Hydroclimatology for

Water Resources Management (CIVL 7740 T05)

Term project report on:

Pre-Seasonal Streamflow Forecasting using a Principal Component Regression Model in the Michipicoten Watershed, Canada

Prepared by:

Setogn Abate Getu

Submitted to: Prof. Donghoon Lee

Submission date: Friday, Dec 12, 202

Abstract

This term project develops a pre-seasonal streamflow forecasting model for the Michipicoten River watershed in Ontario, Canada. Two statistical forecasting models namely Principal Component Regression (PCR) and Multiple Linear Regression (MLR) were developed to predict April–May–June (AMJ) streamflow by using January–February–March (JFM) predictors derived from ERA5 climate reanalysis data and observed streamflow data from Environment and Climate Change Canada for the period 1991–2019. The Predictor variables which were used for this project are basin-averaged precipitation, temperature, snow water equivalent (SWE), potential evapotranspiration (PET), mean sea-level pressure (MSLP), and surface pressure (SP). PCA was used to reduce multicollinearity and high dimensionality, and three PCs were used as a predictor for the PCR model. PCR identified SWE, precipitation, and temperature as the most significant predictors influencing spring streamflow, and SWE, precipitation, and potential evaporation are identified as the most influencer predictors in MLR model. The model results show the results from PCR model indicts better performance with all metrics (e.g. R=0.712) compared to MLR model (R=0.595), demonstrating better stability, lower error, and stronger predictive capability.

Keywords: Seasonal Streamflow Forecasting, Principal Component Regression (PCR), Multiple Linear Regression (MLR), Hydroclimatic Predictors, Michipicoten River Watershed

Table of Contents

List of tables

List of figures

Chapter One

1. Introduction

Accurate seasonal streamflow forecasting is fundamental for hydropower operations, reservoir regulation, ecological planning, and flood mitigation, particularly in snow-dominated regions where water availability depends heavily on winter accumulation and spring melt (Barnett et al., 2005; Stewart et al., 2005). Forecasting streamflow several months before the spring melt remains challenging due to nonlinear interactions among snowpack, temperature, and precipitation, as well as the multicollinearity and high dimensionality inherent in hydroclimatic datasets (Razavi & Coulibaly, 2013; Mishra & Coulibaly, 2009). Accurate seasonal streamflow forecasting is crucial for integrated water resources management in snowmelt-dominated basins. This snowmelt-dominated regime creates both opportunities and challenges for water managers. Generally, snowpack provides a natural water reservoir with inherent predictability, and the complex interactions among climatic drivers introduce substantial forecasting uncertainty. The effective preseason streamflow forecasting is struggling with different factors including the non-linear relationships between snow accumulation, temperature regimes, and melt processes, the high dimensionality and multicollinearity of climate predictor variables, and the need for forecasts with sufficient lead time (2-3 months) to support practical decision-making for water resources management systems. Traditional statistical approaches such as Multiple Linear Regression (MLR) often struggle with highly correlated predictors—common in snowmelt-dominated basins where SWE, temperature, and precipitation are tightly coupled—leading to reduced forecast skill unless dimensionality-reduction techniques like PCA are used (Wold et al., 1987; Wuttichaikitcharoen & Babel, 2014). The Michipicoten River watershed exhibits significant seasonal variability driven by snow accumulation, melt processes, and large-scale climate controls. The Michipicoten River watershed, draining into northeastern Lake Superior in Ontario, Canada, exhibits significant hydroclimatic seasonality characterized by substantial winter snow accumulation followed by pronounced spring melt-driven runoff.

This project aims to overcome these challenges using Principal Component Regression (PCR) and Multiple Linear Regression (MLR). The key motivation is to use early-season (JFM) climate predictors to estimate pre-seasonal (AMJ) streamflow. By transforming correlated predictors into orthogonal components, PCR reduces instability and improves model robustness.

1.2 Research questions for the project

At the end of this project, the following primary questions were addressed:

1. Can Principal Component Regression (PCR) provide better prediction in pre-seasonal (April-June) streamflow forecasting as compared to Multiple Linear Regression (MLR) in the Michipicoten watershed?
2. Which hydroclimatic variables (precipitation, temperature, snow water equivalent, potential evapotranspiration, mean sea-level pressure, and surface pressure) exert the most significant influence on spring streamflow prediction?
3. How do the PCR and MLR models differ in terms of their performance metrics to predict AMJ streamflow in the Michipicoten watershed?

1.3. Objectives

The main objective of this term project is to develop and evaluate PCR and MLR models to forecast pre-seasonal (April–June) streamflow using January–March hydroclimatic variables followed by the following specific objectives:

1.3.1 Specific objectifies

1. To develop PCR and MLR models for predicting April-May-June (AMJ) streamflow using January-February-March (JFM) hydroclimatic predictors derived from ERA5 reanalysis data.
2. Identify and rank the most influential predictors (precipitation, temperature, SWE, PET, MSLP, SP) for the prediction of streamflow.
3. Evaluate model performance using statistical metrics (Pearson correlation, RMSE, MAE, NSE, RPSS)

Chapter Two

2. Literature Review

Seasonal streamflow forecasting is essential for water resources management, hydropower operations, and irrigation planning, particularly in snow-fed basins where spring runoff is largely controlled by snowmelt. Predictability at seasonal timescales depends on antecedent hydrological conditions such as snow water equivalent (SWE) and soil moisture. Arnal et al. (2024) developed FROSTBYTE, a reproducible data-driven workflow that uses SWE measurements as predictors and streamflow observations as predictands, applying quantile mapping to fill SWE gaps and principal component analysis (PCA) to extract independent predictor components for regression-based ensemble hindcasts. Applied to 75 snow-dominated basins across Canada and the USA, their approach demonstrated that late-season snowpack is a key predictor of spring and summer streamflow, while high precipitation during the forecast period can reduce forecast skill. The study also showed that probabilistic forecasts could predict above- and below-normal flows with lead times of up to five months, highlighting the effectiveness of PCA-based regression methods for pre-seasonal streamflow forecasting (Arnal et al., 2024).

Risley et al. (2005) conducted a comprehensive assessment of seasonal forecasting methods—including principal component regression (PCR) and several neural-network approaches—in the Upper Klamath Basin, a snow-dominated watershed like Michipicoten. Their analysis showed that PCR consistently performed well during key spring forecast months and benefited from snowpack-based predictors such as SWE and precipitation. Fleming and Garen (2022) examined how cross-validated principal component regression (PCR), widely used in western North America for operational seasonal streamflow forecasting, performs under increasingly complex predictor datasets and emerging machine-learning extensions. Their results showed that applying cross-validation only to the final supervised-learning step provides reliable error estimates, even when predictors include advanced climate products or nonlinear models. They also mentioned that the finding supports more efficient, robust PCR-based approaches for seasonal water supply forecasting.

The seasonal predictability of streamflow in snowmelt-dominated basins derives from several physical mechanisms that provide memory in the hydrological system. Snowpack accumulation during winter months serves as a natural reservoir whose volume strongly influences subsequent spring runoff (Barnett et al., 2005). This storage mechanism creates a deterministic relationship between winter conditions and spring flows that forms the physical basis for statistical forecasting approaches. Research has demonstrated that snow water equivalent (SWE) measured at or near peak accumulation (typically March) explains 40-70% of the variance in subsequent spring runoff in many mountain basins (Luce et al., 2014). However, this relationship is modulated by spring temperature and precipitation, which affect melt rates and timing. The complex interplay between these factors creates challenges for simple regression approaches and motivates the use of more sophisticated statistical methods that can handle correlated predictors and non-linear relationships (Luce et al., 2014).

Regression models have been widely used for seasonal streamflow forecasting due to their simplicity and interpretability. Multiple Linear Regression (MLR) represents the most straightforward approach, establishing linear relationships between predictors and streamflow. However, MLR's performance is often compromised by multicollinearity among climate variables—a common issue in hydroclimatic datasets where temperature, precipitation, and snowpack are interrelated through physical processes (Wilby et al., 2004). Principal Component Regression (PCR) addresses these limitations by first transforming correlated predictors into orthogonal components. This approach has been successfully applied in various hydrological forecasting contexts. For example, Wuttichaikitcharoen and Babel (2014) demonstrated PCR's superiority over MLR for suspended sediment yield estimation in Thailand, noting its ability to handle correlated predictors while maintaining predictive skill.

Chapter Three

3. Data sources and methodology

3.1. Study area

The Michipicoten River watershed drains into northeastern Lake Superior in Ontario, Canada. The watershed is characterized by boreal forest cover, cold winters, and substantial snow accumulation. The hydrologic regime is dominated by snowmelt, leading to high flows during spring months (AMJ). Streamflow data, which was used for model performance evaluation and predictand, was obtained from Hydrometric Station number 02BD002 from ECCC.

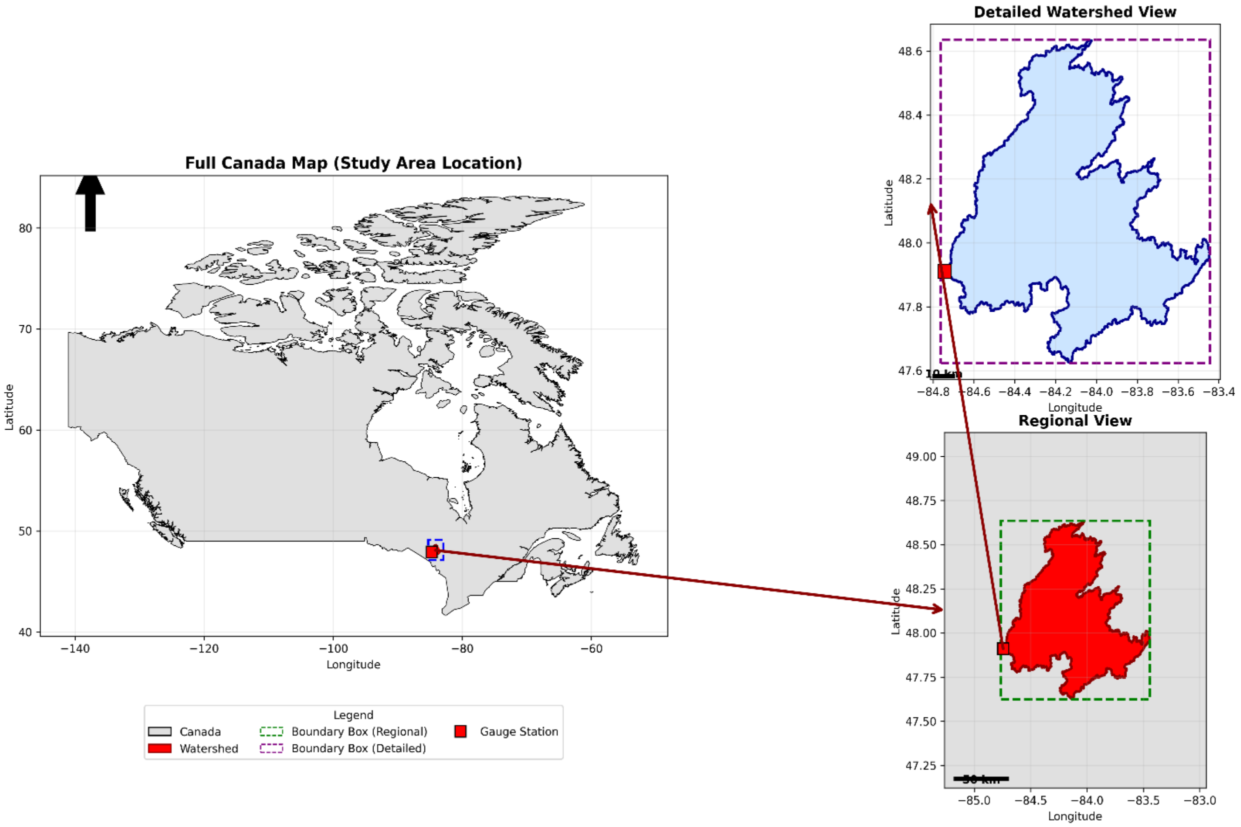


Figure 1 study area

3.2. Data Sources

For this term project, the required climate forcing data or hydrometeorological datasets were downloaded from ERA5, which was produced by ECMWF for the Copernicus Climate Change

Service (C3S), part of the European Union's Copernicus Programme. The ERA5 reanalysis dataset provides comprehensive climate variables at high temporal and spatial resolution.

These original predictors include Precipitation (Pr) Temperature (T), Snow Water Equivalent (SWE), Potential Evapotranspiration (PET), Mean Sea Level Pressure (MSLP) and Surface Pressure (SP). These predictors were aggregated to monthly means for the months JFM, and AMJ streamflow was computed as the seasonal mean. Similarly historical observed streamflow (1991–2019) as the predictand was downloaded for station number 02BD002 from Environment and Climate Change Canada.

3.3. Methods and tools

To achieve the objectives of this term project, the following methodologies and tools were used.

Hydroclimatic variables from January–March was collected from ERA5, clipped to the watershed, and processed and Python was used for data cleaning, aggregation, analysis, and visualization. The collected monthly data were aggregated into the preseason predictor variables. PCA was applied in Python to reduce multicollinearity, and the resulting principal components were used as a predictor for PCR to predict streamflow. Model performance was evaluated using Pearson correlation (r), RMSE, NSE, and MAE, and the most influential predictors were identified for both models. QGIS used for shapefile/watershed preparation, while the Climate Predictability Tool (CPT) was used to predict streamflow using MLR model. Pre-seasonal (April–June) streamflow in the Michipicoten River watershed was forecast using a PCR model by using PCs as predictors, and the most influential predictors were identified for both models. Model performances were evaluated using statistical metrics such as R, RMSE, NSE, and MAE.

3.3.1. Principal Component Regression (PCR)

Principal Component Regression is a two-step multivariate statistical technique designed to address multicollinearity in predictor variables, a common issue in hydroclimatic datasets where climate variables are often interrelated.

PCR combines Principal Component Analysis (PCA) with linear regression: PCA transforms p original correlated predictors $X_1, X_2, \ldots, X_p$ into k uncorrelated principal components $PC_1, PC_2, \dots, PC_k$ (where $k \leq p$), ordered by decreasing variance explained. The principal components are used as predictors in a multiple linear regression model.

Regression Equation (PCR Model):

$$\text{QAMJ}(\text{m3/s}) = \beta_0 + \beta_1 PC_1 + \beta_2 PC_2 + \beta_3 PC_3 + \beta_4 PC_4 + +\beta_5 PC_5 + \beta_6 PC_6 + \varepsilon$$

Where:

- QAMJ: predicted streamflow (discharge) for the months of April, May, and June, measured in cubic meters per second.
- β_0: Intercept of the regression model; represents the mean value of $Q_{AMJ}$when all predictor variables are zero.
- $PC_1, PC_2, \dots, PC_6$: The principal components derived from the original predictor variables.
- $\beta_1, \beta_2, \dots, \beta_6$: Regression coefficients for each principal component; indicate the strength and direction of the relationship between each PC and streamflow.
- ε: Error term; represents the variation in Q_{AMJ}not explained by the regression model.

3.3.2. Multiple Linear Regression (MLR)

Multiple Linear Regression establishes a linear relationship between a dependent variable and multiple independent variables without prior transformation.

Regression Equation (MLR Model):

$$Q_{AMJ} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

Where:

$X1, x2, x3$ … ..Predictor variables (e.g., precipitation, temperature, or other climate indices) selected for the models.

3.4. Methodology and Approach

The approach followed in this project systematically transforms raw hydroclimatic data into the principal components by considering the most variance direction to solve the problems of multicollinearity, high dimensionality, and complex interdependence associated with hydroclimatic variables. This approach includes data preparation, preprocessing, data analysis, dimensionality reduction, identifying the most influencers, model development and evaluation.

To achieve the objectives of the project, JFM climate predictors obtained from ERA5 reanalysis were compiled, basin-averaged, and standardized, while the observed streamflow from ECCC used as predictand for AMJ streamflow prediction. Principal Component Analysis (PCA) was applied to reduce multidimensionality among the predictors and to extract the major modes of variability that were used in developing the Principal Component Regression (PCR) model. In addition, a Multiple Linear Regression (MLR) model was used based on the original predictor variables to compare the performance both models. Python was used extensively for data cleaning, temporal aggregation, PCA computation, PCR modeling and plotting, while QGIS was used for watershed delineation and spatial preparation. The Climate Predictability Tool was used for the development of the MLR model. Model performances were evaluated using statistical metrics, including correlation coefficient (r), root mean square error (RMSE), Nash-Sutcliffe Efficiency (NSE), and mean absolute error (MAE), to quantify predictive ability and identify the most influential hydroclimatic predictors to predict AMJ flows. The figure shown below indicates the general methodological approach for the project.

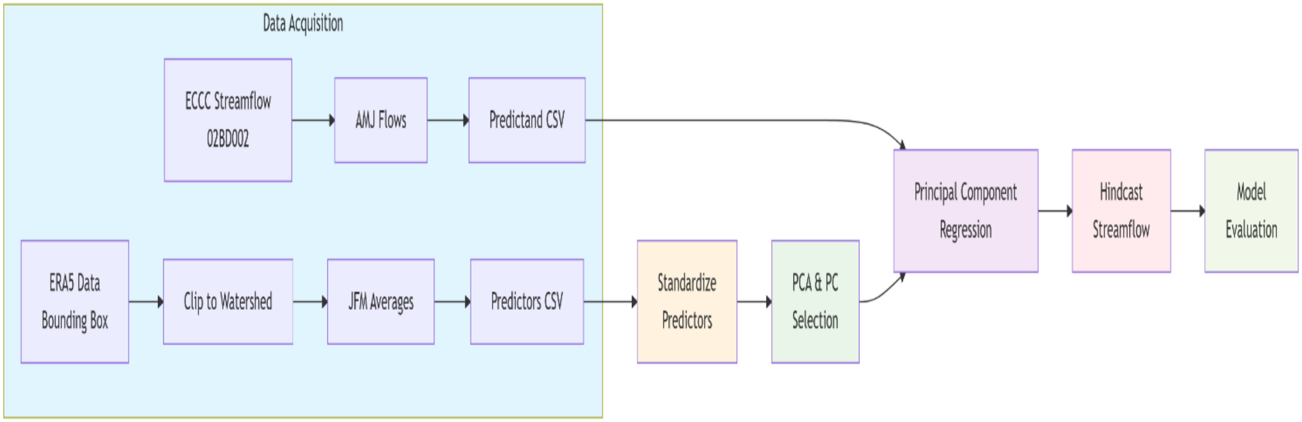


Figure 2: General methodological approach

3.5. Data Analysis

3.5.1. Correlation between predictors and predictand

Evaluating the statistical relationship between hydroclimatic predictors and the target streamflow is a critical preliminary step in developing forecasting models. Correlation analysis provides insight into the strength and direction of linear relationships between variables, helping to identify the most influential predictors and reduce potential redundancy. Pearson correlations and coefficient of determination were computed between all predictors and the target variable (AMJ streamflow). Results showed SWE, precipitation, and temperature had the strongest correlations.

Table 1: Correlation (R) and coefficient of determination(R^2)

Combination	R	R^2
SWE	0.6568	0.4314
SWE+Pr	0.7054	0.4976
SWE+Pr+Temp	0.7249	0.5255
SWE+Pr+Temp+PET	0.732	0.5358
SWE+Pr+Temp+PET+MSLP+SP	0.7535	0.5678

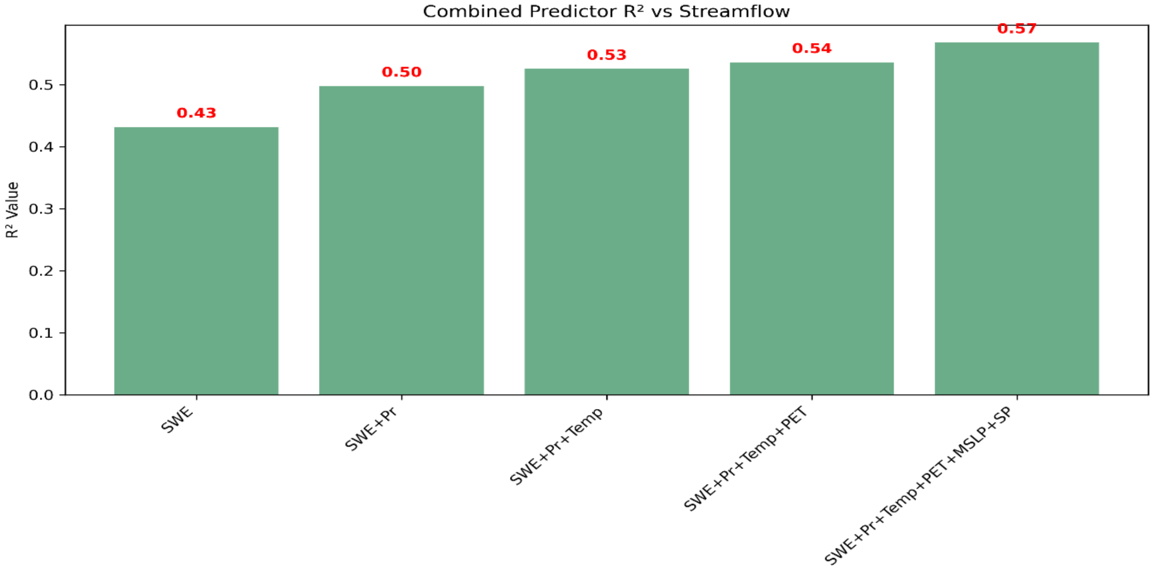


Figure 3: coefficient of determination correlations between combined predictors and streamflow

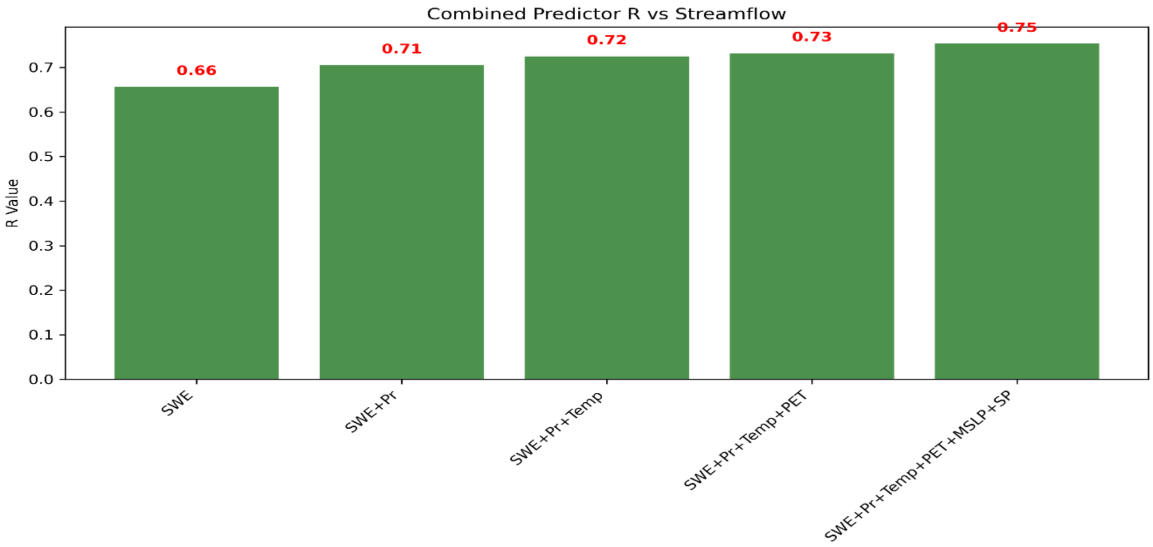


Figure 4: Pearson correlations between combined predictors and streamflow

3.5.2. Correlation matrix heatmap

The correlation matrix heatmap figure shown below for all predictors and predictand summarizes the linear relationships between all predictor variables and the AMJ streamflow predictand. Color intensity indicates the strength and direction of correlation, and the figure shows there is relationship between predictors, and the correlation values demonstrates to use PCR rather than MLR model to reduce the impacts of dimensionality in the prediction of streamflow.

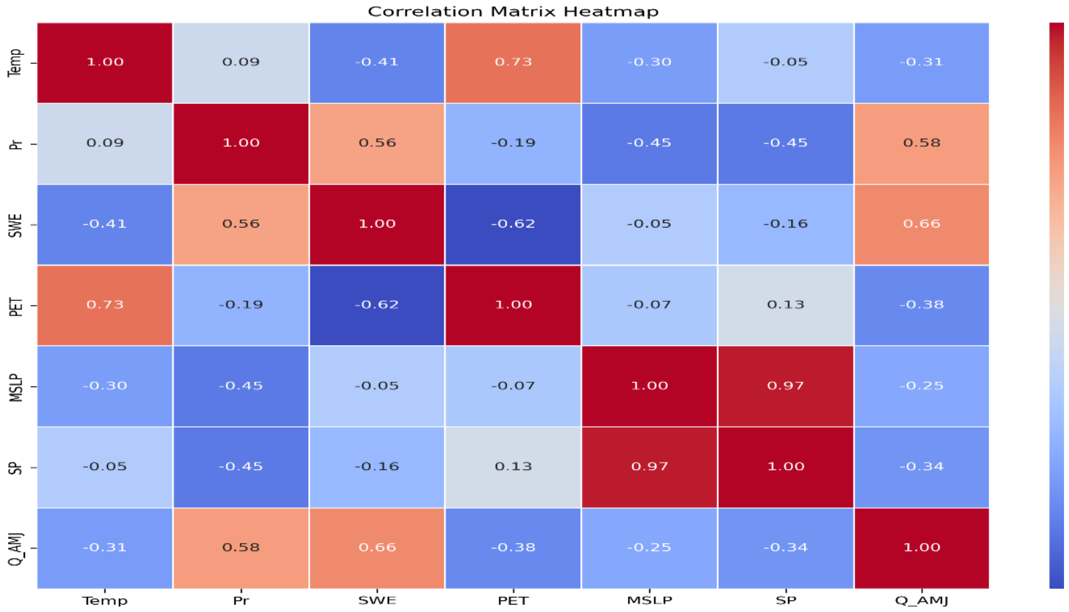


Figure 5: Heatmap correlation matrix for all predictors and predictand

3.5.3. Correlation of each predictor with predictand

The strength and direction of the relationship between each predictor with the target variable (predictand) was evaluated using correlation analysis, providing insight into which predictors are most influential for streamflow forecasting.

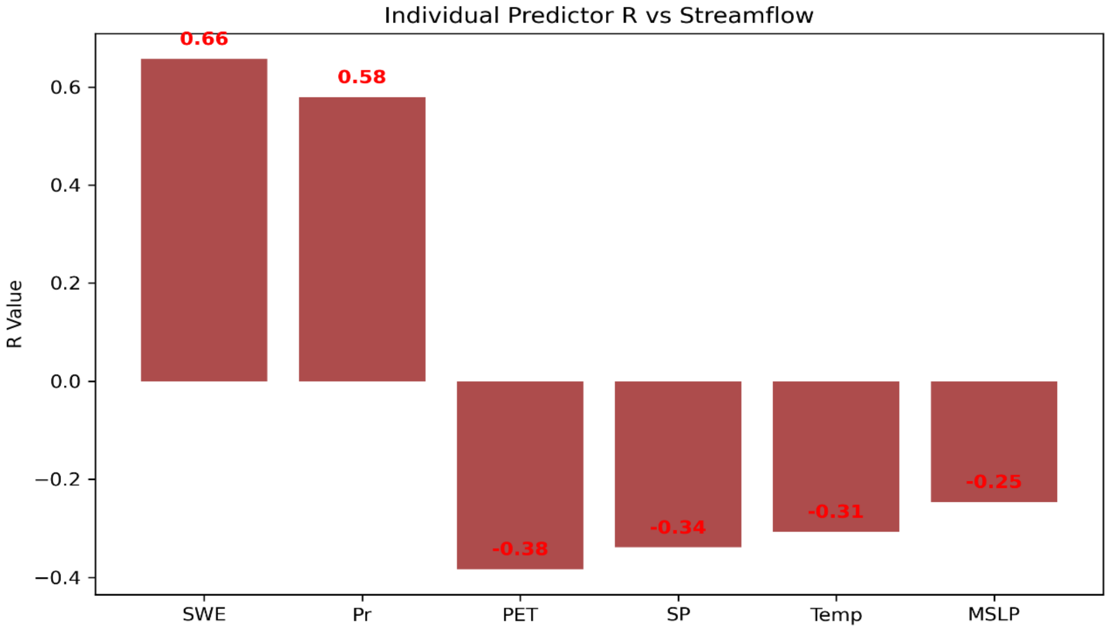


Figure 6: Pearson correlation of each predictor with predictand

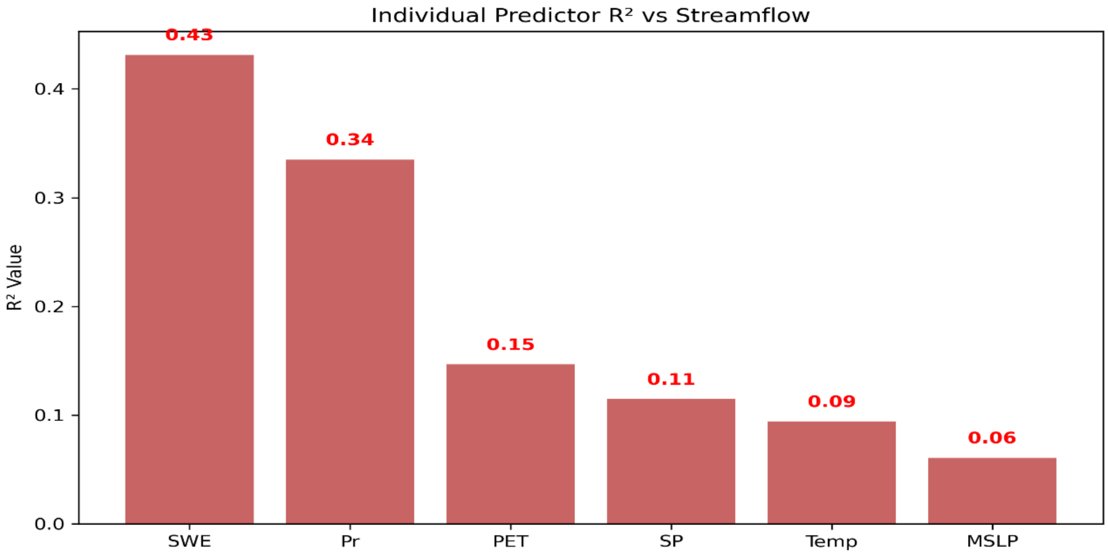


Figure 7: Coefficient of determination correlation of each predictor with predictand

3.5.4. Standardization/Normalization

Standardization was applied to transform each variable to have zero mean and unit variance. Therefore, before the application of PCA all predictors were standardized (mean = 0, variance = 1) to ensure variables with different units contribute equally to the analysis. This step is essential for methods like Principal Component Regression, which are sensitive to the scale of the input data, and the standardization figure is shown below.

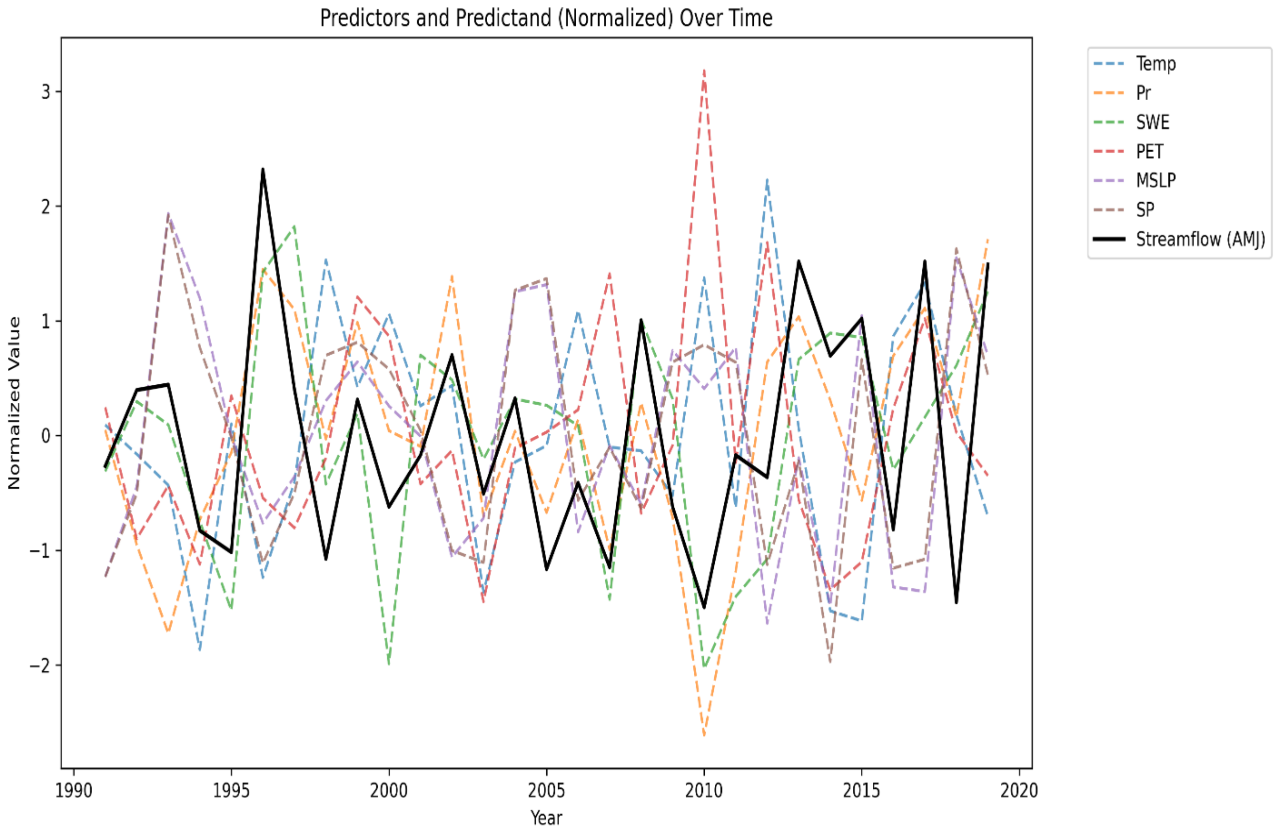


Figure 8: Standardization/Normalization

Chapter Four

4. Results and Discussion

4.1. Total Variance Explained

The cumulative variance explained by each principal component was calculated by applying PCA, providing a measure of the effectiveness of dimensionality reduction in retaining the information from the original predictors. PCA reduces data dimensionality by creating PCs that capture most of the original variance. Based on the 80% as a threshold value three PCs were selected to be used as predictor for PCR model.

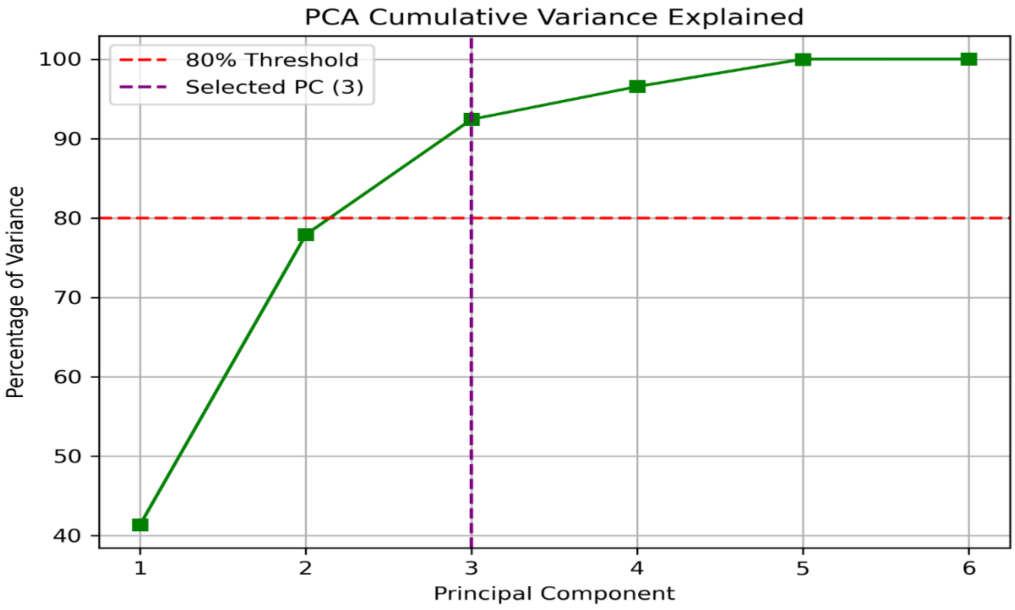


Figure 9: Cumulative total Variance Explained

4.2. Observed and Hindcasted streamflow

The performance of the developed Principal Component Regression (PCR) and Multiple Linear Regression (MLR) models was evaluated through time-series comparison of hindcasted versus observed April–May–June (AMJ) streamflow for the period 1991–2019. The figures shown below indicate that PCR hindcasted (orange line) closely match observed flows (blue line), while MLR hindcasted (green line) deviate more, especially in peak flows. The PCR model successfully captures the interannual variability of spring streamflow throughout the 29-year study period as compared to MLR model. The predicted streamflow (orange line) generally follows the observed streamflow (blue line), with both series exhibiting similar patterns of wet and dry years. This demonstrates the model's ability to reproduce the year-to-year fluctuations that characterize the Michipicoten watershed. The PCR model demonstrates reasonable skill in capturing peak flow events, particularly during high-flow years such as 2002. The predicted peaks generally align in some extent with observed peaks, indicating that the model correctly identifies the timing of snowmelt-driven high flows. However, some overestimation and underestimation are also evident during the study period. The MLR hindcasted model results lacks to match the interannual variability of the observed streamflow, and there are also some both underestimation and overestimation hindcasted streamflow values. In the MLR model, which was developed by CPT, there are some hindcasted values that best matches with some peak values such as for the years 1999 and 2002.

Based on the model results, the regression equations which were used for streamflow hindcast for both models are presented below.

Principal Component Regression equation:

Q_AMJ = 82.361 - 15.087*PC1 - 5.310*PC2 + 8.455*PC3

Multiple Linear Regression equation:

Q_AMJ = 4095.052 - 31.31*T + 11.4*Pr + 0.35*SWE + 10.76*PET - 135.58*MSLP +1.38*SP

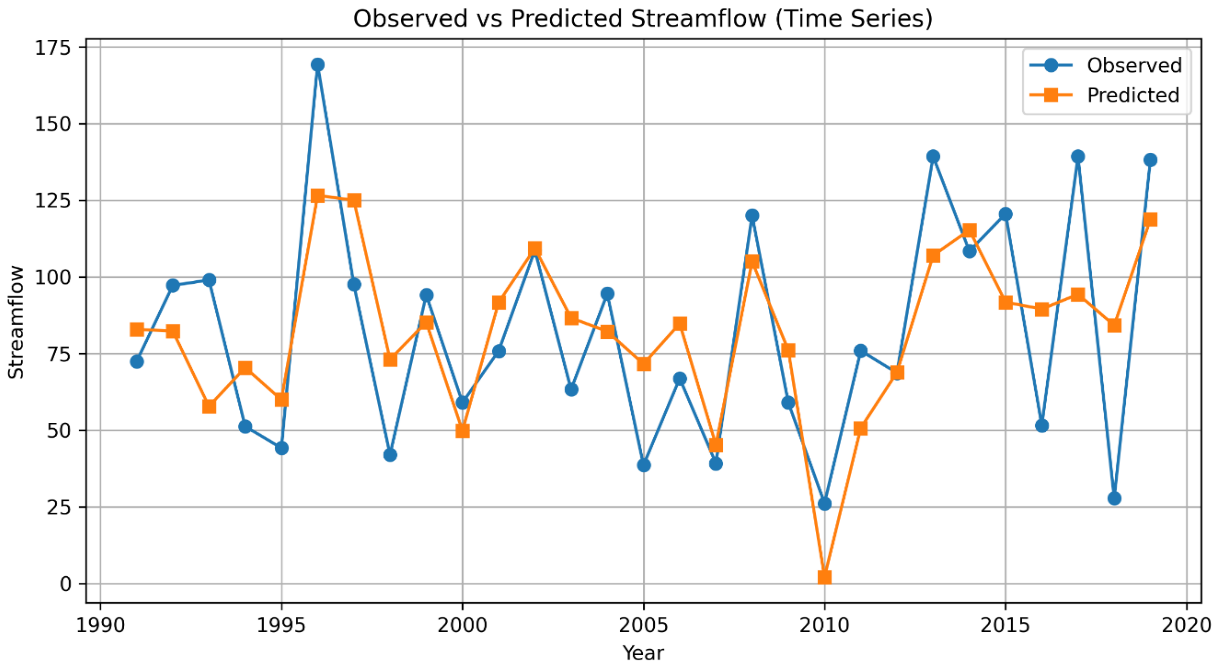


Figure 10: Observed Vs forecasted streamflow for PCR model

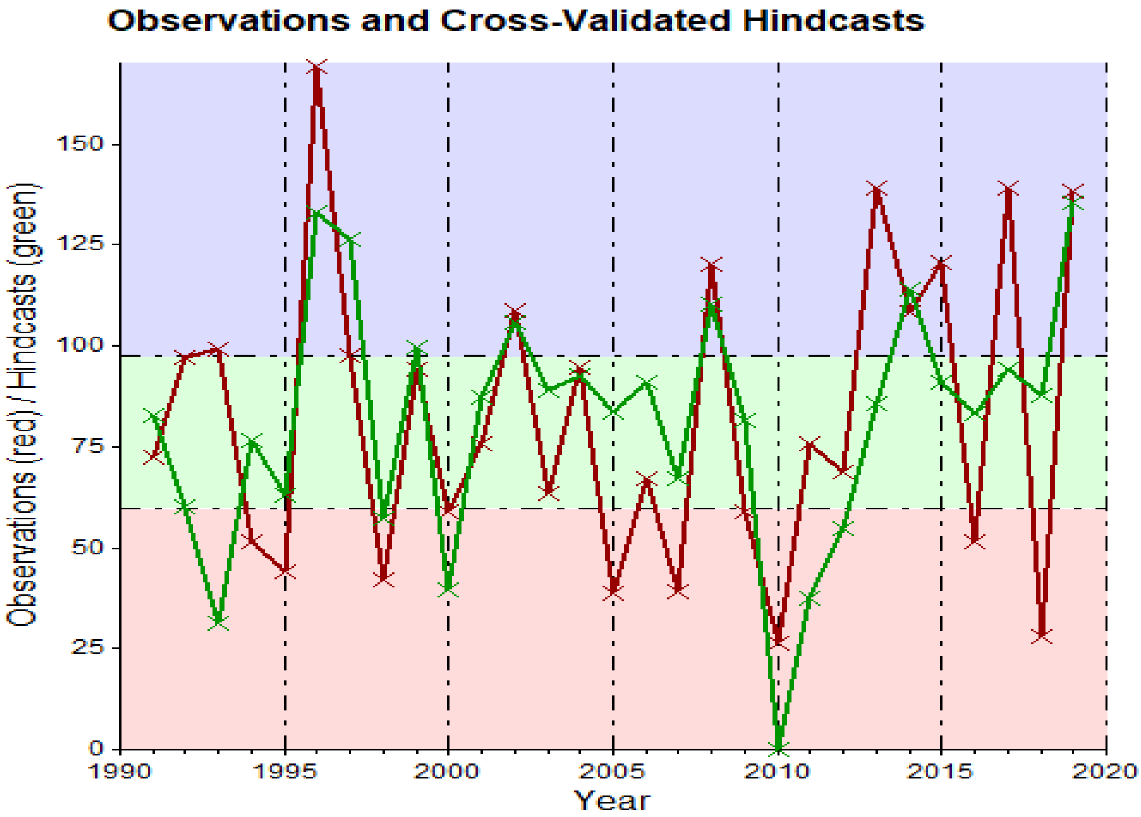


Figure 11: Observed Vs forecasted streamflow for MLR model

4.3.Predictors Influence Ranking for PCR model

The PCR approach provided a more stable and accurate forecast than MLR. The variance explained by the first PCs aligned with the dominant climate signals affecting snow accumulation and melt. Predictor ranking showed SWE as the strongest influence, followed by precipitation and temperature. This ranking shows which original predictors (e.g., SWE, Pr, Temp) have the greatest influence on the predicted streamflow in the PCR model. The figure shown below demonstrates that SWE (Snow Water Equivalent) is ranked highest because it directly represents snowpack, which melts in spring and contributes significantly to streamflow. Then after, Precipitation (Pr) and Temperature (Temp) follow, as they affect snow accumulation and melt rates. Lower-ranked predictors like MSLP and SP have less direct influence in the prediction of hindcast streamflow in the study area.

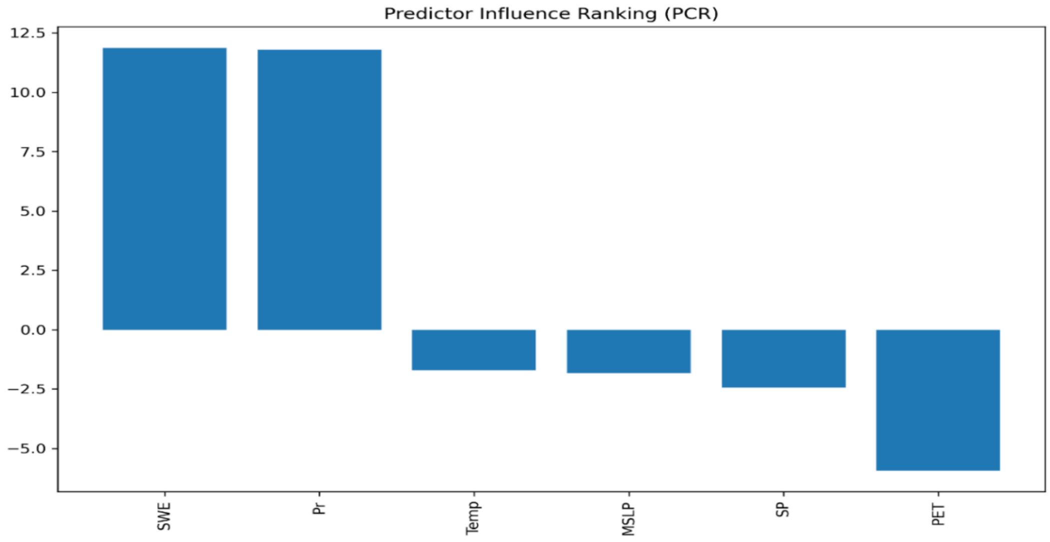


Figure 12: predictor influence ranking for PCR model

4.4. Scatter plots of observed and hindcasted streamflow for PCR and MLR model

The following figures show scatter plots of observed and hindcasted streamflow for PCR and MLR models respectively. In this scatter plots, the model performance can be evaluated by considering the closeness of data points to the best fit line. Based on this, scatter plots of observed and hindcasted streamflow for PCR model is better compared to MLR scatter plot. This indicates that PCR effectively solved multicollinearity problem, and this is also demonstrated in the correlation matrices.

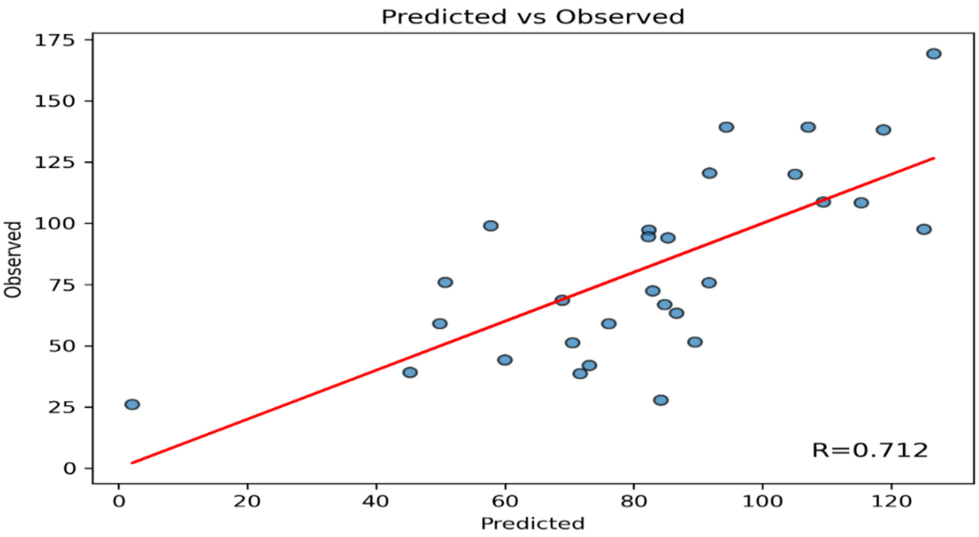


Figure 13: Scatter plots of observed and hindcasted streamflow for PCR model

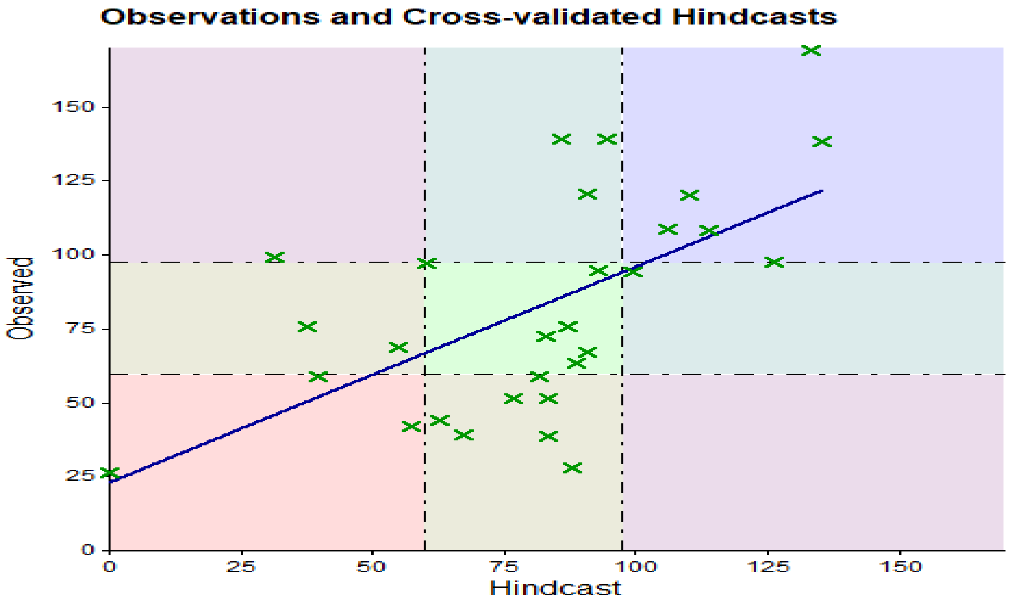


Figure 14: Scatter plots of observed and hindcasted streamflow for MLR model

4.5. Contingency table

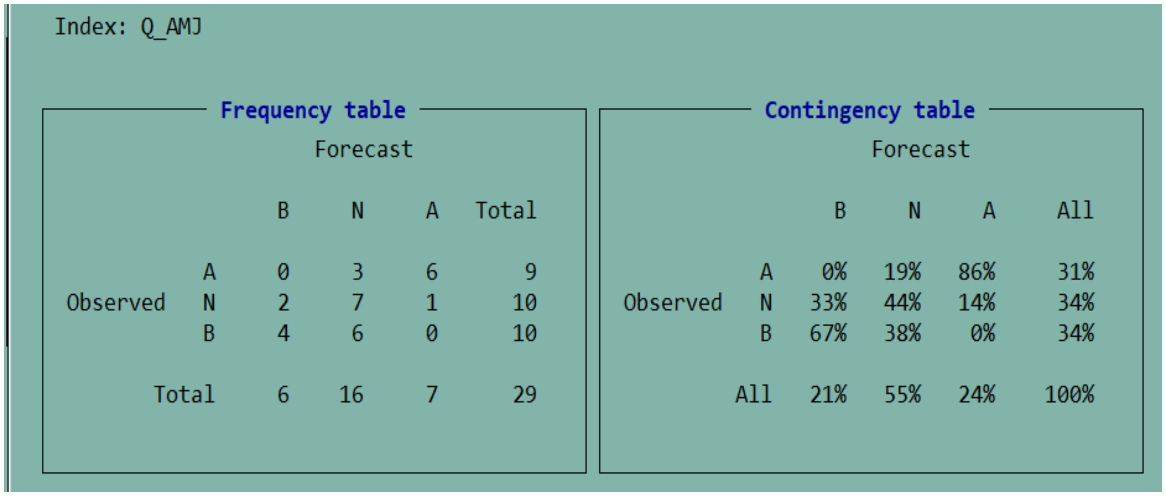
Contingency table is used to evaluate categorical forecast skill (e.g., above-normal, normal, below-normal flows). The contingency table analysis presented in the figure below evaluates the PCR model's skill in correctly classifying AMJ streamflow into three categories. The analysis is based on 29 years of hindcasted streamflow (1991-2019), with flow categories defined relative to historical climatological values. The table presents both frequencies and contingency.

The categorical thresholds can be defined as follows:

1. Below-normal (B): Streamflow below the 33rd percentile of historical AMJ flows
2. Normal (N): Streamflow between the 33rd and 67th percentiles
3. Above-normal (A): Streamflow above the 67th percentile

These tercile-based categories represent equal climatological probabilities (approximately 33% each) and provide a balanced framework for assessing categorical forecast skill across different flow regimes. This contingency table analysis evaluates the Principal Component Regression (PCR) model's skill in categorizing April-May-June streamflow into Below-normal (B), Normal (N), and Above-normal (A) classes for the Michipicoten watershed (1991-2019). The model results demonstrate moderate performance prediction skill. In the table, 67%, 44% and 86% shows the best predictions for each category.

Table 2: Contingency table for PCR model



```
 Index: Q_AMJ

 ─────────────── Frequency table ─────────       ─────────────── Contingency table ─────────
                      Forecast                                       Forecast

                   B     N     A   Total                          B      N      A      All

             A     0     3     6       9                   A     0%    19%    86%     31%
  Observed   N     2     7     1      10        Observed   N    33%    44%    14%     34%
             B     4     6     0      10                   B    67%    38%     0%     34%

       Total       6    16     7      29                 All    21%    55%    24%    100%
```

4.6. Model performance Evaluation

The table below shows that statistical model evaluation metrics demonstrates that the Principal Component Regression (PCR) model significantly outperforms the Multiple Linear Regression (MLR) model across all key performance metrics for predicting April–May–June (AMJ) streamflow in the Michipicoten River watershed. The Pearson correlation coefficient of R = 0.71 indicates a strong positive linear relationship between hindcasted and observed streamflow values. This correlation accounts for approximately 50.7% of the variance in observed streamflow (R^2 = 0.51), which represents moderate to good predictive skill for a seasonal hydrological forecasting model with a three-month lead time.

The reasons for PCR outperformed than MLR model that MLR uses original predictors directly, which were highly correlated (as shown in the heatmap). This leads to multicollinearity, making the model unstable and less reliable. PCR uses uncorrelated PCs, eliminating multicollinearity and resulting in a more stable, accurate model with better model evaluation metrics values.

Table 3: model performance evaluation metrics for both PCR and MLR models

S.no.	Model	Evaluation metrics				
		RMSE	MAE	Pearson R	NSE	RPSS
1	PCR	25.88	22.01	0.71	0.51	0.716
2	MLR	30.72	25.59	0.5953	0.31	

Conclusions

The development and performance evaluation of Principal Component Regression (PCR) and Multiple Linear Regression (MLR) models for pre-seasonal (AMJ) streamflow forecasting using JFM six predictors in the Michipicoten watershed demonstrated good result. PCR provided a significantly better forecast than MLR, mainly due to its ability to transform correlated hydroclimatic predictors into independent principal components by reducing multicollinearity problem which is the primary limitation of traditional regression approaches including MLR.

The PCR model analysis confirmed that SWE, precipitation, and temperature are the dominant climate variables controlling spring runoff in the study area. The most dominant predictors were SWE>PR>TEMP>MSLP>SP>PET for AMJ hindcasted streamflow. The strong influence of SWE aligns with the expectations that winter snowpack serves as a source for spring streamflow in the snow dominated regions. Precipitation and temperature act as secondary controls influencing both snow accumulation and melt rates to feed the river runoff.

Model evaluation metrics (R, RMSE, MAE, NSE, RPSS) collectively show that PCR achieved higher predictive skill, better representation of interannual variability and good hindcasts during both wet and dry years. While MLR captured some peak flows and the model performance was not high because of high correlations between predictors. Overall, the study demonstrates that PCR is the better approach for seasonal streamflow prediction in snow dominated watersheds with high predictor correlation.

Recommendations

To improve the model’s skill to predict preseason streamflow that best matches with the observed streamflow, the following improvements should be done including increase the number of predictors by incorporating additional large scale climate indices finer resolution datasets, integrate statistical models with machine learning models to improve model prediction robustness, incorporate the impacts of uncertainty from various sources in the prediction of streamflow, include other models such as persistence model for comparison and to help for stakeholder water resources managers to make decisions, split the available data for training and prediction and also include the impacts of climate change assessment on the prediction of streamflow.

References

1. Arnal, L., Clark, M. P., Pietroniro, A., Vionnet, V., Casson, D. R., Whitfield, P. H., Fortin, V., Wood, A. W., Knoben, W. J. M., Newton, B. W., & Walford, C. (2024). FROSTBYTE: a reproducible data-driven workflow for probabilistic seasonal streamflow forecasting in snow-fed river basins across North America. Hydrology and Earth System Sciences, 28, 4127–4155.
2. Arnal, L., Clark, M. P., Pietroniro, A., Vionnet, V., Casson, D. R., Whitfield, P. H., Fortin, V., Wood, A. W., Knoben, W. J. M., Newton, B. W., & Walford, C. (2024). FROSTBYTE: A reproducible data-driven workflow for probabilistic seasonal streamflow forecasting in snow-fed river basins across North America. Hydrology and Earth System Sciences, 28, 4127–4155.
3. Barnett, T. P., Adam, J. C., & Lettenmaier, D. P. (2005). Potential impacts of a warming climate on water availability in snow-dominated regions.
4. Fleming, S. W., & Garen, D. C. (2022). Simplified Cross-Validation in Principal Component Regression (PCR) and PCR-Like Machine Learning for Water Supply Forecasting. JAWRA Journal of the American Water Resources Association, 58(4), 528–546.
5. Hydroclimatology for Water Resources Management course materials
6. Luce, C. H., Staab, B., Kramer, M., Wenger, S. J., Isaak, D., & McConnell, C. (2014). Sensitivity of summer stream temperatures to climate variability in the Pacific Northwest.
7. Risley, J. C., Gannett, M. W., Lea, J. K., & Roehl, E. A. Jr. (2005). An Analysis of Statistical Methods for Seasonal Flow Forecasting in the Upper Klamath River Basin of Oregon and California. U.S. Geological Survey Scientific Investigations Report 2005-5177.
8. Wilby, R. L., Wedgbrow, C. S., & Fox, H. R. (2004). Seasonal predictability of the summer hydrometeorology of the River Thames, UK.
9. Wuttichaikitcharoen, P., & Babel, M. S. (2014). Principal component and multiple regression analyses for the estimation of suspended sediment yield in ungauged basins of Northern Thailand.
10. Razavi, T., & Coulibaly, P. (2013). Streamflow prediction in ungauged basins: Review of regionalization methods. Journal of Hydrologic Engineering

Appendixes

Model output results

Table 4: PCA components for PCR model

PC1	PC2	PC3	PC4	PC5	PC6
-0.96633	1.263317	-0.86911	0.002104	0.177793	-0.01864
-0.46151	-0.37577	-1.06134	0.858064	-0.39167	-0.00027
2.427839	-2.18746	0.052448	0.682084	-0.20033	-0.00819
0.932995	-2.24485	-1.15547	-0.934	-0.16395	-0.00306
0.866089	0.764768	-0.62416	-0.8651	-0.29968	0.028743
-2.62359	-0.6275	0.152944	-0.36479	0.787887	-0.00774
-2.09666	-0.84678	0.774719	0.407329	0.218914	0.009638
0.865343	0.54782	0.790592	0.33433	-1.21284	0.00716
0.62203	0.321426	1.650738	-0.50631	0.618799	-0.0192
1.730626	1.440183	0.14434	-0.96662	-0.66058	-0.02062
-0.34788	-0.33595	0.271901	0.644846	-0.22694	0.008139
-1.85317	0.9607	0.490937	-0.21054	-0.13299	0.002599
-1.16679	-0.816	-2.08854	-0.05083	-0.28137	-0.00368
0.975356	-1.28902	0.914764	-0.12444	-0.00426	-0.0131
1.494268	-1.24608	0.612292	0.33985	0.036756	-0.01165
-0.52263	1.291515	0.166443	0.549457	-0.32995	-0.01995
1.492639	1.143997	-1.00723	-0.37922	0.730055	-0.0009
-1.4312	-0.26535	-0.04045	0.548977	-0.04749	-0.01712
0.754161	-1.07309	-0.06807	0.327562	0.293622	-0.0009
4.068464	2.422383	-0.70881	0.710109	1.08341	0.015112
1.683347	-0.66086	-1.15535	-0.52929	-0.33012	0.001504
-0.2166	3.687018	0.332718	0.007775	-0.17679	0.002397
-1.18172	-0.22515	0.665422	-0.15091	-0.27537	0.019416
-2.91775	-0.50984	-1.63004	0.069915	0.310439	0.012061
0.072414	-2.53409	-0.34945	0.125863	0.271056	0.002633
-1.17212	1.770919	-0.13327	-0.06779	-0.29163	0.001258
-1.2096	2.305877	0.706556	0.122071	0.187553	0.004236
1.216483	-1.33032	1.554441	0.10054	-0.06949	0.02394
-1.03448	-1.35182	1.610036	-0.68102	0.379183	0.006175

Table 5: PCA loadings for PCR model

	PC1	PC2	PC3	PC4	PC5	PC6
Temp	0.153432	0.559171	0.422574	0.413789	-0.53218	0.175448
Pr	-0.47886	0.093416	0.598031	-0.63321	-0.05802	-0.00191
SWE	-0.43212	-0.34526	0.415881	0.63761	0.338472	-0.0005
PET	0.339799	0.501139	0.218725	-0.04646	0.763787	0.005188
MSLP	0.439271	-0.45147	0.282935	-0.13435	0.006772	0.71068
SP	0.502002	-0.32315	0.403658	-0.03263	-0.12426	-0.68126

Table 6: PCR comprehensive results for PCR model

Year	Observed	Predicted	Observed Category	Predicted Category
1991	72.36666667	82.88320215	Normal	Normal
1992	97.23333333	82.34480652	Normal	Normal
1993	99	57.79054474	Above	Below
1994	51.26666667	70.43432298	Below	Normal
1995	44.2	59.95606023	Below	Below
1996	169.3666667	126.5678665	Above	Above
1997	97.56666667	125.0397857	Above	Above
1998	42	73.08164566	Below	Normal
1999	94.13333333	85.22752269	Normal	Normal
2000	59.03333333	49.82467166	Below	Below
2001	75.86666667	91.69218656	Normal	Normal
2002	108.7	109.3695964	Above	Above
2003	63.3	86.63731265	Normal	Normal
2004	94.53333333	82.22488943	Normal	Normal
2005	38.63333333	71.61058082	Below	Normal
2006	66.9	84.79563643	Normal	Normal
2007	39.23333333	45.25078721	Below	Below
2008	120.0666667	105.0201559	Above	Above
2009	59	76.105159	Below	Normal
2010	26.16666667	2.125024909	Below	Below
2011	75.93333333	50.70430319	Normal	Below
2012	68.7	68.86509581	Normal	Normal
2013	139.3	107.0113878	Above	Above
2014	108.4	115.3051148	Above	Above
2015	120.6	91.76886893	Above	Normal
2016	51.6	89.51473307	Below	Normal
2017	139.2666667	94.34084074	Above	Normal
2018	27.8	84.21510562	Below	Normal
2019	138.3	118.7594587	Above	Above

Table 7: PCR comprehensive results for MLR model

Year	Observed	Predicted	Observed Category	Predicted Category
1991	72.36666667	91.90287706	Normal	Normal
1992	97.23333333	73.83782378	Normal	Normal
1993	99	55.51081432	Above	Below
1994	51.26666667	70.25407347	Below	Normal
1995	44.2	47.23193101	Below	Below
1996	169.3666667	143.701071	Above	Above
1997	97.56666667	125.1276565	Above	Above
1998	42	48.63312398	Below	Below
1999	94.13333333	103.226166	Normal	Above
2000	59.03333333	46.39928048	Below	Below
2001	75.86666667	83.87404537	Normal	Normal
2002	108.7	106.5877122	Above	Above
2003	63.3	82.88763628	Normal	Normal
2004	94.53333333	86.50393554	Normal	Normal
2005	38.63333333	75.28290425	Below	Normal
2006	66.9	84.15955946	Normal	Normal
2007	39.23333333	59.23391388	Below	Below
2008	120.0666667	108.5169109	Above	Above
2009	59	80.98596244	Below	Normal
2010	26.16666667	15.28590669	Below	Below
2011	75.93333333	45.36714919	Normal	Below
2012	68.7	64.95368728	Normal	Normal
2013	139.3	96.29660743	Above	Normal
2014	108.4	116.898584	Above	Above
2015	120.6	95.52095399	Above	Normal
2016	51.6	84.06300091	Below	Normal
2017	139.2666667	96.11138078	Above	Normal
2018	27.8	75.26048109	Below	Normal
2019	138.3	124.8515174	Above	Above