METEOROLOGICAL DROUGHT FORECASTING USING MACHINE LEARNING TECHNIQUES IN EAST AFRICA

Ali Benard (bernard.ali@aims.ac.rw)
African Institute for Mathematical Sciences (AIMS) Rwanda

Supervised by: Prof. Ernest Fokoué
Rochester Institute of Technology, New York, USA.

May 2020

Submitted in partial fulfilment of the requirements of a Master of Science in Mathematical Sciences at AIMS Rwanda

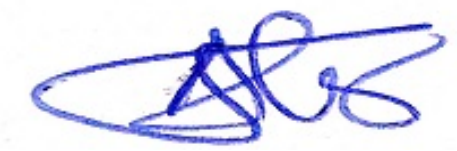
Abstract

There is an increasing frequency of occurrence of drought especially in East Africa and hence has continued to cause negative impacts on lives and livelihoods. For example, the 2010 – 2011 drought in East Africa caused a severe food crisis, documented to have affected approximately 12 million people. Consequently, there is an ever-increasing need for ex-ante early warning systems with the capacity to provide accurate drought forecasts with a lead time. In this study, standardized potential evapotranspiration index (SPEI) was chosen to characterize drought, hierarchical clustering algorithm was used to cluster the stations and variable selection using random forest (VSURF) was used to eliminate redundant variables. Then, multilayer perceptron (MLP), support vector regression (SVR) and XGBoost machine learning techniques were explored and their validation performance in forecasting 1– month in advance SPEI was compared using R^2, RMSE and MAE. The forecast results indicate that XGBoost models were the best for forecasting SPEI_3 and SPEI_6 values.

Keywords: Drought; Drought forecasting; Standardized potential evapotranspiration index (SPEI); Multilayer perceptron (MLP); Support vector regression (SVR); XGBoost

Declaration

I, the undersigned, hereby declare that the work contained in this research project is my original work, and that any work done by others or by myself previously has been acknowledged and referenced accordingly.



Ali Benard, May 2020

List of Abbreviations

CAFEC	Climatically appropriate for existing conditions
CRU	Climate research unit
MAE	Mean absolute error
Max	Maximum
Min	Minimum
ML	Machine learning
MLP	Multi-layer perceptron
NOAA	National oceanic and atmospheric administration
PET	Potential evapotranspiration
RF	Random forest
RMSE	Root mean square error
SPEI	Standardized potential evapotranspiration index
$SPEI_o$	Observed standardized potential evapotranspiration index
$SPEI_p$	Predicted standardized potential evapotranspiration index
$SPEI_3$	3-month timescale standardized potential evapotranspiration index
$SPEI_6$	6-month timescale standardized potential evapotranspiration index
SPI	Standardized precipitation index
SVM	Support vector machine
SVR	Support vector regression
VSURF	Variable selection using random forest
XGBoost	Extreme gradient boosting

Contents

1. Introduction

1.1 Background

Drought is a complex and devastating natural disaster that causes severe social, environmental and economic losses. It is a recurring feature of almost every climatic zone. From previous studies, it is difficult to provide a definitive definition of drought since it seems to vary from study to study, organization to organization and country to country. Most studies have defined drought in terms of the amount of evapotranspiration relative to that of precipitation. Thus, Thornthwaite (1931) has defined drought as a condition in which the amount of water required for transpiration and direct evaporation exceeds the available soil moisture. According to National Oceanic and Atmospheric Administration (NOAA), drought can be caused not only by lack of precipitation and high temperatures but also by overuse and overpopulation. According to the Task Force on Drought Prone Area Programme (DPAP, 1973), the areas which receive rainfall less than 750 mm per annum are classified as drought-prone and those which receive rainfall in the range 750 mm to 800 mm are vulnerable to drought.

Drought can be classified into four main types according to their impact, namely agricultural drought, meteorological drought, hydrological drought and social-economic drought but in most cases these occur together (Mishra and Singh, 2010). Therefore, stream-flow for the hydrological drought, soil-moisture and vegetation for the agricultural drought, evapotranspiration and precipitation for the meteorological drought, and the effects on people's livelihoods for the socio-economic drought. Droughts have very great impacts on agriculture, the economy and the environment. For example, the drought of 2008-2010 affected more than 13 million people in Eastern Africa and the 2010-2011 drought was the most extreme such event in 60 years, leading to a severe food crisis and a famine affecting approximately 12 million people in Eastern Africa (AghaKouchak, 2015). And with the global warming affecting climate, extreme events such as drought have occurred more frequently and become more severe in the past two decades (Dai, 2011). Drought frequency in Eastern Africa has doubled from once every six years to once every three years since 2005 (Guha-Sapir et al., 2004; Ayana et al., 2016; Meier et al., 2007). This increased frequency of droughts has made the need to predict them more urgent, to be able to manage and plan the agricultural resources before the drought ravage.

While droughts have always occurred, as part of the natural variability of the climate, which can lead to a deficit in precipitation, which reduces soil moisture and run-off and therefore a lowered stream-flow, or to higher temperatures, resulting in increased evapo-transpiration, which also reduces soil moisture, human activities too, such as reservoir operations, deforestation and other changes in land use, as well as climate change, can cause changes in hydrological processes and contribute to the frequency and severity of droughts. In general, then, droughts are the result of both natural meteorological conditions and of human activities, between which there can be complicated interactions.

It is very important to be able to forecast droughts, their length and severity, and there are three main kinds of methods that have been used to predict the values of different drought indicators: dynamical, statistical and hybrid methods (Mishra and Singh, 2011; Sylla et al., 2013; Pozzi et al., 2013). Dynamic methods uses the greater computational power and better understanding of the climate, to integrate drought prediction into general circulation models (GCMs). Statistical prediction methods employs empirical relationships of previous records and using the influencing factors as predictors. But in the recent past, hybrid prediction methods which use past statistical information to 'train' the models that are used have led to a dramatic improvement in the range and accuracy of predictions that can be made.

1.2 Problem statement

Since 1900, Over 11 million people have died due to drought and more than 2 billion have been affected by drought worldwide. The Intergovernmental Panel on climate change (IPCC) report states that the duration and intensity of droughts have increased and probably the extent of drought in affected areas globally will increase over the next century. Therefore, drought forecasting is of paramount importance to stakeholders (farmers, the government, etc) for early preparedness on mitigation of drought impacts. Application of advanced machine learning techniques to drought prediction, more accurate forecasts are obtained and hence, this study is motivated by required accurate drought prediction to support early preparedness and resilience to droughts and eventually contribute to environmental, social and economic growth in drought-affected areas in East Africa.

1.3 Scope of the study

The focus of this study is to use modern machine learning methods to provide more accurate predictions of drought in East Africa. While the study recognizes the existence of drought globally, it focuses on East Africa and also while it acknowledges the existence of different types of droughts affecting East Africa, it is restricted to only investigating meteorological drought.

1.4 Aims and objectives of the study

The aim of this study is to use advanced machine learning techniques to forecast the meteorological drought in East Africa.

The specific objectives of the study are:

- Analyze the data from different meteorological stations in East Africa to find clusters with related observations.
- Use the meteorological data from each cluster to identify some key determining factors of meteorological drought in the region.
- Develop and compare the prediction performance of three machine learning techniques for each cluster.

1.5 Justification of the study

It is hoped that the findings of this study will fill the gaps in the information, which has been insufficient, and in the forecasting, which has been inaccurate of meteorological drought in East Africa. The results of this study will be useful for decision-makers, including farmers, governments and other stakeholders, to prepare earlier to mitigate the impacts of drought. Also, the results of this study may likely influence further scholarly research by other researchers who may be interested in this field of study.

2. Literature Review

Previously, there are various research studies which have been undertaken on the drought and its prediction using various scientific methods. Therefore, this chapter presents the related work on the drought, its prediction and how previously, the methods used in this study have been employed.

2.1 Drought

Although drought has been classified as one of the natural disasters, there is no global definition for drought. It is region-specific, showing the differences in climatic characteristics with the inclusion of different physical, social-economic and biological variables, and it is difficult to transfer definitions derived for one region to another (Hadish, 2010). The effect of drought is usually delayed and widely spreads over time and over a larger geographical area than other natural hazards like volcanic eruptions, tornados, earthquakes, etc (Wilhite and Svoboda, 2000). For example, the impact of the agricultural drought is accurately accessed when crops are harvested, this happens after a few months of the drought symptoms (Boken et al., 2005). This has posed a challenge of accurate, timely and reliable estimates of intensity, duration and extent of drought for early preparedness (Ding, 2011).

Drought is generally classified into four categories, namely meteorological, hydrological, agricultural and socio-economic drought. Figure 2.1 shows different categories of drought.

- meteorological drought – precipitation deficiency.
- hydrological drought – surface and/or groundwater deficiency. it associates effects of precipitation shortfalls on surface and groundwater supplies (i.e reservoir, lake levels, streamflow and groundwater).
- agricultural drought – soil water deficiency and it has an impact on food production and vegetation.
- social-economical drought – measures the impact of meteorological, hydrological and agricultural drought on the supply and demand expressed as economical value.

Drought may start as meteorological drought, thereafter become agricultural drought and if there is a deficiency in water storage it becomes hydrological drought (Fleig et al., 2005). This thesis focuses on meteorological drought, which is defined as precipitation deficit (Hayes et al., 1999; Fleig et al., 2005).

2.2 Drought assessment

Drought assessment is to understand the duration, intensity, extent and causes of drought. Several methods have been developed for drought assessment, quantitatively. Generally, drought has been assessed with reference to duration, truncation levels, nature of water deficits and regionalization approaches (Dracup et al., 1980). However, over the advancement of drought assessment techniques, many indices were developed to detect and monitor droughts. So, according to how drought is defined and classified, researchers have attempted to assess the severity of the drought. These studies were classified as either meteorological, hydrological or agricultural drought assessments. This study focuses on how meteorological drought has been previously assessed.

2.2.1 Meteorological drought assessment. Meteorological drought refers to the deficiency of precipitation at a level which would affect the normal societal living. However, definitions vary depending on the base periods and truncation levels to delineate drought periods and rainfall deficiency respectively. Over

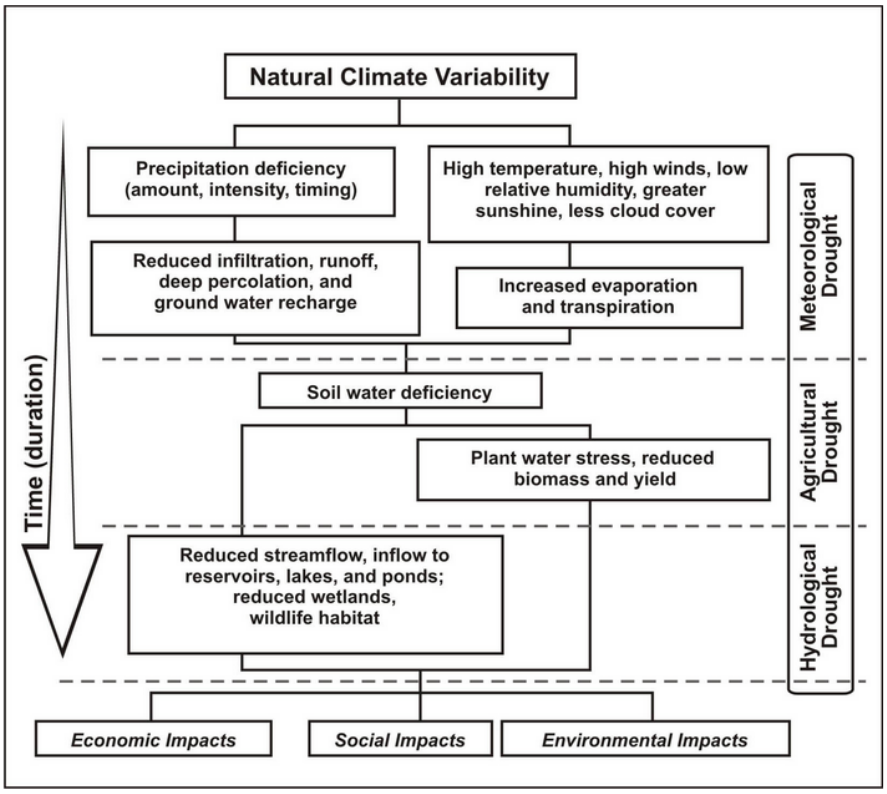


Figure 2.1: Sequence of Impacts of Different categories of Drought,Retrieved 11:08 am, May 8, 2020.

the past few decades, drought indexes based on individual or multiple hydrological and climatic variables have been proposed. Earlier, drought was defined based on daily/monthly precipitation (McGUIRE and Palmer, 1957; Blumenstock, 1942). Later, several drought indexes were developed. The widely used methods for assessing meteorological drought severity index include Aridity index, The deciles, Palmer's Drought Severity Index(PDSI), The Standardized Precipitation Index(SPI) and finally The Standardized Precipitation Evapotranspiration Index(SPEI).

CVS (1969) developed a concept for drought classification using the aridity index. Aridity index is the ratio between annual water deficiency and annual water needed for evapotranspiration expressed as a percentage. According to the Task Force on drought prone areas program(DPAP(1973)), the areas which received rainfall less than 750 mm per annum are classified as prone to drought and those which received rainfall in the range 750 mm to 800 mm are at risk to drought.

Gibbs and Maher (1967) used 'Deciles' to solve the shortcomings of 'per cent of normal' approach. The method is dividing the distribution of the occurrences and a long-term precipitation record into tenths of the distribution. Each category was called 'Decile'.The first decile is the rainfall not exceeded by the first lower 10 per cent of the precipitation occurrences. The second decile is the rainfall not exceeded by the second lower 10 per cent of the precipitation occurrences. The limitation of the decile system is that a long meteorological record is needed to accurately calculate the deciles.

Palmer (1965) developed a comprehensive technique for calculating the drought severity called The Palmer's Drought Severity Index (PDSI). The computations are made with respect to what Palmer termed as the CAFEC values of the hydro-meteorological factors including precipitation, Evapotranspiration and runoff. This method is based on the balancing of water. Palmer used the difference between

actual precipitation and precipitation demand under conditions of an average climate of an area to assess drought severity in space and time. The PDSI is one of the longest-used indexes for drought monitoring (Mishra and Singh, 2010; Dai, 2011; Vicente-Serrano et al., 2011). The shortcoming of PDSI is that it has inherent timescales making it slow to respond to the development and attenuation of drought (Hayes et al., 1999). Also, it cannot be applied uniformly in all agricultural climate zones. In arid and semi-arid zones, it represents hydrological drought whereas, in humid zones, it represents agricultural drought (Ray et al., 2001).

McKee et al. (1993) developed the Standardized Precipitation Index(SPI) to calculate precipitation deficit at different time scales, quantitatively, showing the impact of deficiency on the various water supplies availability. It uses only precipitation data to characterize drought. Earlier studies have shown that areas with a probability of a drought year greater than 20 per cent were classified as prone to drought, areas with the probability of a drought year greater than 40 per cent were classified as prone to chronic drought. However, it does not considers variables related to temperature and since previous studies have already showed that rising temperature is affecting drought. Thus giving this index a shortcoming.

Therefore, to address this problem, the recently developed Standardized Precipitation Evapotranspiration Index (SPEI) (Vicente-Serrano et al., 2010; Beguería et al., 2014) which is an extension of SPI and is designed to account for both precipitation and potential evapotranspiration(PET). It captures the impact of rising temperatures in water demand. SPEI is constructing a water balance equation of supply and demand which is based on the difference between precipitation and evapotranspiration. Evapotranspiration is temperature-based which is calculated using the Thornthwaite equation (Thornthwaite, 1948) but it is underestimated in arid and semi-arid areas and overestimated in humid regions (Jensen et al., 1990). When this happens, the Food Agricultural Organizations of the United Nations (FAO) recommends that it is better to use penman-monteith correction method to calculate potential evapotranspiration. Chen and Sun (2015) showed that SPEI with penman-monteith correction over-weighted Thornthwaite correction in monitoring droughts, specifically in arid regions.

2.3 Meteorological station clustering and Variable selection

In statistics and machine learning, clustering is the task of identifying the set to which a new observation belongs. Clustering is an unsupervised procedure and it involves grouping data points into categories according to their similarity or distance. There are generally two main types of clustering algorithms namely K-means clustering algorithm and Hierarchical clustering. The main difference between the two algorithms is that in K-means clustering since we are starting by choosing the clusters randomly the results produced by multiple running of algorithm might differ but for hierarchical clustering, the results are reproduced. Also, K-means clustering needs prior knowledge of K but in hierarchical clustering, you stop at any number of clusters you find appropriate for interpretation. This study explores both K-means clustering (Hartigan and Wong, 1979) and hierarchical clustering to cluster meteorological stations.

Variable selection is a crucial process for most predictive methods. Most machine learning methods assume that the predictive variables used are not misleading and informative since they treat each predictive variable as all of the equal importance and hence, misleading predictive variables mostly reduce prediction accuracy. Various approaches have been proposed for selecting important variables and have been grouped into three methods namely Generalized linear model (GLM), gradient boosting machines (GBM) and random forest (RF). For GLM, there are several methods available in R (Lumley and Miller, 2009; Chambers and Hastie, 2017) including Anova, bestglm, StepAIC and dropterm. For gradient boosting machine (GBM), variable selection is based on relative influence (Li et al., 2019), they in-

clude important variable based on the predictive accuracy (IVPA) and the unimportant variable based on the predictive accuracy (UVPA). For RF, variable selection methods include Boruta, recursive feature selection (RFS), variable importance (VI), average variable importance (AVI), knowledge informed AVI(KIAVI) and variable selection using random forest (VSURF). This study uses VSURF for variable selection. Genuer et al. (2010) proposed the VSURF and the main advantage of it is that, it is computational intensive in a high dimensional dataset and variable selection based on importance ranking in relation to the response variable.

2.4 Machine learning on drought prediction

Choice and development of an appropriate prediction model is another major challenge in forecasting drought. Previously, researchers attempted to connect data from different sources to reproduce ground-based drought indices by using data-based models. For example, auto-regressive integrated moving average model (ARIMA) (Belayneh et al., 2014) and artificial neural networks (ANN) (Morid et al., 2007). ARIMA is a simple and commonly used method to predict droughts in one station which actually bases on the characteristics of the time series itself and does not account for the effects of other predictors. ANN made greater contributions in fitting the nonlinear relationship between predictors and the target variable in different fields. Cortes and Vapnik (1995) proposed that Support Vector Machine (SVM) which is a popular machine learning tool for classification and regression. The main idea of SVM is to find hyperplane to divide a high-dimensional space into many different classes. In short, provided labeled training data, the algorithm outputs the optimal hyperplane which classifies the new examples. This has been widely used in the prediction of common diseases and in the field of finance. For example, prediction of the case of diabetes and pre-diabetes (Yu et al., 2010) and financial time series forecasting (Okasha, 2014). However, it has less application in the drought prediction. Similarly, Chen and Guestrin (2016) proposed the XGBoost algorithm or Extreme Gradient Boosting which is a successful and powerful tree-based and linear model solver algorithms. It combines the available predictors and trains all the weak into strong learners through the additive training strategy. Its application include disease prediction and diagnosis (Livne et al., 2018) and failures in the banking sector (Carmona et al., 2019) . However, in the field of drought prediction, its strong ability is not yet known.

2.5 Related work on drought prediction

In the recent past, efforts have been made in the drought prediction, for example, using the new machine learning methods like the Support Vector Regression with its concept of structural risk minimization has brought about better drought forecasting, this was discovered in different studies (Belayneh et al., 2014; Khajeh Borj Sefidi and Ghalehnoee, 2016). In remote sensing products (Normalized difference vegetation index(NDVI)) has been applied in predicting drought because of the readily available, consistent temporal and spatial observations at both regional level and global scales (Feng et al., 2019; Oliveira et al., 2014; Asoka and Mishra, 2015). Also, apart from predicting drought based on climatic indicators, other studies have drawn their attention to predicting the impact of drought on the ecosystem (Shafiee-Jood et al., 2014). Other studies have investigated predicting drought by incorporating the human activities which are difficult to quantify like land-use change, deforestation and reservoir operation (Ma et al., 2018; Yuan et al., 2017). Moreover, some studies have really focused on the effects of the drought on the farming sector and most immediate consequences of drought in the fall of crop production (Toulmin, 1986; Olesen et al., 2011). All these efforts have uplifted the drought prediction to a certain level. However, predicting drought is still a big challenge to a climate scientist and other interested groups since drought occurs at different temporal and spatial scales and have a complex origin.

3. Materials and Methods

This research depends on the data from the climate research unit to characterize drought and its prediction. Therefore, this chapter deals with the area of the study, the source of data used, calculation of standardized precipitation and evapotranspiration index (SPEI), clustering of stations and variable selection process. The machine learning techniques required for forecasting drought are all addressed.

3.1 Study area

The East Africa region consists of five countries namely: Kenya, Tanzania, Rwanda, Uganda and Burundi. A large part of East Africa have two distinct rainfall seasons; "short rains" which lasts from October to December(OND) and with "long rains" which last from March to May (MAM). These seasons are connected to the movement of the inter-tropical convergence zone(ITCZ). This region also experienced average surface temperatures increase of between $1°C - 3°C$ in the past 5 decades (Christy et al., 2009). Therefore, this study focuses on meteorological drought forecasting in East Africa. Figure 3.1 shows the map of East Africa.

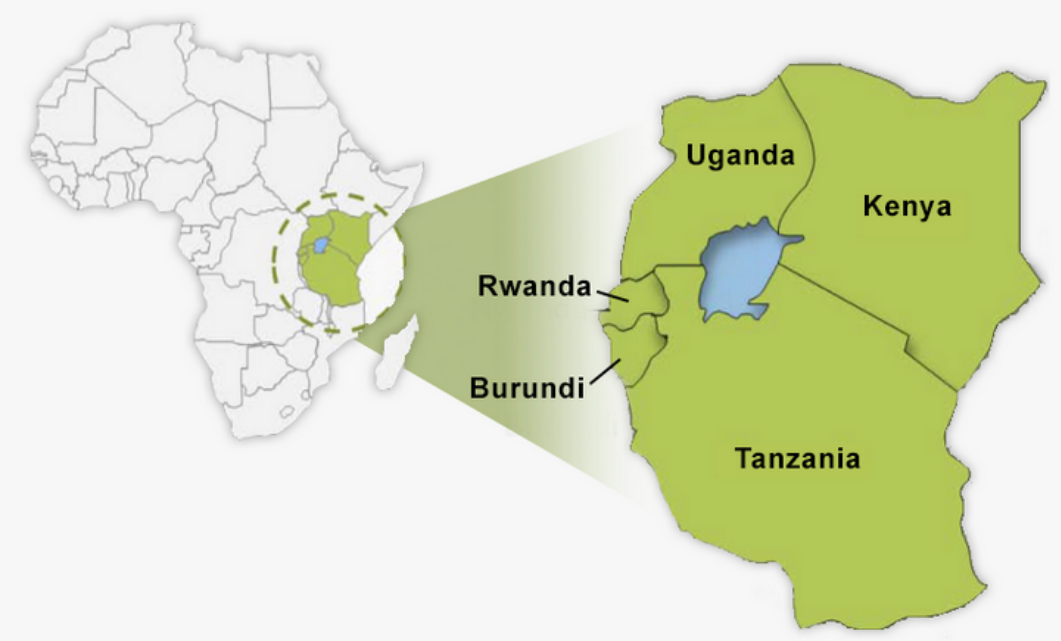


Figure 3.1: The map showing the location of east Africa, Retrieved 22:17 pm, April 20, 2020.

3.2 Data description

This study used the observed climate datasets over East Africa region $(27°E - 40°E, 12°S - 5°N)$. The observed climate data includes the monthly gridded average daily mean temperature $(°C)$, monthly average daily maximum temperature $(°C)$, monthly average diurnal temperature range $(°C)$, precipitation (mm), cloud cover $(\%)$, monthly average daily minimum temperature $(°C)$, number of wet days per month $(days)$ and monthly vapor pressure (hPa). The observed climate data were obtained from

climate research unit dataset (cru-2018) from 1960-2018 (CRU;Mitchell et al. (2004)) with a horizontal resolution of $0.5° \times 0.5°$ degrees.

The data stations were as shown in Table 3.1 and the meaning of abbreviations used as variable names from the data is shown in Table 3.2.

Station name	Latitude (°)	Longitude (°)	Altitude (m)
Mombasa	-4.3	39.8	50
Kigali	-1.8	30.3	1567
Turkana	3.3	35.8	1138
Dodoma	-6.3	35.8	1120
Bujumbura	-3.3	29.3	774
Dar_es_salaam	-6.8	39.3	55
Gitega	-3.3	29.8	1504
Mbarara	-0.8	30.8	1147

Table 3.1: 8 Meteorological observation stations.

Abreviation	Meaning
PRED	precipitation
PET	potential evapotranspiration
CLDD	cloud amount
DTRD	Temperature range
TMPD	Temperature
TMND	minimum temperature
TMXD	maximum temperature
VAPD	vapor pressure
WETD	number of wet days

Table 3.2: Abbreviations and meaning of meteorological variables.

3.3 Study design

The framework for the procedures used in this study is as shown in Figure 3.2.

3.4 Standardized precipitation and evapotranspiration index (SPEI)

All the data was used to calculate the SPEI time series. SPEI was constructed by Vicente-Serrano et al. (2010) and has been widely used in drought analysis. Standardized precipitation and evapotranspiration index (SPEI) characterizes drought by constructing a water balance equation of supply and demand which is based on the difference between precipitation and potential evapotranspiration (PET) (D=precipitation – potential evapotranspiration $= P - PET$). Potential evapotranspiration (PET) shows the amount of evaporation and transpiration when the water supply is enough. A monthly reference for potential evapotranspiration (PET) in the present study was computed using the Thornthwaite

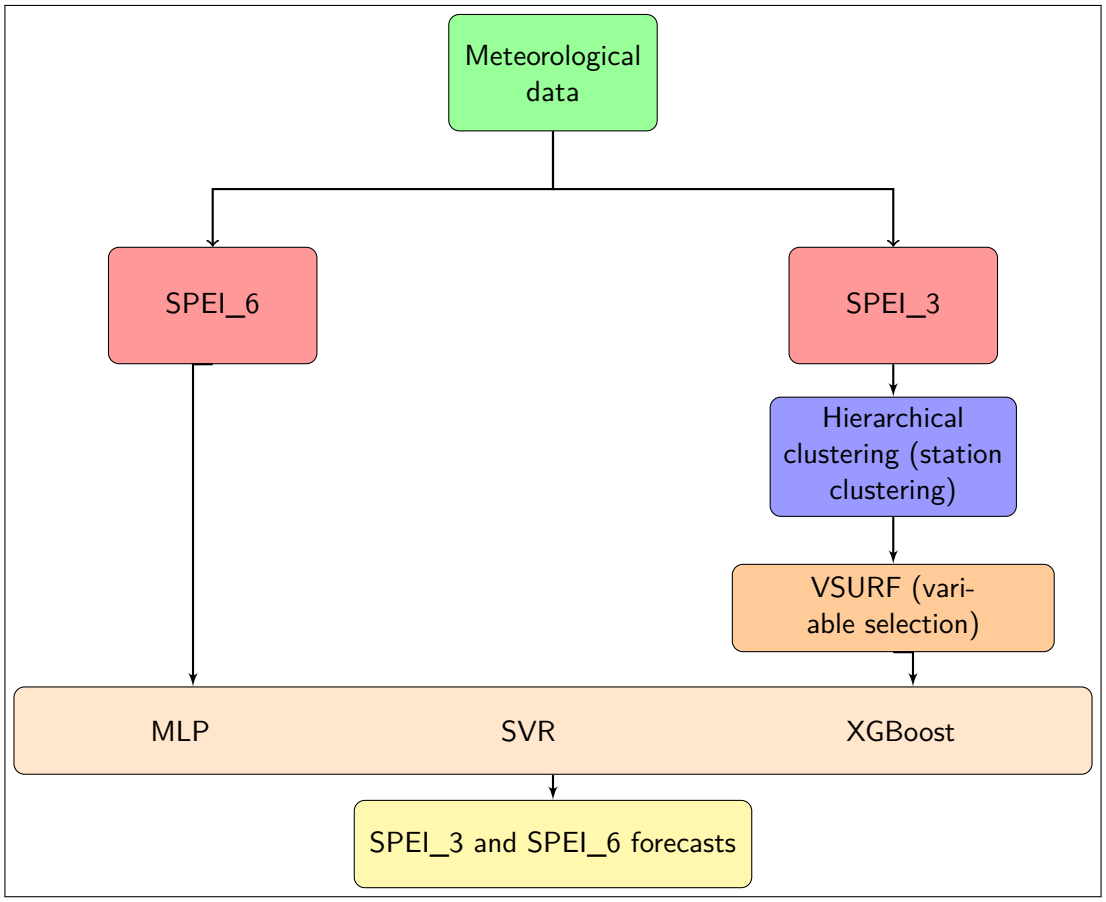


Figure 3.2: The framework of the study.

equation (Thornthwaite, 1948) (mathematics behind this method is presented in Appendix) due to its simplicity and the data constraint.The accumulated water balance in different timescales are constructed as follows:

$$D_n^k = \sum_{i=0}^{k-1} \left(P_{n-i} - PET_{n-i}\right), \qquad n > k, \tag{3.4.1}$$

where n is the calculation frequency and k is the timescale (month). The computation of SPEI needs three parameter distribution and Vicente-Serrano found that the log-logistic distribution is the best since it correlates to the D series. Therefore, log-logistic probability density function is fitted to the sequence as shown in (3.4.2):

$$f(x) = \frac{\beta}{\alpha}\left(\frac{x-\gamma}{\alpha}\right)^{\beta-1}\left[1+\left(\frac{x-\gamma}{\alpha}\right)^{\beta}\right]^{-2}. \tag{3.4.2}$$

Therefore, the probability distribution function of log-logistic distribution of D series is given by:

$$F(x) = \left[1+\left(\frac{x-\gamma}{\alpha}\right)^{\beta}\right]^{-1}, \tag{3.4.3}$$

where β represent the shape, α represent the scale and γ represent the origin parameters for D values between $(\gamma > D < \alpha)$. β, α, and γ are found using the L-moment procedure (Ahmad et al., 1988) as in (3.4.4), (3.4.5) and (3.4.6):

$$\beta = \frac{2w_1 - w_0}{6w_1 - w_0 - 6w_2}, \tag{3.4.4}$$

$$\alpha = \frac{(w_0 - 2w_1)\beta}{\Gamma(1 + \frac{1}{\beta})\Gamma(1 - \frac{1}{\beta})}, \tag{3.4.5}$$

$$\gamma = w_0 - \alpha\Gamma(1 + \frac{1}{\beta})\Gamma(1 - \frac{1}{\beta}), \tag{3.4.6}$$

where Γ is the gamma function of β and w_0, w_1, w_2 can be calculated by probability weighted moments (PWMs) via the L-moment method (Hosking and Wallis, 2005) as shown in (3.4.7):

$$w_i = \frac{1}{n}\sum_{i=1}^{n} x_i \left(1 - \frac{i - 0.35}{n}\right), \tag{3.4.7}$$

where x_i is the ordered random sample of D. The SPEI value can be found as standardized value of $F(x)$ (Yang et al., 2016) as follows:

$$SPEI = W - \frac{c_0 + c_1 W + c_2 W^2}{1 + d_1 W + d_2 W^2 + d_3 W^3}, \tag{3.4.8}$$

$$\text{where, } W = \sqrt{-2ln(P)}, \tag{3.4.9}$$

where P is probability of exceeding a determined D value and $P = 1 - F(x)$ and when P is greater than 0.5, then $P = 1 - P$ and the sign of final SPEI is reversed. And the constants are given by (3.4.10):

$$c_0 = 2.515517, c_1 = 0.802853, c_2 = 0.010328, d_1 = 1.432788, d_2 = 0.189269, d_3 = 0.001308 \tag{3.4.10}$$

Standardized precipitation and evapotranspiration index (SPEI) is a standardized variable with mean 0 and a standard deviation of 1. Therefore, it can be compared at different spaces and times. In addition, it can be calculated at different timescales. For instance, data of the present and past two months can be used to calculate the SPEI_3 for a given month. Other SPEI timescales include SPEI_6, SPEI_12, SPEI_24, etc. Negative and positive values of SPEI represent dry and wet conditions, respectively. The categorization of drought based on SPEI is as shown in Table 3.3, which is recognized and used widely in different regions (Chen and Sun, 2015; Tirivarombo et al., 2018). Moreover, based on specific aims, SPEI can be computed on different timescales, 1- and 6-month SPEI timescales are appropriate for agricultural and meteorological drought while longer timescales are appropriate for hydrological drought. In this study, the 3-month and 6-month timescale SPEIs was used to analyze meteorological drought characteristics. The SPEI was computed using "SPEI" package in R software. Figure 3.3 shows the SPEIs for Dodoma meteorological station where we observe that there is increased frequency of drought occurrences in the last two decades.

Category	Extremely wet	Severely wet	Moderately wet	Normal	Moderately dry	Severely dry	Extremely dry
SPEI	2.00 and above	1.50 –1.99	1.00 – 1.49	-0.99 – 0.99	-1.00 – -1.49	-1.50 – -1.99	-2.00 and below

Table 3.3: Categorization of drought based on SPEI.

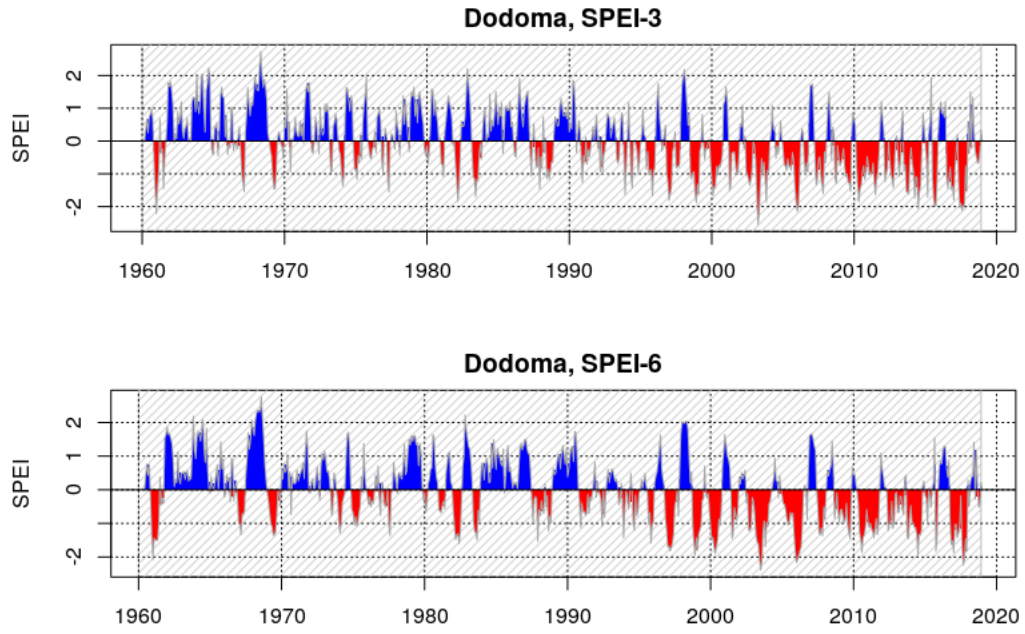


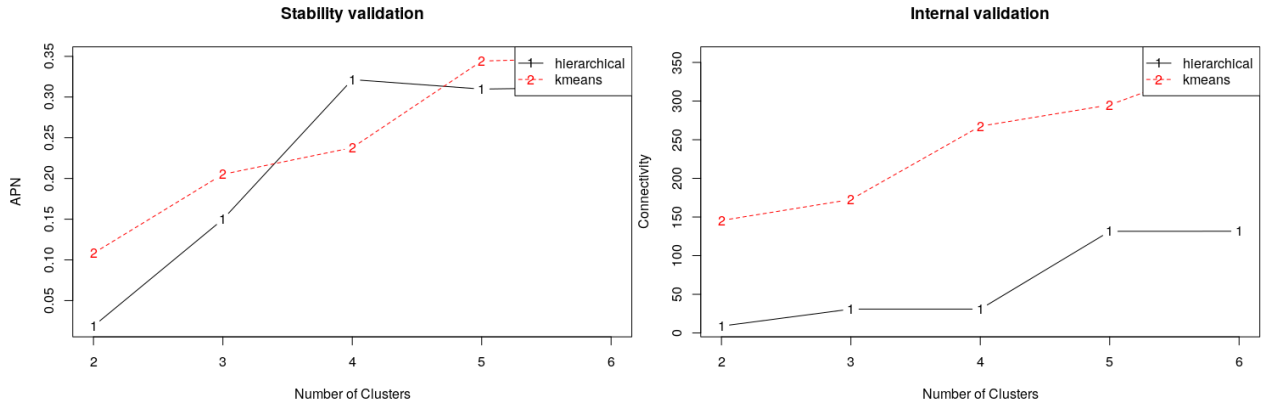
Figure 3.3: The SPEI_3 and SPEI_6 for Dodoma station.

3.5 Meteorological station clustering

East Africa covers an extensive area and hence influenced by different climate patterns. Different meteorological variables maybe having different contributions in the prediction of 3-month and 6-month SPEIs in cold and hot areas or dry and wet areas. Therefore, it is necessary to cluster the stations. There are generally two main types of clustering algorithms namely K-means clustering algorithm and hierarchical clustering. The main difference between the two algorithms is that in K-means clustering since we are starting by choosing the clusters randomly the results produced by multiple running of algorithm might differ but for hierarchical clustering, the results are reproduced. The basic process of hierarchical clustering is that it begins by treating every observation as a separate cluster. It then repeatedly performs the following two steps :(1) identify the two clusters that are closest to each other, and (2) merge the two clusters that are the most similar. This iterative process continues until all clusters have been merged together into one cluster forming a dendrogram. Then the user cuts the tree where it is appropriate.

Also, the main idea of K-means clustering is:

- Randomly choose k groups in the feature plan.
- Group observation by minimizing distance with centroid which yields k groups with n observations.
- Shift the original centroid to the mean of the coordinates within a group.
- Group observation by minimizing the distance according to new centroids. New boundaries are created. Thus, observations will change from one group to another.
- Repeat the process until no observation changes groups.



(a) Stability measure (APN) of cluster validation. (b) Internal measure (connectivity) of cluster validity.

Figure 3.4: Plots from cluster validation.

In this study, based on SPEI_3, cluster validation was done using clvalid package coded in R software for both K-means clustering and hierarchical clustering algorithms, where hierarchical clustering with 2 clusters performs the best, since this pair appears in 2 out of 4 stability measures and in 2 out of 3 internal measures. Figure 3.4 shows validation measures results obtained from clvalid package. Therefore, hierarchical clustering (ward's method) algorithm was used to cluster the meteorological data stations into two groups. Figure 3.5 presents the results attained by hierarchical clustering with 2 clusters.

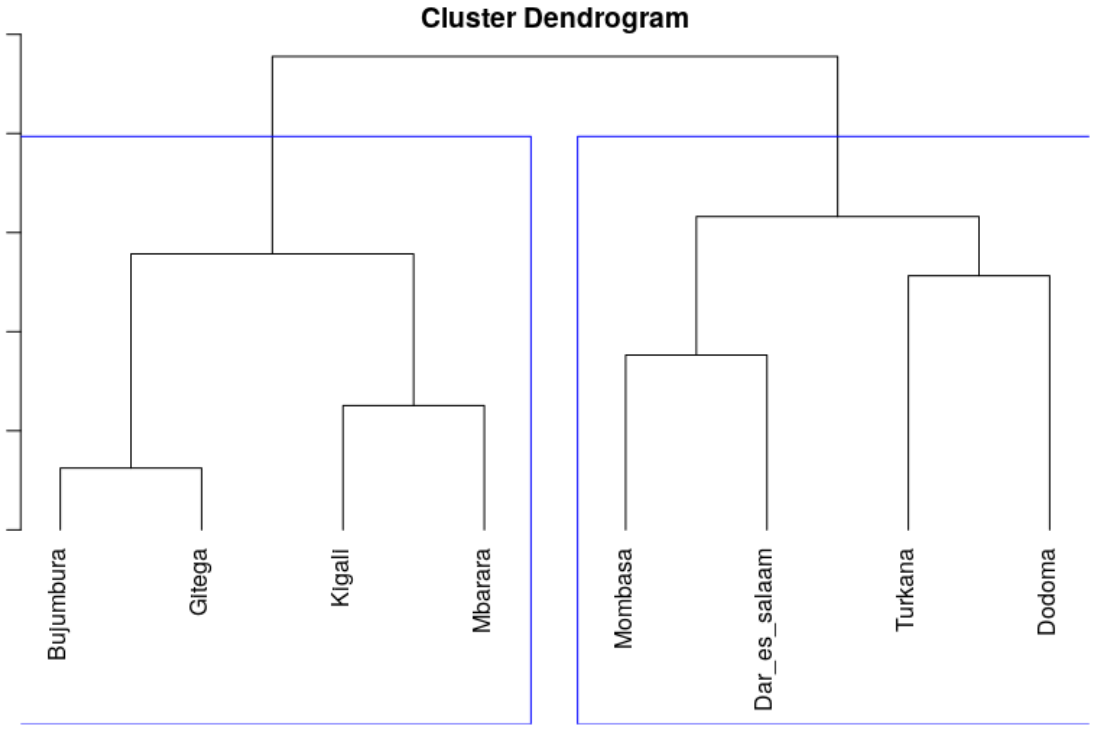


Figure 3.5: Clustering based on hierarchical clustering algorithm.

```
Optimal Scores:
             Score  Method       Clusters
Stability
APN          0.0187 hierarchical 2
AD           2.5989 kmeans       6
ADM          0.1972 hierarchical 2
FOM          0.7368 kmeans       6
Internal
Connectivity 8.8698 hierarchical 2
Dunn         0.1313 hierarchical 3
Silhouette   0.3830 hierarchical 2
```

3.6 Variable selection

From the 9 variables obtained from the data, there could be redundant variables that may be susceptible to over-fitting and increased computational cost problems. In this study, we applied the variable selection using random forest (VSURF) method to remove redundant predictors. This method has been widely applied and it can select optimal variables from a high dimensional dataset for regression purposes. This technique is based on random forest and it involves three steps; threshold, interpretation and prediction step. In threshold step, irrelevant variables are eliminated from the dataset next in interpretation step, variables which are related to the response variable are selected and finally, in the prediction step, redundant variables are eliminated from the set of variables selected in the interpretation step. This procedure was executed before developing the drought prediction models with the response variable being the 3-month standardized precipitation and evapotranspiration index (SPEI_3). This was implemented via the "VSURF" package coded in R software. Drought predictor variables selected by VSURF for two data clusters is shown in Table 3.4. Figure 3.6 shows selected predictive variables for cluster 1 and Figure 3.7 shows selected predictive variables for cluster 2.

Cluster	Selected variables
Cluster 1	PRED, PET, VAPD, WETD, CLDD
Cluster 2	PRED, PET, WETD, TMND, VAPD, TMPD

Table 3.4: Predictor variables selected by VSURF for the two data clusters.

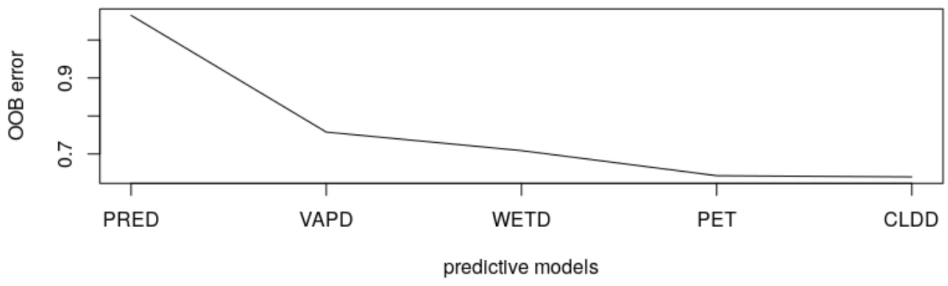


Figure 3.6: Variables for predictive models for cluster 1.

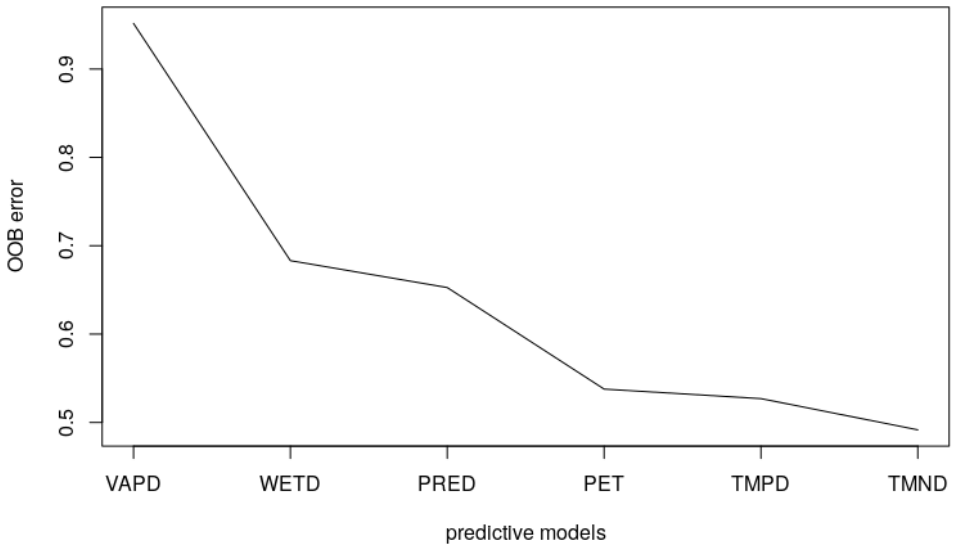


Figure 3.7: Variables for predictive models for cluster 2.

3.7 Machine learning (ML) techniques

3.7.1 Multi-layer perceptron (MLP). A multi-layer perceptron (MLP) is a type of feed-forward artificial neural network. Artificial neural networks are computational models that draws inspiration from biological neural systems. This research study adopted MLP due to its efficiency and popularity. In multi-layer perceptron vertices are arranged in layers and there are generally three layers of nodes consisting of input layer, hidden layer and output layer, in the hidden nodes, there could be one or more layer(s). There is no connection within layer neurons but neurons between layers are connected by weights and errors. Consider the classical case of single hidden layer neural network for regression which is given by (3.7.1):

$$f(\mathbf{x}) = b + \mathbf{W}\sigma(\mathbf{c} + \mathbf{V}\mathbf{x}), \tag{3.7.1}$$

where $\mathbf{x}$ is an input vector(d-vector), $\mathbf{V}$ is an input-to-hidden weights ($k \times d$ matrix), $\mathbf{c}$ is a hidden unit biases (k-vector), σ is an activation function (typical activation functions include ReLU, sigmoid, softmax and hyperbolic tangent), b is an output units biases and $\mathbf{W}$ is hidden-to-output weights ($m \times k$-matrix). Figure 3.8 shows MLP with a single hidden layer architecture and how it works, it has j neurons in the hidden layer.

The working principle of the multi-layer perceptron (MLP) is to minimize the error between the predicted output and the target value by updating the weights i.e if the predicted output is the same as the target then the performance of the model is considered to be satisfactory but if they are different then weights need to be changed to minimize the error. A supervised learning method, namely the Levenberg-Marquardt (LM) back-propagation algorithm is used for training i.e updating weights through back propagation of errors and this is made possible by using sigmoid as a non-linear transfer function. The Levenberg-Marquardt (LM) back-propagation algorithm is efficient and have reduced computational time in training models (Adamowski and Chan, 2011). Inputs and weights are used to work out the

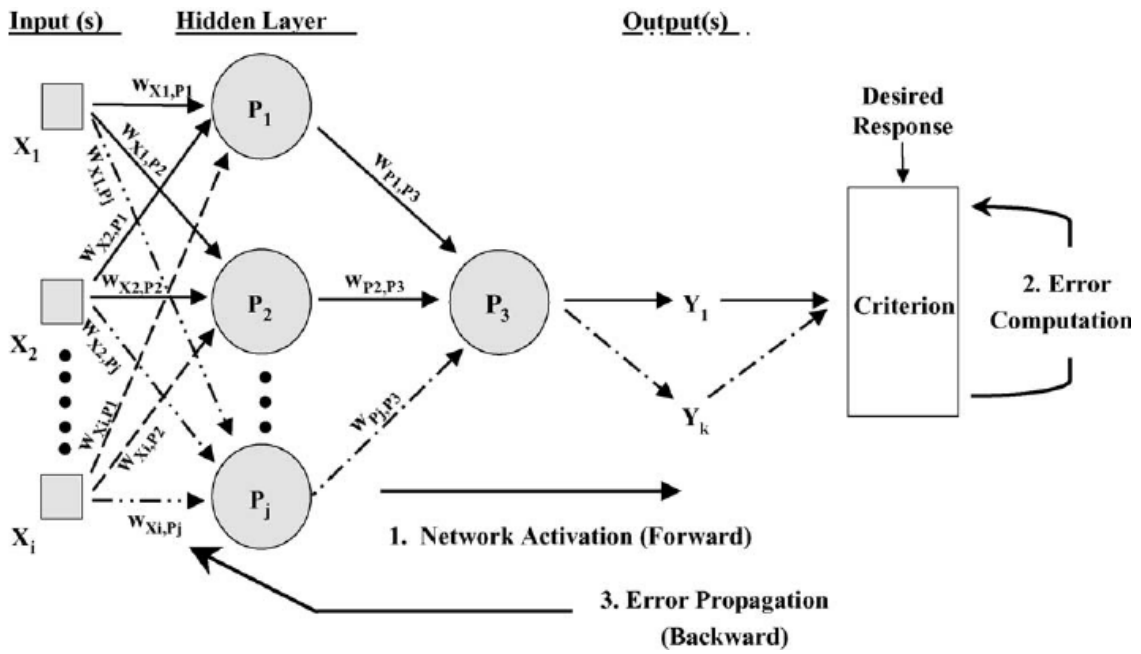


Figure 3.8: MLP with a single hidden layer, Retrieved 11:36 am, May 8, 2020.

activation of any node (i.e weighted sum and transfer function). Normally, the number of hidden neurons for artificial neural network models is selected through trial and error technique. However, a study by Mishra and Desai (2006) discovered that the best number of hidden neurons is $2k + 1$, where k is the number of input neurons and therefore using this proposed method gave $2k + 1$ hidden neurons depending on the number of input variables of the cluster. 75% of the data was used to train the models, while the remaining 25% of the data was used as a testing set.

Alizadeh and Nikoo (2018) utilized five individual artificial intelligence models in order to estimate ground-based standardized precipitation index (SPI) using remote sensing factors and found that multi-layer perceptron (MLP) had the best performance. Hence, the multi-layer perceptron was used in this study to test its performance in forecasting drought in East Africa. multi-layer perceptron (MLP) model was implemented through the "monmlp" package coded in R software.

3.7.2 Support vector regression (SVR). Support vector regression (Cortes and Vapnik, 1995) is a robust algorithm used to estimate a non-linear function. It is an extension of the popular Support vector machine classifiers. SVR deals only with a linear problem but when the system is non-linear, the input vector is mapped implicitly into a high-dimensional feature space (Hilbert space), through a non-linear mapping and then linear regression on this space is conducted. The inner product of this mapping is referred to as kernel function. The main goal of SVR is to find a function $f(\mathbf{x})$ that has utmost ε deviation from the actual targets y_i for all the training data, and is as flat as possible at the same time. Meaning, we don't care about errors as long as they are less than ε, but won't accept any deviation greater than this. The SVR uses ε– insensitive loss function to measure error (Smola and Schölkopf, 2004) as described by:

$$L_\varepsilon(f(\mathbf{x}_i), y_i) = \begin{cases} 0, & \text{for } |f(\mathbf{x}_i) - y_i| < \varepsilon \\ |f(\mathbf{x}_i) - y_i| - \varepsilon, & \text{otherwise} \end{cases}, \tag{3.7.2}$$

The decision function of SVR model based on ε–insensitive loss function is given as:

$$f(\mathbf{x}) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^{\star})\, K(\mathbf{x}_i, \mathbf{x}) + b, \tag{3.7.3}$$

where, l is the number of support vectors, $K(x_i, x)$ is the kernel function with its internal hyper-parameters, α_i are Lagrangian multipliers and b is the bias which is computed as:

$$b = \begin{cases} y_i - \sum\limits_{i=1}^{l} (\alpha_i - \alpha_i^{\star})\, K(\mathbf{x}_i, \mathbf{x}) + \varepsilon, & \alpha_i \in (0, C) \\ y_i - \sum\limits_{i=1}^{l} (\alpha_i - \alpha_i^{\star})\, K(\mathbf{x}_i, \mathbf{x}) - \varepsilon, & \alpha_i^{\star} \in (0, C) \end{cases}, \tag{3.7.4}$$

where the constant $C > 0$ defines the trade-off between the flatness of f and the maximum amount to which deviations greater than ε are tolerated. Solving SVR implies finding the parameters, in its dual formulation by constrained quadratic optimization programming. Figure 3.9 shows the basic idea on how support vector regression works.

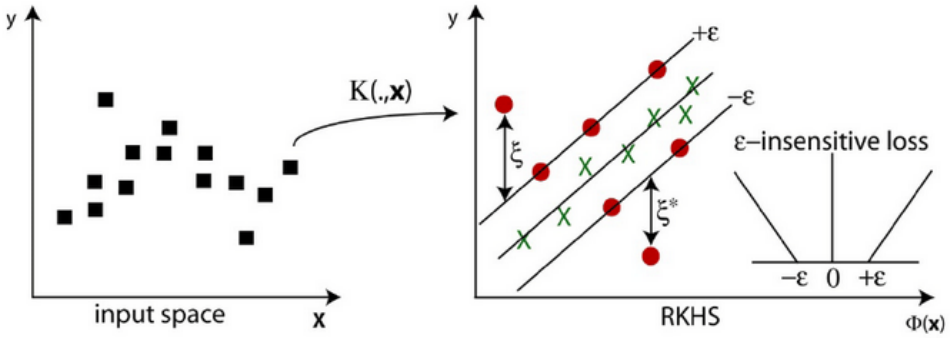


Figure 3.9: SVR maps the input data into the Hilbert space where linear regression is achievable, Retrieved 23:46 pm, May 8, 2020 .

The major benefit of utilizing support vector regression is its non-parametric technique where the output model does not rely on distributions of the underlying independent and dependent variables instead it relies on kernel functions to construct a model. The kernel functions include linear, Gaussian radial basis, polynomial and sigmoid, for this study we selected the Gaussian radial basis function kernel as it is designed to deal with non-linear problems and is commonly used. As a result, the SVR model consisted of three hyper-parameters that were selected: gamma (γ), epsilon (ε) and cost (C). The parameter γ is a constant that reduces the model space and controls the complexity of the solution, while ε is the loss function that describes the regression vector with a reduced set of input data and C is a positive constant for capacity control parameter (Kisi and Cimen, 2011). The optimal hyper-parameters were determined using the 10-fold cross-validation. The data was divided into two, a training set and testing set. 75% of the data was the calibration set while the final 25% of the data was used as the testing set.

Belayneh and Adamowski (2013) used support vector regression to evaluate its performance in forecasting drought in Awash River basin. Therefore, the support vector regression was used in this study to test its performance in forecasting drought in East Africa. Support vector regression model was implemented through the "e1071" package coded in R software.

3.7.3 Extreme Gradient boosting (XGBoost). Boosting is an ensemble technique that tries to create a strong learner based on weak learners by adding models on top of each other through iteration, the errors of the previous model are created by the next predictor until the training is accurately predicted or reproduced by the model. Boosting is called Gradient boosting because we are using the gradient descent algorithm to minimize the losses during new model creation. Figure 3.10 shows how "weak learners" ensembles into a "strong learner".

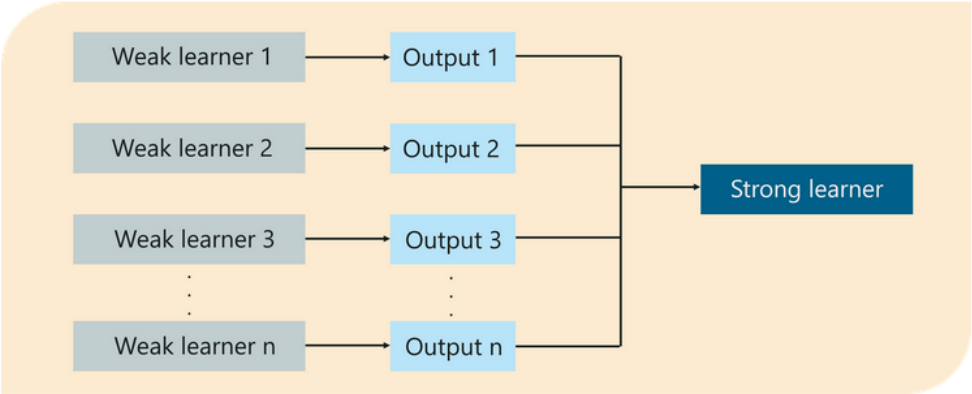


Figure 3.10: Boosting Machine Learning, Retrieved 16:56 pm, May 8, 2020.

XGBoost have a similar framework as gradient boosting but it is used to extremely boost the performance of the model through increased speed and more efficient computation. It has linear model solver and tree learning algorithms. It supports various objective functions including classification, ranking and regression, it has Dmatrix which helps for performance and efficiency and works only with numerical values. Figure 3.11 shows diagrammatically the strengths of XGBoost.

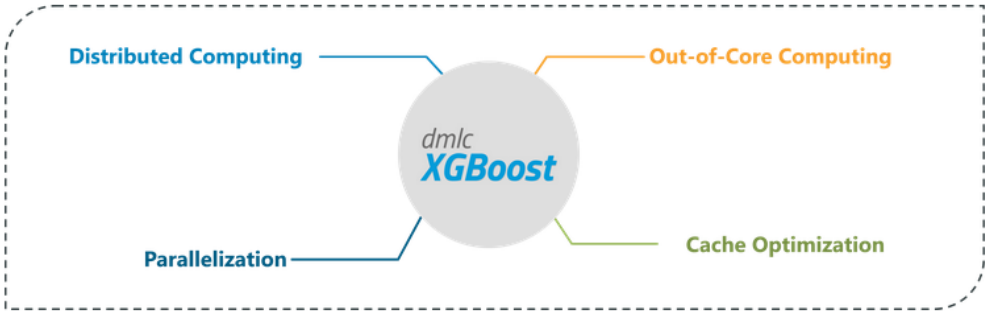


Figure 3.11: XGBoost, Retrieved 2:43 pm, May 8, 2020.

This study used XGBoost because:

- Its fast and accurate execution due to its capacity to do parallel computation on a single machine.
- It has very high predictive power and uses Out-of-Core Computing in analyzing huge datasets.
- In-built cross-validation capacity and regularization to avoid over-fitting.
- It implements distributed computing methods to evaluate complex models.

Due to its parallelism, convenience and impressive predictive accuracy, it has won several machine learning competitions of late (Adam-Bourdarios et al., 2015). Zhang et al. (2019) used the XGBoost technique to evaluate its performance on meteorological drought forecasting based on statistical models in Shaanxi province, China. Therefore, it was used in this study to test its performance in forecasting drought in east Africa. The data was divided into two, a training set and testing set. 75% of the data was the training set while the final 25% of the data was used as the testing set. We trained the model using XGBLinear in the "xgboost" package coded in R software and tuning of parameters was determined using the 10-fold cross-validation.

3.8 Models performance evaluation

After modeling of the Support vector regression (SVR), multi-layer perceptron (MLP) and XGBoost models, the performance of the models in forecasting the monthly SPEI was assessed statistically with key statistical parameters. The parameters used in this study include: the coefficient of determination (R^2) to check the goodness of fit, mean absolute error(MAE) and root mean square error(RMSE) to assess the stability of the models. The calculation of MAE, RMSE and R^2 are shown in (3.8.1), (3.8.2) and (3.8.3) respectively.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|(spei_{pi} - spei_{oi})\right|, \tag{3.8.1}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(spei_{pi} - spei_{oi})^2}, \tag{3.8.2}$$

$$R^2 = \left(\frac{\sum_{i=1}^{n}(spei_{oi} - \overline{spei_o})(spei_{pi} - \overline{spei_p})}{\sqrt{\sum_{i=1}^{n}(spei_{oi} - \overline{spei_o})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(spei_{pi} - \overline{spei_p})^2}}\right), \tag{3.8.3}$$

where n is the number of observations, $spei_{oi}$ and $spei_{pi}$ denote the observed and predicted values of SPEI and $\overline{spei_o}$ and $\overline{spei_p}$ represent the mean of observed and predicted values. Generally, a good model should have lower MAE and RMSE close to 0 while the R^2 values should be close to 1. These values was computed via the "caret" package coded in R software.

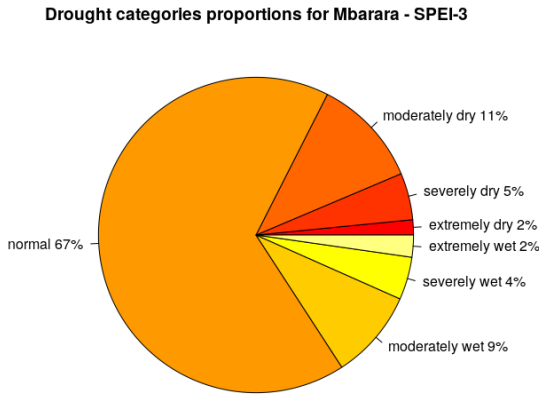
4. Results and Discussion

In this study the ability of the MLP, SVR and XGBoost models to effectively forecast SPEI_3 and SPEI_6 at 1 month lead time was evaluated. The forecasts and comparison of models performance was done for the two stations, the Mbarara station and Turkana station to represent cluster 1 and cluster 2 respectively. While the study was done using all the selected meteorological stations, the results were presented for these two stations in depth.

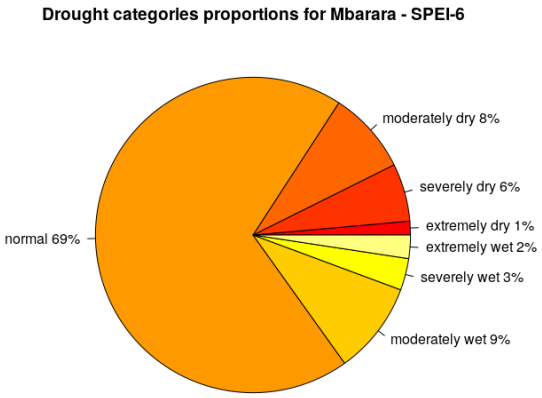
For both stations used in this study, XGBoost presented more accurate forecast results compared to its two counterparts and the predictions across all models for SPEI_6 were more accurate than the results for SPEI_3, the forecast capacity became better with the increase in the time scale because of the increased filter length used in the calculation of SPEI which effectively reduces the noise. In addition, the performance for all the three models was better in the cluster 2 (Turkana) compared to cluster 1 (Mbarara) with XGBoost models presenting the best performance across all the stations.

4.1 Drought descriptive statistics for the stations

During 1960-2018 in both stations, there were 0.71%-13.42% of all months having droughts for two different time scales namely SPEI_3 and SPEI_6. The proportion and the frequency of occurrence of drought categories differed by stations at the same time scales and varied with time scales at the same station as shown in Figure 4.1a – 4.2b. But generally, Turkana station experienced more droughts frequency compared to Mbarara station.

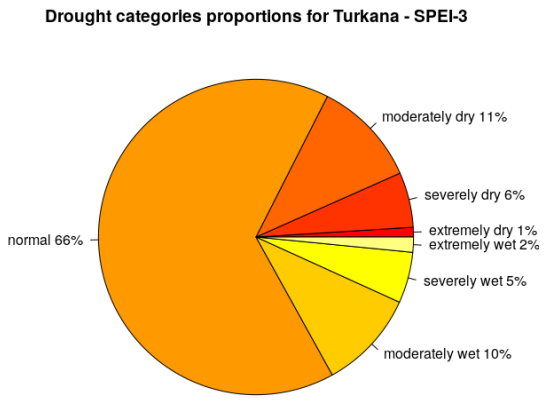


(a) Pie chart showing drought categories proportions for Mbarara station for SPEI_3.

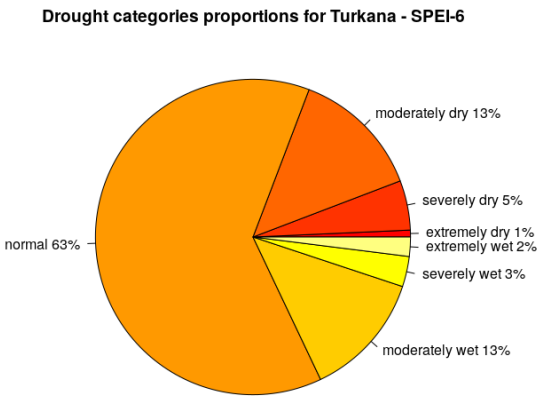


(b) Pie chart showing drought categories proportions for Mbarara station for SPEI_6.

Figure 4.1: Drought categories proportion for Mbarara station.



(a) Pie chart showing drought categories proportions for Turkana station for SPEI_3.



(b) Pie chart showing drought categories proportions for Turkana station for SPEI_6.

Figure 4.2: Drought categories proportion for Turkana station.

For Mbarara station, SPEI_3 and SPEI_6 had a mean of 0.001178 and 0.001016 and standard deviations of 0.9883583 and 0.9864893, respectively. Besides, there were 11.16% and 8.47% of all months during the study period having moderate droughts, 4.80% and 5.63% of all months having severe droughts, and 1.55% and 1.41% of months having extreme droughts for SPEI_3 and SPEI_6, respectively.

For Turkana station, SPEI_3 and SPEI_6 had a mean of 0.0005056 and 0.000409 and standard deviations of 0.9856071 and 0.984511, respectively. Moreover, there were 10.91% and 13.42% of all months during the study period having moderate droughts, 5.67% and 5.08% of all months having severe droughts, and 1% and 0.71% of months having extreme droughts for SPEI_3 and SPEI_6, respectively. The summary statistics for both stations is presented in Table 4.1.

Station	SPEI	Min	Mean	Max	Standard deviation	Moderate drought(%)	Severe drought(%)	Extreme drought(%)
Mbarara	SPEI_3	-2.920878	0.001178	2.980654	0.9883583	11.16	4.80	1.55
	SPEI_6	-2.625647	0.001016	2.701667	0.9864893	8.47	5.63	1.41
Turkana	SPEI_3	-2.3092338	0.0005056	2.8078991	0.9856071	10.91	5.67	1.00
	SPEI_6	-2.298153	0.000409	2.709383	0.984511	13.42	5.08	0.71

Table 4.1: Summary statistics of the two stations based on SPEI with moderate, severe and extreme droughts.

4.2 Prediction of SPEI with its lags

First, we modelled the forecasting of SPEI with its lags using the three machine learning algorithms and found the results of the prediction from the models. For instance, the results from the SPEI_3 forecasts at Mbarara station for an MLP model is 0.4876640, 0.7164997 and 0.5759393 in terms of R^2, RMSE and MAE, respectively. Also, the results for SVR and XGBoost is presented in Table 4.2. This result shows that the prediction capacity of the models using only the lags in forecasting SPEI is low hence indicating that it is necessary to include other meteorological variables which may improve forecasting of SPEI. Hence, in the remaining sections, the modelling of MLP, SVR and XGBoost models includes other meteorological variables selected by VSURF for each cluster. Figure 4.3 shows how the three models forecast SPEI_3 for Mbarara station using only its lags.

Model	R^2	RMSE	MAE
MLP	0.4876640	0.7164997	0.5759393
SVR	0.4960088	0.7020204	0.5625677
XGBoost	0.4768998	0.7241176	0.5871781

Table 4.2: The SPEI_3 forecasts results at Mbarara station for the three models using only its lags.

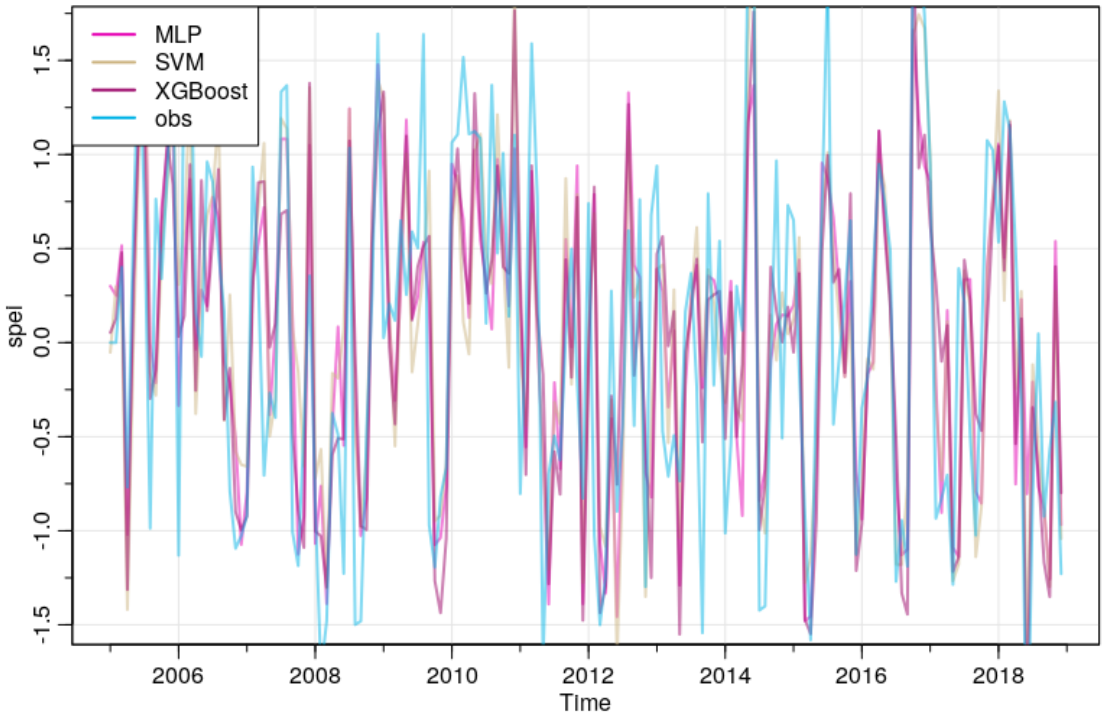


Figure 4.3: SPEI_3 forecasts by own lags using MLP, SVR and XGBoost models at the Mbarara station.

4.3 Target and predictor variables

This study evaluated the ability of the three developed models to forecast SPEI with a 1-month lead time (i.e $SPEI_{n+1}$) between different time scales and across the two stations (Turkana and Mbarara). Therefore, the target variable and input combinations varied based on time scales and cluster in which the station represents. For instance, predicting $SPEI\_3$ in Turkana station, the dependent variable is $SPEI\_3_{n+1}$ and the $SPEI\_3_n, PRED_n, PET_n, WETD_n, TMND_n, VAPD_n, TMPD_n$ were the independent variables. Table 4.3 shows the predicted and the input combinations for different time scales on both stations for all models developed.

Station	Predicted SPEI	Input combination
Mbarara	$SPEI\_3_{n+1}$	$SPEI\_3_n, PRED_n, PET_n, WETD_n, VAPD_n, CLDD_n$
	$SPEI\_6_{n+1}$	$SPEI\_6_n, PRED_n, PET_n, WETD_n, VAPD_n, CLDD_n$
Turkana	$SPEI\_3_{n+1}$	$SPEI\_3_n, PRED_n, PET_n, WETD_n, TMND_n, VAPD_n, TMPD_n$
	$SPEI\_6_{n+1}$	$SPEI\_6_n, PRED_n, PET_n, WETD_n, TMND_n, VAPD_n, TMPD_n$

Table 4.3: The target and predictor variables used to develop models.

4.4 SPEI_3 forecast results

As highlighted earlier SPEI forecasts were conducted and compared for the two stations. This section presents the results from the testing datasets. The SPEI_3 forecasts results for Mbarara station were 0.7160074, 0.5497845 and 0.4202408 in terms of R^2, RMSE and MAE, respectively for MLP model. SVR model presented more accurate results compared to MLP model with 0.8248757, 0.4071635 and 0.3139142 in terms of R^2, RMSE and MAE, respectively. XGBoost model presented the best results with 0.8344541, 0.3992762 and 0.2946675 in terms of R^2, RMSE and MAE, respectively. An R^2 value between $0.30 - 0.50$ indicates a low degree of correlation, a value between $0.50 - 0.70$ indicates a moderate degree of correlation and a value between $0.70 - 0.99$ indicates a high degree of correlation. RMSE and MAE values closer to zero indicates a very stable or high level of precision for models.

Similarly, the results of the SPEI_3 forecasts for Turkana station followed similar pattern shown at the Mbarara station. MLP model results were 0.8747562, 0.3552809 and 0.2526742 in terms of R^2, RMSE and MAE, respectively. Also, the results for the SVR model outperformed the MLP model with 0.89967002, 0.3142822 and 0.2354614 in terms of R^2, RMSE and MAE, respectively. The forecast using XGBoost model yielded the best results with 0.9518374, 0.1971178 and 0.1417616 in terms of R^2, RMSE and MAE, respectively. These results indicate that XGboost outperformed all the other two models across both stations in predicting SPEI_3 based on the three evaluation measurements. These results are presented in Table 4.4. Figure 4.4 and Figure 4.5 show how predicted SPEI_3 by different models mirrors the observed SPEI_3 at Mbarara and Turkana station, respectively.

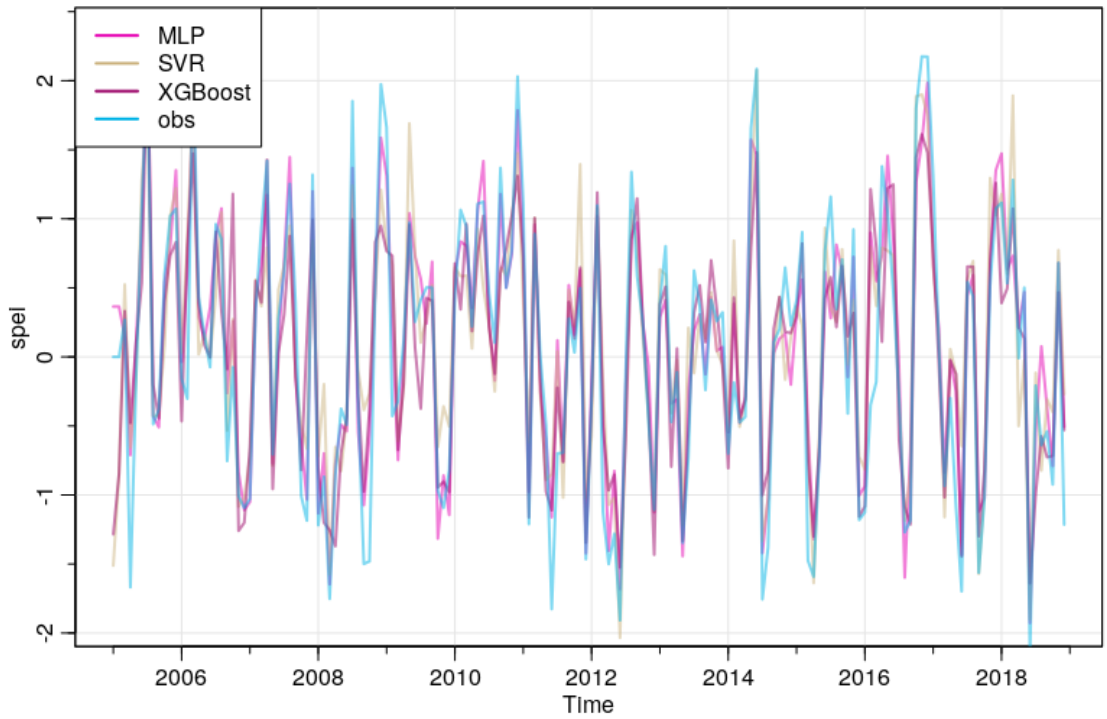


Figure 4.4: SPEI_3 forecasts using MLP, SVR and XGBoost models at the Mbarara station.

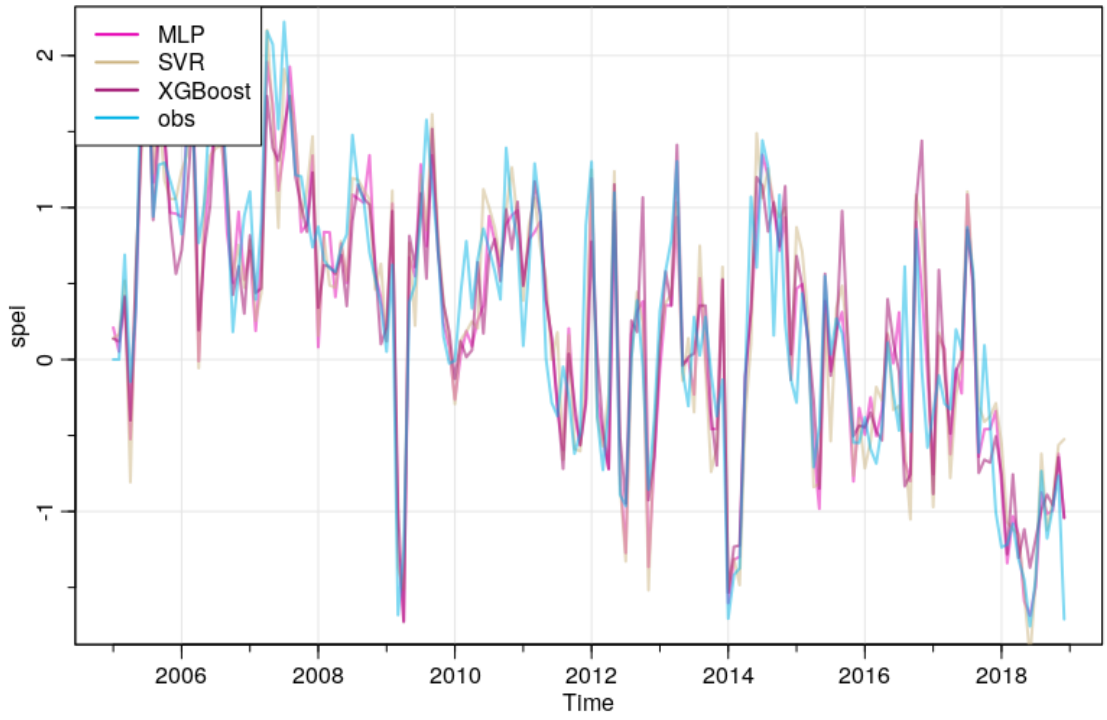


Figure 4.5: SPEI_3 forecasts using MLP, SVR and XGBoost models at the Turkana station.

Model	Predicted	Measure	Mbarara	Turkana
MLP	SPEI_3	R^2	0.7160074	0.8747562
		RMSE	0.5497845	0.3552809
		MAE	0.4202408	0.2526742
	SPEI_6	R^2	0.9103471	0.9361715
		RMSE	0.2857649	0.2520958
		MAE	0.2066440	0.1688453
SVR	SPEI_3	R^2	0.8248757	0.8997002
		RMSE	0.4071635	0.3142822
		MAE	0.3139142	0.2354614
	SPEI_6	R^2	0.9133743	0.9662820
		RMSE	0.2820673	0.1836446
		MAE	0.2103384	0.1388640
XGBoost	SPEI_3	R^2	0.8344541	0.9518374
		RMSE	0.3992762	0.1971178
		MAE	0.2946675	0.1417616
	SPEI_6	R^2	0.9355072	0.9727851
		RMSE	0.2429698	0.1364539
		MAE	0.1881864	0.0961052

Table 4.4: Forecast results for MLP, SVR and XGBoost models.

4.5 SPEI_6 forecast results

Forecast results for SPEI_6 followed the same pattern as the results of SPEI_3. For the Mbarara station, MLP model had forecast results of 0.9103471, 0.2857649 and 0.2066440 in terms of R^2, RMSE and MAE, respectively. SVR model had more accurate results than the MLP model with 0.9133743, 0.2820673 and 0.2103384 in terms of R^2, RMSE and MAE, respectively. XGBoost model presented slightly better results than for the SVR model with a coefficient of determination (R^2) of 0.9355072, a root mean square error of the predicted from the observed SPEI_6 (RSME) of 0.2429698 and a mean absolute error (MAE) of 0.1881864. The forecast results at the Turkana station follows a similar pattern exhibited at the Mbarara station. The results for the Turkana station are presented in Table 4.4.

Figure 4.6 and Figure 4.7 shows how the predicted SPEI_6 by the three models reflects the observed SPEI_6 at Mbarara and Turkana station respectively.

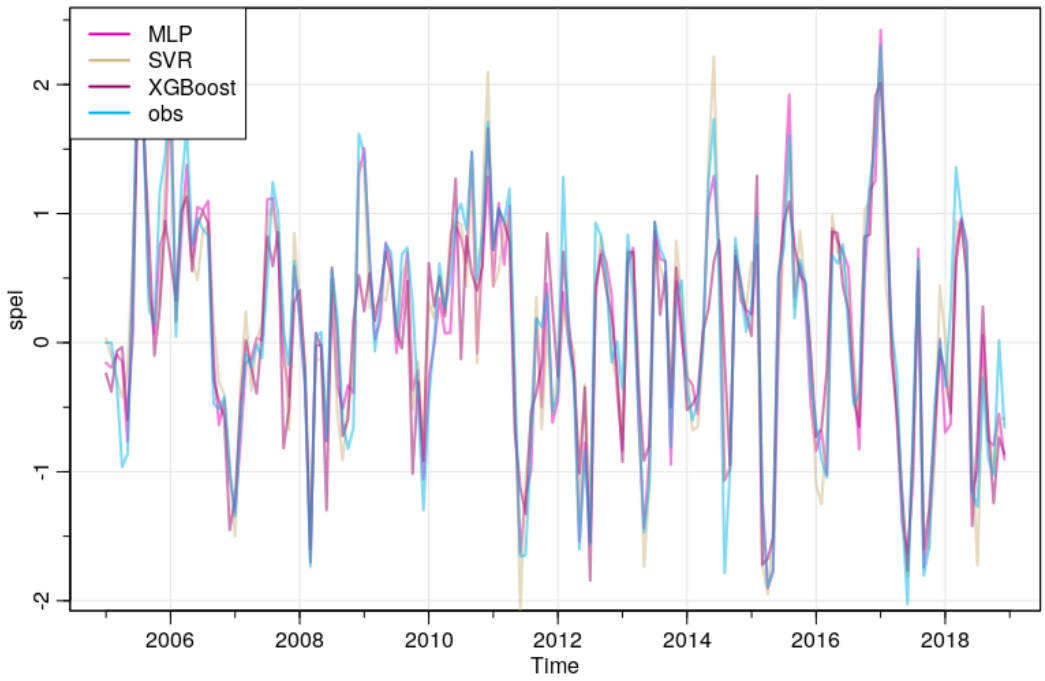


Figure 4.6: SPEI_6 forecasts using MLP, SVR and XGBoost models at the Mbarara station.

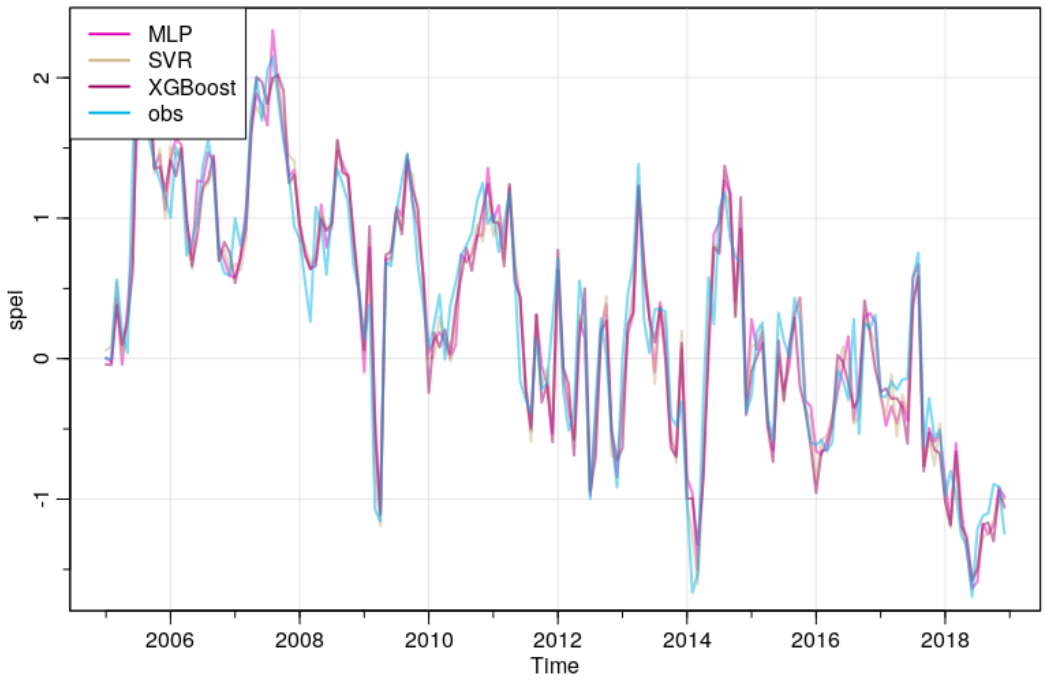
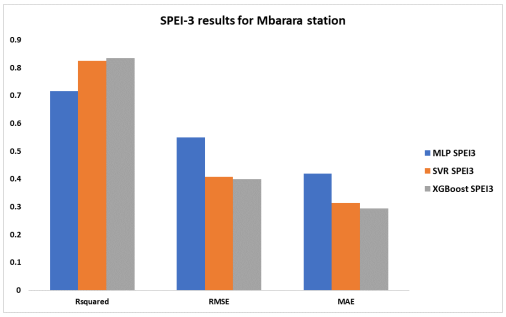
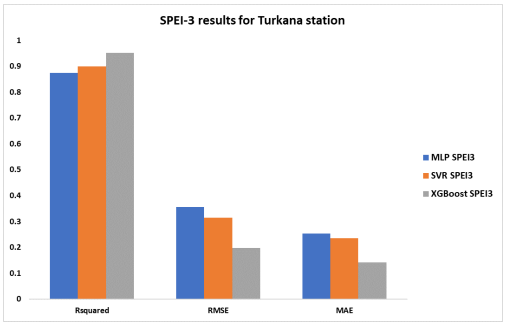


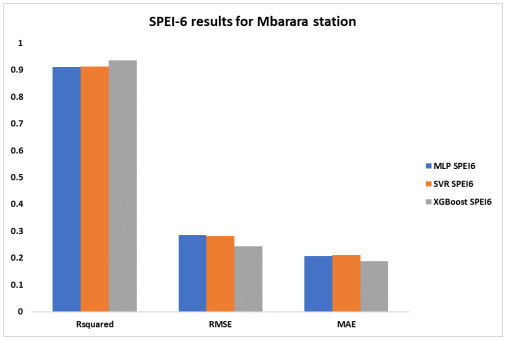
Figure 4.7: SPEI_6 forecasts using MLP, SVR and XGBoost models at the Turkana station.



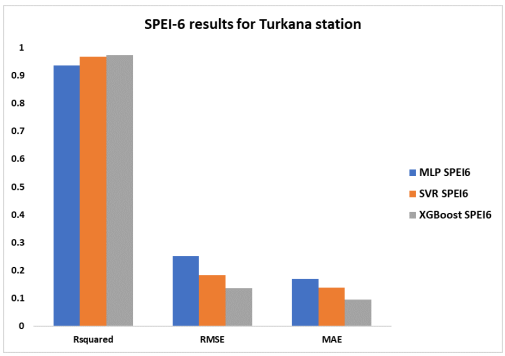
(a) SPEI_3 forecasts results for Mbarara station.



(b) SPEI_3 forecasts results for Turkana station.



(c) SPEI_6 forecasts results for Mbarara station.



(d) SPEI_6 forecasts results for Turkana station.

Figure 4.8: Models performance in terms of R^2, RMSE and MAE.

4.6 Discussion

First, we conducted forecasting of SPEI using lags of itself by the three machine learning algorithms and evaluated the performance of the models. The results indicated that the prediction capacity of the models is low when only the lag of SPEI is used in forecasting SPEI since the R^2 values were far from one and also the RMSE and MAE values were far from zero. This results showed that it is necessary to include other meteorological variables during the forecasting of SPEI.

Therefore, we conducted the selection of predictors for drought forecasting at different stations (clusters) since for machine learning procedures, variable selection is critical because using uncorrelated variables in the training of machines may increase the runtime and the cost of the forecasting process leading to a less accurate generalization of models. We used the selected variables and SPEI for each cluster to train the MLP, SVR and XGBoost models for each station separately and then use the models to forecast.

The results show that MLP, SVR and XGBoost models can be used as a way of predicting the SPEI with lead time in East Africa. Figure 4.4 – 4.7 shows how the SPEI forecasts by the models closely mirror the observed SPEI. This study compared three techniques in drought forecasting and the results indicate that for SPEI_3 and SPEI_6, SVR consistently outperforms MLP. But overall, XGBoost presented the best forecasts as reflected by higher R^2 values and lower RMSE values. Consequently, the MAE values of the XGBoost models were lower compared to the other two models as shown in Figure 4.8. Also, the prediction capacity of the models was better in Turkana station compared to Mbarara station.

In addition, the results from all the models shows that SPEI_6 forecasts were more accurate than SPEI_3 forecasts. As shown in Table 4.4, the results for SPEI_6 are more accurate across all the models. Figure 4.6 and Figure 4.7 shows how SPEI_6 forecasts reflects those of the observed SPEI_6 values very closely. The reason why the forecasting accuracy became better as the timescale increases is due to the increased filter length used in the calculation of SPEI which effectively reduces the noise as the timescale increases.

5. Conclusion and Recommendation

5.1 Conclusion

Efficient, reliable and effective variable selection strategy and forecasting models are useful in monitoring and forecasting drought events. In this study, VSURF approach was used to determine predictive variables from the dataset. Consequently, this study evaluated the ability of the three advanced machine learning techniques (namely MLP, SVR and XGBoost) to forecast meteorological drought in East Africa with a lead time from January 1960 to December 2018 using meteorological drought-related factors. The performance of the models in forecasting SPEI was examined using three evaluation criterions namely coefficient of determination (R^2) to assess goodness of fit, root mean square error (RMSE) and mean absolute error (MAE) to assess the precision of the models.

Based on the high-performance metrics utilizing correlations (R^2) and relatively small forecast errors according to RMSE and MAE, the accuracy of XGBoost model was demonstrated to be a highly powerful tool in comparison to its counterpart models (MLP and SVR). Results also indicated that model performance in predicting SPEI improves with an increase in SPEI timescale as evident by better results for SPEI_6 forecast compared to SPEI_3. Finally, a notable amount of geographic variability in drought models performance was also evident, where a better performance was attained for Turkana station as compared to Mbarara station based on model performance evaluation measures i.e R^2, RMSE and MAE.

This study has limitations. First, because of the difficulty of accessing data other meteorological variables (like relative humidity, wind direction, etc) were not included in the forecasting meteorological drought. Also, rather than using extreme gradient boosting with tree booster that could have performed better, we used XGBoost with a linear booster in order to make the model simple.

This research study provides significantly useful strategies for modelling meteorological drought forecasting which may enable the governments and other relevant stakeholders (including farmers and climate scientists) to make informed decisions on mitigation plans, such as crop and water resource management including dam construction or irrigation operations and other physical and hydrological applications during this phase of continuously changing climate.

5.2 Recommendation

Similar modelling strategy should be emulated in other areas such as forecasting of floods, rainfall and future energy demand. Also, further research should be done in forecasting drought by including more drought-related factors like climate factors (e.g, Oceanic Nino index, Southern Oscillation index, North Atlantic Oscillation, etc) in order to achieve a more accurate performance in drought forecasting.

Appendix

Thornthwaite method

This method has been widely used in calculating the potential evapotranspiration (PET), it is calculated as follows:

$$PET = 16 \times \left(\frac{N}{12}\right) \times \left(\frac{m}{30}\right) \times \left(10 \times \frac{T_i}{I}\right)^a, \tag{.0.1}$$

where m is the number of days in a month, N is the monthly mean sunshine hour, T_i is the monthly mean temperature, I is a cumulative number of 12-month thermal indexes calculated using Equation (.0.3) and a is given by:

$$a = 6.75 \times 10^{-7} \times I^3 - 7.71 \times 10^{-5} \times I^2 + 1.79 \times 10^{-2} \times I + 0.49, \tag{.0.2}$$

$$I = \sum_{i=1}^{12} \left(\frac{T_i}{5}\right)^{1.514}. \tag{.0.3}$$

Acknowledgements

I am thankful to **The Almighty God** for establishing me throughout my academic endeavour. I also sincerely thank my supervisor Prof. Ernest Fokoué and personal thesis tutor Alice Ikuzwe for their immense inspiration, guidance and tireless efforts during the period of research and writing of this thesis, God bless you all. I also acknowledge the authors of books and journals referenced in this study. Lastly, to my fellow students and the entire AIMS–Rwanda community for their genuine and timely contributions, cannot go without mention. Thank you very much indeed.

References

Adam-Bourdarios, C., Cowan, G., Germain-Renaud, C., Guyon, I., Kégl, B., and Rousseau, D. The Higgs machine learning challenge. In *Journal of Physics: Conference Series*, volume 664, page 072015. IOP Publishing, 2015.

Adamowski, J. and Chan, H. F. A wavelet neural network conjunction model for groundwater level forecasting. *Journal of Hydrology*, 407(1-4):28–40, 2011.

AghaKouchak, A. A multivariate approach for persistence-based drought prediction: Application to the 2010–2011 East Africa drought. *Journal of Hydrology*, 526:127–135, 2015.

Ahmad, M., Sinclair, C., and Werritty, A. Log-logistic flood frequency analysis. *Journal of Hydrology*, 98(3-4):205–224, 1988.

Alizadeh, M. R. and Nikoo, M. R. A fusion-based methodology for meteorological drought estimation using remote sensing data. *Remote Sensing of Environment*, 211:229–247, 2018.

Asoka, A. and Mishra, V. Prediction of vegetation anomalies to improve food security and water management in India. *Geophysical Research Letters*, 42(13):5290–5298, 2015.

Ayana, E. K., Ceccato, P., Fisher, J. R., and DeFries, R. Examining the relationship between environmental factors and conflict in pastoralist areas of East Africa. *Science of The Total Environment*, 557:601–611, 2016.

Beguería, S., Vicente-Serrano, S. M., Reig, F., and Latorre, B. Standardized precipitation evapotranspiration index (SPEI) revisited: parameter fitting, evapotranspiration models, tools, datasets and drought monitoring. *International Journal of Climatology*, 34(10):3001–3023, 2014.

Belayneh, A. and Adamowski, J. Drought forecasting using new machine learning methods/Prognozowanie suszy z wykorzystaniem automatycznych samouczących się metod. *Journal of Water and Land Development*, 18(9):3–12, 2013.

Belayneh, A., Adamowski, J., Khalil, B., and Ozga-Zielinski, B. Long-term SPI drought forecasting in the Awash River Basin in Ethiopia using wavelet neural network and wavelet support vector regression models. *Journal of Hydrology*, 508:418–429, 2014.

Blumenstock, G. Drought in the United States analyzed by means of the theory of probability. Technical report, 1942.

Boken, V. K., Cracknell, A. P., and Heathcote, R. L. *Monitoring and predicting agricultural drought: a global study*. Oxford University Press, 2005.

Carmona, P., Climent, F., and Momparler, A. Predicting failure in the US banking sector: An extreme gradient boosting approach. *International Review of Economics & Finance*, 61:304–323, 2019.

Chambers, J. M. and Hastie, T. J. Statistical models. In *Statistical Models in S*, pages 13–44. Routledge, 2017.

Chen, H. and Sun, J. Changes in drought characteristics over China using the standardized precipitation evapotranspiration index. *Journal of Climate*, 28(13):5430–5447, 2015.

Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

Christy, J. R., Norris, W. B., and McNider, R. T. Surface temperature variations in East Africa and possible causes. *Journal of Climate*, 22(12):3342–3356, 2009.

Cook, B. I., Seager, R., and Smerdon, J. E. The worst North American drought year of the last millennium: 1934. *Geophysical Research Letters*, 41(20):7298–7305, 2014.

Cortes, C. and Vapnik, V. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

CVS, S. Some aspects of drought climatology of the dry subhumid zones of south India. *Journal of the Meteorological Society of Japan. Ser. II*, 47(4):239–244, 1969.

Dai, A. Characteristics and trends in various forms of the Palmer Drought Severity Index during 1900–2008. *Journal of Geophysical Research: Atmospheres*, 116(D12), 2011.

Ding, S. *Predicting Dynamics of Vegetative Drought Classes Using Fuzzy Markov Chains*. University of Twente Faculty of Geo-Information and Earth Observation (ITC), 2011.

Dracup, J. A., Lee, K. S., and Paulson Jr, E. G. On the definition of droughts. *Water Resources Research*, 16(2):297–302, 1980.

Feng, P., Wang, B., Li Liu, D., and Yu, Q. Machine learning-based integration of remotely-sensed drought factors can improve the estimation of agricultural drought in South-Eastern Australia. *Agricultural Systems*, 173:303–316, 2019.

Fleig, A., Tallaksen, L., Hisdal, H., and Demuth, S. A global evaluation of streamflow drought characteristics. 2005.

Genuer, R., Poggi, J.-M., and Tuleau-Malot, C. Variable selection using random forests. *Pattern Recognition Letters*, 31(14):2225–2236, 2010.

Gibbs, W. J. and Maher, J. V. RAINFALL DECILES DROUGHT INDDICATORS. 1967.

Guha-Sapir, D., Hargitt, D., and Hoyois, P. *Thirty years of natural disasters 1974-2003: The numbers*. Presses Univ. De Louvain, 2004.

Hadish, L. *Drought risk assessment using remote sensing and GIS: a case study in southern zones, Tigray Region, Ethiopia*. PhD thesis, Addis Ababa Universty, 2010.

Hartigan, J. A. and Wong, M. A. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

Hayes, M. J., Svoboda, M. D., Wiihite, D. A., and Vanyarkho, O. V. Monitoring the 1996 drought using the standardized precipitation index. *Bulletin of the American Meteorological Society*, 80(3): 429–438, 1999.

Hosking, J. R. M. and Wallis, J. R. *Regional frequency analysis: an approach based on L-moments*. Cambridge university press, 2005.

Jensen, M. E., Burman, R. D., and Allen, R. G. Evapotranspiration and irrigation water requirements. ASCE, 1990.

Khajeh Borj Sefidi, A. and Ghalehnoee, M. Analysis of urban growth pattern using logistic regression modeling, spatial autocorrelation and fractal analysis case study: Ahvaz city. *Iran University of Science & Technology*, 26(2):183–194, 2016.

Kisi, O. and Cimen, M. A wavelet-support vector machine conjunction model for monthly streamflow forecasting. *Journal of Hydrology*, 399(1-2):132–140, 2011.

Li, J., Siwabessy, J., Huang, Z., and Nichol, S. Developing an Optimal Spatial Predictive Model for Seabed Sand Content Using Machine Learning, Geostatistics, and Their Hybrid Methods. *Geosciences*, 9(4):180, 2019.

Livne, M., Boldsen, J. K., Mikkelsen, I. K., Fiebach, J. B., Sobesky, J., and Mouridsen, K. Boosted tree model reforms multimodal magnetic resonance imaging infarct prediction in acute stroke. *Stroke*, 49 (4):912–918, 2018.

Lumley, T. and Miller, A. Leaps: regression subset selection. *R Package Version*, 2:2366, 2009.

Ma, F., Luo, L., Ye, A., and Duan, Q. Seasonal drought predictability and forecast skill in the semi-arid endorheic Heihe River basin in northwestern China. *Hydrology and Earth System Sciences*, 22(11): 5697–5709, 2018.

McGUIRE, J. K. and Palmer, W. C. The 1957 drought in the eastern United States. *Mon. Weather Rev*, 85(9):305–314, 1957.

McKee, T. B., Doesken, N. J., Kleist, J., et al. The relationship of drought frequency and duration to time scales. In *Proceedings of the 8th Conference on Applied Climatology*, volume 17, pages 179–183. Boston, 1993.

Meier, P., Bond, D., and Bond, J. Environmental influences on pastoral conflict in the Horn of Africa. *Political Geography*, 26(6):716–735, 2007.

Mishra, A. and Desai, V. Drought forecasting using feed-forward recursive neural network. *Ecological Modelling*, 198(1-2):127–138, 2006.

Mishra, A. K. and Singh, V. P. A review of drought concepts. *Journal of Hydrology*, 391(1-2):202–216, 2010.

Mishra, A. K. and Singh, V. P. Drought modeling–A review. *Journal of Hydrology*, 403(1-2):157–175, 2011.

Mitchell, T. D., Carter, T. R., Jones, P. D., Hulme, M., New, M., et al. A comprehensive set of high-resolution grids of monthly climate for Europe and the globe: the observed record (1901–2000) and 16 scenarios (2001–2100). *Tyndall Centre for Climate Change Research Working Paper*, 55(0): 25, 2004.

Morid, S., Smakhtin, V., and Bagherzadeh, K. Drought forecasting using artificial neural networks and time series of drought indices. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 27(15):2103–2111, 2007.

Okasha, M. K. Using support vector machines in financial time series forecasting. *International Journal of Statistics and Applications*, 4(1):28–39, 2014.

Olesen, J. E., Trnka, M., Kersebaum, K.-C., Skjelvåg, A. O., Seguin, B., Peltonen-Sainio, P., Rossi, F., Kozyra, J., and Micale, F. Impacts and adaptation of European crop production systems to climate change. *European Journal of Agronomy*, 34(2):96–112, 2011.

Oliveira, P. T. S., Nearing, M. A., Moran, M. S., Goodrich, D. C., Wendland, E., and Gupta, H. V. Trends in water balance components across the Brazilian Cerrado. *Water Resources Research*, 50(9): 7100–7114, 2014.

Palmer, W. C. *Meteorological drought*, volume 30. US Department of Commerce, Weather Bureau, 1965.

Pozzi, W., Sheffield, J., Stefanski, R., Cripe, D., Pulwarty, R., Vogt, J. V., Heim Jr, R. R., Brewer, M. J., Svoboda, M., Westerhoff, R., et al. Toward global drought early warning capability: Expanding international cooperation for the development of a framework for monitoring and forecasting. *Bulletin of the American Meteorological Society*, 94(6):776–785, 2013.

Ray, K. S., Shewale, M., et al. Probability of occurrence of drought in various sub-divisions of India. *Mausam*, 52(3):541–546, 2001.

Rodrigues, L. A. Z.-J. F. and do Amaral, J.-A. R. Social tagging for e-learning: an approach based on the triplet of learners, learning objects and tags. 2015.

Shafiee-Jood, M., Cai, X., Chen, L., Liang, X.-Z., and Kumar, P. Assessing the value of seasonal climate forecast information through an end-to-end forecasting framework: Application to US 2012 drought in central Illinois. *Water Resources Research*, 50(8):6592–6609, 2014.

Smola, A. J. and Schölkopf, B. A tutorial on support vector regression. *Statistics and computing*, 14 (3):199–222, 2004.

Sylla, M., Giorgi, F., Coppola, E., and Mariotti, L. Uncertainties in daily rainfall over Africa: assessment of gridded observation products and evaluation of a regional climate model simulation. *International Journal of Climatology*, 33(7):1805–1817, 2013.

Thornthwaite, C. W. The climates of North America: according to a new classification. *Geographical Review*, 21(4):633–655, 1931.

Thornthwaite, C. W. An approach toward a rational classification of climate. *Geographical Review*, 38 (1):55–94, 1948.

Tirivarombo, S., Osupile, D., and Eliasson, P. Drought monitoring and analysis: standardised precipitation evapotranspiration index (SPEI) and standardised precipitation index (SPI). *Physics and Chemistry of the Earth, Parts A/B/C*, 106:1–10, 2018.

Toulmin, C. Drought and the farming sector: Loss of farm animals and post-drought rehabilitation. 1986.

Vicente-Serrano, S. M., Beguería, S., and López-Moreno, J. I. A multiscalar drought index sensitive to global warming: the standardized precipitation evapotranspiration index. *Journal of Climate*, 23(7): 1696–1718, 2010.

Vicente-Serrano, S. M., Beguería, S., and López-Moreno, J. I. Comment on "Characteristics and trends in various forms of the Palmer Drought Severity Index (PDSI) during 1900–2008" by aiguo dai. *Journal of Geophysical Research: Atmospheres*, 116(D19), 2011.

Wilhite, D. A. and Svoboda, M. D. Drought early warning systems in the context of drought preparedness and mitigation. *Early Warning Systems for Drought Preparedness and Drought Management*, pages 1–21, 2000.

Yang, M., Yan, D., Yu, Y., and Yang, Z. SPEI-based spatiotemporal analysis of drought in Haihe River Basin from 1961 to 2010. *Advances in Meteorology*, 2016, 2016.

Yu, W., Liu, T., Valdez, R., Gwinn, M., and Khoury, M. J. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. *BMC Medical Informatics and Decision Making*, 10(1):16, 2010.

Yuan, X., Zhang, M., Wang, L., and Zhou, T. Understanding and seasonal forecasting of hydrological drought in the Anthropocene. *Hydrology and Earth System Sciences*, 21(11):5477, 2017.

Zhang, R., Chen, Z.-Y., Xu, L.-J., and Ou, C.-Q. Meteorological drought forecasting based on a statistical model with machine learning techniques in Shaanxi province, China. *Science of the Total Environment*, 665:338–346, 2019.