TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING
PULCHOWK CAMPUS

Thesis No. 078/MSCCD/005

COMPARATIVE ANALYSIS OF LONG SHORT-TERM MEMORY (LSTM) AND MULTI-LAYER PERCEPTRON (MLP) MODELS FOR RIVER RUNOFF PREDICTION IN THE HINDU KUSH HIMALAYAN REGION

By

Hansal Shrestha

A THESIS
SUBMITTED TO THE DEPARTMENT OF DEPARTMENT OF APPLIED SCIENCE AND CHEMICAL ENGINEERING
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER IN
CLIMATE CHANGE AND DEVELOPMENT ENGINEERING

DEPARTMENT OF APPLIED SCIENCE AND CHEMICAL ENGINEERING
PULCHOWK, LALITPUR

DECEMBER, 2023

COPYRIGHT

DECLARATION

I declare that the work hereby submitted for Master in Climate Change and Development Engineering (MSCCD) at IOE, Pulchowk Campus entitled "Comparative Analysis of Long Short-Term Memory (LSTM) and Multi-Layer Perceptron (MLP) Models for River runoff prediction in The Hindu Kush Himalaya Region" is my own work and has not been previously submitted by me at any university for any academic award. I authorize IOE, Pulchowk Campus to lend this thesis to other institution or individuals for the purpose of scholarly research.

Hansal Shrestha

078MSCCD005

1st December, 2023

TRIBHUVAN UNIVERSITY

INSTITUTE OF ENGINEERING

PULCHOWK CAMPUS

DEPARTMENT OF APPLIED SCIENCE AND CHEMICAL ENGINEERING

The undersigned certify that they have read, and recommended to the Institute of Engineering for acceptance, a thesis entitled " **Comparative Analysis of Long Short-Term Memory (LSTM) and Multi-Layer Perceptron (MLP) Models for River runoff Prediction in The Hindu Kush Himalaya Region** " submitted by **Hansal Shrestha (078MSCCD005)** in partial fulfillment of the requirements for the degree of Master in Climate Change and Development of Tribhuvan University.

Asst. Prof. Neeraj Adhikari
Supervisor
Department of Mechanical
and Aerospace Engineering

Asst. Prof. Dr. Basanta Joshi
External Examiner
Department of Electronics and
Computer Engineering

Prof. Dr. Rinita Rajbhandari
Program Coordinator
Climate Change and Development Program
Department of Applied Science
and Chemical Engineering

Prof. Dr. Hem Raj Panta
Head of Department,
Department of Applied Science
and Chemical Engineering

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Hydrological forecasting in the Hindu Kush Himalayas (HKH) presents special challenges because of the complex interplay between climatic and environmental factors. The quantitative predictive capabilities of two well-established models, Long Short-Term Memory (LSTM) and Multi-Layer Perceptron (MLP), chosen for their proven performance in previous studies, are meticulously compared in this thesis. The analysis uses comprehensive data spanning 2001 to 2013, including discharge records from the Department of Hydrology and Meteorology (DHM), precipitation data from APHRODITE, temperature data from APHRODITE, and snow cover area information from Google Earth Engine with MOD09A1 V6.1. The study employs rigorous evaluation metrics, revealing nuanced insights into the hydrological processes. Contrary to expectations, the MLP model exhibited slight superiority, showcasing a nuanced understanding of the region's complexities. The quantitative assessment, including RMSE (LSTM: 0.2396, MLP: 0.1733), MAE (LSTM: 0.1698, MLP: 0.0841), R^2 Score (LSTM: 0.9976, MLP: 0.9987), and NSE (LSTM: 0.9976, MLP: 0.9987), emphasizes the indispensable role of robust predictive models, showcasing the necessity of reliable models for enhancing accurate river runoff predictions crucial for effective water resource management and flood preparedness in challenging terrains like the HKH.

Keywords

Long Short-Term Memory (LSTM), Multi-Layer Perceptron (MLP), Hindu Kush Himalayan region (HKH)

LIST OF ABBREVATIONS

DD : Data Driven

ML : Machine Learning

HKH : Hindu Kush Himalayan region

SWAT: Soil and Water Assessment Tool

ANN : Artificial Neural Network

WEAP: Water Evaluation And Planning System

GR2M: Gridded Reservoir-Runoff Model

LSTM : Long Short-Term Memory

MLP : Multi-Layer Perceptron

RMSE : Root Mean Squared Error

MAE : Mean Absolute Error

R^2 : R-squared Score

NSE : Nash-Sutcliffe Efficiency

SVM : Support Vector Machines

CHAPTER ONE: INTRODUCTION

1.1 Background

The impacts stemming from the escalating global warming and climate change phenomenon have ushered in a heightened frequency and severity of both drought and flooding events, thereby constituting one of the most formidable challenges confronting our aquatic ecosystems ((UNDP), 2013). A corollary of these environmental shifts is the notable reduction in runoff during dry periods, an occurrence that bears the potential to trigger acute water scarcity, imperiling a spectrum of essential functions encompassing domestic, industrial, hydroelectric, and agricultural irrigation demands (Jain, Goswami, & Saraf, 2009)This cascading effect exerts substantial stress upon existing water infrastructure, straining their capacities beyond designed thresholds. Furthermore, the intricate interplay between global warming and water resources exacerbates uncertainties afflicting the long-term projections governing urban water demand, thereby underscoring the exigency for vigilant planning and management (Urich & Rauch, 2014).

Amidst these unfolding dynamics, the Hindu Kush Himalaya (HKH) region, colloquially dubbed as "The third pole," looms prominently, boasting an immense stockpile of snow resources rivaling its polar counterparts (S. Singh, Bassignana-Khadka, Karky, & Sharma, 2011). However, this very region stands ensnared in the throes of climate change-induced jeopardy. Simultaneously, the repercussive repercussions of reduced dry season runoff engender a palpable crisis, beckoning water scarcity at the doorstep of households, industries, hydroelectric enterprises, and agricultural endeavors (Jain et al., 2009). While the imprint of climate change is discernible across the HKH expanse, the intricate patterns of this transformation evade uniformity, ushering in disparities of direction and magnitude (Pandey, Dhaubanjar, Bharati, & Thapa, 2020). It is this divergent landscape of uncertainties that often yields significant repercussions in domains beyond the immediate purview of the aquatic realm, notably affecting supply chains, operational dynamics, and associated costs a challenge that conventional planning paradigms are ill-equipped to surmount.

In light of these exigencies, the imperative emerges for robust hydrological modeling and incisive climate change impact assessments, functions that substantiate the bedrock of a sustainable watershed management strategy. By unraveling the intricate fabric of

hydrological processes and fostering an astute comprehension of climate-induced transformations, such modeling endeavors engender a pragmatic foundation for both policy formulation and adaptation strategies. It is through these prescient measures that the trajectory of future climate-induced impacts upon water resources may be anticipated, managed, and ultimately navigated with resilience and efficacy.

Many of the studies conducted in the realm of hydrological modeling have traditionally leaned towards the utilization of conceptual degree-day models or physical energy-balance models, as evidenced by the works of (Immerzeel, van Beek, Konz, Shrestha, & Bierkens, 2012; Shrestha, Shrestha, & Babel, 2015; L. Singh & Saravanan, 2020). These established models, while valuable in their own right, operate within the confines of well-defined physical processes. However, the emergence of data-driven (DD) models, epitomized by Machine Learning (ML) techniques, has opened up new avenues for modeling hydrological systems. DD models possess the remarkable ability to replicate intricate non-linear systems by discerning patterns between input and output variables, bypassing the need for an in-depth comprehension of the underlying physical mechanisms (ASCE). This departure from the conventional modeling paradigm signifies a shift towards harnessing the power of artificial intelligence to unravel the complexities of hydrological behavior.

Numerous studies have emphatically substantiated the superiority of ML models over their conventional hydrological counterparts. Research endeavors such as (Uysal, Şensoy, & Şorman, 2016) work on River runoff Model (SRM), (Pradhan, Tingsanchali, & Shrestha, 2020) exploration of Soil and Water Assessment Tool (SWAT), and investigations into Water Evaluation And Planning System (WEAP) and Gridded Reservoir-Runoff Model (GR2M) by (Farfán, Palacios, Ulloa, & Avilés, 2020) have collectively underscored the aptitude of Artificial Neural Network (ANN) models as viable alternatives to traditional methodologies in the realm of hydrological modeling. The culmination of these findings not only solidifies the standing of ANN models within the hydrological domain but also positions them as potential transformative tools, offering simplicity and untapped potential for advancing our understanding of hydrological processes through the integration of artificial intelligence. Consequently, this research initiative seeks to pave the way for enhanced insights into the symbiotic

relationship between AI and hydrology, propelling the field towards innovative realms of predictive accuracy and holistic comprehension.

Adopting physically-based methods could be intimidating for people who are new to hydrological modeling, especially if they do not have access to complex and state-of-the-art climate models. For many, the intricacy and resource-intensiveness of these physical models can be a turnoff. On the other hand, data-driven modeling presents a viable substitute that, with proper application, can lessen the drawbacks of conventional techniques. This claim is especially relevant when considering nations such as Nepal, where the field of hydrology is typified by distinct difficulties and complexities. As a result, the current study is grounded in the fluid hydrological landscape of Nepal and aims to clarify the influence of artificial intelligence (AI)-driven prediction methods on the country's hydrological knowledge. Through utilizing data-driven models, this study aims to open up new avenues for a more readable and perceptive understanding of hydrological processes in Nepal and elsewhere.

1.1.1 Primary Objective:

- To evaluate and compare the performance of Multi-Layer Perceptron (MLP) and Long Short-Term Memory (LSTM) models in predicting daily river runoff in the Lamgtang basin, Central Himalayas.

1.1.2 Specific Objectives:

- To preprocess and transform hydro meteorological data including Snow Cover Area (SCA), temperature, precipitation, and discharge for model input.
- To configure and trail LSTM and MLP models for river runoff prediction, considering different hyperparameters.
- To compare the predictive accuracy of the MLP and LSTM models using performance metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Nash-Sutcliffe Efficiency (NSE).

1.2 Problem Statement

The Hindu Kush Himalayan (HKH) region presents a formidable challenge for hydrological forecasting due to its intricate topography and diverse climatic conditions. Because of the numerous interacting factors, traditional models struggle to capture the

complexities of river runoff in this region. Inaccurate streamflow predictions during snowmelt have serious consequences for water resource management and flood preparedness. Recognizing the limitations of traditional approaches, this study aims to address the pressing issue of insufficient hydrological forecasting in the HKH region. The study compares the performance of two advanced machine learning models in predicting river runoff, Long Short-Term Memory (LSTM) and Multi-Layer Perceptron (MLP). The study makes use of extensive datasets from 2001 to 2012, including discharge, precipitation, temperature, and snow cover area. This research aims to contribute practical insights and solutions to improve the accuracy of hydrological forecasts by delving into the unique hydrological processes of the HKH region, recognizing the critical importance of reliable predictions for effective water resource management in complex terrains.

1.3 Research Questions

1.3.1 How does the performance of the Multi-Layer Perceptron (MLP) model compare to that of the Long Short-Term Memory (LSTM) model in predicting daily river runoff in the Lamgtang basin, Central Himalayas?

1.3.2 How effectively can the MLP model capture the complex non-linear relationships between input variables (Snow Cover Area, temperature, precipitation, and antecedent discharge) and snowmelt runoff?

1.3.3 To what extent does the LSTM model excel in capturing temporal dependencies and long-term patterns in river runoff prediction, compared to the MLP model?

1.3.4 Which model, MLP or LSTM, demonstrates superior performance based on key evaluation metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Nash-Sutcliffe Efficiency (NSE)?

By addressing these research questions, my thesis will delve into the nuanced aspects of MLP and LSTM models for river runoff prediction and contribute valuable insights to the field of hydrological modeling in challenging mountainous environments.

1.4 Significance of the Study

This research has important implications for water resource management in snow-dominated Himalayan basins. Accurate forecasting of river runoff is critical for efficient planning and resource management in regions that rely heavily on snowmelt

for freshwater supply. In predicting runoff, the comparison of Multi-Layer Perceptron (MLP) and Long Short-Term Memory (LSTM) models provides valuable insights into selecting the most reliable and accurate forecasting model. Furthermore, given the Himalayan region's vulnerability to climate change, the findings of the study help to understand and forecast river runoff patterns under changing climate scenarios. This knowledge can be used to inform adaptive strategies and policies to address the effects of climate change.

Furthermore, the evaluation of MLP and LSTM models under various hydrological conditions, including peak flow events, improves operational forecasting systems, allowing for early flood warning and minimizing damage during extreme events. The models' ability to alleviate data scarcity issues in remote Himalayan terrain demonstrates their practical utility, providing reliable predictions despite limited ground observations. Furthermore, the study emphasizes the potential of advanced technologies in addressing critical water resource challenges by utilizing cutting-edge machine learning techniques. The comparison of MLP and LSTM models adds to the existing literature by revealing the advantages and disadvantages of each approach in the context of river runoff prediction.

Finally, the study's findings have practical implications for practitioners, policymakers, and water resource managers, guiding them in the selection of the most appropriate modeling approach for accurate and reliable river runoff forecasts. As a result, informed decision-making is facilitated, and sustainable water resource management practices are promoted in the Himalayan region.

1.5 Scope and Limitations

1.5.1 **Scope:**

In the Langtang basin in the Central Himalayas, the study evaluates the Long Short-Term Memory (LSTM) and Multi-Layer Perceptron (MLP) models for daily river runoff prediction. As model inputs, it focuses on important hydrological variables such as temperature, precipitation, antecedent discharge, and snow cover area. The study is specifically designed to address the unique problems of the Himalayan region, such as

the lack of data. The study also looks into how different hyperparameters affect the performance of the model, providing information on how to optimize settings for precise hydrological forecasting.

1.5.2 **Limitations:**

Several factors temper the study's findings' generalizability and reliability. To begin with, the conclusions are limited to the Langtang basin and its specific hydroclimatic conditions; extrapolating these findings to other regions or diverse hydrological regimes requires cautious validation. The accuracy of the study is dependent on the quality and availability of input data, which includes snow cover area, temperature, precipitation, and discharge data. Errors or limitations in these data sources can have an impact on the MLP and LSTM models' performance. Furthermore, the models' reliance on assumed relationships derived from historical data raises concerns about their ability to adapt to abrupt changes in climate patterns or land use practices, which could affect prediction accuracy.

Despite MLP and LSTM models' powerful capabilities in capturing complex nonlinear relationships, their sensitivity to model complexity and the risk of overfitting highlight the importance of careful tuning to avoid unrealistic predictions. These machine learning models, unlike physically-based models, lack inherent insights into underlying hydrological processes and rely solely on learned patterns, which may limit comprehension. The research's investigation of the impact of hyperparameters acknowledges the need for tailored configurations, but the lack of an uniform guarantee raises concerns about optimal performance under different conditions or datasets.

CHAPTER TWO: LITERATURE REVIEW

2.1 Hydrological Modeling Approaches

Hydrological modeling is the process of simulating the behavior of the water cycle in a given watershed or catchment area. It involves using mathematical and computational methods to represent the movement and distribution of water in various components of the hydrological cycle, such as precipitation, runoff, evaporation, and infiltration. There are several approaches to hydrological modeling, each with its own strengths, limitations, and underlying assumptions. Here, I'll explain three main approaches: conceptual models, physically based models, and data-driven models.

2.1.1 **Conceptual Models:** Conceptual models are simplified representations of the real-world hydrological processes. They are based on empirical relationships and simple equations that describe the movement of water through various components of the hydrological cycle. These models do not require detailed knowledge of the physical characteristics of the watershed. Instead, they rely on calibrated parameters derived from historical data to mimic observed hydrological behavior. Conceptual models (sometimes called gray-box models) consider physical laws but in highly simplified forms. A conceptual model is a descriptive representation of hydrologic system that incorporates the modeler’s understanding of the relevant physical, chemical, and hydrologic conditions (Liu, Wang, Xu, & Duan, 2017).

The use of conceptual hydrological models has various advantages in the assessment of water resources. Notably, their ease of use and simplicity make them accessible instruments for hydrologists and academics. Conceptual models are especially useful in settings with restricted data availability since they require fewer data inputs than their physically based equivalents. These models are invaluable for short assessments and exploratory investigations, giving a realistic option for first hydrological system evaluations. It is, nonetheless, critical to recognize their inherent limitations. Conceptual models may struggle to adequately convey the intricacies of certain physical processes, potentially resulting in forecast inaccuracies. Furthermore, when dealing with ungauged basins or changing environmental circumstances, their application is limited. Another disadvantage of these models is their spatial variability, stressing the importance of carefully considering their scope and applicability in specific hydrological environments.

2.1.2 **Physically Based Models:** Models based on physical processes aim to represent the underlying physical processes of the hydrological cycle. These models include detailed mathematical equations that describe how water moves through various components of the watershed while taking terrain, soil properties, vegetation, and climate inputs into account. Physically based models frequently necessitate detailed input data and are more difficult to set up and calibrate. Physical-based models, on the other hand (also known as white-box models or theoretical models), describe hydrological processes in detail by solving differential equations describing the physical laws of mass, energy, and momentum conservation. These equations are typically solved over some kind of grid structure that represents a spatial domain. Therefore, physically based models are often called distributed hydrological models (Liu et al., 2017).

Physically based hydrological models have significant advantages since they can accurately describe intricate physical processes inside hydrological systems. One major advantage is their ability to mimic a wide range of hydrological conditions, making them adaptable instruments for in-depth analyses. Because they can represent the intricacies involved with changing environmental circumstances, these models are particularly well-suited for long-term studies of climate change implications on hydrological systems. However, it is critical to acknowledge the difficulties involved with physically based models. They are typically data-intensive and computationally demanding, necessitating significant resources for implementation. The setup and calibration processes are difficult, necessitating careful thought and experience. Furthermore, because these models are sensitive to uncertainties in both input data and model parameters, stringent validation and calibration methods are essential.

2.1.3 **Data-Driven Models:** Machine learning models, also known as data-driven models, learn patterns from historical data without explicitly modeling physical processes. These models make use of statistical algorithms to establish relationships between input variables (like precipitation and temperature) and output variables (like runoff). Machine learning models include artificial neural networks (ANN), support vector machines (SVM), decision trees, and other techniques. Though conceptual and physics-based models provided greater accuracy in hydrograph modeling, many issues remained to be addressed by many researchers. Those difficulties include

implementation and calibration difficulty, the vast amount of calibration data and the need of sophisticated tools etc. (Renji Remesan 2015).

Machine Learning (ML) models offer particular advantages in hydrological modeling due to their ability to capture complicated non-linear interactions. Their strength is their ability to manage huge and different datasets, allowing them to accommodate various forms of information. ML models rely less on in-depth knowledge of intricate physical processes, making them especially beneficial for dealing with complex hydrological systems. However, these models are not without limits. The interpretability of their outputs is frequently limited, which can make it difficult to derive important insights. Overfitting is a potential issue, especially when models are not properly regularized to avoid overfitting the training data. Furthermore, the quality and quantity of input data have a substantial impact on the performance of ML models, underlining the need of data preparation and feature selection for optimal results.

Each approach has a place in hydrological modeling, and the approach chosen is determined by the specific research goals, available data, and level of detail required. Many modern studies combine these approaches to capitalize on their respective strengths and achieve more accurate predictions of hydrological processes.

2.2 Machine Learning Models in Hydrological Forecasting

2.2.1 Overview of machine learning algorithms

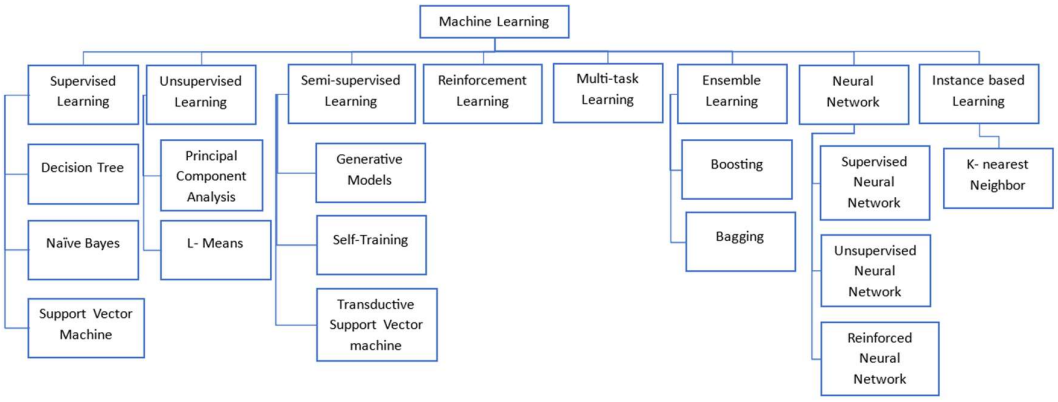

Figure 1-2.2.1: Overview of machine learning algorithms

Source: Machine Learning Algorithms - A Review (Mahesh, 2019)

The overview of machine learning algorithms involves understanding the types of algorithms used for various tasks. Machine learning algorithms can be categorized into three main types: supervised learning, unsupervised learning, and reinforcement learning.

2.2.1.1 Supervised Learning: In this category, the algorithm learns from labeled training data. It is provided with input-output pairs, allowing it to make predictions or classify new, unseen data. Common algorithms include:

- Linear Regression
- Decision Trees
- Random Forest
- Support Vector Machines
- K-Nearest Neighbors
- Neural Networks

2.2.1.2 Unsupervised Learning: Here, the algorithm works with unlabeled data and attempts to find patterns, structures, or relationships within the data. It's often used for clustering and dimensionality reduction. Common algorithms include:

- K-Means Clustering
- Hierarchical Clustering
- Principal Component Analysis (PCA)
- t-Distributed Stochastic Neighbor Embedding (t-SNE)

2.2.1.3 Reinforcement Learning: This type of learning involves training an agent to interact with an environment and learn by trial and error, receiving rewards or penalties for its actions. Common algorithms include:

- Q-Learning
- Deep Q Networks (DQN)
- Policy Gradient Methods
- Proximal Policy Optimization (PPO)

These algorithms are further subdivided into techniques and variations, each suited to a different type of problem or dataset. Machine learning algorithms are widely used for tasks such as classification, regression, clustering, anomaly detection, and others. The algorithm chosen is determined by the nature of the data, the problem at hand, and the desired outcome.

2.3 Overview of MLP and LSTM architectures

2.3.1 Multi-Layer Perceptron

MLPs are a type of artificial neural network with a layered architecture, comprising input, hidden, and output layers. Neurons in each layer are interconnected through weighted connections, and activation functions introduce non-linearity (Gardner & Dorling, 1998). The training process involves forward and backward propagation, adjusting weights via gradient descent. Hyperparameters influence performance, requiring careful tuning.

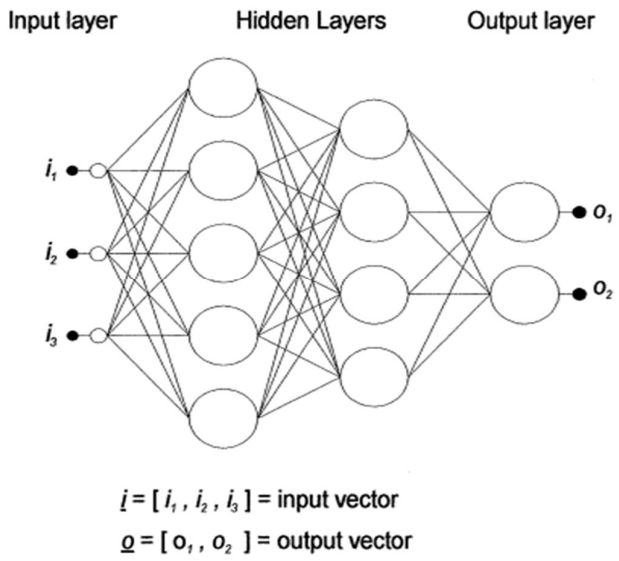

Figure 2-2.3.1: A Multi Layer Perceptron with Two Hidden Layer

Source: ARTIFICIAL NEURAL NETWORKS (THE MULTILAYER PERCEPTRON)—A REVIEW OF APPLICATIONS IN THEATMOSPHERIC SCIENCES (Gardner & Dorling, 1998)

The multilayer perceptron (MLP) is composed of interconnected nodes that model a nonlinear mapping between input and output vectors. The nodes use weights and activation functions to process inputs and produce outputs. This composition of nonlinear functions allows the MLP to approximate complex functions. The common logistic function is often used as the activation function due to its easy-to-compute derivative. The MLP's architecture involves feed-forward processing, passing input through layers of neurons. The input layer transfers data, while hidden layers process information, and the output layer generates the final result. MLPs can be fully connected, with each node linked to all nodes in adjacent layers. They can approximate any smooth function between input and output vectors. Learning occurs through supervised training with training data, adjusting weights to achieve desired input-output relationships. Various algorithms can be used for training. Once trained, MLPs can generalize to new data (Gardner & Dorling, 1998).

2.3.2 Long Short-Term Memory

LSTM was first proposed in the 1990s (Sepp Hochreiter, 1997) but it’s true potential has recently been recognized. The LSTM addresses traditional RNN weaknesses in

learning long-term dependencies. The hydrological community's perspective is integrated to align with deep learning research. The LSTM is a specialized recurrent neural network (RNN) designed to overcome memory limitations in traditional RNNs. Unlike RNNs, LSTMs can remember sequences beyond 10 steps(Kratzert, Klotz, Brenner, Schulz, & Herrnegger, 2018).

In traditional RNNs, a single internal state exists and is recalculated with each step. LSTM, however, introduces a cell state for information storage and gates for information control. These gates include the forget gate, determining forgotten elements, and the input gate, controlling cell state updates. The cell state is updated through element-wise operations based on gate outputs. An output gate controls information flow into a new hidden state. The cell state enables effective learning of long-term dependencies, avoiding gradient issues (Kratzert et al., 2018).

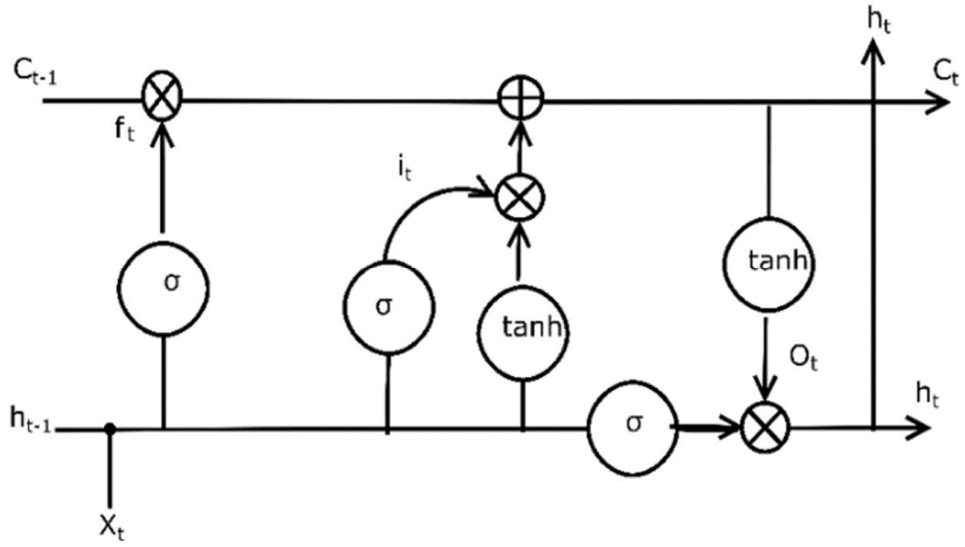

Figure 3-2.3.2: A LSTM Cell

Source: Impact of climate change on river runoff in a Himalayan basin, Nepal (S. Thapa et al., 2021)

The LSTM layer's pseudocode processes input sequences step by step, with stacked layers feeding outputs to the next layer. The final discharge prediction is calculated through a dense layer. This section provides an intricate insight into the LSTM's operations within the context of the study.

Using the notations provided by (Samit Thapa et al., 2020), equations related to LSTM are given below.

Forget gate: $f_t = \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right)$

Input gate: $i_t = \sigma\left(W_i x_t + U_i h_{t-1} + b_i\right)$

Potential Update Vector: $\tilde{c}_t = tanh\left(W_{\tilde{c}} x_t + U_{\tilde{c}} h_{t-1} + b_{\tilde{c}}\right)$

Cell Update: $c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$

Output gate: $O_t = \sigma\left(W_o x_t + U_o h_{t-1} + b_o\right)$

Hidden State: $h_t = tanh\left(c_t\right) \odot o_t$

Output Layer: $y = W_d h_n + b_d$

Sigmoid function: $\sigma(x) = \frac{1}{1 + e^{-x}}$

Tanh function: $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$

2.2.2 Applications of MLP and LSTM in hydrology

Multilayer Perceptron (MLP) and Long Short-Term Memory (LSTM) networks, both being types of artificial neural networks, find valuable applications in the field of hydrology:

2.2.2.1 Applications of MLP in Hydrology:

- **Rainfall-Runoff Modeling**: MLPs are used to simulate the complex relationship between rainfall and runoff. They can detect nonlinearities in the hydrological process, resulting in improved predictive accuracy. (Senthil Kumar, Sudheer, Jain, & Agarwal, 2005).
- **Flood Forecasting**: By analyzing historical rainfall and river discharge data, MLPs can predict flood events, allowing for timely warnings and disaster management. (Widiasari, Nugroho, & Widyawan, 2017).

- **Water Quality Prediction**: Based on various input factors such as land use, weather conditions, and water flow rates, MLPs can predict water quality parameters such as pollutant concentrations. (Najah Ahmed et al., 2019).
- **Reservoir Management**: MLPs assist in optimizing reservoir operations by predicting future water levels, inflow rates, and release strategies, considering various constraints (Baratti et al., 2003).
- **Drought Prediction**: By analyzing climatic and hydrological data, MLPs can forecast drought conditions, aiding water resource planning and management (Rezaeian-Zadeh & Tabari, 2012).

2.2.2.2 Applications of LSTM in Hydrology:

- **Time Series Prediction**: LSTM networks are well-suited for time-series data, making them valuable for predicting hydrological variables like river discharge, groundwater levels, and precipitation (Wang & Lou, 2019).
- **River runoff Modeling**: LSTM can effectively capture the intricate relationship between temperature, snow accumulation, and runoff, crucial for snow-dominated regions (S. Thapa et al., 2021).
- **Flood Inundation Mapping**: LSTM models can simulate flood events and predict their spatial extent, assisting in mapping potential flood-prone areas (Zhou, Wu, Nathan, & Wang, 2022).
- **Climate Change Impact Assessment**: LSTM can analyze long-term hydrological data to project the potential impacts of climate change on water availability, flood frequency, and drought severity (Yang et al., 2023).
- **Streamflow Forecasting**: LSTM networks enable accurate short-term and long-term streamflow forecasting, which aids water resource management and decision-making (Lin et al., 2021).

Both MLP and LSTM can learn complex patterns from historical data, adapt to changing conditions, and provide better predictive capabilities than traditional hydrological models. Their applications help to improve water resource management, disaster preparedness, and long-term hydrological planning.

2.2.3 Studies comparing ML models with conventional models

In this context, the following aspects are typically explored:

- **Model Comparison Methodology**: Researchers conduct comprehensive comparisons by setting up experiments that involve both ML models (such as MLP and LSTM) and conventional hydrological models. They use historical hydrological data and environmental variables as inputs to simulate the behavior of the watershed or catchment of interest (Baratti et al., 2003; Rezaeian-Zadeh & Tabari, 2012; S. Thapa et al., 2021; Samit Thapa et al., 2020).
- **Performance Metrics**: Various performance metrics are employed to evaluate the accuracy and reliability of model predictions. These metrics include Nash-Sutcliffe Efficiency (NSE), Root Mean Squared Error (RMSE), coefficient of determination (R^2), and others. These metrics quantify how well each model reproduces observed hydrological variables like river discharge or groundwater levels (S. Thapa et al., 2021; Samit Thapa et al., 2020).
- **Temporal and Spatial Scales**: The comparison is carried out across different temporal scales (daily, monthly, seasonal) and spatial scales (small catchments to larger basins).(Rezaeian-Zadeh & Tabari, 2012; S. Thapa et al., 2021).
- **Data Availability and Quality**: Researchers explore how well the ML models perform under varying data availability and quality conditions. This includes assessing how sensitive the models are to missing or noisy data, which is a common scenario in hydrology (S. Thapa et al., 2021).
- **Model Complexity and Simplicity**: The comparison considers the complexity of each model. ML models are known for their ability to capture complex non-linear relationships (Robert Abrahart, 2004), while conventional models might rely on simpler conceptual or physical representations. Researchers analyze whether the added complexity of ML models improves their predictive capabilities.
- **Uncertainty Analysis**: The studies examine how both ML models and conventional models handle uncertainties associated with input data, model parameters, and overall model structure (S. Thapa et al., 2021). This helps in understanding the robustness and reliability of predictions under uncertain conditions.

- **Operational Applicability**: Researchers assess the practical usability of ML models compared to conventional models. This includes evaluating factors like computational efficiency, ease of model calibration, interpretability of results, and the level of expertise required for implementation (S. Thapa et al., 2021; Samit Thapa et al., 2020).
- **Generalization**: One of the important aspects explored is how well the models generalize to unseen data. Generalization is crucial for the models to perform well in real-world applications and adapt to changing conditions (S. Thapa et al., 2021).

The outcomes of these studies provide valuable insights into whether ML models can offer improvements over conventional models in hydrological modeling. They contribute to the understanding of where and how ML models excel, where they might have limitations, and how they can complement or replace existing methods. This comparison is essential for guiding the adoption of ML techniques in hydrological research, informing policy decisions, and enhancing water resource management practices.

2.4 River runoff Prediction in Himalayan Basin

2.3.1 Hydroclimatic characteristics of the Himalayas

The Himalayas exhibit distinct hydroclimatic characteristics due to their complex topography, diverse elevations, and unique geographical location. Some key hydroclimatic characteristics of the Himalayas include:

- **Elevation Gradient**: The Himalayas encompass a wide range of elevations, from subtropical foothills to towering peaks. This elevation gradient leads to varying climatic conditions, ranging from warm and humid in the lower regions to cold and snow-dominated in higher altitudes (Ragettli et al., 2015).
- **Precipitation Patterns**: The Himalayas are a major source of moisture for the Indian subcontinent. The region experiences both monsoons and westerlies. During the monsoon season (June to September), heavy rainfall occurs, replenishing rivers and contributing to snow and glacier melt. The westerlies

also bring precipitation in the form of winter snowfall (Kansakar, Hannah, Gerrard, & Rees, 2004).

- **Snow and Glacier Accumulation**: The high elevations result in extensive snow and glacier cover. Snow accumulates during the winter months, while glaciers store ice over centuries. This seasonal snow and glacier melt significantly contribute to river discharge during warmer months (Ageta & Kadota, 1992).
- **Glacial Retreat and Meltwater**: The Himalayan glaciers have been experiencing varying rates of retreat due to global warming. The meltwater from glaciers plays a crucial role in river flow during dry seasons, providing a buffer against water scarcity (Nie et al., 2021).
- **Temperature Variability**: Temperature variations are significant across elevations. Cold temperatures prevail at high altitudes, leading to the formation of glaciers and perennial snow. In lower regions, temperatures are more moderate, influencing the timing and rate of snowmelt (Heynen et al., 2016).
- **Monsoons and Runoff**: Monsoon rains and subsequent snowmelt contribute to the majority of river runoff . The timing and intensity of monsoons impact the hydrological cycle (S. Thapa et al., 2021), affecting water availability for various uses.
- **Flash Floods and Landslides**: Rapid snowmelt and heavy monsoon rains can trigger flash floods and landslides, causing significant damage to infrastructure and communities downstream (Rezaeian-Zadeh & Tabari, 2012).
- **Climate Change Impact**: The Himalayas are highly sensitive to climate change. Rising temperatures can affect snow and glacier dynamics, altering river flow patterns and impacting water availability (S. Singh et al., 2011).

Understanding these hydroclimatic characteristics is essential for effective hydrological modeling, water resource management, and planning adaptation strategies in the face of climate change.

2.3.2 Challenges of snow-dominated hydrology and data scarcity

2.3.2.1 Challenges of Snow-Dominated Hydrology:

- **Complex Hydrological Processes:** Snow-dominated hydrology involves intricate processes like snow accumulation, melting, and runoff, making modeling and prediction challenging (Nie et al., 2021).
- **Variable Snowmelt Timing:** The timing of snowmelt is influenced by a range of factors, including altitude, temperature, and solar radiation, leading to spatial and temporal variability (Nie et al., 2021).
- **Glacier Dynamics:** Glacier contributions to runoff add complexity, as glacier retreat, advance, and melt rates impact downstream water availability (Nie et al., 2021).
- **Uncertain Runoff Patterns:** The combination of rain and snowmelt can lead to unpredictable runoff patterns, causing flooding risks during rapid melt or heavy rainfall events (Nie et al., 2021).

2.3.2.2 Data Scarcity in the Himalayas:

- **Sparse Monitoring Networks:** Limited hydrological monitoring stations in remote Himalayan areas result in sparse data coverage, hindering accurate hydrological assessments (S. Thapa et al., 2021).
- **Inaccessible Terrain:** The rugged and challenging terrain makes installation and maintenance of monitoring equipment difficult, limiting data collection efforts (Samit Thapa et al., 2020).
- **Lack of Long-Term Records:** Short historical records impede the understanding of long-term trends and variations in hydrological parameters(S. Thapa et al., 2021).
- **Seasonal Data Gaps:** Harsh winter conditions can disrupt data collection efforts, leading to seasonal data gaps that hinder comprehensive analysis(S. Thapa et al., 2021).

2.3.3 Previous research on river runoff modeling in the region

Previous research on river runoff modeling in the Himalayan region has largely focused on understanding the complex hydrological processes driven by the unique climatic conditions of the area (Ragettli et al., 2015). These studies have aimed to improve the accuracy of predicting snowmelt runoff, which is crucial for effective water resource management. Researchers have employed various modeling approaches, including both traditional physically-based models and modern data-driven models (Yang et al., 2023).

Physically-based models, such as energy balance models and temperature index models, have been utilized to simulate river runoff by considering factors like energy exchange, temperature variations, and snow accumulation (Hock, 2003; M. Shrestha et al., 2015). While these models provide valuable insights into the physical processes, they can be computationally intensive and require extensive input data, making them less suitable for data-scarce mountainous regions like the Himalayas.

Data-driven models, particularly machine learning approaches like artificial neural networks (ANNs), have gained attention due to their ability to capture complex relationships between inputs and outputs without requiring in-depth understanding of underlying processes. ANN-based models have been applied for river runoff prediction, and some studies have demonstrated their superiority over traditional models. However, the application of ANN models in snowmelt modeling has been relatively limited (Rezaeian-Zadeh & Tabari, 2012; S. Thapa et al., 2021; Samit Thapa et al., 2020).

Research has also highlighted the significance of remotely sensed data, such as snow cover area (SCA) obtained from satellite imagery (S. Thapa et al., 2021), in improving river runoff modeling accuracy. Studies have shown that SCA, in combination with meteorological and discharge data, can enhance the predictive capability of models.

Despite advancements in modeling techniques, challenges remain, particularly in the Himalayan context. Insufficient meteorological stations and underestimated precipitation data pose difficulties in accurately simulating snow accumulation and melting processes (S. Thapa et al., 2021). Moreover, uncertainties arising from the effects of climate change further complicate the modeling process.

In conclusion, previous research on river runoff modeling in the Himalayan region has revealed the complexities of hydrological processes influenced by unique climatic conditions. Physically-based and data-driven modeling approaches have been investigated, with data-driven methods demonstrating the potential for accurate predictions. Incorporating remotely sensed data and leveraging machine learning algorithms have emerged as strategies to improve modeling accuracy, but data scarcity and changing climate patterns remain challenges.

CHAPTER THREE: RESEARCH METHODOLOGY

3.1 Study Area and Data Sources

3.3.1 Description of the Langtang basin

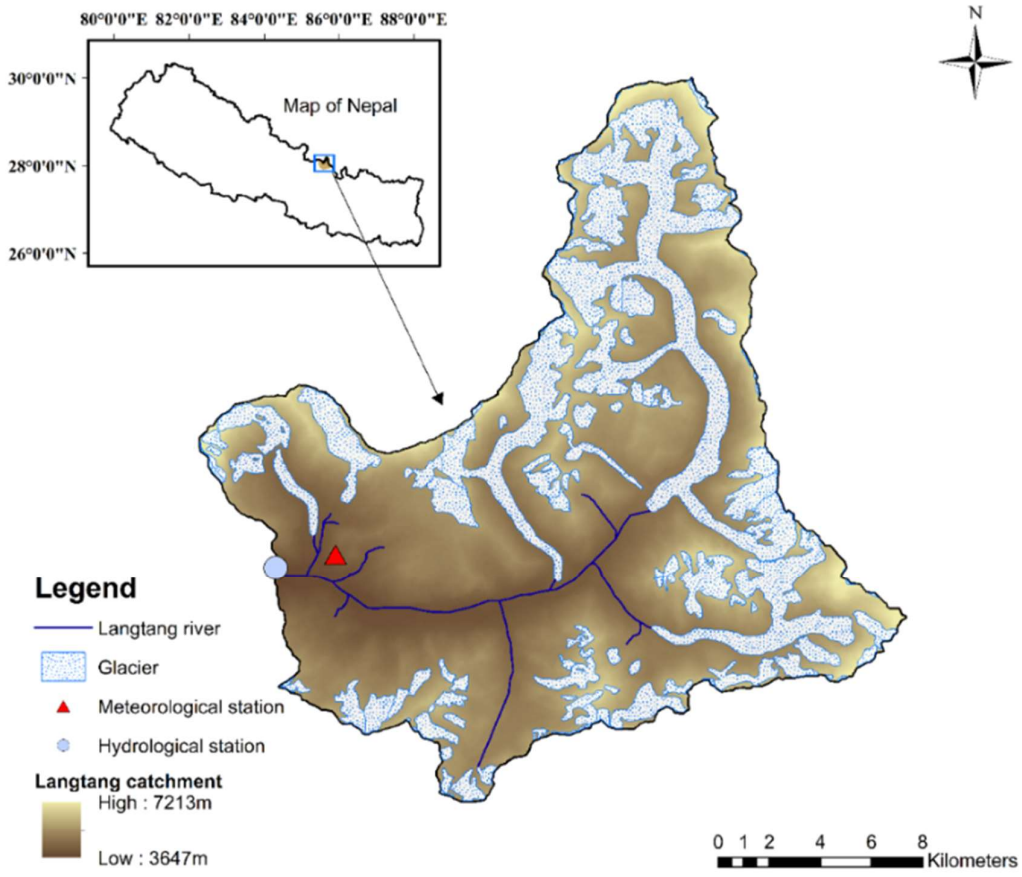

Figure 4-3.1.1: Lantang Basin

Source: Impact of climate change on river runoff in a Himalayan basin, Nepal (S. Thapa et al., 2021)

The Hindu Kush Himalayan (HKH) region, a vast and intricate mountainous environment covering Nepal, Bhutan, India, China, and others, is crucial as the source of major river systems in South Asia. This vast geographical area, noted for its magnificent beauty and ecological richness, acts as an important water tower, supplying vital supplies to downstream regions. The Langtang Valley, located in the central Nepal Himalayas, emerges as an important subregion within this enormous region. The Langtang Valley, known for its alpine majesty, diversified ecosystems, and cultural richness, contributes heavily to the Trishuli River, a tributary of the Ganges. Lantang, home to the Langtang National Park and a variety of ethnic populations, serves as a

focal point for comprehending the delicate interplay of water resources, climatic dynamics, and environmental issues within the framework of the HKH region.

The Langtang basin (shown in Fig. 1) is strategically located within the Central Himalayas, approximately sixty kilometers north of Kathmandu, Nepal. This basin exemplifies the essential features of a snow-dominated Himalayan region. The rationale for this geographic selection stems from its ease of access, which is critical in facilitating comprehensive investigations into Himalayan catchments and their associated data. The Langtang basin emerges as a compelling site for rigorous examination in the context of snow-related water resource management and the exploration of climate change impacts. The basin has a total area of 354.51 square kilometers and a significant glacier area spanning 110 square kilometers. This critical metric, which represents the extent of glacial coverage, is painstakingly calculated from the RGI-GLIMS version 6.0 dataset, meticulously assembled by the (RGI Consortium, 2017).

3.3.2 Source of Input Variables:

3.1.2.1 Snow Cover Area

The snow-covered area data utilized in this study was obtained from the MODIS (Moderate Resolution Imaging Spectroradiometer) dataset, with a specific focus on the MOD10A1 product. This dataset plays a crucial role in determining the extent of snow cover through satellite observations. The MOD10A1 product serves as a valuable resource for deriving accurate snow cover mapping within the study area located in the Himalayan region. This dataset is readily accessible through the National Snow and Ice Data Center (NSIDC) website, a reputable source known for its comprehensive collection of snow and ice-related data. By utilizing the MOD10A1 product, this study was able to capture and analyze the temporal and spatial variations in snow-covered area, a fundamental parameter in understanding hydroclimatic characteristics in the Himalayas (Hall et al, 2016).

3.1.2.2 Hydrometeorological Data

The essential climatic and hydrological data utilized for this investigation were procured from authoritative sources. The Department of Hydrology and Meteorology (DHM) in Nepal facilitated the provision of streamflow data. These datasets were specifically attributed to the reference period spanning from 2001 to 2012. In the context of streamflow data, the Kyangjing hydrological station emerged as a pivotal

data source. This station, positioned at coordinates 28.22° latitude and 85.55° longitude, boasts an elevation of 3800 meters above sea level. Due to the absence of data in the DHM we obtained temperature and precipitation data from Asian Precipitation-Highly-Resolved Observational Data Integration Towards Evaluation (APHRODITE)'s water resource (APHRODITE, 2016).

Table 1-3.1.2.2 Showing Types of Data used and their Sources

S. N	Types of Data	Sources	Resolution	Duration
1	Snow Cover Area	NASA NSIDC DAAC at CIRES	500m*500m	2000-Present
2	Precipitation	APHRODITE - data	0.25 deg	1998-2015
3	Temperature	APHRODITE - data	0.25 deg	1998-2015
4	Discharge	Department of Hydrology and Meteorology	Daily	2001-2013

This study's dataset collection includes a wide variety of hydrological and meteorological variables that are critical for understanding the dynamics of the Hindu Kush Himalayan (HKH) region, notably the Lantang area. The Snow Cover Area data at CIRES is obtained from NASA NSIDC DAAC and provides a detailed perspective at a resolution of 500m x 500m from 2000 to the present. Precipitation and temperature information provided from APHRODITE span the larger Monsoon Asia Region from 1998 to 2015 at a resolution of 0.25 degrees, offering insight into the meteorological conditions influencing the research area. The Discharge dataset, obtained from the Department of Hydrology and Meteorology, focuses on the Lantang region and includes daily resolution data from 2001 to 2013. This extensive collection of datasets serves as the foundation for comparing machine learning models in snowmelt-driven streamflow prediction, providing vital insights to hydrological modeling in difficult terrains.

3.2 Framework

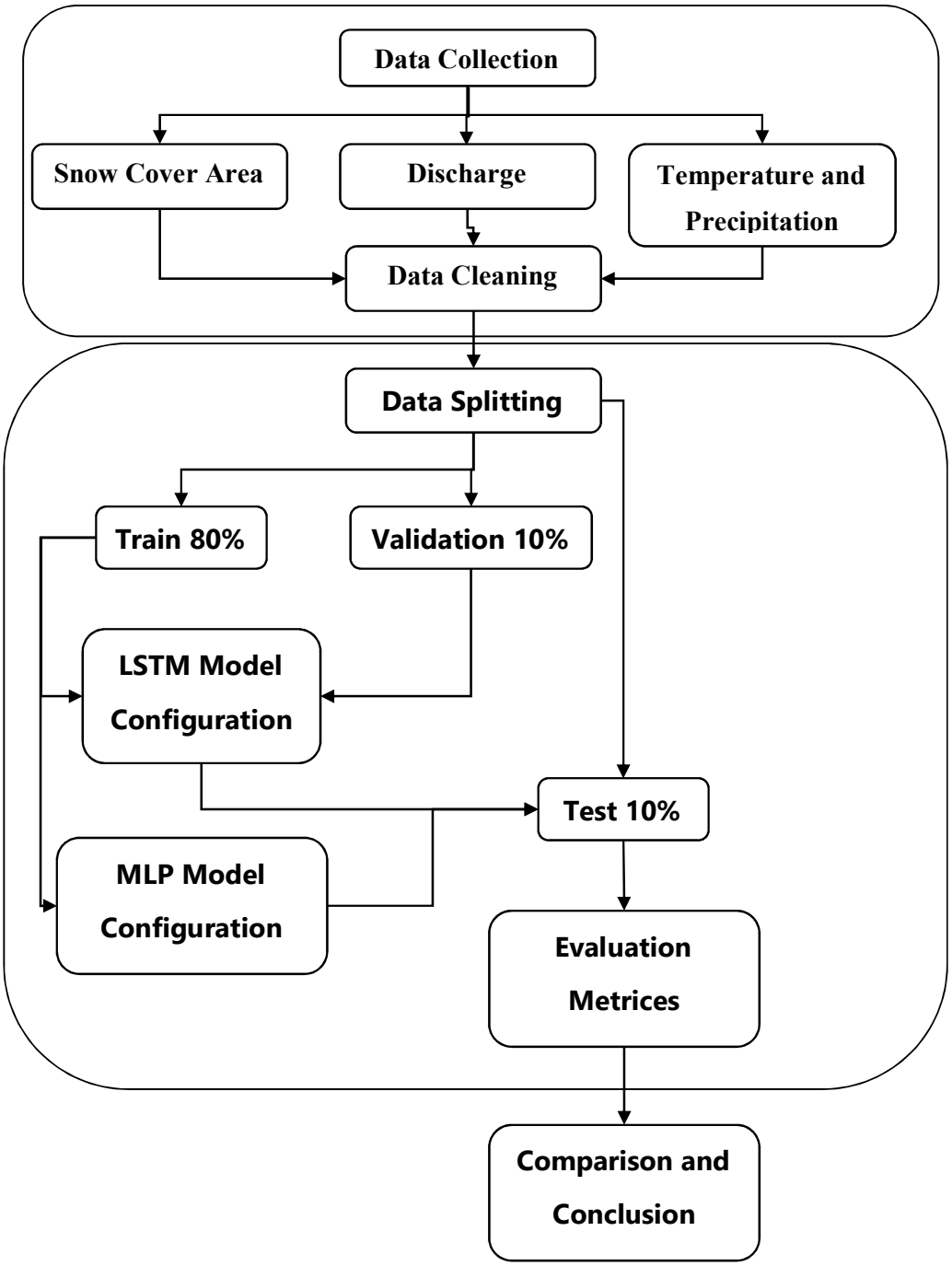

Figure 5-3.5: Research Methodology Framework

3.3 Data Preprocessing and Model Setup

3.3.1 Data cleaning and formatting

The data collection process began by obtaining raw data from various sources, such as sensors, databases, and external datasets, while adhering to a strict timeline. Following that, a thorough data inspection was carried out, thoroughly examining the raw data to identify and correct any errors, missing values, inconsistencies, or outliers. Missing values were handled strategically, with affected rows removed and advanced imputation methods used when necessary (Robert Abrahart, 2004). Outliers were identified using statistical methods and domain knowledge, and decisions were made based on the context of the analysis as to whether to remove, replace, or retain outliers. The data was transformed and standardized to ensure format consistency, including the standardization of dates and the normalization of units and measurement scales for improved comparability. Data types, whether numerical, categorical, or datetime, were confirmed to be suitable for analysis, and the data was formatted into the required structure (e.g., CSV, Excel, database format) for ease of use. To ensure data accuracy, quality assurance measures such as basic statistical checks and cross-checking against external sources or expert knowledge were implemented. The development of visualizations was critical in gaining insights into data distribution, patterns, and potential issues, as well as in identifying anomalies and outliers throughout the extensive data preparation process.

Table 2–3.3.1 Data Preprocessing Process

S.N.	Types of Data	Logic used for Data Cleaning	Data Imputation using SPSS	Unit
1	Snow Cover Area	If, NDSI_Snow_Cover<(33),Return (33), Else (Original Value)	Linear Interpolation	Daily
2	Temperature	Minimum Index for Longitude and Latitude	Linear Interpolation	Daily
3	Precipitation	Minimum Index for Longitude and Latitude	Linear Interpolation	Daily

4	Discharge	None	Linear Interpolation	Daily

The preprocessing of the Hindu Kush Himalayan (HKH) region's climate and hydrological datasets required a meticulous approach to address missing values and ensure data accuracy. Temperature and precipitation data from APHRODITE's Monsoon Asia Region were examined for missing values from 2001 to 2013, and Excel and SPSS were used to identify and evaluate these gaps using linear interpolation. Similarly, discharge data obtained from the Department of Hydrology and Meteorology (DHM) were subjected to the same missing value and date alignment treatment. The snow cover area dataset was subjected to a specific criterion after being derived from MOD10A1.061 Terra Snow Cover Daily Global 500m via Earth Engine. Areas with less than 33% snow cover were excluded, indicating the presence of glaciers. The processing steps ensure the datasets reliability and completeness, laying a solid foundation for subsequent hydrological modeling and analysis.

3.3.2 Division of data into training and testing sets

Dividing data into training and testing sets is a critical step in machine learning and data analysis to evaluate the performance of models on unseen data (Behboudian et al., 2014). The dataset is carefully gathered in the first stage of the data preparation procedure, with an emphasis on correct formatting and cleanliness to provide a strong base for further analysis. After that, an important split ratio decision is made, which establishes the percentage of data allotted to training and testing. Interestingly, 80% of the data is used for training, 10% is used for validation, and 10% is used for testing. When splitting data, the temporal aspect of the data is taken into account to make sure that, in the case of time-series data, the testing set covers later periods and the training set encompasses earlier periods. Effective division requires implementation of the split, whether by manual calculation or programming libraries such as scikit-learn in Python.

A validation set is optionally added to help with model hyperparameter tuning and model selection guidance. Ensuring the representativeness of the training set is crucial for thorough model training, and protecting against data leakage is crucial for preserving the integrity of model evaluation. Transparency and reproducibility require

detailed documentation of the entire data splitting process, including the chosen split ratio and any considerations made. In the end, the suitability of the data partitioning depends on the particular analysis context, the kind of machine learning task that is being performed, and the properties of the dataset itself.

3.3.3 Hyperparameter selection and model configurations

Choosing the right hyperparameters and setting up the model entails identifying and perfecting the salient features of the Multi-Layer Perceptron (MLP) and Long Short-Term Memory (LSTM) models for predicting snowmelt runoff.

An architecture is established for the LSTM model that consists of a single LSTM layer with 50 units, followed by a dense layer with one unit. The loss function is mean squared error, or MSE, and the Adam optimizer is selected. Each of the 50 epochs in the training process has a batch size of 32.

The MLP model, on the other hand, has two hidden layers that cover 100 and 50 units, respectively. There is a maximum of 500 iterations of training, using the Adam optimizer and the MSE loss function.

Together, these hyperparameter selections affect the models' ability to represent intricate temporal patterns in the river runoff data. Achieving optimal predictive performance and enabling a meaningful comparison between the LSTM and MLP architectures within the Model Comparison Framework require careful selection of hyperparameters and model configurations.

3.4 Model Comparison Framework

3.3.1 Evaluation metrics: NSE, R^2, RMSE, MAE

The Model Comparison Framework for evaluating machine learning models using the metrics NSE (Nash-Sutcliffe Efficiency), R^2 (Coefficient of Determination), RMSE (Root Mean Squared Error), and MAE (Mean Absolute Error) involves a systematic process to quantitatively assess the performance of different models (Lamichhane & Shakya, 2019; S. Thapa et al., 2021; Samit Thapa et al., 2020). These metrics are commonly used to measure the accuracy and goodness-of-fit of models in various fields, including hydrology. The evaluation metrics are computed to assess the models' accuracy and generalization abilities after they have been trained. Among the metrics

are Nash-Sutcliffe Efficiency (NSE), R^2 Score, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). The average magnitude of prediction errors is quantified by RMSE, the average absolute error is provided by MAE, and the percentage of the target variable's variance that the model explains is measured by R^2 Score. Furthermore, NSE, a hydrologically significant metric, evaluates the model's performance in comparison to the observed data mean. When compared and the best architecture for river runoff prediction in the context of hydrological modeling is chosen, these metrics provide a thorough understanding of the predictive power of the models.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

where $\bar{y}$ is the mean of the observed values.

$$NSE = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

Similar to R^2 but often used in the context of hydrological modeling.

3.3.2 Sensitivity analysis of input variables

A sensitivity analysis could be carried out by methodically changing one input variable while keeping the values of the others unchanged, and then monitoring any alterations to the model's predictions. This process can help identify significant variables that have a significant impact on the model's performance and provide insights into the behavior of the model under different conditions. Sensitivity analysis can help understand the relative importance of variables like temperature, precipitation, and snow cover in predicting melt-related runoff in the context of hydrological modeling.

CHAPTER FOUR: RESULTS AND DISCUSSION

4.1 Results

4.1.1 Data Correlation:

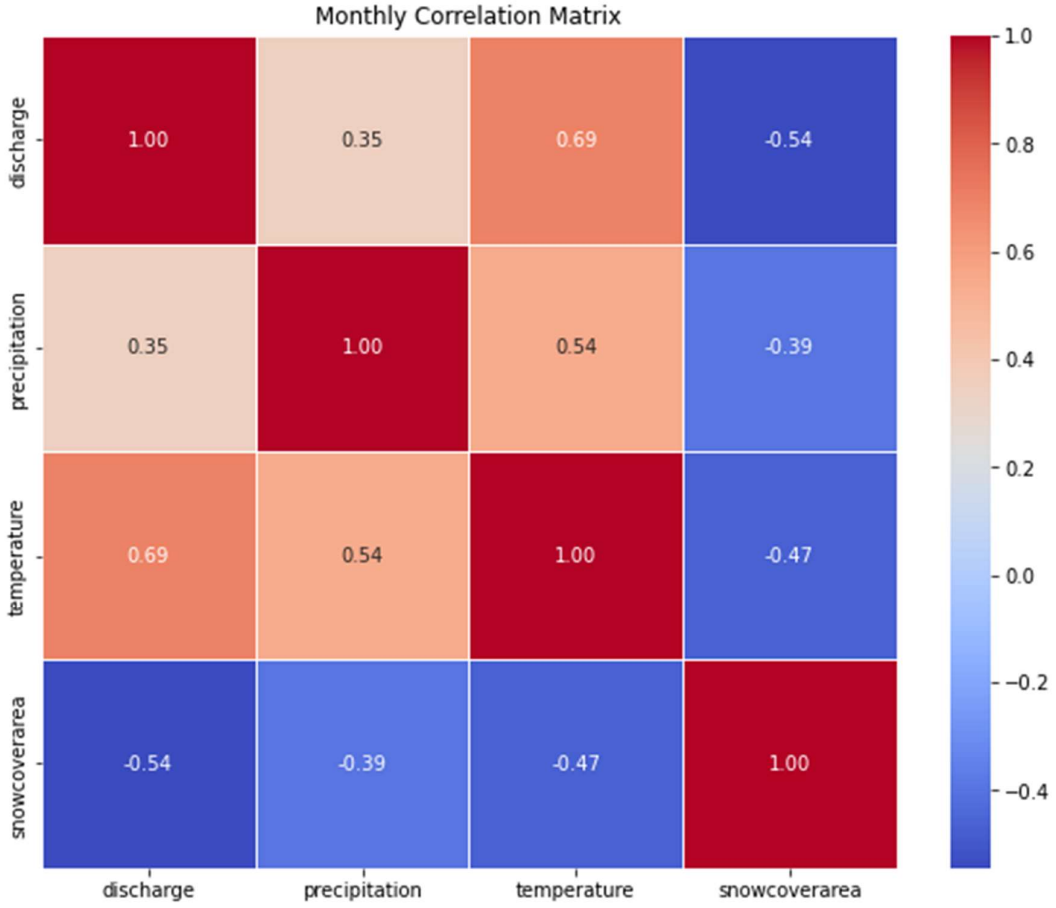

Figure 6-4.1.1: Correlation Matrix Heatmap between Precipitation, Temperature, Snow Cover Area, and Snowmelt Runoff

The relationships between important variables derived from monthly data for each year are visually captured in the correlation matrix heatmap. The following is a representation of the Pearson correlation coefficients: The following graphs show the relationship between discharge and snow cover area: discharge vs. temperature (0.69), discharge vs. precipitation (0.35), precipitation vs. snow cover area (-0.39), precipitation vs. temperature (0.54), and temperature vs. snow cover area (-0.47) respectively. The direction and strength of these correlations are clearly shown by the heatmap. Notably, a direct relationship is indicated by a strong positive correlation between temperature and discharge, whereby higher temperatures are linked to higher discharge. On the other hand, it appears that a higher discharge is associated with a lower amount of snow cover due to the moderately negative correlation between the

two variables. Understanding hydrological patterns and their interdependencies is made easier by the visualization, which offers insightful information about how the variables interact.

4.1.2 Individual Variable Visualizations:

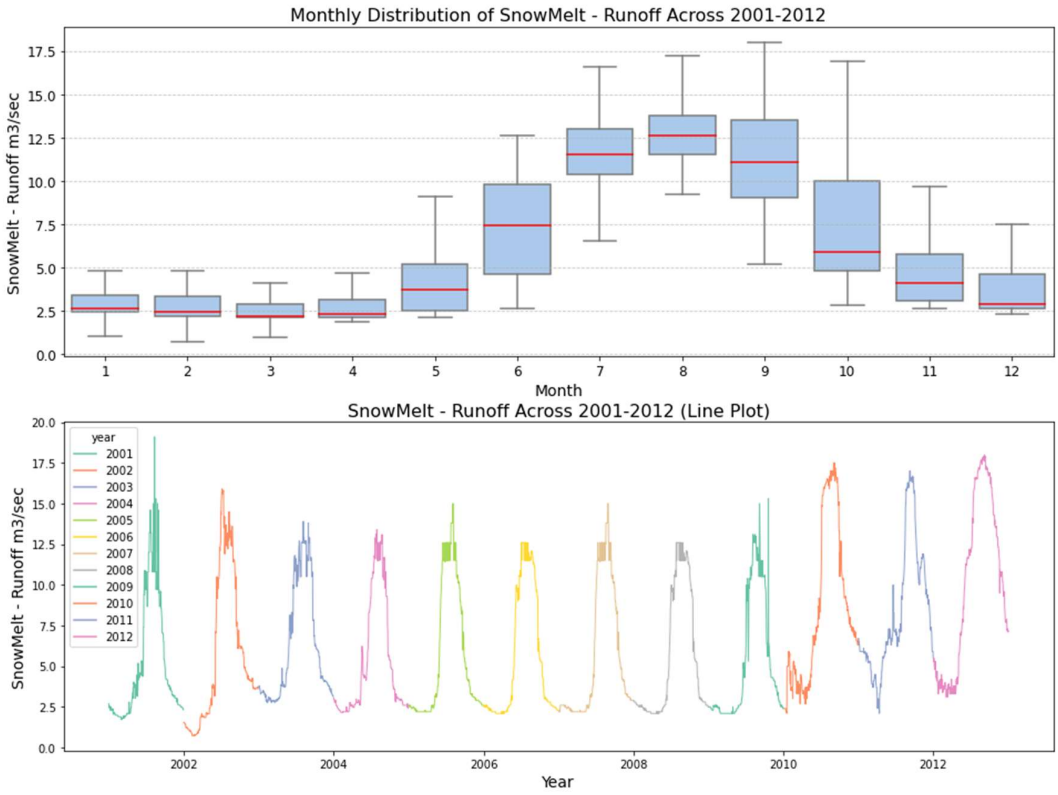

Figure 7-4.1.7(a): SnowMelt-Runoff Data Visualization

The shown Discharge graph depicts the typical flow pattern within the dataset ranging from 2001 to 2013, which was sourced and cleaned from the Department of Hydrology and Meteorology (DHM). The first graph is a box plot that displays the monthly averages for all years, offering a detailed picture of the distribution. Meanwhile, the second graph examines the everyday patterns recorded between 2001 and 2013. Notably, precipitation remains relatively low throughout the first months before steadily increasing until the eighth month, repeating this cyclic pattern over the years.

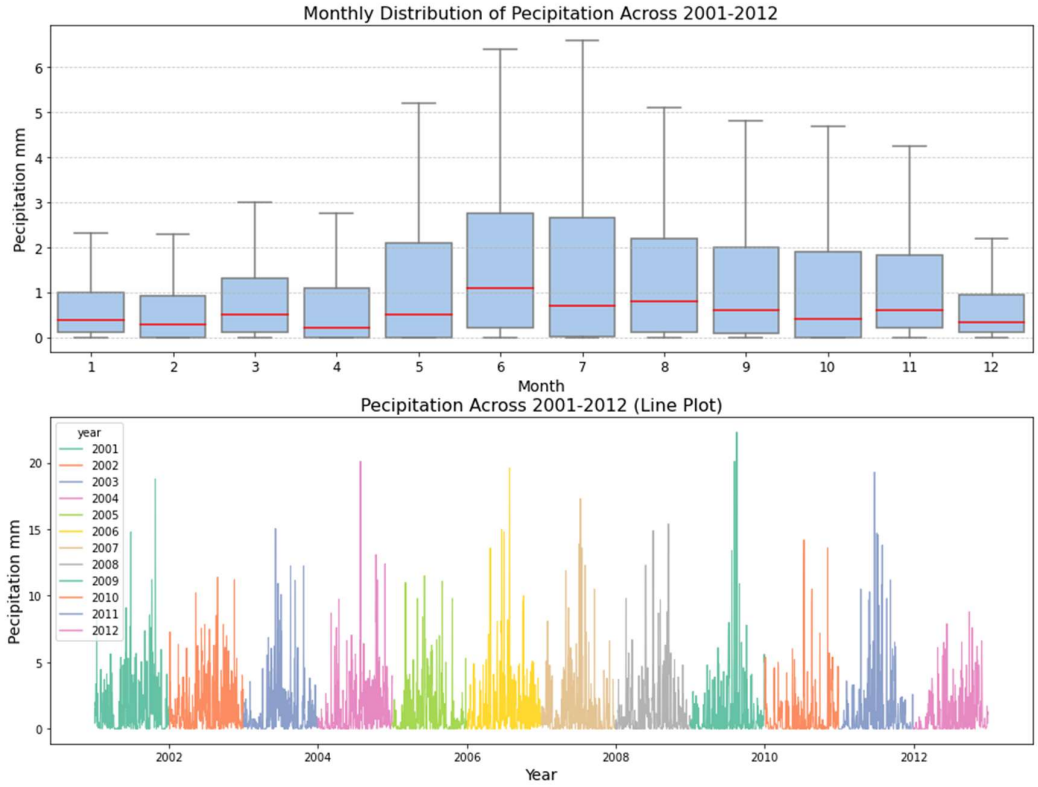

Figure 8-4.1.7(b): Precipitation Data Visualization

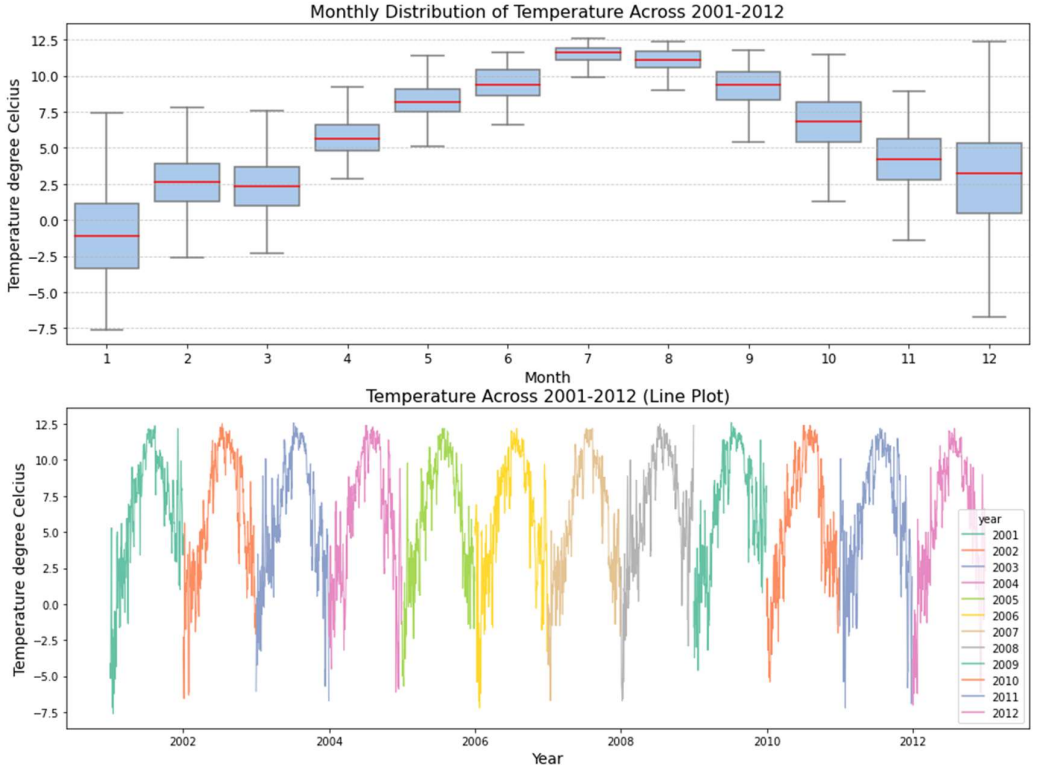

Figure 9-4.1.7(c): Temperature Data Visualization

The Precipitation graph has a comparable structure; however, the data is more unpredictable, with a noteworthy high variance, especially in the 7th and 8th months. The observed discharge pattern reflects the obvious trend of low precipitation during the first months, followed by a rise up to the eighth month, demonstrating a constant link between these variables.

Similarly, the Temperature graph shows a similar structural trend, with smaller variance in the seventh and eighth months but larger variability in the first and last months of the year. This similarity in the observed pattern across the discharge, precipitation, and temperature graphs emphasizes the hydrological system's linked dynamics of these variables.

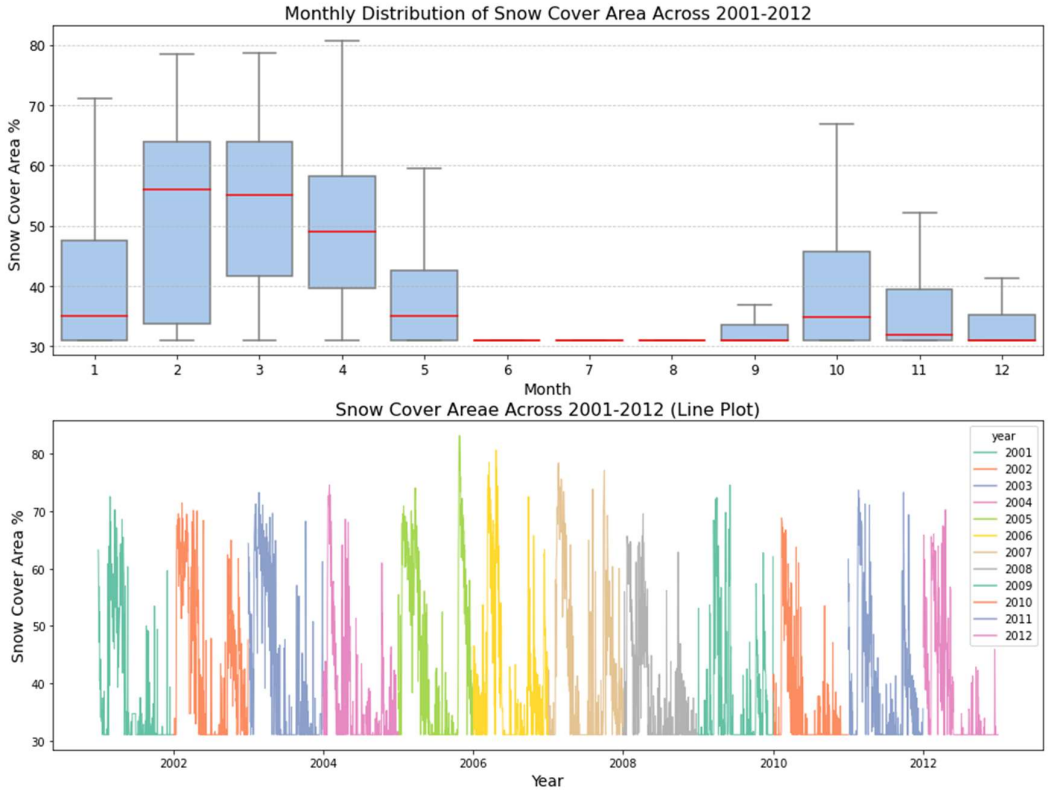

Figure 10-4.1.7(d): Snow Cover Area Data Visualization

The visualization of preprocessed data is presented in the form of a comprehensive graphical representation, which improves understanding of the dataset's characteristics. Each data variable, such as temperature, precipitation, discharge, and snow cover area, is represented by one of two plots. The monthly average is the primary focus, as illustrated by box plots. Each box plot encapsulates the variable's distribution across

different months, revealing seasonal patterns and variations. Individual daily data points are also graphically displayed below each corresponding monthly box plot to provide a more granular view. This dual visualization strategy not only highlights long-term trends and tendencies, but also allows for a closer look at the day-to-day fluctuations within each month. The use of monthly box plots and daily data graphs allows for a more nuanced and detailed analysis of the preprocessed dataset, allowing for a more in-depth knowledge of the hydrological and climatic dynamics within the researched region.

Research has demonstrated that the unique box plot pattern seen in the Lantang region in the 6th, 7th, and 8th months, which displays a rather concentrated range within the 30th percentile, is a result of the area's persistent glaciers (RGI Consortium, 2017). Throughout these particular months, the dynamics of the river runoff are significantly regulated by the continuous glacier cover. Because of their steady and gradual melting, glaciers function as a reservoir, affecting the amount and timing of runoff. As a result, during these months, runoff values tend to cluster within a narrower range, indicating the glacier's stabilizing influence on the hydrological system.

4.1.3 LSTM Model:

The Long Short-Term Memory (LSTM) model was configured with the following hyperparameters:

Number of LSTM Units: 50

Number of LSTM Layers: 1

Learning Rate: 0.001

Number of Epochs: 50

Batch Size: 32

The model was trained using the training dataset with reshaped input data (X_train_reshaped) and target variable (y_train). The validation dataset (X_val_reshaped, y_val) was used to monitor the model's performance during training. The training process involved iterating over the dataset for 50 epochs with a batch size of 32. The Adam optimizer was employed, and the mean squared error (MSE) was used as the loss function.

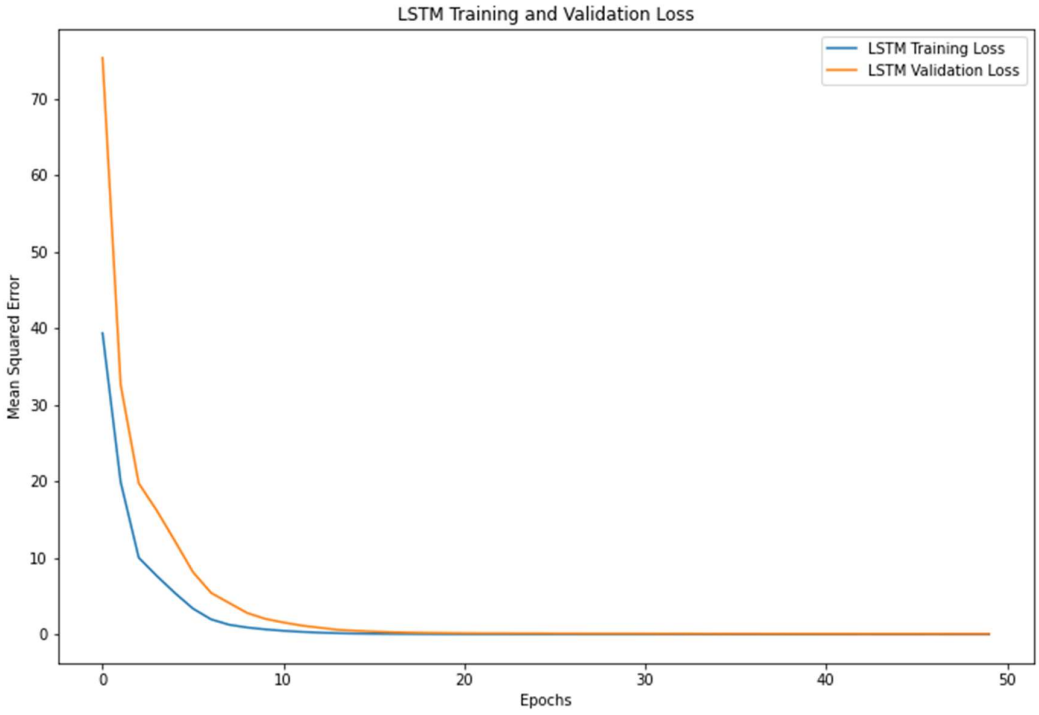

Figure 11-4.5.2: LSTM Training and Validation Loss

The validation loss has a noticeable decreasing trend, indicating effective learning from the training data. Around the 15th epoch, a crucial finding is made: the graph reaches a point of convergence, indicating that the model's performance has stabilized. After this, the validation loss continues to be negligible, indicating that the model has fully reflected the underlying patterns in the data. It is implied that additional epochs do not significantly contribute to further refinement by the consistent minimal loss after convergence. This realization helps determine the ideal training time and makes it easier to comprehend that the model reaches its peak performance early in the training phase. To avoid overfitting, early stopping after convergence may be advantageous, demonstrating a methodical approach to model training and effective use of resources.

4.1.4 MLP Model:

The Multi-Layer Perceptron (MLP) model was configured with the following hyperparameters:

Hidden Layer Sizes: (100, 50)

Activation Function: Rectified Linear Unit (ReLU)

Learning Rate: 0.001

Number of Epochs: 500

Batch Size: 32

The model was trained using the scaled training dataset (X_train_scaled, y_train). The maximum number of iterations (epochs) was set to 500, and the random seed was fixed for reproducibility.

4.1.4.1 Training and Evaluation:

Both models were trained on the respective training datasets, and their performance was evaluated on the testing dataset. The evaluation metrics used to assess the models' performance include:

Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R-squared Score (R^2), Nash-Sutcliffe Efficiency (NSE).

These metrics provide insights into the accuracy and efficiency of the models in predicting snowmelt runoff.

Table 3-4.1.4.1 Evaluation Metrices Table LSTM vs MLP

Description	LSTM	MLP
RMSE	0.234	0.173
MAE	0.170	0.084
R^2	0.9976	0.9987
NSE	0.9976	0.9987

The evaluation metrics offer a thorough analysis of how well the Multi-Layer Perceptron (MLP) and Long Short-Term Memory (LSTM) models predict snowmelt runoff. The Root Mean Squared Error (RMSE) for the Long Short-Term Memory (LSTM) model is 0.234, suggesting a minimal average deviation between the observed and predicted values. A moderate magnitude of errors is indicated by the Mean Absolute Error (MAE) of 0.170, and an exceptionally high level of explained variance is indicated by the R^2 Score of 0.9976. An additional indication of the model's ability to replicate the observed data patterns is the Nash-Sutcliffe Efficiency (NSE) of 0.9976. With an RMSE of 0.173, MAE of 0.084, and R2 Score of 0.9987, the MLP model also performs admirably. The robustness of MLP in capturing the underlying dynamics of river runoff is demonstrated by its NSE of 0.9987. Both models produce highly accurate predictions, with the MLP model producing slightly lower RMSE and MAE and marginally higher R2 and NSE scores, indicating a subtle but significant performance difference between the two architectures.

4.1.4.2 Time Series Plots:

Below are the time series plots illustrating the predicted and actual values of river runoff for both the LSTM and MLP models.

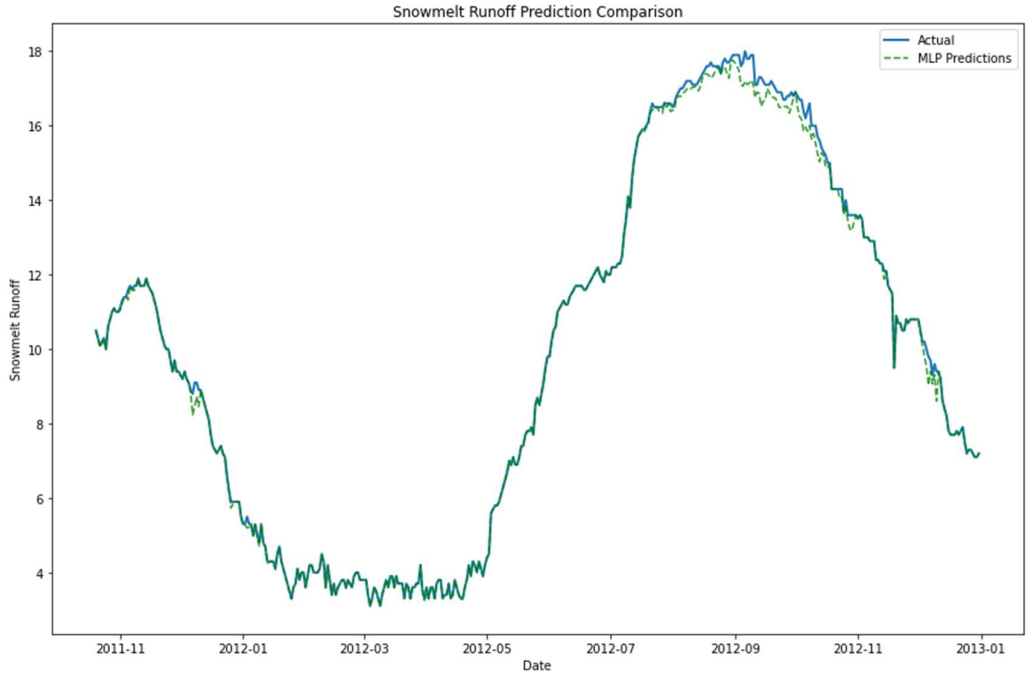

Figure 12-4.1.6(a): River runoff Prediction (Actual Vs MLP Predictions)

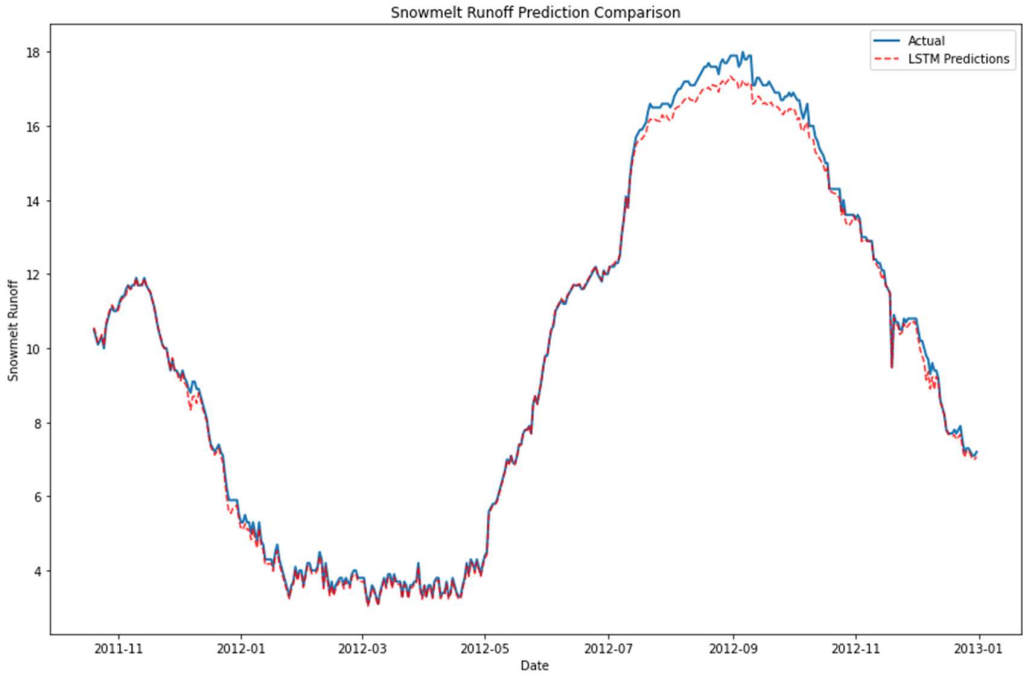

Figure 13-4.1.5(b): River runoff Prediction (Actual Vs LSTM Predictions)

4.1.4.3 Comparative Analysis Plots:

Additionally, the following plots provide a comparative analysis of the LSTM and MLP models, showcasing the actual vs. predicted values.

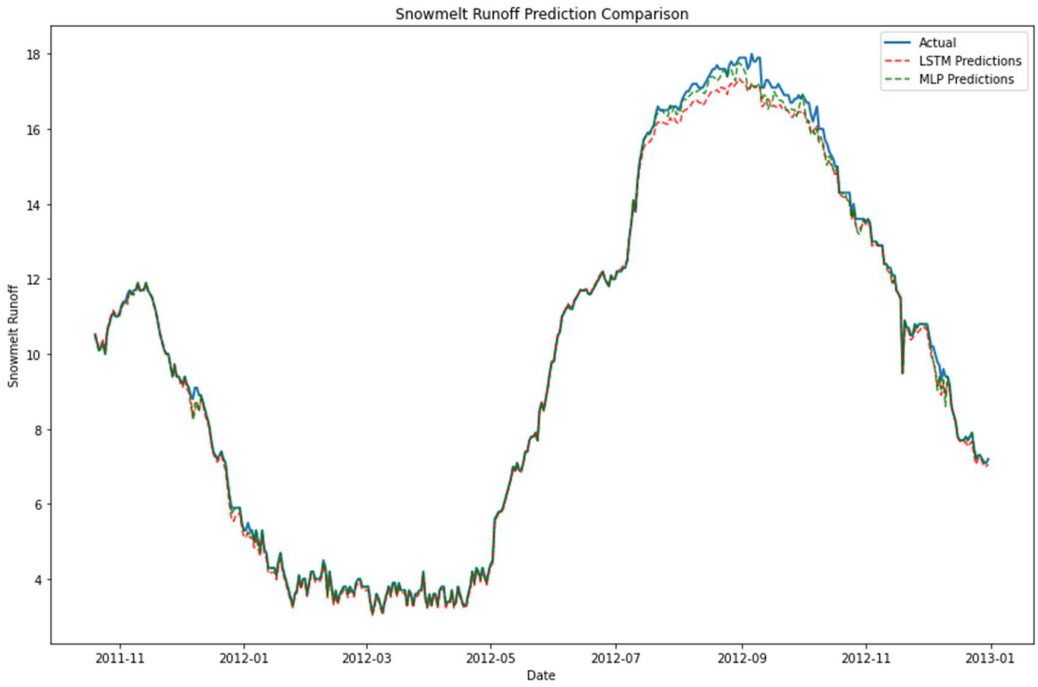

Figure 14-4.1.6: River runoff Prediction (Actual Vs MLP Predictions Vs LSTM Predictions)

There is a significant correlation between our Actual Discharge, LSTM, and MLP predictions. While accuracy varies slightly between data points, both models' predictions roughly match the Actual Discharge. The main predictions made by both models are within small error margins from the actual values.

4.2 Discussion:

The evaluation's findings and visualizations show that the estimates of river runoff are accurate, as evidenced by the relatively low RMSE and MAE values. It is noteworthy, nevertheless, that the MLP model outperformed the LSTM model in terms of RMSE and MAE values, albeit marginally. This shows that the MLP model performed exceptionally well in minimizing the errors between predicted and observed values when predicting river runoff thanks to its feedforward architecture.

High R^2 scores for both models indicated a strong correlation between the observed and predicted values. Furthermore, the NSE values, which show how effective the models

were in simulating variability in river runoff when compared to a simple mean, were nearly equal to unity.

These results imply that the models' ability to effectively capture the underlying patterns in the dataset is influenced by the configurations and hyperparameters chosen. The models' performance over time and the impact of individual variables are clearly depicted by the visualizations.

It is important to emphasize the original prediction that the LSTM model would perform better in river runoff prediction than the MLP model due to its ability to capture temporal dependencies. The results, however, point to a more nuanced reality in which the MLP model performed marginally better according to some evaluation metrics.

CHAPTER FIVE: CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

In conclusion, the comparative analysis of Long Short-Term Memory (LSTM) and Multi-Layer Perceptron (MLP) models for river runoff prediction in the Hindu Kush Himalayan region has provided valuable insights into the complex interplay of hydrological processes. The evaluation metrics and visualizations collectively portray the effectiveness of both models in capturing the intricate dynamics of snowmelt, leading to accurate predictions of runoff.

While the initial expectation was that LSTM, with its inherent ability to capture temporal dependencies, would outperform the MLP model, the results indicated a more nuanced scenario. The MLP model exhibited slightly superior performance in certain evaluation metrics, challenging our preconceived notions and underscoring the importance of empirical evaluation.

The comprehensive analysis of individual variables, including discharge, precipitation, temperature, and snow cover area, further enriched our understanding of the factors influencing snowmelt runoff. The visualizations provided a clear depiction of how each variable contributes to the predictive capabilities of the models.

5.2 Recommendations:

5.2.1 Refinement of Model Architectures:

Further exploration of LSTM and MLP architectures, considering variations in the number of layers, units, and activation functions, could provide insights into optimal configurations for river runoff prediction.

5.2.2 Incorporation of Additional Data Sources:

The inclusion of additional relevant data sources, such as soil moisture or land cover, may enhance the models' ability to capture the complexity of hydrological processes, particularly in the mountainous terrains of the Hindu Kush Himalayan region.

5.2.3 Ensemble Modeling Approaches:

Investigating ensemble modeling approaches, where multiple models are combined, could potentially yield improved predictive performance by leveraging the strengths of different algorithms.

5.2.4 Sensitivity Analysis:

Conducting sensitivity analyses to understand the impact of various hyperparameters on model performance would contribute to the refinement of modeling strategies.

5.2.5 Integration of Climate Change Scenarios:

As climate change continues to influence hydrological patterns, future research should explore the integration of climate change scenarios into predictive models to enhance their adaptability and robustness.

5.2.6 Collaboration with Stakeholders:

Collaboration with local stakeholders, such as water resource management authorities and meteorological agencies, is essential to ensure that predictive models align with practical decision-making needs in the region.

REFERENCES

(UNDP), U. N. D. P. (2013). Water governance in the Arab region managing scarc i ty and s ecuring the f uture.

Ageta, Y., & Kadota, T. (1992). Predictions of changes of glacier mass balance in the Nepal Himalaya and Tibetan Plateau: a case study of air temperature increase for three glaciers. Annals of Glaciology, 16, 89-94. doi:10.3189/1992AoG16-1-89-94

Baratti, R., Cannas, B., Fanni, A., Pintus, M., Sechi, G. M., & Toreno, N. (2003). River flow forecast for reservoir management through neural networks. Neurocomputing, 55(3), 421-437. doi:https://doi.org/10.1016/S0925-2312(03)00387-4

Behboudian, S., Tabesh, M., Falahnezhad, M., & Ghavanini, F. A. (2014). A long-term prediction of domestic water demand using preprocessing in artificial neural network. Journal of Water Supply: Research and Technology-Aqua, 63(1), 31-42. doi:10.2166/aqua.2013.085

Farfán, J. F., Palacios, K., Ulloa, J., & Avilés, A. (2020). A hybrid neural network-based technique to improve the flow forecasting of physical and data-driven models: Methodology and case studies in Andean watersheds. Journal of Hydrology: Regional Studies, 27. doi:10.1016/j.ejrh.2019.100652

Gardner, M. W., & Dorling, S. R. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmospheric Environment, 32(14), 2627-2636. doi:https://doi.org/10.1016/S1352-2310(97)00447-0

Hall, D. K., V. V. Salomonson, and G. A. Riggs. 2016. MODIS/Terra Snow Cover Daily L3 Global 500m Grid. Version 6. Boulder, Colorado USA: NASA National Snow and Ice Data Center Distributed Active Archive Center.

Heynen, M., Miles, E., Ragettli, S., Buri, P., Immerzeel, W. W., & Pellicciotti, F. (2016). Air temperature variability in a high-elevation Himalayan catchment. Annals of Glaciology, 57(71), 212-222. doi:10.3189/2016AoG71A076

Hock, R. (2003). Temperature index melt modelling in mountain areas. Journal of Hydrology, 282(1), 104-115. doi:https://doi.org/10.1016/S0022-1694(03)00257-9

Immerzeel, W. W., van Beek, L. P., Konz, M., Shrestha, A. B., & Bierkens, M. F. (2012). Hydrological response to climate change in a glacierized catchment in the Himalayas. Clim Change, 110(3-4), 721-736. doi:10.1007/s10584-011-0143-4

Jain, S. K., Goswami, A., & Saraf, A. K. (2009). Assessment of River runoff Using Remote Sensing and Effect of Climate Change on Runoff. Water Resources Management, 24(9), 1763-1777. doi:10.1007/s11269-009-9523-1

Kansakar, S. R., Hannah, D. M., Gerrard, J., & Rees, G. (2004). Spatial pattern in the precipitation regime of Nepal. 24(13), 1645-1659. doi:https://doi.org/10.1002/joc.1098

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

Lamichhane, & Shakya. (2019). Integrated Assessment of Climate Change and Land Use Change Impacts on Hydrology in the Kathmandu Valley Watershed, Central Nepal. Water, 11(10). doi:10.3390/w11102059

Lin, Y., Wang, D., Wang, G., Qiu, J., Long, K., Du, Y., . . . Dai, Y. (2021). A hybrid deep learning algorithm and its application to streamflow prediction. Journal of Hydrology, 601, 126636. doi:https://doi.org/10.1016/j.jhydrol.2021.126636

Liu, Z., Wang, Y., Xu, Z., & Duan, Q. (2017). Conceptual Hydrological Models. In Q. Duan, F. Pappenberger, J. Thielen, A. Wood, H. L. Cloke, & J. C. Schaake (Eds.), Handbook of Hydrometeorological Ensemble Forecasting (pp. 1-23). Berlin, Heidelberg: Springer Berlin Heidelberg.

Mahesh, B. (2019). Machine Learning Algorithms -A Review. doi:10.21275/ART20203995

Najah Ahmed, A., Binti Othman, F., Abdulmohsin Afan, H., Khaleel Ibrahim, R., Ming Fai, C., Shabbir Hossain, M., . . . Elshafie, A. (2019). Machine learning methods for better water quality prediction. Journal of Hydrology, 578, 124084. doi:https://doi.org/10.1016/j.jhydrol.2019.124084

Nie, Y., Pritchard, H. D., Liu, Q., Hennig, T., Wang, W., Wang, X., . . . Chen, X. (2021). Glacial change and hydrological implications in the Himalaya and Karakoram. Nature Reviews Earth & Environment, 2(2), 91-106. doi:10.1038/s43017-020-00124-w

Pandey, V. P., Dhaubanjar, S., Bharati, L., & Thapa, B. R. (2020). Spatio-temporal distribution of water availability in Karnali-Mohana Basin, Western Nepal: Climate change impact assessment (Part-B). Journal of Hydrology: Regional Studies, 29, 100691. doi:https://doi.org/10.1016/j.ejrh.2020.100691

Pradhan, P., Tingsanchali, T., & Shrestha, S. (2020). Evaluation of Soil and Water Assessment Tool and Artificial Neural Network models for hydrologic simulation in different climatic regions of Asia. Sci Total Environ, 701, 134308. doi:10.1016/j.scitotenv.2019.134308

Ragettli, S., Pellicciotti, F., Immerzeel, W. W., Miles, E. S., Petersen, L., Heynen, M., . . . Shrestha, A. (2015). Unraveling the hydrology of a Himalayan catchment through integration of high resolution in situ data and remote sensing with an advanced simulation model. Advances in Water Resources, 78, 94-111. doi:https://doi.org/10.1016/j.advwatres.2015.01.013

Renji Remesan , J. M. (2015). Hydrological Data Driven Modelling. Earth Systems Data and Models (ESDM, volume 1). doi:https://doi.org/10.1007/978-3-319-09235-5

Rezaeian-Zadeh, M., & Tabari, H. (2012). MLP-based drought forecasting in different climatic regions. Theoretical and Applied Climatology, 109(3), 407-414. doi:10.1007/s00704-012-0592-3

Robert Abrahart, P. E. K., Linda M. See. (2004). Neural Networks for Hydrological Modeling. 316. doi:https://doi.org/10.1201/9780203024119

Senthil Kumar, A. R., Sudheer, K. P., Jain, S. K., & Agarwal, P. K. (2005). Rainfall-runoff modelling using artificial neural networks: comparison of network types. 19(6), 1277-1291. doi:https://doi.org/10.1002/hyp.5581

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

Shrestha, M., Koike, T., Hirabayashi, Y., Xue, Y., Wang, L., Rasul, G., & Ahmad, B. (2015). Integrated simulation of snow and glacier melt in water and energy balance-based, distributed hydrological modeling framework at Hunza River Basin of Pakistan Karakoram region. 120(10), 4889-4919. doi:https://doi.org/10.1002/2014JD022666

Shrestha, S., Shrestha, M., & Babel, M. S. (2015). Assessment of climate change impact on water diversion strategies of Melamchi Water Supply Project in Nepal.

Theoretical and Applied Climatology, 128(1-2), 311-323. doi:10.1007/s00704-015-1713-6

Singh, L., & Saravanan, S. (2020). Impact of climate change on hydrology components using CORDEX South Asia climate model in Wunna, Bharathpuzha, and Mahanadi, India. Environ Monit Assess, 192(11), 678. doi:10.1007/s10661-020-08637-z

Singh, S., Bassignana-Khadka, I., Karky, B., & Sharma, E. (2011). Climate Change in the Hindu Kush-Himalayas: The State of Current Knowledge.

Thapa, S., Li, H., Li, B., Fu, D., Shi, X., Yabo, S., . . . Zhang, W. (2021). Impact of climate change on river runoff in a Himalayan basin, Nepal. Environ Monit Assess, 193(7), 393. doi:10.1007/s10661-021-09197-6

Thapa, S., Zhao, Z., Li, B., Lu, L., Fu, D., Shi, X., . . . Qi, H. (2020). Snowmelt-Driven Streamflow Prediction Using Machine Learning Techniques (LSTM, NARX, GPR, and SVR). Water, 12(6). doi:10.3390/w12061734

Urich, C., & Rauch, W. (2014). Exploring critical pathways for urban water management to identify robust strategies under deep uncertainties. Water Res, 66, 374-389. doi:10.1016/j.watres.2014.08.020

Uysal, G., Şensoy, A., & Şorman, A. A. (2016). Improving daily streamflow forecasts in mountainous Upper Euphrates basin by multi-layer perceptron model with satellite snow products. Journal of Hydrology, 543, 630-650. doi:10.1016/j.jhydrol.2016.10.037

Wang, Z., & Lou, Y. (2019, 15-17 March 2019). Hydrological time series forecast model based on wavelet de-noising and ARIMA-LSTM. Paper presented at the 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC).

Widiasari, I. R., Nugroho, L. E., & Widyawan. (2017, 2-4 Nov. 2017). Deep learning multilayer perceptron (MLP) for flood prediction model using wireless sensor network based hydrology time series data mining. Paper presented at the 2017 International Conference on Innovative and Creative Information Technology (ICITech).

Yang, S., Tan, M. L., Song, Q., He, J., Yao, N., Li, X., & Yang, X. (2023). Coupling SWAT and Bi-LSTM for improving daily-scale hydro-climatic simulation and climate change impact assessment in a tropical river basin. Journal of

Environmental Management, 330, 117244. doi:https://doi.org/10.1016/j.jenvman.2023.117244

Zhou, Y., Wu, W., Nathan, R., & Wang, Q. J. (2022). Deep Learning-Based Rapid Flood Inundation Modeling for Flat Floodplains With Complex Flow Paths. 58(12), e2022WR033214. doi:https://doi.org/10.1029/2022WR033214

APPENDICES

Appendix A: Data Retrieval for Lantang Region Precipitation

This script in Python makes it easier to extract precipitation data from the APHRODITE dataset for the Lantang region for the year 2012. Using the netCDF4 library, the code retrieves the precipitation variable by using the Lantang region's geographic coordinates. The script retrieves pertinent information for the given location and time period, which extends from the beginning of the dataset to December 31, 2012. To complement the script for 2012, the data retrieval procedure for precipitation in the Lantang region was also manually modified for every year between 2001 and 2011. This customized method guarantees accuracy in obtaining data relevant to the time span of every year. Each year's executions were done separately, producing CSV files with organized names. I named my files consistently, using, for example, 'precipitation_lantang_2001.csv' for 2001. Users can also choose to combine annual data into a single dataset for a more thorough analysis, highlighting the versatility of this manual retrieval process in handling temporal nuances and dataset organization.

```
from netCDF4 import Dataset
import pandas as pd
import numpy as np

data = Dataset(r'F:\Data_New\precipitation\APHRO_MA_025deg_V1901.2012.nc')


lon = data.variables['lon'][:]
lat = data.variables['lat'][:]

lon_lantang = 28.21105
lat_lantang = 85.56713

sq_diff_lat = (lat-lat_lantang)**2
sq_diff_lon = (lon-lon_lantang)**2

min_index_lat = sq_diff_lat.argmin()
min_index_lon = sq_diff_lon.argmin()

precip = data.variables['precip']

starting_date = data.variables['time'].units[14:24]
ending_date = data.variables['time'].units[14:18]+'-12-31'
date_range = pd.date_range(start = starting_date, end = ending_date)
df = pd.DataFrame(0, columns = ['Precipitation'], index = date_range)
dt = np.arange(0, data.variables['time'].size)

for time_index in dt:
    df.iloc[time_index] = precip[time_index, min_index_lat, min_index_lon]

df.to_csv('precipitation lantang 2012.csv')
```

Appendix B: Data Retrieval for Lantang Region Temperature

As per the data retrieval of precipitation data from netCDF4 file, the same process and code is applied to the netCDF4 file for the temperature by only changing the name.

```
from netCDF4 import Dataset
import pandas as pd
import numpy as np

data = Dataset(r'F:\Data_New\temperature\APHRO_MA_025deg_V1901.2012.nc')


lon = data.variables['lon'][:]
lat = data.variables['lat'][:]

lon_lantang = 28.21105
lat_lantang = 85.56713

sq_diff_lat = (lat-lat_lantang)**2
sq_diff_lon = (lon-lon_lantang)**2

min_index_lat = sq_diff_lat.argmin()
min_index_lon = sq_diff_lon.argmin()

temp = data.variables['temp']

starting_date = data.variables['time'].units[14:24]
ending_date = data.variables['time'].units[14:18]+'-12-31'
date_range = pd.date_range(start = starting_date, end = ending_date)
df = pd.DataFrame(0, columns = ['Temperature'], index = date_range)
dt = np.arange(0, data.variables['time'].size)

for time_index in dt:
    df.iloc[time_index] = precip[time_index, min_index_lat, min_index_lon]

df.to_csv('temperature lantang 2012.csv')
```

Appendix C: Data Retrieval for Lantang Region Temperature

This code snippet uses the MODIS satellite imagery to analyze snow cover data through the use of a JavaScript script in the Google Earth Engine environment. The shape file uploaded to the Google Earth Engine cloud defines the region of interest, which is denoted by the variable 'AOI' (Area of Interest). Using 'Map.centerObject(AOI)', the script centers the map on the designated area. 'Map.addLayer(AOI)' adds a layer to visualize the defined AOI.

The code then uses the 'filterDate' function to filter the images based on a specified date range from January 1, 2001, to December 31, 2012, after gaining access to the MODIS snow cover dataset ('MODIS/006/MOD10A1'). The chosen snow cover data is then assigned to the variable'modLSTday'.

The 'ui.Chart.image.series' function is then used by the script to create a time series chart ('ts1'). An overview of the mean snow cover over time in the designated region is given by this chart. The generated time series chart is output for analysis using the 'print(ts1)' statement. By accessing the Task and Graph Subsection, where the data will be obtained, you can also download the CSV file in this chart form.

```
Map.centerObject(AOI);
Map.addLayer(AOI);

var modis = ee.ImageCollection('MODIS/006/MOD10A1');
var mod11a2 = modis.filterDate( '2001-01-01', '2012-12-31');

var modLSTday = mod11a2.select('NDSI_Snow_Cover');

print(modLSTday);


var snowCoverVis = {
  min: 0.0,
  max: 100.0,
  palette: ['black', '0dffff', '0524ff', 'ffffff'],
};


var ts1 = ui.Chart.image.series({
  imageCollection: modLSTday,
  region: AOI,
  reducer: ee.Reducer.mean(),
  scale: 1000,
  xProperty: 'system:time_start'})
  .setOptions({
  title: 'Snow Cover',
  vAxis: {title: 'LST Celsius'}});

print(ts1);
```

Appendix D: River runoff Prediction Comparison

```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import MinMaxScaler
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score
from sklearn.neural_network import MLPRegressor
from keras.models import Sequential
from keras.layers import LSTM, Dense
import matplotlib.pyplot as plt
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score

# Load data from the CSV file
data = pd.read_csv('F:/MSCCD/Semester IV/DATA/climate data allll.csv')
data['date'] = pd.to_datetime(data['date'])  # Convert the 'date' column to
datetime format

# Target variable is in the column 'discharge' and other features are in other
columns
X = data[['discharge','precipitation', 'temperature', 'snowcoverarea']]
y = data['discharge']


# Set the percentage for training and validation data
train_percent = 0.8
val_percent = 0.1
test_percent = 1.0 - train_percent - val_percent

# Calculate the split indices
train_idx = int(len(data) * train_percent)
val_idx = int(len(data) * (train_percent + val_percent))

# Split the data
X_train, X_val, X_test = np.split(X, [train_idx, val_idx])
y_train, y_val, y_test = np.split(y, [train_idx, val_idx])

# Normalize your data
scaler = MinMaxScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_val_scaled = scaler.transform(X_val)
X_test_scaled = scaler.transform(X_test)

# Reshape data for LSTM input (samples, time steps, features)
X_train_reshaped = X_train_scaled.reshape((X_train_scaled.shape[0], 1,
X_train_scaled.shape[1]))
```

```
X_val_reshaped = X_val_scaled.reshape((X_val_scaled.shape[0], 1,
X_val_scaled.shape[1]))
X_test_reshaped = X_test_scaled.reshape((X_test_scaled.shape[0], 1,
X_test_scaled.shape[1]))

# Initialize and train LSTM model
lstm_model = Sequential()
lstm_model.add(LSTM(50, input_shape=(X_train_reshaped.shape[1], X_train_re-
shaped.shape[2])))
lstm_model.add(Dense(1))
lstm_model.compile(optimizer='adam', loss='mse')

# Train the model with validation data
history_lstm = lstm_model.fit(X_train_reshaped, y_train, epochs=50,
batch_size=32, validation_data=(X_val_reshaped, y_val), verbose=1)

# Make predictions on the test set using LSTM
lstm_predictions = lstm_model.predict(X_test_reshaped)
lstm_predictions = lstm_predictions.reshape((lstm_predictions.shape[0],))

# Evaluate the LSTM model
lstm_rmse = np.sqrt(mean_squared_error(y_test, lstm_predictions))
lstm_mae = mean_absolute_error(y_test, lstm_predictions)
lstm_r2 = r2_score(y_test, lstm_predictions)
lstm_nse = 1 - (np.sum((y_test - lstm_predictions)**2) / np.sum((y_test -
np.mean(y_test))**2))

print("LSTM Evaluation Metrics:")
print("RMSE:", lstm_rmse)
print("MAE:", lstm_mae)
print("R^2 Score:", lstm_r2)
print("NSE for LSTM:", lstm_nse)

# Initialize and train MLP model
mlp_model = MLPRegressor(hidden_layer_sizes=(100, 50), max_iter=500, ran-
dom_state=42)
mlp_model.fit(X_train_scaled, y_train)

# Make predictions on the test set using MLP
mlp_predictions = mlp_model.predict(X_test_scaled)

# Evaluate the MLP model
mlp_rmse = np.sqrt(mean_squared_error(y_test, mlp_predictions))
mlp_mae = mean_absolute_error(y_test, mlp_predictions)
mlp_r2 = r2_score(y_test, mlp_predictions)
```

```
mlp_nse = 1 - (np.sum((y_test - mlp_predictions)**2) / np.sum((y_test -
np.mean(y_test))**2))

print("\nMLP Evaluation Metrics:")
print("RMSE:", mlp_rmse)
print("MAE:", mlp_mae)
print("R^2 Score:", mlp_r2)
print("NSE for MLP:", mlp_nse)

# Plot training and validation loss for LSTM model
plt.figure(figsize=(12, 8))
plt.subplot(2, 1, 1)
plt.plot(history_lstm.history['loss'], label='LSTM Training Loss')
plt.plot(history_lstm.history['val_loss'], label='LSTM Validation Loss')
plt.title('LSTM Training and Validation Loss')
plt.xlabel('Epochs')
plt.ylabel('Mean Squared Error')
plt.legend()

# Plot the actual vs predicted values for both LSTM and MLP models with dates
plt.subplot(2, 1, 2)
plt.plot(data['date'].iloc[-len(y_test):], y_test, label='Actual', linewidth=2)
plt.plot(data['date'].iloc[-len(lstm_predictions):], lstm_predictions, la-
bel='LSTM Predictions', linestyle='dashed', alpha=0.8, color = 'red')
plt.plot(data['date'].iloc[-len(mlp_predictions):], mlp_predictions, label='MLP
Predictions', linestyle='dashed', alpha=0.8, color = 'green')
plt.title('Snowmelt Runoff Prediction Comparison')
plt.xlabel('Date')
plt.ylabel('Snowmelt Runoff')
plt.legend()

plt.tight_layout()
plt.show()
```

The performance of the MLP and LSTM models for predicting river runoff is compared in this Python script. Included are the LSTM model's training and validation loss plots as well as a comparison of the actual and predicted values for the MLP and LSTM models with dates. Assessment measures, including RMSE, MAE, R^2 Score, and NSE, are computed for both models, offering valuable perspectives into their forecasting abilities. Depending on the configuration of a given dataset, changes to the names and indices of columns may be required.

Appendix E: Monthly Data Visualization and Correlation Analysis

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

df = pd.read_csv('F:/MSCCD/Semester IV/DATA/climate data allll.csv')


df['date'] = pd.to_datetime(df['date'])

# Extract year and month from the date
df['year'] = df['date'].dt.year
df['month'] = df['date'].dt.month

# Create a new column combining year and month for proper grouping
df['year_month'] = df['date'].dt.to_period('M')

# Aggregate the data for each month across all years for discharge
monthly_aggregated_discharge = df.groupby('month')['discharge'].agg(list)

# Set a simple color for the plot
custom_palette = sns.color_palette("pastel")[0]

# Create a boxplot for each month with a line passing through the center
plt.figure(figsize=(16, 12))
plt.subplot(2, 1, 1)
sns.boxplot(data=monthly_aggregated_discharge.apply(pd.Series).T, color=cus-
tom_palette, showfliers=False, medianprops={'color': 'red'})

# Set plot labels and title
plt.title('Monthly Distribution of SnowMelt - Runoff Across 2001-2012', font-
size=16)
plt.xlabel('Month', fontsize=14)
plt.ylabel('SnowMelt - Runoff m3/sec', fontsize=14)

# Customize the grid and axes
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tick_params(axis='both', labelsize=12)

# Discharge Line Plot
plt.subplot(2, 1, 2)
sns.lineplot(x='date', y='discharge', data=df, hue='year', palette='Set2', lin-
ewidth=1)
plt.xlabel('Year', fontsize=14)
plt.ylabel('SnowMelt - Runoff m3/sec', fontsize=14)
```

```
plt.title('SnowMelt - Runoff Across 2001-2012 (Line Plot)', fontsize=16)

# Aggregate the data for each month across all years for precipitation
monthly_aggregated_precipitation = df.groupby('month')['precipitation'].agg(list)

# Set a simple color for the plot
custom_palette = sns.color_palette("pastel")[0]

# Create a boxplot for each month with a line passing through the center
plt.figure(figsize=(16, 12))
plt.subplot(2, 1, 1)
sns.boxplot(data=monthly_aggregated_precipitation.apply(pd.Series).T, color=cus-
tom_palette, showfliers=False, medianprops={'color': 'red'})

# Set plot labels and title
plt.title('Monthly Distribution of Pecipitation Across 2001-2012', fontsize=16)
plt.xlabel('Month', fontsize=14)
plt.ylabel('Pecipitation mm', fontsize=14)

# Customize the grid and axes
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tick_params(axis='both', labelsize=12)

# Discharge Line Plot
plt.subplot(2, 1, 2)
sns.lineplot(x='date', y='precipitation', data=df, hue='year', palette='Set2',
linewidth=1)
plt.xlabel('Year', fontsize=14)
plt.ylabel('Pecipitation mm', fontsize=14)
plt.title('Pecipitation Across 2001-2012 (Line Plot)', fontsize=16)

# Aggregate the data for each month across all years for temperature
monthly_aggregated_temperature = df.groupby('month')['temperature'].agg(list)

# Set a simple color for the plot
custom_palette = sns.color_palette("pastel")[0]

# Create a boxplot for each month with a line passing through the center
plt.figure(figsize=(16, 12))
plt.subplot(2, 1, 1)
sns.boxplot(data=monthly_aggregated_temperature.apply(pd.Series).T, color=cus-
tom_palette, showfliers=False, medianprops={'color': 'red'})

# Set plot labels and title
plt.title('Monthly Distribution of Temperature Across 2001-2012', fontsize=16)
```

```
plt.xlabel('Month', fontsize=14)
plt.ylabel('Temperature degree Celcius', fontsize=14)

# Customize the grid and axes
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tick_params(axis='both', labelsize=12)

# Discharge Line Plot
plt.subplot(2, 1, 2)
sns.lineplot(x='date', y='temperature', data=df, hue='year', palette='Set2', lin-
ewidth=1)
plt.xlabel('Year', fontsize=14)
plt.ylabel('Temperature degree Celcius', fontsize=14)
plt.title('Temperature Across 2001-2012 (Line Plot)', fontsize=16)

# Aggregate the data for each month across all years for snowcoverarea
monthly_aggregated_sca = df.groupby('month')['snowcoverarea'].agg(list)

# Set a simple color for the plot
custom_palette = sns.color_palette("pastel")[0]

# Create a boxplot for each month with a line passing through the center
plt.figure(figsize=(16, 12))
plt.subplot(2, 1, 1)
sns.boxplot(data=monthly_aggregated_sca.apply(pd.Series).T, color=custom_palette,
showfliers=False, medianprops={'color': 'red'})

# Set plot labels and title
plt.title('Monthly Distribution of Snow Cover Area Across 2001-2012', font-
size=16)
plt.xlabel('Month', fontsize=14)
plt.ylabel('Snow Cover Area % ', fontsize=14)

# Customize the grid and axes
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tick_params(axis='both', labelsize=12)

# Discharge Line Plot
plt.subplot(2, 1, 2)
sns.lineplot(x='date', y='snowcoverarea', data=df, hue='year', palette='Set2',
linewidth=1)
plt.xlabel('Year', fontsize=14)
plt.ylabel('Snow Cover Area %', fontsize=14)
plt.title('Snow Cover Areae Across 2001-2012 (Line Plot)', fontsize=16)
```

Using monthly data from 2001 to 2012, this Python script creates boxplots, line plots, and a correlation matrix heatmap. The distribution and correlations between discharge, precipitation, temperature, and snow cover area are revealed by the visualizations.

NEPAL ENGINEERS' ASSOCIATION

COMPARATIVE ANALYSIS OF LONG SHORT-TERM MEMORY (LSTM) AND MULTI-LAYER PERCEPTRON (MLP) MODELS FOR SNOWMELT RUNOFF PREDICTION IN THE HINDU KUSH HIMALAYAN REGION

Hansal Shrestha[1], Neeraj Adhikari [2,*]

[1] *Pulchowk Campus, Department of Applied Science and Chemical Engineering*

[2] *Tribhuvan University*

**Corresponding email: 078msccd005.hansal@pcampus.edu.np, hansalshrestha123@gmail.com*

Abstract

Hydrological forecasting in the Hindu Kush Himalayas (HKH) presents special challenges because of the complex interplay between climatic and environmental factors. The quantitative predictive capabilities of two well-established models, Long Short-Term Memory (LSTM) and Multi-Layer Perceptron (MLP), chosen for their proven performance in previous studies, are meticulously compared in this thesis. The analysis uses comprehensive data spanning 2001 to 2013, including discharge records from the Department of Hydrology and Meteorology (DHM), precipitation data from APHRODITE, temperature data from APHRODITE, and snow cover area information from Google Earth Engine with MOD09A1 V6.1. The study employs rigorous evaluation metrics, revealing nuanced insights into the hydrological processes. Contrary to expectations, the MLP model exhibited slight superiority, showcasing a nuanced understanding of the region's complexities. The quantitative assessment, including RMSE (LSTM: 0.2396, MLP: 0.1733), MAE (LSTM: 0.1698, MLP: 0.0841), R2 Score (LSTM: 0.9976, MLP: 0.9987), and NSE (LSTM: 0.9976, MLP: 0.9987), emphasizes the indispensable role of robust predictive models, showcasing the necessity of reliable models for enhancing accurate river runoff predictions crucial for effective water resource management and flood preparedness in challenging terrains like the HKH.

Keywords: *Long Short-Term Memory (LSTM), Multi-Layer Perceptron (MLP), Hindu Kush Himalayan region (HKH)*

1. Introduction

The impacts stemming from the escalating global warming and climate change phenomenon have ushered in a heightened frequency and severity of both drought and flooding events, thereby constituting one of the most formidable challenges confronting our aquatic

078MSCCD005_Hansal_Final_Thesis.pdf

ORIGINALITY REPORT

19%
SIMILARITY INDEX

PRIMARY SOURCES

1	elibrary.tucl.edu.np Internet	415 words — 3%
2	www.researchgate.net Internet	115 words — 1%
3	dokumen.pub Internet	103 words — 1%
4	spectrum.library.concordia.ca Internet	76 words — 1%
5	www.mdpi.com Internet	64 words — 1%
6	stackoverflow.com Internet	60 words — < 1%
7	www.coursehero.com Internet	59 words — < 1%
8	ksy1526.github.io Internet	46 words — < 1%
9	link.springer.com Internet	44 words — < 1%
10	dspace.nm-aist.ac.tz Internet	