Daily streamflow forecasting in mixed precipitation/snowmelt driven river basins using Machine Learning

Leo Pham
Michigan State University
What is streamflow?

Streamflow (or discharge m$^3$/s) is volume of water moving down a stream or river per unit of time.

Applications of streamflow forecast

- Flood prediction
- Water management and allocation
- Engineering design and research

Factors that impact streamflow:

- Precipitation
- Snowmelt
- Soil Moisture
- Temperature
Access the capability of a ML method to make streamflow forecast in precipitation/snowmelt dominated river basins with different hydrometeorological characteristics
- A semi-unsupervised ML algorithm within the Decision Tree family
- Uses of an ensemble of uncorrelated trees to yield prediction for classification and regression tasks (Criminisi et al. 2011)

**Hyperparameter | Description**
--- | ---
`mtry` | Number of candidate predictors available for splitting at each node
`sample size` | Number of observations that are drawn for each tree
`n-trees` | Number of trees in the forest
Part of the Columbia River Basin
States intersected: Washington, Oregon, Nevada, Idaho, Utah, Wyoming, and California
Heavily dammed
Have a long history of flooding (Neiman 2011)
Seasonal variation

- Most precipitation in this region occurs in the winter (Nov - March) and the summer (Jun-Aug) tends to be dry
- Mountain snowpack accumulation from winter provides important water storage
- Snowmelt in springtime (Apr-Jul) results in peak in river discharge (Knowles 2011)
Spatial variation

- Coastal region receives more precipitation than inland
- Uneven snow accumulation
Daily precipitation on 2009-01-02

Pacific Northwest Watershed
Data pre-processing and standardization

Predictor selection
1. Daily streamflow \( t-1 \)
2. Daily precipitation \( t-1 \)
3. Sum of precipitation from 3 previous days
4. Daily snow-water equivalent \( t-1 \)
5. \textit{Daily snowmelt} \( t-1 \)
6. Daily temperature max \( t-1 \)
7. Daily temperature min \( t-1 \)
8. Daily temperature range \( t-1 \)
9. Month index
10. Pentad index

Data Preparation

Building Model
Random Forest
Training and calibration

Output
1-7 day streamflow forecasts

Model Evaluation
1. Evaluation
   - Coefficient of determination
   - Root mean squared error
   - Nash-Sutcliffe efficiency
   - Kling-Gupta efficiency

2. Comparison
   - Multiple Linear Regression

Data Source

USGS Gages
Daily streamflow

PRISM AN81D
Daily precipitation

Snow Telemetry Stations
Daily snow-water equivalent
Daily maximum temperature
Daily minimum temperature

Flowchart of the streamflow forecasting
Data Source

USGS Gauges
Daily streamflow

PRISM AN81D
Gridded daily precipitation

SNOTEL Stations
Daily snow-water equivalent
Daily maximum temperature
Daily minimum temperature
Modeling training
and
Hyperparameter tuning

\( mtry = 3 \)

Sample size = bootstrapped sample of n-observations
n-trees = 300

![Random Forest Training Error](image)
Diagnostic results

Overall Performance

- $R^2$ Distribution
- Root Mean Squared Error Distribution
- Nash-Sutcliffe Efficiency Distribution
- Kling-Gupta Efficiency Distribution
Diagnostic results

Spatial variability in performance
Diagnostic results

Seasonal variability in performance

The graph shows the box plots for R² values across different months for one-day forecasts. The x-axis represents the months from January to December, and the y-axis represents the R² values. The box plots are split into two categories: precipitation-driven and snowmelt-driven.
Diagnostic results

Selected station analysis at USGS Gage 14179000

One-day forecast

Permutation Importance of Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>% Included in MAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>stream_1_d_lag</td>
<td>50</td>
</tr>
<tr>
<td>temp_min_1_d_lag</td>
<td>25</td>
</tr>
<tr>
<td>S_W_cqul_1_d_lag</td>
<td>10</td>
</tr>
<tr>
<td>temp_max_1_d_lag</td>
<td>5</td>
</tr>
<tr>
<td>sum_precip_3_d_lag</td>
<td>5</td>
</tr>
<tr>
<td>precip_1_d_lag</td>
<td>5</td>
</tr>
<tr>
<td>pentad_index</td>
<td>0</td>
</tr>
<tr>
<td>snowmelt_1_d_lag</td>
<td>0</td>
</tr>
<tr>
<td>temp_shift_1_d_lag</td>
<td>0</td>
</tr>
<tr>
<td>month_index</td>
<td>0</td>
</tr>
</tbody>
</table>

Prediction vs Observation

Time series between 2016-12-17 and 2017-07-25

Streamflow

2016 2017 2018

January March May July
Diagnostic results

Selected station analysis at USGS Gage 14179000

Three-day forecast
Initial Observations and Moving Forward

Observations
- There is a wide range in the predictive performance of the model across spatial sub-regions and between seasons
- Better performance in sub-regions with higher number of SNOTEL stations
- Model underestimates larger values (higher errors)
- Importance of variables vary with lead time prediction

Moving forward
- Examine outlier gages and impact of anthropogenic activities
- Sub-region analysis
- Consider better representation of precipitation input
- *Extend study period to better model extreme events
- Remove redundant predictor(s)
- Compare the model performance with previous studies


GAGES-II: Geospatial Attributes of Gages for Evaluating Streamflow

Thank you