INDIAN INSTITUTE OF TECHNOLOGY PALAKKAD

Prediction of Streamflow using Machine Learning

Siri Sure
102101034

Guide: Dr. Athira P
Department of Civil Engineering
Indian Institute of Technology Palakkad

May 13, 2025

Overview

1. Introduction
2. Previous work
3. Climate and LULC Impact Using SHAP
4. Seasonal Analysis
5. Model performance
6. Conclusions

Introduction

Overview and Problem Statement

- Predicting streamflow accurately is critical for flood forecasting, reservoir operations, and water management.
- Traditional physically-based models are complex and data-intensive.
- Need for data-driven models that can learn directly from historical data.

Objectives and Motivation

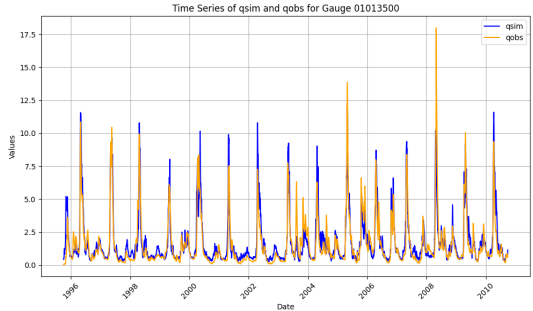
- Develop LSTM-based model for gauged basin streamflow prediction.
- Application of LSTM for unguaged basin prediction.
- Apply **SHAP** to understand how **climate and land use** affect streamflow over time.

Expected Impact

- **Knowledge Transfer:** Learn from data-rich basins and apply to data-scarce ones.
- **Hydrological Insight:** Understand hydrological dynamics using interpretable ML tools.

Recap of previous work

- Initial model (last semester) gave R square of 0.71.
- Added more temporal and static inputs + tuned dropout.
- Final model achieved R square of 0.90 and NSE = 0.8993 on gauged Basin **1013500**.
- Cosine similarity used to find a hydrologically similar ungauged Basin **1030500**.
- Validation on this basin gave NSE = 0.6946 **without using its streamflow data**.
- Confirms model's transferability to ungauged basins.



Feature Inputs for LSTM and SHAP Models

Basins used: 1013500, 1030500, 8029500, 7197000, 5525500

LSTM Model

- **Temporal Features (Daily)**
  - Precipitation
  - Tmax (Daily Maximum Temperature)
  - Tmin (Daily Minimum Temperature)
- **Static / Non-Temporal Features**
  - Elevation, Slope, Area
  - Soil Depth, Soil Type
  - NDVI, Land Cover Class
  - Baseflow Index, Runoff Ratio, Aridity Index
  - Drainage Density, Forest Fraction, Urban Fraction

SHAP Analysis Model

- **Yearly Aggregated Features (1953–2018)**
  - Precipitation (Annual)
  - AET (Actual Evapotranspiration)
  - Tmin, Tmax (Annual Averages)
  - Urban and Agricultural Area Fractions
- **Monthly Aggregated Features (Seasonal SHAP)**
  - Precipitation, AET, Urban
  - Tmin, Tmax (per season)
  - Seasons: Winter, Spring, Summer, Autumn

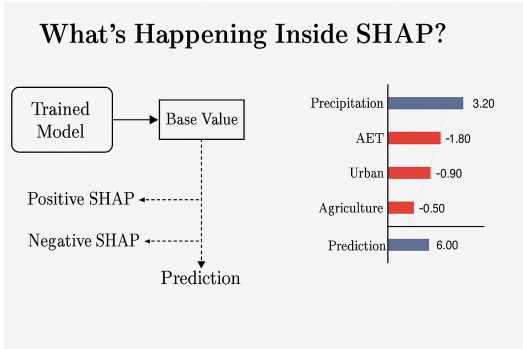
Overview and Motivation

- Investigate the impact of climate variables and land use changes on streamflow and hydrological dynamics over time.
- Apply interpretable machine learning (SHAP) to understand feature contributions and capture hydrological behavior.
- Analyze one representative basin in detail (8029500) and summarize results for others.

What is SHAP and Why Use It?

SHAP (SHapley Additive exPlanations) is a method based on game theory that explains how much each feature (like rainfall, temperature, or urban area) contributed to a prediction.

- SHAP assigns each feature a value that shows:
  - **Magnitude**: How much the feature influenced the prediction.
  - **Direction**: Whether it pushed the prediction higher or lower.
- It provides consistent, local explanations.



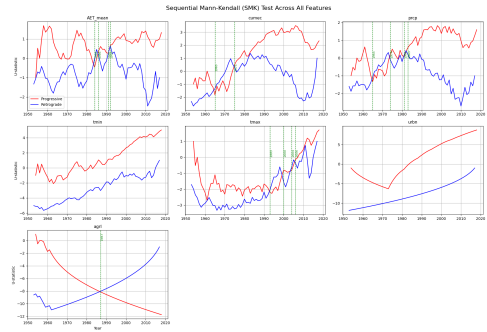
SHAP values visualize feature impact on model predictions, not just average importance.

Why Boosting + SHAP Instead of LSTM?

- LSTM effectively captures temporal patterns but operates as a black box, making interpretation challenging.
- Gradient boosting (e.g., XGBoost) is well-suited for tabular data and provides clearer insights.
- SHAP with XGBoost allows for transparent, feature-level explanations of streamflow dynamics.
- This combination enables understanding of how climate and land features impact hydrological behavior.

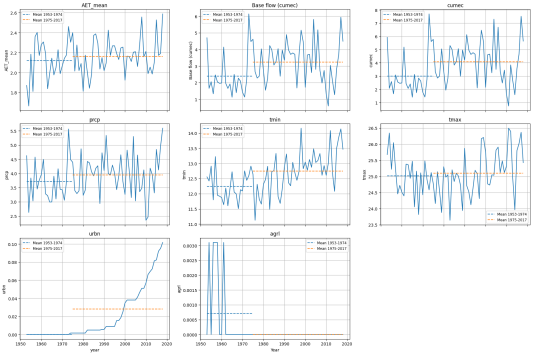
Sequential Mann-Kendall Test for Change Detection

- In hydrology, we often need to test whether a time series (e.g., rainfall, streamflow) is **stationary** or has changed over time.
- The Sequential Mann-Kendall (SMK) test helps detect the **year of change (change point)**.
- It plots:
  - **Progressive (UF)** and **Retrograde (UB)** trend lines
  - Their intersection = **significant shift**



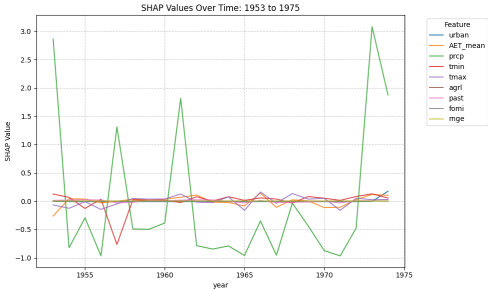
SMK plot showing UF and UB lines across features. Change point identified around 1975.

Mean Comparison Before and After 1975

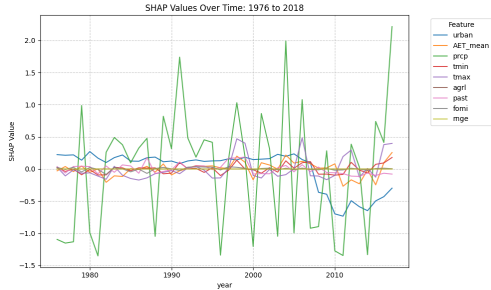


- Increase in Tmin, Tmax, and precipitation.
- Urbanization impact observed post-1975.
- Agricultural land area decreased.
- Higher runoff patterns detected.

SHAP Feature Contributions Over Time



1953–1975

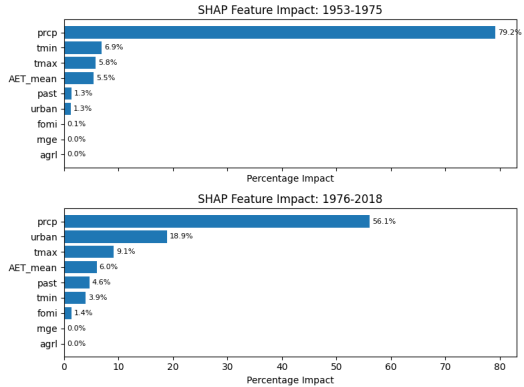


1976–2018

- Precipitation consistently dominant.
- Urbanization influence rises post-1975.

Percentage contributions of key features

- Precipitation remained the most influential driver.
- Urbanization gained significant impact after 1975.
- Climate + Land Use jointly influence streamflow in recent decades.
- Clear shift in dominant features before and after change point.



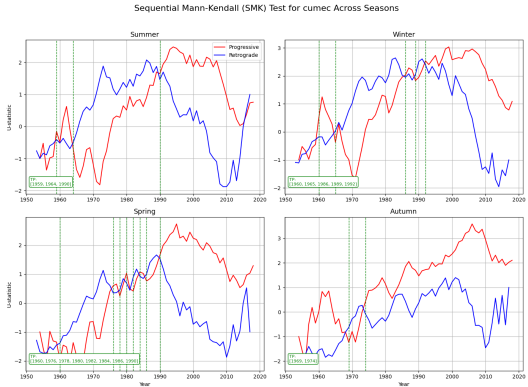
SHAP based feature impact across time

Why Seasonal Analysis?

- Streamflow dynamics vary significantly across seasons.
- Seasonal drivers like **snowmelt, monsoon, evapotranspiration** affect different months differently.
- Aggregating SHAP by season improves:
  - Interpretability of model outputs.
  - Detection of seasonal trends in climate/LULC impact.
- Enables climate-adaptive water resource planning by understanding:
  - What dominates in **Winter** vs. **Summer**, etc.

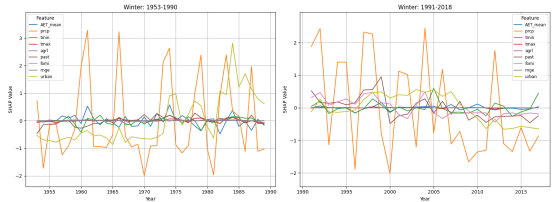
Seasonal Analysis and SMK

- Performed SMK test separately for each season.
- Detected season-specific change points:
  - **Winter:** 1975
  - **Spring:** 1972
  - **Summer:** 1980
  - **Autumn:** 1979
- SHAP was then applied pre/post change point for each season.

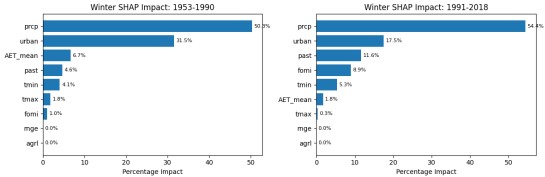


SMK plots for seasonal data (Winter, Spring, Summer, Autumn)

SHAP Seasonal Feature Impact – Winter



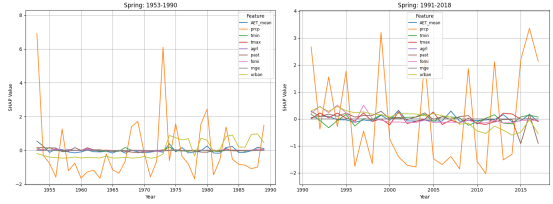
Impact of features on stream flow for winter



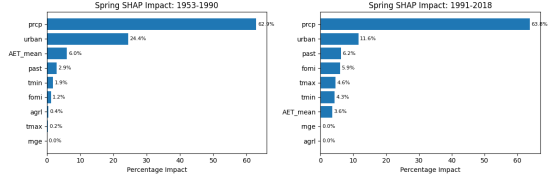
Winter: SHAP feature contributions (before and after the change point)

- **Precipitation** remained the dominant driver, increasing after 1990.
- **Urban influence** dropped significantly, indicating reduced urban impact on winter runoff.
- Soil-related features like **past** and **fomi** gained importance post-change, highlighting stronger soil-driven responses.
- Suggests a shift towards more natural hydrological control during winter flows after 1990.

SHAP Seasonal Feature Impact – Spring



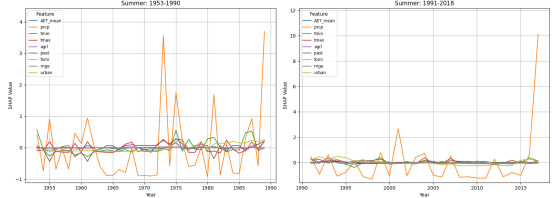
Impact of features on stream flow for Spring



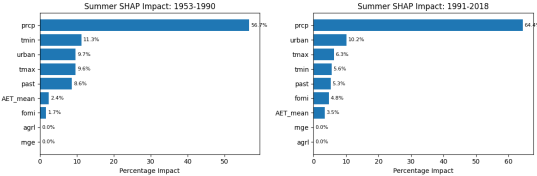
Spring: SHAP feature contributions (before and after the change point)

- Precipitation remained the dominant driver across both periods.
- Post-1975: Urban impact decreased, while contributions from **past** and **fomi** significantly increased.
- This shift indicates a stronger influence of historical land characteristics (past) and soil moisture (fomi) on streamflow during spring.

SHAP Seasonal Feature Impact – Summer



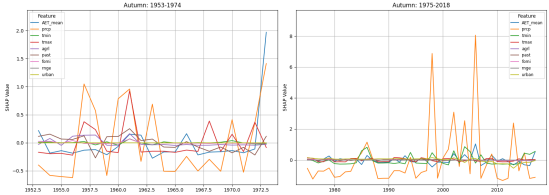
Impact of features on stream flow for Summer



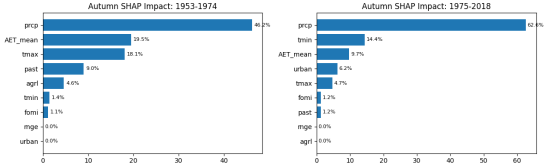
Summer: SHAP feature contributions (before and after the change point)

- Urban land use showed clear monotonic increase in SMK and was among top SHAP contributors.
- Tmax and precipitation continued to dominate, with urban runoff becoming critical post-2000.
- The changes suggest summer runoff became more sensitive to temperature drivers after 1990.

SHAP Seasonal Feature Impact – Autumn



Impact of features on stream flow for Autumn



Autumn: SHAP feature contributions (before and after the change point)

- AET and precipitation displayed trend shifts in SMK plots.
- SHAP values were more evenly distributed among features.
- Weak urban influence but detectable agricultural impact due to harvest-related land changes.
- Reflects increased stormwater sensitivity post-land use change.

SHAP Summary for Other Basins

- **Basin 7197000 (Mixed Humid):**
  - Dominant features: Precipitation, Urban fraction, and Tmax post-1991.
  - Urban SHAP values increased in winter, suggesting runoff influence.
- **Basin 5525500 (Cold Region):**
  - Temperature (Tmax) was most influential, especially in spring and summer.
  - Agriculture influence low and declined after 1990.
- **Observation:**
  - SHAP analysis captured basin-specific patterns in feature importance.
  - Urbanization and temperature emerged as growing contributors post change point.

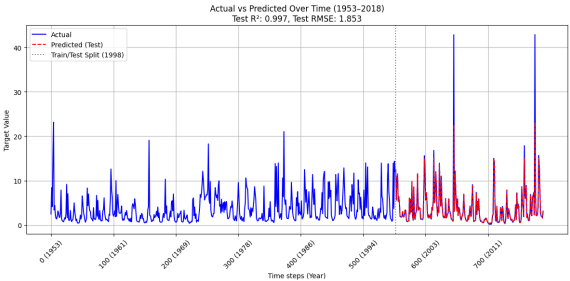
Model Performance Across Basins

- XGBoost models trained using 70% of data and tested on remaining 30%.
- Evaluated on test data (2002–2018) using metrics like R squared, RMSE, Bias.
- Three representative basins tested: Hot Humid, Mixed Humid, and Cold Region.

Basin ID	Climate Zone	Change Point	Top Features After	R squared	RMSE
8029500	Hot Humid	1975	Precip, Urban, AET	0.997	0.86
7197000	Mixed Humid	1991	Precip, Urban, Tmax	0.925	0.965
5525500	Cold Region	1990	Tmax, Precipitation	0.99	0.995

Table: Summary of model performance across three basins

Prediction vs Observed - Basin 8029500 (Hot Humid)



- Model trained with 70% of data, tested on 30% (2002–2018).
- Achieved the best performance among all basins.
- **R squared: 0.997**, **NSE: 0.861**, **RMSE: 0.860**

Insights from Model Results

- **Precipitation** remained the most dominant predictor across all basins and seasons.
- **Urbanization** emerged as a significant contributor post change points - reflecting LULC impact.
- **Seasonal SHAP** analysis highlighted temperature's influence in spring/winter and urban impact in summer.
- **Non-stationarity** detected via SMK helped isolate the impact of changing land/climate patterns.
- Slightly higher RMSE values in Mixed and Cold regions may reflect snowmelt and delayed hydrological responses not fully captured by the model.

Conclusion

- Developed an interpretable streamflow prediction framework using **XGBoost + SHAP**.
- Addressed **non-stationarity** by applying the Sequential Mann-Kendall (SMK) test to identify change points.
- SHAP revealed evolving contributions of features like **precipitation, urbanization, and temperature**.
- Seasonal SHAP analysis showed intra-annual variability - e.g., urban influence in summer, temperature in spring/winter.
- Results validated across diverse climate zones with **R squared ranging from 0.9 to 0.99**.
- The approach provides both **accurate predictions** and **explanations**, aiding future water resource planning.

Thank You!