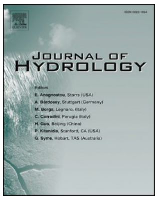
Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol



Research papers

An explainable AI approach for interpreting regionally optimized deep neural networks in hydrological prediction

F. Hosseini [*], C. Prieto, C. Álvarez

Instituto de Hidráulica Ambiental de la Universidad de Cantabria, Santander, Spain

ARTICLE INFO

This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of Fijohn Chang, Associate Editor

Keywords:
Regional rainfall-runoff modeling
Deep neural networks
Hyperparameter optimization
explainable AI (xAI)
Catchment attributes

ABSTRACT

The interpretation of artificial intelligence (AI) and deep learning (DL) model outcomes remains a central challenge in hydrology and rainfall-runoff modeling. This study investigates whether hyperparameter-optimized regional Long Short-Term Memory (LSTM) networks can implicitly learn hydrological processes directly from hydrometeorological data, without access to explicit catchment attributes during training. Specifically, we explore to what extent these models reinforce classical hydrological understanding or reveal new insights through explainable AI (xAI) analyses.

Using hourly precipitation, temperature, and potential evapotranspiration data from 40 humid and flashy catchments in the Basque Country, Spain, we demonstrate that systematically optimized LSTMs exhibit strong generalization and scalability in regional rainfall-runoff modeling. Through a combination of correlation analysis, Random Forest (RF) modeling, Principal Component Analysis (PCA), and SHAP-based feature attribution, we quantify how catchment attributes indirectly influence LSTM performance. This multi-method approach provides a novel framework to assess the hydrological "learning maturity" of deep neural networks in regional hydrology.

The results show that LSTM networks implicitly capture latent catchment characteristics that shape hydrological responses. Catchments with high runoff coefficients and higher mean annual streamflow tend to yield more accurate predictions, while catchments characterized by steep slopes, extreme flow variability, and high precipitation variability pose greater challenges due to their nonlinear hydrological behavior. SHAP analysis confirms that both catchment properties (e.g., average yearly runoff coefficient, precipitation, streamflow) and key LSTM hyperparameters (e.g., input sequence length, hidden size, dropout rate) play critical roles in predictive success, with the latter influencing the models' ability to better generalize and capture extremes.

Furthermore, RF and PCA highlight essential factors influencing model accuracy, including annual precipitation, aridity index, stream density, and land cover, where broadleaf forests improve water retention and urbanization complicates runoff processes. These findings bridge the gap between the "black-box" nature of AI/DL models and hydrological interpretability, offering evidence-based guidelines for practitioners and researchers deploying LSTMs in regional hydrological contexts.

Ultimately, this research underscores the potential of optimized LSTM networks to generalize hydrological processes and reinforce domain knowledge, while also advocating for the integration of catchment attributes in future model designs to enhance predictive robustness. By advancing xAI methodologies for deep learning in hydrology, this study contributes to developing more reliable and interpretable AI-driven solutions for water resource management and flood risk assessment under increasing climate variability.

1. Introduction

The rise of Artificial Intelligence (AI) (Russell & Norvig, 2020) and deep learning (DL) (Goodfellow et al., 2016) has revolutionized numerous fields, including hydrology. Traditional hydrological models, such as conceptual and physically-based approaches, rely on explicitly defining relationships between meteorological variables and hydrological responses based on physical processes (Refsgaard et al., 2022; Beven, 2012). These models incorporate catchment attributes, including climate, topography, geology, land use, and vegetation, to predict

[*] Corresponding author.
E-mail address: farzad.hosseini@alumnos.unican.es (F. Hosseini).

https://doi.org/10.1016/j.jhydrol.2025.133689
Received 26 March 2025; Received in revised form 23 May 2025; Accepted 12 June 2025
Available online 13 June 2025

hydrological processes like runoff and streamflow. However, the increasing availability of Big data (e.g., large hydrometeorological datasets) and advancements in our computational power by Graphics Processing Units (GPUs) have positioned AI/DL models, particularly Deep Neural networks (DNNs) (e.g., Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997)), as powerful alternatives for modeling the complex, nonlinear relationships in rainfall-runoff patterns (Tripathy & Mishra, 2024; Arsenault et al., 2023; Kratzert et al., 2024; Shen & Lawson, 2021).

LSTMs excel in sequential data processing and temporal dependency modeling, often outperforming hydrological traditional models in predictive accuracy (Donnelly et al., 2024). Unlike conventional models that explicitly use catchment attributes, LSTMs implicitly learn latent hydrological features from input-output data (Kratzert et al., 2018). However, the black-box nature of these DNNs has raised concerns about interpretability and trustworthiness. A fundamental question remains: *Do DNNs (*e.g.*, LSTMs) merely mimic observed statistical patterns, or do they capture and learn the underlying physical processes governing catchment behavior*?

Hydrological phenomena are inherently influenced by spatial and temporal variability, requiring models to account for diverse catchment-specific characteristics (Beven, 2020). For DNNs to gain broader acceptance in regional hydrology, they must address two critical aspects: predictability—generating accurate outputs—and understanding—uncovering the physical relationships driving these predictions (Başağaoğlu et al., 2022). While conceptual models explicitly define mechanisms of hydrological behavior, DNNs infer relationships solely from data, emphasizing the need for hybrid approaches that integrate the strengths of both paradigms.

Studies on physics-informed data-driven models have sought to train AI frameworks using physics-based information to enhance interpretability (Donnelly et al., 2024; Xie et al., 2021; Hoedt et al., 2021; Samek et al., 2021; Kraft et al., 2020; Reichstein et al., 2019). Others have employed DNNs (e.g., LSTMs) to improve systematic calibration of traditional hydrological models, providing a bridge between these approaches (Tsai et al., 2021). Nonetheless, a significant gap persists in understanding how catchment-specific attributes influence DNNs' performance, particularly in regional contexts. For example, regional LSTM networks trained on multiple catchments data have been shown to outperform single-catchment models by employing shared regional hydrological information (Hosseini et al, 2025; Kratzert et al., 2024). However, the interplay between DL model performance and physical catchment attributes, such as climate, land cover, and soil type, remains underexplored.

Explainable AI (xAI) techniques (Başağaoğlu et al., 2022) have emerged as valuable tools for overcoming the interpretability challenges of machine learning models in hydrology. These methods aim to elucidate relationships between input features and model predictions, shedding light on the latent knowledge embedded through the training process. For example, Lees et al. (2021) demonstrated that LSTMs could replicate hydrological concepts like soil moisture and snow cover storage, with internal memory states showing strong correlations with these variables. Similarly, Kratzert et al. (2018) observed that LSTMs developed specialized memory cells for snow-driven catchments, effectively mimicking the behavior of conceptual snow storage models. These findings underscore the potential of DL models to internalize hydrological processes when trained on well-prepared datasets, moving beyond mere statistical mimicry to capturing underlying physical phenomena.

However, according to Başağaoğlu et al., 2022, challenges remain. Many xAI techniques rely on historical data, limiting their adaptability to nonstationary conditions caused by climate change and human activities. Additionally, their reliability often depends on the quality of available datasets, which can pose difficulties in data-scarce regions. Hybrid models that combine AI techniques with domain knowledge or physics-based constraints offer a potential solution.

Building on these advancements, this study investigates whether regionally optimized LSTM networks (only trained on hydrometeorological data) can capture the physical characteristics of different catchments without explicitly being exposed to unique catchments' attributes. While previous studies emphasize LSTMs' predictive capabilities, their ability to learn hydrological relationships unique to specific catchments has not been systematically examined. Exploring this ability is essential for bridging the gap between ML efficacy and physical hydrological understanding, especially in regional rainfall-runoff modeling.

We hypothesize that *regionally optimized LSTM networks, trained exclusively on hydrometeorological data while being blind to catchments attributes, are influenced by unique catchment-specific characteristics*. We posit that DNN model performance will vary based on the catchments' attributes and that optimized regional LSTMs can implicitly learn hydrological relationships unique to specific place.

In this study, we propose a novel triple-confirmation explainability framework to investigate whether regionally optimized LSTM networks—trained solely on hydro-meteorological inputs—can learn meaningful latent hydrological representations associated with catchment characteristics. While each of the individual analytical tools employed (correlation analysis, random forest models with SHAP and Gini importance, and principal component analysis) is established, their systematic integration for evaluating implicit catchment-specific feature learning from optimized black-box deep neural networks is new to the field. This approach contributes to the growing demand for trustworthy and explainable AI in hydrology by offering multiple independent lines of evidence to assess whether a neural network has learned hydrologically plausible patterns. To our knowledge, this is the first such attempt to apply an ensemble explainability strategy to pretrained regional LSTM models optimized via large-scale hyperparameter search across a real-world multi-catchment dataset.

This study addresses the following objectives:

1) Correlation Between LSTM Performance and Catchment Attributes: To what extent do catchment attributes, such as climatic conditions, topography, and land use, influence the performance metrics of optimized regional LSTM networks?

2) Latent Learning of Unique Catchment-Specific Features: Can regional LSTMs trained solely on hydrometeorological data implicitly learn catchment-specific features that affect their prediction accuracy in different places?

3) Impact of Hyperparameter Optimization: How does systematic hyperparameter optimization influence LSTM performance across diverse catchments in regional hydrology?

By exploring these questions, this work aims to enhance the interpretability, reliability, and performance of DNNs, specifically LSTMs, in hydrology, contributing to their application in real-world water management. Furthermore, it seeks to uncover latent hydrological insights embedded in DL models, complementing traditional modeling understandings.

2. Method

2.1. Case study and dataset: Basque Country hydrological system

This study focuses on Basque Country, located in north of Spain along the European Atlantic coast (Fig. 1). Covering approximately 4,494 km^2, the studied region is characterized by its humid climatology, abundant water resources, and diverse hydrological behaviors. The area includes 40 catchments ranging from small basins of 4 km^2 to large basins of 1,000 km^2. These catchments, noted for their flashy and humid characteristics, are situated between the Cantabrian Mountains—rising to elevations of 1,300 m—and the Atlantic Ocean Fig. 2.

The region's landscape consists predominantly of grasslands and evergreen forests, benefiting from the warming influence of the Gulf Stream. The climate is humid and temperate, with mean annual

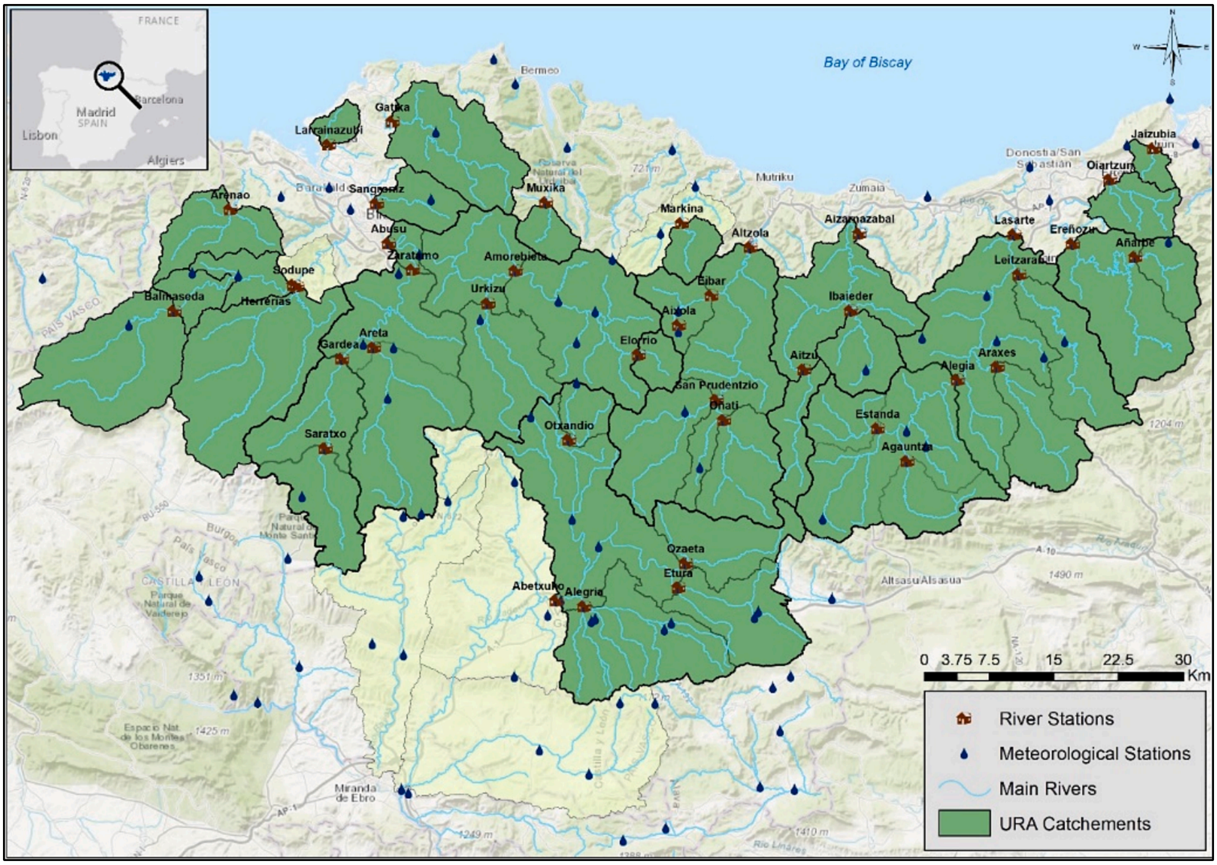


Fig. 1. Study area, including 40 catchments of Basque Country in north of Spain.

temperatures ranging from 9 °C in the mountains to 15 °C in lower regions. Annual precipitation varies between 1,200 mm and 1,600 mm, driven primarily by North Atlantic weather fronts. These climatic and geographic conditions make the region prone to intense rainfall events, rapid runoff, and a heightened risk of flash floods, which makes accurate hourly predictions a vital hydrological task in this region.

For this study, detailed hydrological and physical attributes of the 40 catchments were compiled (Table 1), with definitions provided in Table 2. Most of these attributes—such as land use classes and soil types—were obtained from regional GIS datasets and prior hydrological common projects and studies conducted by IH Cantabria in collaboration with the Basque Water Agency (URA). These datasets were derived from satellite imagery, remote sensing, and field surveys, and are publicly available and documented in previous publications. Meteorological and hydroclimatic variables, including precipitation statistics, runoff coefficient, and aridity index, were computed directly from available timeseries data following standard hydrological practices. These catchment descriptors are essential for interpreting the relationship between physical basin characteristics and the performance of regional LSTM models. Furthermore, the methodological framework we propose is generalizable and can accommodate alternative or simplified attribute sets depending on data availability, making it applicable across different hydrological regions.

The Basque Water Agency (URA), a regional governmental entity, manages water resources and policies in this area. URA has compiled a comprehensive and high-quality dataset of hourly hydro-meteorological timeseries. In this study we utilized data for over 21 years from October 1, 2000 to September 30, 2021. This dataset underpins the development of advanced hydrological models to enhance water management and flood prediction strategies.

- **Hydro-Meteorological Timeseries and Data Splitting**

The training dataset includes precipitation, temperature, potential evapotranspiration (PET), and streamflow. Catchment-scale areal rainfall and temperature were calculated using the inverse distance weighting (IDW) method (Ly et al., 2013), and PET was estimated using the Hargreaves formula (Hargreaves & Allen, 2003).

To ensure rigorous model training and evaluation, the dataset was partitioned into distinct subsets:

1) Training-and-Validation Set (October 1, 2000, to September 30, 2015):
   - Training Period: October 1, 2005, to September 30, 2015.
   - Validation Period: October 1, 2000, to September 30, 2005.

This subset was employed for hyperparameter optimization and deciding on the finally optimized networks.

2) Test Set (October 1, 2015, to September 30, 2021): This subset was withheld during training and validation to provide an unbiased evaluation of the optimized models. Once the model is optimized, it is re-trained on the train set and across 10 different random seeds to ensure robustness, as the initial point of training can still influence the final model's performance. This way, every optimized LSTM concluded in 10 retrained final models that is tested on the test set.

2.2. Model and main setups

This study employed the Multi-Timescale Long Short-Term Memory (MTS-LSTM) network, an advanced deep learning architecture tailored for hydrological predictions at fine temporal resolutions. The MTS-LSTM model, introduced by Gauch et al. (2021) and implemented via the NeuralHydrology Python library (Kratzert et al., 2022), is designed to address the computational challenges associated with hourly

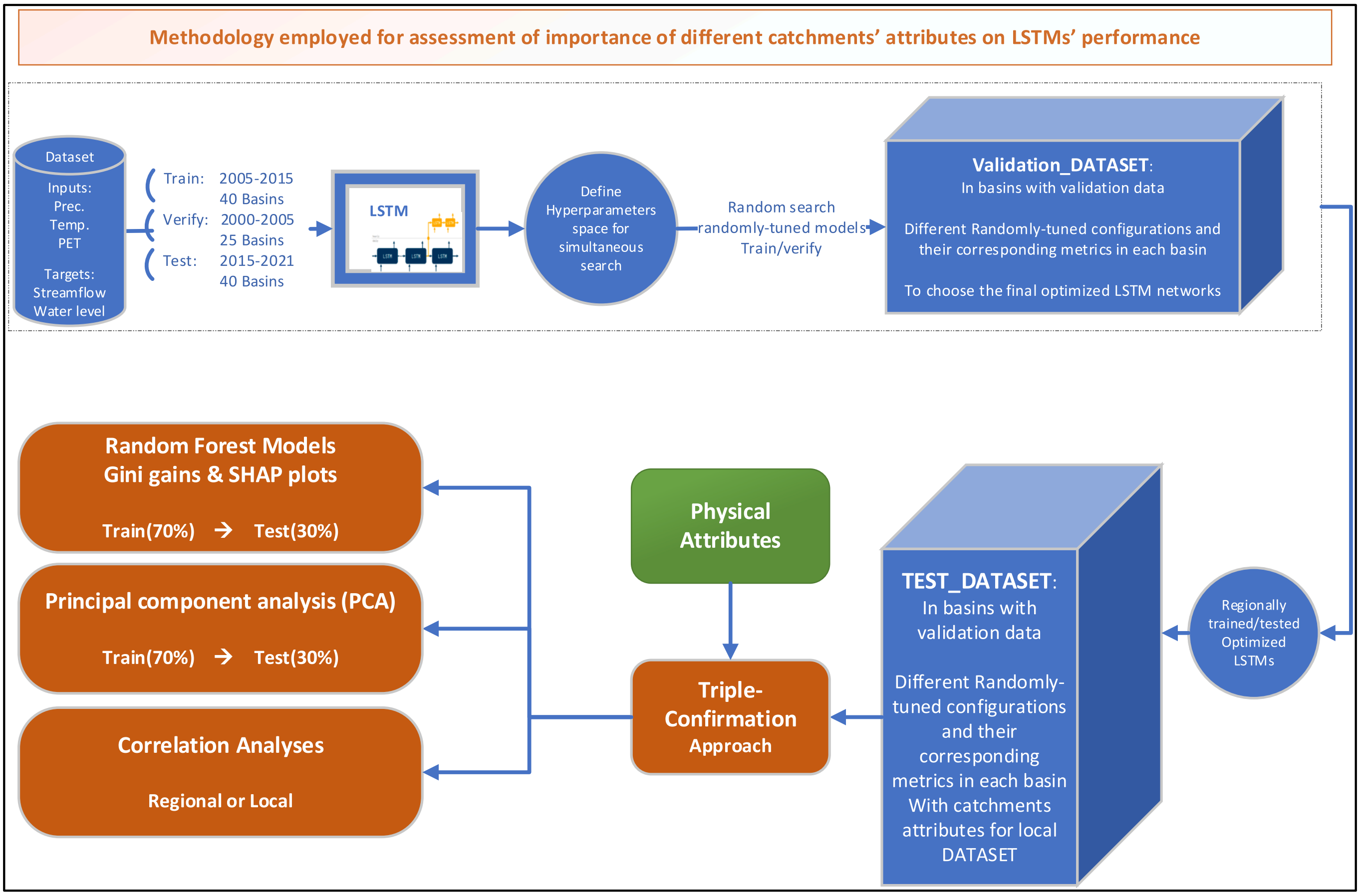

Methodology employed for assessment of importance of different catchments' attributes on LSTMs' performance
Dataset
Inputs:
Prec.
Temp.
PET
Targets:
Streamflow
Water level
Train: 2005-2015 40 Basins
Verify: 2000-2005 25 Basins
Test: 2015-2021 40 Basins
LSTM
Define Hyperparameters space for simultaneous search
Random search randomly-tuned models Train/verify
Validation_DATASET: In basins with validation data
Different Randomly-tuned configurations and their corresponding metrics in each basin
To choose the final optimized LSTM networks
Regionally trained/tested Optimized LSTMs
TEST_DATASET: In basins with validation data
Different Randomly-tuned configurations and their corresponding metrics in each basin With catchments attributes for local DATASET
Physical Attributes
Triple-Confirmation Approach
Random Forest Models Gini gains & SHAP plots
Train(70%) → Test(30%)
Principal component analysis (PCA)
Train(70%) → Test(30%)
Correlation Analyses
Regional or Local

Fig. 2. Methodology for assessing relations between catchments' attributes and LSTMs' test performances.

predictions. By utilizing two parallel LSTM networks—one operating at hourly and the other at daily timesteps—the architecture effectively captures both short-term dynamics and long-term trends while expediting the training process, making it particularly suitable for complex hydrological systems like those of the Basque Country.

For the MTS-LSTM implementation, the inputs were hourly average precipitation, temperature, and Potential Evapotranspiration (PET). These three variables comprehensively reflect the region's humid and temperate climate, providing essential insights into interactions between meteorological conditions and hydrological responses. The target was hourly streamflow in all 40 catchments, simultaneously.

The LSTM models were regionally trained using the comprehensive dataset from 40 catchments, employing the following rigorous workflow:

1. **Hyperparameter Optimization and Training**: All LSTM networks used in this study were regionally optimized through a systematic random search over 1000 hyperparameter configurations, as described in detail in our prior works (Hosseini et al., 2024a, 2024b; 2025). The hyperparameter space included 10 key settings (see Table 3), and a set of regionally optimized 84 configurations were selected based on their high validation performance across Nash–Sutcliffe (NSE) and Kling–Gupta (KGE) efficiency metrics. To ensure robustness against initialization variability, each of these 84 optimized LSTM networks was retrained using 10 different random seeds. This resulted in a comprehensive set of 840 regionally-trained optimized models with high performance, forming a robust and diverse ensemble for downstream explainability analysis. We emphasize that this ensemble serves as the foundation for our triple-confirmation framework and is based on highly optimized and validated deep learning models specifically designed for multi-catchment hydrological modeling.
2. **Evaluation**: The performance of the optimized LSTM networks in this study was evaluated using 14 performance metrics, selected to comprehensively assess different aspects of hydrological model behavior on unseen test data. These metrics include:
   - **Overall performance and error-based metrics**: 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).
   - **Decomposition metrics**: Alpha-NSE and Beta-NSE for evaluating the linearity and bias components of NSE; Beta-KGE for decomposing KGE (Gupta et al., 2012).
   - **Correlation metric**: Pearson's correlation coefficient (Pearson-r).
   - **Flow-segment biases**: %BiasFHV (high flows), %BiasFLV (low flows), and %BiasFMS (mid-segment slope) (Yilmaz et al., 2008).
   - **Peak flow metrics**: Peak-Timing, MAPE_peak (Mean Absolute Percentage Error for peaks), and missed_peaks (Fraction of Missed Peaks) (Kratzert et al., 2019; 2020).

While all 14 metrics were used in our evaluation framework and are presented in the correlation heatmap (Figs. 3 and 4) and overall performance summary (Table 4), the discussion in the results section focused more on the metrics that showed stronger patterns or clearer hydrological implications (e.g., NSE, KGE, %BiasFHV, or Peak-Timing). In analyses such as the Random Forest models (Fig. 5) and SHAP summary plots, only the top 10 most influential metrics (based on model relevance or interpretability) were highlighted for clarity.

We emphasize that hydrological behavior is complex and

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

														Elevation		Land Use									Soil					COND	PERM	HARD
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 %	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).

Table 2
Definitions of the catchments' attributes employed in this study.

Attribute	Definition	Group	Units/ Values
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	Topography	%
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

Table 3
The initial defined hyperparameters space designed for random search.

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

multifaceted, especially in flash-prone basins like those in the Basque Country. Hence, relying on a broad suite of evaluation metrics provides a more complete understanding of model strengths and weaknesses. For instance, while NSE emphasizes high flows, KGE is more sensitive to overall variability; %BiasFHV is essential for assessing peak representation, while %BiasFLV and %BiasFMS help evaluate low and moderate flows. Peak-specific metrics like MAPE_peak and missed_peaks offer insight into how well the models captured flood events — a critical aspect in flood-prone regions. Therefore, although not all metrics were equally emphasized in every figure or analysis, each served a role in the comprehensive evaluation framework. A complete list, including definitions and expected value ranges, is provided in Supplementary document Appendix 01 for transparency and reproducibility.

Moreover, to ensure that regional hydrological behaviors were comprehensively learned, the LSTM networks employed in this analysis were regional in design (meaning that they were trained regionally on data from all 40 catchments). This regional training approach enabled the models to predict streamflow at the outlets of all individual catchments while uncovering insights from the broader dataset. This design choice reflects the study's commitment to developing a scalable and effective framework for regional rainfall-runoff modeling, highlighting the importance of integrating shared and individual catchment dynamics into the predictive process.

2.3. Test_DATASET Setup and compilation

To systematically evaluate the implicit learning of catchments' attributes by the optimized deep neural networks in regional rainfall-runoff modeling, we developed a novel method based on their test performances in different basins. Our primary objective was to determine if the optimized regional LSTMs learned some relevant information about catchments' attributes without direct access to this information during training. Fig. 2 provides an overview of the developed methodology for assessing the hydrological understanding of regionally optimized LSTMs.

To achieve the objectives, this paper 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 84 distinct (in terms of their hyperparameter configurations) regionally optimized LSTM networks for rainfall-runoff modeling and catchment attributes across the 40 studied catchments in Basque Country, Spain. All in all, we analyzed performance of 840 trained LSTMs (84 networks retrained on 10 different random seeds) to study possible relations between catchments attributes and optimized DLs' test performances in different basins.

The test_DATASET consisted of three main group columns:

1) Hyperparameter Configurations: A detailed record of the hyperparameter configurations for all 84 optimized LSTM networks.
2) Catchments attributes (See: Tables 1 and 2).
3) Test Performance Metrics in every catchment: The performance of the 840 optimized LSTM networks in this research was evaluated against observed data from the test set.

Overall, the test_DATASET included up to 67,000 records (comprising hourly and daily predictions on ten different random seeds by 84 distinct hyperparameter-optimized LSTM architectures). This compilation provided a solid foundation for exploring potential correlations and trends between the predictive performance of optimized regional LSTM networks and catchment attributes in the studied region. These records were derived from several models optimized following an exhaustive random search in the hyperparameter space. Each regionally optimized LSTM network in the test_DATASET demonstrated competitive regional accuracy on overall, with some marginal differences. While, all the configured networks exhibited statistically significant varying performance across different locations (See: Hosseini et al., 2024a and 2025 for more details).

2.4. Exploration of implicit learning of catchment attributes

To investigate whether the optimized regional LSTM networks only trained on hydrometeorological timeseries implicitly learned catchment-specific features—despite not being explicitly trained on such attributes—we developed a triple-confirmation interpretability framework. This framework integrates Pearson correlation analysis, Random Forest regression (RF), and Principal Component Analysis (PCA), each contributing a unique perspective to uncover the hidden associations between LSTM performance and catchment attributes. These methods, grounded in explainable AI (xAI), aim to enhance transparency and interpretability in data-driven hydrological modeling (Başağaoğlu et al., 2022).

2.4.1. Pearson correlation analyses

Pearson correlation is a statistical measure of linear association between two continuous variables, calculated as:

$$r = \frac{\sum (x_i - \overline{x})^*(y_i - \overline{y})}{\left(\sum (x_i - \overline{x})^{2*} \sum (y_i - \overline{y})^2\right)}$$

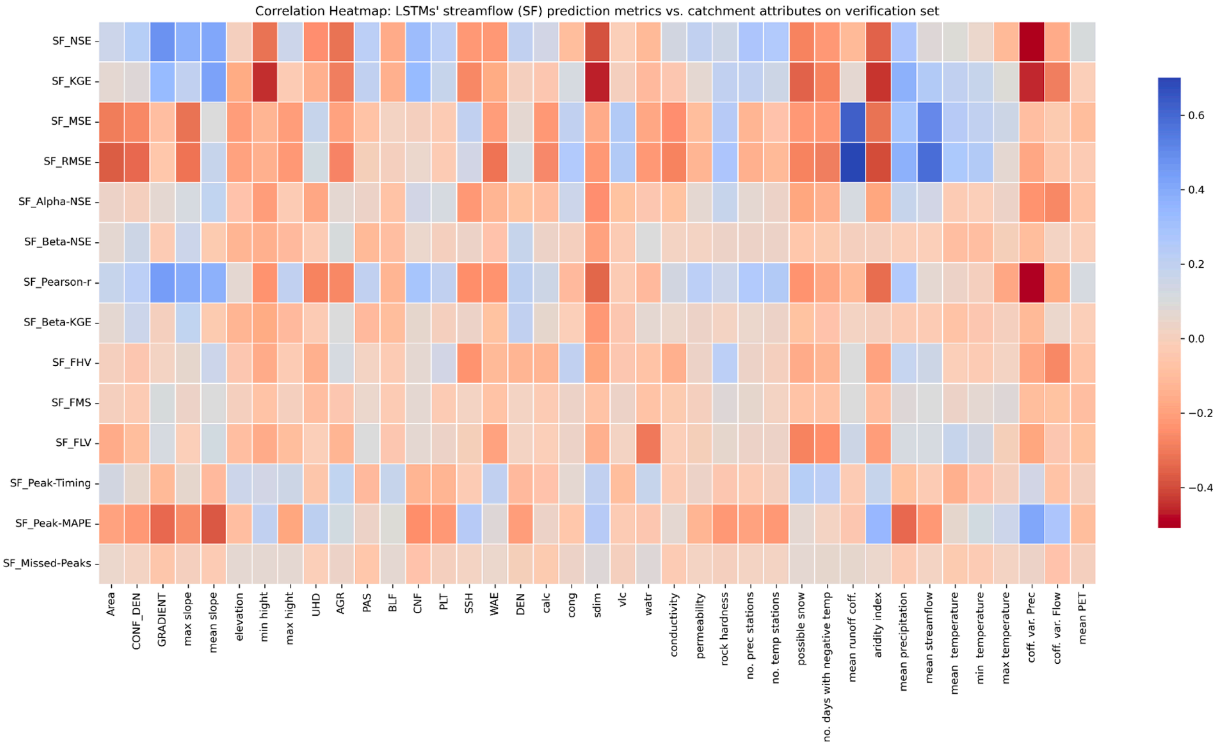


Fig. 3. Presents a correlation heatmap summarizing the relationships between selected catchment attributes and performance metrics across all 40 basins. Only attributes showing moderate correlations in preliminary screening were retained for this visual summary to enhance interpretability. The darker shades in the heatmap highlight stronger correlations, both positive and negative. Axe y demonstrates that different catchments attributes can affect LSTM performance from different perspectives that is confirmed by 14 distinct metrics each having their unique evaluation view point.

Where, r is the Pearson correlation coefficient, and x_i, y_i are paired observations. Values range from -1 (perfect negative correlation) to $+1$ (perfect positive correlation), with 0 indicating no linear association.

We computed the Pearson correlation coefficients between catchment attributes (e.g., area, slope, forest cover, soil type) and models' performance metrics (e.g., NSE, KGE, FHV) across all 40 catchments. Heatmaps were used for visualization, allowing intuitive identification of potentially influential relationships. We used a threshold of $|r| \geq 0.3$ to highlight moderate-to-strong correlations.

This step helped identify attributes that might have been implicitly captured by the LSTM networks through patterns in regional input–output timeseries data, despite not being directly provided.

2.4.2. Catchment-aware random forest model for implicit learning analysis

We then trained Random Forest regression models (RF) (Breiman, 1996), a tree-based ensemble learning method known for its robustness and interpretability. The RFs were used as a *meta*-model to predict different LSTM performance metrics (e.g., NSE) from both LSTM hyperparameters and catchment attributes for different catchments, thus enabling quantitative assessment of implicit learning.

Two complementary key interpretability techniques applied to analyze feature importance in the RF models:

- Gini Importance (Mean Decrease in Impurity): Measures how each feature reduces the impurity (variance) in regression tree splits (Breiman, 1996; Louppe et al., 2013).
- SHAP (SHapley Additive exPlanations) summary plots: A game-theoretic approach that computes the marginal contribution of each feature to model predictions, enabling local (catchment-specific) and global interpretability (Eun Choi et al., 2024; Lundberg & Lee, 2017).

The RFs were trained on 70 % of the test_DATASET and validated on the remaining 30 %. SHAP and Gini scores were then analyzed to:

- Identify key hyperparameters affecting LSTM performance.
- Discover catchment attributes most predictive of performance—indicating implicit learning by LSTMs.

2.4.3. Principal Component Analysis (PCA) for dimensionality reduction

Finally, we used PCA, a dimensionality reduction technique, to further validate whether LSTM performance trends clustered according to latent catchment features. PCA transforms input data into orthogonal principal components that maximize variance, revealing dominant patterns in high-dimensional data (Prieto et al., 2019; 2021; 2022; Jolliffe, 2002).

We applied PCA on the matrix of LSTM test performance metrics across catchments to observe whether performance trajectories aligned with groups of catchments sharing similar physical or hydrological properties. This unsupervised analysis helped visually confirm the emergence of catchment-specific learning patterns.

Together, these three methods provided converging evidence that regionally trained optimized LSTM networks—despite being trained without catchment attribute inputs—learned to reflect physical differences among basins. This finding underscores the capacity of deep learning models to extract latent geographical and hydrological signals from multi-catchment data and contributes to the growing field of interpretable hydrological modeling.

3. Results

We implemented a triple-confirmation analytical approach to investigate the relationships between catchment attributes and the accuracy metrics of optimized regional LSTM configurations on the test set. This approach encompassed Pearson correlation analysis, Random Forest (RF) regression modeling, and Principal Component Analysis (PCA). Together, these methods allowed us to: 1) Examine the associations between catchment attributes and model performance metrics, 2) Evaluate the predictability of LSTM performance using RF models, and 3) Identify the most influential features among the catchments' attributes and hyperparameters.

3.1. Pearson correlation analysis

To explore whether LSTM networks captured physical differences across catchments, we performed a Pearson correlation analysis between model performance metrics and several documented catchment attributes (e.g., topography, climate, land cover). The goal was to determine whether certain physical characteristics consistently influenced LSTM accuracy, despite not being explicitly used during model training.

Fig. 3 presents a Pearson correlation heatmap, illustrating the relationships between catchment attributes and performance metrics for multiple optimized LSTM configurations. While many relationships were weak to moderate, several patterns emerged and are analyzed in more detail below and in Table 4, which lists only the strongest attribute–metric correlations exceeding ± 0.3 threshold. These correlations revealed trends between catchment attributes and specific metrics, providing a clearer understanding of the conditions under which DNNs performed more accurately or struggled. These relationships suggest that regionally trained LSTM networks implicitly encoded catchment characteristics, likely through patterns in precipitation, temperature, and potential evapotranspiration sequences.

We emphasize that the heatmap in Fig. 3 is intended to illustrate general trends and not to suggest uniformly strong correlations. Instead, it visually supports the broader finding that LSTM models display sensitivity to catchment-specific traits even without direct access to those traits during training. The underlying data analysis and correlations are discussed in detail in the following subsections. We also acknowledge that formal significance testing (e.g., p-values for each correlation) and broader grouping of attributes can add statistical rigor. We note its potential value in future works, especially in larger sample contexts with more catchments and different climates.

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 clear 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 slightly diminished performance.

Topographic Influence on Predictive Metrics: Moderate correlations were observed between catchment slope and gradient and metrics like NSE (0.40), KGE (0.42), and Pearson-r (0.44). Steeper catchments likely exhibit distinct runoff patterns with reduced surface retention, aiding DNN model accuracy. However, negative correlations with extreme slopes (e.g., maximum slope attribute: -0.31 for RMSE) suggest that highly steep terrains might introduce noise, complicating model learning due to complex flow dynamics or input data limitations.

Influence of Climate and Variability: The aridity index showed negative correlations with KGE (-0.44) and RMSE (-0.40), indicating that less wet catchments with variable precipitation pose challenges for

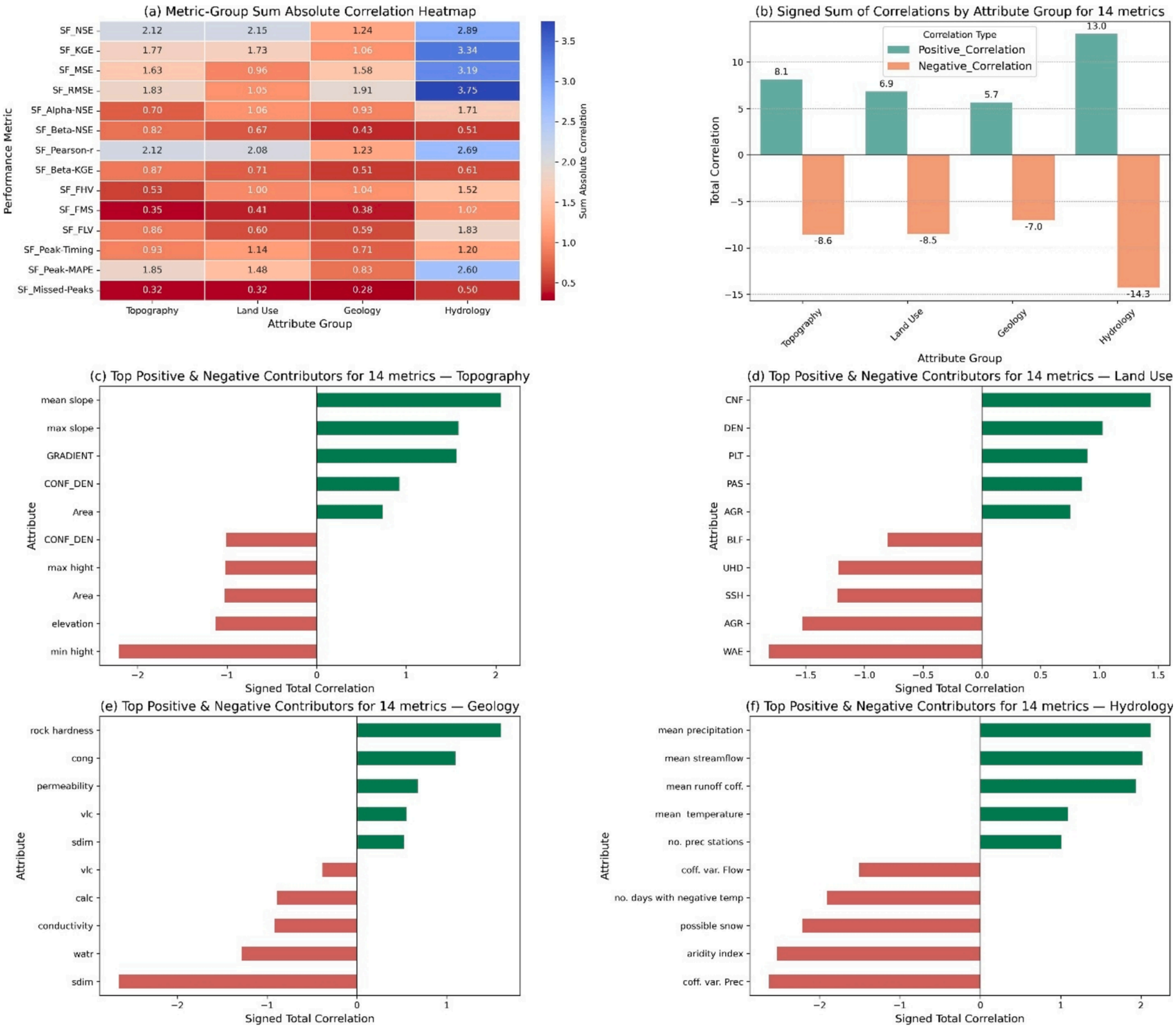


Fig. 4. Aggregated group-level correlation analysis between LSTM performance metrics and catchment attribute categories. (a) Sum absolute correlation heatmap showing the overall strength of association between each of the 14 performance metrics (rows) and the four attribute groups (columns): Topography, Land Use, Geology, and Hydrology. (b) Bar plot of the signed sum of correlations across all metrics for each attribute group, with green bars indicating total positive influence and red bars indicating total negative influence. (c–f) Horizontal bar charts displaying the top five positive (green) and top five negative (red) contributors within each attribute group: (c) Topography, (d) Land Use, (e) Geology, and (f) Hydrology on all 14 metrics. These panels highlight which individual attributes drive LSTM performance most strongly in either direction.

the LSTM models. Such variability reduces the generalizability of hydrological responses, impacting DLs' prediction accuracy. Similarly, the coefficient of variation for precipitation negatively correlated with KGE (−0.45) and Pearson-r (−0.49), emphasizing that stable precipitation regimes facilitate better model generalization and accuracy, whereas irregular patterns increase complexity.

Land Use and Vegetation Cover: Moderate positive correlations between coniferous forest cover and metrics like KGE (0.33) and NSE (0.32) suggest that consistent vegetation patterns may stabilize hydrological responses, improving prediction accuracy. Conversely, agricultural land use showed a negative correlation with NSE (−0.32), likely due to human-induced variability, such as irrigation and land management practices, which disrupt natural flow dynamics.

Geological Characteristics: Negative correlations between sedimentary soils and metrics like NSE (−0.39) and KGE (−0.47) suggest that heterogeneous permeability and storage properties introduce variability, complicating model performance. Similarly, attributes like wetlands (AWE) and water bodies (Watr) correlated positively with FLV (0.31) but negatively with RMSE (−0.31), indicating their role in modulating flow dynamics over time.

Probability of Snowfall and Temperature Variability: The negative correlation of KGE with snowfall probability (−0.35) highlights the challenge of capturing snowmelt dynamics, which introduces delays and variability in runoff. While snowfall is less predominant in Basque Country, its presence in specific basins underscores the complexity of hydrological modeling under such conditions by DNNs. Additionally, temperature variability negatively correlated with NSE, further highlighting the difficulty of handling highly fluctuating thermal conditions.

Aggregated Group-Level Correlation Analysis.

To provide a clearer overview of attribute–metric relationships, we grouped the 40 catchment attributes into four categories—Topography, Land Use, Geology, and Hydrology—and computed the sum absolute correlation between each group and the 14 LSTM performance metrics (Fig. 4a). We further calculated the signed total correlation for each group (Fig. 4b) on all 14 metrics and identified the top five positive and negative contributors within each group (Fig. 4c-f). This aggregated approach amplifies the most relevant signals—e.g., Hydrology exhibits the highest overall correlation magnitude—while preserving the directional information of positive versus negative influences. The detailed breakdown of contributors corroborates and refines our interpretation of

Table 4
The attributes with correlations beyond the ± 0.3 threshold, linking them to their influence on LSTM performance.

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
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

Fig. 3, demonstrating that specific attributes (e.g., mean precipitation, runoff coefficient in Hydrology; mean slope, maximum slope in Topography) are primary drivers of model performance in Basque Country catchments. This new representation strengthens the robustness of our conclusions by showcasing both broad trends and individual variable impacts.

3.2. Random Forest analysis

As stated, to further investigate the factors influencing the performance of optimized LSTM models in regional hydrology, we employed a Random Forest (RF) regression approach. The RF model utilized hyperparameter configurations and catchment attributes as inputs, while performance metrics of the LSTM networks served as the output targets. The dataset was divided into 70 % for training and 30 % for validation, ensuring robust model evaluation.

Fig. 5 displays the validation results of the RF model, illustrating the predicted versus actual performance metrics for streamflow across various catchments. The alignment of data points along the 1:1 line, combined with well-accepted metrics such as near-zero mean squared error (MSE) and R-squared values up to 0.97, demonstrates the trained RF model's capability to predict specific LSTM configurations' performance metrics accurately in different places. This indicates that the trained RF model effectively captured the relationships between catchment attributes, hyperparameter configurations, and the resulting LSTM model performance, underscoring its utility in analyzing complex, nonlinear interactions.

3.2.1. Feature importance analysis by Gini gains

Feature importance rankings derived from the RF model, presented in Fig. 6, reveal the most influential catchment attributes and hyperparameters contributing to regional LSTM model performance in different locations. The Gini gains, used to rank feature importance, highlighted several key findings:

1) **Hydrological Attributes:** Attributes related to hydrological processes—such as mean yearly streamflow, precipitation patterns, and the aridity index—emerged as critical predictors for several metrics, including KGE, Beta-NSE, and Missed-Peaks. For instance: Mean yearly streamflow ranked among the most significant predictors for RMSE and Beta-KGE, highlighting its importance in refining the predictive accuracy of LSTM models. Or, aridity index, which reflects climatic conditions, was also influential.
2) **Land Cover and Vegetation:** Land cover attributes, such as coniferous forest (CNF) and pasture cover (PAS), played a notable role in influencing accuracy metrics. Specifically, CNF ranked highly for metrics like NSE, Alpha-NSE, and Pearson-r, suggesting that vegetation types contributing to hydrological consistency can enhance

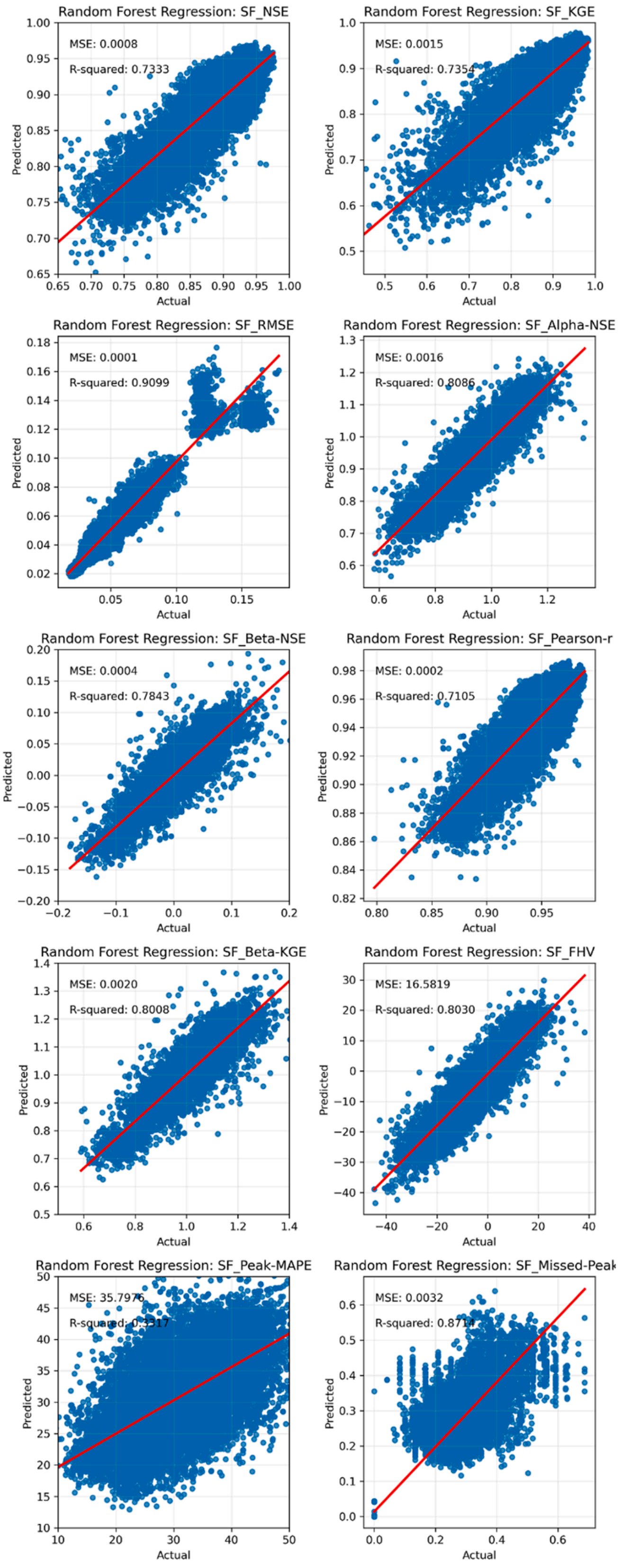


Fig. 5. Random Forest prediction accuracy for different test performance metrics for streamflow. The Random Forest was trained on Hyperparameters, and Attributes as inputs and the metrics as outputs. The figure suggests that a Random Forest model can be trained in a way that accurately predicts the performance outcomes of a configured LSTMs in every catchment by knowing their attributes.

model predictions. In contrast, agricultural land use (AGR) which, also, exhibited negative correlations with performance metrics, demonstrates high Gini values likely due to human-induced variability in hydrological processes.

3) **Topographical and Climatic Factors:** Features such as maximum temperature and maximum slope showed significant importance for specific metrics. Maximum temperature was particularly relevant for metrics like Missed-Peaks, indicating its role in regions with pronounced seasonal variations. Maximum slope showed varying importance, emphasizing the role of topography in shaping hydrological responses.
4) **Hyperparameter Configurations:** Several hyperparameters significantly influenced the RF model's predictive accuracy. However,

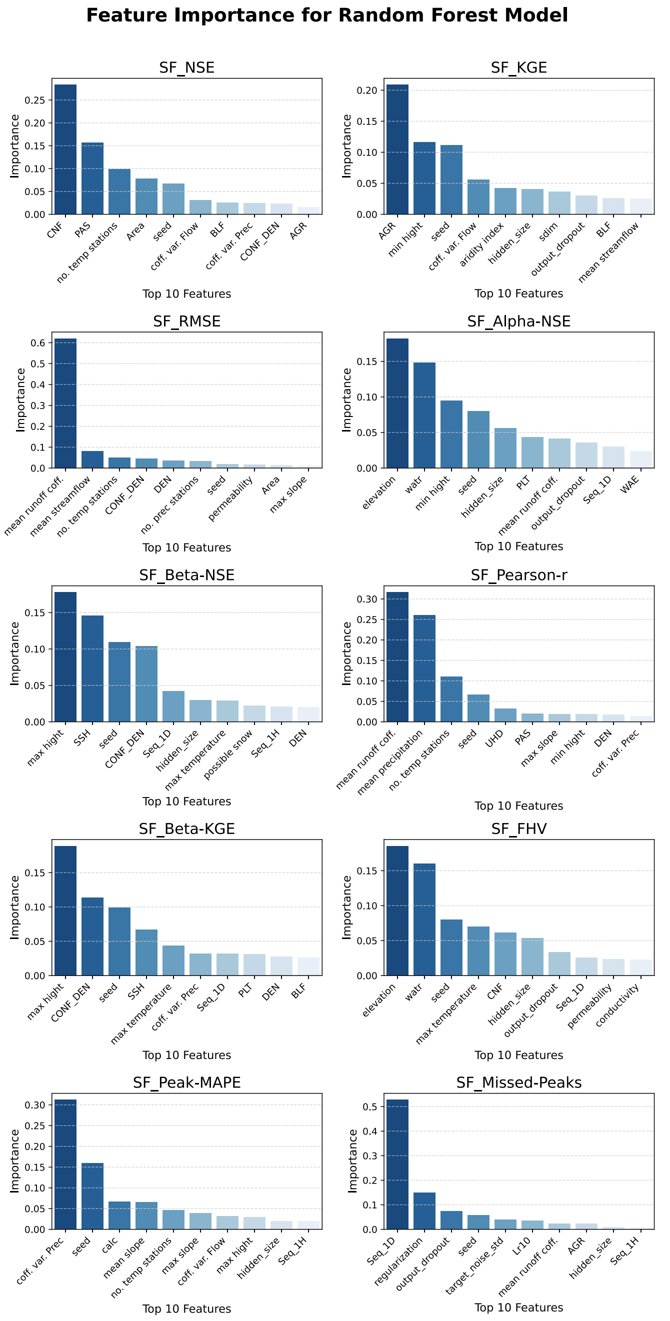


Fig. 6. Feature importance ranking for 10 top features for different target metrics derived from the trained Random Forest model. This figure highlights the relative importance of various features in predicting the test metrics, emphasizing the most influential attributes shaping DNNs' accuracy.

input sequence length, output dropout, and hidden size were critical for predicting metrics across catchments, from Gini gains perspective, indicating the importance of tailoring model configurations to specific hydrological conditions. Among them, the length of the input sequence could have hydrological importance carrying unique hydrological characteristics of the catchments (Hosseini et al., 2024c) in the hydrological domain. Seed values also showed consistent importance, emphasizing the influence of initialization on model outcomes. This finding aligns with the potential of ensemble learning to improve robustness and accuracy in hydrological predictions (Hosseini et al., 2025).

The Gini gains revealed variability in feature importance depending on geographic and climatic contexts. For example, Broadleaf Forest cover (BLF) and the coefficient of variation of flow were more influential in regions with variable flow regimes or significant forest contributions, affecting metrics like Alpha-NSE and Beta-NSE. Or, attributes like max slope and max temperature exhibit greater importance for metrics such as Missed-Peaks and FHV in catchments with distinct topographical and climatic characteristics, suggesting localized factors that affect hydrological responses.

The RF analysis highlights critical catchment attributes and hyperparameters that influence the performance of LSTM models in streamflow predictions. By effectively capturing nonlinear relationships and complex interactions, the RF model offers insights into the underlying drivers of model accuracy. Key takeaways include, 1) the significant role of hydrological attributes, such as streamflow and precipitation patterns, in enhancing model performance; 2) the influence of land cover types and topographical features on predictive accuracy; and 3) the importance of hyperparameter configurations, including input sequence length and seed values, in optimizing LSTM outcomes.

3.2.2. *Explainable AI analysis through SHAP summary plots*

To further explore the implicit learning behavior of optimized regional LSTM models, we employed SHAP (SHapley Additive exPlanations) summary plots—an advanced interpretability tool grounded in cooperative game theory. These plots quantify the contribution of each input feature (both LSTM hyperparameters and catchment attributes) to the model's output, providing both global and local explanations of model behavior across different hydrological contexts.

In this study, SHAP was used to analyze the 20 most influential features across 10 hydrological performance metrics, including NSE, KGE, RMSE, Alpha-NSE, Beta-NSE, Beta-KGE, Pearson-r, %BiasFHV, Peak-MAPE, and Missed-Peaks. This multifaceted evaluation allowed us to observe how various features influenced different aspects of LSTM networks skill—ranging from general accuracy to event-based or high/low flow sensitivity—within the humid and flashy hydrological regime of the Basque Country, Spain.

Each SHAP summary plot presents a horizontal axis representing the SHAP value (i.e., the impact of a feature on the model output), with individual dots denoting specific model instances. The color gradient (typically from blue to red) indicates the feature's actual value (low to high), and the position along the x-axis shows whether that value increased or decreased the predicted performance metric. This dual encoding enables nuanced interpretation, such as identifying whether high values of a given feature consistently lead to higher metrics values. Unlike traditional feature importance measures like Gini gain—which provide only aggregated, global insights—SHAP values enable instance-level (local) explanations and capture nonlinear feature interactions, which are particularly relevant for interpreting complex deep learning architectures like LSTM networks.

Influence of Catchment Attributes on LSTM Performance

The SHAP analysis reveals that the below catchment-specific hydrological and geomorphological characteristics are consistently dominant drivers of model performance across nearly all evaluated metrics:

Yearly mean runoff coefficient, streamflow, and precipitation emerge as the most impactful features, recurrently ranking at the top across NSE, KGE, RMSE, and Pearson-r. Their prominence reflects strong control over discharge dynamics and the data-driven model's capacity to capture basin-specific flow patterns. This aligns with classical hydrological understanding, where catchments with higher runoff generation and precipitation exhibit more predictable flow responses, favoring LSTM learning.

Catchment area and confluence density (CONF_DEN) significantly affect volumetric performance metrics such as NSE and RMSE. Larger catchments with denser drainage networks introduce spatial heterogeneity and complex routing effects, which challenge LSTM's ability to generalize learned patterns.

Slope-related features, including mean slope, maximum slope, and gradient, play a critical role in metrics related to extremes, noticeably for Peak-MAPE, FHV, and Missed-Peaks. These attributes reflect basin steepness, indicative of fast runoff generation and short response times, requiring the LSTM to learn rapid temporal dynamics effectively.

Land cover variables (e.g., pasture (PAS), agriculture (AGR), broadleaf forest (BLF), coniferous forest (CNF), shrubland (SSH), water surfaces (watr), wetlands (PLT), and denuded lands (DEN)) also exhibit substantial influence, particularly on KGE, NSE, and FHV. Their role underscores the effect of land use and vegetation on hydrological processes such as infiltration, evapotranspiration, and storage, which modulate streamflow generation and nonlinear responses to precipitation.

These findings provide empirical, data-driven validation of known hydrological principles, while also highlighting the capability of optimized regional LSTM networks to implicitly learn catchment-specific behaviors without direct access to these attributes during training.

Impact of LSTM Hyperparameters on Model Performance

In addition to catchment attributes, SHAP plots underscore the crucial influence of LSTM hyperparameters on regional model performance:

Sequence lengths daily and hourly (Seq_1D, Seq_1H) is consistently among the top influential hyperparameters across multiple metrics, with highest impact on Missed_peaks. Longer sequence lengths enable the model to capture longer temporal dependencies crucial for hydrograph reproduction; however, this could increase the computational costs.

Hidden size and learning rates are pivotal in shaping the model's capacity to learn complex flow patterns. Larger architectures generally enhance performance in metrics like KGE and FHV but may simultaneously increase vulnerability to errors in extreme flow metrics (e.g., NSE, Peak-MAPE), revealing a trade-off between model complexity and stability.

Dropout rate emerges as a critical regularization mechanism, influencing consistency and robustness across catchments. Properly tuned dropout rates reduce overfitting and enhance generalization, particularly noticeable in NSE, FHV, and KGE.

Batch size and loss function selections demonstrate metric-specific impacts, noticeably on NSE and Missed_peaks, suggesting their role in stabilizing model convergence and handling the variance in streamflow patterns. Larger batch sizes appear to smooth learning but may overlook localized peak behaviors.

Metric-Specific Findings and Hydrological Interpretations

Each performance metric yields distinct insights into how both catchment features and hyperparameters govern LSTM effectiveness:

NSE and KGE (global metrics) are primarily controlled by hydrological attributes—precipitation, area, runoff coefficient, and aridity index—highlighting their central role in shaping overall discharge reproduction. LSTM hyperparameters, particularly input sequence length and hidden size, are secondary but crucial to achieving high predictive skill.

RMSE, being sensitive to large deviations, is heavily influenced by slope, permeability, and confluence density, reflecting the difficulty in predicting extreme flows in steep or highly connected catchments.

Alpha-NSE and Beta-NSE, focusing on high and low flows

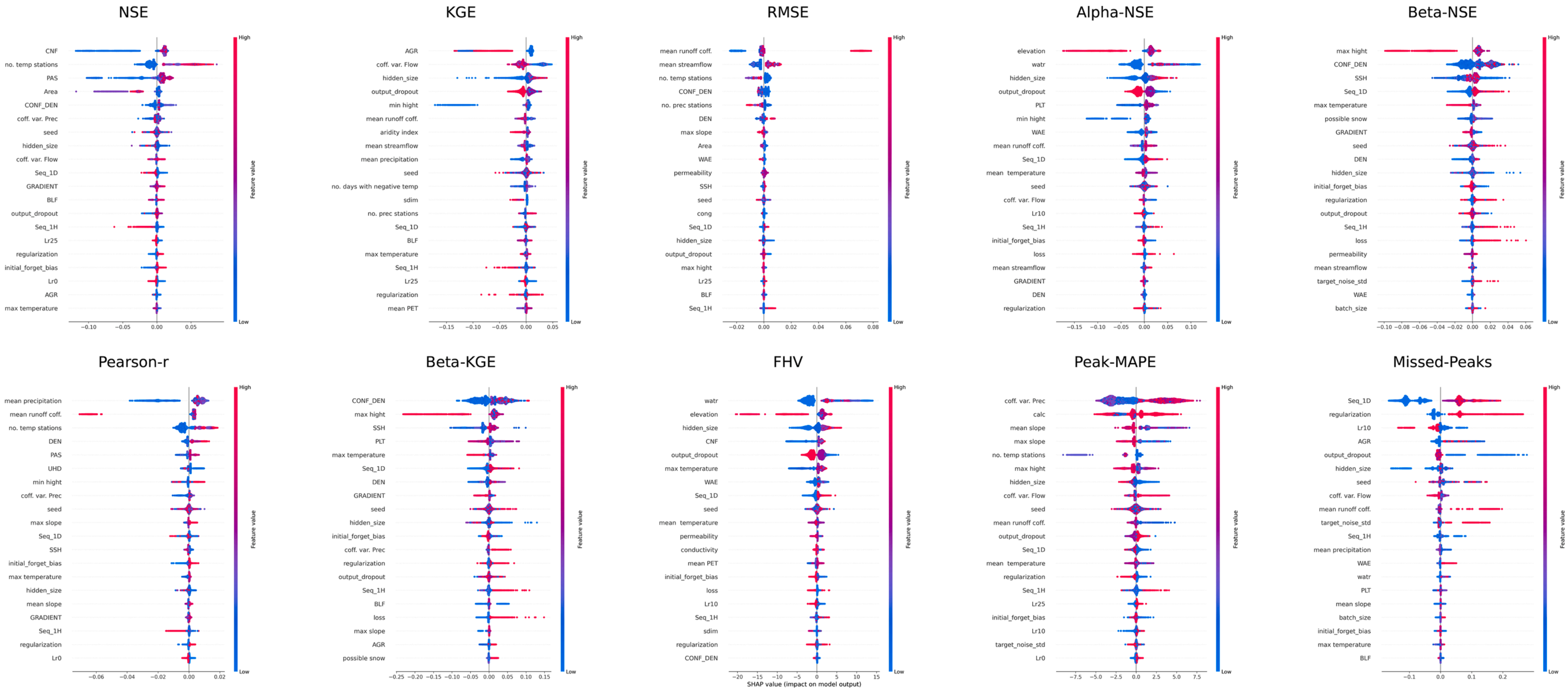


Fig. 7. SHAP summary plots for the 20 most influential catchment attributes and LSTM hyperparameters affecting regional model performance across Basque Country catchments. Each dot represents a single model-catchment instance, with color gradients showing feature magnitudes (red: high, blue: low). The SHAP value axis indicates feature positive and negative contribution to test performance metrics. Aggregated insights highlight dominant hydrological and architectural drivers of LSTM performance across diverse evaluation criteria.

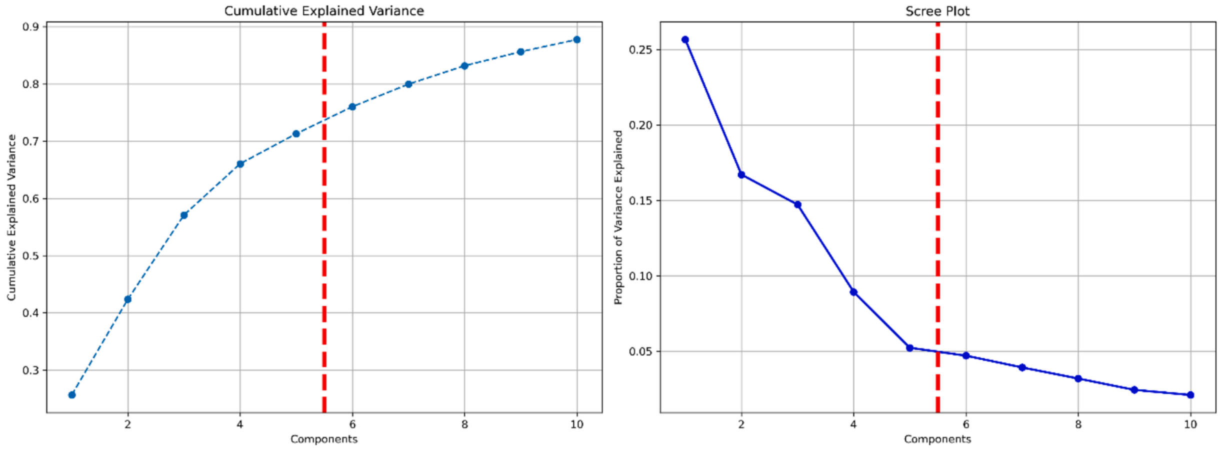


Fig. 8. PCA results applied to the test_DATASET, showing the distribution of principal components. This plot illustrates how much variability is captured by each component, providing insights into the test_DATASET's structure.

respectively, show differential sensitivity to both catchment and LSTM architectural parameters, reflecting the need for targeted model adjustments to handle flow extremes.

FHV and Peak-MAPE, representing high-flow and peak-specific errors, are predominantly controlled by steepness and precipitation extremes, underscoring the need for models that can handle rapid responses and flow surges.

Missed-Peaks, a direct indicator of LSTM capacity to capture critical events, shows notable sensitivity to hyperparameters such as sequence lengths (Seq_1D, Seq_1H), regularization, dropout, hidden size, learning rate, batch size, and initial forget gate bias, demonstrating that architectural design strongly influences extreme event detection capability.

All in all, the ensemble of SHAP plots (Fig. 7) offers a unified framework for understanding the interplay between catchment-specific

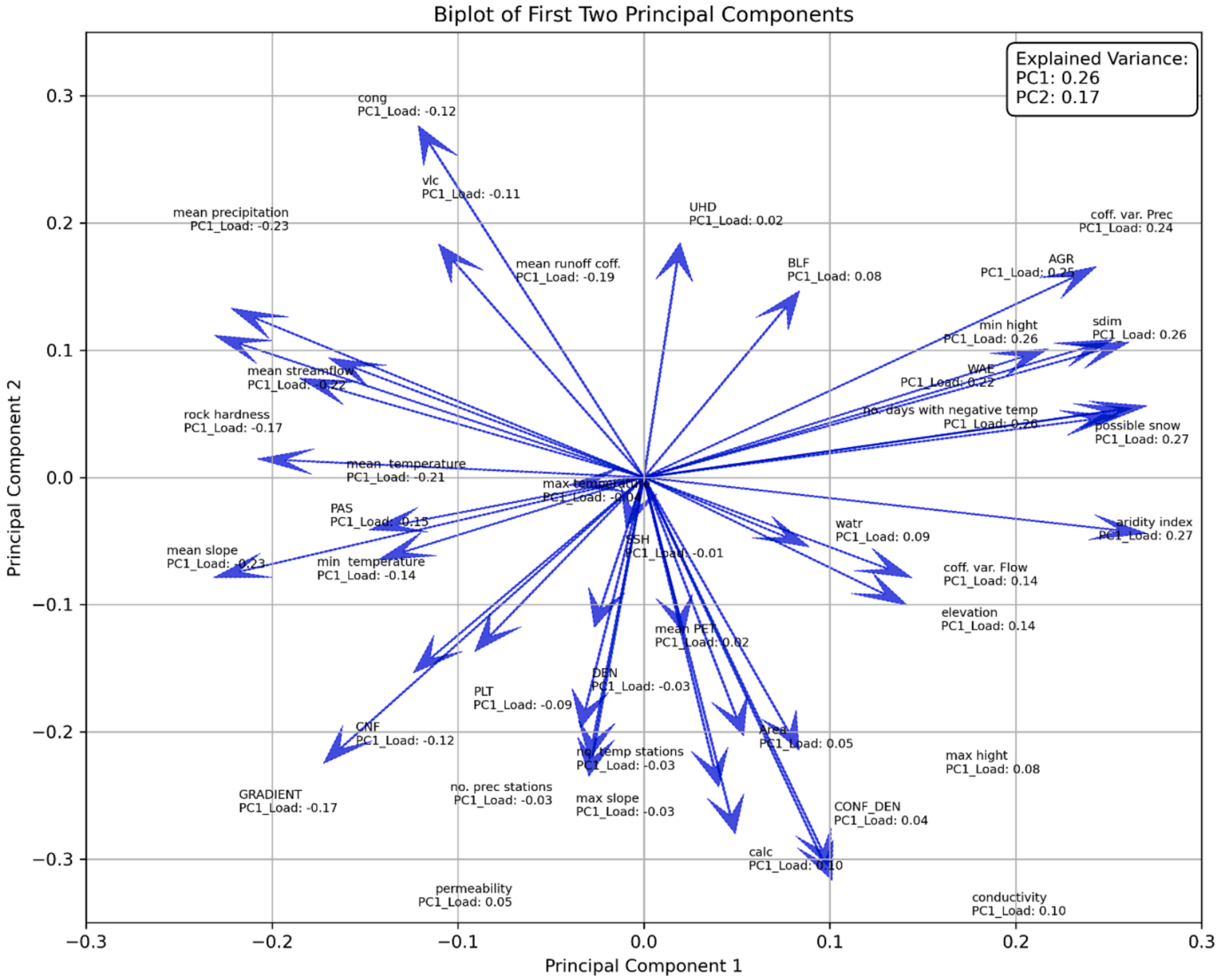


Fig. 9. Biplot of the PCA analysis. This figure displays the principal components in relation to the original features, visually representing how each feature contributes to the principal components and their interactions.

Table 5
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 %

hydrological behavior and LSTM model design. The results convey that hydrological features dominate overall model predictability, while hyperparameter tuning remains essential to optimize performance, particularly for flow extremes and peak events. Noticeably, the fact that LSTMs—trained solely on lumped hydrometeorological timeseries (precipitation, temperature, PET)—implicitly exhibit sensitivity to catchment attributes (not directly provided during training) illustrates the emergence of latent hydrological knowledge within the deep learning framework.

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.

Figs. 8 and 9 illustrate the PCA model results, including the scree plot (Fig. 8) and the biplot analysis (Fig. 9). The scree plot demonstrates that the first ten principal components cumulatively explain around 87.8 % of the test_DATASET's total variance with three first PCs having near 60 %. These 10 components capture the majority of the test_DATASET's variability, thus representing key aspects of the underlying structure. The biplot in Fig. 9 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.

Explained Variance and Component Loadings

Table 5 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.

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 5, the PCA loadings reveal each feature's influence on the principal components:

PC1: Climatic and topographic attributes dominate, explaining 25.7 % of the variance. Features with high positive loadings include the probability of snowfall (0.271), aridity index (0.270), area occupied by sedimentary soils (0.262), elevation (0.255), sedimentary rocks (0.261), and agricultural land area (0.252). Meanwhile, attributes such as average precipitation (−0.231), mean runoff coefficient (−0.185), average temperature (−0.208), and average slope (−0.232) 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) in this region, 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 predominant in PC2, capturing 16.7 % of the variance. Features with high negative loadings include river confluence density (CONF_DEN: −0.243), average soil conductivity (−0.317), area of calcareous rocks (−0.307), catchment size (−0.203), and max slope (−0.235). Conversely, the area occupied by conglomerate rocks (0.276) has a strong positive loading. These loadings suggest that the density of river confluences and soil type diversity are critical factors shaping catchment hydrological behavior in Basque Country basins, with more complex confluence networks likely introducing additional variability in hydrological responses.

PC3: Vegetation cover and land use attributes are influential in PC3, explaining 14.7 % of the variance. High negative loadings include maximum (−0.295) and average (−0.285) elevation, area covered by broadleaf forests (−0.229), and bedrock hardness (−0.228). Average precipitation (−0.218) shows a negative loading, while coefficient of flow variation (0.263) and pasture lands (0.222) and urban area (0.248) have positive loadings. This component suggests that broadleaf forests influence water retention, while urbanization affecting impervious surfaces, complicate runoff dynamics. Loadings for flow variability and urban areas reflect the crucial role of land cover in hydrological behavior in this region.

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 networks.

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 impact retaining water in this region, 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 performance 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 DNN LSTMs, which are critical for managing water resources amid variable climatic scenarios. The results highlight the potential of PCA-informed approaches to optimize DNNs in hydrology, enabling more accurate rainfall-runoff predictions in response to complex, shifting climate patterns.

4. Discussion

4.1. Regional hydrological artificial intelligence

Although the LSTM networks were trained solely on hydrometeorological inputs—without explicitly incorporating catchment attributes such as soil type, land cover, or elevation—the observed correlations between model performance and physical basin characteristics suggest that the models implicitly learned representations of catchment-specific features. In effect, the LSTMs were able to internalize the unique “fingerprint” of each catchment during regional training, using only timeseries input data. This capability likely emerged from the combination of informative input variables and carefully optimized hyperparameters.

Importantly, the results align with established hydrological understanding and complement recent findings by Kratzert et al. (2024) and Heudorfer et al. (2025). The former emphasized the benefits of regional training for enhancing model generalization and the latter concluded that “the superior performance of Entity-Aware deep learning models is primarily driven by information provided by meteorological data, with limited contributions from physiographic static features.”.

Key Findings and Implications from correlation analyses: This analysis underscores that catchments with stable hydrological regimes—characterized by high runoff coefficients, steady precipitation, and certain vegetation types like coniferous forests—are more conducive to accurate predictions by LSTM networks. In contrast, catchments with high climate variability, extreme flow fluctuations, complex geological features, or significant human modifications (e.g., agriculture, controlled flows) challenge DNN model accuracy.

Despite the absence of explicit catchment attributes in the training phase, the optimized regional LSTM models appear to have implicitly learned latent hydrological patterns and what we know from physical hydrology. These patterns likely reflect underlying processes and environmental interactions unique to specific catchments. Overall, this correlation analysis offers critical insights for refining DNNs, guiding future optimization efforts, and improving the understanding of how catchment characteristics influence hydrological predictions in DLs.

Intersection of physical hydrology and AI/DL models

The findings 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 (the catchments' uniqueness paradigm (Beven, 2000, 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.

4.2. Interpreting the hydrological relevance of AI explainability insights

The combined interpretation of Random Forest (RF) Gini gains and SHAP summary plots consistently highlights that a combination of catchment-specific hydrological attributes and AI-related hyperparameters governs LSTM networks performance in regional rainfall-runoff modeling across the Basque Country's humid and flashy catchments. Together, these explainability approaches provide robust and complementary insights into the factors shaping DNN's skill, supporting both hydrological reasoning and AI-specific considerations.

Hydrological attributes related to catchment behavior and rainfall-runoff processes dominate as key factors influencing LSTM accuracy and robustness. Specifically, mean yearly streamflow, precipitation seasonality, precipitation intermittency (prec_cv), and aridity index emerge as top contributors to model skill in both Gini and SHAP analyses. These variables reflect essential hydrological controls: streamflow magnitude and variability act as proxies for catchment responsiveness, hydrological stability, and memory effects, while precipitation regime characteristics inform the timing and intensity of runoff responses. The positive SHAP effects of streamflow on accuracy metrics (e.g., NSE, KGE) suggest that LSTMs perform best where hydrological signals are stable and continuous, while in highly variable or intermittent systems, generalization becomes more challenging. Furthermore, the aridity index highlights model sensitivity to intermediate climatic conditions, where a balance between predictability and hydrological complexity may govern LSTM learning capacity.

The role of land cover attributes, specifically coniferous forest (CNF), pasture (PAS), and agriculture (AGR), underscores the influence of vegetation as well as land use and human intervention on streamflow predictability. Coniferous forests generally stabilize flow regimes, dampening peaks and supporting sustained baseflows, which facilitates model learning. This is confirmed by their positive SHAP associations with model accuracy and negative associations with error metrics (e.g., Missed-Peaks, Bias). In contrast, pasture and agricultural areas, although important, are often associated with higher uncertainty and anthropogenic variability, as indicated by SHAP analyses showing negative influences on model accuracy in many catchments. This suggests that land use heterogeneity and human influences introduce additional complexity that challenges LSTM generalization.

Wetlands (WAE) and water bodies (watr) also exert important influences, pointing to eco-hydrological processes such as time-lagged responses, storage effects, and regulated discharges. These factors introduce delayed and nonlinear dynamics that are difficult for LSTMs to capture, explaining why their presence is often associated with reduced accuracy and higher Missed-Peaks as shown in SHAP summary plots.

Topographical features, especially maximum slope, emerge as important in the context of extreme flow events. Steep catchments, prone to quick runoff and flash floods, require LSTM models to learn fast-response dynamics. SHAP results show that higher slopes are associated with better peak capture (lower Missed-Peaks) but sometimes at the cost of increased bias, reflecting the trade-off between representing extremes and maintaining overall balance.

Hyperparameter Optimization: Hydrologically Informed AI Design

An important and novel insight emerging from both RF Gini gains and SHAP analyses is the critical role of LSTM hyperparameters in shaping model performance, underscoring the need for hydrologically informed AI design. Among these, input sequence length—which determines how much antecedent hydrological memory is provided—-shows one of the highest SHAP impacts across all performance metrics. The consistent positive influence of sequence length on accuracy and error reduction confirms that longer memory windows are beneficial for capturing delayed runoff processes, especially in flashy catchments with short response times. However, SHAP analyses also reveal diminishing returns and even negative impacts for excessively long sequences, signaling potential overfitting or dilution of relevant information. This highlights a critical trade-off between capturing hydrological memory and maintaining model generalization.

Other LSTM architectural hyperparameters, including hidden size and dropout rate, also exhibit substantial influence. Hidden size, which controls model capacity, must be carefully balanced to capture flow variability without overfitting, as evidenced by SHAP's mixed but significant contributions. Similarly, dropout rates, essential for regularization, show dual roles—improving generalization but, when excessively high, hampering the model's ability to learn fine-grained patterns, as revealed in both accuracy and error-based SHAP plots.

Interestingly, random seed (seed), which governs the stochastic initialization of LSTM weights, appears frequently among influential factors in both RF and SHAP results. This underlines the sensitivity of LSTM networks to initialization, a key challenge in deep learning (Sutskever et al., 2013; Glorot and Bengio, 2010). The results support ensemble modeling approaches, where training multiple models with different seeds can reduce variance and improve robustness (Hosseini et al., 2025).

Altogether, this joint SHAP and Gini analysis provides a comprehensive and hydrologically coherent narrative:

- Catchment hydrological attributes, especially those related to flow magnitude, precipitation regime, and aridity, fundamentally control LSTM skill and generalization.
- Land cover and topography modulate flow predictability and extremes, adding complexity that challenges AI models.
- LSTM hyperparameters, specifically input sequence length, hidden size, and dropout, need careful tuning guided by hydrological insights to ensure optimal performance.
- The sensitivity to random seed emphasizes the importance of ensemble learning to achieve stable and generalizable predictions for multi-objective tasks such as regional hydrological rainfall-runoff modeling in different catchments.

4.3. Limitations of the research

Despite the promising results, this study has some limitations that should be acknowledged to guide future research and real-world application:

1. Generalizability Across Hydrological Regimes

The models and findings presented in this study are based on catchments in the Basque Country, Spain, which are predominantly humid and flashy. As such, the generalizability of the interpreted results to drier regions, snow-dominated basins, or basins with different hydrological regimes (e.g., dominant groundwater influence, snowmelt) may be limited. Hydrological regimes vary widely in terms of dominant processes, seasonal variability, and input–output lags, which may affect the ability of LSTM networks to implicitly learn meaningful representations from meteorological data alone. Future work should test the proposed framework in contrasting climatic zones and broader hydrological regimes to assess its transferability and robustness.

2. Data Requirements and Quality Constraints

The training and evaluation of the LSTM models relied on high-quality data, which is critical for capturing fast-response events and hydroclimatic variability in flashy catchments such as Basque Country, Spain. However, such granular data (e.g., hourly) is not always available in many regions, especially in developing countries or sparsely monitored basins. This poses a practical limitation on the applicability of the approach in data-scarce settings. Alternative strategies, such as downscaling or transfer learning from data-rich regions, may be explored to mitigate this challenge.

3. Temporal Non-Stationarity and Model Stability

Hydrological systems are inherently non-stationary due to land use changes, climate variability, and anthropogenic interventions. Deep learning models trained on historical data may degrade in performance over time unless retrained periodically or rely on latest ideas of reinforcement learning by getting updated through the time. However, retraining raises challenges related to data drift, parameter instability, and computational cost. Incorporating temporal adaptation mechanisms (e.g., online learning, domain adaptation) or hybrid physical-AI models may offer a path forward for sustaining model relevance in changing environments.

4.4. Future research directions

Hydrological Regime Sensitivity and Implicit Learning: The key contribution of this study is the demonstration that LSTM networks, when trained regionally on hydrometeorological data, can implicitly learn catchment-specific hydrological behaviors. However, the strength and nature of this implicit learning are likely to vary across different hydrological regimes. In humid and flashy environments like the Basque Country, streamflow responses are more tightly coupled with meteorological drivers, which enhances the learnability of catchment-specific dynamics from timeseries alone. In contrast, arid basins, snow-dominated regions, or areas with complex groundwater–surface water interactions may present additional challenges due to lagged, nonlinear, or less deterministic runoff processes. These differences could limit the transferability of implicit learning frameworks without auxiliary inputs. Future studies should explore how hydrological regime complexity influences the degree to which deep learning models can infer physical characteristics solely from meteorological signals by training them on a much broader datasets from different hydrological regimes such as the CAMELS US catchments.

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 (Klotz et al., 2022), 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.

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. Moreover, further exploration of DL ensemble models, particularly those that combine LSTMs with Transformers (Vaswani et al., 2017), could increase robustness, enabling better performance in regions affected by seasonal shifts and climate-driven hydrological changes.

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. Moreover, 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.

Enhancing Interpretability with Explainable AI (xAI): Future research should prioritize improving the interpretability of DL models through more different and advanced 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. Moreover, the integration of SHAP-based interpretability in this study enhances transparency and provides actionable insights into model design for operational hydrology. Future research should investigate dynamic DL architectures (e.g., attention mechanisms) and hybrid physics-informed neural networks to improve model adaptability and robustness in complex catchments.

5. Conclusion

This study demonstrates that hyperparameter-optimized regionally-trained Long Short-Term Memory (LSTM) networks can effectively learn rainfall-runoff dynamics in diverse catchments of the Basque Country, using only hydrometeorological inputs. Despite being trained without explicit access to catchment-specific attributes, the LSTMs were able to implicitly capture latent unique hydrological signatures of different catchments, achieving robust generalization across humid and flashy basins. Generally, this study posits that regionally-trained optimized LSTMs learned knowledge is aligned with our understanding of physical hydrology.

To assess the internal learning behavior of these models, we introduced a triple-confirmation explainability framework—integrating Pearson correlation analysis, Random Forest modeling (via Gini and SHAP), and Principal Component Analysis. These tools consistently revealed that both catchment characteristics (e.g., runoff coefficient, streamflow magnitude, precipitation variability) and LSTM hyperparameters (e.g., input sequence length, hidden size, dropout) significantly influence model performance.

Catchments with stable hydrological signals were easier to predict, while basins with high variability, steep slopes, or low moisture posed greater challenges. SHAP summary plots further confirmed that larger and wetter basins tended to yield better LSTM accuracy, whereas high precipitation intermittency and topographic extremes negatively affected performance.

From a modeling perspective, input sequence length emerged as a key hyperparameter, with longer memory improving predictions in flashy catchments in general—though with diminishing returns beyond certain thresholds. This highlights the importance of systematic hyperparameter optimization with respect to finding true input sequence length for every catchment when applying deep learning in hydrology (See: Hosseini et al., 2024a, c).

Overall, this study contributes a reproducible methodology for interpreting deep learning models in hydrology, offering practical insights for both AI developers and hydrologists. While our results affirm the potential of regional LSTMs trained solely on meteorological inputs, future work may explore the integration of catchment attributes to further enhance performance in complex or data-scarce regions and broader climates.

These findings support the development of more interpretable, reliable, and operationally useful AI models for hydrological forecasting and water resource management, especially in flash-flood-prone areas like the Basque Country.

6. Codes, data and reproducibility

The dataset and all related codes utilized for this research, along with comprehensive instructions for replicating the experiments, are accessible on our repositories on https://doi.org/10.5281/zenodo.15080569, https://doi.org/10.5281/zenodo.15080600, https://github.com/farzadhoseini/xAI_LSTMs_BasqueCountry, https://github.com/farzadhoseini/ensemble.deep.learning, https://github.com/farzadhoseini/Precise_Tuning_of_Regional_Hydrological_LSTM_Networks. We prioritize transparency and reproducibility so that fellow researchers and practitioners can verify our findings and employ the same codes for hyperparameter optimization, ensemble learning, and Explainable AI techniques of their research and applications.

CRediT authorship contribution statement

F. Hosseini: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **C. Prieto:** Writing – review & editing, Validation, Supervision, Project administration, Conceptualization. **C. Álvarez:** Writing – review & editing, Validation, Supervision, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by the Instituto de Hidráulica Ambiental de la Universidad de Cantabria (IHCantabria), which funded the Ph.D. research and provided the essential computational resources. We extend our sincere gratitude to Dr. Grey Nearing (Google Research) for his invaluable guidance on deep learning methodology, and to the NeuralHydrology Python library team for their open-source contributions. We also thank Dr. Ross Woods (University of Bristol) for his insightful comments on an earlier version of this manuscript, which greatly enhanced its scientific rigor. We are indebted to the Basque Country Water Agency (URA) for granting access to their hydrometeorological dataset, without which this study would not have been possible. Finally, we express our deep appreciation to the two anonymous reviewers and to Professor Dr. Andras Bardossy, Editor-in-Chief of the Journal of Hydrology, for their constructive feedback that substantially improved the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jhydrol.2025.133689.

Data availability

Dataset: https://doi.org/10.5281/zenodo.15080569, https://doi.org/10.5281/zenodo.15080600, Code: https://github.com/farzadhoseini/xAI_LSTMs_BasqueCountry

References

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

Başağaoğlu, H., Chakraborty, D., Lago, C., do, L., Gutierrez, 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. https://doi.org/10.3390/w14081230.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24 (2), 123–140. https://doi.org/10.1007/BF00058655.

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

Beven, K. (2012). Rainfall-runoff modelling: The primer (2nd ed.). John Wiley & Sons. doi:10.1002/9781119951001.

Beven, K.J., 2000. Uniqueness of place and process representations in hydrological modelling. Hydrol. Earth Syst. Sci. 4 (2), 203–213. https://doi.org/10.5194/hess-4-203-2000.

Donnelly, J., Daneshkhah, A., Abolfathi, S., 2024. Physics-informed neural networks as surrogate models of hydrodynamic simulators. Sci. Total Environ. 912, 168814. https://doi.org/10.1016/j.scitotenv.2023.168814.

Eun Choi, J., Won Shin, J., Wan Shin, D., 2024. Vector SHAP values for machine learning time series forecasting. J. Forecast. 44, 635–645. https://doi.org/10.1002/for.3220.

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

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press. ISBN: 978-0262035613. 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.

Gupta, H.V., Clark, M.P., Vrugt, J.A., Abramowitz, G., Ye, M., 2012. Towards a comprehensive assessment of model structural adequacy. Water Resour. Res. 48, W08301. https://doi.org/10.1029/2011WR011044.

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. J. Hydrol. 377 (1–2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003.

Hargreaves, H., ASCE, F., Allen, R., 2003. History and Evaluation of Hargreaves Evapotranspiration Equation. J. Irrig. Drain. Eng. 129 (1), 53–63. https://doi.org/10.1061/(ASCE)0733-9437(2003)129:1(53).

Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory, Neural Computation, 9, 1735–1780, 1997. doi:10.1162/neco.1997.9.8.1735.

Heudorfer, B., Gupta, H.V., Loritz, R., 2025. Are deep learning models in hydrology entity aware? Geophys. Res. Lett. 52. https://doi.org/10.1029/2024GL113036.

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.

Hosseini, F., Prieto, C., Álvarez, C., 2025. Ensemble learning of catchment-wise optimized LSTMs enhances regional rainfall-runoff modelling—Case Study: Basque Country Spain. J. Hydrol. 132269. https://doi.org/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. J. Hydrol. 132003. https://doi.org/10.1016/j.jhydrol.2024.132003.

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

Hosseini, F., Prieto, C., Nearing, G., Alvarez, C., and Gauch, M. (2024c) 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.

Jolliffe, I. T. (2002). Principal component analysis (2nd ed.). Springer.

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., Nearing, G., 2024. HESS opinions: never train a Long Short-Term Memory (LSTM) network on a single basin. Hydrol. Earth Syst. Sci. 28 (17), 4187–4201. https://doi.org/10.5194/hess-28-4187-2024.

Kratzert, F., Gauch, M., Nearing, G., Klotz, D., 2022. NeuralHydrology — A Python library for deep learning research in hydrology. J Open-Source Softw. 7 (71), 4050. https://doi.org/10.21105/joss.04050.

Kratzert, F., Klotz, D., Hochreiter, S., 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. https://doi.org/10.5194/hess-2020-221.

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

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

Lees, T., Reece, S., Kratzert, F., Klotz, D., Gauch, M., De Bruijn, J., Kumar Sahu, R., Greve, P., Slater, L., Dadson, S., 2021. Hydrological concept formation inside long short-term memory (LSTM) networks. Hydrol. Earth Syst. Sci. Discuss. https://doi.org/10.5194/hess-26-3079-2022.

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

Louppe, G., Wehenkel, L., Sutera, A., Geurts, P., 2013. Understanding variable importances in forests of randomized trees. Advances in Neural Information Processing Systems (NeurIPS) 26, 431–439.

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS 2017). doi:10.48550/arXiv.1705.07874.

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. https://doi.org/10.6084/M9.FIGSHARE.1225842.V1.

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

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through. Part I. A conceptual models discussion of principles. J. Hydrol. 10, 282–290. https://doi.org/10.1016/0022-1694(70)90255-6.

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 Resour. Res. 58, e2021WR030705. https://doi.org/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 Resour. Res. 57, e2020WR028338. https://doi.org/10.1029/2020WR028338.

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 Resour. Res. 55, 4364–4392. https://doi.org/10.1029/2018WR023254.

Refsgaard, J.C., Stisen, S., Koch, J., 2022. Hydrological process knowledge in catchment modelling – Lessons and perspectives from 60 years development. Hydrol. Process. 36 (1), e14463. https://doi.org/10.1002/hyp.14463.

Reichstein, M., Camps-Valls, G., Stevens, B., et al., 2019. Deep learning and process understanding for data-driven Earth system science. Nature 566, 195–204. https://doi.org/10.1038/s41586-019-0912-1.

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 Process Mag. 38 (6), 19–48. https://doi.org/10.1109/JPROC.2021.3060483.

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.

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.

Tripathy, K.P., Mishra, A.K., 2024. Deep learning in hydrology and water resources disciplines: concepts, methods, applications, and research directions. J. Hydrol. 628, 130458. https://doi.org/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. Nat. Commun. 12 (1), 5988. https://doi.org/10.1038/s41467-021-26107-z.

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 Res. 30 (1), 79–82.

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. J. Hydrol. 603, 127043. https://doi.org/10.1016/j.jhydrol.2021.127043.

Yilmaz, K.K., Gupta, H.V., Wagener, T., 2008. A process-based diagnostic approach to model evaluation: application to the NWS distributed hydrologic model. Water Resour. Res. 44, W09417. https://doi.org/10.1029/2007WR006716.