DOCTORAL THESIS PROPOSAL

DOCTORAL PROGRAM IN ENGINEERING
PEDAGOGICAL AND TECHNOLOGICAL UNIVERSITY OF COLOMBIA

Title: Computational model for the spatiotemporal prediction of monthly precipitation in mountainous areas using machine learning techniques.

Date: November 20, 2024.
Name of proponent: Manuel Ricardo Pérez Reyes (manuelricardo.perez@uptc.edu.co)
Identification of the proponent: 1057.570.907
Proposer's signature:

Name of advisor: PhD. Marco Javier Suárez Barón (marco.suarez@uptc.edu.co)
Director: 7.167.871
Research group: GALASH
Line of research: Artificial intelligence (AI)

Co-advisor: PhD. Oscar Javier García Cabrejo (oscar.garcia04@uptc.edu.co)
Co-advisor ID: 79.913.080
Research group: GALASH
Line of research: Computational Hydrology

Keywords: Machine learning, Deep learning, Rainfall prediction, Monthly, Time series, Space-time.

Content

List of illustrations

List of tables

1 SUMMARY

This doctoral thesis proposal addresses the critical issue of changing precipitation patterns due to global climate phenomena, which affect the accuracy of existing prediction models. The primary goal is to design and implement a computational model for predicting monthly precipitation in mountainous areas, specifically focusing on the Department of Boyacá, Colombia.

The methodological framework consists of three key components: the state of the art of the research, the project architecture, and a timeline of results. The state of the art provides a solid foundation for the research, while the architecture involves a macro-level design for rainfall prediction, which would incorporate various models depending on their effectiveness. Researchers in the LAGASH group, including undergraduate and master's students, contribute to enriching the data set and the variety of models. The methodological framework's third component is the deliverables schedule, which includes three levels: beginner, intermediate, and advanced, each with specific tasks and expected results to ensure progressive achievement of the project objectives.

The proposal focuses on techniques to improve the accuracy of monthly precipitation predictions in mountainous regions. Data preprocessing techniques are crucial for enhancing prediction accuracy. These techniques include clustering for improved data segmentation, specific preprocessing algorithms, and integrating geographic and zoning variables.

The project architecture integrates several models developed by the Galash research group. Each model is evaluated individually and in combination-standard, hybrid, or stacked-to determine the optimal configuration for performance. This comprehensive approach leverages the strengths of time series analysis, data preprocessing, and hybrid modeling to achieve superior prediction results, providing a robust framework for spatiotemporal prediction of monthly precipitation in mountainous areas.

This research aims to achieve significant social, economic, and technological impacts by providing more reliable rainfall predictions and contributing to effective water resources management by considering the adverse effects of climate change and the complexity of mountainous areas in predictive models. From an economic point of view, it aims to support agricultural planning and disaster preparedness, potentially reducing losses caused by droughts. From a technological point of view, it seeks to advance the field of predictive modeling, providing a model for future research in similar areas. By addressing the challenges of precipitation prediction in mountainous regions with innovative techniques and a comprehensive methodological framework, this proposal aims to contribute substantially to the academic community and practical applications in environmental management and agriculture.

2 PROBLEM STATEMENT

Precipitation is any liquid or ice resulting from the condensation of atmospheric water vapor that falls from a cloud, and it is the ultimate source of freshwater [1]. The primary sources of precipitation include drizzle, rain, sleet, snow, graupel, and hail. Less than 1% of the water on Earth is fresh and accessible, and most of that water is replenished by precipitation [1]. Predicting precipitation is essential for optimal water resource management and informed decision-making [2], [3]. However, climate change and global warming have altered precipitation patterns in various regions [4], [5], [6]. Dry areas become drier, generally in the subtropics, while wet regions become wetter, mainly in mid to high latitudes [7], [8].

Rainfall prediction is inherently complex due to three key factors: its non-linearity, non-stationarity, and stochastic nature. Non-linearity arises because precipitation patterns are influenced by complex interactions among multiple climatic variables, which do not follow direct or predictable relationships, as noted in [9]. Additionally, the non-stationarity of precipitation time series means that the statistical properties of the data change over time, making it difficult to apply traditional models that assume stable or constant behavior. Precipitation variability is also stochastic, presenting high randomness and highly uncertain long-term predictions. These characteristics pose significant challenges for traditional techniques, such as ARIMA models, which struggle to capture these complex and variable dynamics [9]. In this context, artificial intelligence approaches, like neural networks and hybrid models, have proven more effective, as they are better equipped to handle non-linear and non-stationary data, as highlighted in [10]. However, prediction remains a significant challenge due to the constant interaction of these three factors, necessitating advanced techniques to improve accuracy [11].

The department of Boyacá, in Colombia, is a mountainous region with altitudes between 145 and 5,490 meters above sea level, traversed by the eastern range of the Andes. This area has been affected by climatic oscillations that have altered precipitation patterns. According to IDEAM, in the 2014 National Water Study [12], 25% of the municipalities in Boyacá faced drought emergencies, which is reflected in a significant decrease in precipitation and, consequently, in river flows. This affects the water supply in the municipalities of the provinces of Ricaurte, Centro, Tundama, and Sugamuxi for human consumption and agricultural use. Given that the reliance on surface water sources, especially rivers, and streams, for water supply has not changed in the last decade, it is crucial to study and predict the precipitation variations that directly affect water availability in surface channels.

The two factors that define the amount of water in surface streams are the contribution of the aquifers (Base Flow) and the contribution of precipitation. The latter is the factor that presents the most significant spatial and temporal variability due to the climatic oscillations of El Niño and La Niña. Because of this, it is essential to look into how precipitation changes in the Department of Boyacá over time and space. Precipitation behavior forecasting will help bring water balance models closer to reality and help make plans for managing water resources in the Department.

The climate change scenario has significantly impacted the analysis of historical precipitation records, disrupting typical seasonal patterns and leading to non-stationary behaviors such as bimodal precipitation influenced by exogenous factors like the greenhouse effect [13], [14], [15]. While precipitation time series are traditionally considered stationary with constant joint probability distributions, climate change introduces variability, requiring advanced techniques to account for this non-stationarity [16], [17]. The assumption of stationarity, known as the Statistical Homogeneity Hypothesis, no longer applies as greenhouse gas emissions have caused shifts in

atmospheric conditions, making conventional time series methods inadequate [18], [19]. Furthermore, studies show that precipitation exhibits statistical anomalies like heteroscedasticity and long-range dependence, further complicating the use of traditional models [16], [19]. Thus, advanced methodologies, such as downscaling and machine learning, are essential for capturing the complex dynamics of precipitation in a changing climate [16], [17].

Accurate precipitation forecasting is essential for decision-making in water resource management, particularly in regions such as Boyacá, Colombia. The data from the Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM) primarily come from ground-based sensor stations. While these stations provide useful climatic variables, the data are characterized by two major issues: low spatial uniformity and temporal gaps in the records. These challenges hinder the development of robust and precise prediction models, which are crucial for ensuring regional water security. Ensemble learning techniques such as boosting, bagging, and stacking are proposed to address these limitations. These methods combine multiple models to enhance prediction accuracy, particularly in scenarios where data are incomplete or heterogeneous. Research has shown that ensemble techniques are effective in similar environmental prediction applications, improving accuracy by mitigating overfitting and handling the non-linearity of the data [20], [21]. In particular, studies have demonstrated that heterogeneous fusion models, such as stacking, which combine algorithms like XGBoost, Random Forest, and Support Vector Machines (SVM), can achieve better results than single models. This approach would be highly beneficial in the context of precipitation prediction using incomplete or temporally inconsistent data. Additionally, the application of bagging and boosting techniques has been shown to reduce variance and improve the stability of predictive models in noisy or incomplete data environments [22], [23]. These findings suggest that incorporating ensemble methods into prediction models based on ground-based meteorological station data can offer an effective solution to spatial and temporal limitations, thus enhancing decision-making capabilities for water management in Boyacá.

In recent years, machine learning methodologies have been increasingly applied to analyze the information of hydrological variables in a branch within hydrology called hydro-computing. The main reason for using these machine learning procedures to analyze hydrological time series is their robustness in handling non-stationary data sets. Therefore, they have become the fundamental tools for performing this type of study [3], [24], [25]. Another essential reason is that machine learning procedures offer the possibility of finding non-linear relationships between variables and, therefore, in this case, can support the construction of models where additional information related to our variable of interest is integrated. For example, different remote sensors, such as MODIS, measure precipitation-related variables and can be integrated to improve the prediction of this crucial hydrological variable. The study [26] demonstrated the effectiveness of using cloud properties measured by MODIS, such as Cloud Optical Thickness (COT), Cloud Effective Radius (CER), and Cloud Water Path (CWP), to spatially downscale precipitation estimates from the IMERG product of the GPM mission. By applying a Multivariate Linear Regression (MLR) method and a residual correction algorithm, the spatial resolution of the precipitation estimates was increased from 0.1° to 0.01°. This technique resulted in significant improvements in prediction accuracy, with reductions in Root Mean Square Error (RMSE) by up to 75%, Normalized Root Mean Square Error (NRMSE) by up to 79%, and Percent Bias (PB) by up to 98%. Additionally, the correlation between the adjusted predictions and in situ rain gauge measurements improved by up to 20% compared to the initial predictions. These results highlight the potential of MODIS data to enhance precipitation estimates in complex terrains and varying climatic conditions, demonstrating that integrating these sensors into prediction models can provide more accurate and useful estimates for water resource management.

Another essential aspect is the fixed spatial resolution of the maps generated as output from predictive models, which limits their flexibility and practical utility. In particular, limited spatial resolution can compromise the ability of models to accurately predict extreme precipitation events that are also time-scale dependent, such as floods or droughts, which significantly impact public safety, water resource management, and infrastructure development. Improving spatial and temporal resolution is crucial for capturing local variations in precipitation patterns, enabling better preparedness and response to adverse weather events like floods, landslides, and droughts. Moreover, the ability to accurately and detail model these patterns is essential for planning and prioritizing government policies for development and risk mitigation. This type of problem is addressed with downscaling techniques, as seen in [27], [28], and [29], where Generative Adversarial Networks (GAN) are used to estimate the full probability distribution of spatial precipitation patterns with high temporal resolution, providing a more flexible approach tailored to practical needs in land management and public safety.

In conclusion, accurate precipitation prediction in mountainous regions like Boyacá is crucial for sustainable water resource management and mitigating risks from extreme weather events. The challenges posed by climate change, the complexities of mountainous areas, and the limitations of current predictive models highlight the need for advanced methodologies. By leveraging machine learning and downscaling techniques and integrating remote sensing data like MODIS, the model's spatial and temporal resolution can be enhanced, leading to the development of more effective tools for decision-making and ensuring water security in Boyacá.

3 CONCEPTUAL FRAMEWORK

3.1 Climatic variables

To establish the framework of the variables in this proposal, it is crucial to concentrate on the climatic variables identified as the state of the art and form the basis of this study. As a result, the climatic abbreviations identified and named throughout the document will be duly considered. Table 1 below summarizes these key abbreviations, which will be referenced consistently in this proposal's analysis and discussion sections.

Table 1. Climate abbreviations identified.

Abbreviation	Description
PPs	Precipitation Products
IDW	Inverse Distance Weighted
SPI	Standardized Precipitation Index
SPEI	Standardized Precipitation Evapotranspiration Index
CRU	Climate Research Unit
IOD	Indian Ocean Dipole
ENSO	El Niño-Southern Oscillation
PDO	Pacific Decadal Oscillation
SOI	Southern Oscillation Index
SST	Sea Surface Temperature

3.2 External factors of interaction with respect to precipitation

The prediction of precipitation in mountainous regions, such as Boyacá, is influenced by several external factors. These include:

3.2.1 Topography and elevation

Mountainous terrain has a direct impact on precipitation patterns. Precipitation generally increases with altitude, and windward slopes (facing the direction of the wind) tend to receive more rainfall than leeward slopes, due to orographic lift, where moist air rises, cools, and condenses into rain. The slope and aspect (direction of the slope) further contribute to the spatial distribution of precipitation [30], [31].

3.2.2 Local geography

The geographic location plays a crucial role in precipitation variability. In regions like Boyacá, moisture-laden winds, such as the trade winds, and proximity to large water bodies (oceans or lakes) affect the amount of moisture available for rainfall. Latitude, longitude, and the distance from these water sources are important determinants of precipitation distribution [31].

3.2.3 Temporal and seasonal variability

Precipitation in mountainous regions is not uniform throughout the year. Distinct wet and dry seasons, driven by seasonal changes such as monsoons, lead to concentrated periods of rainfall. For instance, in some regions, up to 70-80% of annual precipitation occurs within a few months [32], [30].

3.2.4 Climate Change and Atmospheric Variability

Global phenomena such as El Niño and La Niña significantly influence the quantity of precipitation in mountainous areas by altering regional moisture and temperature patterns. Additionally, climate change is expected to increase the frequency and intensity of extreme precipitation events, which complicates long-term prediction efforts [31], [33].

3.2.5 Data and Technological Limitations

The scarcity of meteorological stations in mountainous areas challenges collecting accurate data. While satellite-based precipitation products offer some solutions, they are often limited in detecting extreme rainfall events due to the spatial heterogeneity and relatively low resolution of satellite sensors [32].

3.3 Preprocessing techniques

In precipitation prediction models, data preprocessing is crucial for enhancing accuracy and reliability. A common technique involves decomposing the original precipitation time series into multiple subcomponents using methods like Empirical Mode Decomposition (EMD) or Wavelet Transform (WT). This approach isolates intrinsic patterns and trends within the data, facilitating more effective modeling of complex, non-linear behaviors. For instance, a study by [9] evaluated various preprocessing techniques, including differencing and spectral analysis, to improve stochastic rainfall forecast models, demonstrating significant improvements in prediction accuracy. Similarly, another research integrated data preprocessing with machine learning models for monthly precipitation forecasting, highlighting the importance of preprocessing in capturing different patterns across multiple stations [34]. Incorporating such preprocessing methods into precipitation prediction models can lead to more accurate and reliable forecasts, essential for effective water resource management and disaster preparedness.

In [35], improved results are obtained for a rainfall prediction model called Boosted Decision Tree Regression (BDTR) through the application of cross-validation and parameter tuning alongside the Autocorrelation Function (AFC). As highlighted in the "Models based on convolutional neural networks (CNN)" section, techniques like incremental Principal Component Analysis (i-PCA) outperform traditional models, particularly in terms of reducing the dimensionality of input variables to improve model performance in precipitation prediction, as well as in [36] is used the Combined Principal Component Analysis (CPCA) to reduce dimensionality and extract critical spatial patterns from the rainfall data. Furthermore, [37] presents the Singular Spectrum Analysis (SSA) as an effective technique, demonstrating superior results in extreme rainfall events due to its ability to handle non-linear components in time series data. In the study of [38], dimensionality reduction techniques, including Principal Component Analysis (PCA) and advanced nonlinear techniques such as isometric mapping and autoencoder, were applied, which significantly improved the generalization capability of time series models for multiscale water flow predictions, demonstrating that these techniques are effective in addressing nonlinearity in hydrological data.

In time series models, techniques such as PCA and Partial Autocorrelation Function (PACF), as discussed in [39], are commonly used to optimize the selection of input lags in data. This enhances correlation functions and subsequently improves the prediction of atmospheric pressures. However, it is essential to note that PCA has shown to be particularly effective in flat regions rather than mountainous areas [40]. In regions where complex terrains dominate, Wavelet Transform (WT) is a widely utilized technique for feature extraction at various time scales, helping capture short- and long-term fluctuations in precipitation patterns. Discrete Wavelet Transform (DWT) further enhances the treatment of non-linearities in the prediction models [9],

[41]. WT-based models have been proven to improve monthly precipitation predictions in mountainous terrains significantly [42].

Another preprocessing method, Empirical Mode Decomposition (EMD), is often applied to decompose non-linear and non-stationary signals. However, Ensemble Empirical Mode Decomposition (EEMD) offers an improved approach by reducing mode mixing, as it decomposes signals into intrinsic mode functions (IMFs) with the addition of white noise [43]. In [44], EEMD was utilized to break down IMF components and residuals, improving model predictions by addressing non-linear and non-seasonal factors within the original signal. More recently, Complementary Ensemble Empirical Mode Decomposition (CEEMD) has been developed to reduce the non-smoothness of time series further and has shown promise when combined with Long Short-Term Memory (LSTM) models to capture the complexity and non-linearity of hydrological factors [45].

Additionally, Kriging and Spatial Interpolation techniques are critical in regions where meteorological data may be sparse due to limited stations. Ordinary Kriging (OK) provides an accurate method to estimate precipitation depth by interpolating data from nearby stations. This method, validated through cross-validation [46], can be especially useful in predicting precipitation in mountainous areas where spatial variability is significant.

3.3.1 Time windows in prediction models

Prediction models are generally classified by the analysis time window, which can be short-term, medium-term, or long-term. However, these classifications depend heavily on data aggregation and the study's context. For example, in [47], the study examines rainfall data spanning 1951 to 2019 to analyze long-term drought patterns using the Standardized Precipitation Index (SPI). Even though the data spans multiple decades, the temporal analysis is performed at seasonal and annual scales to support short-term, medium-term, and long-term planning for drought mitigation. Similarly, in other contexts, a short-term forecast might involve predicting hourly rainfall for immediate flood response. In contrast, long-term forecasts could span years, depending on the granularity of the data and the needs of the analysis. Therefore, defining the short—or long—term prediction window depends on the research objectives and the scale at which the data is aggregated. The time windows in this study will be approximated as follows:

1. Hourly or minute-level forecasting, depending on the immediate needs of the analysis, as in [48].
2. Predictions covering days or weeks are helpful for agricultural and water resource planning.
3. Forecasts covering months, seasons, years, or decades, as demonstrated in [47] for long-term drought prediction.

3.3.2 Time series preprocessing

When time series data is not preprocessed or cleaned (for example, when there is missing data), the regularity of the time series is not maintained, which can lead to errors in the model's parameter fitting [49]. Data preprocessing improves its quality and, therefore, also the models generated from it [50]. According to [51], this phase not only addresses issues like missing data, but also handles outliers, dimensionality reduction, and normalization to ensure that the data meets the input requirements for AI models. Preprocessing, especially for time-series data, is essential to mitigate challenges related to heterogeneity, biases, and redundancies. Moreover, preprocessing techniques such as sensor fusion and data compression enhance the performance

of AI algorithms by reducing resource consumption and improving the overall quality of both input data and model output.

There are various methods for data preprocessing. Data cleaning helps remove noise and fix inconsistencies in the data. Data integration combines information from multiple sources into a unified storage system, like a data warehouse. Data reduction reduces the data size by methods such as aggregation, removing redundant features, or clustering. Data transformations, like normalization, adjust data values to fall within a smaller range, such as 0.0 to 1.0, improving the accuracy and efficiency of algorithms that rely on distance measurements. These techniques can be used together rather than separately. For instance, data cleaning may include transformations to correct inaccuracies, like standardizing the format of date fields [50].

3.3.3 Time lag in prediction models

Adding memory within neural networks can make them dynamic, increasing their capacity and complexity. This is particularly useful in rainfall prediction, where capturing temporal dependencies is critical for improving accuracy. Four main classes of dynamic models stand out according to [52], which have been explicitly applied to rainfall forecasting:

1. Tapped delay line models (TDNN): The network has explicitly available pass-by-pass inputs through a derived delay line. Internal time delay operators make the network dynamic.
2. Context or partial recurrent models: The network retains the previous output of the nodes instead of maintaining the past raw inputs; for example, the output of the hidden layer neurons of a feed-forward network that can be used as inputs to the network along with the true inputs.
3. Fully recurrent models: The network employs full feedback and interconnections between all nodes. The algorithms for training fully recurrent models are much more complex regarding time and storage requirements.
4. Recurrent Neural Network: Once the feedback connections are included, as in (2) and (3), a neural network becomes a recurrent neural network (RNN).

In this context, the Partial Autocorrelation Function (PACF) plays a critical role in optimizing the selection of time lags specifically for rainfall prediction models, as shown in [39], where the PACF helped improve multi-step rainfall forecasts by identifying the most relevant past values of rainfall and climate indices. This fine-tuning of time lags is essential for capturing the irregular nature of rainfall and improving predictive performance.

3.4 Data clustering and aggregation techniques

Data clustering and aggregation techniques are crucial in improving the accuracy of precipitation prediction models, especially in regions with complex terrain such as Boyacá. One essential method is temporal and spatial rainfall disaggregation. By using wavelet decomposition in combination with artificial neural networks (ANNs), researchers can disaggregate rainfall data into finer scales, which improves the capture of short- and long-term variations in precipitation. This combination enhances the performance of models, mainly when applied to regions characterized by heterogeneous rainfall patterns like mountainous areas [53], [54], [55].

Another promising approach is quantile-based spatial clustering, which groups precipitation data based on statistical quantiles. This method is particularly effective in regions that experience significant variability in rainfall distribution, as it captures extremes rather than just central tendencies. When combined with advanced statistical frameworks like Bayesian methods or

Gaussian Mixture Models (GMM), this clustering technique allows for a more nuanced analysis of spatial interdependence, making it ideal for mountainous areas with complex precipitation behaviors [56].

Cluster analysis is also widely applied to understand precipitation patterns. Methods like K-means, Ward's method, and Principal Component Analysis (PCA) help group similar rainfall patterns, making it easier to regionalize predictions in areas with high geographical variability. Non-hierarchical techniques, such as PCA-based clustering, have proven to be particularly effective in improving accuracy compared to hierarchical methods, especially when handling complex datasets with high spatial heterogeneity [57].

3.5 Machine learning models

It is essential to highlight the fundamental concepts of artificial intelligence (AI), which are crucial for discussing the conceptual framework and reviewing previous work. The abbreviations in Table 2, derived from the state-of-the-art research conducted for this doctoral proposal, are used consistently throughout this document:

Table 2. Abbreviations of identified models, techniques, and algorithms.

Model/Algorithm	Description
ACF	Autocorrelation Function
AL	Attention Layer
AM	Attention Mechanism
ANFIS	Neuron Fuzzy Inference System
ANN	Artificial Neural Network
ARIMA	Autoregressive Integrated Moving Average
ATWT	À Trous Wavelet Transform
BBO-ELM	Biogeography-based Extreme Learning Machine
BDTR	Boosted Decision Tree Regression
BLR	Bayesian Linear Regression
BLSTM	Bidirectional LSTM
BRFFNN	Bayesian regularized feed-forward neural network
BRNN	Bayesian regularized neural networks
CEEMD	Complementary ensemble empirical mode decomposition
CEEMDAN	fully adaptive noise ensemble empirical modal decomposition
CPCA	Combined Principal Component Analysis
CVWOTN	Cross-validation without tuning normalization
CVWTN	Cross-validation with tuning normalization
DFR	Decision Forest Regression
DSABASRNN	Dual-Stage Attention-Based Recurrent Neural Network
DWT	Discrete Wavelet Transform
ENN	Elman neural network
FBP	Facebook Prophet
FFNN	Feed-forward neural networks
FFOA	Fruit fly optimization algorithm
GA	Genetic Algorithm
GA-ELM	Genetic Algorithm-based Extreme Learning Machine

GAN	Generative Adversarial Networks
KNN	K-Nearest Neighbors
LGB	Light Gradient Boosting
LR	Linear Regression
LSSVR	Least-squares Support Vector Regression (LS-SVR)
LSTM	Long short-term memory
LSTM-ED	LSTM based encoder-decoder
LSTM-NN	Long short-term memory neural network
LWLR	Local Weighted Linear Regression
MA	Moving Average
MARS	Multivariate Adaptive Regression Splines
ML	Machine Learning
MLP	Multiple Layer Perceptron
MLSTM-AM	Multiscale LSTM with AM
NNR	Neural Network Regression
ORELM	Outlier Robust Extreme Learning Machine
PAC	Partial Autocorrelation
PACF	Partial Autocorrelation Function
poly-MARS	Multivariate adaptive polynomial splines
PSO-ELM	Particle Swarm Optimization (PSO)-based Extreme Learning Machine
RF	Random Forest
RF-R	Random Forest Regression
SARIMA	seasonal ARIMA
SDB	Stepwise decomposition-based
SMLM	Stacking machine learning models
SSA	Singular spectrum analysis
SSA-LSSVR	Singular Spectrum Analysis (SSA) - Least-squares Support Vector Regression (LS-SVR)
SSA-RF	Singular Spectrum Analysis (Data preprocessing method) Random Forest
SVR	Support Vector Regression
TVF-EMD	Time-varying filtering-based empirical mode decomposition
WBBO-ELM	Wavelet-based Biogeography-based Extreme Learning Machine
WDNN	Wavelet-based Deep Neural Network
WELM	Wavelet-based Extreme Learning Machine
WGA-ELM	Wavelet-based Genetic Algorithm (GA)-based Extreme Learning Machine
WORELM	Wavelet-Outlier Robust Extreme Learning Machine
WPSO-ELM	Wavelet and Particle Swarm Optimization (PSO)-based Extreme Learning Machine
WT	Wavelet Transform
WT-ELM	WT Extreme Learning Machines
WT-FFBP-NN	Wavelet Transform Feed-Forward Back-Propagation Neural Network
XGB	Extreme Gradient Boosting
XGR	Gradient Boosting Regression

Given the problem of water supply worldwide and specifically in the municipalities of Boyacá, in Colombia, due to the phenomena of El Niño and La Niña, changes in the climatology of the planet and global warming, methodologies for predicting the amount of water are necessary to direct government development plans in this regard. Currently, there is a branch of artificial intelligence

(AI) that is working quite well to solve prediction problems and this is called Machine Learning, which is defined as "an analytical method that allows a system, by itself -without human intervention and in an automated way-, to learn to discover patterns, trends, relationships in the data and thanks to this knowledge, in each interaction with new information, better perspectives are offered" [58].

Machine learning techniques have been applied extensively to various aspects of water resource management, including forecasting water demand and predicting rainfall behavior. For example, Support Vector Machines (SVM) optimized with genetic algorithms (GA) have shown strong results in forecasting urban and ecological water demand by addressing data scarcity and improving accuracy in predictions [59]. Although these methods have been primarily used for water demand, similar machine learning techniques have also been successfully applied in rainfall prediction. Studies using SVM and ensemble methods in precipitation prediction have demonstrated significant improvements in accuracy by optimizing data-driven models for complex environments such as mountainous regions [60] [61]. These machine learning applications are crucial in refining water resource management by enhancing the ability to predict water supply and rainfall variability. Figure 1 shows the relationships between Shallow Learning (SL) and Deep Learning (DL), with all the internal subtypes within each category, some of which are relevant to this proposal.

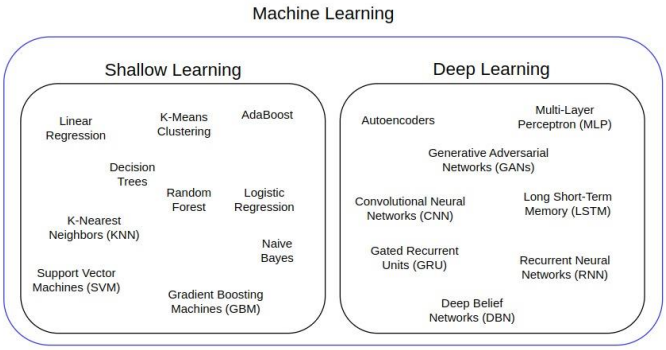

Figure 1. Relationship between shallow learning and deep learning.

As mentioned in the book by [62] and in [63]*, "The quintessential example of a deep learning model is the direct feed-forward deep network or multilayer perceptron (MLP). A multilayer perceptron is just a mathematical function that maps a set of input values to output values. The function is formed by composing many simpler functions."* A technique that has gained considerable traction in recent times is deep learning, which, as we saw in Figure 2 (b), is part of neural networks. Deep learning is distinct from machine learning in that it enables computers to build complex concepts from simpler ones, a process that is particularly useful in real-world scenarios. It Is important to note that the extraction of features in deep learning is done within the model itself and is not a job done by a person; also, at the level of amount of data deep learning techniques scale more in performance over machine learning techniques, as shown in Figure 2.

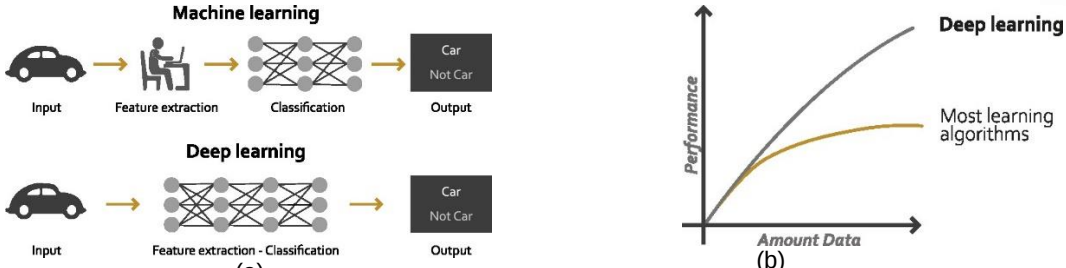

Figure 2. Clear distinctions between machine learning and deep learning. Adapted from: MIT Course: Deep Learning. Understanding these differences is crucial for a comprehensive grasp of the field.
BASICS: INTRODUCTION AND OVERVIEW, LEX FRIDMAN - YOUTUBE.

3.5.1 Models based on convolutional neural networks (CNN)

Neural networks, a key component of deep learning, encompass a variety of architectures, each suited to different types of tasks. The most prominent are convolutional neural networks (CNNs) and recurrent neural networks (RNNs) [64], [65], [66]. CNNs have been particularly successful in tasks involving grid-like data structures, such as images, and played a crucial role in the development of deep learning, with early successes in applications like optical character recognition (OCR) by AT&T in 1990 [62]. This historical context underscores the progress and potential of deep learning in precipitation prediction. For example, a study by [67] developed a rainfall occurrence prediction model using a convolutional neural network (CNN), a prominent machine learning architecture in image recognition. This study created a spatiotemporal data array from time series data of atmospheric variables collected from multiple ground observation sites. This array was then used as input for the CNN, which was trained to classify whether it would rain within the next 30 minutes. The model achieved a detection ratio between 64% and 76% for predictions in three different cities in Japan, demonstrating the utility of CNNs in short-term rainfall forecasting. However, issues like the high false alarm ratio were noted for further refinement.

Another study by [68] applied a deep CNN to predict monthly rainfall for a specific location in eastern Australia. This approach was compared with the Australian Community Climate and Earth-System Simulator-Seasonal Prediction System (ACCESS-S1) and a conventional Multi-Layered Perceptron (MLP). The CNN outperformed both models, showing better results in terms of mean absolute error, root mean square error (RMSE), Pearson correlation, and Nash-Sutcliffe efficiency coefficient. The study highlighted the effectiveness of CNNs, particularly in months with higher annual rainfall averages. The CNN architecture included common layers such as convolutional, clustering, and fully connected layers, which were vital to its performance. Figure 3 illustrates the distribution of these connected layers within the CNN architecture used in the study.

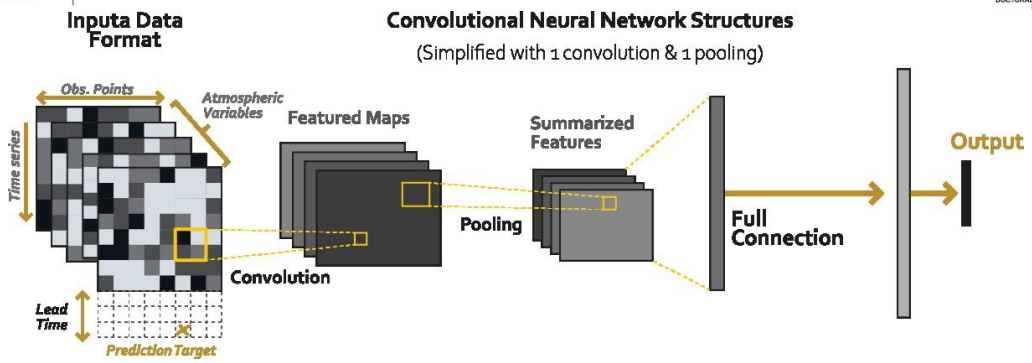

Figure 3. Schematic of the input data matrix and structure of the CNN.
ADAPTED FROM [67].

Figure 3 illustrates the data structure in a convolutional neural network (CNN) for rainfall prediction. The horizontal axis represents observation points over time, the vertical axis shows the time series, and the depth (Z) axis represents the feature space containing various atmospheric variables. The CNN processes this data through convolution and pooling operations, extracting and summarizing critical features before passing them to the fully connected layer to generate the final prediction.

The input data is typically in three dimensions at the mathematical processing level, with L x L pixels and K image channels, usually RGB. The filter also has a 3-dimensional array. The feature extraction can be estimated, and the process, known as the convolution process, is a function of F (·). This process is applied to the entire input range with vertical and horizontal displacement for every specific step, typically one pixel. The convolutional process results in a feature map (L - H + 1) × (L - H + 1) × M, where M is the number of convolutional processes. The size can be controlled $L \times L \times M$ by adding values of around zero from the input data (zero padding). The activation function that converts the convolution results, the most common in convolutional neural networks, is the activation function, which is usually the rectified linear unit (ReLU).

In [69], convolutional neural networks (CNNs) were applied for monthly flow forecasting, demonstrating better performance, lower error, and more excellent stability compared to artificial neural networks (ANNs) and extreme learning machines (ELM). The CNN's ability to automatically extract critical features from various hydrological and atmospheric variables such as rainfall, streamflow, and atmospheric circulation factors through its convolution-pooling mechanism was a key factor in achieving these results. For flood prediction based on monthly rainfall intensity, the study in [70], [38], [69], utilized a dimensionality reduction technique called improved Principal Component Analysis (i-PCA), effectively reducing the data's complexity while preserving relevant features. This method improved the accuracy of flood prediction to 94.24%, outperforming models like LSTM (84.74%) and Explainable Artificial Intelligence (EAI) (86.19%). The i-PCA technique was crucial in determining the most essential features by eliminating irrelevant variables, thereby optimizing the input data for the CNN. This feature selection process allowed the model to focus on the variables that most strongly influenced flood occurrence, making the forecasting more efficient and accurate. Using these optimized features, the model achieved higher accuracy with fewer computational resources, highlighting the importance of careful feature selection and dimensionality reduction in complex hydrological models.

A comparison of machine learning (ML) algorithms for flow forecasting is made in [71], where different models such as CNN, convolutional long short-term memory (ConvLSTM), deep neural

network (DNN), long and short-term memory (LSTM), extreme learning machines (ELM), Multivariate Adaptive Regression Spline (MARS), Extreme Gradient Boosting (XGBoost), and Decision Trees were evaluated for short-term flow prediction. CNN and ConvLSTM stood out at high flow rates, likely due to their ability to capture spatial-temporal patterns effectively. These architectures excel because they can model the complex relationships between multiple variables over time and space, which is critical in inflow forecasting. The challenge in flow prediction is not just the sophistication of the algorithm but instead selecting models that align with the nature of the problem, such as temporal dependencies and sudden flow changes. In this context, CNN and ConvLSTM provide a good balance between capturing local dependencies and managing long-term temporal variations.

3.5.2 Transformer neural networks

Transformer neural networks have consistently outperformed traditional models like LSTMs and RNNs in sequence-based tasks due to their capacity to capture long-range dependencies and process data in parallel, making them both accurate and efficient. Studies have shown that combining Transformer blocks with other models enhances performance in tasks such as image captioning and precipitation forecasting. For example, the Temporal Fusion Transformer (TFT) model demonstrated substantial improvements in predicting extreme seasonal precipitation up to six months ahead, outperforming traditional climatology and ensemble forecasts. Moreover, Transformer-based models such as NowcastingGPT have shown versatility and effectiveness in precipitation nowcasting. These models are up-and-coming for applications in mountainous regions where complex terrain and microclimatic variations challenge traditional forecasting methods. However, the success of Transformer models in such environments relies heavily on data availability and quality, as sufficient high-resolution data is crucial for capturing the spatial and temporal complexity of precipitation in these regions [72].

3.5.3 The impact of deep learning on precipitation prediction models

Currently, among the most relevant precipitation prediction techniques, those that fall into the category of artificial neural networks are highlighted. Studies such as [73] utilize deep neural networks to estimate rainfall from remote sensor data. A study that combines convolutional neural networks with LSTM can be found in [74] in the article titled *Spatiotemporal Convolutional LSTM with Attention Mechanism for Monthly Rainfall Prediction*. This study uses climatological data from CHIRPS, which has a spatiotemporal configuration with a spatial resolution of 0.05° x 0.05°, and temporal resolutions such as daily or monthly. Figure 4 and Figure 5 illustrates the amount of temporal and spatial data used in the model and how precipitation data is processed through spatiotemporal windows.

The use of convolutional networks for sequence-to-sequence spatiotemporal models, known as STConvS2S, for climate prediction is discussed in [75]. It is a 3D architecture-based convolutional neural network (CNN) built as an end-to-end trainable model that satisfies the causal constraint and is not limited by the length of the input sequence for model output. It uses temporal and spatial blocks for sequence handling.

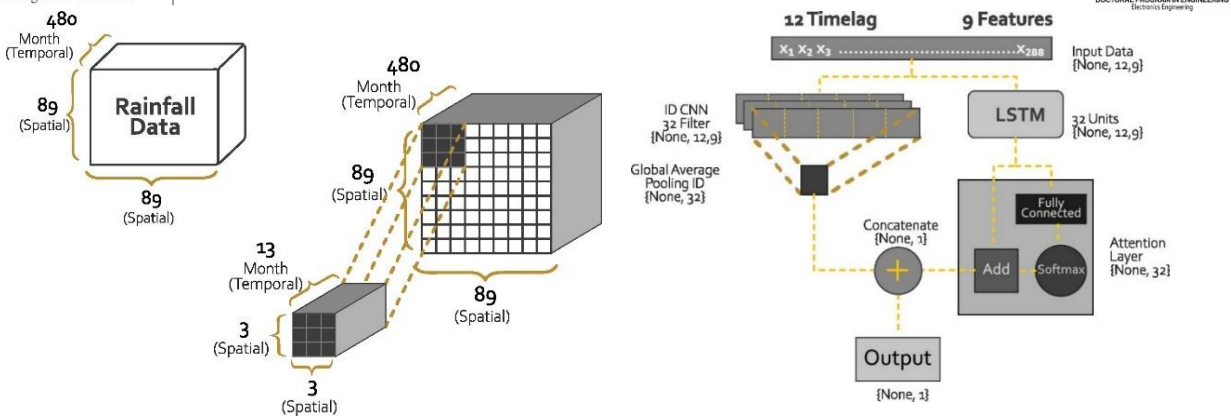

Figure 4. Spatiotemporal data processing and LSTM model.
ADAPTED FROM [74].

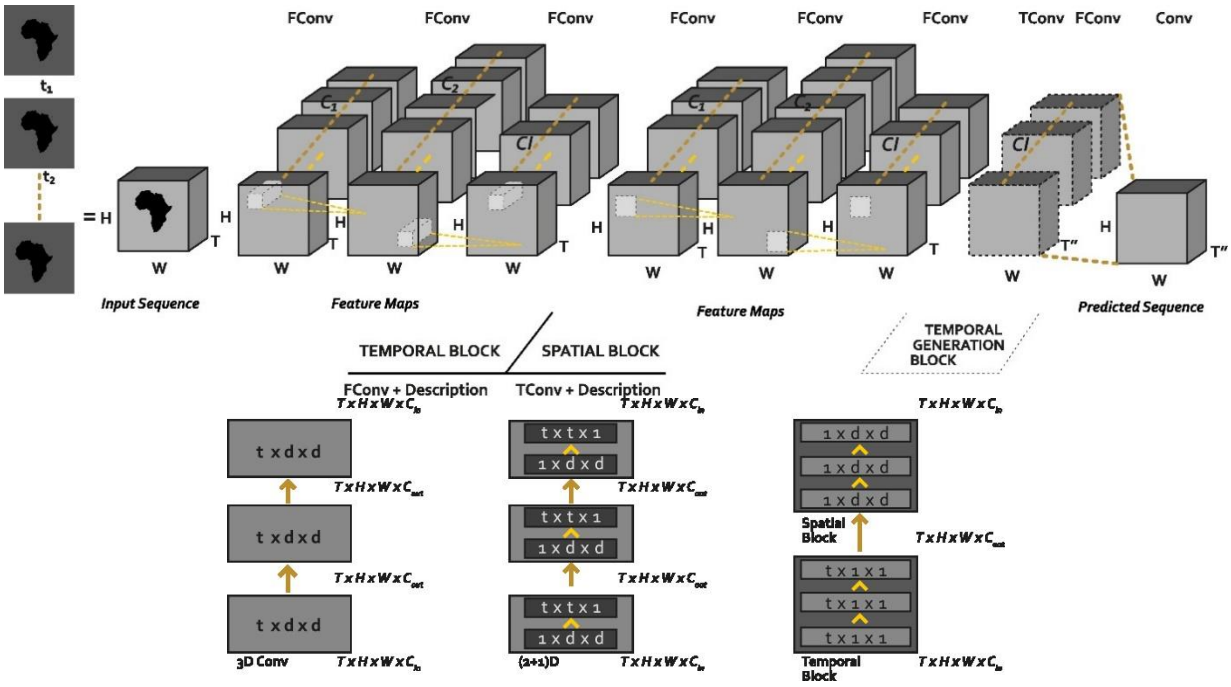

Figure 5. Sequence-by-sequence spatiotemporal convolutional model (STConvS2S).
ADAPTED FROM [75].

With the description of some techniques previously seen at the machine learning level, a description of some previous work on neural networks is provided as a basis for the state of the art and ordered from the oldest to the most recent date:

1. *On the Predictability of Rainfall in Kerala* [76], an adaptive basis function neural network (ABFNN) is used, a variation of the backpropagation (BP) algorithm, with ABFNNN and BPNNN performing better than the former.
2. *A Soft Computing Technique in rainfall forecasting* [77] uses a feed-forward neural network (FFNN), also known as multilayer perceptron (MLP), which improves the prediction error (PE) from 18.3% in the case of a conventional persistence model to 10.2%.
3. *Feed forward Artificial Neural Network model to predict the average summer-monsoon rainfall in India* [78], in this paper, an artificial neural network model with back propagation

learning is implemented to predict the average summer monsoon rainfall in India. It is compared against the persistence prediction model and multiple linear regression (MLR), which have lower prediction errors than artificial neural networks with backpropagation.

4. *Optimized scenario for rainfall forecasting using genetic algorithm coupled with artificial neural network* [79], where artificial neural networks with backpropagation (BP) associated with a genetic algorithm (GA) are used to train and optimize the network.
5. *Rainfall forecasting models using focused time-delay neural networks* [80] use trial-and-error calculated lag times depending on the performance obtained for the artificial neural network models.
6. *Prediction of rainfall time series using modular artificial neural networks coupled with data-preprocessing techniques* [81]. The study treats the input data in three aspects: model inputs, modeling methods and data preprocessing techniques. Uses modular artificial neural networks (MANN) compared with models such as artificial neural networks (ANN), k-nearest neighbors (KNN) and linear regression (LR). There are three coupled preprocessing techniques: average moving (MA), principal component analysis (PCA) and singular spectrum analysis (SSA). The best results are MANN with coupled SSA.
7. *Prediction of rainfall time series using modular soft computing methods* [82]; the modular model is composed of support vector regression (SVR) and artificial neural networks (ANN), which are used to choose the MA and SSA preprocessing methods. The modular models for preprocessing of burial data are divided into three subsets: low, medium and high levels; according to the magnitudes, the two SVR models process the medium and high levels, and both ANN and SVR were used for training and predicting low levels.
8. *Time Series Analysis of Forecasting Indian Rainfall* [83] uses input, hidden and output layers, with one neuron, four and one, respectively, using backpropagation.
9. *Optimized Artificial Neural Networks-Based Methods for Statistical Downscaling of Gridded Precipitation Data* [84], Artificial neural networks (ANN) are used as a technique for downscaling and evaluating a satellite precipitation estimation product (SPE) with a model with an output variable such as precipitation and inputs temperature, MODIS optical cloud and microphysical variables. Three types of algorithms were used to improve the performance of ANNs, which were particle swarm optimization (PSO), imperialistic competitive algorithm (ICA), and genetic algorithm (GA) to examine the efficiency of the networks, the downscaled product was evaluated using 54 rain gauges at a daily scale, the precipitation prediction was more sensitive to cloud optical thickness (COT). Residual correction algorithms significantly improved the accuracy of the final downscaled satellite precipitation. This suggests that combining different inputs to the models improves the prediction quality and that using various algorithms for ANN performance improvement was instrumental in improving the prediction accuracies.
10. *Estimation of rainfall based on MODIS using neural networks.* Since the flooding above problem is directly related to rainfall, the issues of monitoring at specific points versus the study with satellite images of MODIS clouds are related to the issues of the flooding problem [85] contrasted with ground rainfall data, which allows better estimates of rainfall rates and obtain a better estimation effect error accuracy, all this using artificial neural network techniques (ANN), backpropagation algorithm (BPA), genetic algorithm (GA) and remote sensing recovery, which shows a combination of different methods of data origin that are related in the predictive models using the mentioned techniques to improve the estimates.

Table 3 provides a structured summary of scientific studies that use machine learning models for precipitation prediction. Each row corresponds to a different study, organizing key information about the variables used to train the models, the target variables, the types of

machine learning models implemented, and any specific environmental or conditional factors considered in the experiments. In this way, the table allows for a quick comparison of different approaches and techniques in precipitation prediction using machine learning. The input variables are mostly daily or monthly precipitation amounts. Some models include temperature, sunshine, and other environmental data.

Table 3. Comparison of the models.

Ref	Training variables	Target	ML Model	Specific Conditions
[76]	Twelve input nodes, each with the corresponding month's total rainfall.	Amount of rain to be expected in the same month of the fifth year.	Adaptive Basis Function Neural Network (ABFNN).	Factors such as the El-Nino southern oscillations (ENSO) resulting from the pressure oscillations between the tropical Indian Ocean and the tropical Pacific Ocean.
[77]	Precipitation.	The average monsoon rainfall of the year (y+1).	Multilayer Perceptron (MLP).	Focused on the summer monsoon in India.
[78]	Monthly precipitation during the summer (June, July, and August).	Average summer-monsoon rainfall of the year (n+1).	A backpropagation neural network with three layers uses a sigmoidal activation function.	Summer monsoon in India
[79]	Rain gauge station data, primarily accumulated and discrete precipitation measurements, and other factors such as temporal delays in rainfall.	Amount of precipitation at a specific station.	Multilayer feed-forward neural network (MFNN).	The study was conducted at the Parramatta River catchment in Sydney, Australia. Data from 14 rain gauge stations in an urban environment were used over four years (1996-2000) with 5-minute recording intervals.
[80]	Daily precipitation was transformed into monthly, quarterly, semi-annual, and annual series to train the different models.	Future precipitation amount in millimeters.	Focused Time-Delay Neural Network (FTDNN).	The daily rainfall dataset was obtained from the Subang Meteorological Station in Malaysia and covers the period from January 1980 to May 2009.
[81]	Historical values of the precipitation series from previous days or months.	Future precipitation at a specified prediction horizon (e.g., one day or one month in the future).	The modular artificial neural network (MANN) is compared with three benchmark models: artificial neural network (ANN), K-nearest neighbors (K-NN), and linear regression (LR).	Two daily mean rainfall series from Daning and Zhenshui river basins of China, and two monthly mean rainfall series from India and Zhongxian of China, are analyzed.
[82]	Four rainfall time series, consisting of two monthly and two daily rainfalls from different regions, were utilized to evaluate modular models.	Prediction of future precipitation 1, 2, and 3 days in advance.	It evaluates various soft computing models, including artificial neural networks (ANN) and support vector regression (SVR).	"Simulating the response using conventional approaches in modeling rainfall time series is far from a trivial task since the hydrologic processes are complex and involve various inherently complex predictors such as geomorphologic and climatic factors."
[83]	The dataset used for training contains monthly precipitation information normalized in the range [0.2, 0.8].	Future precipitation.	Use a multi-layer feed-forward Artificial Neural Network.	No reference is made to including other weather conditions in the analysis.
[84]	Temperature (ERA5 product), cloud optical thickness (COT), cloud effective radius (CER), and cloud water path (CWP) were obtained from the MODIS sensor.	Precipitation.	Imperialist Competitive Algorithm (ICA), Particle Swarm Optimization (PSO), Genetic Algorithm (GA)	The study is located in Austria's easternmost extension of the Alps, characterized by a warm temperate zone with less than 600 mm of annual precipitation.

[86]	The model inputs include precipitation anomalies and climate indices at different time scales. The components of the decomposed series are used as inputs at each time scale.	Monthly precipitation prediction is obtained from the sum of the forecasted subseries.	Four models were compared: MLSTM-AM, MLSTM, LSTM, and multiple linear regression (MLR).	The study is conducted in the Yangtze River Basin, which has a subtropical monsoon climate and exhibits an uneven spatial and temporal distribution of precipitation.

The prediction of precipitation is a complex challenge due to its stochastic, non-linear, and non-stationary nature. These characteristics complicate accurate modeling of rainfall patterns, as atmospheric conditions can vary unpredictably across different temporal and spatial scales. Advanced models that integrate Artificial Neural Networks (ANNs) with deep learning and hybrid techniques have been developed to address these complexities. For instance, the PERSIANN model (Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks) leverages satellite data and neural networks to estimate real-time precipitation, illustrating the effectiveness of ANNs in this context [87]. Recent studies have also implemented Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to capture complex spatial and temporal patterns in precipitation forecasting, significantly improving predictive accuracy [88]. These advanced approaches enable better adaptation to the inherent variability of meteorological data, providing more robust and precise solutions for precipitation prediction.

3.5.4 Long- and short-term memory (LSTM) and GRU models

In practice, working solely with Simple Recurrent Neural Networks (SimpleRNN) is uncommon due to a problem known as the vanishing gradient. This issue arises when the network struggles to learn long-term dependencies, as the gradient becomes too small, causing the learning process to stall. This effect is similar to what happens in non-recurrent networks, like deep feed-forward networks, where adding more layers eventually makes the network untrainable [89]. This problem was addressed in [90], highlighting three key challenges: the system's ability to store information over long periods, resistance to noise, and trainability through gradient descent. To overcome these challenges, Hochreiter and Schmidhuber introduced the Long Short-Term Memory (LSTM) (see Figure 6) model [91], specifically designed to handle the vanishing gradient, allowing networks to retain information over extended time intervals.

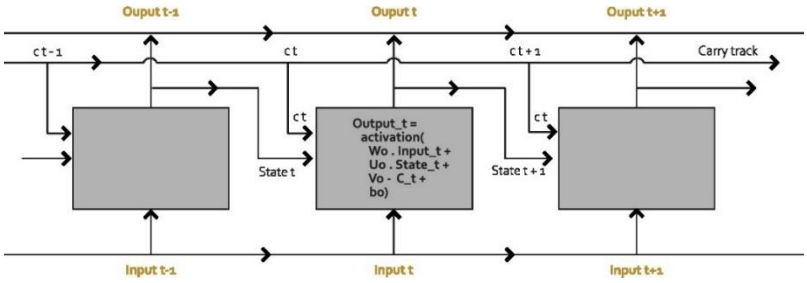

Figure 6. LSMT.
ADAPTED FROM [89].

The iterations, strengths, and weaknesses are outlined in a study on the origins of Long Short-Term Memory networks (LSTMs), which are a type of Recurrent Neural Network (RNN) [92]. LSTMs capture long-term dependencies, making them ideal for sequence prediction tasks. They can stack LSTM layers for increased capacity and utilize multidimensional structures for greater flexibility. In [63], LSTMs are applied to water quality prediction, with an input, hidden, and output layer structure, allowing comparison with other models like Back Propagation Neural Network

(BPNN) and Online Sequential Extreme Learning Machine (OS-ELM). Here, LSTM models achieve superior performance based on the RMSE metric by iterating values for time step selection, the number of hidden layers, and epochs to predict dissolved oxygen (DO) and total phosphorus (TP) values.

The Gated Recurrent Unit (GRU) is like the Long Short-Term Memory (LSTM) model but with a simpler architecture. Proposed initially [93], GRU was designed to adaptively capture dependencies of varying time scales, particularly in tasks involving sequential data such as speech recognition and natural language processing (NLP). Unlike LSTM, GRU lacks a separate memory cell, simplifying its structure using only two gates: the reset and update gates, as shown in Figure 7. This more straightforward design allows GRU to offer advantages in computational efficiency, often converging faster and requiring fewer resources than LSTM, while maintaining comparable performance in many sequence modeling tasks, as demonstrated by empirical evaluations in [93]. In the context of rainfall prediction or other spatiotemporal forecasting tasks, GRU's ability to efficiently model sequential dependencies with fewer computational demands makes it an attractive option when balancing accuracy and resource constraints is crucial.

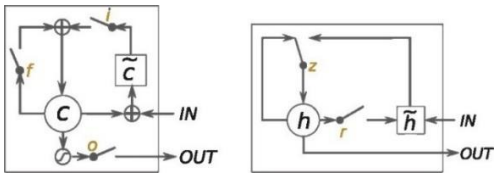

Figure 7. Comparison between LSTM and GRU, (a) i, f and o are the input, forget and output gates, respectively. c and c˜ denote the memory cell and the contents of the new memory cell. (b) r and z are the reset and update gates, and h and h˜ are the activation and candidate activation.
ADAPTED FROM [94].

Regarding the monthly rainfall predictions, some variations of LSTM [86] with attention mechanism (AM) together called multiscale long short-term memory (MLSTM) are made; the above three are compared together with multiple linear regression (MLR) findings becoming the best MLSTM-AM, MLSTM, LSTM, and MLR successively.

3.5.5 Hybrid predictive models

Hybrid models are generally standard models combined with a technique that improves prediction accuracy, thus making them more robust. Traditional hybrid models separately simulate the linear and nonlinear components of the precipitation series to achieve improved precipitation predictions [95].

3.5.5.1 Hybrid precipitation prediction models

Some complementary techniques to the traditional models used to improve precipitation prediction are listed below:

- Variational mode decomposition (VMD) with extreme machine learning (ELM) forming the VMD-ELM model that also attacks the nonlinearity of the phenomenon [96], as shown in Figure 8.
- At [97], the use of hybrid models based artificial neural networks (ANN) and support vector regression (SVR) in the improvement of drought prediction that has stochastic behaviors and high nonlinear effects with the use of two optimization algorithms called Particle Swarm Optimization (PSO) and Response Surface Method (RSM), as we can see in this case, these are hybrid models by combining standard models with optimization algorithms.

In this case, hybrid models SVR-RSM and SVR-PSO were compared, where the SVR-RSM model stands out above the others in accuracy and trend of predictions.

In this way, traditional models are combined with algorithms that form robust models to improve prediction accuracy. Many of these use deep learning techniques, such as ANN, LSTM, and CNN.

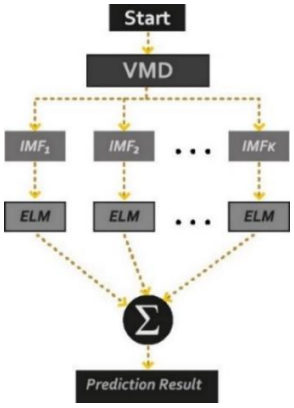

Figure 8. Flow diagram of VMD-ELM prediction model.
ADAPTED FROM [96].

3.5.6 Model coupling in precipitation forecasting

A particular case of coupling model is seen in [98] where using a Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) algorithm that makes a decomposition of the precipitation time series to obtain n components of the IMF to then perform the predictions with the support of the hybrid PSO-ELM model, Particle Swarm Optimization (PSO) can be used to optimize the input weights and thresholds of the Extreme Learning Machine (ELM) and finally perform a reconstruction of the components to obtain a final prediction value as shown in Figure 9.

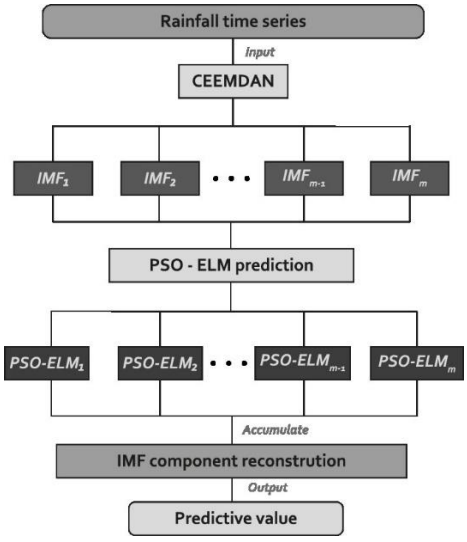

Figure 9. Prediction process of the coupled model CEEMDAN-PSO-ELM.
ADAPTED FROM [98].

3.5.6.1 Stacked rainfall prediction models

At [25], opportunities are seen in the variability of rainfall series, identifying critical factors for improving predictive capability and examining ML models such as RNN. Base models such as KNN, XGB, SVR, and MLP were used in this case. The weights of each model were obtained via quadratic programming and evaluated with metrics: R^2, RMSE and MAE.

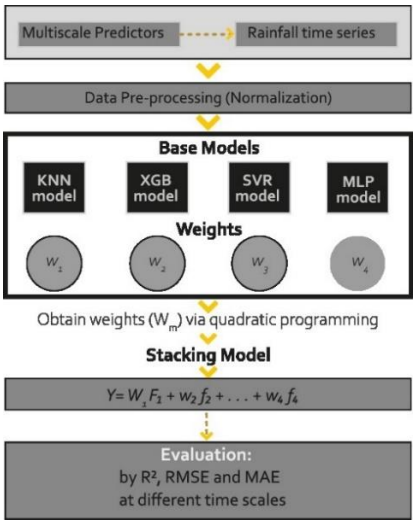

Figure 10. Flowchart of the stacking-based methodology.
ADAPTED FROM [25]

Traditional neural networks cannot process non-stationary or high-frequency abrupt data and the prediction error is usually between 5% and 20% [3]. Using a coupling model to reduce the non-stationarity of the original series has become a new way to increase the accuracy of rainfall prediction. This can be achieved by a non-stationary signal processing method called empirical mode decomposition, the complementary ensemble empirical modal decomposition model (CEEMD), which is adaptive by EMD. The coupling is against neural network models, CEEMD, ELM and optimization algorithms such as Swarm and Fruit Flay Optimization Algorithm (FFOA) to improve accuracy further. CEEMD decomposes the precipitation time series into several intrinsic modal components (IMF Components), a hidden layer feedforward neural network is constructed for each IMF component, and ELM is used for simulation and prediction. Another algorithm used is Drosophila, which optimizes the accumulation of coefficients between IMF components to predict the closest possible value to the true value and improve the prediction accuracy.

3.5.7 Time series models and their relationship with deep learning

A time series is a sequence of observations taken at successive time points, usually at uniform intervals [99]. This data type is fundamental in predictive analytics since it allows modeling patterns over time, such as trends and seasonality, which are essential for making predictions based on historical data [100].

Time series models (TS models) have gained significant importance across various fields, such as medicine, human activity recognition, acoustic scene classification, and cyber security. Traditionally, time series data have been modeled using statistical approaches, like Autoregressive models (AR), Autoregressive Moving Average (ARMA), Autoregressive Integrated Moving Average (ARIMA), and their seasonal variants (SARIMA), which have been highly

effective for capturing linear patterns in the data. However, these methods face limitations when dealing with non-linear, non-stationary, or highly complex time series data, as is often the case in rainfall and other environmental phenomena. In response to these challenges, machine learning-based methods have been increasingly employed for time series prediction and classification tasks. Figure 6 illustrates a comprehensive Deep Learning Framework specifically designed for time series classification (TSC), highlighting the integration of deep learning architectures for handling temporal data [101].

Among machine learning approaches, dynamic time warping (DTW) coupled with nearest neighbor (NN) classifiers has been widely used as a baseline technique for time series classification. However, recent advances have shown that ensemble models, such as random forests and Support Vector Machines (SVM), combined with methods like DTW, can significantly improve performance in terms of accuracy and robustness. Furthermore, as depicted in Figure 11, deep learning models, particularly Deep Convolutional Neural Networks (DCNNs), have demonstrated superior results for time series classification, especially in fields like computer vision, natural language processing, and speech recognition, where temporal data play a crucial role [101].

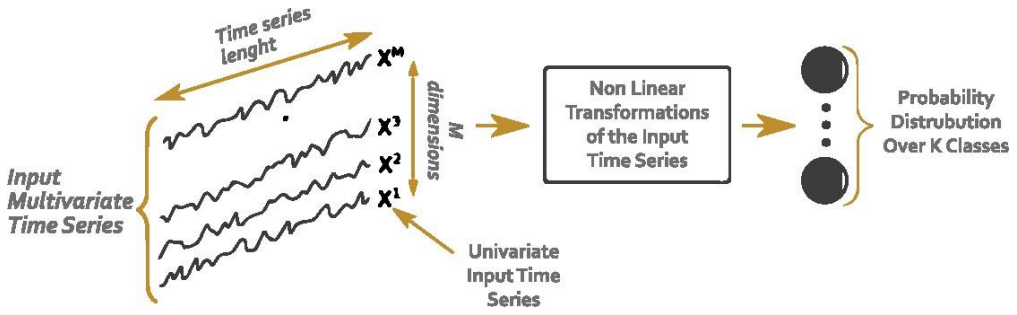

Figure 11. Deep Learning Framework for time series classification.
ADAPTED FROM [101].

Currently, the most prominent algorithms for time series classification include:

1. Multi-Layer Perceptions (MLP)
2. Convolutional Neural Networks (CNN)
3. Echo State Networks (ESN) use recurrent neural networks. They are rarely applied for time series classification for three reasons: one output for each element of the time series, vanishing gradient problem due to long time series training, and difficulty in training and parallelizing.

Recent innovations in deep learning for TSC, such as hierarchical voting systems like HIVE-COTE, have further enhanced the state-of-the-art performance in this area. These models integrate both generative and discriminative approaches, as shown in Figure 12, providing a more comprehensive framework for time series prediction by leveraging various transformations and ensemble techniques.

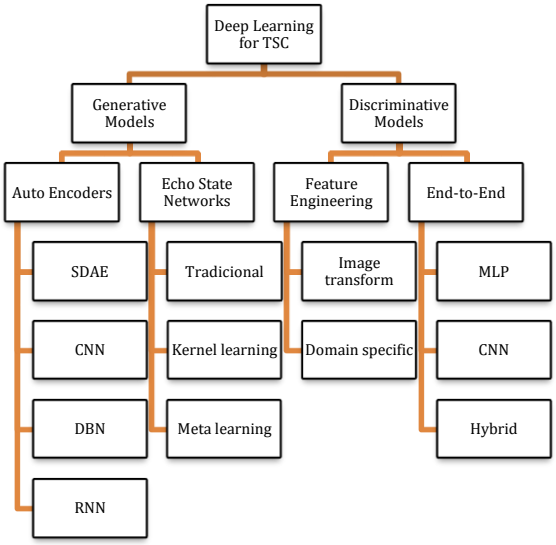

Figure 12. An overview of different deep learning approaches for time series classification. TAKEN FROM [101].

3.5.7.1 Models type white-box machine learning algorithms for time series prediction

3.5.7.1.1 Dynamical Systems using Symbolic Regression

It involves using machine learning to discover mathematical expressions that represent the evolution of complex systems over time. Symbolic regression, as implemented in models like ODEFormer, constructs interpretable models without predefined assumptions by directly inferring differential equations from data, even when data is noisy or irregularly sampled. This approach aligns with time series analysis by modeling temporal dependencies, allowing predictions based on observed patterns and enhancing insights into the mechanisms underlying time-dependent systems [102] [103].

3.5.7.1.2 EXA-GP: Unifying Graph-Based Genetic Programming and Neuroevolution

EXA-GP (Evolutionary eXploration of Augmenting Genetic Programs) is a graph-based genetic programming (GGP) algorithm developed by adapting the EXAMM neuroevolution framework to use genetic programming operations instead of neural and memory cell structures. By leveraging EXAMM's optimizations—like distributed execution, island-based populations, and Lamarckian inheritance—EXA-GP can evolve genetic programs with the same predictive accuracy as recurrent neural networks (RNNs) but with greater interpretability and reduced computational complexity. These results suggest that GP operations could be more effective than recurrent memory cells for time series forecasting, as EXA-GP demonstrates superior performance and explainability over other advanced GP and RNN models in benchmark tests.

3.5.8 Blended learning techniques

Prediction models can be used with ensemble methods, as demonstrated in the state-of-the-art study in [104], which leads to improved accuracy and performance compared to using a single model. In [105], three main classes of ensemble learning techniques are described and illustrated in Figure 13:

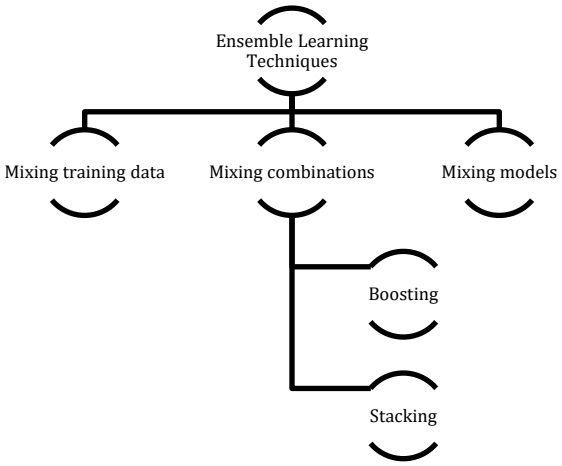

Figure 13. Types of set learning techniques.

The first technique is training data blending or Bagging, which consists of dividing the training data into different subsets so that each classifier captures different behaviors and then combines these classifiers to obtain better accuracy than a large classifier as shown in Figure 14.

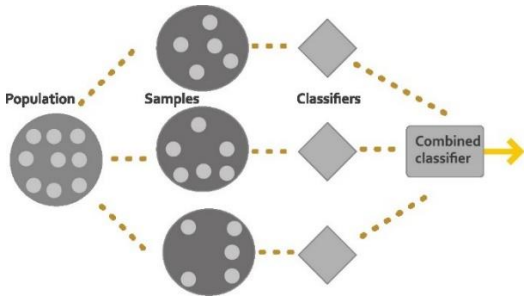

Figure 14. Mixing of training data - Bagging.
ADAPTED FROM [105].

The second technique, called combination mixing, has two subcategories: boosting and stacking, which will be discussed below. A sub-collection of ML trainees is used, trained with a particular subset of training objects; if the model trainee is underperforming, its weaknesses could be emphasized to improve. Different models are trained and stacked on each other, and multiple models are trained to obtain prediction/output apprentices. The first layer is the base, and the second is the meta-learner as shown in Figure 15.

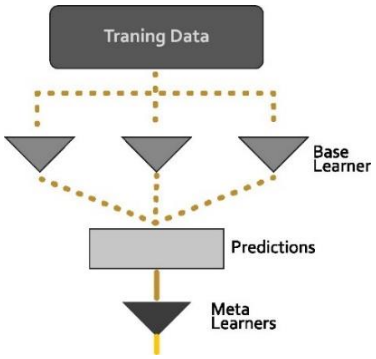

Figure 15. Stacking Learning.

ADAPTED FROM [105].

The third type of ensemble learning method consists of varying the models to obtain better results than with a single model; some models are strong in certain aspects and not in others, and several models complement each other to strengthen the overall model. This method can also apply to model variations, such as configurations or hyperparameters, instead of relying on a single training run, which can obtain greater accuracy and less bias, as shown in Figure 16.

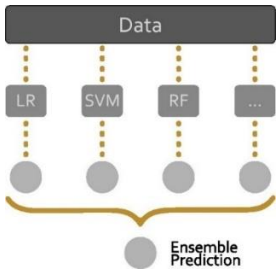

Figure 16. Variation of models.
TAKEN FROM [105].

4 LITERATURE REVIEW AND STATE OF THE ART

4.1 Limitations

Within the limitations of the research, a series of central themes are proposed, which will specify the searches for the closest research concerning this doctoral proposal, as shown in Table 4:

Table 4. Limitations of the study, keywords included.

Keywords included	Reason
Machine learning	Central theme
Precipitation	Central theme
Monthly	Central theme
Prediction	Central theme
Spatiotemporal	Central theme
Time series	Central theme

As shown in Table 5, the constraints are grouped into exclusion criteria to have high-level categories, which makes it easier to perform the information filtering tasks.

Table 5. Limitations of the study: exclusion criteria classification.

Criteria grouping	Description
CE1	Not Extreme Events
CE2	Other geographies, other phenomena not contemplated
CE3	Is not the objective of the proposal

Table 6 shows all the limited topics for the information searches, which are grouped within the exclusion criteria CE3 classification.

Table 6. Limitations of the study, list of exclusion criteria CE3.

Exclusion criteria CE3	
Air pollution	Landside
Aquifer level	Landslide
data intelligence models	monsoon
cyclone	Rainfall
Debris	Runoff
Drought	slope failures
extreme Rainfall	standardized precipitation index
Flood	Streamflow
Groundwater	Typhoon
heavy Rainfall Forecast	water footprint
Landscape	Wildfire

The research window in years is **2020 to 2024**.

4.2 Additional limitations

This research focuses mainly on precipitation variables, selectively incorporating some climatic factors such as temperature, humidity, wind speed, ENSO indices or others, in case they are really necessary, in order to improve the predictive accuracy of the model. This decision is guided by the computational scope of the study, which emphasizes the optimization of machine learning techniques for precipitation prediction in mountainous regions such as Boyacá. The incorporation of more variables would significantly increase the dimensionality of the data, complicating both

the preprocessing and training phases. Although techniques such as feature selection and dimensionality reduction exist to manage such complexity, they could deviate from the main objective of improving ensemble learning methods specific to precipitation.

Moreover, the complex geographical variability of mountainous regions is one of the main objectives of this study. Addressing the spatial heterogeneity of these regions already imposes substantial computational demands, as topography strongly influences precipitation patterns. Limiting the inclusion of additional variables allows the research to focus on refining the spatiotemporal dynamics of precipitation prediction without introducing additional complexities that could dilute the objectives of the study. In addition, the computational cost of the underlying algorithms employed in the models will not be explored, as it is beyond the scope of this study.

Although these limitations may reduce the hydrological scope, they are in line with the project's goal of advancing machine learning methodologies for rainfall prediction. Future research could explore the integration of more climate variables once the foundational model has undergone rigorous performance and adaptability testing in complex terrain.

4.3 Search approach to establish the state of the question

This proposal's state-of-the-art approach will review the bibliographic sources exposed in [106]. It consists of choosing two bibliographic databases and filtering documents through identification, screening, eligibility, and inclusion phases. The identification part consists of article searches in two bibliographic databases, Scopus and Direct Science, based on a series of inclusion and exclusion criteria that will be listed below.

The so-called snowball is developed using the software VOSviewer, which allows aggregating the results of the searches in the databases to obtain views of the data that would improve the searches of articles with terms that are not contemplated but are of great value for the research in progress. Then, duplicates are excluded in both databases for both the original searches and the snowball method.

The next step is to perform the screening part, where articles will be filtered by reviewing titles, abstracts, introductions, and conclusions. We will apply the inclusion criteria shown in Table 7 and Table 8, which show the results of the database searches.

Table 7. Inclusion criteria.

List of inclusion criteria	Description
CI1	machine learning AND precipitation AND monthly
CI2	machine learning AND precipitation AND monthly AND prediction
CI3	machine learning AND precipitation AND monthly AND prediction AND spatiotemporal
CI4	machine learning AND precipitation AND monthly AND prediction AND spatiotemporal AND time series

Table 8. Consultation of Science Direct databases.

Title, abstract, keywords	Range	Result	Type	Database
machine learning AND precipitation AND monthly	2020-2024	74	Research articles, Review articles, Open Access	Science Direct
TI=(machine learning AND precipitation AND monthly) OR AK=(machine learning AND precipitation AND monthly) OR AB=(machine learning AND precipitation AND monthly) AND PY=(2020-2024)	2020-2024	126	Research articles, Review articles, Open Access	Web Of Science

The snowball technique focuses on terms not initially included in the criteria outlined in the previous table. To incorporate these as key terms, the most prominent and highly interconnected nodes are selected, disregarding the color coding of the time scale, as the research window is confined to the past five years.

The results of applying the snowball method can be categorized by the research objective, which is a computational approach with categories such as machine learning (ML), data (data), metric (metric), general, and climatic (climatic). The focus will be on ML; the others are not considered in Table 9.

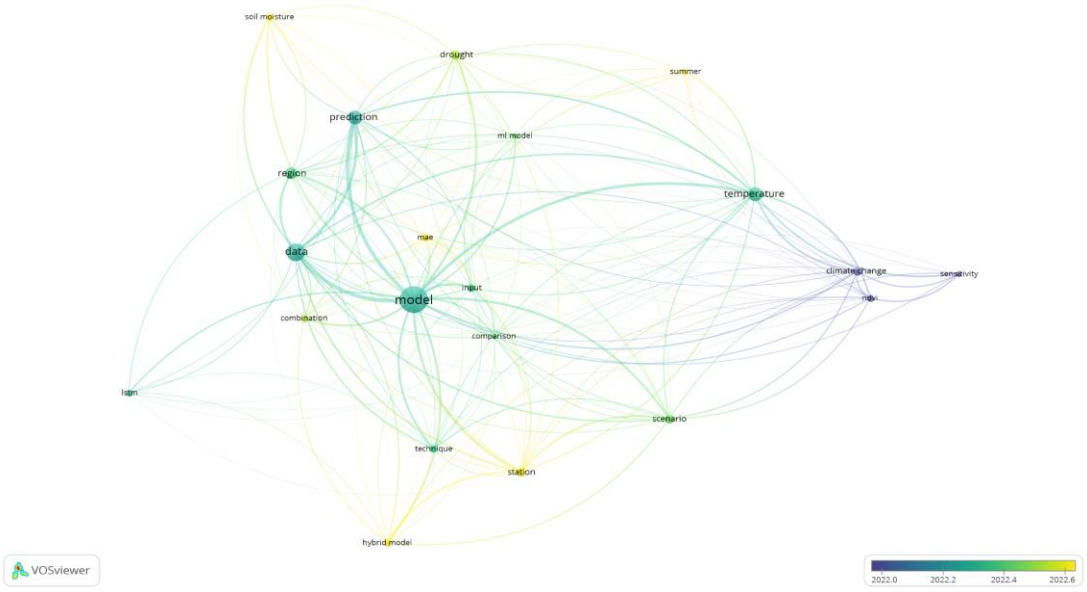

Figure 17. Network diagram – VOSViewer Software.

The terms result of applying the snowball method can be categorized by the research objective which is a computational approach with categories such as machine learning (ML), data (data), metric (metric), general, and climatic (climatic), where the focus is going to be on ML, the others are not considered in the Table 9.

Table 9. VOSviewer result terms.

Term	Computational approach	Date
combination	General	2022-4
comparison	General	2022-3
input	General	2022-4
mae	Metric	2022-6
ml model	ML	2022-3
model	ML	2022-2
technique	ML	2022-4
data	Data	2022-2
drought	Climatic	2022-4
lstm	ML	2022-2
prediction	ML	2022-2
region	General	2022-3
soil moisture	Climatic	2022-6

climate change	Climatic	2022-1
ndvi	Data	2022-1
sensitivity	General	2022-1
summer	Climatic	2022-6
temperature	Climatic	2022-2
hybrid model	ML	2022-6
scenario	General	2022-3
station	Data	2022-6

Filtering only by our focus and techniques and eliminating the repeated word "model" in the terms "hybrid model," "ml model," and "model," we obtain the result shown in Table 10.

Table 10. Filtering of repeated terms.

Term	Computational approach	Date
hybrid model	ML	2022-6
lstm	ML	2022-2
prediction	ML	2022-2

New search conditions are added for the databases based on the terms found at the computational level and machine learning algorithms, i.e., focused on the words LSTM, Prediction, and Hybrid Model, including the keywords Precipitation and Monthly to improve the searches, with data window from 2020 to 2024, as shown in Table 11.

Table 11. Snowball de Science Direct.

Title, resume, key words	Range	Result	Type	Database
precipitation AND monthly AND prediction AND lstm AND hybrid model	2020-2024	2	Research articles, Review articles, Open Access	Science Direct
TI=(prediction AND monthly AND precipitation AND lstm AND hybrid model) OR AK=(prediction AND monthly AND precipitation AND lstm AND hybrid model) OR AB=(prediction AND monthly AND precipitation AND lstm AND hybrid model) AND PY=(2020-2024)	2020-2024	6	Research articles, Review articles, Open Access	Web Of Science

Performing the standard and snowball searches in Web of Science and Science Direct and the different information filters produces the PRISMA diagram shown in Figure 19.

In recent years, there has been growing interest in the scientific community in using machine learning models to predict precipitation in specific geographic areas and time frames. Hybrid and ensemble models have gained particular attention because they combine the strengths of multiple algorithms, improving prediction accuracy and robustness compared to single models. This is especially important in precipitation prediction, where the complexity and variability of weather patterns can benefit from the complementary strengths of different modeling approaches. Figure 18 shows how the number of publications related to precipitation prediction has significantly increased over the past 20 years in Science Direct, Scopus, and Lens. It is observed that before 2014, the number of publications per month did not exceed 10 in any of the three analyzed databases. The search equation used to generate the graph was ('hybrid models' OR 'ensemble models' OR 'machine learning') AND ('precipitation prediction').

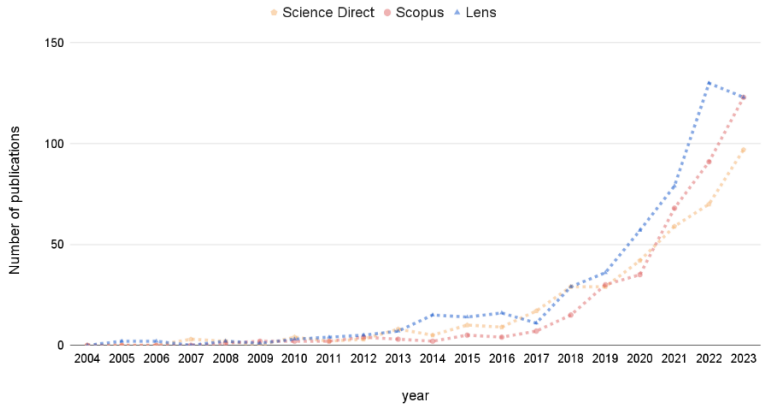

Figure 18. Number of publications per year in Science Direct, Scopus, and Lens.

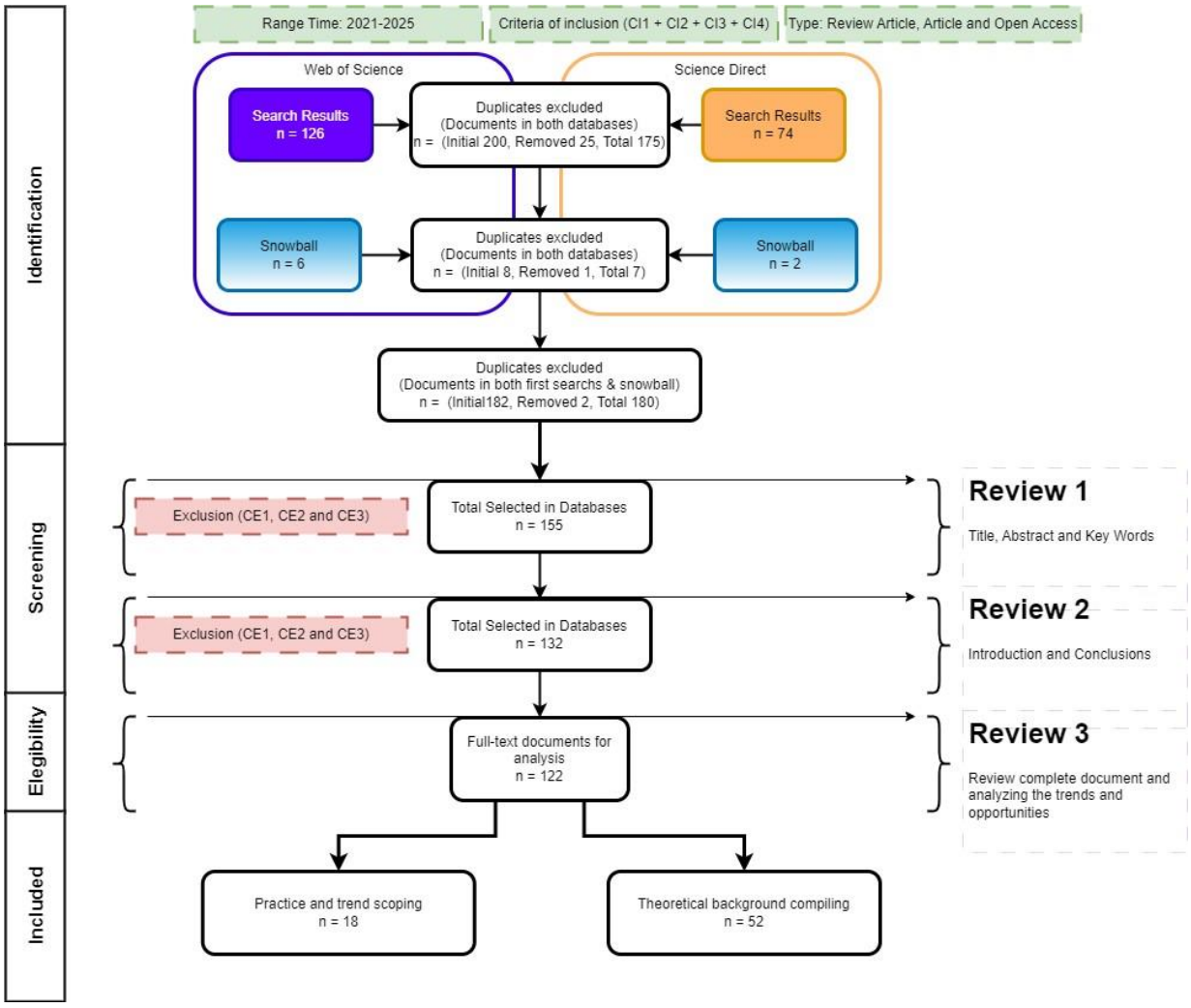

Figure 19. PRISMA protocol for the doctoral proposal.

5 SYNTHESIS OF THE STATE-OF-THE-ART

Table 12 expands the metrics used in the rainfall prediction models, so we will now use these standard terms within the referenced works.

Table 12. Metrics identified.

Metric	Meaning
ACC	Anomaly Correlation Coefficient
Accuracy	Accuracy
AI	accuracy improved
AUC	Area Under the Curve
Bias	Bias or the systematic tendency of a prediction model to overestimate or underestimate true values
Boxplot	Summarizes the distribution of the data, showing the median, quartiles and possible outliers
COV	Covariance
CRPS	Ranked Probability Score (CRPS) CRPS corresponds to the Mean Absolute Error (MAE) for deterministic MPF
CRPSS KS	CRPS Skill Score
Dm	Modified Mean Deviation (Dm) or Willmott's Concordance Index
ERR	Error Rate
FPR	False Positive Rate
IA	Index of Agreement
IQR	90% interquartile range
IQRSS	IQR Skill Score
KGE	Kling-Gupta Efficiency
m	Meter
MAE	Mean Absolute Error
MAPE	Average relative error of a set of data
MARE	Mean Absolute Relative Error
ME	Mean Error
mm	Millimeter
MRI	Mean relative improvement
MSE	Mean Squared Error
NS	Nash Sutcliffe coefficient
NSE	Nash-Sutcliffe Efficiency or Nash–Sutcliffe efficiency coefficient
PCC	Pearson's Correlation Coefficient (r)
Pg Score	Graded Test Pg Score
PI	Permutation Importance
PIT	Continuous, Probability integral transform
PITSS	PIT Area Skill Score
PREC	Precision
r	Pearson's Correlation Coefficient (PCC)
R^2	Coefficient of Determination
RAE	Relative Absolute Error

RB	Relative Bias
RE	Relative error between a single set of simulated data and the real data
RMSE	Root Mean Square Error
RPD	Ratio of Performance to Deviation
RRMSE	Relative Root Mean Square Error
RS	Relative Scoring
RSE	Relative Squared Error
SC	Spearman correlation
Scatter plot	Scatter plot can show how the predictions line up against the observed values
SKP	Skill Percentage
SN	Sensitivity
SP	Specificity
SR	Sinaplots of the rankings
Taylor diagrams	Taylor Diagrams are a graphical tool that combines several error metrics into a single diagram, three key components of the prediction are shown in a Taylor diagram, correlation between observed and predicted values, standard deviation of predictions compared to observations and root mean square error (RMSE)
U	Theil's U; U1 evaluates the overall performance of the model compared to an ideal performance and U2 evaluates the performance of the model compared to a trivial prediction (such as the mean).
VCR	Variance contribution rate

The categorization of "first cousins" is used to discover the state of the art shown in Figure 20, which will be reflected in the colors of the rows of each reference in
Table 13. Different algorithms and techniques are presented together with the variables used in the input of the models and the evaluation metrics, highlighting the best models for predicting monthly precipitation.

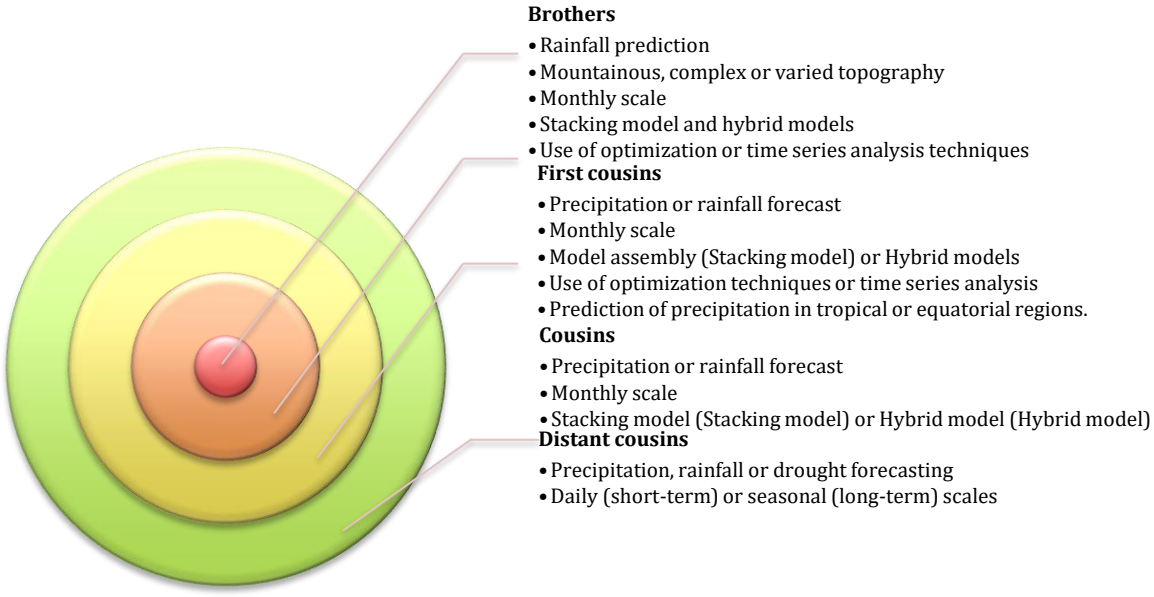

Figure 20. First cousin methodology applied to state-of-the-art classification.

Table 13. State of the art of monthly rainfall prediction models.

Reference	Algorithms	Other methods	Input variables	Delay time	Time interval	Metrics	Best
Proposal doctoral	Hybrid / Stacking	Input time with PAFC, SSA, EEDM, WT and Optimization Methods weights for Stacking Models. Normalization: MinMax Optimization data: - K-Fold Cross-validation	Precipitation, stations in Boyacá and CHIRPS (0.05°x0.05°) Monthly Mountainous zone: Altitude 45-5000 m.	x	Monthly	MSE RMSE MAE NSE R^2 r	y
[41]	ARIMA RF-R LSTM-NN	Normalization: MinMax	Precipitation, stations in Boyacá and CHIRPS (0.05°x0.05°) Monthly, 757 locations in Boyacá, Colombia, from January 1981 to August 2 023. Mountainous zone: Altitude 45-5000 m	12 months	Monthly	RMSE MAE R^2 MSE	LSTM
[107]	Multi-view stacking learning: Decision Tree RF KNN AdaBoost XGBoost LSTM	Multi-variable time series models incorporating lagged meteorological variables were employed to capture the dynamics of monthly rainfall. Normalization: MinMax Optimization data: - Cross-Correlation Function	Monthly rainfall in Rabat, Morocco, from 1993 to 2018.	Lagged meteorological variables Prec-12 Tem Max-11 & 12 Tem Min-4, 11 & 12 Insolation-12	Monthly	RMSE MAE R^2	Combining Decision Tree, KNN, and LSTM to build the meta-base while using XGBoost as the second-level learner
[24]	Stacking ensemble learning: CEEMD-RCMSE Stacking versus CEEMD-LSTM, KNN, RF, SVR and ANN as the base learners. ANN algorithm with strong generalization ability is used as the meta-learner	Adaptive noise-complete ensemble empirical modal decomposition (CEEMD) IMF Decomposition: modal components and one residual component using the CEEMD algorithm.	Precipitation measurements in the 179 Western Gorge, 1960 to 2019.	NA	Monthly	RMSE MAPE R^2	CEEMD296 RCMSE-Stacking
[86].	MLP LSTM MLSTM MLSTM-AM	Multiscale AM Optimization: Wavelet transform	Precipitation, 129 stations http://data.cma.cn/, 40 years 1966-01 to 2005-12.	1-12 months	Monthly	NSE RAE	MLSTM-AM
[37]	LSSVR SSA-LSSVR RF SSA-RF	Singular spectrum analysis, SSA Hybrid Model	1958-2018: Shihmen watershed, 12 rain gauges, rainfall monthly. 1981-2017: Deji Reservoir, watershed, 10 rain gauges, rainfall monthly.	1-month 2-month 3-month	Monthly	RMSE NSE	SSA-LSSVR, 1-month

[25]	KNN XGB SVR MLP Stacking Model (ALL)	Stacking algorithms $y = w_1 f_1 + \ldots w_4 f_4$ Time series analysis: Pearson correlation analysis Optimization data: - Cross-validation Optimization: hyper-parameter optimization	Monthly Precipitation, 9 stations. altitude (m) (4.6, 87.3)	4-month	- Monthly - Annually - Seasonally - Dry and wet months - Months of extreme rainfall	MAE RMSE R^2	Stacking Model
[44]	ARIMA SVR LSTM	- Ensemble empirical mode decomposition (EEMD) and Bayesian model averaging (BMA) - IMFs - The input time with PACF algorithm - Whose weights are calculated by birth-death Markov Chain Monte Carlo algorithm - CP and MW used for MBA prediction intervals	Monthly precipitation series data at Kunming station from January 1951 to December 2020	NA	Monthly	MSE RMSE MAE R^2	EEMD-BMA
[108]	LSTM ARIMA BPNN SVM XGB CEEMDAN-LSTM CEEMDAN-SVM	CEEMDAN can decompose complex data and the strongly fluctuating original data into several smoother component data, reducing the variability and abruptness of the data and decreasing the reconstruction error, thus improving prediction accuracy. IMF decomposition.	Monthly average rainfall data of Lanzhou station from January 1960 to December 2020 were selected for analysis.	NA	Monthly	MSE RMSE MAE R^2	CEEMDAN-LSTM
[109]	ARIMA ANN SVM RF GBR DSABASRNN	Optimization: hyperparameter optimization grid search and random search. Normalization: MinMax Time series analysis: Pearson correlation analysis	Rainfall, Temperature, Humidity and sunshine duration, 25 stations	4-month	Monthly	MAE RMSE R^2 Scatter plot	RF
[110]	LWLR MLP SVM RF	SMLM Precipitation-elevation interpolation method incorporating physiographical factors MLP, SVM, and RF are combined through a meta-learning algorithm with/without rescanning input covariates and K-th iteration. Optimization data: - K-Fold Cross-validation	- 174 stations (rain gauge) of IRIMO with 14 years of data, altitude (m) (-31, 5378) - With sixteen input covariates, including: - Two topographic features - Five cloud properties environmental variables 3 PPs - Inverse distance weighted (IDW)		Monthly	MAE r KGE bias RMSE	LWLR
[39]	ELM DNN GA-ELM BBO-ELM PSO-ELM WELM WDNN WGA-ELM WBBO-ELM WPSO-ELM	Hybrid Model DWT Based Partial Autocorrelation Function (PACF)	1871-2006	1-month 2-month 3-month	Monthly	r NSE MAE RMSE	BBO-ELM
[111]	TVF-EMD-ENN WT-ENN CEEMD-ENN	Decomposition of nonlinear precipitation data into several subcomponents with three methods: 1) empirical mode decomposition based on time-varying filtering (TVF-EMD). 2) wavelet transform (WT)	Monthly precipitation data in this study span from 1961 to 2020, collected from the China Meteorological Data website (http://data.cma.cn/data/cdcdetail/dataCode/SURF_		Monthly	NSE RMSE MAE VCR	TVF-EMD-ENN

		3) complementary ensemble empirical mode decomposition (CEEMD)	CLI_CHN_MUL _MON.htm), covering 60 hydrological years.				
[112]	RF XGB RNN LSTM SMOTE-km-XGB SMOTE-km-RF	The k-means clustering and SMOTE algorithms are combined to augment the sample data and improve forecasting accuracy.	Dataset CN05 0.25°x0.25°, daily data, 1982 a 2015.		Annual Monthly	ACC Pg Score	SMOTE-km-XGB
[113]	WT-FFBP-NN WT-ELM	A wavelet-based multiscale deep learning approach is developed to forecast precipitation using the lagged monthly rainfall.	Indian Metrological Department for each year from 1901 to 2018, 0.25°x0.25°.	ACF CCF	Monthly	RMSE R^2 NSE MAE	WT-ELM
[3]	ELM ELM EMD-HHT CEEMD-ELM CEEMD-ELM-FFOA	- Complementary ensemble empirical mode decomposition (CEEMD) can effectively reduce mode aliasing and white noise interference. - The fruit fly optimization algorithm (FFOA) has better local optimization ability - Intrinsic Mode Functions (IMFs): noise amplitude 0.2, noise logarithm is 50; the decomposition effect is ideal.	Monthly precipitation 1951 to 2020 provided by the National Data Center for Meteorological Sciences and Water Resources Bulletin of Zhengzhoun	NA	Monthly	Taylor diagrams RE MAE RMSE MAPE	CEEMD-ELM-FFOA
[114]	The ensemble learners base learners: MARS poly-MARS RF GBM XGBoost BRNN Mean and median combiners. Stacking with: LR MARS poly-MARS RF GBM XGBoost BRNN	Mean and median combiners.	Monthly data from the PERSIANN and IMERG gridded datasets span 15 years and the entire contiguous United States (CONUS). Gauge-measured precipitation data from the Global Historical Climatology Network monthly database, version 2 (GHCNm).	NA	Monthly	MSE RS	Stacking (regression algorithm) with base learners is a better combination with LR.
[95]	Linear components: Annual-ARIMA Monthly-SARIMA Nonlinear components: SVR, GEP, GMDH. Ensemble models with multiple configurations, hybrid model one linear and one nonlinear model.	The three machine-learning models were combined using a genetic algorithm instead of SVR. Preprocessing configurations and each of the Gene Expression Programming (GEP), Support Vector Regression (SVR) and Group Method of Data Handling (GMDH). Hybrid models use an ensemble of linear and nonlinear algorithms for post-processing the output of forecasting models.	Precipitation data of two weather stations in Iran, 2000–2019.		Annual Monthly	RMSE RRMSE MSE U1 U2 MAE RPD NSE AMAPE APB Dm AI GMER	Monthly: SARIMA-Ensemble-Optimization GA
[115]	Modern hybrid learning machine: - Wavelet and Outlier Robust Extreme Learning Machine (ORELM) models are “Wavelet-Outlier Robust Extreme Learning Machine (WORELM)	Normalization: MinMax - The optimal mother wavelet and the best decomposition level of the wavelet model are computed. A trial-and-error procedure is used to choose the optimal number of hidden layer neurons and the best activation function of the ORELM model. - The regularization parameter of the	Ardabil meteorological station, 1976 to 2020.	WORELM (14), which contained time-series lags (t-1), (t-2), (t-3) and (t-12).	Monthly	r MARE VAF NSE	Hybrid WORELM

		ORELM model is also optimized - Using the autocorrelation function (ACF), the most influencing lags of the time-series data are detected, and 14 WORELM models are developed.					
[116]	- Multilayer perceptron neural networks (MLP-NN). - Deep neural networks using TensorFlow.	- Optimization of neural networks using the Multi-Particle Collision Algorithm (MPCA). - Adam optimization	Wind components (u, v) at 850 hPa and 500 hPa levels, air temperature at 2 meters, specific humidity at 850 hPa, and monthly precipitation amount. Data from 1980 to 2019, where 1980-2016 was used for training and generalization, and 2017-2019 for testing		Monthly	RMSE COV ME	Deep neural networks using TensorFlow
[117]	- Fine Decision Tree - Course K-Nearest Neighbors (CKNN) - Gaussian Support Vector Machines (GSVM) - Neural Network	- Data preprocessing	The dataset includes minimum and maximum temperatures, humidity, wind speed, and wind direction from 2007 to 2018.		Daily	ERR Accuracy SN SP PREC FPR AUC	Course K-Nearest Neighbors
[35]	BLR DFR NNR BDTR	Optimization data: - Cross-validation Time series analysis: - AFC Normalization: MinMax	Rainfall, 10 stations, altitude (m) (-14, 2062)		- Daily AFC - Weekly AFC - 10 Days AFC - Monthly AFC	MAE RMSE RAE RSE R^2	BDTR
[118]	MLP CNN GRU BLSTM LSTM BLSTM+GRU	Hybrid models Normalization: MinMax	The station at Simtokha (1997 – 2017): Maximum Temperature Minimum Temperature Rainfall Relative Humidity Sunshine Wind Speed Altitude (m) (2248, 2648)	12 months	Monthly	MAE RMSE R^2 r	BLSTM+GRU
[119]	LSTM, ELM, EMD, ELM-EMD Input t-3, t-2, t-1 Output t+1 In different combinations.	Optimization: ACF for optimum lag times. Hybrid model ELM-EMD	1960 – 2020 Monthly average precipitation	1-3 months	Monthly 3-months 6-months	MSE RMSE MAE NSE OI R^2 Scatter plot Taylor Radar Boxplot	For SPI-1, LMST For SPI-6, ELM-EMD
[120]	LGB XGB	Deterministic evaluation and probabilistic evaluation with BJP (Bayesian Joint Probability) calibration. The Gibbs sampling based on Markov chain Monte Carlo is the core of BJP to establish the Bayesian parameters inference and generate climatological reference distribution.	CMA China Meteorological Administration precipitation data 1981-2015	12 months	Monthly	RB CRPS PIT IQR CRPSS PITSS IQRSS	Ensemble monthly precipitation forecasts (MPFs) generated from GCMs.
[121]	FBP MLP SVR LGB SVR XGB	All models trained 1911/1961 to 2005 and predict 2006-2015 Case 1: a.FBP Model was trained from 1911. b.FBP Model was trained in 1961.	Monthly observed rainfall (1911–2015) over the Brisbane River catchment varied from nil to 1360 mm, with an annual average	NA	Monthly	r ACC IA MAE	FBP

	RF Stacking Models (5)	Case 2: a. Additional regressor was the arithmetic mean of the best five GCMs b. FBP with all eight GCMs as eight individual regressors.	rainfall of 628 mm				
[122]	LR RF GBM XGB FFNN BRFFNN poly-MARS		PERSIANN Data, 0.25°x0.25° spatial resolution transform daily to monthly precipitation, 2001-2015. Total monthly precipitation data from the Global Historical Climatology Network, 2001 to 2015. Elevation data Amazon Web Services (AWS) Terrain Tiles.	NA	Monthly	RS SC PI MRI SR	XGB
[123]	North American multi-model ensemble (NMME): GEM-NEMO NASA-GEOSS2S CanCM4i COLA-RSMAS-CCSM4 ANN RF	Multi-model ensemble forecasting of monthly precipitation: Models; Artificial neural network (ANN) and random forest (RF) algorithms for post-processing the output of forecasting models.	ERA5 data for estimating monthly precipitation, 1°×1°, 1982 to 2007.	Lead Times 1-4, 5-9, 9,12 Months	Monthly	RMSE KGE NSE r	RF

6 HYPOTHESIS

Applying machine learning models, combined with time series analysis and data preprocessing methods, will significantly improve the accuracy of monthly precipitation predictions in mountainous areas. This research addresses the limitations of traditional models by more effectively capturing the spatial and temporal variations in precipitation, considering the geographical complexity of the Boyacá region.

7 OVERALL OBJECTIVE

To optimize a monthly computational model for spatiotemporal precipitation prediction in mountainous areas, improving its accuracy through the use of hybridization and ensemble machine learning techniques.

7.1 SPECIFIC OBJECTIVES

1. Generate a dataset by extracting information features from heterogeneous sources, facilitating its integration for pattern discovery and exploration.
2. Develop baseline models for spatiotemporal monthly precipitation prediction using the generated dataset and applying various machine learning approaches.
3. Propose a variation of models using hybridization and ensemble techniques, aimed at improving the accuracy of spatiotemporal monthly precipitation predictions, and compare them with the baseline models.
4. Evaluate the accuracy and efficiency of the proposed predictive models through specific quality and performance metrics, comparing the results with those obtained from the baseline models.

8 METHODOLOGICAL FRAMEWORK

The methodological development is framed by three main features: state-of-the-art research methodology, project architecture and a deliverables schedule.

8.1 State-of-the-art research

The section discusses the state-of-the-art search methodology supported by the PRISMA protocol and the snowball technique in more detail.

8.2 Project architecture

At a more organizational and technical level, an architecture was proposed at the macro project level for rainfall prediction, as shown in Figure 21. Here, we can see the contribution of the different researchers of the Galash research group within the models to be evaluated within the macro project. We will find the various models developed and centralized to have them available in standard, hybrid, or stacking evaluations.

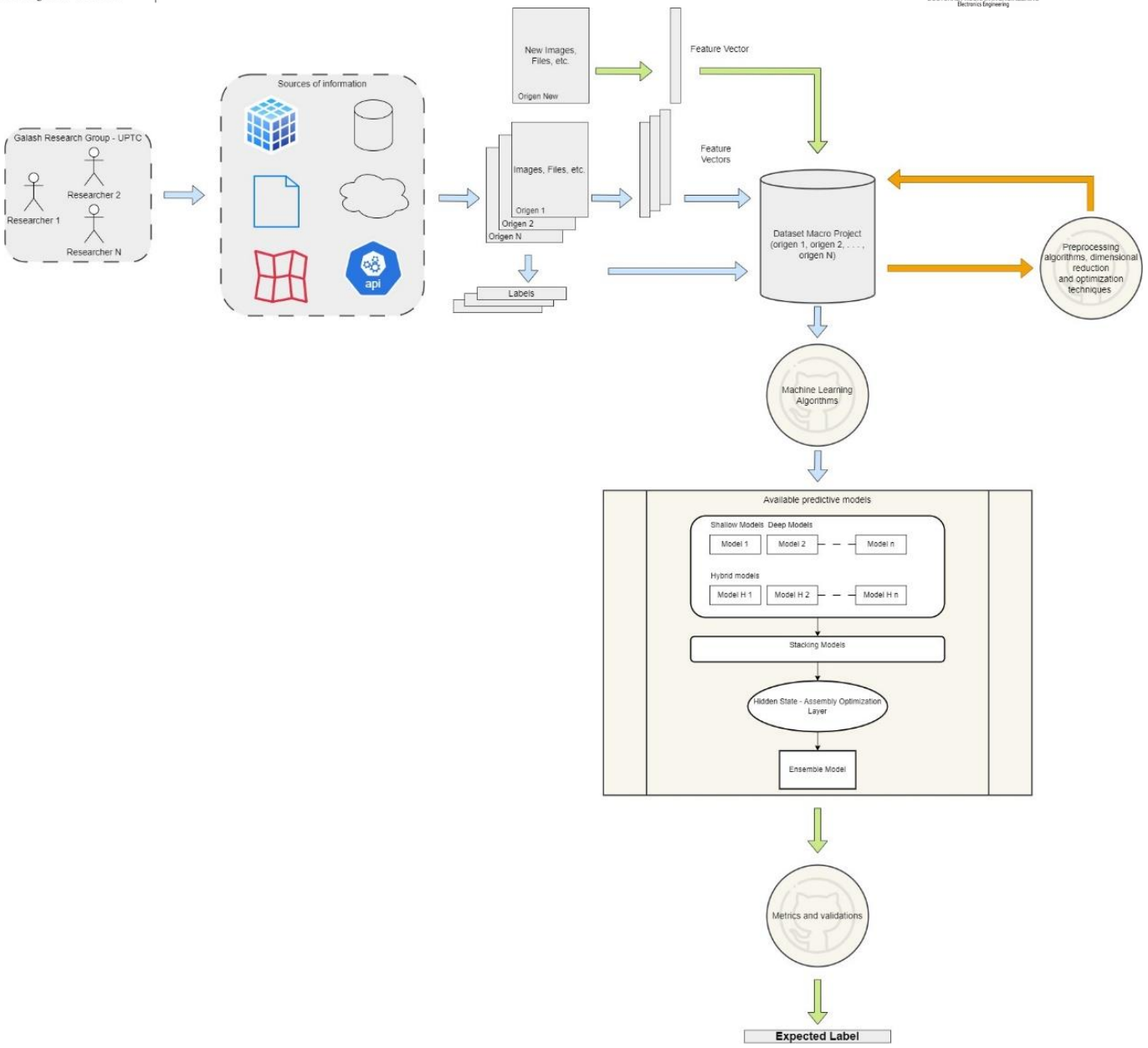

Figure 21. The architecture of the Project Framework for the Galash Research Group.

8.3 Framework

In the context of this doctoral proposal, a task-based framework has been formulated to facilitate the development of the project at hand. The framework, divided into three phases, encompasses tasks that will be meticulously displayed on a calendar with corresponding expected outcomes. The overarching objective is to achieve the proposed goals over time, with the advanced level serving as the ultimate target within the work plan. As we progress through the levels, the complexity and requirements of the tasks will progressively elevate to ensure the attainment of achievable goals.

Figure 22. Framework by levels.

8.4 Evaluation Metrics

Several widely accepted performance metrics will be used to ensure a comprehensive and scientifically sound evaluation of the machine learning models developed for monthly precipitation prediction in mountainous areas. These metrics have been selected based on their frequent application in climate and hydrological prediction models, as shown in
Table 13, summarized in Figure 23, is similar to the spatiotemporal nature of precipitation prediction, and the most relevant metrics will be used in this regard.

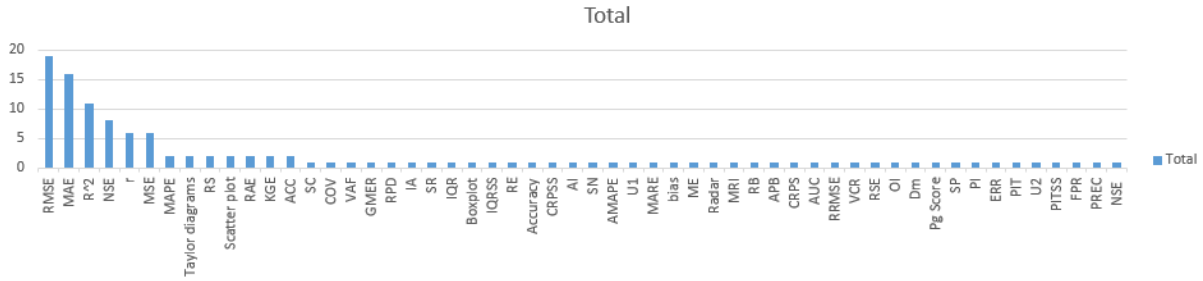

Figure 23. Grouping of the use of metrics used in the studies in Table 13.

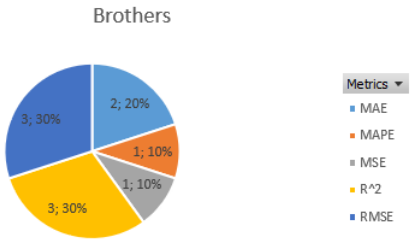

Figure 24. Distribution of use of Brother's category metrics from Table 13.

Other model agnostic techniques could also be considered to quantify the uncertainty of model predictions, such as conformal prediction, which has been successfully used to provide confidence intervals for time series prediction currently available in Python packages.

Root Mean Square Error (RMSE)

RMSE is a standard metric used to evaluate the accuracy of regression models, particularly in climate and precipitation prediction. RMSE emphasizes and penalizes significant errors more heavily, making it suitable for capturing extreme variability in monthly precipitation data. Given the unpredictable and fluctuating nature of precipitation in mountainous regions, RMSE helps quantify the performance of models that may exhibit high variability. RMSE is frequently applied in hydrological studies and climate prediction models.

8.4.1 Mean Absolute Error (MAE)

MAE measures the average magnitude of errors in predictions, offering a straightforward and interpretable metric. Unlike RMSE, MAE treats all errors equally, providing a balanced model performance evaluation without disproportionately weighting large errors. This makes MAE useful for applications where moderate deviations in monthly precipitation predictions are expected but not critical to penalize.

8.4.2 Coefficient of Determination (R^2)

The R^2 or Coefficient of Determination is a fundamental metric used to evaluate the quality of regression models. It measures the proportion of the variance in the dependent variable explained by the model's independent variables. An R^2 value close to 1 indicates that the model explains most of the variability in the observed data, making it a helpful tool for understanding how well a model captures the overall trend in monthly precipitation predictions. R^2 is frequently applied in climate and hydrological studies due to its ability to assess the overall predictive accuracy of models. However, it may be less effective when outliers influence the model's predictions.

8.4.3 Nash-Sutcliffe Efficiency (NSE)

NSE is a specialized metric widely used in hydrological modeling to evaluate the performance of predictive models. It measures the relative magnitude of residual variance compared to the variance of the observed data. An NSE value closer to 1 indicates that the model predictions are highly accurate. As in this study, NSE is particularly suitable for spatiotemporal predictions where spatial and temporal variations are considered.

8.4.4 Pearson Correlation Coefficient (r)

The Pearson Correlation Coefficient (r) measures the strength and direction of the linear relationship between two variables. It is widely employed in weather and precipitation prediction models to assess the correlation between predicted and observed values. The r value ranges from -1 to 1, with values close to 1 indicating a strong positive correlation. In such studies, r is precious for understanding how well the model predicts the linear trend of monthly precipitation data, offering insight into the relationship between predicted and observed values in mountainous regions. However, like R^2, r can be sensitive to extreme values and outliers in the dataset.

8.4.5 Mean Squared Error (MSE)

The Mean Squared Error (MSE) is a commonly used metric for evaluating the performance of regression models by calculating the average of the squared differences between predicted and observed values. MSE emphasizes more significant errors due to the squaring of residuals, making it particularly sensitive to outliers. This sensitivity makes MSE a valuable tool for evaluating models where minimizing large deviations, such as those occurring in extreme precipitation events, is critical. In the context of climate modeling and precipitation predictions, MSE helps quantify model accuracy, with smaller values indicating more precise predictions. However, MSE may not interpret errors intuitively as the output units are squared, necessitating careful consideration when comparing different models.

9 SCHEDULE

Table 14. Development schedule.

Levels and Activities	Time (in months) 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
Beginner level	
LB.1: Study of available data	---
LB.2: Adaptation of the data for monthly single-point forecasts	---
LB.3: Choice of monthly single-point forecasting model	---
LB.4: Development of the single-point monthly forecasting model	---
LB.5: Metrics and evaluation	---
Intermediate level	
LM.1: Adaptation of the data for the clustering technique.	---
LM.2: Choice and application of hybridizations to two monthly prediction models.	--- ---
LM.3: Precipitation model in series of months, with average, maximum and minimum values.	---
LM.4: Exploration and application of interpolation techniques between points.	---
LM.5: Exploration and application of monthly recalibration techniques for the models.	---
LM.6: Adaptation of models to take into account geographical areas (basic).	---
Advanced level	
LA.1: Precipitation model in series of months, with moving average.	---
LA.2: Implementation of coupling model with the two hybrid models.	--- ---
LA.3: Adaptation of improvements for the prediction distribution solution.	--- ---
LA.4: Adaptation for forecast windows of one month or more.	--- ---
LA.5: Adaptation of the model for geographic zones by localities	--- ---
LA.6: Design and implementation of model deployment scheme and user interfaces.	--- ---
LA.7: Exploration and implementation of forecast resolution enhancements	--- ---
Complementary tasks	
TC.1: Organization, documentation and presentation of final results	--- ---
Expected outputs	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
1: Dataset with the available data in a format suitable for working with ML models.	
2: Source code used to build the Dataset	
3: Source code used for model construction	
4: Source code used for metrics construction	
5: Source code used for building clustering models	
6: Source code used for building hybrid models	
7: Source code used for interpolation model building	
8: Source code used for building re-calibration models	
9: Source code used for building geographic zone models	
10: Source code used for moving average series of months	
11: Source code used for construction of the fitting model	
12: Source code of improvements of the forecast distribution solution	
13: Source code for adapting forecast windows of one month or longer	
14: Source code of model adaptation for geographic zones by locations	
15: Source code for model deployment scheme and user interfaces	
16: Predictive model consumption interface	
17: Prediction resolution enhancement source code	
18: Document with final results and presentation	

10 EXPECTED OUTPUTS

The development of this proposal will produce the following deliverables.

10.1 Human resources training

- The doctoral training of the author.

10.2 Technology development and product innovation

- Heterogeneous source dataset: A comprehensive dataset generated from diverse sources, integrated and preprocessed for use in machine learning models for monthly precipitation prediction.
- Optimized machine learning (ML) model: A high-performance ML model specifically tailored for predicting monthly precipitation in mountainous regions, validated through extensive testing and comparison with reference models.
- AI/ML-based architecture for spatiotemporal predictions: A flexible, scalable architecture designed to support the deployment of spatiotemporal precipitation prediction models, incorporating hybridization and ensemble techniques to improve prediction accuracy.

10.3 New knowledge

- Two articles in scientific journals.

10.4 Social knowledge appropriation products

- One International Conference.

11 BUDGET

Table 15. Budget.

OVERALL PROPOSAL BUDGET BY SOURCE OF FINANCING (IN COP PESOS)					
ITEMS	**RESOURCES**				**TOTAL**
	UPTC		**Own resources**		
	Cash	**Equipment - staff**	**Cash**	**Equipment - staff**	
Fees (Professional Services)		$19.200.000			$19.200.000
Equipment Purchase				$6.000.000	$6.000.000
Software			$400.000		$400.000
Printing and Publications				$15.000.000	$15.000.000
Per Diem and Travel Expenses	$8.000.000		$12.000.000		$20.000.000
Registration: academic events	$6.000.000				$6.000.000
TOTAL	$14.000.000	$19.200.000	$12.400.000	$21.000.000	$66.600.000

12 DISAGGREGATED BUDGET

Table 16. Disaggregated budget.

DESCRIPTION FEE EXPENSES (IN COP PESOS)							
Type of Researcher	**Researcher Profile**	**Dedication Hours/week**	**RESOURCES**				**TOTAL**
			UPTC		**Own resources**		
			Cash	**Staff**	**Cash**	**Staff**	
Advisor	Senior	1		$100.000			$100.000
Co-advisor	Senior	1		$100.000			$100.000
TOTAL		192					$19.200.000

DESCRIPTION PURCHASE OF EQUIPMENT (IN COP PESOS)						
Equipment	**Justification**	**RESOURCES**				**TOTAL**
		UPTC		**Own resources**		
		Cash	**Equipment - staff**	**Cash**	**Equipment - staff**	
1 computer	For data analysis and machine learning model generation.				$6.000.000	$6.000.000
TOTAL						$6.000.000

SOFTWARE DESCRIPTION (IN COP PESOS)						
Equipment	**Justification**	**RESOURCES**				**TOTAL**
		UPTC		**Own resources**		
		Cash	**Equipment - staff**	**Cash**	**Equipment - staff**	
Anaconda	Development environment				$0	$0
Python3	Development of ML models				$0	$0
Jupiter Notebooks	Notebook development environment				$0	$0
Zotero	Bibliographic manager				$400.000	$400.000
TOTAL						$400.000

DESCRIPTION OF PRINTED MATERIAL AND PUBLICATIONS (IN COP PESOS)				
Item	**Justification**	**RESOURCES**		**TOTAL**
		UPTC	**Own resources**	
1	Article publication		$5.000.000	$5.000.000

2	Article publication		$10.000.000	$10.000.000
			TOTAL	$15.000.000

DESCRIPTION OF PER DIEMS AND TRAVEL EXPENSES (IN COP PESOS)							
Place / No. of trip	Justification	Tickets	Hotel	Days	**RESOURCES**		
					UPTC	**Own resources**	**TOTAL**
					Cash	**Cash**	
Chile/1	International Internship	$2.500.000	$9.500.000	90		$12.000.000	$12.000.000
Europe/2	International Internship	$2.500.000	$3.500.000	30	$6.000.000		$6.000.000
Colombia/1	International Internship	$500.000	$1.500.000	5	$2.000.000		$2.000.000
				TOTAL	$2.000.000	$12.000.000	$20.000.000

DESCRIPTION REGISTRATION TO ACADEMIC EVENTS (IN PESOS COP)		
Item	**Justification**	**Value**
National Conference	Results dissemination	$2. 000.000
International Conference	Results dissemination	$4.000.000
	TOTAL	$6.000.000

13 BIBLIOGRAPHY

[1] National Aeronautics and Space Administration, “Understanding Earth: Whats Up With Precipitation? | Precipitation Education.” Accessed: Jun. 03, 2024. [Online]. Available: https://gpm.nasa.gov/education/articles/understanding-earth-whats-precipitation

[2] Z. M. Yaseen, S. O. Sulaiman, R. C. Deo, and K.-W. Chau, “An enhanced extreme learning machine model for river flow forecasting: State-of-the-art, practical applications in water resource engineering area and future research direction,” *J. Hydrol.*, vol. 569, pp. 387–408, Feb. 2019, doi: 10.1016/j.jhydrol.2018.11.069.

[3] X. Zhang and X. Wu, “Combined Forecasting Model of Precipitation Based on the CEEMD-ELM-FFOA Coupling Model,” *Water*, vol. 15, no. 8, p. 1485, Apr. 2023, doi: 10.3390/w15081485.

[4] M. Ohba, “Chapter 2 - Precipitation under climate change,” in *Precipitation*, J. Rodrigo-Comino, Ed., Elsevier, 2021, pp. 21–51. doi: https://doi.org/10.1016/B978-0-12-822699-5.00002-1.

[5] D. C. Sohoulande Djebou and V. P. Singh, “Impact of climate change on precipitation patterns: a comparative approach,” *Int. J. Climatol.*, vol. 36, no. 10, pp. 3588–3606, 2016, doi: 10.1002/joc.4578.

[6] K. E. Kunkel *et al.*, “Probable maximum precipitation and climate change,” *Geophys. Res. Lett.*, vol. 40, no. 7, pp. 1402–1408, 2013, doi: 10.1002/grl.50334.

[7] K. E. Trenberth, “Changes in precipitation with climate change,” *Clim. Res.*, vol. 47, no. 1–2, pp. 123–138, Mar. 2011, doi: 10.3354/cr00953.

[8] M. R. Islam and M. Khan, “Forest Fires and Anthropogenic CO2,” in *The Science of Climate Change*, John Wiley & Sons, Ltd, 2019, pp. 37–102. doi: https://doi.org/10.1002/9781119522850.ch3.

[9] I. Ebtehaj, H. Bonakdari, M. Zeynoddin, B. Gharabaghi, and A. Azari, “Evaluation of preprocessing techniques for improving the accuracy of stochastic rainfall forecast models,” *Int. J. Environ. Sci. Technol.*, vol. 17, no. 1, pp. 505–524, Jan. 2020, doi: 10.1007/s13762-019-02361-z.

[10] V. Nourani, S. Uzelaltinbulat, F. Sadikoglu, and N. Behfar, “Artificial Intelligence Based Ensemble Modeling for Multi-Station Prediction of Precipitation,” *Atmosphere*, vol. 10, no. 2, Art. no. 2, Feb. 2019, doi: 10.3390/atmos10020080.

[11] R. Lee and J. Liu, “iJADE WeatherMAN: a weather forecasting system using intelligent multiagent-based fuzzy neuro network,” *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, vol. 34, no. 3, pp. 369–377, Aug. 2004, doi: 10.1109/TSMCC.2004.829302.

[12] IDEAM, “Estudio Nacional del Agua.” May 2015.

[13] “The Impact of Sea Surface Temperature Biases on North American Precipitation in a High-Resolution Climate Model in: Journal of Climate Volume 33 Issue 6 (2020).” Accessed: Sep. 19, 2024. [Online]. Available: https://journals.ametsoc.org/configurable/content/journals$002fclim$002f33$002f6$002fjcli-d-19-0417.1.xml?t:ac=journals%24002fclim%24002f33%24002f6%24002fjcli-d-19-0417.1.xml

[14] F. Raymond, A. Ullmann, P. Camberlin, P. Drobinski, and C. C. Smith, “Extreme dry spell detection and climatology over the Mediterranean Basin during the wet season,” *Geophys. Res. Lett.*, vol. 43, no. 13, pp. 7196–7204, 2016, doi: 10.1002/2016GL069758.

[15] M. Hope, “Accidental activist,” *Nat. Clim. Change*, vol. 5, no. 11, pp. 974–974, Nov. 2015, doi: 10.1038/nclimate2856.

[16] “Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling - Fowler - 2007 - International Journal of Climatology

- Wiley Online Library." Accessed: Sep. 19, 2024. [Online]. Available: https://rmets.onlinelibrary.wiley.com/doi/10.1002/joc.1556

[17] P. C. D. Milly, K. A. Dunne, and A. V. Vecchia, "Global pattern of trends in streamflow and water availability in a changing climate," *Nature*, vol. 438, no. 7066, pp. 347–350, Nov. 2005, doi: 10.1038/nature04312.

[18] D. B. Preston, "Review of Spectral Analysis and Time Series," *Technometrics*, vol. 25, no. 2, pp. 213–214, 1983, doi: 10.2307/1268567.

[19] D. KOUTSOYIANNIS, "Climate change, the Hurst phenomenon, and hydrological statistics," *Hydrol. Sci. J.*, vol. 48, no. 1, pp. 3–24, Feb. 2003, doi: 10.1623/hysj.48.1.3.43481.

[20] A. Mosavi, F. Sajedi Hosseini, B. Choubin, M. Goodarzi, A. A. Dineva, and E. Rafiei Sardooi, "Ensemble Boosting and Bagging Based Machine Learning Models for Groundwater Potential Prediction," *Water Resour. Manag.*, vol. 35, no. 1, pp. 23–37, Jan. 2021, doi: 10.1007/s11269-020-02704-3.

[21] J. Xu *et al.*, "A Spatial Downscaling Framework for SMAP Soil Moisture Based on Stacking Strategy," *Remote Sens.*, vol. 16, no. 1, p. 200, Jan. 2024, doi: 10.3390/rs16010200.

[22] K. Halder *et al.*, "Application of bagging and boosting ensemble machine learning techniques for groundwater potential mapping in a drought-prone agriculture region of eastern India," *Environ. Sci. Eur.*, vol. 36, no. 1, p. 155, Sep. 2024, doi: 10.1186/s12302-024-00981-y.

[23] Y. Zhang, J. Liu, and W. Shen, "A Review of Ensemble Learning Algorithms Used in Remote Sensing Applications," *Appl. Sci.*, vol. 12, no. 17, p. 8654, Aug. 2022, doi: 10.3390/app12178654.

[24] X. Zhang, K. Wang, and Z. Zheng, "A novel integrated learning model for rainfall prediction CEEMD-FCMSE -Stacking," *EARTH SCIENCE INFORMATICS*, vol. 15, no. 3. SPRINGER HEIDELBERG, TIERGARTENSTRASSE 17, D-69121 HEIDELBERG, GERMANY, pp. 1995–2005, Sep. 2022. doi: 10.1007/s12145-022-00819-2.

[25] J. Gu, S. Liu, Z. Zhou, S. R. Chalov, and Q. Zhuang, "A Stacking Ensemble Learning Model for Monthly Rainfall Prediction in the Taihu Basin, China," *Water*, vol. 14, no. 3, p. 492, Feb. 2022, doi: 10.3390/w14030492.

[26] S. Stathopoulos, A. Gemitzi, and K. Kourtidis, "Statistical Downscaling of Remote Sensing Precipitation Estimates Using MODIS Cloud Properties Data over Northeastern Greece," *Remote Sens. Earth Syst. Sci.*, vol. 7, no. 2, pp. 113–122, Jun. 2024, doi: 10.1007/s41976-024-00107-1.

[27] S. Scher and S. Peßenteiner, "Technical Note: Temporal disaggregation of spatial rainfall fields with generative adversarial networks," *Hydrol. Earth Syst. Sci.*, vol. 25, no. 6, pp. 3207–3225, Jun. 2021, doi: 10.5194/hess-25-3207-2021.

[28] Y. Zhang, J. Li, and D. Liu, "Spatial Downscaling of ERA5 Reanalysis Air Temperature Data Based on Stacking Ensemble Learning," 2024.

[29] S. Welten *et al.*, "Synthetic rainfall data generator development through decentralised model training," *J. Hydrol.*, vol. 612, p. 128210, Sep. 2022, doi: 10.1016/j.jhydrol.2022.128210.

[30] S. S. Dagne, Z. R. Roba, M. B. Moisa, K. T. Deribew, D. O. Gemeda, and H. H. Hirpha, "Rainfall prediction for data-scarce areas using meteorological satellites in the case of the lake Tana sub-basin, Ethiopia," *J. Water Clim. Change*, vol. 15, no. 5, pp. 2188–2211, May 2024, doi: 10.2166/wcc.2024.636.

[31] S. Ranhao, Z. Baiping, and T. Jing, "A Multivariate Regression Model for Predicting Precipitation in the Daqing Mountains," *Mt. Res. Dev.*, vol. 28, no. 3/4, pp. 318–325, Aug. 2008, doi: 10.1659/mrd.0944.

[32] C. Chen *et al.*, "Performance of Multiple Satellite Precipitation Estimates over a Typical Arid Mountainous Area of China: Spatiotemporal Patterns and Extremes," *J. Hydrometeorol.*, vol. 21, no. 3, pp. 533–550, Mar. 2020, doi: 10.1175/JHM-D-19-0167.1.

[33] L. Wang *et al.*, "Precipitation–altitude relationships on different timescales and at different precipitation magnitudes in the Qilian Mountains," *Theor. Appl. Climatol.*, vol. 134, no. 3–4, pp. 875–884, Nov. 2018, doi: 10.1007/s00704-017-2316-1.

[34] M. R. Nikpour, S. Abdollahi, H. Sanikhani, J. Raeisi, and Z. M. Yaseen, "Coupled data pre-processing approach with data intelligence models for monthly precipitation forecasting," *Int. J. Environ. Sci. Technol.*, vol. 19, no. 12, pp. 11919–11934, Dec. 2022, doi: 10.1007/s13762-022-04395-2.

[35] W. M. Ridwan, M. Sapitang, A. Aziz, K. F. Kushiar, A. N. Ahmed, and A. El-Shafie, "Rainfall forecasting model using machine learning methods: Case study Terengganu, Malaysia," *Ain Shams Eng. J.*, vol. 12, no. 2, pp. 1651–1663, Jun. 2021, doi: 10.1016/j.asej.2020.09.011.

[36] C. Ocampo-Marulanda, C. Fernández-Álvarez, W. L. Cerón, T. Canchala, Y. Carvajal-Escobar, and W. Alfonso-Morales, "A spatiotemporal assessment of the high-resolution CHIRPS rainfall dataset in southwestern Colombia using combined principal component analysis," *Ain Shams Eng. J.*, vol. 13, no. 5, p. 101739, Sep. 2022, doi: 10.1016/j.asej.2022.101739.

[37] P. O. Bojang, T.-C. Yang, Q. B. Pham, and P.-S. Yu, "Linking Singular Spectrum Analysis and Machine Learning for Monthly Rainfall Forecasting," *Applied sciences-basel*, vol. 10, no. 9. MDPI, St Alban-Anlage 66, ch-4052 Basel, Switzerland, May 2020. doi: 10.3390/app10093224.

[38] F. Ghobadi, A. S. Tayerani Charmchi, and D. Kang, "Feature Extraction from Satellite-Derived Hydroclimate Data: Assessing Impacts on Various Neural Networks for Multi-Step Ahead Streamflow Prediction," *Sustainability*, vol. 15, no. 22, p. 15761, Nov. 2023, doi: 10.3390/su152215761.

[39] R. Kumar, M. P. Singh, B. Roy, and A. H. Shahid, "A Comparative Assessment of Metaheuristic Optimized Extreme Learning Machine and Deep Neural Network in Multi-Step-Ahead Long-term Rainfall Prediction for All-Indian Regions," *Water Resour. Manag.*, vol. 35, no. 6, pp. 1927–1960, Apr. 2021, doi: 10.1007/s11269-021-02822-6.

[40] T. O. Muslim *et al.*, "Investigating the Influence of Meteorological Parameters on the Accuracy of Sea-Level Prediction Models in Sabah, Malaysia," *Sustainability*, vol. 12, no. 3, p. 1193, Feb. 2020, doi: 10.3390/su12031193.

[41] J. S. Niño Medina, M. J. Suarez Barón, and J. A. Reyes Suarez, "Application of Deep Learning for the Analysis of the Spatiotemporal Prediction of Monthly Total Precipitation in the Boyacá Department, Colombia," *Hydrology*, vol. 11, no. 8, p. 127, Aug. 2024, doi: 10.3390/hydrology11080127.

[42] X. He, H. Guan, X. Zhang, and C. T. Simmons, "A wavelet-based multiple linear regression model for forecasting monthly rainfall," *Int. J. Climatol.*, vol. 34, no. 6, pp. 1898–1912, May 2014, doi: 10.1002/joc.3809.

[43] T. Wang, M. Zhang, Q. Yu, and H. Zhang, "Comparing the applications of EMD and EEMD on time–frequency analysis of seismic signal," *J. Appl. Geophys.*, vol. 83, pp. 29–34, Aug. 2012, doi: 10.1016/j.jappgeo.2012.05.002.

[44] S. Luo *et al.*, "Forecasting of monthly precipitation based on ensemble empirical mode decomposition and Bayesian model averaging," *Front. EARTH Sci.*, vol. 10, Aug. 2022, doi: 10.3389/feart.2022.926067.

[45] X. Zhang, X. Wu, S. He, and D. Zhao, "Precipitation forecast based on CEEMD–LSTM coupled model," *Water Supply*, vol. 21, no. 8, pp. 4641–4657, Dec. 2021, doi: 10.2166/ws.2021.237.

[46] P. M. A. Castellanos, A. C. Ortegon, and H. F. G. Sierra, "Evaluation of Simple Space Interpolation Methods for the Depth of Precipitation: Application for Boyacá, Colombia," *Adv. Sci. Technol. Eng. Syst. J.*, vol. 5, no. 6, pp. 1322–1327, Dec. 2020, doi: 10.25046/aj0506157.

[47] H. N, S. A. Ahmed, S. Kumar, and A. M, "Computation of the spatio-temporal extent of rainfall and long-term meteorological drought assessment using standardized precipitation index over Kolar and Chikkaballapura districts, Karnataka during 1951-2019," *Remote Sens. Appl. Soc. Environ.*, vol. 27, p. 100768, Aug. 2022, doi: 10.1016/j.rsase.2022.100768.
[48] F. R. Adaryani, S. Jamshid Mousavi, and F. Jafari, "Short-term rainfall forecasting using machine learning-based approaches of PSO-SVR, LSTM and CNN," *J. Hydrol.*, vol. 614, p. 128463, Nov. 2022, doi: 10.1016/j.jhydrol.2022.128463.
[49] A. Adineh, Z. Narimani, and S. Satapathy, "Importance of data preprocessing in time series prediction using SARIMA: A case study," *Int. J. Knowl.-Based Intell. Eng. Syst.*, vol. 24, pp. 331–342, Jan. 2021, doi: 10.3233/KES-200065.
[50] J. Han, M. Kamber, and J. Pei, *Data mining, concepts and techniques*, Third. ELSEVIER.
[51] A. Tawakuli, B. Havers, V. Gulisano, D. Kaiser, and T. Engel, "Survey:Time-series data preprocessing: A survey and an empirical analysis," *J. Eng. Res.*, Mar. 2024, doi: 10.1016/j.jer.2024.02.018.
[52] J. Abbot and J. Marohasy, "Application of artificial neural networks to rainfall forecasting in Queensland, Australia," *Adv. Atmospheric Sci.*, vol. 29, no. 4, pp. 717–730, Jul. 2012, doi: 10.1007/s00376-012-1259-9.
[53] D. P. Chowdhury and U. Saha, "Improvement of extreme rainfall characteristics for disaggregation of rainfall using MMRC with machine learning based DBSCAN clustering algorithm," *Earth Sci. Inform.*, vol. 17, no. 4, pp. 2849–2868, Aug. 2024, doi: 10.1007/s12145-024-01309-3.
[54] P. Deka and U. Saha, "Introduction of k-means clustering into random cascade model for disaggregation of rainfall from daily to 1-hour resolution with improved preservation of extreme rainfall," *J. Hydrol.*, vol. 620, p. 129478, May 2023, doi: 10.1016/j.jhydrol.2023.129478.
[55] S. Kim, O. Kisi, Y. Seo, V. P. Singh, and C.-J. Lee, "Assessment of rainfall aggregation and disaggregation using data-driven models and wavelet decomposition," *Hydrol. Res.*, vol. 48, no. 1, pp. 99–116, Feb. 2017, doi: 10.2166/nh.2016.314.
[56] C. Gaetan, P. Girardi, and V. M. Musau, "Spatial quantile clustering of climate data," *Adv. Data Anal. Classif.*, Feb. 2024, doi: 10.1007/s11634-024-00580-y.
[57] X. Gong and M. B. Richman, "On the Application of Cluster Analysis to Growing Season Precipitation Data in North America East of the Rockies," *J. Clim.*, vol. 8, no. 4, pp. 897–931, 1995, doi: 10.1175/1520-0442(1995)008<0897:OTAOCA>2.0.CO;2.
[58] T. M. Mitchell, *Machine Learning*. in McGraw-Hill International Editions. McGraw-Hill, 1997. [Online]. Available: https://books.google.ca/books?id=EoYBngEACAAJ
[59] Y. Zhang and A. Ye, "Machine learning for precipitation forecasts post-processing — Multi-model comparison and experimental investigation," *J. Hydrometeorol.*, Oct. 2021, doi: 10.1175/JHM-D-21-0096.1.
[60] H. Ahmadi, B. Aminnejad, and H. Sabatsany, "Application of machine learning ensemble models for rainfall prediction," *Acta Geophys.*, vol. 71, no. 4, pp. 1775–1786, Aug. 2023, doi: 10.1007/s11600-022-00952-y.
[61] J. Du, Y. Liu, Y. Yu, and W. Yan, "A Prediction of Precipitation Data Based on Support Vector Machine and Particle Swarm Optimization (PSO-SVM) Algorithms," *Algorithms*, vol. 10, no. 2, p. 57, May 2017, doi: 10.3390/a10020057.
[62] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.
[63] Y. Wang, J. Zhou, K. Chen, Y. Wang, and L. Liu, "Water quality prediction method based on LSTM neural network," in *2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, Nanjing: IEEE, Nov. 2017, pp. 1–5. doi: 10.1109/ISKE.2017.8258814.

[64] G. Ciaburro and B. Venkateswaran, *Neural Networks with R: Smart models using CNN, RNN, deep learning, and artificial intelligence principles*. Packt Publishing Ltd, 2017.
[65] "Evaluation of Adversarial Training on Different Types of Neural Networks in Deep Learning-based IDSs." Accessed: Sep. 17, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9297344
[66] K. Gurney, *An Introduction to Neural Networks*. London: CRC Press, 2017. doi: 10.1201/9781315273570.
[67] S. Kim, T. Suzuki, and Y. Tachikawa, "Rainfall occurrence prediction with convolutional neural network," *J. Jpn. Soc. Civ. Eng. Ser B1 Hydraul. Eng.*, vol. 76, no. 2, p. I_379-I_384, 2020, doi: 10.2208/jscejhe.76.2_I_379.
[68] A. Haidar and B. Verma, "Monthly Rainfall Forecasting Using One-Dimensional Deep Convolutional Neural Network," *IEEE Access*, vol. 6, pp. 69053–69063, 2018, doi: 10.1109/ACCESS.2018.2880044.
[69] X. Shu, W. Ding, Y. Peng, Z. Wang, J. Wu, and M. Li, "Monthly Streamflow Forecasting Using Convolutional Neural Network," *Water resources management*, vol. 35, no. 15. Springer, Van godewijckstraat 30, 3311 gz dordrecht, netherlands, pp. 5089–5104, Dec. 2021. doi: 10.1007/s11269-021-02961-w.
[70] M. Jehanzaib, M. Ajmal, M. Achite, and T.-W. Kim, "Comprehensive Review: Advancements in Rainfall-Runoff Modelling for Flood Mitigation," *Climate*, vol. 10, no. 10, p. 147, Oct. 2022, doi: 10.3390/cli10100147.
[71] S. Ghimire, Z. M. Yaseen, A. A. Farooque, R. C. Deo, J. Zhang, and X. Tao, "Streamflow prediction using an integrated methodology based on convolutional neural network and long short-term memory networks," *Sci. Rep.*, vol. 11, no. 1, p. 17497, Sep. 2021, doi: 10.1038/s41598-021-96751-4.
[72] C. Meo *et al.*, "Extreme Precipitation Nowcasting using Transformer-based Generative Models," Mar. 06, 2024, *arXiv*: arXiv:2403.03929. Accessed: Nov. 09, 2024. [Online]. Available: http://arxiv.org/abs/2403.03929
[73] Y. Tao, X. Gao, A. Ihler, K. Hsu, and S. Sorooshian, "Deep neural networks for precipitation estimation from remotely sensed information," in *2016 IEEE Congress on Evolutionary Computation (CEC)*, Vancouver, BC, Canada: IEEE, Jul. 2016, pp. 1349–1355. doi: 10.1109/CEC.2016.7743945.
[74] R. Fredyan and G. Kusuma, "Spatiotemporal convolutional LSTM with attention mechanism for monthly rainfall prediction," *Commun. Math. Biol. Neurosci.*, 2022, doi: https://doi.org/10.28919/cmbn/7761.
[75] R. Castro, Y. M. Souto, E. Ogasawara, F. Porto, and E. Bezerra, "STConvS2S: Spatiotemporal Convolutional Sequence to Sequence Network for weather forecasting," *Neurocomputing*, vol. 426, pp. 285–298, Feb. 2021, doi: 10.1016/j.neucom.2020.09.060.
[76] N. S. Philip and K. B. Joseph, "On the Predictability of Rainfall in Kerala - An Application of ABF Neural Network," in *Computational Science - ICCS 2001*, vol. 2074, V. N. Alexandrov, J. J. Dongarra, B. A. Juliano, R. S. Renner, and C. J. K. Tan, Eds., in Lecture Notes in Computer Science, vol. 2074. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, pp. 400–408. doi: 10.1007/3-540-45718-6_44.
[77] S. Chattopadhyay and M. Chattopadhyay, "A Soft Computing Technique in rainfall forecasting," 2007.
[78] S. Chattopadhyay, "Feed forward Artificial Neural Network model to predict the average summer-monsoon rainfall in India," *Acta Geophys.*, vol. 55, no. 3, pp. 369–382, Sep. 2007, doi: 10.2478/s11600-007-0020-8.
[79] M. Nasseri, K. Asghari, and M. J. Abedini, "Optimized scenario for rainfall forecasting using genetic algorithm coupled with artificial neural network," *Expert Syst. Appl.*, vol. 35, no. 3, pp. 1415–1421, Oct. 2008, doi: 10.1016/j.eswa.2007.08.033.

[80] K. K. Htike and O. O. Khalifa, "Rainfall forecasting models using focused time-delay neural networks," in *International Conference on Computer and Communication Engineering (ICCCE'10)*, Kuala Lumpur, Malaysia: IEEE, May 2010, pp. 1–6. doi: 10.1109/ICCCE.2010.5556806.

[81] C. L. Wu, K. W. Chau, and C. Fan, "Prediction of rainfall time series using modular artificial neural networks coupled with data-preprocessing techniques," *J. Hydrol.*, vol. 389, no. 1–2, pp. 146–167, Jul. 2010, doi: 10.1016/j.jhydrol.2010.05.040.

[82] C. L. Wu and K. W. Chau, "Prediction of rainfall time series using modular soft computing methods," *Eng. Appl. Artif. Intell.*, vol. 26, no. 3, pp. 997–1007, Mar. 2013, doi: 10.1016/j.engappai.2012.05.023.

[83] V. Singh, "Time Series Analysis of Forecasting Indian Rainfall," vol. 3, no. 1, p. 4, Apr. 2014.

[84] A. H. Salimi, J. Masoompour Samakosh, E. Sharifi, M. R. Hassanvand, A. Noori, and H. Von Rautenkranz, "Optimized Artificial Neural Networks-Based Methods for Statistical Downscaling of Gridded Precipitation Data," *Water*, vol. 11, no. 8, p. 1653, Aug. 2019, doi: 10.3390/w11081653.

[85] NASA, "MODIS Web." modis.gsfc.nasa.gov, Nov. 2023. [Online]. Available: https://modis.gsfc.nasa.gov/about/

[86] L. Tao, X. He, J. Li, and D. Yang, "A multiscale long short-term memory model with attention mechanism for improving monthly precipitation prediction," *J. Hydrol.*, vol. 602, p. 126815, Nov. 2021, doi: 10.1016/j.jhydrol.2021.126815.

[87] K. Hsu, X. Gao, S. Sorooshian, and H. V. Gupta, "Precipitation Estimation from Remotely Sensed Information Using Artificial Neural Networks," *J. Appl. Meteorol.*, vol. 36, no. 9, pp. 1176–1190, Sep. 1997, doi: 10.1175/1520-0450(1997)036<1176:PEFRSI>2.0.CO;2.

[88] M. Sadeghi *et al.*, "PERSIANN-CNN: Precipitation Estimation from Remotely Sensed Information Using Artificial Neural Networks–Convolutional Neural Networks," *J. Hydrometeorol.*, vol. 20, no. 12, pp. 2273–2289, Dec. 2019, doi: 10.1175/JHM-D-19-0110.1.

[89] F. Chollet, *Deep Learning with Python*. Manning, 2017.

[90] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994, doi: 10.1109/72.279181.

[91] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[92] R. C. Staudemeyer and E. R. Morris, "Understanding LSTM -- a tutorial into Long Short-Term Memory Recurrent Neural Networks," Sep. 12, 2019, *arXiv*: arXiv:1909.09586. Accessed: Dec. 16, 2023. [Online]. Available: http://arxiv.org/abs/1909.09586

[93] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the Properties of Neural Machine Translation: Encoder-Decoder Approaches," Oct. 07, 2014, *arXiv*: arXiv:1409.1259. doi: 10.48550/arXiv.1409.1259.

[94] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," Dec. 2014.

[95] L. Parviz, K. Rasouli, and A. Torabi Haghighi, "Improving Hybrid Models for Precipitation Forecasting by Combining Nonlinear Machine Learning Methods," *Water resources management. Springer, van godewijckstraat 30, 3311 gz dordrecht, netherlands*. SPRINGER, VAN GODEWIJCKSTRAAT 30, 3311 GZ DORDRECHT, NETHERLANDS, May 26, 2023. doi: 10.1007/s11269-023-03528-7.

[96] G. Li, X. Ma, and H. Yang, "A Hybrid Model for Monthly Precipitation Time Series Forecasting Based on Variational Mode Decomposition with Extreme Learning Machine," *Information*, vol. 9, no. 7. MDPI, St Alban-Anlage 66, ch-4052 Basel, Switzerland, Jul. 2018. doi: 10.3390/info9070177.

[97] J. Piri, M. Abdolahipour, and B. Keshtegar, "Advanced Machine Learning Model for Prediction of Drought Indices using Hybrid SVR-RSM," *WATER RESOURCES MANAGEMENT*, vol. 37, no. 2. SPRINGER, VAN GODEWIJCKSTRAAT 30, 3311 GZ DORDRECHT, NETHERLANDS, pp. 683–712, Jan. 2023. doi: 10.1007/s11269-022-03395-8.

[98] X. Zhang, D. Zhao, T. Wang, X. Wu, and B. Duan, "A novel rainfall prediction model based on CEEMDAN-PSO-ELM coupled model," *Water supply*, vol. 22, no. 4. Iwa publishing, Republic-Export bldg, units 1 04 & 1 05, 1 Clove Crescent, London, England, pp. 4531–4543, Apr. 2022. doi: 10.2166/ws.2022.115.

[99] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*. in Wiley Series in Probability and Statistics. Wiley, 2015. [Online]. Available: https://books.google.com.co/books?id=rNt5CgAAQBAJ

[100] C. Chatfield, *The Analysis of Time Series: An Introduction, Sixth Edition*. in Chapman & Hall/CRC Texts in Statistical Science. CRC Press, 2016. [Online]. Available: https://books.google.com.co/books?id=qKzyAbdaDFAC

[101] H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Deep learning for time series classification: a review," *Data Min. Knowl. Discov.*, vol. 33, no. 4, pp. 917–963, Jul. 2019, doi: 10.1007/s10618-019-00619-1.

[102] G. Kronberger, L. Kammerer, and M. Kommenda, "Identification of Dynamical Systems using Symbolic Regression," vol. 12013, 2020, doi: 10.1007/978-3-030-45093-9.

[103] S. d'Ascoli, S. Becker, A. Mathis, P. Schwaller, and N. Kilbertus, "ODEFormer: Symbolic Regression of Dynamical Systems with Transformers," Oct. 09, 2023, *arXiv*: arXiv:2310.05573. Accessed: Nov. 09, 2024. [Online]. Available: http://arxiv.org/abs/2310.05573

[104] F. Zennaro *et al.*, "Exploring machine learning potential for climate change risk assessment," *Earth-Sci. Rev.*, vol. 220, p. 103752, Sep. 2021, doi: 10.1016/j.earscirev.2021.103752.

[105] A. Kumar and M. Jain, "Why Ensemble Techniques Are Needed," in *Ensemble Learning for AI Developers: Learn Bagging, Stacking, and Boosting Methods with Use Cases*, Berkeley, CA: Apress, 2020, pp. 1–10. doi: 10.1007/978-1-4842-5940-5_1.

[106] O. I. Higuera Martínez, L. Fernández-Samacá, and L. F. Serrano Cárdenas, "Trends and opportunities by fostering creativity in science and engineering: a systematic review," *Eur. J. Eng. Educ.*, vol. 46, no. 6, pp. 1117–1140, Nov. 2021, doi: 10.1080/03043797.2021.1974350.

[107] M. El Hafyani, K. El Himdi, and S.-E. El Adlouni, "Improving monthly precipitation prediction accuracy using machine learning models: a multi-view stacking learning technique," *Front. Water*, vol. 6, May 2024, doi: 10.3389/frwa.2024.1378598.

[108] Z. Shen and W. Ban, "Machine learning model combined with CEEMDAN algorithm for monthly precipitation prediction," *Earth science informatics*. Springer Heidelberg, Tiergartenstrasse 17, D-69121 Heidelberg, Germany, Apr. 26, 2023. doi: 10.1007/s12145-023-01011-w.

[109] Z. Zhou, J. Ren, X. He, and S. Liu, "A comparative study of extensive machine learning models for predicting long-term monthly rainfall with an ensemble of climatic and meteorological predictors," *Hydrological processes*, vol. 35, no. 11. WILEY, 111 RIVER ST, HOBOKEN 07030-5774, NJ USA, Nov. 2021. doi: 10.1002/hyp.14424.

[110] O. Zandi, B. Zahraie, M. Nasseri, and A. Behrangi, "Stacking machine learning models versus a locally weighted linear model to generate high-resolution monthly precipitation over a topographically complex area," *Atmospheric Res.*, vol. 272, p. 106159, 2022, doi: https://doi.org/10.1016/j.atmosres.2022.106159.

[111] C. Song, X. Chen, P. Wu, and H. Jin, "Combining time varying filtering based empirical mode decomposition and machine learning to predict precipitation from nonlinear series," *J. Hydrol.*, vol. 603, p. 126914, 2021, doi: https://doi.org/10.1016/j.jhydrol.2021.126914.

[112] T. Tang, D. Jiao, T. Chen, and G. Gui, "Medium- and Long-Term Precipitation Forecasting Method Based on Data Augmentation and Machine Learning Algorithms," *IEEE journal of selected topics in applied earth observations and remote sensing*, vol. 15. IEEE-Inst Electrical Electronics Engineers Inc, 445 Hoes Lane, Piscataway, Nj 08855-4141 USA, pp. 1000–1011, 2022. doi: 10.1109/JSTARS.2022.3140442.

[113] P. K. Yeditha, G. S. Anusha, S. S. S. Nandikanti, M. Rathinasamy, and P. Kucera, "Development of Monthly Scale Precipitation-Forecasting Model for Indian Subcontinent using Wavelet-Based Deep Learning Approach," *WATER*, vol. 15, no. 18, Sep. 2023, doi: 10.3390/w15183244.

[114] G. Papacharalampous, H. Tyralis, N. Doulamis, and A. Doulamis, "Ensemble Learning for Blending Gridded Satellite and Gauge-Measured Precipitation Data," *Remote Sens.*, vol. 15, no. 20, Oct. 2023, doi: 10.3390/rs15204912.

[115] F. Esmaeili, S. Shabanlou, and M. Saadat, "A wavelet-outlier robust extreme learning machine for rainfall forecasting in Ardabil City, Iran," *Earth science informatics*, vol. 14, no. 4. Springer Heidelberg, Tiergartenstrasse 17, D-69121 Heidelberg, Germany, pp. 2087–2100, Dec. 2021. doi: 10.1007/s12145-021-00681-8.

[116] J. A. Anochi, V. A. de Almeida, and H. F. de Campos Velho, "Machine Learning for Climate Precipitation Prediction Modeling over South America," *Remote Sens.*, vol. 13, no. 13, Art. no. 13, Jan. 2021, doi: 10.3390/rs13132468.

[117] R. E. N. Macabiog and J. C. Dela Cruz, "Rainfall Predictive Approach for La Trinidad, Benguet using Machine Learning Classification," in *2019 IEEE 11th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*, Nov. 2019, pp. 1–6. doi: 10.1109/HNICEM48295.2019.9072761.

[118] M. Chhetri, S. Kumar, P. Pratim Roy, and B.-G. Kim, "Deep BLSTM-GRU Model for Monthly Rainfall Prediction: A Case Study of Simtokha, Bhutan," *Remote Sens.*, vol. 12, no. 19, p. 3174, Sep. 2020, doi: 10.3390/rs12193174.

[119] O. Coskun and H. Citakoglu, "Prediction of the standardized precipitation index based on the long short-term memory and empirical mode decomposition-extreme learning machine models: The Case of Sakarya, Turkiye," *Physics and chemistry of the earth*, vol. 131. Pergamon-elsevier science ltd, The Boulevard, Langford Lane, Kidlington, Oxford Ox5 1Gb, England, Oct. 2023. doi: 10.1016/j.pce.2023.103418.

[120] Y. Li *et al.*, "Deterministic and probabilistic evaluation of raw and post-processing monthly precipitation forecasts: a case study of China," *Journal of hydroinformatics*, vol. 23, no. 4. Iwa Publishing, Republic-Export Bldg, Units 1 04 & 1 05, 1 Clove Crescent, London, England, pp. 914–934, Jul. 2021. doi: 10.2166/hydro.2021.176.

[121] M. M. Hossain, A. H. M. F. Anwar, N. Garg, M. Prakash, and M. Bari, "Monthly Rainfall Prediction at Catchment Level with the Facebook Prophet Model Using Observed and CMIP5 Decadal Data," *Hydrology*, vol. 9, no. 6, p. 111, Jun. 2022, doi: 10.3390/hydrology9060111.

[122] G. Papacharalampous, H. Tyralis, A. Doulamis, and N. Doulamis, "Comparison of Machine Learning Algorithms for Merging Gridded Satellite and Earth-Observed Precipitation Data," *Water*, vol. 15, no. 4, p. 634, Feb. 2023, doi: 10.3390/w15040634.

[123] M. Pakdaman, I. Babaeian, and L. M. Bouwer, "Improved Monthly and Seasonal Multi-Model Ensemble Precipitation Forecasts in Southwest Asia Using Machine Learning Algorithms," *Water*, vol. 14, no. 17. MDPI, St Alban-Anlage 66, Ch-4052 Basel, Switzerland, Sep. 2022. doi: 10.3390/w14172632.