# Predicting the Amount of Water Resources in Lake Tahoe

Ayaka Abe

## ABSTRACT

Predicting the amount of water in a reservoir is important for proper water resources managements. However, predicting the flow of water requires local geological understandings and numerical simulation as a hydrological system is often complex. In this project, I used regression models to predict the Lake Tahoe's water level from precipitation and snow fall data to see how much closer I can model the hydraulic system around the Lake Tahoe. The results show that the Kernel ridge with linear kernel shows the best fit to the observed lake water level based on the inputs of precipitation, snow fall, and water discharge data. For future development, additional data set such as temperature, humidity, sun radiation, and precipitation at other stations around the Lake Tahoe, may improve the prediction.

## 1. INTRODUCTION

Water resources management is essential to optimize the water supply, hydropower generation, and flood management (Bonakdari et al. 2019). Especially, forecasting lake water levels is important as lakes play an important role as a freshwater reservoir.

California draught December 2011 to March 2017 was the driest since record keeping began in 1895 (Los Angels Times 2017, PPIC 2016). The combination of high temperatures and low winter precipitation was the main cause of this severe water deficit (PPIC 2016). Most of California's water supply originates in northern California, where the biggest reservoir in northern California, the Lake Tahoe, is located. The lake contains an average of 37 trillion gallons of water (U.S. Forest Service 2019). Therefore, to predict Lake Tahoe's water level helps California's water management.

Water is charged to the ground as rain fall or is stored as snow pack, then the water is stored in underground as underground water, then it flows out to a river or to a lake, then the water evaporate to the air and form a cloud, then fall to the ground as rain. The precipitation and the snow melt water are the inputs to the geological systems, the outputs are the lake water level or the river water flow rate. Those hydrological modeling is mostly done by numerical simulations, however, there are also research that try to model the hydraulic system with ML technique. In this project, I use ML regression models to predict the Lake Tahoe's water level from precipitation and snow fall data to see how much closer I can model the hydraulic system around the Lake Tahoe.

## 2. RELATED WORK

One of the early research utilizing a machine learning technique is the one done by Watts (1995). The author built regression models to predict the monthly water level changes in the Closed Basin Division, in the San Luis Valley, Colorado, with the features including elapsed time, cosine and sine functions with an annual period, streamflow depletion, electrical use for agricultural purposes, runoff into the closed basin, precipitation, and air temperature.

One of the difficulties of modeling water level changes is that there is a time delay between input (rain/snow fall) and output (water level response) because rain fall water is stored as snow pack and/or underground water before it flows out to a lake. Also, there are seasonal effects of climate on the models as well. This research handles annual fluctuation of the water level by including a cosine and sine functions with an annual period and remove the seasonal effect by taking account to the response delay.

There are some other examples utilizing the multiple-linear regression. Kühn and Schöne (2017) applied the linear regression to predict the water level response induced by water productions by a geothermal power plant. Bonakdari et al. (2019) investigated the forecast of the lake water level data of the Great Lakes Basins with four different methods (MPMR, GPR, RVM, ELM), different time delays of historical lake level with varied features and evaluated the models. Nowadays, other approaches using neural network, SVR, and other regression models are getting more common.

## 3. DATASET

The water level of Lake Tahoe (USGS 2019a) is the daily observations at midnight measured at a station. The station is operated by the California Department of Water Resources. I retrieved the data of the first day of a month and reduced them to a monthly data. Precipitation and snow fall data are obtained from the Western Regional Climate Center (2019). Those data are the average value over a month therefore the datasets are monthly data. Water discharge data is obtained at the Truckee River (USGS 2019b) as the Truckee River is the main outlet from the Lake Tahoe and flowing into Pyramid Lake in the Great Basin (Water Education Foundation 2019).

The data used in this project is the precipitation and the water discharge data from January 1910 and the water level from January 1911. The data points are in total 1308 (109 years times 12 months). Since I have small number of data, I applied k-fold cross validation to fully utilize the dataset. The data set was divided into train dataset and test data set. The number of training set is 1303, which corresponds to the monthly data from Jan 1911 to Dec 2010 (100 years). The number of the test set is 119 which is monthly data from Jan 2011 to Nov 2019 (9 years and 11 months).

**Table 1: Dataset used in this project**

| Dataset | Units | Data range |
|---|---|---|
| Water Level of Lake Tahoe | in | Monthly data from Dec 1900 to Sep 1957<br>Daily data from Oct 1957 to Nov 2019 |
| Precipitation | in/month | Monthly total data from Jan 1910 to Nov 2019 |
| Snow Fall | in/month | Monthly total data from Oct 1909 to Nov 2019 |
| Water Discharge | ft$^3$/second | Monthly mean data from Jun 1909 to Nov 2019 |

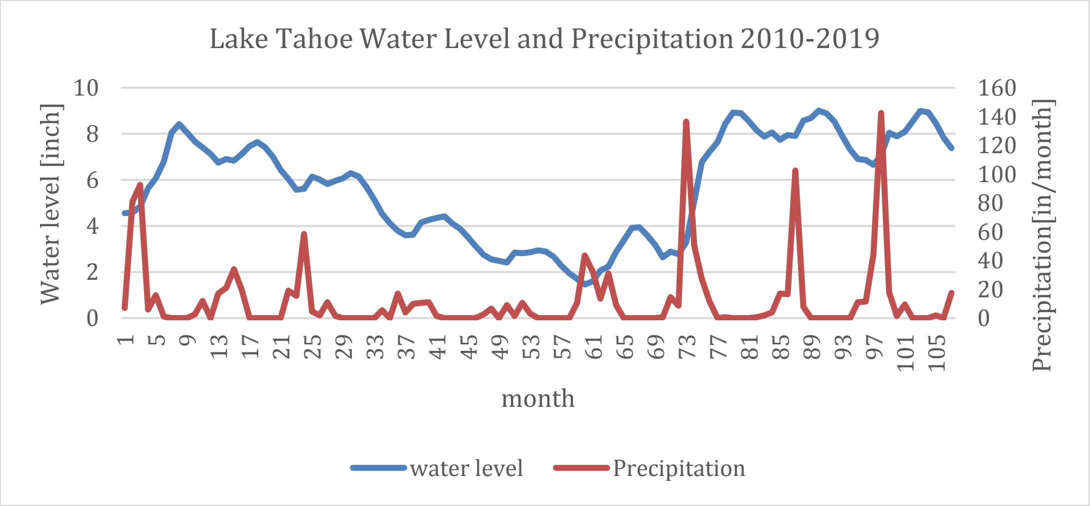

**Figure 1: Lake Tahoe Water Level and Precipitation (Jan 2010 – Nov 2019)**

## 4. FEATURES

The mass conservation equation to model the amount of water in the lake is expressed as below:

$$w(t) = w(t-1) + \Delta w_{in}(t) - \Delta w_{out}(t)$$

where $w(t)$ is the mass of water in the lake at month t, $\Delta w_{in}(t)$ is the mass of water flowing into the lake, and $\Delta w_{out}(t)$ is the mass of water flowing out of the lake.

The features to the models are the water discharge of current month, 12 months of precipitation data, and 12 months of snow fall data. One of the challenging points of modeling water level changes is that there is a time delay between input (rain/snow fall) and output (water level response) because rain fall water is stored as snow pack and/or underground water before it flows out to a lake. To incorporate this response delay, I included 12 months of data to model a lake water level in a certain month. A feature vector is expressed as

$$\phi^{(i)} = [w(t-1), d(t), r(t), r(t-1), \ldots, r(t-11), s(t), s(t-1), \ldots, s(t-11)]$$

where d = 26, r(t) is the precipitation data in a month t, and s(t) is the snow fall data in a month t.

## 5. METHODS

This is a regression problem, so my primary target is to achieve a reasonable fit by a linear regression model.

$$h_\theta(\mathrm{x}) = \theta_0 + \theta_x\phi_1 + \cdots + \theta_d\phi_d$$

where $\theta_i$'s are the parameters $\phi_i$'s are the features described in the Feature section, and d is the number of features used to predict a lake water level at a month t.

Additionally, I compared other available regression models in scikit-learn (Pedregosa et al. 2011) to the linear regression model. The regression models used in this project are the support vector regression with the linear kernel, the decision tree regression, the kernel ridge regression, the k-nearest neighbors, and multi-layer Perceptron regressor. The same features and train/val/test dataset are used.

### 5.1 Model Selection

The suitable model is selected by the k-fold cross validation. The train dataset is divided into k = 10 sets which is a common choice, and each mode is evaluated by the mean squared error.

### 5.2 Feature selection

Water discharge data, precipitation of the past 12 months, and snow fall data of the past 12 months are the features included in the model. To understand which data have stronger correlation to the output, I used a feature selection technique implemented in Scikit Learn (Pedregosa et al. 2011).

In the feature selection, the correlation between each feature and the output is computed by ((X[:, i] - mean(X[:, i])) * (y - mean_y)) / (std(X[:, i]) * std(y)), then the features are ranked by the correlation (Pedregosa et al. 2011).

## 6. EXPERIMENTS/RESULTS/DISCUSSION

### 6.1 Feature Selection

There are in total 26 features in the model. I applied the univariate linear regression tests using each feature and ranked the features based on the correlation between each feature and the target as described in Pedregosa et al. (2011).

The most correlated feature is the precipitation data half a year ago. The top 6 most correlated features are the precipitation and snow fall data between 5-7 months before the month when the lake water level is measured. This could reflect the local hydrological system around the Lake Tahoe. It would typically take around half a year for water to saturate into the ground from the ground surface to flow out to the Lake Tahoe. Also, snow pack that accumulates in winter slowly melts through summer with feeding the reservoir. This also causes half a year delay from snow fall to the lake water level response.

**Table 2: Top 6 most correlated features**

| | |
|---|---|
| 1 | Precipitation 6 month before |

| | |
|---|---|
| 2 | Snow fall 6 month before |
| 3 | Precipitation 5 month before |
| 4 | Precipitation 7 month before |
| 5 | Snow fall 5 month before |
| 6 | Snow fall 7 month before |

### 6.2 Model Selection

Five regression models are compared by the k-fold cross validation with k=10. The models are evaluated by the error measured by averaging the mean squared error regression losses over the k holdings. Each model was optimized with varied parameters beforehand.

The best model among these is Kernel Ridge regression with slightly smaller average mean squared error (0.0856) than the linear regression model. After the cross validation, the Kernel Ridge regression model was trained with the entire training set again and used to predict the lake water level from Jan 2010 to Nov 2019 as the test dataset (Figure 2). The error on the test set is 0.04540.

**Table 3: The average mean squared error regression loss with all the features and selected features**

| Model | Average mean squared error regression loss | |
|---|---|---|
| | With 25 features | With top 6 features |
| Linear regression | 0.085606 | 0.17400722 |
| Support vector regression with linear kernel | 0.085558 | 0.17390467 |
| Decision tree regression | 0.1962321 | 0.30998383 |
| Kernel ridge regression | 0.085601 | 0.17390467 |
| Multi-layer Perceptron regressor | 0.09123215 | 0.18269063 |

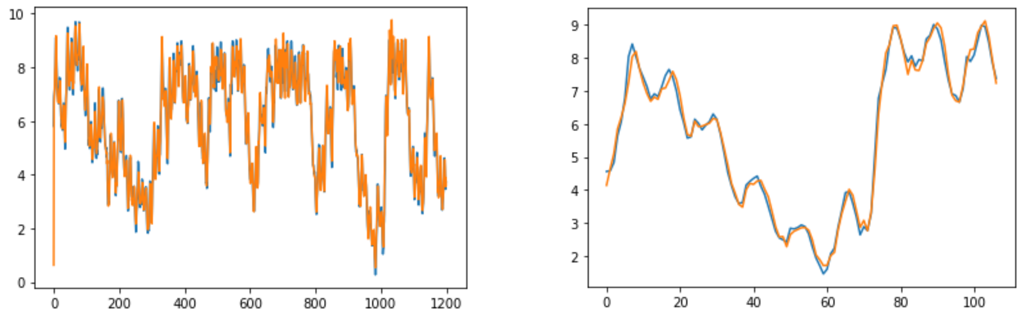

**Figure 2: Kernel ridge regression results. (Left) Train with error 0.08560 (blue: observed, orange: trained), (Right) Test with error 0.04540 (blue: observed, orange: predicted)**

The modeled results match very well to the observed data as shown in Figure 2. However, this model predicts the lake water level one month later, which means we can predict the water level only one month ahead.

In the next experiments, I used those models to understand how farther months we can predict the water level with these models. To predict the water level at M months ahead, I modified the feature vector to estimate the water level at month (t+M) where M = 0, 1, …, 5

$$\phi^{(i)}(t+M) = [w(t-1), d(t), r(t), r(t-1), \ldots, r(t-11), s(t), s(t-1), \ldots, s(t-11)\,]$$

using the lake water level, the water discharge, precipitation, and snow fall data M month before the predicted month. Table 4 shows the results. In all the cases, the Kernel Ridge has the smallest error among these models. Also, the prediction error increases as it predicts later month.

**Table 4: The average mean squared error regression loss of the prediction for M months ahead.**

| Model | Average mean squared error regression loss of the prediction for M months ahead | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| Linear regression | 0.08561 | 0.19315 | 0.32636 | 0.44292 | 0.53753 | 0.63852 |
| Support vector regression with linear kernel | 0.08556 | 0.19847 | 0.33945 | 0.45790 | 0.54911 | 0.65777 |
| Decision tree regression | 0.19623 | 0.45903 | 0.69082 | 0.91972 | 1.11523 | 1.10458 |
| Kernel ridge regression | 0.08560 | 0.19282 | 0.32593 | 0.44253 | 0.53727 | 0.63831 |
| Multi-layer Perceptron regressor | 0.09123 | 0.19620 | 0.33405 | 0.46594 | 0.54342 | 0.64720 |

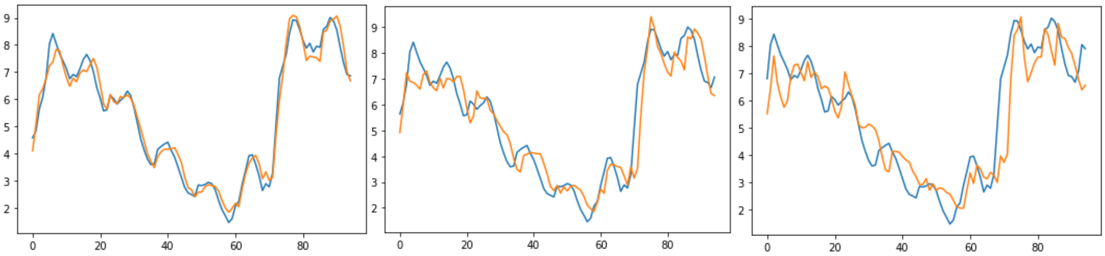

**Figure 3: The Kernel ridge regression results on the test set predicting 2 months (left), 4 months (middle), and 6 months (right) ahead the input data (blue: observed, orange: predicted).**

## 7. Discussion and Conclusion

Linear regression, SVR with linear kernel, and Kernel ridge with linear kernel models predicted the water level well. This would mean that the hydraulic system around the Lake Tahoe would be explained by a linear model in respect to the features. Kernel ridge shows the best fit to the system over all. The Kernel ridge uses linear least squares with L2-norm regularization with kernels. The kernel used in this project is a linear kernel, so the regularization worked well to prevent overfitting to the train set, which helps to show the better fit to the test set than the linear regression and SVR models. Non-linear kernels did not work well with this problem. I have tried 'linear', 'poly', 'rbf', 'sigmoid' with the SVR and the Kernel ridge, but the liner kernel shows the best fit to the validation set.

For future development, additional data set such as temperature, humidity, sun radiation, and precipitation at other stations around the Lake Tahoe, may improve the prediction. The features in this project include snow fall data, therefore, the temperature data would affect how quick snow melts and humidity would affect the vaporization from the lake and the soil in the basin.

## 8. Link to the project codes

The link to the project codes is:

https://www.dropbox.com/s/tqjh10spnjrf1jy/FinalReport_AyakaAbe.zip?dl=0

## REFERENCES

Bonakdari, Hossein, Isa Ebtehaj, Pijush Samui, and Bahram Gharabaghi. 2019. "Lake Water-Level Fluctuations Forecasting Using Minimax Probability Machine Regression, Relevance Vector Machine, Gaussian Process Regression, and Extreme Learning Machine."

Kühn, Michael and Tim Schöne. 2017. “Multivariate Regression Model from Water Level and Production Rate Time Series for the Geothermal Reservoir Waiwera (New Zealand).” Pp. 571–79 in *Energy Procedia*. Vol. 125. Elsevier Ltd.

Los Angels Times. 2017. “Gov. Brown Declares California Drought Emergency Is over - Los Angeles Times.” Retrieved December 9, 2019 (https://www.latimes.com/local/lanow/la-me-brown-drought-20170407-story.html).

Pedregosa, Fabian, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, and Bertrand Thirion. 2011. *Scikit-Learn: Machine Learning in Python*. Vol. 12.

PPIC. 2016. “California’s Latest Drought - Public Policy Institute of California.” Retrieved December 9, 2019 (https://www.ppic.org/publication/californias-latest-drought/).

U.S. Forest Service. 2019. “Lake Tahoe Basin Mgt Unit - About the Area.” Retrieved December 9, 2019 (https://www.fs.usda.gov/main/ltbmu/about-forest/about-area).

USGS. 2019a. “USGS Current Conditions for USGS 10337000 LAKE TAHOE A TAHOE CITY CA.” Retrieved November 28, 2019 (https://waterdata.usgs.gov/nwis/dv?cb_00065=on&format=gif_default&site_no=10337000&referred_module=sw&period=&begin_date=2017-11-17&end_date=2019-11-17).

USGS. 2019b. “USGS Surface Water Data for USA: USGS Surface-Water Monthly Statistics.” Retrieved November 30, 2019 (https://waterdata.usgs.gov/nwis/monthly?referred_module=sw&amp;site_no=103366092&amp;por_103366092_102819=172692,00060,102819,1990-06,2019-10&amp;format=html_table&amp;date_format=YYYY-MM-DD&amp;rdb_compression=file&amp;submitted_form=parameter_selection_).

Water Education Foundation. 2019. “Truckee River.” Retrieved November 28, 2019 (https://www.watereducation.org/aquapedia/truckee-river).

Watts, Kenneth R. 1995. *Regression Models of Monthly Water-Level Change In and Near The Closed Basin Division of the San Luis Valley, South-Central Colorado Water-Resources Investigations Report 93-4209 Prepared in Cooperation with The*.

Western Regional Climate Center. 2019. “TAHOE CITY, CA, Total of Precipitation (Inches), Monthly Precipitation Listings, Monthly Totals.” Retrieved (https://wrcc.dri.edu/cgi-bin/cliMAIN.pl?ca8758).