Geospatial data applications
GEOG 495/595
Theo Ross & Ian Van Dusen

gage-guesser
https://github.com/ianvd/gage-guesser

A machine learning model comparing SNOTEL, Daymet, and satellite data sets to best estimate discharge in the Mount Rainier Nisqually River basin

Introduction.

Rivers are essential to communities across the planet. For both human and natural needs, fluctuations in the discharge of rivers have consequences. Understanding and predicting river discharge is a fundamental goal of hydrology and globally, the understanding of river discharge patterns vary significantly with infrastructure. In the United States, the USGS provides excellent open-access gage data on all major rivers in the lower 48. Data access and temporal resolution have improved greatly with many gage sites providing sub-hourly measurements.

The upcoming launch of the NASA Surface Water and Ocean Topography (SWOT) satellite provided inspiration for the project, as there is increasing interest in providing discharge estimates for global rivers. We were curious about the level of accuracy to which we could make predictions in a data-rich region. The SWOT satellite may prove to provide unprecedented measurement of surface water discharge. Our findings for multiple data sets could be used to complement and improve the temporal resolution of the SWOT mission.

The area of interest for this project is the Nisqually River basin off of the Nisqually glacier of Mt. Rainier. This region, near Paradise, Washington, is ideal for many reasons - the proximity of the gage to the source of melt water, the accessibility of multiple data sets, and the relatively simple hydrologic system. We chose to monitor and compare in situ, interpolated, and remote data via SNOTEL, Daymet, and satellite datasets respectively. For this project, we set out to create a machine model to accurately predict discharge of rivers based on a number of environmental and climate parameters. We compared multiple data sources to find the most accurate and accessible predictor of discharge for the Nisqually river.

Methods

We focused on predicting discharge values with three different sets of training data. First, we used data derived from a Snowpack Telemetry site (SNOTEL) within the basin. Next, we used interpolated weather data (Daymet) to retrieve data within the basin. Lastly, we collected MODIS and GPM satellite data within the basin. All three datasets provided daily values of climate metrics, however, satellite data proved to have more missing values than the other datasets due to limitations of cloud cover when collecting surface metrics such as temperature and snow cover.

In terms of the amount of training data required to make a reliable model, we decided that having at least five years of daily climate data for each dataset would provide an adequate amount of data and allow the model to predict patterns beyond seasonality. For the SNOTEL and satellite data, we collected daily data from 2013-2021. For Daymet, we collected data from 2013-2018 as the data provider has not updated the repository for the years 2019-present.

Additional data collection for this project included gathering discharge values from the USGS river gage: 12082500 for 2013-2021 using an API called dataretrieval produced by the USGS.

SNOTEL data was collected from the USDA's National Water and Climate Center portal and included the daily climate variables: precipitation accumulation (from start of water year), precipitation increment and snow adjusted precipitation increment, maximum, minimum, and average air temperature, and soil moisture percent at depths of 2", 4", 8", and 20". Additionally, the snow variables: snow depth, snow density, snow water equivalent, and snow rain ratio were used. This data was then joined with daily discharge data to create a Pandas DataFrame of all daily variables. Following this step, the DataFrame was cleaned of no data values as the machine learning approach used requires fully populated training data.

In order to collect Daymet data, we found a web service API produced by the ORNL DAAC that allows batch ordering of Daymet data as well as other NASA data. This process required a bit of modification to the examples provided on their repo page, and we ended up looping through each year for each variable in order to not exceed the order limit for the HTTP GET call. In the end, we downloaded daily climate data of precipitation, snow water equivalent, and maximum and minimum temperature. As noted in our code, after collecting this data we exported it to csv format in order to be able to rerun the jupyter notebook file without having to reinitiate the data retrieval process every time. As done previously, this DataFrame was merged with the discharge values.

Satellite data was collected using Google Earth Engine. We clipped climate data for two MODIS datasets: Daily Snow Cover 500m and Daily Land Surface Temperature and Emissivity 1km to our imported watershed shapefile. To collect daily precipitation data, we used the Global Precipitation Measurement (GPM) constellation of satellites to download sub hourly precipitation values for our study area. This data was then resampled to daily values within our jupyter notebook file. The full code for clipping and exporting this data from GEE can be found

in the README file on the github repository. The satellite data was then joined to the discharge data to create a DataFrame that could be used as training and testing data.

To create the machine learning model for each dataset, we used a Random Forest model from the Scikit-Learn Python library. With more time to complete the project, we would have liked to experiment with other models as well. Within the model we split the data into 80% training and 20% testing data and used a total number of 100 decision trees.

Results

For each machine learning model, we were able to test the mean squared error and compare them in the table below (Table 1). Looking at the results, the SNOTEL data provided substantially better results than the Daymet or Satellite data. (Table 1). Given that discharge ranges 9746 cfs in the SNOTEL dataset, 9453 cfs in the Daymet dataset, and only 4936 cfs in the satellite dataset, it makes the low mean squared error for the SNOTEL model even more impressive. The variation of this range is due to the exclusion of dates with no data values for the datasets, and the satellite dataset had the highest number of no data values. Nonetheless, this means that the error of the SNOTEL model is only 1.08%, while the error for Daymet is 3.82%, and of the satellite data is 6.55%. Figures 1-3 below help to visualize the predicted discharge values in comparison to the observed discharge values, and include a separate variable showing the difference between these two values.

	Mean Squared Error (cfs)
SNOTEL	105.09
Daymet	361.46
Satellite	323.50

Table 1 - Mean squared error for each machine learning model

Extreme values Removed	**Mean Squared Error (cfs)**
SNOTEL	7.60
Daymet	201.18
Satellite	240.92

Table 2 - Mean squared error after filtering for discharge outliers

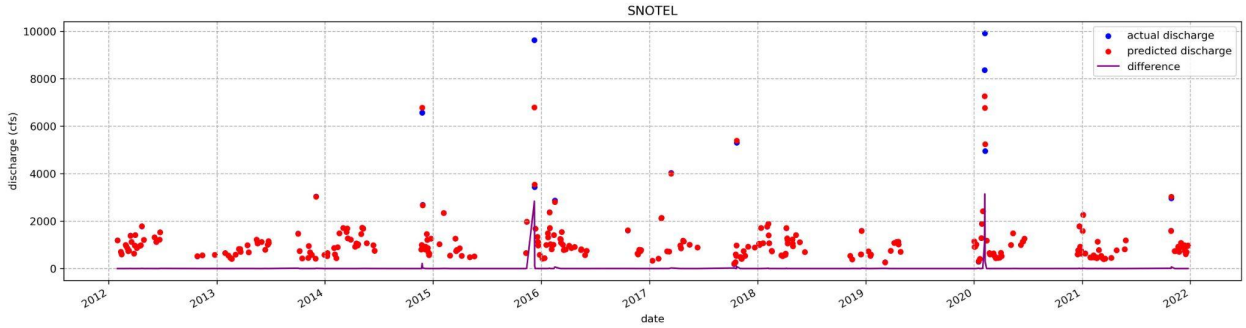

Figure 1: SNOTEL model predicted vs. observed discharge values

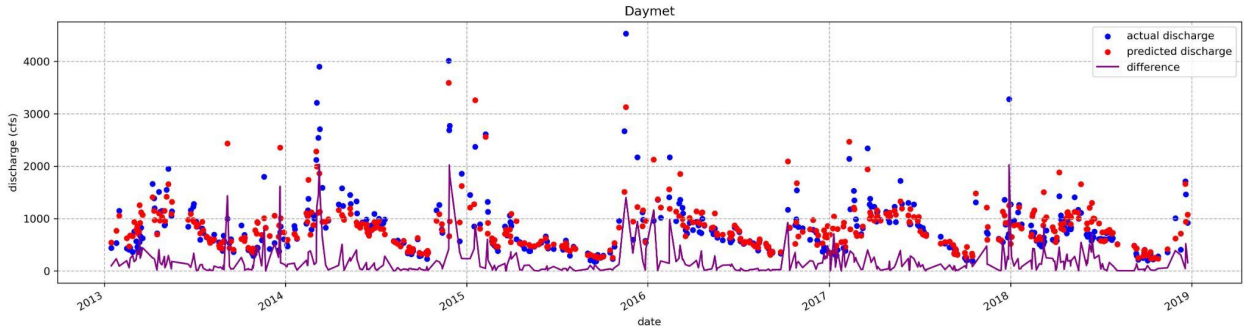

Figure 2: Daymet model predicted vs. observed discharge values

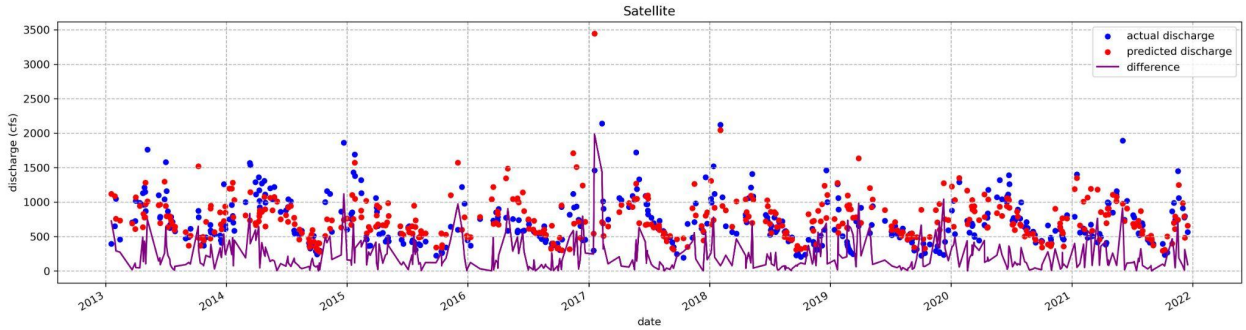

Figure 3: Satellite model predicted vs. observed discharge values

After running the models and looking at the error values, we were interested in how much extreme weather events were impacting our model and if these were the hardest to predict. Looking at the discharge figure (Figure 4), multiple spikes are clearly visible, that vary greatly from discharge averages. In order to understand how much these values were influencing our models' accuracy, we decided to create new models for each dataset, excluding dates with discharge values higher than 95% of the average and less than 5% of the average. The resulting table (Table 2) displays the error of each dataset after performing this filtering on the training and testing data. This clearly illustrates that the most difficult discharge values to predict are these types of weather anomalies.

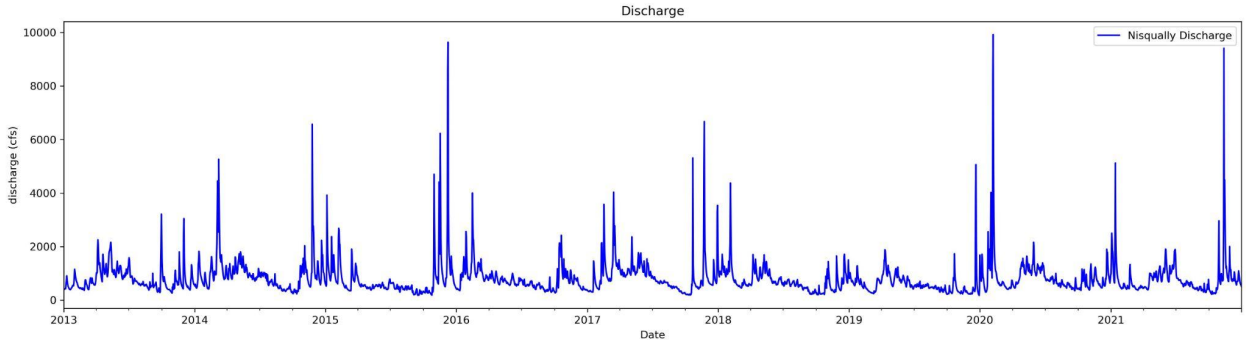

Figure 4: Daily discharge observed at the Nisqually river gage from 2013-2021

Discussion.

In the 'gage-guesser' project we prioritized creating a concise completed project, even though there were some scientific questions that we failed to answer in the scope of this project. If we expanded this from a class project, and towards publication, we could answer additional questions we failed to address.

The first difficulty of this project was that we had a circular plan in nature. We created a machine learning model to estimate the discharge of a river. However, we used the discharge from gage data to train our model. In this case, the estimation of discharge may be less useful, because the gage data is accessible year round at sub-hourly frequency. However, this could be useful for predicting flooding or extreme weather events, although our model is weakest for extreme events. The creation of such a model would be most useful to rivers and basins that do not already have accurate gage readings. Our model was designed off of the gage so its accuracy of prediction is dependent upon a quality data source. If we had more time, we would like to compare the accuracy of the model on significantly less gage data. For example, if we had discharge data for the first of every month, would these findings still prove to be useful? It is unrealistic to have 5 years of discharge data on a river without a gage, but is it possible that our model could prove useful for a monthly data set? This project, at least furthers the question, to finding how much data is enough to be 'close enough'.

One of the major takeaways was the accuracy of the SNOTEL dataset, especially after removing the extreme values. We created a model that can predict the discharge of the river based predominantly on soil moisture, snow coverage, and precipital data. The SNOTEL data set proved to be much more accurate than the satellite and Daymet sources. Does this mean that to understand specific watershed trends, physical insitu models may be a much more cost effective plan than satellite missions? It depends on the scope of the project, but this study demonstrated that SNOTEL datasets can prove to be extremely useful for estimating the discharge of a small hydrologic system. For future work, we would like to combine the data of multiple SNOTEL sites to predict discharge in a larger system, and to monitor the accuracy of our predictions. This project is certainly an excellent start to understanding the power of SNOTEL data, which is abundant in the United States, and the Pacific Northwest in particular.

Finally, the Machine Learning model used was chosen out of comfortability and ease of access. The model trained itself off a variety of variables and our data had significant annual trends. Could it have been more accurate if we spent more time training the seasonality of the data set? We attempted to add in the Month as a variable, but there are surely additional ways we could have assisted the model to more accurate predictions. Additionally, we would like to test out additional Machine learning models outside of Random Forest to compare effectiveness.

This project was our first exposure into the world of hydrologic modeling and we have learned significant steps into what is most valuable when predicting future discharge events. I look forward to further study, and appreciate the accessibility of water and environmental data in the United States. GitHub proved to be a useful tool for collaboration, as we could push and pull Jupyter Notebooks, datasets, and files with ease.