Seasonal Streamflow Drought Forecasting based on Pattern Recognition Concepts using Statistical and Machine Learning Approaches

by

Arundi Maheka Ruwanpathirana

A thesis presented to Lakehead University

in fulfilment of the

thesis requirement for the degree of

Master of Applied Science

in

Civil Engineering Department

Thunder Bay, Ontario, Canada, 2024

I hereby declare that I am the sole author of this thesis.

I authorize Lakehead University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

Arundi Maheka Ruwanpathirana

I further authorize Lakehead University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Arundi Maheka Ruwanpathirana

Lakehead University requires the signatures of all persons using or photocopying this thesis.

Please sign below and give the address and date.

Abstract

Understanding and forecasting drought events is crucial for effective water resource management and mitigation planning. Forecasting droughts is challenging due to their inherently complex patterns and dependencies. However, there is a tendency for droughts to occur during specific seasons or times of the year and exhibit distinct seasonal variability. This research focuses on analyzing seasonal drought patterns using a grouped data concept, where similar data points are aggregated into groups to represent distinct hydrological drought conditions.

The objective is to develop a methodology that can effectively recognize and predict droughts based on these grouped streamflow data sets. In the proposed study exploratory data analysis techniques are used to recognize the seasonal patterns within the data to extract meaningful drought patterns from the streamflow data. The study employed a combination of statistical methods and machine learning techniques, including Markov models and Long Short-Term Memory models (LSTM), to forecast the grouped seasonal streamflow data. A Markov model is employed to model the transition probabilities among hydrological drought states, capturing the temporal dependencies in streamflow behaviour. Subsequently, a Hidden Markov model (HMM) is utilized to employ the underlying states (or underlying drought levels) in observed streamflow data. To further enhance forecasting capabilities, monthly and weekly LSTM networks are utilized to learn long-term sequential dependencies and forecast future streamflow drought patterns.

The study area was selected as the Palliser Triangle, the driest region in Canada. A total of 25 river stations (catchment area ranging from 319 to 47,800 km^2) were chosen, representing a range of river capacities: low flow (annual runoff range from 0 to 50 mm), medium flow ((annual runoff range from 50 to 175 mm), and high flow (annual runoff more than 175 mm) The monthly flow sequences of these rivers displayed the coefficient of variation ranging from 0.61 to 3.84, skewness from 0.57 to 8.39 and lag-1 autocorrelation from 0.2 to 0.63. In view of the highly skewed nature of monthly flows, the Box-Cox transformation was applied to normalize the data sequences and the normalization parameter λ ranged from -0.96 to 0.16. The Box-Cox transformation proved powerful for the normalization of flow data sets, which provided a strong platform for the analysis and forecasting of hydrologic droughts. The model results revealed that the discrete Markov model performed best for medium-flow rivers, achieving an average forecast accuracy of 65%, and the Hidden Markov model demonstrated superior performance for both low-flow and high-flow rivers, with an average forecast accuracy of 74%. The LSTM model showed consistent performance across all river types, providing monthly forecasts with approximately 80% accuracy and weekly forecasts with an impressive 90% average accuracy. Though the Markov models were found less

reliable and efficient in hydrologic drought forecasting, they offer valuable insights into the passages of drought passing through various states of severity. The outcome of this study is expected to provide valuable insights into the complex relationships between seasonal drought patterns and ultimately support more effective water resource management and drought preparedness strategies.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Dr. Umed Panu, Professor in the Department of Civil Engineering, for his unwavering encouragement, guidance, and financial support throughout my research journey. His insightful feedback has been invaluable to the successful completion of this thesis.

I am also profoundly thankful to Dr. Garima Bajwa, Professor in the Department of Computer Science, for her kind support and guidance in the machine learning methods that were crucial to my research. Her patience and willingness to share her knowledge was essential in overcoming many challenges along the way.

My heartfelt thanks go to my husband, Chamath for his endless patience and continuous support and to my parents for their constant encouragement. Their understanding and encouragement have been my pillars of strength during this journey.

Finally, I would like to extend my thanks to everyone who contributed to the completion of this thesis. Their help and support, whether big or small, have not gone unnoticed, and I am truly grateful.

To

Everyone, who trying to make this world a better place!

TABLE OF CONTENTS

List of Tables

List of Figures

Abbreviations

PDSI	Palmer Drought Severity Index
AWC	Available Water Content
CMI	Crop Moisture Index
SPI	Standardized Precipitation Index
SOI	Southern Oscillation Index
AR(p)	Autoregressive model
MA(q)	Moving Average model
ARMA(p,q))	Auto-Regressive Moving Average
ARIMA (p,d,q)	Autoregressive Integrated Moving Average
SARIMA	Seasonal Autoregressive Integrated Moving Average
PDI	Palmer Drought Index
HMM	Hidden Markov Model
FFT	Fast Fourier Transform
WT	Wavelet Transform
HOST	Harmonic Oscillator Seasonal Trend
ANN	Artificial Neural Networks
SVM	Support Vector Machine
KNN	K-nearest neighbours
RNNs	Recurrent Neural Networks
LSTM	Long Short-Term Memory
DMSNN	Direct Multistep Neural Network
RMSNN	Recursive Multistep Neural Network
SPEI	Standardized Precipitation Evapotranspiration Index
MLPNN	Multilayer Perceptron Neural Network
RF	Random Forest method
MI	Mutual Information method
SVR	Support Vector Regression
SVPDI	Standardized Vapor Pressure Deficit Drought Index
W-kNN	Weighted kNN
ENSO	El Nino Southern Oscillation
WFL	Wavelet Fuzzy Logic
ANFIS	Adaptive Neuro-Fuzzy Inference System
DLNN	Deep Learning Neural Network
FRBS	Fuzzy Rule-Based System
DT	Decision Tree
SRI	Standardized Runoff Index
SSA	Singular Spectrum Analysis

Chapter I: Introduction

1.1 Introduction to Hydrological Droughts

Hydrological drought is a prolonged period of natural water scarcity that is characterized by reduced streamflow, groundwater levels, and reservoir storage that can occur anywhere on the globe. It is a slow-onset disaster which can have severe implications for water supply, agriculture, health, economies, energy, and the environment. Unlike other types of droughts, such as meteorological droughts, which are defined by a lack of precipitation, hydrological droughts arise from a prolonged scarcity of precipitation, surface runoff, and other natural water resources. Hydrological drought reflects the combined impact on all-natural water resources and can persist even after normal rainfall patterns resume. It is important to understand that hydrological drought is different from aridity, which pertains to areas that are consistently dry and where water scarcity is a constant issue even under normal conditions (Beran and Rodier,1985). Unlike aridity, hydrological droughts are temporary periods of abnormally low water in rivers and reservoirs that can occur in any climate. These hydrological droughts are particularly challenging because they affect the water resources on which societies and ecosystems depend, often leading to long-term consequences.

1.2 Current Issues in Hydrological Droughts

Even though many people have less understanding of the severity of hydrological droughts, it affects people all around the globe more than any other natural disaster (Hewitt, 1997). Even in very recent decades, severe hydrological drought events in East Africa, Texas, and California have caused major losses in water supply causing huge damage to human life and the global economy (Hao et al., 2018).

Hydrological droughts can happen due to various natural causes arising from periodic weather patterns. Besides natural causes, droughts are caused by altered weather patterns which affect air circulation through the atmosphere. When there is an anomaly in surface temperatures, particularly

over the sea, air circulation patterns are altered which, in turn, changes how and where precipitation falls around the globe. Furthermore, the global human population continues to balloon leading to intensive agricultural practices requiring increased water demand to sustain humans as well as agricultural practices. Due to this increasing water demand, hydrological droughts have attracted more attention recently in the world. A study (Wada et al., 2013) estimated that between 1960 and 2010, the human consumption of water increased the frequency of drought in North America by 25%. Other than these causes, deforestation and soil degradation are human-made causes of droughts. When forests and vegetation disappear, the water cycle is fed by less water which makes the water cycle weak, and regions become more vulnerable to droughts.

Several hypotheses explain the combined atmospheric, air, and ground interface effects that result in hydrological drought conditions. One hypothesis is that the rainfall is reduced by the over-seeding of clouds by dust particles, while there are other hypotheses related to the Albedo increase. When the surface of Earth has a higher albedo, it means that more incoming solar radiation is being reflected into space, and less is being absorbed by the surface. Increased Albedo can reflect the heat away from the earth thus increasing subsidence and in turn reducing rainfall (Beran and Rodier, 1985). Biogenic aerosols, which include particles from natural sources such as industrial plants, trees, and marine organisms, can act as cloud condensation nuclei. An increase in Albedo can decrease the availability of biogenic nuclei for raindrops. These processes can increase the vulnerability to droughts.

1.2.1 Drought Definitions and Types

Human factors, such as water demand and water management, can exacerbate the impact that drought has on a region. Because of the interplay between a natural drought event and various human factors, drought can mean different things to different people, and it is defined in various ways that reflect different perspectives and interests.

The earliest known study on hydrological droughts was done by Palmer (1965), who stated that drought can be defined according to the varying interests in various fields, but basically, it is a prolonged and abnormal natural water deficiency period. Kumar and Panu (1997) defined hydrological drought based on a functional relationship between two key factors, the estimated

water demand, and the expected water supply. Wilhite and Glantz (1985) described hydrological drought as a “creeping phenomenon” such that it is difficult to identify how critical is a drought because it is challenging to spot the start and the end of a drought. For enhanced understanding, they clustered and defined drought into four types based on the kind of deficiency of natural water, namely meteorological, agricultural, hydrologic, and socio-economic. Due to climatic differences, what might be considered a drought in one location of the country may not be a hydrological drought in another location. There are many definitions and explanations for each type of drought, but basically, meteorological drought occurs when the amount of precipitation is less than the amount (or state) considered normal. Agricultural drought refers to a situation where the amount of moisture in the soil no longer meets the needs of a particular crop. Hydrologic drought occurs when surface and subsurface water supplies are below normal. Socioeconomic drought refers to the situation that occurs when physical water shortages begin to affect people.

It is crucial to understand and analyze droughts for forecasting purposes. Forecasting hydrological droughts is an important part of early warning systems of droughts, as they provide lead time to planners for threat responses, which helps minimize hydrological drought impact risk. It has a great impact on water availability and, therefore is particularly important for ensuring food and water security. Effective forecasting systems can give enough lead time to adequately plan for water storage, identify alternative sources of freshwater, implement new (water-saving) agricultural practices, and import food and water, if necessary.

This research introduces a novel approach to forecasting hydrological droughts based on a pattern-based concept. Chapter II provides a literature review of existing hydrological drought forecasting methods. Chapter III outlines the research objectives and scope. Chapter IV develops methodologies to identify and predict seasonal drought patterns within streamflow data. Chapter V details the selection of study areas and data collection. Chapter VI applies the developed models to these selected watersheds. Chapter VII discusses the results obtained from the models, and Chapter 8 concludes with the final findings and offers recommendations for future research.

Chapter II: Review of Literature

This chapter provides an overview of existing methods for drought forecasting, introducing various parameters used to measure droughts both qualitatively and quantitatively. It discusses how these parameters are incorporated into different forecasting methods, outlining the advantages and disadvantages of each approach/method. Additionally, the chapter explores how these methods have evolved over time to improve the forecast of drought parameters, highlighting advancements in forecasting techniques.

2.1 Existing Methods for Hydrological Drought Forecasting

2.1.1 Drought Variables and Parameters Used for Drought Forecasting

Drought forecasting methods are used to predict the hydrological conditions or drought conditions in a certain area by examining current meteorological/hydrological information along with past weather patterns (Maija Bertule, 2017). There are various methods which are used in drought forecasting, including those based on historical correlations between drought variables, stochastic processes and most recently, machine learning methods.

Drought forecasting methods require input data for analysis. The input data requirements change according to the type of drought being considered and/or analyzed. The variables and associated types of droughts include:

(i) Precipitation for meteorological drought analysis because meteorological droughts are caused by a lack of precipitation,
(ii) Stream flow, reservoir, and lake level data for hydrological drought analysis, and
(iii) Soil moisture and crop yield for agricultural droughts.

Furthermore, based on the combinations of the above inputs (precipitation, temperature, stream flow, and soil moisture) several drought indices have been derived in recent decades to study

agricultural droughts that can be used for forecasting (Mishra and Singh, 2010). The following are some of the commonly used drought indices for drought forecasting.

1. Palmer Drought Severity Index (PDSI):

The meteorological drought index was developed by Palmer (1965) to measure the departure of moisture supply. This index is based on the supply and demand concept of the water balance equation, considering more than just the precipitation deficit at specific locations. Palmer developed and tested using the regional climates of the U.S. The PDSI is calculated based on the monthly precipitation and temperature data, as well as the local available water content (AWC) of the soil. From the inputs, all basic terms of the water balance equation can be determined, including evapotranspiration, soil recharge, runoff, and moisture loss from the surface layer. The PDSI scale ranges from + 4 to - 4 indicating extreme wet conditions and extreme dry conditions, respectively. Even though PDSI is the most popular and commonly used indicator for drought classification, there are a few drawbacks such as difficulty due to the requirement of AWC of the soil and not including snow and human factors (irrigation) of the area.

2. Crop Moisture Index (CMI):

The CMI was developed by Palmer (1968) due to some of the drawbacks associated with PDSI becoming apparent. It is a weekly calculated short-term drought index especially suited to drought impacts on agriculture, which responds quickly to rapidly changing conditions. It is calculated by subtracting the difference between potential evapotranspiration and moisture, to determine any deficit.

3. Standardized Precipitation Index (SPI):

The SPI was developed by McKee et al. (1993) based on the standardized value of the precipitation. It is simply the number of standard deviations that observed cumulative precipitation deviates from the climatological average. Since the precipitation is typically not normally distributed, a transformation was applied to the distribution until it fitted to a normal probability distribution. It can be applied for one or more than one month, and a functional and quantitative definition of drought can be established for each time scale.

In the arena of hydrological drought, Yevejevich (1967) introduced the concept of truncation level to describe the drought in which the time series of drought variable (such as streamflow) is truncated at the demand level. All the episodes (termed as run) below the truncation level are designated as drought and above the truncation level are non-drought or wet periods. Five parameters are involved in characterizing droughts (as shown in Figure 2.1) which are duration, areal extension, magnitude (or intensity), probability of occurrence, and time of initiation/termination. These parameters provide a comprehensive framework for understanding and measuring droughts. Duration refers to the length of time a drought persists, while areal extension describes the geographical area affected by the drought. In the parlance of hydrologic drought, magnitude refers to the cumulative shortage of water below the truncation level during the drought duration. The intensity is expressed as the ratio of magnitude to duration. The probability of occurrence indicates the likelihood of a drought happening in each area. Lastly, initiation and termination denote the onset and end of a drought on the time horizon.

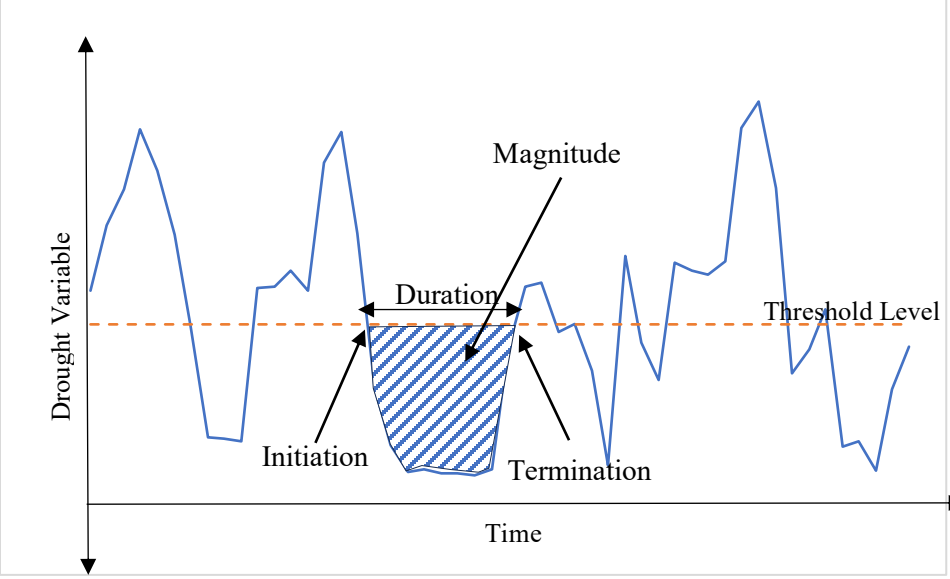


Figure 2.1 Parameters to define the characteristics of droughts.

In the study of drought forecast, various existing methods are employed to forecast the key parameters identified by Yevjevich as above. The following section briefly discusses different methodologies as well as their applications in drought forecasting.

2.1.2 Regression Analysis

Linear regression analysis is one of the earliest and most widely used forecasting approaches for drought data. Regression analysis helps to study the relationship between variables by analyzing the influence of independent variables on the dependent variable. Generally, multiple regression is used for drought analysis and independent variables are for which data is available and dependent data is the value that needs to be predicted. The formula for the regression analysis can be presented as follows.

$$Y_i = f(X_{i,}\, \beta) + e_i \quad \text{.....(2.01)}$$

The dependent and independent parameters respectively for i = 1, 2, 3, …, N with β as an unknown variable, f is the function to combine the independent parameter and the coefficients of variables, and e_i is the error term. The drought parameter is a variable which has a significant contribution to the occurrence of droughts. The dependent variable is a drought quantifying parameter, for example, a drought index, whereas the independent variable is an explanatory variable for the drought quantifying parameter (i.e., precipitation, stream flow, and soil moisture).

Some applications of regression analysis for drought forecasting are discussed below. Kumar and Panu (1997) developed a regression model using the stepwise regression procedure to predict the grain yield of a main crop as an agricultural drought quantifying parameter. Also, variables affecting the grain yield in the region (explanatory variables) were identified for agricultural drought forecasting. Based on the deviation of estimated grain yield from the long-term mean grain yield, the degree of drought severity as being non, mild, moderate, or severe was determined (Kumar and Panu, 1997). A study (Leilah and Al-Khateeb, 2005) analyzed the correlation and regression to clarify the relationship between wheat grain yield and its components under drought conditions. They found that the most effective variable that affects the grain yield under drought conditions, and identified the type of breeding materials which give more grain yield under drought conditions. Meng et al. (2016) developed a logistic regression model using the SPI of the previous season and the Southern Oscillation Index (SOI) to give the probability of drought occurrence in the current season and reported that the impact of SPI of the previous season, and SOI of current season drought substantially varies from region to region, and season to season. This study also showed stronger drought persistence during the summer compared to other seasons. In other

words, the probability of fall drought occurrence is closely related to summer moisture conditions (Meng et al., 2016).

Although regression analysis has been a commonly used method, there are some limitations as described below (Mishra and Singh, 2010).

All calculations are done in regression models assuming linearity between the predictand and predictors. Therefore, this type of model has less capability for long-term forecasting and can only be recommended for short-term forecasting such as agricultural droughts. Since the regression model only considers linearity and thus non-linearity and stochastic behaviour of the drought data can be captured from the error term, therefore, it is difficult to understand the hidden mechanism and correlation with the past values of the same variables. To address issues arising from stochasticity and internal correlation of a variable, researchers have utilized time series models which utilize lagged versions of variables for forecasting droughts as described below.

2.1.3 Time Series Models

A stochastic process is a family of random variables $\{X, \theta\}$, indexed by a parameter θ, where θ belongs to some indices set of time. It is simply a collection of random variables which describe the evolution of a system. There are many methods to analyze a stochastic process depending on the nature of a process. In time series models, the forecast of drought is done solely based on verified historical data of the same time series, hence the dependency of variables is the governing factor.

The first applied time series model for drought forecast was the Autoregressive model with a memory of p-lagged data (AR(p)), which predicts the future time series depending only on the current and past values of the time series. Autoregressive models for hydrologic forecasting were developed by Thomas and Fiering (1962), Yevjevich (1967), Box and Jenkins (1970) and Salas et al., (1980). Since 1970, many researchers have developed more defined models to get improved estimations for parameters based on the initial work done on autoregressive models. Since autoregressive models predict the future based on the memory of the current and past values, they give more accurate results when the time series data have fewer variations. In the dry season, when the drought values are governed by smaller flows (such as groundwater influences), an AR(p)

model is used. Instead of lagged values, the time series data can be forecasted solely based on the past errors in time series data, and this concept was also applied in the Moving Average model with q lagged errors (MA(q)). When time series have high variations such as during the wet season where drought data is governed by high flows (such as precipitation), the MA(q) model is used to capture variations and to adjust the forecast. To consider both high and low flow behaviour in drought time series data, AR(p) and MA(q) models were combined to formulate an Auto-Regressive Moving Average (ARMA(p,q)) model. In the ARMA(p,q) model, non-stationarity is removed by periodic standardization of daily, weekly, and monthly series but the required number of parameters becomes very large. To alleviate this situation, an Autoregressive Integrated Moving Average (ARIMA (p,d,q)) where non-stationarity series is differenced by the order of 'd' to obtain stationarity series. ARIMA models were developed accordingly to measure the non-stationarity within the seasons and thus Seasonal Autoregressive Integrated Moving Average (SARIMA) models were introduced. Since SARIMA models consider seasonality as a parameter, they are significantly more powerful than ARIMA models in forecasting complex time series that exhibit cyclicity and periodicity behaviours.

Depending on the foregoing observations, a drought quantifying parameter of a time series can be used for drought forecasting with time series models. Rao and Padmanabhan (1984) investigated the yearly and monthly PDI (Palmer Drought Index) series and characterized them via valid stochastic models which may be used to forecast and simulate the PDI series. Results revealed that PDI series can be forecasted with reasonable accuracy, one to several months ahead. However, the yearly forecasts were found to be much less accurate (Rao and Padmanabhan, 1984). A linear stochastic model (ARIMA and SARIMA) was used by Mishra and Desai (2005) to develop drought severity-area-frequency curves and to analyze the temporal variation of drought using the SPI for different time scales. The SPI is computed by fitting a Gamma probability density function to the frequency distribution of precipitation summed over the time scale of interest. The frequency curves thus developed by Mishra and Desai (2005) were used to predict short-term droughts and hence applied for drought preparedness plans. A comparative analysis of ARIMA and SARIMA models for drought forecasting indicated that ARIMA models can be used with reasonable accuracy for a 1–2-month lead time (Mishra and Desai, 2005). The capability of SARIMA models in streamflow forecasting was analyzed by Modarres (2007) who found that both observed as well as forecasted streamflow showed a drought period with different severity in the lead time. Results

also revealed the significance of the selection of a suitable threshold value in time series modelling (Modarres, 2007). Another study in Guanzhong Plain, China (Han et al., 2013) developed adequate linear stochastic models ARIMA and SARIMA to predict drought using SPI as the drought index and indicated that the forecasting accuracy of the ARIMA model increases with the time scale and are more powerful models in short-term drought forecasting.

Even though time series models gave much stronger and more flexible output based on the past data, they are still considering only the linear relation between the predictor and predicant values. Nonlinear and more complex relationships such as soil moisture relationship with precipitation and evapotranspiration cannot be predicted accurately from these models. Time series models give more accurate results when there is a large amount of data available. Also, it is always easier to predict with a higher confidence level with shorter lead times. Another main problem with time series models is that there are many types of time series models, and the reliability of results mainly depends on the capability and understanding of the user in the identification of the best adequate model amongst the candidate models (Mishra and Singh, 2010).Hydrological systems are always subjected to variations due to changes in weather patterns and land use practices. Therefore, it is very difficult to predict the exact behaviour of hydrologic processes using these models, and an uncertainty factor is always linked with the outcome. Therefore, probability models are used to deal with the complexity and uncertainty of a hydrologic process.

2.1.4 Probability Models

2.1.4.1 Markov Chain Models

Markov chain models are the most used probability models for drought forecasting. In Markov chain models, the probability of the next state is only dependent on the current state, not any previous states which implies independence of variables. This can be denoted as:

Probability $(X_t = a_j/(X_{t-1} = a_i, X_{t-2} = a_k, X_{t-3} = a_{1,..}, X_0 = a_0)) = \text{Probability } (X_t = a_j/X_{t-1} = a_i)$ ……...(2.02)

The conditional probability, Prob $(X_t = a_j / X_{t-1} = a_i)$, gives the probability that the process at time t will be in "state j" given that at time t-1, the process was in "state i". The term Prob $(X_t = a_j/X_{t-1} = a_i)$ is commonly called one-step transition probability. That is, it is the probability that the

process makes the transition from state a_i to state a_j in one time step, which is commonly denoted by "p_{ij}".

The Markov model is very important in the context of drought forecasting due to its ability to capture temporal dependencies and probabilistic transitions between different states of drought conditions. A Markov model is generally a mathematical model that follows the principles of a Markov process. It is a stochastic model that represents a system where the future state depends only on the current state, irrespective of any previous history. In other words, the Markov property assumes that the system has no memory of how it arrived at its current state. A Markov chain model is a specific type of Markov model that represents a system with discrete states and discrete time steps. It consists of a set of states and probabilities of transitioning from one state to another. The transition probabilities are typically represented in a transition matrix, where each element denotes the probability of moving from one state to another in a single time step.

The Markov models for drought analysis were first applied by Gabriel and Neumann in 1962. In their study, a Markov chain model was found to fit the daily rainfall data of Tel Aviv for the mid-winter period. Various aspects of rainfall occurrence patterns could be mathematically derived from the Markov chain, and these models were also found to fit observed data (Neumann, 1961). A non-homogeneous Markov chain model was used by Lohani and Loganathan (1997) based on the Palmer Drought Severity Index to calculate the steady monthly class probabilities for various weather classes for the Tidewater region in Virginia; where the stochastic behaviour of droughts was characterized as an early warning system in the form of all possible sequences of drought progression that was useful for drought management (Lohani and Lognathan, 1997). The first-order Markov chains to forecast drought conditions for future months are based on the current drought class described by the Palmer drought index (Lohani et al., 1998). In 2003, Steinemann (2003) used homogeneous Markov chain models and proposed a drought trigger preparedness plan at the basin scale to characterize the probabilities of drought transition. The seasonal SPI was forecasted (Cancelliere et al., 2007) by computing transition probabilities from one current drought condition to another in the future based on the statistics of the underlying monthly precipitation. The Markov Chain model was applied by Yang et al., (2018) for SPI values to define the spatial effects resulting from the spatial heterogeneity and the dependency. They defined local and spatial transition matrices to measure the spatial heterogeneity and the dependency other than the

traditional Markov transition matrices and showed that there is spatial heterogeneity in both 3-month and 9-month SPI values.

2.1.4.2 Hidden Markov Model (HMM)

HMM offers a unique way to understand sequential data by assuming that observations are influenced by hidden states. At its core, an HMM consists of two main parts: hidden states and observed emissions. Hidden states represent the underlying dynamics of the system, while emissions are the measurable outcomes associated with each state. The transition probabilities between states capture how the system evolves, while emission probabilities determine the likelihood of observing certain outcomes given a state.

Mohammed et al. (2021) discussed the use of a hybrid model called HMM-GA (Hidden Markov model-genetic algorithm) for forecasting the standard precipitation index (SPI) in the Bisha Valley, Saudi Arabia. The results showed a reasonable accuracy of 95%, implying the appropriateness of the HMM model as a tool for drought mitigation and warning systems Mohammed et al. (2021).

Even though Markov models are effective in quantifying uncertainty associated with hydro-meteorological factors that cause droughts, such models are argued to have the following disadvantages in drought modelling:

1. An apparent performance: The Markov chain model for forecasting a drought class may not be more informative than the historical proportion of the drought class. An analysis done by Silva and Estacio, 2020 showed at least 1000-year database would be necessary to guarantee forecast accuracy superior 63% - 75%. Even if such large databases may be made available by paleoclimatic studies, the hypothesis of stationarity of precipitation must be verified, as the pattern of drought class transitions may change over the centuries.
2. Less memory property: Markov models assume that the future state only depends on the current state and not on the past states. This low memory property limits the ability of the model to capture long-term dependencies or trends in the drought data. As a result, Markov models may not adequately capture the persistence or duration of drought events.
3. Stationarity assumption: Markov models often assume that the underlying data-generating process is stationary, meaning that the transition probabilities between states remain constant over time. However, drought conditions may exhibit temporal variations and non-

stationarity due to climate change or other factors. Ignoring non-stationarity can lead to inaccurate forecasts and limit the ability of such models to capture changing drought patterns.

4. Limited representation of spatial and temporal dynamics: Markov models typically treat each location or grid cell independently, ignoring spatial dependencies and interactions between neighbouring regions. Drought events can exhibit spatial coherence and propagation, which is not adequately captured by simple Markov models. Similarly, Markov models may not fully account for seasonal or inter-annual variations in drought patterns, limiting their ability to capture complex temporal dynamics.
5. Sensitivity to initial conditions: The performance of Markov models can be sensitive to the initial conditions or the starting state of the model. Small variations in the initial state can lead to significantly different forecast outcomes. This sensitivity can introduce uncertainty and affect the reliability of drought forecasts.

Transforming the time domain data into frequency domain data can aid in identifying different frequencies and periodic patterns present in the data. The Fast Fourier Transform (FFT) and the Wavelet Transform (WT) are the most used methods for frequency domain transformations and these approaches are invaluable for characterizing seasonality, reducing noise, and handling non-stationarity in time series. By extracting features from the frequency domain, such as FFT coefficients or WT features, one can enhance the performance of forecasting models and gain insights into the dynamic behaviour of signals. Incorporating FFT and WT analysis is particularly crucial when dealing with complex signals that exhibit diverse frequency characteristics, offering a comprehensive and nuanced view of the underlying data structure.

2.1.5 Fast Fourier Transform and Wavelet Transform

The Fast Fourier Transform (FFT) and the Wavelet Transform (WT) can be applied to analyze and extract information from streamflow time series data.

The FFT is a mathematical technique used to transform a time-domain signal into its frequency-domain representation by decomposing signals into sinusoidal components. In the context of streamflow forecast, FFT can be used to identify the dominant periodic components within the

streamflow data which in turn can help discern and identify the underlying patterns in the streamflow time series data. FFT especially helps detect the seasonal component or periodicity component of the data and filter out the noise embedded in the streamflow data. Almedeij (2016) employed Fourier spectral analysis to examine the cyclic structure of drought patterns and developed a long-term periodic model for drought forecasting using SPI values in an arid region of Kuwait. The study used the periodogram technique to identify periodicities in the drought data, revealing cycles of 12, 14, 19, 26, 31, 43, 64, 103, and 258 months. A new tool called the Harmonic Oscillator Seasonal Trend (HOST) model was introduced (Raczyński and Dyer, 2023) to assess temporal patterns of streamflow drought and used the Fast Fourier Transform to extract the first five harmonics of the streamflow signals. The best-fit model from the study resulted in an accuracy in the range of 70-90% and was able to capture annual precipitation variation and longer-term inertia. A comparative analysis of the performance of models trained with FFT-based features versus traditional time domain features was conducted and showed that FFT-based feature engineering is a promising approach to enhance the performance of time-series forecasting models (Galan-Sales and Jiménez, 2023).

Wavelet transform is a unique approach to analyze signals and extract meaningful information at different scales and resolutions by breaking down signals into wavelets, which are small, localized waveforms that capture both time and frequency information simultaneously. A comparison conducted by (Chong et al., 2019) revealed that when it comes to the temporal resolution for detecting both trends and periodic patterns, the Fourier transform seems less promising. Wavelet analysis can be applied to capture both high and low-frequency components in streamflow data with their temporal changes. These models demonstrate an improved performance compared to traditional FFT methods. The fundamental concept behind wavelet transform is decomposition and reconstruction which involves breaking down a signal into its constituent wavelets at different scales and positions, enabling the analysis of both high and low-frequency components with high precision. This decomposition process allows for the efficient representation of signals with both localized and global features (Sharma and Panu, 2024). A study used wavelet transform to analyze trends and significant periodicities in drought variables such as SPI and drought duration, severity, and peak in different regions of India which showed a significant decreasing trend in northeast India (Sharma et al., 2022). The study by Chong et al. (2022) analyzed the spatiotemporal

variability of droughts using wavelet transform and standardized precipitation indices (SPIs). The findings showed a declining trend in droughts from east to west in Sabah, Malaysia.

However, wavelets can possess several disadvantages that can limit their efficacy in forecasting tasks. One major drawback is their limited adaptability, as wavelet transforms are often tailored to specific types of data or signal analysis, making them less versatile across diverse forecasting scenarios. Additionally, wavelets may struggle to capture complex nonlinear relationships present in some datasets, leading to less accurate forecasts in such cases. Another challenge is the manual selection of wavelet parameters, including basis functions and decomposition levels, which can be a subjective and time-consuming process requiring extensive experimentation.

For such limitations, machine learning models offer several advantages over wavelet transforms for forecasting tasks. These models are highly flexible and adaptable, capable of learning complex patterns and capturing nonlinear relationships in the data. With the ability to automatically learn relevant features from the input data, machine learning models reduce the need for manual feature engineering, making them well-suited for high-dimensional or unstructured data.

2.1.6 Machine Learning Models

Machine learning is a useful approach for analyzing complicated and data-rich occurrences such as droughts and it has shown promise in drought forecasting due to its ability to capture complex patterns and relationships in time series data. Deep learning models such as Artificial Neural Networks (ANN) and hybrid models were found to be significantly powerful in understanding drought dynamics and predicting its impacts (Gyaneshwar et al., 2023).

Drought is a complex phenomenon influenced by multiple interacting factors. Machine learning models, such as ANN, Support Vector Machine (SVM), and K-nearest neighbours (KNN) models can capture nonlinear relationships and interactions among various meteorological, hydrological, and environmental variables. The ability of these models to capture nonlinear relationships enables them to uncover hidden patterns and dependencies that may not be easily captured by linear models (Shiri et al., 2021).

Drought forecasting often involves analyzing many variables, including precipitation, streamflow, temperature, soil moisture, vegetation indices, and more. Machine learning models can efficiently

handle high-dimensional data and automatically identify relevant features for predictive purposes. They can also handle missing data and outliers, allowing for more robust and accurate forecasts (Hao et al., 2019).

Drought is a time-dependent phenomenon, and machine learning models can effectively model temporal dynamics. Models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks excel at capturing temporal dependencies and patterns in time series data. Such models can learn from historical drought patterns to make accurate forecasts for future time steps. Machine learning models can handle complex data relationships and interactions, and can identify nonlinear and higher-order dependencies, as well as capture spatial and temporal correlations in drought data. This is particularly useful for capturing the spatial coherence and propagation of drought events across different locations (Shi et al., 2020).

Machine learning models can adapt and learn from new data. They can continuously update their internal parameters and improve their performance as new observations become available. This adaptability is particularly valuable in drought forecasting, where environmental conditions and climate patterns can change over time (Feng et al., 2021).

Another main advantage of machine learning models is that they can effectively integrate data from multiple sources, such as satellite imagery, climate models, ground-based observations, and remote sensing data. By incorporating diverse data types and sources, machine learning models can provide a more comprehensive and accurate understanding of drought dynamics. Drought forecasting involves inherent uncertainties due to the complexity and variability of environmental systems. Machine learning models can handle noisy and uncertain data, as well as quantify and propagate uncertainty through their forecasts. This allows for a more reliable assessment of drought conditions and associated risks (Zhang et al., 2022).

It is important to note that the choice of machine learning model depends on the specific characteristics of the data, the available predictors, and the desired forecasting horizon as follows,

1. Characteristics of the data: Understanding the characteristics of the data is crucial for selecting an appropriate machine learning model. This includes examining the nature of the input data, such as its format (e.g., time series, spatial data), the presence of any patterns or dependencies, the data size, and the availability of historical records.

2. Available predictors: The predictors or features are the variables or measurements used to make forecasts. In the context of drought forecasting, these could include meteorological data (temperature, rainfall, humidity), soil moisture levels, streamflow, remote sensing data, or any other relevant environmental or climatic variables. The choice of predictors depends on their availability, reliability, and their known or suspected relationships with drought conditions.
3. Desired forecasting horizon: The forecasting horizon refers to the length of time into the future for which forecasts are sought. Drought forecasting can involve short-term forecasts (e.g., drought severity of the next week) or long-term forecasts (e.g., seasonal or multi-year drought outlook). Different machine learning models may have varying capabilities and performance for different forecasting horizons.

The best fit for a specific drought forecasting task can be found by considering these factors. Some commonly used models for drought forecasting include regression models (e.g., linear regression, support vector regression), neural networks (e.g., artificial neural network model, recurrent neural networks), and more advanced techniques like deep learning models.

2.1.6.1 Artificial Neural Network Model

The most used machine learning model for drought forecasting is the Artificial Neural Network (ANN). It is a computational model inspired by the structure and functioning of biological neural networks, such as the human brain. ANNs are a subset of machine learning algorithms that are designed to learn and recognize patterns from data.

An ANN consists of three main layers, which are the input, hidden and output layers. Those layers are interconnected by artificial neurons, which are also called nodes or units. The most common type of ANN is the feedforward neural network, where information flows in one direction, from the input layer through one or more hidden layers to the output layer. Each neuron receives input signals, performs a computation, and produces an output signal that is transmitted to the neurons in the next layer. The connections between neurons are represented by weights, which determine the strength of the signal transmitted from one neuron to another. During the training process, the weights are adjusted based on the input data and the desired output, allowing the ANN to learn and adapt to the underlying patterns in the data. The computation within each neuron involves two main steps: a linear combination of inputs and an activation function. The linear combination

involves multiplying the input signals by their corresponding weights and summing them. The activation function introduces non-linearity to the output of the neuron, enabling the network to learn complex relationships and patterns in the data. For forecasting purposes, a 3-layer feed-forward ANN is used which consists of an input layer, a hidden layer, and the output layer. In the input layer, a linear combination is built between inputs, which are drought-quantifying parameters. In the hidden layer, nonlinear relationships are calculated using the activation function and, in the output, a layer is used for forecasting different lead times (Mishra and Singh, 2010).

ANNs can have different architectures and configurations depending on the problem they are designed to solve. A comparative study on the Recursive Multistep Neural Network (RMSNN) and Direct Multistep Neural Network (DMSNN) for drought forecasting found that RMSNN was useful for short-term drought forecasting, while DMSNN was useful for long-term drought forecasting (Mishra and Desai, 2006). ANN was applied to predict drought using SPI by Morid et al. (2007) and identified that a modeller is not required to completely define the intermediate interactions between inputs and outputs (physical processes). This property makes ANNs particularly suited for analyzing complicated processes such as drought forecast, in which the interactions with the output of several input variables must be investigated. Using the Standardized Precipitation Evapotranspiration Index (SPEI), a Multilayer Perceptron Neural Network (MLPNN) algorithm was applied for drought forecasting with results showing a potential capability (Ali et al., 2017). A study introduced a machine learning framework based on the recurrent neural network called "DroughtCast" to predict the United States Drought Monitor Index using the simulated meteorological parameters and the satellite-observed soil moisture data (Brust et al., 2021). The DroughtCast model gave promising results up to 12-week lead time and identified precipitation, soil moisture and temperature as the most important variables for drought forecasting. An ANN model was found better for drought forecasting in semi-arid areas than arid areas, where precipitation has less correlation with the stream flow (Tareke and Awoke, 2023).

Although the ANN model provides more promising results, it typically requires a large amount of high-quality training data to perform well. Insufficient or low-quality data can lead to overfitting or poor generalization, resulting in inaccurate forecasts. Drought data, particularly long-term reliable records, may be limited, which can pose challenges for ANN model development. Also are complex models with multiple layers of interconnected neurons. This complexity can make it

difficult to interpret the internal workings of the model and understand the underlying relationships between inputs and outputs. Traditional Artificial Neural Networks (ANNs) may struggle to capture long-term dependencies in sequential data due to the vanishing gradient problem. This issue arises during the training of deep neural networks, where the gradients of the loss function concerning the weights decrease exponentially as they are back propagated through the network. The vanishing gradient problem becomes particularly pronounced in deep networks with many layers, which is often the case in traditional ANNs. This lack of interpretability can limit the usefulness of ANNs in providing actionable insights for drought management. Sometimes ANN models tend to overfit and train from the noise data rather than learning the hidden patterns in the time series data.

2.1.6.2 Long Short-Term Memory (LSTM) Model

Long Short-Term Memory (LSTM) network is a Recurrent Neural Network (RNN) which was updated to address the challenge of capturing long-term dependencies in sequences, a hurdle that traditional ANNs often struggle with due to the vanishing gradient problem. The inclusion of memory cells and gating mechanisms in LSTMs enables the network to selectively retain and utilize information across different time steps, facilitating the capture of intricate temporal patterns. LSTMs are inherently designed for sequential data processing, excelling in tasks such as time series forecasting, natural language processing, and speech recognition where the temporal order of input data is crucial. Gating mechanisms in LSTMs, such as the forget gate and input gate, allow for controlled information flow, ensuring that the network can adapt to varying time scales and effectively handle both short-term and long-term dependencies. The ability to mitigate the vanishing gradient problem, coupled with state memory and multivariate time series handling, renders LSTMs more versatile and powerful in capturing the complexities of sequential data compared to traditional ANNs.

LSTM was introduced by Hochreiter and Schmidhuber (1997) to resolve the vanishing problem in Recurrent Neural Networks. The vanishing gradient problem arises when back-propagating errors pile up through many layers of a neural network, especially in the case of deep networks or networks with long sequences. The gradients can become extremely small, causing the weights to receive very small updates, and leading to slow or stalled learning. The LSTM architecture introduced memory cells and gating mechanisms to mitigate this issue and enable the network to

capture long-term dependencies more effectively. Selvaraj et al. (2023) forecasted the long-term rainfall and Standardized Precipitation Index for drought estimation using LSTM with more than 99% accuracy. Another study used a combination of the Random Forest (RF) method, attention mechanism, and LSTM network to identify the important input values from the monthly average precipitation values to estimate the precipitation value for the next month with less than 10% error (Li et al., 2022). An LSTM and a hybrid model based on a Regularized Extreme Learning Machine were applied to predict the drought level using the Multivariate Standardized Streamflow Index (MSSI) by Hameed et al. (2023) and found that the hybrid model produced more accurate results compared to LSTM. On the other hand, the use of LSTM to predict monthly rainfall values in Ghana appeared to outperform when compared to the standard SVR and RF forecasting models (Dotse, 2023). The research by Coşkun and Citakoglu, (2023) utilized the LSTM deep learning algorithm to predict meteorological drought in Sakarya province (Türkiye) using the standardized precipitation index (SPI) at 1, 3, and 6-month time scales. The LSTM model achieved the best results for the SPI-1-month time scale and the SPI-3-month time scale.

While LSTM networks have demonstrated success in capturing long-range dependencies in sequential data but have encountered several limitations. LSTMs can be computationally complex and require substantial training time, particularly for large datasets, making them resource-intensive. Despite such disadvantages, LSTMs remain valuable in various applications, and ongoing research continues to explore enhancements in the realm of hydrological drought forecasting using seasonal grouped data.

2.1.6.3 Support Vector Machines (SVM)

Support Vector Machines (SVM) is a popular machine learning algorithm used for both classification and regression tasks. It is effective in solving both linearly separable and non-linearly separable problems. SVMs are based on the concept of finding an optimal hyperplane that maximally separates different classes or approximates a regression line with the maximum margin. SVMs can handle non-linearly separable data by using the kernel trick. The kernel trick involves the use of a kernel function, which computes the inner product between pairs of transformed data points in the higher-dimensional feature space. The key idea is that this inner product can be computed directly in the original feature space without explicitly transforming the data.

SVMs have been employed to predict drought indices, such as the Standardized Precipitation Index (SPI) or the Palmer Drought Severity Index (PDSI). SVM models trained on historical climate data have been used to forecast these indices, providing valuable information about future drought conditions. A study conducted on the domestic water supply in Tehran, Iran applied SVM to predict the seasonal standardized precipitation index using the Mutual Information (MI) method as the feature selection method. Different Kernel functions were compared, and the linear kernel function was found to outperform the other kernel functions in SPI forecast (Zahraie and Nasseri, 2011). The Godavari River basin in India upgraded the performance of the SVM by applying the sequential minimal optimization method for SPI forecast and observed that the Pearson Universal Kernal function was found to give the best results (Pande et al., 2023) in the long forecast of SPI and concluded that the performance of different Kernel functions increased when predicting long term drought conditions.

SVM has been applied to classify drought severity levels based on meteorological and hydrological variables. SVM models have been used to accurately categorize drought conditions into different severity classes, such as mild, moderate, severe, or extreme drought. Schumacher (2020) trained SVM with precipitation, evapotranspiration, and soil moisture, and applied it to predict flash droughts up to 6-week lead time with reasonable accuracy. Another study in 2021 by Kolachian and Saghafan (2021) classified the standardized hydrological drought index into 9 classes using Support Vector Regression (SVR) and showed that classification by SVR has higher accuracy when the number of classes is lower.

There are, however, some drawbacks of SVMs such as being computationally demanding, especially when dealing with large datasets or high-dimensional feature spaces. Training an SVM requires solving a quadratic optimization problem, which can become time-consuming and memory-intensive for large-scale applications. Also, the selection of appropriate values for the hyper parameters is crucial for the optimal solution.

2.1.6.4 k-means Clustering and K-Nearest Neighbour Classification

The k-means clustering is an unsupervised machine learning algorithm used for partitioning data into distinct groups, or clusters, based on similarity patterns. The fundamental principle behind k-means is to iteratively assign data points to clusters in a way that minimizes the within-cluster sum of squares and effectively finds centroids that represent the center of each cluster. The algorithm

operates by initializing cluster centroids randomly and then iteratively refining them until convergence.

A combination of multi-timescale meteorological drought indices and the Self Organizing Maps-k-means clustering technique has been used to map future drought conditions in China for the period 2021-2050 (Peng et al., 2021); and results have indicated that the differences between precipitation and potential evapotranspiration are expected to increase throughout China, leading to more severe, frequent, and longer duration droughts in most parts of the country during the next three decades. Gamelin et al. (2022) successfully used the k-means clustering technique to identify known short-term and long-term drought events based on the Standardized Vapor Pressure Deficit Drought Index (SVPDI).

The K-Nearest Neighbors (K-NN) algorithm is a supervised algorithm for labelling an unknown data point given existing labelled data. The K-NN algorithm assumes that similar things exist in proximity. The K-NN captures the idea of similarity or closeness by calculating the distance between points on a graph. The nearness of points is typically determined by using distance algorithms such as the Euclidean distance formula based on the parameters of the data. The algorithm will classify a point based on the labels of the K nearest neighbour points, where the value of K can be specified.

The K-NN algorithm using temperature and precipitation data was applied to classify or predict agricultural drought events for five crop districts in Saskatchewan, Canada. Crop production quantities within the data period (1975-2002) were considered and based on the mean of the crop quantities, data were classified into two classes, which are drought and no drought. Then the drought class was reclassified to different drought levels starting from moderate to very severe droughts based on the mean values within the drought class. The overall accuracy was 83% for predicting non-drought events and 71% for predicting drought events (Boken and Haque, 2007). In 2019, a study in Hokkaido, Japan applied self-organizing maps (SOM) in combination with k-means clustering to classify the anomalies in weather patterns. Results gave more accuracy for a 3-day lead time but suggested further improvement of the method (Nguyen-Le and Yamada, 2019).

Weighted K-nearest neighbours (W-KNN) is an extension of the traditional K-NN algorithm where the contribution of each neighbour to the classification (or regression) of a query point is weighted based on its distance from the query point. In traditional K-NN, all neighbours within the k nearest

neighbours are given equal importance in determining the class (or value) of the query point. However, in Weighted KNN, the influence of each neighbour is adjusted by assigning weights inversely proportional to their distances from the query point. W-KNN is a novel approach in drought forecasting, but it has been applied in other domains such as gesture recognition, fault detection, and communication signal modulation recognition. A gesture recognition method called Wi-NN captures human body action information, achieving gesture recognition without the need for additional devices. A classification accuracy of 93% has been reported by Yajun et al. (2023) on the original data once such data was processed and incorporated into the weighted KNN. Additionally, the W-KNN approach improved the state separation ability of the fault detection study especially in cases where the operating data has a heterogeneous distribution (Qian, 2022). Another study (Yulin, 2022) proposed a pattern recognition method for communication-modulated signals based on weighted KNN where it was concluded that feature extraction significantly improved the recognition accuracy and computing efficiency of weighted KNN.

The K-nearest neighbour algorithm is one of the first algorithms that one learns in data science, and it is useful in classification problems due to its simplicity, quick calculation time and accuracy. However, as the size of the dataset differs, KNN becomes increasingly inefficient, compromising overall model performance.

2.1.6.5 Hybrid Models

Some studies have explored hybrid models that combine multiple machine-learning techniques or integrate machine learning-with other modelling approaches. For example, combining SVM and WT, or combining ANNs with Markov chains or hidden Markov models, has been proposed to improve drought forecasting accuracy.

A study by Belayneh and Adamowski (2013) compared hybrid models based on ANN, SVM and coupled wavelet-ANN revealed the highest accuracy for the coupled wavelet-ANN model with a 1 to 3-month lead time for the Awash River basin in Ethiopia. A Wavelet-Boosting Support Vector Regression model was applied for SPEI as the drought quantification parameter for drought forecasting at Langat River Basin, Malaysia. The results of the study showed a higher accuracy when WT was adapted for de-noising data compared to the support vector regression model (Fung et al., 2020). The Hybrid model from the nonlinear autoregressive model with the neural networks model was found to give better results than individual models analyzing the SPI for the Bursa

region in Northern Antonia (Evkayaa and Kurnaz, 2020). The study by Ozger et al. (2022) used El Nino Southern Oscillation (ENSO) and persistence as the drought parameters to develop the Wavelet Fuzzy Logic (WFL) hybrid model which produced more accurate results in drought forecasting compared to an artificial neural network (ANN) model and a coupled wavelet and ANN (WANN) model. Six machine learning models (Adaptive Neuro-Fuzzy Inference System (ANFIS), Artificial Neural Network (ANN), Deep Learning Neural Network (DLNN), Fuzzy Rule-Based System (FRBS), Support Vector Machines (SVM), and Decision Tree (DT)) were applied to the Han River basin in South Korea to calculate the Standardized Runoff Index (SRI) and found that the fuzzy rule-based network provided better performance compared to neural-based networks (Jehanzaib et al., 2021). A review of hybrid models used for drought forecasting by Alawsi et al., (2022) briefly described data preprocessing and the advantages and disadvantages of the different hybrid models. As for future research directions, data pre-treatment techniques, such as Singular Spectrum Analysis (SSA) and empirical mode decomposition were proposed.

Choosing the most suitable machine learning algorithm for drought forecasting can be challenging. Various algorithms have their strengths and weaknesses, and the selection of an appropriate algorithm for a specific drought forecasting task requires a thorough understanding of data characteristics, underlying assumptions, requirements of models, and reasonable experimentation.

It is important to assess the strengths, limitations, and assumptions of different models with specific characteristics of the data and the forecasting requirements to make an informed choice. Additionally, model evaluation and validation techniques are necessary to assess the accuracy and reliability of the chosen model for drought forecasting.

2.2 Discussion on Future Direction

The performance and accuracy of forecasting models can be influenced by several criteria. The understanding of the drought data can be improved by recognizing groups, classes, or patterns in the data set based on seasons. This is a more effective and less complex approach to drought forecasting rather than dealing with complicated models. The ability to predict, mitigate, and adapt to drought conditions can be enhanced by identifying and characterizing distinct patterns or categories of drought. This approach in drought forecasting can offer certain advantages over

regular drought forecasting methods. Some potential advantages of the seasonal pattern approach are:

1. Capturing nonlinear relationships: Drought patterns often involve complex, nonlinear relationships between meteorological variables, soil moisture, and other factors. Traditional methods, such as statistical regression models, may assume linear relationships and struggle to capture these nonlinearities. Seasonal pattern recognition techniques, on the other hand, can identify intricate patterns and dependencies that exist within seasons, allowing for more accurate modelling and forecasting. The drawbacks of linear models have been described by Tong (1983), who pointed out their inadequacy in predicting time series data.
2. Stability of the confidence boundary: Grouping data based on seasons can contribute to the stability of confidence boundaries throughout the group by providing an aggregated perspective rather than individual values for each element. By aggregating data within each season, the inherent variability within smaller timeframes is smoothed out, leading to more stable estimates of drought conditions. This reduction in variability contributes to greater stability in confidence boundaries, as they are based on averaged values across multiple observations within each season (Panu, 2000).
3. Incorporating spatial and temporal patterns: Droughts can exhibit spatial and temporal variability, with patterns emerging at different scales. Pattern recognition techniques can account for these patterns and capture the spatial and temporal dependencies present in the data (Unny et al., 1981). This allows for the development of spatially explicit models that consider regional variations and temporal dynamics, enhancing the understanding and forecast of drought patterns at different scales.
4. Uncovering novel insights: Pattern recognition approaches can uncover previously unrecognized patterns and associations in the data. By applying advanced data mining or machine learning techniques, researchers may discover new relationships, dependencies, or indicators of drought occurrence that were not apparent using traditional methods (Panu et al., 1978). These novel insights can lead to improved understanding and forecast of droughts, enabling more effective mitigation and adaptation strategies.

With the increasing application of pattern recognition approaches in computer science, image processing, speech recognition, data mining, and many others, future research will likely predict droughts based on their spatiotemporal patterns. This research focuses on pattern recognition in drought data to identify meaningful patterns and predict drought conditions based on those patterns. By analyzing and interpreting these patterns, the study aims to enhance the accuracy and effectiveness of drought forecasting.

In the next chapter (Chapter III), the research objectives and scope are outlined. Chapter IV will focus on methods for recognizing patterns within drought data and utilizing these patterns to make accurate drought forecasts.

Chapter III: Objectives

Analyzing drought impacts in Canada is vital for safeguarding agriculture, ensuring food security, and maintaining ecosystem health. A thorough analysis of drought can be gained through identifying heterogeneous patterns within drought data as droughts often manifest in diverse ways across different seasons, with varying intensities, durations, and impacts. Discerning the hidden seasonal patterns enables us to understand the dynamics of drought behavior which can enhance the predictability of droughts. The main objectives of this thesis are to set the foundation for seasonal drought streamflow forecasting in Canada based on pattern recognition concepts using statistical and machine learning approaches on a monthly time scale as follows.

1. To establish operational definitions of hydrological drought and describe qualitative and quantitative measures for drought assessment as it relates to drought-impacted sectors.

2. To discuss on the existence of heterogeneous Groups or Patterns within streamflow data.

3. To develop a measure of disorder and identification of similar behaviour within monthly streamflow patterns to recognise the seasonal drought patterns.

4. To analyze of seasonal drought patterns to identify the characteristics of each pattern.

5. To cluster of seasonal drought patterns to find the variation of drought levels within drought patterns.

6. To forecast seasonal drought patterns based on pattern recognition concepts using statistical and machine learning approaches.

    i. To forecast drought levels based on the seasonal drought patterns using the straightforward probabilistic framework of Discrete Markov Model.

ii. To forecast drought levels based on the Hidden Markov Model because it enhances the Discrete Markov Model performance by capturing more dynamic features of the streamflow drought patterns.

iii. To forecast drought levels based on the Long Short-Term Memory (LSTM) model to learn dynamic features in both underlying long-term and short-term dependencies in monthly streamflow drought patterns.

iv. To forecast drought levels based on the LSTM model by utilizing weekly data to analyze and capture detailed features of streamflow drought patterns by utilizing weekly data.

7. To conduct comparative analyses of three models on the differences in predicting drought patterns and model performances model for forecasting drought patterns based on different performance assessment metrics.

Chapter IV: Development of Methodology

This chapter describes the methods used to recognize drought patterns within streamflow data. It explains how these drought patterns are classified into different drought levels, and outlines the methods employed to predict droughts based on these identified patterns.

4.1 Assessment of the Completeness and Quality of Streamflow Datasets

To effectively model and analyze streamflow data, the assessment of the completeness and quality of the streamflow datasets before use is crucial. It is because without a thorough assessment of the streamflow datasets, the reliability and accuracy of the model results may be compromised.

The initial step of assessing the streamflow datasets involves ensuring that there are no gaps or missing records in the data set. In cases where necessary, the missing data values could be supplemented using values that would have minimal impact on the underlying probability distribution of the dataset. Following the data infilling, an evaluation must be executed to verify that the dataset adheres to the conditions of weak stationarity. This confirmation is crucial to ensure precise and consistent statistical evaluations.

4.1.1 Infilling of the Missing Data

Frequently, it is typical to come across gaps or temporal periods where records are missing in the hydrometric (streamflow) datasets that were regularly maintained by the Water Survey of Canada. These missing streamflow records needed correction before could be applied to any analysis. As pointed out by Tallaksen et al. (2004), even a minor number of missing data values could greatly diminish the significance of summarized statistical information. To address this issue, two methods can be applied to infill the gaps in the data set and the selection of such a method is based on the length of the gap. The procedure involves a Linear Interpolation method for gaps shorter than ten

days and an Analogue River Ratio method for gaps longer than ten days up and to a year, as described below.

4.1.1.1 Linear Interpolation Method

Missing data were filled after plotting the relevant hydrological year and calculating the trend line for the hydrograph. The hydrograph was plotted indicating the streamflow value (m^3/s) against each month and the trend line which gives the relationship between each streamflow value and the month was found using the remaining data of the year. Tallaksen et al. (2004) suggested that for short gaps of less than ten days, manual infilling using the Linear Interpolation method is appropriate, as long as there are no evident signs of flood or drought events.

4.1.1.2 The Analogue River Ratio Method

The Analogue River Ratio Method was used to in-fill missing data segments longer than ten days and up to a year. Tallaksen et al. (2004) described this method as straightforward, utilizing a nearby gauging station (analogue) to manually estimate missing flow values at the target station. The analogue station is recommended to be an upstream or downstream station of the target river or a station at the nearest catchment. This method relies on an established relationship, typically a ratio, between the flows at the target and analogue stations, which is then used to estimate the missing data.

To perform the infilling, the target river and the nearest adjacent river(s) should ideally have over 30 years of flow data. This data is used to develop a daily relationship (Equation (4.01)) for each day of the year. According to Brase and Brase (2006), a sample size of more than 30 years is generally used to approximate normality. The relevant equation is provided below.

$$x_i = \frac{\sum x_d}{\sum a_d} * a_i \quad \text{.....(4.01)}$$

Where x_i represents an estimate of the missing value on the i^{th} day and $\sum x_d$ gives the summation of known daily streamflow (m^3/s) values over d-length days of the target station and a_i is the known daily streamflow (m^3/s) value of the i^{th} day in the adjacent station and $\sum a_d$ gives the summation of daily streamflow (m^3/s) over d-length days of the adjacent station. Length d is ten days which starts from the nearest day where there are no missing data for both stations over a d-period.

4.1.2 Quality of the Data Set

It is important to assess the quality of the streamflow datasets. Assessing precision involves examining the variability and repeatability of the data while evaluating consistency entails checking for any inconsistencies or outliers within the dataset. A study (Vivian, 2022) showed that data quality for hydrological modelling is crucial, pre-treatment, including outlier removal should be done before using data for hydrological modelling as after removing the outliers in the study, the model classification performance increased up to at a 95% confidence level.

It is necessary that the time series data must follow stationarity for statistical modelling of the time series. Stationarity refers to a property of a time series where its statistical properties, such as mean, variance, and autocorrelation, remain constant over time. But in hydrology, only the assumption of weak stationarity is acceptable, which includes both the first moment, the mean, and the second moment, the variance is constant in time. Achieving this involves checking and adjusting the average and variability across various time intervals and eliminating trends by differencing the data. Overall, assessing the completeness and quality of streamflow datasets is essential for ensuring the reliability and accuracy of streamflow modelling.

This assessment involves checking for any missing data points or gaps in the time series, as well as evaluating the accuracy, precision, and consistency of the data through techniques such as sensitivity analysis, calibration, and validation. This assessment is necessary to identify any potential issues or limitations with the datasets that could impact the reliability and accuracy of streamflow modelling.

4.2 Normalization of the Data Set

Normalization of streamflow data is a necessary preprocessing step when applying streamflow data for hydrological analysis. It is necessary to ensure that the underlying distribution of the data is known, which makes the behaviour of the data set visible making it useful for model applications.

4.2.1 Need for Normalization

The original streamflow data for most rivers possesses a skewed distribution. Transforming original streamflow data to normalized data can help to improve the stability, efficiency, fairness, and interpretability of models. Those advantages can be described according to the following factors.

1. Improving the Model Performance: Many mathematical theories and statistical techniques assume that the data follow a normal distribution or some other specific distribution. Most of them assume the normal distribution. Normalization of the data can help meet these assumptions, potentially leading to better model performance and more accurate forecasts.
2. Outlier Handling: There are extreme values or outliers in the hydrological data which usually result from catastrophic natural events such as extreme rainfall and extreme snowfall events, cyclones/hurricanes, or large forest fires and these can have a significant impact on hydrological analyses. While the traditional approach involves removing outliers, in hydrology outliers can be considered as part of the hydrograph, which includes potentially valuable information (Panu and Ng, 2017). However, normalization can help mitigate the influence of outliers by compressing extreme values and making them less extreme in the transformed data.
3. Convergence and Stability: In machine learning methods, many optimization algorithms are used such as gradient descent, and those optimization methods that converge faster and are more stable when dealing with normalized data. This can help the model reach a solution more efficiently and prevent convergence issues that might arise from unscaled or poorly scaled data sets.

When normalizing streamflow data, it is important to consider the distribution of the data and the specific needs of the modelling approach. Common normalization techniques can be varied from simple transformations such as min-max scaling, z-score standardization, and logarithmic transformations to more advanced transformations such as Box-Cox transformation (Box and Cox, 1964). In this thesis, the Box-Cox transformation (one of the most common transformations that has been successfully used in the field of hydrology) has been utilized.

4.2.2 The Box-Cox Transformation

The Box-Cox transformation (Box and Cox, 1964) is a mathematical technique that transforms the original data into an approximate normal distribution. It is particularly useful when dealing with data that exhibits heterogeneity, that is, streamflow data exhibits unequal temporal variability across different levels of streamflow data, making streamflow data laden with non-stationarity. At a given location, these varying levels of variability throughout the year could occur due to seasonal variations. The Box-Cox transformation can help stabilize the variance, making it more consistent across the data range. This can be particularly important when conducting statistical analyses that assume constant variance or stationarity. Streamflow data often exhibit significant skewness, with most values clustered at lower flow rates and a few extremely high values during flood events. This skewness can make statistical analysis challenging. The Box-Cox transformation can help reduce the skewness, making the data more symmetric and closer to a normal distribution.

There are two conditions for using this transformation, which is that the values should be positive values of continuous variables. Both conditions are satisfied with monthly streamflow values. Box-Cox transformation uses an exponent of a variable named lambda (λ) for the transformation, and λ can vary from -5 to 5. All values of λ need to be considered and the optimal value of λ for the data should be selected such that the optimal value is the one which results in the best approximation of a normal distribution. The Box-Cox transformation of the y variable has the following form:

$$y(\lambda) = \begin{cases} \frac{x^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(y) & \text{if } \lambda = 0 \end{cases} \quad \text{.....(4.02)}$$

Where x is the original data or monthly streamflow values in m^3/s, and $y(\lambda)$ is the transformed data. The parameter λ determines the type of transformation applied to the data. If the optimal λ is zero, data needs a natural logarithmic transform and when optimal λ is one then data is already normalized, it will be reduced by one unit keeping the shape of the distribution as same as the original distribution. To find the optimal value of λ, the maximum value of the likelihood function should be calculated. The likelihood function assumes that the original data follows a normal distribution after the Box-Cox transformation. It is expressed as:

$$L(\lambda) = -\frac{n}{2} Ln\left(\sigma_{\lambda}^2\right) + (\lambda - 1) \sum_{i=1}^{n} (y_i) \quad \text{.....(4.03)}$$

where n is the number of data points, y_i is the i[th] data point and σ_λ^2 is the estimate of the least squares variance using the transformed y variables for a specific λ value.

The likelihood function L(λ) quantifies how well the transformed data fits the assumed normal distribution after the Box-Cox transformation, for a specific value of λ. The value of λ that maximizes this likelihood function, corresponds to the best transformation for the original data.

In practice, maximizing L(λ) can be done by numerical optimization methods using different software packages like R or Python libraries such as SciPy. In this thesis, the MATLAB software is used to perform Box-Cox transformations and estimate the optimal λ value based on the likelihood function. By selecting a value of λ that maximizes the likelihood function, the original streamflow data can be effectively transformed to make it more closely resemble a normal distribution.

4.3 Recognition of Patterns within Monthly Streamflow Data

Streamflow data, which is typically measured as discharge over time, contains a combination of information. The meaningful insights can be extracted from the data by recognizing the underlying patterns within the data set.

4.3.1 Existence of Groups or Patterns within Streamflow Data

Streamflow data provides an integrated measure of the water cycle, reflecting the combined effects of various hydrological processes such as precipitation, infiltration, and evapotranspiration within a watershed. Therefore, streamflow can exhibit similar characteristics due to factors such as seasonal changes and climate oscillations. These factors can result in heterogeneous groups or clusters of streamflow based on similar behaviours. For instance, during wet seasons or periods of snowmelt, streamflow levels tend to rise, forming clusters of high-flow data. Conversely, during dry seasons or drought conditions, streamflow levels decrease, resulting in clusters of low-flow data.

The presence of heterogeneous groups in streamflow data was first documented by Panu and Afza (1993), and Goodier and Panu (1994). Identifying the structure of groups and efficiently capturing

heterogeneity in streamflow data based on pattern recognition concepts was introduced by Panu et al., (1978). This approach can provide main two advantages over univariate models, (1) Increasing the stability of the confidence boundary throughout the group and (2) Predicting the level of the drought for the entire group instead of predicting the drought level for a month/week/day. In addition to the main advantages, there are several benefits such as leading to improved model performance because the models can learn more efficiently from data with consistent behaviors. This can be particularly important when the original dataset, particularly streamflow data has a lot of fluctuations that might make it harder for the model to discern meaningful patterns. Also, grouping data based on similarity can result in more interpretable models. Instead of dealing with individual data points, the model is working with aggregated groups that can represent specific hydrological conditions such as seasons, which can make it easier to understand and communicate the results of the models. In this section, exploratory data analysis techniques are used to identify seasonal patterns and variations within the data to extract meaningful drought patterns from the streamflow sequences.

4.3.2 Streamflow Drought Index (SDI) based Drought Levels (DL)

The Streamflow Drought Index (SDI) is used to get a qualitative measurement of the drought level for a month. It analyzes streamflow data to determine how much a streamflow value deviates from historical patterns and assigns a drought level based on the variation. As an initial step of SDI calculations, the streamflow data is normalised and then the SDI value for each month is obtained as follows.

$$SDI_i = \frac{y_i - \bar{y}}{Sd} \quad \text{.....(4.04)}$$

Where y_i is the normally transformed streamflow data for ith month and $\bar{y}$ and Sd are the mean and standard deviation respectively for normally transformed streamflow data of a specific time period being analyzed (Nalbantis and Tsakiris, 2009). For each month based on the SDI values, the drought level can be found as follows. This process transforms the normalised streamflow data into standardized units known as standard deviations from the mean. The SDI indicates how many standard deviations the observed streamflow is from the mean streamflow for a given time. As only droughts are considered in this research, only negative values, which indicate below-average

streamflow are categorized as shown in Table 4.1 and all positive values indicating above-average streamflow are considered as no drought states.

Table 4.1 Drought levels based on SDI values.

State	Drought Level	Criterion	Probability (%)	Threshold Values of Normally Distributed Flow
0	No Drought	SDI ≥ 0.0	50	Q50
1	Mild Drought	-1.0 ≤ SDI < 0.0	34.1	Q66
2	Moderate Drought	-1.5 ≤ SDI < - 1.0	9.2	Q90
3	Severe Drought	-2.0 ≤ SDI < - 1.5	4.4	Q95
4	Extreme Drought	SDI < -2.0	2.3	Q97.5

The threshold values for each drought level in the flow were determined based on the empirical rule of the normal distribution. This rule, which states that approximately 68%, 95%, and 99.7% of data within a normal distribution fall within one, two, and three standard deviations from the mean, respectively, was used to classify the flow data into different drought levels. By applying this statistical method, the thresholds were systematically established to reflect the variability and extremity of the flow conditions associated with each drought level (Pramanik et al., 2009). The drought level or the drought state of each month used to calculate the drought level of the group is described in the following sections.

4.3.3 Drought Level (DL) of the Group

As discussed in the section 4.3.1, there is heterogeneity across the streamflow data and can be grouped based on the similarity within a group. A cluster of data points which shows a similar behaviour is called a "group" or a "pattern" from hereafter. To characterize the overall drought level across the group, a common practice which involves calculating the mean drought level (DL) for the group was considered. This mean drought level represents the average drought severity throughout the analyzed timeframe. Therefore, the group drought level was calculated as follows.

$$\text{The drought level of the group or } (DL_{Group}) = \frac{1}{L}\sum_{i=1}^{L} DL_i \quad \ldots\ldots\ldots\ldots (4.05)$$

In which, L is the optimal length (or the optimal number of months) in a group. To simplify the process of finding the optimal drought length, a hydrological year was divided into equal-length groups. By consistently grouping the time periods, each group contains the same number of observations or the same number of monthly drought levels. This uniformity allows for straightforward calculation of the mean drought level (DL) for each group. The mean DL represents the average drought severity within each group, providing a single value that summarizes drought conditions over that period. This approach not only streamlines calculations but also ensures comparability across groups, making it easier to analyze and interpret drought patterns throughout the year. It is noted that such groups in a monthly streamflow time series could comprise of time duration of 2, 3, 4, 6, and 12 months and thus yielding respectively 6, 4, 3, 2, and 1 group per year. For brevity in demonstrating the methodology in this thesis, only a group length of 3 months is considered and hence only four patterns per year are considered.

4.3.4 Shannon Entropy Calculations to Recognize (or Categorise) Groups

Shannon Entropy (Information Entropy) is a concept from information theory that can be used to measure the disorder or uncertainty within a group of data. A group with high Shannon entropy indicates a high degree of disorder or heterogeneity indicating the data points within the group exhibit significant variability. Conversely, a group with low Shannon entropy suggests a high level of order or homogeneity indicating that the data points are relatively similar, with minimal variation. Such a concept based on entropy consideration can be applied to measure the disorder of data within a group to quantify the homogeneity of the data within each group and conversely to measure the heterogeneity between groups.

Shannon Entropy is expressed as follows (Shannon, 1948).

$$H(X) = -\sum_{i=1}^{n} p(X_i) log_2\, p(X_i) \quad \text{.....(4.06)}$$

Where: H(X) is the entropy of the data measured in bits, n is the number of data points and $p(X_i)$ is the probability of the i^{th} data point. In the context of pattern recognition in hydrology, entropy can be used to assess the different patterns by measuring entropy within the group (intra-entropy) and measuring entropy among various groups (inter-entropy).

A group consists of monthly normalised streamflow values that behave similarly. In reality, a hydrological year might have seasons of varying lengths, leading to a different number of seasons each year. However, in this thesis, as stated earlier, all hydrological years are assumed to have the same number of seasons with identical durations.

4.3.4.1 Intra-Entropy (Within-Group Entropy):

Intra-entropy measures the level of disorder or uncertainty within individual groups. It quantifies how well data points within a cluster are grouped or how similar they are to each other. Lower intra-entropy indicates that the data points within a cluster are more similar indicating a high degree of order and homogeneity within the group. Intra-entropy helps assess the cohesion or compactness of clusters. When clustering is effective, data points within a cluster should be more similar to each other, resulting in a low value of intra-entropy.

Intra-entropy for a specific group was calculated using the entropy expression as described in equation (4.06) and the drought level (DL) of each month within a group is considered as a data point (X). The entropy is computed based on the probability of each drought level within the group.

4.3.4.2 Inter-Entropy (Between-Groups Entropy):

Inter-entropy measures the level of disorder or uncertainty between groups. It quantifies how distinct or dissimilar different groups are from each other. Higher inter-entropy indicates that groups are well-separated and distinct from each other, signifying a clear distinction and heterogeneity between them.

Inter-entropy was calculated considering an entire group as a data point(X). The drought level of a group can be found by equation (4.05). After calculating the group drought levels, the probability of occurrence of each drought level can be found and used in equation (4.06) to obtain the inter entropy.

The crucial aspect of evaluating the optimal structures of groups or patterns lies in achieving a balance between internal coherence and external distinctiveness. This implies that the optimal structure of groups has a minimum intra entropy, ensuring data points within a group are cohesive, and have a maximum inter entropy, making groups externally distinct. In this study entropy calculations for different group structures are carried out and the optimal structure of a group/pattern can be denoted as the length of L months.

4.3.5 Moving-Average Calculations

After finding the optimal structure or the length of a group as L months, Moving Average is used to find the months that belong to a group and the seasonality hidden within a hydrological year. Moving average is a simple and effective method to reduce the fluctuations or randomness within the data, making it easier to see underlying trends or patterns. It involves calculating the average value of a set of data points within a sliding window as it moves through the streamflow time series. In this case, the moving window is taken as L months. This moving window calculates a new series $z_1, z_2, z_3, \ldots, z_n$ where z_i is the average for streamflow values of i^{th} to i+(L-1) consecutive months as described below.

$$z_i = \frac{\sum_{i=1}^{(w-1)+i} x_i}{w} \quad \ldots\ldots(4.07)$$

Where w is the moving window length (w is equal to L), x_i is the streamflow value (m^3/s) for the month i and z_i is the moving average for starting from i^{th} to $(w-1)+i^{th}$ month.

4.3.5.1 De-seasonalization by Moving Average

This is a technique used in hydrological time series analysis to separate seasonal variations from the time series to reveal the underlying trends and patterns. After calculating the moving average of the monthly streamflow data, the mean of the moving average from each z value is subtracted to isolate the seasonal components. After finding the seasonal components of the data points in all the years of the study period, the average seasonal component for a month can be calculated by getting the average of the specific month over all years. By comparing the seasonal component for each month, hidden seasons or groups within a hydrological year (a hydrological year in the North American context runs from October first of a year to September 30 of the following year) can be indicated. The seasonal decomposition of the j^{th} month in the hydrological year can be described as follows,

$$\text{Seasonal components of the } j^{th} \text{ month} = s_j = \left(\frac{1}{n}\sum_{i=1}^{n} z_{j,i}\right) - \bar{z}_j \quad \ldots\ldots(4.08)$$

Where j is the month of the hydrological year and ranges from 1, 2, 3, …, 12 and n is the total number of years in the study period, which should be a minimum of 30 years. As this determines the seasons in a year and the groups are based on these seasons, the groupings or the patterns identified from entropy calculations can be considered seasonal groups or seasonal patterns. To

identify the different clusters within a given seasonal group, a K-means clustering algorithm can be employed.

4.3.6 K-means Clustering Algorithm

To recognize the optimal number of clusters of patterns within a given seasonal pattern in all the years, K-means clustering is used. It can discover natural groupings or clusters within a dataset. A data point will be a seasonal pattern which was identified from Moving average de-seasonalization and hence the data set is L dimensional.

The basis of the K-means clustering algorithm is to group similar patterns or groups based on their similar features, which in this case are streamflow values. It is an unsupervised machine learning algorithm with a goal to assign each group to one of the K clusters in such a way that the within-cluster variance is minimized, leading to compact and well-separated clusters.

As shown in Figure 4.1, the process begins by randomly initializing K centroids. Each data point is then assigned to the nearest centroid, forming several K clusters. After the initial assignment, the centroids of these clusters are recalculated as the mean of all data points within each cluster. The assignment and centroid update steps are repeated iteratively, such that data points are reassigned to the nearest centroid, and centroids are recalculated. This process continues until convergence, which occurs when the centroid assignments no longer change, or the centroids stabilize. The final output is K clusters with minimized within-cluster variance, ensuring that data points within each cluster are as similar as possible, while clusters are as distinct as possible from each other.

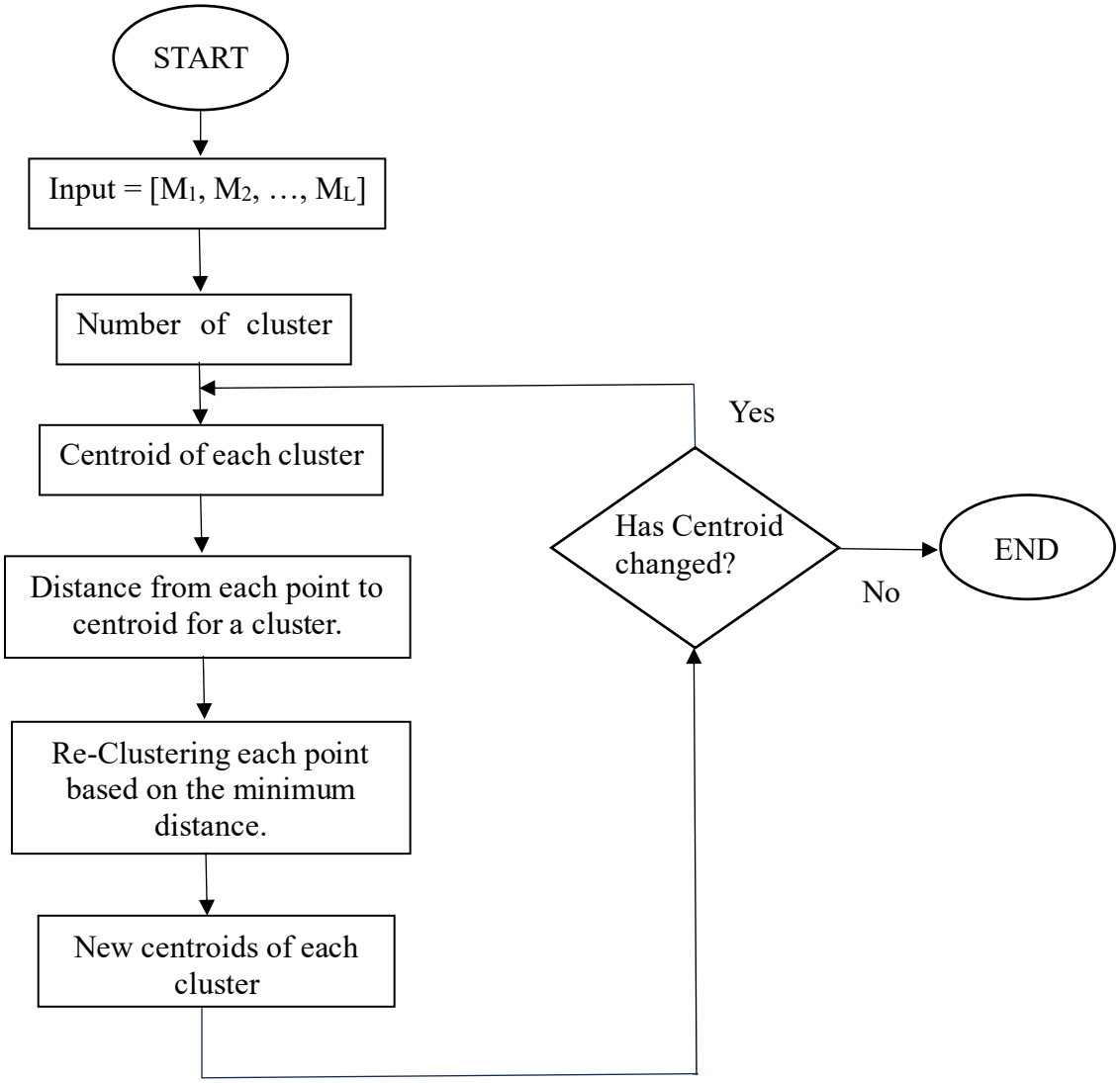


Figure 4.1 Flow chart for k means clustering algorithm.

4.3.6.1 The Silhouette Scores to find the Optimal Number of Clusters

The silhouette score is a metric used to evaluate the optimal number of clusters *K* within different numbers of clusters. It measures how similar an object is to its cluster (cohesion) compared to other clusters (distinction). The silhouette score ranges from -1 to 1, where a score close to +1 indicates that the data point is well-clustered, with instances within the same cluster being close to each other and far from instances in other clusters. A score around 0 indicates overlapping clusters,

where data points may be close to the decision boundary between clusters. A score close to -1 suggests that data points may have been assigned to the wrong cluster.

In this thesis to compute a silhouette score for a single data point, two distance parameters should be calculated for a seasonal pattern of L monthly normalised streamflow values. The first one is the average distance from a data point to all other points in the same cluster, denoted as "*a"*, representing cohesion. The second parameter is the average distance from a data point to all points in the nearest neighbouring cluster (the cluster other than its own) is taken as "*b"*, which represents the separation. Then the silhouette score *S* for a data point is computed as follows.

$$S = \frac{b-a}{\max(a,b)} \quad \text{.....(4.09)}$$

The average silhouette score across all data points should be calculated to compute the silhouette score for the entire dataset for a given number of clusters. A range from 2-10 is taken as the number of clusters and the silhouette score for different numbers of clusters is calculated. The optimal number of clusters can be selected as the number of clusters which gives the maximum silhouette score.

4.4 Forecast of the Seasonal Drought Level

As seasonal patterns within values of streamflow are recognized, it becomes crucial to predict/forecast future patterns using the recognized historical patterns. Predicting these patterns offers several benefits over simply predicting monthly flow values, as outlined in section 4.3.1 of this chapter. These patterns are interdependent due to the natural interdependencies of dry and wet seasons throughout a hydrological year. This dependency allows for the application of dependent models, such as Markov models, which can effectively predict future drought states depending on the current conditions. This section delves into the exploration and development of three distinct models for the forecast of the subsequent pattern: the discrete-time Markov model, the Hidden Markov Model, and the Long Short-Term Memory model. Each of these models brings its unique approach to forecast/forecasting future patterns in streamflow, providing a comprehensive analysis of their predictive capabilities of drought patterns.

4.5 MODEL 1: Development of Discrete Markov Model (DMM) for Drought Forecasting

In this section, the process of predicting seasonal patterns in drought data using the Markov model theorem is explored and/or developed.

As described in the introduction, a Markov Model is a stochastic process that behaves according to Markov property. The base of the Markov Model undergoes transitions from one state to another in a state space. The core concepts include states, which represent the possible conditions (e.g., different levels of drought severity), and transition probabilities, which quantify the likelihood of moving from one state to another. A crucial feature of Markov Model is the less memory or lag 1 memory property, meaning the probability of transitioning to any future state depends solely on the present state and not on the sequence of events that preceded it. By establishing a solid understanding of these concepts, one can proceed to analyze how they can be applied to accurately predict patterns in drought data. This involves defining appropriate states for drought conditions, estimating the transition probabilities based on historical data, and using these probabilities to forecast/predict future drought seasonal patterns.

4.5.1 State Space Models

In state space consideration with a Discrete Markov Model is a non-empty set with given values associated with the process in the set. Dealing with discrete Markov Model, the state space will either be finite or countably infinite. In the context of drought forecasting using streamflow data with a Discrete Markov Model, the state space represents the collection of all possible drought levels for groups or seasonal patterns (DL-Group) that the system can be in at any given time. These states are based on the Streamflow Drought Index (SDI) as discussed in section 1.3.3. A time step t in the monthly streamflow time series corresponds to a specific range of SDI values, indicating a specific Drought Level (DL_t). The average of the drought levels of the individual time steps within the group (length of the group is L) indicates the drought level for each group (DL-Group). The number of groups within a year can be changed according to the length of the group. Figure 4.2 shows how the Drought Level of each month and the Drought Level of each group

change when the length of the group is 3 months. The monthly and group indices are respectively denoted by i and j to show how those indices change throughout a given year n.

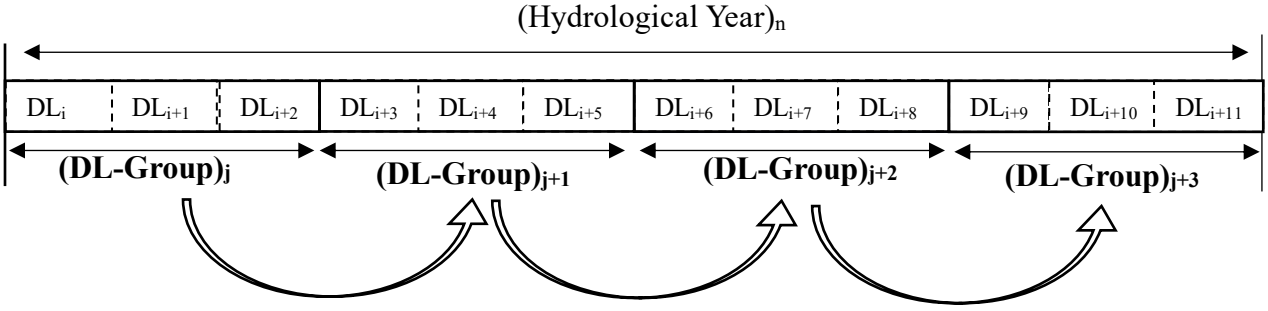


Figure 4.2 Variations in interdependence among Drought Level of the group (DL-Group) when the group length is 3 months.

4.5.2 Markov Property

A stochastic process is a Markov Model if it satisfies the Markov Property as follows.,

$$P\,[X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \ldots, X_1 = i_1, X_0 = i_0] = P\,[X_{n+1} = j \mid X_n = i] \quad \text{.....(4.10)}$$

for all states $i_0, i_1, \ldots, i_{n-1}$, i, j and $n \geq 0$.

This equation, known as the Markov Property, describes that a stochastic process, is a Markov Model given that the future position of states has no dependence on the past position of state transitions, but only depend on the current state. In other words, what happens with the next transition between states is independent of previous movements of states implying that the probability of moving from the current drought level to the next drought level depends solely on the current drought state. It does not consider how the system reached that state. For group drought forecasting, the group drought level of the group at time T (DL-Group$_T$), only depends on the group drought level of the group at time T-1 (DL-Group$_{T-1}$).

4.5.3 Discrete Markov Model (DMM)

Markov Model are often referred to as discrete Markov Model. When a Markov Model is discrete, it usually has a finite number of outcomes that may occur with equal or different probabilities. The discrete-time Markov Model deals with a finite or countable number of states, S = {1, 2, 3, . . . , n}in which transitions can occur between them. Depending on the factors that are being dealt with, there may not be any transitions that occur between states or there may be numerous transitions. The total sum of the transitions leaving one state is always one in a discrete Markov Model. Transitions between states occur at random, which is how a Markov Model relates to stochastic processes.

4.5.4 Transition Probability

Transition probability is the probability of moving from one state to another in a finite number of transitions. The transition probabilities are denoted by the following:

$$P_{ij}^{m} = P\{X_{n+m} = j \mid X_0 = i\} \quad \text{.....(4.11)}$$

Where i is the state that is started in and after m number of transitions, the process ends in state j. When a state moves from one to another, its probability, denoted by P_{ij}, is the likelihood of state i moving to state j. These probabilities, P_{ij}, are called transition probabilities. Markov property assumes these transitions occurred completely at random. Previous transitions made are independent of future transitions and do not influence the outcome. These probabilities are independent of the states it was in before and just focus on the current position of the state.

The transition probabilities can be denoted with a transition matrix. This matrix captures the probabilities of transitioning from one state to another between time steps. A transition matrix is a square matrix used to describe the transitions of a Markov Model. The element in the i^{th} row and j^{th} column of the matrix represents the probability of transitioning from state i to state j in a time step.

		Future State				
		1	2	3	……	j
Current State	1	P_{11}	P_{12}	P_{13}	……	P_{1j}
	2	P_{21}	P_{22}	P_{23}	……	P_{2j}
	3	P_{31}	P_{32}	P_{33}	……	P_{3j}
	……	……	……	……	……	……
	i	P_{i1}	P_{i2}	P_{i3}	……	P_{ij}
	……	……	……	……	……	……

Figure 4.3 Transition matrix denoting the probabilities of each future state given the current state.

4.5.5 Predicting the Future Group Drought levels using Discrete Markov Model

When the sequence of the group drought levels (DL-Group) is known, the transition matrix can be calculated considering the next group drought level given the current group drought level for each time step T. By looking at the row of the current state in the matrix, the highest probable future state, which is the most likely next step can be identified. This method predicts the future by implementing a probabilistic estimate based on historical seasonal patterns. Even though this model gives simple drought forecast, it does not capture the dynamics of drought conditions. There are two main disadvantages of this model as follows.

1. By converting streamflow values directly into drought levels, a lot of features in the streamflow can be lost. As streamflow data provides a continuous measure of water flow, it contains rich information about the dynamics of the hydrological conditions. By converting these continuous values into categorical drought levels, much of this detailed variability is lost which might be important for accurate forecasting and understanding of drought dynamics.
2. As this model only provides the average drought level for the entire group, it fails to reflect the variability of drought levels of the months within the group. This approach can smooth

out important peaks and troughs in drought severity, potentially leading to less accurate forecasts and a poorer understanding of drought dynamics over time.

Using a Hidden Markov Model (HMM) can address the disadvantages of the simpler Discrete-Markov Model (DMM) model by incorporating both observed data (streamflow values) and hidden states (drought severity levels).

4.6 MODEL 2: Development of Hidden Markov Model (HMM) for Drought Forecasting

A Hidden Markov Model is a way to understand a continuous and non-stationary system that changes over time. The basis of the HMM model is that there is an underlying system that changes into different states, and it controls the observable output of the system. The hidden Markov model was developed by Baum in 1966. The HMM assumes these states change randomly, but only according to the Markov property. The states are directly not visible, but the output or the emissions of those states are visible, called observations. The HMM helps to understand the hidden underlying mechanism of the states by looking at the observations.

4.6.1 General Form of a Hidden Markov Model

There are 5 elements consisting of HMM and it can be denoted as (S, O, A, B, π).

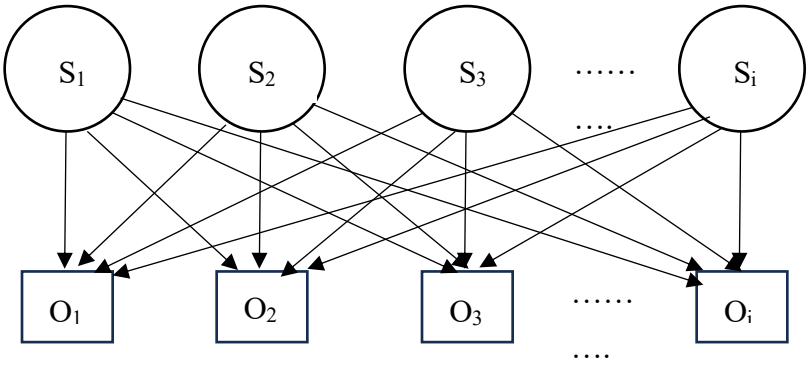


Figure 4.4 General form of a Hidden Markov Model.

The HMM procedure can be described as follows. Let us consider a hidden Markov model with that:

- There is a N discrete number of states, and the set of states can be denoted as $S=\{S_1, S_2,.., S_N\}$. These are the underlying states through which the system transitions. They are not directly observable but influence the observable outputs.
- Observations can have both discrete and continuous values. M possible observable symbols or outcomes can be denoted as $\mathbf{O}=\{O_1, O_2,..., O_M\}$, observations are distributed according to the hidden state distribution.
- Transition probabilities are the probability of a hidden state transitioning from one state to another state and are defined by the $N \times N$ transition matrix $\mathbf{A}= [a_{ij}]$, where a_{ij} represents the probability of transitioning from state s_i to state s_j.
- Emission probabilities, which are the probability of an observation being generated from a hidden state and defined by the $N \times M$ emission matrix $\mathbf{B} =[b_{ij}]$, where b_{ij} represents the probability of emitting symbol o_j from state s_i.
- Initial state probabilities are the probabilities of the system starting in each of the hidden states represented by the initial state distribution $\boldsymbol{\pi} =[\pi_1, \pi_2, ..., \pi_N]$, where π_i is the probability of starting in state s_i. There can be $\pi_j =0$, meaning that they cannot be initial hidden states.

Altogether, the model parameters of HMM can be denoted as $\boldsymbol{\lambda} = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$ as **A** given by the transition matrix, **B** is given by the emission matrix and $\boldsymbol{\pi}$ is the initial probability distribution. These parameters allow HMM to capture the system dynamics and predict the future of the system behaviour accordingly.

4.6.2 Fundamental Problems of HMM

The Hidden Markov Model should be characterized into three fundamental problems (Rabiner, 1989).

1. Likelihood problem: Given model parameters $\lambda = (A, B, \pi)$ and an observation sequence $O = \{O_1, O_2,..., O_M\}$, what is the probability of observing that sequence given the model? In other words, what is $P(O \mid \lambda)$?

2. Decoding problem (Finding the Best State Sequence): Given an HMM model (model parameters λ) and a sequence of observations O, what is the most likely sequence of hidden states that could have generated those observations?

3. Learning Problem (Parameter Estimation): Given a sequence of observations (O) and a set of possible HMMs, how the parameters can be adjusted of the model λ to maximize the probability of observing the sequence?"

To solve each of those questions, there are well-established algorithms which are the Forward-Backward Algorithm, Viterbi Algorithm and Baum-Welch Algorithm for each question respectively.

4.6.2.1 Problem 1: Likelihood problem (finding the probability of an observation):

This problem calculates how likely it is that the observed sequence could have been generated by the specific given HMM or P ($\mathbf{O} \mid \boldsymbol{\lambda}$). This step is crucial to evaluate a known HMM model or model parameters. AN efficient algorithm called the forward-backwards algorithm is used to solve this. There are two processes in this algorithm which are forward $\alpha_i(t)$ and backward algorithm $\beta_i(t)$.

4.6.2.1.1 Forward Algorithm $\alpha_i(t)$:

As the model parameters give the sequence of the states and how the system is changing, forward variable $\alpha_i(t)$ assesses the total probability of ending up in state S_i at time t, given the observation which can be denoted as:

$$\alpha_i(t) = P(O_1, O_2, .. , O_t ; S_t = i \mid \boldsymbol{\lambda}) \quad \text{.....(4.12)}$$

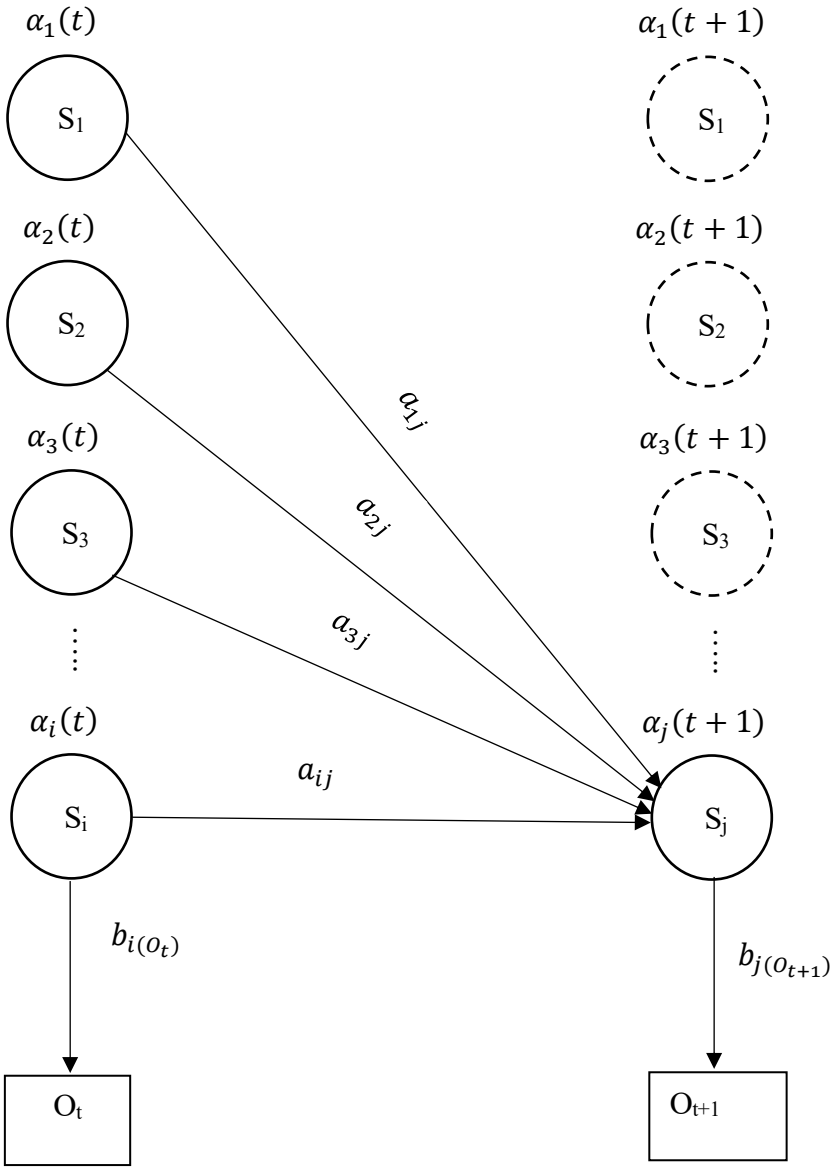


Figure 4.5 Visualizing the computation of a single element $\alpha_j(t+1)$ in the trellis by summing all the previous values $\alpha_i t$, weighted by their transition probabilities a_{ij} and multiplying by the observation probability $b_{i(ot)}$.

The term $\alpha_i(t)$ is the forward variable and represents the probability of being in state i after seeing the first t observations, given the model parameters **λ**. The value of each cell $\alpha_i(t)$ is computed by summing over the probabilities of every path that could lead to ith hidden state in the time step t.

The forward variable at each time t can be calculated inductively as the forward variable at time t+1 (α_j(t+1)) can be denoted as follows,

$$\alpha_j(\mathrm{t}+1) = \sum_{i=1}^{N} \alpha_i(t) a_{ij}\, b_{j(O_{t+1})} \quad \text{.....(4.13)}$$

Where the hidden state at time step t+1 is j and a_{ij} gives the probability of transitioning from i^{th} state to j^{th} state and $b_{j(O_{t+1})}$ gives the probability of emitting the given observation by the state j at the time step t+1. The forward algorithm can be solved as the forward variable at the initial state is given by the initial probability distribution π.

4.6.2.1.2 *Backward Algorithm $\beta_i(t)$:*

The backward procedure calculates the probability of the partial observation sequence from t + 1 to the end, given the model μ and state si at time t. The backward variable βi(t) is defined as:

$$\beta_i(t) = P(O_{t+1}, O_{t+2}, \ldots, O_T \mid S_t = i, \lambda) \quad \text{.....(4.14)}$$

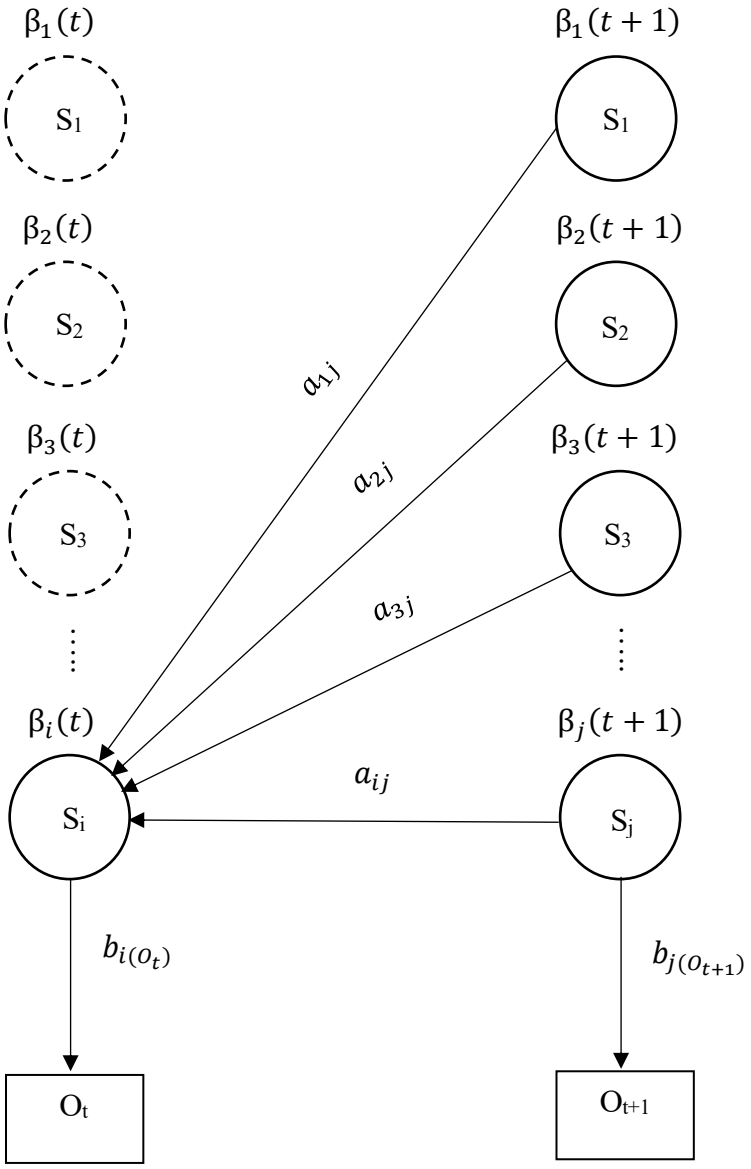


Figure 4.6 The computation of $\beta_i(t)$ by summing all the successive values $\beta_j(t+1)$ weighted by their transition probabilities a_{ij} and their observation probabilities $b_{j(ot+1)}$.

The β is the probability of seeing the observations from time t +1 to the end, given that the system is in state i at time t (and given the model parameters **λ**). The value of each cell $\beta_i(t)$ is computed by summing over the probabilities of every path that could lead to ith hidden state in the time step t.

induction computation for the forward procedure can also be performed in the reverse order.

$$\beta_i(t) = \sum_{j=1}^{N} \beta_j(t+1) a_{ij}\, b_{j(o_{t+1})} \quad \text{.....(4.15)}$$

Where the hidden state at time step t+1 is j and a_{ij} gives the probability of transitioning from i^{th} state to j^{th} state and $b_{j(o_{t+1})}$ gives the probability of emitting the given observation by the state j at the time step t+1. The backward algorithm can be solved as the backward variable at the end state is equal to one.

4.6.2.2 Problem 2: Decoding problem (Finding the best state sequence)

The second problem is to find the best state sequence given a model and the observation sequence or $P(\boldsymbol{S} \mid \mathbf{O}, \boldsymbol{\lambda})$. For this Viterbi algorithm is used to find the most probable path through the hidden states of the model that explains the observed data. The Viterbi algorithm is designed to evaluate recursively to find the most likely state sequence, $\mathbf{S} = \{S_1, S_2, .., S_N\}$ for a given observation sequence $\mathbf{O} = \{O_1, O_2, ..., O_M\}$ by selecting the most probable state sequence, which can be denoted as:

$$\arg\max P(\boldsymbol{S} \mid \mathbf{O}, \boldsymbol{\lambda}) \quad \text{.....(4.16)}$$

The Viterby probability value of each state (j)at time step t can be denoted as ($v_t(j)$) and it represents the probability that the HMM is in state j after seeing the first t observations and passing through the most probable state sequence $S_1,...,S_{t-1}$, given. Formally, each state expresses the probability:

$$v_t(j) = \max_{S_1,..,S_{t-1}} P(s_1, s_2, \ldots s_{t-1}; o_1, o_2, \ldots, o_t, s_t = j \mid \boldsymbol{\lambda}) \quad \text{.....(4.17)}$$

Here $\max_{S_1,..,S_{t-1}}$ signifies that the most probable path at t is represented by taking the maximum of overall possible previous state sequences max $s_1, s_2, \ldots s_{t-1}$. As with the other dynamic programming algorithms, Viterbi probability for each hidden state j and time t is filled recursively. It is calculated similarly to the forward-backwards algorithm, except that the forward-backwards algorithm uses summing over previous states while the Viterbi algorithm uses the argument of maximization. For a given state S_j at time t, the value $v_t(j)$ is computed as

$$v_t(j) = \max_{i=1:N} v_{t-1}(i)\, a_{ij} b_{jO_t} \quad \text{.....(4.18)}$$

The three factors that are multiplied in the above equation for extending the previous paths to compute the Viterbi probability at time t are $v_{t-1}(i)$ the previous Viterbi path probability from the

previous time step, a_{ij}; the transition probability from previous state qi to current state qj, b_{jO_t} the state observation likelihood of the observation symbol O_t given the current state j.

4.6.2.3 Problem 3: Learning problem (Parameter estimation)

The learning problem in Hidden Markov Models (HMMs) is all about training the model, which makes this problem more advanced than evaluation and decoding problems. Unlike the evaluation and decoding problems that focus on using a trained model or known model parameters, the learning problem deals with setting the model parameters of the HMM $\boldsymbol{\lambda} = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$ based on a set of observed sequences. The goal of this question is to find the set of parameters $\boldsymbol{\lambda} = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$ that maximizes the probability of the observed sequence ($\mathbf{O}$) given the model ($P(\mathbf{O} \mid \boldsymbol{\lambda})$). This is challenging as the hidden states are not directly observable. As a solution for this issue, the Baum-Welch Algorithm is used because it uses an iterative technique to solve the learning problem in HMMs. It works in a step-by-step manner to improve the model parameters based on the observed data.

Baum-Welch uses the Expectation Maximization method to find the optimal model parameters. It works iteratively to improve the likelihood of $P(\mathbf{O}|\boldsymbol{\lambda})$. It uses the same forward-backward algorithm but with the Expectation Maximization (E & M) method. The initial transition matrix and the emission matrix are assumed randomly and use Expectation Maximization for the convergence.

4.6.2.3.1 Expectation Step

Taking the model parameters $\boldsymbol{\lambda} = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$ the probability of being in state j at time t, which is denoted by $\gamma_t(i)$ as follows.

$$\gamma_t(i) = P(S_t = j \mid O, \boldsymbol{\lambda}) \quad \text{.....(4.19)}$$

This means that:

$$\gamma_t(i) = \frac{\alpha_i(t)\ \beta_i(t)}{P(\boldsymbol{O} \mid \boldsymbol{\lambda})} \quad \text{.....(4.20)}$$

The probability ξ_t as the probability of being in state i at time t and state j at time t+1, given the observation sequence and of course the model:

$$\varepsilon_t(i,j) = P(S_t = i, S_{t+1} = j \mid \mathbf{O}, \boldsymbol{\lambda}) \quad \text{.....(4.21)}$$

This can be simplified using the forward and backward variables as follows.

$$\varepsilon_t(i,j) = \frac{\alpha_i(t) a_{ij} b_{j(O_{t+1})} \beta_j(t+1)}{\sum_{j=1}^{N} \alpha_j(t)\beta_j(t)} \quad \text{.....(4.22)}$$

Given the above definitions, an initial model **λ** is ready and it can be used to run the training data **O** through the current model to estimate the expectations of each model parameter. Then the model is changed to maximize the values of the paths that are used as follows.

4.6.2.3.2 Maximization Step

Here the calculated $\gamma_t(i)$ and $\varepsilon_t(i,j)$ is used to recompute new A and B probabilities.

$$a_{ij}{}' = \frac{\text{expected number of transitions from state i to j}}{\text{expected number of transitions from state i}}$$

$$a_{ij}{}' = \frac{\sum_{t=1}^{T-1} \varepsilon_t(i,j)}{\sum_{t=1}^{T-1} \sum_{k=1}^{N} \varepsilon_t(i,k)} \quad \text{.....(4.23)}$$

$$b_j(\mathrm{v}_k)' = \frac{\text{expected number of times in state j and observing symbol } \mathrm{v}_k}{\text{expected number of times in state j}}$$

$$a_{ij}{}' = \frac{\sum_{t=1\, s.t\, O_t = v_k}^{T} \gamma_t(i)}{\sum_{t=1}^{T} \gamma_t(i)} \quad \text{.....(4.24)}$$

The Expectation-Step and Maximization-Step are repeated until a stopping criterion is met.

4.6.3 Development of the Hidden Markov Model for Drought Forecasting

Observations of the HMM for drought forecasting are seasonal patterns or groups, making one observation L-dimensional. As there are 5 levels of drought in SDI values according to Table 4.1, the number of hidden states of the HMM is taken as five. Since the drought data or the observations are standardised and normally distributed values (SDI values), observations are Gaussian

distributed. Using the Baum Welch algorithm, the optimal model parameters were calculated and with the trained model (optimal parameters), the future group drought data is predicted. A trained model can identify the set of observations or groups governed by each state. Then for each hidden state, the model estimates the mean (μ) and the covariance (Σ) of the dependent observation. The SDI values (O_{T+1}) of the future groups are calculated after finding (O_{T-j}) of the nearest past group. The nearest past group is selected according to two conditions: the first condition is that it should be generated by the same hidden state as (O_T) of the current hidden state of the group and the second condition is it should have the nearest values to the current group. Once the earliest group (O_{T-j}) is identified, the differential changes in the SDI values from this past group to the subsequent group (O_{T-j+1}) were calculated. Then the calculated differential change was added to the observations of the current group(O_T) SDI to predict the value of the future group as denoted in the equation below.

$$O_{T+1} = O_T + (O_{T-j+1} - O_{T-j}) \quad \text{.....(4.25)}$$

Even though HMM can identify the hidden state distribution and how the observations are changing accordingly, it primarily relies on the immediate previous state for predicting future states and observations. This makes them less effective in capturing long-term dependencies and trends in the data, which are crucial for accurate drought forecast. To avoid this, the Long Short-Term Memory (LSTM) Model was used to maintain information over long periods, allowing them to learn dependencies from distant past data.

4.7 MODEL 3: Development of the Long Short-Term Memory Model (LSTM) for Drought Forecast

Long Short-Term Memory (LSTM) are Recurrent Neural Networks type designed to handle sequential data with short- and long-range dependencies more effectively than standard RNNs.

4.7.1 Artificial Neural Networks Model

An Artificial Neural Networks Model, as introduced in section 2.2.6.1 of chapter 2, consists of three main layers, which are the input, hidden and output layers. Those layers are interconnected

by artificial neurons, which are also called nodes or units. The most common type of ANN is the feedforward neural network (Figure 4.7) and information flows in one direction, from the input layer through one or more hidden layers to the output layer. Each neuron receives input signals, performs a computation, and produces an output signal that is transmitted to the neurons in the next layer. The connections between neurons are represented by weights, which determine the strength of the signal transmitted from one neuron to another. During the training process, the weights are adjusted based on the input data and the desired output, allowing the ANN to learn and adapt to the underlying patterns in the data.

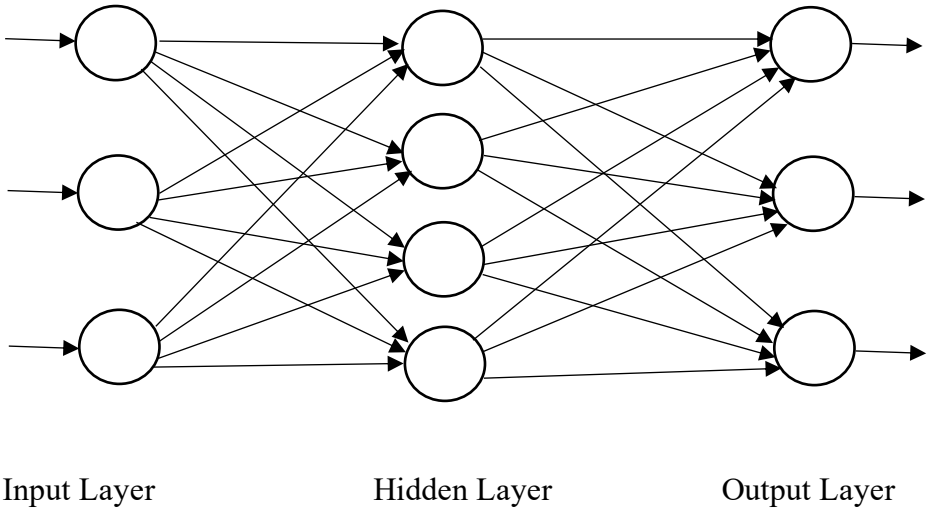


Figure 4.7 A general structure of a feed-forward Artificial Neural Network including the input layer, the hidden layer and the output layer.

4.7.2 Recurrent Neural Networks

Recurrent Neural Networks mimics the way the human brain works. The thinking process of the human brain does not reset every moment, and it is connected to each moment. As that, Recurrent Neural Networks have loops that allow them to remember information from previous steps. This is the difference between Recurrent Neural Networks and traditional Neural Networks. Traditional neural networks cannot use the context from earlier sequences to help understand later sequences.

The loop in the Recurrent Neural Networks makes it much more complex than a traditional Neural Network, but they are not that different from regular neural networks connected as shown in Figure 4.8.

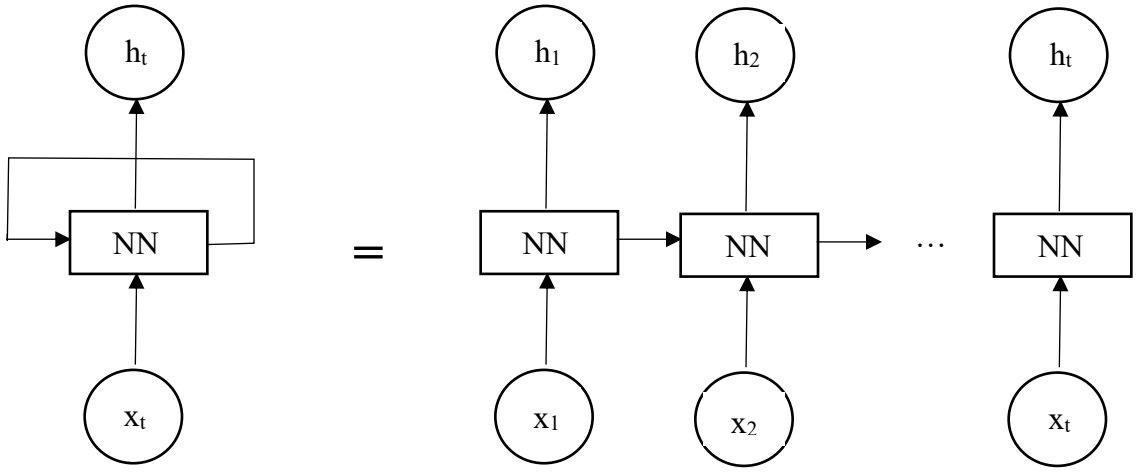


Figure 4.8 An extracted Recurrent Neural Networks.

The seasonal pattern (x_t) of (Streamflow Drought Index) SDI values and NN is a Neural Network h_t is the hidden state of the current time step. This chain-like structure shows that Recurrent Neural Networks (RNNs) are closely related to sequence, making RNN the ideal neural network architecture for handling sequential types of data, where the order of data points is crucial. Even though theoretically RNN can handle both short-term and long-term sequential data effectively, practically RNN does not have the ability to learn very long-term dependencies. Long- Short-Term Memory (LSTM) models are specially designed RNNs to handle both short-term and long-term dependencies (Hochreiter and Schmidhuber, 1997).

4.7.3 LSTM Network Architecture

The LSTMs have the same chain structure as RNN, but the repeating module is more complex. LSTM have a basic unit named hidden unit that repeats within a hidden layer as shown in Figure 4.9. Instead of having a single neural network layer, each hidden unit contains four neural network layers that interact uniquely. These layers work together to manage and control the flow of information, allowing the network to remember and use important details over long sequences of data.

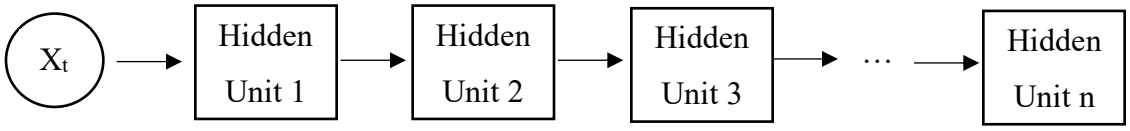


Figure 4.9 One hidden layer in LSTM.

A hidden unit of an LSTM consists of four interacting neural network layers (three Sigma neural networks and one tanh neural network) as shown in Figure 4.10. In this figure, the circles represent pointwise operations, such as vector addition and multiplication, while the boxes represent learned neural network layers. Cell state (C_t) and the hidden state (h_t) are the memory storages of the LSTM. Cell state is the long-term memory storage, and it keeps the information across multiple time steps while hidden state (or the short-term memory storage) capture the information of the current time step and carry forward the relevant information to the next step.

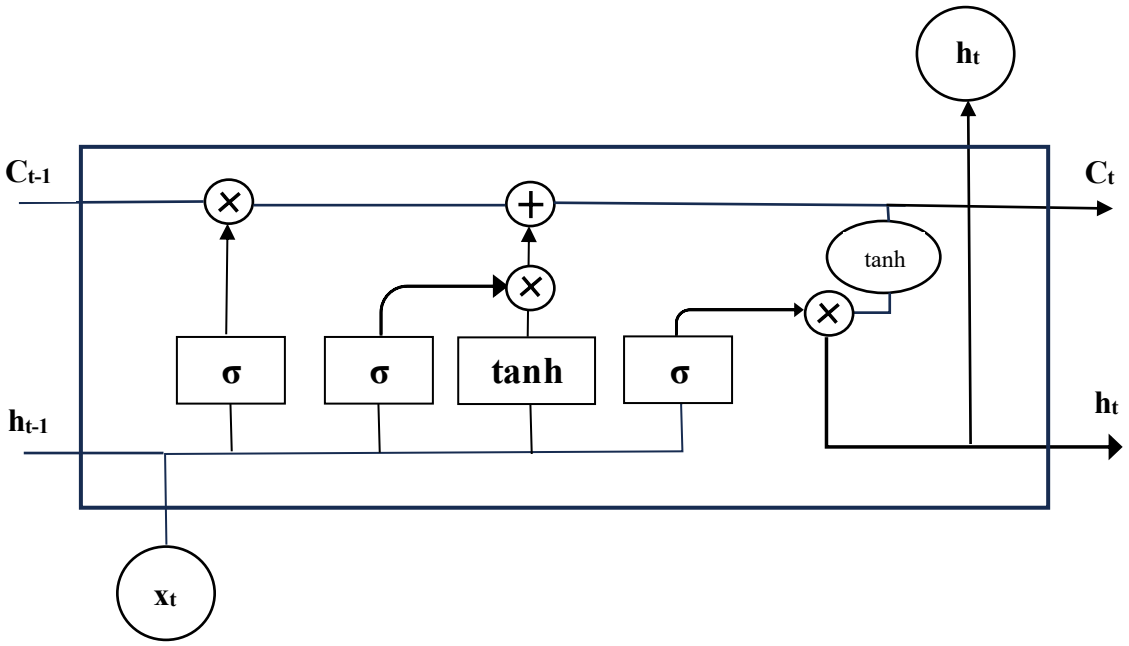


Figure 4.10 Illustration of the structure of a hidden unit.

The selection of the relevant information within the data is carefully operated by a structure called a “gate”. Gates are mechanisms that control the flow of information and decide what should be

removed or considered. There are three gates in an LSTM and each gate consists of a sigmoid neural network layer (σ) and a pointwise multiplication operation.

$$\sigma = \frac{1}{1+e^{-x}} \quad \text{.....(4.26)}$$

The sigmoid layer outputs values between 0 and 1, determining how much information should be passed through. This allows the LSTM to selectively keep or discard information as needed.

The other neural network layer in the gates is the tanh neural network layer (tanh) and it can be defined as follows.

$$\tanh = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \quad \text{.....(4.27)}$$

Unlike the sigmoid function, which outputs values between 0 and 1, the tanh function outputs values between -1 and 1. This can compress the linear inputs into nonlinear outputs which range between -1 and 1. During training, the LSTM model goes through several iterations, adjusting its internal parameters (weights and biases) in sigmoid and tanh functions to minimize the error between its predictions and the actual data. Both sigmoid and tanh neural networks are essential for gates to learn complex patterns within the data to predict the streamflow drought patterns.

4.7.4 Forget-gate, Input-gate, and Output-gate.

In LSTM, the cell state and hidden state are regulated by three gates; the Forget Gate, the Input Gate, and the Output Gate and these gates manipulate the information to determine what should be deleted, added and carried forward. Figure 4.11 illustrates how the information is processed through gates.

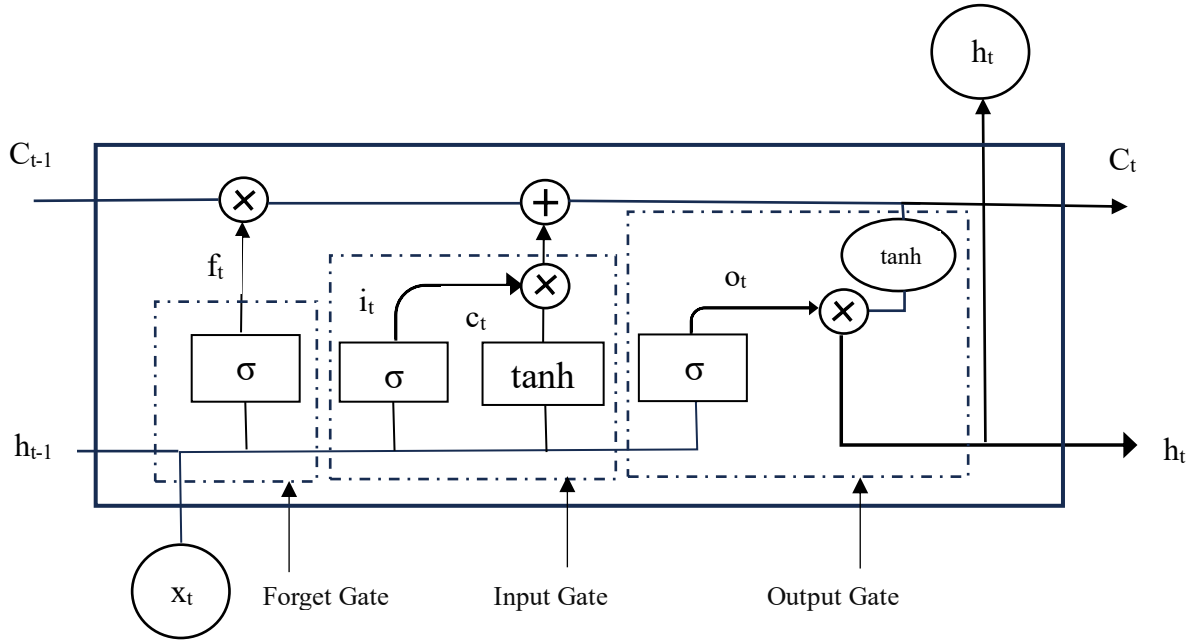


Figure 4.11 Three gates in a hidden unit to manipulate information.

The first gate in the LSTM is the forgot gate which is a sigmoid layer. It decides what information to throw away from the cell state. It considers both the previous hidden state h_{t-1} and the current input x_t, and outputs a number between 0 and 1 for each number in the cell state C_{t-1}.

$$f_t = \sigma(W_f.[h_{t-1}, x_t] + b_f) \quad \text{.....(4.28)}$$

Where, W_f and $b_f$are weight matrix and bias gate for the forget gate.

The next stage is the Input gate where the new information that will be stored in the cell state is decided. The input gate has two parts. First, a sigmoid layer called the "input gate layer" selects the values to update. Next, a tanh layer creates a vector of new candidate values which is denoted by c_t.

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i) \quad \text{.....(4.29)}$$

$$c_t = tanh(W_c.[h_{t-1}, x_t] + b_c) \quad \text{.....(4.30)}$$

Where W_i , b_i and W_i , b_i are weight matrix and bias gate for the input gate and candidate list respectively. After the forget gate and input gate, the cell state is updated with the relevant information after discarding outdated or irrelevant information. The new cell state C_t can be given as:

$$C_t = f_t * C_{t-1} + i_t * c_t \quad \text{.....(4.31)}$$

Finally, the output layer filters out the data to determine which parts of the cell state will be used to compute the hidden state. Then the cell state is put through the tanh operator to transform it to values between -1 and 1.

$$o_t = \sigma(W_o.[h_{t-1}, x_t] + b_o) \quad \text{.....(4.32)}$$

$$h_t = o_t * \tanh(C_t) \quad \text{.....(4.33)}$$

Where, W_o and b_o are weight matrix and bias gate for the output gate.

This process ensures that the LSTM outputs a relevant and filtered version of the cell state, tailored to the current context and ready for the next step in the sequence.

4.7.5 LSTM for Drought Forecast

The parameters of an LSTM model, such as in the forget, input, and output gates, relate to the physical properties of drought patterns and their forecast in several ways. Firstly, the forget gate decides which information from the previous cell state should be discarded or retained. This is crucial for managing long-term dependencies in drought data as drought conditions often persist over long periods. The forget gate helps the LSTM retain information about past conditions that are still relevant, such as prolonged periods of low streamflow or sustained high streamflow, which have long term dependencies within the data set. Then the input gate determines which new information should be added to the cell state (long term memory) and it controls the extent to which new observations affect the hidden state (short term memory) of the network. It allows the model to understand the short-term influences such as seasonal variations and sudden anomalies such as temporary reduction in streamflow patterns, which could impact drought conditions. Both the forget gate and input gate help to understand the dynamic changes within the streamflow patterns while the output gate decides what information from the forget gate and input gate should

be used to predict the next hidden state which gives the next streamflow pattern. The output gate helps in interpreting the future drought states based on the past information. It translates the long term and short-term memory of the model (both cell state and hidden state) into actionable forecasts about future drought conditions, considering both long-term trends and recent changes. All weights are adjusted iteratively to get the best forecast results from those three gates during the training phase

LSTM model trained for seasonal patterns which include grouped SDI values. For this model, firstly the monthly streamflow data was used by grouping it to a length of L months. Secondly, the model is used for weekly data instead of monthly data. This choice is made to evaluate the model behaviour for different data set sizes. Therefore, the length of the group containing weekly data is 4.33L (= 13 weeks) instead of L. It is concluded that the availability of a larger dataset is important for training the LSTM model effectively, as LSTMs typically require substantial amounts of data to capture temporal dependencies accurately. A study conducted in 2020 indicated that the optimal number of hidden layers for streamflow forecasting ranges between 2 to 4 (Apaydin et al., 2020). Given that this thesis focuses on more complex drought data, characterized by significant variations across seasons, four hidden layers were chosen for the LSTM model. This selection ensures that the model can effectively capture and process the intricate patterns present in the data for improved forecasting accuracy. The LSTM is designed with 4 hidden layers each containing 128 hidden units as streamflow data can contain complex dependencies. This number of hidden units and layers is structured to effectively capture the temporal dynamics of the drought index.

4.8 Model Performance Metrics

To evaluate the performances of each model for predicting the streamflow drought index of the future group, three main metrics including Mean Squared Error (MSE), R-squared (R^2) and mismatch method have been used.

The Nash-Sutcliffe efficiency is a widely used metric in hydrological modeling to assess the performance of a model relative to observed data (Mohammadi, 2021). However, there has been increasing interest in exploring alternative metrics, such as mean squared error and R-squared (R^2), to evaluate the performance of machine learning models in the field of hydrology. R^2 and Mean

squared error can give better insight into the predictive capabilities of the model compared to the Nash-Sutcliffe efficiency (Plevris et al., (2022)), as they provide a more comprehensive understanding of correlation of forecasted values with observation values and how the error distributed through out the forecast values. Since this research based on statistical and machine learning approaches of drought forecasting, R^2 and Mean squared error were considered as the model performance metrices.

4.8.1 Mean Squared Error (MSE)

MSE measures the average squared difference between the predicted values and the actual values. Lower MSE values indicate better model performance.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_{Actual,i} - y_{Predicted,i})^2 \quad \text{.....(4.34)}$$

Where N is the number of samples, $y_{Actual,i}$ and $y_{Predicted,i}$ are the actual and the predicted SDI values for i^{th} sample respectively. If MSE is close to 0, it indicates that the model is accurately predicting the target variable.

4.8.2 R-squared / R^2

R-squared measures the proportion of the variance in the target variable that is predictable from the independent variables. Higher R^2 values indicate better model performance, with a maximum value of 1.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_{Actual,i} - y_{Predicted,i})^2}{\sum_{i=1}^{N}(y_{Actual,i} - \bar{y})^2} \quad \text{.....(4.35)}$$

4.8.3 Mismatch Accuracy

This approach compares the predicted drought level groups with the actual drought level groups. The actual and predicted SDI values of a group were converted into their respective drought level to find the drought level of the group. For each group, the predicted drought level to the actual drought level is compared. If they match, a value of 1 is assigned; if they do not match, a value of

0 is assigned. The percentage accuracy of correctly predicting the drought level of the group is calculated below,

$$\text{Accuracy} = (\text{Number of correct forecasts/Total number of forecasts}) \times 100\% \quad .....(4.36)$$

This method provides a straightforward way to measure how well the model predicts the drought level of the group based on the streamflow drought index (SDI).

4.9 Concluding Remarks

In this chapter, the initial step involved data preprocessing to thoroughly understand the behaviour and characteristics of the dataset. Subsequently, patterns and groups within the data are identified through various techniques, including entropy calculations, moving average de-seasonalization, and k-means clustering. These identified patterns were then utilized to evaluate and compare the performance of three different forecast models, providing a comprehensive analysis of their effectiveness in capturing the underlying trends within the data. The forecast of streamflow drought patterns is critical for effective water resource management and mitigating the impacts of drought. In this research, three different models which are, Discrete Markov Model, Hidden Markov Model, and Long Short-Term Memory (LSTM) networks were utilized to predict drought levels based on the seasonal pattern of Streamflow Drought Index (SDI). Each model offers unique advantages in capturing the complex temporal dynamics of drought.

The Discrete Markov model provides a straightforward probabilistic framework to predict future drought states based on the current state, making it useful for short-term drought forecasting. The Hidden Markov Model enhances this approach by considering the same hidden states as a discrete Markov model, but it can capture more features to see how the streamflow drought patterns behave dynamically, offering a more accurate forecast. The LSTM model has the ability to learn features in both long-term and short-term dependencies in streamflow data, thereby providing highly accurate forecasts.

The next chapter (Chapter V) will describe the selection of study areas and watersheds for applying the models. This chapter will detail the process of choosing these regions and how they will be used to recognize and predict drought patterns using the developed models.

Chapter V: Study Area and Data Assembly

This chapter will provide a detailed description of the study area used in the evaluation of each model for forecasting droughts. The study area was selected after a detailed examination of its suitability for evaluating the ability of each model to forecast droughts. This examination included a thorough analysis of the hydrological characteristics of the area using flow values of the rivers to understand the factors influencing regional drought occurrences. The data selection process involved collecting and analyzing different monthly time series datasets from various hydrometric stations distributed across the area to identify the monthly streamflow data sets that are suitable for model application qualitatively and quantitatively.

5.1 Selection of the Study Area

The study area was specifically selected from semi-arid regions in Canada, where drought conditions are both unique and frequent. The driest part of Canada is known as the Palliser Triangle, a vast semi-arid region located in the southern parts of Alberta and Saskatchewan, and a small section of southwestern Manitoba as shown in Figure 5.1. The Palliser Triangle is notable for its distinctive environmental challenges, especially the recurrent droughts that have historically affected the region.

The origins of the Palliser Triangle can be traced back to the mid-19th century, when Captain John Palliser, a British explorer, conducted extensive fieldwork in the region between 1857 and 1860 (Villmow, 1956). Palliser observations and reports highlighted the distinctively dry character of the area, which was at odds with the prevailing perceptions of the Canadian Prairies as a prosperous agricultural heartland (Villmow, 1956). He recognized a roughly triangular area that he felt to be poorly suited for farming and the area was named Palliser Triangle. Despite these observations, Palliser-triangle and the surrounding prairie area have become large farmland with leading crops of canola, spring wheat and lentils today. However, a primary issue in this region is the persistent occurrence of drought, which significantly impacts agricultural productivity and livelihoods of the area. Droughts have been a recurring feature in the Palliser Triangle, with some of the most severe

events occurring in the late 19th and early 20th centuries (Rannie, 2006). This makes an accurate drought forecast crucial for managing the risks and impacts of drought in regions.

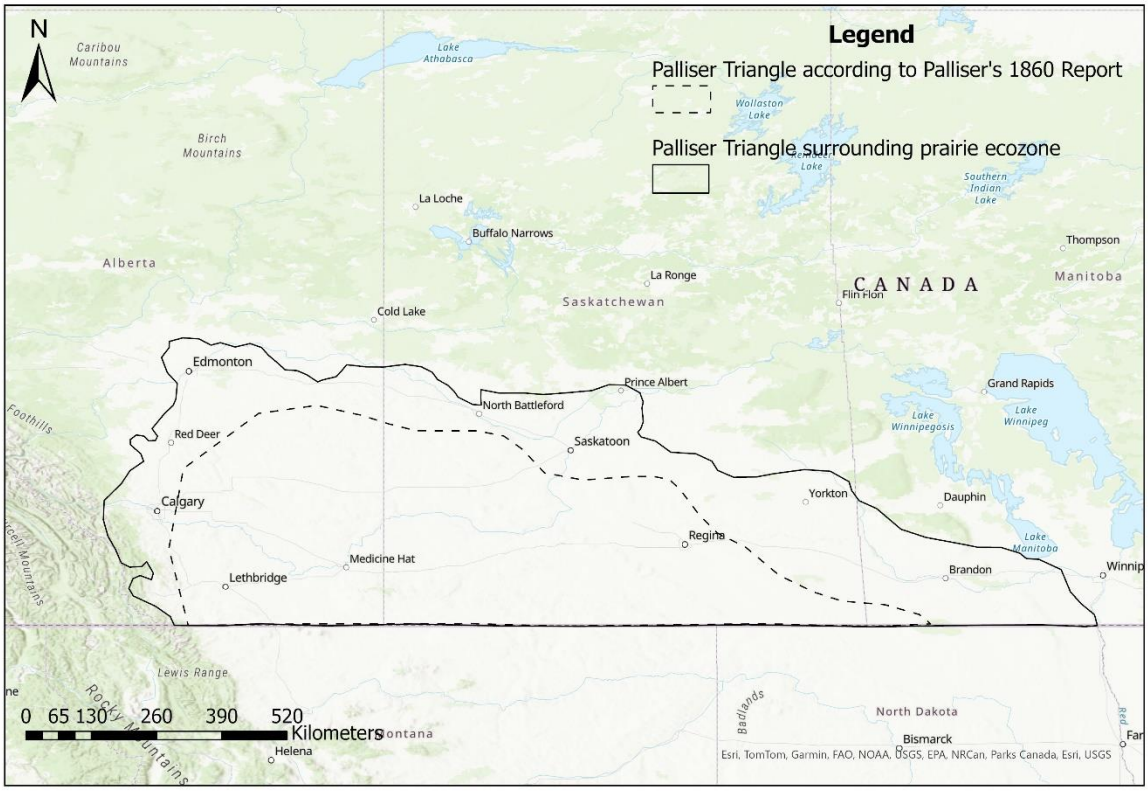


Figure 5.1Palliser triangle and surrounding prairie area.

5.2 Droughts in the Palliser Triangle

There are frequent droughts in the Palliser Triangle and due to climate change in recent years, they have become more frequent and more severe (Safarianzengir et al., 2022). Palliser triangle is a main agricultural region producing major crops such as wheat and canola that are heavily impacted by droughts.

The significant drought features in this area include the high precipitation variability in different seasons and high evapotranspiration rate due to the high temperature. Even though there are several rivers and lakes in this area, they are vulnerable to dwindling water levels due to reduced precipitation and increased evapotranspiration (Safarianzengir et al., 2022). This leads to the drying of lakes and creeks, groundwater depletion, and salinization of arable lands, with far-reaching consequences for agriculture, ecosystems, and local communities within the area

(Safarianzengir et al., 2022). Therefore, it is very important to analyse this area in order to develop effective drought monitoring techniques.

5.2.1 The Palliser Triangle

Palliser Triangle is located between coordinates of 49°N to 51°N and 100°W to 114°W, but this boundary can fluctuate based on the climatic cycle (Burnett, 2007). Climatological features of this area include an average of approximately 352 mm of annual precipitation and an annual snowfall of 101 cm (Burnett, 2007). Main climatological seasons are characterized by long, dry summers that approximately last from April to October and relatively short, cold winters from November to March. (Last, 1994). The temperature varies from -5°C to -15°C in the winter season and the summer season temperature often exceeds 30 °C. Due to the high temperature and the low humidity, the evapotranspiration rate is very high which makes the annual water deficit, or the amount of the potential evapotranspiration exceeds the precipitation by 524 mm (Villmow, 1956; Burnett, 2007).

The soil types of this region are primarily brown chernozemic and dark brown chernozemic which is representative of the semi-arid prairie environment (Villmow, 1956). The grassland of the Palliser Triangle consists of native species adapted to dry conditions, such as blue grama, needle-and-thread, and western wheat grass. The basic ecosystem of the area is characterized by a lack of surface water features, with few permanent streams and many playa lakes or basins that only hold water seasonally (Last, 1994). Those lakes have high variability in water levels due to the high evapotranspiration rate.

Despite the low precipitation in the Palliser Triangle, several significant rivers flow through this region, playing crucial roles in its hydrology and ecosystems. The South Saskatchewan River, originating in the Canadian Rockies, flows through Alberta and Saskatchewan before joining with the Bow River near Medicine Hat, Alberta. Another important river, the Red Deer River, meanders through central Alberta, eventually joining the South Saskatchewan River near the Saskatchewan border. In the southwestern reaches, the Oldman River winds through Alberta, adding to the flow of the South Saskatchewan River. Further east, the Assiniboine River crosses southern Saskatchewan and Manitoba, eventually merging with the Red River. The Bow River itself begins

in the Rocky Mountains and flows through Calgary before merging with the South Saskatchewan River. Other than these rivers, the Palliser Triangle region contains several long, continuous and evenly distributed streams with flow records that are suitable for reliable use in early warning systems for drought monitoring.

5.3 Selection of Streamflow Stations

The streamflow data for the Palliser Triangle area were extracted from HYDAT (Hydrologic Data), which is the streamflow data resource managed by the Water Survey of Canada. Over 100 streamflow stations were assessed for completeness and statistical validity. To ensure data completeness, monthly data gaps were first cross-referenced with daily data for infilling purposes. For gaps of fewer than 10 days, linear interpolation was employed, while for gaps exceeding 10 days but less than a year, the Analogue River Ratio method (Tallaksen, 2004) was used. Furthermore, the general rule for a sample size is to be greater than 30 years to yield stable estimates of statistics, viz., mean, variance, skewness and autocorrelations (Brase et al., 2006) Therefore, the streamflow stations were selected based on two criteria. The first criterion is that the station has no missing data or missing data only up to one year and the second criterion is that there are more than 30 years of available data, to obtain estimates of the aforementioned statistics.

5.3.1 Filling the Missing Data using the Analogue River Ratio Method

The Analogue River Ratio Method, as described by Tallaksen et al. (2004), was used to fill in missing data segments for up to a year. This method involves using the ratio between the target station and a nearby gauging station (analogue) to manually estimate missing flows at the target station. The analogue station was chosen as either an upstream or downstream station of the target river or a station in the nearest catchment. The filled missing data were then checked to ensure they fell within the 95% confidence limit of the mean streamflow values from the last 5 years and the next 5 years of the target river. If the filled data did not lie within the 95% confidence boundary, another station was selected according to the Analogue River Ratio Method. This process was repeated until the filled data lay within the 95% confidence boundary. This process is described in the Figure 5.2 below.

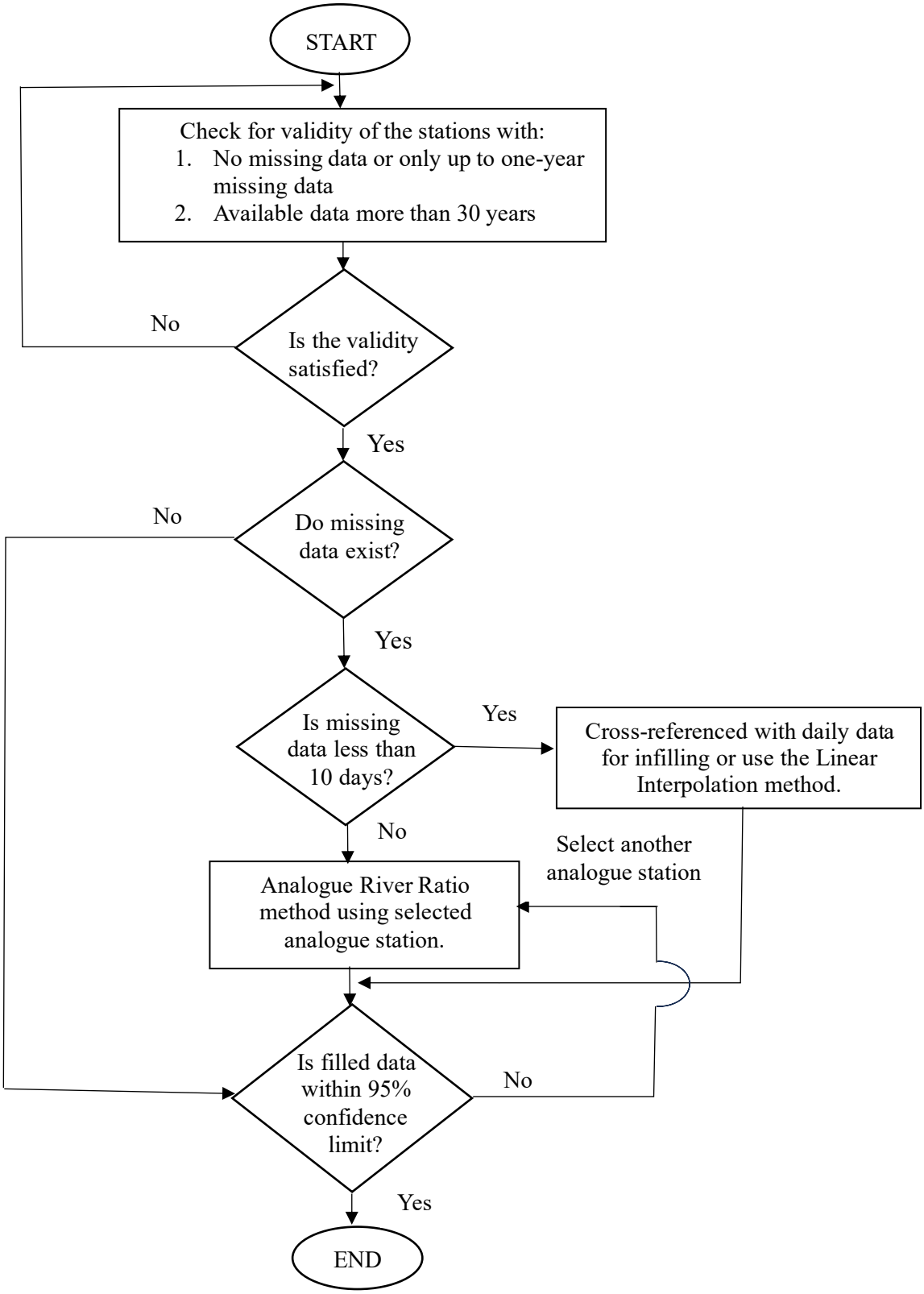


Figure 5.2 Flow chart for selection of streamflow stations.

5.3.2 Selection of Streamflow Datasets

After careful consideration, only 25 streamflow stations were deemed suitable for further analysis and model application. Appendix- A provides additional details on the Analogue River Ratio Method for infilling missing records within the 95% confidence limit. The relevant characteristics of all such stations are summarized in Table 5.1 below. A common period of 48 years (October 1, 1973, to September 30, 2022) was selected for all these stations. Catchments sizes of the selected stations range from 319 to 131,000 km^2.

A study conducted in 2020 introduced three clusters of rivers in the Canadian prairies based on the cumulative annual runoff of each river (Whitfield et al., 2020). The cumulative annual runoff was calculated as the flow per basin unit area. The first and most common type is low-flow capacity rivers, characterized by infrequent flow days and low total annual runoff, ranging from 0 to 50 mm. The second type is medium-flow capacity rivers, which have more days with flow and a slightly greater annual runoff, ranging from 48 to 175 mm. The least common third type consists of high-flow capacity rivers, which have the fewest days without flow, include perennial streams, and exhibit much greater annual runoff, exceeding 173 mm.

Low-flow capacity rivers are very common in the Canadian prairies and tend to dry out during the dry season, making it difficult to find small rivers with continuous flow throughout the year. Medium-flow capacity rivers are more prone to droughts than high rivers. Therefore, as shown in Table 5.1, three low-flow rivers, sixteen medium-flow capacity rivers, and six high-flow rivers within the Palliser Triangle were selected for analysis. This selection ensures a representative sample of rivers with varying susceptibility to drought conditions. Figure 5 shows the locations of the small, medium and large flow rivers.

Table 5.1 Basic hydrological characteristics of selected hydrometric stations within the study area.

No	Streamflow station name	Selected Period	Prov	Station ID	Latitude (N)	Longitude (W)	Area (km^2)
1	Pembina River at Neche^{+}	1974-2022	MB	05OC004	48°59'22"	97°33'24"	8,480
2	Souris River near Sherwood^{+}	1974-2022	SK	05ND007	48°59'24"	101°57'28"	23,100
3	East Poplar River at International Boundary^{+}	1976-2022	SK	11AE003	48°59'58"	105°24'32"	1,400
4	Little Saskatchewan River near Minnedosa^{++}	1974-2022	MB	05MF001	50°21'31"	99°54'27"	2,610
5	Whitemud River at Westbourne^{++}	1974-2022	MB	05LL002	50°08'05"	98°35'02"	6,360
6	Moose Jaw River near Burdick^{++}	1974-2022	SK	05JE006	50°24'01"	105°23'52"	9,230
7	North Saskatchewan River at Prince Albert^{++}	1974-2022	SK	05GG001	53°12'12"	105°46'21"	131,000
8	North Saskatchewan River near Deer Creek^{++}	1974-2022	SK	05EF001	53°31'23"	109°37'04"	57,200
9	Qu'appelle River near Lumsden^{++}	1974-2022	SK	05JF001	50°39'01"	104°52'00"	17,500
10	South Saskatchewan River at Saskatoon^{++}	1974-2021	SK	05HG001	52°08'26"	106°38'39"	141,000
11	Swift Current Creek below Rock Creek^{++}	1974-2022	SK	05HD036	49°50'40"	108°28'46"	1,430
12	Assiniboine River at Kamsack^{++}	1974-2022	SK	05MD004	51°33'53"	101°54'59"	13,000
13	Blindman River near Blackfalds^{++}	1974-2022	AB	05CC001	52°21'25"	113°47'42"	1,800
14	Battle River near Ponoka^{++}	1974-2022	AB	05FA001	52°39'47"	113°34'53"	1,820
15	Red Deer River at Drumheller^{++}	1974-2022	AB	05CE001	51°28'02"	112°42'41"	24,900
16	Medicine River near Eckville^{++}	1975-2022	AB	05CC007	52°19'10"	114°20'39"	1,920
17	Milk River at Milk River^{++}	1974-2022	AB	11AA005	49°08'36"	112°04'54"	2,720
18	Red Deer River near Bindloss^{++}	1974-2022	AB	05CK004	50°54'09"	110°17'58"	47,800
19	South Saskatchewan River at Medicine Hat^{++}	1974-2022	AB	05AJ001	50°02'31"	110°40'39"	56,400
20	Crowsnest River at Frank^{+++}	1974-2022	AB	05AA008	49°35'50"	114°24'38"	403
21	Bow River at Calgary^{+++}	1974-2022	AB	05BH004	51°03'00"	114°03'05"	7,870
22	Oldman River near Lethbridge^{+++}	1974-2022	AB	05AD007	49°42'33"	112°51'46"	17,000
23	Highwood River near the Mouth^{+++}	1974-2022	AB	05BL024	50°46'57"	113°50'38"	3,950
24	Belly River near Mountain View^{+++}	1974-2022	AB	05AD005	49°05'58"	113°41'51"	319
25	St. Mary River at International Boundary^{+++}	1974-2022	AB	05AE027	49°00'43"	113°17'58"	1,210

NOTE: + indicates low flow; ++ medium flow, and +++ high flow rivers. “Prov” indicate the province of the streamflow station.

The following Figure 5.3 shows the location of each station according to the station number listed in Table 5.1.

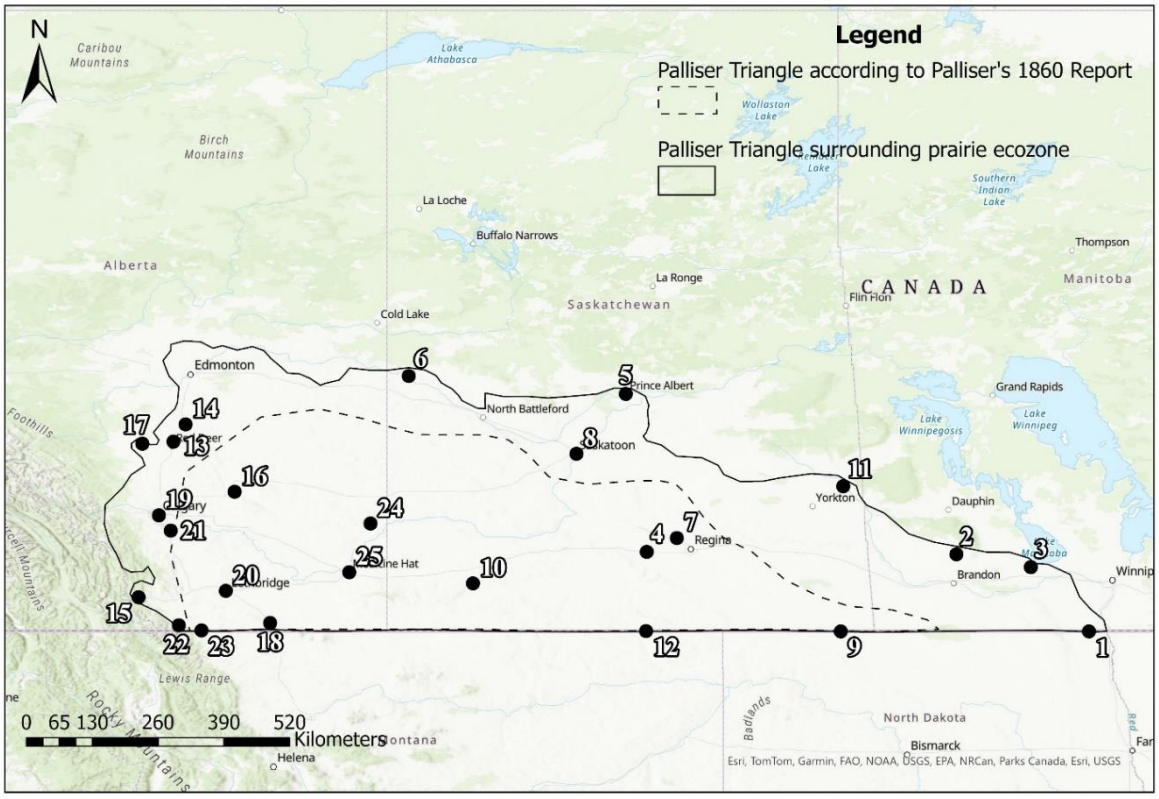


Figure 5.3 Topographic map of Palliser triangle illustrating the locations of hydrometric stations selected for model application.

The following Table (Table 5.2) provides the statistical characteristics of the selected streamflow stations, including the optimal lambda(λ)value for the Box-Cox transformation of each station, as well as the mean, standard deviation, and skewness of both monthly and weekly raw streamflow data distribution.

Table 5.2 Statistical characteristics of monthly and weekly flow sequences of streams under study

*No	Monthly Data					Weekly Data				
	Lambda	Mean (m^3s^{-1})	C_V	Cs	ρ_1	Lambda	Mean (m^3s^{-1})	C_V	Cs	ρ_1
1	0.11	11.30	2.20	3.94	0.49	0.12	11.18	2.59	5.55	0.84
2	0.08	5.11	3.84	9.21	0.61	0.04	5.12	4.34	11.99	0.86
3	-0.44	0.24	3.54	8.39	0.20	-0.07	0.27	5.82	18.02	0.30
4	-0.05	4.91	1.70	3.75	0.54	-0.04	4.94	1.88	4.30	0.86
5	-0.13	7.28	1.94	3.68	0.55	-0.04	4.92	1.87	4.24	0.89
6	0.05	2.61	3.66	6.60	0.28	-0.01	2.61	4.54	8.95	0.74
7	-0.52	239.74	0.66	1.85	0.62	-0.55	239.57	0.77	3.03	0.82
8	-0.62	218.14	0.63	2.10	0.60	0.08	218.41	0.73	3.26	0.80
9	-0.46	6.99	2.20	5.58	0.36	-0.46	6.99	2.67	8.52	0.80
10	0.16	209.45	0.64	2.33	0.66	0.07	205.56	0.73	3.57	0.87
11	-0.09	1.54	2.20	7.49	0.28	0.16	1.05	2.44	6.94	0.55
12	-0.09	10.29	2.37	4.01	0.48	-0.06	10.29	2.94	6.43	0.59
13	-0.13	2.74	1.88	3.70	0.26	-0.17	2.74	2.75	7.56	0.46
14	-0.05	1.92	1.90	4.28	0.36	-0.04	1.92	2.56	8.40	0.58
15	-0.21	50.75	1.03	2.47	0.55	-0.26	50.83	1.21	3.79	0.76
16	-0.15	4.15	1.67	2.92	0.37	0.06	4.21	2.31	5.34	0.53
17	0.25	9.15	0.93	0.57	0.72	0.19	9.18	1.04	1.47	0.82
18	-0.06	56.29	1.04	2.34	0.55	-0.10	56.39	1.23	3.85	0.76
19	-0.50	156.94	1.03	3.16	0.51	-0.40	156.92	1.18	4.06	0.83
20	-0.41	4.59	1.09	2.27	0.60	-0.43	4.60	1.19	2.84	0.88
21	-0.96	86.19	0.61	2.16	0.63	-0.94	86.31	0.66	2.91	0.89
22	-0.29	65.24	1.46	3.73	0.50	-0.26	65.26	1.67	5.10	0.79
23	-0.22	18.71	1.55	3.74	0.48	-0.23	18.73	1.82	6.37	0.73
24	-0.24	8.32	1.16	2.23	0.60	-0.25	8.34	1.28	2.84	0.85
25	-0.15	18.08	1.18	2.63	0.59	-0.18	18.12	1.29	3.13	0.88

NOTE: *The numbers presented correspond to the same streamflow data outlined in Table 5.1.*

C_V indicates Coefficient of Variation; C_S skewness, and ρ_1 autocorrelation coefficient at lag 1.

The Table 5.2 indicate the statistical characteristics of the monthly and weekly raw streamflow data sets. Here, lambda represents the optimal value used in the Box-Cox normalization transformation, ensuring the data achieves normality. C_V refers to the coefficient of variation, calculated by dividing the standard deviation by the mean, providing insight into the relative variability of the data. C_S denotes the skewness, indicating the asymmetry of the data distribution.

Lastly, "ρ_1" represents the autocorrelation coefficient at lag 1, measuring the correlation between current and previous time steps in the dataset.

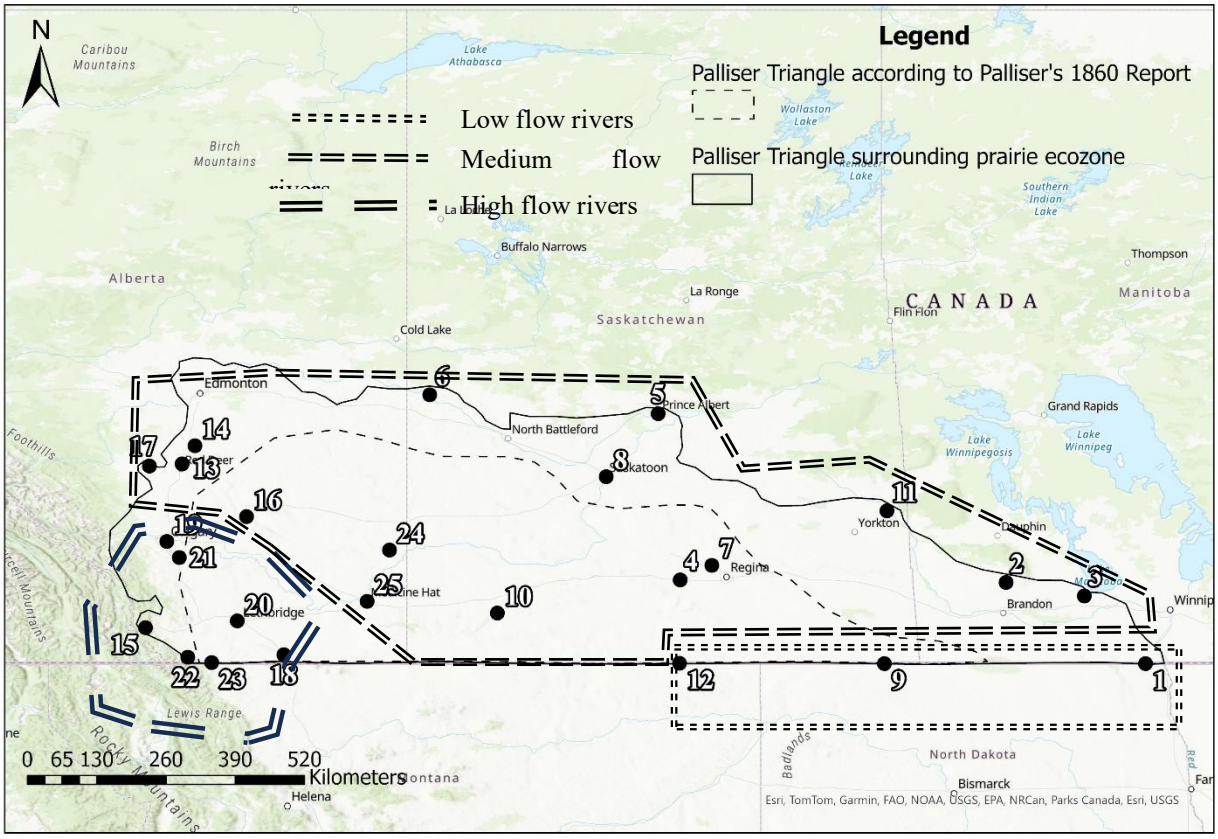


Figure 5.4 Low flow, medium flow and high flow river locations.

The next chapter (Chapter VI) will involve applying the developed models to the selected watersheds. It will detail how the models are implemented in these specific regions to recognize and predict drought patterns based on the previously discussed methodologies.

Chapter VI: Application of Models

In this chapter, the efficacy of the proposed models is evaluated by applying them to the selected 25 watersheds known for their drought-prone characteristics. As discussed in Chapter V, these watersheds were specifically chosen from the Palliser Triangle area in Canada where drought conditions are unique and frequent. A monthly dataset spanning 48 hydrological years starting from October 1, 1973, to ending in September 30, 2022 was chosen for analysis from each streamflow station. The models were applied to all low-flow, medium-flow and high-flow per unit area generating capacity of rivers within the Palliser Triangle. The application of models on specific rivers representing the flow generation capacity based on flow rate per unit area basis as low-flow river (the East Poplar River), the medium-flow river (the South Saskatchewan River), and the high-flow river (the Highwood River) is presented. A detailed analysis demonstrates how models perform across different drought levels and flow characteristics.

6.1 Implementation of Streamflow Data for Proposed Models

After carefully selecting 25 streamflow watersheds, the chosen data was used to evaluate the three models which are the Discrete Markov Model, the Hidden Markov Model, and the Long Short-Term Memory (LSTM) model. Before applying these models, the data underwent a transformation process to ensure it was suitable for modelling. This transformation involved normalization to ensure that the data was statistically suitable. In turn, the data was analysed using different techniques to identify the meaningful drought patterns or groups within the streamflow data.

Figure 6.1 illustrates the process of model application. Initially, the data must be transformed to suit the models being used. Subsequently, drought patterns need to be recognized before applying the models to ensure accurate forecasts.

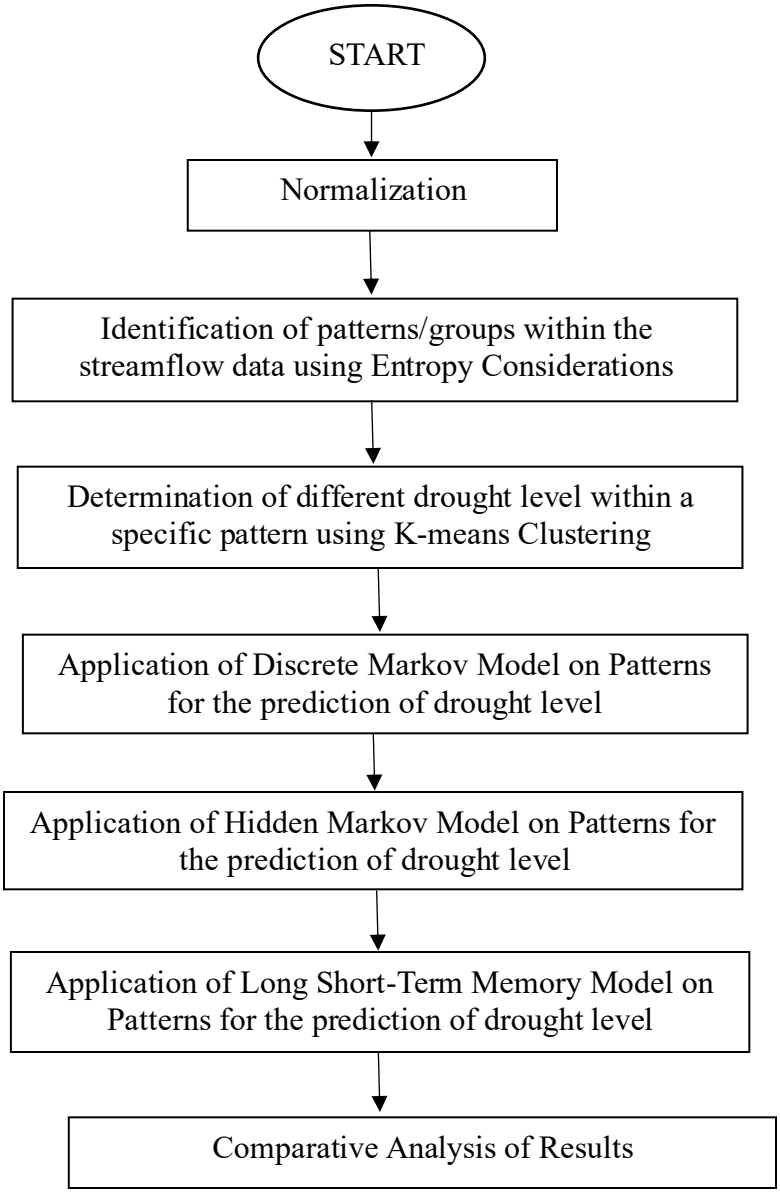


Figure 6.1 Flow chart for model application.

6.2 Normalization of Data Sets

Before applying the models, it is necessary to transform the data to a near-normal distribution to ensure suitability for modelling. This transformation was conducted using the Box-Cox normal transformation (Eq. 4.02 in Chapter IV), which stabilized variance and made the data closer to a normal distribution.

The Rankit plot, also known as the normal probability plot, was used to visually evaluate if the transformed data were normally distributed. In this plot, the quantiles of the transformed data are plotted against the theoretical quantiles of a normal distribution. If the data follows a normal distribution, the Rankit plot is an approximate straight line close to the expected normal distribution line.

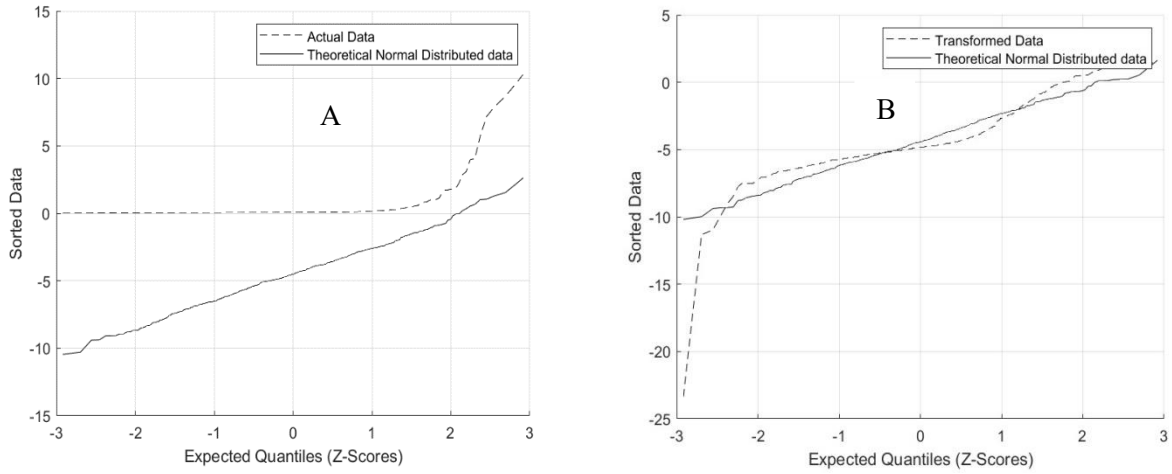


Figure 6.2 Rankit normalization comparison for actual monthly streamflow data (A) and transformed monthly streamflow data (B) - East Poplar River at International Boundary.

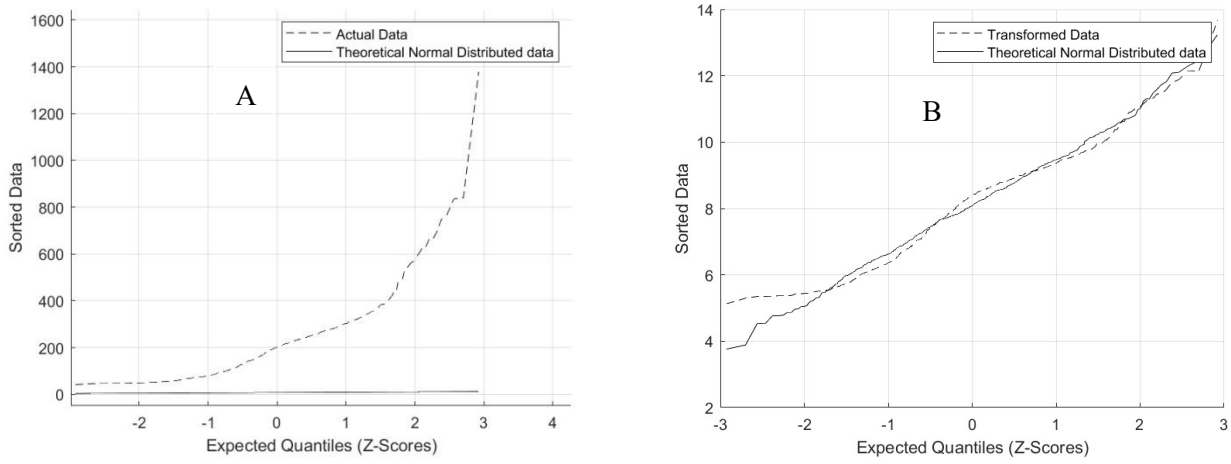
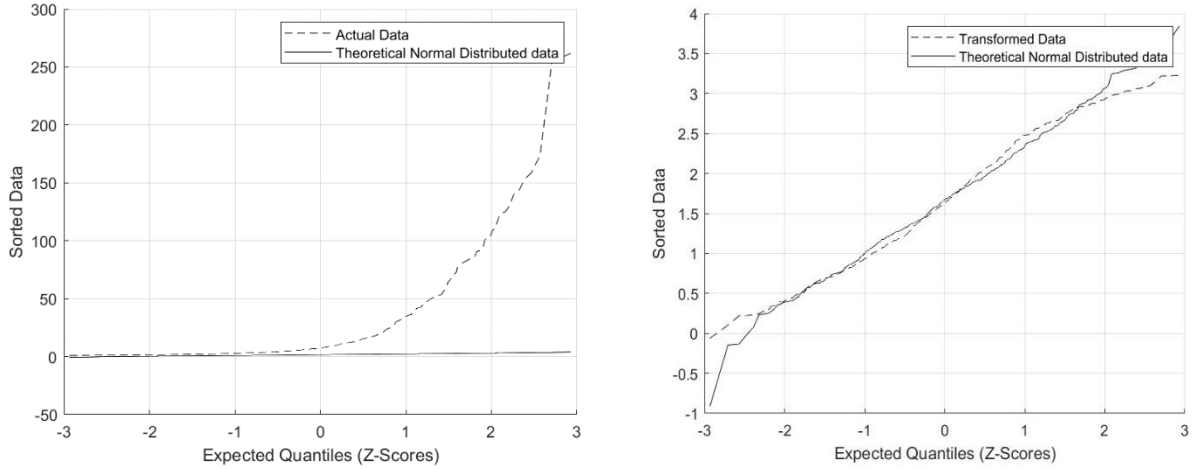


Figure 6.3 Rankit normalization comparison for actual monthly streamflow data (A) and transformed monthly streamflow data (B) - South Saskatchewan River at Saskatoon station.



A B

Figure 6.4 Rankit normalization comparison for actual monthly streamflow data (A) and transformed monthly streamflow data (B) - Highwood River near its mouth.

Rankit plot illustrating a comparison between the theoretical normal distribution and the actual monthly streamflow data is shown in A while B represents a comparison between the theoretical normal distribution and transformed data for the East Poplar River (Figure 6.2), South Saskatchewan River at the Saskatoon station (Figure 6.3) and Highwood River (Figure 6.4). Those plots visually demonstrate how well the data conforms to a normal distribution after the

transformation process. As shown in the Figure 6.2 for low flow East Poplar river, the Box-Cox transformation can produce negative values, even if the original data is positive. This can occur when using negative λ values, especially for right-skewed data. When the optimal λ value determined by the Box-Cox transformation is negative, the transformation is a reciprocal transformation. This can result in negative values, especially for smaller values in the original data. Appendix B provides similar Rankit plots for all 25 streamflow stations, allowing for a comparative analysis of normalization across different watersheds.

6.3 Recognition/Classification of Drought Patterns/Groups

The optimal patterns (or groups) within a streamflow data were recognized/classified using entropy considerations. The goal was to achieve minimal intra-entropy, or entropy within the groups, to ensure that each group is as internally cohesive as possible, and maximal inter-entropy, or entropy among the groups, to ensure that the groups are externally distinguishable. While a hydrological year can have groups of varying lengths, this research focused on groups of similar lengths for ease of operations. Therefore, the length of a group can be 2 months, 3 months, 4 months, or 6 months. Intra- and inter-entropy calculations were performed for these different lengths of groups based on the Streamflow Drought Index (SDI). Using the Shannon Entropy (equation 3.06), the intra-entropy was calculated considering the DL (Drought Level) of each

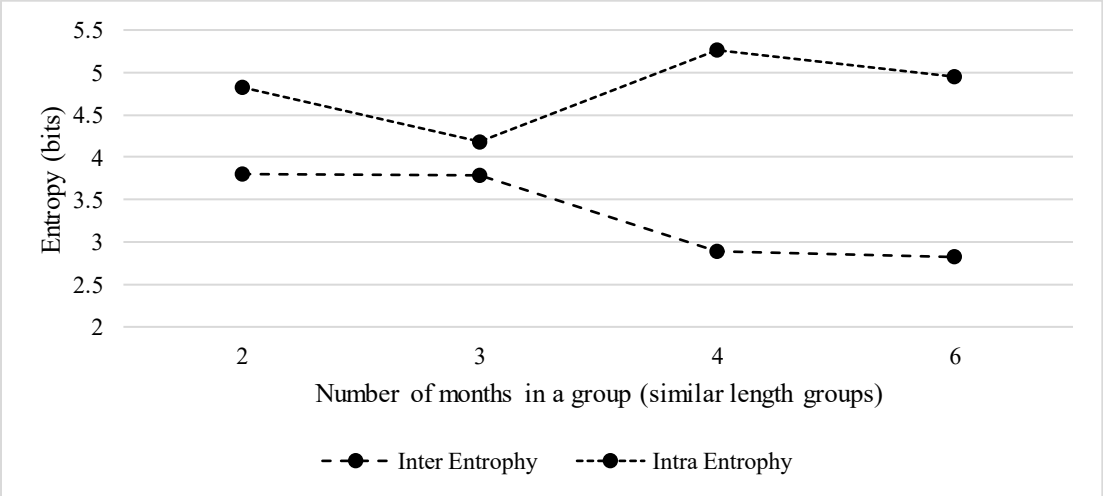


Figure 6.5 Intra and inter entropy calculation for groups with different length for East Poplar River at International Boundary.

month within a group and inter-entropy was calculated considering the DL for the entire group of patterns.

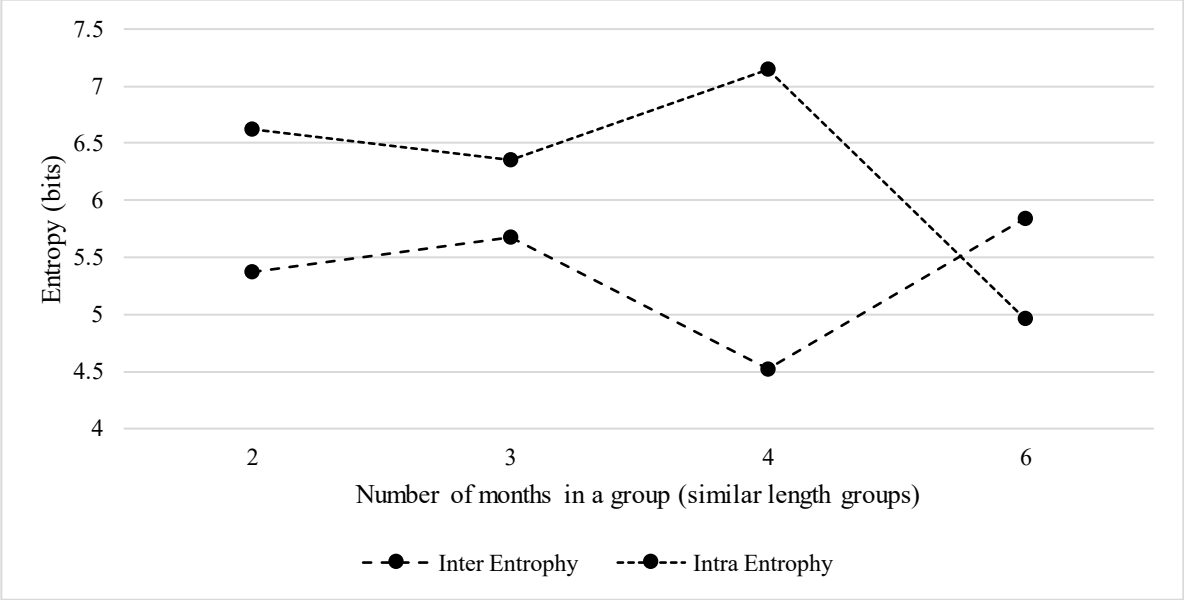


Figure 6.6 Intra and inter entropy calculation for groups with different length for South Saskatchewan River at Saskatoon station.

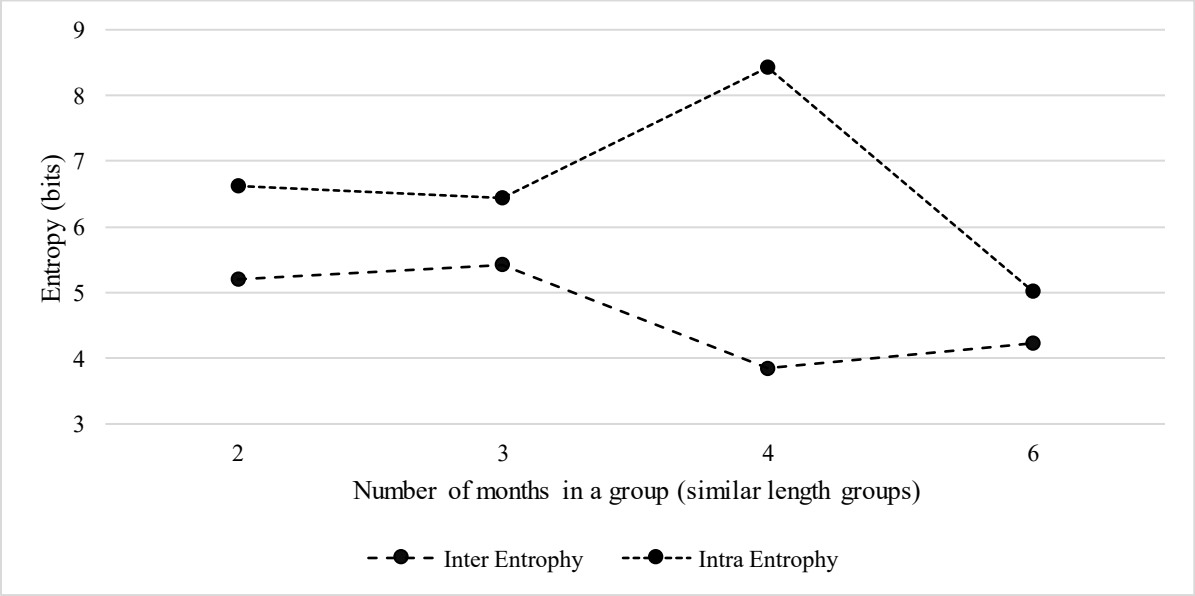


Figure 6.7 Intra and inter entropy calculation for groups with different length for Highwood River near its mouth.

Figures 6.5, 6.6 and 6.7 illustrate the results of intra- and inter-entropy calculations for various group lengths for the East Poplar River, South Saskatchewan River, and Highwood River respectively. The figures indicate that a 6-month group length is optimal for identifying patterns, effectively dividing a hydrological year into two distinct groups: wet and dry. Since the research specifically focuses on drought conditions, only the dry group is considered for further analysis. Consequently, the next best grouping length is selected as 3 months to critically evaluate the droughts for the South Saskatchewan River. Appendix B provides entropy calculations for all 25 streamflow stations, with the majority showing that 3 months is the optimal group length. Therefore, a 3-month grouping is chosen for subsequent analyses as shown in the below equation,

$$\text{Pattern or group} = [M_1, M_2, M_3] \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (6.01)$$

Given that the data is divided into 3-month groups within a hydrological year, resulting in four distinct groups per year, these groups can be referred to as seasonal patterns or seasonal groups. Such a segment of 3 months representing a specific season or quarter of the year is designated as Season 1 from October to December, Season 2 from January to March, Season 3 from April to June, and Season 4 from July to September. The analysis shows streamflow variations according to seasons throughout the year and each seasonal pattern for i^{th} year can be demonstrated below,

$$\text{Seasonal Pattern } j = SP_{i,j} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (6.02)$$

Where i indicate i^{th} year and varies from 1 to 48^{th} year representing the hydrological year starting from 1974 to 2022. Seasonal pattern number within a given year is indicated by j which varies from 1 to 4 as follows.

$$j = 1 = \text{Seasonal pattern 1 (October 1, to December 31)} \quad \ldots\ldots\ldots\ldots\ldots\ldots (6.03)$$

$$j = 2 = \text{Seasonal pattern 2 (January 1, to March 31)} \quad \ldots\ldots\ldots\ldots\ldots\ldots (6.04)$$

$$j = 3 = \text{Seasonal pattern 3 (April 1, to June 30)} \quad \ldots\ldots\ldots\ldots\ldots\ldots (6.05)$$

$$j = 4 = \text{Seasonal pattern 4 (July 1, to September 30)} \quad \ldots\ldots\ldots\ldots\ldots\ldots (6.06)$$

6.4 Characteristics of Seasonal Patterns

After identifying the structure of the seasonal patterns, with each pattern consisting of three months (e.g., Season 1: October 1, to December 31, Season 2: January 1, to March 31, etc.), a moving average technique is applied to understand the characteristics of these patterns. Using a moving average window of three months, the streamflow data can be smoothed to provide a better representation. This smoothing process helps to reduce short-term fluctuations and highlight long-term patterns in the data. After applying the smoothing technique, the average value for each month over the 48 years can be calculated. This allows for an examination of how the streamflow values for each month have changed, on average over the years, providing insights into characteristics of seasonal patterns.

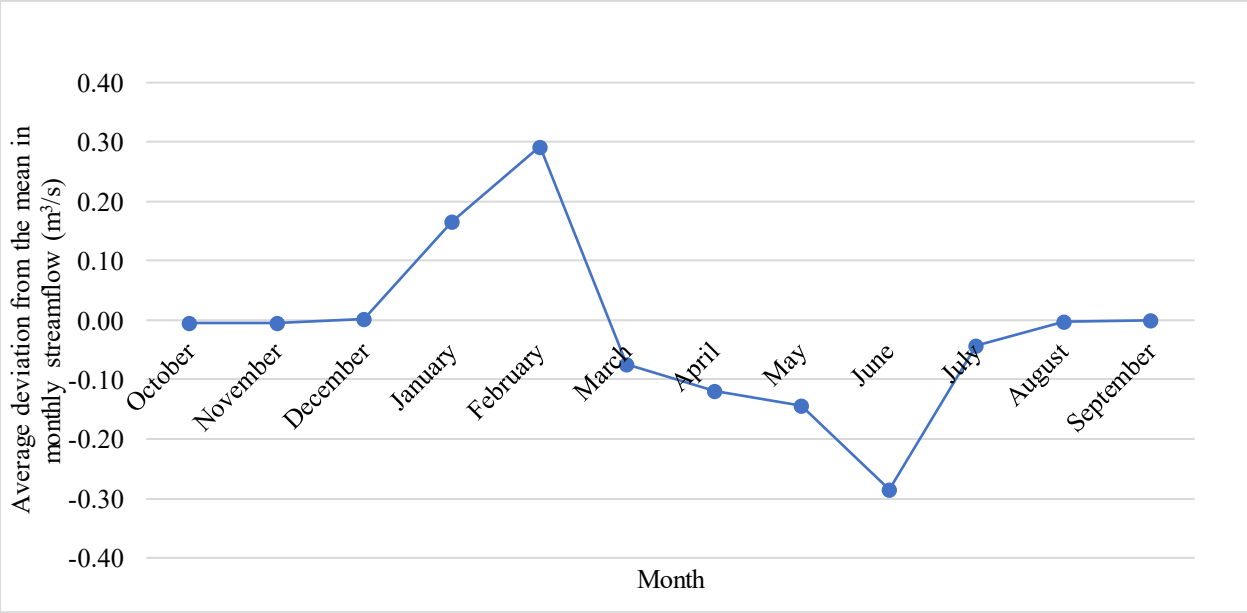


Figure 6.8 Average deviation from the mean in monthly streamflow for East Poplar River.

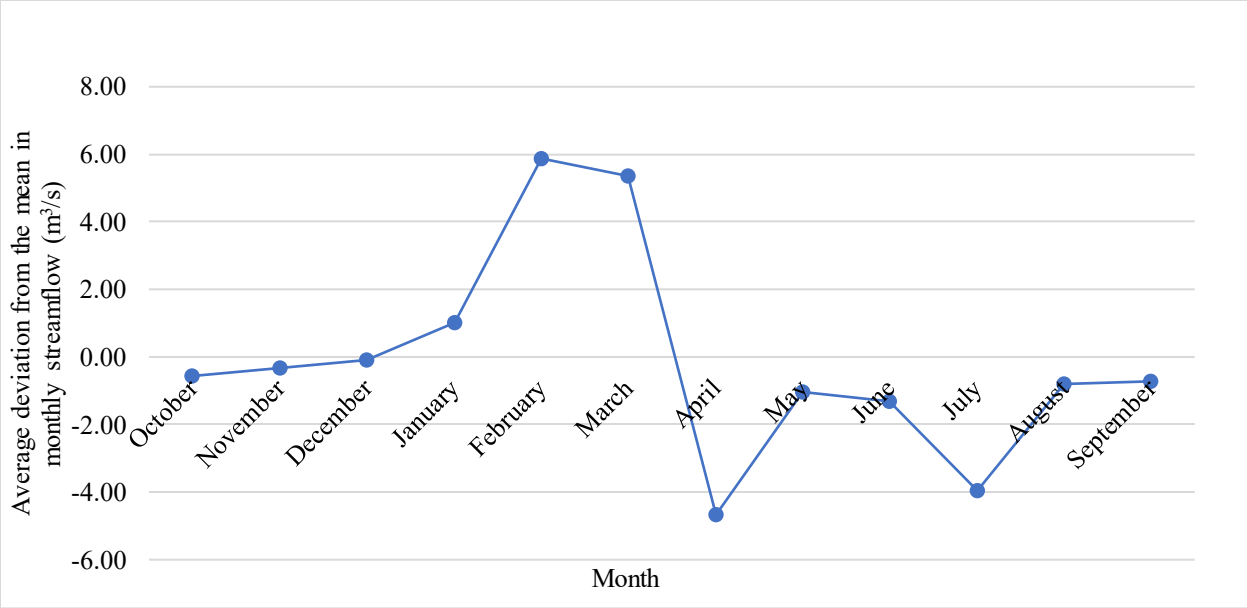


Figure 6.9 Average deviations from the mean in monthly streamflow for South Saskatchewan River.

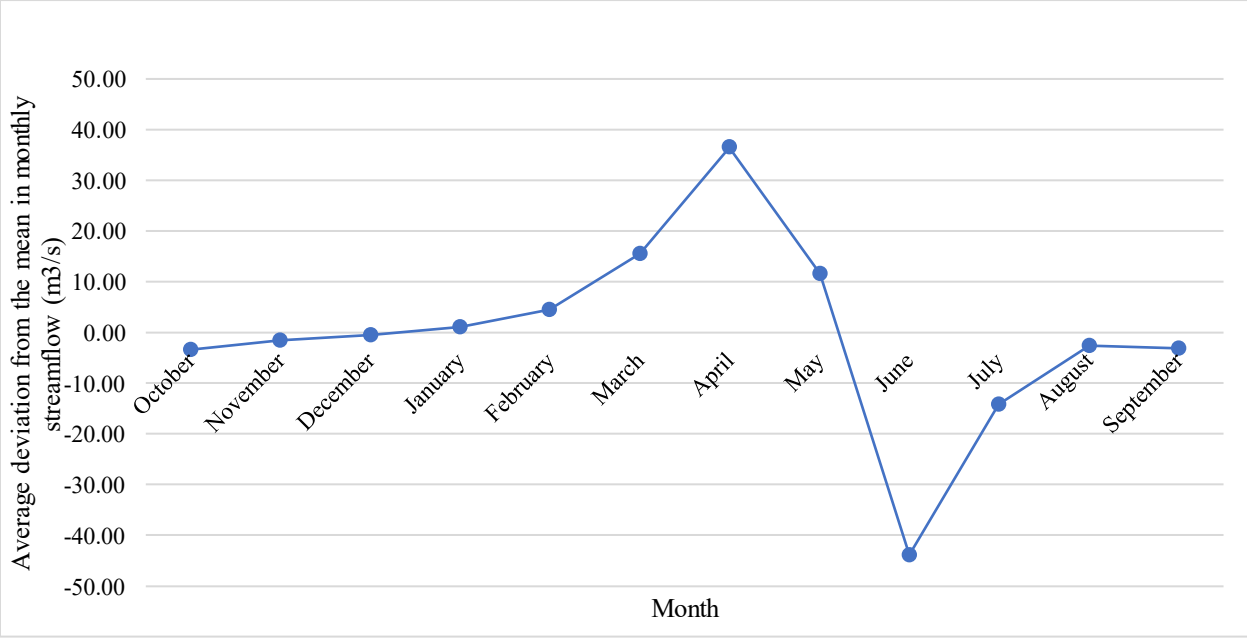


Figure 6.10 Average deviation from the mean in Monthly streamflow for Highwood River

As seen in Figures 6.8 to 6.10, Seasonal Pattern 1, which consists of October, November, and December, on average characterized by a less severe drought pattern. Following this, Seasonal Pattern 2, encompassing January, February, and March, transitions into wet months, indicating a shift from mild drought levels to no drought conditions. However, Seasonal Pattern 3, which includes April, May, and June, exhibits a more severe drought level for both low-flow and medium-flow rivers throughout all three months. This signifies a distinct transition from no drought to a more severe drought condition. For high-flow rivers, Seasonal Pattern 3 initially shows no drought, but by June, drought conditions start to emerge. Subsequently, July, August, and September, representing Seasonal Pattern 4, return to a mild drought pattern across low, medium, and high flow-producing capacity based on flow rate per unit area of rivers.

From a visual analysis of Figures 6.8 to 6.10, it is apparent that there are four seasons within a year, indicating two wet seasons and two dry seasons. However, when examining high-flow rivers, it becomes evident that these rivers tend to have just two main seasons over a year: one dry season and one wet season. Such seasonal pattern differs from those observed in low flow and medium flow producing capacity based on flow rate per unit area of rivers, which exhibit more distinct seasonal variations with separate dry and wet periods.

Furthermore, as the entropy calculations provide 3-month patterns to be the optimal size, 4 patterns or groups per hydrological year are considered as 4 different seasonal patterns. Appendix B provides the Moving Average plots for all 25 streamflow stations.

6.5 Clustering the Seasonal Drought Patterns

As a hydrological year is divided into four seasonal patterns, the optimal number of clusters within each seasonal pattern for 48 years was determined using K-means clustering. As mentioned in the characteristics of patterns, on average, Seasonal Pattern 1 is characterized by a mild drought pattern, Seasonal Pattern 2 exhibits no drought, Seasonal Pattern 3 experiences a more severe drought pattern, and Seasonal Pattern 4 returns to a mild drought pattern. However, in different years, these drought patterns can deviate from such average levels. To analyze such variations, a K-means clustering algorithm was used, providing an insightful analysis of how drought levels can change within a seasonal pattern.

For instance, the seasonal pattern 1 which comprises the months from October to December for all 48 years was selected, and K-means clustering was performed to find the optimal number of clusters within patterns of Season 1. The optimal clustering for patterns of each season was determined using the Silhouette score. This score is the metric used to evaluate the quality of clustering by measuring how similar each data point is to its cluster compared to other clusters. It ranges from -1 to 1 and when a score is closer to 1 indicates that the data point is well-matched to its cluster and is distinct from other clusters. A score around 0 suggests that the data point lies on the boundary between clusters and a negative score indicates that the data point may have been incorrectly assigned to its cluster.

To find the optimal number of clusters using the Silhouette score, the average silhouette score for different numbers of clusters (K) was calculated. The optimal number of clusters is the one that maximizes the average silhouette score, indicating well-defined and well-separated clusters. For the silhouette score calculation, only drought data was considered. Drought data refers to the normalised streamflow data below a specified threshold level, which in this case is defined as the mean value. By focusing on data below this threshold, the clustering analysis specifically targets periods of lower streamflow, providing insights into drought patterns within each seasonal pattern.

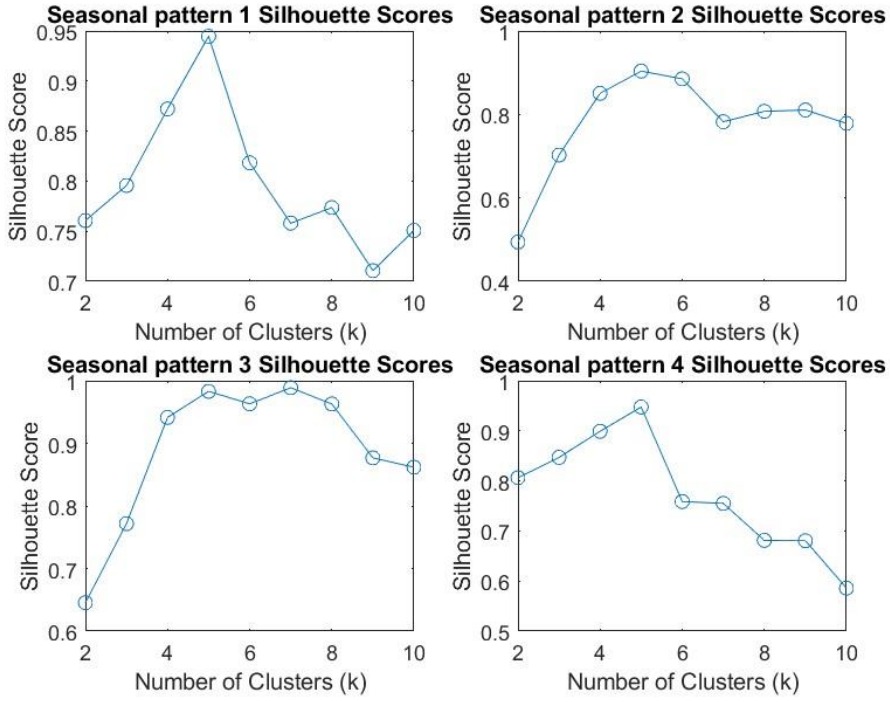


Figure 6.11 Silhouette score for seasonal patterns in East Poplar River at International Boundary.

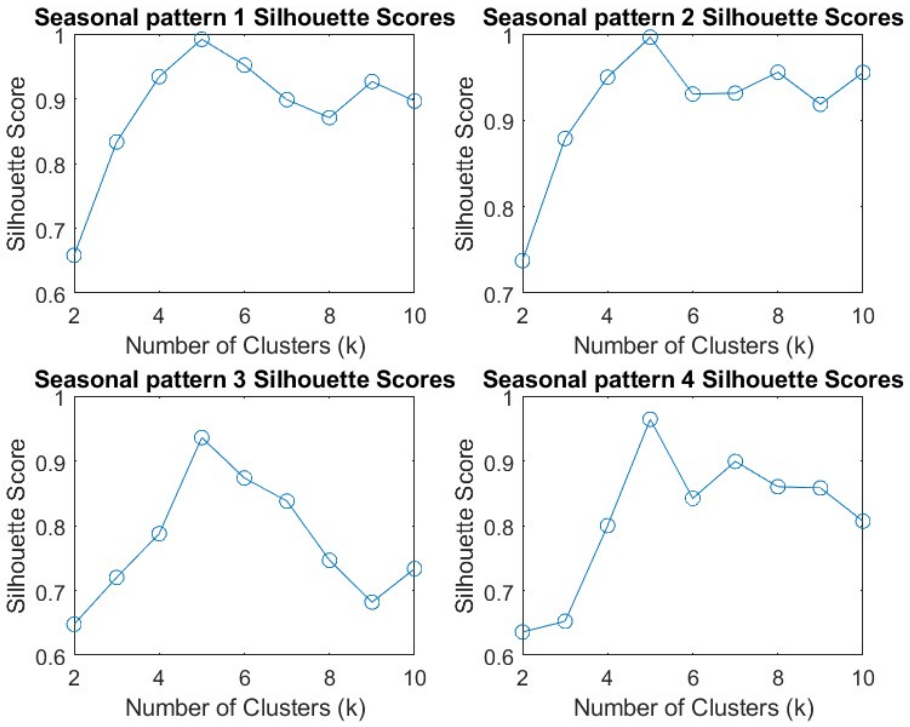


Figure 6.12 Silhouette score for seasonal patterns in South Saskatchewan River at Saskatoon Station.

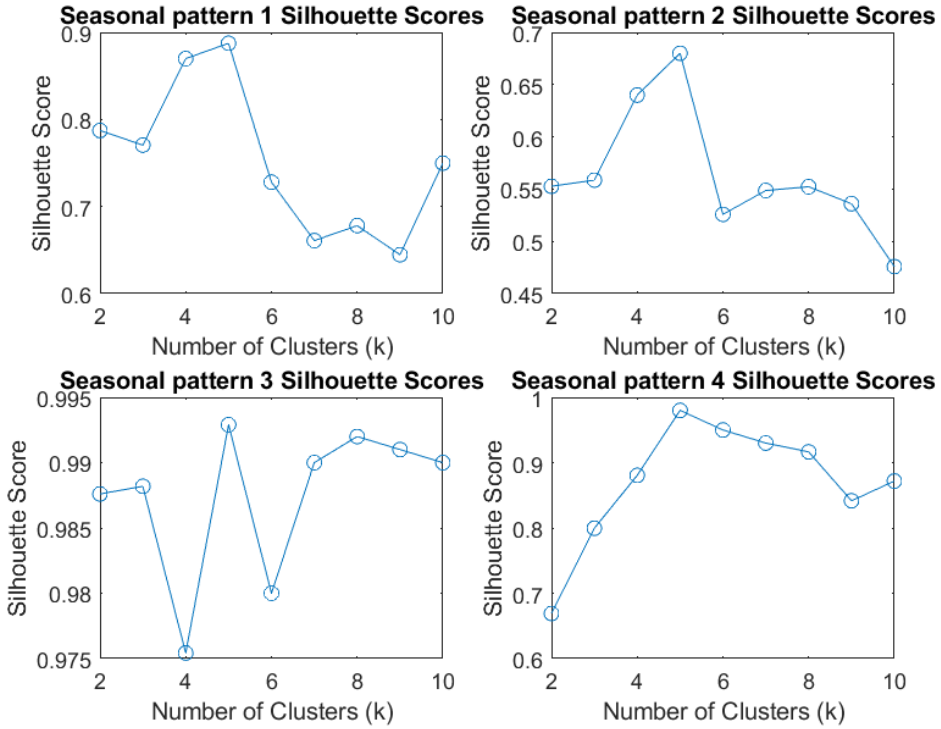


Figure 6.13 Silhouette score for seasonal patterns in Highwood River near its mouth.

As seen in Figures 6.11, 6.12 and 6.13, all four seasonal patterns exhibit an optimal number of 5 drought clusters. This finding aligns with the Streamflow Drought Index (SDI), which categorizes monthly drought data into 5 clusters based on their severity. Therefore, in this thesis, seasonal patterns are categorized into five drought levels: no drought(D0), mild drought(D1), moderate drought(D2), severe drought(D3), and extreme severe drought(D4). Consequently, the grouped drought data can also be classified into 5 different drought severity classes according to the Streamflow Drought Index of each group as shown in the table below.

Table 6.1 Drought levels based on SDI values for seasonal patterns/groups.

State	Drought Level	Criterion for mean SDI value
0	No Drought	SDI ≥ 0.0
1	Mild Drought	-1.0 ≤ SDI < 0.0
2	Moderate Drought	-1.5 ≤ SDI < - 1.0
3	Severe Drought	-2.0 ≤ SDI < - 1.5
4	Extreme Drought	SDI < -2.0

Appendix B provides the Silhouette scores for seasonal patterns of all 25 streamflow stations, with the majority showing 5 clusters within the seasonal patterns. After recognizing the seasonal patterns or groups within the data and identifying the drought level of each seasonal pattern, these patterns were used in applications to three models which are the Discrete Markov model, the Hidden Markov model, and the Long Short-Term Memory (LSTM) model. The models were evaluated on their performance in predicting the drought level of seasonal patterns for low-flow, medium and high-flow rivers.

6.6 Applications of the Model 1: Discrete Markov Model (DMM)

The Discrete Markov model was applied to each river, to predict the drought level of the seasonal patterns. There are four distinct types of seasonal patterns namely SP_j, j=1, 2, 3, and 4. Each type of Pattern (SP_j, has 48 members) and thus there are 192 SP_j. Each type of SP_j can belong to one of five drought categories (or levels). To assign a specific SP_j, to a specific drought level (or state), the mean Streamflow Drought Index of the pattern is used. Based on the classification of a specific SP_j to a specific drought level (or state), a transition matrix can be developed describing the

probability of transition from one drought level (or state) to another drought state (or level) for predictive purposes such as Markov state forecast.

It is crucial to select an appropriate portion of the dataset for training and testing to ensure the model is trained effectively while also allowing it to be tested on a broader range of drought-level variations. This balanced approach helps the model learn from diverse patterns during training and accurately assess its performance across different drought levels during testing. Recent studies have shown that the model performance for statistical and machine learning models for streamflow forecasting can be affected by the different training and testing data splits (Feng et al., 2020). There are studies done to show that a 70:30 training-testing data split can offer promising results for predicting hydrological variables using machine learning models (Rahimi & Ebrahimi, 2023). The study has shown that streamflow modeling 70:30 could give more accurate results than 90:10 or 80:20, as the smaller testing data size, in the latter approaches may not be sufficient to adequately assess the generalization capability of the model. The study discusses the importance of training and testing data to ensure that the testing dataset is large and varied enough for a more comprehensive evaluation of how well the model performs across different drought levels and conditions. This confirms that the model is not just overfitting to the training data but can also accurately predict drought patterns in new, unseen data (Rahimi & Ebrahimi, 2023). Therefore, this study also follows the 70:30 training and testing split to train and test the models.

To build the transition matrix, 70% of the data was used for training, which amounts to 135 seasonal patterns. The remaining 30% of the data, comprising 57 seasonal patterns, was used for forecast based on the transition matrix.

6.7 Applications of the Model 2: Hidden Markov Model (HMM)

The Hidden Markov Model was applied to predict the drought level of the seasonal patterns as described in the methodology chapter.

Instead of focusing on the drought level of the group, which was based on the average streamflow drought index of the group, this method directly employs the streamflow drought index (SDI) values for each seasonal pattern. Consequently, a seasonal pattern contains three monthly streamflow drought indices to the model to train on the patterns of these indices and to predict the

SDI values for the next seasonal pattern (or the next three months). Using the predicted SDI values, the drought level for the upcoming seasonal pattern can then be calculated.

With 48 years of data, resulting in 192 seasonal patterns, 70% of the patterns (135 patterns) were used for training the HMM, while the remaining 30% (57 patterns) were reserved for testing. This approach ensures that the model is well-trained before being tested on the 57 seasonal patterns to evaluate its forecast accuracy.

6.8 Applications of the Model 3: Long Short-Term Memory (LSTM) Model

Similar to other model applications, the same approach is used for the LSTM model, where the streamflow drought indices for each seasonal pattern are employed. The model is trained on these indices to predict the streamflow drought indices for the next seasonal pattern or the next three months, allowing for the calculation of the drought level of the upcoming seasonal pattern based on the predicted values.

The LSTM model was applied to predict the drought level of the seasonal patterns with 48 years of data. Accordingly, 70% of the seasonal patterns which is 135 patterns were used for training the LSTM, while the remaining 30% (57 patterns) were used for testing.

6.8.1 Selection of Model Parameters for Monthly LSTM Model

First, the LSTM model was applied to predict values of the SDI using the monthly data. The Adam optimizer (for brevity also known as (Adaptive Moment Estimation) is a popular optimization algorithm used in the training of deep neural networks, including Long Short-Term Memory models. The Adam optimizer provides several advantages over traditional gradient descent algorithms for complex time series forecasting, such as an ability to adaptively adjust the learning rate for each parameter based on the estimated mean and variance of the gradients (Kingma & Ba, 2015). Therefore, to predict SDI values, the Adam optimizer was chosen for the LSTM model.

To get a balanced pace for the optimization process and to ensure that the model can make significant progress in learning without overshooting the optimal parameters, the learning rate and the number of iterations (epochs are the same as iteration number) were selected as 0.01 and 800.

The training plots for a low-flow river (East Poplar River), medium-flow river (South Saskatchewan River), and high-flow river (Highwood River) are shown in the figures to follow.

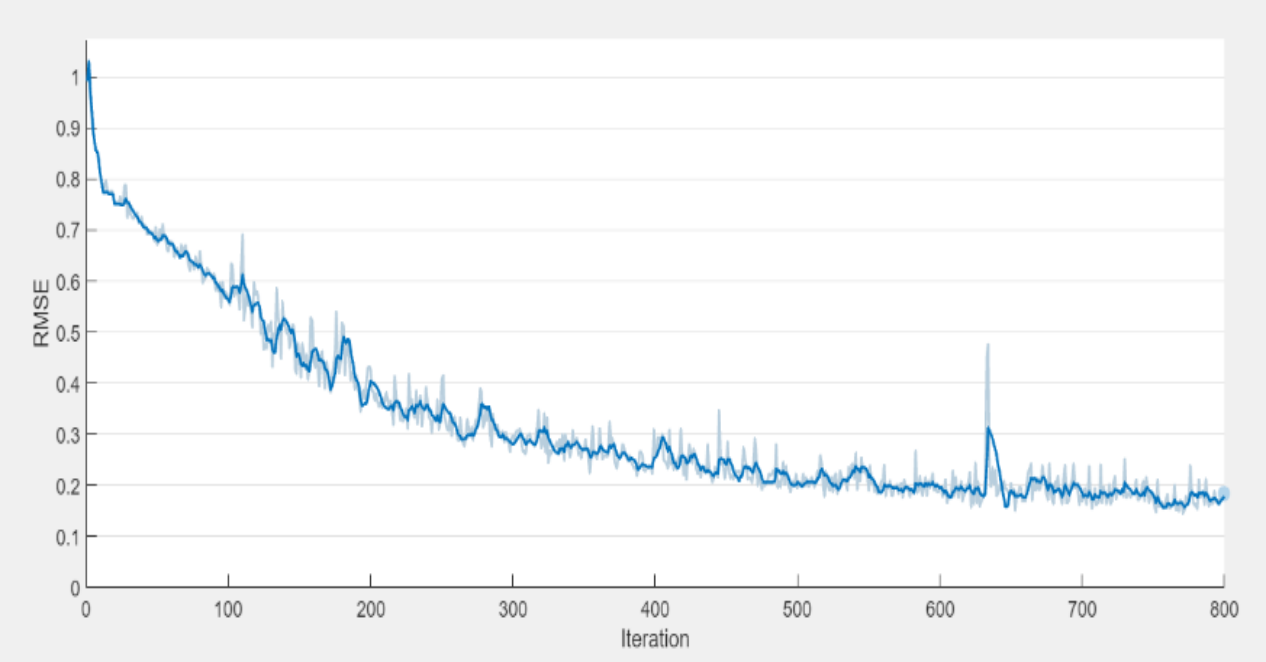


Figure 6.14 Training plot for the East Poplar River.

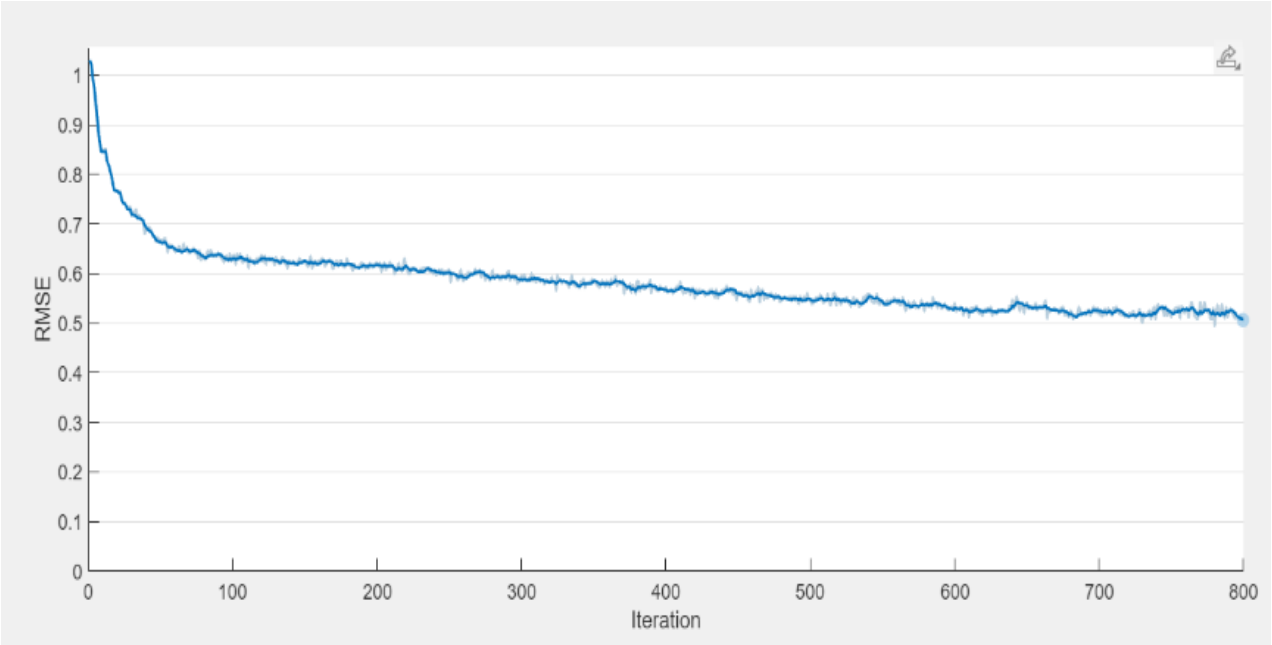


Figure 6.15 Training plot for the South Saskatchewan River.

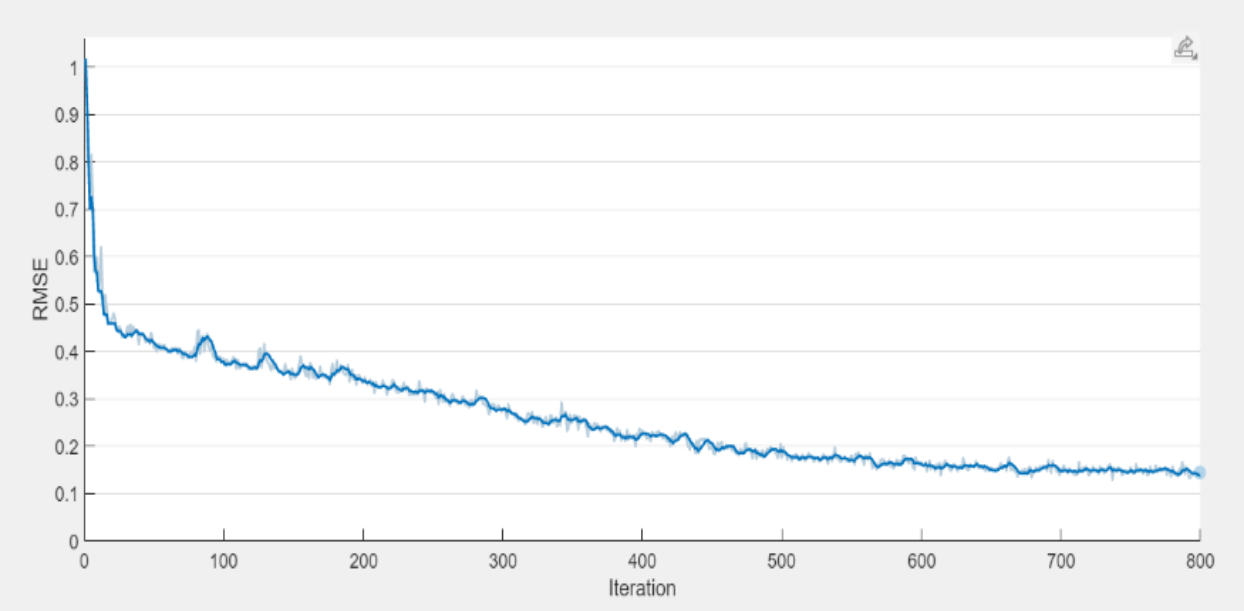


Figure 6.16 Training plot for the Highwood River.

As seen in Figures 6.14, 6.15, and 6.16, both low flow and high flow rivers require approximately 800 iterations to reach optimization, whereas medium flow rivers achieve optimization within around 600 iterations. This is because low flow and high flow rivers exhibit significant flow variations and more complex patterns compared to medium flow rivers. As a result, they require more iterations to achieve optimization. Medium flow rivers, with less variation and simpler patterns, reach optimization more quickly.

LSTM models have demonstrated the ability to capture the complex non-linear relationships inherent in hydrological time series data (Zhang et al., 2018). However, a key concern in the application of LSTM models for streamflow forecast is the requirement for large amounts of training data (Liu et al., 2019). Studies have shown that the reliability of LSTM models for streamflow forecast can be influenced by the frequency of the input data (Kwon et al., 2023). Specifically, research has indicated that LSTM models trained on weekly data may provide more reliable results compared to those trained on monthly data (Kwon et al., 2023). Therefore, the LSTM model was also applied to the weekly data set of rivers within the project study area.

6.8.2 Selection of Model Parameters for Weekly LSTM Model

The LSTM model was applied using weekly data instead of monthly data. Therefore, three-month seasonal patterns for 48 years span into 192 weekly seasonal patterns. Instead of having 3 elements in the drought pattern vector (Eq. 5.01), the seasonal patterns have 13 elements corresponding to weekly streamflow data for three months as shown below.

$$\text{Weekly pattern/group} = [W_1, W_2, W_3, W_4, W_5, W_6, W_7, W_8, W_9, W_{10}, W_{11}, W_{12}, W_{13}] \ldots. (6.06)$$

Accordingly, 70% of the seasonal patterns which is 135 patterns were used for training the LSTM model, while the remaining 30% (57 patterns) were used for testing the LSTM model.

Similar to monthly LSTM model, the weekly LSTM model was optimized using the Adam optimizer. However, due to a more detailed distribution in the weekly data, the number of iterations was reduced to 500 from 800 as was used for the monthly LSTM. The learning rate was kept consistent at 0.01 for both models. The training plots for a low-flow (East Poplar River), medium-flow river (South Saskatchewan River), and high-flow river (Highwood River) for the weekly LSTM model are shown in Figures 6.17, 6.18 and 6.19.

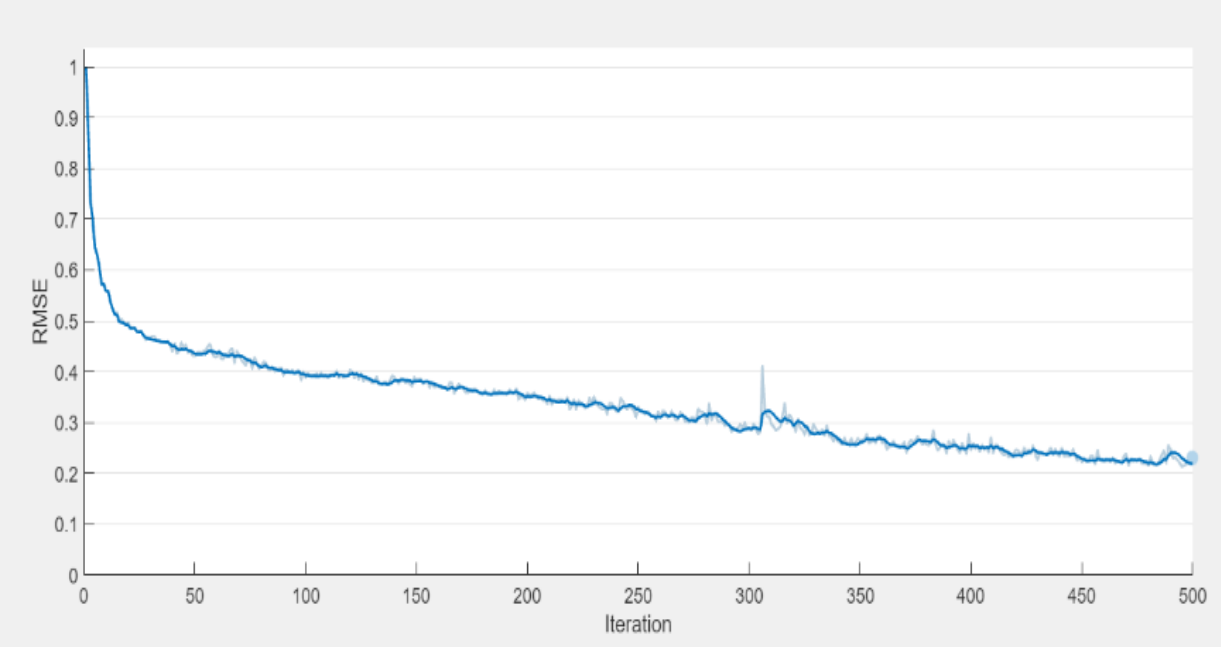


Figure 6.17 Weekly LSTM model training plot for East Poplar River.

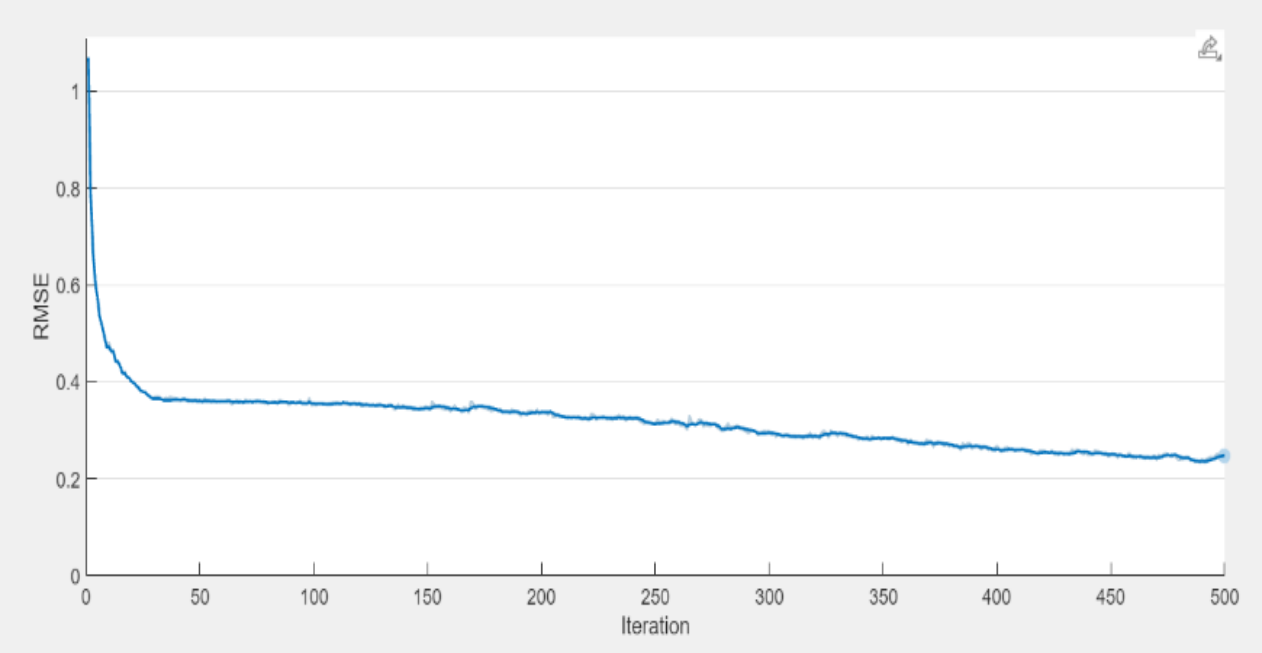


Figure 6.18 Weekly LSTM model training plot for South Saskatchewan river

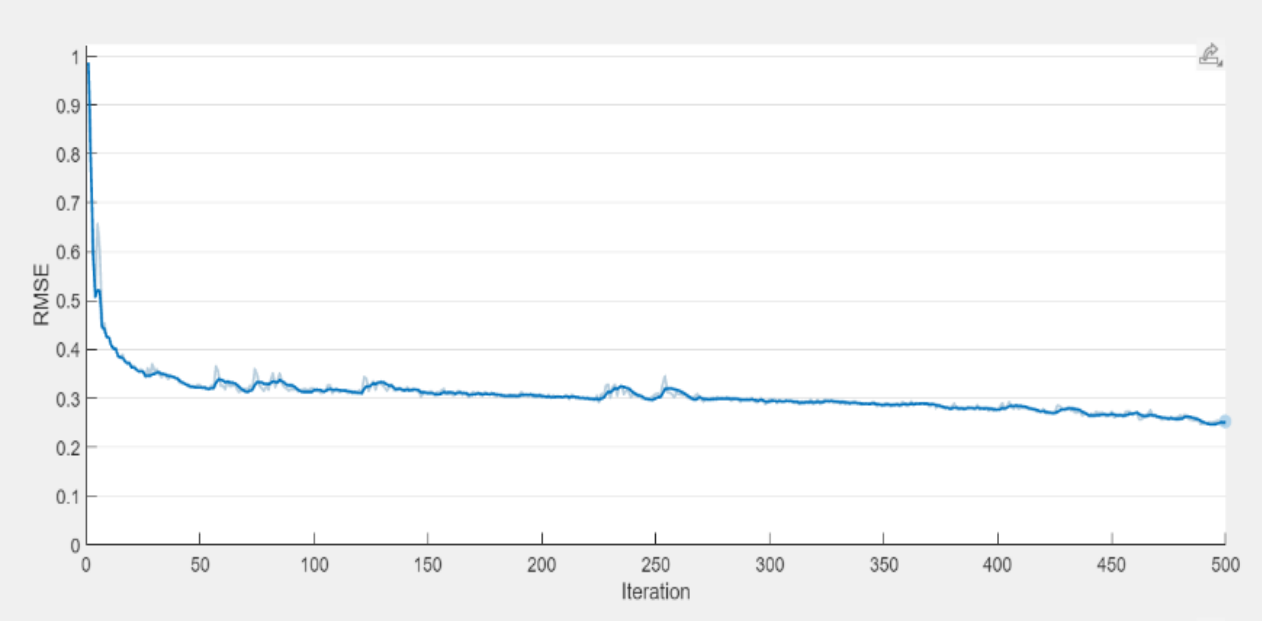


Figure 6.19 Weekly LSTM model training plot for Highwood River.

When examining the training plots for the weekly LSTM model, it is observed that the model achieves optimization within fewer iterations (500) compared to the monthly LSTM model (800). This is because the weekly LSTM can capture more refined details in weekly patterns, allowing it to converge more quickly than the monthly LSTM, which processes less granular data.

In the next chapter, the results obtained from applying different models to all selected rivers are summarized in tables and exhibited in figures. Additionally, model evaluation measures such as Mean Squared Error (MSE) and R-squared (R^2) are compared for the Hidden Markov model (HMM), and the Long Short-Term Memory (LSTM) model. The comparison also includes an analysis of how all three models (Discrete Markov Model (DMM), HMM, and LSTM) predict the drought level by examining the mismatch in drought levels to check the accuracy of model forecasting. This comprehensive analysis highlights the performance and effectiveness of each model in predicting drought levels across different river types.

Chapter VII: Results and Discussion

The results of the proposed discrete Markov model, Hidden Markov model, and LSTM model are presented, compared, and discussed. Graphical analysis includes visual representations of forecasts and observed data, while statistical analysis involves evaluating model performance using measures such as Mean Squared Error (MSE) and R-squared. The models were then evaluated to assess their performance in predicting each drought level. Through such analyses, the chapter provides a comprehensive comparison of the effectiveness of models in predicting drought levels across the selected rivers. The graphical presentation of results for all models on specific rivers representing the low-flow (East Poplar River), the medium-flow river (South Saskatchewan River), and the high-flow river (Highwood River) are presented in this chapter while the model applications for all 25 watersheds are summarised in the tables. Graphical presentations of results corresponding to 25 watersheds are presented in Appendix C.

7.1 Results of the Discrete Markov Model

The proposed discrete Markov model predicts the drought level based on the most probable next drought level given the current drought level using the transition matrix. The transition matrix for East Poplar River (low flow river), South Saskatchewan River (medium flow river) and Highwood River (high flow river) is shown in the below Tables 7.1, 7.2 and 7.3, respectively. The transition matrices were calculated using 70% of the data set, which is 135 seasonal patterns.

Table 7.1 Transition matrix for the East Poplar River.

		Future Drought level				
		1	2	3	4	5
Current Drought Level	1	0.457	0.514	0.014	0.000	0.014
	2	0.340	0.630	0.020	0.000	0.010
	3	0.250	0.500	0.250	0.000	0.000
	4	1.000	0.000	0.000	0.000	0.000
	5	1.000	0.000	0.000	0.000	0.000

Table 7.2 Transition matrix for the South Saskatchewan River.

Current Drought Level	Future Drought level 1	2	3	4	5
1	0.696	0.275	0.000	0.029	0.000
2	0.419	0.372	0.116	0.093	0.000
3	0.100	0.400	0.200	0.300	0.000
4	0.077	0.385	0.231	0.308	0.000
5	1.000	0.000	0.000	0.000	0.000

Table 7.3 Transition matrix for the Highwood River.

Current Drought Level	Future Drought level 1	2	3	4	5
1	0.485	0.485	0.030	0.000	0.000
2	0.380	0.360	0.200	0.060	0.000
3	0.667	0.000	0.000	0.250	0.083
4	1.000	0.000	0.000	0.000	0.000
5	1.000	0.000	0.000	0.000	0.000

The future drought level was predicted as the most probable next drought level given that the current drought level is known. Based on the transition matrix of the East Poplar River (Table 7.1), given that the current drought levels are 1, 2, 3, 4, and 5, then the most probable future drought levels are 2, 2, 2, 1, and 1, respectively. Based on the transition matrix for the South Saskatchewan River in Table 7.2, when the current drought levels are 1, 2, 3, 4, and 5, the most probable future drought levels are 1, 1, 2, 2, and 1, respectively. Accordingly, using the transition matrix of the Highwood River in Table 7.3, given that the current drought levels are 1, 2, 3, 4, and 5, then the most probable future drought levels are (1 or 2), 1, 1, 1, and 1, respectively.

As shown in the transition matrices, when the current drought level is severe or extreme, the most probable next drought level is no drought. This indicates that within a 3-month period, severe drought levels are likely to transition to no drought in the following 3 months. This pattern arises because severe drought durations in the Palliser Triangle area are typically short-term, lasting less than a month.

The model comparison for all 25 selected stations was conducted using mismatch evaluations, as described in the methodology (Chapter 4 section 4.8.3). The Mismatch evaluation measures the accuracy of models in predicting the correct drought level by calculating the percentage of correct forecasts. Model forecast results are summarized in Table 7.4, providing a detailed comparison of the performance of each model in predicting drought levels across the selected watersheds.

Table 7.4 Model 1- Mismatch accuracy of Discrete Markov forecast for 25 streamflow stations

Streamflow station name	Province	Type of the river flow	Discrete Markov Model Forecast Accuracy
Pembina River at Neche	MB	Low	61%
Souris River near Sherwood	SK	Low	49%
East Poplar River at International Boundary	SK	Low	43%
Little Saskatchewan River near Minnedosa	MB	Medium	80%
Whitemud River at Westbourne	MB	Medium	61%
Moose Jaw River near Burdick	SK	Medium	73%
North Saskatchewan River at Prince Albert	SK	Medium	63%
North Saskatchewan River near Deer Creek	SK	Medium	53%
Qu'appelle River near Lumsden	SK	Medium	56%
South Saskatchewan River at Saskatoon	SK	Medium	69%
Swift Current Creek below Rock Creek	SK	Medium	71%
Assiniboine River at Kamsack	SK	Medium	73%
Blindman River near Blackfalds	AB	Medium	51%
Battle River near Ponoka	AB	Medium	49%
Red Deer River at Drumheller	AB	Medium	75%
Medicine River near Eckville	AB	Medium	73%
Milk River at Milk River	AB	Medium	49%
Red Deer River near Bindloss	AB	Medium	71%
South Saskatchewan River at Medicine Hat	AB	Medium	61%
Crowsnest River at Frank	AB	High	55%
Bow River at Calgary	AB	High	49%
Oldman River near Lethbridge	AB	High	43%
Highwood River near the Mouth	AB	High	78%
Belly River near Mountain View	AB	High	65%
St. Mary River at International Boundary	AB	High	57%

As seen in Table 7.4 above, low-flow rivers have the lowest average accuracy of 51% in drought-level forecasts with a range varying from 43% to 61%. In contrast, the medium and high-flow rivers yield more satisfactory results, with the medium rivers averaging an accuracy of 64% with a range varying from 49% to 80%, and the large rivers averaging an accuracy of 58% with a range varying from 43% to 78% as mentioned in the following Table 7.5.

Table 7.5 Average mismatch accuracy of Discrete Markov Model

River flow type	Average accuracy	Maximum Accuracy	Minimum Accuracy
Low flow	51%	61%	43%
Medium flow	64%	80%	49%
High flow	58%	78%	43%

The rivers which yield a forecast accuracy of more than 75% are the Little Saskatchewan River, Red Deer River and Highwood River which exhibit the two highest probable drought levels in the transition matrix. This can be seen when considering the Little Saskatchewan River transition matrix shown in Table 7.6 below.

Table 7.6 Transition matrix for the Little Saskatchewan River.

		Future Drought level				
		1	2	3	4	5
Current Drought Level	1	0.483	0.417	0.100	0.000	0.000
	2	0.396	0.396	0.146	0.063	0.000
	3	0.381	0.190	0.381	0.000	0.048
	4	0.600	0.000	0.000	0.400	0.000
	5	1.000	0.000	0.000	0.000	0.000

The presence of these two highly probable drought levels increases the likelihood of accurately predicting the next drought level, resulting in improved forecast performance. Such an observation helps explain the higher accuracy observed for the Little Saskatchewan River, Red Deer River, and Highwood River. However, aside from such cases, the model does not demonstrate good accuracy across other rivers, indicating its inability to consistently predict different drought levels.

The discrete Markov model yields better results for medium flow rivers compared to low and high flow rivers due to the relatively low fluctuations in medium flow rivers. These rivers tend to exhibit consistent patterns, with the same seasonal variations repeating throughout the years. This stability allows the Markov model to effectively capture and predict the flow behavior. In contrast, low and high flow rivers experience significant variations across different seasons and years, lacking a repeating pattern, which makes it more challenging for the Markov model to accurately predict their flow behavior.

The model can also be evaluated concerning its predictive ability for different drought levels which are: No Drought (D0), Mild Drought (D1), Moderate Drought (D2), Severe Drought (D3), and Extreme Drought (D4). The following Table 7.7 provides a summary of results for all 25 watersheds.

Table 7.7 Mismatch accuracy for Discrete Markov Model forecast for different drought levels

Streamflow station name	River flow capacity (flow/unit area)	Forecast Accuracy of Drought Level				
		D0	D1	D2	D3	D4
Pembina River at Neche	Low	60%	56%	NA	NA	NA
Souris River near Sherwood	Low	55%	44%	0%	NA	NA
East Poplar River	Low	34%	52%	0%	NA	NA
Little Saskatchewan River	Medium	69%	95%	0%	NA	NA
Whitemud River at Westbourne	Medium	55%	43%	0%	NA	NA
Moose Jaw River near Burdick	Medium	50%	60%	0%	NA	NA
North Saskatchewan River at PA*	Medium	55%	50%	100%	NA	NA
North Saskatchewan River near DC*	Medium	52%	47%	75%	NA	NA
Qu'appelle River near Lumsden	Medium	40%	50%	20%	100%	NA
South Saskatchewan River at SA*	Medium	67%	60%	33%	NA	NA
Swift Current Creek	Medium	71%	33%	0%	NA	NA
Assiniboine River at Kamsack	Medium	79%	63%	0%	NA	NA
Blindman River near Blackfalds	Medium	44%	44%	50%	NA	NA
Battle River near Ponoka	Medium	39%	56%	50%	50%	0%
Red Deer River at Drumheller	Medium	63%	50%	0%	NA	NA
Medicine River near Eckville	Medium	56%	82%	0%	0%	NA
Milk River at Milk River	Medium	44%	67%	20%	NA	NA
Red Deer River near Bindloss	Medium	100%	45%	0%	NA	NA
South Saskatchewan River at MH*	Medium	33%	40%	100%	NA	NA
Crowsnest River at Frank	High	50%	29%	100%	NA	NA
Bow River at Calgary	High	50%	44%	67%	50%	0%
Oldman River near Lethbridge	High	48%	36%	0%	NA	NA
Highwood River	High	50%	50%	67%	NA	NA
Belly River	High	33%	71%	100%	NA	NA
St. Mary River	High	45%	43%	100%	NA	NA

Note: PA* = Prince Albert, DC* = Deer Creek, SA* = Saskatoon, MH* = Medicine Hat, and NA = No Available Occurrences.

The following table 7.8 shows the summary of the Average Drought Level forecasts for each drought level. This summary provides an overview of how well the model predicts different drought levels across all low-flow, medium-flow, and high-flow rivers.

Table 7.8 Average mismatch accuracy for drought level forecast - Discrete Markov Model.

River flow rate capacity per unit area	Average Mismatch Accuracy for Drought Level Forecast (%)				
	D0	D1	D2	D3	D4
Low flow Rivers	50%	51%	0%	NA	NA
Medium flow Rivers	57%	55%	28%	50%	0%
High flow Rivers	46%	46%	72%	50%	0%

A review of low-flow rivers indicates that the model struggles to accurately predict moderate drought level (D2) and the average forecast accuracy for both no drought (D0) and mild drought (D1) levels is 50% indicating lower ability in forecast. For medium-flow rivers, the model shows accuracy in predicting no drought (D0) with an average accuracy of 57% and mild drought level (D1) with an average accuracy of 55%. These results are slightly higher than low flow rivers. The moderate drought level (D2) forecast for medium-flow rivers is still not successful because the average accuracy is 28%. But for North Saskatchewan and South Saskatchewan rivers show 100% accuracy in moderate drought-level forecast (Table 7.7). Both these rivers are large rivers which show more high flow river behaviour. Also, the high-flow rivers show higher accuracy in the forecast of moderate drought levels with an average accuracy of 72% (Table 7.7). No drought(D0) and mild drought level (D1) accuracies are slightly less than medium-flow rivers with a 46% average for both drought levels.

One plausible reason for such behaviour could be the inability of the model to capture variations within the patterns. Particularly for low and medium-flow rivers where flow levels exhibit significant variations from season to season. In contrast, high-flow rivers tend to maintain more consistent flow levels throughout the year, resulting in less pronounced flow variation. Consequently, the model tends to accurately predict severe drought levels in high-flow rivers but fails to do so for extremely severe drought levels. Overall, the model does not accurately capture the intricate variations within the seasonal patterns, impacting its predictive performance across different drought levels.

The inability of the Discrete Markov Model to capture the variations in drought patterns may rise from two different reasons. First, the transition matrix only indicates the likelihood of different future drought levels based on the known current drought level. By predicting only, the highest

probable drought level, the model disregards other possible scenarios such that, while less likely, could still occur. Therefore, this approach does not account for variability in drought levels for forecasting.

The second reason is discrete Markov model categorize drought data into distinct states or levels. This discretization process simplifies the variability within the seasonal patterns into an average level, which can result-in loss of crucial information. Specifically, the model may lose valuable details about the temporal variability, such as the intensity and duration of drought conditions and local anomalies that impact forecasts. By averaging these variations into discrete categories, the model potentially be missing out on important information or features within the patterns that could enhance forecast accuracy.

To address this limitation, the Hidden Markov Model (HMM) was applied using Streamflow Drought Index values in the seasonal pattern instead of using the drought level of the seasonal pattern. Unlike the Discrete Markov Model, the HMM can directly handle Streamflow Drought Index values compared to getting an average drought level within the pattern.

The HMM does not calculate the average drought level as the state; instead, it assumes that there are hidden states associated with the streamflow, corresponding to the five different drought levels. The model attempts to map these hidden drought states to the seasonal drought patterns, allowing it to understand the transitions between each drought state. By identifying these hidden states, the HMM forecasts the seasonal drought patterns based on the Markov Property, which relies on the assumption that future states depend only on the current state, not on the sequence of events that preceded it. This approach enables the model to capture the complex dynamics within the data and make more accurate drought predictions.

7.2 Hidden Markov Model

7.2.1 Evaluation of Results of HMM obtained during the Training Phase

The proposed Hidden Markov Model was applied to each streamflow station to predict values of SDI. For each river case, three-month-length seasonal patterns were used to represent the groups. Results from the training phase and the testing phase of the model were compared to evaluate the ability of the model to train on known data to predict unknown data sets. The predicted versus actual SDI values were compared using the Mean Squared Error (MSE) and R-squared (Table 7.9). Such a comparison helps assess the accuracy and reliability of the model in predicting drought indices across different streamflow stations.

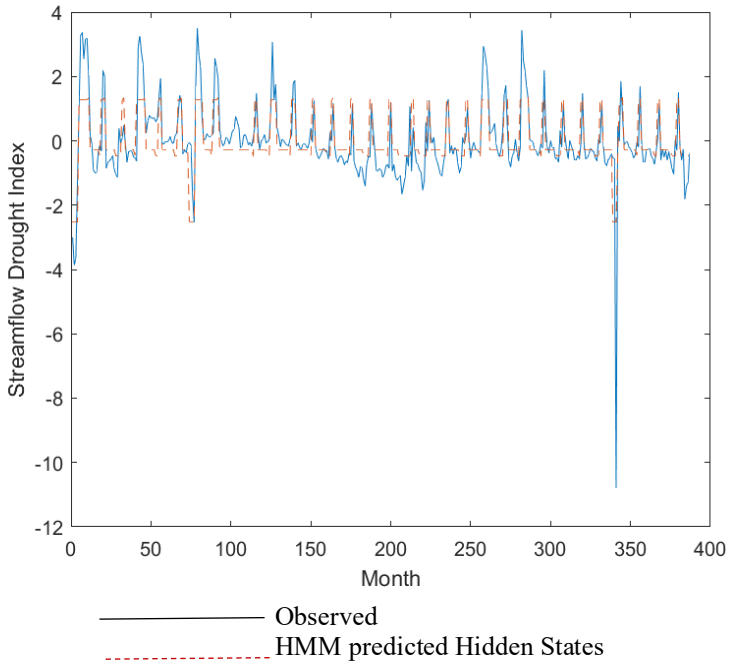


Figure 7.1 Hidden Markov Model for the East Poplar River (Training).

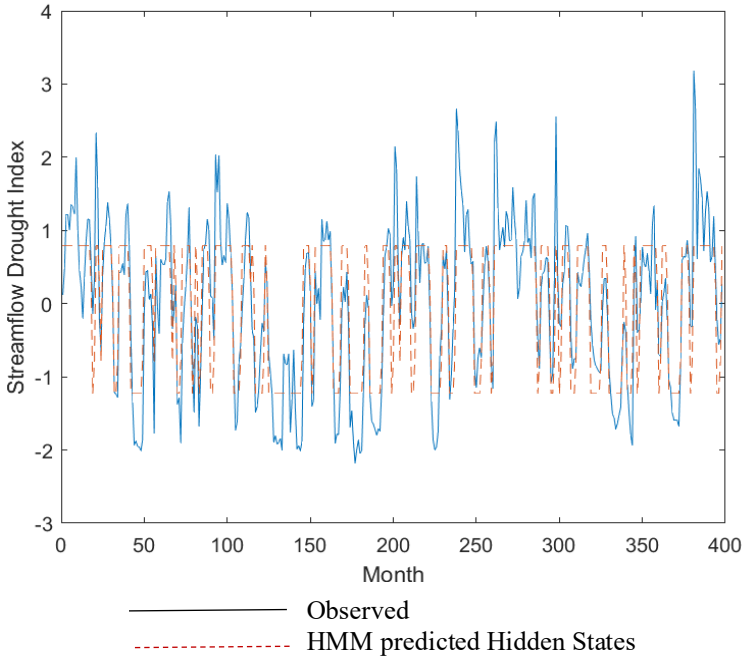


Figure 7.2 Hidden Markov Model for South Saskatchewan River (Training).

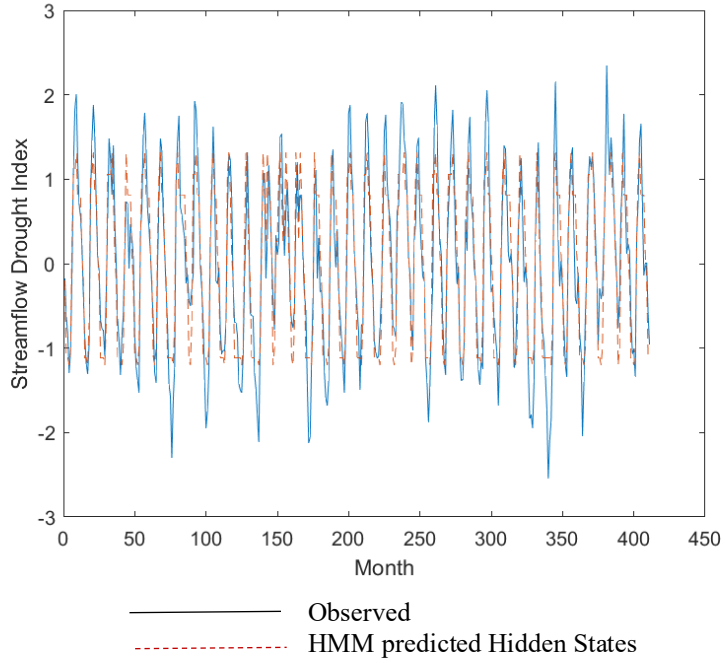


Figure 7.3 Hidden Markov Model for Highwood River (Training)

When training the Hidden Markov Model (HMM), it seems to only predict the hidden states of the dataset. This is because HMMs are specifically designed to identify the most likely sequence of hidden states that could have predicted the observed states. Essentially, HMMs focus on learning how the hidden states are changing with the observations. This means that the primary output of an HMM during the training process is the sequence of hidden states, the transition probabilities between hidden states and the emission probabilities that relate hidden states to observed data/states. This process ensures the model captures the underlying patterns in the data, which can then be used to forecast future occurrences of drought states.

The performance of the model during the training phase was evaluated using statistical measures such as MSE and R^2 for all streamflow stations. These metrics provided insights into the ability of a model to accurately predict values of SDI as summarized in Table 7.9 below.

Table 7.9 Summary statistics of the Hidden Markov Model (Training data).

Streamflow station name	River flow capacity (flow/unit area)	MSE	R^2
Pembina River at Neche	Low	0.37018	0.68132
Souris River near Sherwood	Low	0.47019	0.63534
East Poplar River at International Boundary	Low	0.44990	0.59861
Little Saskatchewan River near Minnedosa	Medium	0.37008	0.68313
Whitemud River at Westbourne	Medium	0.34753	0.70722
Moose Jaw River near Burdick	Medium	0.37079	0.68923
North Saskatchewan River at Prince Albert	Medium	0.28000	0.71506
North Saskatchewan River near Deer Creek	Medium	0.20175	0.80595
Qu'appelle River near Lumsden	Medium	0.21018	0.80760
South Saskatchewan River at Saskatoon	Medium	0.36100	0.70065
Swift Current Creek below Rock Creek	Medium	0.45348	0.63238
Assiniboine River at Kamsack	Medium	0.34618	0.70473
Blindman River near Blackfalds	Medium	0.33619	0.70833
Battle River near Ponoka	Medium	0.58980	0.49932
Red Deer River at Drumheller	Medium	0.37997	0.67684
Medicine River near Eckville	Medium	0.54581	0.54624
Milk River at Milk River	Medium	0.15827	0.85098
Red Deer River near Bindloss	Medium	0.32053	0.72452
South Saskatchewan River at Medicine Hat	Medium	0.64641	0.48941
Crowsnest River at Frank	High	0.37437	0.67468
Bow River at Calgary	High	0.16988	0.80130
Oldman River near Lethbridge	High	0.37611	0.68244
Highwood River near the Mouth	High	0.20230	0.81015
Belly River near Mountain View	High	0.29393	0.73857
St. Mary River at International Boundary	High	0.30313	0.73090

Note: PA* = Prince Albert, DC* = Deer Creek, SA* = Saskatoon, MH* = Medicine Hat, and NA = No Available Occurrences.

During the training phase, the HMM model exhibits slightly higher accuracy for large flow rivers, with an average value of 0.73 for the R^2 compared to medium flow and low flow rivers, which

have average R^2 values of 0.68 and 0.63, respectively. Additionally, the forecast accuracy of the HMM model for each drought level can be evaluated to gain a more detailed understanding of its performance. Table 7.10 provides a summary of the results for No Drought (D0), Mild Drought (D1), Moderate Drought (D2), Severe Drought (D3), and Extreme Drought (D4) levels.

Table 7.10 Mismatch accuracy of the Hidden Markov Model results for different drought level forecast (Training)

Streamflow station name	River flow capacity (flow/unit area)	Drought Level forecast				
		D0	D1	D2	D3	D4
Pembina River at Neche	Low	100%	100%	100%	100%	100%
Souris River near Sherwood	Low	84%	78%	0%	0%	67%
East Poplar River	Low	74%	91%	0%	NA	50%
Little Saskatchewan River	Medium	74%	100%	0%	0%	0%
Whitemud River at Westbourne	Medium	90%	65%	27%	0%	0%
Moose Jaw River near Burdick	Medium	85%	50%	43%	NA	0%
North Saskatchewan River at PA*	Medium	96%	57%	100%	0%	NA
North Saskatchewan River near DC*	Medium	86%	73%	5%	0%	0%
Qu'appelle River near Lumsden	Medium	92%	82%	25%	NA	NA
South Saskatchewan River at SA*	Medium	86%	49%	80%	0%	NA
Swift Current Creek	Medium	76%	71%	14%	0%	0%
Assiniboine River at Kamsack	Medium	71%	91%	14%	0%	0%
Blindman River near Blackfalds	Medium	80%	91%	42%	0%	0%
Battle River near Ponoka	Medium	81%	68%	43%	0%	0%
Red Deer River at Drumheller	Medium	94%	72%	14%	0%	0%
Medicine River near Eckville	Medium	92%	94%	0%	0%	NA
Milk River at Milk River	Medium	100%	77%	67%	0%	NA
Red Deer River near Bindloss	Medium	97%	93%	20%	0%	0%
South Saskatchewan River at MH*	Medium	74%	92%	0%	0%	0%
Crowsnest River at Frank	High	61%	100%	0%	0%	NA
Bow River at Calgary	High	85%	88%	0%	0%	NA
Oldman River near Lethbridge	High	76%	90%	75%	0%	0%
Highwood River	High	98%	69%	91%	0%	0%
Belly River	High	99%	98%	0%	0%	NA
St. Mary River	High	93%	86%	0%	0%	0%

Note: PA* = Prince Albert, DC* = Deer Creek, SA* = Saskatoon, MH* = Medicine Hat, and NA = No Available Occurrences.

As the Hidden Markov Model (HMM) tries to learn the hidden states (or the hidden drought conditions) during training, this table shows how accurately the model has identified those hidden drought levels.

Table 7.11 Average mismatch accuracy of forecast for various Drought Levels – HMM (Training)

River flow rate capacity per unit area	Average Mismatch Accuracy for Drought Level Forecast (%)				
	D0	D1	D2	D3	D4
Low flow	86%	90%	33%	50%	72%
Medium flow	86%	77%	31%	0%	0%
High flow	85%	89%	28%	0%	0%

From a review of Table 7.11, it is apparent that for low-flow rivers, the HMM is capable of predicting no drought (D0), mild drought (D1), moderate drought (D2), severe drought (D3), and extreme drought (D4) levels with an average accuracy of 86%, 89% 33%, 50% and 72% respectively. However, for medium-flow rivers, the model can only predict no drought (D0), mild drought (D1), and moderate drought (D2) with lower accuracies of 86%, 77%, and 31%, respectively. It fails to identify severe (D3) and extreme drought (D4) levels in medium flow as well as high flow rivers. For high-flow rivers, the model predicts no drought (D0) and mild drought (D1) levels with an average accuracy of 85% and 89%, respectively. However, the accuracy drops to 28% for moderate drought (D2), and the model fails to identify severe drought (D3) and extreme drought (D4) levels, with 0% accuracy for both.

As discussed above, low-flow rivers experience the greatest variations in their flow levels from season to season, while medium-flow rivers exhibit less variation. In contrast, high-flow rivers generally maintain more consistent flow levels throughout the year, leading to less pronounced variations. The hidden state prediction results indicate that the model struggles to capture higher drought levels when the data has lower variations. This means the model is better at identifying higher drought levels in low-flow rivers compared to high-flow rivers, making the training phase of the model more reliable for low-flow rivers. However, even in low-flow rivers, the model faces challenges in capturing more complex drought patterns, such as extreme or severe droughts. Then to evaluate the performance of the trained model on the testing data, the trained model was applied to the testing data and the results are discussed below.

7.2.2 Evaluation of Results of the HMM obtained during the Testing Phase

The testing results for the selected low-, medium-, and high-flow rivers are presented in figures and summarized in tables below. For the East Poplar River (low flow river), South Saskatchewan River (medium flow river), and Highwood River (high flow river), Figures 7.4, 7.5 and 7.6 illustrate a visual representation of the performance concerning the predicted versus observed drought levels. The summary Table 7.12 lists MSE and R^2 values for each river, offering a quantitative assessment of the accuracy of the model.

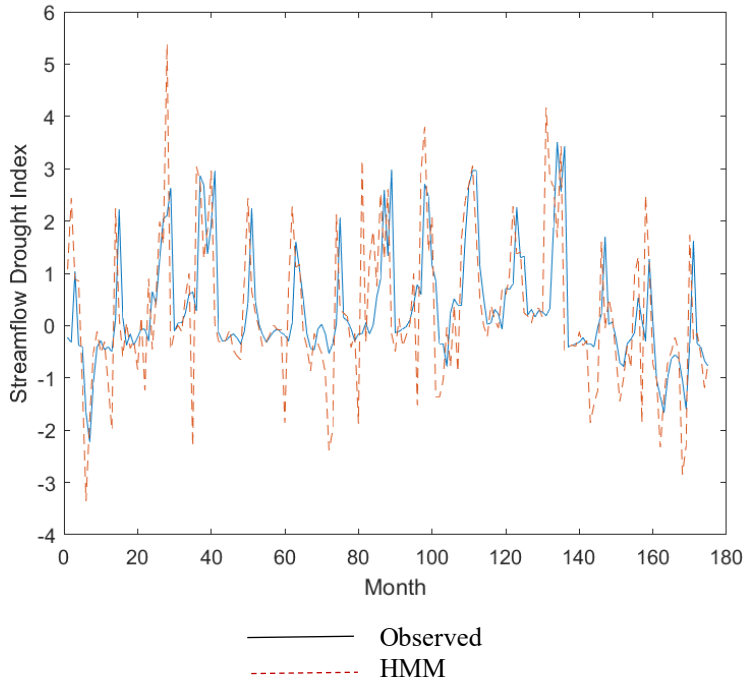


Figure 7.4 Hidden Markov Model for East Poplar River (Testing)

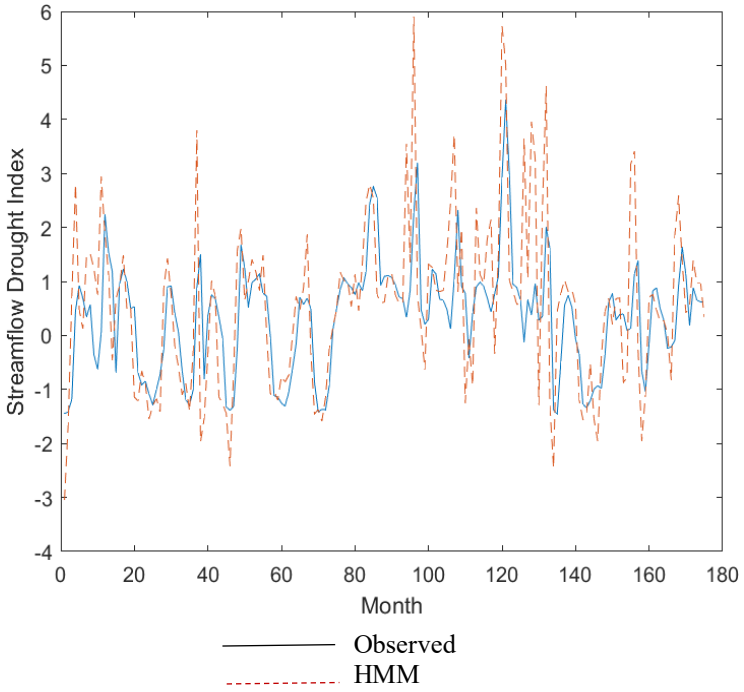


Figure 7.5 Hidden Markov Model for South Saskatchewan River (Testing)

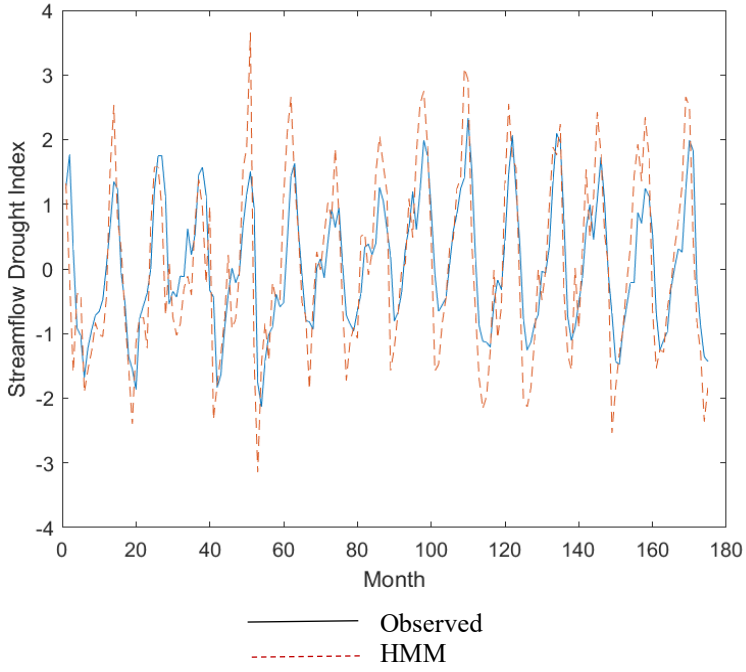


Figure 7.6 Hidden Markov Model for Highwood River (Testing)

As expected, the test results appear to be poorer than the training results, indicating the limitations of the model to forecast new unseen data. However, the forecast results tend to be consistently

higher or lower than the actual results, which still makes it a useful model for drought forecasting. Although it doesn't perfectly match real-world data, the model tends to forecast more extreme drought patterns for dry conditions and more extreme no-drought patterns for non-drought conditions. This tendency can be advantageous in designing water management structures, as it provides a safety margin, ensuring that these structures are better prepared for potential extreme conditions. This characteristic suggests that while the model may not perfectly match the observed data, it can reliably indicate the general trend and severity of droughts. The statistical summary of the HMM model is provided in Table 7.12 below:

Table 7.12 Summary statistics for Hidden Markov Model (Testing data)

Streamflow station name	River flow capacity (flow/unit area)	MSE	R^2
Pembina River at Neche	Low	0.58410	0.53574
Souris River near Sherwood	Low	0.56701	0.38350
East Poplar River at International Boundary	Low	1.08055	0.46469
Little Saskatchewan River near Minnedosa	Medium	1.06400	0.45119
Whitemud River at Westbourne	Medium	1.06050	0.36771
Moose Jaw River near Burdick	Medium	1.04311	0.36791
North Saskatchewan River at Prince Albert	Medium	0.58646	0.60247
North Saskatchewan River near Deer Creek	Medium	0.56785	0.68940
Qu'appelle River near Lumsden	Medium	1.42901	0.45508
South Saskatchewan River at Saskatoon	Medium	1.45755	0.40349
Swift Current Creek below Rock Creek	Medium	1.84620	0.26826
Assiniboine River at Kamsack	Medium	1.02980	0.51596
Blindman River near Blackfalds	Medium	1.25580	0.32295
Battle River near Ponoka	Medium	1.60967	0.42813
Red Deer River at Drumheller	Medium	1.18937	0.38614
Medicine River near Eckville	Medium	1.33911	0.35929
Milk River at Milk River	Medium	1.22800	0.41387
Red Deer River near Bindloss	Medium	1.16561	0.48063
South Saskatchewan River at Medicine Hat	Medium	1.42110	0.30933
Crowsnest River at Frank	High	0.63944	0.60345
Bow River at Calgary	High	0.58632	0.64331
Oldman River near Lethbridge	High	0.75275	0.52027
Highwood River near the Mouth	High	0.65028	0.66733
Belly River near Mountain View	High	0.83696	0.59489
St. Mary River at International Boundary	High	1.04066	0.44371

Note: PA* = Prince Albert, DC* = Deer Creek, SA* = Saskatoon, MH* = Medicine Hat, and NA = No Available Occurrences.

As seen in Table 7.12, the MSE and R^2 values indicate that the Hidden Markov Model (HMM) predicts better results for both low- and high-flow rivers compared to medium-flow rivers. This suggests that the model performs better with highly variable data. As discussed earlier, medium-flow rivers exhibit the lowest variability within their data, maintaining a more constant pattern throughout the seasons. During the testing phase, the model attempts to identify the hidden state, or drought level, within the seasonal patterns and predicts the seasonal pattern based on the highest drought level within that hidden state. This approach explains why the forecast results of the model tend to be either significantly higher or lower than the actual values, as it emphasizes extreme patterns over more moderate variations. However, when evaluating drought level forecasts for each watershed as shown in below Table 7.13, the model exhibits notable discrepancies.

Table 7.13 Hidden Markov Model forecast results for different drought levels (Testing).

Streamflow station name	River flow capacity (flow/unit area)	Drought Level forecast accuracy (%)				
		D0	D1	D2	D3	D4
Pembina River at Neche	Low	84%	77%	NA	NA	NA
Souris River near Sherwood	Low	89%	42%	100%	NA	NA
East Poplar River	Low	82%	65%	0%	NA	NA
Little Saskatchewan River	Medium	86%	50%	NA	NA	NA
Whitemud River at Westbourne	Medium	70%	47%	0%	NA	NA
Moose Jaw River near Burdick	Medium	76%	35%	50%	50%	NA
North Saskatchewan River at PA*	Medium	79%	63%	100%	NA	NA
North Saskatchewan River near DC*	Medium	79%	32%	33%	100%	NA
Qu'appelle River near Lumsden	Medium	88%	57%	75%	0%	NA
South Saskatchewan River at SA*	Medium	98%	33%	67%	NA	NA
Swift Current Creek	Medium	91%	67%	NA	NA	NA
Assiniboine River at Kamsack	Medium	80%	58%	100%	NA	NA
Blindman River near Blackfalds	Medium	76%	79%	50%	NA	NA
Battle River near Ponoka	Medium	76%	71%	44%	33%	100%
Red Deer River at Drumheller	Medium	76%	62%	NA	NA	NA
Medicine River near Eckville	Medium	73%	56%	100%	NA	NA
Milk River at Milk River	Medium	82%	32%	38%	100%	NA
Red Deer River near Bindloss	Medium	77%	65%	50%	NA	NA
South Saskatchewan River at MH*	Medium	79%	52%	0%	NA	NA
Crowsnest River at Frank	High	88%	83%	40%	NA	NA
Bow River at Calgary	High	86%	68%	24%	NA	100%
Oldman River near Lethbridge	High	87%	78%	0%	NA	NA
Highwood River	High	85%	57%	25%	NA	NA
Belly River	High	78%	59%	14%	NA	NA
St. Mary River	High	81%	65%	14%	NA	NA

Note: PA* = Prince Albert, DC* = Deer Creek, SA* = Saskatoon, MH* = Medicine Hat, and NA = No Available Occurrences.

When considering the average forecast results for low, medium and high-flow rivers as seen in the Table 7.14, it seems that all three types of rivers have similar type average results.

Table 7.14 Average mismatch accuracy for Drought Level forecast – HMM(Testing)

River flow rate capacity per unit area	Average Mismatch Accuracy for Drought Level Forecast (%)				
	D0	D1	D2	D3	D4
Low flow	85%	61%	50%	NA	NA
Medium flow	80%	54%	54%	57%	100%
High flow	85%	68%	20%	NA	100%

This average result shows that the model achieves over 80% accuracy in predicting no drought (D0) levels and 100% accuracy for extreme severe drought (D4) forecasts for all low, medium and high-flow rivers. This is due to the tendency of the model to predict values that are either significantly higher or lower than the observed data. While the model excels in predicting extreme scenarios, its accuracy reduces for mild to severe drought levels, as it tends to overestimate drought conditions in these categories. This overestimation results in lower accuracy for the less extreme drought levels, highlighting a limitation in the ability of the model to accurately capture and differentiate between varying degrees of drought severity.

This limitation could occur because the model attempts to interpret the hidden states of seasonal patterns and adjusts its forecasts based on the maximum values associated with each hidden state. As the model tries to map the observed data to these underlying states, it may lean toward predicting higher or lower drought levels, depending on the extremes it identifies within each hidden state. This tendency to align forecasts with the maximum values of the hidden states can lead to higher accuracy for extreme scenarios but lower accuracy for mild to severe drought levels, where the model might overestimate the drought severity.

Table 7.15 Comparison between average mismatch accuracy for Drought Level forecast for Discrete Markov Model and Hidden Markov Model

Model	River flow rate capacity per unit area	Average Mismatch Accuracy for Drought Level Forecast (%)				
		D0	D1	D2	D3	D4
Discrete Markov Model	Low flow	86%	90%	33%	50%	72%
	Medium flow	86%	77%	31%	0%	0%
	High flow	85%	89%	28%	0%	0%
Hidden Markov Model	Low flow	85%	61%	50%	NA	NA
	Medium flow	80%	54%	54%	57%	100%
	High flow	85%	68%	20%	NA	100%

When comparing the average accuracy for drought level forecasting between the Discrete Markov Model and Hidden Markov Model as in Table 7.15, it is observed that the forecast accuracy for no drought levels is the same for both models. However, the accuracy for mild drought levels has decreased with the Hidden Markov Model, while the forecast accuracy for more severe drought levels has increased. This indicates that the HMM can overcome the limitations of the Discrete Markov model by capturing more complex and less frequent extreme severe drought levels more effectively. The ability of the HMM to handle these severe conditions better highlights its advantage in modeling and predicting extreme drought scenarios.

However, the Hidden Markov Model (HMM) struggles to capture less severe drought conditions accurately, as it tends to overestimate extreme values. This issue arises because the model forecasts based on the extreme values of the hidden states, which can lead to exaggerated predictions for severe drought conditions.

The LSTM model can address the limitations of the Hidden Markov Model (HMM) by offering a more nuanced approach to forecasting drought conditions. Unlike the HMM, which relies on hidden states and tends to overestimate extreme values, LSTM networks can capture complex temporal dependencies and variations in the data over time. By learning from a sequence of past observations, LSTM models can better account effect from both extreme and mild drought conditions happened in the past, providing more balanced and accurate forecasts. This helps the LSTM model avoid the tendency to overestimate drought severity and provides more balanced and accurate forecasts across all drought levels.

7.3 Long Short-Term Memory Model for the Monthly Data

7.3.1 Evaluation of Results of LSTM obtained during the Training Phase

The proposed Long Short-Term Memory Model was applied to each streamflow station to predict the values of SDI. For each case, three-month-length seasonal patterns were used to represent the groups. Results from the training and testing phases of the model were compared to evaluate the ability of the model to train on known data to generate unknown data sets as shown in the following figures. The predicted versus actual values of SDI were compared using MSE and R^2 (Table 7.16). This comparison helps assess the accuracy and reliability of the model in predicting drought indices across different streamflow stations.

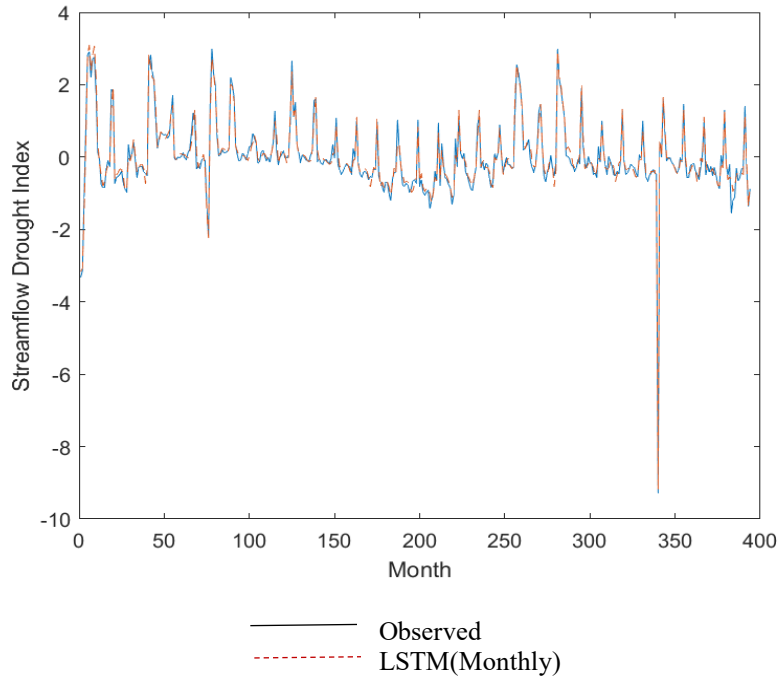


Figure 7.7 Long Short-Term Memory Model for monthly data in East Poplar River (Training).

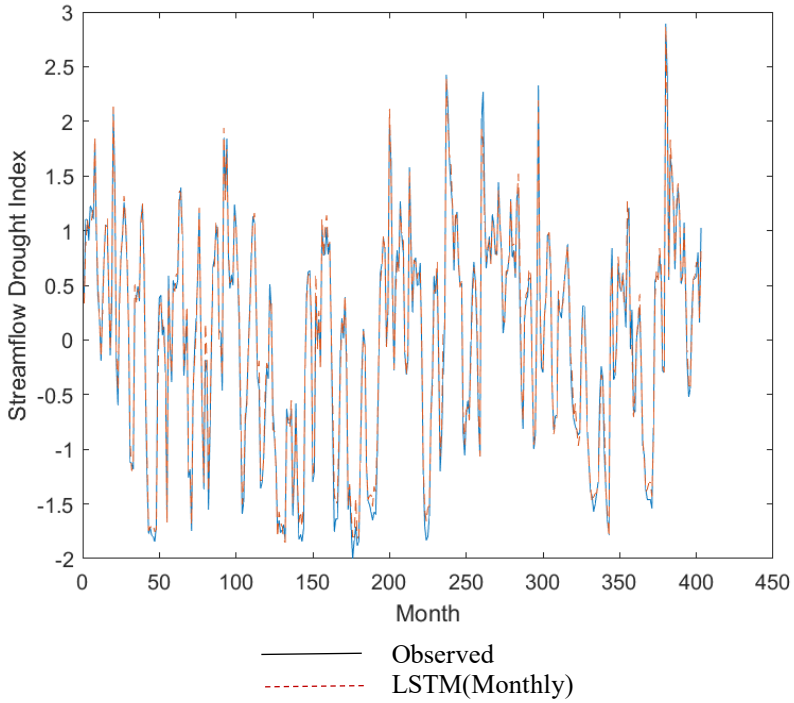


Figure 7.8 Long Short-Term Memory Model for monthly data in South Saskatchewan River (Training).

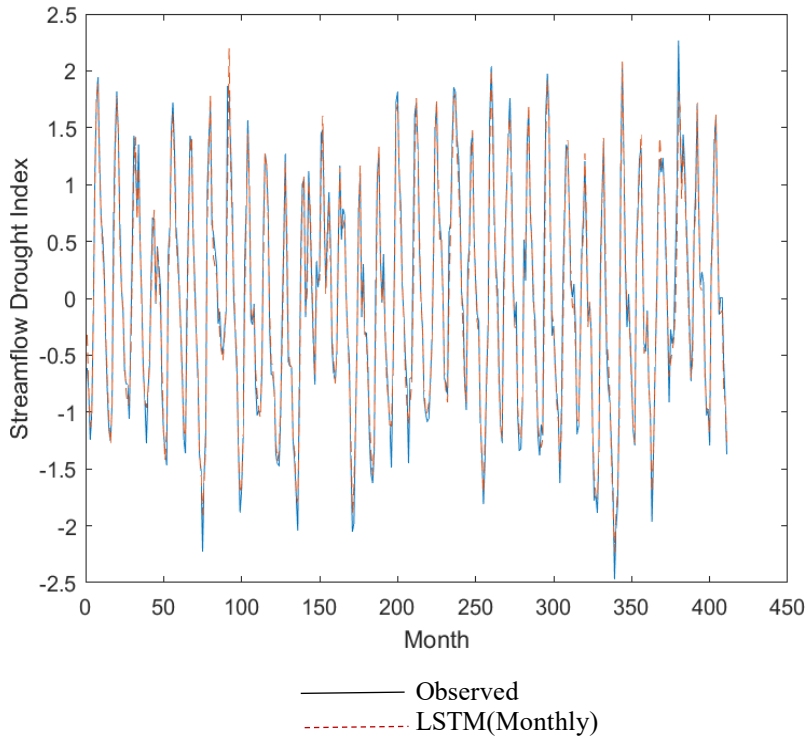


Figure 7.9 Long Short-Term Memory Model for monthly data in Highwood River (Training).

The LSTM model training results closely match the observed patterns, demonstrating the ability of the model to capture both linear and nonlinear patterns within the data. It can effectively capture both long-term and short-term dependencies or patterns, making it a reliable tool for predicting and analyzing seasonal patterns and drought levels in streamflow data. Table 7.16 summarizes the statistical results of the LSTM model for monthly data.

Table 7.16 Summary statistics for Long Short-Term Memory Model for monthly data (Training)

Streamflow station name	River flow capacity (flow/unit area)	MSE	R^2
Pembina River at Neche	Low	0.01489	0.98677
Souris River near Sherwood	Low	0.01610	0.98540
East Poplar River at International Boundary	Low	0.02815	0.98340
Little Saskatchewan River near Minnedosa	Medium	0.01161	0.98971
Whitemud River at Westbourne	Medium	0.00954	0.99065
Moose Jaw River near Burdick	Medium	0.01141	0.98874
North Saskatchewan River at Prince Albert	Medium	0.01448	0.98572
North Saskatchewan River near Deer Creek	Medium	0.01368	0.98674
Qu'appelle River near Lumsden	Medium	0.01848	0.98155
South Saskatchewan River at Saskatoon	Medium	0.01072	0.99055
Swift Current Creek below Rock Creek	Medium	0.01262	0.98806
Assiniboine River at Kamsack	Medium	0.00851	0.99189
Blindman River near Blackfalds	Medium	0.01161	0.98937
Battle River near Ponoka	Medium	0.01237	0.98772
Red Deer River at Drumheller	Medium	0.01460	0.98604
Medicine River near Eckville	Medium	0.01516	0.98652
Milk River at Milk River	Medium	0.01738	0.98401
Red Deer River near Bindloss	Medium	0.01485	0.9859
South Saskatchewan River at Medicine Hat	Medium	0.01100	0.98943
Crowsnest River at Frank	High	0.00896	0.99145
Bow River at Calgary	High	0.01150	0.99000
Oldman River near Lethbridge	High	0.01188	0.99015
Highwood River near the Mouth	High	0.02204	0.98006
Belly River near Mountain View	High	0.01156	0.98931
St. Mary River at International Boundary	High	0.01566	0.98516

The LSTM model demonstrates exceptional performance with an average R^2 value of 0.99 across low, medium, and high-flow rivers, indicating high accuracy in forecasts during training. This suggests the model effectively captures the patterns in the data and provides reliable forecasts. Furthermore, Table 7.17 shows how the training phase of the LSTM model can forecast the drought levels.

Table 7.17 Mismatch accuracy for monthly LSTM Model forecast results for different drought levels (Training)

Streamflow station name	River flow capacity (flow/unit area)	Drought Level Forecast Accuracy (%)				
		D0	D1	D2	D3	D4
Pembina River at Neche	Low	92%	98%	56%	100%	100%
Souris River near Sherwood	Low	93%	95%	88%	75%	100%
East Poplar River	Low	95%	95%	100%	NA	100%
Little Saskatchewan River	Medium	93%	98%	85%	80%	NA
Whitemud River at Westbourne	Medium	93%	98%	100%	67%	NA
Moose Jaw River near Burdick	Medium	93%	91%	88%	100%	100%
North Saskatchewan River at PA*	Medium	89%	98%	85%	75%	NA
North Saskatchewan River near DC*	Medium	94%	98%	92%	75%	NA
Qu'appelle River near Lumsden	Medium	92%	93%	84%	NA	NA
South Saskatchewan River at SA*	Medium	100%	97%	86%	89%	NA
Swift Current Creek	Medium	90%	98%	75%	100%	100%
Assiniboine River at Kamsack	Medium	94%	94%	78%	100%	100%
Blindman River near Blackfalds	Medium	91%	98%	89%	100%	100%
Battle River near Ponoka	Medium	94%	98%	91%	100%	NA
Red Deer River at Drumheller	Medium	93%	92%	89%	60%	NA
Medicine River near Eckville	Medium	92%	94%	93%	50%	NA
Milk River at Milk River	Medium	86%	97%	70%	75%	NA
Red Deer River near Bindloss	Medium	91%	98%	67%	60%	NA
South Saskatchewan River at MH*	Medium	93%	98%	100%	100%	50%
Crowsnest River at Frank	High	88%	96%	100%	100%	NA
Bow River at Calgary	High	88%	98%	82%	100%	100%
Oldman River near Lethbridge	High	93%	98%	80%	50%	NA
Highwood River	High	91%	98%	100%	75%	NA
Belly River	High	89%	98%	85%	100%	NA
St. Mary River	High	94%	98%	88%	75%	NA

Note: PA* = Prince Albert, DC* = Deer Creek, SA* = Saskatoon, MH* = Medicine Hat, and NA = No Available Occurrences.

The forecast accuracies for each drought level are very high with a minimum forecast accuracy of 50%, as shown in Table 7.17. When considering the average forecast accuracy for each drought level, it is evident that the model can predict each drought level with even greater accuracy.

Table 7.18 Average mismatch accuracy of
Drought Level forecast– Monthly LSTM (Training)

River flow rate capacity per unit area	Average Mismatch Accuracy of of Drought Level Forecast (%)				
	D0	D1	D2	D3	D4
Low flow	93%	96%	81%	88%	100%
Medium flow	92%	96%	86%	82%	90%
High flow	91%	98%	89%	83%	100%

The average forecast accuracies for all drought levels are higher than 80% (Table 7.18), indicating that the model performs well in identifying different drought conditions and provides reliable forecasts across various levels for all low, medium and high-flow rivers. The trained model was then tested to evaluate its performance on the test data.

7.3.2 Evaluation of Results of LSTM obtained during the Test Phase

The test results for the selected low-flow, medium, and high-flow rivers are presented in Figures 7.10 to 7.12 and summarized in Table 7.19 below. For the East Poplar River (low flow river), South Saskatchewan River (medium flow river), and Highwood River (high flow river), figures 7.10 to 7.12 illustrate the predicted versus actual drought levels, providing a visual representation of the model's performance. The summary Table 7.19 lists the MSE and R^2 values for each river, offering a quantitative assessment of the accuracy of the model.

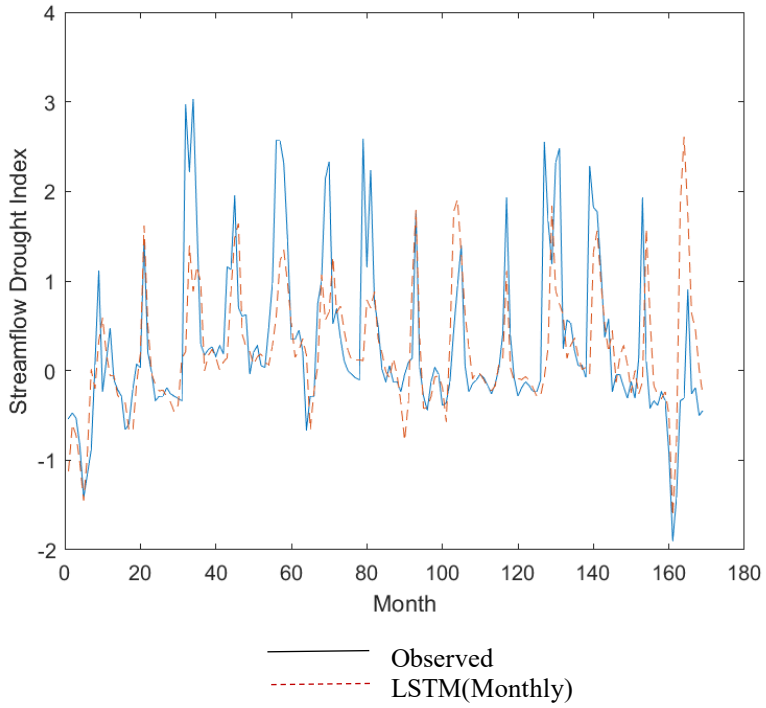


Figure 7.10 Long Short-Term Memory Model for monthly data in East Poplar River (Testing).

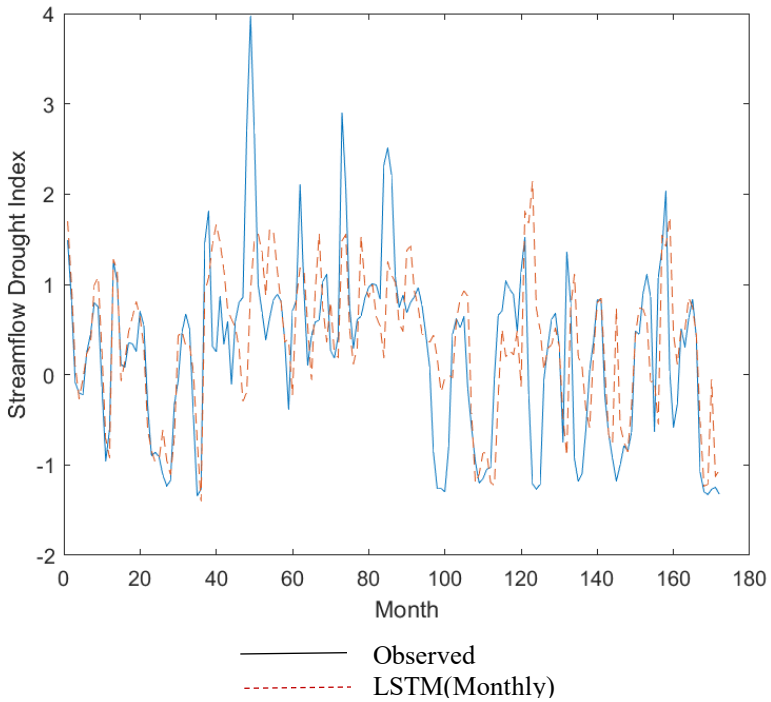


Figure 7.11 Long Short-Term Memory Model for monthly data in South Saskatchewan River (Testing).

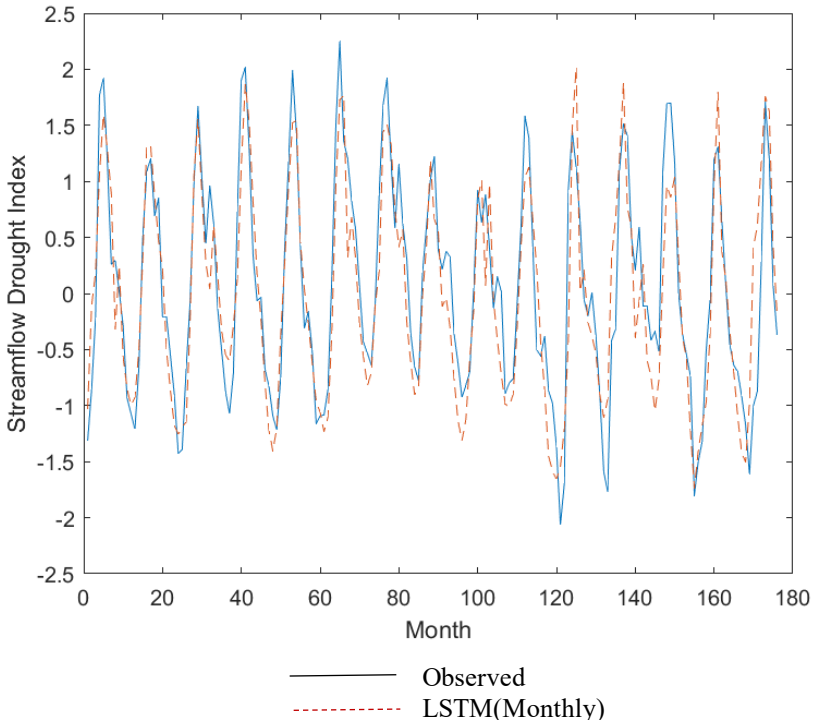


Figure 7.12 Long Short-Term Memory Model for monthly data in Highwood River (Testing)

Table 7.19 displays the statistical measurements of forecasts for the LSTM model using monthly data for the testing phase.

Table 7.19 Summary statistics for Long Short-Term Memory Model for monthly data (Testing)

Streamflow station name	River flow capacity (flow/unit area)	MSE	R^2
Pembina River at Neche	Low	0.33068	0.52056
Souris River near Sherwood	Low	0.61005	0.33877
East Poplar River at International Boundary	Low	0.56557	0.34090
Little Saskatchewan River near Minnedosa	Medium	0.31732	0.66202
Whitemud River at Westbourne	Medium	0.40658	0.57703
Moose Jaw River near Burdick	Medium	0.80325	0.25159
North Saskatchewan River at Prince Albert	Medium	0.22971	0.75472
North Saskatchewan River near Deer Creek	Medium	0.33625	0.66278
Qu'appelle River near Lumsden	Medium	0.59509	0.49162
South Saskatchewan River at Saskatoon	Medium	0.65540	0.36081
Swift Current Creek below Rock Creek	Medium	0.70714	0.29517
Assiniboine River at Kamsack	Medium	0.24932	0.74080
Blindman River near Blackfalds	Medium	0.46340	0.43569
Battle River near Ponoka	Medium	0.61737	0.49738
Red Deer River at Drumheller	Medium	0.30736	0.60733
Medicine River near Eckville	Medium	0.49164	0.47552
Milk River at Milk River	Medium	0.24037	0.73999
Red Deer River near Bindloss	Medium	0.33139	0.59670
South Saskatchewan River at Medicine Hat	Medium	0.56914	0.32083
Crowsnest River at Frank	High	0.21495	0.78525
Bow River at Calgary	High	0.32765	0.72430
Oldman River near Lethbridge	High	0.34178	0.47775
Highwood River near the Mouth	High	0.20009	0.79442
Belly River near Mountain View	High	0.27669	0.76162
St. Mary River at International Boundary	High	0.31139	0.67487

Even though the LSTM model resulted in highly accurate forecasts with an average of 0.99 (R^2) for the training data, the testing data forecast results are much lower with an average of 0.56 (R^2) measurement. As observed, the LSTM model performs slightly better for high-flow and few medium-flow rivers such as the North and South Saskatchewan rivers, which are medium but have large basins and show more high-flow river characteristics. However, the model faces difficulties in capturing the high variability of the low-flow rivers. Following Table 7.20 shows how the model performed in predicting each drought level for 25 watersheds.

Table 7.20 Mismatch accuracy for monthly LSTM Model forecast results for different drought levels (Testing)

Streamflow station name	River flow capacity (flow/unit area)	Drought Level (Forecast Average (%))				
		D0	D1	D2	D3	D4
Pembina River at Neche	Low	87%	67%	NA	NA	NA
Souris River near Sherwood	Low	83%	45%	0%	NA	NA
East Poplar River	Low	94%	67%	50%	NA	NA
Little Saskatchewan River	Medium	95%	86%	NA	NA	NA
Whitemud River at Westbourne	Medium	85%	65%	0%	NA	NA
Moose Jaw River near Burdick	Medium	87%	67%	0%	NA	NA
North Saskatchewan River at PA*	Medium	97%	91%	50%	NA	NA
North Saskatchewan River near DC*	Medium	91%	75%	100%	NA	NA
Qu'appelle River near Lumsden	Medium	73%	73%	33%	0%	NA
South Saskatchewan River at SA*	Medium	92%	80%	17%	NA	NA
Swift Current Creek	Medium	87%	57%	NA	NA	NA
Assiniboine River at Kamsack	Medium	91%	67%	50%	NA	NA
Blindman River near Blackfalds	Medium	79%	72%	0%	NA	NA
Battle River near Ponoka	Medium	68%	78%	38%	0%	0%
Red Deer River at Drumheller	Medium	88%	76%	NA	NA	NA
Medicine River near Eckville	Medium	92%	94%	93%	50%	NA
Milk River at Milk River	Medium	90%	52%	63%	NA	NA
Red Deer River near Bindloss	Medium	87%	68%	NA	NA	NA
South Saskatchewan River at MH*	Medium	85%	80%	0%	NA	NA
Crowsnest River at Frank	High	85%	71%	50%	NA	NA
Bow River at Calgary	High	92%	84%	29%	0%	0%
Oldman River near Lethbridge	High	77%	79%	NA	NA	NA
Highwood River	High	85%	72%	33%	NA	NA
Belly River	High	83%	70%	33%	NA	NA
St. Mary River	High	81%	78%	40%	NA	NA

Note: PA* = Prince Albert, DC* = Deer Creek, SA* = Saskatoon, MH* = Medicine Hat, and NA = No Available Occurrences.

The following Table 7.21 shows the average drought level forecasts for the testing phase of the LSTM model.

Table 7.21 Average Drought Level forecast – Monthly LSTM (Testing)

River flow rate capacity per unit area	Average Mismatch Accuracy for Drought Level Forecast (%)				
	D0	D1	D2	D3	D4
Low flow	88%	60%	25%	NA	NA
Medium flow	87%	74%	37%	17%	0%
High flow	84%	76%	37%	0%	0%

When considering the average drought level forecast, the LSTM model performs well in accurately predicting no drought (D0) levels. However, its accuracy decreases for higher severity drought levels such as moderate drought (D2), severe drought (D3) and extreme severe drought (D4) across all three types of rivers. This suggests that while the model is effective at identifying the absence of drought, it faces challenges in accurately predicting more severe drought conditions.

Even though the model shows high accuracy during the training phase, it does not perform as well on the testing data, particularly struggling to capture more severe drought levels. This is a sign of overfitting of the model. Overfitting occurs when the model becomes too complex and learns the specific details and noise within the training data, rather than capturing the underlying patterns that are generalizable to new, unseen data. This can happen due to several reasons, such as an complexity of the data series or insufficient training data. As a result, the model is unable to accurately identify the less frequently occurring, more complex patterns, such as severe and extreme-severe drought levels, due to insufficient data. The limited dataset may not provide enough examples for the model to learn all the patterns effectively.

Transitioning from monthly to weekly data could help address this issue by offering a more detailed and refined view of temporal variations within the hydrological system, potentially allowing the model to capture these complex patterns more accurately. Weekly data may allow the model to detect short-term fluctuations and trends that monthly data might miss, potentially improving its ability to predict extreme drought conditions. Therefore, the same LSTM model was trained using the weekly data. The predicted weekly data was converted to monthly data by aggregating the values of every four weeks to represent one month.

7.4 Long Short-Term Memory Model for Weekly data

7.4.1 Evaluation of Results of LSTM obtained during the Training Phase

The proposed Long Short-Term Memory Model was applied to each streamflow station to predict the value of SDI. Therefore, instead of three-month-long seasonal patterns, 13-week seasonal patterns were used to represent the groups. This approach allows for a finer resolution in capturing the temporal dependencies and variations within the streamflow data, potentially leading to more accurate forecasts of drought levels. The predicted versus actual SDI values were compared using MSE and R^2, as presented in Table 7.22. This comparison helps assess the accuracy and reliability of the model in predicting drought indices across different streamflow stations.

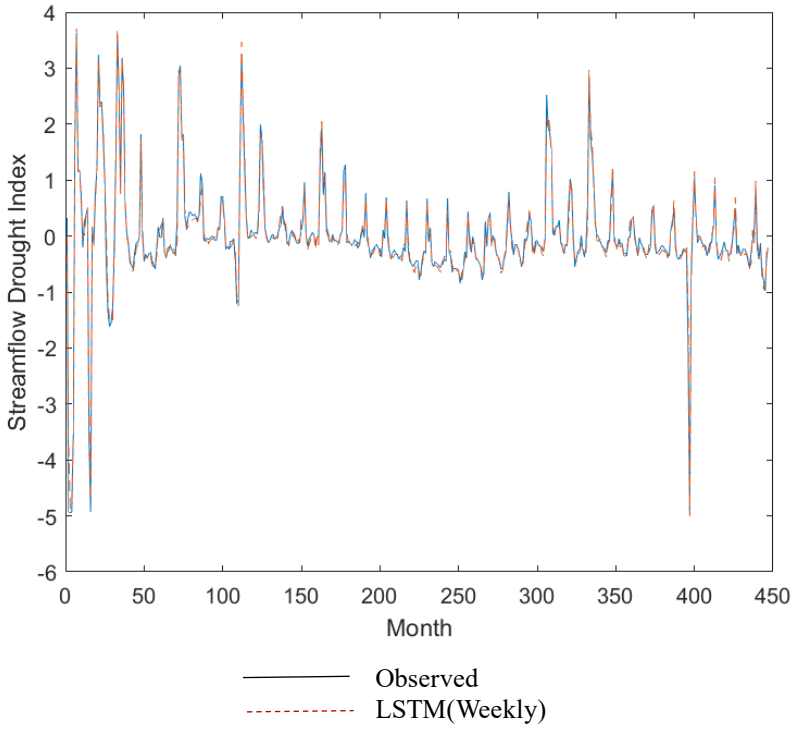


Figure 7.13 Long Short-Term Memory Model for weekly data in East Poplar River (Training).

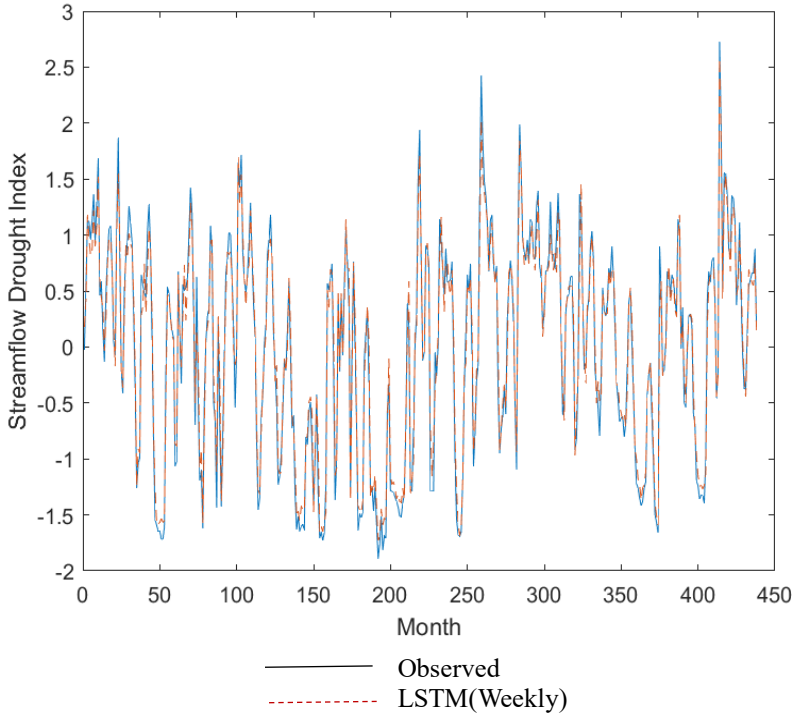


Figure 7.14 Long Short-Term Memory Model for weekly data in South Saskatchewan River (Training).

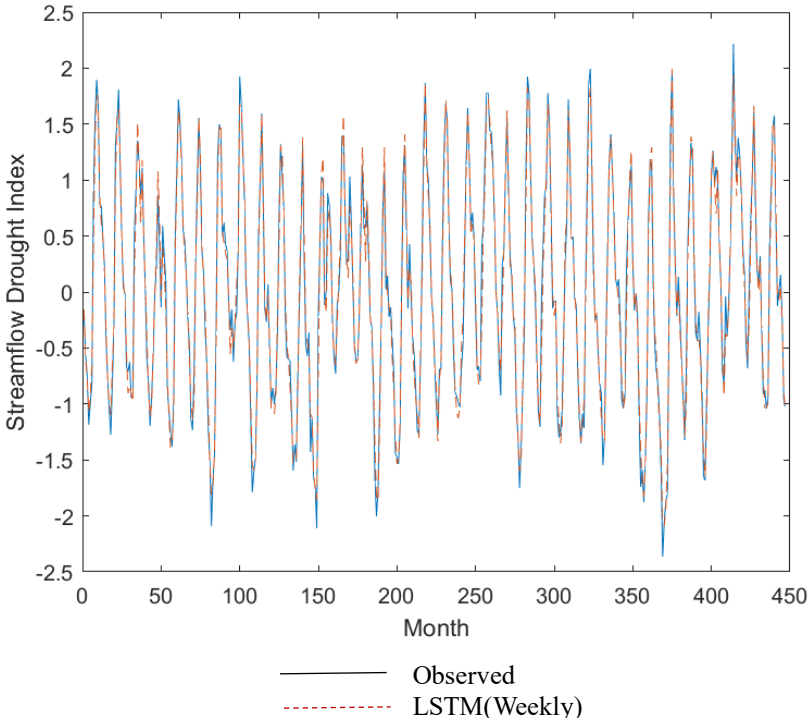


Figure 7.15 Long Short-Term Memory Model for weekly data in Highwood River (Training).

Table 7.22 Summary statistics for Long Short-Term Memory Model for weekly data (Training).

Streamflow station name	River flow capacity (flow/unit area)	MSE	R^2
Pembina River at Neche	Low	0.0010	0.99943
Souris River near Sherwood	Low	0.02067	0.97900
East Poplar River at International Boundary	Low	0.01725	0.98041
Little Saskatchewan River near Minnedosa	Medium	0.01106	0.98863
Whitemud River at Westbourne	Medium	0.01930	0.97959
Moose Jaw River near Burdick	Medium	0.02863	0.97158
North Saskatchewan River at Prince Albert	Medium	0.02364	0.97424
North Saskatchewan River near Deer Creek	Medium	0.02685	0.96977
Qu'appelle River near Lumsden	Medium	0.02281	0.97740
South Saskatchewan River at Saskatoon	Medium	0.02738	0.97370
Swift Current Creek below Rock Creek	Medium	0.03832	0.95619
Assiniboine River at Kamsack	Medium	0.00627	0.99890
Blindman River near Blackfalds	Medium	0.01984	0.97724
Battle River near Ponoka	Medium	0.01290	0.98575
Red Deer River at Drumheller	Medium	0.02109	0.97954
Medicine River near Eckville	Medium	0.03752	0.96277
Milk River at Milk River	Medium	0.01781	0.98155
Red Deer River near Bindloss	Medium	0.02203	0.98352
South Saskatchewan River at Medicine Hat	Medium	0.03026	0.96908
Crowsnest River at Frank	High	0.01653	0.98373
Bow River at Calgary	High	0.02003	0.98015
Oldman River near Lethbridge	High	0.01638	0.98235
Highwood River near the Mouth	High	0.01930	0.97959
Belly River near Mountain View	High	0.02050	0.97769
St. Mary River at International Boundary	High	0.01978	0.98052

The training results using weekly data show an average R^2 of 0.98, which is slightly lower than the R^2 of 0.987 achieved with the monthly data. However, this difference is minimal and can be considered almost similar. When considering the forecast accuracy for each drought level, the

weekly data results align closely with the monthly data, maintaining high accuracy in drought level forecast.

Table 7.23 Mismatch accuracy of weekly LSTM Model forecast results for different drought levels (Training)

Streamflow station name	River flow capacity (flow/unit area)	Drought Level Forecast Average (%)				
		D0	D1	D2	D3	D4
Pembina River at Neche	Low	92%	98%	76%	100%	100%
Souris River near Sherwood	Low	93%	95%	91%	75%	100%
East Poplar River	Low	95%	95%	100%	NA	100%
Little Saskatchewan River	Medium	93%	98%	85%	80%	NA
Whitemud River at Westbourne	Medium	93%	98%	100%	67%	NA
Moose Jaw River near Burdick	Medium	93%	91%	88%	100%	100%
North Saskatchewan River at PA*	Medium	89%	98%	85%	75%	NA
North Saskatchewan River near DC*	Medium	94%	98%	92%	75%	NA
Qu'appelle River near Lumsden	Medium	95%	95%	88%	NA	NA
South Saskatchewan River at SA*	Medium	100%	99%	86%	89%	NA
Swift Current Creek	Medium	90%	98%	75%	100%	100%
Assiniboine River at Kamsack	Medium	94%	94%	78%	100%	100%
Blindman River near Blackfalds	Medium	91%	98%	89%	100%	100%
Battle River near Ponoka	Medium	94%	98%	91%	100%	NA
Red Deer River at Drumheller	Medium	93%	92%	89%	60%	NA
Medicine River near Eckville	Medium	92%	94%	93%	88%	NA
Milk River at Milk River	Medium	86%	97%	70%	75%	NA
Red Deer River near Bindloss	Medium	91%	98%	87%	60%	NA
South Saskatchewan River at MH*	Medium	93%	98%	100%	100%	50%
Crowsnest River at Frank	High	88%	96%	100%	100%	NA
Bow River at Calgary	High	88%	98%	85%	100%	100%
Oldman River near Lethbridge	High	93%	98%	80%	88%	NA
Highwood River	High	95%	98%	100%	75%	NA
Belly River	High	92%	98%	85%	100%	NA
St. Mary River	High	94%	98%	90%	100%	NA

Note: PA* = Prince Albert, DC* = Deer Creek, SA* = Saskatoon, MH* = Medicine Hat, and NA = No Available Occurrences.

The drought level forecast indicates that the LSTM model trained on weekly data can capture the drought levels effectively as well as the training phase of LSTM for monthly data. Then the trained model was applied to the testing phase to evaluate the model performance on the testing data set.

Table 7.24 Average mismatch accuracy of Drought Level forecasts - Weekly LSTM (Training)

River flow rate capacity per unit area	Average Mismatch Accuracy of Drought Level Forecast (%)				
	D0	D1	D2	D3	D4
Low flow	93%	96%	89%	88%	100%
Medium flow	93%	97%	87%	85%	90%
High flow	92%	98%	90%	94%	100%

When considering training, the weekly LSTM model performs well in accurately predicting all drought levels, with all forecast averages exceeding 85% accuracy. This indicates the ability of the model to capture and accurately represent complex drought patterns, highlighting its effectiveness in understanding and forecasting varying drought conditions. The trained model was then evaluated on testing data to determine the model performance on unseen data.

7.4.2 Evaluation of Results of LSTM obtained during the Testing Phase

The trained model was then applied to predict the test data. As weekly data were used to train and predict, the monthly value was obtained by averaging four consecutive weekly values. This approach ensures that the model captures finer temporal patterns, leading to more accurate monthly forecasts. The predicted versus actual SDI were compared using MSE and R^2, as shown in Table 7.25.

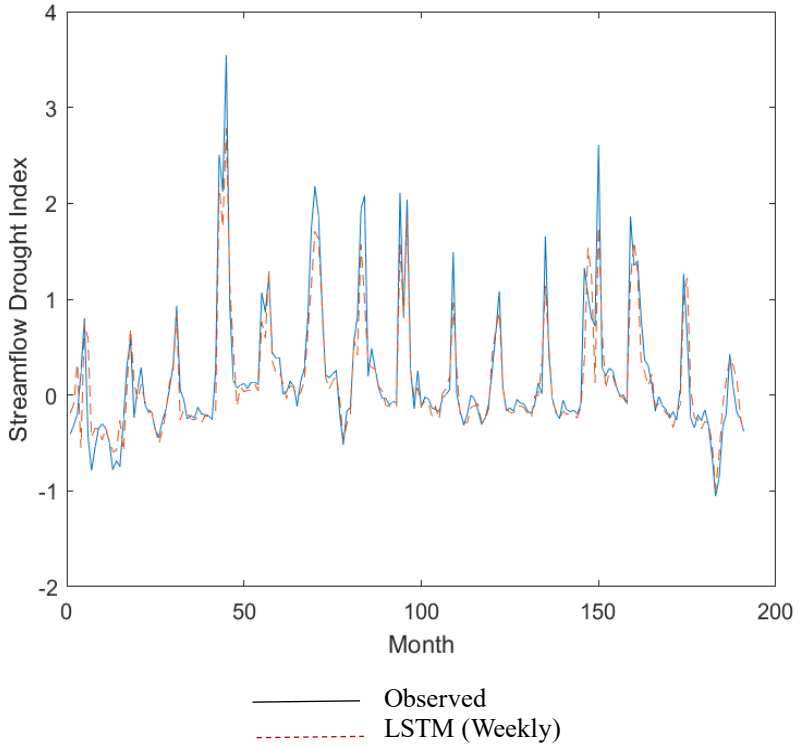


Figure 7.16 Long Short-Term Memory Model for weekly data in East Poplar River (Testing).

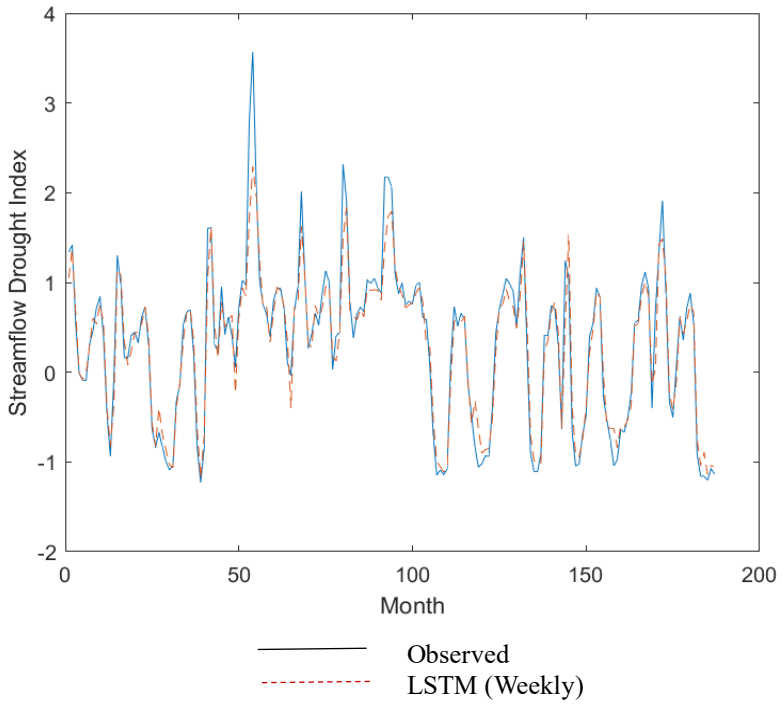


Figure 7.17 Long Short-Term Memory Model for weekly data in South Saskatchewan River (Testing).

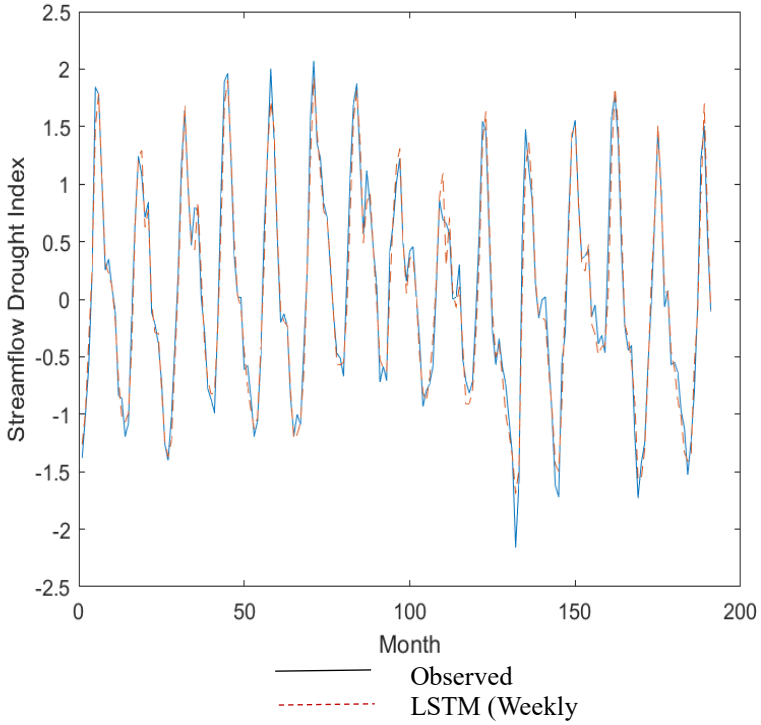


Figure 7.18 Long Short-Term Memory Model for weekly data in Highwood River (Testing).

When considering the forecast accuracy of testing phase with the weekly data, the LSTM model shows the capability of capturing both short-term and long-term relationships as the extended weekly patterns can uncover hidden dependencies within the data, leading to more accurate forecasts as well as in training phase.

Considering the testing results of both the monthly and weekly LSTM models, it is clear that the weekly LSTM provides superior forecasting performance compared to the monthly LSTM. This improvement can be attributed to the expanded pattern provided by the weekly data, which allows the LSTM model to capture finer details and nuances in the patterns. As a result, the model gains a better understanding of the data, leading to more accurate predictions.

Table 7.25 shows the summary of statistical measurements, including MSE and R^2 values, for all 25 watersheds.

Table 7.25 Summary statistics for Long Short-Term Memory Model for weekly data (Testing data)

Streamflow station name	River flow capacity (flow/unit area)	MSE	R^2
Pembina River at Neche	Low	0.02554	0.95962
Souris River near Sherwood	Low	0.03411	0.95724
East Poplar River at International Boundary	Low	0.06555	0.87732
Little Saskatchewan River near Minnedosa	Medium	0.06878	0.90207
Whitemud River at Westbourne	Medium	0.05355	0.93685
Moose Jaw River near Burdick	Medium	0.08090	0.92969
North Saskatchewan River at Prince Albert	Medium	0.04346	0.94384
North Saskatchewan River near Deer Creek	Medium	0.03192	0.95590
Qu'appelle River near Lumsden	Medium	0.07684	0.92439
South Saskatchewan River at Saskatoon	Medium	0.03054	0.95981
Swift Current Creek below Rock Creek	Medium	0.08636	0.84334
Assiniboine River at Kamsack	Medium	0.00871	0.99926
Blindman River near Blackfalds	Medium	0.14615	0.78890
Battle River near Ponoka	Medium	0.10274	0.91130
Red Deer River at Drumheller	Medium	0.02480	0.96281
Medicine River near Eckville	Medium	0.08101	0.86588
Milk River at Milk River	Medium	0.03908	0.95470
Red Deer River near Bindloss	Medium	0.02768	0.96946
South Saskatchewan River at Medicine Hat	Medium	0.06853	0.89240
Crowsnest River at Frank	High	0.02154	0.97837
Bow River at Calgary	High	0.06984	0.93931
Oldman River near Lethbridge	High	0.05927	0.88773
Highwood River near the Mouth	High	0.02575	0.97058
Belly River near Mountain View	High	0.03619	0.96182
St. Mary River at International Boundary	High	0.02487	0.97458

The test results reveal significantly higher forecast accuracy (Table 7.25) for the weekly LSTM model with an R^2 of 0.93, compared to an R^2 of 0.56. for the monthly LSTM model. This suggests

that using weekly data has substantially improved the performance of the model in the testing phase.

Table 7.26 Mismatch accuracy of weekly LSTM Model forecast results for different drought levels (Testing)

Streamflow station name	River flow	Drought Level Forecast Averages (%)				
		D0	D1	D2	D3	D4
Pembina River at Neche	Low	98%	91%	NA	NA	NA
Souris River near Sherwood	Low	98%	80%	100%	NA	NA
East Poplar River	Low	95%	95%	100%	NA	100%
Little Saskatchewan River	Medium	97%	75%	NA	NA	NA
Whitemud River at Westbourne	Medium	96%	100%	NA	NA	NA
Moose Jaw River near Burdick	Medium	93%	88%	73%	NA	NA
North Saskatchewan River at PA*	Medium	94%	96%	100%	NA	NA
North Saskatchewan River near DC*	Medium	100%	93%	85%	88%	NA
Qu'appelle River near Lumsden	Medium	100%	97%	100%	90%	NA
South Saskatchewan River at SA*	Medium	100%	95%	100%	NA	NA
Swift Current Creek	Medium	100%	80%	NA	NA	NA
Assiniboine River at Kamsack	Medium	94%	94%	78%	100%	100%
Blindman River near Blackfalds	Medium	86%	85%	100%	NA	NA
Battle River near Ponoka	Medium	95%	100%	71%	67%	NA
Red Deer River at Drumheller	Medium	97%	100%	NA	NA	NA
Medicine River near Eckville	Medium	79%	94%	NA	NA	NA
Milk River at Milk River	Medium	97%	100%	100%	NA	NA
Red Deer River near Bindloss	Medium	94%	100%	NA	NA	NA
South Saskatchewan River at MH*	Medium	97%	90%	NA	NA	NA
Crowsnest River at Frank	High	100%	85%	80%	NA	NA
Bow River at Calgary	High	100%	93%	85%	NA	100%
Oldman River near Lethbridge	High	93%	96%	NA	NA	NA
Highwood River	High	97%	92%	78%	NA	NA
Belly River	High	100%	93%	93%	90%	NA
St. Mary River	High	97%	92%	100%	NA	NA

Note: PA* = Prince Albert, DC* = Deer Creek, SA* = Saskatoon, MH* = Medicine Hat, and NA = No Available Occurrences.

The following Table 7.27 summarizes the Average Drought Level forecasts. This summary provides an overview of how well the model predicts different drought levels.

Table 7.27 Drought level forecast averages – Weekly LSTM (Testing)

River flow rate capacity per unit area	Average Mismatch Accuracy of Drought Level forecast (%)				
	D0	D1	D2	D3	D4
Low flow	97%	89%	100%	NA	100%
Medium flow	95%	93%	NA	86%	100%
High flow	98%	92%	87%	90%	100%

When considering the average drought level forecast, the weekly LSTM model performs exceptionally well, with all drought level forecasts achieving more than 85% accuracy. Notably, the model performs particularly well in predicting extreme drought levels, indicating its strong capability in handling severe drought conditions. This demonstrates that the weekly LSTM model is superior in forecasting seasonal drought patterns compared to all other models because it effectively captures the complexity of drought conditions with high accuracy.

7.5 Comparison between the Hidden Markov model and LSTM models

Table 7.28 summarizes the statistical evaluation and compares how the Hidden Markov Model, LSTM model for the monthly data, and LSTM model for the weekly data handle the complexity of drought data. The MSE and R^2 values for both the training and testing phases for each model are indicated.

After predicting the values of the monthly SDI, the drought index for each group was calculated as the average of the SDI of relevant months within the group. The actual drought levels were then compared to the predicted drought levels of the group using the mismatch evaluation method to assess the accuracy of each model. This provides a qualitative evaluation of the model forecasts by assessing how well each model captures drought conditions. Table 7.29 below shows this qualitative evaluation of the drought level forecast for each model.

Table 7.28 Summary statistics for the training and testing phases of each model.

Streamflow station name	Hidden Markov Model				LSTM - Monthly				LSTM - Weekly			
	Training		Testing		Training		Testing		Training		Testing	
	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2
Pembina River at Neche^{+}	0.37	0.68	0.58	0.54	0.01	0.99	0.33	0.52	0.01	1.00	0.03	0.96
Souris River near Sherwood^{+}	0.47	0.64	0.57	0.38	0.02	0.99	0.61	0.34	0.02	0.98	0.03	0.96
East Poplar River at International Boundary^{+}	0.45	0.60	1.08	0.46	0.03	0.98	0.57	0.34	0.02	0.98	0.07	0.88
Little Saskatchewan River near Minnedosa^{++}	0.37	0.68	1.06	0.45	0.01	0.99	0.32	0.66	0.01	0.99	0.07	0.90
Whitemud River at Westbourne^{++}	0.35	0.71	1.06	0.37	0.01	0.99	0.41	0.58	0.02	0.98	0.05	0.94
Moose Jaw River near Burdick^{++}	0.37	0.69	1.04	0.37	0.01	0.99	0.80	0.25	0.03	0.97	0.08	0.93
North Saskatchewan River at Prince Albert^{++}	0.28	0.72	0.59	0.60	0.01	0.99	0.23	0.75	0.02	0.97	0.04	0.94
North Saskatchewan River near Deer Creek^{++}	0.20	0.81	0.57	0.69	0.01	0.99	0.34	0.66	0.03	0.97	0.03	0.96
Qu'appelle River near Lumsden^{++}	0.21	0.81	1.43	0.46	0.02	0.98	0.60	0.49	0.02	0.98	0.08	0.92
South Saskatchewan River at Saskatoon^{++}	0.36	0.70	1.46	0.40	0.01	0.99	0.66	0.36	0.03	0.97	0.03	0.96
Swift Current Creek below Rock Creek^{++}	0.45	0.63	1.85	0.27	0.01	0.99	0.71	0.30	0.04	0.96	0.09	0.84
Assiniboine River at Kamsack^{++}	0.35	0.70	1.03	0.52	0.01	0.99	0.25	0.74	0.01	1.00	0.01	1.00
Blindman River near Blackfalds^{++}	0.34	0.71	1.26	0.32	0.01	0.99	0.46	0.44	0.02	0.98	0.15	0.79
Battle River near Ponoka^{++}	0.59	0.50	1.61	0.43	0.01	0.99	0.62	0.50	0.01	0.99	0.10	0.91
Red Deer River at Drumheller^{++}	0.38	0.68	1.19	0.39	0.01	0.99	0.31	0.61	0.02	0.98	0.02	0.96
Medicine River near Eckville^{++}	0.55	0.55	1.34	0.36	0.02	0.99	0.49	0.48	0.04	0.96	0.08	0.87
Milk River at Milk River^{++}	0.16	0.85	1.23	0.41	0.02	0.98	0.24	0.74	0.02	0.98	0.04	0.95
Red Deer River near Bindloss^{++}	0.32	0.72	1.17	0.48	0.01	0.99	0.33	0.60	0.02	0.98	0.03	0.97
South Saskatchewan River at Medicine Hat^{++}	0.65	0.49	1.42	0.31	0.01	0.99	0.57	0.32	0.03	0.97	0.07	0.89
Crowsnest River at Frank^{+++}	0.37	0.67	0.64	0.60	0.01	0.99	0.21	0.79	0.02	0.98	0.02	0.98
Bow River at Calgary^{+++}	0.17	0.80	0.59	0.64	0.01	0.99	0.33	0.72	0.02	0.98	0.07	0.94
Oldman River near Lethbridge^{+++}	0.38	0.68	0.75	0.52	0.01	0.99	0.34	0.48	0.02	0.98	0.06	0.89
Highwood River near the Mouth^{+++}	0.20	0.81	0.65	0.67	0.02	0.98	0.20	0.79	0.02	0.98	0.03	0.97
Belly River near Mountain View^{+++}	0.29	0.74	0.84	0.59	0.01	0.99	0.28	0.76	0.02	0.98	0.04	0.96
St. Mary River at International Boundary^{+++}	0.30	0.73	1.04	0.44	0.02	0.99	0.31	0.67	0.02	0.98	0.02	0.97
NOTE: + indicates low flow; ++ medium flow, and +++ high flow rivers.												

Table 7.29 Drought forecast results of the training and testing phases of each model.

Streamflow station name	Discrete Markov Model Forecast	Hidden Markov Model		LSTM - Monthly data		LSTM - Weekly data	
		Forecast (Train)	Forecast (Test)	Forecast (Train)	Forecast (Test)	Forecast (Train)	Forecast (Test)
Pembina River at Neche+	61%	98%	81%	95%	80%	96%	95%
Souris River near Sherwood+	49%	72%	80%	96%	75%	97%	95%
East Poplar River at International Boundary+	43%	81%	73%	95%	81%	95%	86%
Little Saskatchewan River near Minnedosa++	80%	68%	73%	96%	92%	96%	89%
Whitemud River at Westbourne++	61%	69%	71%	97%	78%	98%	97%
Moose Jaw River near Burdick++	73%	65%	63%	95%	78%	95%	90%
North Saskatchewan River at Prince Albert++	63%	79%	73%	94%	93%	94%	95%
North Saskatchewan River near Deer Creek++	53%	65%	59%	98%	85%	98%	95%
Qu'appelle River near Lumsden++	56%	80%	75%	94%	69%	96%	94%
South Saskatchewan River at Saskatoon++	69%	65%	85%	97%	81%	97%	99%
Swift Current Creek below Rock Creek++	71%	67%	85%	96%	80%	96%	95%
Assiniboine River at Kamsack++	73%	71%	76%	96%	85%	97%	95%
Blindman River near Blackfalds++	51%	77%	76%	96%	75%	97%	90%
Battle River near Ponoka++	49%	69%	68%	99%	64%	99%	93%
Red Deer River at Drumheller++	75%	75%	69%	94%	83%	98%	98%
Medicine River near Eckville++	73%	78%	68%	95%	83%	97%	87%
Milk River at Milk River++	49%	85%	58%	89%	73%	89%	98%
Red Deer River near Bindloss++	71%	85%	71%	93%	78%	98%	97%
South Saskatchewan River at Medicine Hat++	61%	69%	64%	98%	81%	98%	94%
Crowsnest River at Frank+++	55%	63%	78%	96%	75%	97%	92%
Bow River at Calgary+++	49%	65%	75%	95%	78%	98%	91%
Oldman River near Lethbridge+++	43%	77%	81%	96%	78%	97%	94%
Highwood River near the Mouth+++	78%	82%	68%	97%	76%	97%	93%
Belly River near Mountain View+++	65%	82%	61%	96%	71%	97%	95%
St. Mary River at International Boundary+++	57%	73%	66%	96%	77%	97%	95%
NOTE: + indicates low flow; ++ medium flow, and +++ high flow rivers.							

The following discussions are made according to the summarized results provided by all four models during the training and testing phases.

The Discrete Markov model provides better results for high and medium-flow rivers compared to low-flow rivers. This suggests that while the model performs well in predicting drought conditions for rivers with higher or more consistent flow, it faces challenges with the variability and complexity of patterns often associated with low-flow rivers. As a solution to address these limitations, the Hidden Markov Model (HMM) was applied as it is designed to better capture more complex patterns in the data using emission matrix and it can understand the underlying hidden states (underline drought levels) using transition matrix.

From the summary results of the Hidden Markov Model, it is apparent that the model performs well for low-flow rivers compared to medium and high-flow rivers during the training phase. However, during the testing phase, the model demonstrates higher accuracies in predicting both no drought (D0) and extreme severe drought (D4) across all flow categories (low-, medium-, and high-flow rivers. Plus, for intermediate drought levels, the Hidden Markov Model did not yield good results. This limitation arises because the model focuses on identifying hidden states and making forecasts based on the extreme values of each drought level (each hidden state).

To address this limitation and capture the more complex relationships within seasonal patterns, the LSTM model can be utilized. Unlike the Markov Models, the LSTM model can account for both long term and short-term temporal dependencies and variations, offering a more comprehensive approach to predicting a broader range of drought conditions.

First, the LSTM model was applied using monthly data, and it achieved highly accurate results during the training phase for all low-, medium-, and high-flow rivers. However, during the testing phase, the model could primarily predict lower drought levels, such as no drought (D0) and mild drought (D1), with considerable accuracy. However, it struggled to accurately predict more severe drought levels for all types of low-, medium-, and high-flow rivers. A reason for such an issue is due to a lack of sufficient data to identify the complex hidden relationships within the data. To address this issue, the same LSTM model was applied to weekly data instead of monthly data. This approach aimed to expand the distribution and provide a more detailed view of temporal variations, thereby improving the ability of the model to capture and predict a broader range of drought levels, including more severe conditions.

The LSTM model using weekly data yields equally good results, with average drought level forecast accuracy exceeding 85% for both the training and testing phases across all low-flow, medium-flow, and high-flow rivers. This consistency indicates that the LSTM model trained on weekly data effectively captures the complex patterns and dependencies across different river flow conditions, making it a successful tool for predicting droughts.

Drought data are inherently complex due to their intricate patterns and variability. This complexity makes accurate forecast challenging, as it requires capturing subtle and often non-linear relationships within the data. Compared to the Discrete Markov model, the Hidden Markov Model, and the LSTM model; the LSTM model delivered the highest accuracy in predicting drought data. This superior performance highlights the LSTM unique ability of the model to effectively capture and model complex patterns within drought data. Its advanced architecture allows it to learn both short-term and long-term dependencies, making it particularly adept at handling the intricate temporal dynamics of streamflow patterns. As a result, LSTM models demonstrate a clear advantage in accuracy for streamflow forecasts, showcasing their strength in dealing with sequential data.

7.5.1 Evaluation of Reliability of LSTM Model

Unlike traditional models that struggle to capture complex, nonlinear patterns, LSTM networks can learn to recognize and remember relevant past information, allowing them to make more accurate forecasts. Several key advantages of the LSTM are presented as follows,

1. Capture Long-Term Dependencies: LSTMs are designed to retain information over long sequences, making them effective at learning and remembering long-term dependencies.
2. Adapt to Complex Patterns: The model analyse temporal patterns through three gates in order to identify the complex dependencies within the data.
3. Improved Performance in Sequence Forecast: Due to their ability to learn both short term and long-term patterns, LSTMs provide better performance in sequence forecast tasks.

Therefore, the LSTM- model based on weekly data developed in this thesis were selected as the best models for seasonal drought pattern forecasting.

Chapter VIII: Conclusion and Recommendations

8.1 Conclusions

This study has investigated the concept of grouping streamflow data based on similar behaviours and predicting hydrological droughts based on such recognized patterns or groups within streamflow data. Entropy calculations were employed to identify similar behaviours and determine the optimal way to group the data. The characteristics of such groups were analyzed as seasonal patterns using the Moving Average. Such seasonal patterns were clustered using K-means clustering into different drought levels (namely: no drought, mild drought, moderate drought, severe drought and extreme severe droughts). After determining the drought levels of seasonal patterns, such recognized patterns were applied to three different models for drought level forecast such as the Discrete Markov Model, the Hidden Markov Model, and the LSTM model. For the evaluation of models for their suitability in drought forecast, the study involved streamflow data sets from 25 watersheds prone to drought in the driest part of Canada.

All three models; the Discrete Markov Model, the Hidden Markov Model, and the LSTM model offer distinct advantages and disadvantages. Each model has its strengths, such as the simplicity and interpretability of the Discrete Markov Models or the ability to handle hidden states in the Hidden Markov Model. However, the LSTM model stands out with superior overall results. The LSTM model excels in capturing complex temporal patterns and long-term dependencies, making it the most effective for accurately predicting streamflow data in this study. Its advanced architecture allows it to handle intricate variations and dependencies, leading to the most reliable forecasting performance among the models evaluated.

8.2 Forecasting Seasonal Drought Pattern using Discrete Markov Model

The Discrete Markov Model was used to predict the drought level based on the streamflow drought index of each seasonal pattern or group. In this approach, forecasts were made by determining the most probable next drought level given the current drought level through a transition matrix. The results of the model indicate that it forecasts drought levels more accurate for medium flow rivers. This improved performance is largely due to the repeating patterns observed in these rivers. The Markov Model, which excels in understanding transitions between states based on historical data, benefits from the regularity of these patterns. The consistent and predictable nature of medium flow rivers allows the model to better grasp the underlying process and make more precise forecasts. Consequently, the ability of the model to accurately predict drought levels in medium flow rivers is enhanced by the stable and repetitive flow characteristics that align well with the probabilistic approach of the model.

However, the transition matrix only indicates the likelihood of different future drought levels based on the known current drought level. By predicting only the highest probable drought level, the model disregards other possible scenarios that, while less likely, could still occur. This approach does not account for uncertainty or variability in forecasts. It may miss out on less likely but significant drought levels that could have substantial impacts. Ignoring such likelihoods of occurrences can lead to a lack of preparedness for less probable but possible drought events.

As a solution, instead of relying solely on a univariate dataset like streamflow, incorporating additional parameters such as precipitation and soil moisture levels can enhance the predictive capability of the model. Including such additional factors provides a more comprehensive view of the hydrological system and improves the ability of the model to account for various scenarios and uncertainties.

Additionally, the transition matrices could be expanded to more advanced forms to better capture the behaviour of each seasonal pattern. Incorporating other statistical measures, such as variation and skewness within the seasonal patterns, can provide a more nuanced understanding of the data. These measures help in identifying and modelling the subtle differences and shifts in drought conditions, leading to more accurate and comprehensive forecasts. By integrating such advanced statistical techniques, the model can better account for the complexities and variability within seasonal patterns.

8.3 Forecasting Seasonal Drought Pattern using Hidden Markov Model

While comparing the summarized statistics, it became evident that the Hidden Markov Model (HMM) primarily focuses on predicting the hidden states of the observations rather than the actual observations themselves during the training phase. A hidden state in an HMM is an internal, unobservable condition or state that influences the observable data. Although these hidden states are not directly observed, they represent underlying processes or patterns that affect the observable outcomes.

The training process utilizes the Viterbi algorithm, which is designed to identify the most probable sequence of hidden states that best explains the observed data. This algorithm emphasizes uncovering the underlying dynamics that drive the observable outcomes rather than directly predicting the specific values of the observations. Hidden states play a crucial role in this process by representing the internal conditions that influence the data, and the model captures the transitions between these states as well as their impact on the observed data through emission probabilities.

In the testing phase, the Hidden Markov Model provide estimates of future seasonal patterns based on such states. The results from the Hidden Markov Model (HMM) demonstrate its capability to effectively forecast high variability in streamflow data, yielding better performance for both low flow and high flow rivers. The model leverages extreme values identified in each hidden state to predict drought patterns. However, this approach can result in forecasts that are more extreme than the actual observed outcomes. The strength of HMM lies in its ability to capture and predict the complex dynamics and variability within the data, but this sensitivity to extreme values can lead to forecasts that exaggerate the severity of drought conditions compared to the real data. This tendency can be advantageous for disaster management, as it often predicts more extreme scenarios, potentially providing a more cautious approach to risk assessment and management. The behaviour of the model rises from its tendency to predict observations based on the extreme values associated with the hidden states. By focusing on such extremes, the HMM prioritize scenarios that represent the worst-case conditions, which can be useful for planning and preparedness. However, this approach also means that the model may not accurately capture the more moderate, typical observations, leading to forecasts that skew towards the extremes of the drought levels.

The model performed better for low-flow rivers during the training phase, suggesting that it is more effective at identifying the hidden states associated with highly variable distributions. However, in the testing phase, the HMM showed improved performance for all low-flow, medium-flow and high-flow rivers. This indicates that while the model may have initially excelled at learning patterns from the more complex flow rivers, it ultimately generalized better to the different types of rivers when applied to new data. This shift in performance highlights the capability of the model to adapt and provide useful forecasts even for more complex scenarios.

8.4 Forecasting Seasonal Drought Pattern using the LSTM Model

Overall, the LSTM model demonstrated significantly higher accuracy compared to statistical models, such as the Discrete Markov Models and Hidden Markov Models. This superior performance is attributed to the advanced ability of LSTM to capture both short-term and long-term dependencies within the data, as well as its capacity to handle complex temporal patterns and variations. While the statistical models provide valuable insights, their more simplistic approach and reliance on discrete state transitions or aggregated data often limit their predictive accuracy.

The LSTM model produced similar results for the training phase when using both monthly and weekly data but exhibited differing performance in the testing phase. During training, the ability of the model to fit the data was comparable because it learned the underlying patterns effectively in both scenarios.

The LSTM model ultimately provided the most accurate forecasts for drought patterns due to its ability to capture both long-term and short-term dependencies in the data. By effectively learning from these temporal relationships, the LSTM model achieved superior results in predicting drought conditions. The weekly data-based LSTM model delivered the best performance, as the finer granularity of weekly data allowed the model to identify more detailed and refined patterns in drought conditions. This expanded input enabled the model to discern subtle variations and trends that were less apparent with monthly data, enhancing its overall forecasting accuracy. This arises because the weekly data, with its finer temporal resolution, captures more detailed and frequent variations in the streamflow data. This granularity allows the LSTM model to learn and generalize complex, short-term patterns more effectively. In contrast, the monthly data aggregates

observations over a longer period, potentially smoothing out critical short-term fluctuations and reducing the ability of the model to capture and predict these nuances.

8.5 Limitations of LSTM Models

The structure of the developed LSTM model is specifically suited for drought forecast in rivers within the Palliser Triangle area. As a result, for watersheds in other regions or responses to significant changes within the Palliser Triangle, such as urbanization or deforestation, the current model structure may not be appropriate. Therefore, it is crucial to adapt and reconstruct the LSTM model to account for these varying conditions and ensure its effectiveness across different environments and scenarios.

8.6 Recommendations for future works

It should be noted that this study is based on specific model architectures. To achieve more accurate and reliable results, future research should consider exploring several additional concepts.

1. Utilize Multivariate Models: As this study only used streamflow data, it is recommended to incorporate multiple variables such as precipitation, soil moisture and temperature data to capture a more comprehensive view of the factors influencing drought conditions. This approach especially could be helpful for the Discrete Markov Model.

2. Explore Hybrid and Ensemble Models: This study applied three different models separately but combining those modelling approaches to leverage the strengths of various techniques can improve forecast accuracy.

3. Investigate Different Lengths of Seasonal Patterns: In this study, only similar-length segments were considered for simplicity. But with climate change, the seasons have varied with different durations and lengths. Based on such an observation, it is recommended to experiment with seasonal patterns of varying lengths to capture a broader range of seasonal variations.

References

Alawsi, M.A., Zubaidi, S.L., Al-Bdairi, N.S.S., Al-Ansari, N. and Hashim, K., 2022, Drought Forecasting: A Review and Assessment of the Hybrid Techniques and Data Pre-Processing. Hydrology, vol 9, pp 115.
DOI: https://doi.org/10.3390/hydrology9070115

Ali, Z., Hussain, I., Faisal, M., Nazir, H.M., Hussain, T., Muhammad, Y., Shoukry, A.M., and Gani, S.H., 2017, Forecasting Drought Using Multilayer Perceptron Artificial Neural Network Model. Advances in Meteorology.
DOI: https://doi.org/10.1155/2017/5681308

Almedeij, J., 2016, Long-term periodic drought modeling. Stoch Environ Res Risk Assess, vol 30, pp 901–910.
DOI: https://doi.org/10.1007/s00477-015-1065

Alquraish M, Abuhasel K.A, Alqahtani A.S. and Khadr M., 2021, SPI-Based Hybrid Hidden Markov–GA, ARIMA–GA, and ARIMA–GA–ANN Models for Meteorological Drought Forecasting. Sustainability. vol 13(22), pp12576.
DOI: https://doi.org/10.3390/su132212576

Apaydin, H., Feizi, H., Sattari, M.T., Colak, M.S., Shamshirband, S. and Chau, K.W., (2020), Comparative Analysis of Recurrent Neural Network Architectures for Reservoir Inflow Forecasting. Water. Vol 12(5), pp 1500.
DOI: https://doi.org/10.3390/w12051500

Baum, L. E., 1966, Statistical inference for probabilistic functions of finite-state Markov chain. Annals of Mathematical Statistics, vol 37(6), pp 1554-1563.
DOI: https://doi.org/10.1214/aoms/1177699147

Belayneh, A., and Adamowski, J., 2013, Drought forecasting using new machine learning methods. Journal of Water and Land Development, no. 18, pp 3–12.

Beran, M., and Rodier, J., 1985, Hydrological aspects of drought. Paris, France: UNESCO.

Boken, V.K., Haque, C.E., and Hoogenboom, G., 2007. Agricultural drought prediction using pattern recognition, 2007. Annals of Arid Zone, vol 46(2), pp 133-144.

Box, G.E.P. and Cox, D.R., 1964, An Analysis of Transformation. Journal of the Royal Statistical Society, Series B (Methodological), vol 26, pp 211-252.

Brase, C. H., and Brase, C. P., 2006, Understanding basic statistics (4th ed.). Boston, MA: Houghton Mifflin Co.

Brust C, Kimball J.S., Maneta M.P., Jencso K. and Reichle R.H., 2021, DroughtCast: A Machine Learning Forecast of the United States Drought Monitor. Front. Big Data, vol 4:773478,
DOI: https://doi.org/10.3389/fdata.2021.773478

Burnett, L. D., (2007). Palliser Triangle. Retrieved from Encyclopedia of Saskatchewan: https://esask.uregina.ca/entry/palliser_triangle.html#:~:text=It%20is%20a%20natural%20grassland,annual%20snowfall%20is%20101%20cm.

Cancelliere, A., Di Mauro, G., Bonaccorso, B. and Rossi, G., 2007, Drought Forecasting Using the Standardized Precipitation Index. Water Resources Management, vol 21, pp 801-819.
DOI: https://doi.org/10.1007/s11269-006-9062-y

Chong, K., Lai, S., and El-Shafie, A., 2019, Wavelet transform based method for river stream flow time series frequency analysis and assessment in tropical environment. Water Resources Management, vol 33(6), pp 2015-2032.

Chong, K.L., Huang, Y., Ali, N.A.A, and El-Shafie, A., 2021, Spatiotemporal Variability Analysis of Standardized Precipitation Indexed Droughts Using Wavelet Transform. Journal of Hydrology. vol 605. pp 127299.
DOI: 10.1016/j.jhydrol.2021.127299

Coşkun, Ö., and Citakoglu, H., 2023, Prediction of the standardized precipitation index based on the long short-term memory and empirical mode decomposition-extreme learning machine models: The Case of Sakarya, Türkiye. Physics and Chemistry of the Earth, Parts A/B/C, vol 131, pp 103418.
DOI: 10.1016/j.pce.2023.103418.

Dotse, S.Q., 2024, Deep learning–based long short-term memory recurrent neural networks for monthly rainfall forecasting in Ghana, West Africa. Theor Appl Climatol, vol 155, pp 3033–3045.
DOI: https://doi.org/10.1007/s00704-023-04773

Evkaya, O., and Kurnaz, F., 2020, Forecasting drought using neural network approaches with transformed time series data. Journal of Applied Statistics, vol 48(14), pp 2591-2606.

Feng, D., Lawson, K., and Shen, C., 2021, Prediction in ungauged regions with sparse flow duration curves and input-selection ensemble modeling, Geophysical Research Letters, vol 48(16).
DOI: 10.48550/arXiv.2011.13380

Feng, X., Merow, C., Liu, Z., Park, D.S., Roehrdanz, P.R., Maitner, B., Newman, E.A., Boyle, B.L., Lien, A., Burger, J.R., Pires, M.M., Brando, P.M., Bush, M.B., McMichael, C.N.H., Neves, D.M., Nikolopoulos, E.I., Saleska, S.R., Hannah, L., Breshears, D.D., Evans, T.P., Soto, J.R., Ernst, K.C. and Enquist, B.J., 2021, How deregulation, drought and increasing fire impact Amazonian biodiversity. Nature, vol 597(7877), p 516-521.
DOI: 10.1038/s41586-021-03876-7

DOI: K., Huang, Y., and Chai-Hoon, C., 2020, Seasonal Hydrological Drought Indicator for Tropical Drought Identification. International Journal of Environmental Science and Development, vol 11, pp 99-105.
DOI: 10.18178/ijesd.2020.11.2.1233

Gabriel, K.R. and Neumann, J., 1962, A Markov Chain Model for Daily Rainfall Occurrence at Tel Aviv. Quarterly Journal of the Royal Meteorological Society, vol 88, pp 90-95.
DOI: http://dx.doi.org/10.1002/qj.49708837511

Galan-Sales, F., Jiménez, P., García, M., and Romera, J.L., 2023, An Approach to Enhance Time Series Forecasting by Fast Fourier Transform.
DOI: 10.1007/978-3-031-42529-5_25.

Gamelin, B., Rao, V., Bessac, J., and Altinakar, M., 2022, Machine Learning Methods with the Standardized VPD Drought Index to Identify and Assess Drought in the United States, EGU General Assembly, Vienna, Austria, EGU22-13100.
DOI: https://doi.org/10.5194/egusphere-egu22-13100

Goodier, C., and Panu, U.S., 1994, Infilling Missing Monthly Streamflow Data Using a Multivariate Approach, Stochastic and Statistical Methods in Hydrology and Environmental Engineering. Water Science and Technology Library, vol 10/3. Springer, Dordrecht.
DOI: https://doi.org/10.1007/978-94-017-3083-9_15

Gyaneshwar, A., Mishra, A., Chadha, U., and Vincent, P., 2023, A contemporary review on deep learning models for drought forecast. Sustainability, vol 15(3), pp 1234-1254.
DOI: 10.3390/su15076160

Hameed, M.M., Mohd Razali, S.F., Wan, W.H.M., Ahmad, A.M., and Yaseen, Z.M., 2023, Deep learning versus hybrid regularized extreme learning machine for multi-month drought forecasting: A comparative study and trend analysis in tropical region. Heliyon. vol 10(1), e22942.
DOI:: 10.1016/j.heliyon. 2023.e22942

Han, P., Wang, P., Tian, M., Zhang, S., Liu, J., and Zhu, D., 2013, Application of the ARIMA Models in Drought Forecasting Using the Standardized Precipitation Index. In: Li, D., Chen, Y. (eds) Computer and Computing Technologies in Agriculture VI. CCTA 2012. IFIP Advances in Information and Communication Technology, vol 392. Springer, Berlin.

Hao, Z., Hao, F., Singh, V., and Zhang, X., 2018, Changes in the severity of compound drought and hot extremes over global land areas. Environmental Research Letters, vol 13.
DOI: 10.1088/1748-9326/aaee96

Hao, Z., Singh, V. P., and Xia, Y., 2018, Seasonal Drought Prediction: Advances, Challenges, and Future Prospects. Reviews of Geophysics, vol 56, pp 108-141.
DOI: https://doi.org/10.1002/2016RG000549

Hewitt, K., 1997, Regions of Risk: A Geographical Introduction to Disasters (1st ed.). Routledge, London.
DOI: https://doi.org/10.4324/9781315844206

Hochreiter, S., and Schmidhuber, J., 1997, Long short-term memory. Neural Computation, vol 9(8), pp 1735-1780.
DOI: https://doi.org/10.1162/neco.1997.9.8.1735

Jehanzaib, M., Idrees, M., Kim, D., and Kim, T., 2021, Comprehensive evaluation of machine learning techniques for hydrological drought forecasting. Journal of Irrigation and Drainage Engineering, vol 147(7).
DOI: https://doi.org/10.1061/(asce)ir.1943-4774.0001575

Kingma, D.P. and Ba, J., 2015, Adam: A Method for Stochastic Optimization. Conference Paper at the 3rd International Conference for Learning Representations, San Diego.
DOI: arXiv:1412.6980ss

Kolachian, R., and Saghafan, B., 2021, Hydrological drought class early warning using support vector machines and rough sets. Environmental Earth Sciences.
DOI: 80. 10.1007/s12665-021-09536-3

Kumar, V., and Panu, U.S., 1997, Predictive assessment of the severity of agricultural droughts based on agro-climatic factors. Journal of the American Water Resources Association, vol 33(6), pp 1255-1264.
DOI: https://doi.org/10.1111/j.1752-1688.1997.tb03550.x

Kwon, Y., Cha, Y., and Park, Y., 2023, Assessing the impacts of dam/weir operation on streamflow predictions using LSTM across South Korea. Sci Rep, vol 13, pp 9296.
DOI: https://doi.org/10.1038/s41598-023-36439-z

L. R. Rabiner, 1989, A tutorial on hidden Markov models and selected applications in speech recognition, Proceedings of the IEEE, vol. 77, no. 2, pp. 257-286.
DOI: 10.1109/5.18626

Last, W.M., Teller, J.T. and Forester, R.M., 1994, Paleohydrology and paleochemistry of Lake Manitoba, Canada: the isotope and ostracode records. J Paleolimnol, vol 12, pp 269–282.
DOI: https://doi.org/10.1007/BF00678025

Leilah, A.A. and Al-Khateeb, S., 2005, Statistical analysis of wheat yield under drought conditions. Journal of Arid Environments, vol 61, pp 483-496.
DOI: 10.1016/j.jaridenv.2004.10.011

Li, J., Cai, Y., Lu, H., and Jin, L., 2022, Deep learning forecast model of drought and flood in summer based on random forest and attention mechanism. Research Square.
DOI: https://doi.org/10.21203/rs.3.rs-1619962/v1

Liu P, Wang J, Sangaiah AK, Xie Y and Yin, X., 2019, Analysis and Prediction of Water Quality Using LSTM Deep Neural Networks in IoT Environment. Sustainability, vol 11(7), pp 2058.
DOI: https://doi.org/10.3390/su11072058

Lohani, V., and Loganathan, G., 1997, An early warning system for drought management. Journal of American Water Resources Association, vol 33(6), pp 1375-1386.

Lohani, V., Loganathan, G., and Mostaghim, S., 1998, Long-Term Analysis and Short-Term Forecasting of Dry Spells by Palmer Drought Severity Index. Hydrology Research 1 February 1998; vol 29 (1), pp 21–40.
DOI:: https://doi.org/10.2166/nh.1998.0002

McKee, T.B., Doesken, N.J. and Kleist, J., 1993, The Relationship of Drought Frequency and Duration to Time Scales. 8th Conference on Applied Climatology, Anaheim, pp 179-184.

Meng, L., Ford, T., and Guo, Y., 2017, Logistic regression analysis of drought persistence in East China. International Journal of Climatology, vol 37(3), pp 1444-1455.
DOI: https://doi.org/10.1002/joc.4789

Mishra, A., and Desai, V., 2006, Drought forecasting using feed-forward recursive neural network. Ecological Modelling, vol 198(1-2), pp 127-138.
DOI: 10.1016/j.ecolmodel.2006.04.017

Mishra, A., Singh, V., and Desai, V., 2007, Drought characterization: A probabilistic approach. Stochastic Environmental Research and Risk Assessment, vol 21(1), pp 41-55.

Mishra, A.K. and Desai, V.R., 2005, Drought Forecasting Using Stochastic Models. Stochastic Environmental Research and Risk Assessment, vol 19, pp 326-339.
DOI: https://doi.org/10.1007/s00477-005-0238-4

Mishra, A.K. and Desai, V.R., 2005, Spatial and Temporal Drought Analysis in the Kansabati River Basin, India. International Journal of River Basin Management, vol 3, pp 31-41.
DOI: https://doi.org/10.1080/15715124.2005.9635243

Mishra, A.K. and Singh, V.P., 2010, A Review of Drought Concepts. Journal of Hydrology, vol 391, pp 202-216.
DOI: https://doi.org/10.1016/j.jhydrol.2010.07.012

Modarres, R., 2007, Streamflow drought time series forecasting. Stochastic Environmental Research and Risk Assessment, vol 21(1), p 223-233.
DOI: 10.1007/s00477-006-0058-1

Morid, S., Smakhtin, V. and Bagherzadeh, K., 2007, Drought Forecasting Using Artificial Neural Networks and Time Series of Drought Indices. International Journal of Climatology, vol 27, pp 2103-2111.
DOI: https://doi.org/10.1002/joc.1498

Nalbantis, I., Tsakiris, G., 2009, Assessment of Hydrological Drought Revisited. Water Resour Manage, vol 23, pp 881–897.
DOI: https://doi.org/10.1007/s11269-008-9305-1

Nguyen-Le, D., and Yamada, T., 2019, Using weather pattern recognition to classify and predict summertime heavy rainfall occurrence over the Upper Nan River Basin, Northwestern Thailand. American Meteorological Society, vol 100(3), pp 345-360.

Özger, M., Mishra, A. K. and Singh,V. P., 2012, Long Lead Time Drought Forecasting Using a Wavelet and Fuzzy Logic Combination Model: A Case Study in Texas. J. Hydrometeor., vol 13, pp 284–297.
DOI: https://doi.org/10.1175/JHM-D-10-05007.1

Palmer, W. C.,1965, Meteorological Drought. Office of Climatology Research Paper No. 45, Washington DC: US Weather Bureau.

Palmer, W.C., 1968, Keeping Track of Crop Moisture Conditions, Nationwide: The New Crop Moisture Index. Weatherwise, vol 21, pp 156-161.
DOI: https://doi.org/10.1080/00431672.1968.9932814

Pande C. B., Kushwaha N. L., Orimoloye I. R., Kumar R., Abdo H. G., Tolche A. D. and Elbeltagi A., 2023, Comparative Assessment of Improved SVM Method under Different Kernel Functions for Predicting Multi-scale Drought Index, Water Resources Management, vol. 37. no. 3. pp 1367-1399.

Panu, U. S. and Ng, W.W., 2017, Outlier Analysis and Infilling of Missing Records in Hydrologic Data. Chapter 16 in the Handbook of Applied Hydrology (New Edition -2017). pp 16-1 to 16-7, McGraw-Hill, New York.

Panu, U. S., and Afza, N., 1993, Entropic evaluation of streamflow data infilling procedures. In Proceedings of Stochastic and Statistical Methods in Hydrology and Environmental Engineering (pp. 410-412).

Panu, U., Unny, T.E, and Ragade, R., 1978, A feature forecast model in synthetic hydrology based on concepts of pattern recognition. Water Resources Research, vol 14(2), pp 335-344.

Panu, U.S and Unny, T.E., 1980, Extension and application of feature prediction model for synthesis of hydrologic records. Water Resources Research, Vol 16, No.1, pp 77-96.

Panu, U.S., Khalil, M., and Elshorbagy, A., 2000, Streamflow Data Infilling Techniques Based on Concepts of Groups and Neural Networks. In: Govindaraju, R.S., Rao, A.R. (eds) Artificial Neural Networks in Hydrology. Water Science and Technology Library, vol 36. Springer, Dordrecht.
DOI: https://doi.org/10.1007/978-94-015-9341-0_13

Poornima, S., Pushpalatha, M., Jana, R.B., Patti, L.A., 2023, Rainfall Forecast and Drought Analysis for Recent and Forthcoming Years in India. Water. 2023; vol 15(3), pp 592.
DOI: https://doi.org/10.3390/w15030592

Qian, T. S. (2022). A weighted kNN fault detection based on multistep index and dynamic neighborhood scale under complex working conditions. IEEE Access, vol 10, pp 49183-49192.

Raczyński, K., and Dyer, J. (2023). Harmonic oscillator seasonal trend (HOST) model for hydrological drought pattern identification and analysis. Journal of Hydrology, vol 620, 129514,
DOI: https://doi.org/10.1016/j.jhydrol.2023.129514

Rahimi, M., and Ebrahimi, H., 2023, Data driven of underground water level using artificial intelligence hybrid algorithms. Sci Rep 13, 10359.
DOI: https://doi.org/10.1038/s41598-023-35255-9

Rannie, W. F. (2006). A comparison of 1858-59 and 2000-01 drought patterns on the Canadian Prairies. Canadian Water Resources Journal, vol 31(3), pp 263–274.

Rao, A., and Padmanabhan, G. (1984). Analysis and modeling of Palmer's drought index series. Journal of Hydrology, 71(1-4), 211-229.
DOI: https://doi.org/10.1016/0022-1694(84)90212-9

Safarianzengir, V., Fatahi, A., Sobhani, B., and Doumari, S. A. (2022). Temporal and spatial analysis and monitoring of drought (meteorology) and its impacts on environment changes in Iran. Atmospheric Science Letters, vol 23(5), e1080.
DOI: https://doi.org/10.1002/asl.1080

Salas, J.D., Delleur, J.W., Yevjevich, V. and Lane, W.L. (1980) Applied Modeling of Hydrologic Time Series. Water Resources Publications, Littleton.

Schumacher, D. (2020). On the application of support vector machines to flash drought forecasting. University of Wisconsin-Madison.

Shannon, C.E. (1948), A Mathematical Theory of Communication. Bell System Technical Journal, 27: 379-423.
DOI: https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Sharma T.C., Panu U.S., 2024, Current State of Advances in Quantification and Modeling of Hydrological Droughts. Water. vol 16(5), pp 729.
DOI: https://doi.org/10.3390/w16050729

Sharma, H., Joshi, N., and Suryavanshi, S. (2022). Wavelet transform-based trend analysis of drought variables over homogenous monsoon regions of India. In Advances in Hydrology and Climate Change (p. 556). New York: Apple Academic Press.

Sharma, T. C., and Panu, U. S. (2010). Analytical procedures for weekly hydrological droughts: A case of Canadian rivers. Hydrological Sciences Journal, vol 55(1), pp 120-136.

Shi, K., Touge, Y., and Kazama, S. (2020). Evaluating spatiotemporal patterns and trends of drought in Japan associated with global climatic drivers. Natural Hazards and Earth System Sciences.
DOI: https://doi.org/10.5194/nhess-2020-416

Shiri, N., Shiri, J., Yaseen, Z., Kim, S., Chung, M., Nourani, V., and Kermani, M. (2021). Development of artificial intelligence models for well groundwater quality simulation: Different modelling scenarios. PLOS ONE, vol 16(11), pp 1-24.

Silva, S., and Estacio, A. (2020). Statistical uncertainty in drought forecasting using Markov Chains and the Standard Precipitation Index (SPI). Revista Brasileira de Climatologia, vol 24, pp 471-493.

Steinemann, A. (2003). Drought indicators and triggers: A stochastic approach to evaluation. Journal of Climate, vol 16(20), pp 3296-3307.

Tallaksen, L., and van Lanen, H. A. J. (Eds.). (2004). Hydrological drought: Processes and estimation methods for streamflow and groundwater. Amsterdam: Elsevier.

Tareke, K., and Awoke, A. (2023). Hydrological drought forecasting and monitoring system development using artificial neural network (ANN) in Ethiopia. Heliyon, vol 9(3), e14354. DOI: https://doi.org/10.1016/j.heliyon.2023.e13287

Tong, H. (1983). Threshold models in non-linear time series analysis. New York: Springer-Verlag.

Unny, T.E., Panu, U.S., MacInnes, C.D. and Wong, A.K.C., 1981, Pattern Analysis and Synthesis of Time-Dependent Hydrologic Data. Advances in Hydroscience, vol 12, pp 195-295. DOI:10.1016/B978-0-12-021812-7.50009-2

Villmow, J. R. (1956). The nature and the origin of the Canadian dry belt. Annals of the Association of American Geographers, vol 46(2), pp 211–232.

Vivian, E. F. (2022). Data test and pre-treatment for hydrological modelling and applications. Journal of Engineering in Agriculture & the Environment, vol 8(1), pp 1-13.

Wada, Y., Beek, L., Wanders, N., and Bierkens, M. (2013). Human water consumption intensifies hydrological drought worldwide. Environmental Research Letters, vol 8(3), pp 1-14.

Whitfield, P. H., Shook, K., and Pomeroy, J. (2020). Spatial patterns of temporal changes in Canadian Prairie streamflow using an alternative trend assessment approach. Journal of Hydrology, vol 591, pp 125573.

Wilhite, D., and Glantz, M. (1985). Understanding: The drought phenomenon: The role of definitions. Water International, vol 10(3), pp 111-120.

Yajun, Z. B. (2023). Wi-NN: Human gesture recognition system based on weighted KNN. Applied Sciences, vol 13(8), pp 3743.

Yang, W., Deng, M., Tang, J., and Gin, R. (2018). On the use of Markov chain models for drought class transition analysis while considering spatial effects. Natural Hazards, vol 92(2), pp 1029-1048.

Yevjevich, V. (1967). An objective approach to definitions and investigations of continental hydrological droughts. Hydrology Papers, Colorado State University.

Yulin, L. G. (2022). Application and effectiveness of weighted KNN in pattern recognition of communication modulated signals. In Proceedings of the IEEE 4th International Conference on Civil Aviation Safety and Information Technology (pp. 744-748). Dali, China.

Zahraie, B., and Nasseri, M. (2011). Basin-scale meteorological drought forecasting using a support vector machine (SVM). International Conference on Drought Management Strategies in Arid and Semi-Arid Regions.

Zeng, P., Sun, F., Liu, Y., Wang, Y., Li, G., and Che, Y., 2021, Mapping future droughts under global warming across China: A combined multi-timescale meteorological drought index and SOM-Kmeans approach. Weather and Climate Extremes, vol. 31, no. 100304.
DOI: 10.1016/j.wace.2021.100304

Zhang, J., Zhu, Y., Zhang, X., Ye, M., and Yang, J. (2018). Developing a Long Short-Term Memory (LSTM) based model for predicting water table depth in agricultural areas. Journal of Hydrology, vol 556, pp 918-929.

Zhang, L., Liu, Y., Ren, L., and Teuli, A. (2022). Analysis of flash droughts in China using machine learning. Hydrology and Earth System Sciences, vol 26(12), pp 3241–3261.

APPENDICES

Appendix-A: **Infilling of the Missing Data**

Appendix-B: **Recognizing Seasonal Drought Patterns**

Appendix-C: **Forecasting using Markov Models and LSTM Models**

APPENDIX – A

Infilling of the Missing Data

This appendix also demonstrates that the filled data for the selected 25 streamflow stations are within the 95% confidence boundary, ensuring their reliability for further analysis.

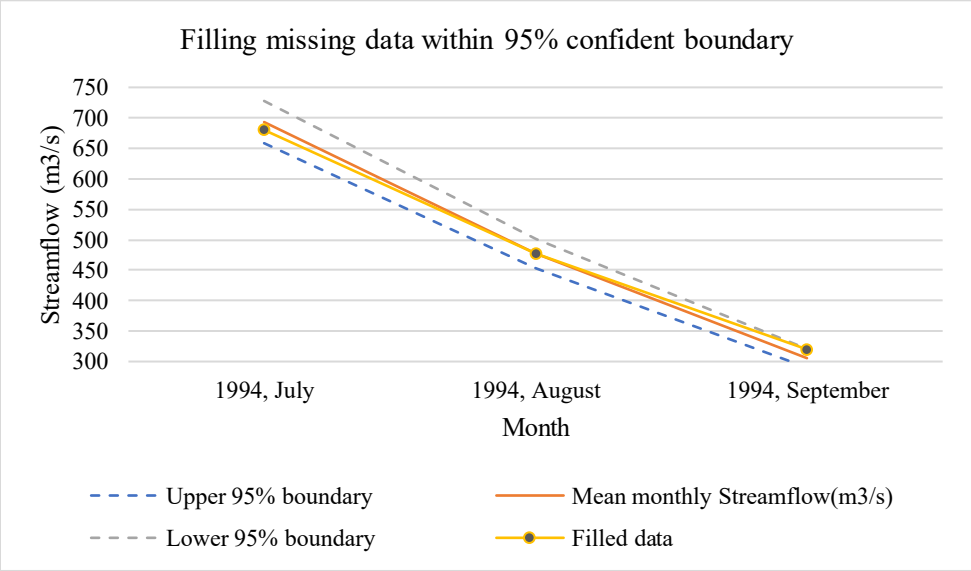


Figure A.1 Filling North Saskatchewan River at Prince Albert station data using North Saskatchewan River near Deer Creek station data.

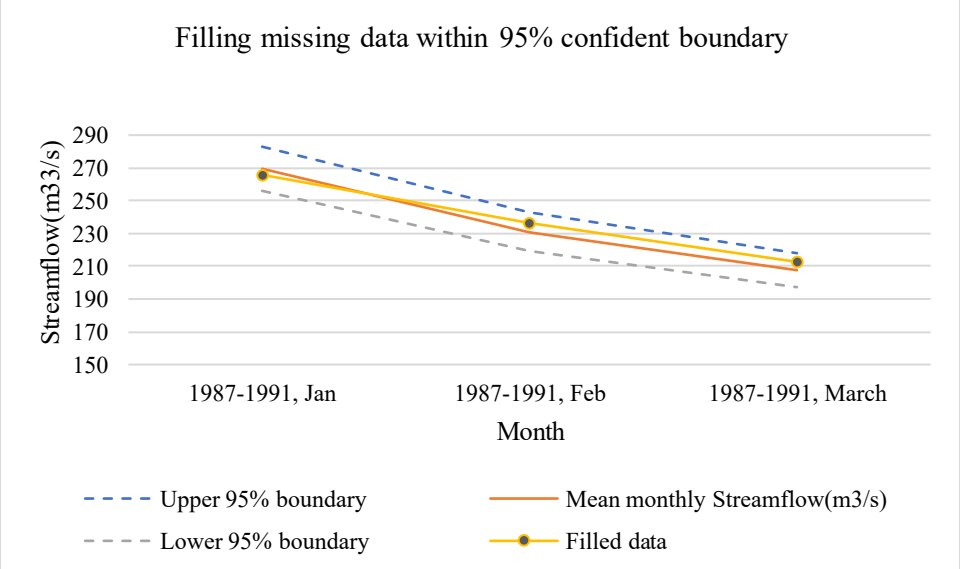


Figure A.2 Filling South Saskatchewan River at Saskatoon station data using South Saskatchewan River at St. Louis station data.

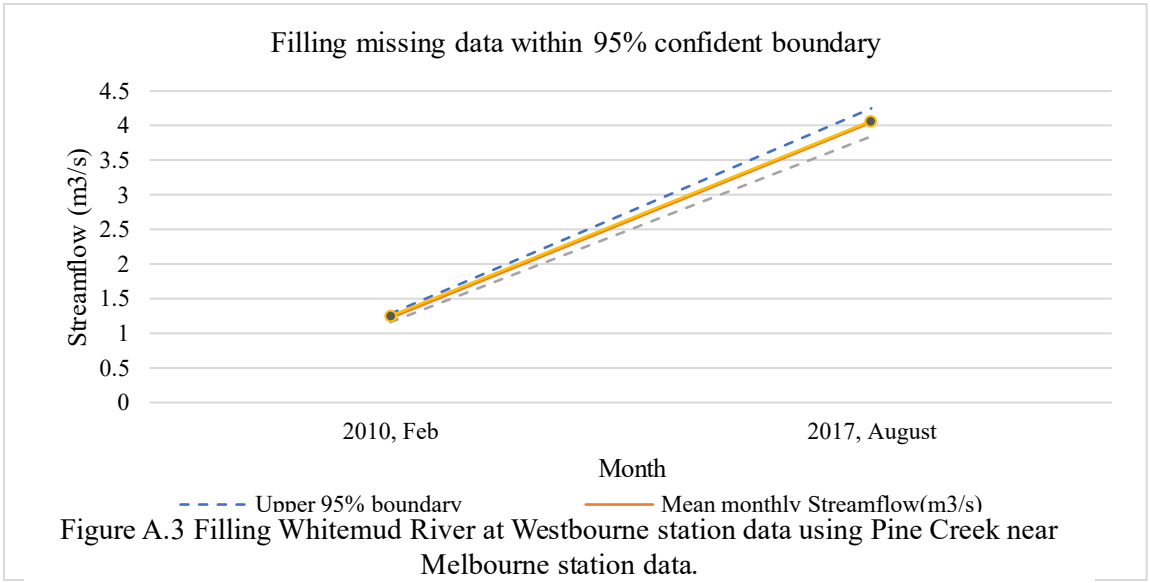


Figure A.3 Filling Whitemud River at Westbourne station data using Pine Creek near Melbourne station data.

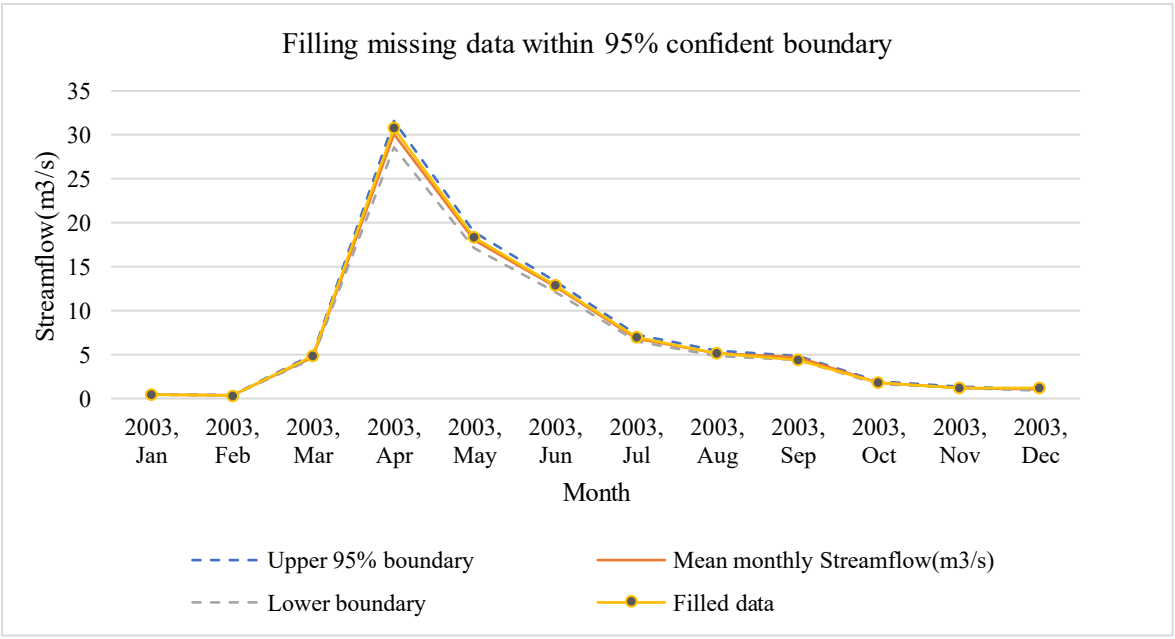


Figure A.4 Filling Pembina River at Neche station data using Pembina River near Windygates station data.

APPENDIX - B

Recognizing Seasonal Drought Patterns

The following figures show the Rankit comparison plots for normalized data in selected 25 watersheds, where the Box-Cox transformed data is compared with theoretically normally distributed data. This comparison helps in assessing how well the data conforms to a normal distribution after the Box-Cox transformation.

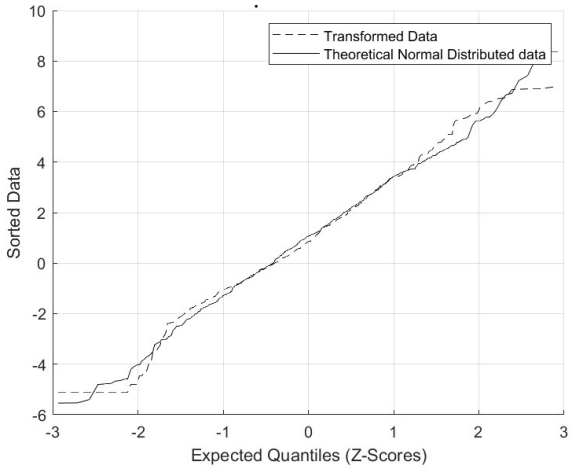


Figure B.01 Rankit Normalization Comparison for Pembina River at Neche

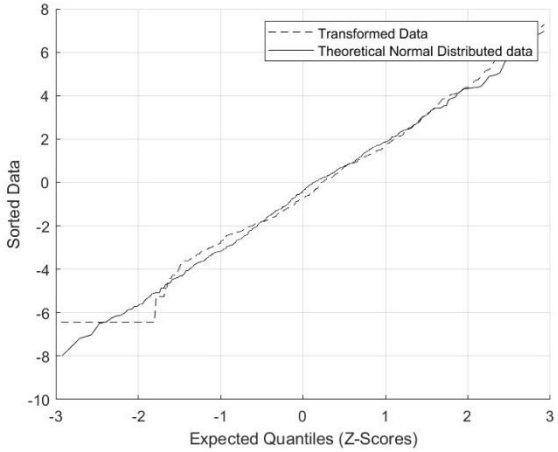


Figure B.02 Rankit Normalization Comparison for Souris River near Sherwood

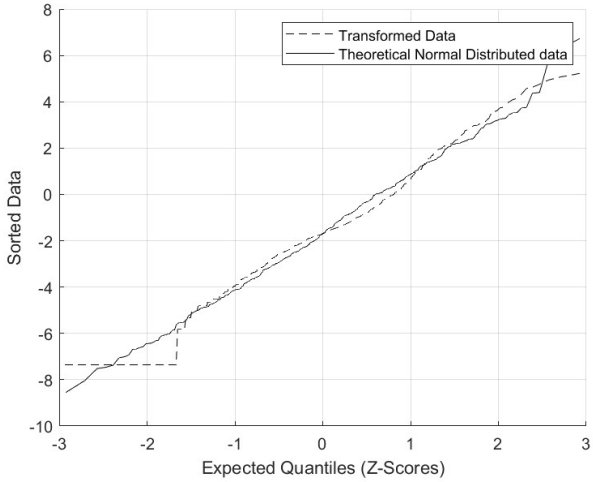


Figure B.113 Rankit Normalization Comparison for Moose Jaw River near Burdick

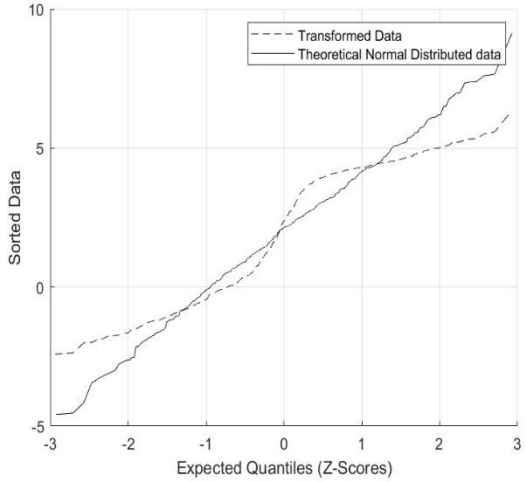


Figure B.04 Rankit Normalization Comparison for Little Saskatchewan River near Minnedosa

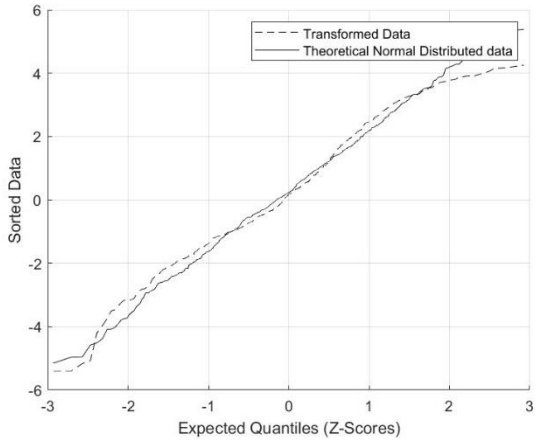


Figure B.05 Rankit Normalization Comparison for Assiniboine River at Kamsack

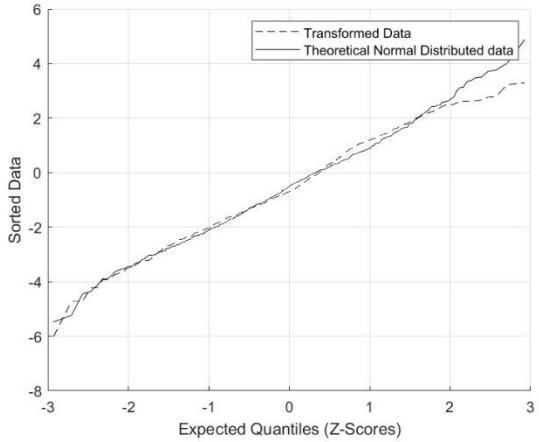


Figure B.06 Rankit Normalization Comparison for Battle River near Ponoka

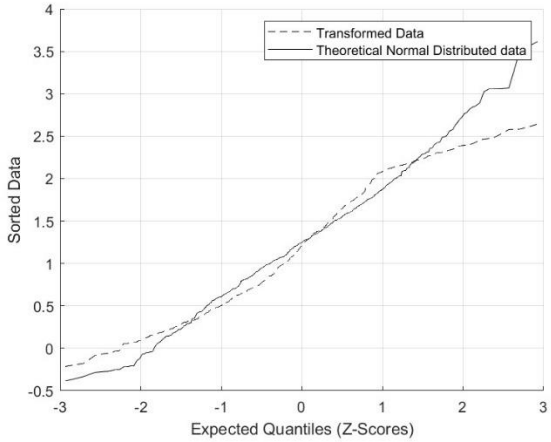


Figure B.07 Rankit Normalization Comparison for Belly River near Mountain View

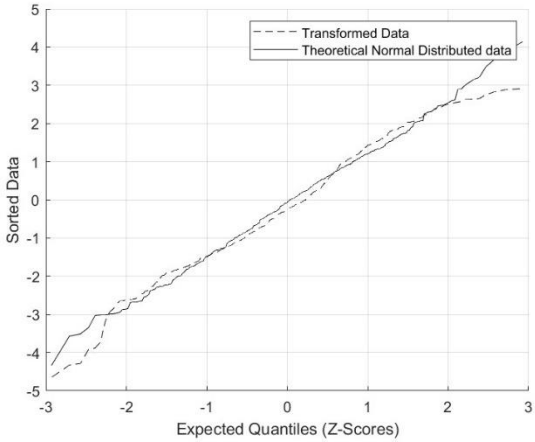


Figure B.08 Rankit Normalization Comparison for Blindman River near Blackfalds

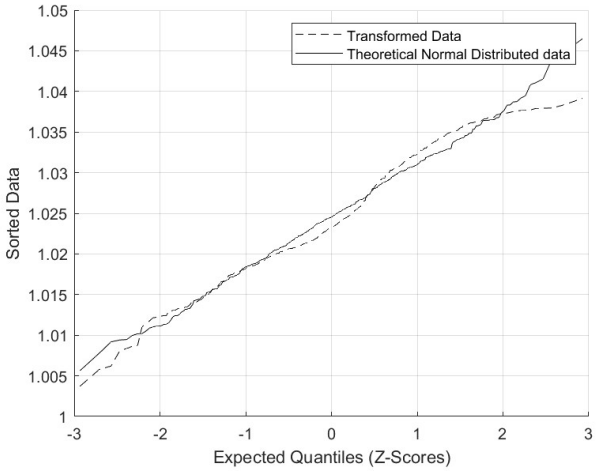


Figure B.09 Rankit Normalization Comparison for Bow River at Calgary

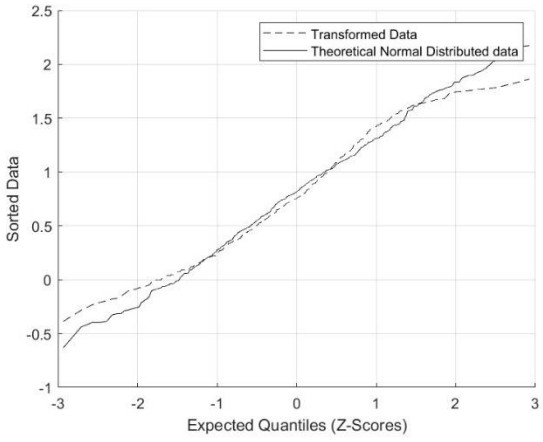


Figure B.10 Rankit Normalization Comparison for Crowsnest River at Frank

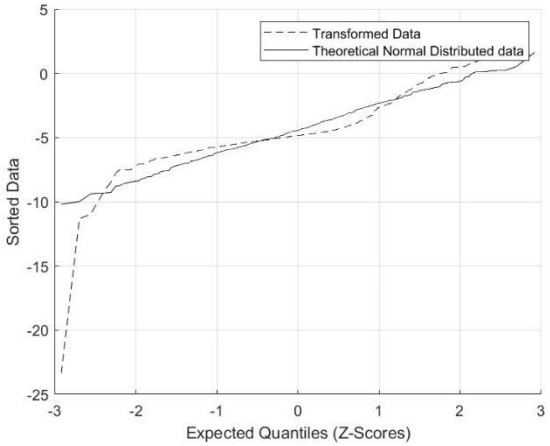


Figure B.11 Rankit Normalization Comparison for East Poplar River at International Boundary

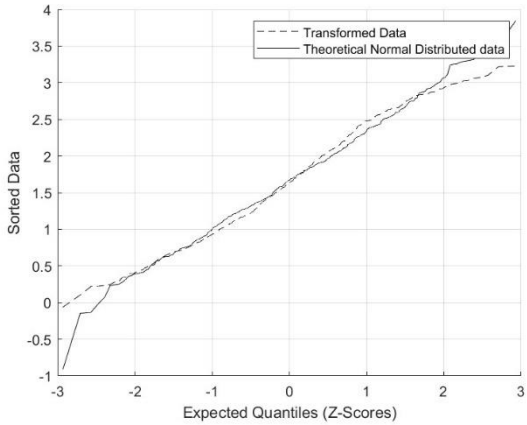


Figure B.12 Rankit Normalization Comparison for Highwood River near the Mouth

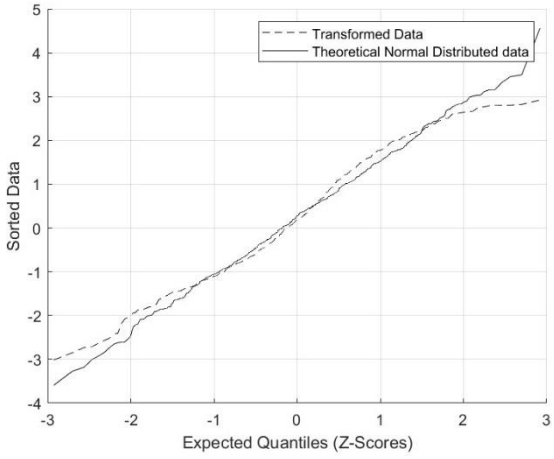


Figure B.13 Rankit Normalization Comparison for Medicine River near Eckville

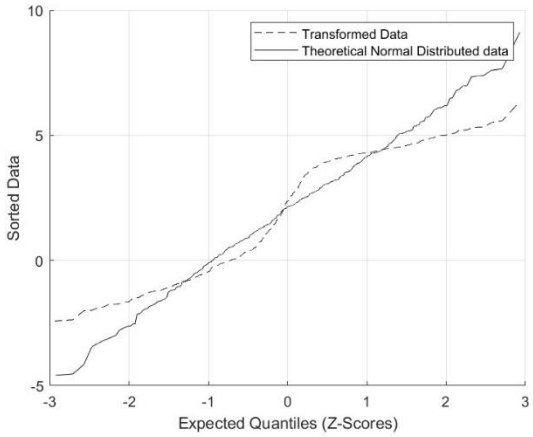


Figure B.14 Rankit Normalization Comparison for Milk River at Milk River

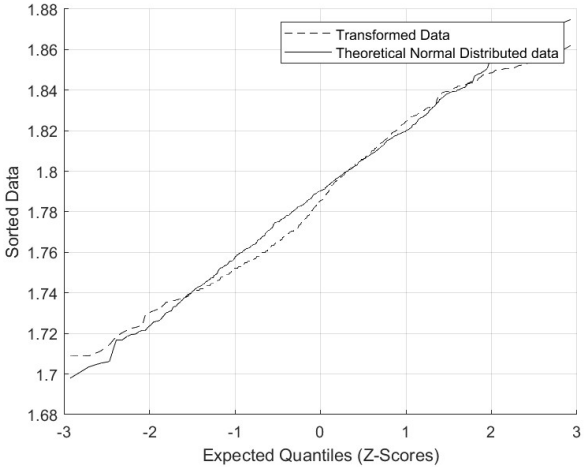


Figure B.15 Rankit Normalization Comparison for North Saskatchewan River at Prince Albert

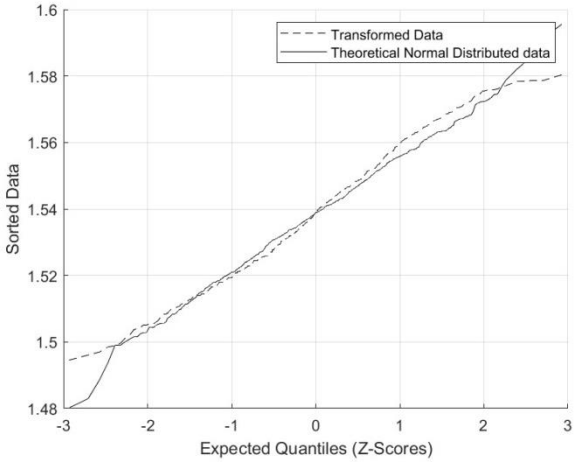


Figure B.16 Rankit Normalization Comparison for North Saskatchewan River near Deer Creek

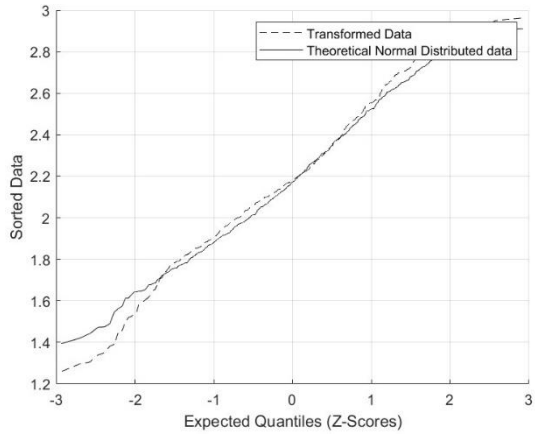


Figure B.17 Rankit Normalization Comparison for Oldman River near Lethbridge

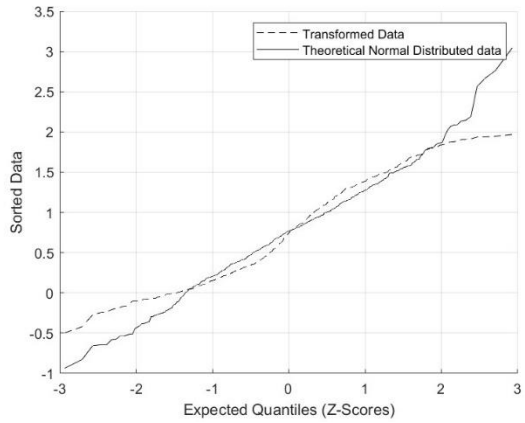


Figure B.18 Rankit Normalization Comparison for Qu'appelle River near Lumsden

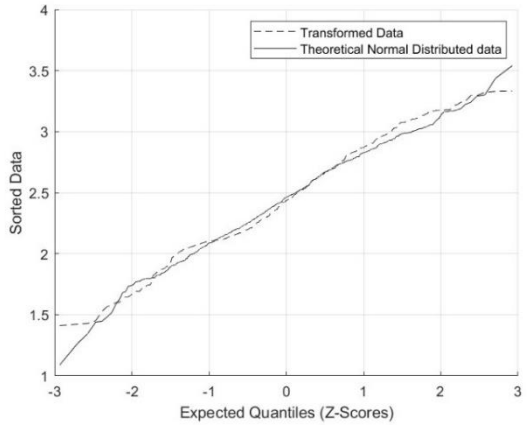


Figure B.19 Rankit Normalization Comparison for Red Deer River at Drumheller

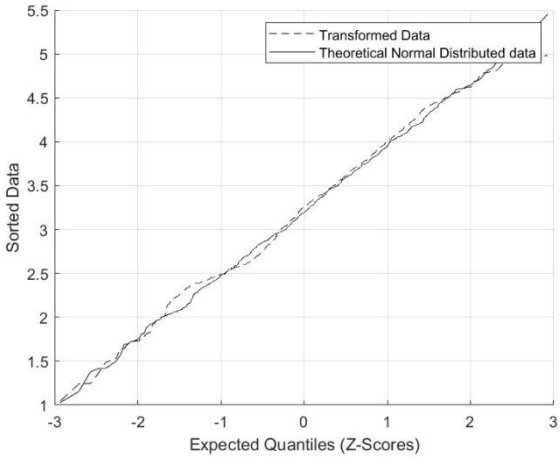


Figure B.20 Rankit Normalization Comparison for Red Deer River near Bindloss

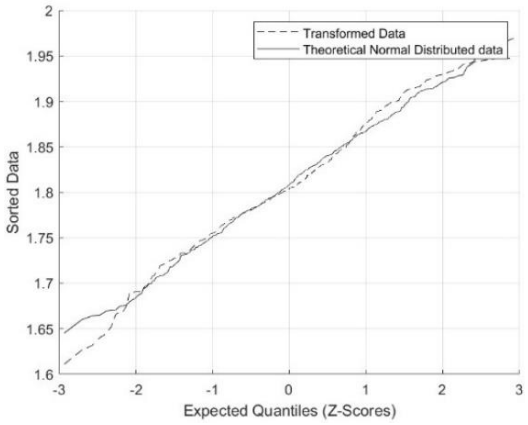


Figure B.21 Rankit Normalization Comparison for South Saskatchewan River at Medicine Hat

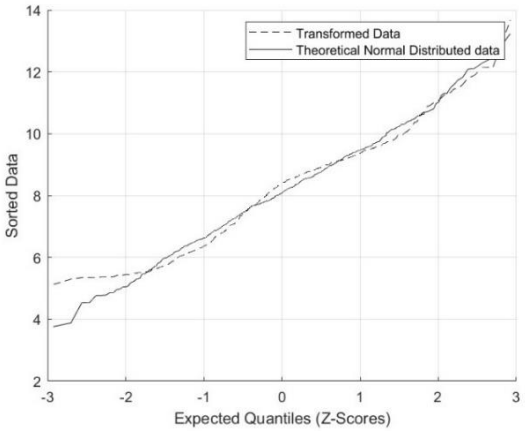


Figure B.22 Rankit Normalization Comparison for South Saskatchewan River at Saskatoon

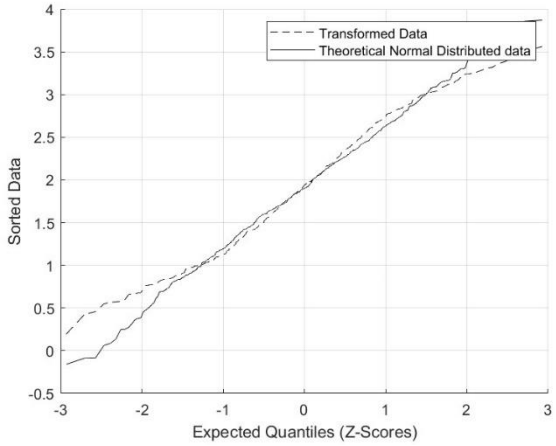


Figure B.23 Rankit Normalization Comparison for St. Mary River at International Boundary

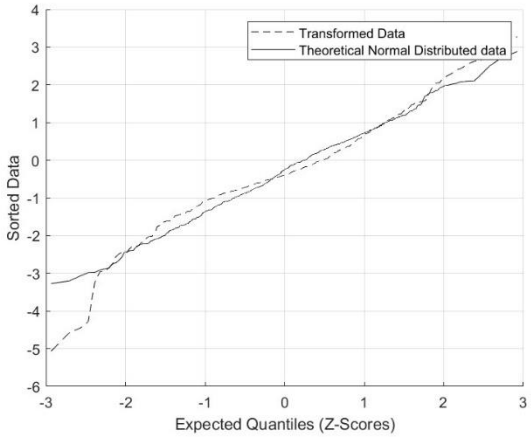


Figure B.24 Rankit Normalization Comparison for Swift Current Creek below Rock Creek

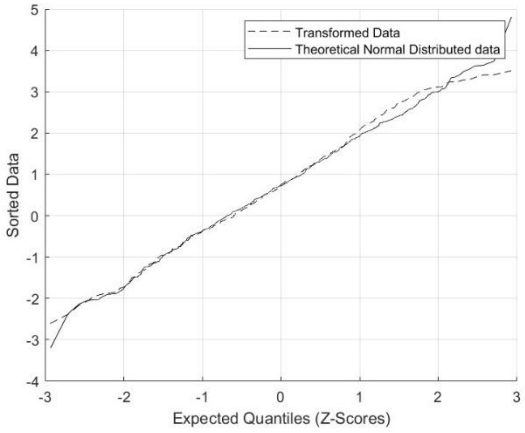


Figure B.25 Rankit Normalization Comparison for Whitemud River at Westbourne

Shannon Entropy Calculations

The following figures illustrate the entropy calculations conducted in this study. The primary objective was to achieve minimal intra-entropy, or entropy within the groups, to ensure that each group remains as internally cohesive as possible. Simultaneously, the goal was to maximize inter-entropy, or entropy among the groups, to ensure that the groups are externally distinguishable from one another.

Intra- and inter-entropy calculations were performed for these different group lengths of 2 months, 3 months, 4 months, or 6 months based on the Streamflow Drought Index (SDI). The intra-entropy was calculated using Shannon Entropy, considering the Drought Level (DL) of each month within a group.

Based on the entropy calculations, it appears that some rivers naturally form 6-month groups, resulting in two similar groups per year—typically classified as either wet or dry groups, or more specifically as drought and no drought groups. However, since this research focuses exclusively on drought conditions, it will prioritize the next optimal entropy calculation results, which correspond to 3-month groupings. This choice of 3-month groups divides the hydrological year into four distinct seasonal patterns, allowing for a more detailed and focused analysis of drought patterns throughout the year.

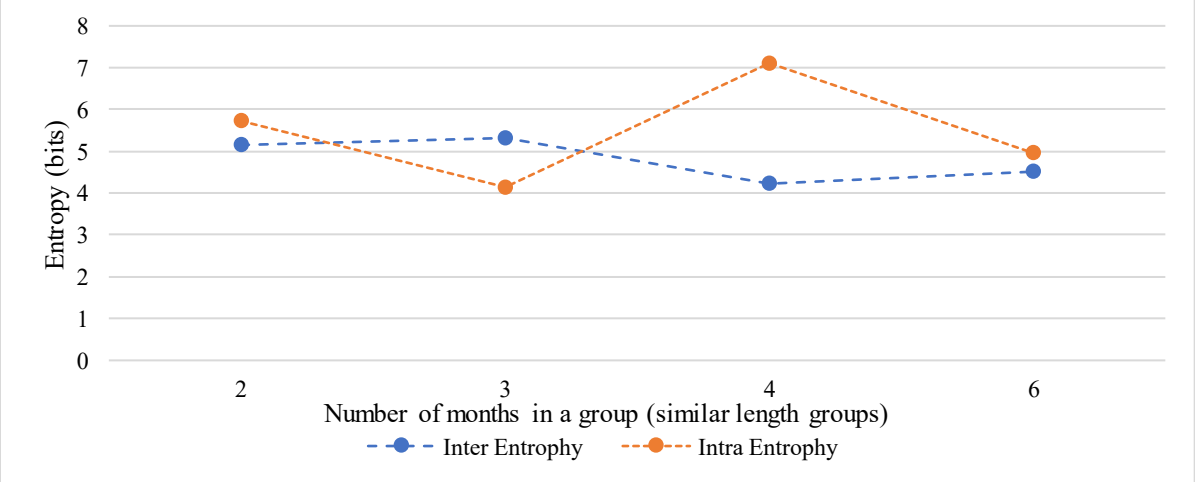


Figure B.26 Intra and Inter entropy calculation for groups with different lengths for Pembina River at Neche

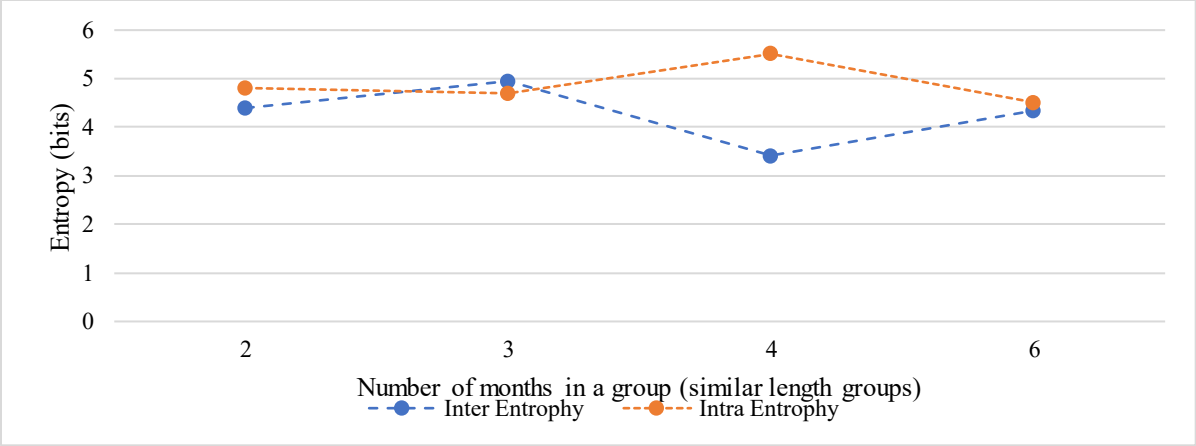


Figure B.27 Intra and Inter entropy calculation for groups with different length for Souris River near Sherwood

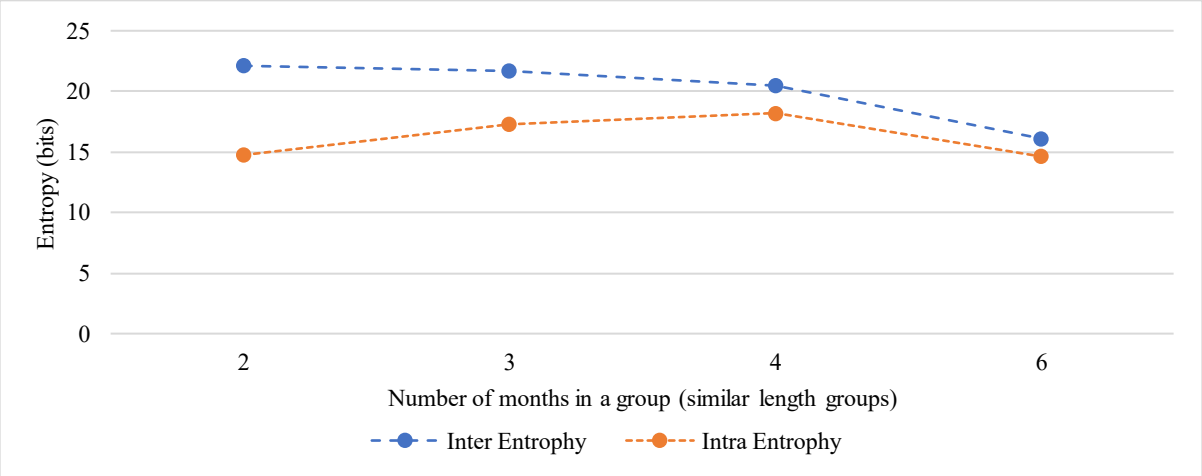


Figure B.28 Figure 0.24 Intra and Inter entropy calculation for groups with different length for Assiniboine River at Kamsack

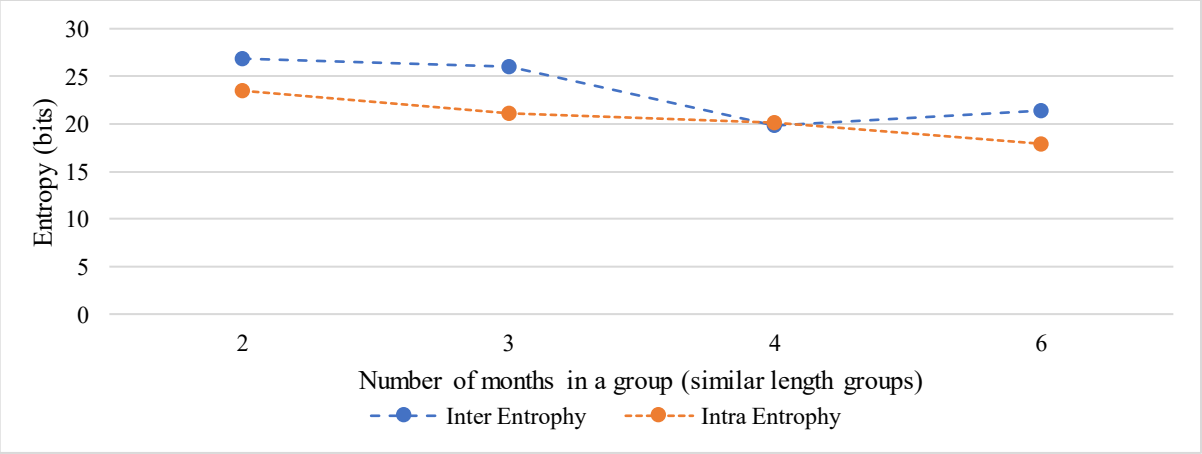


Figure B.29 Intra and Inter entropy calculation for groups with different lengths for Battle River near Ponoka

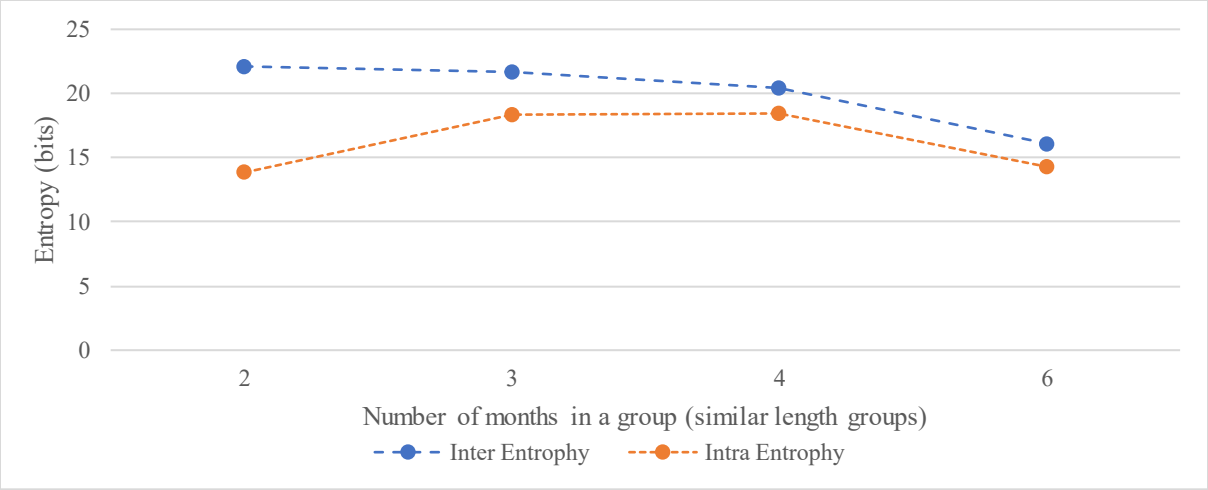


Figure B.30 Intra and Inter entropy calculation for groups with different lengths for Belly River near Mountain View

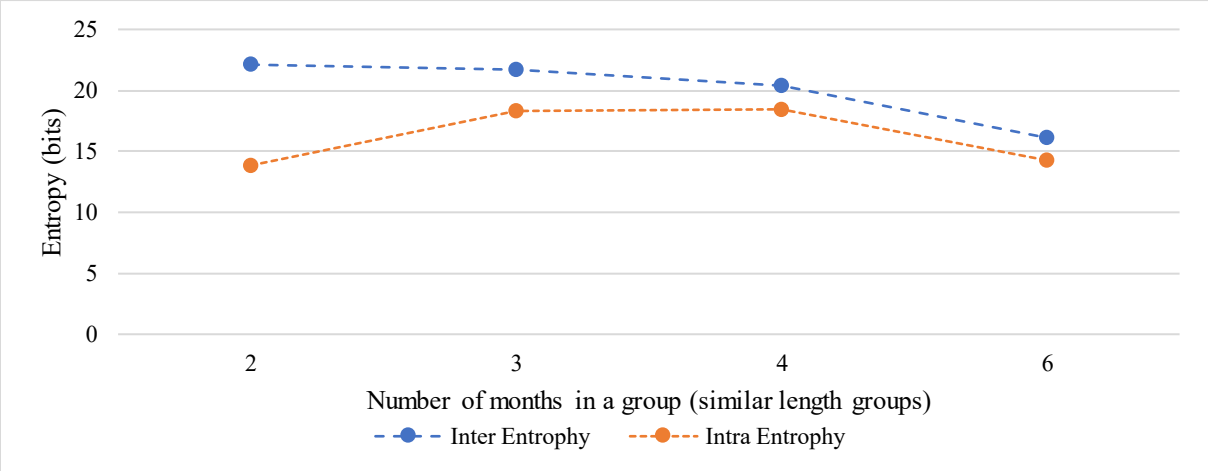


Figure B.31 Intra and Inter entropy calculation for groups with different lengths for Blindman River near Blackfalds

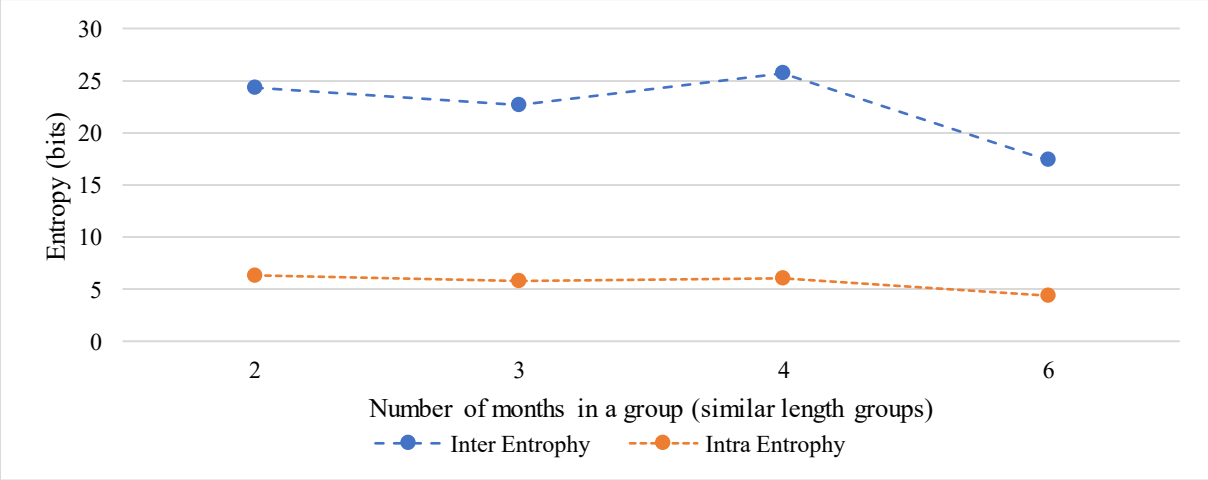


Figure B.32 Intra and Inter entropy calculation for groups with different lengths for Bow River at Calgary

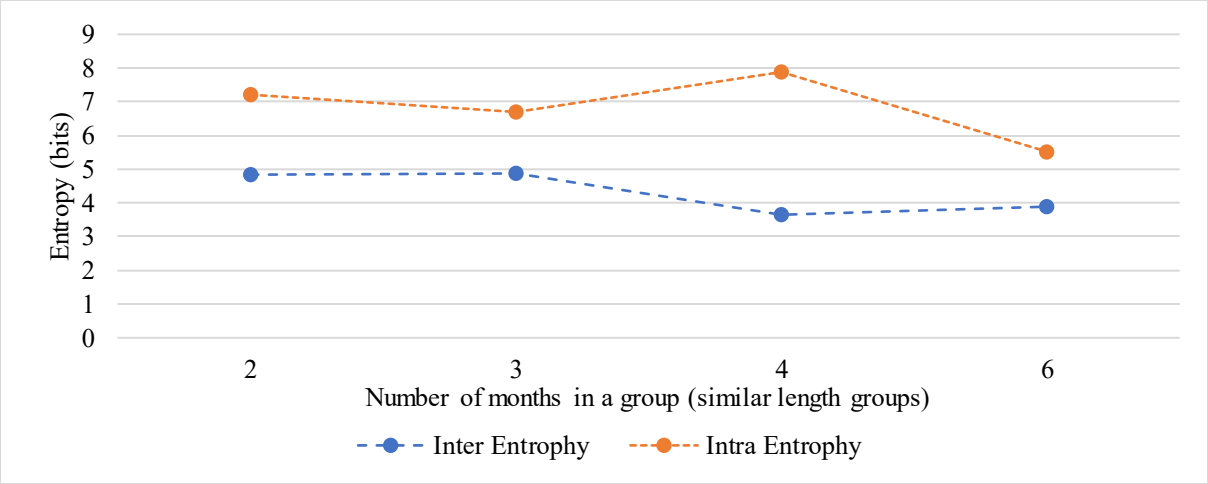


Figure B.33 Intra and Inter entropy calculation for groups with different lengths for Crowsnest River at Frank

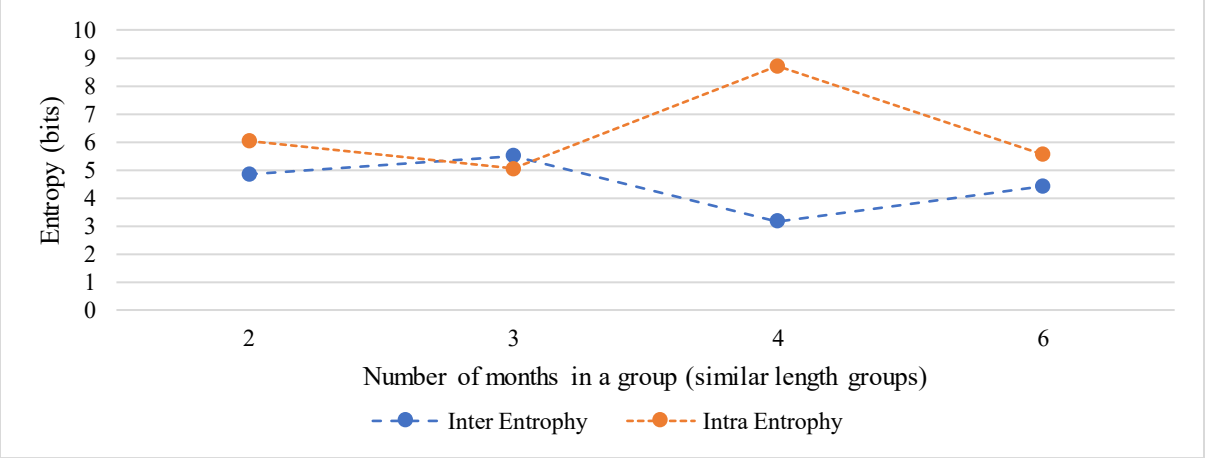


Figure B.34 Intra and Inter entropy calculation for groups with different lengths for Little Saskatchewan River near Minnedosa

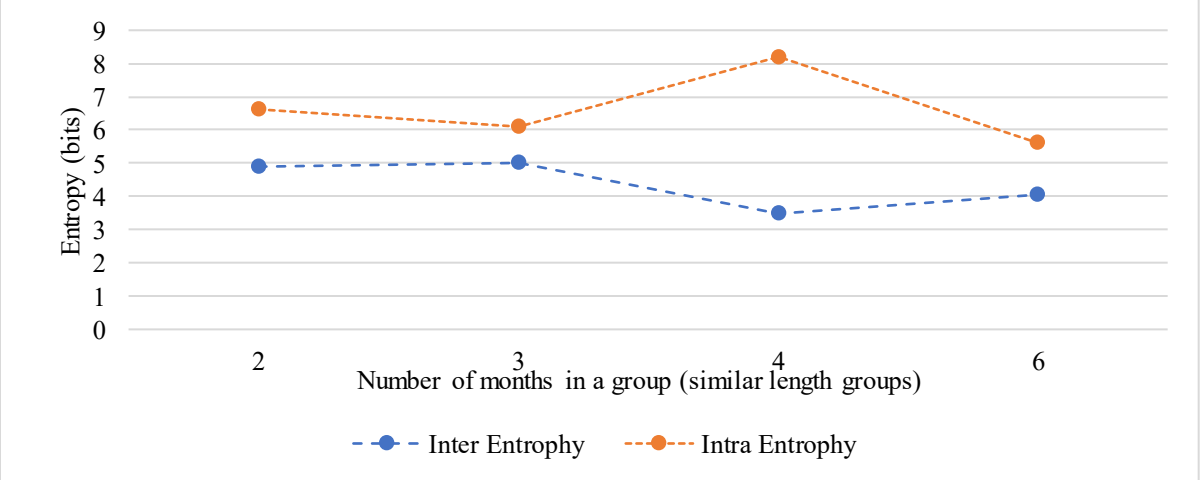


Figure B.35 Intra and Inter entropy calculation for groups with different lengths for Medicine River near Eckville

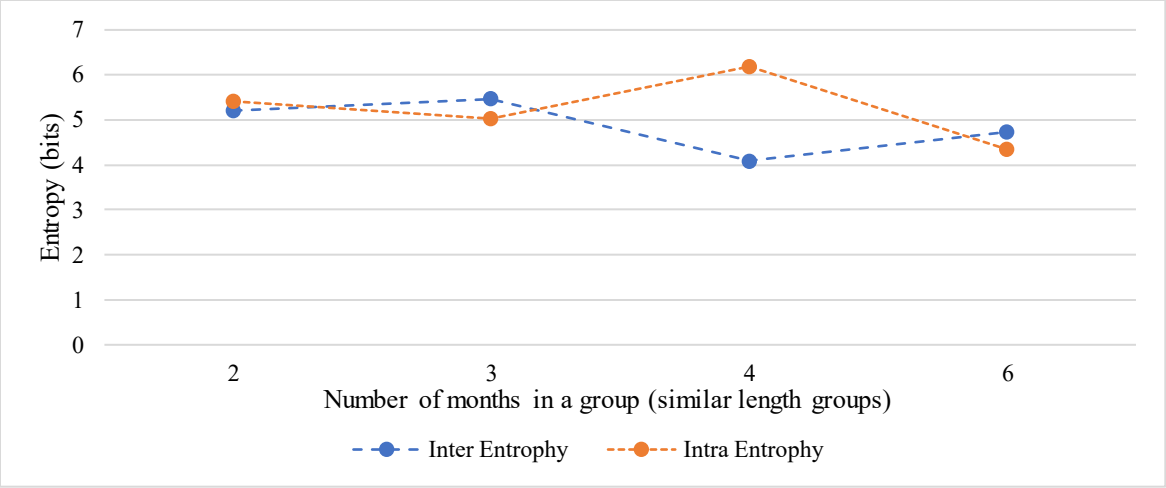


Figure B.36 Intra and Inter entropy calculation for groups with different lengths for Milk River at Milk River

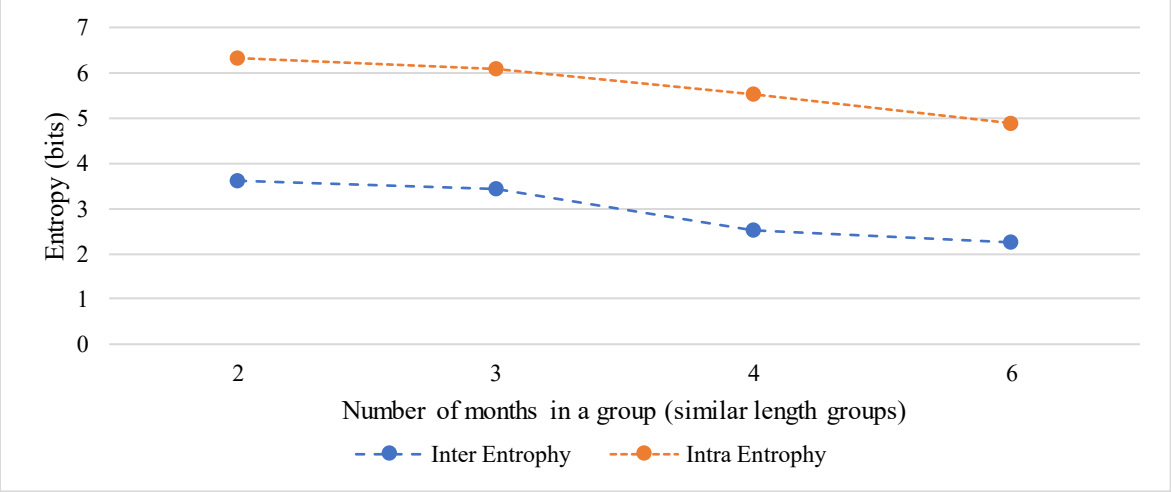


Figure B.37 Intra and Inter entropy calculation for groups with different lengths for Moose Jaw River near Burdick

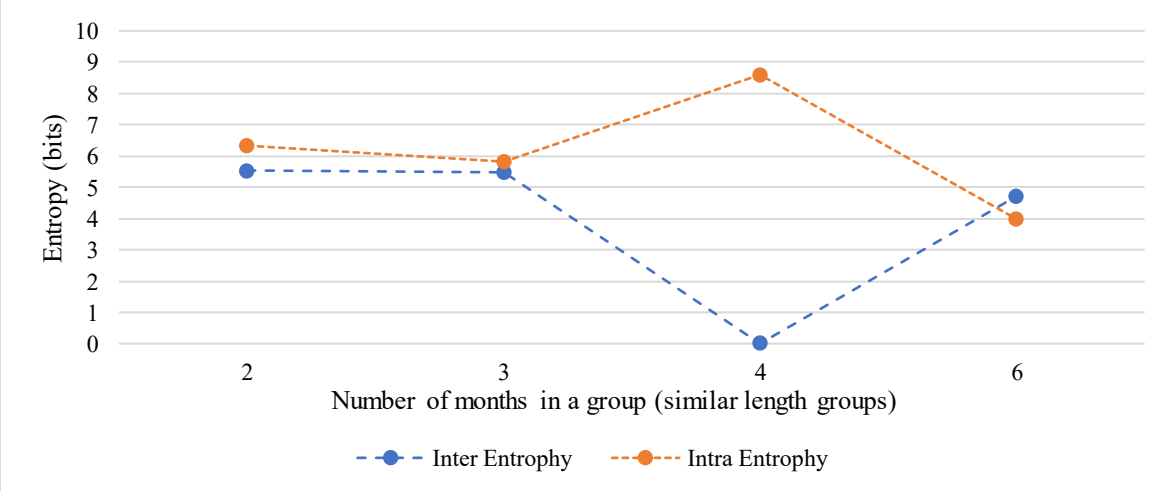


Figure B.38 Intra and Inter entropy calculation for groups with different lengths for the North Saskatchewan River at Prince Albert

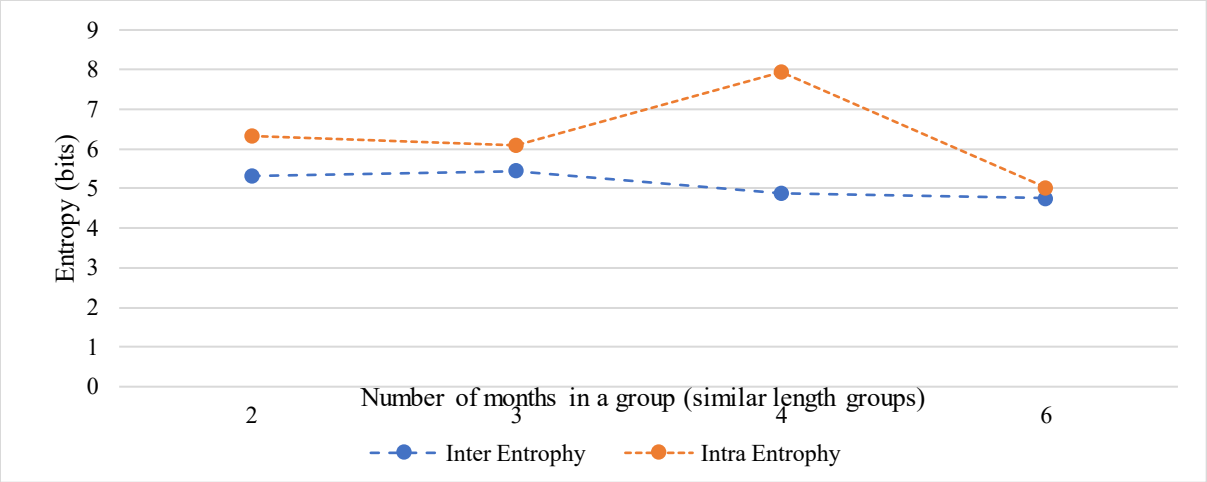


Figure B.39 Intra and Inter entropy calculation for groups with different lengths for the North Saskatchewan River near Deer Creek

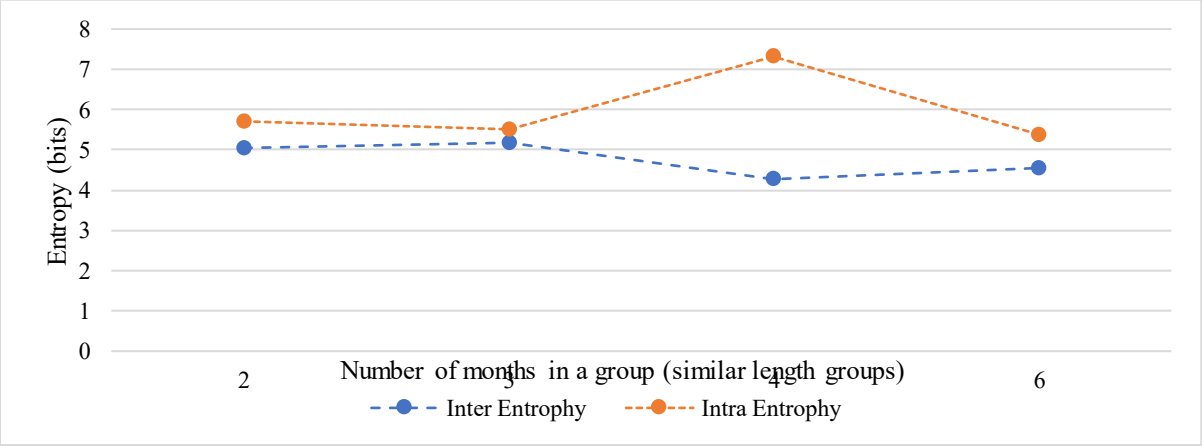


Figure B.40 Intra and Inter entropy calculation for groups with different lengths for Oldman River near Lethbridge

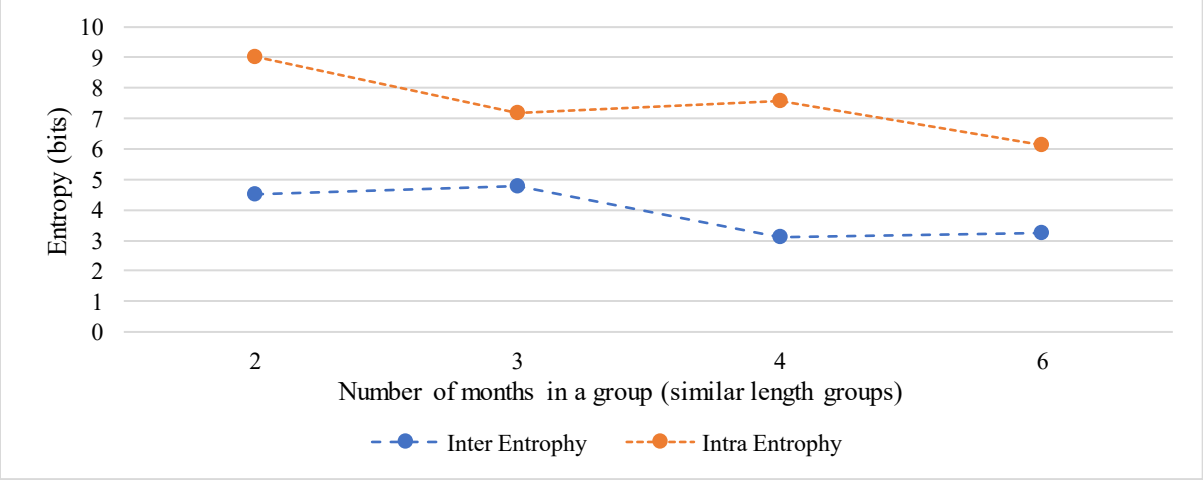


Figure B.41 Intra and Inter entropy calculation for groups with different lengths for Qu'Appelle River near Lumsden

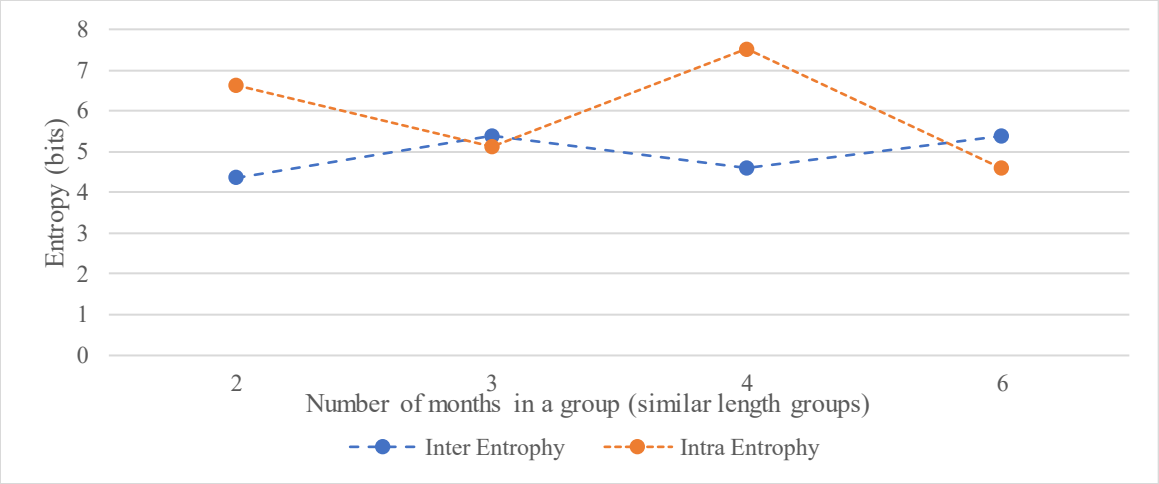


Figure B.42 Intra and Inter entropy calculation for groups with different lengths for Red Deer River at Drumheller

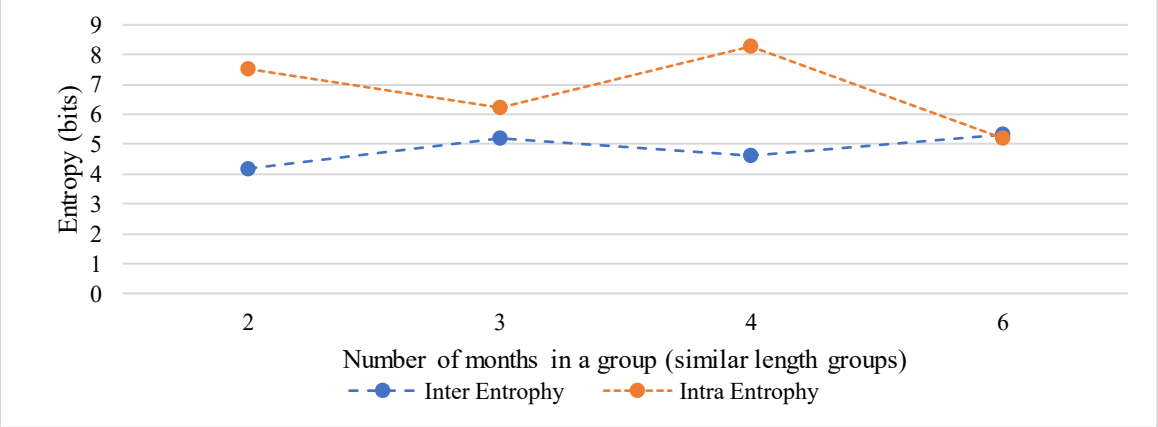


Figure B.43 Intra and Inter entropy calculation for groups with different lengths for Red Deer River near Bindloss

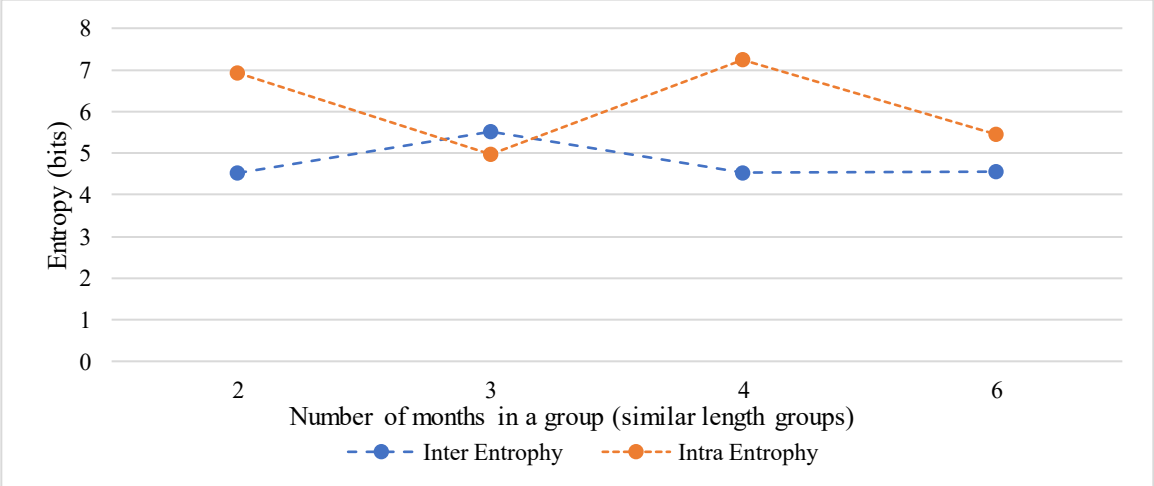


Figure B.45 Intra and Inter entropy calculation for groups with different lengths for the South Saskatchewan River at Medicine Hat

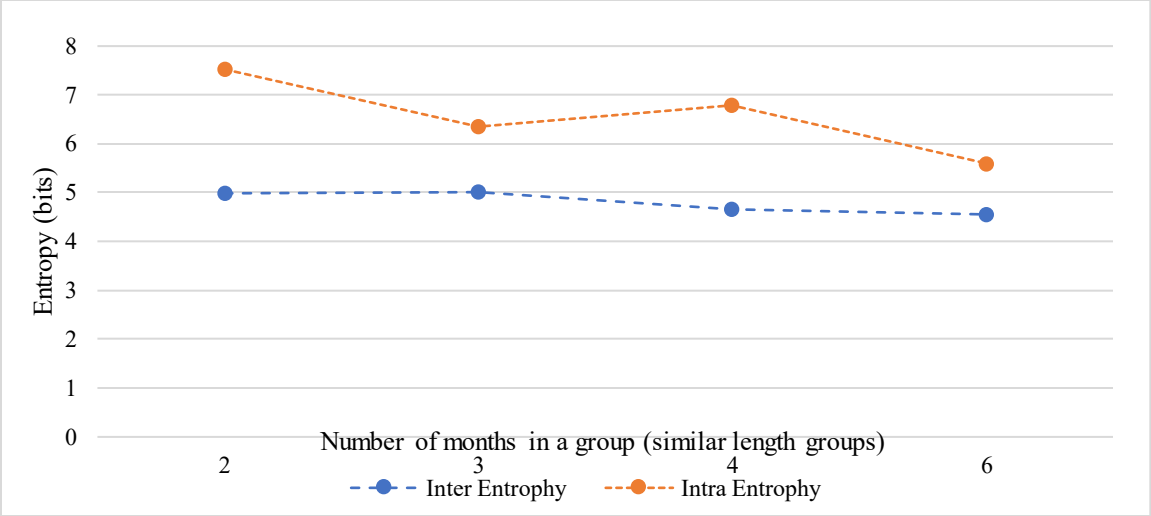


Figure B.46 Intra and Inter entropy calculation for groups with different lengths for Swift Current Creek below Rock Creek

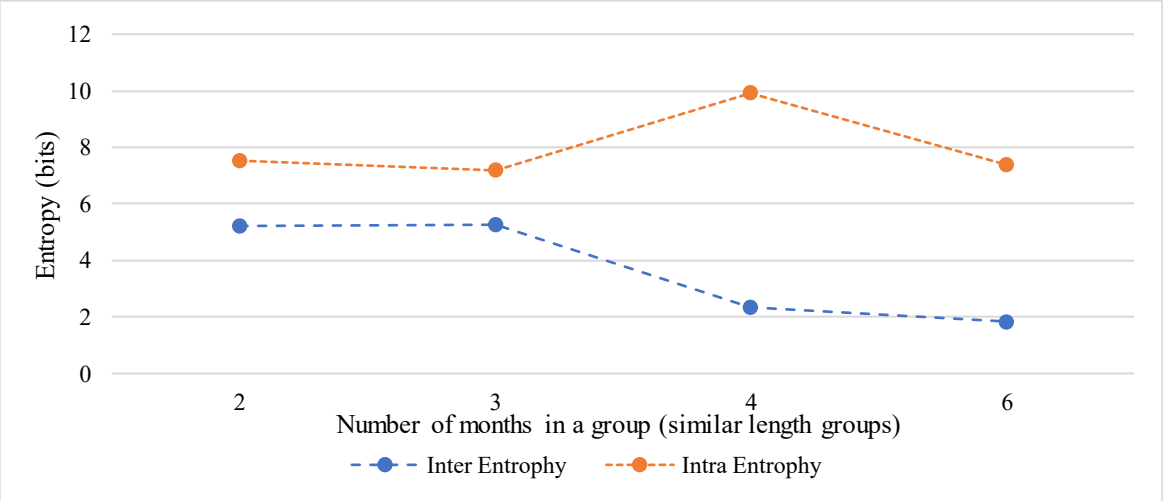


Figure B.47 Intra and Inter entropy calculation for groups with different lengths for Whitemud River at Westbourne

Moving Average Calculations

The summarized graph of all 25 watersheds reveals distinct seasonal drought characteristics. On average, the months of October, November, and December typically exhibit no drought or mild drought conditions. From January to March, the conditions generally remain drought-free. However, the period from April to June shows a transition from no drought to more severe drought levels, indicating the onset of drier conditions. Conversely, the months from July to September display a shift from severe drought back to mild drought conditions, suggesting a gradual easing of drought severity as the hydrological year progresses.

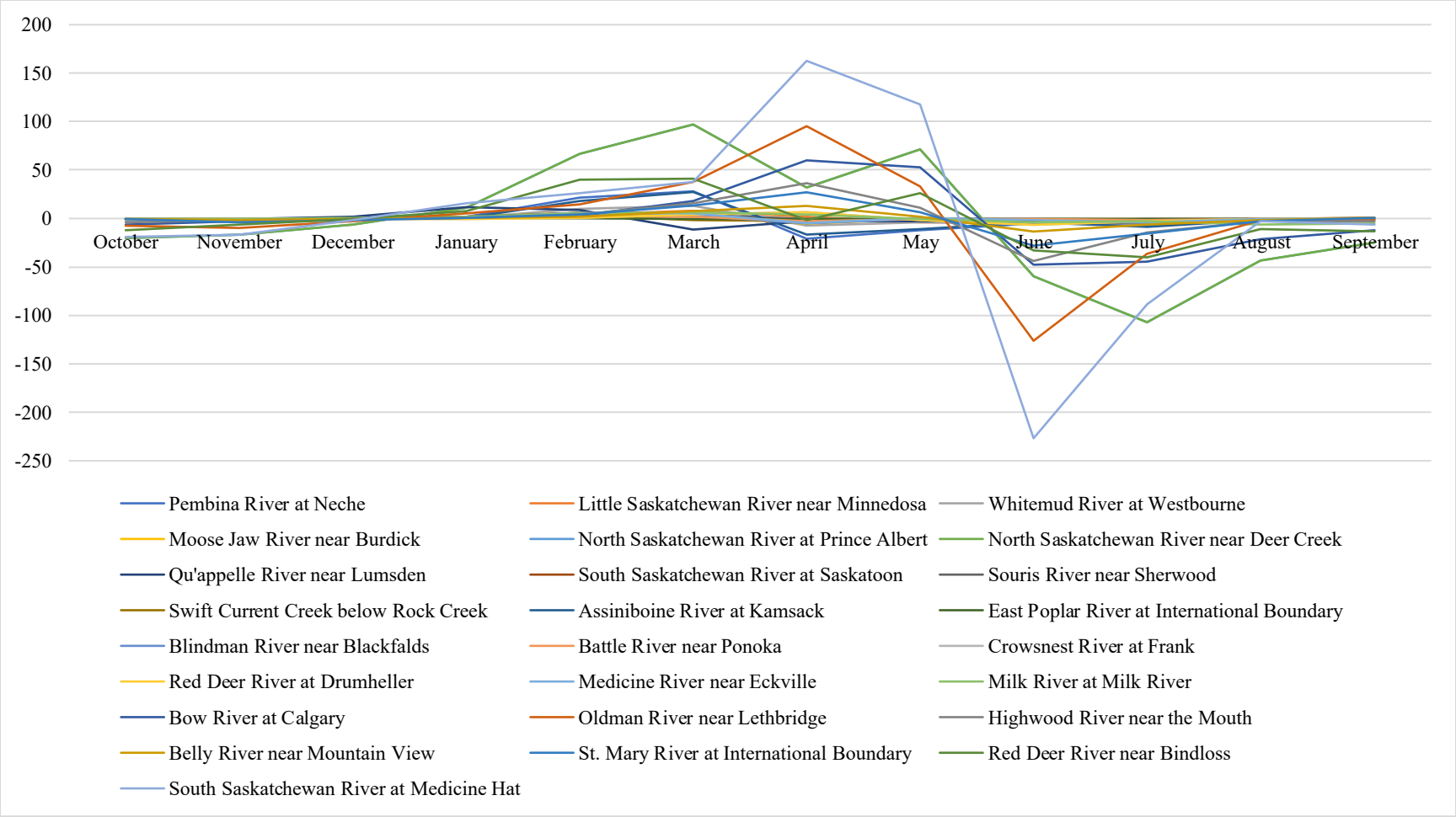


Figure B.48 Average Monthly streamflow variations for selected 25 watersheds

K-means clustering

The average characteristics of the seasonal patterns are as, Seasonal Pattern 1 is generally characterized by mild drought conditions, Seasonal Pattern 2 exhibits no drought, Seasonal Pattern 3 experiences more severe drought conditions, and Seasonal Pattern 4 returns to a mild drought pattern. However, these drought patterns can vary significantly from year to year. To analyze these deviations of drought levels within the average patterns, K-means clustering was applied to identify distinct clusters within each seasonal pattern. This clustering approach provides a detailed analysis of how drought levels can change within a seasonal pattern. The silhouette scores, calculated for the 25 watersheds, are shown below.

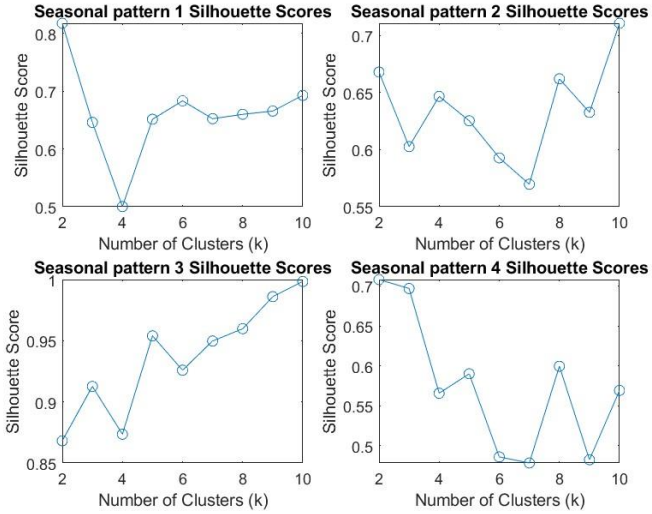


Figure B.49 Silhouette Score for seasonal patterns in Assiniboine River at Kamsack

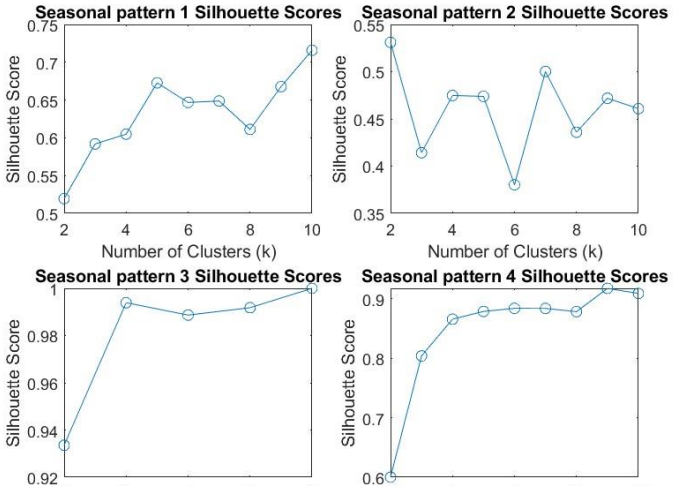


Figure B.50 Silhouette Score for seasonal patterns in Belly River near Mountain View

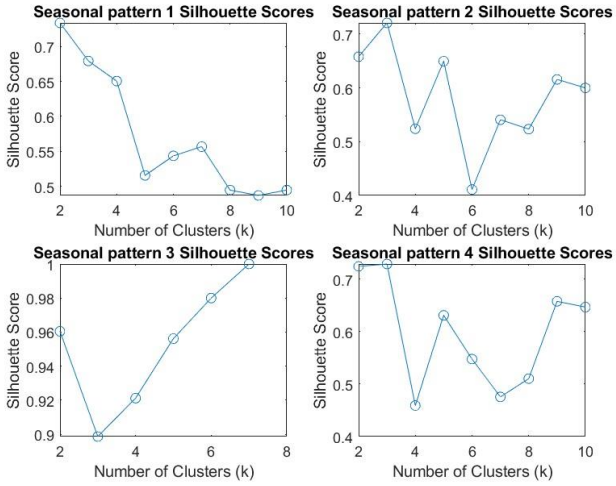


Figure B.51 Silhouette Score for seasonal patterns in Blindman River near Blackfalds

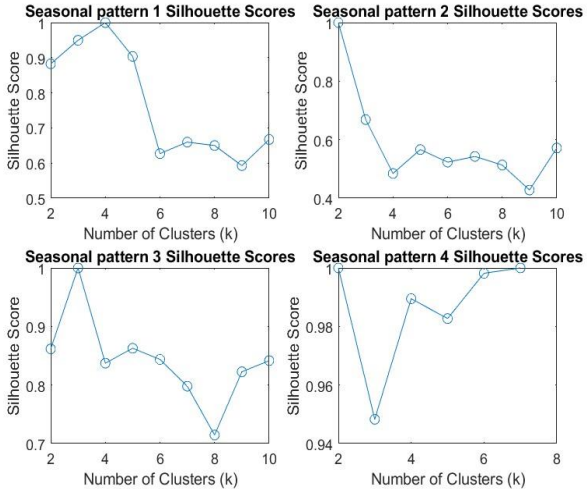


Figure B.52 Silhouette Score for seasonal patterns in Bow River at Calgary

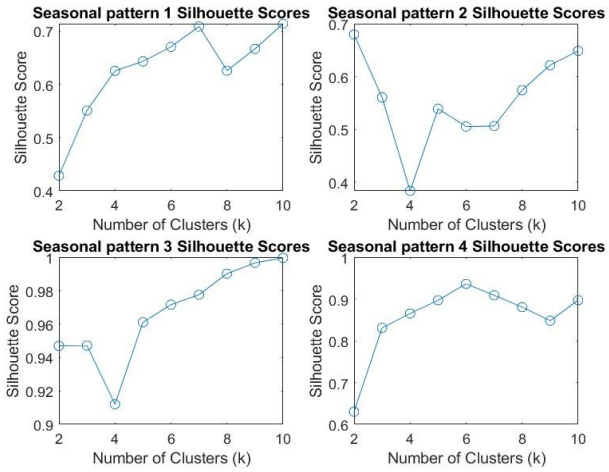


Figure B.53 Silhouette Score for seasonal patterns in Crowsnest River at Frank

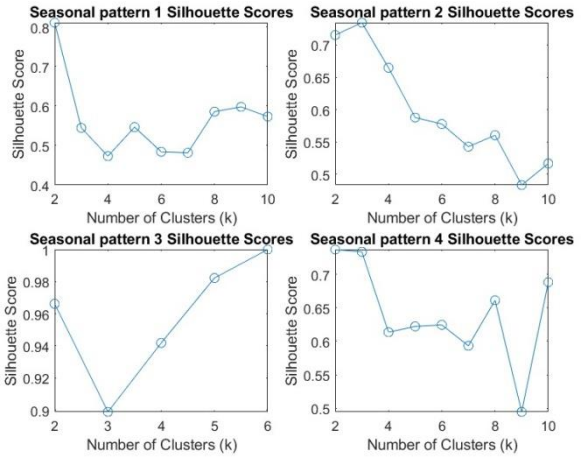


Figure B.54 Silhouette Score for seasonal patterns in Little Saskatchewan River near Minnedosa

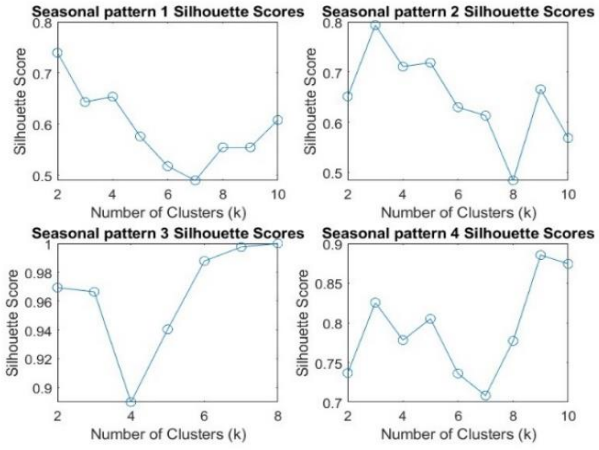


Figure B.55 Silhouette Score for seasonal patterns in Medicine River near Eckville

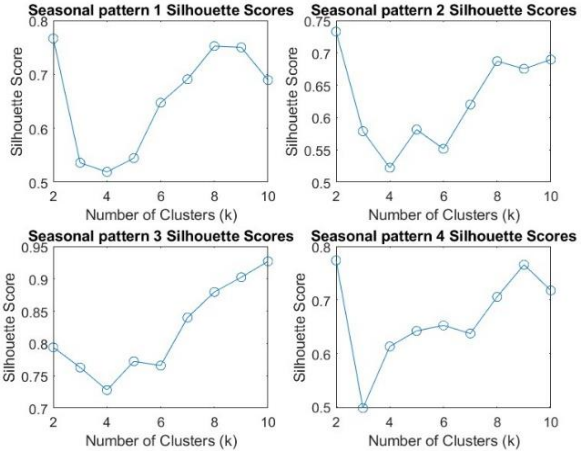


Figure B.56 Silhouette Score for seasonal patterns in Moose Jaw River near Burdick

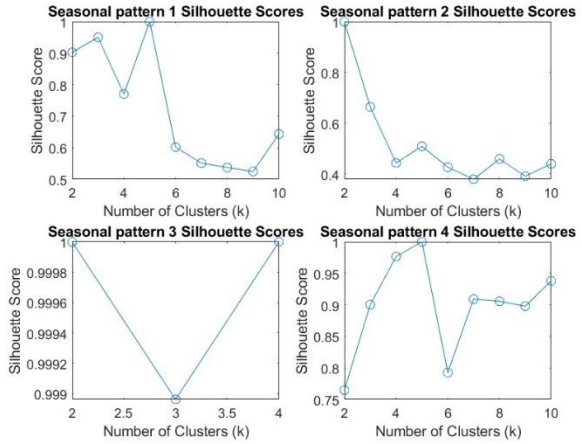


Figure B.571 Silhouette Score for seasonal patterns in North Saskatchewan River near Deer Creek

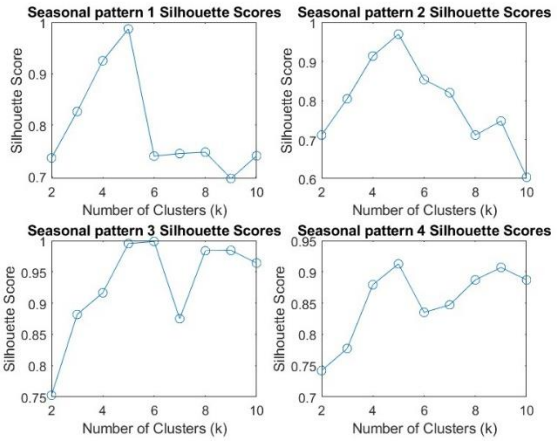


Figure B.58 Silhouette Score for seasonal patterns in North Saskatchewan River at Prince Albert

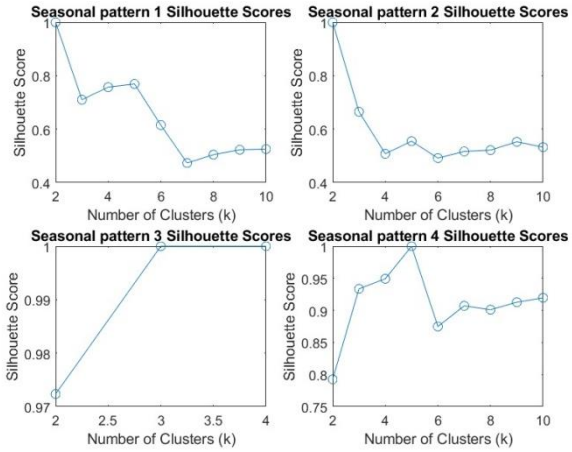


Figure B.59 Silhouette Score for seasonal patterns in Oldman River near Lethbridge

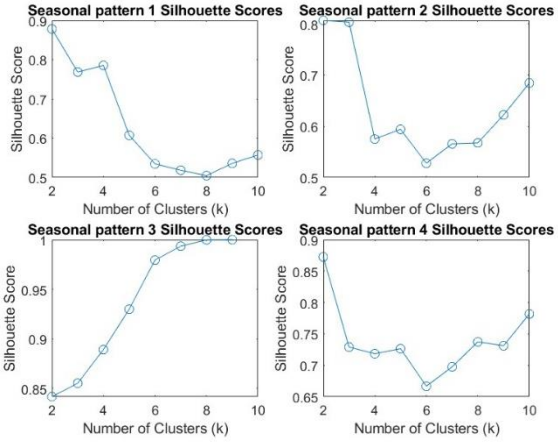


Figure B.60 Silhouette Score for seasonal patterns in Pembina River at Neche.

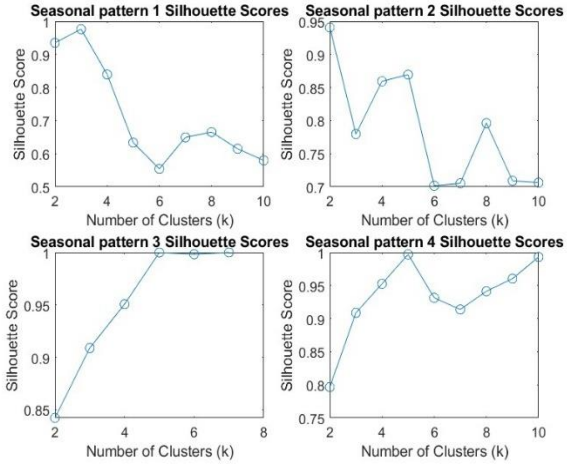


Figure B.61 Silhouette Score for seasonal patterns in Red Deer River at Drumheller.

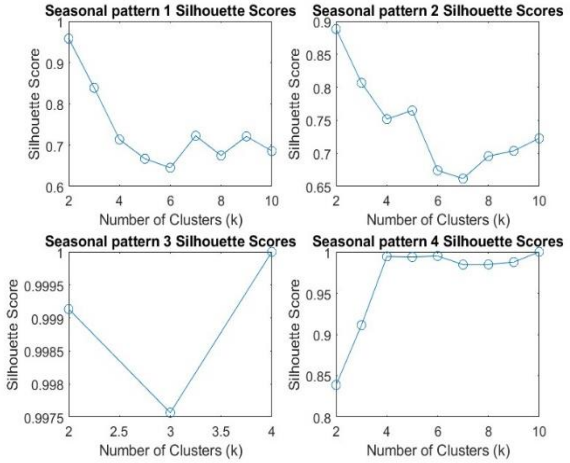


Figure B.62 Silhouette Score for seasonal patterns in Red Deer River near Bindloss.

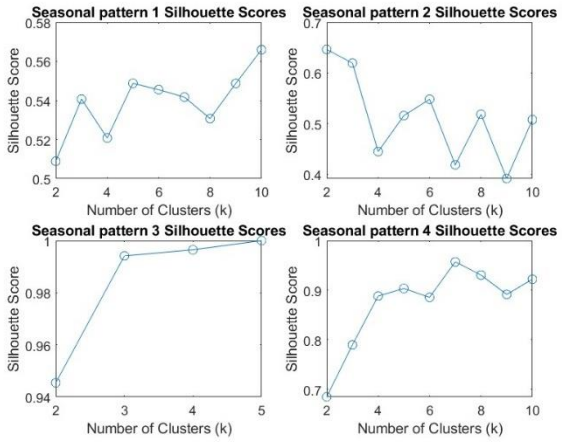


Figure B.63 Silhouette Score for seasonal patterns in Qu'appelle River near Lumsden

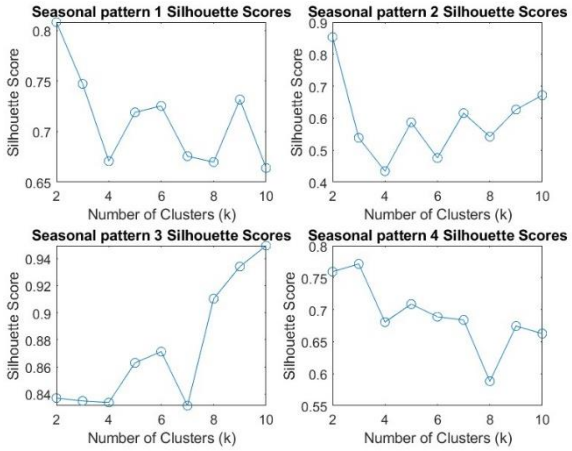


Figure B.64 Silhouette Score for seasonal patterns in Souris River near Sherwood

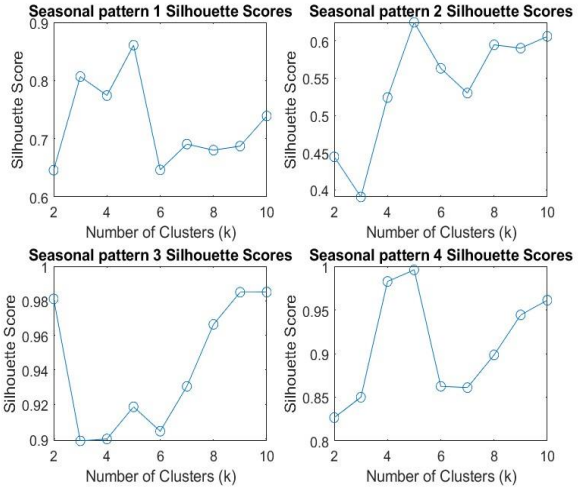


Figure B.65 Silhouette Score for seasonal patterns in South Saskatchewan River at Medicine Hat

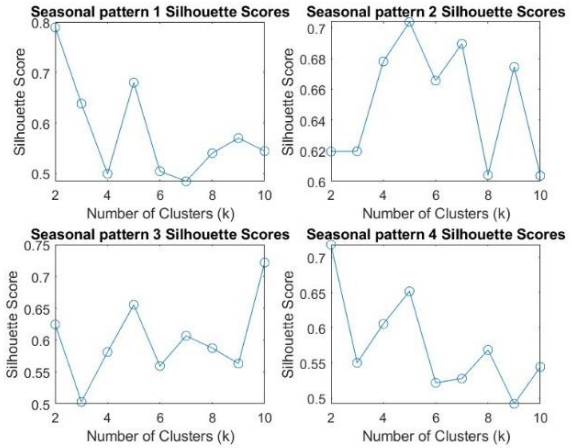


Figure B.66 Silhouette Score for seasonal patterns in St. Mary River at International Boundary

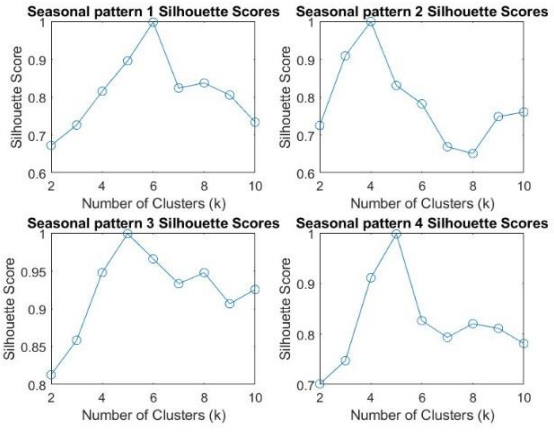


Figure B.67 Silhouette Score for seasonal patterns in Swift Current Creek below Rock Creek

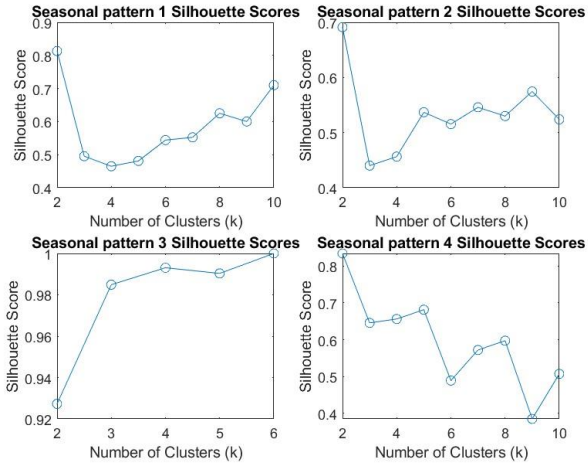


Figure B.68 Silhouette Score for seasonal patterns in Whitemud River at Westbourne

APPENDIX - C

Forecasting using Markov Models and LSTM Models

Hidden Markov Model Forecast Results

The following figures illustrate the forecast results of the Hidden Markov Model (HMM) for all 25 watersheds. These figures reveal a consistent trend where the HMM model tends to predict values that are either higher or lower than the actual observations. This tendency indicates that the model might be overestimating or underestimating the drought levels, leading to a systematic deviation from the observed data. This behaviour suggests that while the HMM can identify general patterns, it may struggle with accurately capturing the nuances of drought severity, resulting in forecasts that are not fully aligned with observed conditions.

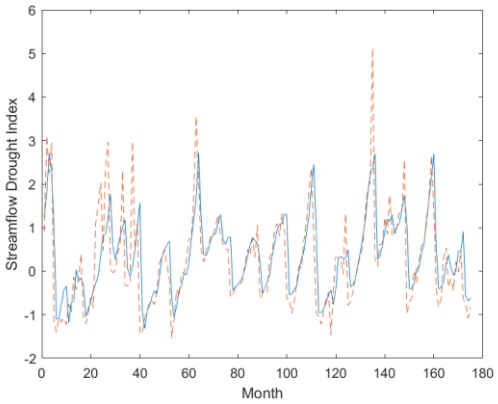


Figure C.02 HMM prediction results for Souris River

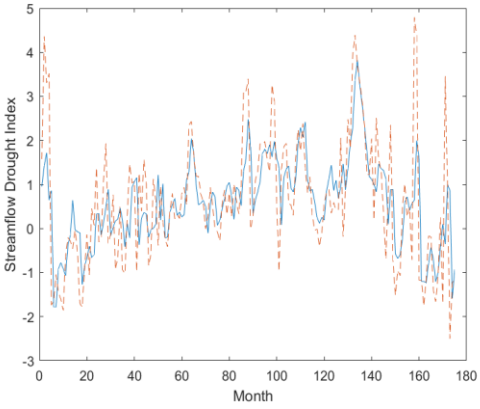


Figure C.01 HMM prediction results for Pembina River.

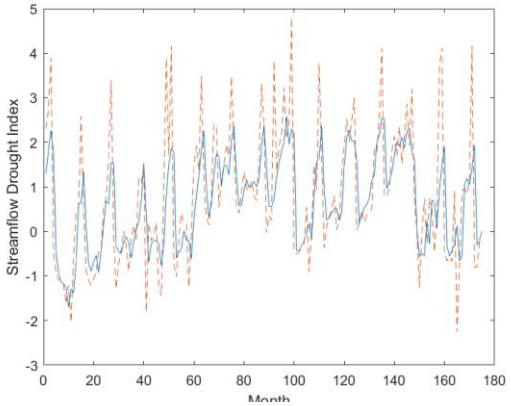


Figure C.03 HMM prediction results for Assiniboine River

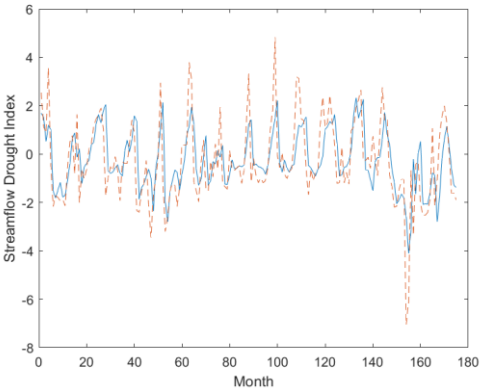


Figure C.04 HMM prediction results for Battle River

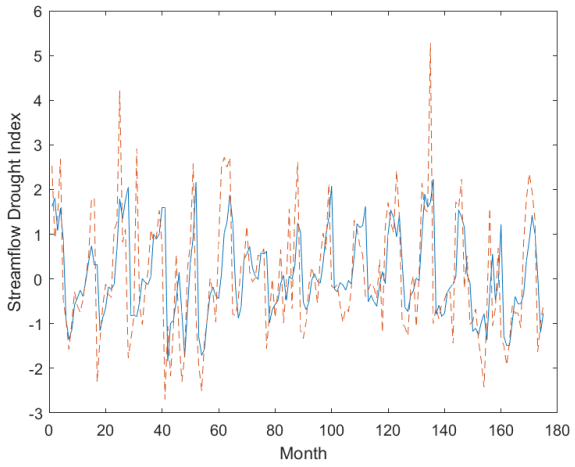


Figure C.05 HMM prediction results for Blindman River

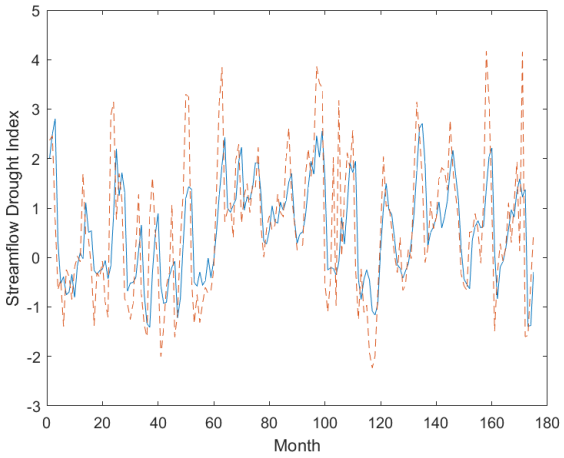


Figure C.06 HMM prediction results for Little Saskatchewan River

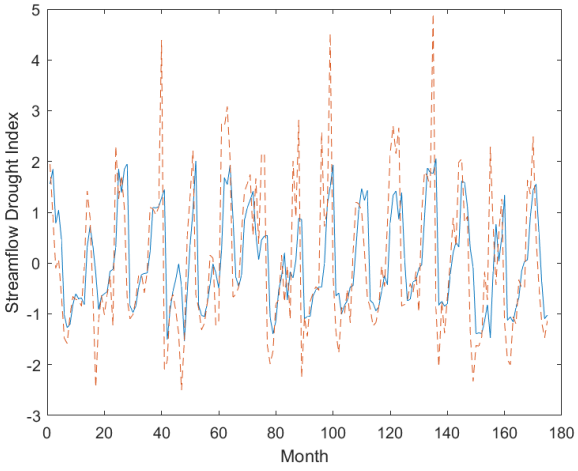


Figure C.07 HMM prediction results for Moose Jaw River

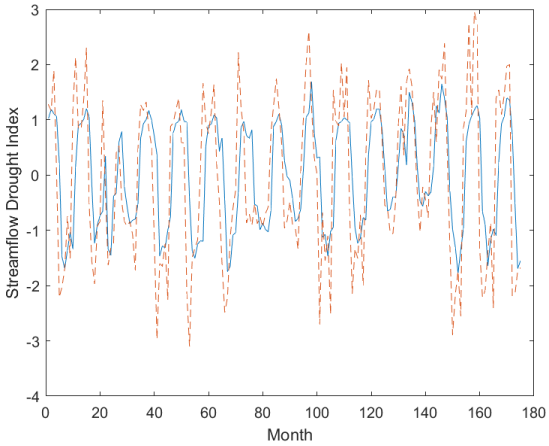


Figure C.08 HMM prediction results for Milk River

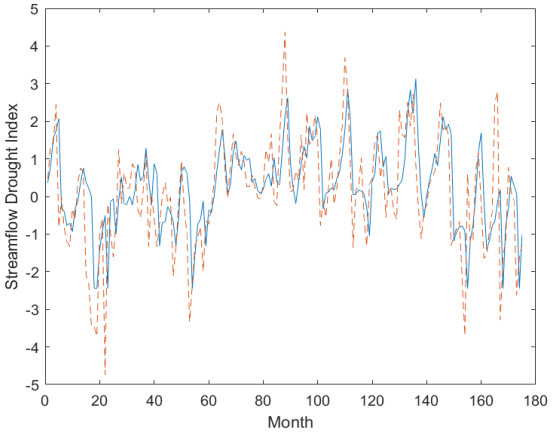


Figure C.09 HMM prediction results for Medicine River

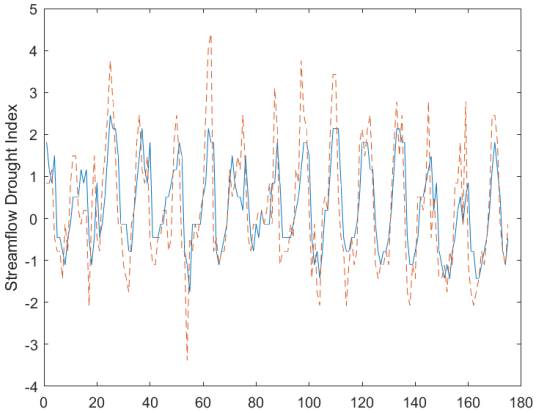


Figure C.10 HMM prediction results for North Saskatchewan River

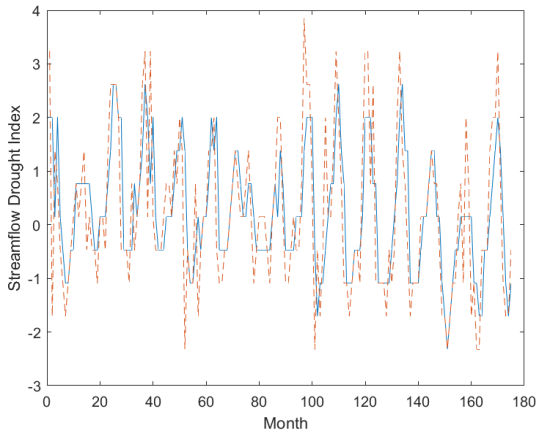


Figure C.11 HMM prediction results for North Saskatchewan River near Deer Creek

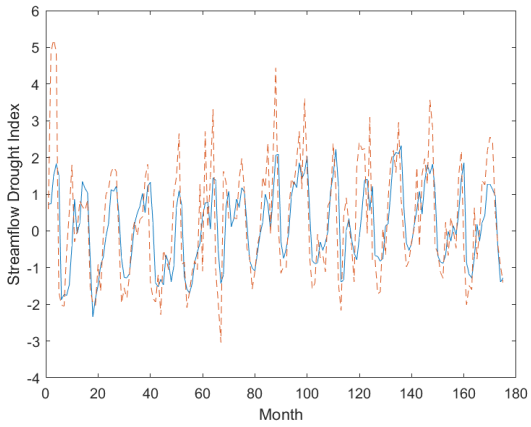


Figure C.12 HMM prediction results for Qu'appelle River

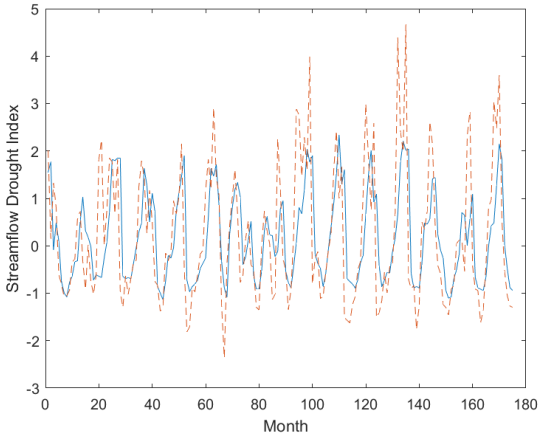


Figure C.13 HMM prediction results for Red Deer River

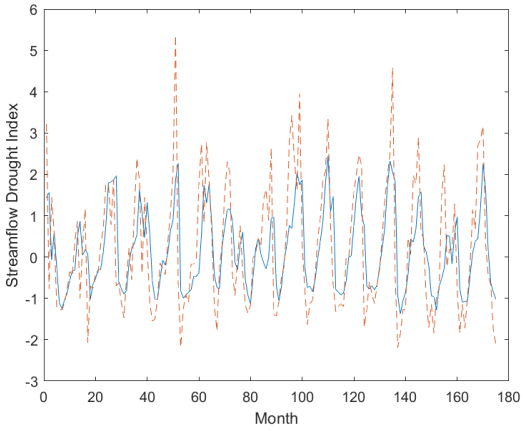


Figure C.14 HMM prediction results for Red Deer River

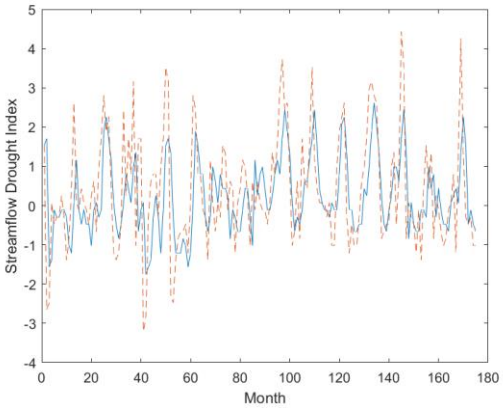


Figure C.15 HMM prediction results for South Saskatchewan River at Medicine Hat

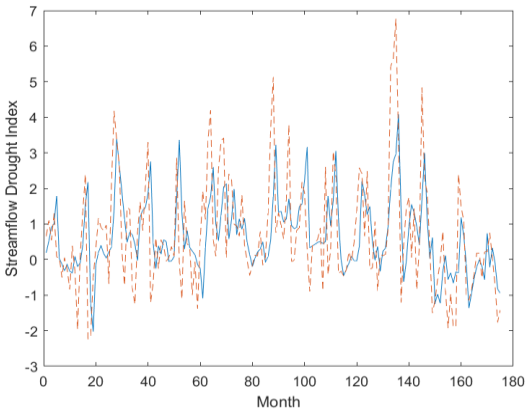


Figure C.16 HMM prediction results for Swift Current Creek

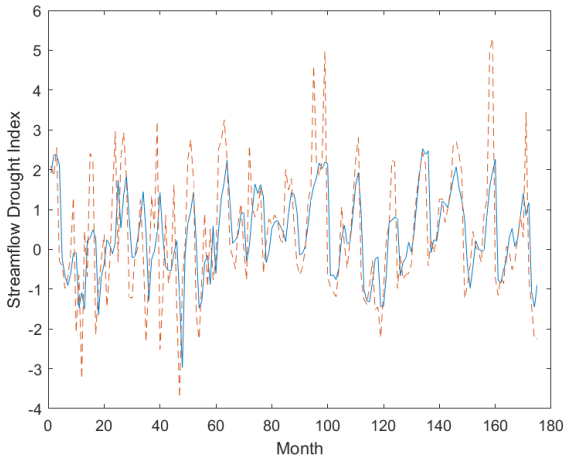


Figure C.17 HMM prediction results for Whitemud River

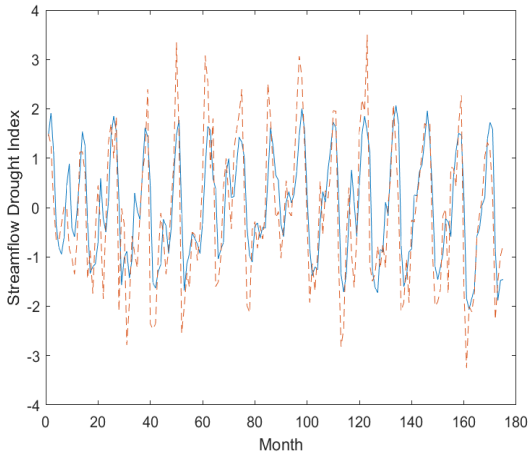


Figure C.18 HMM prediction results for Belly River

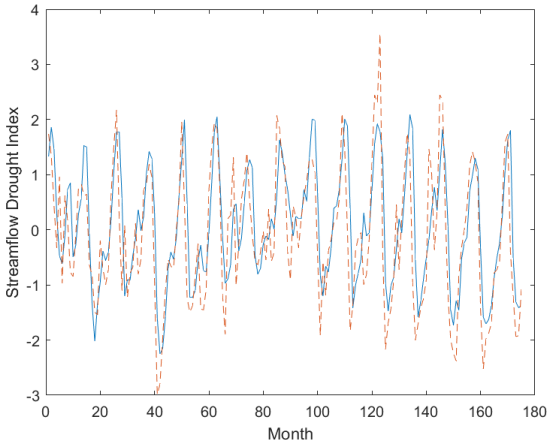


Figure C.19 HMM prediction results for Crowsnest River

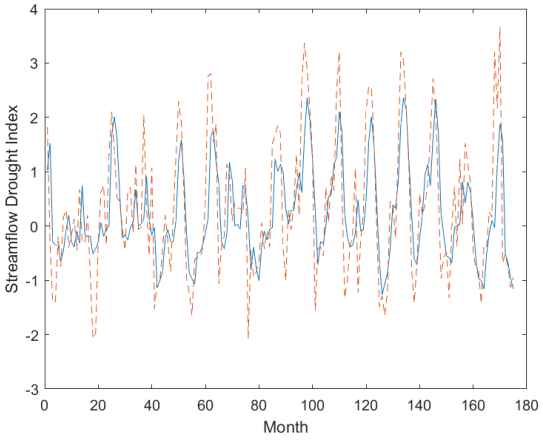


Figure C.20 HMM prediction results for Oldman River

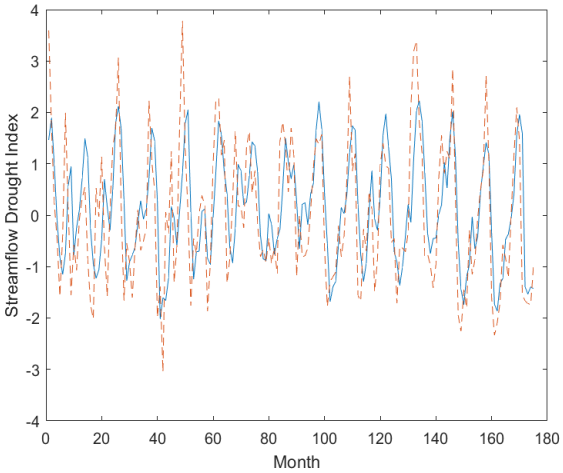


Figure C.21 HMM prediction results for St. Mary River

Monthly LSTM Model Forecast Results

The following figures present the monthly LSTM forecast results for the selected 25 watersheds. These figures demonstrate that the LSTM model performs better than the HMM model in predicting drought levels. The LSTM model's forecasts align more closely with the actual observations, indicating a higher accuracy in capturing the nuances of monthly streamflow data. This improved performance suggests that the LSTM model is more effective at handling the complexities of drought patterns compared to the HMM, providing more reliable and accurate forecasts.

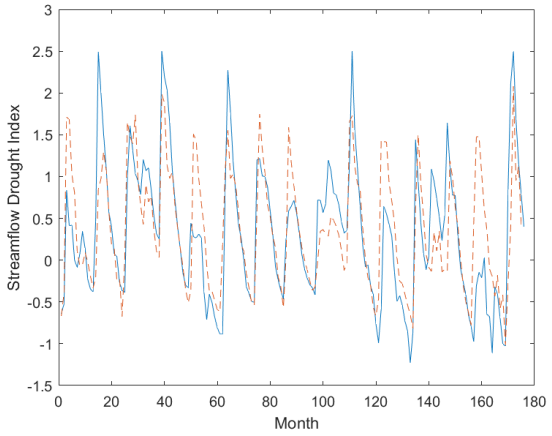


Figure C.22 Monthly LSTM prediction results for Pembina River

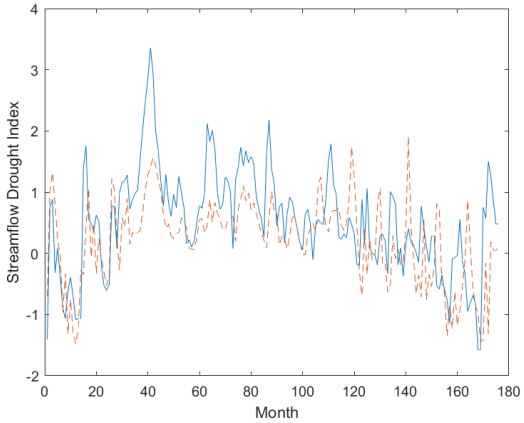


Figure C.23 Monthly LSTM prediction results for Souris River

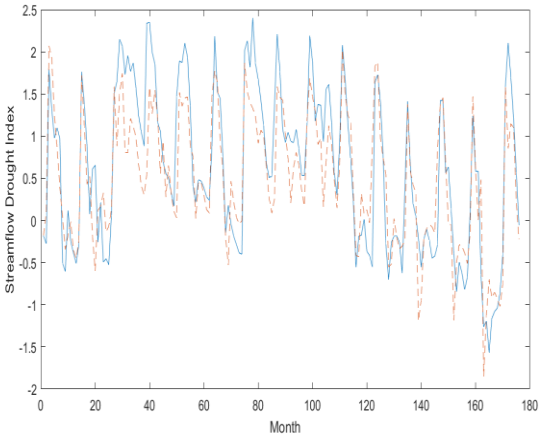


Figure C.24 Monthly LSTM prediction results for Assiniboine River

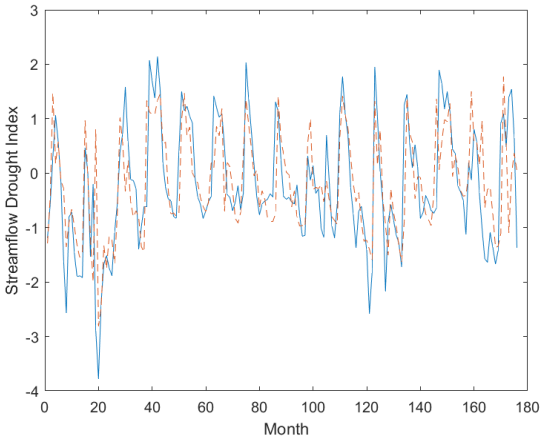


Figure C.25 Monthly LSTM prediction results for Battle River

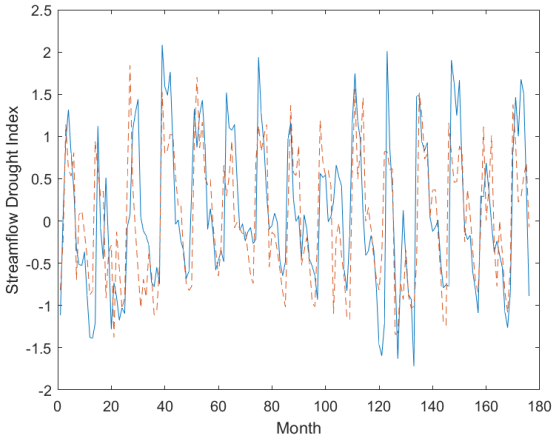


Figure C.26 Monthly LSTM prediction results for Blindman River

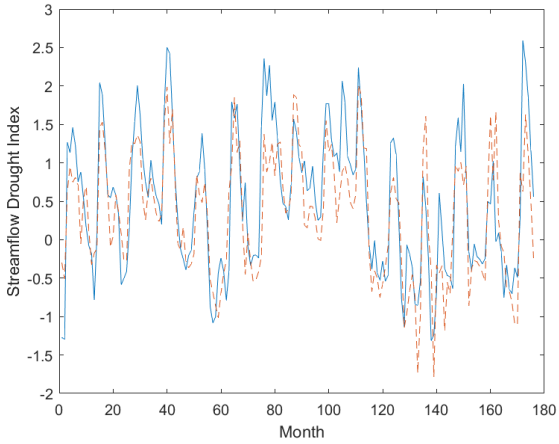


Figure C.27 Monthly LSTM prediction results for Little Saskatchewan River

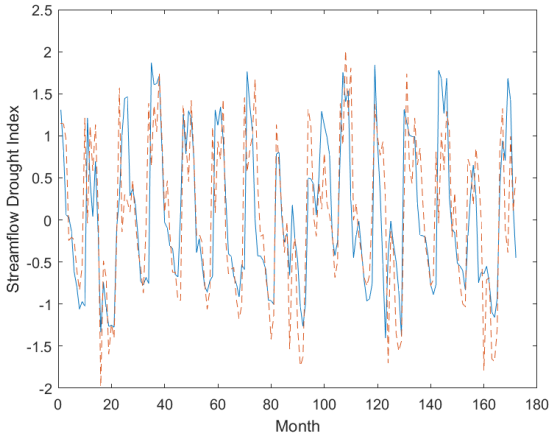


Figure C.28 Monthly LSTM prediction results for Medicine River

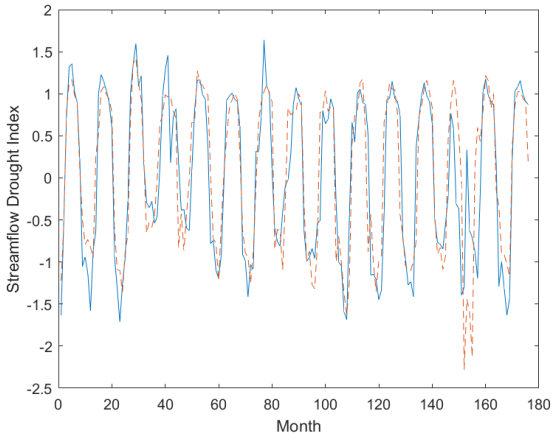


Figure C.29 Monthly LSTM prediction results for Milk River

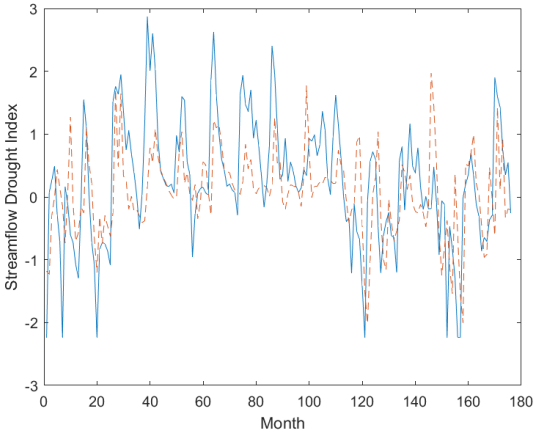


Figure C.30 Monthly LSTM prediction results for Moose Jaw River

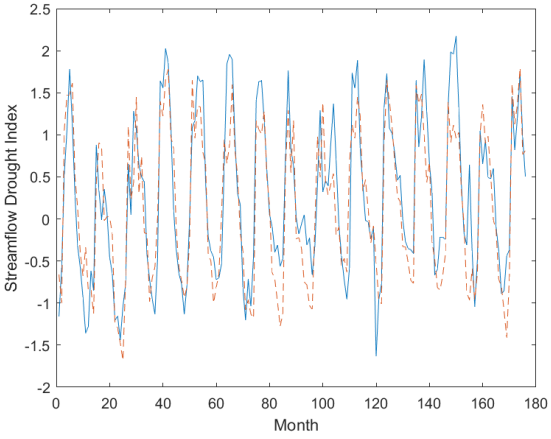


Figure C.31 Monthly LSTM prediction results for North Saskatchewan River at Prince Albert

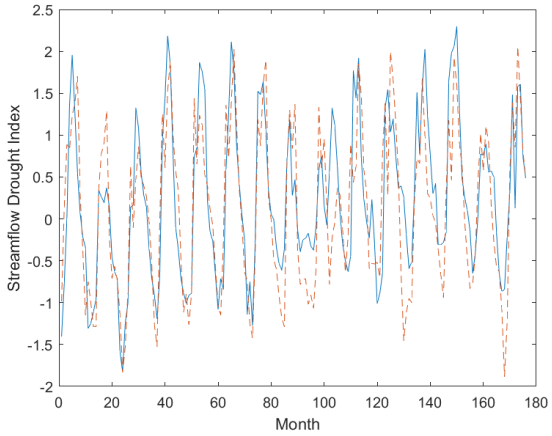


Figure C.32 Monthly LSTM prediction results for North Saskatchewan River near Deer Creek

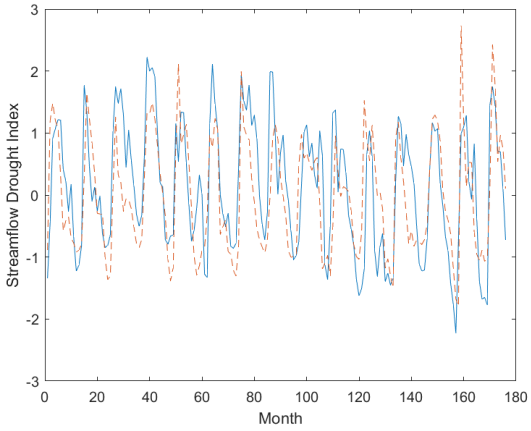


Figure C.33 Monthly LSTM prediction results for Qu'appelle River

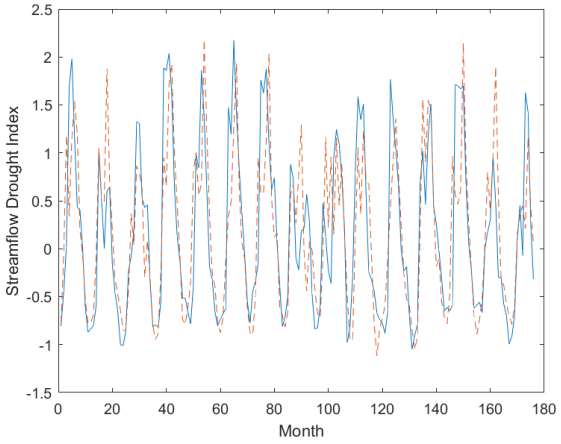


Figure C.34 Monthly LSTM prediction results for Red Deer River

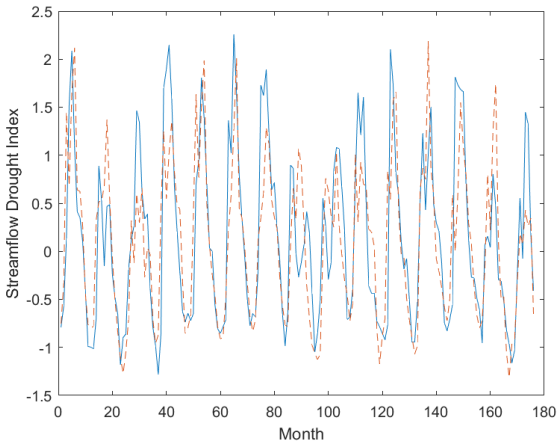


Figure C.35 Monthly LSTM prediction results for Red Deer River near Bindloss

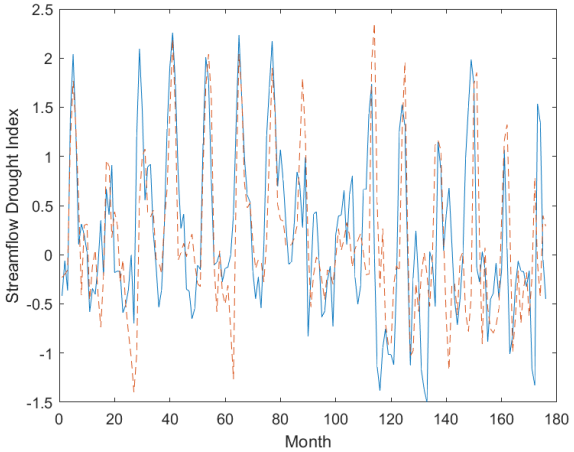


Figure C.36 Monthly LSTM prediction results for South Saskatchewan River at Medicine Hat

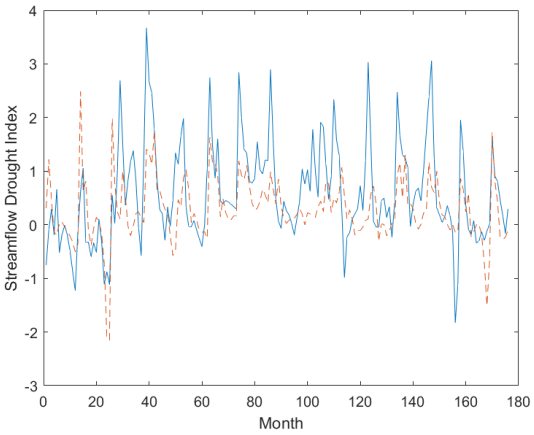


Figure C.37 Monthly LSTM prediction results for Swift Current Creek

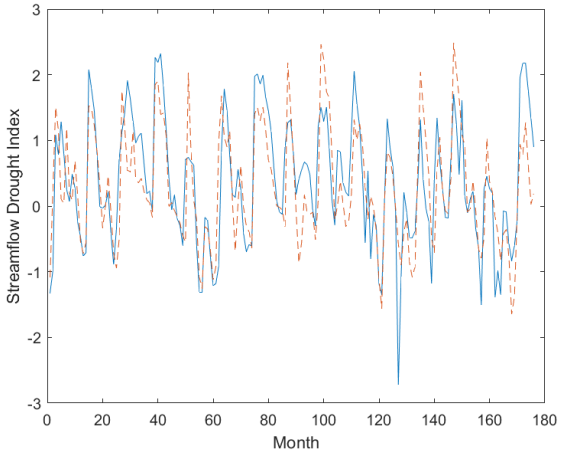


Figure C.38 Monthly LSTM prediction results for Whitemud River

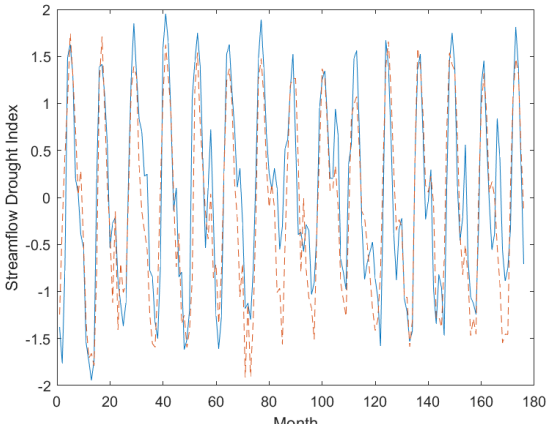


Figure C.39 Monthly LSTM prediction results for Belly River

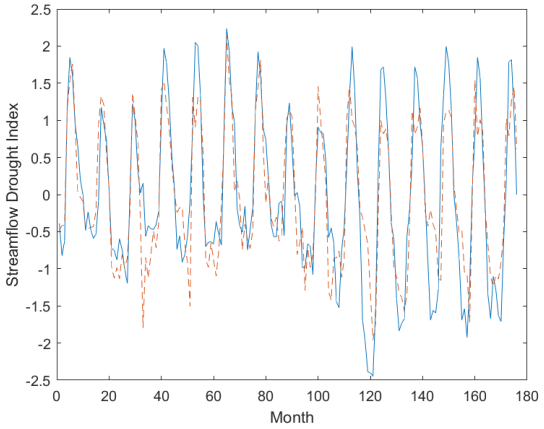


Figure C.40 Monthly LSTM prediction results for Bow River

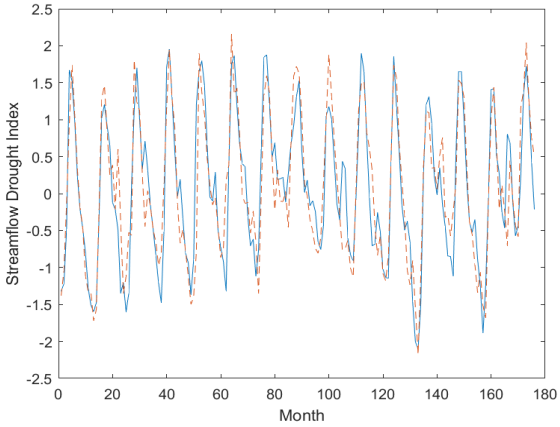


Figure C.41 Monthly LSTM prediction results for Crowsnest River

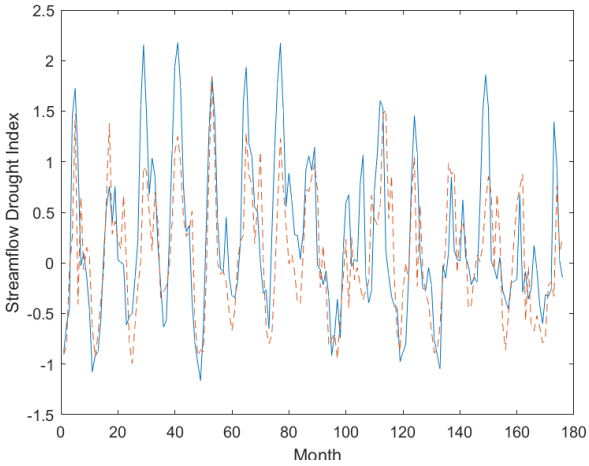


Figure C.42 Monthly LSTM prediction results for Oldman River near Lethbridge

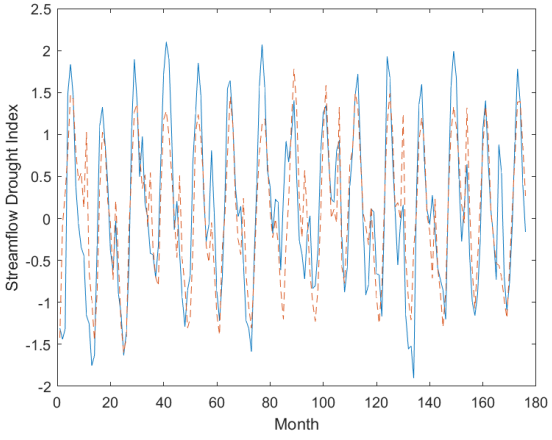


Figure C.43 Monthly LSTM prediction results for St. Mary River at International Boundary

Weekly LSTM model forecast results

The following figures present the weekly LSTM forecast results for 25 selected watersheds. These figures demonstrate that the weekly LSTM model outperforms the monthly LSTM model in predicting drought levels. The weekly model shows higher accuracy in capturing the nuances and variations within the data, indicating that LSTM models benefit from a larger dataset to enhance precision. Despite the complexity of the drought data, the weekly LSTM model excels in providing accurate forecasts, effectively handling even the most intricate drought patterns. This improvement highlights the advantage of using more granular data for better capturing the detailed fluctuations in drought conditions.

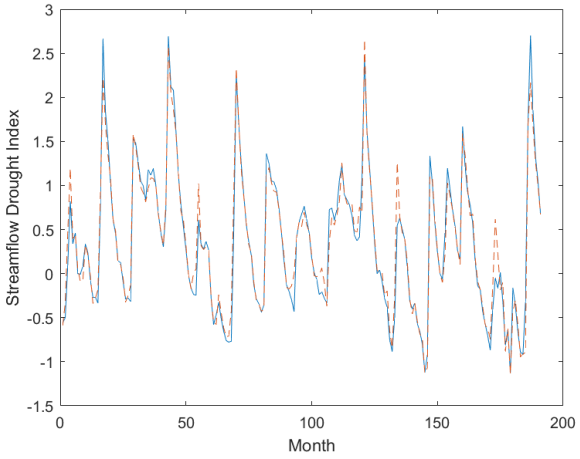


Figure C.44 Weekly LSTM prediction results for Pembina River

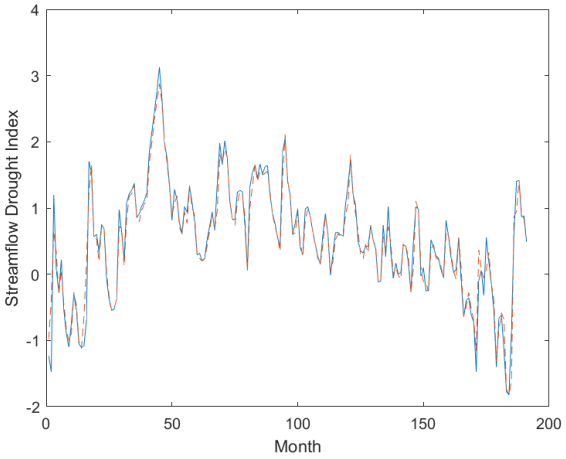


Figure C.45 Weekly LSTM prediction results for Souris River

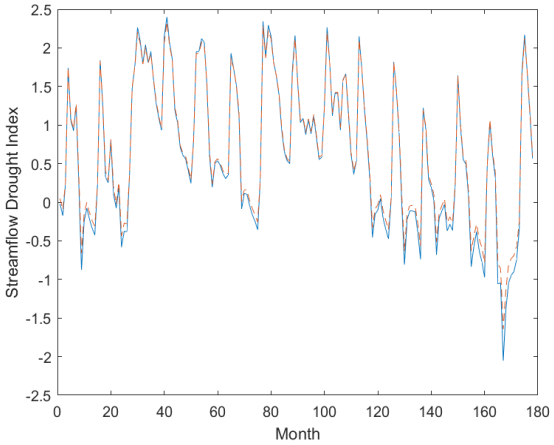


Figure C.46 Weekly LSTM prediction results for Assiniboine River

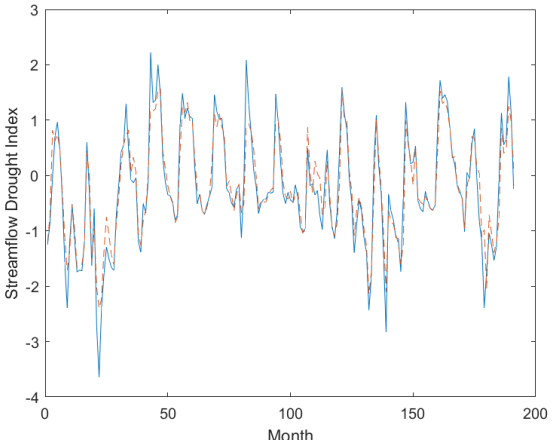


Figure C.47 Weekly LSTM prediction results for Battle River

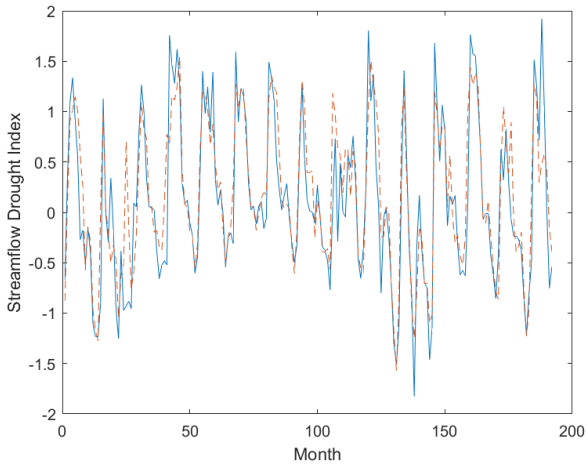


Figure C.48 Weekly LSTM prediction results for Blindman River

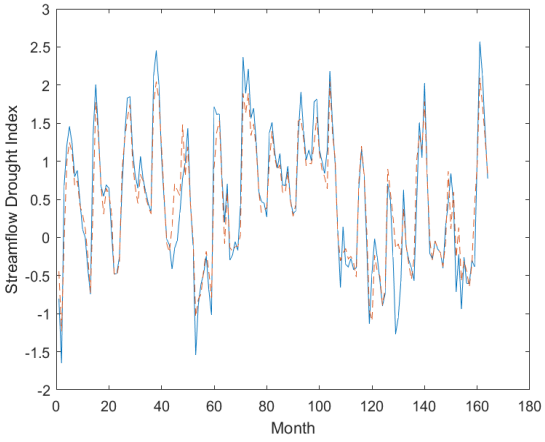


Figure C.49 Weekly LSTM prediction results for Little Saskatchewan River

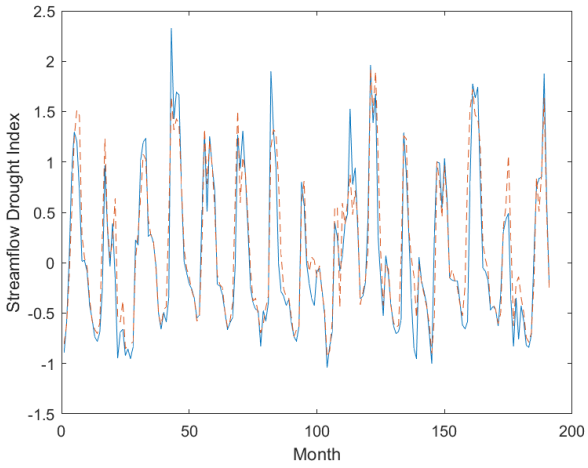


Figure C.50 Weekly LSTM prediction results for Medicine River

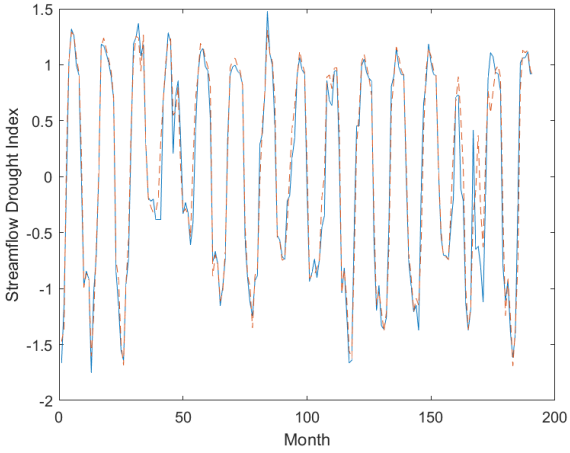


Figure C.51 Weekly LSTM prediction results for Milk River at Milk River

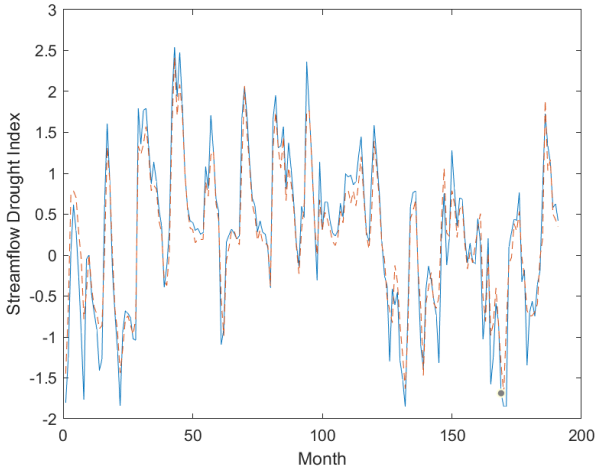


Figure C.52 Weekly LSTM prediction results for Moose Jaw River

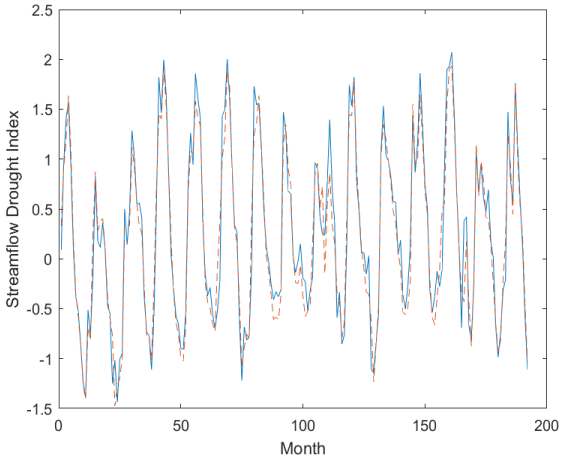


Figure C.53 Weekly LSTM prediction results for North Saskatchewan River at Prince Albert

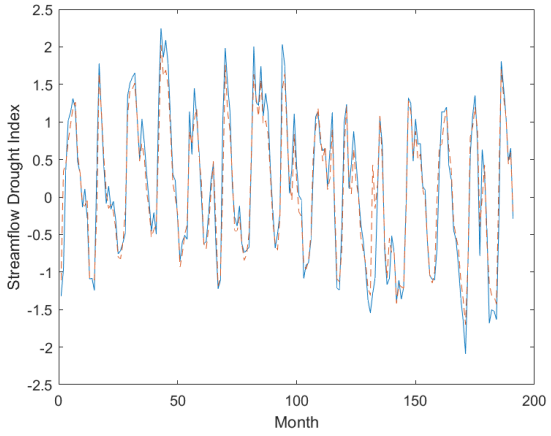


Figure C.54 Weekly LSTM prediction results for Qu'appelle River

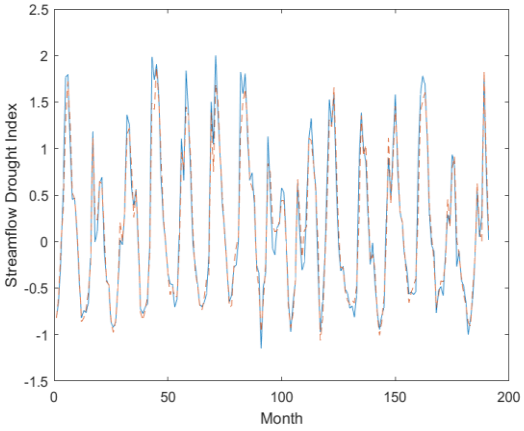


Figure C.55 Weekly LSTM prediction results for Red Deer River at Drumheller

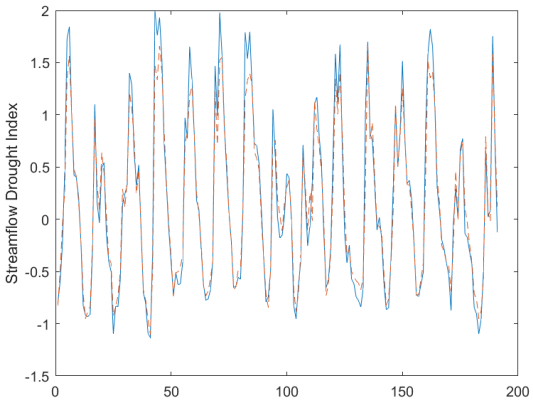


Figure C.56 Weekly LSTM prediction results for Red Deer River near Bindloss

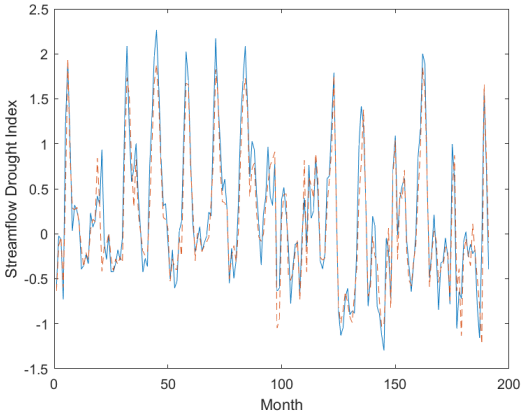


Figure C.57 Weekly LSTM prediction results for South Saskatchewan River at Medicine Hat

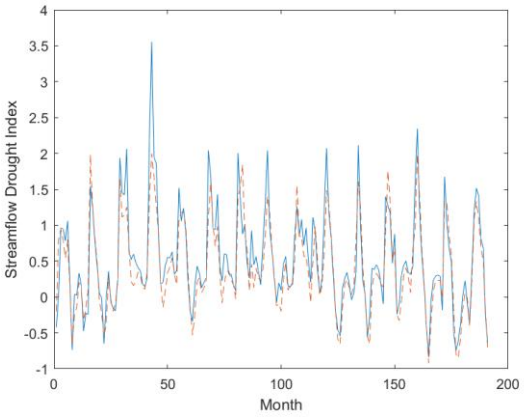


Figure C.58 Weekly LSTM prediction results for Swift Current Creek below Rock Creek

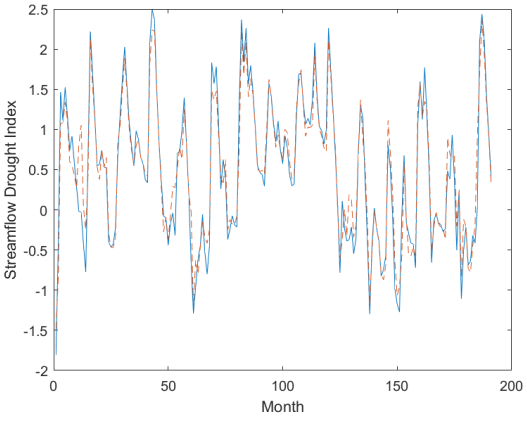


Figure C.59 Weekly LSTM prediction results for Whitemud River at Westbourne

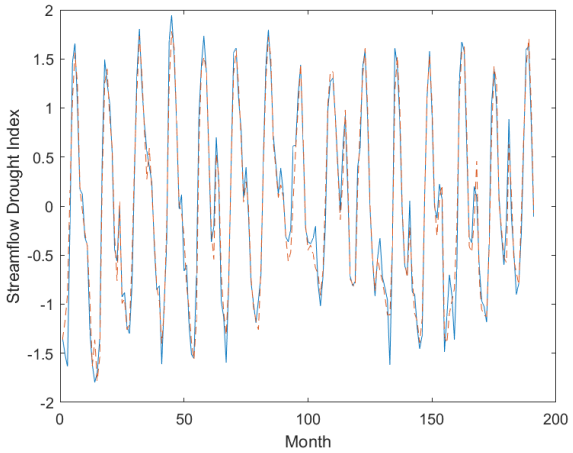


Figure C.60 Weekly LSTM prediction results for Belly River

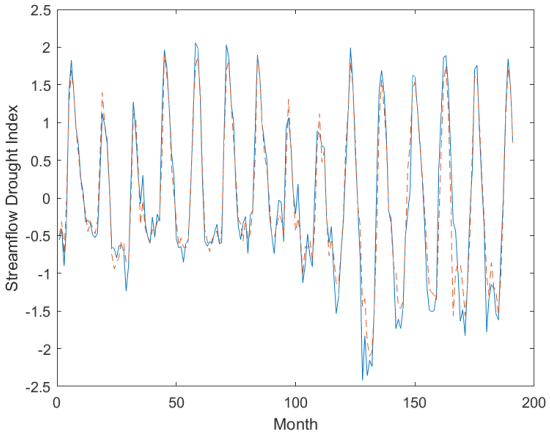


Figure C.61 Weekly LSTM prediction results for Bow River

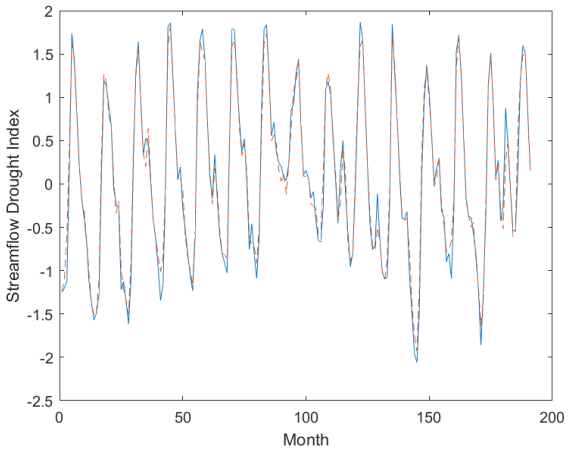


Figure C.62 Weekly LSTM prediction results for Crowsnest River at Frank

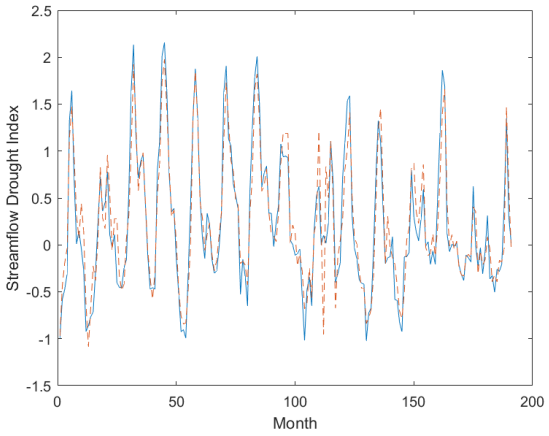


Figure C.63 Weekly LSTM prediction results for Oldman River

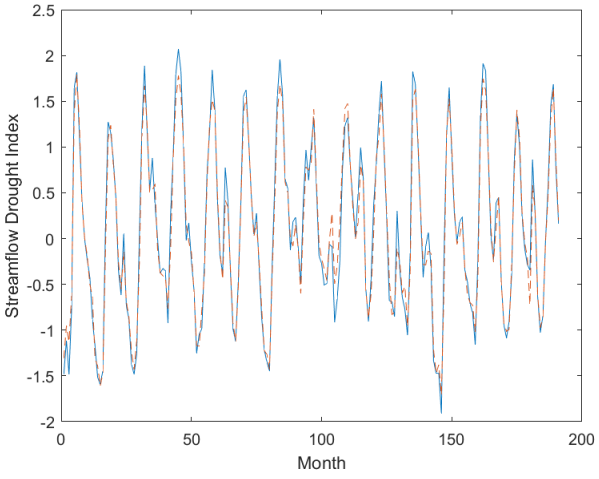


Figure C.64 Weekly LSTM prediction results for St. Mary River at International Boundary