Rachel Ha (z5161472)
Eu Shaun Lim (z5156345)

Introduction and Motivation

Pakistan is located on a great landmass north of the Tropic of Cancer. Four seasons happen throughout the year, which are the cool, dry winter from December to February to hot, dry spring from March to May, rainy season or the Southwest Monsoon period which occurs in the summer from June to September and the retreating monsoon period happening from October to November. The heavy rainfall that occurs during the Southwest Monsoon can cause extreme flooding and this leads to drowning or ingestion of toxic water from the flood, which in turn causes deadly diseases, such as malaria, jaundice, gastro intestinal infections like typhoid and cholera (Pakistan Today 2018). One such massive flood happened in 2010, where nearly all of Pakistan was caught in heavy monsoon rain, affecting approximately 18 million people with a death toll of nearly 2,000 (Scott 2011). Therefore, the objective of this project is to improve the given rainfall forecast models in Pakistan, with the hope that flood detection will be improved and thus more lives will be saved.

Literature Review

Many studies have been done on rainfall forecasting using various methodologies. One such study focuses on using the global sea-surface temperature (SST) and mean sea-level pressure (SLP) to predict the summer rainfall in Pakistan using Multiple Linear Regression (MLR) and Principal Component Regression (PCR) (Adnan et al. 2017). Observation data from 1961 to 2004 were used as the training set, while data from 2005 to 2014 were the testing sets. Using MLR, the initial predictors for the model are the regions in Pakistan over the sea as the ocean has a much higher heat capacity than land. Stepwise regression is used while selecting predictors to ensure that multicollinearity does not happen and only predictors that are 5% significant were chosen in the final model. For PCR, the predictors are transformed using Principal Component Analysis (PCA) which removes multicollinearity in the predictors. PCA is combined with regression models using the training data to create a PCR model. It was found that both SLP and SST are significant in predicting rainfall in Pakistan, and that both MLR and PCR accurately captured the rainfall patterns. However, PCR outperformed MLR in the various evaluation methods such as correlation coefficient, root mean square error (RMSE), mean absolute error (MAE) and mean bias. More importantly, MLR made prediction errors for anomaly rainfalls which PCR predicted fairly well.

Chai et al. (2017) utilises Artificial Neural Networks (ANN) to predict the rainfall in Kuching, Malaysia. The observed data is collected from 2009 to 2013 and precipitation is classified into light, moderate, heavy and very heavy. Back Propagation Neural Network (BPNN) and Radial Basis Function Neural Network (RBFN) were used to forecast rainfall. BPNN produced inconsistent results compared to RBFN, hence more training and validation are required for BPNN in order to generate a more accurate result. Chai et al. (2017) suggests that the inconsistent results observed in BPNN are due to random assignment of the weights in the network as compared to RBFN which involves the adjustment of "spread" of the network in the single-hidden-layer experiment. BPNN was more sensitive with the different number of hidden neurons used compared to the RBFN model. BPNN achieved a higher accuracy than RBFN with the use of 10 hidden neurons, but RBFN model can be trained faster than BPN model. By comparing the mean squared error (MSE) for both models, BPNN with 10 hidden neurons has a lower MSE than RBFN, indicating that BPNN performs better than RBFN.

Nair et al. (2017) also uses ANN to predict the summer monsoon rainfall over the Indian region, but instead of looking at observed data, they used forecasted data from Global Climate Models (GCM). Predictors for the ANN are the precipitation variables from the GCMs, and the target is the observed rainfall. BPNN is used with one hidden layer in the structure, and double-cross-validation is employed to find the optimal number of hidden neurons and prevent overfitting. Pre-processing was done by randomizing the data to avoid selection bias before applying Min-Max transformation. During each iteration of development, RMSE and MAE is evaluated while increasing hidden neurons. Nair et al. (2017) noticed that after a certain point, RMSE and MAE in the testing data increases while in the training data the errors are unchanged. The best ANN model is chosen based on the smallest RMSE and MAE values. It is then evaluated by comparing against the ensemble mean of each member of the GCMs. It was found that the ANN model was very efficient and predicted rainfall anomalies quite accurately. In addition, it also predicted the rainfall spread close to the observed data during the monsoon months.

Software and Data Description

The software that we intend to use are R/RStudio and Python/Jupyter Notebook. The data consists of the observed rainfall data and the forecasted rainfall data from various forecasting models. The observed data spans from January 2019 to March 2019 while the forecasted data contains data from the last 4 months (February 2019 - June 2019). There are 84 stations across Pakistan that record these data independently at 3 hour intervals. There is a dictionary that contains the latitude, longitude and height above sea level for each of the stations. For standardization purposes, the date and time are recorded in Coordinated Universal Time (UTC). The observation data is stored in daily and hourly formats, where the hourly data is stored at 3 hour intervals starting at 0000 UTC (5am PKT).

Activities and Schedule

Week 1-3	Choose project and complete proposal.
Week 4	Pre-processing phase: - Understanding the data via data visualization - Analyze location of each station and corresponding data - Handling missing values, normalization of data - Choosing models: ANN, Random Forest, Support Vector Machine, Logistic Regression
Week 5-7	Run and evaluate models used. - Start with daily observation datasets - Move on to hourly / forecasted once familiar with data - Evaluate using RMSE, MAE - Hyperparameter tuning via cross validation
Week 8-10	Communicate findings via report / presentation slides

References

Adnan, M, Rehman, N, Ali, S, Mehmood, S, Mir, KA, Khan, AA and Khalid, B 2017, ‘Prediction of summer rainfall in Pakistan from global sea-surface temperature and sea-level pressure’, *Weather*, vol. 72, no. 3, pp. 76-84.

Chai, S, Wong, W and Goh, K 2017, 'Rainfall Classification for Flood Prediction Using Meteorology Data of Kuching, Sarawak, Malaysia: Backpropagation vs Radial Basis Function Neural Network', *International Journal of Environmental Science and Development*, vol. 8, no. 5, pp. 385-388.

Nair, A, Singh, G, and Mohanty, UC 2017, ‘Prediction of Monthly Summer Monsoon Rainfall Using Global Climate Models Through Artificial Neural Network Technique’, *Pure and Applied Geophysics*, vol. 175, no. 2018, pp. 403-419.

Pakistan Today 2018, ‘With the arrival of monsoon, threat of diseases increases’, viewed 20 June 2019, https://www.pakistantoday.com.pk/2018/07/15/with-the-arrival-of-monsoon-threat-of-diseases-increases/.

Scott, M 2011, ‘Heavy Rains and Dry Lands Don’t Mix: Reflections on the 2010 Pakistan Flood’, *Earth Observatory*, viewed 21 June 2019, https://earthobservatory.nasa.gov/features/PakistanFloods.