San Lorenzo River Flood Prediction

Santa Cruz, California

Economic Impact of Flooding on Vulnerable Communities

Floods are becoming more frequent and more catastrophic with changing climate. Can machine learning be a useful tool for predicting flood peaks?

Data Sources

USGS Streamflow (gages 11161000 in Santa Cruz; 11160500 at Big Trees)

Features: date, flow (cfs), stage (feet)

CDEC precipitation (at Ben Lomond, station BLN)

Features: date, event-based rain tip event

Landsat-derived Normalized Difference Water Index (NDWI)

Downloaded from climateengine.org

NOAA and NIDIS drought index for Santa Cruz County

Calculated Flood Stage Category

San Lorenzo River at Big Trees

CATEGORY	STAGE
Major Flooding	21.76 ft
Moderate Flooding	19.5 ft
Minor Flooding	16.5 ft
Action	14 ft

San Lorenzo River at Santa Cruz

CATEGORY	STAGE
Major Flooding	25 ft
Moderate Flooding	23.33 ft
Minor Flooding	20.55 ft
Action	18 ft

Data Cleaning

Missing Data

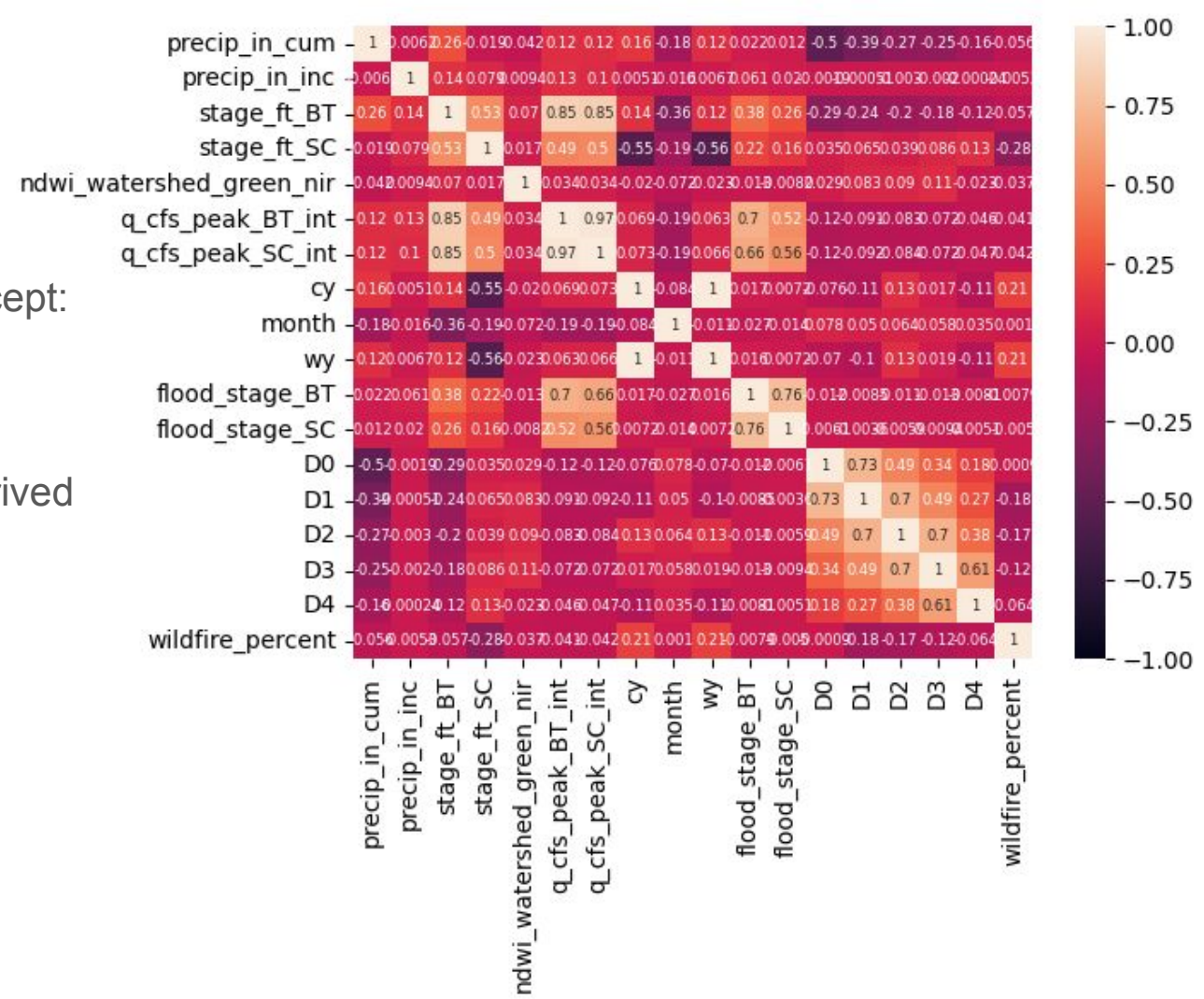
- Streamflow and stage linearly interpolated
- Event-based precipitation converted to cumulative and incremental precip, assuming no missing

No outliers or duplicates suggesting data issues

Data Summary

Correlation generally low, except:

- Drought and precip
- Drought and stage
- Flow and stage (one derived from the other)

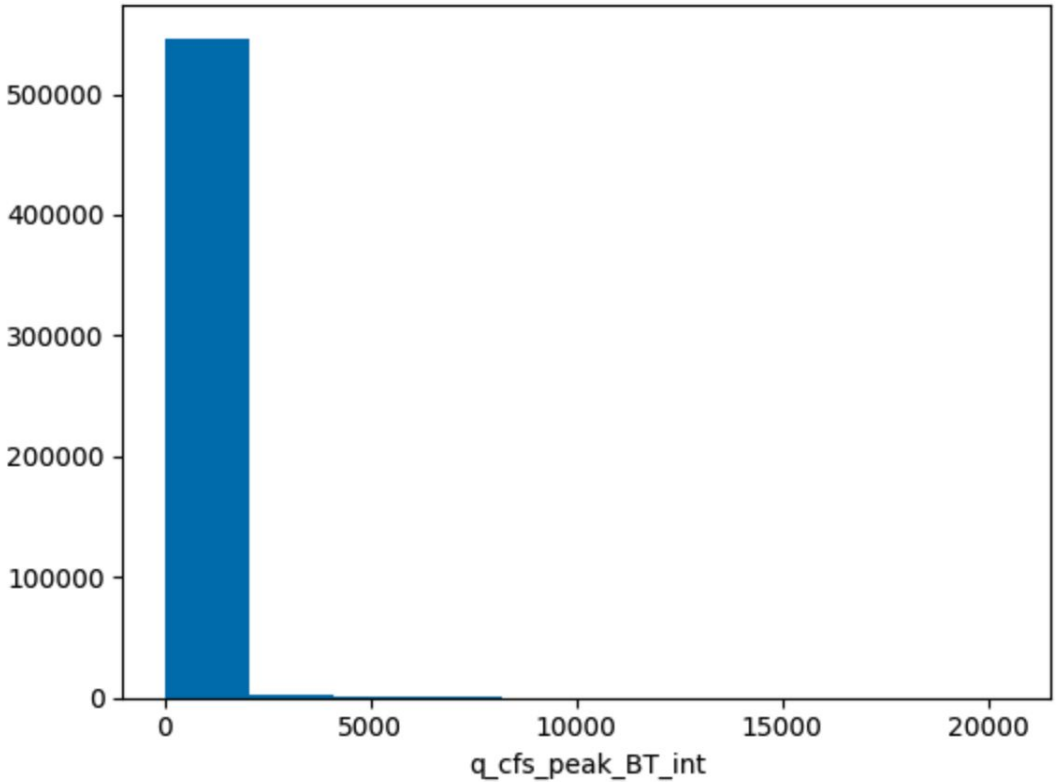


Data Summary

Data highly unbalanced

Most flows very low and relatively few peak flood events - which are the interesting datapoints

For each model, test-train split (0.25/0.75) was stratified by flood stage category



Feature Engineering

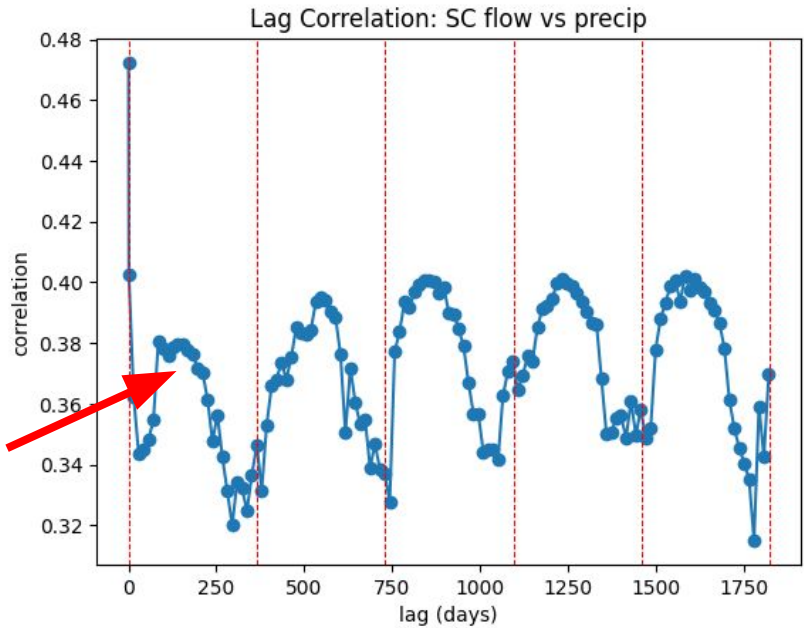
Stream runoff (ie how much water is in the creek) depends on what came before, including rain, drought periods, and soil moisture content

Added features which compute data lag

(precip and drought)

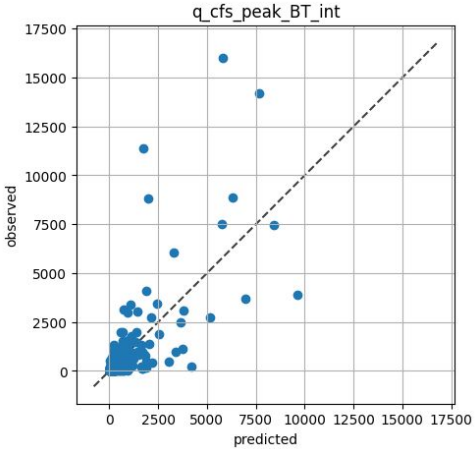
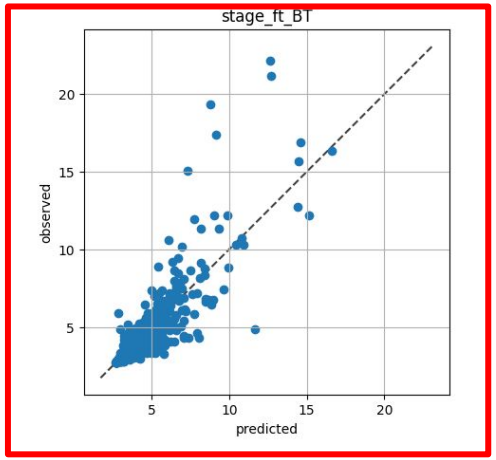
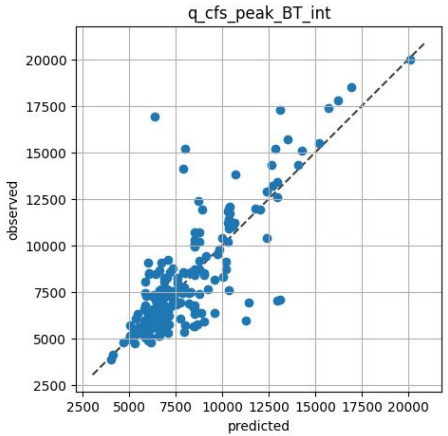
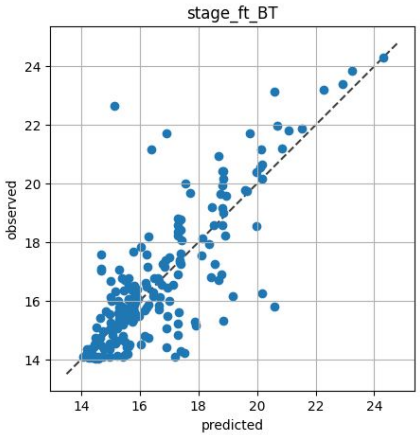
Calculated lag correlation up to 6 months

(where correlation is high)



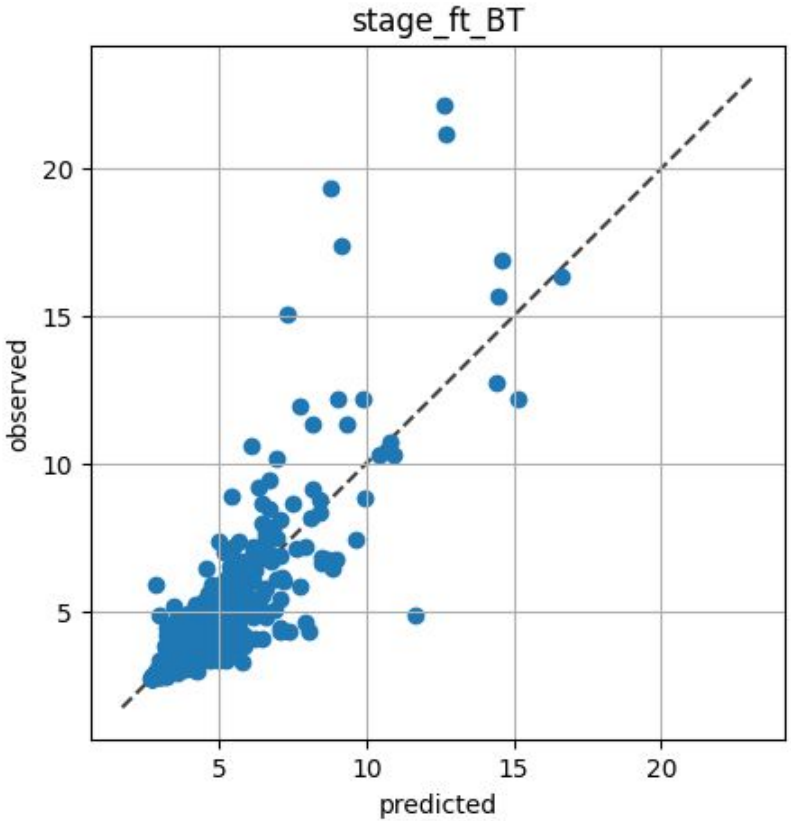
Model: Random Forest Regression @ Big Trees

(Best model)

	Flow (cfs)	Stage (ft)
All flows		
Floods only		

Model: Random Forest Regression @ Big Trees

Best model: Stage for all flows

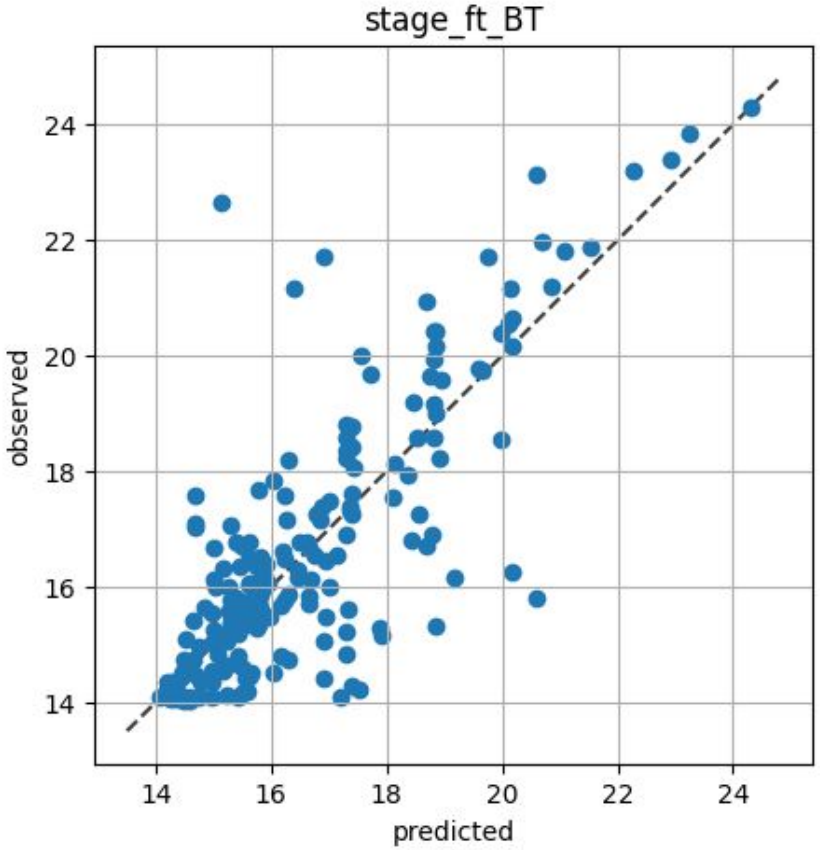


Best model score, but missing peaks

Train score: 0.95
Test score: 0.77

Model: Random Forest Regression @ Big Trees

Stage for floods only

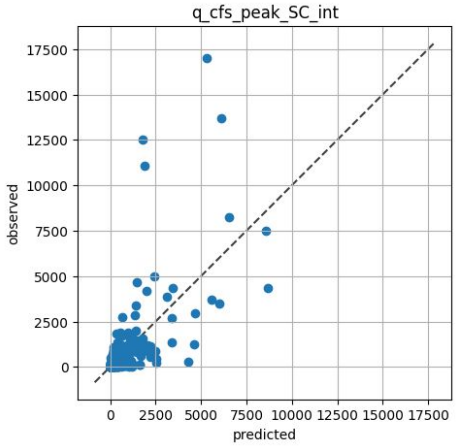
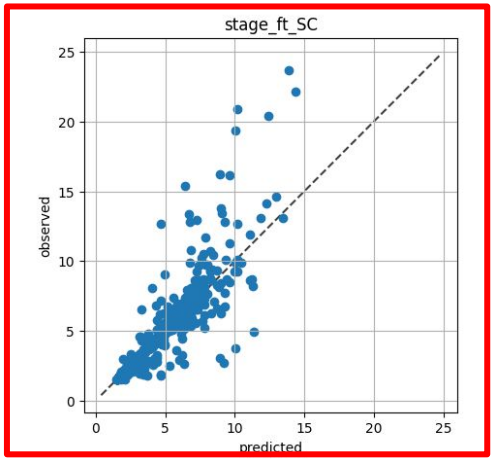
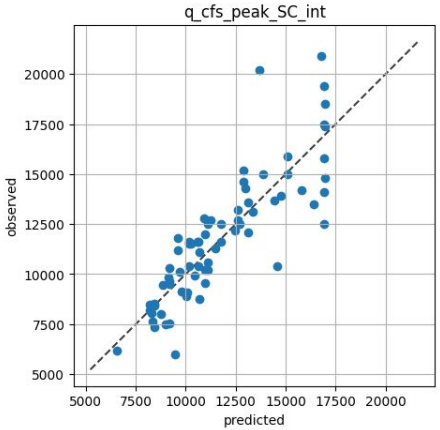
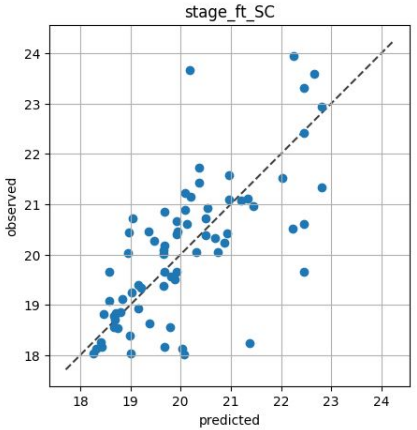


Better peak performance, misses in lower stage

Train score: 0.90
Test score: 0.64

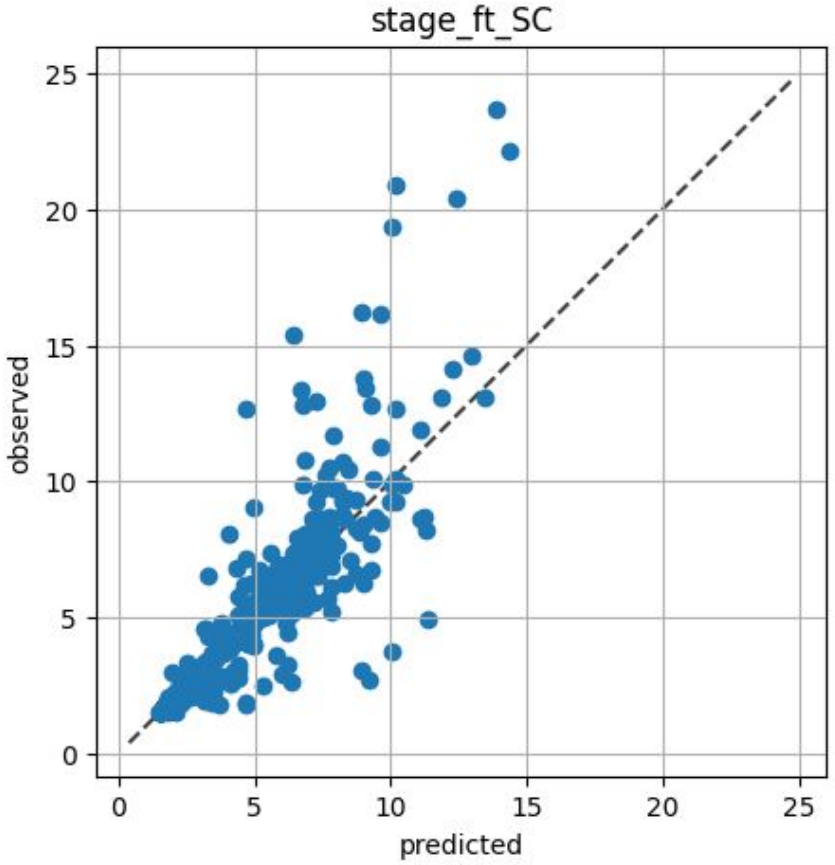
Model: Random Forest Regression @ Santa Cruz

(Best model)

	Flow (cfs)	Stage (ft)
All flows		
Floods only		

Model: Random Forest Regression @ Santa Cruz

Best model overall: Stage for all flows



Best model score of all RFR models, but clear peak flow misses - don't consider score alone

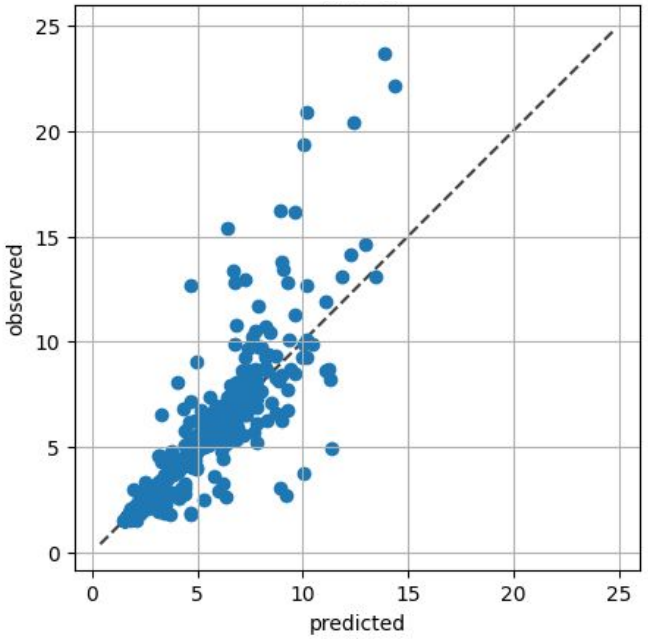
Train score: 0.98
Test score: 0.83

Model:
Neural Net better performance?

Santa Cruz stage, all flows models

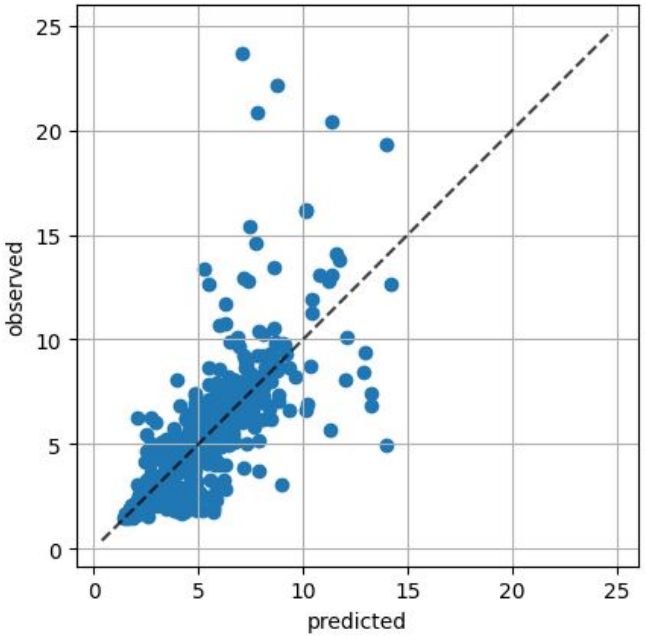
(No)

RFR



MSE: 1.01

Neural Net



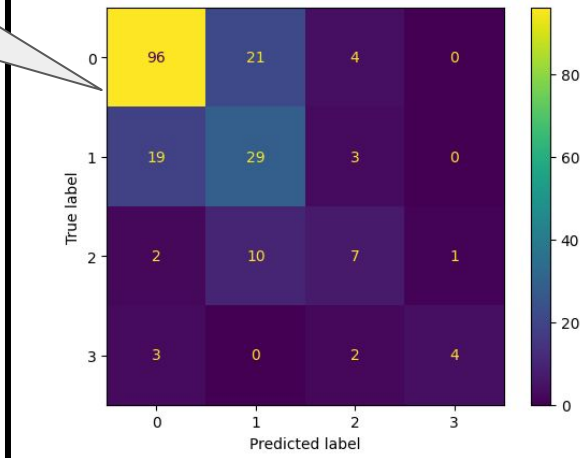
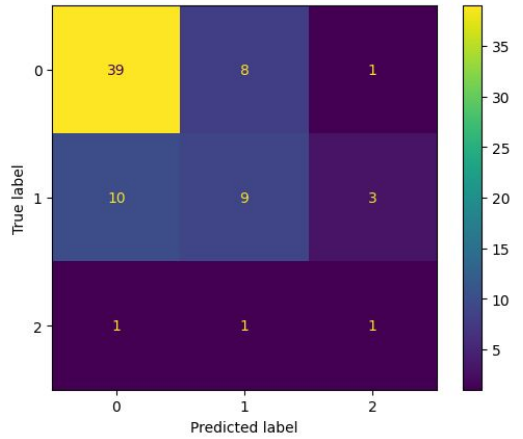
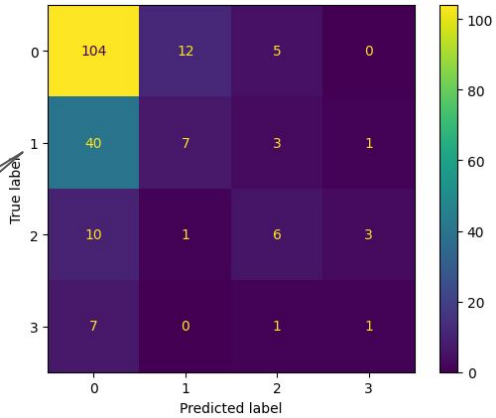
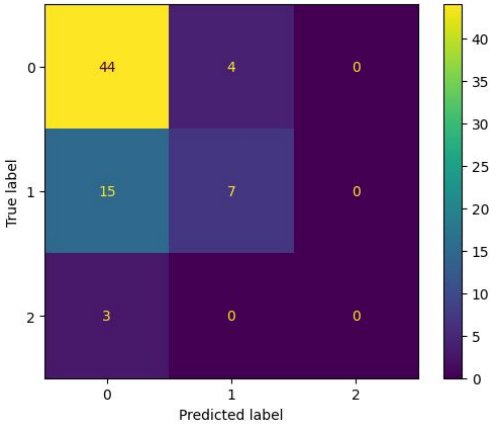
MSE: 1.66

Model: Classifiction (Floods only)

Random Forest Classification (RFC)
vs
Logistic Regression (LR)

More symmetrical (under- and over-predict)

Mostly under-predict

	Big Trees	Santa Cruz
RFC		
LR		

Summary - Feature Importance

What can model results tell us about hydrologic process?

- Precip (cumulative and incremental important)
- NDWI (proxy for soil moisture) important particularly for flood only models
- Small, but measurable impact of wildfires in 2017 and 2020
- Shorter term lags in precip more important but some significance to all precip lag - complete precip record important for model learning