Fuzzy Systems and Data Mining VII
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA210206

Responses of Climate Indicators to Droughts in SF Bay Area

Patrick Li[1 a] and Gang Li [b]
[a] *Basis Independent Silicon Valley, High School Student*
[b] *Panyi Technology Inc, San Jose, VP of Engineering*

Abstract. Droughts have appeared in many different regions that increase the chances of wildfires and other health risks like heat strokes. With satellite imaging and more collections on the Google Earth Engine (GEE) library, more information is available to discover trends. This study analyzes different causes and signs of historical drought data in the San Francisco Bay Area then uses several machine learning tools to model the drought.

Keywords. Google Earth Engine, Drought, Climate Change, Cloud Computing, Large-Scale Data Mining.

1. Introduction

The research in the paper was done on Google Earth Engine (GEE) [1] because of its big data parallel computing power. GEE combines multi-petabyte satellite imagery and geospatial datasets with planetary-scale analysis capability [1] making it a good platform to research global and regional climate change. Several GEE datasets were used to analyze the drought in California San Francisco Bay Area with hopes of providing insight on certain climate change indicators. The results might affect this technology innovation center and provide the fundamentals for policy makers on future droughts.

2. Related Works

There are a lot of researches to analyze and predict droughts in different regions and using different methods including machine learning. Peiyu Lai et al. [2] analyzed seasonal indicators related to droughts in Southwest China. Their work performed comprehensive survey on changes of variety of variables including surface water area, vegetation, meteorological factors and human activities, and whether these changes match the start and end of drought period on record without providing a mathematical modeling for predicting drought. Researches [3] have been made with Internet of Things (IoT) sensors that transfer meteorological information to a fog layer. Theses sensors include temperature, humidity, precipitation, and water levels. This fog layer would compress the data and then transfer it to the cloud for computations. This work's main focus is on building infrastructure for real-time data collection and computing. Another

[1] Corresponding Author, Patrick Li, Basis Independent Silicon Valley, High School Student; Email: patrick.h.li@gmail.com

study [4] uses the Canadian Earth System Model (CanESM2) produced by ClimEx project by Ouranos with the Canadian Regional Climate Model (CRCM5) as well as CanESM2-LE's monthly sea level pressures, to create accurate drought predictions when combined with artificial neural networks. Data captured with satellites have also been used. Their major contribution is applying an Artificial Neural Network drought model to two European domains, Munich and Lisbon. Sumin *et al.* in their research [5], the Scaled Drought Condition Index (SDCI), Standard Precipitation Index (SPI), and topographic characteristics were used in convolutional long short-term memory (convLSTM) and random forest models to generate predictions. Their area of study focused on East Asia.

The study in this paper focuses in San Francisco Bay Area, one of the most populated and important high-tech economic regions, the water shortage has become more frequent in recent years impacting both daily life and economic activities. This area is not studied in related researches. The method proposed uses readily available massive satellite and geo-sensing data to extract information including permanent water area (PWA), seasonal water area (SWA), temperature, precipitation, and drought area and index. The study of this research tries to use these publicly available, valuable datasets to do the data mining, avoiding the need for extra sensor installation or data collection which were performed by some related researches [3]. The extracted data is analyzed against the drought historic records. This data is also feed into the proposed model using a linear regression to train and predict drought patterns. The result is evaluated with correlation coefficient and shows competitive or better performance compared to some earlier researches as in [5] for example.

3. Proposed Methodology

3.1. Datasets Used

In this study, we choose to use the satellite imagery and geospatial datasets readily available in Google Earth Engine (GEE) database. The datasets contain massive global data and we utilize GEE parallel computing capability to extract the data specifically for our region of interest (RoI) – San Francisco Bay Area, and use the powerful GEE API to reduce the multi-dimensional imagery and geospatial data to several easy-to-process indicators such as surface water area, temperature and precipitation. We call this step data extraction stage with GEE. The program is written in JavaScript using GEE Code Editor platform seen in Table 1.

Table 1. Experiment Details

	Tools/Library	**Language**	**Platform/Framework**
Data Extraction Stage	Gee Code Editor	JavaScript	Google Earth Engine
Data Condition Stage	Pandas, Numpy, Jupyter	Python	None
Training Stage	Keras, XGBRegressor	Python	XGBoost/Tensorflow

The datasets include JRC Yearly Water Classification History v1.2 [6], TIGER: US Census Counties 2018 [7], GRIDMET DROUGHT: CONUS drought indices [8], and ERA5 Monthly aggregates – Latest climate reanalysis produced by ECMWF / Copernicus Climate Change Service [9]. Most of the data are from satellite remote sensing images and contain huge amount of information. A single step of data processing usually takes hours even with power of GEE parallel computing. Once the data is

processed and extracted, a model using decision-tree ensemble model is constructed with XGBoost to compute predictions, the linear regression and squared error objective were used to train the model.

The JRC Yearly Water dataset [6] contains occurrence of permanent surface water (PSW), seasonal surface water (SSW) and No-Water. The GEE satellite image of geographic region map is split into small tiles, and each tile is labeled with the occurrence of either PSW, SSW or No-Water. The tiles marked with PSW are integrated into total Permanent Water Area (PWA) in the region of interest (RoI). The integration is calculated by GEE Reducer function parallel computing engine. Similarly, SSW labels are used to compute SWA, and No-Water for No-Water Area. In order to focus our analysis on SF bay area, only the areas within the 6 counties of the SF Bay area were included as RoI: Almeda, Contra Costa, Marin, Napa, San Francisco, and Santa Clara. The outlines of these counties were extracted from the TIGER: US Census Counties 2018 dataset [7] as a polygon into GEE and used as a region of interest (RoI) for the data reducer. Figure 1 shows graphs of the areas of No-Water Area, SWA, and PWA which are blue, orange, and gray respectively. The areas are in kilometers squared for each year from 1985 to 2018.

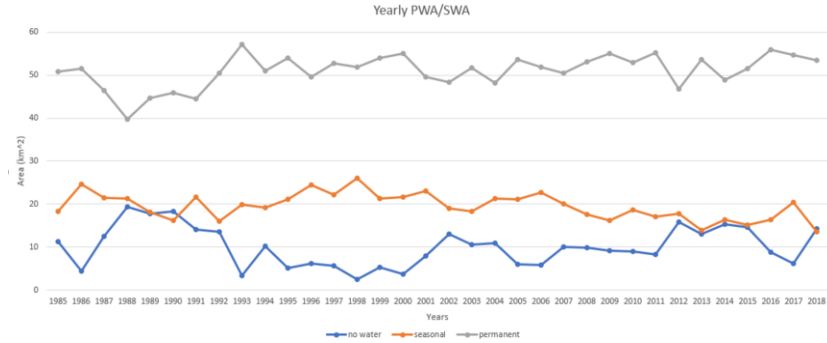

Figure 1. Yearly PWA, SWA and No-Water Area in SF bay area.

ERA5 monthly aggregates dataset [9] contains several global climate parameters, among which the 2-meter (2m) air temperature and total precipitation data were used to analyze the climate change. The monthly air temperature in Santa Clara County were extracted, integrated, averaged and shown in Figure 2. Similarly, monthly precipitation in Santa Clara County change were shown in Figure 3.

To find the drought period and area, the GRIDMET DROUGHT: CONUS drought indices dataset [8] was used. It contains the Evaporative Drought Demand Index (EDDI) around every 5 days from 1985 to 2020. Using the monthly 30-day average, the EDDI drought index from 1985 to 2020 was calculated. Index less than -1.3 is considered a moderate drought, -1.99 to -1.6 is a severe drought, and -2.0 or below is an extreme drought. EDDI was used to mark drought locations [7] in Santa Clara County, and the locations were then integrated to compute the monthly drought area in Figure 4. The drought period could then be identified and compared to PWA and SWA change in Figure 1. A drought is expected to be an increase in drought area, the No-Water Area, and monthly temperature, and drop in precipitation, PWA and SWA. However, many factors lead to the result of a drought and some droughts with changes in one factor are less noticeable than ones with large differences.

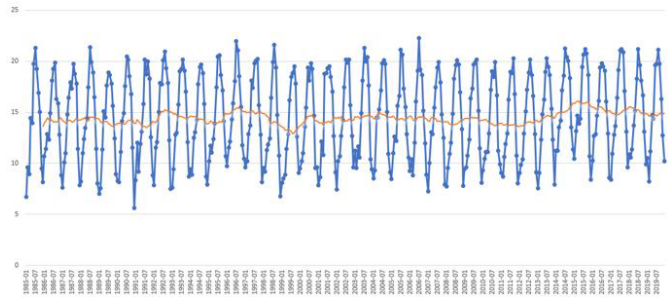

Figure 2. Monthly Temperature (blue curve) and 12-month moving average (orange curve) in Santa Clara County

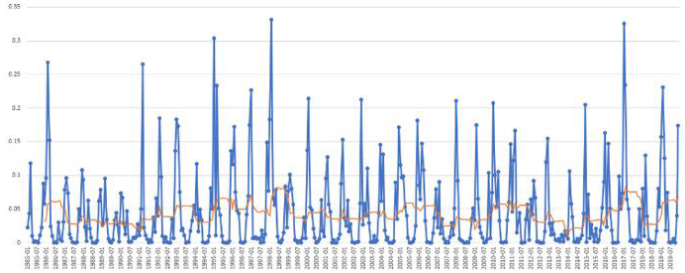

Figure 3. Monthly Precipitation (blue curve) and 12-month moving average (orange curve) in Santa Clara County

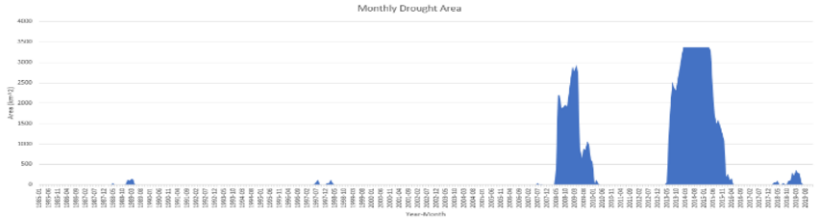

Figure 4. Monthly Drought Area in Santa Clara County

3.2. Proposed Model

This paper's interest is to find the correlation between the drought occurrence time and pattern with respect to the geo-sensing and climate data such as surface water area, temperature and precipitation. A model was made with decision-tree ensemble in XGBoost's library, and uses the collected data to train itself. XGBoost is highly effective and widely used machine learning method [10]. The training process is fast and easy to fit variety of target data. The tree ensemble structure is also stable and fast to train and fine-tune whenever new data are added or revised. Since this research is an on-going work, the XGBoost method is chosen so that more data could be easily added and evaluated. The XGBoost is not like Convolutional Neural Network (CNN) or Recurrent Neural Network (RNN) which is hard to interpret, the trained model can actually provide many insights into the mechanism under study, for example the most important or influential input factors to the result. At the end of this section, the experiment of another popular RNN model is also provided.

The input data includes time series (1985-2018) of temperature, precipitation, PWA, SWA and No-Water Area. Month as numbers (1-12) is also added as input to include the

impact of seasonal change. The target ground truth is the calculated drought area as in Figure 4. Temperature and precipitation data is pre-processed using 12-month moving average to remove intrinsic fluctuation noises. Since all this data has different unit and range, they are normalized to facilitate the model training. The normalized data is converted to pandas DataFrame format before fed into the model. Normalized example data is shown in Table 2. This step is the data conditioning stage. The code is written in Python and uses libraries listed in Table 1.

Table 2. Normalized input data and drought ground truth.

MONTH	TEMP	RAINX100	PWA	SWA	NOWATER	DROUGHT
1	6.7224	2.2597	50.8733	18.2422	11.2207	0.0
2	9.6151	4.3679	50.8732	18.2422	11.2207	0.0
3	8.9493	11.7706	50.8732	18.2422	11.2207	0.0
4	14.438	0.9999	50.8732	18.2422	11.2207	0.0
5	13.951	0.1271	50.8732	18.2422	11.2207	0.0

The model is constructed in XGBoost and trained with objective of squared error to the ground truth. The model parameter colsample_bytree is set to 0.6 such that 60% of the data is used for training to avoid overfitting. Maximum tree depth, regulation weight alpha, number of estimators are adjusted to find the optimal model setting. The model is constructed with XGBRegressor() function and trained with fit() procedure. In order to evaluate the accuracy of the model prediction, correlation coefficient (r) is used as statistical metric.

$$r = \frac{n(\Sigma y\hat{y}) - (\Sigma y)(\Sigma \hat{y})}{\sqrt{[n\Sigma y^2 - (\Sigma y)^2][n\Sigma \hat{y}^2 - (\Sigma \hat{y}^2)]}}$$

n is the number of the samples, y and $\hat{y}$ are the values of reference and predicted drought area. The modeling and training program is written in Python using XGBoost library package as listed in Table 1. We call this step modeling and training stage.

We also evaluated a RNN model which is widely used for analyzing time series sequential data. The model is constructed using 3 layers of Long Short-Term Memory (LSTM) with 20 units each and a Dense fully-connected layer at the end to output the predicated drought area value. The input data is re-arranged with 5 features (PWA, SWA, No-Water Area, Temperature, Precipitation) and 10 time-steps as a 2-dimensional array before fed into the model. "Adam" was used as the optimizer for training with learning rate set at 0.001. To facilitate the training, we added dropout after each layer to avoid overfitting. LSTM models are hard to train and are prone to weight explosions, so we used gradient clipping with a value of 0.2 and normalized the input data by scaling it to ±0.25 range. Different LSTM cell units, learning rates, dropout rates and clipping values are experimented with. However, the training of this model was not stable and was hard to converge within a reasonable number of epochs. Based on evaluations above, as well as other advantages of XGBoost described at the beginning of this section, XGBoost is chosen in this research.

4. Analysis and Experiment

4.1. Analysis

There is no obvious visual trending correlation between the extracted data and the drought record. Any single input data does not show strong correlation to start or end of

the two drought periods: May 2008 – June 2010, May 2013 – April 2014 shown in Figure 4.

For example, PWA shows the permanent water levels had actually risen slightly during drought period of May 2008 – June 2010 which is counter-intuitive, while drought period of May 2013 – April 2014 could be seen from a peak in No-Water Area or drop in PWA. No-Water Area shows some peaks before the drought period, while the time interval length from the peak to the actual drought start varies. On the other hand, some data correlation could be observed. The average data of No-Water Area was 11.056 km^2 with PWA being 53.629 km^2, during the drought No-Water Area increased to 14.923 km^2 with the PWA dropping to 50.180 km^2.

Looking at the temperature data itself in Figure 2, it is hard to identify the drought period. Even though the drought period of May 2013 - April 2014 can be seen from the slight increase of temperature in the winter from 8.099°C to 8.913°C and increase in summer from 20.253°C to 21.202°C, the drought of May 2008 - June 2010 could not be identified, which means that there is not a strong relationship between droughts and the temperatures data alone during them. We see similar characteristics in precipitation data.

The monthly temperature (as in Figure 2), monthly precipitation (as in Figure 3), PWA, SWA and No-Water Area (as in Figure 1), and the drought levels were all gathered into a dataset and put into a model. Because the temperatures were constantly changing through the different seasons in the year the model produced with the data would not be have seen a pattern of sudden spikes of temperature. The solution was to take the moving average in intervals of 12 months to smooth the data points and make it easier to spot periods of drought. The same method of taking the moving average was used on the monthly precipitation for the model. Figures 2 and 3 shows the actual data as blue lines and the calculated average as the orange line. Month as number (1-12) is also used as input to take the seasonal change into account.

4.2. Experiment and Results

Three model parameters were adjusted in order to find the model with the most accuracy, which was tested by comparing the correlation coefficient of the model. These three were the max depth of the model (*max_depth*), L1 regularization (*alpha*), and number of estimators (*n estimators*). The *max_depth* controls how many levels there are in the model, because the XGBoost model uses decision-trees, the larger the max depth, the more complex the model is. The *alpha* value controls L1 regularization on the weights so weights would not have too much of an influence on a result to avoid overfitting and to make the model more conservative. The *n_estimator* value is the number of decision trees that the model uses together to make a prediction. Larger *n_estimator* results in more complex model. We choose [3,8] for *max_depth*, [10-30] for *alpha* and [5-30] for n_*estimator* as data range in this experiment. After testing different values, increasing the *max_depth* of the model increased the correlation coefficient *(r)* up to 0.99478 with a *max depth* of 8, however this could be the cause of overfitting. Changing the *alpha* values had little effect as the correlation coefficient did not change much. The last parameter changed was the *n_estimators*. Like the *max_depth*, increasing this value would also improve correlation coefficient because there are more decision trees together. While the parameters changed, the feature importance were also captured, this would give what the model deemed the biggest contributor to its prediction. All correlation values above 0.96 had temperature as their most important feature, with high f scores as shown in Figure 5. As the final result, we choose the model with less complexity while

correlation coefficient *(r)* is above *0.96*: *max_depth=6, alpha=15, n_estimator=26, r=0.966*. The calculated drought area output from this model vs. the ground true is shown in Figure 6.

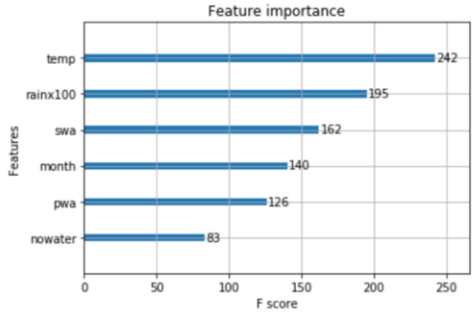

Figure 5. Feature Importance in XGBoost model.

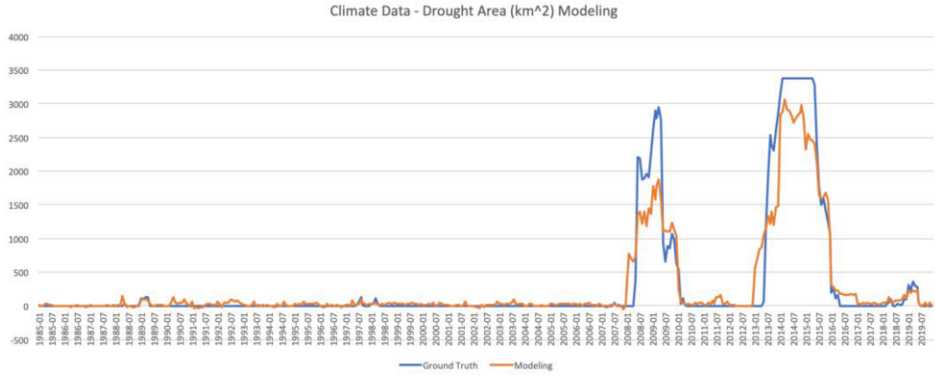

Figure 6. Drought Model vs Drought Ground Truth.

In practical application, predicting future drought based on past and current climate data could enable policy maker to take action in advance. To meet this purpose, the target is shifted 2 years in time axis and similar model is trained. As shown in Figure 7, the model could predict the drought occurrence 2 years in advance.

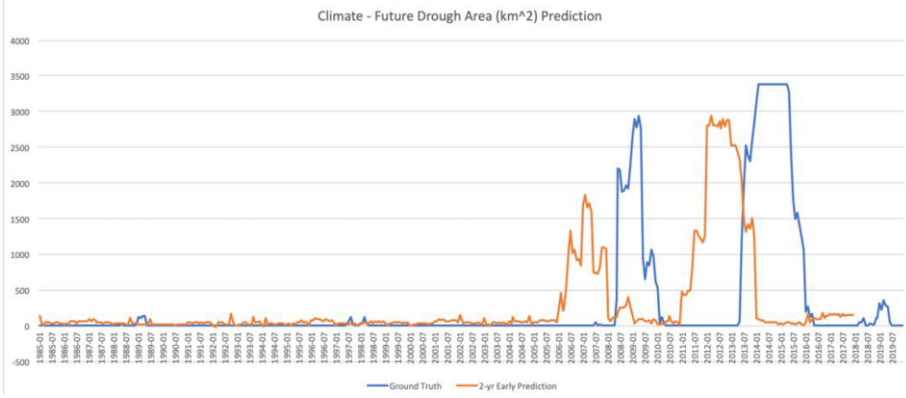

Figure 7. Drought 2-yr Early Prediction vs Drought Ground truth.

5. Conclusion and Future Work

In this work, the author extracted several climate data from GEE satellite imagery and geospatial datasets, including surface water area, precipitation, temperature and drought area in SF bay area. The climate data vs. drought area/occurrence is modeled using decision tree ensemble with XGBoost, trained with linear regression and squared error objective. With time shift in target data, the model is also able to predict the drought occurrence in advance. The model prediction shows good results in terms of r (around 0.96), competitive to other research results [5] (r is around 0.9). Since climate change has long term and complicated impact, more advanced models like RNN/LSTM would be explored to add the accumulation effect, and the prediction would be more robust and precise. More data would be incorporated in the future work, such as vegetation area, cloud area, and possibly the human activity data to explore the more complex drought occurrence mechanism. Due to the fact that the climate change is in larger geographic scale, the author also plans to expand the region of interest to include more surrounding counties' impact and model more accurate models.

References

[1] Google Earth Engine Documents, https://developers.google.com/earth-engine/
[2] Peiyu Lai, Miao Zhang *et al.* , "Responses of Seasonal Indicators to Extreme Droughts in Southwest China", in remote sensing, March 3rd, 2020
[3] Amandeep Kaur and Sandeep K. Sood, "Artificial Intelligence-Based Model For Drought Prediction and Forecasting", The British Computer Society 2019, Advance Access publication on 17 November 2019
[4] Elizaveta Felsche and Ralf Ludwig, "Applying machine learning for drought prediction using data from a large ensemble of climate simulations", Natural Hazard and Earth System Science, 15 April 2021
[5] Sumin Park, Jungho Im, Daehyeon Han and Jinyoung Rhee, "Short Term Forecasting of Satellite-Based Drought Indices Using Their Temporal Patterns and Numerical Model Output", MDPI Remote Sensing, 24 October 2020
[6] JRC Yearly Water Classification History, https://developers.google.com/earth-engine/datasets/catalog/JRC_GSW1_2_YearlyHistory
[7] TIGER: US Census Counties 2018, https://developers.google.com/earth-engine/datasets/catalog/TIGER_2018_Counties
[8] GRIDMET DROUNT: CONUS drought indices, https://developers.google.com/earth-engine/datasets/catalog/GRIDMET_DROUGHT
[9] ERA5 Monthly aggregates – Latest climate reanalysis produced by ECMWF/Copernicus Climate Change Service, https://developers.google.com/earth-engine/datasets/catalog/ECMWF_ERA5_MONTHLY#:~:text=ERA5%20MONTHLY%20provides%20aggregated%20values,10m%20v%2Dcomponent%20of%20wind.&text=All%20other%20parameters%20are%20provided%20as%20monthly%20averages.
[10] Tianqi Chen, Carlos Guestrin, "XGBoost: A Scalable Tree Boosting System", Cornell University arXiv.org, 10 Jun 2016