2.3 Statistical Analyses

Seasonal values of MEI and DMI were used as the predictors to classify seasons into the three rainfall classes. The total data set is divided into 75 % for training the model and 25 % for testing model performance. Quadratic discriminant analysis (QDA) and classification trees were selected for the analyses. A random forest model also was applied to investigate the reliability of a cross-validated statistical forecast tool based on an advance estimate of MEI and DMI. We used R programming language to carry out the statistical analyses. R packages: caret, tree, randomForest, fitdistriplus, devtools and quantreg are used for the studies.

2.3.1 Quadtratic Districriminant Analysis (QDA)

The mathematical formulation of QDA can be derived from Bayes theorem assuming that observations from each class are drawn from a Gaussian distribution ((James, Witten, Hastie, & Tibshirani, 2013; Löwe, Madsen, & McSharry, 2016).
The prior probability π_k represents the randomly chosen observation coming from kth class with density function$f_k(x)$. Bayes theorem states that

$$Pr(Y = k|X = x) = -\frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)} \qquad \text{Eq.(1)}$$

In Eq (1), the posterior probability $Pr(Y = k|X = x)$ indicates that observation $X = x$ belongs to the kth class. For p predictors, the multivariate Gaussian distribution density function is defined for every class k (Eq.(2)).

$$f_k(x) = -\frac{1}{(2\pi)^{p/2}|\sum_k|^{1/2}}\ exp\left(-\frac{1}{2}(x-\mu_k)^T\sum_k^{-1}(x-\mu_k)\right) \qquad \text{Eq.(2)}$$

In Eq.(2), $\sum_k$ is the covariance matrix and μ_x is the mean vector. The covariance matrix ($\sum_k$) and mean (μ_x) for each class are estimated from the training data set (Eq.(3), Eq.(4)).

$$\mu_k = -\frac{1}{N_k}\sum_{i:y_i=k} x_i \qquad \text{Eq.(3)}$$

$$\sum_k = -\frac{1}{(N_k-1)}\sum_{i:y_i=k}(x_i-\mu_k)^T(x_i-\mu_k) \qquad \text{Eq.(4)}$$

Substituting a Gaussian density function for the k[th] class (Eq.(2) into Bayes theorem and taking the log values, the quadratic discriminant function is derived (Eq.(5)). Prior probabilities for class k (π_k) is calculated by the frequency of data points of class k in the training data (Eq.(6)). For a total number of N points in the training observations, N_k is the number of observations belong to kth class.

$$\delta_k(x) = -\frac{1}{2}\,(x-\mu_x)^T \sum_k^{-1}(x-\mu_x\,)\; + log\;\pi_k \qquad \text{Eq.(5)}$$

$$\pi_k = -\frac{N_k}{N} \qquad \text{Eq.(6)}$$

Covariance, mean and prior probability values are inserted into the discriminant function ($\delta_k(x)$) together with the state variables (Eq.(5)). The corresponding class is selected according to the largest value of the function. The number of parameters to be estimated for the QDA model for k classes and p predictors is $k.p.(p+1)\,/\,2$. For this study, the QDA model output is the probability that an observation of a climate category will fall into each of the rainfall classes.

2.3.2 Classification Tree model

For the classification tree model the predictor space is divided into non-overlapping regions ($R_1..R_j$). A classification tree predicts each observation as belonging to the most commonly occurring class of the training data regions (James et.al., 2013). Recursive binary splitting is used to grow the classification tree.

Classification error rate, Gini index and cross-entropy are typically used to evaluate the quality of particular split (James et.al., 2013), and in our study we used the first two indices. Classification error rate (E) gives fraction of observation that do not belong to the most commonly occurring class of the training data regions (Eq.(7)). However, for the tree-growing, the Gini index (G) is considered as the criterion for splitting into regions (Eq.(8))

$$E = 1 - max_k(\hat{p}_{mk}) \qquad \text{Eq.(7)}$$

$$G = \sum_{k=1}^{K} \hat{p}_{mk}\,(1-\hat{p}_{mk}) \qquad \text{Eq.(8)}$$

In Eq.(8) and Eq.(8) , $\hat{p}_{mk}$ represents the fraction of observations in the m^{th} class that belong to the k^{th} class. The Gini index is considered as a measure of node purity of the tree model, since small values of the index indicate that node has a higher number of observations from a single class.

The complexity of the trees are adjusted using a pruning process to produce more interpretable results. Complex trees reduces training error by overfitting the training data. Simple trees can be interpreted well, however, selecting a model which can find the pattern of data is important. In order to achieve the low classification error (training error + testing error), pruning technique is used. First, grow the very large tree, and sub tree is obtained by removing the weak links of the tree. Using tuning parameter to examine the trade-off between complexity of tree and the training error, and defining minimum samples for a node, maximum depth of the tree, and maximum number of terminal nodes are some of the pruning methods (Analytical Vidhya Team, 2016). For this study, we defined the maximum number of nodes to obtain the simple tree (pruned tree).

Tree models give the probability that an observation falls into each of the three rainfall classes. The predicted class is assigned based on the highest probability. Tree models handle ties of probability values by randomly assigning the class.

2.3.3 Random Forest

A random forest is an ensemble learning method used for classification and regression problems. The method is based on a multitude of decision trees based on training data with the final model as the mean of the ensemble (Breiman, 2001). Individual trees are built on a random sample of the training data with several predictors from the total number of predictors. Individual trees are built from the bootstrapped training data set.

There are some features, which can be tuned to make the better performed random forest model. Maximum number of predictors from the total predictors for individual trees, maximum number of trees, maximum node size of the trees and minimum sample leaf size are some of these features (Analytical Vidhya Team, 2015). In our study, we use the maximum number of trees as the main tuning parameters.

In a random forest model the importance of the variable is measured as the decrease in node impurity from the splits over the variable. This value is calculated by averaging the Gini index over the multitude of trees with a larger value indicating high importance of the predictor (James et.al., 2013).

References:

Analytical Vidhya Team. (2015). Tunning the parameters of your Random Forest model. Retrieved March 12, 2018, from https://www.analyticsvidhya.com/blog/2015/06/tuning-random-forest-model/

Analytical Vidhya Team. (2016). A Complete Tutorial on Tree Based Modeling from Scratch. Retrieved March 12, 2018, from https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/

Breiman, L. (2001). Randomforest2001. *Machine Learning, 45(1)*, 5–32. https://doi.org/10.1017/CBO9781107415324.004

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *Springer Texts in Statistics An Introduction to Statistical Learning - with Applications in R*. https://doi.org/10.1007/978-1-4614-7138-7

Löwe, R., Madsen, H., & McSharry, P. (2016). Objective classification of rainfall in northern Europe for online operation of urban water systems based on clustering techniques. *Water (Switzerland)*, *8*(3). https://doi.org/10.3390/w8030087