Prediction of Rainfall in California

Swarna Sinha[1]
Stanford University, Stanford, CA, 94305

I. Introduction

CALIFORNIA'S ongoing drought has recently resulted in some daunting repercussions, including increased incidence of wildfire and a lack of water supplies for agriculture throughout the state. The latter is of additional concern since California is the largest agricultural producer in the country in terms of total output and exports. This study aims to utilize measured data to predict rainfall in Fresno, CA, which is one example of an agricultural region in central California that is threatened by drought. Farmers face challenging decisions every year, not knowing whether to pack up, downsize, or continue as usual with their crops.

A simple tool, developed by machine learning, could be highly useful if it predicts rainfall accurately even one month or a year in advance. This project will explore the machine learning element of this problem, applying elastic net, support vector machine, and random forest regression methods to available data. The resulting precipitation models will be evaluated on a testing data set and compared.

II. Related Work

Climate modeling, while a useful tool, simulates highly nonlinear systems with limited long-term success due to the chaotic nature of these systems. Lately, machine learning has been integrated with climate modeling in a new field coined as "climate informatics [1]." The range of potential applications of machine learning is very broad, ranging from implementation in physics-based climate simulations to approaches involving measured data only.

Local weather is invariably linked to large-scale, global trends and the resulting physical simulations are computationally expensive. The timescales of certain climate trends, such as ice ages, are also extremely long, reaching back to historical times where we have no data. Machine learning has been implemented in climate modeling in various ways. It has been used to augment historical climate data as well as simulations.

Monteleoni et. al. have forecasted weather data for the Intergovernmental Panel on Climate Change (IPCC), running several different climate models simultaneously and adaptively weighting the results with a machine learning approach. This significantly bolstered accuracy and confidence in the predictions [2]. Alternatively, machine learning can be implemented on weather forecasts to derive other useful metrics, such as cloud cover and solar power generation in the future. Least-squares regression and support vector machines had considerable success in this application [3].

Again, a significant challenge in the current problem is data availability. High-quality observations only extend to the 1950s, at earliest. Another relatively uncommon approach is to run a weather simulation and train unknown physical parameters with the available observations [4]. A variant of this approach has been created by Rasouli et. al., using a climate model along with local and global observations [5]. This study will execute a similar technique, developing a model with machine learning approaches that may later incorporate features provided by simulations. Direct climate modeling will not be utilized for the scope of this project, since this requires expensive computational resources.

III. Dataset and Features

Data for Fresno, CA and some global parameters were acquired from multiple sources. The National Oceanic and Atmospheric Administration's (NOAA) publicly accessible database, the National Centers for Environmental Information (NCEI), contains monthly statistics for Fresno, CA [6]. Table 1 provides a list of stations across Fresno, CA that have measured different subsets of features.

Table 1: NCEI Data Stations for Fresno, CA

CLOVIS 1.3 NE	FRESNO 3.4 SSE	FRESNO 7.1 E
CLOVIS 2.9 N	FRESNO 4.3 NNW	FRESNO YOSEMITE INTERNATIONAL
FRESNO 2.4 NW	FRESNO 5 NE	FRIANT GOVERNMENT CAMP
FRESNO 3 NW	FRESNO 5.9 NNE	SANGER 6.8 N

[1] Graduate Student, Department of Aeronautics & Astronautics

Significant pre-processing was required to include all of the available features from the different stations for every monthly data point. Any features measured at more than one station at any given time were averaged. For the poster session, a much smaller set of features had been incorporated. Therefore, the results in this report may appear significantly different. Table 2 summarizes the list of features used to predict total monthly precipitation in inches. Note that two global parameters, atmospheric CO_2 emissions (millions of pounds per year) and temperature (°F), have also been included to very roughly account for global effects. These were downloaded from the Earth Policy Institute (EPI) [7]. The local depth to water table from the surface (feet) of Fresno, CA was acquired from the city of Fresno's website [8]. Figure 1 shows a subset of standardized features.

Table 2: List of Features, 1985 to 2009

Local =		**Global =**		Monthly mean of max. soil temperature for grass (°F)
Cooling Degree Days (season-to-date)				Monthly avg. temperature (°F)
No. days with min. temp <= 0 °F				Monthly max. temperature (°F)
No. days with min. temp <= 32 °F				Monthly min. temperature (°F)
No. days with max. temp ≥ 32 °F				Total monthly wind movement over evaporation pan (mi)
No. days with max. temp ≥ 70 °F				Depth to Water Level (ft)
No. days with max. temp ≥ 90 °F				Monthly Avg. Wind Speed (mph)
Total Monthly Evaporation				Global Temperature (°F)
Heating Degree Days (season-to-date)				Global Million Tons of Carbon
Monthly mean of min. soil temperature for grass (°F)				Time (years)

Finally, the response variable of interest is the total monthly precipitation in inches. The following analysis includes precipitation data with the full set of features in Table 2 from 1985 to 2009. Additional precipitation measurements extend to the present day in 2016.

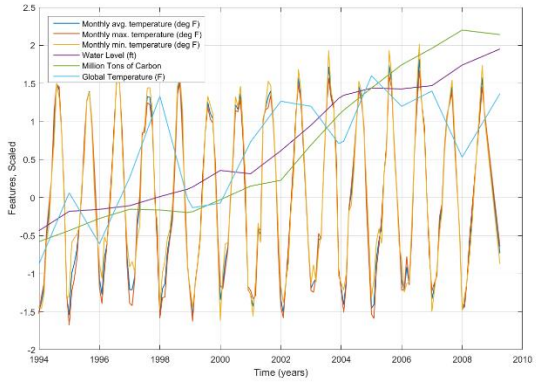

Figure 1: Subset of Features

IV. Methods

This section briefly describes the machine learning approaches under consideration: Elastic Net Regression , Support Vector Regression, and Random Forest Regression. All methods employ 10-fold cross-validation and standardize the features prior to analysis. Available MATLAB toolboxes were used to implement these approaches.

A. Elastic Net Regression

This technique is like least-squares regression, but with a penalty term containing a weighted sum of the L1- and L2-norms of the weights, $\vec{\beta}$. The N observations are the monthly measurements of the response variable and features, denoted as $\vec{y}$ and $\bar{x}$, respectively. $\overrightarrow{x_j}$ is a measured feature such as depth to water table, average local temperature, etc. and $j = 1, \ldots, p$. The optimization approach is the following:

$$F_{EN} = argmin_{\beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \left(\alpha \sum_{j=1}^{p} |\beta_j|^2 + (1-\alpha) \sum_{j=1}^{p} |\beta_j| \right)$$

The penalty term encourages feature selection, driving weights to zero for features that do not affect the response variable. This is useful because it may tell us which features affect rainfall the most. α may be adjusted anywhere from 0 to 1. α closer to 1 behaves more like the Lasso Method, while α near 0 behaves more like Ridge Regression. The latter drives weights to zero more readily than the former. The Elastic Net Method is convenient since it is a compromise between the two. The algorithm takes a first guess of the weights, $\widehat{\beta_j}$, using a standard least-squares fit. We can define the initial constraint, $t_0 = \sum_{p=1}^{j} \widehat{\beta_j}$, and the shrinkage parameter, $s = t/t_0$. As $s \to 0$, the weights go to zero. The algorithm increments over decreasing values of t and produces a solution for $\vec{\beta}$ at each increment. Some constant λ is chosen to proportionately shrink the coefficients at each increment. This technique is also known as "soft-thresholding."

Given a set of solutions, 10-fold cross validation is used to determine the λ that produces minimum mean-squared error (MSE). In summary, the data set is divided into 10 subsets. Using all subsets except the 10th subset, the algorithm produces many solutions for $\vec{\beta}$ over a range of λ. Then, for each λ, the MSE is computed over the points in the 10th subset. The λ that produces minimum MSE thus determines the ideal weights [9].

B. Support Vector Machine (SVM) Regression

This method extends support vector machine classification to continuous problems. The goal is to find a function $f(x)$ that predicts the response variable within a designated error, ε. We minimize the following dual formula for nonlinear SVM regression:

$$L(\alpha) = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} (w_i - w_i^*)(w_j - w_j^*)\, G(x_i, x_j) + \varepsilon \sum_{i=1}^{N} (w_i + w_i^*) - \sum_{i=1}^{N} y_i (w_i - w_i^*)$$

subject to

$$\sum_{i=1}^{N} (w_i - w_i^*) = 0, \quad \forall i: 0 \le w_i \le C, \quad \forall i: 0 \le w_i^* \le C$$

where the Gram matrix, G, is comprised of Gaussian kernels and C is a cost parameter.

$$G(x, z) = \varphi^T(x)\varphi(z), \quad \varphi(x) = \exp(-\frac{\|x - \mu\|^2}{2\sigma^2})$$

The Karush-Kuhn-Tucker (KKT) complementary conditions are applied to close the model, and a sequential minimal optimization (SMO) algorithm solves for the Lagrange multipliers, w_i. Any points with non-zero Lagrange multipliers, or points that are not close to the predictions, are support vectors that define the function $f(x)$. Therefore, $f(x)$ only depends on a subset of the training data.

$$f(x) = \sum_{i=1}^{N} (w_i - w_i^*)\, G(x_i, x) + b$$

SVM regression is a robust tool for many applications [9]. Therefore, it is reasonable to test it on the weather data.

C. Random Forest Regression

The random forest algorithm is more commonly applied to classification problems, where there is a much larger selection of boosting techniques. For regression, the random forest algorithm fits a number of randomly generated decision trees on subsets of the data and uses averaging to improve accuracy and prevent over-fitting. Each tree is generated by splitting the feature set into subsets. The process continues until a subset at a given node equals the response variable. Least-squares boosting is applied. At every step, the ensemble fits a new learner to the difference between the observed response and the aggregated prediction of all prior learners. Mean-squared error is minimized for every fit [9].

In this application, the MATLAB function is applied for various numbers of learners (NL), or decision trees. A larger number of learners should result in lower training error, as will be shown in the results. The prediction, based on a feature set, is simply the average of predictions over all learners.

V. Results & Discussion

The data was partitioned into training and training sets, spanning from 1985-2004 and 2005-2009, respectively. The training set contains 225 points while the testing set contains 40 points. The error is taken as the mean absolute error between the predictions $f(x)$, and original data, y, where there are N observations.

$$error = \frac{1}{N}\sum_{i=1}^{N}|y_i - f(x_i)|$$

A model for monthly precipitation is created for every algorithm that is executed on the training data. The training data is resubsituted into the model, and the error from this prediction is the training error. The testing error results from error between original precipitation data and predictions from inputting unseen testing data into the model. These results are summarized in Table 3.

Table 3: Error Summary of Different Models (Units: inches, monthly precipitation)

Method	Parameter	Training Error	Testing Error
Elastic Net	α = 0.2	0.504	0.800
	α = 0.5	0.505	0.808
	α = 0.8	0.508	0.918
	α = 1	0.530	0.620
SVR		0.522	0.656
Random Forest	NL = 10	0.562	0.940
	NL = 50	0.366	0.860
	NL = 100	0.286	1.028
	NL = 150	0.237	1.060

Elastic net regression shows the least testing error for α=1. SVM regression is the next best method, while the random forest algorithm has substantially larger testing errors, but the smallest training errors. It appears that the random forest algorithm does not generalize as well to unseen data, but perhaps setting a larger number of learners would rectify this. Figure 2 shows plots of total monthly precipitation for the different algorithms. Figure 2(a) shows results for α=1 and Figure 2(d) shows how the error for random forest regression behaves with an increasing number of learners. As can be seen on the plots, elastic net and SVM regression predict the regular precipitation spikes accurately, but always underpredict any extremes. Random forest is capable of predicting isolated, larger spikes in precipitation, even if its overall testing error is larger.

Table 4: Weights for Dominant Features in Elastic Net Regression, α = 1

Feature	Weight, β_j
Monthly Avg. Wind Speed (mph)	0.28
No. days with max. temp ≥ 70 °F	-0.05
No. days with max. temp ≥ 90 °F	-0.04
Total Monthly Evaporation	-0.26
Monthly mean of max. soil temperature for grass (°F)	-0.05
Monthly max. temperature (°F)	-0.05
Monthly min. temperature (°F)	0.25
Global Temperature (°F)	-0.03
Time (years)	-0.06

Results from elastic net regression are shown in Table 4. Any weights that are negligibly small or zero have been omitted. Interestingly, the weights show intuitive trends. Monthly precipitation decreases with evaporation, the number of days with max. temperatures greater than 70 or 90 degrees, mean max. soil temperature for grass, global temperature, monthly max. temperature, and time. However, a higher averge wind speed and min. temperature both result in more rainfall.

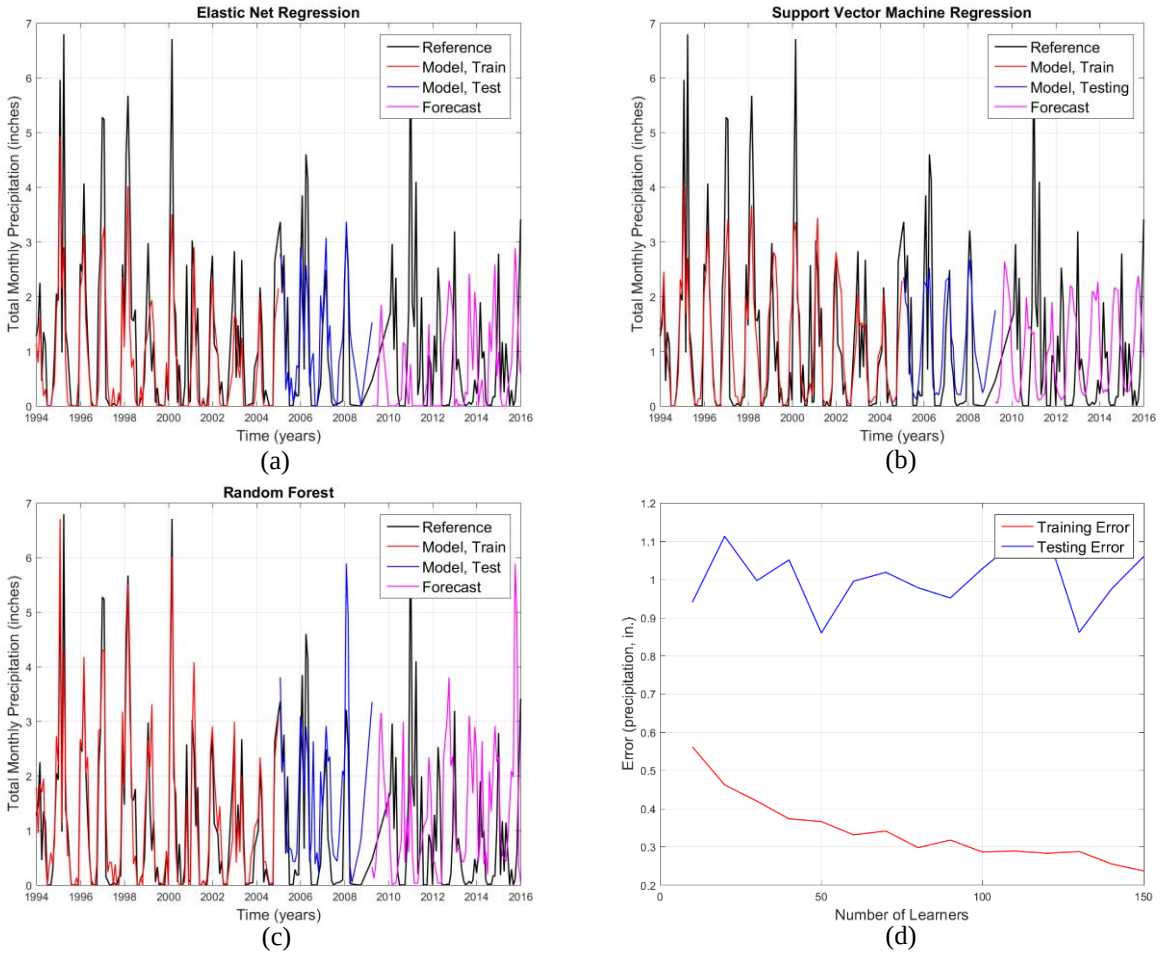

Figure 2: Monthly Precipitation Predictions vs. Time for (a) Elastic Net Regression for α=1, (b) SVM Regression, and (c) Random Forest Regression with 50 Learners, with (d) Random Forest Error vs. No. of Learners

The images contain a third prediction called "Forecast." This is generated for years 2009-2015 by using features from 2002-2008. Recall that we only have a full feature set up until 2009. The only updated feature is time, in years. While this is obviously not the best feature set for future predictions, the "Forecast" curve is a placeholder for future work. Recommendations for improving this curve are provided below.

VI. Conclusion & Future Work

This study compared elastic net, SVM, and random forest regression methods on set of measured climate data for Fresno, CA to predict monthly precipitation. Elastic net and SVM regression were found to be the most effective methods. However, the study is limited to the testing and training data and is currently unable to reasonably predict the future.

For further work, predicting a weather forecast in the future, where we have no data, is the next step. This requires simulation of the features in the data set. For some features, such as depth to water table, this is a fairly straightforward extrapolation. However, average wind speed, temperatures, evaporation, and other features require more sophisticated simulations. A more holistic analysis of the relevant global parameters is also required. This forecast model could be tested on the data from 2009-2015, as shown above, to verify the results.

References

[1] Arindam Banerjee, and Claire Monteleoni. "Climate Change: Challenges for Machine Learning." (n.d.): n. pag. NIPS, 2014. Web.

[2] Monteleoni, Claire, et al. "Tracking climate models." Statistical Analysis and Data Mining 4.4 (2011): 372-392.

[3] Sharma, Navin, et al. "Predicting solar generation from weather forecasts using machine learning." Smart Grid Communications (SmartGridComm), 2011 IEEE International Conference on. IEEE, 2011.

[4] Monteleoni, Claire, Gavin A. Schmidt, Francis J. Alexander, Alexandru Niculescu-Mizil, Karsten Steinhaeuser, Michael Tippett, Arindam Banerjee et al. "Climate informatics." (2013).

[5] Rasouli, Kabir, William W. Hsieh, and Alex J. Cannon. "Daily streamflow forecasting by machine learning methods with weather and climate inputs." Journal of Hydrology 414 (2012): 284-293.

[6] "National Centers for Environmental Information (NCEI)." National Centers for Environmental Information (NCEI). Web. 11 Dec. 2016. <https://www.ncdc.noaa.gov/>.

[7] Data Center | EPI. Web. 16 Dec. 2016. <http://www.earth-policy.org/data_center/>.

[8] "Historical Water Level." Historical Water Level. City of Fresno. Web. 16 Dec. 2016. <http://www.fresno.gov/Government/DepartmentDirectory/PublicUtilities/Watermanagement/watersource/historicalwaterlevel.htm>.

[9] Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. “The elements of statistical learning: data mining, inference and prediction.” The Mathematical Intelligencer, 27(2), 83-85. (2005).