# Helping the Red Cross predict flooding in Togo

**Abstract- This project aims to help the Red Cross predict flooding occurrences in Togo due to overflow in the Nangbeto Dam. Flooding is a result of both high flow rate and the water level in the dam at any point in time. This project focuses specifically on predicting the flow rate in the dam using precipitation data from eight locations around the country. A Lasso model and cross validation were employed to evaluate the significance of the predictors and capture the variance of flow rate.**

## I. Background and Significance

Every year, the Red Cross spends time and money to help populations affected by floods in Togo. The unpredictable nature of flooding makes it difficult for the Red Cross to identify when and where to distribute resources prior to the flood event. As a result, even more time and money is spent to distribute the resources after infrastructure is impacted (e.g. roads closures). This project is a partnership with the Red Cross to help the organization predict flooding in the Nangbeto Dam, which is a result of both high flow rate and a high water level. While the latter can easily be measured, the flow rate depends on many variables. Using time-series precipitation data from eight locations around the country [1], a Lasso model was employed with cross validation to predict the flow rate of the dam. Previously, models have had little success in predicting flow rates in the dam.

## II. Methods

### A. Data Collection

Nine data files were provided by the Red Cross. The first eight contained daily precipitation measurements [units: $mm$] from 08/13/2012 to 12/31/2015 for the following locations: Aledjo, Anie Mono, Blitta, Goubi, Nangbeto, Sokode, Sotoboua, and Tchamba. The ninth file contained measurements of the flow rate at the Nangbeto Dam [units: $m^3/s$] from 08/13/2012 to 12/31/2015.


**Fig. 1** *Map of Togo with data recording sites marked [2].*

### B. Data Visualization

In order to visualize the data, a plot of precipitation and flow rate at each location on each branch of the river was created. The results are shown in Figure 2, and two main things were observed. There is a clear seasonality pattern in both precipitation and flow rate and peaks always appear in the months of August and September. Additionally, these seasonality patterns are consistent across the 8 locations.

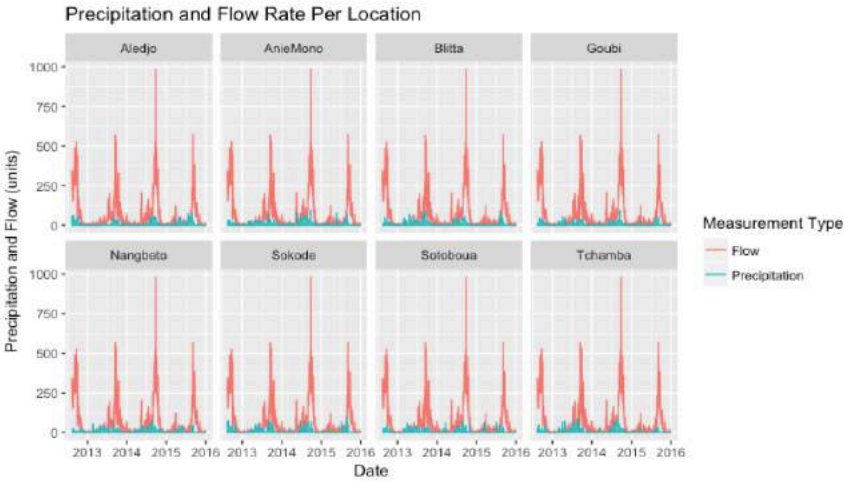


**Fig. 2** *Precipitation at each location and Flow Rate at the Nangbeto Dam over time.*

**C. Variable Creation**

The data provided included precipitation at each location, flow rate at the dam, and dates. Additional variables were created in each of the data sets to represent potential influential factors for flow rate. Given time series data, previous days' values for precipitation and flow rate were used as predictors to inform the model's predictions. Hence, the following variables were created:

- Location variable
- Lag variables: 14 lags for the precipitation of each of the 14 preceding days, and 1 Lag for flow rate of the previous day
- Date variables: a season binary variable for the months of Aug-Nov (1=Aug-Nov, 0=otherwise), a month variable (1-12), and a September binary variable (1=Sept, 0=otherwise)

After these variables were created, the precipitation data sets were joined together by rows in long format, and the resulting data set was then joined together by columns to the flow rate data set.

**D. Analytic Methods**

The lasso model was selected because it succeeds with correlated predictors and offers an interpretable model. Correlation has been observed in the data. The precipitation at each location along the river is correlated both spatially and temporally because weather occurs over expansive area that would include multiple precipitation recording sites. The shrinkage method of lasso regression is appropriate as there are many predictors within the dataset. The main goals of the work were to create an effective and understandable model and lasso regression allows for ease of interpretability of the factors, or predictors, that play a role in the model [3].

Two approaches were used, *Approach 1* consisted of training on the entire data set and *Approach 2* consisted of implementing a training set (years 2012-2014) and a test set (year 2015). Note that because the data is a highly correlated time-series, it was necessary to split the train and test set by years to retain the order of the dates and information. As it was only 3 years worth of data that was provided, Approach 1 utilizes all of the data provided to fit the model. However, the only metric available to evaluate the performance of Approach 1 is the training error rate because it trains on the entire dataset. With this, it is also possible that the model may be overfit.

To get a more representative depiction of the model's predictive performance, the model that implements a training and a testing set is essential. Approach 2 provides a more accurate assessment of the lasso model's performance on an "unknown" data set.

Various lasso models were trained using each approach and implementing various representations of the date. Initially these models contained 14 precipitation lag variables, 8 location variables, and 12 month variables. These produced similar results to the final selected model. In favor of a more interpretable and parsimonious model, variables with near-zero coefficients were dropped (e.g. lag8-lag14) and created a season variable to more concisely represent the month variables with the four largest coefficients (e.g. months Aug-Nov). As September had a large coefficient, the effect of the addition of this variable was considered through comparing the results of models trained with and without this predictor. The final model selected yielded the lowest MSE and highest $R^2$.

## III. Results

*Model 1* was achieved using Approach 1 and gave an $R^2$ of 0.8787 and a training MSE of 1520.8. Approach 2 created *Model 2* and gave a test $R^2$ value of 0.8802 and a test MSE of 734.5.

Based its highest $R^2$, lowest MSE and model parsimony, the final model was chosen to be Model 2. It uses as predictors the 7 precipitation lag variables ($lag_{p0}$ to $lag_{p7}$), the flow rate of

the previous day ($lag_{f1}$), the binary season variable for the months of August through November ($season$), and the binary September variable ($Sept$).

**Fitted Model 1**:

$$\widehat{flowrate} = 1.32 + 0.0566lag_{p0} + 0.0017lag_{p1} + 0.5lag_{p2} + 0.62lag_{p3} + 0.21lag_{p4} + 0.02lag_{p5} + 0.21lag_{p6} + 0.05lag_{p7} + 0.8lag_{f1} + 8.24season + 41.95Sept$$

**Fitted Model 2**:

$$\widehat{flowrate} = 1.43 + 0.05lag_{p0} + 0.00048lag_{p1} + 0.5lag_{p2} + 0.62lag_{p3} + 0.2lag_{p4} + 0.017lag_{p5} + 0.2lag_{p6} + 0.0456lag_{p7} + 0.8lag_{f1} + 8.16season + 41.86Sept$$

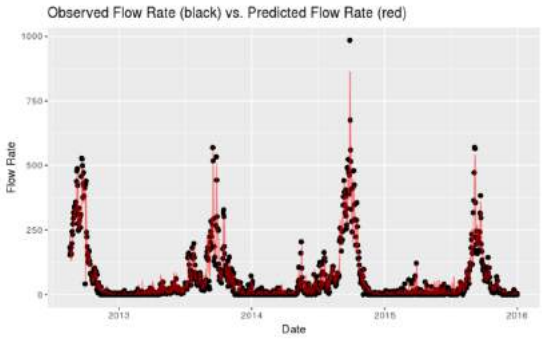


**Fig. 3 Fitted Model 1**

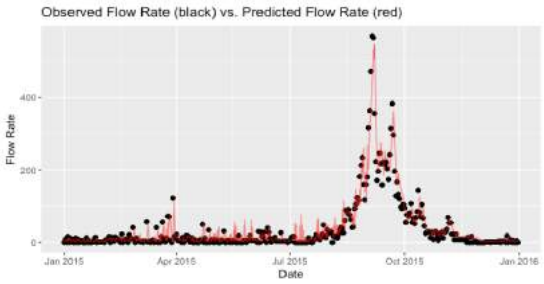


**Fig. 4 Fitted Model 2**

## IV. Discussion/Conclusions

The chosen model, used by implementing cross validation on Lasso with a validation set approach, displayed good predictive power for the year 2015 with an $R^2$ value of 0.8802 and a MSE of 734.5. In order to further test the model, data from 2016 (if available) could be used as another test set. This would further demonstrate the model's predictive power. With constant changes in the climate, the model will likely need to be updated in the future to maintain its predictive power.

As the model is intended to be applied by non-statisticians, the Lasso model was selected for its ease of interpretation and its ability to perform variable selection when many predictors are available. Regularization methods, including the Lasso, trade some of their predictive power to achieve this level of interpretability. For example, ARIMA models may present a worthwhile pathway to explore if higher-resolution data becomes available as ARIMA models could enable the model to predict for the flow rate of the river farther into the future than the Lasso model presented here. The ultimate goal of the work is to predict flooding. In this approach, the Lasso model created is able to predict the flow rate of the water. It may be valuable to explore models which trade better overall predictions for accuracy in the prediction of values of interest (high flow rates, as these correspond with flooding conditions).

## V. References

[1] “World Bank Projects.” World Bank Projects | Aquaveo.com.

[2] Google Maps. (2017). Red Cross Data: Precipitation and Flow Rate. Retrieved from https://tinyurl.com/TogoMap

[3] James, Gareth, et al. An Introduction to Statistical Learning with Applications in R. Springer, 2017.