Estimating the effect of climate change on global and local sea level rise

Mala Alahmadi
malamer@stanford.edu

Jan Kolmas
kolmas@stanford.edu

December 12, 2015

Abstract

In this CS 229 final project, we apply several machine learning algorithms and error analysis methods to predict the local sea level rise in the San Francisco Bay and global sea level rise based on recorded local and global atmospheric and other conditions between years 1900 and 2014.

1 Introduction

scientists and researchers have not yet arrived at a standard approach in predicting sea level rise due to the many uncertainties and the availability of data. Moreover, the sea level has been rising in more alarming way since the rate of rise has been accelerating in recent years [1]. This aspect of climate change is critical to predict, as almost half of the world's population lives in coastal regions. Knowing how much the levels will rise can inform governments and other interested entities in disaster prevention, real estate evaluation and public safety. In addition, sea level rise varies greatly with geography due to various global processes, such as ocean currents. Estimating the sea level rise in a specific location is the figure of merit, which is of interest of local decision makers. We have selected the San Francisco Bay Area as location of interest to our study, as it is a coastal region with a large population. In this project we predict the amount of sea level rise globally and in San Francisco due to climatic aspects using machine learning algorithms with different climatic inputs such as global CO_2, snow melt and temperature.

2 Related Work

Most researches explored the global sea level rise predictions using yearly data. In a paper published in the *Science* journal, Rahmstorf relates annual global sea level rise to temperature [2]. Furthermore, several researchers have presented similar semi-empirical approach where they simply relate past annual sea level rise to temperature or radiative forcing first then use IPCC projections to extrapolate through the 21st [3]. Such researches yield a wide range of global sea level rise predictions for the year 2100 ranging from 30 cm to 180 cm. In the most recent Intergovernmental Panel on Climate Change (IPCC) report, global mean sea level was predicted by using different models in which isostatic and tectonic effects correction were applied [1], and it yielded a result closer to the lower range of the above sea level rise predictions, that could be the result of not including some important features such as snowmelt. The semi-empirical model would be a practical approach if more features are included rather than just temperature. One study describes a somewhat comprehensive approach that uses global coupled atmosphere-ocean general circulation models in which different components such as temperature, glacial, steric and ice sheet are included to make regional projections of future sea level change [4]. Another study by Yin is along the same lines of the previous study in which in which land, ocean, sea-ice and atmosphere systems are incorporated within the climate model to predict sea level rise on the northeast coast of the United States [5]. Some of the studies mentioned above have incorporated glacial isostatic adjustment (GIA) models [6]. All these climatic models and technical methods are good but substantial work still need to be done in order to create a more reliable model for sea level rise prediction.

3 Data

After an initial research of the data, we anticipated that all the feature and target data could be collected from National Oceanic and Atmospheric Administra-

Table 1: All data collected for the sea level rise analysis. Local features are specific to the San Francisco bay.

	Start	End	Frequency	Source
Local sea level	1898	2012	monthly	University of Hawaii
Global snow balance	1979	2014	monthly	NASA EarthData
Local precipitation	1900	2014	monthly	WRCC
Local temperature	1914	2014	monthly	WRCC
Global temperature index	1880	2014	monthly	NOAA
Global Heat flux	1979	2014	monthly	NASA EarthData
Global CO_2 concentration	1870	2013	yearly	EPI
Global sea level	1880	2012	yearly	NOAA-CSIRO
Global population	1800	2014	yearly	World bank
Local population	1970	2014	yearly	St.Louis federal reserve

Table 2: Number of examples

	Not including satellite data		Including satellite data	
	Training	Testing	Training	Testing
Monthly	1010	178	347	61
Seasonal	337	59	116	20
Yearly	84	15	29	5

tion (NOAA). However, in order to guarantee a sufficient time span, we had to combine data from many different sources. The main data resources besides NOAA were the National Aeronautics and Space Administration (NASA), University of Hawaii Sea Level Center (UHSLC), Earth Policy Institute (EPI), and Western Regional Climate Center (WRCC). The preprocessing that was performed on the data consisted of importing from CSV, text and GRIB formats to MATLAB, bridging over missing values, interpolating and averaging to get monthly, seasonal and yearly data for all features. Finally, the data was normalized by subtracting the mean and dividing by the variance, to conform with the expectation of the learning algorithms see Figure 1.

We had two targets for our analysis. The first is the local sea level rise. Monthly data for San Francisco is available from the University of Hawaii Sea Level Center. Local and global features were both used to estimate the local sea level rise. The second target was the global mean sea level rise, for which the data is available from the NOAA.

Besides improving our model by virtue of having more features, we were interested in looking at which features have the biggest impact on the prediction, and whether the local features have any impact at all. In the final prediction, only a subset of the available features was used.

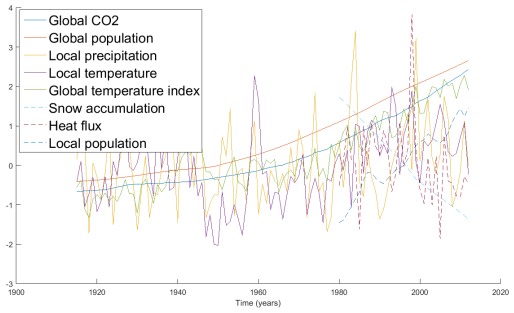

Figure 1: Normalized inputs.

All the data collected, targets and features alike, are presented in Table 1. It could be seen from Table 2 that in order to take into account some of the important features such as snowmelt, which its data was measured by satellite, the number of our training examples reduces significantly making it a huge disadvantage to use such satellite features.

4 Methods

We have started with running linear regression on the data. While this simple algorithm worked well for yearly data (8% average test error - see next section for definition of the average test error), it failed to capture the variations associated with seasonal and

monthly sea level. Therefore, we decided to select the best prediction from three machine learning algorithms: random forest, support vector regression and neural network.

In order to enable prediction for year 2020, having features from 2013 latest, we had to look 7 years back in our modeling. To mitigate this rather arbitrary assumption, we decided to use data from a span s of years from year $Y - s$ to year Y as predictors for the target in year $Y + 7$. For example, the estimate of the sea level rise in 2020 was predicted using data from 2008 through 2013. The span was used as an additional parameter when training the algorithms and reducing testing error.

4.1 Random forest

Random forest is an algorithm that can be used both for classification and regression. It is based on a group of n_{trees} randomly grown decision trees, which output either a class for classification problems or a continuous variable for regression problems. Each tree is created by splitting the feature set randomly into m_{try} subsets based on an attribute value test. This process is then repeated recursively on each derived subset, until the subset at a node has the same value of the target variable. The random aspect of the algorithm should help prevent overfitting. To make a prediction based on a feature set, the set gets fed to each tree in the model, and the output of the algorithm is the average of predictions over all the trees.

In our application, we have used a pre-compiled function [7], taking the training feature matrix, the training target vector and the parameters n_{trees} and m_{try} as inputs. The function output a model of the decision trees grown, which was then used by another pre-compiled function to make a prediction on a testing feature matrix.

4.2 Support vector regression

Support vector regression is an extension of the support vector machine algorithm to continuous problems.

The model produced by support vector machine depends only on a subset of the training data, because the cost function for building the model does not care about training points that lie beyond the margin. Only several data points, called support vectors, constitute the model. Analogously, the model produced by SVR depends only on a subset of the training data, because the cost function for building the model ignores any training data close to the model prediction.

The SVR optimization problem is:

$$\underset{\mathbf{w},\mathbf{b}}{minimize}\, ||\mathbf{w}||^2$$

$$s.t. - C \leq y_i - K(w, x_i) - b \leq C$$

Where C is a cost parameter and $K(x, y)$ is the kernel function. In our analysis, we have used a Gaussian kernel with parameter λ.

$$K(x, y) = exp(-\lambda\, ||x - y||^2)$$

We have used a custom implementation by Ronnie Clark [8], which we have modified.

4.3 Neural network

A neural network for regression is a collection of linear perceptrons, where each neuron is a function, which outputs a weighed linear combination of its inputs. The network is then created by feeding the output of some neurons to the inputs of others. In our case, we have used a two layer feed-forward network. The first layer, called hidden layer contained five neurons, parametrized by weights $\mathbf{w}$ and offset b. The output layer contained a single neuron with equivalent parameters. This output layer took the outputs of the hidden neurons as its inputs, and in turn output the predicted value of the target variable.

In the MATLAB Neural Network Toolbox implementation that we employed, the parameters $\mathbf{w}$ and b of all neurons were randomly initialized and then varied, until the training error was minimized. Having a trained neural network, we could then simply feed the features to the network and obtain a prediction.

5 Analysis

When training the algorithms, we varied the parameters both by computerized iterative search and by manual search. For example, we saw that for neural network layer sizes beyond 5, there was little improvement in testing error, but the computation time was getting longer. Therefore, we have decided to use 5 neurons for our neural network hidden layer.

Our situation was specific in having highly correlated data points, because they came from a time series. Therefore, we could not do cross-validation, but we withheld the latest 15% of the inputs to be a testing matrix, and trained the algorithms only on the first 85%. We have chosen this selection because restricting the training set further, to 70% for example, would yield relatively bad testing errors and prevent us from taking into account the climatic changes in the last decades of the 20th century.

Our figure of merit, measuring the performance of the algorithms was the average percent testing error, which we have defined as follows:

$$\epsilon_{test} = 100\frac{1}{n}\sum\left|\frac{y_{ref} - y_{predict}}{y_{ref}}\right|$$

where n is the number of training samples, y_{ref} is the reference target vector and $y_{predict}$ is the prediction given by the trained model.

We have performed feature selection by manually removing individual features to see, which ones have the most influence on the testing error. Features, that did not change or even exacerbated the testing error were excluded.

The random forest algorithm never predicted a higher sea level than the highest value in the training set. This is logical, since the algorithm takes an average of predictions of individual trees, which can only go as high as the highest prediction that it has seen.

Initially, the random forest was overfitting, which was solved by setting the number of trees to 500. The neural network was overfitting as well, but this behavior was mitigated by specifying a validation fraction. The training algorithm from the MATLAB Neural Network toolbox sets this fraction of the training set aside as validation set, and when the error relative to this validation set stops improving, it stops the training and returns the model.

The testing error varied greatly depending on the frequency of the data. SVR and neural network were able to capture the periodicity of seasonal and monthly data, but random forest could not. The performances of the three algorithms on two scopes (local and global) and using three data frequencies (monthly, seasonal, yearly) are plotted in Table 3.

The best predictions were then selected for both local and global scopes. The parameters and performance of these algorithms is seen in Table 4, and the predicted values, compared to the real values, including the prediction to 2020, are plotted in Figures 2 and 3.

6 Conclusion

Based on our analysis, the best performing algorithms were support vector regression for the San Francisco sea level and neural network for the global sea level. As seen in Table 4, the predicted sea level rise from 2013 to 2020 is 12.2mm in San Francisco and 18.2mm globally.

Future extension of this project could follow multiple avenues. It would be interesting to try different algorithms and additional features to see if it is possible to create a better prediction, especially for the local sea level. With advancing time, more emphasis can be put on satellite data such as heat flux and snowmelt, especially for global predictions, because the reason for not using it was the lack of data points, as the measurements started in 1979.

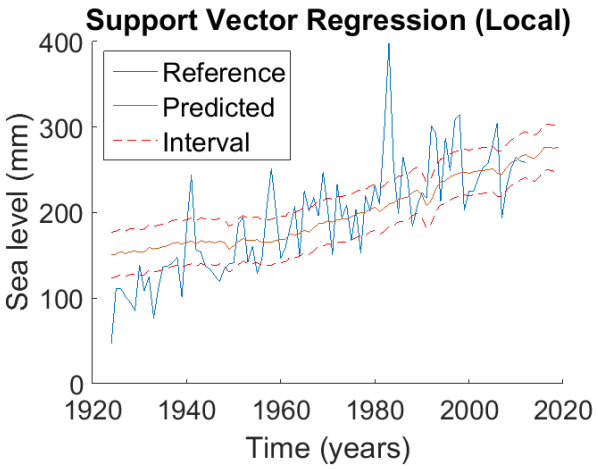

Figure 2: Prediction of sea level rise in San Francisco.

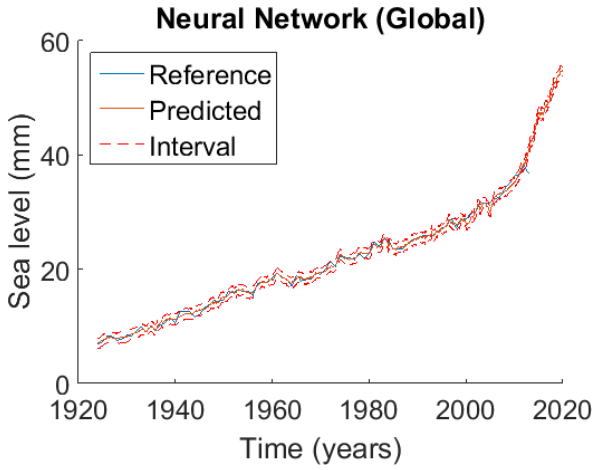

Figure 3: Prediction of global sea level rise.

Table 3: Percent test error

		Random forest	SVR	Neural network
Local	Monthly	20.4	27.2	30.4
	Seasonal	13.6	16.1	22.3
	Yearly	8.16	7.80	15.8
Global	Monthly	14.4	11.1	2.09
	Seasonal	14.8	6.63	10.0
	Yearly	14.7	2.66	5.46

Table 4: Best predictions

Scope	Local (San Francisco)	Global
Frequency	yearly	monthly
Algorithm	SVR	Neural network
Average percent error	7.8%	2.09%
Features	Global CO_2	Global CO_2
	Global population	Global population
	Local precipitation	Global temperature index
	Local temperature	
Parameters	$s = 3$	$s = 3$
	$\lambda = 0.002$	
	$C = 80$	
Sea level rise prediction from 2013 to 2020	$\mathbf{+12.2 \pm 26.7mm}$	$\mathbf{+18.2 \pm 0.9mm}$

References

[1] M. Masson-Delmotte, V., a. Schulz, J. Abe-Ouchi, a. Beer, J.F. Ganopolski, E. González Rouco, K. Jansen, J. Lambeck, T. Luterbacher, T. Naish, B. Osborn, T. Otto-Bliesner, R. Quinn, M. Ramesh, X. Shao Rojas, and a. Timmermann. Information from Paleoclimate Archives. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, pages 383–464, 2013.

[2] Stefan Rahmstorf. A semi-empirical approach to projecting future sea-level rise. *Science*, 315(5810):368–370, 2007.

[3] Robert J. Nicholls and Anny Cazenave. Sea-level rise and its impact on coastal zones. *Science*, 328(5985):1517–1520, 2010.

[4] A.B.A. Slangen, C.A. Katsman, R.S.W. van de Wal, L.L.A. Vermeersen, and R.E.M. Riva. Towards regional projections of twenty-first century sea-level change based on ipcc sres scenarios. *Climate Dynamics*, 38(5-6):1191–1209, 2012.

[5] Jianjun Yin, Michael E Schlesinger, and Ronald J Stouffer. Model projections of rapid sea-level rise on the northeast coast of the United States. *Nature Geosci*, 2(4):262–266, apr 2009.

[6] John A. Church and Neil J. White. A 20th century acceleration in global sea-level rise. *Geophysical Research Letters*, 33(1):n/a–n/a, 2006. L01602.

[7] abhirana. Google Code. https://code.google.com/p/randomforest-matlab/.

[8] Ronnie Clark. Matlab Central. http://www.mathworks.com/matlabcentral/fileexchange/43429-support-vector-regression.