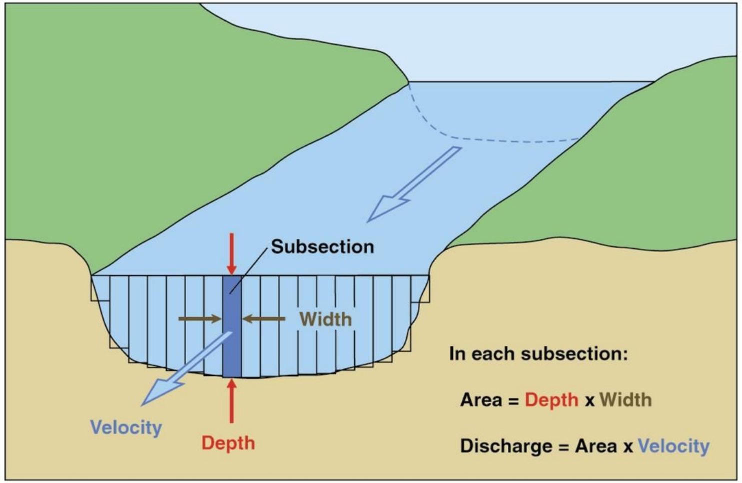
Daily streamflow forecasting in mixed precipitation/snowmelt driven river basins using Machine Learning

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.



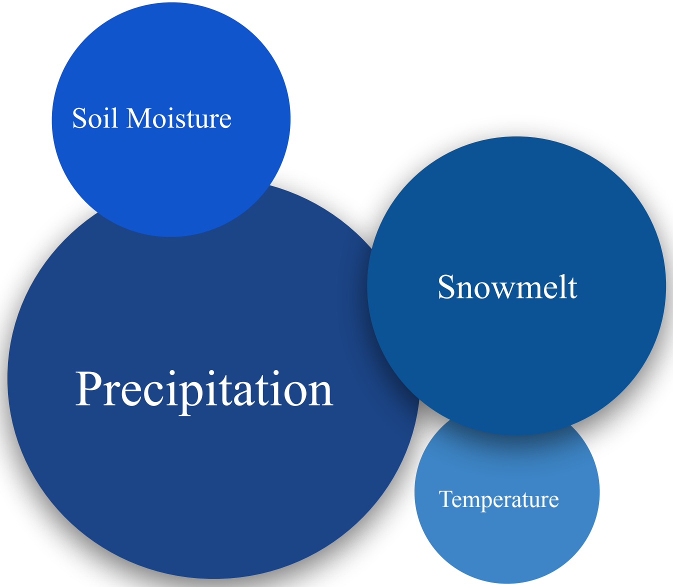
m^3/s m^2 m/s

Applications of streamflow forecast

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

(US Geological Survey)

Factors that impact streamflow



Soil Moisture

Precipitation

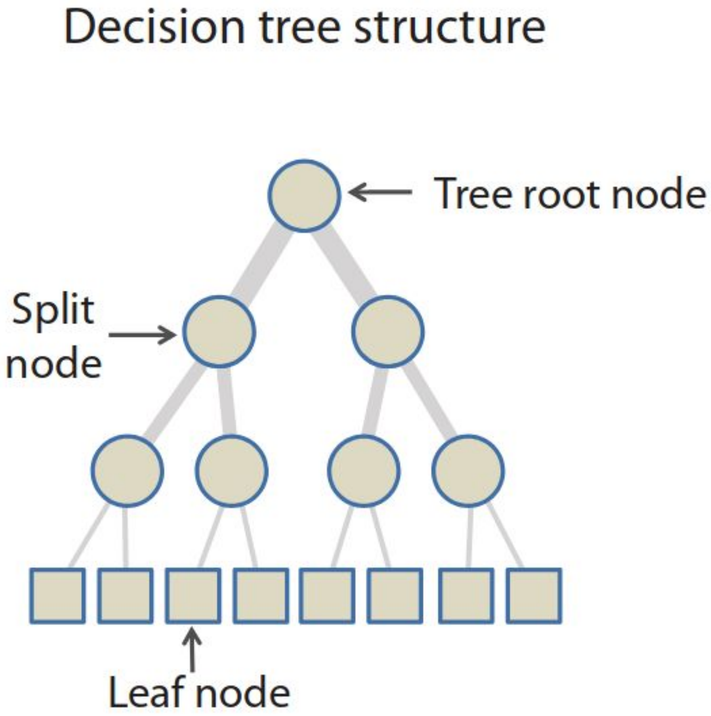
Snowmelt

Temperature

Study Objective

Access the capability of a ML method to make streamflow forecast in precipitation/snowmelt dominated river basins with different hydrometeorological characteristics

Random Forest

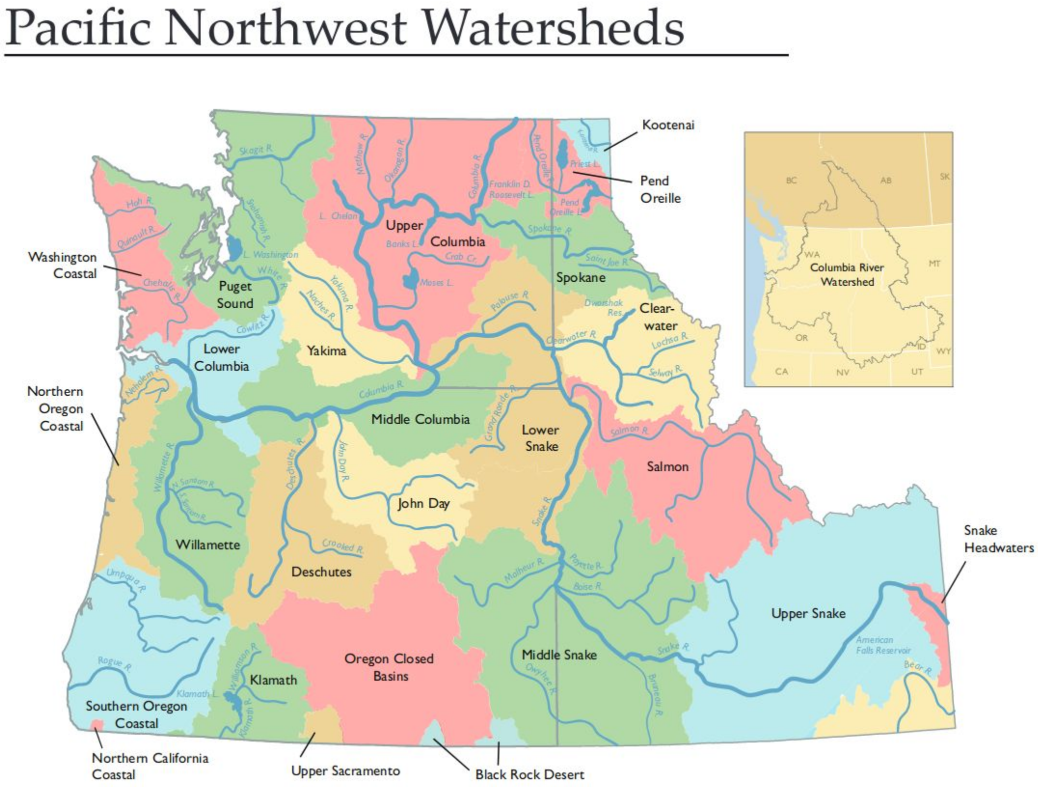


Sonka, based on (Criminisi et al., 2011)

- 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

Study Area



Pacific Northwest Watersheds

Portland State University, Department of Geography

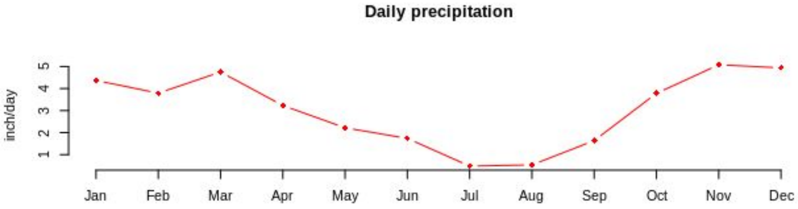
Part of the Columbia River Basin

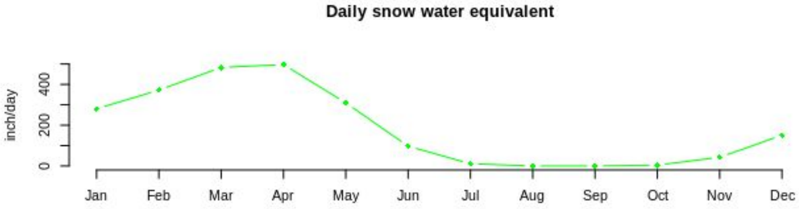
States intersected: Washington, Oregon, Nevada, Idaho, Utah, Wyoming, and California

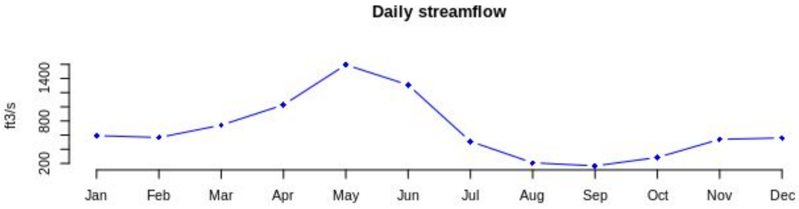
Heavily dammed

Have a long history of flooding (Neiman 2011)

Climate characteristics



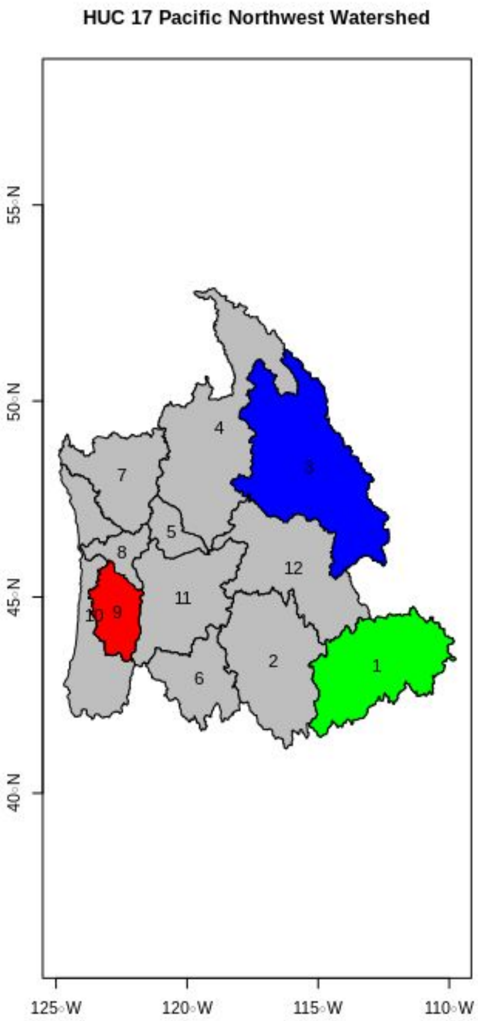


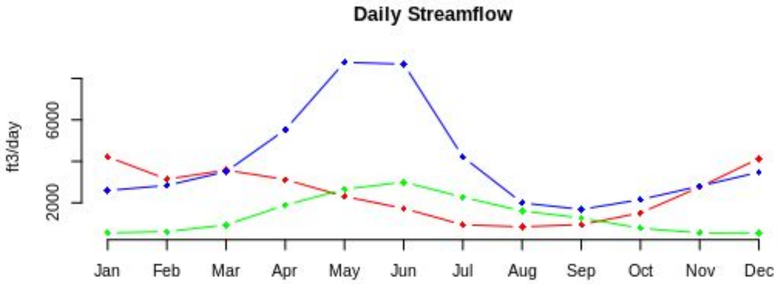


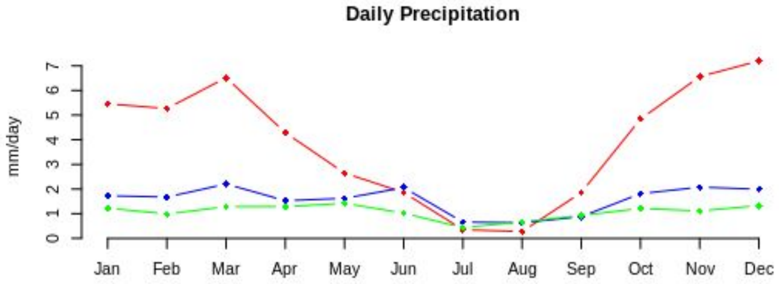
Seasonal variation

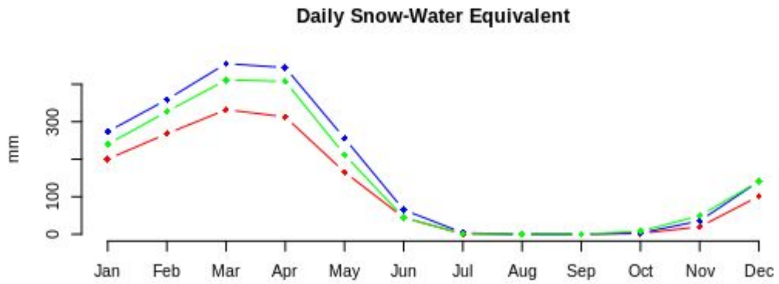
- Most precipitation in this region occurs in the winter (Nov - March) and the summer (Jun-Aug) tends to be dry
- Mountain snowpack accumulation from winter provides important water storage
- Snowmelt in springtime (Apr-Jul) results in peak in river discharge (Knowles 2011)

Climate characteristics





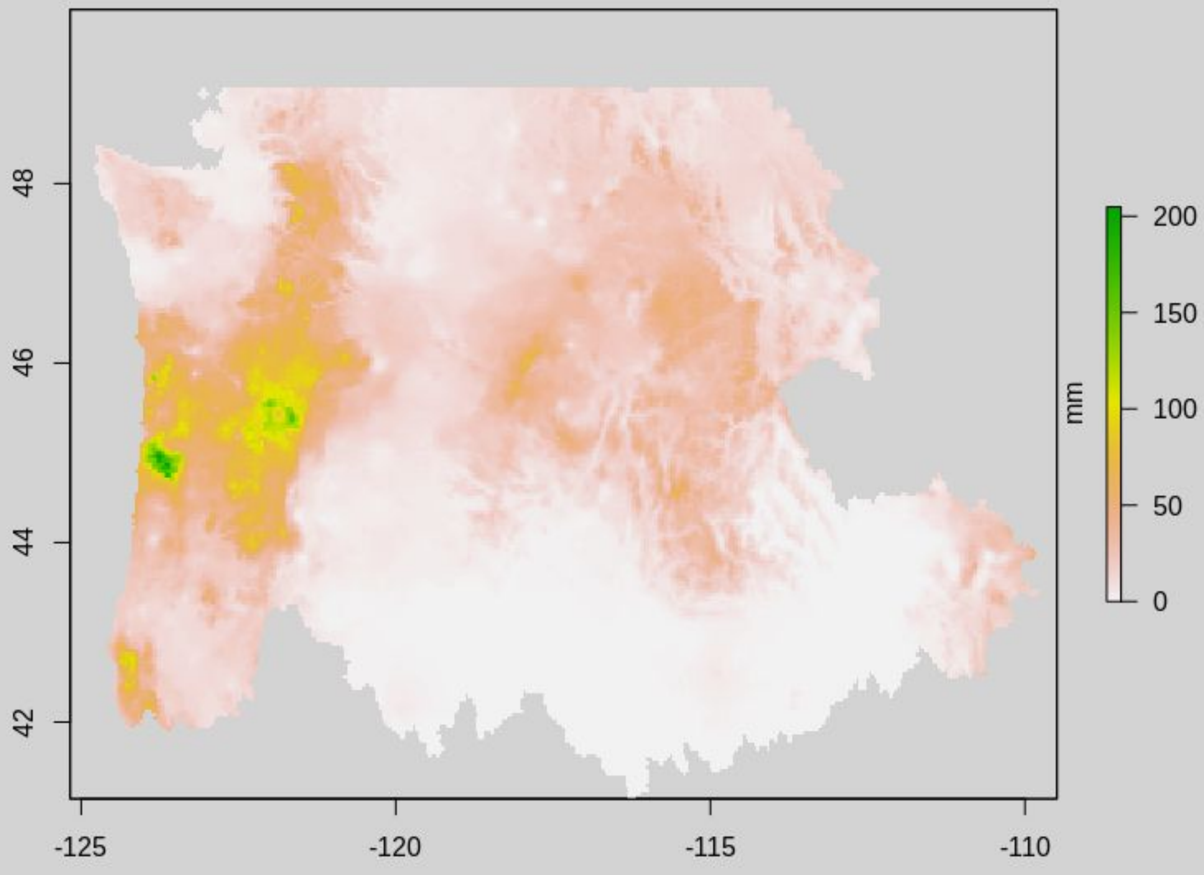




Spatial variation

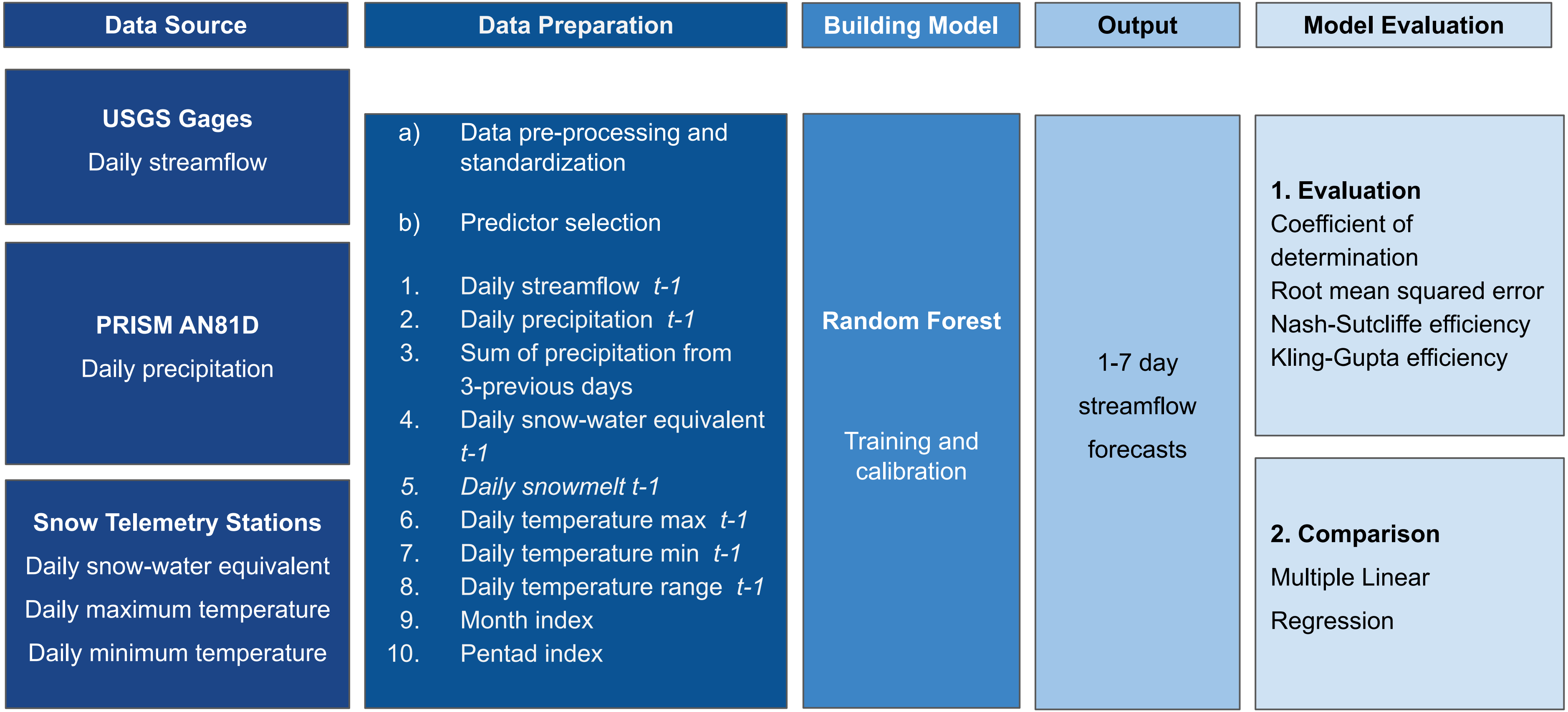
- Coastal region receives more precipitation than inland
- Uneven snow accumulation

Daily precipitation on 2009-01-02



Pacific Northwest Watershed

Flowchart of the streamflow forecasting



Data Source

USGS Gages

Daily streamflow

PRISM AN81D

Daily precipitation

Snow Telemetry Stations

Daily snow-water equivalent

Daily maximum temperature

Daily minimum temperature

Data Preparation

a) Data pre-processing and standardization

b) 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. *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

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 Gauges

Daily streamflow

PRISM AN81D

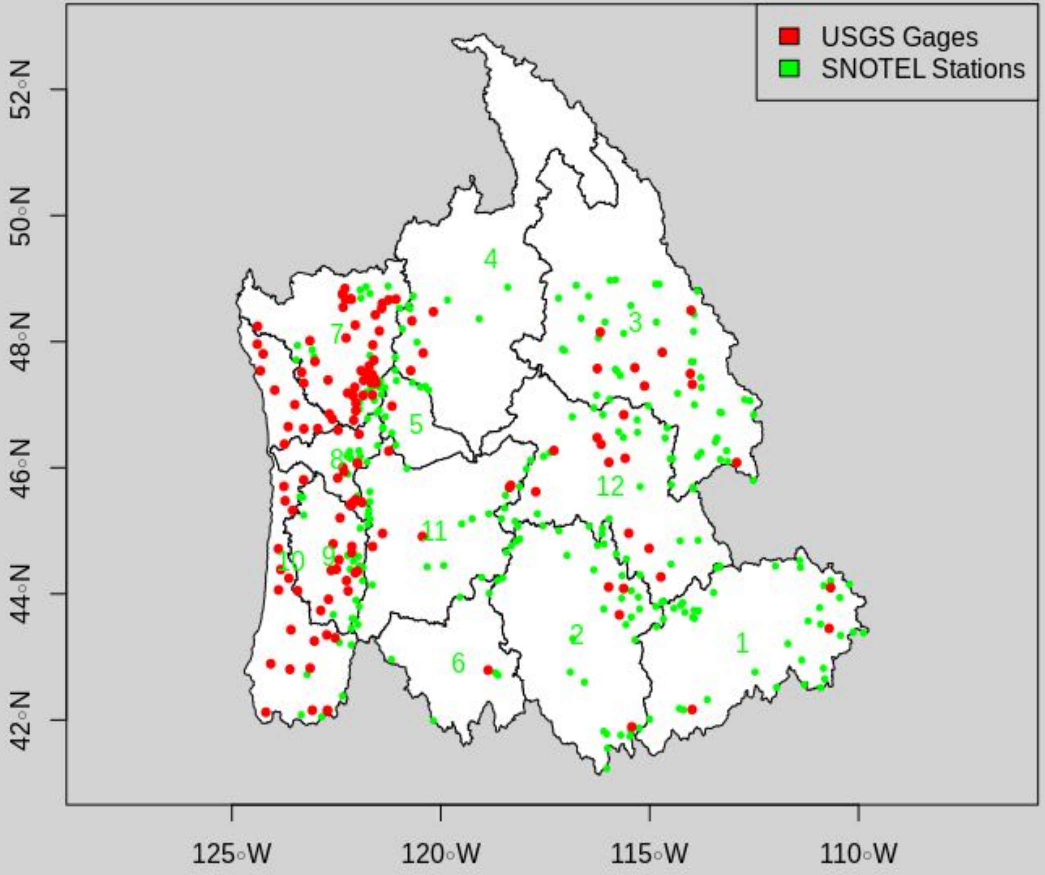
Gridded daily precipitation

SNOTEL Stations

Daily snow-water equivalent

Daily maximum temperature

Daily minimum temperature

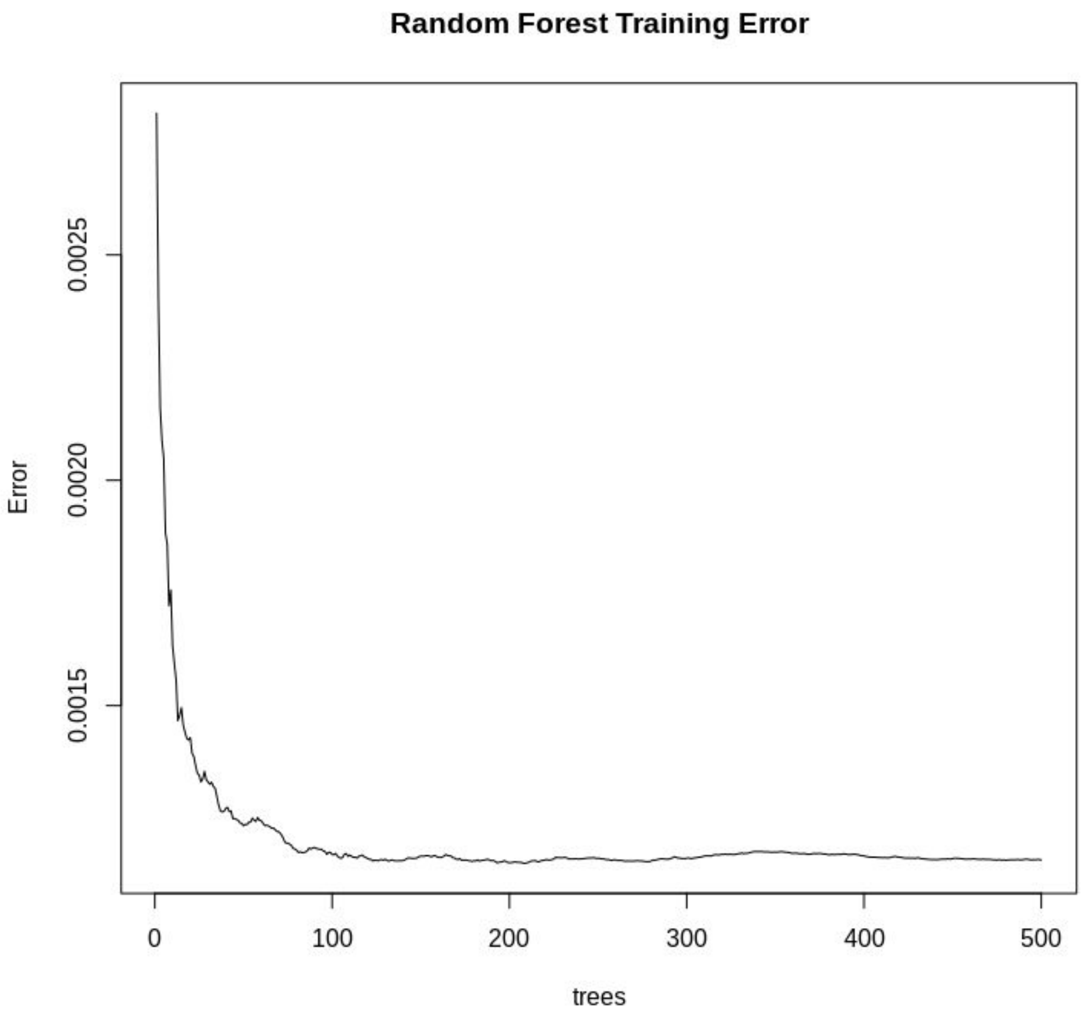


Modeling training and Hyperparameter tuning

mtry = 3

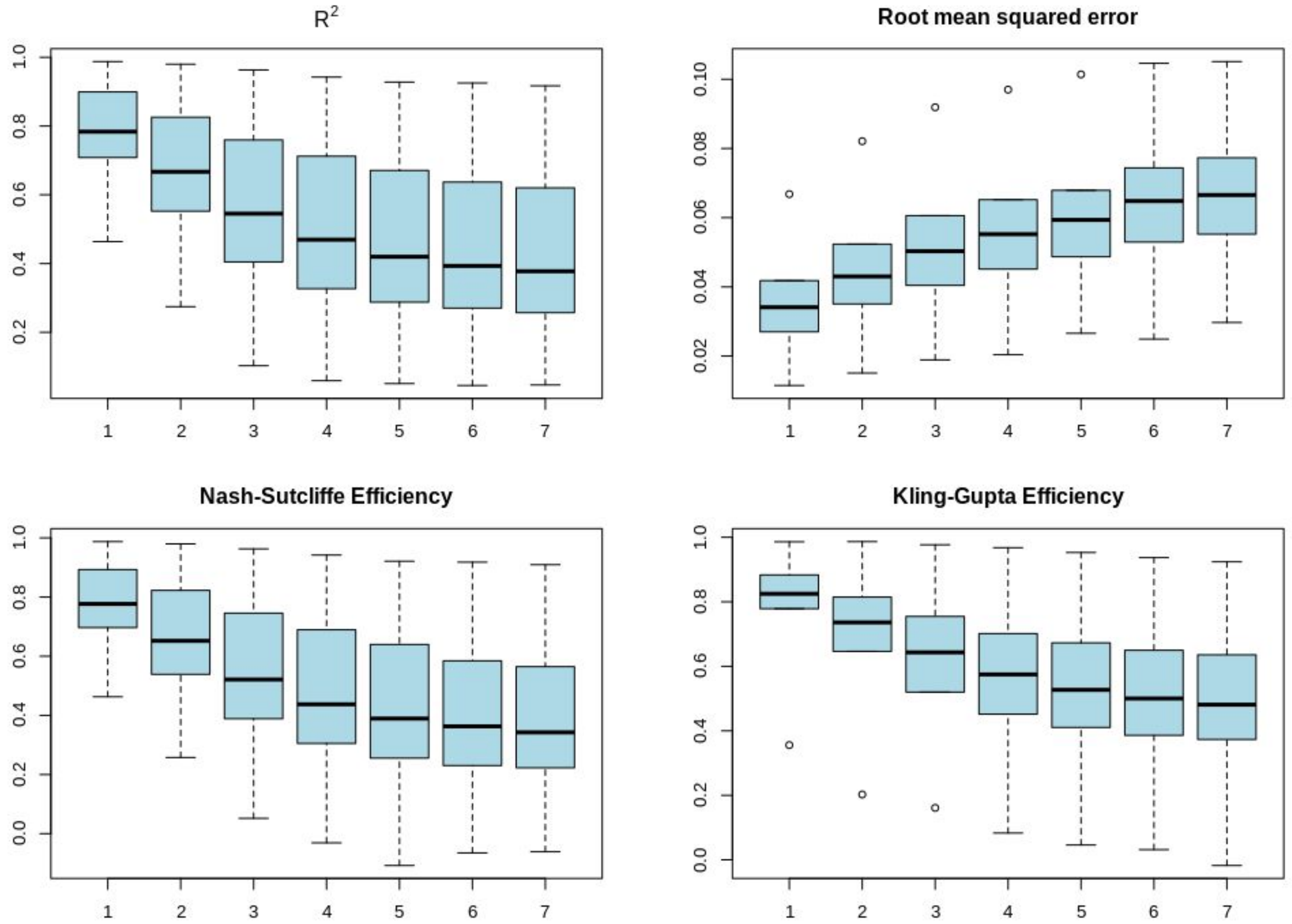
Sample size = bootstrapped sample of n-observations

n-trees = 300



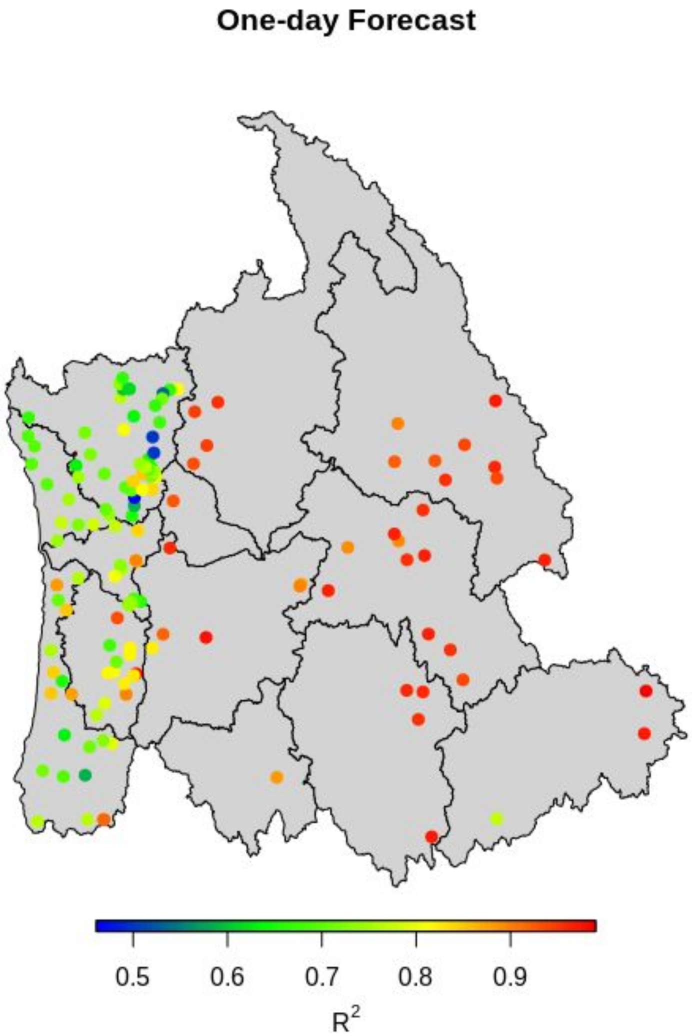
Diagnostic results

Overall Performance



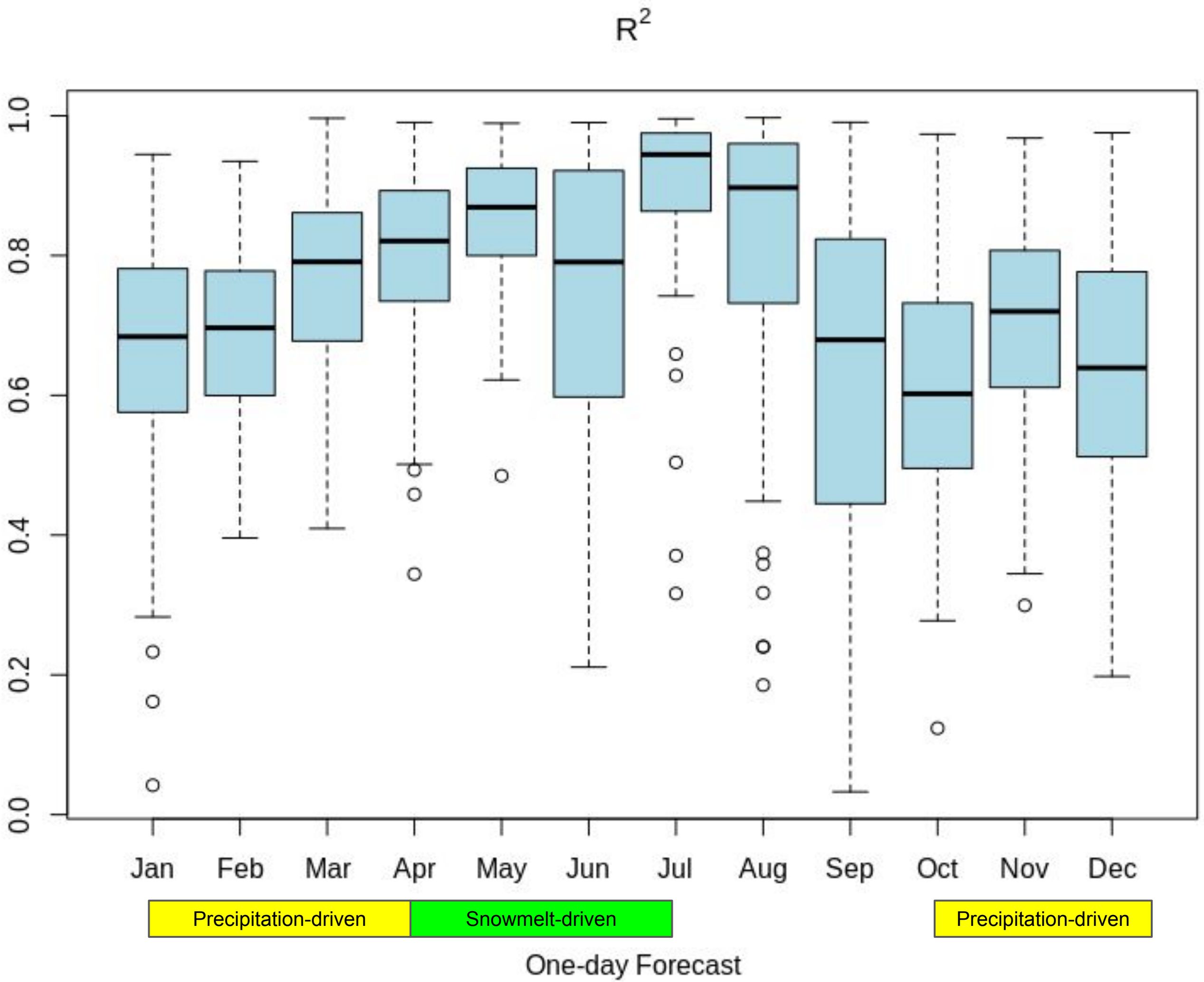
Diagnostic results

Spatial variability in performance



Diagnostic results

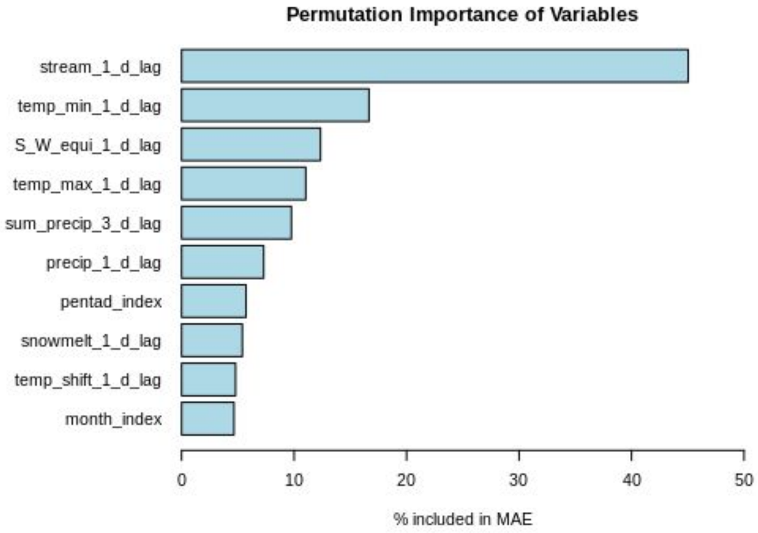
Seasonal variability in performance

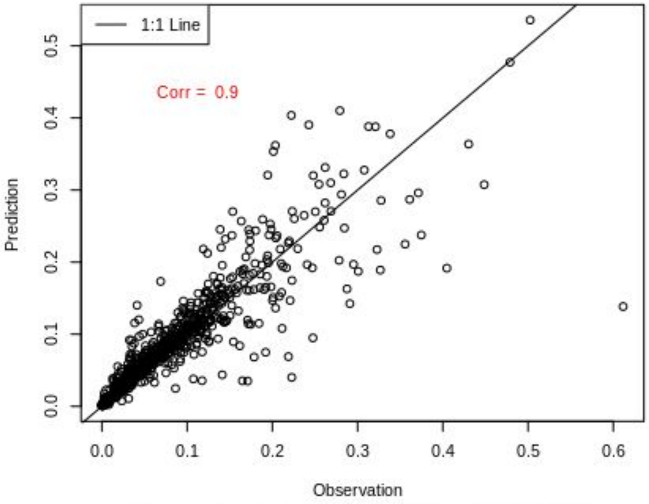


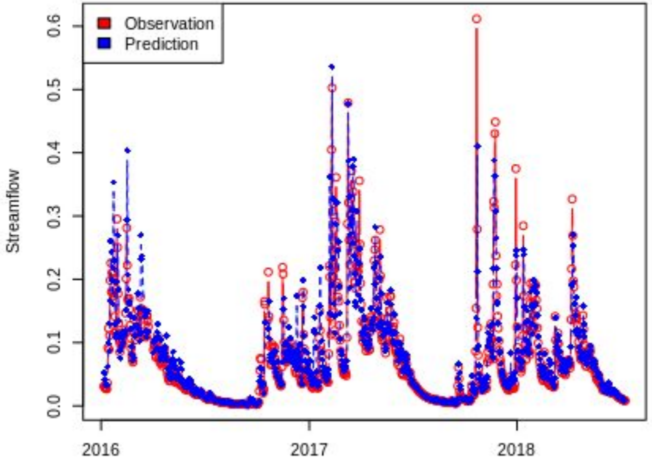
Diagnostic results

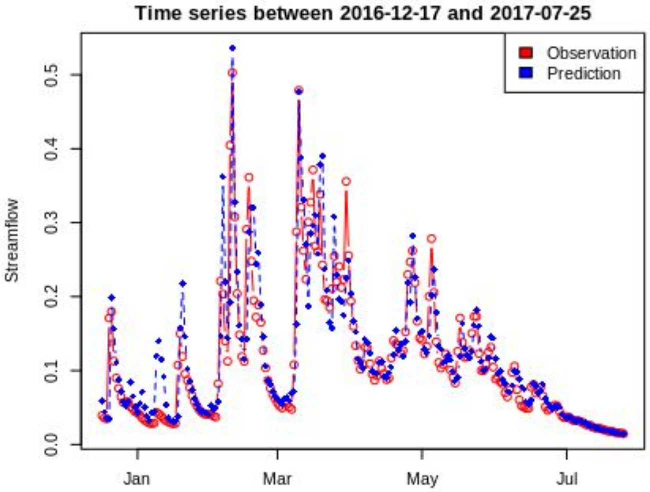
Selected station analysis at USGS Gage 14179000

One-day forecast





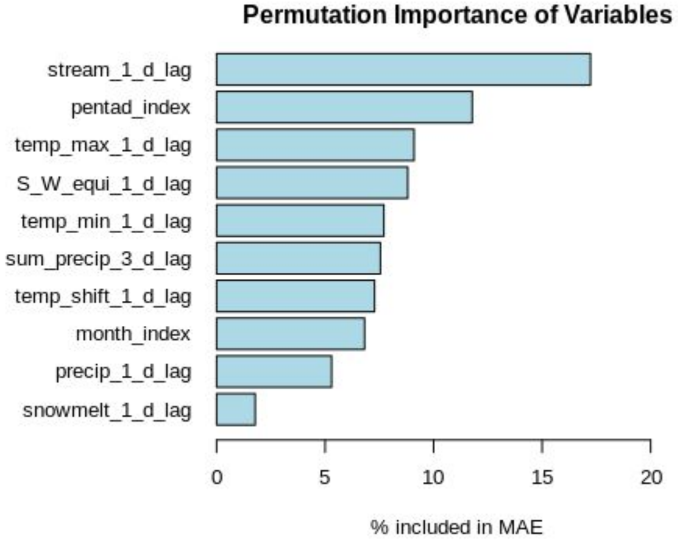


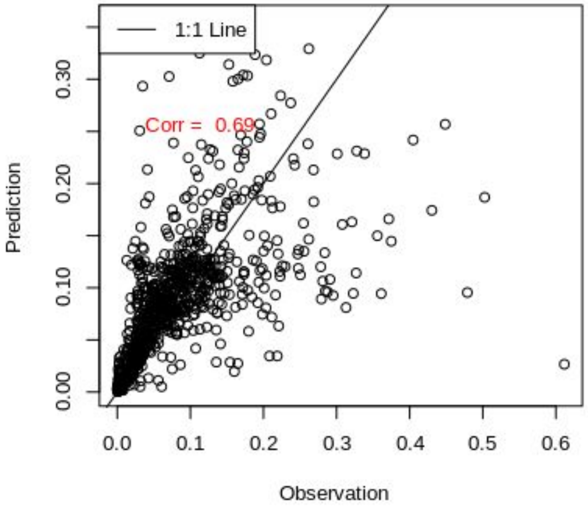


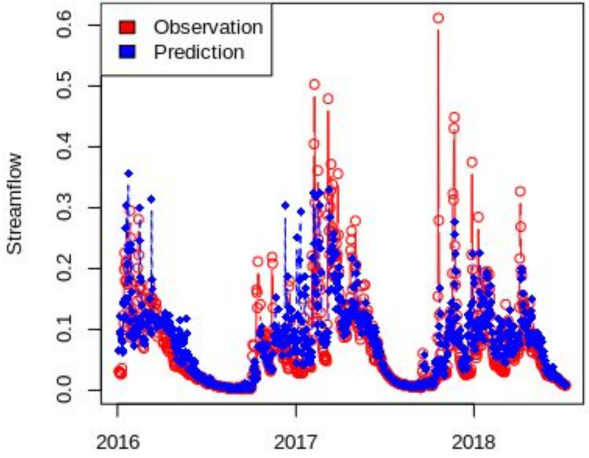
Diagnostic results

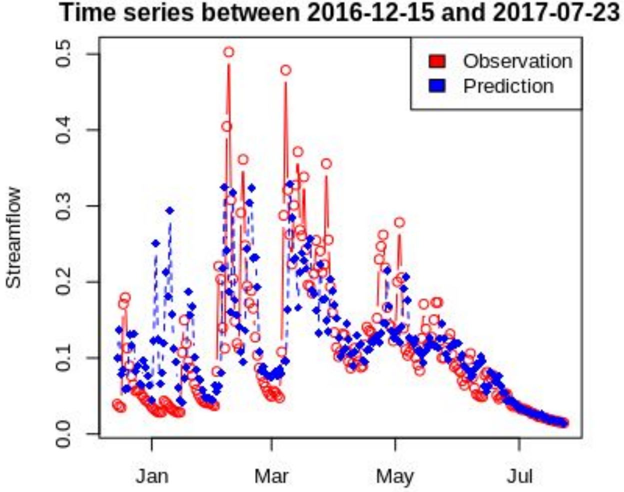
Selected station analysis
at USGS Gage
14179000

Three-day forecast


Permutation Importance of Variables






Time series between 2016-12-15 and 2017-07-23

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

References

Criminisi A., Shotton J., and Konukoglu E. Decision forests for classification, regression,density estimation, manifold learning and semi-supervised learning. Technical ReportMSR-TR-2011-114, Microsoft Research, Ltd., Cambridge, UK, 2011.

Knowles, N., Dettinger, M. D., & Cayan, D. (2001). Trends in Snowfall Versus Rainfall for the Western United States, 1949-2001. Energy, 19(April 2007), 4545–4559.

Liaw, A., Wiener, M., 2002. Classification and regression by random forest. R News 2, 18–22

A. Kalra, S. Ahmad, and A. Nayak, “Increasing streamflow forecast lead time for snowmelt-driven catchment based on large-scale climate patterns,” Advances in Water Resources, vol.

GAGES-II: Geospatial Attributes of Gages for Evaluating Streamflow

Neiman, P. J., Schick, L. J., Ralph, F. M., Hughes, M., & Wick, G. A. (2011). Flooding in Western Washington: The Connection to Atmospheric Rivers*. Journal of Hydrometeorology, 12(6), 1337–1358. https://doi.org/10.1175/2011JHM1358.1

Thank you