Capstone 3 Project Proposal

By Kealie Pretzlav

Problem Statement

Floods occur every year all over the United States, impacting human life, local economics, and infrastructure. Climate change projections indicate that floods will likely increase in frequency and magnitude in the coming years. Management of flood infrastructure, early warning systems, and dam spillway operations could all greatly benefit from flood likelihood predictions. Long-term flood probabilities over the next several decades can help prioritize flood infrastructure investment and aid in the design of robust mitigation strategies. I will combine historical weather predictions, and precipitation records to predict flood in the San Lorenzo River and to answer the following questions:

Can machine learning be used to accurately predict flooding in the near future on the San Lorenzo River? Can it predict whether the river will flood or not (ie categorical model for flood stage)? Or, can I accurately predict the numerical flow rate? Which gage locations produce a better model and what does that say about the hydrologic processes in those locations?

To answer these questions, I will build a model predicting streamflow on the San Lorenzo River with historical data. Once validated, the model could then be used to simulate a range of what if scenarios about climate change, short-term weather forecasts, and basin properties.

Datasets

To complete this project, I will use three primary data sources, USGS flow records of the San Lorenzo River, historical rainfall data, and long-term projections of precipitation.

San Lorenzo River Flow

The USGS has an extensive network of historical stream gages all over the country. Data collection techniques are highly standardized and generally produce very high quality data. USGS Flow records are the national standard for collecting, calibrating, and cleaning historical flow data. The San Lorenzo River has one of the longest records in the state of California, beginning in 1936, which will provide an excellent long-term historical record for characterizing climate change.

15-minute data (1987 to present) and daily data (1936 to present) at two locations:
In Santa Cruz:
https://waterdata.usgs.gov/monitoring-location/11161000/#parameterCode=00065&period=P7D&showMedian=false

At Big Trees:
https://waterdata.usgs.gov/monitoring-location/11160500/#parameterCode=00065&period=P7D&showMedian=false

Historical Precipitation Record

There are two precipitation gages with long-term records in the San Lorenzo River watershed:
Ben Lomond:
https://cdec.water.ca.gov/dynamicapp/staMeta?station_id=BLN

Landsat NDWI

Normalized Difference Water Index (NDWI) is a measure of the water or moisture content from satellite imagery. Data was downloaded from climateengine.org, as median NDWI (Green/NIR) over the whole watershed. Watershed boundary from: https://purl.stanford.edu/cp847hc0870

Santa Cruz County Drought Index

from https://www.drought.gov/historical-information?dataset=0&selectedDateUSDM=20240618 for Santa Cruz County

Watershed Burn Percent

There were wildfires in the San Lorenzo watershed in water years 2017 and 2020. I found estimates of percent of watershed burned in each fire and added that to the data. The same burn percentage was added for each row in the relevant water year as the fires occurred prior to the wet period.

Data Preprocessing

Each dataset was downloaded, cleaned and resampled to 15 minute timestep as necessary, and merged on timestamp.

Missing data:

- Missing streamflow, stage, and flood stage category was linearly interpolated, max consecutive duration of missing streamflow was 4.8 days for Santa Cruz station and 1 day for Big Trees
- In the conversion from rain gage tipping events to timeseries precipitation record, missing precipitation was either set to zero for incremental rainfall or filled forward for cumulative rainfall.

There were no outliers or duplicates indicating a data problem.

Data Exploration

The more correlated features were:

- Extreme drought conditions (D0) and cumulative precipitation
- Extreme drought conditions (D0) and stage
- Wildfire percent and stage
- Month and stage
- Year and stage

Data Feature Engineering

Because watershed hydrologic response depends very heavily on what came before, I calculated several time-lagged features of drought index and precipitation. To see how far back to lag the data, I calculated the correlation coefficients of flow with drought index and precipitation. The calculations indicated that I should lag the data by 4 months. Correlation values were low for NDWI. I then computed lag in precipitation and drought index features for various timesteps from 1 day to 4 months.

Modeling Methods

Each model uses a train/test split of 0.75/0.25. Because the data is unbalanced, with relatively few large streamflow or stage values, I stratified the test and train datasets following the flood stage categories. Each model selection exercise uses a grid search cross-validation to find the best-fit hyperparameters.

I modeled flood stage, streamflow, and flood stage category for daily data; 15-minute data had performance issues on my personal computer and didn’t provide much additional prediction quality. I also ran models which limited the data to flood stage conditions (flood stage category above 1) to create a more balanced dataset. Because river stage is at baseflow or below flood stage >90% of the time, classification models were predicting deceptively “good” scores when only considering lower flows.

Modeling Results - Random Forest Regression

Unsurprisingly, streamflow prediction for all flows, including summer baseflow overfits the training data and generally misses the peak flows.

Target	Train Score	Test Score	Best Params	Important Features (top 3)	Comments
BT Flow (all)	0.90	0.56	Depth: 8 Estimators: 50	Inc precip, Cum Precip Precip lag 1 day	Under-predicts peaks
BT Flow (floods only)	0.89	0.65	Depth: 10 Estimators: 250	Cum precip, NDWI, Precip lag 1 days	Under-predicts peaks, but better performance than using all flow
SC Flow (all)	0.85	0.49	Depth: 6 Estimators: 50	Inc Precip, Cum precip, Precip lag 1 day,	Under-predicts peaks
SC Flow (floods only)	0.85	0.73	Depth: 6 Estimators: 1000	Cum precip, NDWI, Precip lag 5 days	Under-predicts peaks, but better performance than using all flow
BT stage (all)	0.95	0.77	Depth: 40 Estimators: 300	Inc precip, Precip lag 1 day, Cum precip	Some under-predicting for flow peaks, but surprisingly good performance overall
BT stage (floods only)	0.90	0.64	Depth: 12 Estimators: 50	Cum precip, NDWI, Precip lag 1 days	Surprisingly good model performance, largest peak predicted almost spot on
SC stage (all)	0.98	0.83	Depth: 35 Estimators: 100	Cum precip, Precip lag 1 day, Inc precip	Generally pretty good model performance, slightly under-predicting peak stage
SC stage (floods only)	0.84	0.47	Depth: 12 Estimators: 50	NDWI, D0 Precip lag 5 days	Under- and over-predicting

Generally, the regression models perform poorly for extreme values, suggesting more data is needed, or that something fundamental in the underlying processes for when a river floods is missing. Hydrologically-speaking, flow in a river accumulates over the rainy season. Many watersheds won’t have significant runoff until a certain precipitation threshold is met. As a result, flow/stage response to a precipitation sequence early in the season will be very different than late-season precipitation. NDWI was an attempt to characterize the accumulated moisture in the watershed (i.e. how “full” is the watershed), but it is undoubtedly an imperfect measure, particularly as it is remotely sensed via satellite data and therefore is obscured by cloud cover when rain occurs. Another proxy for watershed moisture may product better results.

Interestingly, the flow models for the Santa Cruz station generally have poorer performance than the Big Trees station. This is likely because the contributing watershed at the Big Trees location is closer to a natural, unmodified watershed than the Santa Cruz station watershed, where water is diverted for drinking water, agriculture, and other industrial uses.

The overall best-performing model is the Santa Cruz stage model using all stage values. Although the model overfits the training dataset some, it is not as extreme as the other models, and the test dataset has the highest score of all the RFR models. Figure 1 and 2 show feature importance and predicted vs observed values.

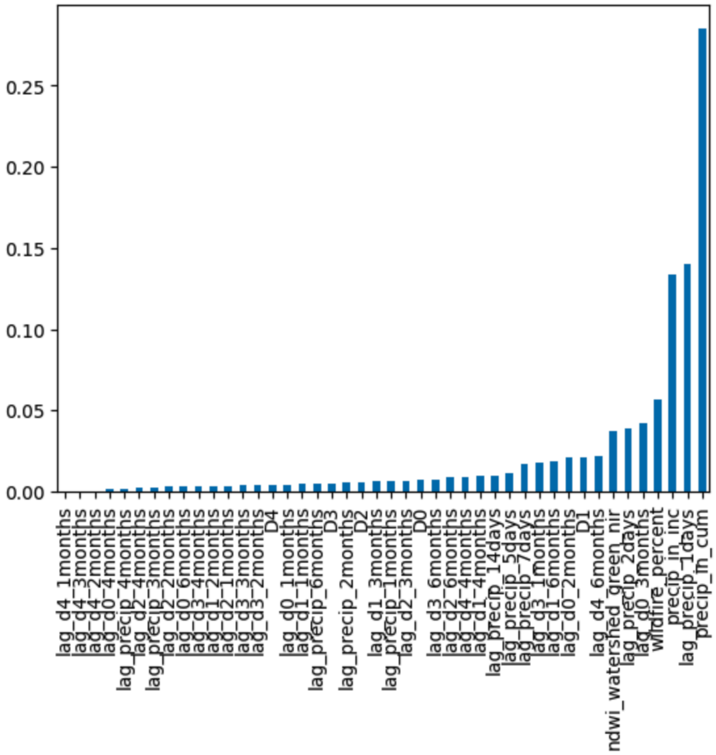

Figure 1. Feature importance for the Santa Cruz stage RFR model, all flows

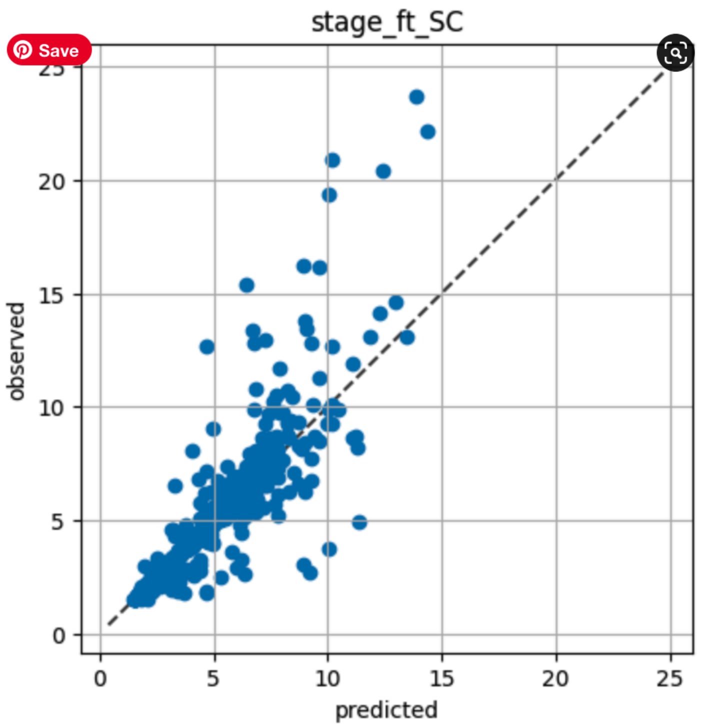

Figure 2. Predicted vs observed for Santa Cruz stage data, all flows

Model Results - Neural Net

I also wanted to see if a Neural Net model produced better results than RFR. I ran the Santa Cruz all flows stage model with the same stratified train/test split data through a Neural Net model with 5 layers, ran for 200 epochs. The resulting prediction still missed the peak stage values, with a mse of 1.66, compared to the equivalent RFR model which has a mse of 1.01.

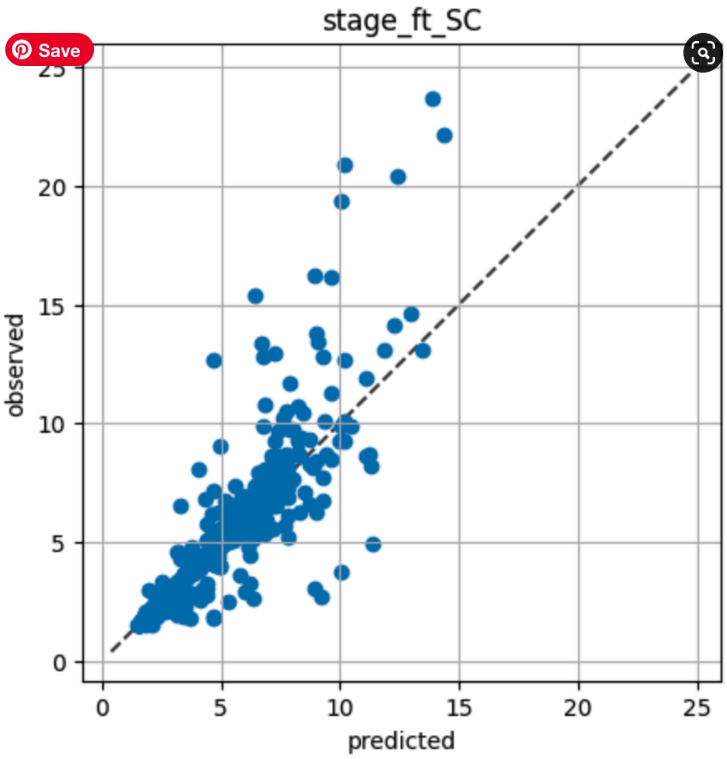

Figure 3. Predicted vs observed Santa Cruz stage data for the neural network model

Modeling Results - Flood Stage Classification

Next, I trained several classification models on the flood stage categories, from 0 (below flood stage) to 4 (major flooding). Classification models of all flood stage (including baseflow) are highly unbalanced, with > 95% of data in the below flood stage class. So all classification models were trained using data above flood stage only (categories 1 - 4).

I trained Random Forest Classification and Logistic Regression models for both gage locations (Santa Cruz and Big Trees). Results are summarized in the table below.

Target (all floods only)	**Model**	**Train Score**	**Test Score**	**Best Params**	**Important Features (top 3)**
BT Flood Stage	RFC	0.93	0.67	Depth: 15, Estimators: 200	Cum Precip, Inc Precip, Precip lag 14 days
SC Flood Stage	RFC	0.89	0.67	Depth: 8 Estimators: 50	Cum precip, D0, Lag precip 1 day

BT Flood Stage	LR	0.67	0.58	C: 2, Max_iter: 1000	N/A
SC Flood Stage	LR	0.71	0.70	C: 0.5, Max_iter: 1000	N/A

Because even the floods only data is unbalanced, it is important to evaluate model performance more heavily on test score rather than train score. Interestingly, the RFC models mis-classify flood stages in both directions (under- and over-predicting) where the LR models tend to mostly over-predict flood stage category.

Interpretation of Feature Importance

For all of the random forest models, relative feature importance is an important model validation and interpretation tool. Most of the models had cumulative or incremental precipitation in the top 3 features. Interestingly, all of the floods only stage and flow models also had NDWI - a good proxy for watershed moisture retention - as one of the most important features. This suggests that NDWI may be a good estimate of how wet a watershed already is in a given time, and how likely a given rainfall amount is to produce a large discharge event. Notably, NDWI is available on a roughly weekly timescale, and missing for periods with heavy cloud cover; a denser data availability would likely improve model performance. Many of the models also had the lag of precipitation from 1 to 5 days, similarly imparting a short-term memory to the model. The data were simply shifted back by the lag time window, but a cumulative sum of precipitation in that time window may also provide the model with extra information and could be used in future model revisions.