treated the three learning rates with the same importance as the other hyperparameters during the random search. This selection process involved an initial manual trial-and-error phase, conducted in parallel with consultations from experts' knowledge on other datasets. These hyperparameters significantly influence the performance of MTS-LSTM in predictions for our dataset.

The selected tuned hyperparameters in this research and their definitions include:

Length of Input Sequence:

The configured LSTM networks operate in a sequence-to-value mode, meaning that to predict a single discharge value, the model requires information from the preceding n−1 timesteps, along with the meteorological data for the target time. Consequently, the input sequences consist of n timesteps. In other words, the input sequence length is a hyperparameter representing the number of consecutive samples fed into the LSTM network. Each sample thus includes n input meteorological data points, allowing the network to predict the final unknown target value (streamflow or water level). Kratzert et al. (2018) set this value as a fixed 365 days to effectively capture the dynamics of a full annual cycle, in line with the hydrological concept of the water year. As a cognitive bias, mainly in hydrology, researchers have often considered this hyperparameter to be less critical for tuning compared to other hyperparameters.

In our study, we used two distinct hyperparameters: **Hourly Input Sequence Length** and **Daily Input Sequence Length**, which define the length of input sequences in terms of hourly and daily timesteps for the MTS-LSTM model, respectively. This input sequence length acts as a window size through which the network views the data to learn patterns. We found that tuning this window size is crucial for each catchment in hydrology data first in our initial try-and-errors. So, we decided to tune both daily and hourly input sequences. Later, the importance of this hyperparameter to get tuned was observed after the systematic random search hyperparameter optimization.

Table 2. Ranges of the input sequence lengths

sequence length daily (days)	146, 182, 365, 730, 1095
sequence length hourly (hours)	168, 336, 504, 672, 1344, 2016, 4032, 6720, 8064, 8760

Based on hydrological literature, such as Beven (2020), we recognized that each catchment has unique characteristics that affect water routing to the water basin's outlet. We hypothesized that the input sequence length, which dictates how the LSTM network processes data, might hold hydrological significance. Therefore, we chose to tune this hyperparameter over various sub-yearly to multi-year ranges. We speculated that the deep learning model could learn crucial water routing patterns within these windows, both at the catchment level and across the entire region. We decided to consider a range for this hyperparameter based on our understanding of hydrology and computational costs. Specifically, we set ranges from a hydrological perspective, including 2 weeks, 1 month, 2 months, 3 months, 6 months, 9 months, and up to 3 years (Table A.1.1).

Mini-Batch or Batch Size: This hyperparameter represents the number of samples shown to the model per back-propagation step during training. Back-propagation is the process by which the model updates its weights and biases based on the error of its predictions. In simpler terms, batch size is the number of data points the model processes before updating its internal parameters (weights and biases of the neural network). For instance, if the batch size is 64, the model will process 64 samples, then back-propagate the sum of errors from those 64 predictions through the network to modify its prediction parameters. However, the whole process is not that simple, and several other factors, such as vanishing gradient, could have adverse effects. To clearly understand the complex term of batch size, hydrologists interested in AI models are encouraged to refer to machine learning bible texts such as those by Russell and Norvig (2020).

Moreover, in general and from a machine learning perspective, larger batches optimize hardware utilization and yield stable gradient estimates but demand more memory and may risk overfitting. Conversely, smaller batches, while less hardware-efficient, serve as a regularization tool by offering noisier gradients that aid in generalization and prevent local minima.

After each round of training, the loss function is calculated based on the simulated and observed runoff for these samples. For example, Kratzert et al. (2018) used a batch size of 512, comprising one discharge value of a given day and the meteorological input of the preceding days. Later, Kratzert et al. (2024) adjusted this value to 256 following a new grid search. According to AI literature, batch sizes are typically set as a power of 2 (e.g., 16, 32, 64, 128, 256, 512).

Our trials indicated that increasing the batch size beyond 256 does not improve accuracy, and high accuracies could still be achieved with a batch size of 32. Additionally, our experiments suggested an inter-relationship between the length of the input sequence and batch size. Consequently, we decided to set the range for random search to 32, 64, 128, and 256.

Hidden Size: This hyperparameter represents the number of cells in the LSTM and significantly influences the LSTM's capacity to capture temporal dependencies in the data (Kratzert et al., 2018). It is important to note that increasing the hidden size is costly and increases the deep neural network's memory requirements. However, a larger memory is not always necessary, depending on the domain and specific concepts. Moreover, as demonstrated in this paper, there should be a balance between different hyperparameter configurations, and the assumption that higher memory always results in higher accuracy is not always correct. According to AI literature, hidden size values should also be a power of 2. Therefore, we decided to set the range for random search to 16, 32, 64, 128, and 256.

Initial Forget Gate Bias: This hyperparameter crucially impacts the decision-making process of the forget gate in LSTM cells. Proper initialization is vital to counteract issues like vanishing gradients and facilitate effective gradient flow across multiple timesteps (Hoedt et al., 2021; Gauch et al., 2021; Greff et al., 2017; Jozefowicz et al., 2015; Gers et al., 1999). While removing the forget gate is not viable due to its role in retaining pertinent information over time (Jozefowicz et al., 2015; Greff et al., 2017), initializing its bias to a small positive value has been proposed to address vanishing gradient challenges (Gers et al., 1999; Gauch et al., 2021; Hoedt et al., 2021). Gauch et al. (2021) consistently utilized this bias initialization across all MTS-LSTM models. We tested different values during our try and error and discovered that even negative values even have a positive effect on the accuracy at the end. At the end of random search, our idea of considering negative values worked and we had some highly accurate configurations with negative values for the forget gate. So, based on these considerations, we define the range for this hyperparameter as: -3, -1, 0, 1, 3.

Loss Function: The loss function quantifies the difference between the model's predictions and the observed values, playing a pivotal role in training LSTM networks through backpropagation by calculating and minimizing the network error. In this study, two common loss functions, NSE (from a hydrological perspective) and RMSE (reflecting the general machine learning viewpoint), were considered. While prior literature has favored NSE as the superior choice (Kratzert et al., 2024; 2019; 2018), our experimentation revealed nuanced outcomes, indicating the significance of other network hyperparameters on the loss function. Consequently, we advocate for researchers to employ random search to tune this hyperparameter effectively.

Regularization Term: The primary function of this hyperparameter is to enforce consistency across timescales in predictions through loss regularization (Gauch et al., 2021). In the context of the MTS-LSTM model, which simultaneously generates predictions at multiple timescales, ensuring coherence and alignment across these predictions is paramount. Gauch et al. (2021) operationalized this notion of consistency by defining predictions as coherent when the mean of hourly predictions matches the daily prediction for each day, a concept explicitly integrated into the loss function. Drawing from the catchment-averaged NSE loss introduced by Kratzert et al. (2019), the MTS-LSTM loss function incorporates contributions from individual timescales, enhancing the model's capacity for consistent and accurate predictions. In our study, the hyperparameter setting for this term could be the "Regularization: tie frequencies" option or None.

Dropout Rate: The Dropout Rate is a technique employed to mitigate overfitting in neural networks. It operates by randomly deactivating a fraction of input units, effectively reducing the network's reliance on specific features and promoting more robust learning (Kratzert et al., 2018; Gauch et al., 2021; Klotz et al., 2022). In practical terms, during each training step, a specified proportion of neurons are temporarily ignored, helping prevent the model from memorizing noise or idiosyncrasies in the training data. Dropout rate encourages broader learning and enhances model generalization, making it beneficial for its application to hydrological modeling tasks. In our study, we explored a range of Dropout rates during random search, including 0, 0.2, and 0.4, to evaluate their impact on model performance and robustness.

Standard Target Noise: The Standard Target Noise hyperparameter involves augmenting the output values during model training with relative noise characterized by a specified standard deviation. The noise added follows a Gaussian (normal) distribution centered around zero mean, where the standard deviation determines the spread of this distribution. In our context, we specified the standard deviation values that best represent the variability observed in hydrological datasets. This technique, as discussed by Klotz et al. (2022) and Gauch et al. (2021), aims to enhance model generalization and resilience to data variations. In our study, informed by experimentation and literature review, and considering potential errors inherent in hydrological datasets, we determined the range for this hyperparameter to be: 0, 0.01, 0.02, 0.05, 0.1.

Learning Rates: This hyperparameter dictates the extent to which the neural network weights adjust during optimization to minimize the loss function, underscoring its significance to both the optimizer and the loss function itself (Russell & Norvig, 2020; Goodfellow et al., 2016). In our approach, we employ three learning rate hyperparameters with scheduled adaptations at 10 and 25 epochs. An epoch denotes a full iteration of the training data set, encompassing all necessary iterations for the model to process every data point once. This scheduling strategy, as discussed by Gauch et al. (2021) and Nearing et al. (2021), balances initial rapid progress with precise fine-tuning later in training. For random search in this study, the defined learning rates were selected to cover a range that balances between effective learning and stable convergence (Table A.1.2). These values were selected to strike a balance between effective learning and stable training, aligning with typical choices in machine learning that fit well within our hyperparameter space.

Table 3. Learning rates ranges and schedules for random search

Learning rates	
Lr0	1e-3, 1e-2, 5e-2
Lr10	5e-4, 1e-3, 5e-3
Lr25	1e-4, 1e-3

While other hyperparameters, such as optimizer type, LSTM head, and output activation function, are also adjustable, we fixed them as “Adam”, “Regression”, and “Linear”, respectively (See: Keras Documentation, latest version for more on: https://keras.io), and focused on the 10 selected hyperparameters stated; however, it is valuable if in future studies researchers take them into account to find their tuning importance on final predictions. “Random Seed” is, also, another hyperparameter that can influence the outcome, but we decided not to emphasize on it during hyperparameter optimization and allowed the models to choose random seeds for each random experiment to add more randomness to the random search space for robustness of the process. Later, for robustness of the tuned networks, we trained them on 10 different random seeds. Furthermore, we set the number of epochs to 50 for all experiments both during hyperparameter optimization and final train/test, which were confirmed in our experience to be sufficient for training and did not result in overfitting through the TensorBoard module of Python.

4.2.1.2. The employed hyperparameters values in the literature

Kratzert et al. (2019) extended their previous studies and conducted systematic hyperparameters tuning using a grid search. They focused on four hyperparameters: hidden size, dropout rate, length of the input sequence, and the number of stacked LSTM layers. The tuned LSTM configurations featured 256 hidden cells and a single fully connected layer with a dropout rate of 0.4. In the latest version of an existing tuned regional LSTM architecture on CAMELS-US, Kratzert et al. (2024) configured their network as follows: hidden size: 256; batch size: 256; dropout rate: 0.4; input sequence length: 365 days; learning rates of 1e-3, 5e-4, and 1e-4 adapting at epochs of 20 and 25 in 30 final epochs of learning; and loss function: NSE.

Gauch et al. (2021) employed a two-stage systematic hyperparameters tuning approach for their multi-timescale LSTM (MTS-LSTM) architecture. Due to computational constraints, they performed a grid search on a confined hyperparameter space, focusing on learning rate and batch size. In the first stage, Gauch et al. (2021) optimized several network configurations (including regularization term, hidden size, input sequence length, and dropout rate) over 30 epochs with a batch size of 512 and a learning rate schedule starting at 0.001, reducing to 0.0005 after 10 epochs and further to 0.0001 after 20 epochs. The configuration with the best median metrics—an average of both daily and hourly values on all catchments—was selected. In the second stage, they tuned the learning rate and batch size by fixing other hyperparameters. However, they did not perform systematic hyperparameter tuning on the

architectural configuration of the naïve LSTM networks, and relied on the architecture tuned by Kratzert et al. (2019).

Mai et al. (2022) employed manual hyperparameter tuning, adjusting hyperparameters. They used an input sequence length of 365 days; the same as Kratzert et al. (2018). This constant value was an intentional decision by Kratzert et al. (2018) “in order to capture at least the dynamics of a full annual cycle”. However, one year later, Kratzert et al. (2019) changed this value to 270 days performing a limited systematic hyper-tuning using grid search. This is in contrast to using 365 days which is typical for research employing LSTMs in hydrology with respect to the hydrological water year.

In a study focusing on soil moisture prediction using LSTMs, Feng et al. (2020) manually tested various hyperparameters employing a batch size of 100, a hidden size of 256, and an input sequence length of 365 days. The latter was substantially longer than in a previous soil moisture prediction case (Fang et al., 2017) on the same dataset, where 30 or 60 days were used. Respecting traditional hydrology, Feng et al. (2020) justified a longer instance for the input sequence length to represent catchment snow and subsurface storage processes, which need longer-term memory compared to surface soil moisture. Similarly, Rahmani et al. (2021) selected hyperparameters through multiple trial-and-errors. Ouyang et al. (2021) took inspiration from hyperparameters similar to those manually tuned by Feng et al. (2020). The hyperparameters were: a batch size of 100, an LSTM input sequence length of 365, a hidden size of 256, and a dropout rate of 0.5.

In another experiment using LSTMs for differentiable parameters learning (dPL) to calibrate traditional hydrological models, Tsai et al. (2021) manually tuned hidden size and batch size using one year of data. They experimented with hidden sizes of 64, 256, and 1280, used a batch size of 300 instances and a training input sequence length of 240 days. They set the dropout rate of the network to 0.5.

Klotz et al. (2022) performed a more focused search considering six hyperparameters. To balance their computational resources and search depth, they followed three steps: First, they informally identified sensible general presets. Second, they trained networks for different combinations of four hyperparameters: hidden size, standard target noise, the number of densities (density heads are included to account for prediction uncertainty), and dropout rate. Third, they selected the best-performing architecture and refined it through more searches to determine the optimal settings for two hyperparameters of batch size and learning rate.

4.2.1.3. Selection of Hyperparameters to be tuned

When we set out to design the hyperparameter space for this study, our approach was methodical and informed by multiple sources. Specifically, we drew from three key sources, as outlined in the method section 2.3, briefly:

Empirical Testing: We conducted several try-and-error experiments in a sample catchment of our case study. This involved systematically testing the performance of different

hyperparameters to understand their impact on model performance. During these tests, for example, we found that the Length of Input Sequence can have a crucial hydrological impact on the final performance.

Expert Consultation: We consulted with the developers of NeuralHydrology to gain insights into optimal hyperparameter values for hydrological LSTM networks, including their experiences from other datasets on CARAVAN (Kratzert et al., 2023). However, we found discrepancies in some hyperparameter values based on our findings in our representative catchment. For instance, while usually LSTM users do not tune Input Sequence Length and consider 365 days as a reasonable choice, our experimentation revealed that the performance of the trained networks is very sensitive to this hyperparameter. Consequently, we argue that Input Sequence Length should be systematically tuned in hydrological applications, as it may have significant hydrological implications (Hosseini et al., 2024).

Literature Review: We performed an in-depth analysis of several key papers (The main text and also specifically, focusing on their attachments on hyper-tuning approach) that applied LSTMs in rainfall-runoff modeling and some other hydrological domains. By examining how these studies tuned their networks, we identified common practices and effective ranges for various hyperparameters. This review shaped the foundation of our hyperparameter selection.

Additionally, we considered the definitions and theoretical underpinnings of different gates and hyperparameters of LSTMs as described by Hochreiter and Schmidhuber (1997) and Kratzert et al., (2018). These foundational works provided crucial insights into the functioning and optimization of LSTM networks. We also referred to bibles of machine learning such as the great books of: Goodfellow et al., (2016) and Russell & Norvig, (2020).

Finally, we balanced these considerations with Occam's razor, ensuring our hyperparameter space was both computationally feasible and effective, prioritizing simplicity and minimizing unnecessary complexity. This way we have been able to pay the high computational costs of simultaneously optimizing 10 distinct hyperparameters with two learning rate schedules through 1000 random searches.

4.2.2. Random search

Our ultimate hyperparameter space had 5,400,000 possible configurations (Table 4), which, in itself, was unprecedented in the literature employing LSTMs for hydrological modeling. To investigate the hyperparameter space effectively, we conducted an exhaustive random experiment, including 1000 search iterations, inspired by the findings of Bergstra & Bengio (2012) to employ the random search method.

Moreover, we sought to determine the optimal number of random searches required for achieving satisfactory network performance within our resource constraints. Bergstra & Bengio (2012) suggested that conducting a larger number of random searches can lead to better outcomes. We designed our experiment to explore the impact of varying numbers of random searches on tuned networks' prediction accuracy. We selected 2 distinct numbers of

random searches: 100 and 1000. While we used 100 searches to create a reference point within feasible computational resources, we decided to conduct 1000 searches so as to investigate the potential for further improvement in regional prediction accuracy, albeit at a higher computational cost. By systematically varying the number of random searches while keeping other experimental conditions constant, we sought to determine the relationship between search effort and network performance in our hydrological prediction modeling tasks.

Table 4. The defined hyperparameters space designed for random search and the 2 final best-performing configurations on the validation set after 100 (Regional Optimal - RO) and 1000 (Enhanced Regional Optimal - ERO) random searches

Hyperparameter		Range	ERO	RO
hidden size		16, 32, 64, 128, 256	32	16
batch size		32, 64, 128, 256	64	256
output dropout		0, 0.2, 0.4	0.2	0.2
initial forget bias		-3, -1, 0, 1, 3	-	3
learning rates	Lr0	1e-3, 1e-2, 5e-2	0.01	0.01
	Lr10	5e-4, 1e-3, 5e-3	0.001	0.0005
	Lr25	1e-4, 1e-3	0.0001	0.001
target noise std		0, 0.01, 0.02, 0.05, 0.1	-	0.02
loss function		NSE, RMSE	NSE	NSE
seq length daily		146, 182, 365, 730, 1095	1095	365
seq length hourly		168, 336, 504, 672, 1344, 2016, 4032, 6720, 8064, 8760	336	168
regularization		tie_frequencies, None	tie_frequencies	tie_frequencies

We conducted our experiment using state-of-the-art machine learning frameworks and our available computational resources. Leveraging Python's scikit-learn (Pedregosa et al., 2011) for configuration management, we systematically explored our high-dimensional hyperparameter space to identify optimal configurations. By defining the hyperparameters space and using a parameter sampler, we generated 1000 random hyperparameter configuration sets (which we called the "randomly-tuned configurations"). We arbitrarily chose 100 of these configurations and trained them at first step. Later, we trained the remaining 900 configurations to increase the number of searches to 1000. The random search phase involved training and validating the performance of the randomly-tuned configurations on the training-and-validation set. We enabled the MTS-LSTM architecture to use random seeds for weight initialization during random search, increasing randomness as much as possible and avoiding fixed seed limitations. This allowed for broader exploration of possibilities and reduced computation constraints.

4.2.3. Post-random search

After conducting extensive training and validation spanning 46,000 minutes—equivalent to approximately 32 days—we scrutinized a total of 1000 randomly experiences of training and validation. The resulting DATASET encapsulated 1000 hyperparameter configurations and their respective validation metrics derived from 25 catchments with validation data. Subsequent outlier analysis led to the manual exclusion of 24 randomly-tuned configurations with average regional performance metric values below 0.5. Furthermore, 382 configurations failed to pass the training phase. Following these filtering steps, 100 random searches yielded to 74 viable configurations and 1000 random searches included 594.

We selected the 2 finally optimized configurations denoted as "**Regional Optimal (RO)**" and "**Enhanced Regional Optimal (ERO)**" models, respectively after 100 and 1000 random searches in the hyperparameter space. The final configuration settings of the 2 optimized networks are shown in Table 4. The finally hyper-tuned configurations were selected based on their highest average regional validation performance metrics of the 2 targets for 100 and 1000 random searches. The 2 finally optimized configurations underwent rigorous retraining 10 times on fixed-but-randomly-chosen random seeds. The trained networks were tested to evaluate their performance on all 40 catchments each resulting in 10 simulations for every target in every catchment. A flowchart schematic is provided in Figure 8 to summarize the whole method from data preprocessing to selection of the optimized models, retraining and test evaluation.

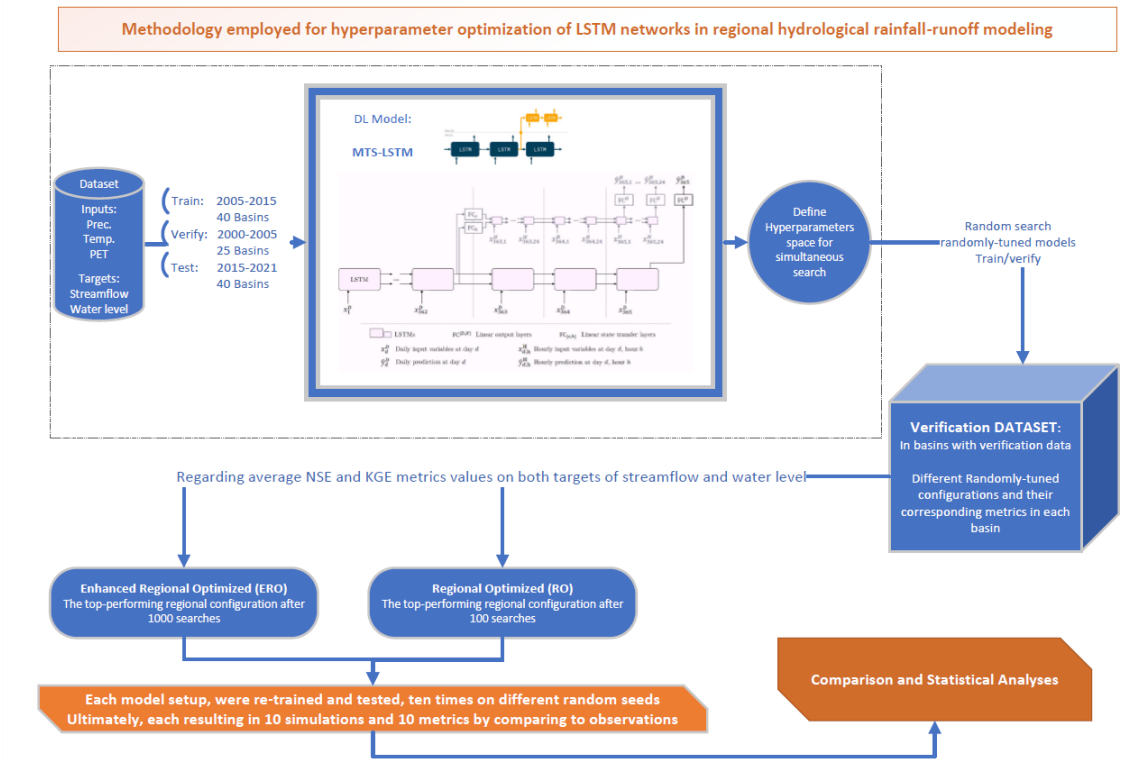


Figure 8. Methodology designed and employed in this thesis for hyperparameter optimization of LSTM networks in regional hydrological rainfall-runoff modeling; the MTS-LSTM schematic is from the reference paper: (Gauch et al., 2021)

4.2.4. Performance evaluation

Following the testing of the two regionally optimized RO and ERO networks on the test set, we compared their predictions with observations using several accuracy metrics to assess their overall and catchment-specific performance. To ensure reliable and accurate predictions several evaluation metrics have been presented. These metrics provide a detailed analysis of the models' accuracy, bias, and overall reliability, enabling a thorough evaluation of their predictive capabilities. Additionally, we evaluated the networks based on convergence speed and computational efficiency, regarding increase of number of random search iterations from 100 to 1000.

Moreover, to investigate whether the observed performance disparities between the RO and ERO networks were statistically significant or merely random occurrence, we adopted a P-value threshold of < 0.05. This allowed us to test the null hypothesis that there are no significant differences between the two optimized configurations across catchments. Detailed analyses were performed on ten individual performance metrics for simulations of RO and ERO networks across each catchment. We utilized Wilcoxon signed-rank, ANOVA, and Mann–Whitney U tests to evaluate differences in paired performance metrics, group means, and independent samples, respectively. This comprehensive approach enabled us to rigorously assess the impact of hyperparameter variations on model performance, providing insights into the effectiveness of the optimized models and the significance of different configurations in regional LSTM prediction outcomes.

4.3. Results

4.3.1. Overall test accuracy of optimized networks

In alignment with the first and second objectives of this study, simultaneous systematic hyperparameter optimization using random search identified 2 fine-tuned configurations (RO and ERO) that resulted in highly accurate hourly predictions both for streamflow and water level targets on overall, as illustrated in Table 5. The table represents minimum, maximum, average and median of all metrics for each configuration in 40 different catchments on 10 different random seeds. The ERO model simulations achieved regional average NSE metrics of 0.892 and 0.914, respectively for streamflow and water level, along with average KGE metrics of 0.872 and 0.915. Similarly, the RO model exhibited average NSE values of 0.895 and 0.908, and average KGE values of 0.873 and 0.900 for the 2 different targets, respectively. Moreover, the high average and median values in comparison to minimum metrics values in the table confirms that both chosen configurations after 100 and 1000 random searches demonstrated high levels of accuracy in 40 different catchments. It should be noted that minimum values are the minimum of all prediction metrics in all catchments on the 10 different random seeds. And as catchment-scale results demonstrate (Sec. 3.2), even in water basins that these minimums are coming from, we observe high values for the same metrics

on different random seeds. These results indicate accurate and precise regional performance for both optimized RO and ERO configurations.

Table 5. Overall regional performance metrics in all 40 catchments on 10 different random seeds by Enhanced Regional Optimal (ERO) and Regional Optimal (RO) optimized networks for streamflow and water level predictions on the test set. The table shows that both optimized networks of RO and ERO demonstrated highly accurate predictions in general, from a regional perspective and by aggregated metrics on the whole region.

Overall Performance Metrics in 40 catchments on 10 seeds on the Test set for the 2 optimized networks

Model	Enhanced Regional Optimal (ERO)				Regional Optimal (RO)			
Target	Streamflow		Water level		Streamflow		Water level	
Metric	NSE	KGE	NSE	KGE	NSE	KGE	NSE	KGE
Max	0.969	0.970	0.974	0.970	0.968	0.964	0.950	0.962
Average	0.892	0.918	0.872	0.920	0.895	0.913	0.873	0.906
median	0.903	0.930	0.891	0.933	0.902	0.927	0.888	0.922
Min	0.744	0.814	0.643	0.776	0.722	0.772	0.607	0.767

Although Table 5 presents similar overall regional aggregated metrics for both RO and ERO optimized networks, the granularity of catchment-specific analysis reveals nuanced ERO model's advantages. We need to state that Table 5 wants to verify accuracy of both optimized networks (RO and ERO) that were configured by the proposed method. In our analysis of catchment-scale test performance (See: Sec. 3.2), we found that the ERO network generally outperformed the RO network across more catchments. By analyzing the results at a finer scale, we demonstrate that the ERO model achieves higher accuracy and reliability in larger number of locations. This underscores the importance of evaluating regional models not solely on aggregated metrics but also based on their ability to perform well across diverse catchment characteristics and respecting the "uniqueness of the place" paradigm (Beven, 2020).

Figure 9 depicts the frequency distribution of NSE and KGE metrics for ERO and RO models across 40 different catchments on the 10 different random seeds. In each subplot, the histograms illustrate the distribution of metrics values, while the Kernel Density Estimation (KDE) curves provide a smoother presentation of the data distribution. The blue histograms and KDE curves represent the performance of ERO model, while the pale red counterparts depict RO model. Wherever the 2 colors fit on each other, we see it in dark red for frequency distributions. Generally, the ERO optimized network exhibits a higher frequency of higher metric values, as indicated by the larger blue areas in the histograms. Furthermore, the KDE curves for ERO model are a bit shifted towards the right, closer to the maximum metrics value of 1, suggesting overall outperformance of ERO compared to RO model in more catchments.

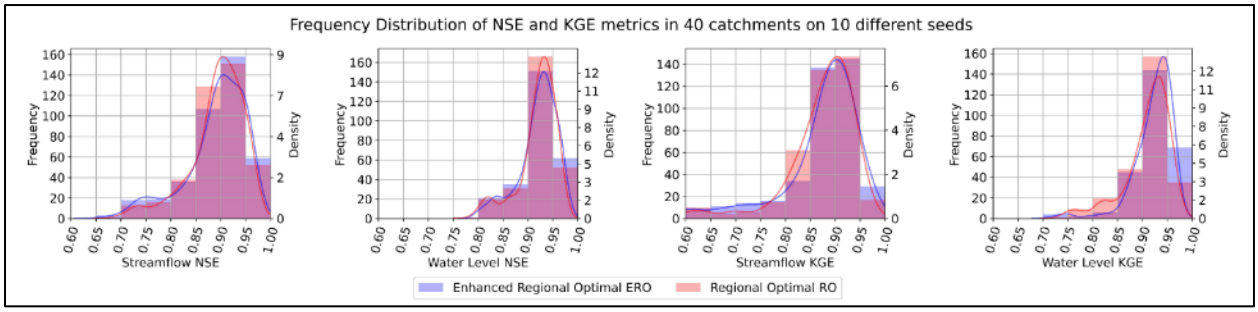


Figure 9. Illustration of the frequency distribution of RO and ERO performance metrics in 40 catchments on the 10 different seeds. The figure confirms overall outperformance of ERO in more locations compared to RO.

In addition to evaluating the predictive performance of the ERO and RO networks, we also assessed their computational efficiency in terms of training time. The ERO model, optimized through 1,000 random searches, required a total training time of approximately 81 hours across 10 instances for the 10 random seeds, whereas the RO model, found after 100 random searches, required approximately 86 hours. Despite the ERO model's increased complexity due to its extensive hyperparameter tuning, it demonstrated higher computational efficiency with a shorter overall training time at the end. Overall, the comparison of computational efficiency provides valuable insights for practical implementation, indicating that the ERO model offers a balanced trade-off between more accurate predictive performance and lower computational time demands. This makes it suitable for diverse real-world applications where both accuracy and resource management are critical.

4.3.2. Evaluating Catchment-Scale Performance of Optimized Networks

Figure 10, depicts a deep analysis comparing the nuanced performance differences between RO and ERO optimized MTS-LSTM networks in every catchment. The figure exhibits box plots of NSE and KGE metrics for streamflow and water level on all 10 random seeds in different catchments for each model in parallel. As is seen, on general we observe blue colors to exhibit higher accuracy in favor of ERO model in comparison to red colors for RO in more locations. The figure visualizes overall outperformance of ERO model in several water basins for different targets and different performance metrics. However, it shows that RO model still did not lose the competition; there are some specific catchments that RO outperforms meaningfully. The outperformance of RO in some specific places is an important note that we will discuss later.

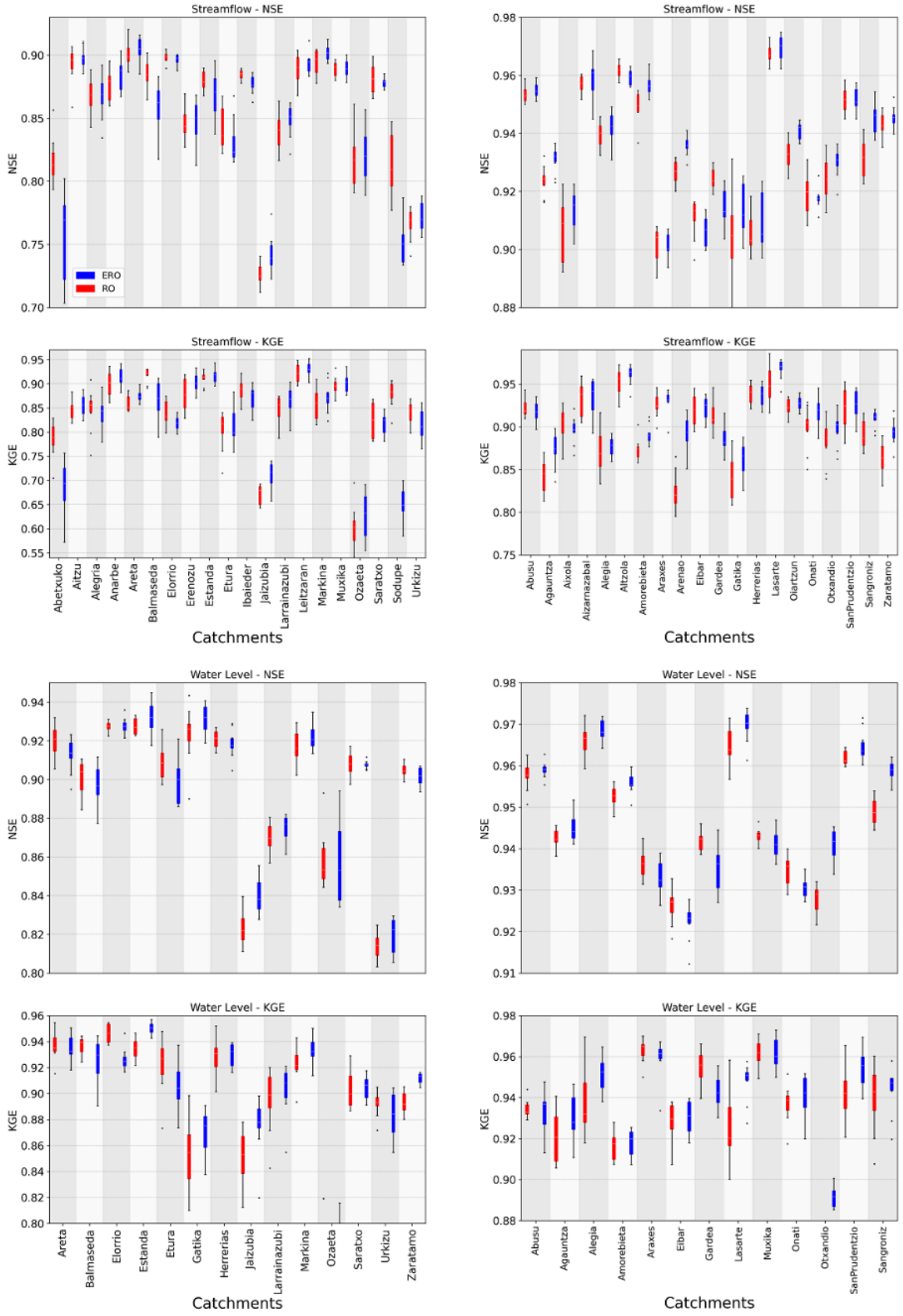


Figure 10. Illustration of NSE and KGE performance metrics distributions of RO and ERO models on 10 different seeds in every catchment. This figure shows where and how much each of the 2 optimized networks of RO and ERO, outperformed each other. In general, ERO outperformed in more locations; however, RO has its merits in some specific catchments.

When we zoom in each catchment on Figure 10, we observe that RO outperformed in some specific water basins that are known for their data deficiency (Sodupe and Urkizu), presence of reservoirs (Abetxuko, Balmaseda, Ibaieder, and Eibar), the snowy catchment of Etura, and the 2 catchments of Gardea and Estanda. It should be noted that only in 3 of the aforementioned places, including Abetxuko, Balmaseda and Sodupe, ERO, underperformed significantly. We know that Abetxuko has 2 large reservoirs, Balmaseda also has one, and Sodupe suffers bad quality data and several missing data records. The specific outperformance of RO reveals its different learning habits during the same training approach and emphasizes on the importance of hyperparameter optimization of regional LSTMs regarding the uniqueness of the place. This aspect is discussed further in the discussion section.

For a deeper understanding of the performance of RO and ERO models, we plotted the cumulative distribution functions of all 10 simulations for each model in all 40 catchments in Figure 11 to inspect which of them outperformed on general and in more locations. The blue color represents the performance of ERO model, while the red counterparts depict RO model. The plots demonstrate that although RO and ERO are extremely competitive from different aspects and in different catchments, ERO that was identified after 1000 random searches is more accurate regarding its predictions in general and in more water basins. Specifically, ERO outperformed in water level predictions and for the KGE metrics in several places.

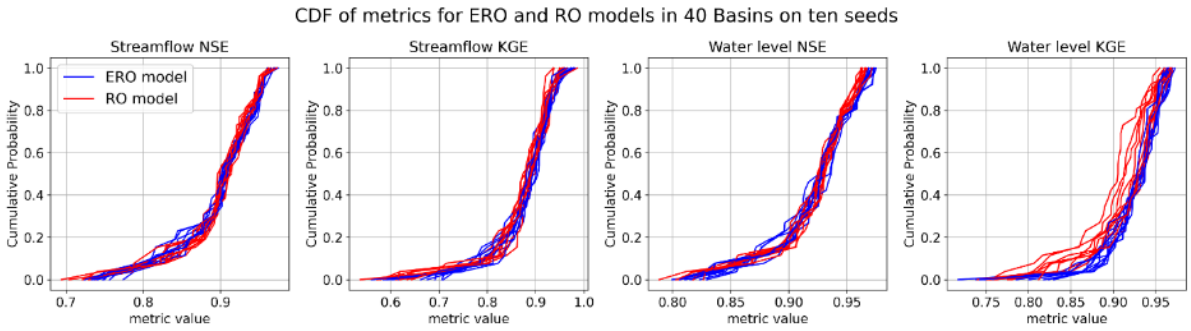


Figure 11. CDF of NSE and KGE metrics for 10 simulations of every model for the 2 targets in 40 Catchments. The plots show that ERO network is considered as a better optimized network, specifically, for water level predictions and KGE metric.

4.3.3. Significant Disparities in simulations of the optimized networks

To evaluate the differences in performance metrics between the 2 optimized MTS-LSTM networks (RO and ERO) across each catchment, we employed 3 statistical tests: Wilcoxon signed-rank, Mann–Whitney U, and ANOVA. These tests were applied to the results of 10 different simulations for each network configuration. The findings, which illustrate the disparities between the performance metrics of the RO and ERO networks, are presented in Figure 12. A difference was considered statistically significant if the P-value was less than 0.05.

Figure 12. Results of statistical analyses comparing the performance metrics of the ERO and RO models across different catchments. The analysis utilized the Wilcoxon signed-rank, ANOVA, and Mann-Whitney U tests to assess the significance of differences between the models' performance. Statistically significant results are indicated where P value < 0.05, demonstrating the impact of hyperparameter optimization on prediction accuracy across varying catchment conditions. In the figure, SF refers to streamflow and WL refers to Water level.

The 3 different statistical tests identified statistically significant differences in the performance metrics of RO and ERO at least on one target and one metric, with the exception of 11 catchments at the bottom of Figure 12 where no statistically significant differences were identified. These differences were observed across various catchment types and were not limited to specific locations. However, in some instances, disparities were more pronounced, particularly across both targets and the 2 metrics. The results of the statistical tests rejecting the null hypothesis - that speculated differences between the 2 distinct RO and ERO configurations are merely random occurrence, demonstrate that the 2 optimized configurations underwent the same training approaches and fed with the same training input, though both having high accuracy in several place; performed statistically differently in several catchments at least on one target or one metric. This confirms that disparities in performance metrics of different models in some water basins are not random occurrences but likely relate to the hyperparameter configuration settings and their learning skills.

4.3.4. Relation between number of random searches and accuracy

The analysis of the post-random search validation DATASET reveals significant insights pertaining to the third objective. As the number of random search iterations increased from 100 to 1000, a clear trend emerges, indicating an overall enhancement in identifying better-performing configurations that yield more precise predictions in more locations.

Figure 13 depicts the frequency distribution of overall regional validation NSE and KGE metrics for all succeeded randomly-tuned configurations across 25 catchments for the 2 targets. In each subplot, the histograms illustrate the distribution of metrics, curves provide a smoother presentation of the regional metrics distribution. The blue histograms and curves represent the performance of 594 succeeded randomly-tuned configurations after 1000 random searches, while the red counterparts depict the outcomes of 74 succeeded randomly-tuned configurations after first 100 random searches. Notably, increasing the number of searches, resulted in finding several better-performing configurations for different targets and on different metrics.

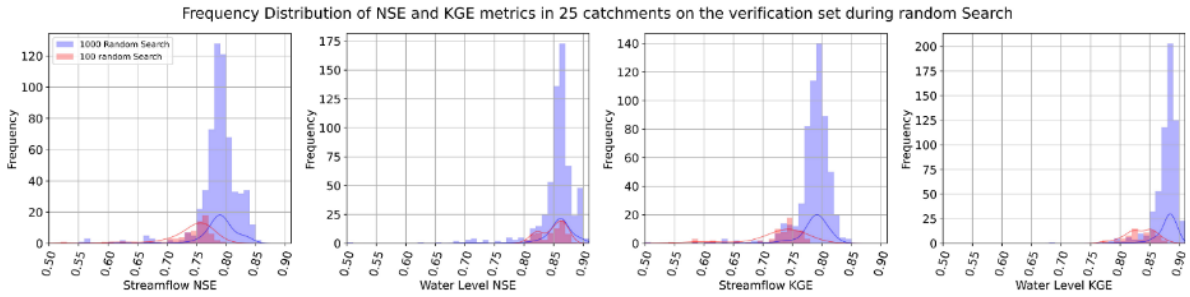


Figure 13. Illustration of the frequency distribution of validation performance metrics for randomly-tuned configurations after 100 and 1000 random searches.

There is an improvement in streamflow predictions for the overall regional maximum NSE score, which increased from 0.840 to 0.855, and the maximum regional KGE score, which improved from 0.828 to 0.846 with the transition from 100 to 1000 random searches. Particularly, for streamflow predictions, the average regional KGE score had an improvement from 0.780 to 0.785. Furthermore, the analysis of average and median performance metrics demonstrates consistency in the chance of finding better-performing configurations. As evidenced by the stable trends observed in the test metrics across different search iterations, it becomes evident that increasing the search volume systematically enhances the likelihood of identifying configurations that yield improved performance. This consistency in performance enhancements underscores the robustness of the random search approach in effectively exploring the hyperparameter space and uncovering configurations that lead to more accurate regional models with higher metrics in more places.

4.4. Discussion

4.4.1. Efficiency and efficacy of random search in hyperparameter optimization of regional LSTM networks

Consistent with Bergstra & Bengio's (2012) assertion, our findings corroborate the effectiveness of exhaustive random search in yielding mature network configurations. These optimized configurations ultimately contribute to enhanced prediction accuracy across a broader range of hydrological catchments. As shown in the results section, the performance metrics of both RO and ERO models indicate high accuracy. These findings confirm our hypothesis that random search method combined with simultaneous systematic hyperparameters tuning is an efficient and effective method for precise hyperparameter optimization of regional hydrological LSTM networks.

Furthermore, the findings suggest that increasing the number of random searches consistently results in performance improvements both regionally and in several places; although it cannot be asserted everywhere - we can find exceptions in some catchments. The ERO model demonstrates higher overall accuracy both regionally and in more catchments; this is consistent with the claim regarding the relation between the number of search efforts and predictive accuracy. However, random search always provides the possibility of finding a well-performing configuration promptly by chance at any time during the search. A well-balanced search in the hyperparameter space can ensure that an adequate number of trials, even as few as 100, can lead to discovering a configuration with high regional accuracy.

4.4.2. Performance metrics values interpretation

High accuracy of identified configurations has been confirmed by their high values for the NSE and KGE performance metrics on the 2 different targets of streamflow and water level. The reliability of NSE and KGE metrics in assessing accuracy is affirmed by the insightful work of Gauch et al. (2023). In their study, they established that these 2 metrics are robust indicators of overall and high-flow hydrograph quality, although the efficacy in assessing low-flow quality may be limited. Additionally, the stated research highlights the alignment of the quantitative metrics with human preferences, as hundreds of participants tended to favor machine learning models based on both visual judgments and quantitative assessments. Given our focus on generating accurate hourly predictions in flashy catchments, these findings show the importance of relying on NSE and KGE metrics to evaluate the accuracy of the optimized regional LSTM networks. Moreover, we considered 2 metrics to increase robustness of the evaluations.

Achieving highly accurate predictions in many catchments by the RO network that emerged only after 100 searches, highlights the efficiency of random search method for hyperparameter optimization of LSTM networks in regional hydrological predictions. This

finding, also, underscores the effectiveness of 3 key elements of the method, including meticulous design of the hyperparameter spaces, integration of hydrological domain expertise, and implementation of simultaneous random search to optimize several hyperparameters of LSTMs. Regarding the literature, it should be noted that this was the first try to simultaneously tune a vast number of hyperparameters for a regional hydrological prediction LSTM network using random search. We strongly suggest repeating this interesting experiment on other datasets and time steps to both confirming the method's applicability and help improve the existing configurations in the literature for more accuracy in predictions.

4.4.3. Complexity of post-random search configuration selection

In light of Bergstra & Bengio's (2012) observations on random search, we must reconsider the evaluation of validation performance of the randomly-tuned configurations when dealing with a large number of trials, each claiming superiority post-random search, regionally or in some specific places. In other words, after random search we are facing several configurations that each has something to say, in terms of high performance regionally or in some catchments on the validation set. However, in this research we decided to simply choose the configurations having the highest overall average NSE and KGE performance metrics on both targets. Our experiment demonstrated that ERO emerged as the best average regional model among the 1000 randomly-tuned configurations. However, another optimized configuration, RO, demonstrated its efficacy after just 100 searches with its specific architectural settings outperforming in some place compared to ERO. This contradiction is more obvious in specific catchments that (1) lack data quality (e.g., Sodupe and Gardea); or (2) suffer anthropogenic fingerprints (e.g., Abetxuko and Balmaseda having large dams).

If we define a "mature configuration" as an optimized regional hyperparameter setting that result in more accurate predictions across multiple catchments, then we can consider the RO model as a "premature" regional version of all possible best-performing configurations such as ERO. While the RO model may lack the comprehensive understanding we seek, it has still managed to identify some crucial anomalies and outperformed in specific locations, which is significant from a hydrological perspective. We think that this unique learning skill or "learning habits" of the RO network should not be overlooked. In other words, it appears that each of these 2 optimized network settings, RO and ERO, with their different hyperparameter configurations, has its own strengths and weaknesses. Therefore, it might be beneficial to explore a hybrid combination of different network architectures rather than relying solely on one configuration setting for such a complex task of regional hydrological prediction. This could pave the way for future research.

4.4.4. Learning maturity of different optimized regional configurations

Outperformance of the premature RO configuration in certain catchments suggests the need for alternative approaches in hydrology when training regionally-working artificial intelligence networks (AIs), highlighting an interaction between hyperparameter

configurations and the corresponding networks' "learning maturity" and learning habits, from a regional perspective. The statistically significant differences between the performance metrics of RO and ERO in many water basins, moreover, suggests meaningful distinctions in their learning approach and habits. Reviewing the specific learning behavior of RO and ERO networks reveals a sort of unification between their learning habits and the architecture (hyperparameters). In other words, every configuration setup develops a unique architectural network that affects what to learn and what not to. Such unique skills should be evaluated after training and validating a regional network regarding performance of the trained network in different places from a hydrological viewpoint.

Furthermore, the analysis of the behaviors of RO and ERO suggests that networks with optimized hyperparameters and well-designed architectures exhibit more efficient learning patterns in terms of what we wanted from them. To clarify, levels of learning maturity are what we define and force our AI networks during hyperparameter optimization. All of our decisions on how to tune a deep learning neural network such as LSTMs directly or indirectly inject biases to the AIs. For example, a network that accurately predicts in more catchments is what we actually wanted to train as a regional model in this research; but an optimized network (RO) that accurately functions in some specific water basins, has already learned something that we did not specifically want it to learn. The RO network learned patterns in places that we used to think prediction is a cumbersome task there due to some reasons such as data deficiency or human intervention – common hydrological biases coming from our traditional knowledge working with conventional hydrological models.

This observation underscores the importance of optimizing the architecture (hyperparameter configurations) to facilitate effective learning and enhance the network's ability to intelligently capture more and more complex patterns within the hydrological data. In other words, the time has come that we learn from our trained AIs and let them have more freedom during training instead of confining their performance to our traditional biases (e.g., 365 days of water year for the length of the input sequence). Additionally, this observation emphasizes the significance of meticulous and systematic hyperparameter optimization in achieving high accuracy and robustness across various hydrological tasks and datasets.

The following question remains: "If there are crucial deficiencies in training data in some specific catchments, why a premature network (RO) could capture them better?" Although we do not know the answer yet, our findings seem to reinforce the idea that hyperparameters have significant influence on deep learning models' learning skills and habits (Russell & Norvig, 2020; Goodfellow et al., 2016; Bergstra & Bengio, 2012).

Regarding the rank of the premature RO among the 1000 randomly-tuned configurations on the post-random search validation DATASET, 39 other randomly-tuned versions exhibited higher average regional performance metrics than RO but did not surpass ERO. It is essential to acknowledge that our selection of the best-performing configuration post-random search among these randomly-tuned variations was still influenced by human biases; we simply selected the ones with highest overall average regional performance metrics without giving any weights to their catchment-scale performance in different locations. Our decision to prioritize the intelligent network with the best overall regional performance from a human

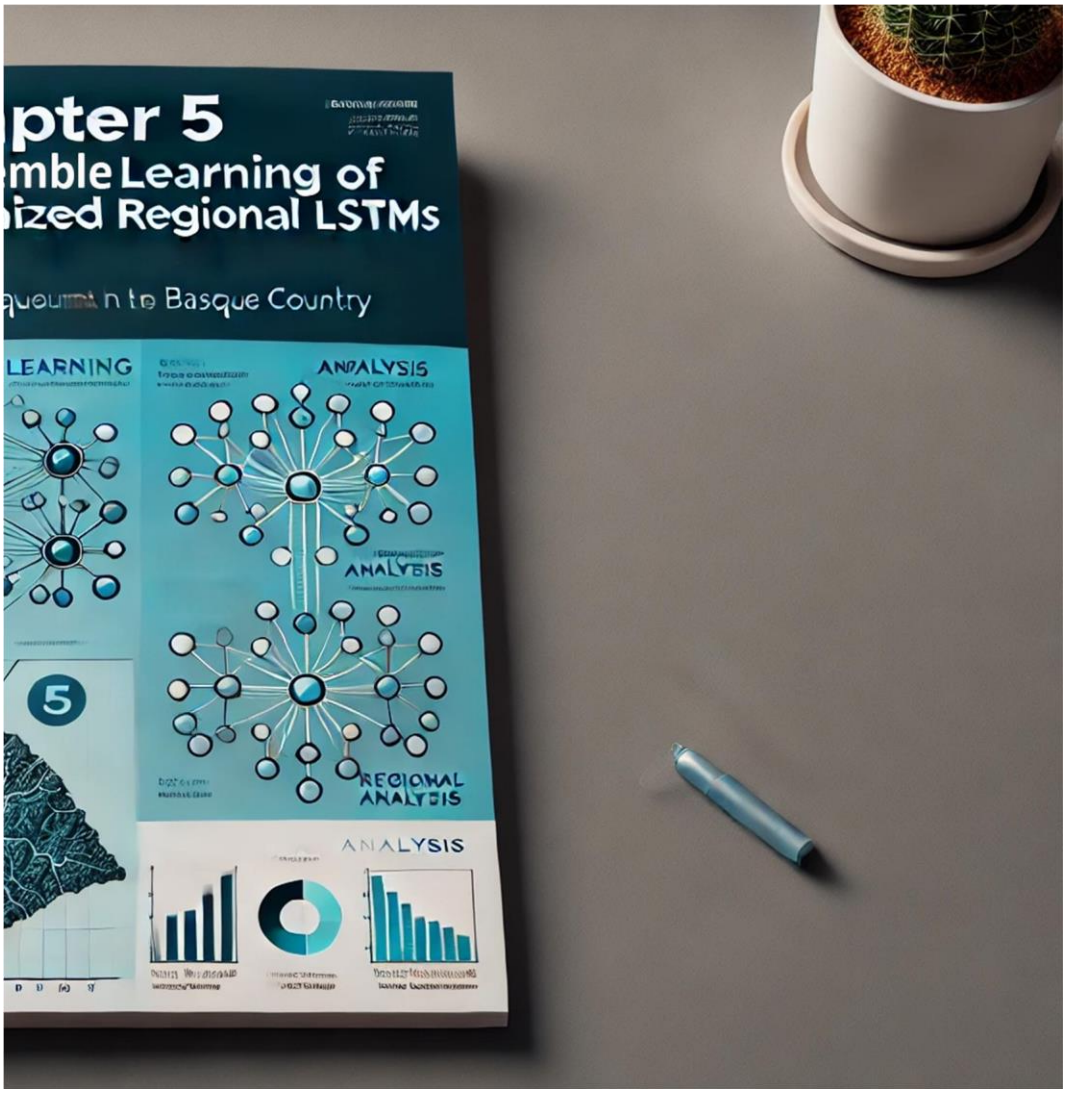
point of view injects a subjective element into the evaluation of several unknown processes, potentially obscuring the actual capabilities of the training AIs.

In the introduction, we posed the common adoption of a 365-day for the length of the input sequence of LSTM networks to help LSTMs “in order to capture at least the dynamics of a full annual cycle (Kratzert et al. 2018).”. This conventional practice, flowed by many researchers, aims to facilitate the learning process of the advanced deep learning models, allowing them to capture intricate patterns within the complex Earth system that may elude human comprehension. However, our experimentation with the ERO model unveiled a departure from this norm. Unlike the RO model, which performed optimally with a 365-day input sequence length, the more mature version of ERO required a longer and specific sequence length spanning 3 years (1095 days). This unexpected finding challenges traditional beliefs and may carry particular significance given the flashy nature of the studied catchments in Basque Country. In that, conventional hydrological wisdom may suggest a shorter input sequence length would have sufficed for our case study; however, simultaneous systematic hyperparameter optimization claimed it was not the case.

Chapter V

This chapter is an edited version of two Research papers:

1. Hosseini, Farzad, Prieto, Cristina, & Álvarez, Cesar. (2024). Hyperparameter optimization of regional hydrological LSTMs by random search: A case study from Basque Country, Spain. Journal of Hydrology, 132003. doi: 10.1016/j.jhydrol.2024.132003.
2. Hosseini, Farzad, Prieto, Cristina, & Álvarez, Cesar. (2024). Alpine-Peaks Shape of Optimized Configurations Post Random Search in Regional Hydrological LSTMs, Proceeding of the 10th IEEE edition of the International Conference on Optimization and Applications (ICOA), Almeria, Spain, 2024, pp. 1-5, doi: 10.1109/ICOA62581.2024.10754182.



Ensemble Learning of Optimized Regional LSTMs

“Everything should be made as simple as possible, but no simpler (Albert Einstein).”

5.1. Introduction

Despite becoming a cornerstone in hydrological modeling, the performance of Long Short-Term Memory (LSTM) networks can vary considerably across different catchments, which presents challenges in regional rainfall-runoff modeling. While LSTMs often exhibit strong generalization capabilities on a regional scale, they may underperform in specific catchments, undermining their reliability for crucial applications like flood resilience and mitigation (Beven, 2020; Prieto et al., 2020). This inconsistency highlights the need to optimize models based on their performance at the catchment level, where underperformance in even a single catchment can erode stakeholders' confidence. Embracing the "uniqueness of the place" paradigm requires enhancing model accuracy in underperforming catchments to ensure both overall and catchment-scale reliability.

Traditional hydrology often attributes a model's underperformance in specific locations to factors like snow, reservoirs, or underground flows. However, in the context of modern hydrological deep learning models (DLs) such as LSTMs, which are designed to uncover latent features from extensive datasets (Donnelly et al., 2024a), such explanations are no longer sufficient. This discrepancy raises a crucial question: why do intelligently trained regional LSTM models still struggle to achieve acceptable results in certain locations (Beven, 2020)?

A significant challenge in regional hydrological modeling is the tendency to evaluate regional models based on median performance metrics across the whole region. While this approach can indicate general model effectiveness, it can also mask substantial deficiencies at the catchment level. As Valiela (2000) points out, regional comparative studies often suffer from the drawback that their "conclusions are valid only for the dataset on aggregate." In practice, a single-configuration LSTM network trained on aggregated data from multiple catchments may perform well on average but fail in specific locations, underscoring the need for more sophisticated approaches to improve model reliability across diverse environments.

The objective in developing a regional model should be to achieve high-performance metrics across the majority of catchments rather than relying solely on regional median metrics, which can obscure critical issues. It is essential to focus on individual catchment performances and their true significance in different locations, as highlighted by concepts like "Multi-Objective Recommendations" (Zheng & Wang, 2021) or "Model soups" (Wortsman et al., 2022). Ignoring poor performance in specific catchments, even if they are considered outliers, should not be acceptable in regional hydrological predictions using deep learning.

Hyperparameter optimization is critical for improving model performance. Proper network design and architectural complexity are fundamental to enhancing a neural network's learning capacity (Russell & Norvig, 2020; Goodfellow et al., 2016; Shalev-Shwartz & Ben-David, 2014; Sutskever et al., 2013; Glorot & Bengio, 2010). Optimizing these aspects through hyperparameter tuning is essential for maximizing the predictive potential of deep neural networks, particularly in the hydrology domain.

Conventional methods to optimize one regional network for the whole region often overlook the unique characteristics of individual catchments. In Chapter 2, we demonstrated that different optimized regional LSTM networks exhibit unique strengths and weaknesses specific to each catchment, highlighting the complexity and importance of selecting the best-performing configuration for each location. Ensemble methods, widely used in various disciplines (Opitz & Maclin, 1999; Polikar, 2006), including hydrology (Prieto et al., 2021, 2022; Höge et al., 2018; Carneiro et al., 2022), could offer a solution to the challenges posed by regional variability in LSTM performance in regional hydrology. This chapter hypothesizes that an ensemble of regionally optimized LSTM configurations (what we term it: "Ensemble Learning") might achieve higher learning capacity and prediction accuracy than a single regionally optimized configuration.

The primary objective of this chapter is to develop and test different ensemble learning strategies to enhance the accuracy and robustness of regional LSTM-based rainfall-runoff models across multiple catchments. By combining several optimized regional LSTM networks, each tailored to specific catchment conditions, we aim to create a more resilient and reliable prediction system. This approach also aligns with the "uniqueness of the place" paradigm (Beven, 2020), which emphasizes the importance of considering local conditions in hydrological modeling.

Ensemble methods in ML can be implemented using various techniques, each with its own advantages and trade-offs. Bagging (Bootstrap Aggregating) and Voting are of common methods for ensemble methods in ML/DL (Breiman, 1996; Dietterich, 2000). Bagging involves training multiple models on different subsets of the training data and averaging their predictions, which helps reduce variance and prevent overfitting (Breiman, 1996). Voting combines predictions from multiple models using majority voting or averaging, offering a simple yet effective way to improve performance (Polikar, 2006; Opitz & Maclin, 1999).

In this chapter, we will explore these ensemble techniques, focusing on their application to LSTM networks for regional rainfall-runoff modeling. We will generate three different ensembles of regional optimized LSTMs following the random search we performed. We will introduce a novel approach, termed "catchment-wise optimized ensemble," which selects the best-performing regional models for each catchment and combines them to form an ensemble that maximizes performance across the region regarding the general ideas of Begging and Voting ensemble methods in MLs.

The objectives of this chapter are threefold:

1) To implement and evaluate three different ensemble learning methodologies to mitigate cognitive bias in determining hyperparameter configurations for regional hydrological LSTM networks.

2) To benchmark the performance of ensemble learning approaches against the best-performing single configurations at the regional level from Chapter 04.

3) To compare the performance of ensemble learning approaches to study their pros and cons.

By addressing these objectives through our case study, we aim to enhance hourly streamflow and water level prediction accuracy and robustness across 40 catchments of the flashy, humid region of Basque Country, Spain. We aim to improve prediction accuracy in as many catchments as possible including those with challenging predictions by single-configuration optimized regional LSTMs from Chapter 01 (e.g., human-intervened catchments). Moreover, this research seeks to ensure that regional DLs achieve high performance not just on aggregate metrics, but across multiple individual catchments, aligning with the practical needs of hydrology.

5.2. Method

5.2.1. Hyperparameter optimization and ensemble development technique

In this research, we aimed to optimize the hyperparameters of the MTS-LSTM network to enhance its performance in predicting hourly streamflow and water level measurements in as many locations as possible. For this aim, we considered 10 distinct hyperparameters, including the learning rate, which was scheduled twice during the optimization process (See: Table 4 in Chapter 4).

We used the random search method (Bergstra & Bengio, 2012) to optimize these hyperparameters simultaneously. Through 1000 iterations of random searches in the hyperparameter space, we generated the post-random search validation DATASET (with capital letters). The DATASET consisted of the randomly-tuned configurations and their corresponding validation performance metrics for each target in 25 catchments having validation data.

The post-random search validation DATASET developed through an exhaustive random search for hyperparameter optimization represents a crucial step in this research. The final DATASET included the 594 randomly-tuned configurations (which successfully completed both training and validation phases) and their corresponding validation metrics for 25 out of the 40 catchments (those with available validation data).

In Chapter 3, analyzing the DATASET, we identified the configurations with the highest overall regional validation performance metrics after 100 and 1000 search iterations as the best-performing regionally optimized networks, termed “Regional Optimal” (RO) and “Enhanced Regional Optimal” (ERO) networks (See: Table 4 in Chapter 4). We re-trained this network 10 times with different random seeds to increase the robustness of the predictions. Each of these 10 re-trained networks was tested on the test set, resulting in 10 distinct prediction timeseries for each of the 40 catchments. The performance metrics of these predictions, compared with observations, served as our benchmarks to evaluate the potential performance improvements achieved by the ensemble learning approaches, both at the regional level and on a catchment-by-catchment basis.

We explored different ensemble learning approaches for selecting configurations rather than relying solely on the regionally best-performing configuration (ERO network). Ultimately, we adopted 3 approaches for this purpose:

1) **Top 10 Configs**: We selected the top ten regionally best-performing configurations on the validation set after 1000 random searches; this ensemble included ERO.

2) **Catchment-wise Configs**: Recognizing the uniqueness of the catchments in shaping their hydrological behaviors, we chose the best-performing regional configurations for each catchment individually, regarding its validation metrics. Since validation data was available for only 25 catchments, this approach yielded 23 unique configurations, with some overlap in certain cases at the end.

3) **K-means Configs**: To minimize cognitive bias in the configuration selection process, we employed a K-means Clustering (MacQueen, 1967) unsupervised machine learning model. This model was trained on the normalized post-random search validation DATASET to select an ensemble of best-performing regional configurations. After experimenting with different numbers of clusters, we converged on 8 configurations chosen by the K-means Clustering model, representing a cluster with the highest overall average metrics in several tries.

Despite potential overlaps, these 3 approaches resulted in a total of 37 distinct regional configurations for re-training and testing. Table 6 presents the settings of these 37 configurations and the methods used to select them. Additionally, the overall regional rank of each configuration after 1000 random searches in the post-random search validation DATASET (sorted by the highest overall regional validation performance metrics) is seen that will be discussed in detail later.

Figure 14 presents a flowchart summarizing the entire methodology, from data preprocessing to hyperparameter optimization, and from final configuration selection for ensemble learning to re-training and model evaluation.

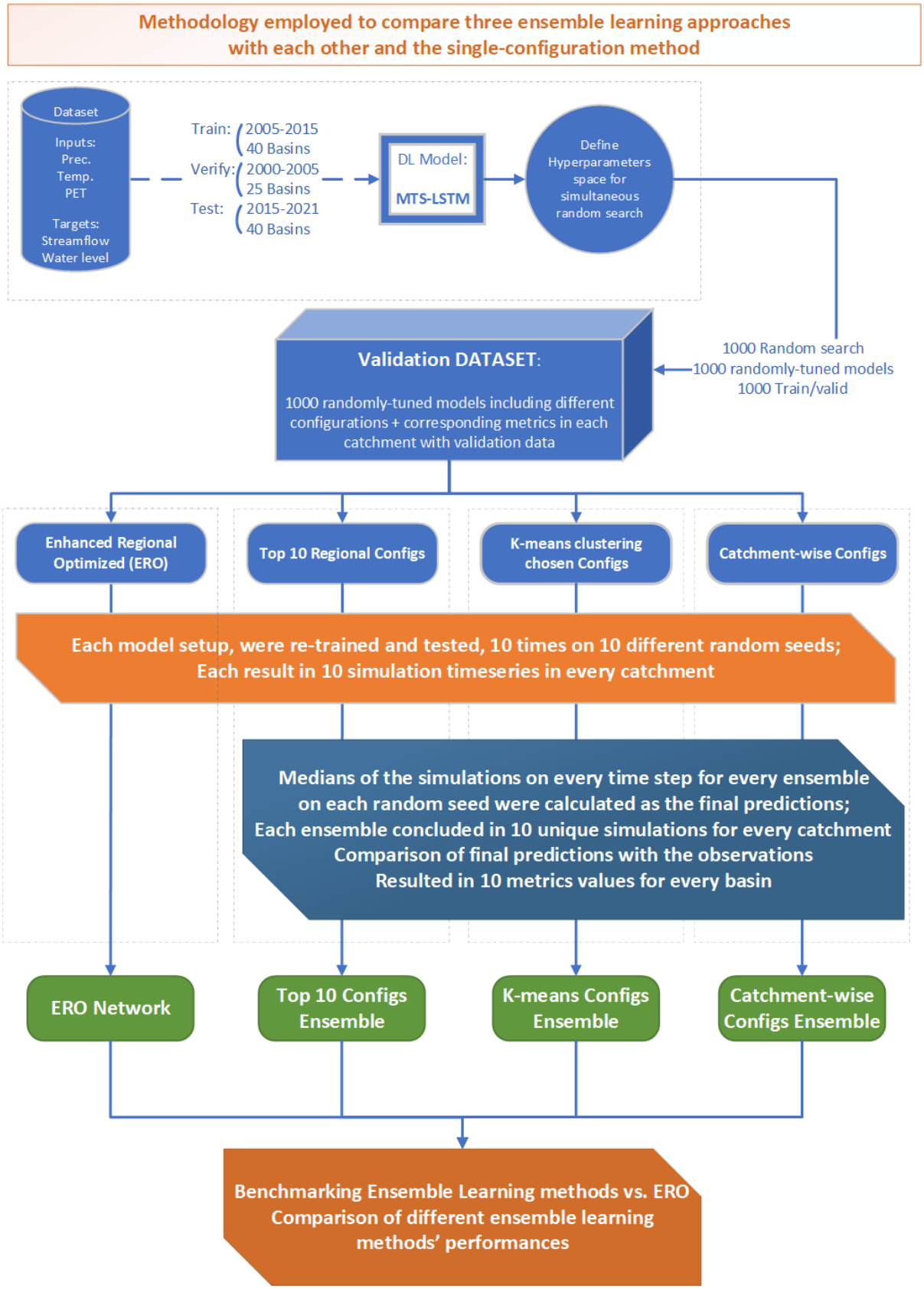


Figure 14. Methodology designed and employed in this thesis for hyperparameter optimization of LSTM networks by random search method and configuration selection post-random search to develop 3 final ensemble deep learning approaches compared with single-configuration approach in regional hydrological rainfall-runoff modelling.

Table 6. Different hyperparameter configurations for the 3 ensemble learning methods employed in this study. The table also shows a snapshot of the post-random search validation DATASET, consisting of randomly-tuned networks with different hyperparameter configurations and their overall aggregated average regional validation performance metrics. However, in the original post-random search validation DATASET, we see the metrics for every catchment.

Randomly-tuned Networks	Hyperparameters												Post-Random Search Configuration Selected ensembles	Best Performance Overall or in a basin	Rank	Regional validation metrics						364 Final Re-trained Networks
	Sequence Length		Batch size	Target noise std	Learning rates			loss	Hidden size	Output Dropout rate	Initial Forget bias	Regularization				Stream flow		Water Level		Overall	Train conflict	
	Dalily	Hourly			Lr0	Lr10	Lr25									NSE	KGE	NSE	KGE			
1005142928	1095	336	64	0	0.01	0.001	0.0001	NSE	32	0.2	0	Yes	Top10Configs/ERO	Overall	1	0.839	0.828	0.897	0.905	0.867		10
2108152814	1095	4032	64	0	0.001	0.0005	0.0001	RMSE	128	0.2	0		Top10Configs	Overall	2	0.854	0.824	0.884	0.902	0.866		10
1005101004	1095	336	128	0.01	0.001	0.005	0.0001	NSE	32	0.2	3	Yes	Top10Configs/Catchment_wise	Overall/Balmaseda	3	0.836	0.823	0.902	0.901	0.865		10
3107112133	1095	4032	128	0.02	0.001	0.0005	0.001	RMSE	64	0.2	3	Yes	Top10Configs	Overall	4	0.855	0.822	0.882	0.896	0.864		10
2804180425	1095	504	64	0	0.01	0.005	0.0001	NSE	32	0.2	-3	Yes	Top10Configs	Overall	5	0.824	0.818	0.895	0.909	0.861		10
3107061251	1095	2016	128	0.05	0.001	0.0005	0.0001	NSE	128	0.4	-3		Top10Configs/K-means	Overall	6	0.837	0.840	0.863	0.903	0.861		10
2807224240	1095	6720	64	0.02	0.01	0.001	0.0001	RMSE	128	0.2	3	Yes	Top10Configs	Overall	7	0.834	0.846	0.861	0.902	0.861	Yes	5
2804012925	1095	504	128	0.05	0.01	0.001	0.0001	NSE	128	0.4	0	Yes	Top10Configs/Catchment_wise	Overall/Sangroniz	8	0.830	0.814	0.894	0.905	0.861		10
905140944	1095	336	128	0	0.01	0.005	0.001	NSE	128	0.4	3	Yes	Top10Configs	Overall	9	0.829	0.818	0.892	0.901	0.860	Yes	9
908110638	1095	1344	128	0.05	0.01	0.001	0.001	NSE	64	0.4	0	Yes	Top10Configs/Catchment_wise	Overall/Onati	10	0.846	0.824	0.876	0.893	0.860		10
1608204529	1095	2016	64	0.05	0.001	0.005	0.001	RMSE	64	0.4	3	Yes	K-means	Overall	29	0.843	0.822	0.869	0.887	0.855		10
1005022449	730	504	128	0.01	0.01	0.005	0.001	NSE	32	0.2	0	Yes	Catchment_wise	Erenozu	30	0.810	0.811	0.897	0.903	0.855		10
2707071603	1095	504	128	0	0.01	0.005	0.0001	NSE	64	0.4	3	Yes	Catchment_wise	Anarbe, Herrerias	36	0.838	0.818	0.872	0.890	0.855		10
1808192306	1095	2016	256	0.02	0.001	0.001	0.0001	NSE	64	0.2	3	Yes	K-means	Overall	44	0.839	0.813	0.872	0.887	0.853		10
2008082456	1095	2016	256	0.05	0.001	0.001	0.0001	NSE	32	0.2	3	Yes	K-means	Overall	63	0.841	0.801	0.868	0.889	0.850		10
105104537	365	8759	128	0	0.001	0.005	0.001	NSE	32	0.4	3	Yes	Catchment_wise	Aizarnazabal	70	0.813	0.801	0.903	0.879	0.849		10
2108105805	730	6720	64	0	0.001	0.0005	0.0001	RMSE	128	0	-3	Yes	Catchment_wise	Alegia	74	0.797	0.809	0.886	0.903	0.849		10
1508015650	730	504	256	0.05	0.001	0.0005	0.0001	NSE	64	0.4	3	Yes	Catchment_wise	Agauntza	92	0.800	0.817	0.875	0.898	0.848		10
205011029	365	8759	128	0	0.001	0.005	0.0001	RMSE	32	0.4	3	Yes	Catchment_wise	Otxandio	122	0.821	0.794	0.893	0.865	0.843		10
2407192914	730	4032	256	0.02	0.001	0.0005	0.001	NSE	64	0.2	3	Yes	Catchment_wise	Muxika	124	0.798	0.794	0.882	0.899	0.843		10
2808094043	730	672	128	0	0.001	0.001	0.001	NSE	16	0.2	1		Catchment_wise	Leitzaran	134	0.806	0.823	0.859	0.882	0.842		10
1908205136	730	504	256	0.05	0.001	0.0005	0.0001	NSE	64	0	0		Catchment_wise	Altzola	136	0.792	0.817	0.868	0.892	0.842		10
1108202004	1095	2016	64	0	0.01	0.005	0.0001	RMSE	32	0.2	3		K-means	Overall	144	0.830	0.790	0.869	0.877	0.841		10
2308074224	730	336	128	0.05	0.001	0.0005	0.0001	NSE	64	0.2	1		Catchment_wise	Aitzu	156	0.788	0.805	0.864	0.897	0.839		10
2008204655	1095	2016	64	0.05	0.01	0.001	0.0001	RMSE	16	0.2	3		K-means	Overall	165	0.826	0.789	0.861	0.880	0.839		10
2708221140	730	336	256	0.1	0.001	0.0005	0.001	NSE	32	0	1	Yes	Catchment_wise	Lasarte, Amorebieta	167	0.789	0.819	0.865	0.882	0.839		10
2507010123	365	672	128	0.05	0.001	0.005	0.001	RMSE	64	0.2	0	Yes	Catchment_wise	Zaratamo	169	0.796	0.799	0.871	0.891	0.839		10
2607202253	1095	2016	128	0.05	0.001	0.005	0.001	NSE	32	0.4	-3	Yes	K-means	Overall	183	0.828	0.799	0.848	0.878	0.838		10
2807122044	146	168	128	0.02	0.01	0.0005	0.0001	NSE	128	0.4	0	Yes	Catchment_wise	Gardea	217	0.791	0.793	0.871	0.885	0.835		10
2008130710	146	336	64	0	0.001	0.005	0.0001	RMSE	64	0	0		Catchment_wise	Saratxo	321	0.779	0.778	0.872	0.886	0.829		10
2808020247	365	168	128	0	0.01	0.001	0.001	NSE	64	0	0		Catchment_wise	Elorrio	341	0.771	0.790	0.864	0.890	0.829		10
1308142418	1095	2016	128	0	0.001	0.005	0.001	NSE	32	0.4	-3	Yes	K-means	Overall	356	0.807	0.776	0.844	0.882	0.827		10
1108224759	1095	6720	128	0	0.001	0.001	0.001	NSE	128	0.2	3	Yes	Catchment_wise	Estanda	391	0.802	0.772	0.854	0.876	0.826		10
1708232434	365	504	64	0.05	0.001	0.001	0.001	NSE	32	0.2	3	Yes	Catchment_wise	Abusu	437	0.772	0.780	0.863	0.875	0.823		10
1308172158	1095	8064	256	0.02	0.001	0.0005	0.001	NSE	64	0.4	-3		Catchment_wise	Aixola	537	0.778	0.802	0.793	0.842	0.804	Yes	10
3004061929	730	504	64	0.02	0.01	0.001	0.001	RMSE	16	0.4	3	Yes	Catchment_wise	Urkizu	548	0.784	0.730	0.870	0.816	0.800		10
2408005852	365	168	64	0.1	0.001	0.0005	0.001	NSE	16	0.4	0		Catchment_wise	Oiartzun	559	0.767	0.760	0.786	0.851	0.791		10

5.2.2. Re-training, test, and ensembles predictions

We proceeded to re-train each of the 3 configuration ensembles 10 times, using 10 different random seeds, consistent with the approach used for the ERO model. Despite some limited training challenges that will be discussed later, we completed the entire re-training process. In total, the Top 10 Configs, K-means Configs, and Catchment-wise Configs ensembles resulted in 364 final re-trained networks (refer to Table 1).

We then tested all these re-trained networks on the previously unseen test set and extracted their simulation timeseries in every catchment. Next, we calculated the medians of predicted timeseries for every time step across different networks within each ensemble on every random seed. This approach yielded 10 final prediction timeseries per random seed for each of the 3 ensemble learning approaches in every catchment. Using these ensemble prediction timeseries, we calculated performance metrics by comparing them with observations, generating 10 metrics values for each configuration ensemble on every random seed in every catchment for every target. This mirrored the approach used for evaluating the ERO model performance and allowed us to benchmark the final performances.

5.2.3. Evaluation and benchmarking the results

We benchmarked the performance of the Top 10 Configs, K-means Configs, and Catchment-wise Configs ensembles against the optimized networks from Chapter 3, assessing their performance both regionally and at the catchment scale using different accuracy metrics. This evaluation aimed to discern whether observed differences in performance metrics among the ensemble learning methods and the ERO network were statistically significant or merely random variations. Using a criterion of P-value < 0.05 to reject the null hypothesis, we conducted detailed analyses of the 10 individual performance metrics for simulations across each catchment. We employed three statistical tests—Wilcoxon signed-rank, ANOVA, and Mann–Whitney U—to assess differences among group means, paired performance metrics, and independent groups, respectively. This systematic evaluation framework provided comprehensive insights into the effectiveness of ensemble learning methods, highlighting their potential for improving LSTM predictions in regional hydrological applications by capturing different aspects of prediction accuracy and revealing the true impact of variations in hyperparameter configurations.

5.3. Results

5.3.1. Benchmarking ensemble learning versus RO and ERO networks

Figure 15 illustrates the enhancement in final predictions achieved through ensemble learning of different regionally optimized configurations. Specifically, the minimums and averages of the test performance metrics from the 3 ensemble methods consistently outperformed ERO outcomes across all catchments for the NSE metric. For the KGE metric, particularly in water level predictions, we observed that while K-means Configs did not universally outperform ERO, the other two ensemble learning methods exhibited strong outperformance.

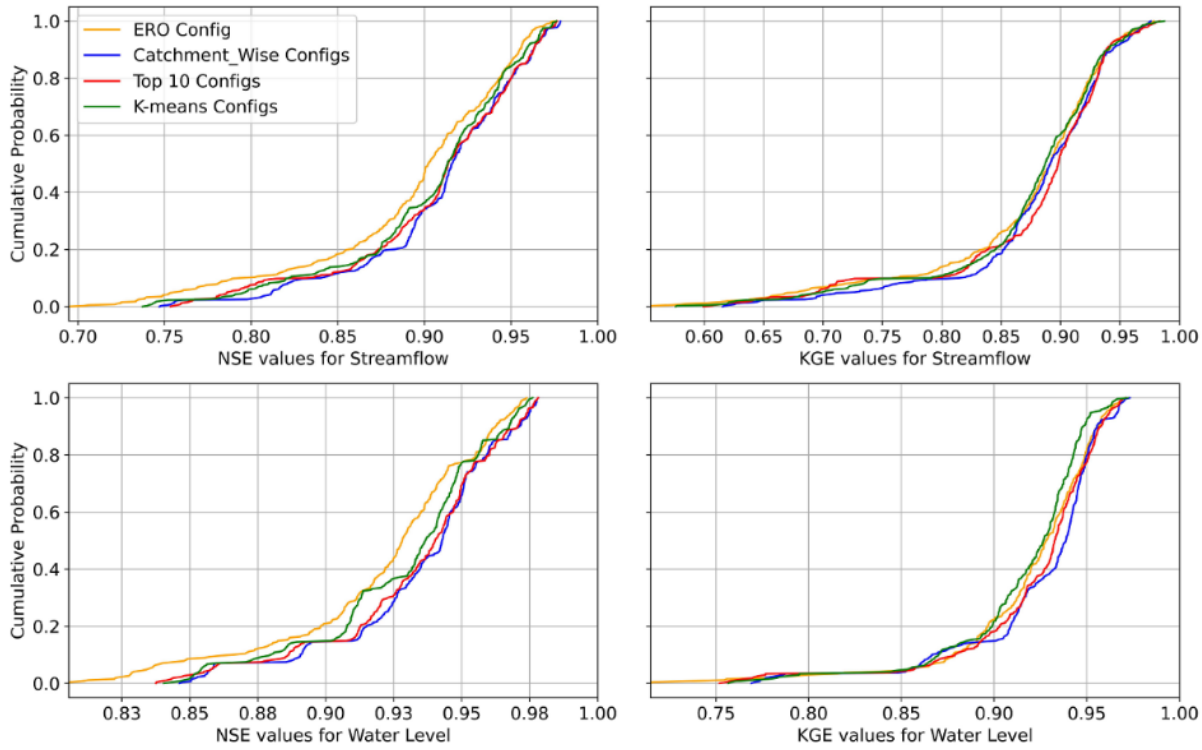


Figure 15. Benchmarking performance metrics of the 3 ensemble learning approaches versus ERO network. The plot depicts CDFs for all 10 different simulations of every method on 10 different random seeds in all 40 catchments.

To gain deeper insights into the catchment-scale performance of the 3 configuration ensembles, we plotted detailed box plots depicting each approach's prediction performance across the 10 random seeds. Figures 16 to 22 present the results for different performance metrics, allowing for an in-depth examination of the learning behaviors exhibited by each approach across different catchments. The figure highlights the outperformance of the ensemble learning approaches in different water basins over the ERO model.

The figures are benchmarking different ensemble learning methods versus RO and ERO performance across various catchments. Box plots demonstrate every method's performance metrics range on the 10 different random seeds. The upper four plots show streamflow performance, while the lower four illustrate water level performance. Streamflow tests cover all 40 catchments, whereas water level tests are limited to 27 catchments due to data availability. Overall, the Catchment-wise Configs ensemble outperformed other methods in more catchments, especially in locations where other methods and ERO showed poor performance. This saying is more accurate for NSE metric and therefore for high flows.

Moreover, across all water basins, Catchment-wise Configs ensemble consistently outperforms the ERO configuration in terms of NSE and KGE metrics for both streamflow and water level targets. This outperformance is particularly evident in catchments where the ERO model encounters prediction complications, such as Abetxuko, Balmaseda, Jaizubia, Ozaeta, Sodupe, and Urkizu. Some of these catchments, like Abetxuko and Balmaseda, are affected by human interventions such as presence of reservoirs, while others, like Ozaeta, Sodupe, Jaizubia, and Urkizu, suffer from data quality deficiencies.

Overall, the figures illustrate a notable trend indicating the outperformance of Catchment-wise Configs among the 3 ensemble learning methods from a catchment-scale perspective in several locations. The blue box plots associated with this configuration ensemble often exhibit narrower lengths and higher values, indicative of more accurate and robust predictions. The narrower lengths suggest greater consistency and robustness in predictions across the 10 random seeds, highlighting the reliability of the Catchment-wise ensemble method.

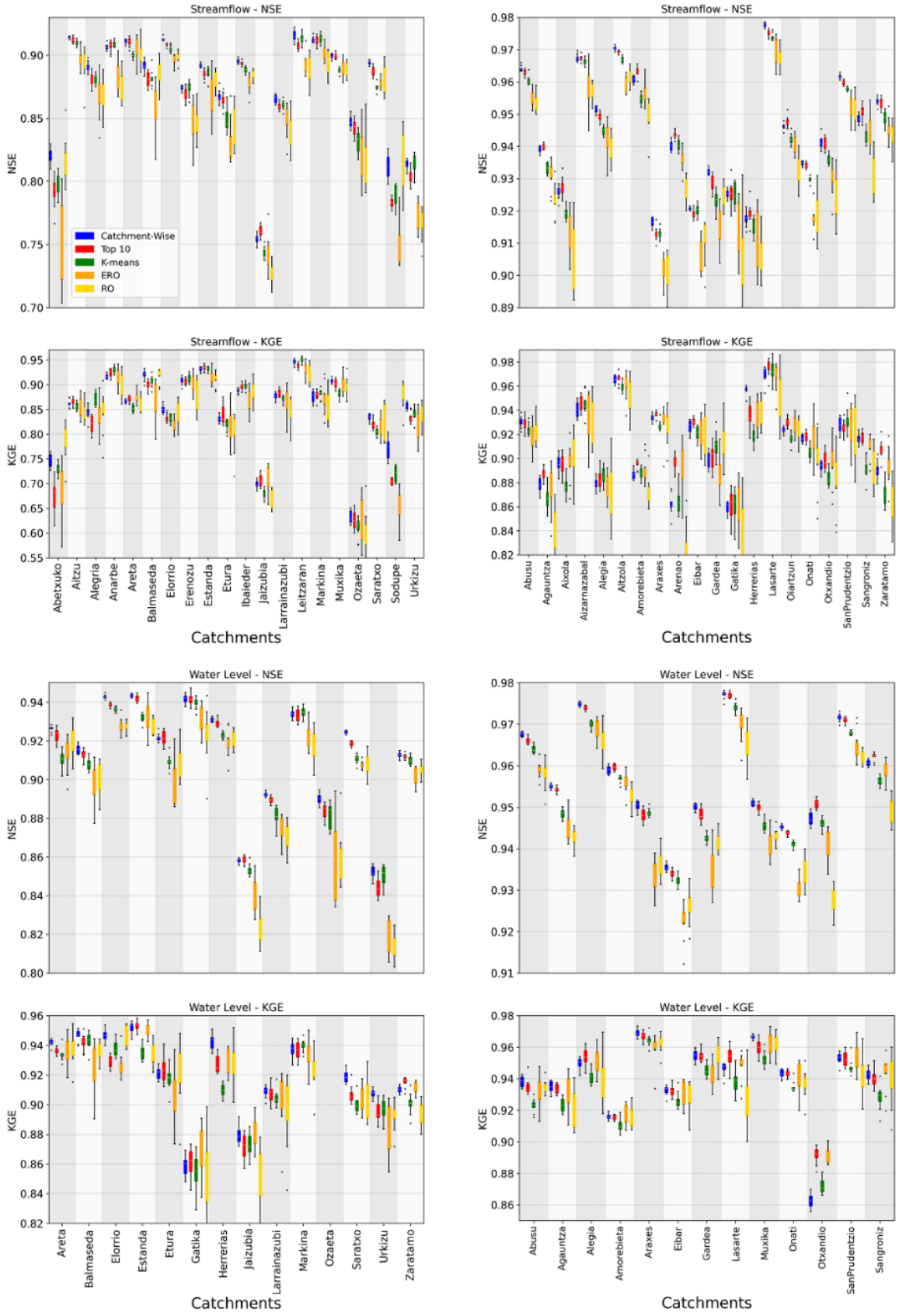


Figure 16. NSE and KGE test metrics

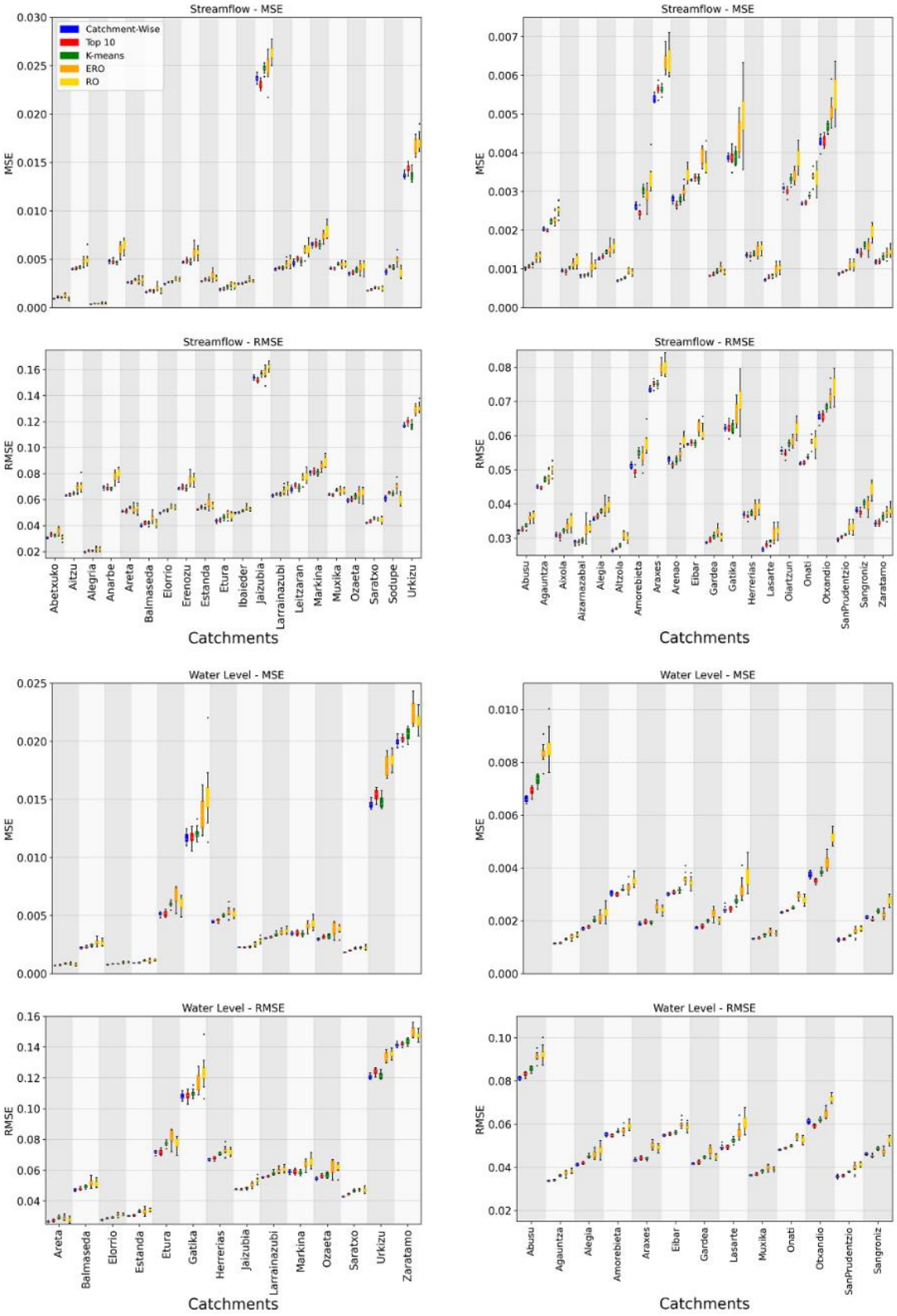


Figure 17. MSE and RMSE test metrics

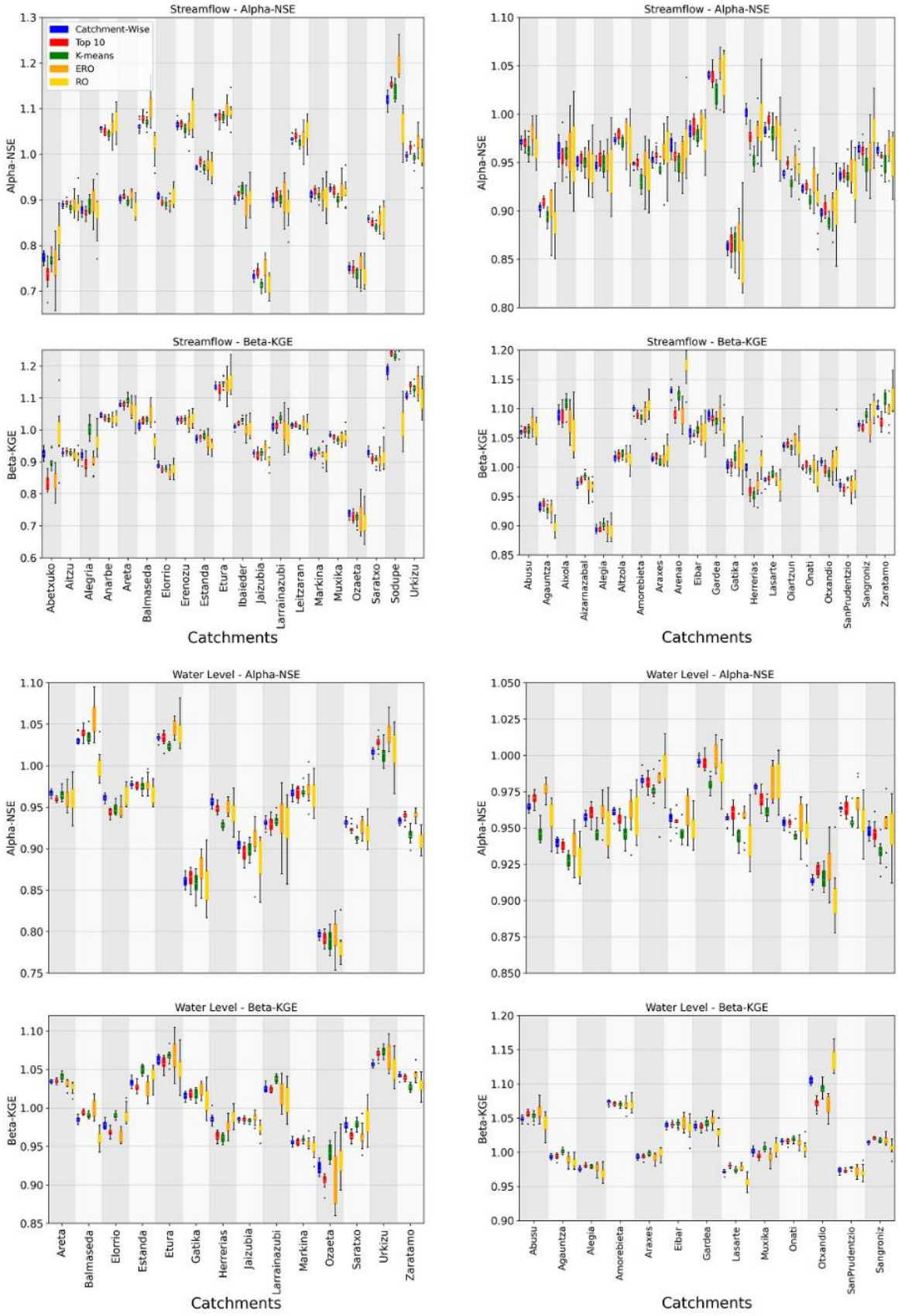


Figure 18. Alpha-NSE and Beta-KGE test metrics

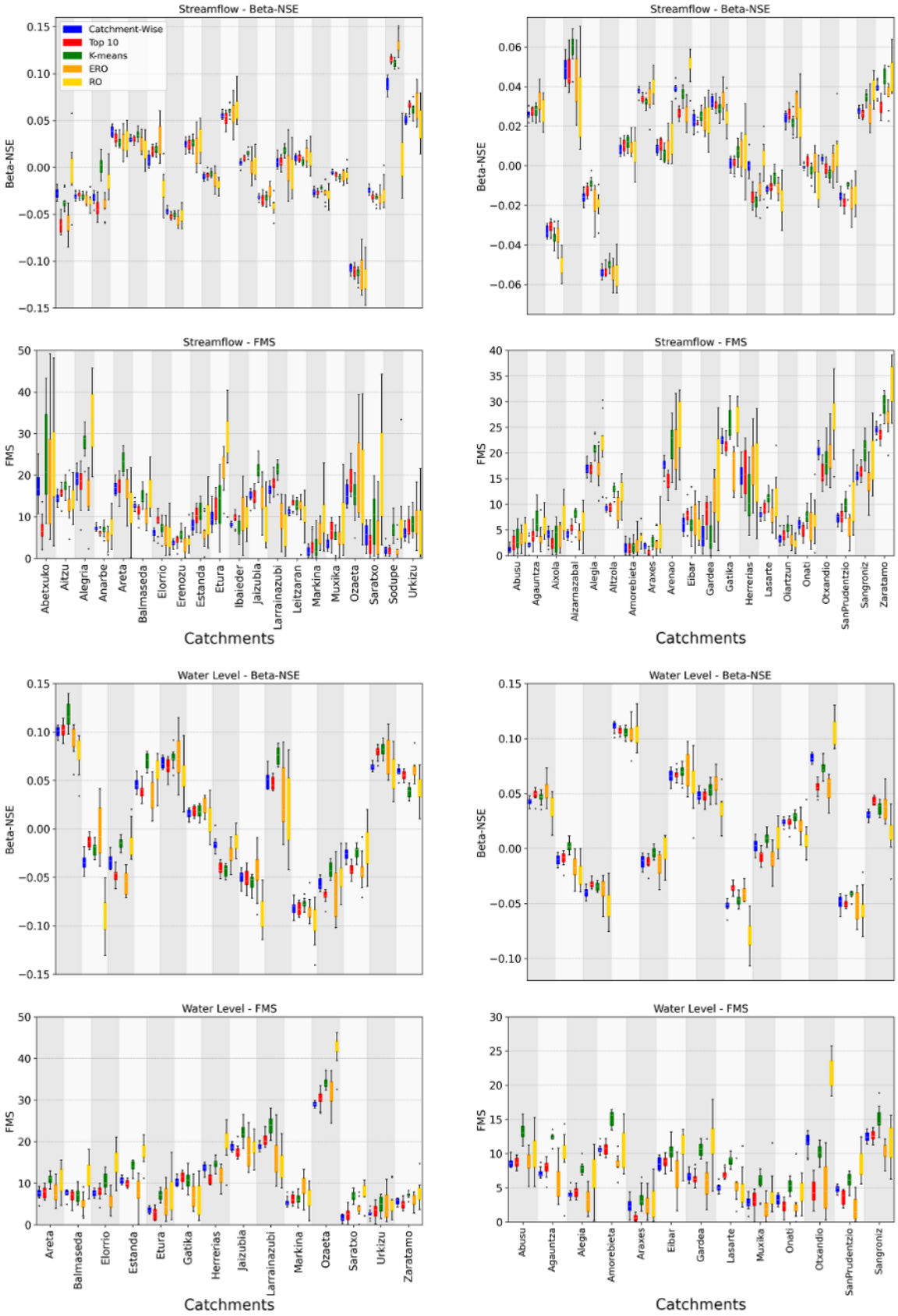


Figure 19. Beta-NSE and FMS test metrics

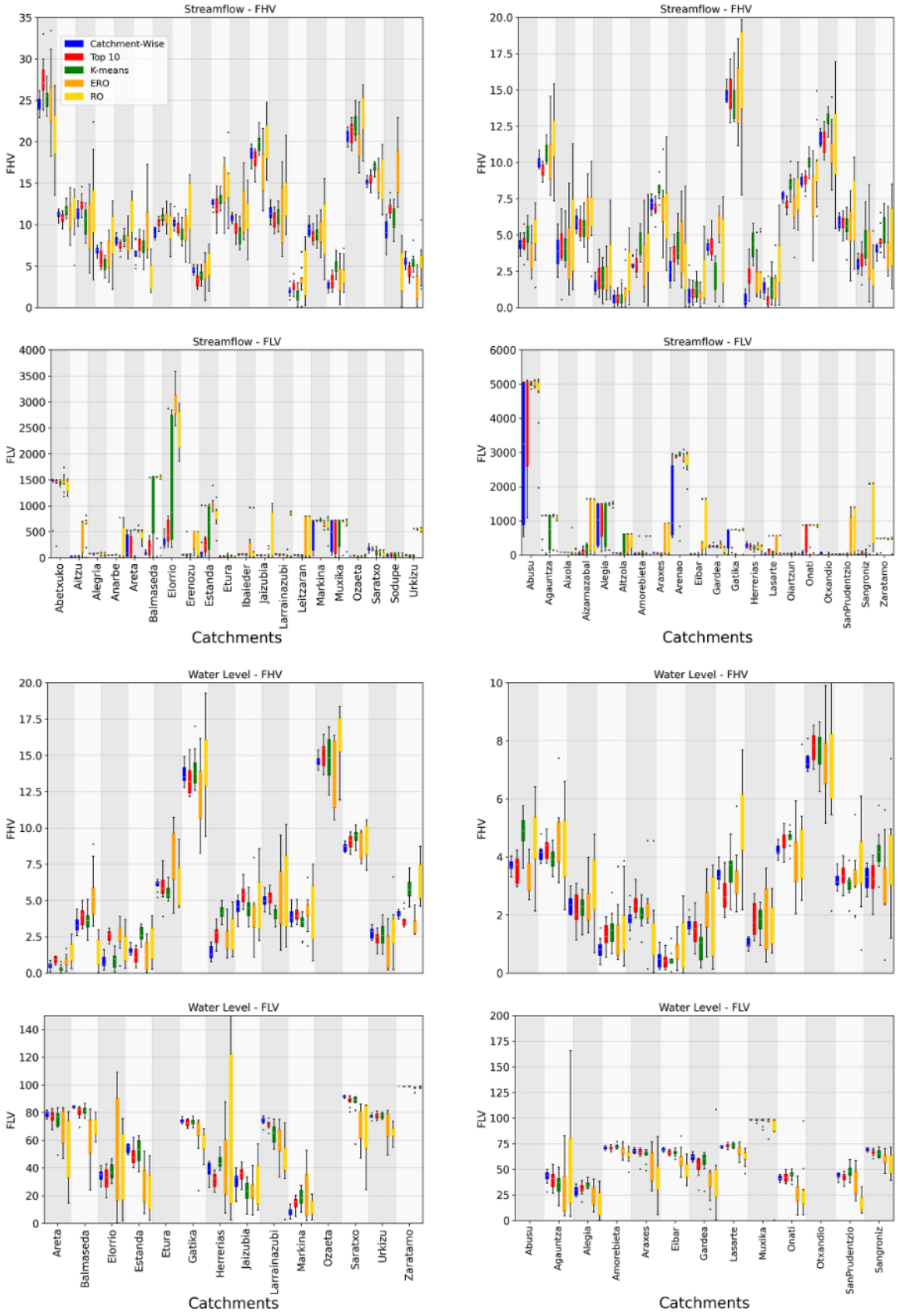


Figure 20. FHV and FLV test metrics

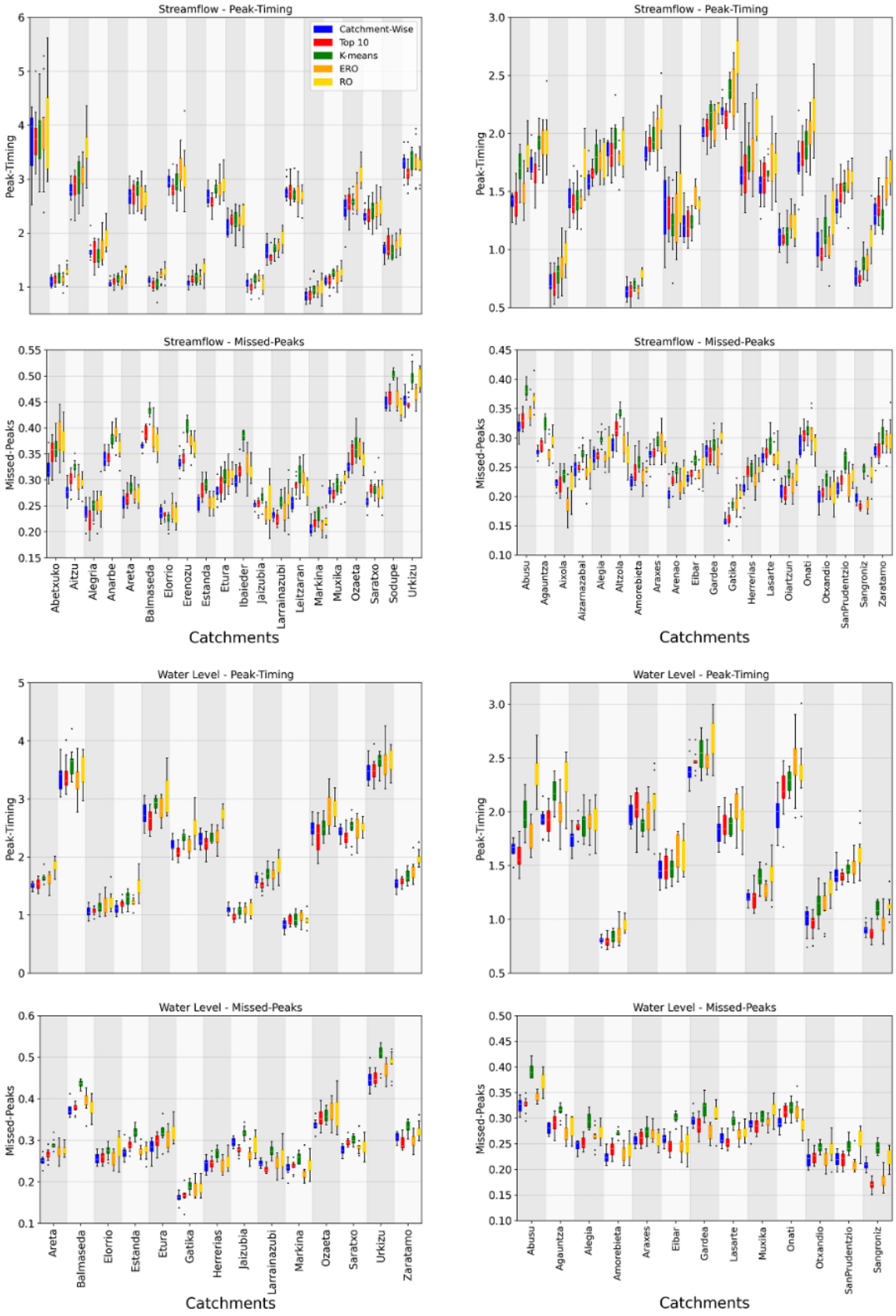


Figure 21. Peak-timing and Missed-Peaks test metrics

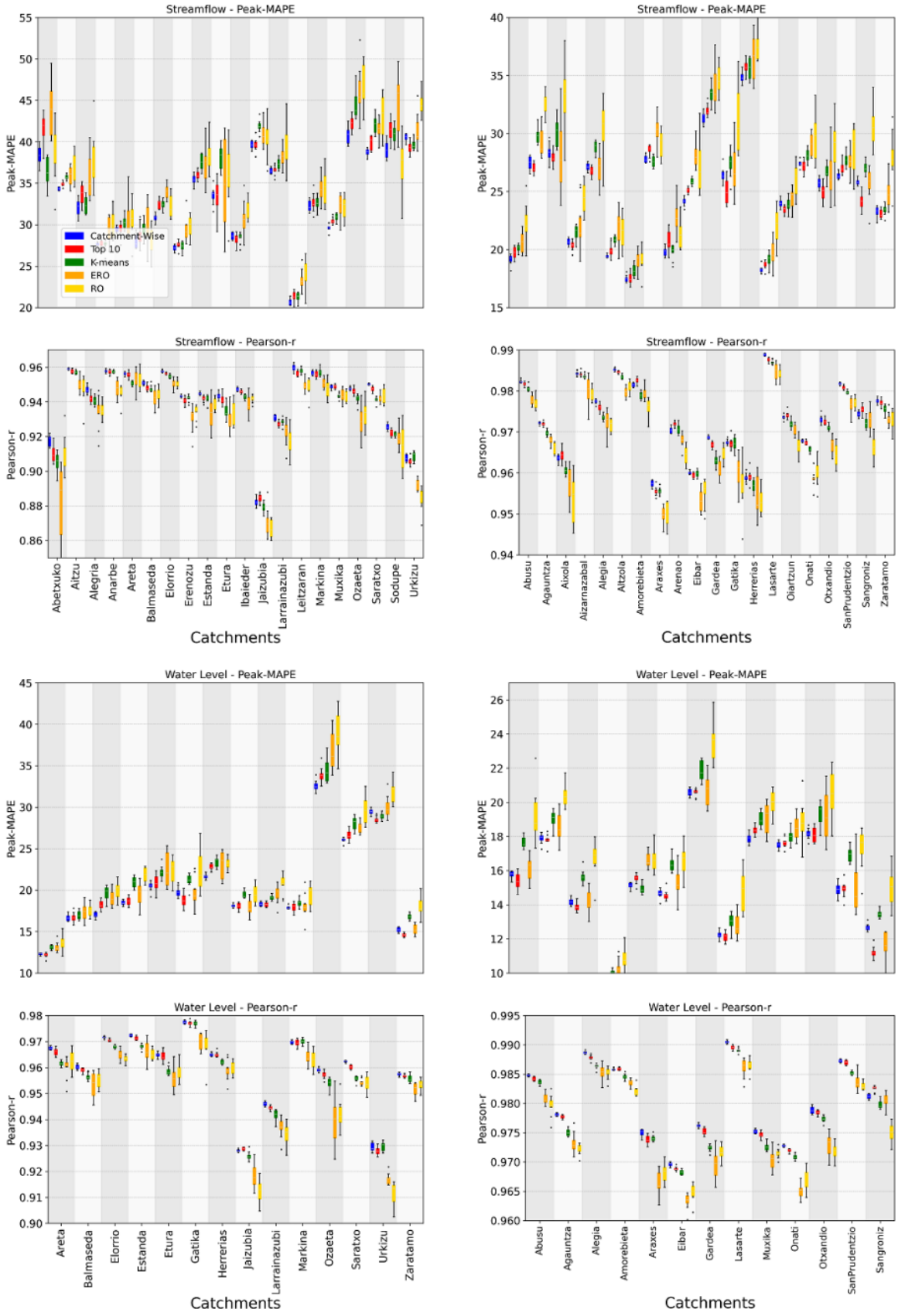


Figure 22. Peak-MAPE and Pearson-r test metrics

5.3.2. Significant disparities in simulations of catchment-wise method

To study the disparities among the catchment-wise ensemble learning method and the 3 other approaches, we conducted statistical analyses using 3 different statistical tests. This section presents the results of these tests to ascertain how significantly different the predictions of catchment-wise ensemble are. To reject the null hypothesis, which claims there is no statistically significant difference, we considered a P-value of less than 0.05.

The results, illustrating the disparities between the Catchment-wise Configs ensemble and other approaches, are depicted in Figure 23. Specifically, the figure confirms the presence of significant statistical differences between the Catchment-wise ensemble learning method and the single configuration method (ERO network), particularly for the NSE metric and high flows.

Moreover, the figure demonstrates that the Catchment-wise Config ensemble outcomes were also statistically significantly different in several instances compared to the other ensemble learning approaches of Top 10 Configs and K-means Configs. This indicates that the different configuration ensemble types exhibited different learning behaviors and provides further evidence of the outperformance of the Catchment-wise ensemble learning approach in more locations.

In summary, the statistical tests, which reject the null hypothesis that differences between unique configurations are random, reveal that various regionally optimized configurations—despite using the same training methods and input data—showed statistically significant performance differences in several catchments, particularly regarding the NSE metric and high flows. This affirms that variations in model performance across certain catchments are not arbitrary but likely result from differences in initial settings and the hyperparameter configurations that affect their learning capabilities. This underscores the importance of meticulous hyperparameter optimization to maximize the performance of LSTMs in regional hydrology.

5.3.3. Hydrographs confirm outperformance of ensemble learning

The sample hydrographs in Figure 24 showcasing some events in different catchments vividly illustrate the advancements achieved through ensemble learning techniques. A comparative analysis of the hydrographs reveals that the Catchment-wise Configs ensemble effectively mitigated the overestimation tendencies observed by other methods for these events. By integrating multiple optimized regional LSTM networks tailored to the unique hydrological characteristics of individual catchments, this ensemble approach significantly improved prediction accuracy.

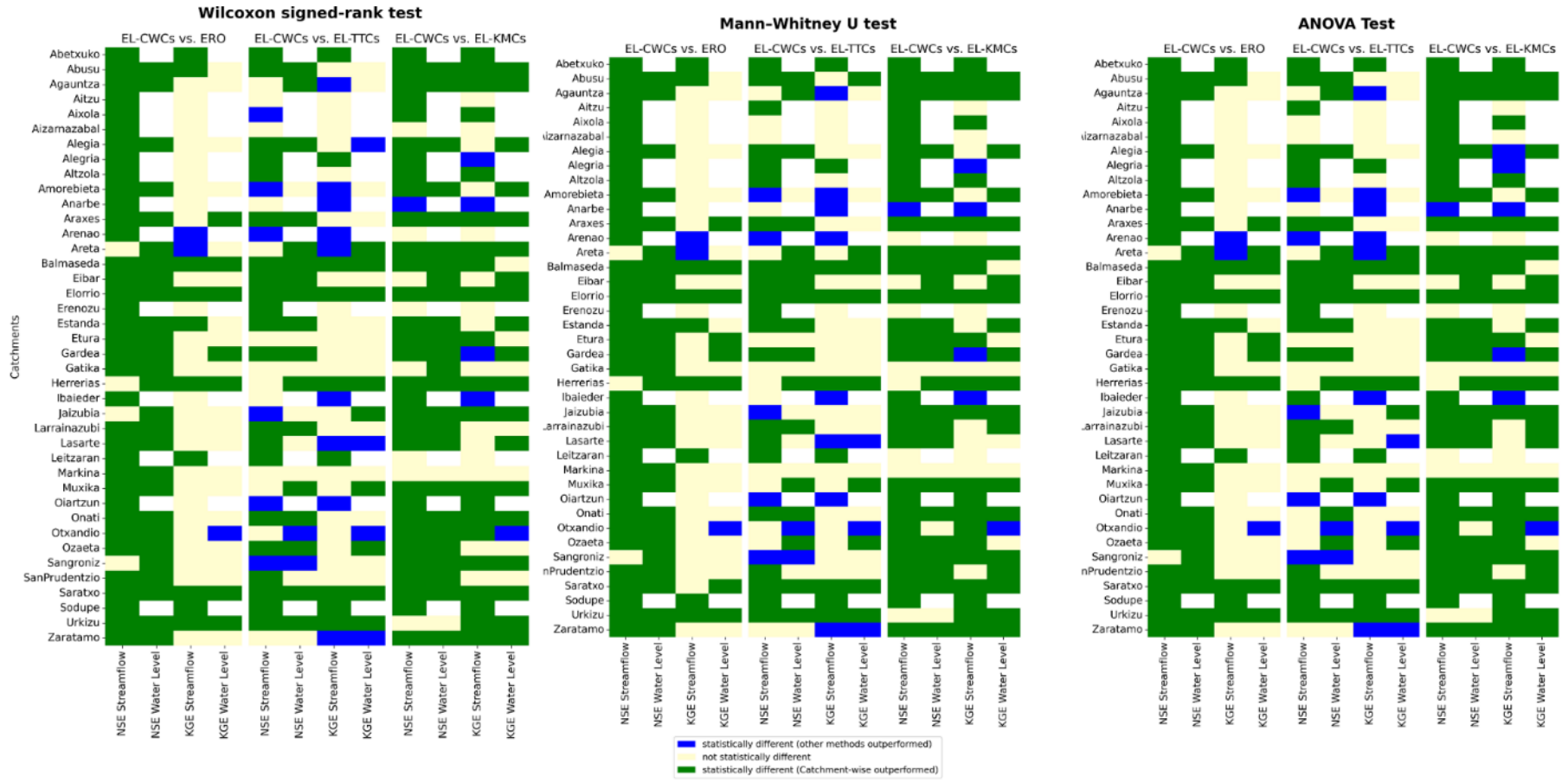


Figure 23. Results of three statistical tests (Wilcoxon signed-rank, Mann–Whitney U, and ANOVA) reveal statistically significant differences between the performance metrics of Catchment-wise Configurations and three other approaches in several instances. The differences are notably better compared to the ERO network. Additionally, the results confirm that the Catchment-wise approach outperformed in more catchments compared to the other two ensemble learning methods.

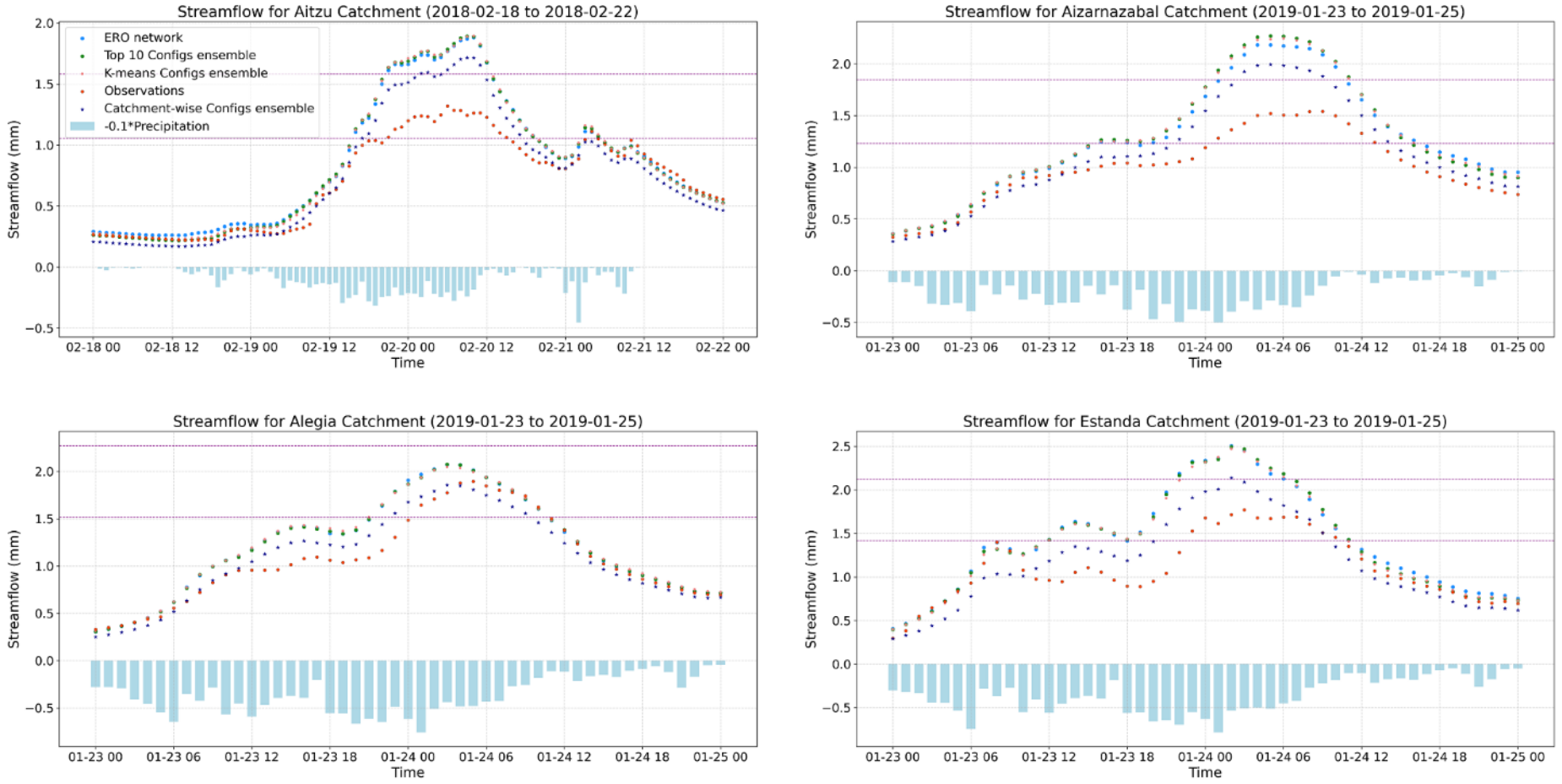


Figure 24. Sample hydrographs showcasing observed streamflow and LSTM predictions for various events and catchments. The predictions compare the observations against the Enhanced Regional Optimal (ERO) network and the three ensemble learning methods: (1) catchment-wise configurations, (2) Top-10 configurations, and (3) K-means configurations. The catchment-wise ensemble significantly reduces overestimation tendencies of LSTMs during peak flow events, yielding predictions that better align with observed values. These findings underscore the outperformance of Catchment-wise Configs ensemble in addressing hydrological variability and enhancing predictive accuracy for regional applications.

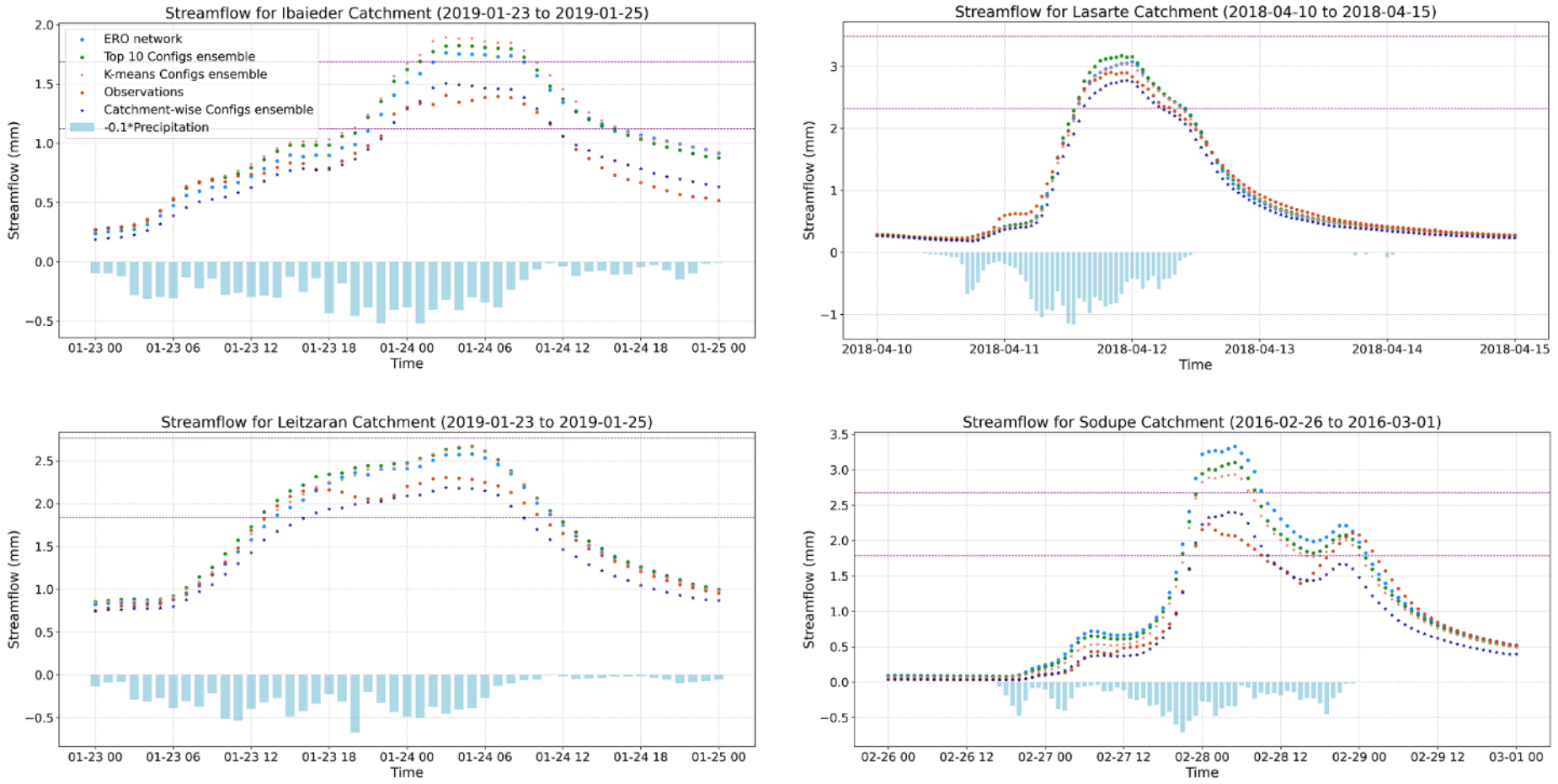


Figure 24 (continued). Sample hydrographs showcasing observed streamflow and LSTM predictions for various events and catchments. The predictions compare the observations against the Enhanced Regional Optimal (ERO) network and the three ensemble learning methods: (1) catchment-wise configurations, (2) Top-10 configurations, and (3) K-means configurations. The catchment-wise ensemble significantly reduces overestimation tendencies of LSTMs during peak flow events, yielding predictions that better align with observed values. These findings underscore the outperformance of Catchment-wise Configs ensemble in addressing hydrological variability and enhancing predictive accuracy for regional applications.

Moreover, the hydrographs underscore the nuanced performance of ensemble learning in handling variability. While other methods exhibited occasional overestimations or underestimations during peak flow events, the catchment-wise ensemble demonstrated a consistent reduction in prediction error. This alignment with observed values highlights the method's capacity to generalize across diverse catchments while preserving the integrity of localized features. Such improvements not only enhance predictive reliability but also provide actionable insights for hydrological applications, especially in flood management and resource allocation in humid flashy catchments like those of the Basque Country.

5.4. Discussion

5.4.1. Regional hydrological artificial intelligent agents

Considering optimized regional prediction LSTM networks as hydrological "artificial intelligent agents" (Russell & Norvig, 2020), this experiment emphasizes the importance of configuration selection in deep neural networks and their corresponding learning habits during the training process. Every deep learning model learns something based on its learning maturity. If we define the learning maturity of a regional hydrological prediction deep learning network as its potential to accurately predict in as many catchments as possible, then "what is learned" might align with our initial goals for the trained networks. This way, if not being aware, during network design and hyperparameter optimization, we may insert our cognitive bias influencing the final performance of the trained models. In general, a fits-all artificial intelligent regional network tries to learn general rules as much as possible and may decide to ignore some unique features in some places. However, some other networks may focus on patterns that we did not intend for them to learn (e.g., human influences in Abetxuko catchment in our case study having two large reservoirs). In regional hydrological prediction by deep neural networks, we encounter the concept of learning maturity level of different designed networks, especially when multiple regional configurations claim to be the best post-random search hyperparameter optimization on the validation set. This study provided evidence that the complexity of deciding on the most mature version among a group of regionally optimized networks can be solved by an ensemble learning approach.

Visualizing the hyperparameter optimization process as an Alpine landscape offers a compelling analogy. Just as each mountain peak offers a unique viewpoint of the surrounding landscape, every configuration in the hyperparameter space provides distinct insights into the behavior of regional hydrological prediction LSTM networks. Our experiment highlights the effectiveness of ensemble learning in navigating this rugged terrain in the hyperparameters space. Rather than fixating on a single peak configuration, ensemble learning harnesses the collective wisdom of diverse configurations, akin to exploring multiple viewpoints from different peaks. This approach significantly enhances prediction accuracy compared to traditional methods, where a solitary configuration is deemed the best-performing regional model on the validation set (e.g., RO and ERO networks). By aggregating predictions from multiple configurations and considering their medians on every time step, the model achieves

a broader understanding of the landscape, resulting in more accurate predictions across a variety of terrains.

Furthermore, the innovative approach of considering the medians of time series predictions at each time step, generated by different the regionally optimized configurations within each ensemble, fosters a form of “democracy in decision-making” among artificial intelligence agents- akin to a “wisdom of the crowd” principle (Surowiecki, 2004). By prioritizing predictions that gain unanimous acceptance across multiple configurations with distinct viewpoints within every ensemble, this method reduces the impact of poor learning habits by some configurations in some places and ultimately enhances prediction accuracy and robustness in all locations.

Our experiment corroborates findings from Wortsman et al. (2022), highlighting the efficacy of ensemble learning methods in enhancing accuracy and robustness. That paper claims that while fine-tuned models may appear to fail to some extent, “averaging the weights of multiple fine-tuned models with different hyperparameter configurations often improves accuracy and robustness.” However, our different approach, centered on the “Alpine-peaks shape” of optimized hyperparameter configurations, offers a novel and more straightforward perspective on ensemble learning in regional hydrological prediction by deep learning models.

5.4.2. Catchment-scale performance evaluation of regional models

Table 7 compares the overall regional performance of the 3 ensemble learning approaches and the ERO network. Analyzing the table confirms that however, the minimum and average performance metrics of ERO were noticeably lower than those of the ensemble learning methods, still by naked eyes we cannot easily discover if and how much meaningfully ensemble learning methods outperform single-configuration of the ERO network. This is in line with what we presented in the introduction. Figures 18 to 23 in the results in conjunction to Table 4, clearly provide evidence that only reporting median (and even average) of test performance metrics in all catchments is not adequate for regional comparative studies (See: Valiela, 2000). This highlights the need for catchment-scale performance evaluations, as demonstrated by this study. This includes plotting results for each catchment and performing statistical tests on the distributions of performance metrics across different approaches and random seeds in each catchment.

Furthermore, Figure 25 offers a more comprehensive comparison of the cumulative distribution functions of various learning approaches for NSE and KGE metrics across the 40 URA catchments on the 10 random seeds. While Catchment-wise Configs ensemble learning generally outperforms Top 10 Configs, a detailed analysis is necessary to determine which approach exhibits higher accuracy and robustness.

Table 7. Comparison of the overall regional performance metrics (average) of different configuration ensembles and the ERO network. Although the Catchment-wise Configurations ensemble slightly outperformed from a regional perspective, the regional metric values are very close to each other that makes hardship to discover which approach outperformed. Even, this proximity can lead to the initial misinterpretation that all four approaches have similar outcomes. However, catchment-scale studies reveal that this conclusion is not true. This table underscores the need to evaluate the performance of different models in specific locations for regional comparative studies to determine if and how one regional model outperforms the others.

Configuration	Streamflow								Water Level							
	NSE				KGE				NSE				KGE			
	Max	Average	Median	Min	Max	Average	Median	Min	Max	Average	Median	Min	Max	Average	Median	Min
Catchment_Wise	0.979	0.910	0.916	0.747	0.976	0.880	0.891	0.615	0.978	0.934	0.943	0.846	0.973	0.924	0.939	0.769
Top_10	0.977	0.906	0.914	0.753	0.983	0.877	0.898	0.599	0.978	0.933	0.940	0.838	0.971	0.922	0.933	0.752
K_means	0.976	0.903	0.913	0.737	0.987	0.872	0.887	0.576	0.976	0.929	0.937	0.840	0.972	0.916	0.927	0.757
ERO	0.975	0.892	0.902	0.695	0.978	0.869	0.891	0.555	0.974	0.921	0.928	0.806	0.973	0.919	0.928	0.715

* Higher values are darker green

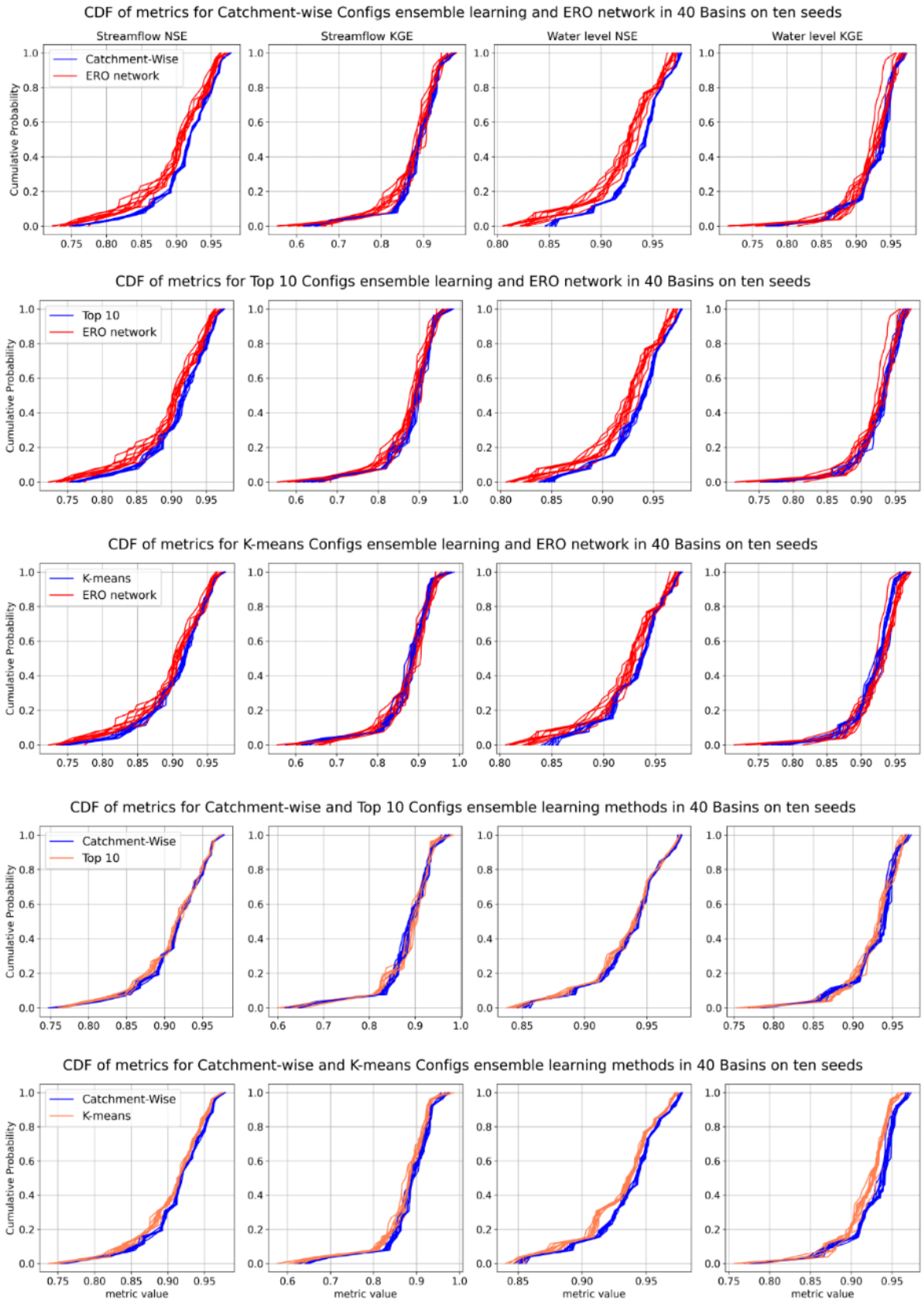


Figure 25. Cumulative distribution function plots for simulations of different configuration settings on 10 different random seeds in 40 URA catchments. The top 3 rows, compare each ensemble learning method with the ERO network prediction outcomes. The 2 lower rows, demonstrate catchment-wise ensemble learning method versus Top 10 and K-means Configs.

Evaluating the nuance improvements in predictions for regional networks proves challenging, especially when different methods demonstrate high accuracy in certain areas and overall. For instance, the benchmarking optimized ERO network, boasting NSE and KGE values up to 0.97 in some locations, sets a high standard in hydrology. Consequently, regional comparisons should focus on areas where improvements are feasible, rather than universally across all catchments or comparing the regional median or average of the performance metrics in the whole area.

5.4.3. Catchment-wise hyperparameter optimization

Configured LSTM networks for hydrological predictions operate in a sequence-to-value mode, where input sequences consist of meteorological inputs from preceding time steps along with the corresponding target discharge values. In our case, the MTS-LSTM network employs 2 distinct hyperparameters for hourly and daily input sequence lengths, as presented in Table 1.

A significant finding of this research is the outperformance of Catchment-wise Configs ensemble over the other ensemble learning methods, particularly the Top 10 Configs. Figures 4, and 5 support this claim. Noticeably, a closer examination of the peak-configurations within the Catchment-wise Configs ensemble (Table 1) reveals specific characteristics, particularly in the length of input sequences, which can hold hydrological significance (Hosseini et al., 2024b).

According to Table 1, while all configurations presented in Top 10 Configs and K-means Configs ensembles employ a fixed 3-year daily input sequence length, the Catchment-wise Configs exhibit variability in sequence durations tailored to individual water basins. This observation underscores the importance of adjusting input sequence durations based on the "uniqueness of the place" paradigm (Beven, 2020). The finding highlights that each catchment may require specific daily and hourly input sequence durations tailored to its unique attributes. In other words, the true input sequence duration of a catchment may encompass crucial hydrological information related to water movement over short and long-term periods within the catchment.

Moreover, the regional daily sequence duration corresponds with the maximum local daily sequence values across all water basins among the Catchment-wise Configs, suggesting a convergence of hydrological characteristics. Overall, the presence of tailored input sequences among the Catchment-wise Configs ensemble underscores the importance of considering catchments' uniqueness in hyperparameter optimization.

5.4.4. pros and cons of Top 10 Configs and K-means Configs ensembles

The K-means clustering method, though competing the ERO network in several place, did not perform as well compared to the other 2 ensemble learning methods. However, according to Figure 15, this approach still outperformed the ERO configuration on the NSE metric for

both streamflow and water level predictions. But the same level of success was not observed for the KGE metric, especially for water level.

On the other hand, the Top 10 Configs ensemble proved to be reliable and provided a straightforward method that was competitive (specifically in terms of computational costs) with Catchment-wise Configs in several catchments but not everywhere. Although Catchment-wise Configs appeared to offer greater accuracy and robustness in more places at the end, particularly in capturing the nuances of individual catchments, the Top 10 Configs ensemble still yielded useful results and in some limited locations outperformed. This suggests that simply considering some of the top-performing regional configurations can yield in a high level of learning maturity by the final ensembles.

While Top 10 and K-means Configs ensembles have their own strengths, it is essential to consider their limitations and the potential for improvement. One promising avenue could be the exploration of alternative approaches, such as clustering catchments based on their geo-hydrological attributes and selecting some of the peak-configurations for each cluster for the final ensemble learning method. By tailoring configurations to specific clusters of catchments, we may potentially enhance prediction accuracy and robustness while reducing computational complexity. However, implementing such an approach would require careful consideration of various factors, including the selection of clustering criteria, the number of clusters, and the practical feasibility of implementing customized configurations for different clusters.

In summary, while Catchment-wise Configs ensemble may offer the highest accuracy and granularity in prediction, the Top 10 Configs ensemble presents a pragmatic and efficient alternative for regional prediction tasks when dealing with a large number of catchments (e.g., CAMELS-type datasets (Addor et al., 2017), or Caravan (Kratzert et al., 2023)). Further research exploring innovative approaches to configuration selection post-random search, such as clustering-based methods, holds promise for improving the performance and applicability of ensemble learning techniques in regional hydrological prediction. These insights could ultimately contribute to development of more effective and reliable hydrological prediction systems, with implications for various water resource management applications.

5.4.5. Disparities in the learning skills of different configuration ensembles

Bergstra & Bengio (2012) underscored the significance of acknowledging uncertainties when determining the best model, particularly in scenarios with a relatively limited validation set. They noted that the uncertainty associated with selecting the optimal model might outweigh the uncertainty linked to assessing the test set performance of any single model. This highlights the importance of considering both model structure error and model selection uncertainty. Model structure error arises from the inherent limitations and assumptions within the chosen model architecture, while model selection uncertainty pertains to the variability in performance across different networks due to hyperparameter tuning. Thus, it is imperative to account for both types of uncertainty when presenting the uncertainty

surrounding the top model identified by a search algorithm. This approach can ensure a more robust and reliable evaluation of model performance, applicable across various experiments wherein multiple models achieve comparable performance post-random search hyperparameter optimization.

The results of the current experiment confirm this notion, particularly when focusing on the overall average performance rank of each of the selected configurations on the validation set in the ensembles. As seen in Table 3, only 3 of the Catchment-wise Configs were among the Top 10 Configs, with the majority of them ranking much lower from this perspective (having higher overall regional validation performance metrics in the DATASET post-random search). However, despite this initial regional performance ranking, the ensemble of Catchment-wise Configs ultimately outperformed.

This discovery challenges a conventional belief to choose the best-performing regional configurations and ignore the potential learning capacity of the others. While Catchment-wise Configs and Top 10 Configs appeared to be competitive on overall, the statistical tests demonstrate the Catchment-wise Configs ensemble's outperformance.

These disparities in performance highlight the complexity of configuration selection in regional hydrological prediction and the importance of considering various factors beyond just counting on the overall regional performance metrics. Factors such as catchment uniqueness, learning behaviors of the configurations in different places, and robustness of the predictions on different random seeds should all be taken into account when evaluating and selecting the most suitable configuration ensemble for regional hydrological prediction tasks. This nuanced approach ensures a more comprehensive understanding of model performance and helps mitigate uncertainties associated with configuration selection, ultimately leading to more reliable and accurate predictions.

5.4.6. Significance of ensemble deep learning in real-life practice

Our research has primarily focused on advancing hyperparameter optimization for deep learning models in regional hydrology. We recognize the critical need to translate these advancements into practical applications for environmental management and water resources planning. We present a robust framework for optimizing LSTMs, significantly enhancing the precision and reliability of hydrological predictions by ensemble learning method. Accurate predictions are crucial for effective water resource management and flood forecasting, specifically in the flashy URA catchments, which are vital for environmental protection and sustainable development in the studied region.

Ensemble learning of regional LSTM networks for hourly predictions can directly assist various environmental management practices in flashy humid catchments. Enhanced prediction accuracy can lead to better flood risk assessments and timely warnings, allowing for proactive measures to protect communities and ecosystems. Improved water availability predictions support efficient water resources allocation, aiding in the maintenance of aquatic habitats and preventing over-extraction from water bodies.

Our ensemble learning approach for hyperparameter optimization ensures that hydrologists can deploy advanced AI models with improved performance in regional hydrology addressing the “uniqueness of the place” paradigm, thereby enhancing environmental management and planning. Integrating these optimized models into existing hydrological workflows will provide stakeholders with more reliable predictions, ultimately leading to better environmental protection strategies. Also, collaboration with local environmental agencies will tailor these models to specific regional needs, ensuring that our research has a tangible impact.

5.4.7. Limitations and challenges in training ensemble configurations

In our research on ensemble learning with several optimized regional LSTMs, we encountered complications while re-training 3 of the selected final catchment-wise configurations. However, the eight configurations in the K-means Configs ensemble were re-trained without issues across all random seeds. Re-training the Catchment-wise Configs ensemble presented challenges, particularly for the configuration associated with the Aixola water basin. This specific configuration demanded high computational resources due to large sequence lengths and batch size, including addressing CUDA memory inadequacy issues. But we overcame this conflict at the end and re-trained the Aixola specific configuration on all random seeds.

Similarly, during the re-training of models in the Top 10 Configurations ensemble, we encountered limited complications with two networks (common with catchment-wise configs for Sangroniz and Onati catchments) failing to re-train on certain fixed random seeds. Despite our several efforts, six experiments did not finish re-training, including models 2807224240 and 905140944, which failed on 5 and 1 of the 10 fixed random seeds, respectively (refer to Table 1 – “Train conflicts”).

Nevertheless, our developed method, which considers medians of the predicted timeseries in every time step by different configurations in every ensemble on every random seed, addressed this computational issue effectively. This approach proved robust, as it did not require the completion of all experiments to function reliably. This resilience is a strength of the proposed method, demonstrating that it remains effective and reliable even when some configurations do not run due to computational constraints.

Chapter VI

A subset of this chapter was presented at EGU24 meeting:

1. Hosseini, F., Prieto, C., Nearing, G., Alvarez, C., and Gauch, M. (2024) Hydrological Significance of Input Sequence Lengths in LSTM-Based Streamflow Prediction, EGU General Assembly 2024, Vienna, Austria, 14–19 Apr 2024, EGU24-571, doi: 10.5194/egusphere-egu24-571.

Significance of Different Hyperparameters

during Optimization of Regional LSTM Networks

6.1. Introduction

One of the central challenges in applying Deep Learning (DL) models, particularly Long Short-Term Memory (LSTM) networks, to rainfall-runoff modeling lies in understanding how hyperparameters influence their predictive accuracy for streamflow and water level. This challenge becomes even more pronounced when scaling these models to a regional level, where the diverse characteristics of multiple catchments introduce complexities that hinder generalization. Although LSTMs have been increasingly utilized in hydrology, the influence of specific hyperparameters on model performance across varying hydrological conditions remains insufficiently explored. A key question for advancing DL-based hydrological modeling is: How can we systematically evaluate the impact and relative importance of hyperparameters in regional rainfall-runoff modeling, especially when addressing the heterogeneity of catchment attributes? Answering this question is crucial for improving the robustness and generalizability of LSTM models across diverse hydrological settings.

The importance of hyperparameter optimization in ML/DL has been extensively documented in the broader artificial intelligence (AI) literature. Seminal work by Bergstra and Bengio (2012) underscored the efficacy of random search over grid search, highlighting its ability to explore hyperparameter space more comprehensively by focusing less on hyperparameters with limited importance. In the context of hydrology, Kratzert et al. (2018; 2019) demonstrated the predictive outperformance of LSTM networks for rainfall-runoff modeling, often surpassing traditional hydrological models. However, much of the focus in the literature has centered on model architecture and data preprocessing, with less attention given to the systematic fine-tuning of hyperparameters, which is critical for optimizing DL model performance across different hydrological conditions.

Existing studies tend to either concentrate on single-objective optimizations or adopt domain-centric solutions that do not fully account for the complexities inherent in regional hydrological modeling, such as catchment heterogeneity and uniqueness (Beven, 2020). A significant knowledge gap remains in understanding how different hyperparameter settings affect model performance under varying hydrological conditions. In many cases, hyperparameter tuning is treated as a technical step, with insufficient consideration of its hydrological implications. This limits our ability to interpret DL model's behavior in real-world applications. With advances in computational infrastructure, there is now a growing opportunity to uncover how these so-called black-box models interpret latent information in hydrological data and, in turn, to leverage their grasped knowledge understanding to outperform traditional models in addressing complex prediction tasks.

This chapter aims to address the aforementioned knowledge gaps by exploring the hydrological significance of hyperparameters in regional LSTM networks. Specifically, we investigate how different hyperparameter configurations affect LSTM performance in predicting streamflow and water levels across 25 validation catchments in the Basque Country, Spain, following the comprehensive random search methodology outlined in Chapter 3.

By leveraging advanced ML techniques such as Random Forest Regression (RF) and Principal Component Analysis (PCA), this study hypothesizes that certain hyperparameters exert a greater influence on model performance during the optimization process, with their importance potentially varying depending on the unique hydrological characteristics of each catchment.

The objectives of this chapter are twofold:

1. To quantify the relative importance of 10 unique hyperparameters (refer to Chapter IV, Table 4) in shaping the performance of optimized DL models.

2. To investigate whether the impact of these hyperparameters varies with catchment characteristics, and to assess whether any of these hyperparameters hold specific hydrological significance.

This investigation aimed to provide a deeper understanding of the hyperparameter optimization process for regional LSTM networks and contribute to the development of more robust and accurate predictive models. The findings of this study not only highlight the importance of systematic hyperparameter optimization in regional hydrology but also pave the way for future research on the interaction between model hyperparameters and hydrological processes.

6.2. Method

To systematically evaluate the impact of hyperparameters on the performance of MTS-LSTM networks in regional rainfall-runoff modeling, we employed an exhaustive random search strategy. Our primary objective was to determine how the optimization of 10 distinct hyperparameters influences LSTM predictions across 25 validation catchments located in the Basque Country, Spain. These 10 hyperparameters were chosen based on their potential to significantly affect model performance, identified through trial and error, as well as expert consultation with experienced LSTM users, as discussed in Chapter 3. The selected hyperparameters included learning rate (with two scheduling stages), dropout rate, batch size, two input sequence lengths (daily and hourly), hidden size, standard target noise, loss function, and a regularization term—each of which is critical to LSTM training and prediction accuracy in regional hydrological contexts (see Chapters 3 and 4 for more detailed discussions). Figure 26 provides an overview of the developed methodology for assessing the influence of these hyperparameters.

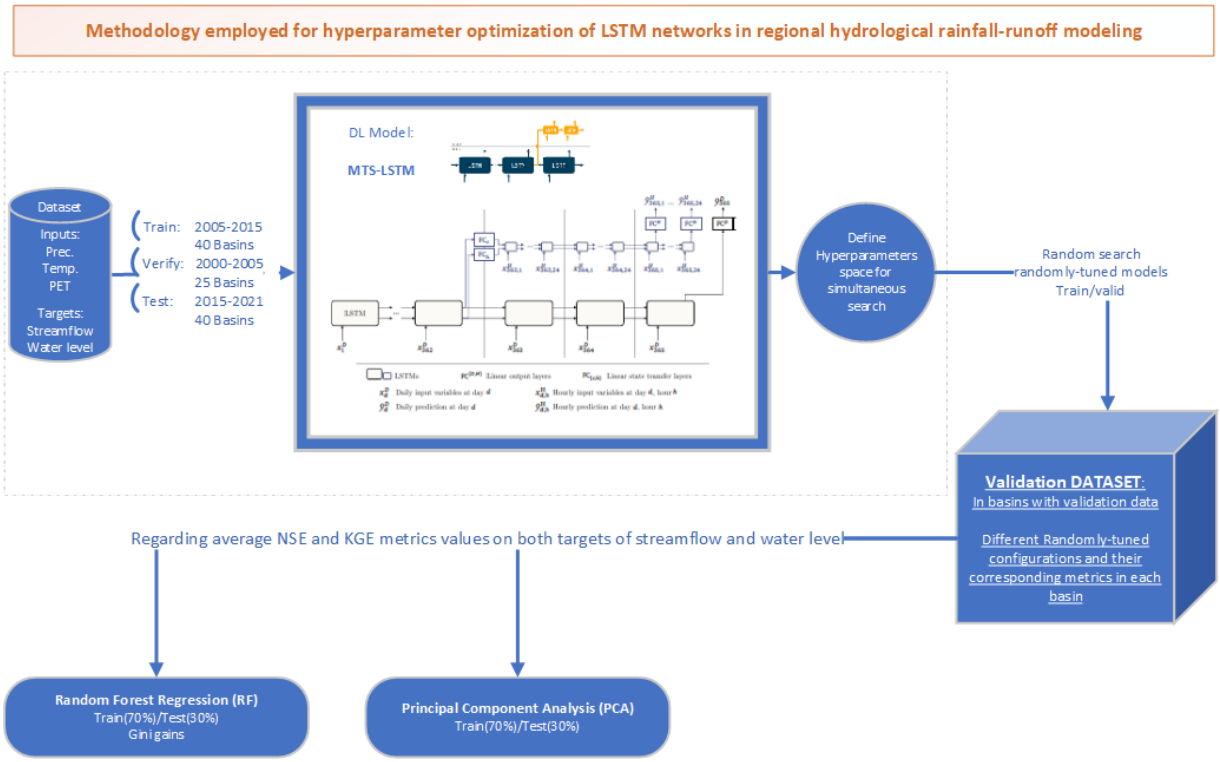


Figure 26. Methodology for Assessing Hyperparameter Importance During Optimization

6.2.1. Generating the Post-Random Search Validation DATASET

To facilitate our analysis, we generated a comprehensive Post-Random Search Validation Dataset (referred to as "val_DATASET" with capital letters), consisting of the hyperparameter configurations of all randomly-tuned LSTMs during the random search process and their corresponding validation performance metrics across the 25 catchments. The performance metrics used in this phase were the Nash–Sutcliffe (NSE) and Kling–Gupta (KGE) efficiencies, calculated for the two targets of streamflow and water level. These metrics were selected as they capture key aspects of hydrological prediction accuracy, including bias, correlation, and variability. This val_DATASET serves as the foundation for understanding how different hyperparameter configurations influence DL models' performance at both regional and catchment scales. To further enrich the val_DATASET, we incorporated attributes of each catchment along with a specific catchment code for every validation catchment.

The random search approach (Bergstra & Bengio, 2012) was employed to avoid the limitations inherent in grid search, where fixed intervals might overlook optimal configurations. Random search allowed for a broader exploration of the hyperparameter space, crucial for evaluating the relative importance of individual hyperparameters. The search involved 1000 different hyperparameter configurations, each of which was applied to the MTS-LSTM network, with the performance metrics for each configuration being recorded.

As described in Chapter 4, the random search resulted in 594 successful experiments out of 1000 iterations. Consequently, the final regional val_DATASET included 1188 records (594 configurations * 2 target types: daily and hourly), representing the average overall

performance across the entire region. The final local val_DATASET contained 29,700 records (594 configurations * 25 catchments * 2 target types: daily and hourly), capturing detailed performance metrics for each validation catchment for every randomly-tuned networks.

6.2.2. Random Forest Model for Hyperparameter Impact Analysis

To quantify the influence of each hyperparameter on model performance, we employed a Random Forest Regression (RF) model (Breiman, 2001). RF was selected for its robustness in handling complex variable interactions and its capability to assess feature importance using Gini gain scores.

We trained separate RF models using 70% of both the regional and local val_DATASETs, allowing the models to learn the intricate relationships between hyperparameter settings of randomly-tuned networks, catchment attributes, and corresponding validation metrics for the configured LSTM networks' prediction performance. The remaining 30% of the val_DATASETs were reserved for testing and validating the trained RF models, ensuring that the models could generalize to unseen hyperparameter configurations and predict their likely performance metrics.

Feature Importance and Gini Gain

Once the RF models were trained, we extracted Gini gain scores to assess feature importance. Gini gain quantifies the contribution of each feature (in this case, the hyperparameters) in accurately predicting the performance metrics of an LSTM network. By analyzing and plotting the Gini gains, we identified the hyperparameters with the most significant impact on the LSTM model's capacity to predict streamflow and water levels accurately in different catchments.

Three Random Forest Approaches

To achieve a comprehensive understanding of hyperparameter importance, we employed three distinct RF modeling approaches, tailored to both regional and local val_DATASETs:

Regional Random Forest Model:

This model was trained exclusively on the overall regional performance metrics (average NSE and KGE values for both targets) for each hyperparameter configuration. The goal here was to evaluate hyperparameter importance across the entire region, without accounting for individual catchment variations. The RF was trained and validated on the regional val_DATASET, focusing on identifying the hyperparameters that most broadly influenced model performance at the regional level.

Catchment-Aware Random Forest Model:

In this approach, the RF model was trained on hyperparameter configurations and performance metrics specific to each catchment, using a catchment code as an identifier. This allowed the model to account for performance variability between different catchments. The

catchment-aware RF model was trained and validated on the local val_DATASET, enabling a deeper analysis of how hyperparameters affect model performance at the catchment scale.

Attribute-Aware Random Forest Model:

This model built on the catchment-aware approach by incorporating detailed catchment attributes (e.g., catchment area, elevation, land use) along with hyperparameter settings. This extension enabled the model to capture how hyperparameter configurations and specific catchment characteristics interact to influence the final performance of LSTM models. The attribute-aware RF model was trained and validated on the local val_DATASET, providing a more nuanced view of the relationship between hydrological context and hyperparameter tuning.

For each of these models, we compared the RF's predictions on the 30% test set to the actual performance metrics derived from the validation process after training the corresponding configured network. Additionally, we visualized the Gini gains for the most influential features, offering a detailed interpretation of each hyperparameter's relative importance in shaping model performance.

6.2.3. Principal Component Analysis (PCA) for Dimensionality Reduction

In addition to the RF analysis, we employed Principal Component Analysis (PCA) to further explore the intricate relationships between hyperparameters and model performance during the random search process. PCA is a statistical technique that transforms a dataset into a set of orthogonal components, facilitating better visualization and interpretation of the influence of each feature (hyperparameters in this case) and, most importantly, to reduce redundancy.

PCA is widely used for dimensionality reduction, feature extraction, and data compression (Prieto et al., 2019; 2021; 2022; Bengio et al., 2013; Jolliffe, 2002). It operates by identifying principal components—combinations of the original features that capture the greatest variance within the dataset. In the context of hyperparameter impact analysis, PCA helps reveal complex relationships and dependencies between hyperparameters and their effects on model performance by analyzing the Post-Random Search Validation DATASET. Specifically, PCA allows for the visualization of the hyperparameter space, providing insights into which hyperparameters have the greatest influence on the optimized LSTM model's performance during random tuning. Here, our objective was to assess the importance of various hyperparameters in rainfall-runoff modeling from a regional hydrological perspective.

We applied PCA to both the regional and local val_DATASETs, focusing on the first 10 principal components, which explained the majority of variance in the data. For interpretability, we generated biplots for the first two principal components, which highlighted the directions and magnitudes of each hyperparameter's contribution. These biplots offer a clear visual representation of how different hyperparameters influence model performance, providing insights into their impact on the randomly-tuned LSTM model's predictive accuracy.

By combining RF and PCA, a complementary approach to analyze the influence of hyperparameters and their possible interactions on the optimized LSTM networks for regional rainfall-runoff modeling was employed. RF provides a quantitative assessment of individual hyperparameter importance, highlighting the most influential factors that dominate the learning process and significantly impact DL model performance. Subsequently, PCA was used to reduce dimensionality, offering a broader perspective on how hyperparameters collectively influenced model behavior, including their redundancy and potential interdependencies. While RF and PCA were applied independently, their combined insights offer valuable guidance for hydrologists when optimizing LSTM networks for hydrological predictions.

6.3. Results

6.3.1. Random Forest models Results

Overall, the trained Random Forest (RF) models provide valuable insights into the effects of various hyperparameters and catchment attributes on the predictive accuracy of Long Short-Term Memory (LSTM) networks in regional rainfall-runoff modeling. The test results of these RF models show highly accepted values for MSE (near zero) and R-squared (near 1) scores between the actual performance metrics of the randomly-tuned LSTM networks and the RF-predicted values across different catchments using the val_DATASET. This relationship is illustrated through scatterplots (Figures 27, 29, and 31) and their relevance feature importance plots (Figures 28, 30, and 32) to enhance the visualization of results. The scatterplots reveal the relationship between the actual performance metrics achieved by the randomly-tuned LSTM networks and the RF models' predictions of likely accuracy for each configured network across different catchments, considering hyperparameter settings, water basin codes, and catchment attributes.

As shown in these figures, R-squared scores approach 1.0 and MSE values are near zero for catchment-aware RFs (either aware of water basin codes or detail catchments' attributes), especially for the NSE metric. The KGE outcomes display slightly lower accuracy, likely due to the nature of the KGE metric, which emphasizes a balanced assessment of correlation, bias, and variability. For regional RF (Figure 31), the overall performance metrics, MSE values are close to zero, but R-squared scores are less satisfactory. Nevertheless, this does not diminish the conclusion that RF models are effective in predicting the potential performance metrics of different randomly-tuned LSTM networks. As seen in the regional RF figure, while the RF predictions of regional overall metrics are not exact, particularly as indicated by R-squared scores, they still capture the general trend effectively, as reflected in the MSE values, also the predicted metrics values are in a well-accepted range.

The corresponding feature importance plots for the RF models highlight the contribution of different hyperparameters and catchment characteristics in shaping the configured LSTM networks' ability to accurately predict streamflow and water levels across the region. By comparing actual performance against RF predictions, these results provide a clearer

understanding of which hyperparameters exert the most influence on the LSTM models' accuracy at both regional and catchment levels. In next subsections, we will discuss each of these trained RFs in more details.

6.3.1.1. Catchment-wise Random Forest Trained on Local val_DATASET with Catchment Codes

Figure 27 illustrates the performance of the catchment-wise RF model trained on the local val_DATASET, which takes LSTM hyperparameters and catchment codes as inputs. The RF model exhibits high predictive accuracy, achieving MSE scores of 0.006, 0.002, 0.009, and 0.001, and R-squared scores of 0.88, 0.99, 0.70, and 0.99 for streamflow NSE, water level NSE, streamflow KGE, and water level KGE, respectively. These scores indicate that the RF model can reliably estimate the performance metrics (NSE and KGE) of a randomly-tuned hyperparameter setting for both targets of streamflow and water level across different catchments with high accuracy.

The scatterplots in Figure 27 exhibit a close alignment with the 1:1 line between the actual performance of the configured LSTM networks in various water basins and the RF model's predictions, underscoring the RF model's ability to predict the LSTM networks' accuracy in a catchment-aware manner. This alignment signifies that the RF model can accurately capture the influences of catchment-specific configurations, confirming its robustness in assessing LSTM network performance across varying catchment conditions.

Figure 28 provides the Gini gain scores (feature importance) for this catchment-wise RF model, highlighting the relative influence of different input features on model performance. The catchment code stands out as the most significant feature, explaining 67%, 61%, 98%, and 99% of the variance in predictive accuracy for streamflow NSE, streamflow KGE, water level NSE, and water level KGE, respectively. This finding underscores the pivotal role of catchment-specific attributes in shaping model outcomes, which supports the notion that hydrological performance of LSTMs is highly contingent on unique catchment characteristics and location.

Additional hyperparameters such as input sequence length and hidden size also play a role but are comparatively secondary. For example, input sequence length contributes 15% and 12% of the variance for streamflow NSE and KGE (daily data) and 5% for streamflow KGE (hourly data), while hidden size accounts for 4% of the variance for streamflow KGE. These results suggest that while LSTM hyperparameters can influence performance, the intrinsic characteristics tied to catchment codes are primary drivers of predictive success in regional hydrological modeling.

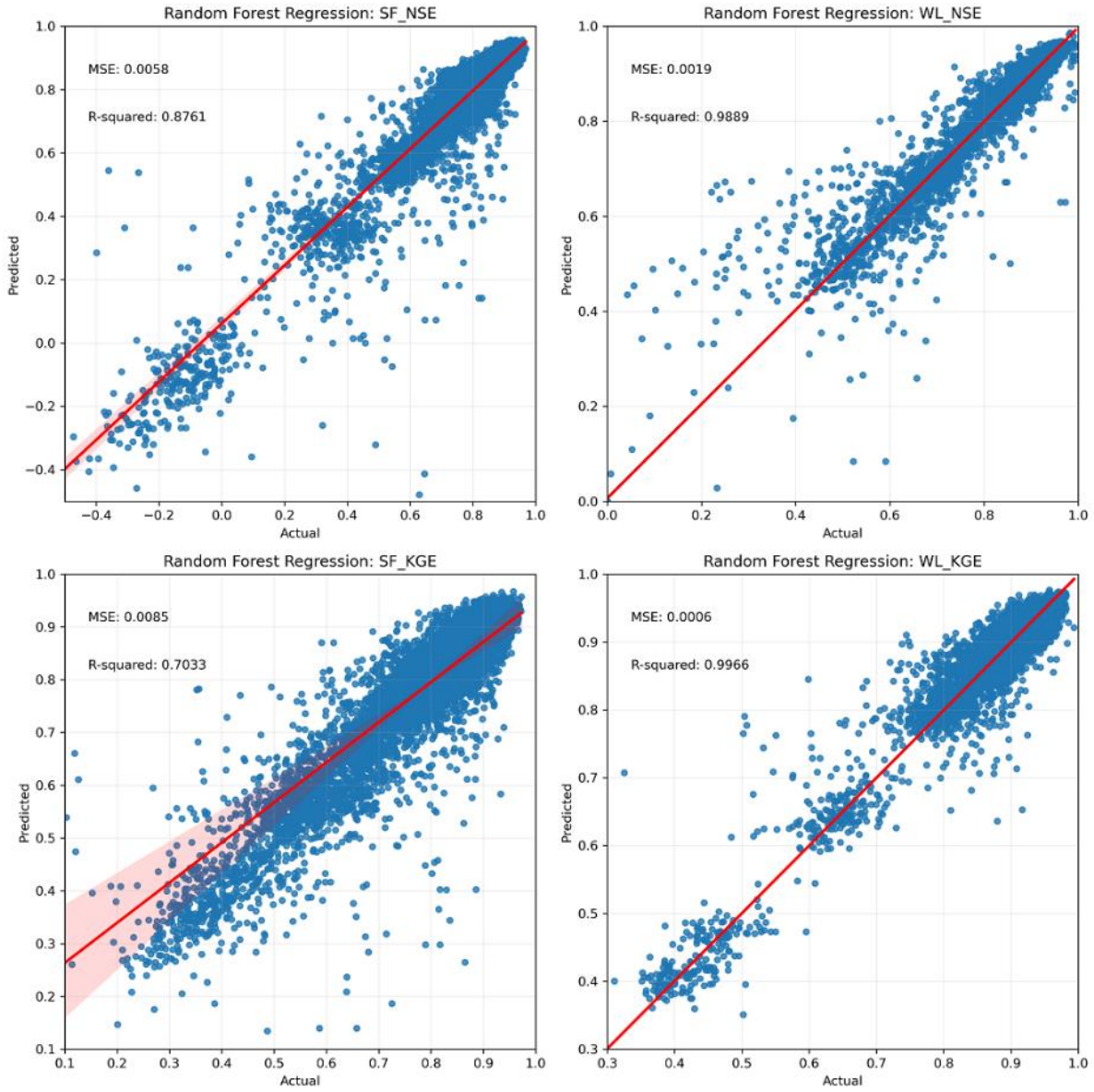


Figure 27. Local Random Forest model applied on local val_DATASET trained on hyperparameters configurations and catchment codes as inputs and performance metrics of each setting in every catchment. SF: streamflow; WL: water level; NSE and KGE: Nash–Sutcliffe and Kling–Gupta efficiency performance metrics.

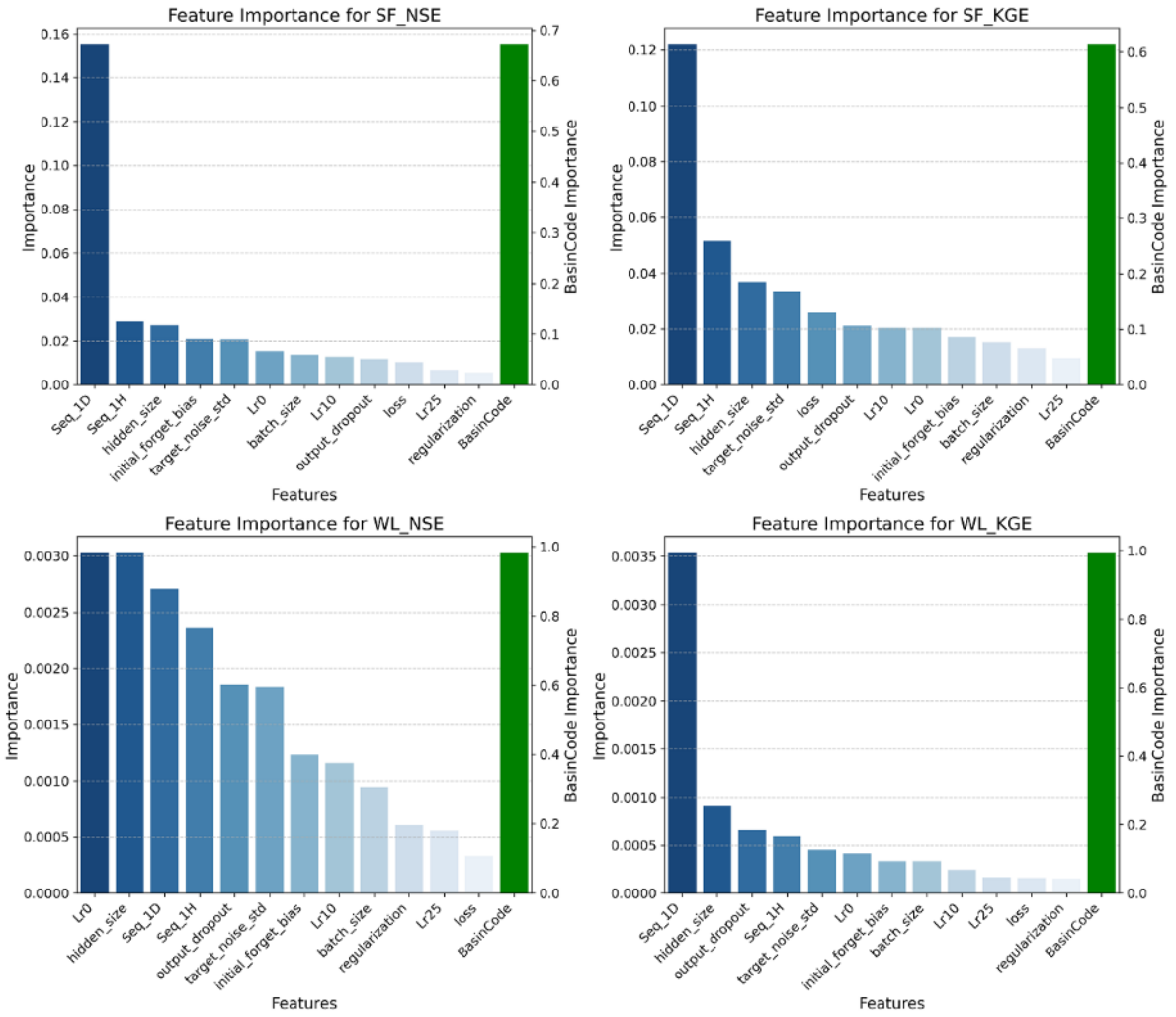


Figure 28. Gini gains show feature importance for the local Random Forest model applied on local val_DATASET trained on hyperparameters configurations and catchment codes as inputs and performance metrics of each setting in every catchment. SF: streamflow; WL: water level; NSE and KGE: Nash–Sutcliffe and Kling–Gupta efficiency performance metrics.

6.3.1.2. Attribute-wise Random Forest Trained on Local val_DATASET with Catchment Attributes

Figure 29 extends the analysis by including detailed catchment attributes alongside LSTM hyperparameters in the Attribute-wise RF model trained on LSTM hyperparameters and detailed catchments attributes as inputs. The RF model demonstrates high predictive accuracy, achieving MSE scores of 0.006, 0.002, 0.009, and 0.001, and R-squared scores of 0.88, 0.99, 0.70, and 0.99 for streamflow NSE, water level NSE, streamflow KGE, and water level KGE, respectively (very similar to RF model aware of catchment codes instead of attributes). These scores reflect the model's robustness in estimating the performance metrics for different catchments. The close alignment with the 1:1 line between predicted and actual performance metrics (as shown in Figure 28) highlights the Attribute-wise RF model's reliability, underscoring the combined influence of hyperparameter configurations

and catchment-specific physical attributes on LSTM accuracy in regional hydrological predictions.

During training, although the actual catchment attributes were not directly exposed to the original randomly-tuned LSTM models, they were aware of catchment codes, as discussed previously in section 6.3.1.1, where the catchment code demonstrated high importance. This indirect awareness may contribute to the Attribute-wise RF model's ability to accurately predict performance metrics based on physical characteristics and LSTM configurations.

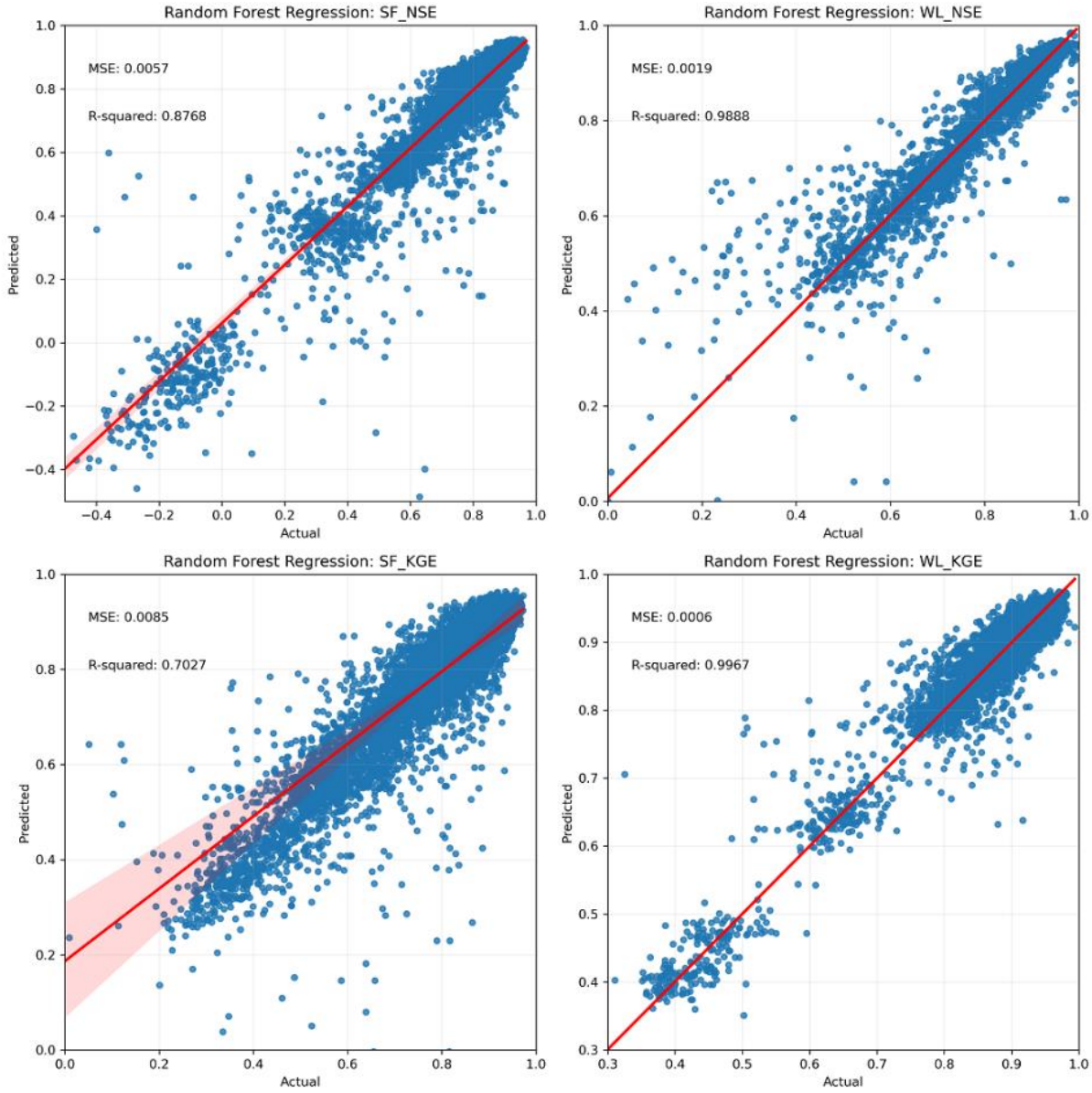


Figure 29. Local Random Forest model applied on val_DATASET trained on hyperparameters configurations and catchments attributes as inputs and performance metrics of each setting in every catchment. SF: streamflow; WL: water level; NSE and KGE: Nash–Sutcliffe and Kling–Gupta efficiency performance metrics.

Figure 30 displays the corresponding Gini gain scores (feature importance) for this attribute-wise RF model, identifying significant catchment attributes that impact LSTM prediction accuracy for hydrological metrics in the humid flashy catchments of Basque Country region. Notably, average yearly potential evapotranspiration accounts for 38% of the

variance in streamflow NSE, while the yearly coefficient of variation of precipitation explains 15% and 20% of the variance for streamflow NSE and KGE, respectively. Additionally, the number of days with negative temperatures per year contributes 19% of the variance for streamflow KGE, suggesting that factors like evapotranspiration and precipitation variability substantially drive streamflow prediction accuracy within this region.

For water level predictions, different catchment attributes emerge as significant influencers. Average yearly precipitation explains 35% and 34% of the variance for water level NSE and KGE, respectively. Other influential factors include average gradient (15% and 17% variance for water level NSE and KGE, respectively) and the surface area covered by wetlands and water ecosystems (18% and 16% variance for water level NSE and KGE, respectively). These results indicate that factors associated with precipitation and landscape features play a crucial role in shaping the predictive performance of LSTM models for water level metrics.

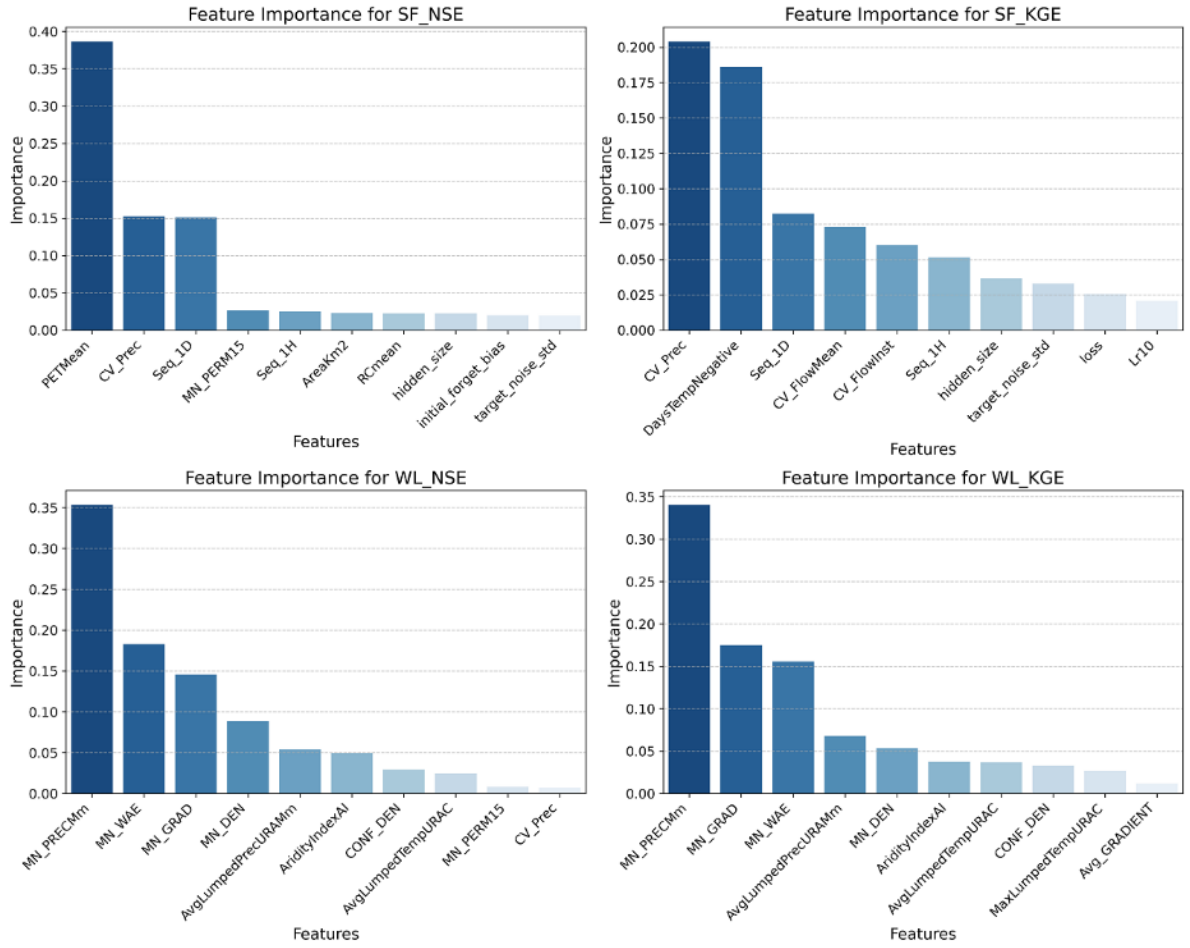


Figure 30. Gini gains show feature importance for the local Random Forest model applied on val_DATASET trained on hyperparameters configurations and catchments attributes as inputs and performance metrics of each setting in every catchment. SF: streamflow; WL: water level; NSE and KGE: Nash–Sutcliffe and Kling–Gupta efficiency performance metrics.

Among the LSTM hyperparameters, sequence length retains moderate importance. Specifically, sequence length (daily) explains 15% and 8% of the variance for streamflow NSE and KGE, respectively, while sequence length (hourly) contributes about 5% of the variance for streamflow KGE. Although hyperparameters like input sequence length hold some relevance, the hydrometeorological data's inherent information on catchment attributes

tends to have a more dominant influence on the LSTM networks' performance predictions across the region. This outcome highlights the significant impact of environmental and catchment-specific characteristics over some LSTM configurations in regional rainfall-runoff modeling accuracy.

6.3.1.3. Regional Random Forest Trained on Regional val_DATASET

Figure 31 presents the results of the regional RF model, which was trained on the regional val_DATASET using only LSTM hyperparameters as inputs to predict the regional average performance metrics for each randomly-tuned LSTM configuration. The RF model shows acceptable predictive accuracy, achieving MSE scores of 0.001 across all metrics (streamflow NSE, water level NSE, streamflow KGE, and water level KGE) and R-squared scores of 0.3, 0.6, 0, and 0.6, respectively. While the R-squared values are relatively low, the general trends between predicted and actual performance metrics remain well captured, as demonstrated in the close alignment between predicted and actual values on the scatterplots. The low R-squared scores are attributed to small variations between predicted and actual values across numerous checkpoints, which does not undermine the regional RF model's capacity to capture overarching performance trends, as evidenced by the consistent MSE values and the well-aligned predicted metrics ranges.

Figure 32 provides a breakdown of the Gini gain scores (feature importance) for this regional RF model, identifying key hyperparameters that influence predictive accuracy across the region from a broader perspective. The most impactful hyperparameters include:

1) **Length of the input sequence**
- **Daily input sequence length**: Accounts for 23%, 16%, 9%, and 15% of the variance in predictive performance for streamflow NSE, streamflow KGE, water level NSE, and water level KGE, respectively.
- **Hourly input sequence length**: Explains around 10% of the variance for streamflow NSE, streamflow KGE, and water level NSE, proving particularly relevant in humid and flashy catchments of the Basque Country.

2) **First learning rate**: Contributes 15%, 11%, 23%, and 14% variance for streamflow NSE, streamflow KGE, water level NSE, and water level KGE, respectively.
3) **Hidden size**: Explains 12%, 11%, 14%, and 17% of the variance for streamflow NSE, streamflow KGE, water level NSE, and water level KGE, respectively.
4) **Output dropout**: Holds a 15% variance for water level KGE.
5) **Initial forget gate bias**: Accounts for approximately 10% variance for streamflow NSE and water level NSE.

These hyperparameters exhibit consistent importance across both streamflow and water level metrics, underscoring their role in guiding the LSTM networks toward optimized performance at a regional level. This insight suggests that even when direct catchment-specific attributes are excluded, certain LSTM configuration hyperparameters substantially impact regional predictions, enhancing model accuracy across varied hydrological metrics. The results underscore that the RF model's ability to accurately predict general trends in

regional metrics, informed solely by hyperparameter configurations, is highly promising for regional applications.

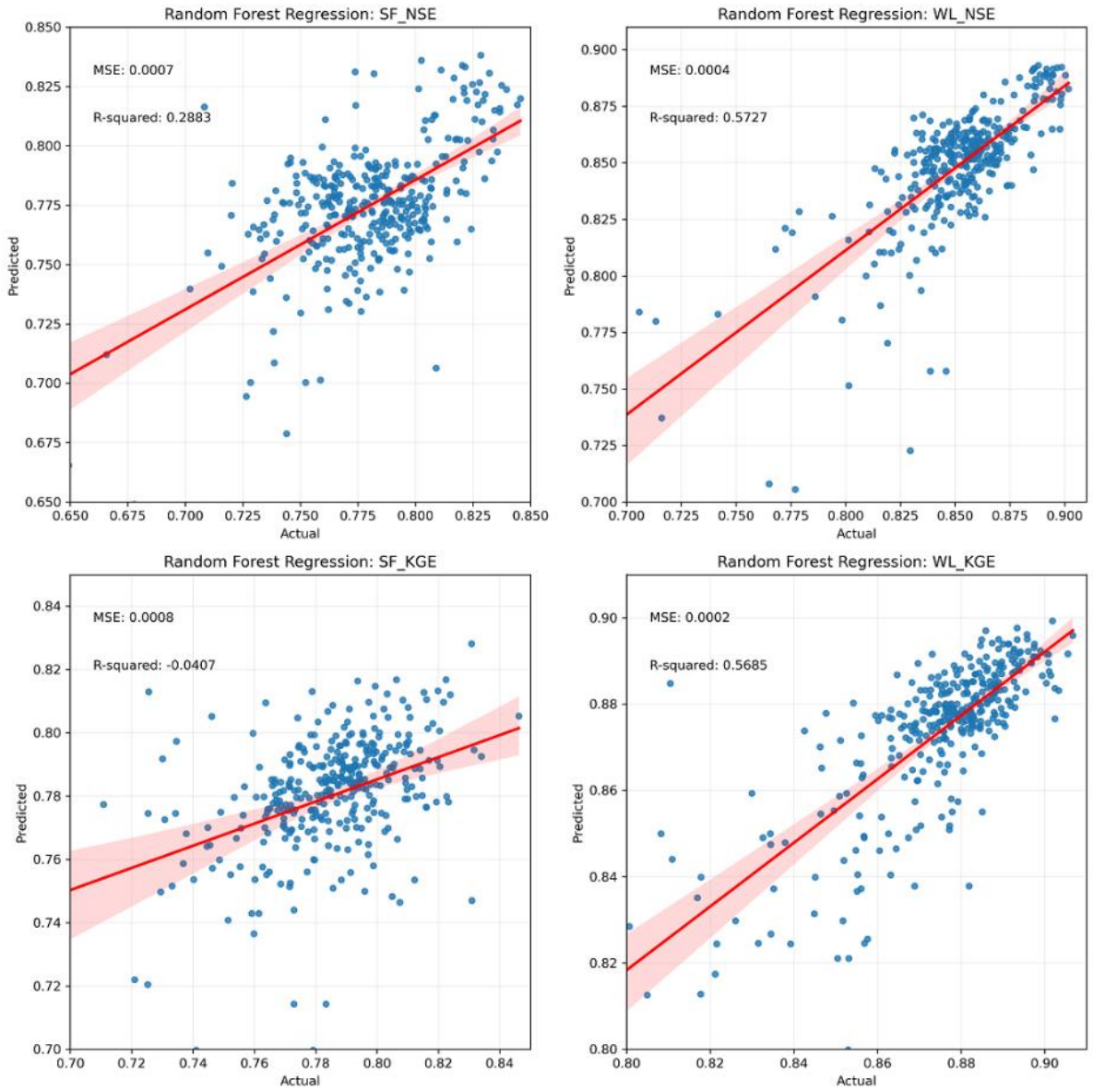


Figure 31. Regional Random Forest model applied on val_DATASET trained on hyperparameters configurations as inputs and the overall average regional performance metrics of each setting in the whole region. SF: streamflow; WL: water level; NSE and KGE: Nash–Sutcliffe and Kling–Gupta efficiency performance metrics.

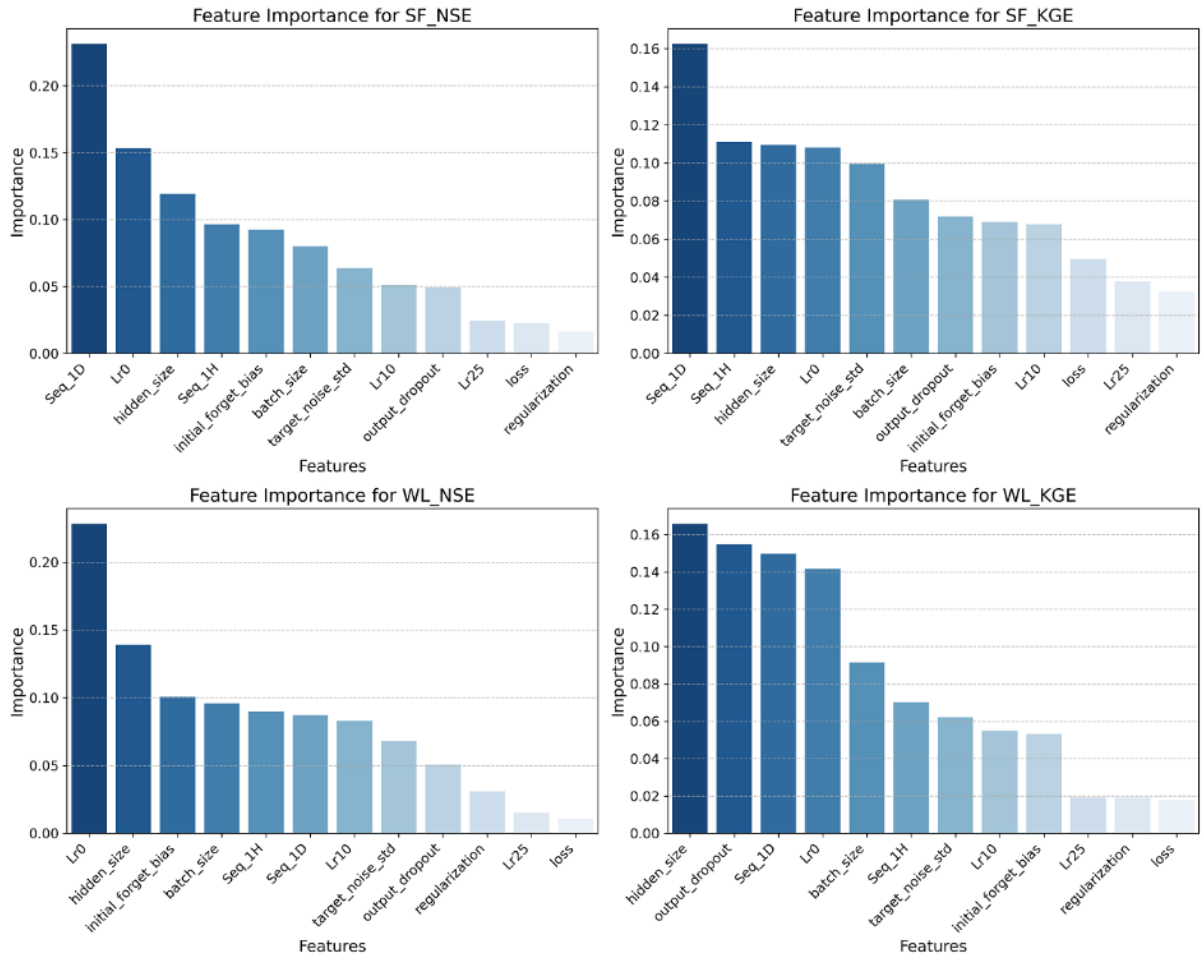


Figure 32. Gini gains show feature importance for the regional Random Forest model applied on val_DATASET trained on hyperparameters configurations as inputs and the overall average regional performance metrics of each setting in the whole region. SF: streamflow; WL: water level; NSE and KGE: Nash–Sutcliffe and Kling–Gupta efficiency performance metrics.

6.3.2. Principal Component Analysis

To further analyze the impact of hyperparameter configurations on the performance of LSTM networks, we applied Principal Component Analysis (PCA) as a dimensionality reduction technique. This approach enabled us to condense the multi-dimensional hyperparameter space (12 in our case) into a lower-dimensional representation while retaining the essential patterns that explain the variability in the predictive performance metrics.

Figures 33 and 34 present the scree plots for the PCAs applied to the regional and local val_DATASETs, respectively. These scree plots illustrate both the cumulative and proportional variance explained by the first ten components of the PCA models, revealing the extent to which the variability in LSTM performance can be attributed to combinations of hyperparameters. The plots are very the same for the two different val_DATASETs, suggesting the same findings by both regional and local PCAs.

The first principal component (PC1) explains approximately 13% of the total variance for both the regional and local PCAs. Among the set of hyperparameters, the input sequence

lengths for hourly data exhibit the highest loadings in PC1, with a value of around -0.5. The load of the input sequence length for daily data in PC1 is also significant at -0.34. This strong association indicates that properly tuning these input sequence length hyperparameters is crucial for enhancing regional hydrological DL model performance.

Moreover, the relatively low percentage of variance explained by PC1 (13%) suggests that no single hyperparameter overwhelmingly dominates model performance. Instead, performance is driven by a combination of factors, with different hyperparameters contributing incrementally to the variability in the final prediction performance metrics. The subsequent components explain smaller percentages of variance, providing a more nuanced understanding of the complex interactions between hyperparameters and the complexity of hyperparameter optimization of regional hydrological LSTMs.

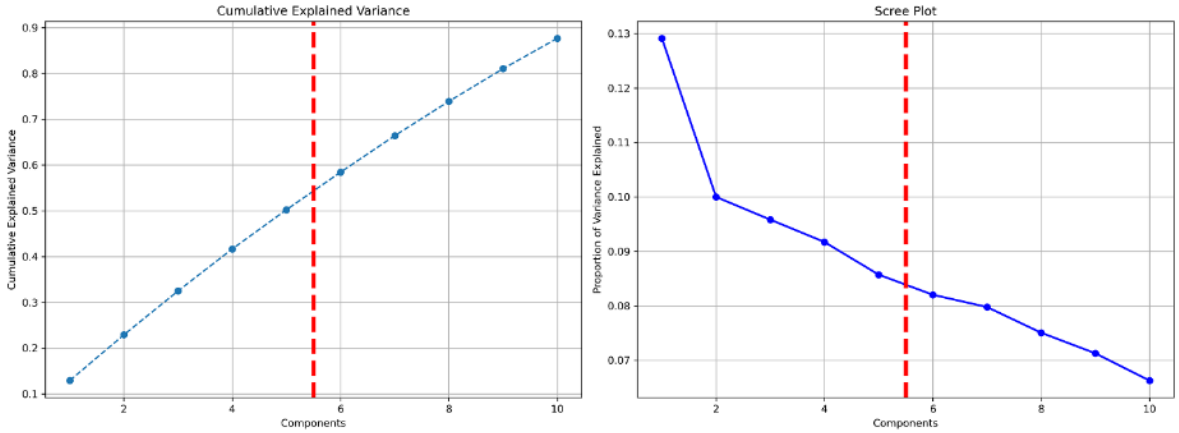


Figure 33. Scree Plots of the Cumulative and Proportional Explained Variances for the Components of PCA applied on regional val_DATASET

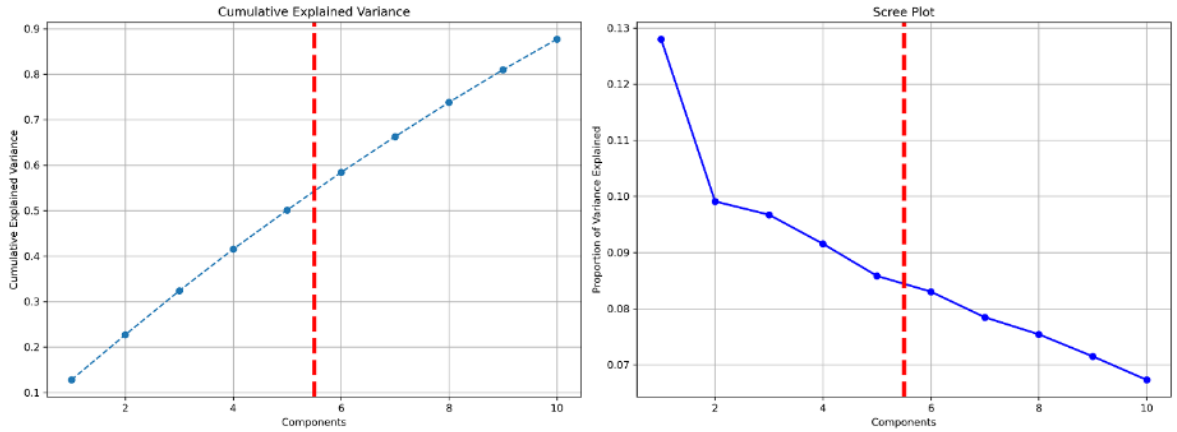


Figure 34. Scree Plots of the Cumulative and Proportional Explained Variances for the Components of PCA applied on local val_DATASET

Table 8 summarizes the detailed explained variance and variable loadings for each principal component, allowing us to identify the specific hyperparameters that most strongly influence each component. For example, in the PCA applied to the regional dataset, PC2 shows notable contributions from hyperparameters related to loss (0.685) and the regularization term (-0.613), indicating their significant roles in shaping model performance, particularly in the context of loss minimization strategies and temporal dependencies in data.

Table 8. PCA Analysis - Explained Variances and Variable Loadings for Principal Components

PCA applied on the Regional val_DATASET														
PCA Components	Hyperparameters												Explained Variance Ratio	Cumulative Variance
	Seq_1D	Seq_1H	batch_size	target_noise_std	Lr0	Lr10	Lr25	loss	hidden_size	output_dropout	initial_forget_bias	regularization		
PC1	-0.340	-0.466	0.112	0.383	0.373	0.050	-0.087	-0.154	0.444	-0.339	-0.022	-0.150	0.129	0.129
PC2	-0.038	0.044	0.051	-0.134	0.091	0.202	-0.101	0.685	0.180	0.195	-0.039	-0.613	0.100	0.229
PC3	0.330	0.010	0.184	-0.133	0.451	0.486	-0.024	-0.100	0.224	0.325	0.387	0.290	0.096	0.325
PC4	0.198	0.098	0.443	-0.160	0.141	-0.469	-0.615	0.115	-0.033	-0.256	0.169	0.060	0.092	0.417
PC5	0.468	0.381	-0.291	0.440	0.157	0.196	-0.214	-0.215	-0.126	-0.243	-0.079	-0.348	0.086	0.502
PC6	0.149	0.078	0.471	-0.118	0.138	0.146	0.107	-0.104	0.018	0.010	-0.818	0.057	0.082	0.584
PC7	0.201	0.123	0.562	0.383	-0.238	-0.142	0.522	0.030	0.072	-0.028	0.304	-0.183	0.080	0.664
PC8	0.322	0.033	-0.303	-0.376	0.305	-0.317	0.464	0.121	0.303	-0.383	-0.016	0.007	0.075	0.739
PC9	0.067	0.171	-0.144	0.120	-0.094	-0.421	-0.138	-0.237	0.577	0.560	-0.135	-0.069	0.071	0.811
PC10	-0.481	0.627	0.000	0.234	0.418	-0.102	0.120	0.217	-0.052	0.016	-0.004	0.266	0.066	0.877

PCA applied on the Local val_DATASET														
PCA Components	Hyperparameters												Explained Variance Ratio	Cumulative Variance
	Seq_1D	Seq_1H	batch_size	target_noise_std	Lr0	Lr10	Lr25	loss	hidden_size	output_dropout	initial_forget_bias	regularization		
PC1	-0.334	-0.482	0.132	0.363	0.409	0.069	-0.067	-0.184	0.411	-0.316	-0.046	-0.161	0.128	0.128
PC2	0.089	-0.063	0.318	-0.279	0.255	0.343	-0.031	0.615	0.178	0.253	0.171	-0.350	0.099	0.227
PC3	0.137	-0.075	0.270	-0.171	0.274	0.236	0.018	-0.355	0.101	0.249	0.361	0.643	0.097	0.324
PC4	0.330	0.130	0.293	-0.170	0.099	-0.397	-0.664	-0.001	0.050	-0.378	0.079	-0.008	0.092	0.415
PC5	0.418	0.281	0.333	0.196	0.277	0.231	0.191	-0.208	-0.106	-0.011	-0.603	-0.108	0.086	0.501
PC6	0.413	0.175	-0.480	0.276	0.056	0.448	-0.130	-0.111	0.058	-0.204	0.416	-0.203	0.083	0.584
PC7	0.295	0.134	0.042	-0.085	-0.065	-0.416	0.569	-0.054	0.568	-0.085	0.195	-0.119	0.078	0.663
PC8	0.078	-0.191	-0.393	-0.658	0.080	0.224	0.033	-0.012	0.185	-0.335	-0.381	0.141	0.075	0.738
PC9	0.061	-0.243	0.370	-0.065	-0.240	0.178	0.349	0.053	-0.428	-0.581	0.249	-0.006	0.072	0.810
PC10	0.105	-0.082	-0.289	0.054	0.677	-0.365	0.219	0.262	-0.404	-0.045	0.092	0.118	0.067	0.877

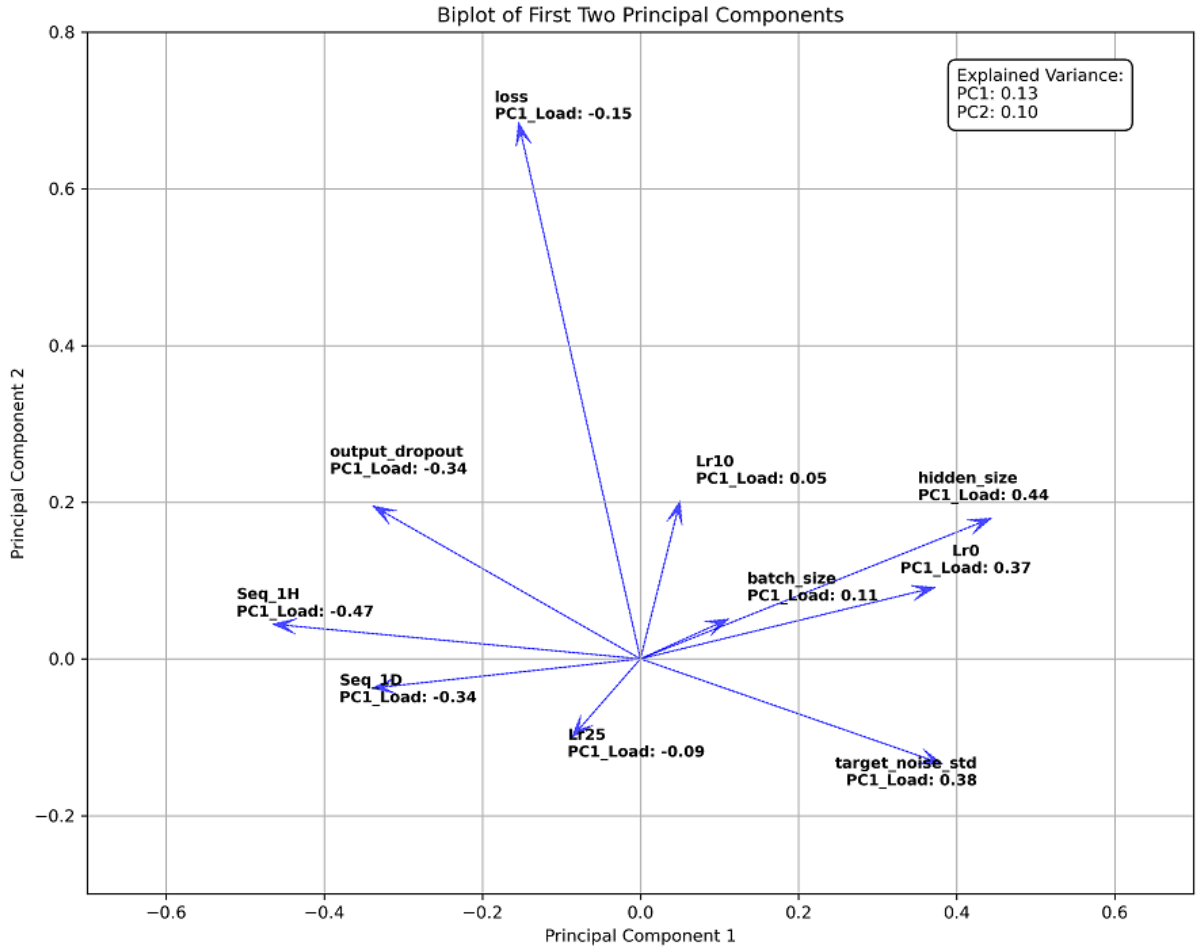


Figure 35. Biplots of Principal Component Analysis (PCA) applied on the regional val_DATASET for Hyperparameters and Validation's Metrics.

The analysis of PCA loadings, visualized in Figures 35 and 36, reveals that the loadings for the first two principal components are similar for both the regional and local val_DATASETs. This consistency suggests that, regardless of the scale (regional or local), the same set of

hyperparameters tends to drive model performance. Specifically, hyperparameters such as daily input sequence length, the first learning rate, and hidden size consistently rank among the top loadings for both val_DATASETs, underscoring their overall importance in the modeling process.

Additionally, the loadings associated with PC3, PC4, and PC5 in both datasets highlight the importance of hyperparameters such as hourly sequence length, batch size, the scheduled second and third learning rates, initial forget gate bias, dropout rate, and standard target noise deviation, indicating their potential impact on model robustness and generalization capabilities. This suggests that careful consideration of different hyperparameters can enhance LSTM performance under varying hydrological conditions.

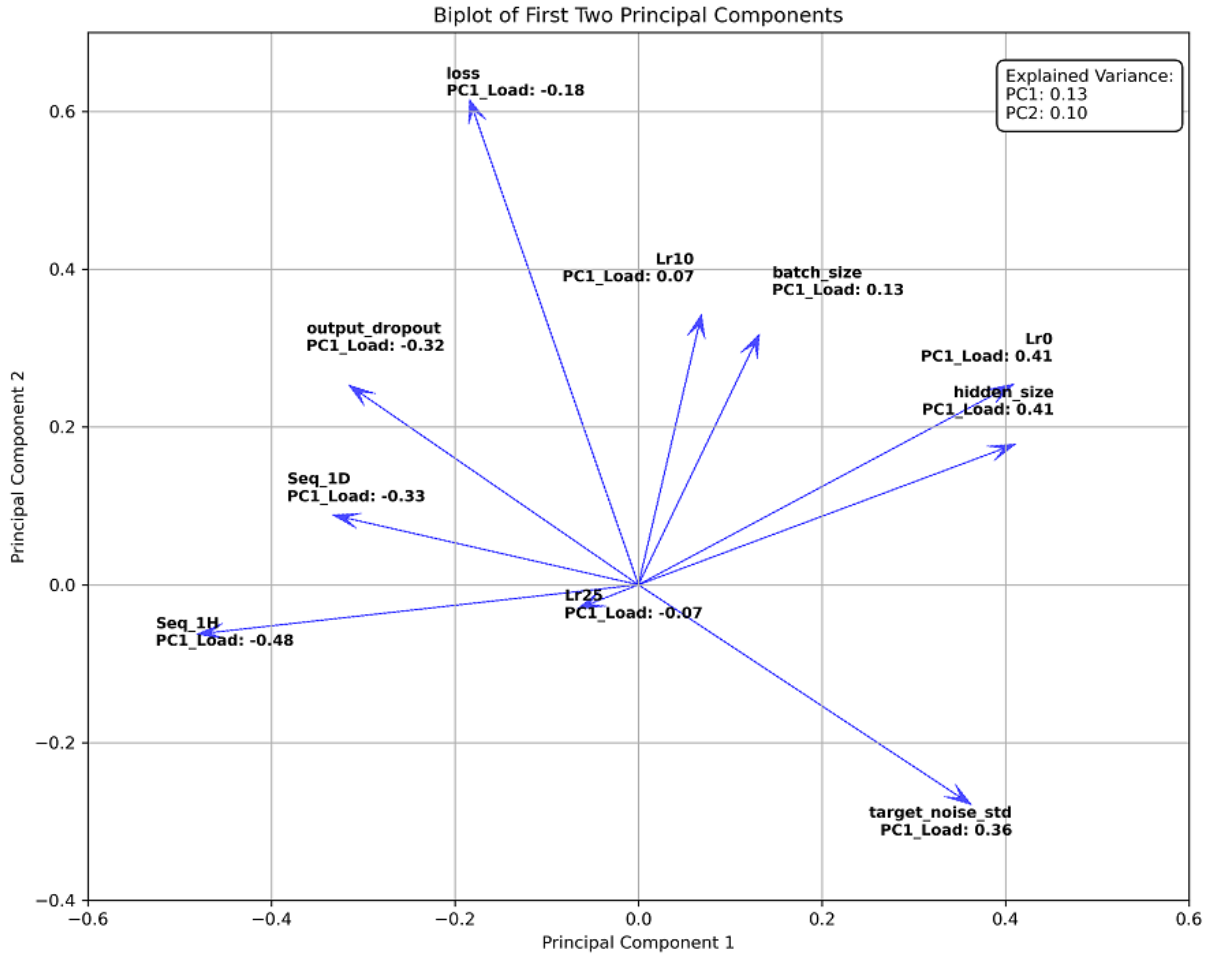


Figure 36. Biplots of Principal Component Analysis (PCA) applied on the local val_DATASET for Hyperparameters, catchments attributes and Validation's Metrics.

6.4. Discussion

Overall, the results indicate that carefully tuned LSTM networks, guided by systematic hyperparameter optimization through random search, efficiently capture the complex dynamics of hydrological systems, leading to accurate catchment-scale predictions for streamflow and water level using regional deep learning models. The significant roles of water

basin codes and detail catchment attributes underscore the spatial variability in model performance, a factor that must be considered when deploying LSTM networks for regional hydrological forecasting. These findings highlight the potential of optimized LSTM configurations to offer precise and robust predictions, adaptable to the unique hydrological characteristics and variability of different catchments, thus enhancing the accuracy of regional forecasts.

The results also underscore the stability and consistency of certain key hyperparameters in shaping LSTM networks' behavior across spatial scales. Principal component analysis (PCA) findings reinforce that, while catchment-specific characteristics are crucial, certain hyperparameters can be optimized to improve model performance at both local and regional levels. Some hyperparameters may even hold hydrological significance; for example, the input sequence length can provide meaningful insights into hydrological processes.

The combined insights from the Random Forest (RF) models and PCA offer a comprehensive understanding of the roles that hyperparameters and catchment attributes play in regional hydrological deep learning modeling. RF analysis demonstrates the ability to predict the performance of various LSTM configurations with high accuracy, especially when incorporating catchment-specific information. In parallel, PCA reveals how distinct hyperparameters consistently drive the performance of optimized deep learning models, enabling more targeted optimization in hydrological modeling.

By integrating Random Forest and PCA findings, this analysis provides a holistic view of how model configurations and catchment attributes influence predictive performance. This approach establishes a strong foundation for the hyperparameter optimization and deployment of optimized regional LSTM networks to meet the challenges of complex hydrological forecasting tasks. In the next subsection, we will discuss the findings in more detail.

6.4.1. Interpretation of the outcomes

The scatterplots (Figures 27, 29, and 31) display strong regression alignments, demonstrating a high degree of agreement between the actual performance metrics of the configured LSTM networks and the predictive estimates provided by the trained Random Forest (RF) models based on hyperparameter configurations and catchment-specific information. This close alignment suggests that regional hydrological LSTM networks are highly sensitive not only to hyperparameter optimization but also to the unique geohydrological characteristics of each catchment.

These results underscore the remarkable capacity of LSTM networks, and by extension, deep learning models, to encode and learn catchment-specific hydrological behaviors. This learning can occur either directly during training from catchment attributes or indirectly through latent patterns embedded in the hydrometeorological input data. In our study, we adopted the latter approach by training LSTMs solely on lumped precipitation, temperature, and evapotranspiration data from all 40 catchments in the Basque Country. This method

enabled the models to implicitly learn critical geohydrological relationships between input features and target variables, specifically streamflow and water level at the catchment outlets. The ability of deep learning models to generalize hydrological predictions across regions with diverse hydrological characteristics is a significant advantage, paving the way for more adaptable and efficient modeling frameworks.

The insights derived from the Gini gains analysis can significantly inform model optimization strategies. For instance, the identified dominance of daily input sequence length in both streamflow and water level predictions indicates that this hyperparameter plays a crucial role in model performance. A focused exploration of input sequence length within the broader hyperparameter space could lead to more optimal model configurations. This suggests that LSTM models are highly sensitive to the temporal window of input data, with longer or more appropriate input sequences enabling the model to capture longer-term dependencies in hydrological processes, such as seasonal shifts or prolonged droughts and floods.

Moreover, the higher performance of water level metrics compared to streamflow metrics, as evidenced by the scatterplots, can be attributed to several factors, including the uncertainties inherent in the rating curves used for converting water level records into streamflow estimates. Rating curves, which depict the relationship between water level (stage) and streamflow, are subject to variability influenced by numerous factors such as changes in channel geometry, sediment deposition, and vegetation growth. These factors introduce significant uncertainties in the flow estimates derived from water level data, potentially leading to discrepancies in streamflow predictions.

In contrast, water level predictions may be more reliable in specific contexts due to the direct measurement of stage, which is often less susceptible to the uncertainties associated with flow estimation. Additionally, water levels can be influenced by a variety of hydrological processes that may exhibit more stability over time compared to the dynamic nature of streamflow, which is heavily affected by immediate factors like rainfall events and watershed responses.

Furthermore, LSTM networks may possess an inherent advantage in modeling the continuous nature of water level fluctuations, which often follow predictable patterns influenced by factors such as precipitation events and seasonal changes. This capacity enables the models to capture temporal dependencies in water level data more effectively, ultimately resulting in improved predictive accuracy.

The differences in performance between streamflow and water level metrics underscore the importance of understanding the underlying processes that contribute to each variable. Additionally, they highlight the need for refining rating curve methodologies and improving calibration techniques to enhance the accuracy of streamflow predictions. Continued research in this area could lead to the development of more robust hydrological modeling frameworks that leverage both water level and streamflow data for improved rainfall-runoff forecasting.

6.4.2. Role of catchments attributes

Our investigation deepened with the application of an attribute-aware Random Forest (RF) model, which incorporated detailed catchment attributes. The Gini gains from the attribute-aware RF model offered critical insights into how LSTM networks interpret latent hydrological processes. These Gini gains provide a quantitative measure of how much each attribute contributed to the model's decision-making, offering a window into the hydrological comprehension embedded within the LSTM architecture.

As illustrated in Figure 30, the Gini gains for the attribute-aware RF model highlight distinct patterns in the LSTM predictions for both streamflow and water level metrics. For streamflow predictions, the model emphasized attributes such as yearly average potential evapotranspiration, the coefficient of variation of precipitation, the count of days with negative temperature, and daily input sequence length as critical contributors. These attributes are aligned with well-understood hydrological processes: potential evapotranspiration captures the balance between water input and loss in a catchment, while the variability in precipitation and temperature extremes significantly influences streamflow variability.

On the other hand, for water level predictions, the RF model assigned notably higher Gini gains to attributes such as average annual precipitation, average catchment gradient, and the percentage of wetlands and water ecosystem coverage. These attributes are critical in determining the storage and release dynamics within a water basin, with wetlands and gradients playing significant roles in controlling water retention and flow velocities. The high Gini gain associated with these attributes suggests that LSTM networks can capture and integrate catchment-scale hydrological features that influence water level more than short-term meteorological fluctuations.

This behavior underscores that deep learning LSTM networks possess the capability to acquire hydrologically meaningful insights by leveraging latent information in large datasets. With sufficient data, these models can identify and respond to the diverse hydrological processes unique to each catchment, allowing for more accurate generalizations across different catchments. Moreover, the distinction between the key attributes influencing streamflow and water level predictions reflects the complexity of hydrological systems, where different drivers dominate different aspects of the water cycle.

6.4.3. Possible hydrological meaning of input sequence length hyperparameters

Among the numerous hyperparameters examined in our experiments, the daily Input Sequence Length emerged as a crucial determinant of predictive accuracy. Both the Random Forest (RF) models and Principal Component Analysis (PCA) consistently identified this

hyperparameter as having a significant influence, surpassing many other hyperparameters in terms of importance. This observation held true across different configurations, whether we applied a catchment-aware or attribute-aware RF model, which had access to catchment codes or attributes (Figures 27 and 29), or a regional RF model blinded to catchment-specific information (Figure 31). The consistent importance of the daily Input Sequence Length by RF models underscores its central role in shaping the performance of LSTM models for regional hydrological predictions.

In Figures 28, 30, and 32, the Gini gains attributed to the Input Sequence Length hyperparameter are evident in catchment-aware, attribute-aware, and regional RF models, respectively. The catchment-aware and attribute-aware RF models, which benefit from detailed knowledge of individual catchments, highlight the nuanced impact of Input Sequence Length on different LSTM configurations. Meanwhile, the regional RF model, which operates on aggregated regional data, reaffirms the importance of this hyperparameter across a broader spatial context, regardless of unique catchment characteristics.

This consistent prominence of Input Sequence Length in RF models suggests that the temporal context—that is, the length of the input sequence used to train the LSTM models—plays a critical role in enabling the model to capture complex hydrological processes. The significance of this hyperparameter likely stems from its ability to help LSTM models learn long-term dependencies in hydrological data, which are essential for accurately predicting hydrological behaviors such as streamflow and water levels across different catchments. In regional hydrological modeling, longer input sequences may allow the LSTM networks to capture patterns linked to seasonal shifts, prolonged precipitation events, or droughts, which are key in catchment-scale water cycle processes.

The importance of the daily Input Sequence Length is further reinforced by PCA results, particularly in the first principal component (PC1), which emphasizes its central role in shaping LSTM predictions (Figures 35 and 36). This influence was observed to surpass that of other hyperparameters, as shown in RF models. Both daily and hourly Input Sequence Lengths were identified as the most influential components on PC1 and PC2, highlighting their substantial impact on model performance at both local and regional scales. This underscores the fundamental importance of configuring the Input Sequence Length to optimize LSTM models for regional hydrological predictions.

Interestingly, the daily Input Sequence Length was more influential than the hourly sequence length in our case study, despite the fact that our modeling focused on hourly rainfall-runoff processes in the flashy catchments of the Basque Country. This contradictory finding points to the existence of latent, long-term patterns in hydrometeorological data that LSTM models must learn in order to provide accurate regional understandings. These latent patterns are critical for capturing catchment-specific behaviors that may not be immediately apparent from short-term, high-resolution data alone.

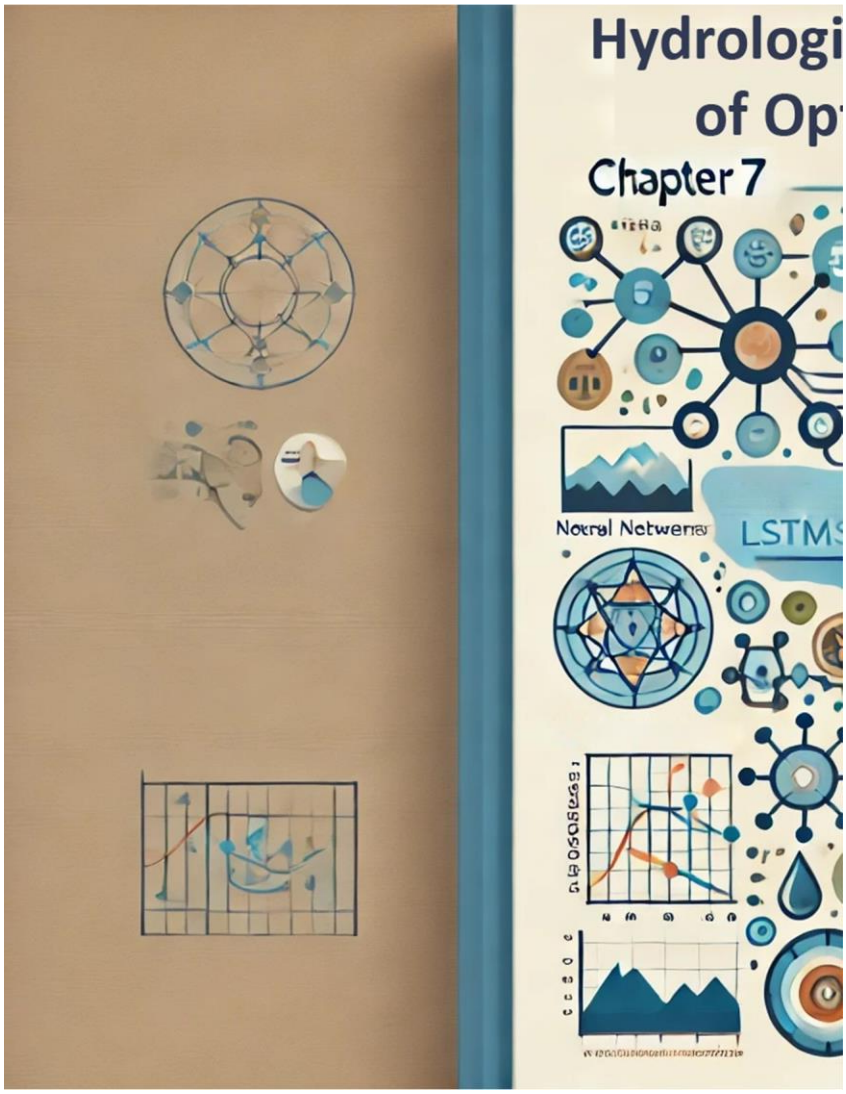
This emphasis on long-term patterns aligns with some established hydrological theories, such as the “Old Water Paradox” (Kirchner, 2003). The paradox refers to the observation that even during significant rainfall events, much of the water observed in runoff is “old water” that has been stored in the catchment for extended periods. In this context, the high

importance of daily Input Sequence Length suggests that LSTM networks might be implicitly learning the storage and slow-release processes that govern the movement of this "old water" through the catchment. This concept could be termed the "catchments' fingerprints", referring to the unique temporal patterns of water movement and storage within each catchment that the LSTM network learns through carefully optimized input daily sequence lengths.

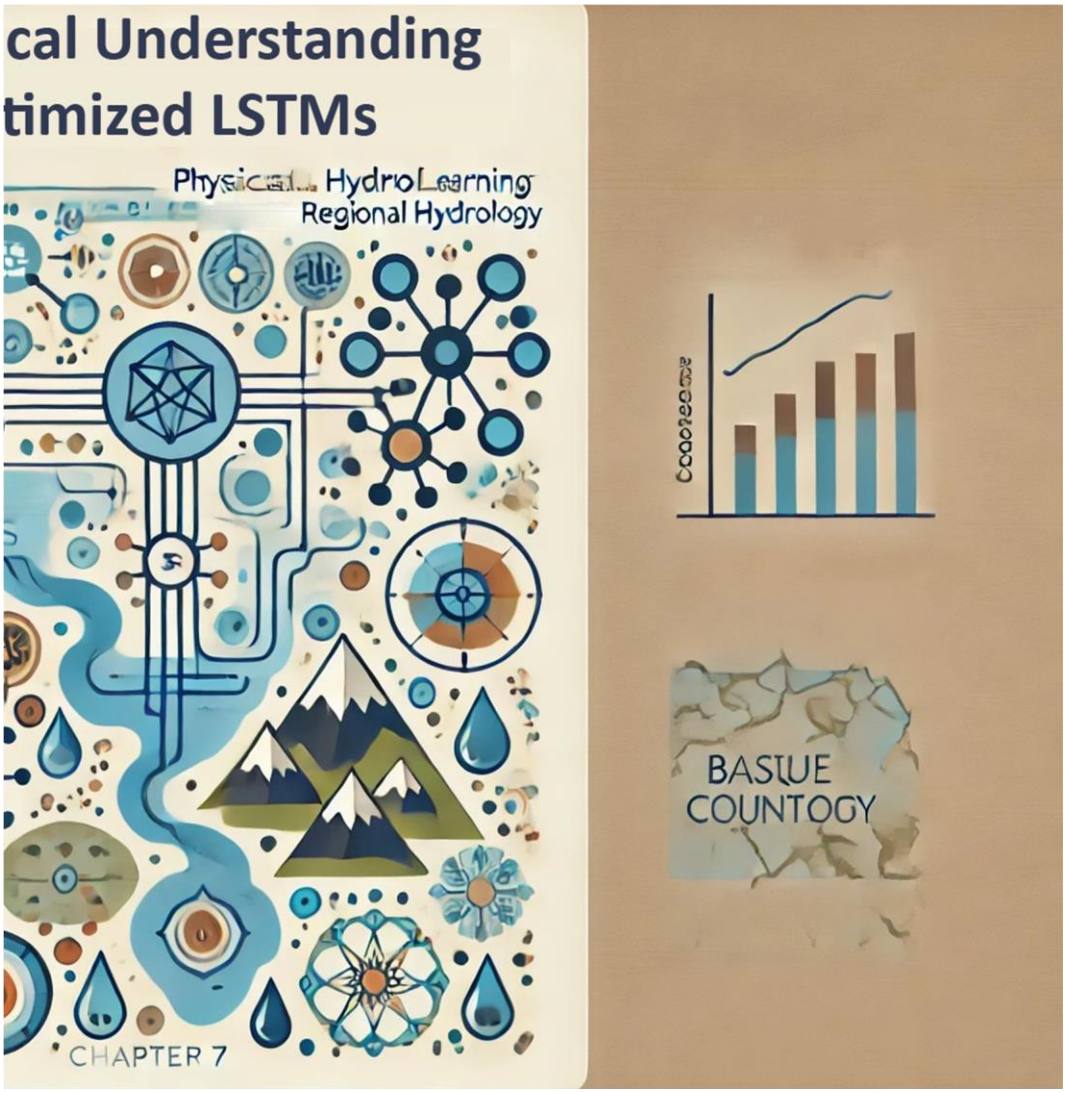
While traditional hydrology has recognized the importance of temporal windows—such as the "hydrological water year"—for understanding long-term catchment behavior, our findings indicate that the optimal input sequence length for LSTM networks need not conform to predefined periods, such as 365 days. Instead, data-driven approaches to tuning Input Sequence Length during LSTM optimization may uncover more suitable timeframes that better capture the complex, nonlinear relationships between meteorological inputs and hydrological responses, especially in flashy or rapidly responding catchments like those in the Basque Country.

The prominence of the daily Input Sequence Length highlights the need for careful tuning of this hyperparameter when applying LSTM models to hydrological prediction tasks. While setting this length arbitrarily (e.g., to match the length of a calendar year) may seem convenient, it is clear from our findings that tailoring the sequence length to capture the specific temporal dynamics of the catchments involved leads to more accurate, reliable, and robust predictions. This suggests that hyperparameter tuning should be approached with a greater focus on capturing the temporal dependencies inherent in hydrological systems, which vary across regions and scales.

In summary, the Input Sequence Length emerges as a critical factor in optimizing LSTM models for regional hydrological applications. Its ability to capture long-term dependencies in the data allows the DL model to learn and predict complex hydrological processes more effectively, potentially unlocking deeper insights into how water moves through landscapes. By carefully tuning this hyperparameter, practitioners can ensure that their models are equipped to handle the unique temporal dynamics of the regions they are studying, leading to more accurate and reliable hydrological forecasts.



Chapter VII



Performance analysis of

optimized rainfall-runoff modeling LSTMs

on hydrological data

7.1. Introduction

A central question in hydrological modeling with deep learning (DL) models and Long Short-Term Memory (LSTM) networks is: *How well do these AI-driven models capture and reflect the underlying physical characteristics of different catchments*? While LSTMs have demonstrated high effectiveness in rainfall-runoff modeling, it remains unclear whether these models are simply learning statistical patterns from the input meteorological data to "mimic target values" or if they are implicitly capturing deeper hydrological understanding of latent relationships within large datasets, linked to specific catchment attributes. This chapter aims to explore whether there are correlations between the performance of optimized LSTM networks and the physical characteristics of the catchments they were trained on.

The use of LSTMs in hydrology has grown significantly in recent years, with studies demonstrating their ability to model non-linear relationships between meteorological variables and hydrological responses such as streamflow (Kratzert et al., 2024). Despite these advances, a gap persists in our understanding of how catchment-specific attributes influence the performance of regional LSTM networks. Most LSTM-based studies focus on single catchments and improving predictive accuracy through better data preprocessing and architectural modifications, often without delving deeply into how DL models' performance varies across different catchments with unique physical and hydrological characteristics. Even in regional studies, the interaction between regionally configured LSTMs and catchment-specific features remains underexplored. While many in the hydrological community recognize DL models as effective tools for predictions, they also acknowledge that predictability and understanding are distinct tasks in real-world applications. Similar to conceptual models, which rely on hypotheses about processes and mechanisms governing hydrological behavior, DL models depend fundamentally on the quality and scope of the data. Each approach has its strengths and limitations, highlighting the importance of complementary approaches.

Traditionally, hydrological models rely on explicit understanding of catchment attributes—such as climate, topography, geology, land uses, and vegetation—to predict runoff and other hydrological processes (Beven, 2012). These attributes are critical for understanding catchment behavior. However, the way LSTM networks, even when not directly accessing such attributes, handle latent hydrological features encoded in hydro-meteorological inputs (like precipitation, temperature, and potential evapotranspiration) remains largely unexamined. This gap is particularly significant as several studies emphasize the importance of linking machine learning models with physical understanding (Reichstein et al., 2019), with the field of explainable AI (XAI) in DLs offering new methods to address this challenge (Samek et al., 2021).

Recent advancements in XAI have sought to open the black-box of DL models in hydrology, aiming to enhance the transparency and trustworthiness of these models (See review by Başağaoğlu et al., 2022). For instance, XAI approaches have been instrumental in revealing the internal logic of AI-based decisions by explaining the importance and influence of input

features on model predictions. By making AI more interpretable, these methods enhance the accountability and reliability of AI models, enabling hydrologists to trust and improve the decisions made by these models. Or, some texts argue that that a large number of accurate AI models can exist for the same problem, with some of these models being interpretable (Başağaoğlu et al., 2022). This highlights the potential to use multiple XAI methods to identify the best-performing interpretable model tailored to specific hydrological problems and catchments.

For example, XAI techniques are increasingly being used to extract the inner workings of LSTM networks and their relationship with hydrological phenomena. Lees et al. (2021) demonstrated that LSTMs could replicate hydrological concepts such as soil moisture and snow cover storage, with good correlation between the model's internal memory states and these real-world variables. This capability to decode LSTM representations offers hydrologists a novel lens through which they can understand both the strengths and limitations of DL models in the context of physical hydrology. Furthermore, Kratzert et al. (2018a) provided evidence that LSTMs internally learn to represent patterns consistent with known hydrological processes; in snow-driven catchments, for example, LSTMs develop specialized memory cells that mimic conceptual snow storages with annual dynamics, similar to process-based catchment models. This reinforces the idea that DL models do not merely recognize statistical patterns but also have the potential to internalize complex hydrological behaviors over time.

While XAI presents promising opportunities, challenges remain, particularly with respect to using these models across different spatial scales and varying data availability. Başağaoğlu et al. (2022) emphasize that XAI models can be applied at scales ranging from watersheds to continents, provided there is sufficient high-quality data. However, where data is scarce, domain knowledge becomes essential to guide the interpretation of XAI models, particularly in groundwater predictions or catchments with limited measurements. In such scenarios, a hybrid approach, combining XAI models with physics-based models, can enhance prediction accuracy (Başağaoğlu et al., 2022). Additionally, one limitation of current XAI models is that they are largely non-interventional, relying heavily on historical data. In rapidly changing hydroclimatic systems influenced by climate change or human activities, XAI models may need to be retrained with new data when unprecedented events occur in none of the catchments (Başağaoğlu et al., 2022).

Moreover, we advocate for the initial idea (Kratzert et al., 2018; Shen et al., 2018) that novel hydrological insights can be uncovered by analyzing the knowledge embedded in DL models trained on vast, readily available hydrological datasets. In this chapter, we explore these ideas further by investigating the possible relationships between optimized regional LSTM performance in different catchments and catchment-specific attributes, aiming to bridge the gap between machine learning efficacy and physical hydrological understanding. This exploration is essential not only to improve model performance but also to enhance the interpretability of AI models in hydrological applications.

Here, we hypothesize that regionally optimized LSTM networks, trained exclusively on hydrometeorological data without direct access to catchment attributes, are nonetheless

influenced by these characteristics across different locations. We posit that the performance of these rainfall-runoff DL models in predicting streamflow will vary based on catchment-specific attributes. Furthermore, we aim to determine whether optimized regional LSTMs can capture hydrological relationships unique to different catchments, suggesting that these models learn latent hydrological features through their input-output data during training.

The objectives of this chapter are threefold:

1) Investigate the correlations between the predictive performance metrics of hyperparameter-optimized regional LSTM networks and the physical and hydrological attributes of the catchments, quantitatively assessing how these attributes influence model performance.
2) Explore whether regional LSTMs, trained solely on hydro-meteorological input data (precipitation, temperature, and potential evapotranspiration), can implicitly learn catchment-specific features that enhance prediction accuracy, despite lacking direct access to these attributes during training.
3) Examine the impact of precise hyperparameter optimization on the performance of different regional LSTMs in their predictions at various locations.

By conducting a thorough examination, we aim to bridge the gap between AI-driven hydrological modeling and the physical processes governing catchment behavior. Through this analysis, we seek to enhance the understanding of optimized regional LSTMs in hydrology and explore their potential to inform real-world water management practices by identifying critical catchment attributes that may affect model performance. This work contributes to the broader goal of integrating machine learning approaches with domain-specific hydrological knowledge to develop more robust and interpretable models that can ultimately facilitate the learning of hydrological insights from our intelligent agents.

7.2. Method

To achieve the objectives, this chapter employed a systematic approach to mine and comprehensively analyze what is referred to as the "**test_DATASET** (with capital letters)." The test_DATASET consists of streamflow test performance metrics from several regional optimized rainfall-runoff LSTM networks, their hyperparameter configurations, and catchment attributes across the 40 studied catchments in the Basque Country, Spain. This section outlines the methodology adopted to investigate these relationships, forming the foundation of our findings.

7.2.1. Test_DATASET Setup and Compilation

We began by curating a comprehensive test_DATASET that included hyperparameter-optimized regional LSTM networks' performance metrics across different locations, alongside relevant catchment attributes. This test_DATASET was compiled from several fine-tuned regional LSTM models developed during our research, all trained and tested across 40 catchments in the Basque Country, Spain. Each regional LSTM network was optimized to predict two targets—streamflow and water levels—on an hourly timestep based on three meteorological inputs: precipitation, temperature, and potential evapotranspiration (PET).

The **test_DATASET** consisted of three main group columns:

1) **Hyperparameter Configurations**: A detailed record of the hyperparameter configurations for all optimized LSTM networks.
2) **Test Performance Metrics in every catchment**: We evaluated different performance metrics in this chapter (refer to Chapter III: 3.6.1.)
3) **Catchment Attributes**: Detail hydrological and physical attributes of the 40 catchments were gathered (refer to Chapter III: Table 1). The definitions of these attributes can be found in Table 9. This data is essential for understanding the relationship between catchment characteristics and regional model performance.

Table 9. Definitions of the catchments' attributes employed in this study.

Attribute	Definition	Group	Units
Area	Contributing area to the downstream end of the segment	Topography	km^2
CONF_DEN	Number of rivers confluences by catchment area	Topography	Number/km^2
GRADIENT	Mean gradient through the reach (vertical change/horizontal length)	Topography	ratio
max slope	max slope of catchment	Topography	°
mean slope	average slope of catchment	Topography	°
elevation	Average catchment elevation upstream the river reach	Topography	m
min height	min catchment eleveation	Topography	m
max height	max catchment eleveation	Topography	m
UHD	Surface occupied by urban areas upstream the river reach	Land Uses	%
AGR	Surface occupied by agricultural land upstream the river reach	Land Uses	%
PAS	Surface occupied by pasture upstream the river reach	Land Uses	%
BLF	Surface occupied by broadleaf forest upstream the river reach	Land Uses	%
CNF	Surface occupied by coniferous forest upstream the river reach	Land Uses	%
PLT	Surface occupied by plantations upstream the river reach	Land Uses	%
SSH	Surface occupied by moors, heathland, scrub and shrubs upstream the river reach	Land Uses	%
WAE	Surface occupied by wetlands and water ecosystems upstream the river reach	Land Uses	%
DEN	Surface occupied by denuded areas upstream the river reach	Land Uses	%
calc	Area occupied by calcareous rocks upstream the river reach	Geology	%
cong	Area occupied by conglomerate rocks upstream the river reach	Geology	%
sdim	Area occupied by sedimentary rocks upstream the river reach	Geology	%
vlc	Area occupied by volcanic rocks upstream the river reach	Geology	%
watr	Area occupied by wetlands and water associated ecosystems upstream the river reach	Geology	%
conductivity	Average soil conductivity upstream the river reach (derived from geology variables). Reaches with MN_watr and MN_othe=1 this value have 0 for this field	Geology	Class: 1 - 5
permeability	Average terrain permeability upstream the river reach (derived from geology variables). Reaches with MN_watr and MN_othe=1 this value have 0 for this field	Geology	Class: 1 - 5
rock hardness	Average soil hardness upstream the river reach (derived from geology variables). Reaches with MN_watr and MN_othe=1 this value have 0 for this field	Geology	Class: 1 - 5
no. prec stations	Nomber of stations participated in calculating lumped prec values for basins	Hydrology	Number
no. temp stations	Nomber of stations participated in calculating lumped temp values for basins	Hydrology	Number
possible snow	Percentage of number of days with negative temp on total number of days	Hydrology	%
no. days with negative temp	Number of days with negative temp in the dataset	Hydrology	Number
mean runoff coeff.	yearly average runoff cofficient	Hydrology	dimensionless
aridity index	Aridity Index	Hydrology	dimensionless
mean precipitation	yearly average precipitation	Hydrology	mm
mean streamflow	yearly average streamflow	Hydrology	mm
mean temperature	yearly average temperature	Hydrology	°C
min temperature	yearly min temperature	Hydrology	°C
max temperature	yearly max temperature	Hydrology	°C
Coeff. var. Prec	Cofficient of variation of precipitation	Hydrology	dimensionless
Coeff. var. Flow	Cofficient of variation of streamflow	Hydrology	dimensionless
mean PET	average potential evapotranspiration	Hydrology	mm

Overall, the test_DATASET included 67,040 records (comprising hourly and daily predictions on ten different random seeds by 84 distinct optimized single-configuration MTS-LSTM architectures in terms of their hyperparameters), providing a solid foundation for exploring potential correlations and trends between the predictive performance of optimized regional LSTM networks and catchment attributes. These records were derived from all models optimized during this PhD research. Among them, 37 optimized networks were presented in Chapters 4 and 5 in detail (Table 6). Additionally, we intentionally included 47 extra optimized regional LSTM networks from our initial trial-and-error phase, prior to the final exhaustive random search, to enhance the robustness of our investigation.

Each regionally optimized single-configuration LSTM network in the test_DATASET demonstrated competitive regional accuracy overall, with some marginal differences. However, as discussed in Chapters 4 and 5, these configured networks exhibited statistically significant varying performance across different locations, with some catchments showing underperformance. In Chapter 5, we proposed the use of ensemble learning with catchment-wise optimized regional LSTMs to enhance accuracy in diverse locations. Nevertheless, we think that the overall accuracy of these individual optimized regional LSTMs justifies their inclusion in the test_DATASET.

All LSTM networks used in this analysis were regional in design, meaning they were trained on data from all 40 catchments simultaneously. This regional setup approach allowed the models to predict streamflow and water levels at the outlets of individual catchments, facilitating the extraction of latent features from the combined data and leveraging interconnections between catchments during the training process.

The LSTM networks utilized hourly precipitation, temperature, and potential evapotranspiration (PET), which was calculated using the Hargreaves and Allen's (2003) equation, as input data. Although these models did not directly incorporate catchment attributes—such as soil type, land cover, or elevation—any observed correlation between model performance and catchment characteristics would suggest that the optimized LSTMs effectively learned implicit representations of these attributes. This hidden knowledge was likely derived from the input variables, particularly PET, and the relationship between streamflow and water level targets, often represented by human-defined rating curves. In this study, we had access to both targets from the outset, and rating curves were developed by the Basque Country water agency in accordance with established methods in hydrology and hydraulics.

7.2.2. Exploration of Catchment Attribute-Performance Relationships

To investigate potential correlations between catchment attributes and the performance of the optimized regional LSTM networks, we created heatmaps that visually represented the relationships between 14 streamflow prediction performance metrics and various catchment attributes. Each cell in these heatmaps quantified the degree of association between a specific performance metric for the optimized networks in different locations, such as Nash-Sutcliffe Efficiency (NSE) or Kling-Gupta Efficiency (KGE), and a particular catchment attribute, such as area, slope, soil type, or land cover, across all 40 catchments. These heatmaps provided an initial overview of potential trends, offering insights into how the physical and hydrological characteristics of different catchments might influence LSTM model performance.

To delve deeper into these relationships, we focused on the highest correlations observed in the heatmaps. This step aimed to identify the most influential catchment attributes and their roles in shaping the accuracy of the rainfall-runoff LSTM networks' predictions. By examining these high-correlation pairs, we determined which catchment attributes consistently contributed to accurate hydrological predictions across various catchments.

Building on this initial analysis, we applied Random Forest (RF) and Principal Component Analysis (PCA) models to the test_DATASET, following the methodology outlined in Chapter 6, this time with focus on test results. The RF model was trained on 70% of the test_DATASET and tested on the remaining portion. In this setup, the target variables were the 14 metrics of the optimized regional LSTM models in each catchment, while the input features were the hyperparameter configurations of LSTMs and catchment attributes. This approach leveraged the learning capabilities of the RF to assess which catchment attributes significantly impacted the accuracy of the optimized regional LSTM predictions.

Moreover, by analyzing the corresponding feature importance Gini scores learned by the RF model, we quantified the relative importance of each catchment attribute. This analysis provided a detailed understanding of how specific catchment characteristics, such as topography, climate, and land use, contributed to the optimized regional LSTM model's ability to accurately predict runoff from rainfall in different locations. This approach not only validated the findings from the heatmaps but also offered a robust, quantitative assessment of the influence of catchment attributes on model performance, highlighting the potential for LSTMs to implicitly learn and adapt to catchment-specific features, even without direct access to these attributes during training.

Additionally, PCA was applied to the test_DATASET to examine if catchment attributes can influence the final average NSE and KGE performance of the optimized LSTMs in each catchment. This analysis aimed to uncover the unique characteristics of each catchment and their potential impacts on model performance.

7.3. Results

As stated, we employed three main analytical approaches: correlation analysis, Random Forest (RF) modeling, and Principal Component Analysis (PCA). These techniques aimed to look at the relationships between catchment attributes and the accuracy metrics of optimized LSTM models on the test set, as well as to assess the predictability of performance using RF models and identify key feature importance among the catchments' attributes.

7.3.1. Pearson Correlation Analysis

To explore potential relationships between catchment attributes and the performance of various optimized regional LSTMs across different catchments, we computed correlation coefficients. This analysis allowed us to detect possible associations between specific catchment features (e.g., climate, land use, elevation, and drainage area) and the performance metrics of regional LSTMs in different locations to infer hydrological understanding of optimized LSTM networks on Basque Country dataset.

Figure 37 presents a Pearson correlation heatmap that displays the relationships between catchment attributes and 14 test performance metrics for several LSTM networks that were optimized and trained at different stages of the research. This heatmap provides valuable insights into potential relationships between catchment's unique characteristics and the accuracy of the LSTMs. As stated before, the LSTM networks were trained without direct knowledge of these catchment attributes; instead, they relied on input features like lumped precipitation, temperature, potential evapotranspiration, and streamflow or water level as targets from all 40 catchments, simultaneously.

We explored the correlations by focusing on attributes with correlation values higher than 0.3 and lower than -0.3, as presented in Table 10. These correlations provide deeper insights into how catchment attributes influence LSTM performance, and by extension, highlight important hydrological factors that may affect model accuracy. The extracted correlations from Table 10 reveal possible trends between catchment attributes and LSTM performance metrics, providing a clearer picture of the specific conditions under which the DL models performed better or worse. The following presents noticeable findings based on both positive and negative correlations in more detail.

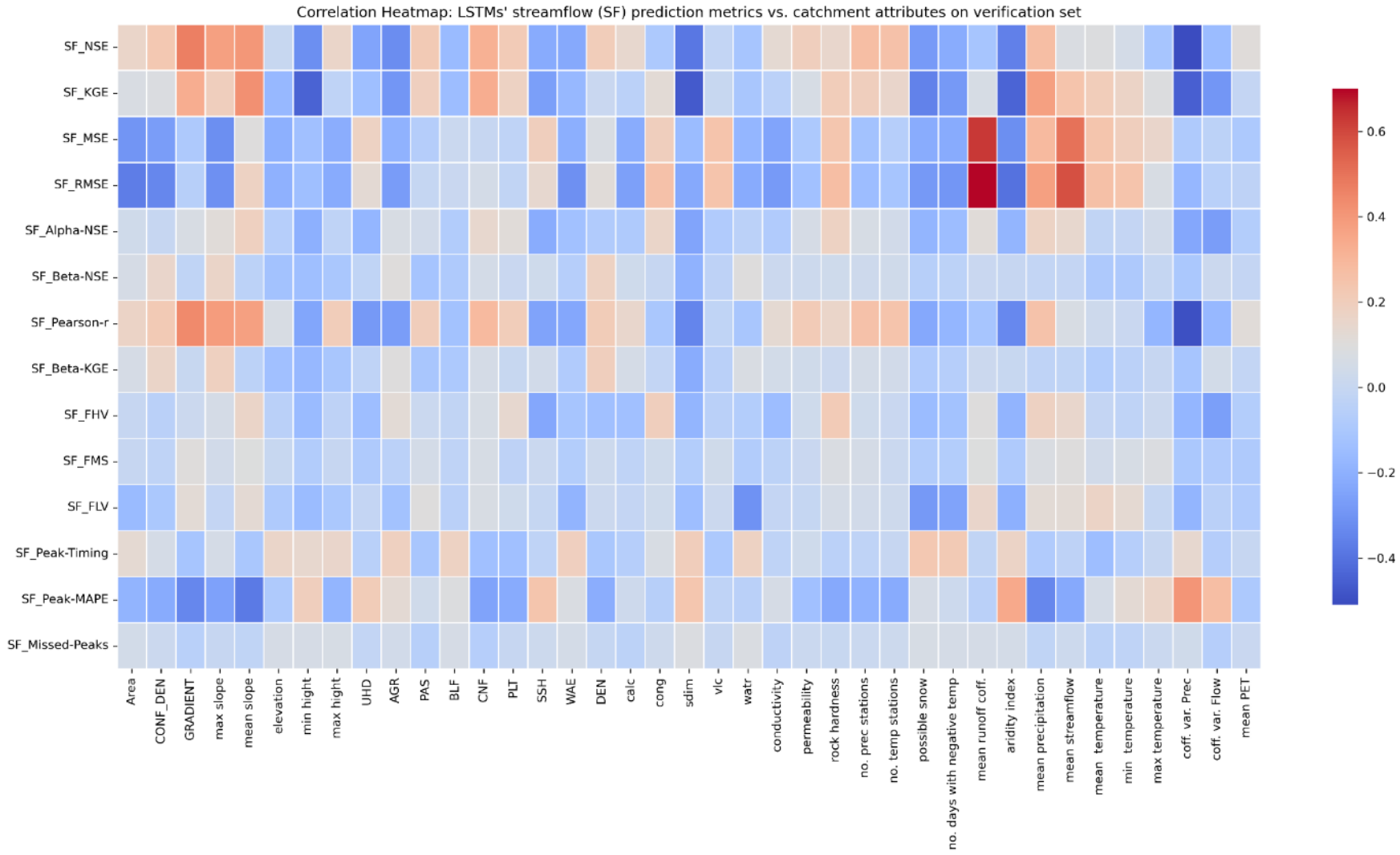


Figure 37. Correlation heatmap showing the relationships between catchment attributes and streamflow (SF) test metrics. Darker colors represent stronger correlations, with both positive and negative relationships between various features and metrics.

1. Runoff Coefficient and yearly Streamflow Impact on Prediction Accuracy

High positive correlations between MSE and RMSE test metrics and the **average runoff coefficient** (0.63 and 0.70, respectively) as well as **mean yearly streamflow** (0.50 and 0.58, respectively) suggest that catchments with higher runoff coefficients tend to have a bit larger absolute errors in these metrics. Since MSE and RMSE are not dimensionless and scale with streamflow magnitude, larger streamflow naturally leads to higher absolute error values. However, this does not necessarily indicate lower relative predictive accuracy, as normalized performance metrics (e.g., NSE or KGE) may provide a different perspective on model effectiveness.

Given that all models in this study are optimized and demonstrate acceptable accuracy, this suggests that catchments with high runoff coefficients—such as those in the Basque Country—exhibit clearer hydrological signals, making flow dynamics easier for LSTM networks to capture. However, at both very high and very low runoff coefficient values, LSTMs may struggle to accurately learn the trends, potentially leading to diminished performance.

2. Topographic Influence on Predictive Metrics

Catchment **slope** and **gradient** correlate moderately with several accuracy metrics, including NSE (0.40), KGE (0.42), and Pearson-r (0.44). This suggests that steeper catchments with higher gradients could improve model accuracy, possibly due to more distinct runoff patterns in these regions. Steep terrains tend to have concentrated water flow with less surface retention, which may lead to more defined hydrological responses that the LSTM models can better capture. However, a negative correlation with maximum slope (-0.31 for RMSE) suggests that extremely steep topographies could introduce noise or outliers. This may be due to the complexities of flow dynamics in steep catchments or limitations in the accuracy of input data, which could challenge the model's ability to learn effectively. Therefore, moderate slopes seem to enhance predictive accuracy, while extreme slopes may complicate predictions.

3. Influence of Climate and Variability

The **aridity index** exhibits negative correlations with several metrics (e.g., KGE at -0.44, RMSE at -0.40), indicating that drier, and more arid catchments with variable precipitation pose challenges for the LSTM model. These conditions often result in inconsistent hydrological responses that reduce the model's ability to generalize, impacting prediction accuracy. Similarly, the **coefficient of variation for precipitation**, moderately negatively correlated with KGE (-0.45) and Pearson-r (-0.49), suggests that high precipitation variability adds complexity to the hydrological processes. Stable precipitation regimes likely facilitate the model's generalization capabilities, resulting in higher prediction accuracy. In contrast, irregular rainfall patterns reduce the predictability of runoff and streamflow, making it harder for the LSTM models to accurately capture hydrological dynamics.

4. Land Use and Vegetation Cover

Coniferous forest cover (CNF) shows a moderate positive correlation with metrics like KGE (0.33) and NSE (0.32), suggesting that catchments with coniferous vegetation may exhibit

hydrological consistency beneficial to LSTM predictions. The likely impact of such vegetation cover on evapotranspiration and soil stability could contribute to steadier flow patterns that enhance model performance. In contrast, **agricultural land use (AGR)** shows a moderate negative correlation with NSE (-0.32), likely due to human alterations in hydrological flows, such as irrigation or land management practices, which add variability to streamflow responses and diminish model accuracy.

5. Geological Characteristics

Presence of **sedimentary soils (sdim)** correlates negatively with NSE and KGE (-0.39 and -0.47), suggesting that sedimentary formations introduce variability, possibly due to differences in permeability and water storage properties that affect flow dynamics. This variability likely challenges the LSTM model's ability to capture consistent hydrological responses. Similarly, **water bodies (Watr) and wetland areas (AWE)**, correlated with FLV (0.31) and negatively with RMSE (-0.31), may add flow variability that impacts prediction accuracy, as these features can modulate streamflow responses over time.

6. Probability of Snowfall and Temperature Variability

The negative correlation of KGE with **snowfall probability** (-0.35) suggests that snow-related processes introduce complexities in streamflow modeling, possibly due to the delayed runoff associated with snowmelt. Although the Basque Country basins are not predominantly snowy, some basins exhibit snowfall patterns, and the LSTM models have identified relationships between snowfall probability and streamflow, highlighting the challenge of capturing these dynamics without explicit snow data. **Temperature variability** also impacts performance, as evidenced by the negative correlation with NSE, indicating the model's difficulty in handling catchments with significant temperature fluctuations.

Table 10. High Values (-0.3< or >0.3) in the Correlation heatmap showing the relationships between catchment attributes and test metrics.

Metric	Attribute	Correlation
SF_RMSE	mean runoff coff.	0.70
SF_MSE	mean runoff coff.	0.63
SF_RMSE	mean streamflow	0.58
SF_MSE	mean streamflow	0.50
SF_NSE	GRADIENT	0.47
SF_Pearson-r	GRADIENT	0.44
SF_KGE	mean slope	0.42
SF_Peak-MAPE	coff. var. Prec	0.41
SF_NSE	mean slope	0.40
SF_Pearson-r	max slope	0.39
SF_Pearson-r	mean slope	0.38
SF_KGE	mean precipitation	0.38
SF_NSE	max slope	0.38
SF_RMSE	mean precipitation	0.38
SF_Peak-MAPE	aridity index	0.35
SF_KGE	GRADIENT	0.34
SF_KGE	CNF	0.33
SF_NSE	CNF	0.32
SF_FLV	watr	-0.31
SF_RMSE	max slope	-0.31
SF_RMSE	WAE	-0.31

Metric	Attribute	Correlation
SF_MSE	max slope	-0.32
SF_NSE	min hight	-0.32
SF_MSE	aridity index	-0.32
SF_NSE	AGR	-0.32
SF_Pearson-r	aridity index	-0.33
SF_RMSE	CONF_DEN	-0.34
SF_Peak-MAPE	mean precipitation	-0.34
SF_Peak-MAPE	GRADIENT	-0.34
SF_Pearson-r	sdim	-0.34
SF_KGE	possible snow	-0.35
SF_NSE	aridity index	-0.36
SF_RMSE	Area	-0.37
SF_Peak-MAPE	mean slope	-0.38
SF_NSE	sdim	-0.39
SF_RMSE	aridity index	-0.40
SF_KGE	aridity index	-0.44
SF_KGE	min hight	-0.45
SF_KGE	coff. var. Prec	-0.45
SF_KGE	sdim	-0.47
SF_Pearson-r	coff. var. Prec	-0.49
SF_NSE	coff. var. Prec	-0.51

Overall, the correlation analysis reveals that catchments with stable hydrological regimes—characterized by higher runoff coefficients, steady precipitation, and specific vegetation types (e.g., coniferous forests)—tend to support improved LSTM model accuracy. Conversely, catchments with high climate variability, arid conditions, complex geological features, or significant human modifications (like agriculture) are associated with reduced prediction accuracy. These findings highlight the importance of hydrological stability in enhancing LSTM performance and suggest areas for model improvement.

This analysis further suggests that, despite not having direct access to catchment attributes, the LSTM models might have implicitly “learned” latent hydrological patterns, reflecting distinctive behaviors across catchments. These learned features may represent underlying processes or environmental connections unique to specific catchments. All in all, this correlation analysis offers critical insights for model development, optimization, and further studies focused on understanding how specific catchment attributes influence prediction performance in LSTM models.

7.3.2. Random Forest Analysis

Figure 38 displays the validation results of the Random Forest (RF) model trained on the test_DATASET, where hyperparameter configurations for each optimized LSTM network, along with catchment attributes, were used as inputs. The performance metrics of each model configuration across various catchments served as the targets for training the RF

model. The RF model was trained on 70% of the dataset and validated on the remaining 30%, ensuring a robust performance assessment. As seen in Figure 38, the alignment of predicted and actual values along the 1:1 line—together with well-accepted MSE (near zero) and R-squared (up to 0.97) values—confirms the RF model's capability to approximate the performance metrics of optimized LSTM networks across different catchments by utilizing their attributes and model configurations.

Figure 39 presents the corresponding feature importance rankings (Gini gains) derived from the RF model trained on the test_DATASET. This analysis highlights the relative importance of various catchment attributes in predicting the accuracy metrics for each optimized LSTM configuration, focusing specifically on streamflow (SF) predictions. Key observations from the Gini gains can be summarized as below:

Overall: The RF model reveals that attributes related to hydrological processes, such as precipitation patterns, mean yearly streamflow, and aridity index, are instrumental in enhancing model accuracy, especially for metrics like KGE, Beta-NSE, and Missed-Peaks. For example, mean streamflow emerged as a prominent predictor for RMSE and Beta-KGE, underscoring the importance of hydrological attributes in refining the predictive power of the LSTM models.

Influence of Catchment Attributes: The feature importance analysis indicates that specific catchment characteristics significantly influence the RF model's ability to predict the performance of optimized LSTM networks across diverse catchments. Noticeably, features such as CNF (coniferous forest cover) and PAS (pasture cover) rank highly in importance, particularly for metrics such as NSE, Alpha-NSE, and Pearson-r. The presence of certain land cover types, as well as climatic monitoring, appear to be critical factors influencing model performance.

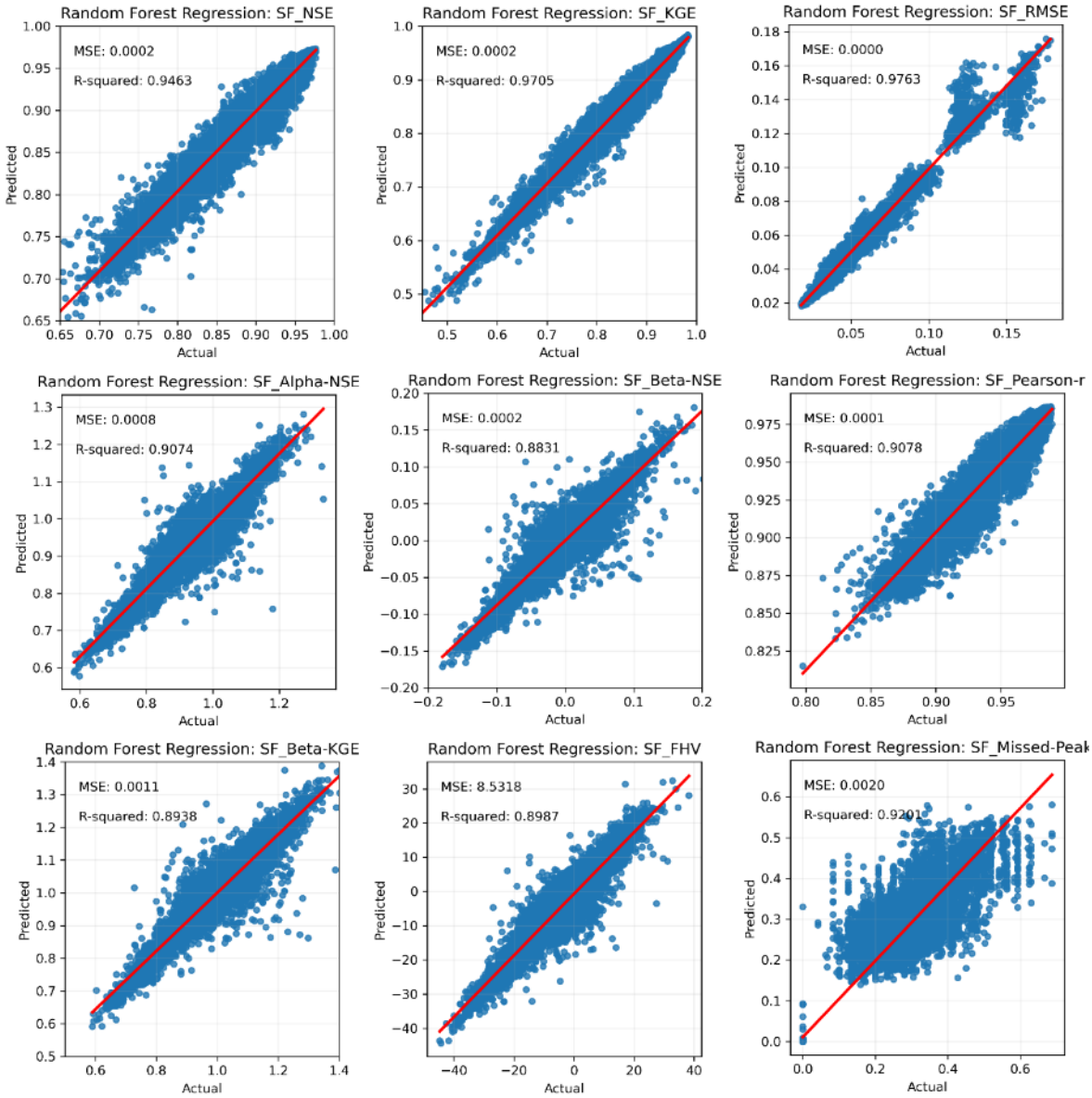


Figure 38. Random Forest prediction accuracy for different test performance metrics for streamflow and water level. The Random Forest was trained on Hyperparameters, Attributes, and Basin code as inputs and the metrics as outputs. The figure suggests that a Random Forest model can be trained in a way that accurately predicts the outcomes of the optimized LSTMs in every catchment by knowing their attributes.

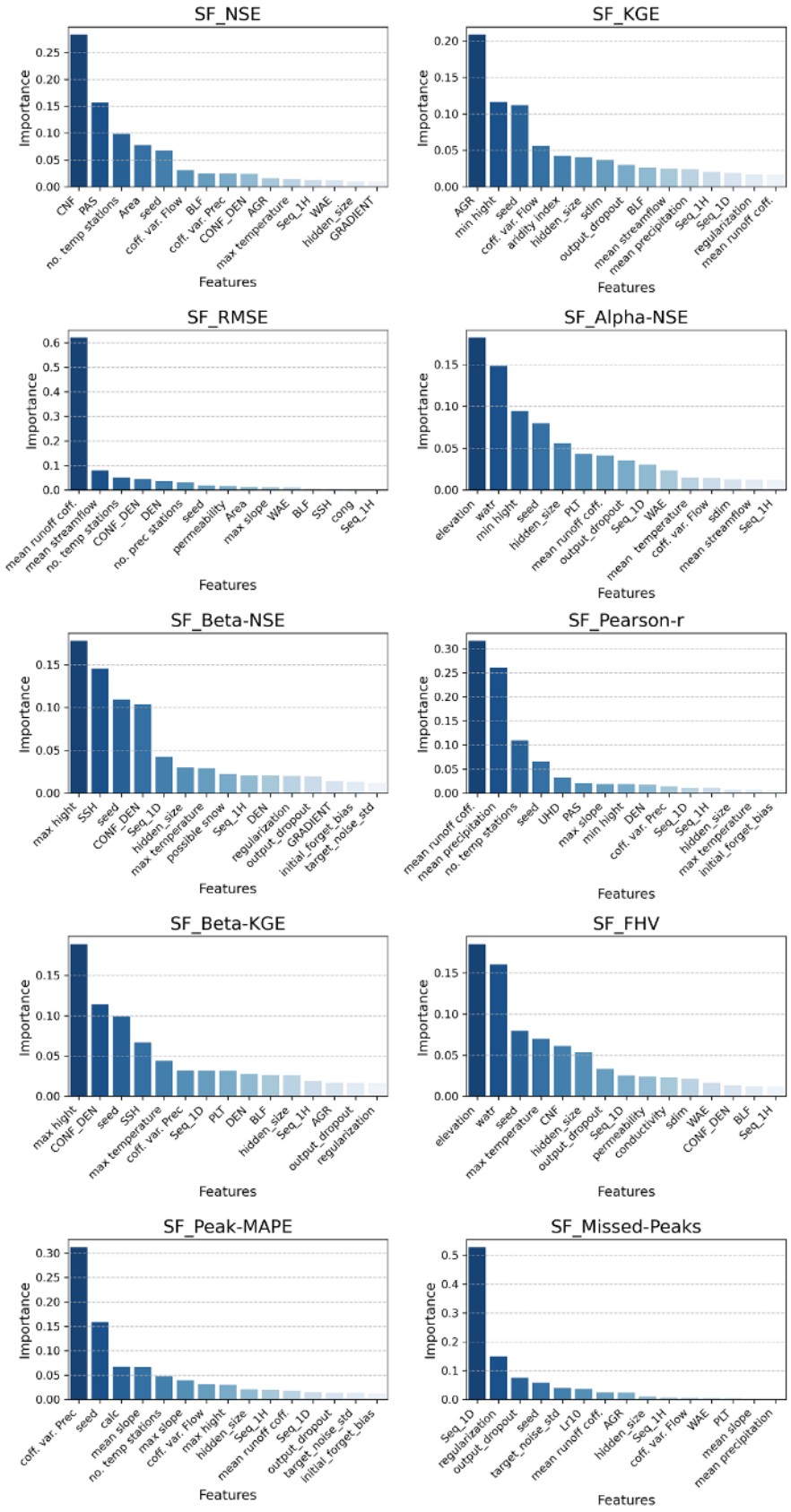


Figure 39. Feature importance ranking for 15 top features for every target metric derived from the Random Forest model. This figure highlights the relative importance of various features in predicting the test metrics, emphasizing the most influential attributes.

Hyperparameters Influence: Certain hyperparameters also play a significant role in RF model performance. For instance, seed values, which control random initialization, were consistently influential across multiple accuracy metrics, including NSE, KGE, and Missed-Peaks. Other hyperparameters, such as output dropout, input sequence length, regularization, and hidden size, also show considerable importance in accurately predicting LSTMs' performance metrics across various catchments. This highlights the role of model configuration in determining performance.

Catchment-Specific Insights: The Gini gains reveal variability in feature importance across different catchments, suggesting that specific attributes may carry more weight depending on geographic or climatic contexts. For example, attributes like max temperature and max slope exhibit greater importance for metrics such as Missed-Peaks and FHV in certain catchments, indicating that distinct topographical and climatic factors affect hydrological responses. Furthermore, attributes like BLF (Surface occupied by broadleaf forest) and coefficient of variation of flow were notable for Alpha-NSE and Beta-NSE, reflecting their relevance in regions with broadleaf forest contributions or variable flow conditions.

In summary, the RF analysis provides valuable insights into the factors influencing the accuracy of LSTM models for streamflow predictions, offering advantages over linear or stepwise approaches by effectively capturing non-linear relationships and complex interactions between hyperparameters. By examining Gini gains, we identify key features, such as land cover, climate-related attributes, and specific hyperparameters, that enhance model performance. This analysis not only helps to pinpoint critical attributes for optimized LSTM configurations but also offers targeted strategies for improving hydrological predictions across diverse catchments by LSTMs.

7.3.3. Principal Component Analysis (PCA)

A Principal Component Analysis (PCA) was conducted on the test_DATASET, incorporating catchment attributes alongside the average NSE and KGE test performance metrics derived from optimized LSTM networks for streamflow across different catchments.

Figures 40 and 41 illustrate the PCA model results, including the scree plot (Figure 40) and the biplot analysis (Figure 41). The scree plot demonstrates that the first ten principal components cumulatively explain around 87.8% of the dataset's total variance with three first PCs having near 60%. These 10 components capture the majority of the dataset's variability, thus representing key aspects of the underlying structure. The biplot in Figure 41 visualizes the relationships between the first two principal components and the original features, indicating which catchment attributes contribute most significantly to each component. This figure shows a biplot from the local PCA analysis, illustrating how original catchment features contribute to the principal components and interact with one another. It visually represents the contributions of individual features to the first few principal components and highlights their relationships in the context of catchment hydrology.

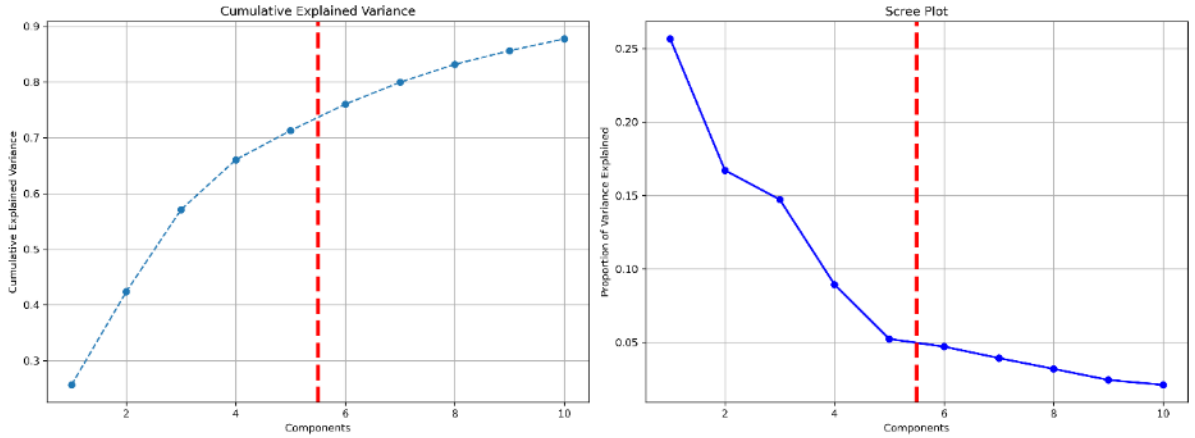


Figure 40. PCA results applied to the test_DATASET, showing the distribution of principal components. This plot illustrates how much variability is captured by each component, providing insights into the test_DATASET's structure.

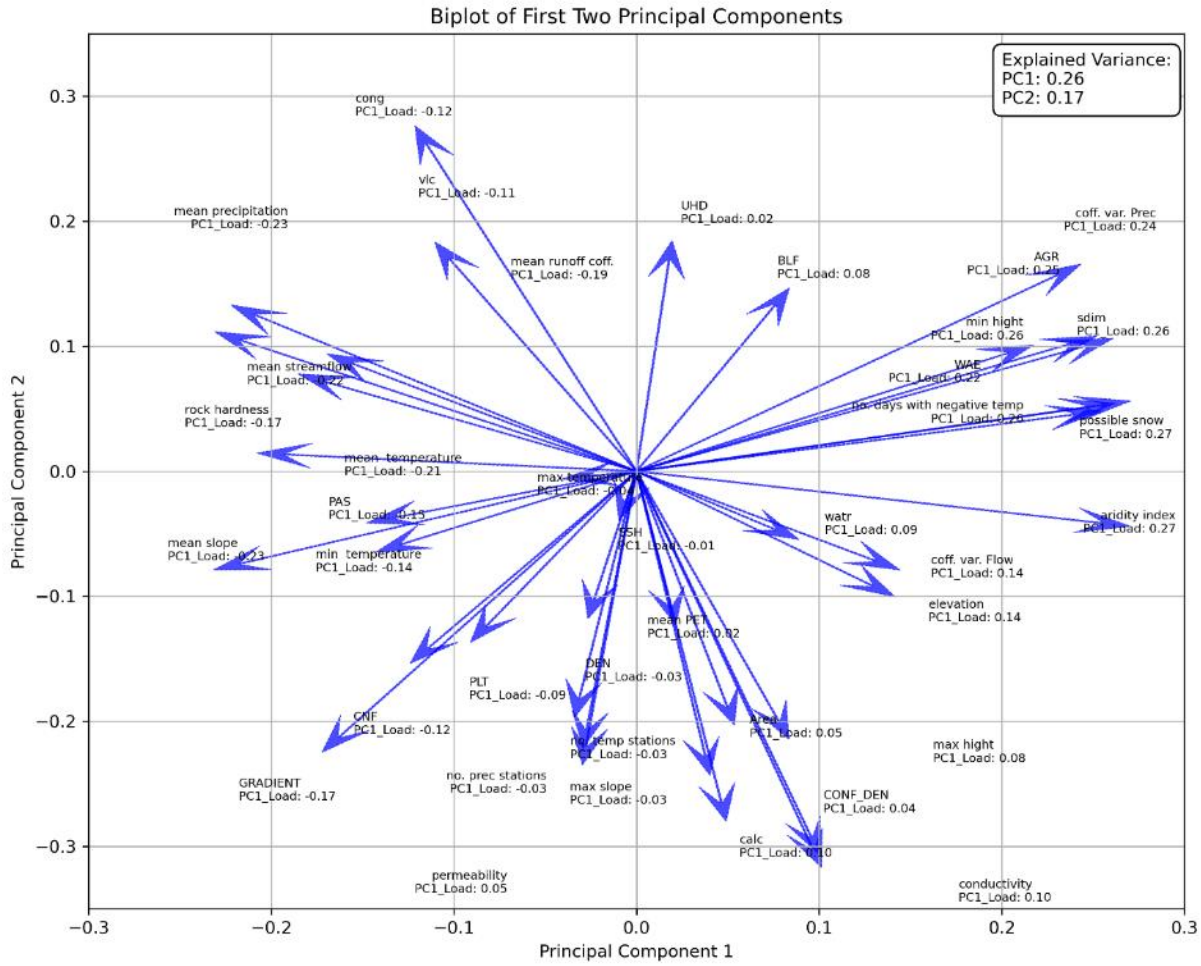


Figure 41. Biplot of the PCA analysis. This figure displays the principal components in relation to the original features, visually representing how each feature contributes to the principal components and their interactions.

Explained Variance and Component Loadings

Table 11 displays the component loadings and explained variance ratios, illustrating each principal component's contribution to the total variance. These loadings help identify key catchment features that influence the hydrological performance metrics of the optimized regional LSTMs.

Table 11. PCA components' loads and the explained variance ratios

Attributes	Principal components									
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Area	0.054	-0.203	-0.099	0.355	-0.247	-0.061	-0.053	-0.024	0.066	-0.036
CONF_DEN	0.040	-0.243	-0.068	0.344	-0.157	-0.054	-0.019	-0.127	0.058	-0.100
GRADIENT	-0.172	-0.225	-0.096	-0.106	-0.011	-0.018	-0.048	0.084	0.255	0.012
max slope	-0.030	-0.235	-0.108	0.118	0.188	0.247	-0.020	0.046	0.437	-0.144
mean slope	-0.232	-0.079	-0.198	-0.140	-0.094	-0.079	-0.137	0.036	0.006	0.007
elevation	0.142	-0.100	-0.295	-0.112	0.110	0.169	0.008	0.109	0.064	-0.037
min hight	0.255	0.048	-0.133	-0.126	0.119	0.107	0.117	-0.027	-0.128	-0.026
max hight	0.083	-0.214	-0.285	0.016	0.084	0.157	-0.011	0.092	0.111	0.057
UHD	0.019	0.185	0.248	0.158	-0.119	0.190	0.076	-0.052	-0.050	0.279
AGR	0.252	0.108	-0.053	0.008	-0.082	-0.120	0.101	0.143	-0.005	0.026
PAS	-0.148	-0.041	0.222	-0.121	-0.214	0.191	-0.056	0.088	0.153	-0.151
BLF	0.084	0.147	-0.229	0.091	0.067	0.306	-0.086	0.318	-0.117	-0.233
CNF	-0.124	-0.154	0.053	-0.247	-0.222	-0.262	0.056	-0.297	0.122	0.117
PLT	-0.091	-0.137	-0.111	0.036	0.278	-0.148	-0.272	-0.103	-0.288	0.324
SSH	-0.010	-0.039	0.151	0.298	0.390	-0.124	0.065	-0.119	0.060	-0.094
WAE	0.218	0.101	-0.090	0.044	-0.091	-0.248	-0.176	0.107	0.069	0.258
DEN	-0.027	-0.118	-0.068	0.032	0.337	0.028	0.451	-0.326	0.110	0.153
calc	0.100	-0.307	0.123	-0.178	0.068	-0.057	0.037	0.101	-0.131	-0.016
cong	-0.121	0.276	-0.102	0.145	-0.139	0.145	-0.095	-0.118	0.149	0.079
sdim	0.261	0.106	-0.050	0.035	-0.031	-0.093	0.056	0.067	-0.023	0.123
vlc	-0.110	0.183	-0.094	0.159	0.241	-0.237	0.176	-0.001	0.009	-0.276
watr	0.089	-0.055	-0.118	0.250	-0.061	0.028	-0.378	-0.389	-0.130	-0.114
conductivity	0.101	-0.317	0.128	-0.153	0.031	0.077	0.023	0.111	-0.017	0.079
permeability	0.049	-0.280	-0.114	-0.175	0.044	-0.113	-0.123	0.183	-0.292	0.031
rock hardness	-0.170	0.094	-0.228	0.006	0.150	-0.244	-0.124	-0.015	-0.166	-0.322
no. prec stations	-0.029	-0.222	-0.042	0.279	-0.147	-0.034	0.172	0.203	-0.141	0.099
no. temp stations	-0.034	-0.197	-0.078	0.326	-0.220	-0.094	0.224	0.090	-0.121	0.098
possible snow	0.271	0.056	-0.166	-0.002	-0.071	-0.075	0.021	-0.090	-0.006	0.019
no. days with negative temp	0.262	0.054	-0.186	-0.009	-0.072	-0.093	0.024	-0.052	0.010	0.078
mean runoff coff.	-0.185	0.078	-0.116	0.078	0.069	0.207	0.321	0.112	-0.211	0.292
aridity index	0.270	-0.044	0.143	0.124	0.036	0.098	-0.037	-0.009	0.042	-0.046
mean precipitation	-0.231	0.112	-0.218	-0.094	-0.053	-0.044	-0.046	-0.069	-0.036	0.025
mean streamflow	-0.222	0.133	-0.222	-0.001	0.015	0.051	0.112	0.017	-0.117	0.152
mean temperature	-0.208	0.015	0.232	0.147	-0.011	0.014	-0.126	0.162	-0.131	0.050
min temperature	-0.143	-0.065	0.187	0.163	0.070	0.080	-0.076	0.206	-0.270	-0.126
max temperature	-0.041	-0.006	0.114	0.127	0.366	-0.066	-0.350	0.142	0.250	0.386
coff. var. Prec	0.243	0.166	0.136	-0.007	0.119	0.050	-0.151	0.040	0.013	0.051
coff. var. Flow	0.144	-0.079	0.263	-0.054	0.105	-0.127	0.117	-0.120	-0.208	-0.242
mean PET	0.021	-0.123	-0.027	-0.059	-0.006	0.483	-0.140	-0.425	-0.290	0.045
Explained Variance Ratio	25.7%	16.7%	14.7%	8.9%	5.2%	4.7%	3.9%	3.2%	2.5%	2.1%
Cumulative Variance	25.7%	42.4%	57.1%	66.1%	71.3%	76.0%	80.0%	83.2%	85.6%	87.8%

Explained Variance: The first principal component (PC1) captures approximately 25.7% of the variance, with PC2 and PC3 accounting for 16.7% and 14.7%, respectively. The first five components collectively account for 71.3% of the variance, while the first ten components cover around 87.8%, providing a comprehensive view of the test_DATASET.

Component Loadings: According to Table 11, the PCA loadings reveal each feature's influence on the principal components:

PC1: Climatic and topographic attributes dominate, explaining 23.9% of the variance. Features with high positive loadings include the probability of snowfall (0.270), aridity index (0.270), area occupied by sedimentary soils (0.261), elevation (0.255), and agricultural land area (0.252). Meanwhile, attributes such as average precipitation (-0.230), mean runoff coefficient (-0.184), average temperature (-0.208), and average slope (-0.231) show significant negative loadings. This pattern suggests that hydrological dynamics are strongly influenced by both climatic (e.g., aridity, precipitation) and topographical factors (e.g., elevation, slope), highlighting complex hydrological responses in relation to these variables. High aridity and elevation values correspond to increased runoff variability, while higher precipitation and temperature appear to dampen streamflow variability.

PC2: Geological and morphological characteristics are prominent in PC2, capturing 16.5% of the variance. Features with high positive loadings include river confluence density (CONF_DEN: 0.284), average soil conductivity (0.267), area of calcareous rocks (0.257), catchment size (0.249), and max slope (0.232). Conversely, the area occupied by conglomerate rocks (-0.232) has a strong negative loading. These loadings suggest that the density of river confluences and soil type diversity are critical factors shaping catchment hydrological behavior, with more complex confluence networks likely introducing additional variability in hydrological responses.

PC3: Vegetation cover and land use attributes are influential in PC3, explaining 13.8% of the variance. High loadings include maximum (0.258) and average elevation (0.276), area covered by broadleaf forests (0.247), and bedrock hardness (0.236). Average precipitation (0.225) shows a positive loading, while coefficient of flow variation (-0.274) and areas of pasture (-0.229) and urban land (-0.214) have negative loadings. This component suggests that broadleaf forests enhance water retention, while urbanization may increase impervious surfaces, complicating runoff dynamics. Negative loadings for flow variability and urban areas reflect the crucial role of land cover in hydrological behavior.

Other Components: Attributes such as soil conductivity, permeability, and anthropogenic influences further contribute to variance in the higher components. These factors highlight the impact of land use and soil properties on natural hydrological processes, modifying water storage and flow patterns.

Interpretation of PCA Outcomes

The PCA results reveal the interconnectedness of catchment attributes in driving hydrological behavior:

Component 1 (PC1): Highlights the combined influence of climatic and topographical factors, suggesting that aridity and elevation enhance runoff variability while precipitation and temperature have a buffering effect that facilitate predictions by LSTM architectures.

Component 2 (PC2): Emphasizes the importance of catchment morphology and geology. Catchments with higher confluence densities and unique soil types exhibit increased variability in hydrological responses, reflecting the complex geological landscape's impact that makes it harder for LSTMs to accurately predict runoff from rainfall events.

Component 3 (PC3): Underlines the significance of vegetation cover and land use. Broadleaf forests help retain water, while urban areas disrupt natural runoff processes. Flow variability is a distinguishing factor, with specific land cover types amplifying or mitigating runoff responses. These behaviors can indirectly affect LSTMs' performance in different locations and catchments.

The PCA outcomes underscore the intricate nature of hydrological dynamics in catchments, where climate, topography, geology, and land use interact in complex ways to shape runoff and streamflow behavior. By interpreting optimized LSTM results through PCA, we gain insights into these interrelationships, which can guide future modeling efforts and improve rainfall-runoff predictions under changing environmental conditions. These findings are especially valuable for enhancing regional hydrological deep learning (DL) models, like LSTMs, which are critical for managing water resources amid variable climatic scenarios. The results highlight the potential of PCA-informed approaches to optimize DL models, enabling more accurate rainfall-runoff predictions in response to complex, shifting climate patterns.

7.4. Discussion

7.4.1. Intersection of physical hydrology and AI/DL models

The findings in Chapter 7 reveal the intricate relationships between catchment attributes and the performance metrics of optimized regional Long Short-Term Memory (LSTM) networks in rainfall-runoff modeling. Despite the absence of explicit catchment features in the training phase, the LSTM networks showcased a remarkable ability to capture complex latent relationships inherent in hydrometeorological data. This capability underscores the potential of deep learning (DL) techniques and hyperparameter optimization to interpret underlying hydrometeorological dynamics, aligning with advances in machine learning that highlight deep neural networks' strengths in identifying patterns within extensive datasets, often exceeding the capabilities of traditional models.

Moreover, integrating both timeseries data and catchment attribute information into the training process could expedite model convergence and improve the fitting process, leading to computational efficiency gains. By incorporating both hydrometeorological and environmental features, DL models can uncover critical relationships earlier, facilitating more efficient training and enhancing predictive accuracy. This advocates for a multifaceted approach to DL model training that incorporates diverse factors instead of relying solely on historical timeseries data.

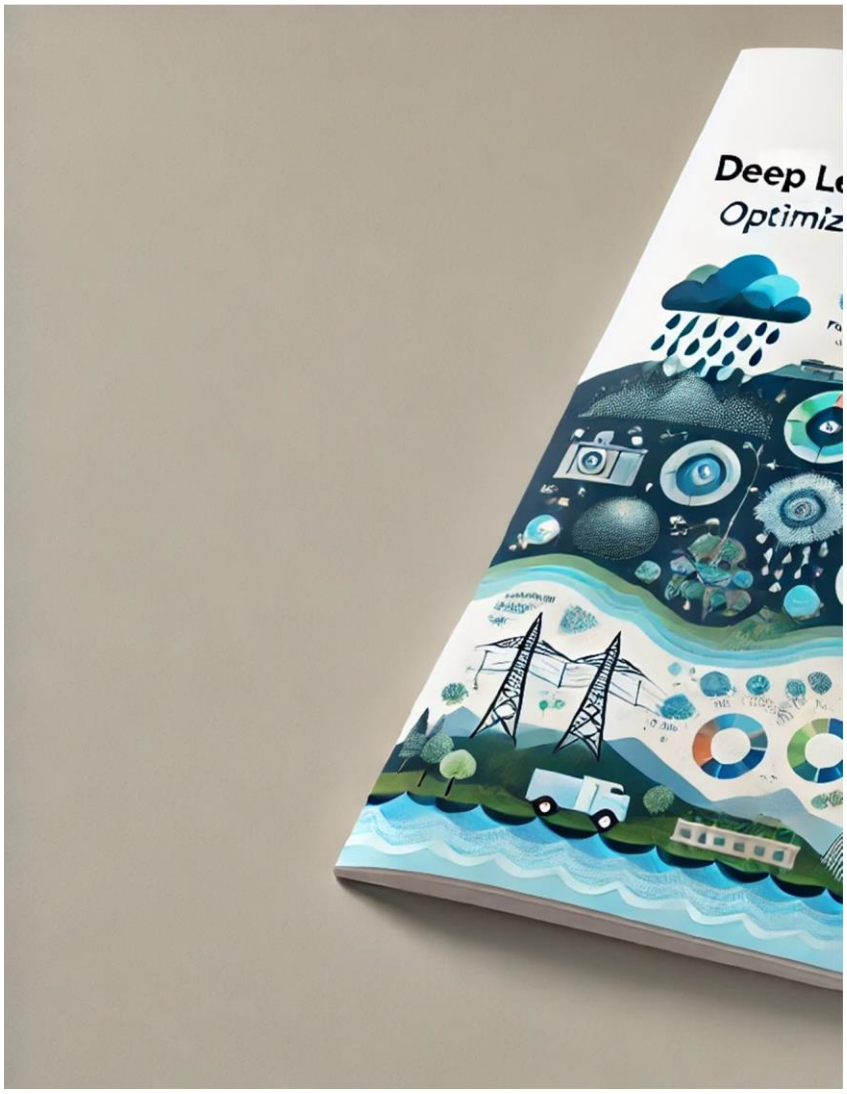
Additionally, the LSTM networks' demonstrated ability to discern relationships from extensive hydrometeorological datasets highlights their potential as powerful modeling tools

in hydrology. These findings suggest that LSTMs are effective in capturing nonlinear relationships and temporal dependencies often present in hydrological systems, reinforcing the cautions presented by Kratzert et al. (2024) against training LSTMs exclusively on single catchments. Training on data from multiple catchments is essential for capturing the variety of hydrological behaviors across geographic and climatic contexts, helping to avoid overfitting to individual water basin characteristics (catchments' uniqueness paradigm, Beven, 2020) and enhancing model robustness and generalizability.

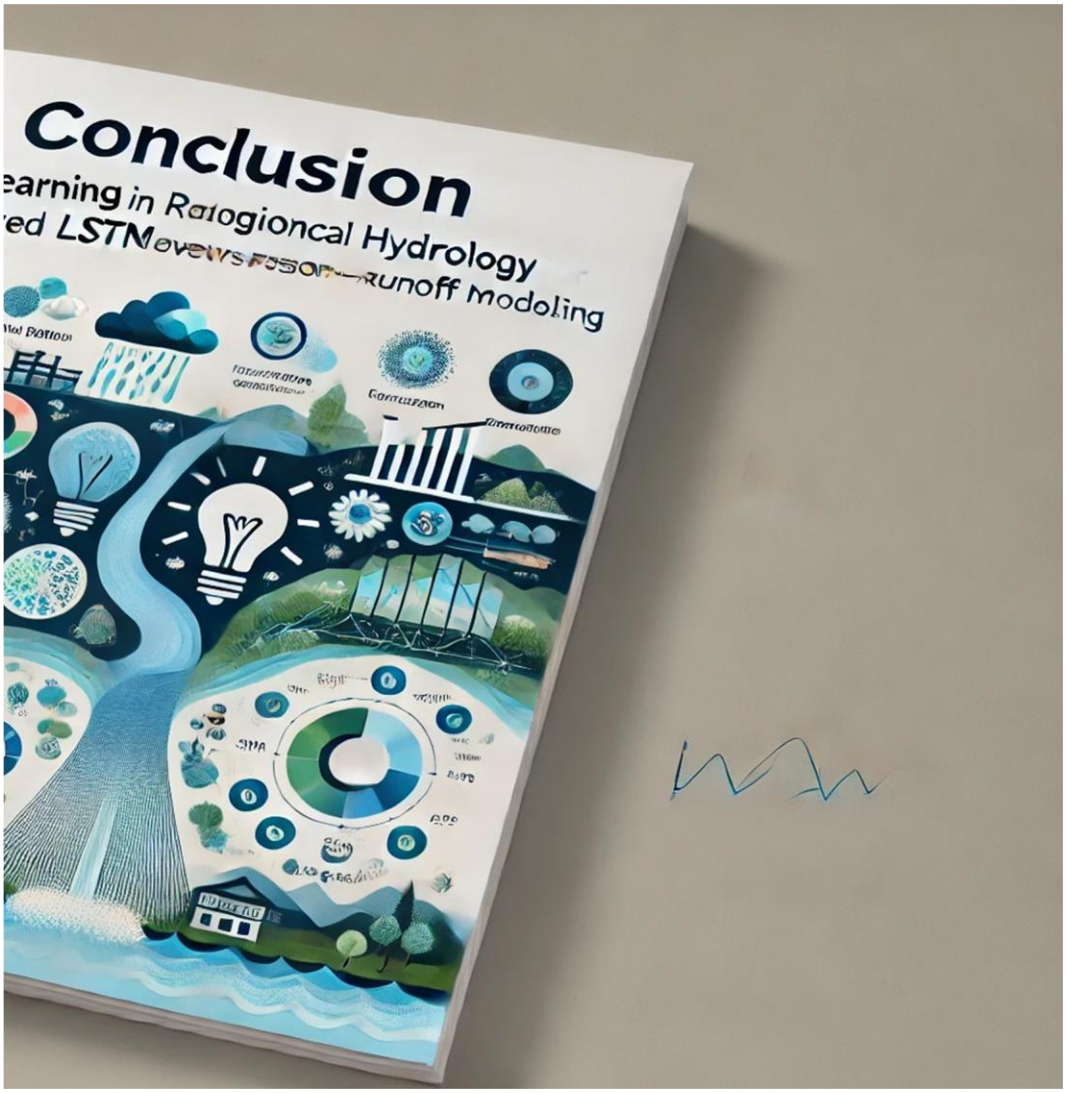
The implications of these results extend beyond model performance to water resource management strategies. An improved understanding of catchment attributes and hydrological responses aids informed decision-making in water resource management, particularly amidst the era of climate variability. As hydrological extremes become more frequent, accurate predictions are crucial for effective management and mitigation.

Furthermore, the inclusion of rating curves—describing the relationship between streamflow and water level—enhances the LSTM's ability to interpret catchment dynamics. Our two-target LSTM setting, which predicts both streamflow and water level, leverages rating curves that encapsulate vital information about riverbed morphology, hydraulic structures, and catchment responses to precipitation. By analyzing latent information in these curves in terms of relations between streamflow and water levels, LSTMs can implicitly capture unique catchment attributes through the patterns within rating curves, ultimately learn unique catchments' characteristics. This ability to learn implicit representations from environmental data underscores the potential of DL models in hydrology, especially where direct information on catchment characteristics may be sparse or inconsistent.

In summary, while optimized regional LSTMs reveal intricate relationships among catchment attributes and their performance, combining timeseries data with catchment characteristics during training is essential for maximizing predictive power of LSTMs. This holistic approach to hydrological DL modeling fosters more effective water resource management strategies in the face of evolving environmental challenges. Future research should prioritize refining integrated modeling approaches, exploring ensemble techniques, and enhancing interpretability to further exploit DL capabilities in hydrology.



Conclusions



Overall Findings

and

Future Works

This thesis provides a detailed study of Long Short-Term Memory (LSTM) networks in regional rainfall-runoff modeling, examining their potential to enhance the predictive accuracy of streamflow and water level in multiple catchments. The study's findings reinforce the value of hyperparameter optimization and ensemble learning to improve LSTM performance across diverse hydrological settings. In addition, it highlights the importance of nuanced, data-driven approaches that account for regional climate variability, catchment-specific attributes, and the unique hydrological responses characteristic of areas like the Basque Country, Spain. This concluding section synthesizes key findings from each chapter and suggests future research directions to advance the role of AI-driven models in hydrology and rainfall-runoff modeling, particularly in the new era that climate change poses evolving challenges in front of us.

Key Findings

Optimizing LSTMs in Regional Hydrological Modeling by Random Search

The optimized LSTM models developed in this thesis demonstrate high predictive accuracy in regional rainfall-runoff modeling. In particular, Chapter 4 confirmed that systematic hyperparameter optimization via random search could yield robust regional models. This approach, tested across 40 catchments, enabled the creation of the Regional Optimal (RO) and Enhanced Regional Optimal (ERO) models. Despite requiring only 100 iterations, the RO model achieved high overall Nash-Sutcliffe Efficiency (NSE) and Kling-Gupta Efficiency (KGE) scores. Subsequent refinement with 1,000 iterations in the ERO model further validated the efficacy of random search, demonstrating that carefully designed search spaces can produce models with high predictive accuracy while balancing computational efficiency.

The research underscores the significance of simultaneously tuning multiple hyperparameters, especially in diverse regional settings. While increased search iterations tend to improve accuracy, they also demand greater computational resources. The RO and ERO models exemplify the trade-off between efficiency and model maturity (generating more accurate predictions in as many places as possible), with the RO model proving particularly useful in cases where computational resources are constrained. This balance is critical for practical hydrological applications, where resources for extensive model training may be limited, yet accuracy remains paramount.

Moreover, this chapter discovered statistically significant differences between the performance of two regionally optimized networks of RO and ERO in different locations. This suggests different hyperparameter configured LSTM networks, learn and perform differently in different catchments. This finding opened the road for next step in Chapter 5 to test possibility of prediction improvements by ensembles of different regionally optimized LSTM configurations.

Ensemble Learning of optimized regional LSTMs, enhanced predictions

Chapter 5 explored ensemble learning, aiming to boost the accuracy of rainfall-runoff models by leveraging multiple optimized regional LSTM configurations. The Catchment-wise Configs ensemble, which tailors hyperparameters to each catchment's unique hydrological characteristics, achieved the highest performance. This finding illustrates the benefits of accounting for catchment-specific variability in ensemble learning, as such approaches offer resilience against overfitting while adapting to diverse hydrological processes. The Catchment-wise ensemble approach effectively captures short-term water retention and travel times in flashier, humid catchments—characteristics often encountered in the Basque Country—demonstrating that model adaptability can significantly enhance predictive outcomes in regions with distinct hydrological regimes, even though at an aggregated scale they are homogeneous. Moreover, ensemble deep learning resolved the conflict of prediction in intervened catchments under human fingerprints (e.g., those having reservoirs and human-managed flow regimes.)

Importance of different LSTM hyperparameters in Regional Adaptability

Chapter 6 applied machine learning techniques, including Random Forest (RF) regression and Principal Component Analysis (PCA), to investigate the influence of individual hyperparameters on LSTM performance after 1000 random experiments. The analysis revealed that certain hyperparameters, such as input sequence length, have a significant impact, varying based on the hydrological characteristics of the catchment. This finding underscores the importance of adaptive optimization strategies that account for local variability in regional models. The thesis highlights that a one-size-fits-all optimization approach for deep learning models often falls short in hydrology. However, treating it as an ensemble of different optimized networks—each tailored to specific conditions—can effectively address the diversity of catchment characteristics and ensure optimal model performance.

By elucidating the impact of hyperparameters in different contexts, this thesis highlights the importance of adaptive modeling frameworks that consider both local and regional characteristics. These insights can inform future model improvements in regions where environmental factors vary significantly across space, underscoring the need for nuanced tuning strategies to achieve reliable predictions.

Implicit Learning of Hydrological Features

Chapter 7 investigated the capability of regionally optimized LSTM networks to learn latent hydrological features solely from hydrometeorological data, without explicit access to catchment-specific attributes. The results revealed that optimized LSTMs can effectively capture implicit hydrological dynamics, suggesting that these models can generalize well in regions with complex hydrological responses, such as the flash flood-prone Basque Country catchments. This capability is advantageous in areas where detailed catchment data may be unavailable, as it offers a scalable approach for hydrological modeling across regions with limited data resources for future research.

The study also highlighted the significant influence of catchment-specific attributes on LSTM performance. Attributes such as catchment area, mean slope, land use, and stream density were found to affect the accuracy of LSTMs' predictions, particularly in scenarios requiring regional model generalization. For instance, in catchments with steep slopes and higher stream densities, more accurate predictions were achieved. The findings underline the importance of tailoring model configurations to reflect the hydrological characteristics of individual catchments, ensuring improved prediction capabilities and accuracy.

The LSTMs' implicit learning of hydrological processes underscores the potential of DL models to function as data-driven hydrology tools, providing insights into complex catchment behaviors that may otherwise require extensive data collection. This research makes a contribution moving towards the development of adaptable, scalable DL frameworks that can generalize across catchments and aid in the creation of effective, AI-based hydrological models suitable for diverse applications.

Future Research Directions

Expanding Hyperparameter Optimization Techniques: While random search proved effective in hyperparameter optimization, future studies could explore more sophisticated techniques, such as Bayesian optimization, to refine model tuning further. Incorporating methods like clustering-based optimization or alternative ensemble strategies may identify better configurations that balance accuracy with computational efficiency. Uncertainty quantification methods, including Bayesian LSTMs, Monte Carlo dropout, and Mixture Density Networks, could also enhance prediction reliability, an important consideration as climate variability introduces heightened risk in regions prone to hydrological extremes.

Real-Time Adaptation and Dynamic Model Tuning: The ongoing challenges posed by climate change call for hydrological models capable of real-time adaptation. Future research should prioritize developing AI-driven frameworks that can dynamically adjust model structures and hyperparameters in response to changing environmental conditions with regard to AI approaches such as reinforcement learning. Such adaptive frameworks would enhance resilience, enabling accurate predictions in areas experiencing altered precipitation, streamflow variability, and increased extreme weather events. These real-time systems could also deepen our understanding of how catchments respond to climate variability, offering critical insights for managing flood and drought risks.

Hybrid Modeling Approaches: Integrating LSTM models with traditional physically-based hydrological models present an opportunity to leverage the strengths of both approaches. Hybrid models combining physically-based principles with data-driven insights could improve accuracy across multiple timescales and catchment types, extending the applicability of LSTMs to varied hydrological environments. Further exploration of DL ensemble models, particularly those that combine LSTMs with Transformers, could increase robustness, enabling better performance in regions affected by seasonal shifts and climate-driven hydrological changes.

Expanding Hyperparameter Exploration Using RF-Guided Optimization: Chapter 6 demonstrated the potential of RF regression in identifying influential hyperparameters post-random search. Expanding this approach to guide broader hyperparameter exploration may reduce computational costs, especially when optimizing models to account for climate-induced hydrological shifts. By using RF models to inform random search strategies, researchers could achieve a more targeted exploration of hyperparameter spaces, improving model accuracy while maintaining computational efficiency.

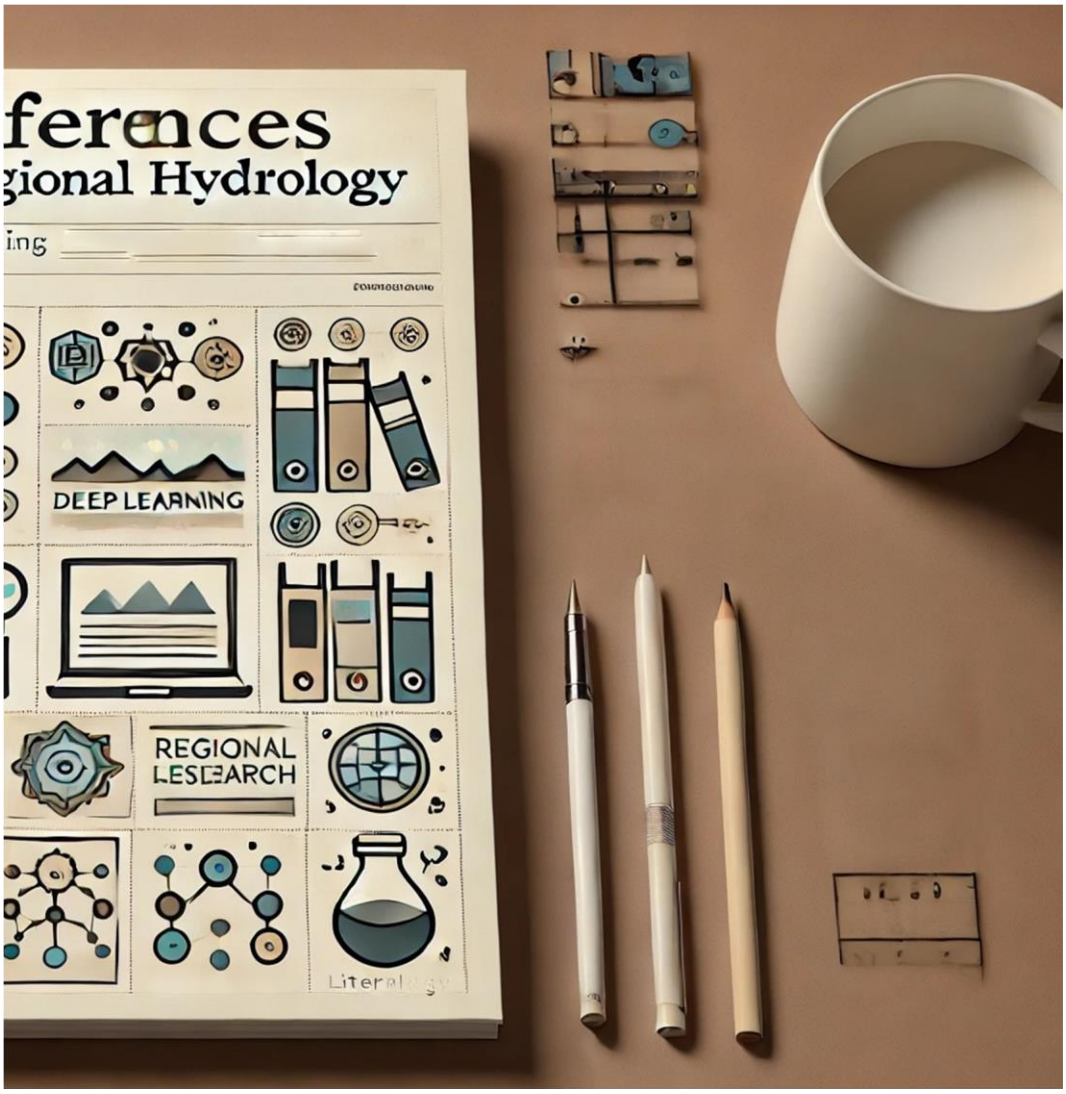
Extending Model Validation Across Diverse Regions: To validate the robustness of optimized LSTM frameworks, future studies should apply these methods in regions with diverse hydrological and climatic characteristics. Evaluating model flexibility in different environments is essential as climate change amplifies hydrological extremes. Integrating DL models with real-time environmental data from remote sensing and Internet of Things (IoT) technologies could further enhance adaptability, enabling timely responses to climate-affected water resource needs.

Integrating Climate Modeling with LSTM Networks for Water Resource Management: In light of climate change, combining LSTM hydrological models with climate projections can enhance predictions of floods and droughts. Research on frameworks that incorporate climate model data into regional LSTM forecasts could advance water resource management, supporting adaptive strategies and proactive responses to shifting hydrological risks.

Enhancing Interpretability with Explainable AI (XAI): Future research should prioritize improving the interpretability of DL models through Explainable AI (XAI) techniques, fostering collaboration between AI specialists and hydrologists. By clarifying how DL models weigh different features and adjust to new conditions, XAI could increase trust in AI-driven hydrological forecasting, particularly in applications sensitive to climate-induced variability.

Finally, this thesis demonstrated the capabilities of optimized LSTM networks for regional hydrological rainfall-runoff in humid flashy contexts. Through systematic hyperparameter optimization, ensemble learning, and implicit feature learning, this research establishes a framework for accurate rainfall-runoff predictions. These findings provide valuable guidance for future AI-driven hydrological research, with significant implications for water resource management, flood risk mitigation, and environmental sustainability.

References


ferences
gional Hydrology
ing
DEEP LEARNING
REGIONAL
LESEARCH

Abrahart, R. J., Anctil, F., Coulibaly, P., Dawson, C. W., Mount, N. J., See, L. M., Shamseldin, A. Y., Solomatine, D. P., Toth, E., & Wilby, R. L. (2012). Two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting. Progress in Physical Geography: Earth and Environment, 36(4), 480–513. doi: 10.1177/0309133312444943

Addor, N., Newman, A. J., Mizukami, N., and Clark, M. P. (2017). The CAMELS data set: catchment attributes and meteorology for large-sample studies, Hydrology and Earth System Science, 21, 5293–5313, doi: 10.5194/hess-21-5293-2017

Ahmadi, S. M., Balahang, S., & Abolfathi, S. (2024). Predicting the hydraulic response of critical transport infrastructures during extreme flood events. Engineering Applications of Artificial Intelligence, 133, 108573. doi: 10.1016/j.engappai.2024.108573

Andersen, J., Refsgaard, J. C., & Jensen, K. H. (2001). Distributed hydrological modelling of the Senegal River basin – Model construction and validation. Journal of Hydrology, 247, 200–214.

Anderson M. G., (2005). Editor. Encyclopedia of Hydrological Sciences. John Wiley & Sons, Chichester, UK, 5 volumes

Arsenault, R., Martel, J.-L., Brunet, F., Brissette, F., and Mai, J., (2023). Continuous streamflow prediction in ungauged basins: long short-term memory neural networks clearly outperform traditional hydrological models, Hydrol. Earth Syst. Sci., 27, 139–157, doi: 10.5194/hess-27-139-2023

Arsenault, R., Brissette, F., & Martel, J. L. (2018). The hazards of split-sample validation in hydrological model calibration. Journal of Hydrology, 566(September), 346–362. doi: 10.1016/j.jhydrol.2018.09.027

Bales, r. C. (2015). Hydrology, floods and droughts | overview (g. R. North, j. Pyle, & f. B. T.-e. Of a. S. (second e. Zhang (eds.); pp. 180–184). Academic press. doi: 10.1016/b978-0-12-382225-3.00166-3

Başağaoğlu, H., Chakraborty, D., Lago, C. do, Gutierrez, L., Şahinli, M. A., Giacomoni, M., Furl, C., Mirchi, A., Moriasi, D., & Şengör, S. S. (2022). A Review on Interpretable and Explainable Artificial Intelligence in Hydroclimatic Applications. Water, 14(8), 1230. doi: 10.3390/w14081230

Breiman, L. (1996). Bagging Predictors. Machine Learning, 24(2), 123–140. doi: 10.1023/A:1018054314350

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13(Aug), 281-305. https://www.jmlr.org/papers/v13/bergstra12a.html

Bergström, S. (1991). Principles and confidence in hydrological modelling. Nordic Hydrology, 22, 123–136.

Beven, K. (2020). Deep learning, hydrological processes and the uniqueness of place. Hydrological Processes, 34(16), 3608–3613. doi: 10.1002/hyp.13805

Beven, K. J. (2018). On hypothesis testing in hydrology: Why falsification of models is still a really good idea. Wiley Interdisciplinary Reviews Water, 5(3), e1278. doi: 10.1002/wat2.1278

Beven, K. J. (2012). Rainfall-runoff modelling: The primer. Chichester, West Sussex: Wiley-Blackwell.

Beven, K. (2002). Towards an alternative blueprint for a physically based digitally simulated hydrologic response modelling system. Hydrological Processes, 16(2), 189–206.

Beven, K. J. (2000). Uniqueness of place and process representations in hydrological modelling. Hydrology and Earth System Sciences, 4(2), 203–213. doi: 10.5194/hess-4-203-2000

Beven, K. (1989). Changing ideas in hydrology – The case of physically based models. Journal of Hydrology, 105, 157–172.

Blöschl, G., Bierkens, M. F. P., Chambel, A., Cudennec, C., Destouni, G., Fiori, A., Kirchner, J. W., McDonnell, J. J., Savenije, H. H. G., Sivapalan, M., Stumpp, C., Toth, E., Volpi, E., Carr, G., Lupton, C., Salinas, J., Széles, B., Viglione, A., Aksoy, H., … Zhang, Y. (2019). Twenty-three unsolved problems in hydrology (UPH) – a community perspective. Hydrological Sciences Journal, 64(10), 1141–1158. doi: 10.1080/02626667.2019.1620507

Blöschl, G. and Sivapalan, M. (1995), Scale issues in hydrological modelling: A review. Hydrol. Process., 9: 251-290. 10.1002/hyp.3360090305

Box, G. E. P. (1976). Science and Statistics. Journal of the American Statistical Association, 71(356), 791. doi: 10.2307/2286841

Box, G. E. P., & Jenkins, G. M. (1970). Time series analysis, forecasting and control. Holden-Day Inc.

Burnash, R. J. C., Ferral, R. L., McGuire, R. A. (1973). A generalized streamflow simulation system: Conceptual modeling for digital computers. Joint Federal-State River Forecast Center., United States., & California

Carneiro, T., Rocha, P., Carvalho, P., Fernández-Ramírez, L., (2022). Ridge regression ensemble of machine learning models applied to solar and wind forecasting in Brazil and Spain, Applied Energy, Volume 314, 118936, ISSN 0306-2619, doi: 10.1016/j.apenergy.2022.118936

Chow, V. T., Maidment, D. R., & Mays, L. W. (1988). Applied hydrology. New York: McGraw-Hill.

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., Wagener, T., & Hay, L. E. (2008). Framework for understanding structural errors (FUSE): A modular framework to diagnose differences between hydrological models. Water Resources Research, 44, W00B02.

Coleman, J. S. M., & Law, K. T. B. T.-R. M. in E. S. and E. S. (2015). Meteorology. Elsevier. doi: 10.1016/B978-0-12-409548-9.09492-6

De la Fuente, L., Ehsani, E., Gupta, H., Condon, L., (2022). Hydro-LSTM: A Hydrological Approach to LSTM ML based Modeling. (HydroML Symposium 2022 – Penn State University)

Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. In Multiple Classifier Systems (Vol. 1857, pp. 1–15). Springer. doi: 10.1007/3-540-45014-9_1

Donnelly, J., Daneshkhah, A., & Abolfathi, S. (2024a). Forecasting global climate drivers using Gaussian processes and convolutional autoencoders. Engineering Applications of Artificial Intelligence, 128, 107536. doi: 10.1016/j.engappai.2023.107536

Donnelly, J., Daneshkhah, A., & Abolfathi, S. (2024b). Physics-informed neural networks as surrogate models of hydrodynamic simulators. Science of The Total Environment, 912, 168814. doi: 10.1016/j.scitotenv.2023.168814

Ershadi, A., McCabe, M., Walker, J., & Evans, J. (2011). Evaluation of energy balance, combination and complementary schemes for estimation of evaporation. In S. W. Franks, E. Boegh, E. Blyth, D. M. Hannah, & K. K. Yilmaz (Eds.), Proceedings of IAHS Lead Symposia (Vol. 344, pp. 52 - 56). IAHS Press

Fang, K., Shen, C., Kifer, D., & Yang, X. (2017). Prolongation of SMAP to Spatio-temporally Seamless Coverage of Continental US Using a Deep Learning Neural Network. Geophysical Research Letters, 44, 11030–11039. doi: 10.1002/2017GL075619

Fang, K., Kifer, D., Lawson, K., & Shen, C. (2020b). Evaluating the potential and challenges of an uncertainty quantification method for long short-term memory models for soil moisture predictions. Water Resources Research, 56, e2020WR028095. doi: 10.1029/2020WR028095

Feng, D., Fang, K., & Shen, C. (2020a). Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales. Water Resources Research, 56, e2019WR026793. doi: 10.1029/2019WR026793

Fenicia, F., Kavetski, D., and Savenije, H. H. G. (2011), Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development, Water Resour. Res., 47, W11510, doi: 10.1029/2010WR010174

Fenicia, F., J. J. McDonnell, and H. H. G. Savenije (2008), Learning from model improvement: On the contribution of complementary data to process understanding, Water Resour. Res., 44, W06419, doi: 10.1029/2007WR006386

Ferré, t. P. A., & warrick, a. W. (2005). Infiltration (d. B. T.-e. Of s. In the e. Hillel (ed.); pp. 254–260). Elsevier. doi: 10.1016/b0-12-348530-4/00382-9

Fu, J., Chu, J., Guo, P., & Chen, Z. (2019). Condition Monitoring of Wind Turbine Gearbox Bearing Based on Deep Learning Model. IEEE Access, 7, 57078–57087. 10.1109/ACCESS.2019.2912621

Frame, J.M., Kratzert, F., Ullrich, P.A., Gupta, H.V., & Nearing, G.S. (2022). On strictly enforced mass conservation constraints for modeling the rainfall-runoff process.

Frame, J., Kratzert, F., Klotz, D., Gauch, M., Shelev, G., Gilon, O., Qualls, L. M., Gupta, H. V., and Nearing, G. S. (2021a). Deep learning rainfall-runoff predictions of extreme events, Hydrol. Earth Syst. Sci. Discuss., doi: 10.5194/hess-2021-423

Frame, J.M., Kratzert, F., Raney, A., Rahman, M., Salas, F.R., and Nearing, G.S.. (2021). " Post-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. doi: 10.1111/1752-1688.12964

Gharari, S., Gupta, H. V., Clark, M. P., Hrachowitz, M., Fenicia, F., Matgen, P., & Savenije, H. H. G. (2021). Understanding the information content in the hierarchy of model development decisions: Learning from data. Water Resources Research, 57, e2020WR027948. doi: 10.1029/2020WR027948

Gauch, M., Kratzert, F., Gilon, O., Gupta, H., Mai, J., Nearing, G., et al. (2023). In defense of metrics: Metrics sufficiently encode typical human preferences regarding hydrological model performance. Water Resources Research, 59, e2022WR033918. doi: 10.1029/2022WR033918

Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Lin, J., and Hochreiter, S. (2021). Rainfall–runoff prediction at multiple timescales with a single Long Short-Term Memory network, Hydrol. Earth Syst. Sci., 25, 2045–2062, doi: 10.5194/hess-25-2045-2021

Gers, F. A., Schmidhuber, J., and Cummins, F. (1999). Learning to forget: continual prediction with LSTM, IET Conference Proceedings, pp. 850–855.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. The MIT Press. Retrieved from https://www.deeplearningbook.org

Glorot, X.; & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, in Proceedings of Machine Learning Research, 9:249-256 Available from https://proceedings.mlr.press/v9/glorot10a.html

Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., and Schmidhuber, J. (2017). LSTM: A Search Space Odyssey, IEEE Transactions on Neural Networks and Learning Systems, 28, 2222–2232, doi: 10.1109/TNNLS.2016.2582924

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

Gupta, H. V., Clark, M. P., Vrugt, J. A., Abramowitz, G., & Ye, M. (2012). Towards a comprehensive assessment of model structural adequacy. Water Resources Research, 48, W08301. doi: 10.1029/2011WR011044

Hanrahan, G. (2012). Chapter 1 - Introduction to Environmental Chemistry (G. B. T.-K. C. in E. C. Hanrahan (ed.); pp. 3–30). Academic Press. doi: 10.1016/B978-0-12-374993-2.10001-9

Hargreaves, H. & Allen. R. (2003). History and Evaluation of Hargreaves Evapotranspiration Equation. Journal of Irrigation and Drainage Engineering, 129(1), 53–63. doi: 10.1061/(ASCE)0733-9437(2003)129:1(53)

Harman, C., & Troch, P. A. (2014). What makes Darwinian hydrology "Darwinian"? Asking a different kind of question about landscapes. Hydrology and Earth System Sciences, 18(2), 417–433. doi: 10.5194/hess-18-417-2014

Harte, J. (2002). Toward a Synthesis of the Newtonian and Darwinian Worldviews. Physics Today, 55(10), 29–34. doi: 10.1063/1.1522164

Hashemi, R., Brigode, P., Garambois, P.-A., and Javelle, P. (2022). How can we benefit from regime information to make more effective use of long short-term memory (LSTM) runoff models?, Hydrol. Earth Syst. Sci., 26, 5793–5816, doi: 10.5194/hess-26-5793-2022

Hasiotis, S. T., Kraus, M. J., & Demko, T. M. (2007). CHAPTER 11 - Climatic Controls on Continental Trace Fossils (W. B. T.-T. F. MILLER (ed.); pp. 172–195). Elsevier. doi: 10.1016/B978-044452949-7/50137-6

Hawkins, R. H., Hjelmfelt, A. T., Jr., Zevenbergen, A. W., (1985). Runoff Probability, Storm Depth, and Curve Numbers. Journal of Irrigation and Drainage Engineering, 111(4), 330–340. doi: 10.1061/(ASCE)0733-9437(1985)111:4(330)

Hansen, J. R., Refsgaard, J. C., Hansen, S., & Ernstsen, V. (2007). Problems with heterogeneity in physically based agricultural catchment models. Journal of Hydrology, 342(1–2), 1–16.

Hilbert, D., 1900. Mathematische Probleme. Nachrichten von der Königlichen Gesellschaft der Wissenschaften zu Göttingen – mathematisch-Physikalische Klasse. 253–297.

Hillel, d. (2008). Soil-water dynamics. Soil in the environment, 91–101. doi: 10.1016/b978-0-12-348536-6.50012-5

Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation. 9(8), 1735–1780 (1997)

Hoedt, P.-J., Kratzert, F., Klotz, D., Halmich, C., Holzleitner, M., Nearing, G. S., Hochreiter, S., and Klambauer, G. (2021). MC-LSTM: Mass- Conserving LSTM, in: Proceedings of the 38th International Conference on Machine Learning, edited by Meila, M. and Zhang, T., vol. 139 of Proceedings of Machine Learning Research, pp. 4275–4286, PMLR, http://proceedings.mlr.press/v139/hoedt21a.html

Hofmann, T., Schölkopf, B., & Smola, A. J. (2008). Kernel methods in machine learning. The Annals of Statistics, 36(3). doi: 10.1214/009053607000000677

Höge, M., Scheidegger, A., Baity-Jesi, M., Albert, C., & Fenicia, F. (2022). Improving hydrologic models for predictions and process understanding using Neural ODEs. Hydrology and Earth System Sciences Discussions, 2022, 1–29. doi: 10.5194/hess-2022-56

Höge, M., Wöhling, T., & Nowak, W. (2018). A primer for model selection: The decisive role of model complexity. Water Resources Research, 54, 1688–1715. doi: 10.1002/2017WR021902

Hopmans, J. W. (2011). 2.05 - Infiltration and Unsaturated Zone (P. B. T.-T. on W. S. Wilderer (ed.); pp. 103–114). Elsevier. doi: 10.1016/B978-0-444-53199-5.00031-2

Hosseini, F., Prieto, C., & Álvarez, C. (2025). Ensemble learning of catchment-wise optimized LSTMs enhances regional rainfall-runoff modelling – case Study: Basque Country, Spain. Journal of Hydrology, 646, 132269. doi: 10.1016/j.jhydrol.2024.132269

Hosseini, F., Prieto, C., & Álvarez, C. (2024a). Hyperparameter optimization of regional hydrological LSTMs by random search: A case study from Basque Country, Spain. Journal of Hydrology, 132003. doi: 10.1016/j.jhydrol.2024.132003

Hosseini, F., Prieto, C., Nearing, G., Alvarez, C., and Gauch, M. (2024b) Hydrological Significance of Input Sequence Lengths in LSTM-Based Streamflow Prediction, EGU General Assembly 2024, Vienna, Austria, 14–19 Apr 2024, EGU24-571, doi: 10.5194/egusphere-egu24-571

Hosseini, F., Prieto, C., & Álvarez, C. (2024C), Alpine-Peaks Shape of Optimized Configurations Post Random Search in Regional Hydrological LSTMs, 2024 10th IEEE edition of the International Conference on Optimization and Applications (ICOA), Almeria, Spain, 2024, pp. 1-5, doi: 10.1109/ICOA62581.2024.10754182

Hrachowitz, M., & Clark, M. P. (2017). HESS opinions: The complementary merits of competing modelling philosophies in hydrology. Hydrology and Earth System Sciences, 21, 3953–3873.

Jain, A., & Kumar, A. M. (2007). Hybrid neural network models for hydrologic time series forecasting. Applied Soft Computing, 7(2), 585–592. doi: 10.1016/j.asoc.2006.03.002

Jiang, P., Shuai, P., Chen, X., (2022). Hydrological Model Calibration Using Knowledge-Informed Deep Learning at Coal Creek Watershed, CO. (HydroML Symposium 2022 – Penn State University)

Jiang, S., Zheng, Y., & Solomatine, D. (2020). Improving AI system awareness of geoscience knowledge: Symbiotic integration of physical approaches and deep learning. Geophysical Research Letters, 46, e2020GL088229. doi: 10.1029/2020GL088229

Jozefowicz, R., Zaremba, W., and Sutskever, I. (2015). An Empirical Exploration of Recurrent Network Architectures, in: Proceedings of the 32nd International Conference on Machine Learning, edited by: Bach, F. and Blei, D., vol. 37 of Proceedings of Machine Learning Research, pp. 2342–2350, PMLR, Lille, France.

Ivezic, V., D. Bekic, and R. Zugaj. (2016). A review of procedures for water balance modelling, Journal of Environmental Hydrology, Vol. 25, Paper 4.

Jakeman, A. J., & Hornberger, G. M. (1993). How much complexity is warranted in a rainfall-runoff model? Water Resources Research, 29(8), 2637–2649.

Jang, J. -S. R., ANFIS: adaptive-network-based fuzzy inference system, in IEEE Transactions on Systems, Man, and Cybernetics, vol. 23, no. 3, pp. 665-685, May-June 1993, doi: 10.1109/21.256541

Karpatne et al., (2017). Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data, in IEEE Transactions on Knowledge and Data Engineering, vol. 29, no. 10, pp. 2318-2331, 1 Oct. 2017, doi: 10.1109/TKDE.2017.2720168

Kavetski, D. (2018). Parameter Estimation and Predictive Uncertainty Quantification in Hydrological Modelling. In Q. Duan, F. Pappenberger, J. Thielen, A. Wood, H. L. Cloke, & J. C. Schaake (Eds.), Handbook of Hydrometeorological Ensemble Forecasting (pp. 1–42). Springer Berlin Heidelberg. doi: 10.1007/978-3-642-40457-3_25-1

Kavetski, D., and Fenicia, F. (2011a), Elements of a flexible approach for conceptual hydrological modeling: 2. Application and experimental insights, Water Resour. Res., 47, W11511, doi: 10.1029/2011WR010748

Kavetski, D., Fenicia, F., and Clark, M. P. (2011b), Impact of temporal data resolution on parameter inference and model identification in conceptual hydrological modeling: Insights from an experimental catchment, Water Resour. Res., 47, W05501, doi: 10.1029/2010WR009525

Khoshkalam, et al., (2021). H23A-01 - Improving Daily Streamflow Forecasting Systems in Data-scarce Regions with A Long Short-term Memory Model. AGU Fall meeting, 12-16 Dec 2021

Khosravi, K., Rezaie, F., Cooper, J. R., Kalantari, Z., Abolfathi, S., & Hatamiafkoueieh, J. (2023). Soil water erosion susceptibility assessment using deep learning algorithms. Journal of Hydrology, 618, 129229. doi: 10.1016/j.jhydrol.2023.129229

Kirchner, J. W. (2003). A double paradox in catchment hydrology and geochemistry. Hydrological Processes, 17(4), 871–874. doi: 10.1002/hyp.5108

Kirchner, J. W. (2006). Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. Water Resources Research, 42, W03S04.

Klemeš, V. (1986). Operational testing of hydrological simulation models. Hydrological Sciences Journal, 31(1), 13–24. doi: 10.1080/02626668609491024

Kling, H., Fuchs, M., & Paulin, M. (2012). Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. Journal of Hydrology, 424-425, 264-277.

Klotz, D., Kratzert, F., Gauch, M., Keefe Sampson, A., Brandstetter, J., Klambauer, G., Hochreiter, S., and Nearing, G. (2022). Uncertainty estimation with deep learning for rainfall–runoff modeling, Hydrol. Earth Syst. Sci., 26, 1673–1693, doi: 10.5194/hess-26-1673-2022

Knoben, W. J. M., Freer, J. E., & Woods, R. A. (2019). Technical note: Inherent benchmark or not? Comparing nash-sutcliffe and kling-gupta efficiency scores. Hydrology and Earth System Sciences, 23(10), 4323–4331. doi: 10.5194/hess-23-4323-2019

Konapala, G., Kao, S. C., Painter, S. L., & Lu, D. (2020). Machine learning assisted hybrid models can improve streamflow simulation in diverse catchments across the conterminous US. Environmental Research Letters, 15, 104022.

Kraft, B., Jung, M., Körner, M., & Reichstein, M. (2020). Hybrid modeling: Fusin of a deep learning approach and a physics-based model for global hydrological modeling. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLIIIB2- 2020 2020 XXIV ISPRS Congress (2020 edition), 1537–1544.

Kratzert, F., Gauch, M., Klotz, D., and Nearing, G., (2024). HESS Opinions: Never train an LSTM on a single basin, Hydrol. Earth Syst. Sci. Discuss., doi: 10.5194/hess-2023-275

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), 61. doi: 10.1038/s41597-023-01975-w

Kratzert, F., Gauch, M., Nearing, G., & Klotz, D. (2022). NeuralHydrology — A Python library for Deep Learning research in hydrology. Journal of Open-Source Software, 7(71), 4050. doi: 10.21105/joss.04050

Kratzert, F., Klotz, D., Gauch, M., Klingler, C., Nearing, G., and Hochreiter, S. (2021a). Large-scale river network modeling using Graph Neural Networks, EGU General Assembly 2021, online, 19–30 Apr 2021, EGU21-13375, doi: 10.5194/egusphere-egu21-13375

Kratzert, F., Klotz, D., Hochreiter, S., and Nearing, G. S. (2021b). A note on leveraging synergy in multiple meteorological data sets with deep learning for rainfall–runoff modeling, Hydrol. Earth Syst. Sci., 25, 2685–2703, doi: 10.5194/hess-25-2685-2021

Kratzert, F., Gauch, M., Nearing, G., Hochreiter, S., Klotz, D., (2021c). Niederschlags-Abfluss-Modellierung mit Long Short-Term Memory (LSTM). Österr Wasser- und Abfallw 73, 270–280. doi: 10.1007/s00506-021-00767-z

Kratzert, F., Klotz, D., Klambauer, G., Nearing, G., and Hochreiter, S. (2020a). The performance of LSTM models from basin to continental scales, EGU General Assembly 2020, Online, 4–8 May 2020, EGU2020-8855, doi: 10.5194/egusphere-egu2020-8855

Kratzert, F., Klotz, D., Hochreiter, S., and Nearing, G. S., (2020) A note on leveraging synergy in multiple meteorological datasets with deep learning for rainfall-runoff modeling, Hydrol. Earth Syst. Sci. Discuss., doi: 10.5194/hess-2020-221

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019a). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. Hydrology and Earth System Sciences, 23(12), 5089–5110. doi: 10.5194/hess-23-5089-2019

Kratzert, F., Herrnegger, M., Klotz, D., Hochreiter, S., & Klambauer, G. (2019b). Neural hydrology–interpreting LSTMs in hydrology. W Same G Montavon A Vedaldi L.K. Hansen & K-R (Eds.), Müller In Explainable AI: Interpreting, explaining and visualizing deep learning, (pp. 347–362). Cham, Switzerland: Springer.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018a). Rainfall--runoff modelling using Long Short-Term Memory (LSTM) networks. Hydrology and Earth System Sciences, 22(11), 6005–6022. doi: 10.5194/hess-22-6005-2018

Kratzert, F., Herrnegger, M., Klotz, D., Hochreiter, S., Klambauer, G., (2018b). Do internals of neural networks make sense in the context of hydrology? American Geophysical Union, Fall Meeting 2018.

Kratzert, F., Klotz, D., Herrnegger, M. and Hochreiter, S. (2018c). A glimpse into the Unobserved: Runoff simulation for ungauged catchments with LSTMs, Workshop on Modeling and Decision-Making in the Spatiotemporal Domain, 32nd Conference on NeuralInformation Processing Systems (NeuRIPS 2018), Montréal, Canada.

Krause, P., Boyle, D. P., & Bäse, F. (2005). Comparison of different efficiency criteria for hydrological model assessment. Advances in Geosciences, 5, 89-97.

Kurth, T., Treichler, S., Romero, J., Mudigonda, M., Luehr, N., Phillips, E., ... & Houston, M. (2018, November). Exascale deep learning for climate analytics. In SC18: International Conference for High Performance Computing, Networking, Storage and Analysis (pp. 649-660). IEEE. doi: 10.1109/SC.2018.00054

Ledley, t. S. (2003). Energy balance model, surface (j. R. B. T.-e. Of a. S. Holton (ed.); pp. 747–754). Academic press. doi: 10.1016/b0-12-227090-8/00150-0

Lees, T., Reece, S., Kratzert, F., Klotz, D., Gauch, M., De Bruijn, J., Kumar Sahu, R., Greve, P., Slater, L., and Dadson, S. (2021). Hydrological Concept Formation inside Long Short-Term Memory (LSTM) networks, Hydrol. Earth Syst. Sci. Discuss. [preprint], doi: 10.5194/hess-2021-566

Legates, D. R., & McCabe, G. J. (1999). Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water Resources Research, 35(1), 233-241.

Li, K., Huang, G., Wang, S., Razavi, S., & Zhang, X. (2022a). Development of a joint probabilistic rainfall-runoff model for high-to-extreme flow projections under changing climatic conditions. Water Resources Research, 58, doi: 10.1029/2021WR031557

Li, J., Hsu, K., Jiang, A. (2022b). Rainfall–runoff modelling using attention-based model. (Conf. at HydroML_Symposium 2022 – Penn State University)

Li, W., Kiaghadi, A. & Dawson, C. (2021). High temporal resolution rainfall–runoff modeling using long-short-term-memory (LSTM) networks. Neural Comput & Applic 33, 1261–1278. doi: 10.1007/s00521-020-05010-6

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

Littlewood, I. G. (2002). Improved unit hydrograph characterisation of the daily flow regime (including low flows) for the River Teifi, Wales: towards better rainfall-streamflow

models for regionalisation. Hydrology and Earth System Sciences, 6(5), 899–911. doi: 10.5194/hess-6-899-2002

Ly, S., Charles, C., & Dégre, A. (2013). Different methods for spatial interpolation of rainfall data for operational hydrology and hydrological modeling at watershed scale: a review. Biotechnologie, Agronomie, Société et Environnement, 17, 392-406. doi: 10.6084/M9.FIGSHARE.1225842.V1

Ma, K., Feng, D., Lawson, K., Tsai, W.-P., Liang, C., Huang, X., et al. (2021). Transferring hydrologic data across continents – leveraging data-rich regions to improve hydrologic prediction in data-sparse regions. Water Resources Research, 57, e2020WR028600. doi: 10.1029/2020WR028600

MacQueen, J. (1967). Some Methods for classification and Analysis of Multivariate Observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297). Oakland, CA, USA.

Mahdian, M., Noori, R., Salamattalab, M. M., Heggy, E., Bateni, S. M., Nohegar, A., Hosseinzadeh, M., Siadatmousavi, S. M., Fadaei, M. R., & Abolfathi, S. (2024). Anzali Wetland Crisis: Unraveling the Decline of Iran's Ecological Gem. Journal of Geophysical Research: Atmospheres, 129(4). doi: 10.1029/2023JD039538

Mai, J., Shen, H., Tolson, B. A., Gaborit, É., Arsenault, R., Craig, J. R., Fortin, V., Fry, L. M., Gauch, M., Klotz, D., Kratzert, F., O'Brien, N., Princz, D. G., Rasiya Koya, S., Roy, T., Seglenieks, F., Shrestha, N. K., Temgoua, A. G. T., Vionnet, V., & Waddell, J. W. (2022). The Great Lakes Runoff Intercomparison Project Phase 4: The Great Lakes (GRIP-GL). Hydrology and Earth System Sciences Discussions, 2022, 1–54. doi: 10.5194/hess-2022-113

Maier, H. R., & Dandy, G. C. (2000). Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. Environmental Modelling & Software, 15(1), 101–124. doi: 10.1016/S1364-8152(99)00007-9

Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (1993). Forecasting: Methods and applications. John Wiley & Sons.

Malthus, T. R., & Stimson, S. C. (2018). An Essay on the Principle of Population: The 1803 Edition. Yale University Press. (1st ed. was published in 1798.)

Mann, H. B., & Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. The Annals of Mathematical Statistics, 18(1), 50–60. http://www.jstor.org/stable/2236101

Miller, D. H. B. T.-I. G. (Ed.). (1977). Chapter XV - Groundwater and its Outflows into Local Ecosystems. In Water at the Surface of the Earth (Vol. 21, pp. 392–422). Academic Press. doi: 10.1016/S0074-6142(08)60493-3

Mitchell, T. M. (1980). The need for biases in learning generalizations. Technical Report. Rutgers University. Retrieved from https://www.cs.cmu.edu/~tom/pubs/NeedForBias_1980.pdf

Moore, R. J. (2007). The PDM rainfall-runoff model. Hydrology and Earth System Sciences, 11(1), 483–499. doi: 10.5194/hess-11-483-2007

Moore, D. S. (2006). Introduction to the practice of statistics (5th ed.). W.H. Freeman and Co.

Moshe, Z., Metzger, A., Kratzert, F., Morin, E., Nevo, S., Elidan, G., and Elyaniv, R., (2020). HydroNets: Leveraging River Network Structure and Deep Neural Networks for Hydrologic Modeling , EGU General Assembly 2020, Online, 4–8 May 2020, EGU2020-4135, doi: 10.5194/egusphere-egu2020-4135

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I - A discussion of principles. Journal of Hydrology, 10(3), 282-290. doi: 10.1016/0022-1694(70)90255-6

Nearing, G. S., Klotz, D., Sampson, A. K., Kratzert, F., Gauch, M., Frame, J. M., Shalev, G., and Nevo, S. (2021). Technical Note: Data assimilation and autoregression for using near-real-time streamflow observations in long short-term memory networks, Hydrol. Earth Syst. Sci. Discuss. doi: 10.5194/hess-2021-515

Nearing, G.S., F. Kratzert, A.K. Sampson, C.S. Pelissier, D. Klotz, J.M. Frame, C. Prieto, and H.V. Gupta. (2020a). "What Role Does Hydrological Science Play in the Age of Machine Learning?" Water Resources Research 57: e2020WR028091. doi: 10.1029/2020WR028091

Nearing, G. S., Ruddell, B. L., Bennett, A. R., Prieto, C., & Gupta, H. V. (2020b). Does information theory provide a new paradigm for earth science? Hypothesis testing. Water Resources Research, 56, e2019WR024918. doi: 10.1029/2019WR024918

Nearing, G.S., F. Kratzert, D. Klotz, P.J. Hoedt, G. Klambauer, S. Hochreiter, H. Gupta, S. Nevo, and Y. Matias. (2020c). "A Deep Learning Architecture for Conservative Dynamical Systems: Application to Rainfall-Runoff Modeling." AI for Earth Sciences Workshop at NEURIPS 2020.

Nevo, S., Morin, E., Gerzi Rosenthal, A., Metzger, A., Barshai, C., Weitzner, D., Voloshin, D., Kratzert, F., Elidan, G., Dror, G., Begelman, G., Nearing, G., Shalev, G., Noga, H., Shavitt, I., Yuklea, L., Royz, M., Giladi, N., Peled Levi, N., … Matias, Y. (2021). Flood forecasting with machine learning models in an operational framework. Hydrology and Earth System Sciences Discussions, 2021, 1–31. doi: 10.5194/hess-2021-554

Opitz, D., & Maclin, R. (1999). Popular Ensemble Methods: An Empirical Study. Journal of Artificial Intelligence Research, 11, 169–198. doi: 10.1613/jair.614

Ouyang, W., Lawson, K., Feng, D., Ye, L., Zhang, C., & Shen, C. (2021). Continental-scale streamflow modeling of basins with reservoirs: Towards a coherent deep-learning-based strategy. Journal of Hydrology, 599, 126455. doi: 10.1016/j.jhydrol.2021.126455

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F.

d\textquotesingle Alché-Buc, E. Fox, & R. Garnett (Eds.), Advances in Neural Information Processing Systems (Vol. 32). Curran Associates, Inc.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825-2830.

Perrin, C., Michel, C., & Andréassian, V. (2001). Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments. Journal of Hydrology, 242, 275–301.

Perrin, C., Michel, C., & Andréassian, V. (2003). Improvement of a parsimonious model for streamflow simulation. Journal of Hydrology, 279(1), 275–289. doi: 10.1016/S0022-1694(03)00225-7

Pham, B. T., Luu, C., Phong, T. Van, Trinh, P. T., Shirzadi, A., Renoud, S., Asadi, S., Le, H. Van, von Meding, J., & Clague, J. J. (2021). Can deep learning algorithms outperform benchmark machine learning algorithms in flood susceptibility modeling? Journal of Hydrology, 592, 125615. doi: 10.1016/j.jhydrol.2020.125615

Prieto, C., Le Vine, N., Kavetski, D., Fenicia, F., Scheidegger, A., & Vitolo, C. (2022). An exploration of Bayesian identification of dominant hydrological mechanisms in ungauged catchments. Water Resources Research, 58, doi: 10.1029/2021WR030705

Prieto, C., Kavetski, D., Le Vine, N., Álvarez, C., & Medina, R. (2021). Identification of dominant hydrological mechanisms using Bayesian inference, multiple statistical hypothesis testing, and flexible models. Water Resources Research, 57, e2020WR028338. doi: 10.1029/2020WR028338

Prieto, C., Patel, D., and Han, D. (2020). Preface: Advances in flood risk assessment and management, Nat. Hazards Earth Syst. Sci., 20, 1045–1048, doi: 10.5194/nhess-20-1045-2020

Prieto, C., Le Vine, N., Kavetski, D., García, E., & Medina, R. (2019). Flow prediction in ungauged catchments using probabilistic random forests regionalization and new statistical adequacy tests. Water Resources Research, 55, 4364–4392. doi: 10.1029/2018WR023254

Planck, M., & Ogg, A. (1990). Treatise on thermodynamics. New York: Dover publications.

Polikar, R. (2006). Ensemble Based Systems in Decision Making. IEEE Circuits and Systems Magazine, 6(3), 21–45. doi: 10.1109/MCAS.2006.1688199

Rahmani, F., Lawson, K., Appling, A., Oliver, S., Shen, C. (2021). Process learning of stream temperature modelling using deep learning and big data - Journal Article DP - Earth and Space Science Open Archive doi: 10.1002/essoar.10509644.1

Rasheed, Z., Aravamudan, A., Zhang, X., Anagnostopoulos , G., Nikolopoulos, E., (2022). Using ML and satellite precipitation data for flood prediction. (Conf. at HydroML_Symposium 2022 – Penn State University)

Reed, S., Koren, V., Smith, M., Zhang, Z., Moreda, F., & Seo, D. J. (2004). Overall distributed model intercomparison project results. Journal of Hydrology, 298, 27–60.

Refsgaard, J. C., Stisen, S., & Koch, J. (2022). Hydrological process knowledge in catchment modelling – Lessons and perspectives from 60 years development. Hydrological Processes, 36(1), e14463. doi: 10.1002/hyp.14463

Refsgaard, J. C., & Henriksen, H. J. (2004). Modelling guidelines – Terminology and guiding principles. Advances in Water Resources, 27(1), 71–82.

Refsgaard, J. C., & Knudsen, J. (1996). Operational validation and intercomparison of different types of hydrological models. Water Resources Research, 32(7), 2189–2202.

Refsgaard, J. C. (1996). Terminology, modelling protocol and classification of hydrological model codes. In M. B. Abbott & J. C. Refsgaard (Eds.), Distributed hydrological modelling (pp. 17–39). Kluwer Academics Publishers.

Reichstein, M., Camps-Valls, G., Stevens, B. et al. Deep learning and process understanding for data-driven Earth system science. Nature 566, 195–204 (2019). doi: 10.1038/s41586-019-0912-1

Robertson, D. M., Perlman, H. A., & Narisimhan, T. N. (2022). Hydrological Cycle and Water Budgets (T. Mehner & K. B. T.-E. of I. W. (Second E. Tockner (eds.); pp. 19–27). Elsevier. doi: 10.1016/B978-0-12-819166-8.00008-6

Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., ... & Bengio, Y. (2022). Tackling climate change with machine learning. ACM Computing Surveys (CSUR), 55(2), 1-96.

Rosbjerg, D., & Rodda, J. (2019). IAHS: a brief history of hydrology. History of Geo- and Space Sciences, 10(1), 109–118. doi: 10.5194/hgss-10-109-2019

Rumelhart, D., Hinton, G. & Williams, R. Learning representations by back-propagating errors. Nature 323, 533–536 (1986). doi: 10.1038/323533a0

Russell, S. J., & Norvig, P. (2020). Artificial intelligence: A modern approach. (4th ed.). Boston: Pearson. ISBN 13: 978-1-292-40113-3

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K.-R. (2021). Explaining deep neural networks and beyond: A review of methods and applications in bioinformatics, signal and image analysis. IEEE Signal Processing Magazine, 38(6), 19-48. doi: 10.1109/JPROC.2021.3060483

Schaefli, B., & Gupta, H. V. (2007). Do Nash values have value? Hydrological Processes: An International Journal, 21(15), 2075–2080

SCS. (1972). National Engineering Handbook, Section 4. Hydrology, Soil Conservation Service, US Department of Agriculture: Washington, DC.

Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding Machine Learning: From Theory to Algorithms. Cambridge: Cambridge University Press

Shen, H., Tolson, B. A., & Mai, J. (2022). Time to update the split-sample approach in hydrological model calibration. Water Resources Research, 58, e2021WR031523. doi: 10.1029/2021WR031523

Shen, C. and Lawson, K. (2021). Applications of Deep Learning in Hydrology. In Deep Learning for the Earth Sciences (eds G. Camps-Valls, D. Tuia, X.X. Zhu and M. Reichstein). doi: 10.1002/9781119646181.ch19

Shen, C., Chen, X., and Laloy, E., (2021). Editorial: Broadening the Use of Machine Learning in Hydrology. Frontiers in Water 3:681023. doi: 10.3389/frwa.2021.681023

Shen, C. (2018). A transdisciplinary review of deep learning research and its relevance for water resources scientists. Water Resources Research, 54, 8558–8593. doi: 10.1029/2018WR022643

Shen, C., Laloy, E., Elshorbagy, A., Albert, A., Bales, J., Chang, F.-J., Ganguly, S., Hsu, K.-L., Kifer, D., Fang, Z., Fang, K., Li, D., Li, X., & Tsai, W.-P. (2018b). HESS Opinions: Incubating deep-learning-powered hydrologic science advances as a community. Hydrology and Earth System Sciences, 22(11), 5639–5656. doi: 10.5194/hess-22-5639-2018

Sherman, L. K. (1932). Streamflow from rainfall by unitgraph method. English News Record, 1008, 501–505.

Stisen, S., Koch, J., Sonnenborg, T. O., Refsgaard, J. C., Bircher, S., Ringgaard, R., & Jensen, K. H. (2018). Moving beyond runoff calibration – Multi-variable optimization of a surface-subsurface-atmosphere model. Hydrological Processes, 32(17), 2654–2668.

Smith, L. (2015). Hydrogeology. Elsevier. doi: 10.1016/B978-0-12-409548-9.09469-0

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. In Advances in neural information processing systems (pp. 2951-2959). https://dash.harvard.edu/handle/1/11708816

Stransky, D., Bares, V., & Fatka, P. (2007). The effect of rainfall measurement uncertainties on rainfall–runoff processes modelling. Water Science and Technology, 55(4), 103–111. doi: 10.2166/wst.2007.100

Surowiecki, James. The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations. Doubleday, 2004.

Sutskever, I., Martens, J., Dahl, G. & Hinton, G.. (2013). On the importance of initialization and momentum in deep learning. Proceedings of the 30th International Conference on Machine Learning, in Proceedings of Machine Learning Research, 28(3):1139-1147 Available from https://proceedings.mlr.press/v28/sutskever13.html

Tetzlaff, D., Seibert, J., McGuire, K.J., Laudon, H., Burns, D.A., Dunn, S.M. and Soulsby, C. (2009), How does landscape structure influence catchment transit time across different geomorphic provinces?. Hydrol. Process., 23: 945-953. 10.1002/hyp.7240

Tripathy, K. P., & Mishra, A. K. (2024). Deep learning in hydrology and water resources disciplines: concepts, methods, applications, and research directions. Journal of Hydrology, 628, 130458. doi: 10.1016/j.jhydrol.2023.130458

Tsai, W.-P., Feng, D., Pan, M., Beck, H., Lawson, K., Yang, Y., Liu, J., & Shen, C. (2021). From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling. Nature Communications, 12(1), 5988. doi: 10.1038/s41467-021-26107-z

Tsai, W. P., Fang, K., Ji, X., Lawson, K., and Shen, C. (2020). Revealing causal controls of storage-streamflow relationships with a data-centric Bayesian framework combining machine learning and process-based modeling. Front. Water 2:583000. doi: 10.3389/frwa.2020.583000

Valiela, I. (2001). Doing Science Design, Analysis and Communication of Scientific Research. Oxford University Press. doi: 10.1071/pc010219

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I., (2017). Attention Is All You Need. NeurIPS Proceedings, 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Jarrod Millman, K., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C., Polat, ̇I., Feng, Y., Moore, E. W., Vand erPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., 415 Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and Contributors, S. . .: SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python, Nature Methods, 17, 261–272, doi: 10.1038/s41592-019-0686-2, 2020

Wang, D., and Tang, Y. (2014), A one-parameter Budyko model for water balance captures emergent behavior in darwinian hydrologic models, Geophys. Res. Lett., 41, 4569– 4577, doi: 10.1002/2014GL060509

Western, A. W., Grayson, R. B., & Blöschl, G. (2002). Scaling of Soil Moisture: A Hydrologic Perspective. Annual Review of Earth and Planetary Sciences, 30(1), 149–180. 10.1146/annurev.earth.30.091201.140434

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. Biometrics Bulletin, 1(6), 80–83. doi: 10.2307/3001968

Willmott, C. J., & Matsuura, K. (2006). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research, 30(1), 79-82.

Wortsman, M., Ilharco, G., Gadre, S.Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A.S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S. & Schmidt, L.. (2022). Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. Proceedings of the 39th International Conference on Machine Learning, in Proceedings of Machine Learning Research 162:23965-23998 Available from https://proceedings.mlr.press/v162/wortsman22a.html

WMO, 2006. Technical Regulations III, Hydrology, no 49.

WMO. (1994). Guide to hydrological practices. WMO-No. 1968 (5th ed.). World Meteorological Organization.

Xie, K., Liu, P., Zhang, J., Han, D., Wang, G., & Shen, C. (2021). Physics-guided deep learning for rainfall-runoff modeling by considering extreme events and monotonic relationships. Journal of Hydrology, 603, 127043. doi: 10.1016/j.jhydrol.2021.127043

Yilmaz, K. K., Gupta, H. V., and Wagener, T. (2008). A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model. Water Resources Research, 44, W09417. doi: 10.1029/2007WR006716

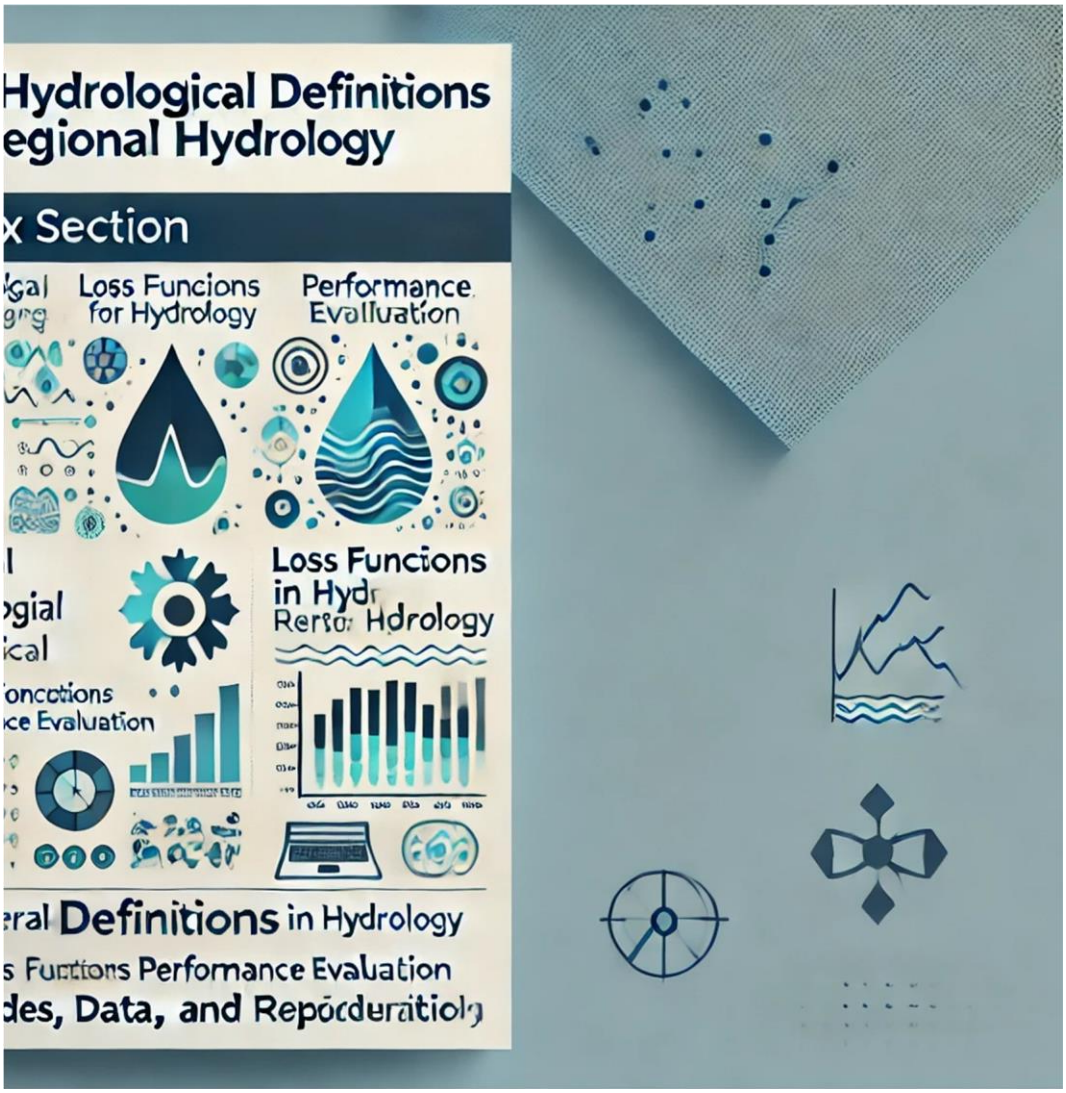
Zheng, Y., & Wang, D. (2021). Multi-Objective Recommendations: A Tutorial (arXiv:2108.06367v2 [cs.IR]). Retrieved from doi: 10.48550/arXiv.2108.06367

Zohuri, B. (2018). Chapter 5 - First Law of Thermodynamics (B. B. T.-P. of C. Zohuri (ed.); pp. 119–163). Elsevier. doi: 10.1016/B978-0-12-814519-7.00005-7

Zounemat-Kermani, M., Batelaan, O., Fadaee, M., & Hinkelmann, R. (2021). Ensemble machine learning paradigms in hydrology: A review. Journal of Hydrology, 598, 126266. doi: 10.1016/j.jhydrol.2021.126266


General
in R
Appendi
General Hydrolo
Genera
Hydrolo
Hydroi
Perforan
• Gene
• Los
• Co

Appendixes



Appendix 01:

General Hydrological Definitions in hydrological modeling

Appendix 02:

Loss Functions in Hydrology for Performance Evaluation

Appendix 03:

Codes, Data and reproducibility

Appendix 01: General Hydrological Definitions in hydrological modeling

Appendix 01.01. Hydrological Definitions

Regarding the interdisciplinary approach of this thesis, to facilitate a common language between hydrologists and AI/DL scientists, we present some definitions and key concepts in hydrology that we think are important to understand in this domain.

Natural Water Cycle: The natural water cycle describes the continuous movement of water on, above, and below the Earth's surface. Key processes include evaporation, transpiration, condensation, precipitation, infiltration, runoff, and percolation. This cycle is driven by solar energy and involves various components such as the atmosphere, oceans, lakes, rivers, and land surfaces (Robertson et al., 2022; Zohuri, 2018). Figure 42 depicts the natural water cycle and its different components.

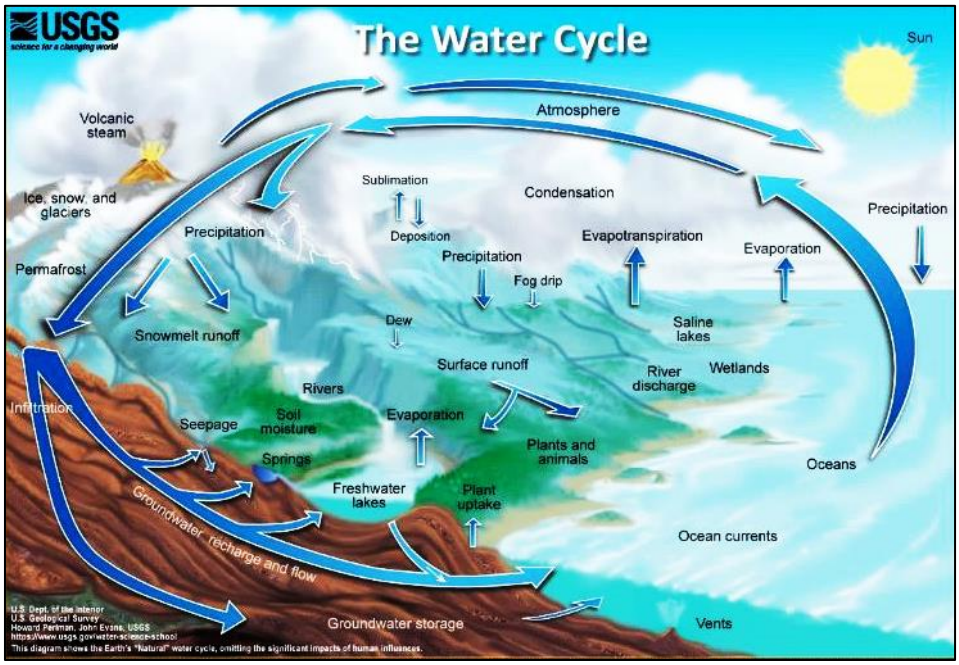


Figure 42. A schematic of the Natural Water Cycle provided by © USGS

Water Year: In hydrology, a water year is a 12-month period used to aggregate and analyze water-related data, such as precipitation, streamflow, and other hydrological variables. It typically spans from October 1 to September 30 of the following year, aligning with the hydrological water cycle and seasonal patterns in many regions. This standard period facilitates consistent comparisons and evaluations of water availability, flood risks, and drought conditions across years. The use of a water year helps in understanding long-term trends and planning for water resource management and flood forecasting in traditional hydrology.

Catchment: A catchment, also known as a drainage water basin or watershed, is an area where precipitation collects and drains into a common outlet like a river mouth or reservoir. It is defined by its drainage divide, such as ridges or hills, and includes both surface and subsurface water sources (Beven, 2012; Chow et al., 1988). Figure 43 illustrates a sample catchment in the Basque Country, located in north of Spain, showing its borders and river network.

If a catchment is not affected by human intervention (e.g., dams or water transfer projects), it is termed a natural catchment; otherwise, its streams and water flows could be significantly altered by human activities (the so-called anthropogenic fingerprints). For example, the construction of a dam in a catchment area creates an artificial reservoir that can regulate water flow, reduce downstream flood peaks, and alter seasonal flow patterns. Similarly, water transfer projects that divert water from one catchment to another can impact the natural hydrological cycle, affecting both local and downstream water availability and ecological health. These human interventions underscore the need for careful management and consideration of both natural and anthropogenic factors when assessing water resources and planning for sustainable water management.

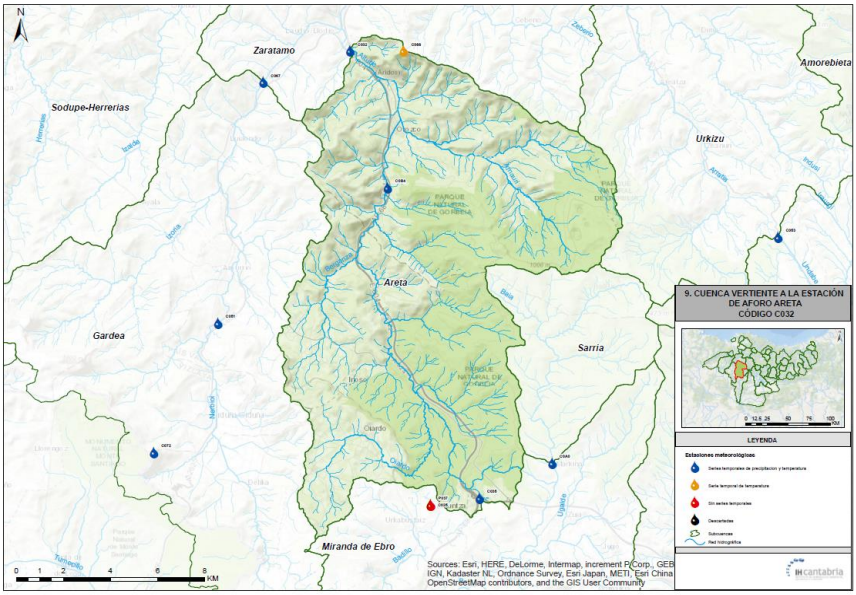


Figure 43. Areta catchment in Basque Country, Spain. The catchment at its outlet, flows to Zaratamo catchment as seen

Catchment Attributes: These are static characteristics such as area, slope, topography, drainage density, climate (precipitation, evapotranspiration, temperature), and geology (soil types, permeability, infiltration rates). These attributes are typically measured in-situ or via remote sensing (Addor et al., 2017).

Catchment Topography: Each catchment has distinct topographical features such as slope, shape, and river network. These characteristics influence how water moves within the catchment and ultimately affects runoff patterns. The interplay between topography and hydrological behavior is critical for understanding streamflow and flood dynamics.

Catchment Geology and Soil Characteristics: Soil properties are essential in hydrology due to their close inter-relations with runoff generation. Soils with lower infiltration rates,

such as clay, result in higher surface runoff compared to soils with higher infiltration rates, like sand (Table 12). When calibrating rainfall-runoff models, catchments with similar soil characteristics often have comparable model parameters. Key soil attributes include infiltration rates and saturation capacity, which are vital for understanding water absorption and runoff (Ferré & Warrick, 2005; Beven, 2012; Chow et al., 1988).

Table 12. Basic infiltration rates for various soil types. (https://www.fao.org)

Soil type	Basic infiltration rate (mm/hour)
sand	less than 30
sandy loam	20 - 30
loam	10 - 20
clay loam	5 - 10
clay	1 - 5

Soil type, such as clay, sandy, silty, peaty, chalky, and loamy, determines various properties including infiltration rate. For instance, sand has a high infiltration rate, whereas clay has a significantly lower rate.

Land Cover refers to the physical material on the landscape surface, encompassing vegetation, bare ground, water bodies, and urban areas. It represents semi-static features of a catchment that can change seasonally or due to human activities and climate change. Different types of land cover affect hydrological processes differently, influencing infiltration rates and runoff generation (Beven, 2012). For instance, dense forest canopies reduce flood runoff by promoting infiltration and slowing down surface flow compared to bare ground. Conversely, increased impervious surfaces, such as urban infrastructure, typically lead to higher runoff and elevated flood risks. Understanding the interplay between land cover and soil permeability is crucial for accurate streamflow and flood predictions, as these factors significantly impact hydrological behavior of catchments.

Runoff is the flow of water over the land surface, occurring when rainfall exceeds the soil's infiltration capacity. It is an essential component of the hydrological cycle and is measured in millimeters to quantify the depth of water running off a catchment area (Beven, 2012; Chow et al., 1988).

Streamflow refers to the flow of water in rivers, streams and water networks in a catchment. It includes contributions from surface runoff, subsurface runoff, and groundwater. Streamflow is also measured in millimeters and indicates the volume of water flowing past a point in a river or stream (Beven, 2012).

Water Level measures the height of water in bodies such as lakes, rivers, or reservoirs. Monitoring water levels helps assess water availability and dynamics, influencing water resource management and ecosystem health. Water level is measured in flow gauges by using

various methods such as pressure sensors, float-operated devices, or radar and ultrasonic sensors. These instruments provide continuous data on water elevation, which is crucial for predicting floods, managing water supplies, and understanding hydrological patterns.

Rating Curves are empirical relationships that link water stage (level) to discharge (streamflow). Developed through field measurements, they enable the estimation of streamflow from water levels, crucial for flood forecasting and water resources management. Rating curves inherently include information about catchment attributes such as channel geometry, slope, and roughness. They reflect the physical characteristics of the river or stream channel and its interactions with flow conditions. By accounting for variations in these attributes, rating curves help in accurately translating water stage readings into streamflow estimates, providing valuable insights into hydrological processes and aiding in effective water management.

Evaporation is the process where solar energy heats water, turning it into vapor that rises into the atmosphere (Robertson et al., 2022).

Transpiration is the release of water vapor from plant surfaces during photosynthesis, facilitated by stomata in leaves. It is a key component of the water cycle and is influenced by latent heat flux (Hanrahan, 2012; Ledley, 2003).

Evapotranspiration is the combined process of evaporation and transpiration. It includes Actual Evapotranspiration (AET), the actual water vapor released, and Potential Evapotranspiration (PET), the maximum potential vapor flux under ideal conditions (Hasiotis et al., 2007). PET can be estimated using formulas like Hargreaves and Allen's (2003) equation (2):

$$PET = 0.0023 \,.\, Ra. \sqrt{(T_{max} - T_{min}).(T_{mean} + 17.8)}$$

Equation 2

Here, Ra equals the extraterrestrial radiation and T_{min}, T_{max}, and T_{mean} are the minimum, maximum, and average temperatures of a certain period, respectively.

Condensation is the process where water vapor in the atmosphere cools down and transforms into liquid droplets, forming clouds (Chow et al., 1988).

Precipitation is any form of water falling from the atmosphere to the Earth's surface (e.g., rain, snow, sleet, and hail). It is measured using rain gauges to quantify the amount of water received over a specific area and period (Robertson et al., 2022; Stransky et al., 2007).

Infiltration is the process where water penetrates the soil. The rate at which this occurs is known as the infiltration rate, which determines how much water the soil can absorb before excess runoff begins (Ferré & Warrick, 2005; Chow et al., 1988).

Percolation refers to the downward movement of water through soil and rock layers, reaching groundwater (Miller, 1977).

Subsurface Water exists below the ground surface and includes groundwater, aquifers, and underground streams. The upper surface of the saturated zone is known as the water table (Smith, 2015; Bales, 2015).

Water Storage refers to spaces that retain water, such as lakes, reservoirs, glaciers, and aquifers. Both natural and human-made reservoirs play a significant role in managing water resources and influencing downstream water patterns (Beven, 2012).

Hydro-Geo-Meteorological Data: Rainfall-runoff models rely on a diverse range of hydro-geo-meteorological data (e.g., temperature, precipitation, evapotranspiration, solar radiation, wind speed, slope, and land cover). Hydrological prediction models are either calibrated in traditional hydrological approaches or optimized and trained in hydrological DL models. Once developed, the models are verified using hydrological observation data records such as streamflow measurements, water levels, soil moisture, and groundwater tables. Accurate integration and representation of these data types are essential for model performance and reliable predictions.

Data is collected using various methods, including automatic, semi-automatic, and manual instruments at different stations, as well as advanced technologies like radar and satellite remote sensing (Anderson, 2005). Despite these advancements, historical datasets often suffer from gaps and uncertainties, highlighting the need for ongoing improvements in data quality and model development. Meteorological data, encompassing physical parameters such as precipitation, temperature, dew point, wind speed, and radiation, are directly measured by instrumentation (Coleman & Law, 2015). Hydrological data, defined by the World Meteorological Organization (WMO, 2006), describe various aspects of the water cycle and can be measured in-situ, via satellite, or estimated through equations, such as streamflow from water level rating curves. Key hydrological indices include flow metrics like hourly, daily, and annual flows, high and low flow frequencies, and timing of flow events (Beven, 2012).

Appendix 01.02. Fundamental hydrological perspectives that draw perceptual models

To facilitate a common language between hydrologists and AI/DL scientists, we present some well-known hydrological fundamental that we think are important to understand in this domain. These terms and definitions and concepts provide a foundational understanding of hydrology, essential for interpreting rainfall-runoff modeling and relevant hydrological studies. Truly understanding these definitions later could aid us interpret the predictive outcomes of our new generation of AI/DL models.

Newtonian approaches in hydrological modeling: The Newtonian approach to hydrologic science is grounded in the development of “physically-based” models derived from Newtonian first principles, particularly the conservation equations. This approach emphasizes the use of experimental, field, and modeling-based research to capture the key hydrological processes at the catchment scale. Although these models often assume that processes can be

effectively upscaled through appropriate parameter values, the focus remains on identifying these values through observations, field experiments, and optimization techniques. However, it is important to note that this method bears little resemblance to Newton's original scientific practices.

Water Balance Law: This principle states that the total inflows (e.g., precipitation, snowmelt) must equal the total outflows (e.g., evaporation, runoff) plus any changes in storage. This balance is crucial for understanding water flow and storage within a hydrological system (Rosbjerg & Rodda, 2019; Ivezic et al., 2016) as the foundation of traditional hydrological models.

Energy Balance Law: The First Law of Thermodynamics, or the law of energy conservation, applies to hydrology through the energy balance method. In this approach, only sensible heat flux is considered, with evapotranspiration as the residual term in the energy balance equation (Ershadi et al., 2011).

Newton's Gravity Law: Water naturally flows from upstream to downstream, a fundamental observation that should be considered in DL models to ensure if they respect this basic physical principle. Verification of models against this principle could be beneficial.

Darcy's Law: This law describes the capacity of a porous medium to transmit water, relevant for understanding soil-water relationships and groundwater-surface water exchanges (Robertson et al., 2022; Hillel, 2008). The Richards' equation is used for water movement in unsaturated soils (Hopmans, 2011; Ferré & Warrick, 2005).

Darwinian Hydrology: This approach focuses on complex interdependencies and patterns in hydrological systems, drawing parallels with ecological principles. It aims to understand watershed behavior through observable structures and historical patterns rather than purely mechanical explanations (Harte, 2002; Harman & Troch, 2014). Key strategies include developing simple, falsifiable models and identifying patterns and principles in hydrology.

Logistic Equilibrium Hypothesis: The logistic growth model, initially used in population studies, has been adapted to hydrology. It describes how runoff behaves similarly to population growth, reaching a saturation point and then declining (Malthus & Stimson, 2018; Wang & Tang, 2014).

The Old Water Paradox: This paradox highlights the puzzling observation that catchments can retain "old water"—water that has been stored in the subsurface for extended periods—yet release it rapidly during storm events. This challenges traditional hydrological models, which often assume a quicker turnover of water within a catchment. The paradox suggests that our definitions and models of baseflow and stormflow may need revisiting, especially as advanced tools like DL models could offer new insights into the underlying processes (Kirchner, 2003). Research into this paradox often involves the use of isotopic tracers, which can differentiate between old and new water sources, providing a more detailed understanding of catchment hydrodynamics and the pathways through which water travels before being discharged.

Unit Hydrograph Theory (UH): This theory treats catchments as linear systems, routing runoff based on a unit-response function. It generates hydrographs representing the cumulative effect of rainfall over time (Sherman, 1932; Littlewood, 2002).

Curve Number Method (SCS-CN): This method estimates the rainfall-runoff coefficient based on precipitation and antecedent soil moisture. It is based on an empirical proportional hypothesis where the rate of actual evaporation is proportional to runoff (SCS, 1972; Wang & Tang, 2014; Hawkins et al., 1985).

Appendix 02: Loss Functions in Hydrology for Performance Evaluation

Appendix 02.01. Nash-Sutcliffe Efficiency (NSE)

The Nash-Sutcliffe Efficiency (NSE) is a widely used metric in hydrological modeling to assess the agreement between observed and simulated data, providing valuable insights into the dynamics of a hydrological system (Nash & Sutcliffe, 1970). It serves as a robust measure of model performance, allowing hydrologists to evaluate how well a model captures the patterns and dynamics of streamflow. The simplicity and ease of interpretation have made NSE a popular error function, particularly for analyzing high-flow conditions (Arsenault et al., 2018; Gupta et al., 2009).

The NSE metric, represented by Equation 3, quantifies the relative magnitude of the residual variance with respect to the observed variance, ranging from negative infinity to 1. A value of 1 indicates a perfect match between observed and simulated data. The equation involves calculating the squared differences between observed and simulated values, which are then normalized by the observed variance. This normalization accounts for the variability in the observed data. Consequently, NSE provides a valuable measure of the model's performance in replicating observed streamflow across various hydrological analyses and conditions.

$$NSE = 1 - \frac{\sum_{i=1}^{n}\left(Q_{obs,i} - Q_{sim,i}\right)^2}{\sum_{i=1}^{n}\left(Q_{obs,i} - \bar{Q}_{obs}\right)^2}$$

Equation 3

In the equation, $\mathrm{Q_{obs,i}}$, $\mathrm{Q_{sim,i}}$ represent the observed and simulated streamflow or water level, respectively, at a specific simulation hour denoted by the index i. $\bar{Q}_{obs}$ denotes the mean of the observed streamflow values.

By comparing the squared differences between observed and simulated values to the observed variance, NSE provides a quantitative assessment of how well the model reproduces the observed streamflow dynamics. Higher values of NSE indicate better agreement between observed and simulated data, implying a higher level of model performance in capturing the hydrological system's behavior.

Appendix 02.02. Kling-Gupta Efficiency (KGE)

The Kling-Gupta Efficiency (KGE) metric is widely used for evaluating the performance of hydrological models. It considers three essential components: correlation coefficient (r), bias term (β), and variability ratio (α). KGE values range from negative infinity to 1, where a value of 1 indicates a perfect match between observed and simulated data, and higher values generally signify better model performance.

Gupta et al. (2009) developed the KGE metric to address limitations in the Nash-Sutcliffe Efficiency (NSE), particularly in accurately representing both high and low flows. NSE, while commonly used, tends to underestimate runoff peaks due to its sensitivity to large runoff values and the overall underestimation of flow variability. The three components contributing to NSE—linear correlation, bias, and flow variability—can have varying impacts across different catchments and years. In regions with high flow variability, the bias has a reduced influence on NSE, leading to an underestimation of peak flows. This is because, in such cases, the slope of the regression between simulated and observed values is often less than one, which systematically underestimates peaks.

To mitigate these issues, the KGE criterion was introduced. It assigns equal importance to the correlation (r), bias (β), and variability (α) components, leading to a more balanced and comprehensive assessment of model performance. Optimizing for KGE improves bias and variability measures, although it may slightly reduce correlation. The formula for KGE is provided below:

Equation 4 represents the KGE metric, which calculates the KGE value based on the correlation coefficient (r), flow variability error (α), and bias term (β). The terms α and β are calculated as the ratios of the standard deviations (σ_{sim} and σ_{obs}) and means (μ_{sim} and μ_{obs}) of the simulated and observed data, respectively.

$$KGE = 1 - \sqrt{((r-1)^2 + (\alpha-1)^2 + (\beta-1)^2)}, \alpha = \left(\frac{\sigma_{sim}}{\sigma_{obs}}\right), \beta = \left(\frac{\mu_{sim}}{\mu_{obs}}\right)$$

Equation 4

In the equation, r represents the linear correlation between observations and simulations, α is a measure of the flow variability error, β represents the bias term, σ_{sim} corresponds to the standard deviation in simulations, σ_{obs} represents the standard deviation in observations, and μ_{sim} and μ_{obs} represent the simulation mean and observation mean, respectively.

The KGE metric is frequently used as an objective function in hydrological modeling to verify the effectiveness of calibration techniques (Gupta et al., 2009; Knoben et al., 2019). In an inter-comparison study by Mai et al. (2022), KGE and its three components were employed for streamflow calibration and validation. The study used KGE's Euclidean distance from its ideal point in the untransformed criteria space, ensuring optimal performance is represented by a maximum KGE value of 1, consistent with NSE. Table 13, adapted from Mai et al. (2022), categorizes the ranges for KGE components to qualify performance as excellent, good, medium, or poor.

Table 13. The qualification metrics of KGE components; Mai et al., 2022

	poor performance	medium performance	good performance	excellent performance
KGE_α	$(-\infty,0.70)$	[0.70,0.80)	[0.80,0.90)	[0.90,1.0]
KGE_β	$(-\infty,0.70)$	[0.70,0.80)	[0.80,0.90)	[0.90,1.0]
KGE_r	$(-\infty,0.70)$	[0.70,0.80)	[0.80,0.90)	[0.90,1.0]
KGE	$(-\infty,0.48)$	[0.48,0.65)	[0.65,0.83)	[0.83,1.0]

Appendix 02.03. Mean Squared Error (MSE)

Mean Squared Error (MSE) is a metric used to measure the average squared difference between observed and simulated values, providing an overall assessment of the model's predictive accuracy (Makridakis et al., 1993). By averaging the squared differences, MSE emphasizes the significance of large errors and penalizes them more heavily than smaller errors. Lower MSE values indicate better model performance, with zero representing a perfect match between observed and simulated data.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Q_{obs,i} - Q_{sim,i}\right)^2$$

Equation 5

MSE is particularly applicable in hydrological studies focusing on streamflow prediction. It captures the overall magnitude of errors in the model predictions, enabling an assessment of the average discrepancy between observed and simulated streamflow values (Legates and McCabe, 1999).

Appendix 02.04. Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) is derived from MSE by taking the square root of the averaged squared differences, resulting in a metric that is in the same unit as the original data (Willmott and Matsuura, 2006). RMSE provides a measure of the average magnitude of the differences between observed and simulated values and is widely employed for model evaluation and comparison.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Q_{obs,i} - Q_{sim,i}\right)^2}$$

Equation 6

RMSE is particularly useful in assessing the standard deviation of the errors between observed and simulated streamflow, representing the typical magnitude of the residuals. Lower RMSE values indicate better model performance, with zero indicating a perfect match between observed and simulated data.

The utilization of MSE and RMSE in hydrology has gained significant recognition and widespread adoption. These metrics are valued for their capacity to offer objective and quantitative measures of model performance in streamflow prediction studies (Krause et al., 2005). MSE and RMSE facilitate the evaluation and comparison of various modeling approaches, assisting researchers in comprehending the accuracy and reliability of their predictions. In addition to evaluating models' performance, RMSE was employed as an objective function, besides NSE, for training some of the DL models in this research. This

approach allowed for the optimization of the models' performance based on the RMSE criterion, further enhancing their predictive capabilities.

Appendix 02.05. Alpha-Nash-Sutcliffe Efficiency (Alpha-NSE)

The alpha-NSE component represents the linear correlation between simulated and observed values, providing an assessment of the model's ability to capture the overall trend and variability of the observed data. A high alpha-NSE value indicates a strong linear relationship between the model outputs and observed values, indicating good agreement in terms of the overall pattern (Gupta et al., 2009). Alpha-NSE extends the NSE to focus on high flows and extreme events, serving as a measure of the model's performance in capturing peak flows. Positive Alpha-NSE values indicate satisfactory simulation of high flows, while negative values indicate poor performance in capturing extreme events.

According to Gupta et al. (2009), the alpha NSE decomposition quantifies the relative variability between simulated and observed time series. This analysis evaluates the variability of the model's simulations compared to the observed data. The decomposition is calculated using the formula $\alpha = \frac{\sigma_{sim}}{\sigma_{obs}}$, where α represents the alpha NSE decomposition, σ_{sim} denotes the standard deviation of the simulated time series, and σ_{obs} represents the standard deviation of the observed timeseries.

Appendix 02.06. Beta-Nash-Sutcliffe Efficiency (Beta-NSE)

The beta-NSE is another component of the NSE criterion that complements the alpha-NSE by capturing the systematic deviation or bias between simulated and observed values (Gupta et al., 2009). A positive beta-NSE indicates that the model consistently overestimates the observed values, while a negative beta-NSE suggests consistent underestimation. Ideally, a model with a beta-NSE value close to zero indicates unbiased predictions.

Understanding the beta-NSE component is crucial for identifying and correcting systematic errors in model predictions. By analyzing the bias, insights into potential model deficiencies can be gained, and the model's performance can be improved by adjusting relevant parameters or model structures. The beta NSE decomposition is calculated using the formula:

$$\beta = \frac{(\mu_{sim} - \mu_{obs})}{\sigma_{obs}}$$

Equation 7

Here, β represents the beta NSE decomposition, μ_{sim} and μ_{obs} denote the means of the simulated and the observed time series, respectively, and σ_{obs} represents the standard deviation of the observed time series.

Appendix 02.07. Beta-Kling-Gupta Efficiency (Beta-KGE)

Beta-KGE is a complementary metric that assesses model performance by considering both bias and variability (Kling et al., 2012). Similar to beta-NSE, beta-KGE accounts for the bias between simulated and observed values (Gupta et al., 2009). It provides a measure of the systematic deviation in model predictions. By incorporating beta-KGE in the evaluation process, we can effectively assess both the accuracy and precision of the model outputs. It enables the evaluation of bias-related issues while improving the representation of flow variability in the model.

Beta-KGE is an extension of the Kling-Gupta Efficiency (KGE) that specifically focuses on capturing low flow conditions (Kling et al., 2012). Positive values of beta-KGE indicate good performance in simulating low flows, while negative values suggest poor performance in capturing low flow conditions. The beta term of the KGE represents the fraction of the means between the simulated (sim) and observed (obs) time series (Gupta et al., 2009). This metric quantifies the difference in average values between the simulated and observed data.

The beta KGE term is calculated using the Equation 7. Here, β denotes the beta KGE term, μ_{sim} represents the mean of the simulated time series, and μ_{obs} represents the mean of the observed time series. The beta KGE term provides insights into the model's ability to capture the average behavior of the observed streamflow. Furthermore, beta-KGE focuses on the temporal aspects of model performance. It evaluates how well the model captures the timing and variability of the observed streamflow, particularly in relation to low flow conditions (Kling et al., 2012).

$$\beta = \frac{\mu_{sim}}{\mu_{obs}}$$

Equation 8

Appendix 02.08. Pearson's Correlation Coefficient (Pearson-r)

Pearson's Correlation Coefficient (Pearson-r) is a widely used statistical metric that quantifies the linear relationship between observed and simulated values. It is calculated by dividing the covariance of the two datasets by the product of their standard deviations. Pearson-r measures the strength and direction of the linear association between the observed and simulated data points.

In hydrological studies, Pearson-r is commonly employed to assess the linear correlation between observed and simulated streamflow values. It helps researchers assess the accuracy of their models in capturing the trends and variations observed in the real-world hydrological processes. The coefficient ranges from -1 to 1, where a value of 1 indicates a perfect positive linear relationship. This means that as the observed streamflow increases, the simulated streamflow also increases proportionally. On the other hand, a value of -1 represents a perfect negative linear relationship, implying that as the observed streamflow increases, the

simulated streamflow decreases proportionally. A Pearson-r value close to 0 suggests a weak or no linear correlation between the datasets, indicating that the observed and simulated streamflow values are not linearly related.

By evaluating Pearson-r, this study aimed to determine how well the simulated streamflow captured the variations and trends observed in the real-world data. A high Pearson-r value indicates a strong linear correlation, suggesting that the model accurately reproduces the observed hydrological behavior. Conversely, a low Pearson-r value suggests a weak linear relationship, indicating potential discrepancies between the simulated and observed streamflow patterns.

Appendix 02.09. High-segment volume (%BiasFHV)

High-segment volume (%BiasFHV) is a metric that assesses the model's ability to capture the volume of streamflow during high-flow periods. It quantifies the percentage bias between the observed and simulated high-flow volumes. Positive values indicate an overestimation of high-flow volumes, while negative values indicate an underestimation. %BiasFHV represents the difference between the simulated and observed values at a specified fraction of upper flows in the flow duration curve. By evaluating peak flow estimation, %BiasFHV provides insights into the model's accuracy in estimating high flows and enhances the evaluation methodology.

$$\%BiasFHV = \frac{\sum_{h=1}^{H}(Q_{sim,h} - Q_{obs,h})}{\sum_{h=1}^{H} Q_{obs,h}} * 100$$

Equation 9

Where $Q_{sim,h}$ and $Q_{obs,h}$ are the simulations, the observations and H is the upper fraction of flows of the FDC (Fraction of upper flows to consider as peak flows of range]0,1[, in this research: 0.02).

Appendix 02.10. Low-segment volume (%Bias FLV)

Low-segment volume (%BiasFLV) is a metric similar to %BiasFHV but focuses on the model's performance in capturing the volume of streamflow during low-flow periods. It quantifies the percentage bias between the observed and simulated low-flow volumes. Positive values indicate an overestimation of low-flow volumes, while negative values indicate an underestimation. %BiasFLV evaluates the difference between the simulated and observed values at a specified fraction of lower flows in the flow duration curve. By assessing low-flow estimation, %BiasFLV enhances the evaluation methodology and contributes to a comprehensive understanding of the model's performance in capturing different flow characteristics.

$$\%BiasFLV = -1 * \frac{\sum_{l=1}^{L}[log(Q_{sim,l}) - log(Q_{sim,L})] - \sum_{l=1}^{L}[log(Q_{obs,l}) - log(Q_{obs,L})]}{\sum_{l=1}^{L}[log(Q_{obs,l}) - log(Q_{obs,L})]} * 100$$

Equation 10

Where are the simulations, the observations and L is the lower fraction of flows of the FDC (Fraction of lower flows to consider as low flows of range]0,1[, in this research: 0.3).

Appendix 02.11. Mid-segment slope (%Bias FMS)

Mid-segment slope (%BiasFMS) measures the model's performance in capturing the slope of the rising and falling limbs of the hydrograph, excluding the extreme high and low flows. It quantifies the percentage bias in the slope between the observed and simulated hydrographs. Positive values indicate an overestimation of the slope, while negative values indicate an underestimation. %BiasFMS evaluates the difference between the logarithmic values of simulated and observed flows at the lowest and highest exceedance probabilities within the midsegment of the flow duration curve. By examining the transition between high and low flows, %BiasFMS contributes to a comprehensive understanding of the model's ability to capture different flow regimes.

$$\%BiasFMS = \frac{\left|\log(Q_{sim,lower}) - \log(Q_{sim,upper})\right| - \left|\log(Q_{obs,lower}) - \log(Q_{obs,upper})\right|}{\left|\log(Q_{obs,lower}) - \log(Q_{obs,upper})\right|} * 100$$

Equation 11

Where $Q_{sim,lower/upper}$ corresponds to the FDC of the simulations at the lower and upper bound of the middle section and similarly $Q_{obs,lower/upper}$ for the observations. (Lower is the lower bound of the middle section in range]0,1[, in this research: 0.2; Upper is the upper bound of the middle section in range]0,1[, in this research: 0.7)

Appendix 02.12. Mean difference in Peak Flow Timing (Peak-Timing)

Peak-Timing is a metric used to assess the model's ability to accurately predict the timing of peak flow events. It quantifies the time lag between the observed and simulated peak flows, providing insights into the model's performance in capturing the temporal aspect of hydrological processes. A small-time lag indicates good agreement in capturing the timing of peak flows, while a large time lag suggests a deviation from the observed timing.

To evaluate the consistency in timing between observed and simulated peak flows, the mean difference in peak flow timing is calculated. This analysis utilize the SciPy's find_peak function (Virtanen et al., 2020), which identifies peaks in the observed time series. Those observed peaks with a prominence value less than the standard deviation of the observed time series are discarded (Kratzert et al., 2020). This step helps filter out smaller peaks that may be affected by noise or variability. Subsequently, an iterative process is implemented to ensure well-defined and separated peaks for analysis. The lowest peaks are successively removed until the remaining peaks have a minimum distance of 100-time steps between them, following the methodology outlined by Kratzert et al. (2020). This ensures that only prominent and distinct peaks are included in the analysis.

Once the observed peaks are determined, the corresponding peaks in the simulated time series are identified within a specified window size (window) centered around each observed peak. The window size depends on the temporal resolution of the time series, such as '1D' for daily or '1H' for hourly data, with default values of 3 and 12, respectively. By comparing the observed and simulated peaks within this window, the absolute time differences between them are calculated. Finally, the mean of these differences across all peaks is computed, providing the mean peak time difference as the resulting metric. This study took advantage of NeuralHydrology library to calculate Peak-Timing metric.

The mean peak time difference metric serves as an indicator of the model's ability to accurately reproduce the timing of peak flow events. A smaller mean difference signifies a closer alignment between observed and simulated peaks, indicating better performance in capturing the temporal dynamics of the hydrological system (Kratzert et al., 2020).

Appendix 02.13. Mean Absolute Percentage Error for peaks (MAPE_peak)

MAPE_peak is a metric used to evaluate the accuracy of peak flow predictions by comparing the observed and simulated peak flow values. It provides a quantitative measure of the relative deviation between the observed and simulated peaks, expressed as a percentage.

To calculate MAPE_peak, the scipy.find_peaks function is utilized to identify peaks in the observed time series. This function identifies local maxima in the time series, considering them as potential peaks. By using this approach, prominent peaks in the observed flow data

are identified, while smaller fluctuations and noise are filtered out. The resulting indices of the observed peaks are then used to subset both the observed and simulated flow data (Kratzert et al., 2020).

Next, the observed peak flows and the corresponding simulated peak flows are extracted from the respective time series. These flows represent the magnitudes of the peak events in the hydrological system. The MAPE metric is then calculated as the mean absolute percentage error between the observed peak flows and the corresponding simulated peak flows. The formula for MAPE is as follows:

$$MAPE\_peak = \frac{1}{P}\sum_{p=1}^{P} \left| \frac{(Q_{sim,p} - Q_{obs,p})}{Q_{obs,p}} \right| * 100$$

Equation 12

Where $Q_{sim,p}$ and $Q_{obs,p}$ are the simulated and the observed peaks, respectively, and P is the number of peaks.

The MAPE_peak metric provides insights into the model's performance in accurately predicting the magnitudes of peak flow events. A lower MAPE value indicates a smaller relative deviation between the observed and simulated peak flows, suggesting better agreement. Conversely, a higher MAPE value suggests larger discrepancies between the observed and simulated peak flows.

By utilizing the MAPE_peak metric, the model's ability to reproduce the magnitudes of peak flow events can be quantitatively evaluated. This evaluation provides valuable information for understanding the model's accuracy in capturing extreme hydrological events, which are crucial for various water resources management applications, especially flood forecasting in flashy catchments.

Appendix 02.14. Fraction of Missed Peaks (missed_peaks)

The missed_peaks metric is used to quantify the fraction of peaks in the observed time series that are not captured in the simulated time series within a specified window. It provides an assessment of the model's ability to accurately reproduce peak flow events.

To calculate the missed_peaks metric, the scipy.find_peaks function is utilized to identify peaks in both the observed and simulated time series. Peaks are identified as local maxima in the time series, considering them as potential peaks. The metric focuses on peaks above a certain flow percentile, defined by the percentile parameter (ranging from 0 to 100).

Next, a window of a specified size, determined by the window parameter, is considered on each side of the observed peak. This window is centered around the observed peak and is used to search for the corresponding simulated peak. The total window length to find the peak in the simulations is centered at the observed peak. The default window size depends on the temporal resolution of the time series, such as '1D' for daily and '1H' for hourly data, with different default values used compared to the peak-timing metric for '1D'.

The number of peaks in the observed time series that do not have a corresponding peak in the simulated time series within the specified window is counted. This count represents the missed peaks. Finally, the missed_peaks metric is calculated as the fraction of missed peaks relative to the total number of observed peaks.

The missed_peaks metric provides insights into the model's performance in capturing the occurrence of peak flow events. A lower fraction of missed peaks indicates a better agreement between the observed and simulated peak flows, suggesting improved performance. Conversely, a higher fraction of missed peaks suggests that the model fails to capture a significant number of peak flow events. By utilizing the missed_peaks metric, the model's ability to reproduce peak flow events can be assessed, allowing for a more comprehensive evaluation of its performance in capturing extreme hydrological events.

It is worth noting that the missed_peaks metric considers peaks above a specified flow percentile, which allows for a focus on the more significant peak events in the time series. Additionally, the use of a window around the observed peak provides flexibility in assessing the agreement between the observed and simulated peaks, considering their temporal proximity.

Appendix 03: Codes, Data and reproducibility

The codes and dataset utilized in this study, along with comprehensive instructions for replicating the experiments, are accessible on our repositories on:

1. https://doi.org/10.5281/zenodo.13220528 (URA hourly Dataset)
2. https://zenodo.org/records/13220701
3. https://doi.org/10.5281/zenodo.13236262
4. https://github.com/farzadhoseini/Precise_Tuning_of_Regional_Hydrological_LSTM_Networks
5. https://github.com/farzadhoseini/ensemble.deep.learning
6. https://github.com/farzadhoseini/Ph.D.Thesis.Codes

We prioritize transparency and reproducibility so that fellow researchers and practitioners can verify our findings and employ the same codes for hyperparameter optimization and ensemble learning of their research and applications.