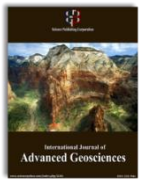
International Journal of Advanced Geosciences, 6 (2) (2018) 195-199

International Journal of Advanced Geosciences

Website: www.sciencepubco.com/index.php/IJAG
doi: 10.14419/ijag.v6i2.14150
Research paper



Data mining techniques for rainfall prediction in the Tepi region of Ethiopia

D. Sivanesan [1*], M. Javed Idrisi [2]

[1] Department of Information Systems, Mizan Tepi University, Ethiopia
[2] Department of Mathematics, Mizan Tepi University, Ethiopia
**Corresponding author E-mail: sivanesansalem@gmail.com*

Abstract

Agriculture depends mainly on the rainfall especially in countries like Ethiopia (Africa) as irrigation system is not much in practice. One of the main reasons is because of its natural topography. Though there are many factors that affect the agricultural yield, it is appropriate to consider the main factor rainfall that decides about the food production. The prediction of the rainfall can be done by using different techniques like regression analysis, clustering, artificial neural network (ANN) and fuzzy logic. Therefore, the significance of this research is essential for the Tepi region in the south west part of Ethiopia (SNNPR) where agriculture is the main occupation of the people living here. This research is first of its kind conducted in this region, and this paper shows the result related with the rainfall prediction by using LR – Linear Regression technique for the early prediction of the next consecutive three (3) years based on the previous available rainfall data.

Keywords: *Data Mining; Rainfall Prediction; Linear Regression; Agriculture.*

1. Introduction

The occurrence of extremely heavy rainfall during a specific short period causes very great effect on the human life such as flood, and when there is insufficient rainfall in long period causes drought (Özlem, 2012). Thus, rainfall estimation is very important in terms of effects on human life, water resources, and water usage areas. However, rainfall affected by the geographical and regional variations and features is very difficult to estimate.

Agriculture forms as the main economy in most of the world countries. Irrigation is considered as the prime factor for agricultural crop production and that irrigation is depending on the rainfall in that region (Navid et.al., 2018). Basically there are two approaches for predicting rainfall. Emperical and Dynamical approach. Emperical – based on the analysis of historical data of the rainfall. This data with its relationship with atmosphere, wind, soil temperature, surface temperature and oceanic variables are used for detailed analysis related to climate and rainfall prediction. Dynamical - The physical models that are based on systems of equations used for predicting the global climate system (Navid et.al., 2018). The numerical rainfall forecasting method is used in this dynamical approach (Ismail et.al., 2009).

Regression is a statistical empirical technique used in many fields including rainfall prediction. The regression analysis includes the parametric methods like linear, multi linear, and logistic regression. For estimation and prediction analysis Non-parametric methodologies like additive models, projection pursuit and multivariate adaptive regression are also used (Paras et.al., 2012). The future events are estimated in statistical analysis by using regression models based on the available previous data. The other methods like trend extraction, curve fitting are used to estimate the future behavior of the time series and fitting the future data to the trend. Regression techniques are widely used in the many areas including business, the social and behavioral sciences, the biological sciences, climate prediction. Regression analysis includes parametric and non-parametric methodologies such as linear and logistic regression under parametric and projection pursuit, additive models, multivariate adaptive regression etc. under Non-parametric methodologies. They also been applied on problems related with prediction and estimation (Paras et.al., 2012). For predicting the gold price using linear regression a model was proposed by the author by using the parameters like inflation, money supply and concluded that the performance of MLR (Multiple Linear Regression) is better than Naïve method of prediction (Ismail et.al., 2009).

In statistical analysis especially in the prediction methods, the Support vector machines (Radhika et.al., 2009) are also used for a set of supervised learning methods, by creating decision making system that is trying to predict new values. The forecasting of the climate can be simply done (Paras et.al., 2012) by regression techniques. In the researches that are carried out recently, using of data mining process is increasing in the field of hydrology. The studies have been conducted in many areas (Roz, 2011). In the present days, by using Artificial Intelligent methods the rainfall in the future can be predicted. The study conducted on examining the effects of El Ni˜no–southern oscillation (ENSO) and the Indian Ocean Sea Surface Temperature (SST) on the rainfall variability in the country of Sudan (the largest country in Africa). Two types of quantitative rainfall prediction models are developed and compared by the author (Yassin et.al., 2002). The variability in the rainfall in different parts of the world is influenced by ENSO is revealed by many authors in their studies (Parthasarathy et.al., 1984; Janowiak, 1988; Parthasarathy et.al., 1988; Allan, 1990; Tapley, 1990; Lough, 1992, 1997; Hastenrath et al., 1993, 1995; Camberlin, 1995, 1997; Gingras et.al., 1995; Seleshi et.al., 1995; Glantz, 1996; Piechota et.al., 1996; Sun et.al., 1997; Chiew et.al., 1998; Nazemosadat et.al., 2000; Sewell et.al., 2001).

The main cause for the occurrence of the drought is of the below normal precipitation that are affected by various natural phenomena. The notable climatic variation from one year to another is of southern Oscillation Index (SOI) (Panu et.al., 2002). The influence of seasonal rainfall pattern in Zimbabwe is by the SOI the difference between normalized sea level pressure of Darwin (in Australia) and Tahiti (in mid pacific). Next Darwin sea level pressure is influencing the seasonal rainfall in Zimbabwe (Manatsa et.al., 2007). To forecast the (SMR) the rainfall over Thailand, SMR method by using multiple linear regression and local polynomial-based nonparametric approaches. The factors that are considered are for predictions are SST, sea level pressure (SLP), wind speed, EiNino Southern Oscillation Index (ENSO), and IOD. The outcome of the experiments indicated that the correlation between observed data and forecasted rainfall data was 0.6 (Nkrintra et.al., 2005). Data mining process was used by proposing a new model for estimating the rainfall in Esparto. The author in his research used the monthly rainfall data of Senirkent, Uluborlu and E˘girdir stations. It is found that the relative error of the model used by him was 0.7% (Ozlem, 2012). The procedure for forecasting rainfall amount involve certain steps that includes Data collection, data preprocessing, data selection, reduction of explanatory predictor, model building by using the regression and finally the validity check (Neha, 2012). For the prediction of the rainfall, the MPR technique,is a better way to describe complex nonlinear I/P-O/P relationship. Then the accuracy is compared between the MPR and MLR technique (Wint, 2008).

Data Mining in weather prediction by using Naïve Bayes and C4.5 decision tree algorithm proved that accuracy was 88.2% for C4.5 decision tree algorithm and it was 54.8% for Naïve Bayes while classifying the instances (Fahad et.al. 2016). The sahel rainfall over Sudan, northern Ethiopia and Eritrea are influenced by the tropical Indian Ocean SST (Sea Surface Temperature) (Camberlin, 1995). It is determined that the most appropriate algorithm was multilinear regression among the models that were developed by means of different data-mining algorithms (Özlem, 2012). The different methods such as support vector machine, Fuzzy logic, Back Propagation Neural Network were used and the author has got results that are significant in the prediction analysis (Nitin, 2016). Three years data related with the rainfall in different cities were collected and concluded that decision tree is suitable for multi variable analysis and weather prediction (Patil 2017). The hourly rainfall was predicted in time efficiently by means of C4.5 and CART decision tree algorithms. In his findings, CART gives slightly better performance than C4.5 (Soo-Yeon, 2012). The model based on decision tree data mining prediction algorithm is better than ANN is revealed by his research (Ramsundram et.al., 2016).

The paper is organized in the following manner: Section II explains about the Objective of conducting this research. Section III discusses about the factors that are influencing the rainfall. Section IV describes about the methodology used in this research and Section V about results and discussions. Section VI gives the conclusion and future enhancements in this research.

2. Objectives of the research

The major problem which is directly connected with agricultural production is the quantity of rainfall the land receives in its region. Today the amount of rainfall received throughout the world is decreasing every year that has been proved in many researches. We are now in a stage of identifying the reasons for the rainfall reduction as it varies from place to place and thereby taking immediate measures to increase the rainfall same like before.

For this purpose, we first need to collect rainfall data pertinent to the region that we select, analyze the data and apply some statistical techniques to predict about the future rainfall. This analysis will in turn help the government to make strict policies and procedures related to the environment protection (deforestation) and impose them on the agricultural sectors. Therefore the ultimate goal of this study shows how the prediction is done by using the selected technique. It also helps the farmers in choosing alternative vegetation at the time of drought, creates them awareness to save and utilize the water efficiently as they will be informed well in advance about the amount of rainfall that they may receive in the near future based on the report generated from the prediction analysis.

3. Factors influencing rainfall

There are many factors that decide the amount of rainfall received in a particular region. Oceanic factors, area covered by forest, the type of tree that is grown in the surrounding area, air temperature, vapor pressure, cloud cover etc. Since the location that we choose for the research is a region is not near the sea shore, we are not using oceanic data. The Deforestation is happening because of converting the forest into coffee and tea estates. In Tepi region, mostly coffee are planted which means the total area cultivated for coffee is equivalent to that of the deforested area. Tea and coffee plants, cashew nut tree will not cultivate clouds that give rainfall. Another factor that affects the rainfall is planting of some type of trees like Eucalyptus tree which can be found with considerable amount in this region. It is also to be noted that Coffee, Tea are considered as cash crops yielding more economy for the country.

4. Methodology

Rainfall Prediction using Linear Regression

In statistical analysis regression attempts to determine the strength of the relationship between a dependent variable which is normally denoted as y and a series of other changing variables called as independent variables. There are only two variables used in simple regression. One variable is independent and the other variable is dependent.

The relationship is mentioned as y=a+bx. This is called as deterministic model (Navid, 2018). In the above equation, y= dependent variable; x= independent variable; a,b= regression parameters

Data Collection It is the first step for data mining. The Rainfall dataset is collected from SPICE RESEARCH CENTER – TEPI (a Government office) who are recording rainfall data twice a day. They record many parameters like air temperature, wind data, soil temperature, surface temperature etc.

Data Cleaning is one of the challenging tasks in data mining. The data collected from the research center had some noisy data and there are some missing values, wrong data and some unwanted data. The data have to be cleaned by filling missing values and removing the inappropriate data.

Data selection is the next step after the data cleaning. Identifying the data which are necessary for our research work is the significance of this step. After this the predictors that have high inter correlation with other parameters are reduced, as the presence of many inter correlated explanatory variables may substantially decrease the sampling accuracy of the regression coefficients, and it will degrade the predictive model ability.

Data Modelling The step after the reduction explanatory predictors is building the model with the use of training set data. The technique that is used for our work is linear regression technique.

5. Results and discussions

On the basis of data available we may predict the average rainfall for the next three years i.e. 2018, 2019 and 2020. The average rainfall (AR) in mm for each month since 2013 to 2017 is given in Table 1. To predict the rainfall we adopt the linear regression technique. We assume the linear regression equation as $R = \alpha + \beta Y$, where R is the average rainfall, Y is the corresponding year and α, β are the regression parameters. The regression parameters α and β are given by

$$\alpha = \frac{\sum R}{n}\left[\frac{\sum YR - (\sum Y)^2}{(\sum Y)^2 - n\sum Y^2}\right] \text{ And } \beta = \frac{\sum Y \sum R - n\sum YR}{(\sum Y)^2 - n\sum Y^2}.$$

Therefore, after obtaining the values of α and β and substituting in R = α + β Y, we have a linear regression equation (here n = 5). Using this linear regression equation we can predict the average rainfall in the next three years shown in Fig. 1 and Fig. 2. In all figures, the available data is shown by blue dots while predicted values are shown by red dots. In Fig. 1(a), the red dot corresponding to the year 2018 shows the increase in the average rainfall with respect to the previous year 2017. Similarly, the red dots corresponding to the years 2019 and 2020 represents the decrease in the average rainfall with respect to the previous years 2018 and 2019 respectively. Our observations on the basis of the linear regression technique are given in the observation Table i.e. Table 1, where RI, RD and NR means rainfall increases, rainfall decreases and no rainfall respectively with respect to the previous year and overall assumption (OA) represents the rainfall for the next three coming years.

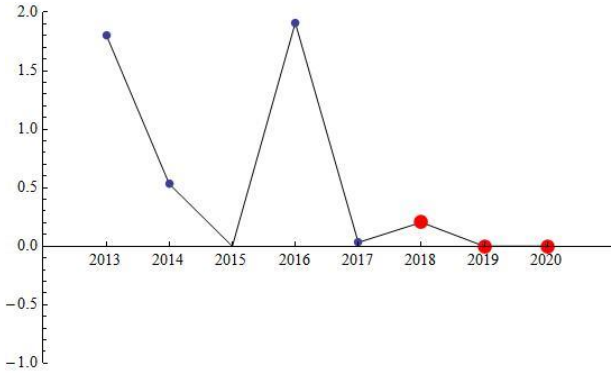


Fig. 1: A) AR in the Month of January.

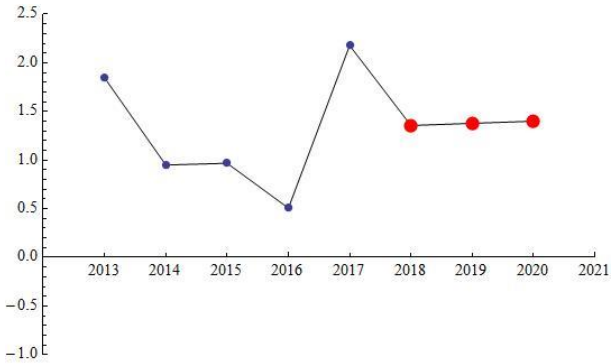


Fig. 1: B) AR in the Month of February.

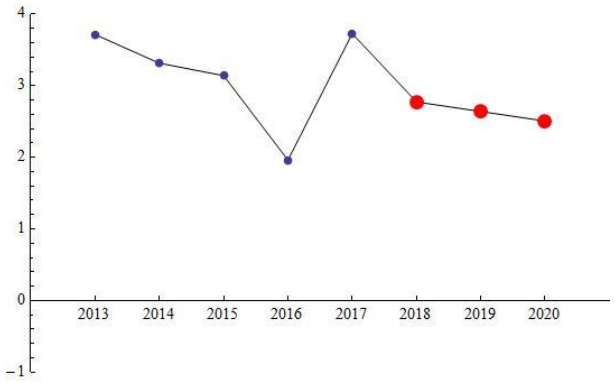


Fig. 1: C) AR in the Month of March.

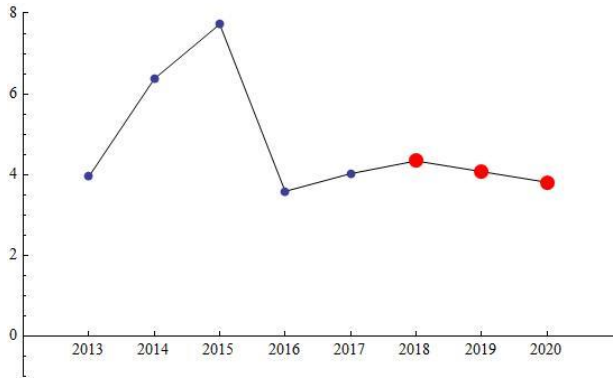


Fig. 1: D) AR in the Month of April.

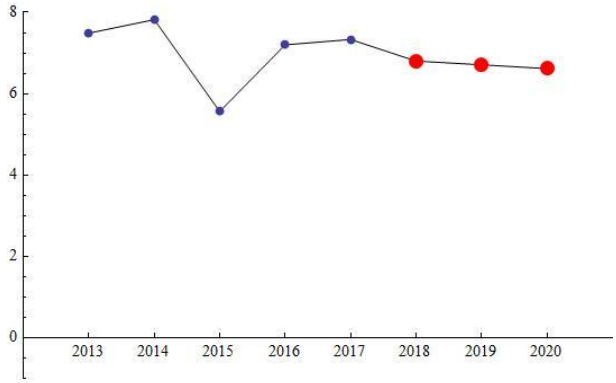


Fig. 1: E) AR in the Month of May.

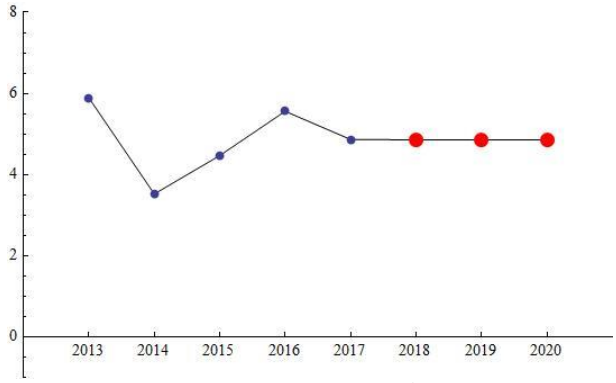


Fig. 1: F) AR in the Month of June.

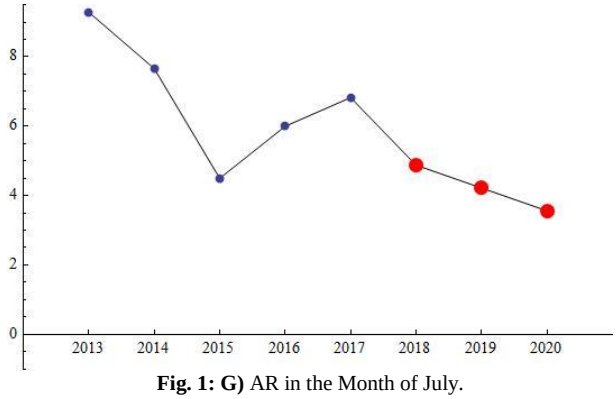


Fig. 1: G) AR in the Month of July.

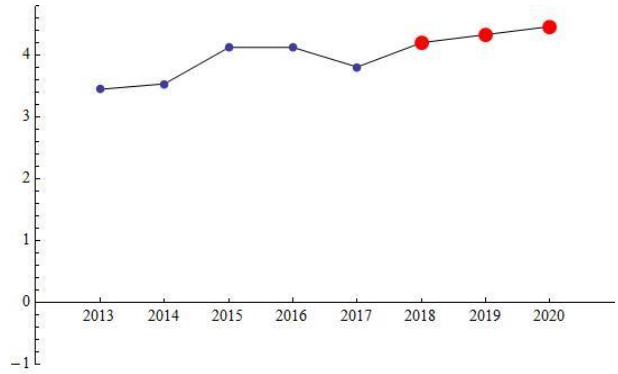


Fig. 1: H) AR in the Month of August.

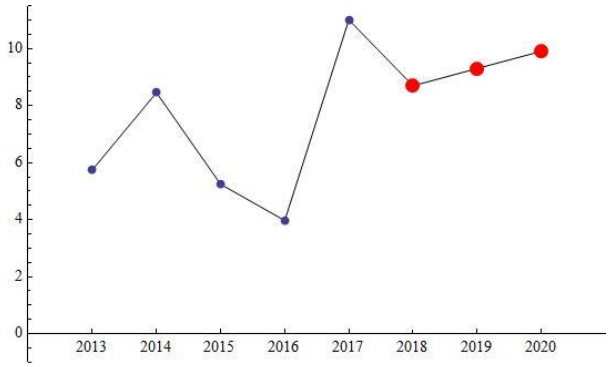


Fig. 1: G) AR in the Month of September.

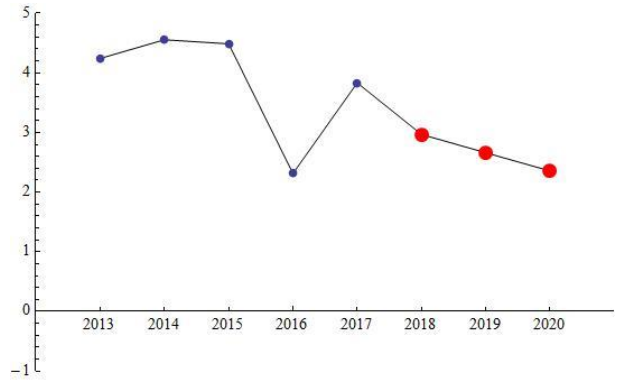


Fig. 1: H) AR in the Month of October.

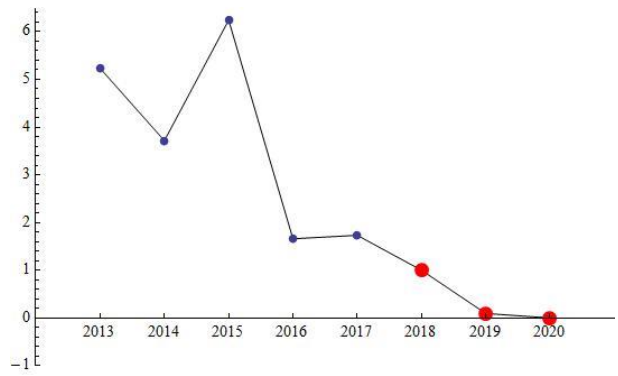


Fig. 1: I) AR in the Month of November.

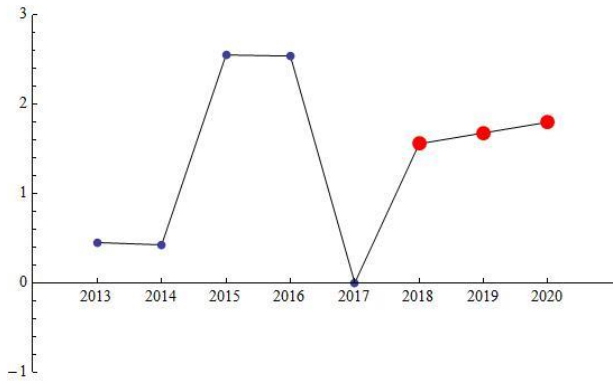


Fig. 1: J) AR in the Month of December.

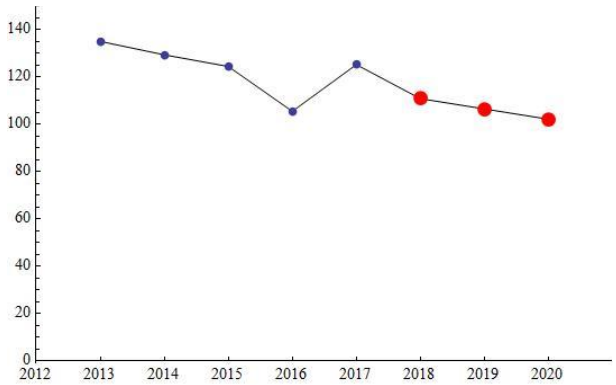


Fig. 2: Predicted AR in the Years 2018, 2019 and 2020.

Table 1: Observation Table

Year	2018	2019	2020
January	RI	NR	NR
February	RD	RI	RI
March	RD	RD	RD
April	RI	RD	RD
May	RD	RD	RD
June	RD	RD	RD
July	RD	RD	RD
August	RI	RI	RI
September	RD	RI	RI
October	RD	RD	RD
November	RD	RD	NR
December	RI	RI	RI
OA	RD	RD	RD

Note: RI: Rainfall Increases, RD: Rainfall Decreases, NR: No Rainfall, OA: Overall Assumption.

6. Conclusion and future enhancements

From our observation, we come to a conclusion that there is a reduction in the rainfall for the next consecutive years starting from 2018 to 2020. It is also to be noted that amount of rainfall received is decreasing every year. Therefore this research has its significance in making the government organizations, NGOs and agriculturists living in this region to take precautionary measures to face the consequences that may arise because of this rainfall reduction. This research can be extended with more oceanic parameters, and other parameters like air temperature, surface temperature, and soil temperature etc. The other aspect of taking this research further is by comparing the results by adopting different methodology and algorithms like Clustering, ANN and Fuzzy logic. This will give a new vision for selecting the appropriate algorithm based on the different parameters that are considered for our prediction analysis.

References

[1] Allan, R.J., Pariwono, J.I.: Ocean–atmosphere interactions in low-latitude Australia. International Journal of Climatology, 10, 145-178 (1990). https://doi.org/10.1002/joc.3370100204.

[2] Camberlin, P.: June–September rainfall in northeastern Africa atmospheric signals over the tropics: a zonal perspective. International Journal of Climatology, 15, 773-783 (1995). https://doi.org/10.1002/joc.3370150705.

[3] Camberlin, P.: Rainfall anomalies in the source region of the Nile and their connection with the Indian summer monsoon. Journal of Climate, 10, 1380–1392 (1997). https://doi.org/10.1175/1520-0442(1997)010<1380:RAITSR>2.0.CO;2.

[4] Chau, K.W., Muttil, N.: Data mining and multivariate statistical analysis for ecological system in coastal waters. Journal of Hydroinformatics, 9 (4), 305-317 (2007). https://doi.org/10.2166/hydro.2007.003.

[5] Chiew, F.H.S., Piechota, T.C., Dracup, J.A., McMahon, T.A.: El Ni˜no/southern oscillation and Australian rainfall, streamflow and drought: links and potential for forecasting. Journal of Hydrology, 204, 138–149 (1998). https://doi.org/10.1016/S0022-1694(97)00121-2.

[6] Damle, C., Yalcin, A.: Flood prediction using time series data mining. Journal of Hydrology, 333, 305-316 (2007). https://doi.org/10.1016/j.jhydrol.2006.09.001.

[7] Erol, K.M., Ozlem, T., Ugur, E.K.: Data mining process for integrated evaporation model. Journal of Irrigation and Drainage Engineering, https://doi.org/10.1061/(ASCE)0733-9437(2009)135:1(39).

[8] Fahad, S., Karthick, S., Malathi, D., Sudarsan, J.S., Arun, C.: Analysis of Data Mining Techniques for Weather Prediction, Indian Journal of Science and Technology, 9(38), 1-9 (2016)

[9] Gingras, D., Adamowski, K.: The impact of El Ni˜no southern oscillation on central Canadian floods and droughts. Canadian Journal of Civil Engineering 22, 834–837 (1995). https://doi.org/10.1139/l95-092.

[10] Glantz, M.H.: Currents of Change: El Ni˜no's Impact on Climate and Society. Cambridge University Press (1996).

[11] Hastenrath, S., Nicklis, A., Greischar, L.: Atmospheric–hydrospheric mechanisms of western Indian Ocean. Journal of Geophysical Research, 98(20), 219–235 (1993).

[12] Hastenrath, S., Greischar, L., Van, H. J.: Prediction of the summer rainfall over South Africa. Journal of Climate, 8, 1511–1518 (1995). https://doi.org/10.1175/1520-0442(1995)008<1511:POTSRO>2.0.CO;2.

[13] Ismail, Z., Yahya, A., Shabri, A.: Forecasting Gold Prices Using Multiple Linear Regression Method. American Journal of Applied Sciences, 6(8), 1509-1514 (2009). https://doi.org/10.3844/ajassp.2009.1509.1514.

[14] Janowiak, J.E.: An investigation of inter annual rainfall variability in Africa. Journal of Climate, 1, 240-255 (1988). https://doi.org/10.1175/1520-0442(1988)001<0240:AIOIRV>2.0.CO;2.

[15] Lough, J.M.: Variation of sea surface temperature off northeastern Australia and association with rainfall in Queensland: 1956–1987. International Journal of Climatology, 12, 765–782 (1992). https://doi.org/10.1002/joc.3370120802.

[16] Lough, J.M.: Regional indices of climate variation: temperature and rainfall in Queensland, Australia. International Journal of Climatology, 17, 55–66 (1997). https://doi.org/10.1002/(SICI)1097-0088(199701)17:1<55::AID-JOC109>3.0.CO;2-Z.

[17] Manatsa, D., Chingombe, W., Matsikwa, H., Matarira, C.H.: The superior influence of the Darwin Sea Level Pressure anomalies over ENSO as a single drought predictor in Southern Africa. Theoretical and Applied Climatology, https://doi.org/10.1007/s00704-007-0315-3.

[18] Navid, M., Niloy, N.: Multiple Linear Regressions for Predicting Rainfall for Bangladesh. Communications, 6 (1), 1-4 (2018).

[19] Nazemosadat, M.J., Cordery, I.: On the relationships between ENSO and autumn rainfall in Iran. International Journal of Climatology, 20, 47–61 (2000). https://doi.org/10.1002/(SICI)1097-0088(200001)20:1<47::AID-JOC461>3.0.CO;2-P.

[20] Neha, K., Ruchi, D.: Climatic Assessment of Rajasthan's Region for Drought with Concern of Data Mining Techniques. International Journal of Engineering Research and Application, 2(5), 1695-1697 (2012).

[21] Nitin, M., Dhawal, H.: A Survey on Rainfall Prediction Techniques. International Journal of Computer Application. 6(2), (2016).

[22] Nkrintra, S., Balaji, R., Martyn, C., Kumar, K. K.: Seasonal Forecasting of Thailand Summer Monsoon Rainfall. International Journal of Climatology, American Meteorological Society 25(5), 649-664 (2005).

[23] Özlem, T.: Monthly Rainfall Estimation Using Data-Mining Process. Applied Computational Intelligence and Soft Computing, 2012 (Article ID 698071), 6 pages (2012).

[24] Panu, U.S., Sharma, T. C.: Challenges in drought research: some perspectives and future directions, Hydrological Sciences Journal, 47(S), S19-S30 (2002).

[25] Paras, Sanjay, M.: A Simple Weather Forecasting Model Using Mathematical Regression in Bangladesh. Research Journal of Extension Education Special Issue 1, 161-168 (2012).

[26] Parthasarathy, B., Pant G.B.: The spatial and temporal relationships between Indian summer monsoon rainfall and the southern oscillation. Tellus A, 36, 269–277 (1984).

[27] Parthasarathy, B., Diaz, HF., Eischeid, J.K.: Prediction of all-Indian summer monsoon rainfall with regional and large scale parameters. Journal of Geophysical Research, 93 (D5), 5341-5350 (1988). https://doi.org/10.1029/JD093iD05p05341.

[28] Patil, R.V., Sannakki, S.S., Rajpurohit, V.S.: A Survey on Classification of Liver Diseases using Image Processing and Data Mining Techniques. International Journal of Computer Sciences and Engineering, 5(3), 29-34 (2017).

[29] Piechota, T.C., Dracup, J.A.: Drought and regional hydrologic variation in the United States: associations with the El Ni˜no–southern oscillation. Water Resources Research, 32(5), 1359–1373 (1996). https://doi.org/10.1029/96WR00353.

[30] Radhika, Y., Shashi, M.: Atmospheric Temperature prediction using support vector machines. International Journal of Computer Theory and Engineering, 1 (1), 55-58 (2009). https://doi.org/10.7763/IJCTE.2009.V1.9.

[31] Ramsundram, N., Sathya, S., Karthikeyan, S.: Comparison of Decision Tree Based Rainfall Prediction Model with Data Driven Model Considering Climatic Variables. Irrigation and Drainage Systems Engineering, https://doi.org/10.4172/2168-9768.1000175.

[32] Roz, E. P.: Water quality modeling and rainfall estimation: a data driven approach, M.S.thesis. University of Iowa, Iowa city, Iowa, USA (2011).

[33] Seleshi, Y., Demaree, G.R.: Rainfall variability in the Ethiopian and Eritrean highlands and its links with the southern oscillation index. Journal of Biogeography, 22, 945–952 (1995). https://doi.org/10.2307/2845995.

[34] Sewell, R.D., Landman, W.A.: Indo-Pacific relationships in terms of sea-surface temperature variations. International Journal of Climatology, 21, 1515–1528 (2001). https://doi.org/10.1002/joc.548.

[35] Soo-Yean, J., Sharad, S., Byunggu, Y., Dong, H.J.: Designing a Rule-Based Hourly Rainfall Prediction Model. https://doi.org/10.1109/IRI.2012.6303024.

[36] Sun, H., Furbish, D.: Annual precipitation and river discharges in Florida in response to El Ni˜no and La Ni˜na sea surface temperature anomalies. Journal of Hydrology, 199, 74–87 (1997). https://doi.org/10.1016/S0022-1694(96)03303-3.

[37] Tapley, T.D., Waylen, P.R.: Spatial variability of annual precipitation and ENSO events in western Peru. Hydrological Sciences Journal, 35, 429–446 (1990). https://doi.org/10.1080/02626669009492444.

[38] Wint, T.Z., Thinn, T.N.: Empirical Statistical Modeling of Rainfall Prediction over Myanmar. World Academy of Science, Engineering and Technology, 2(10), 3418-3421 (2008).

[39] Yassin, Z.O., Asaad, Y. S.: qualitative rainfall prediction models for central and southern Sudan using el nino–southern oscillation and Indian ocean sea surface temperature indices. International Journal of Climatology, 22, 1861–1878 (2002). https://doi.org/10.1002/joc.860.