MAUSAM, **73**, 1 (January 2022), 19-26

551.572 : 551.509.314 (540.65)

Monthly rainfall hind-cast using machine learning algorithms for Coimbatore, Tamil Nadu

S. KOKILAVANI, V. GEETHALAKSHMI*, S. P. RAMANATHAN, GA. DHEEBAKARAN,

N. K. SATHYAMOORTHY and N. MARAGATHAM

Agro Climate Research Centre, Tamil Nadu Agricultural University,

Coimbatore – 641 003, India

**Crop Management, Tamil Nadu Agricultural University,*

Coimbatore – 641 003, India

(*Received 8 June 2020, Accepted 29 September 2021*)

e mail : kokilavani.s@tnau.ac.in

सार – वर्तमान समय में विश्व जलवायु परिवर्तन के कारण वर्षा के पूर्वानुमान की सटीकता महत्वपूर्ण है। इस शोध पत्र में चार अलग-अलग मशीन लर्निंग एल्गोरिदम का उपयोग करके एक दृष्टिकोण प्रस्तुत किया गया है, जैसे : डिसीजन ट्री रिग्रेशन (DTR), ग्रेडिएंट बूस्टिंग (GB), एडा बूस्ट (AB) और रैंडम फॉरेस्ट रिग्रेशन (RFR) तकनीकों का उपयोग करके वर्षा पूर्वानुमान के कार्य में सुधार किया जा सकता है। जब ऐतिहासिक परिघटनाओं को मॉडल में दर्ज किया जाता है और यह महसूस करने के लिए मान्य किया जाता है कि आउटपुट हिंड -कास्ट के रूप में संदर्भित ज्ञात परिणामों के लिए कितना उपयुक्त है। 42 साल (1976 से 2017) की अवधि में ऐतिहासिक मासिक मौसम प्राचलों को एग्रो क्लाइमेट रिसर्च सेंटर, तमिलनाडु कृषि विश्वविद्यालय से एकत्र किया गया । वैश्विक जलवायु चालक जैसे: दक्षिणी दोलन सूचकांक और हिंद महासागर द्विध्रुव सूचकांकों को ऑस्ट्रेलिया के मौसम विज्ञान ब्यूरो से प्राप्त किया गया। प्रशिक्षण सेट के रूप में उपयोग किए जाने वाले डेटा की अतिरेक प्रकृति को समाप्त करके 42 वर्षों की अवधि के लिए मूल डेटा के 90 प्रतिशत पर केन्द्रक पहचान (जो अद्वितीय विशिष्ट विशेषताओं के साथ पंक्तियों का चयन करता है) के लिए K- साधन एल्गोरिथ्म को नियोजित किया गया। प्राप्त परिणामों ने 89.2 प्रतिशत के साथ प्रदर्शन के मामले में अन्य एल्गोरिदम पर RFR की सर्वोच्चता और उल्लेखनीय कार्य शक्ति का संक्रे त दिया। 2015 से 2017 तक मासिक वर्षा के लिए अनुमानित और प्रेक्षित मानों के लिए निर्धारण गुणांक (R2) 0.8 पाया गया।

ABSTRACT. Due to current world climate change, the accuracy of predicting rainfall is critical. This paper presents an approach using four different machine learning algorithms, *viz.*, Decision Tree Regression (DTR), Gradient Boosting (GB), Ada Boost (AB) and Random Forest Regression (RFR) techniques to improve the rainfall forecast performance. When historical events are entered into the model and get validated to realise how well the output suits the known results referred as Hind-cast. Historical monthly weather parameters over a period of 42 years (1976 to 2017) were collected from Agro Climate Research Centre, Tamil Nadu Agricultural University. The global climate driver's *viz.*, Southern Oscillation Index and Indian Ocean Dipole indices were retrieved from Bureau of Meteorology, Australia. K-means algorithm was employed for centroid identification (which select the rows with unique distinguished features) at 90 per cent of the original data for the period of 42 years by eliminating the redundancy nature of the datawhich were used as training set. The result indicated the supremacy and notable strength of RFR over the other algorithms in terms of performance with 89.2 per cent. The Co-efficient of Determination (R2) for the predicted and observed values was found to be 0.8 for the monthly rainfall from 2015 to 2017.

Key words – Rainfall, K-means, Decision tree regression, Gradient boosting, Ada boost, Random forest regression, Hindcast.

1. Introduction

Even today, weather prediction remains a research subject in which scientists and mathematicians work to create a model or algorithm that forecasts the weather accurately (Karthick *et al.*, 2018). Rainfall is a result of complex natural processes and becomes one of the main concerns in meteorological services (Purnomo *et al.*, 2017). Due to the arbitrary characteristics of rainfall series, they are often labelled as a stochastic process.

Rainfall prediction nowadays is a gruelling chore which is taking into the contemplation of most of the major world-wide authorities. Forecasting of rainfall is useful to warn about natural disaster such as flood and to plan ahead activity such as cropping pattern scheduling.

According to Shrivastava *et al*. (2012), there are two main approaches in rainfall forecasting, numerical and statistical methods. The performance of the numerical method depends on the initial condition, which is inherently incomplete. The method is poor for long-range prediction. On the other hand, the statistical method is widely used for long-term rainfall predicted and stated that statistical method performance were successful in normal monsoon rainfall but fail in extreme monsoon years.

Machine learning algorithms are now successfully used to identify, cluster or reduce the dimensionality of large collections of particularly high-dimensional input data tasks (Marsland, 2014). Abhisek *et al*. (2012) stated that for meteorology predictions, ANNs pictured as alternative method which opposed to traditional method, are based on self-adaptive mechanisms that learn from examples and capture functional relationships between data, even if the relationships between the data is unknown or difficult to describe. Deep learning is essentially a series of multilayer architecture that are trained. The main changes which impact the model are weight and the learning rate of the layers. Making a good prediction of climate is always a major task nowadays because of the changes in climate.

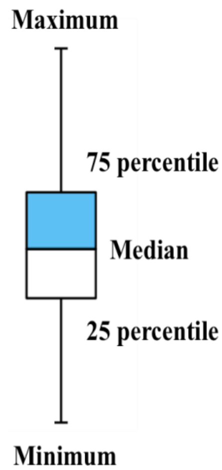
Surajit and Manojit (2007) used artificial neural network as a soft computing technique to anticipate the average monsoon rainfall over India and results have been compared with those obtained through conventional techniques.

Considering the significance of rainfall in many decision making processes, the present study aims to find out the relationship between large scale climatic signals and regional rainfall using machine learning algorithms.

2. Materials and method

In this paper, the data set under consideration contains monthly weather observations over a period of 42 years (1976 to 2017) obtained from Agro Climate Research Centre, Tamil Nadu Agricultural University, Coimbatore-3. The geo coordinates for Coimbatore is 11°N, 77° E and 427 MSL.

2.1. *Box and whiskers plot*



Box and whiskers plot are generally used to depict the descriptive statistics of a given set of data. The two whiskers (vertical lines) in top and bottom represent maximum and minimum value in the data. Upper bound of the box represents third quartile or seventy-fifth percentile. In the same way the lower bound of the box represents first quartile or twenty-fifth percentile while, the central line represents the median. The rainfall variability for the study region was depicted using this plot.

Feature	*Description*
Max. temp	Maximum temperature in degree celcius
Min. temp	Minimum temperature in degree celcius
Rainfall	The amount of rainfall recorded for the day in mm
Wind speed	Wind speed observation in KMPH
RH1	Atmospheric Humidity (per cent) at 7.22 am
RH2	Atmospheric Humidity (per cent) at 14.22 am
SOI	Southern Oscillation Index
IOD	Indian Ocean Dipole/Dipole Mode Index

2.2. *ENSO linked climate variability*

El Niño is one of the important factors leading to climate variability. Operationally ENSO conditions are defined based on sea surface temperature variations and their persistence along the equatorial Pacific Ocean. The National Oceanic and Atmospheric Administration (NOAA) define El Niño and La Niña events based on a threshold of +/- 0.5 °C for the Oceanic Niño Index (ONI) (3 months running mean of SST anomalies over equatorial eastern Pacific) (http://ggweather.com/enso/oni.htm).

2.3. *IOD linked climate variability*

This is an index based on SST anomaly difference between the Eastern and Western Tropical Indian Ocean. The index, its impact on the adjoining continental rainfall, interactions with El Nino Southern Oscillation (ENSO) and tele-connections can all be obtained from the IOD home page http://www.jasstec.go.jp/frsgc/research/d1/iod/. The ENSO and IOD data were collected for the corresponding monthly weather observations used for the study period of 42 years (1976 to 2017).

2.4. *Principal Component Analysis*

Principal Component Analysis (PCA) is a technique of extracting necessary variables from a huge set of variables. It extracts low dimensional set with a motive to capture the maximum amount of information. With few variables, visualization becomes more significant. Principal Component Analysis was performed to isolate independent factors that significantly explain the variation of a dependent variable (Wasik *et al*., 2017).

2.5. *Data pre-processing*

Data pre-processing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent and/or lacking in certain behaviors or trends and is likely to contain many errors. Isolation Forest was used to remove noise present in the data.

2.6. *K-means centroid*

The K-Means algorithm is commonly used in data mining, which is a basic clustering technique. It is a non-supervised learning algorithm used to address the well-known problem of the cluster (Jigui *et al*., 2008). Partition-based clustering is a way to cluster large data sets that first contain a number of objects and then divide these objects into a number of groups and each group contains similar data points (Neha and Vidhyavathi, 2015). Through the iterative, converging method, the K-means algorithm classifies the data into different clusters of K. The generated clusters of K-Means are independent. The K-Means clustering algorithm works in two different parts. Firstly, a K-value is chosen, where K is the number of clusters. Another part is to take each data point to the nearest centre into account (Fahim *et al*., 2006). In K-Means algorithm, the data groups are created before calculating the distance between centroid to each data point and this process continues a number of times until each data points are purely group (Nazeer and Sebastian, 2009). So the time complexity of the K-Means clustering algorithm is O (mkt). Where 'm' is the data points, 'k' is the initial centroids, 't' is the number of iterations (Zakir Hossain *et al*., 2019).

2.7. *Models*

Machine learning applications are used in retail, finance, military and health industries, etc. and researchers create various algorithms using knowledge from different fields of study to achieve their target (Neto and Castro, 2017). These algorithms can be used to build models, which acquire knowledge from previous data. For the research, different classifiers were picked, each belonging to different model families. All the classifiers were implemented using scikit-learn**.**

2.8. *Decision tree regression*

A decision tree is a tree in which each node displays a function (attribute), each branch displays a decision (rule) and each leaf denotes an outcome (categorical or continues value) (Jadhav and Channe, 2016). As decision trees imitate the reasoning of the human level, it is so easy to grab the data and make some good interpretations. The decision tree makes explicit all possible alternatives and traces each alternative to its conclusion in a single view, to make easy comparison among the various alternatives (Anyanwu *et al*., 2009).

2.9. *Gradient boosting or boosted decision tree regression*

The key idea behind gradient boosting machines algorithm is to create the new base-learners to be maximally correlated with the loss function's negative gradient, connected to the entire ensemble. This high unique property makes the machines for gradient boosting up in the era for any specific data-driven objective. It is relatively easy to implement boosting algorithms, allowing one to explore with various model designs (Johnson and Zhang, 2012).

2.10. *Adaptive Boosting (AdaBoost)*

The meta algorithm in combining with many other learning algorithms is called Adaptive Boosting (AdaBoost) algorithm. It is used to enhance classification efficiency and uses the nesting operator and it has a sub-process. The learner provided in its sub process is made use to generate a better model. The skill of the ensemble model increases the precision of classification by having more than one classifier. The ensemble model facilitates decision-making by integrating the effects of their classification techniques. AdaBoost relates to weight within the training package. At the beginning point, the weights are evenly distributed, while for the other training the weights are increased to indicate that they are not properly classified. Weak learners are trained with hard examples by this approach. The merit is that this approach will enhance the precision of the classification (Narayanan and Govindarajan, 2016).

2.11. *Random forest regression*

Random forest is an aggregation approach (Brieman, 2001) and is known to be one of the most accurate general purpose techniques. Random forests are a mixture of tree predictors such that each tree depends on the values of a

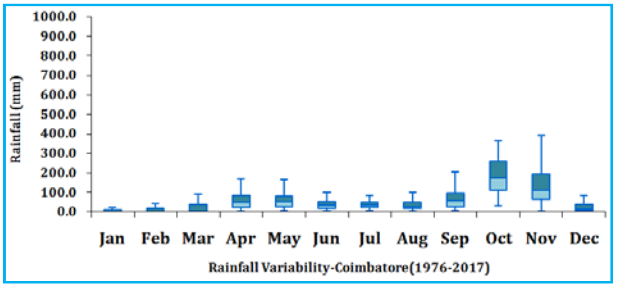


Fig. 1. Rainfall variability at Coimbatore

TABLE 1

Weightage assigned for different components using PCA

Algorithm	Weightage assigned from Component 0 to 5
Decision Tree Regression	[0.48,0.29,0.10,0.08,0.05]
Gradient Boosting	[0.30,0.25,0.18,0.16,0.11]
Ada Boosting	[0.45,0.33,0.05,0.13,0.04]
Random Forest Regression	[0.33,0.29,0.16,0.14,0.08]

random vector sampled autonomously and with the same distribution for all trees in the forest. The generalization error of a forest in tree classifiers depends on the strength of the individual trees in the forest and the association between them. A different subset of the training data is selected, with replacement, to train each tree (Jehad Ali *et al.*, 2012).

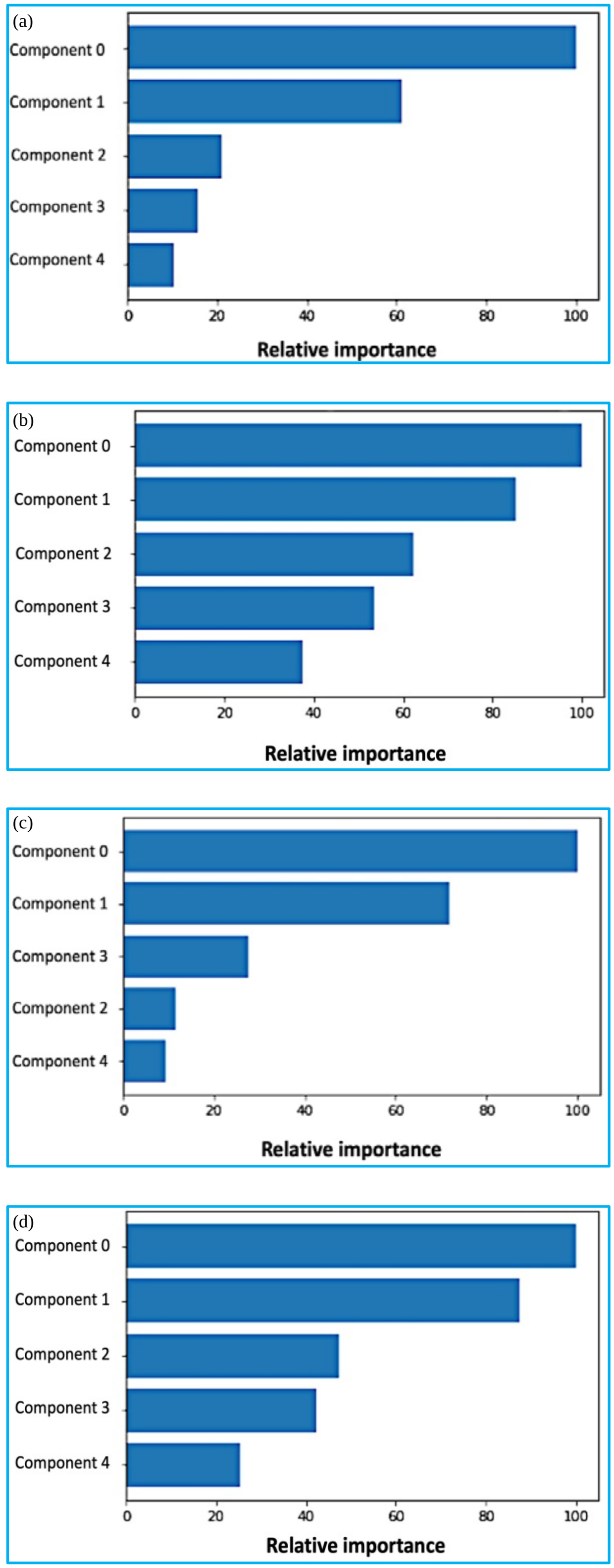
3. Results and discussion

3.1. *Temporal rainfall distribution*

A temporal analysis of the rainfall distribution for Coimbatore showed that it is not homogeneous over the period of time analyzed (1976 to 2017). The seasonality feature could be seen during pre-monsoon (April, May) and North East Monsoon (October and November) season prominently in the given time scale (Fig. 1). Similar findings were observed by Colomo *et al*. (2019).

3.2. *Principal Component Analysis (PCA) for the features*

PCA saved the most pertinent information from the multidimensional dataset and at the same time lessens its dimension. Among the features employed in the study, the PCA has reduced seven features to five features and assign different weightage for the components in the four



Figs. 2(a-d). Variable Importance (a) Decision Tree Regression, (b) Ada Boost Decision Tree Regression, (c) Gradient Boosting Regression and (d) Random Forest Regression

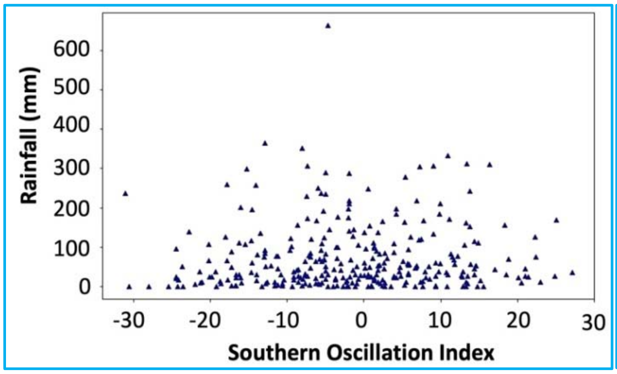


Fig. 3(a). Southern Oscillation *vs* rainfall

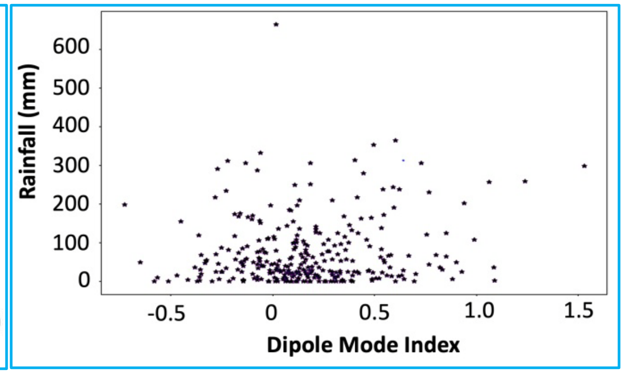


Fig. 3(b). Dipole Mode Index *vs* rainfall

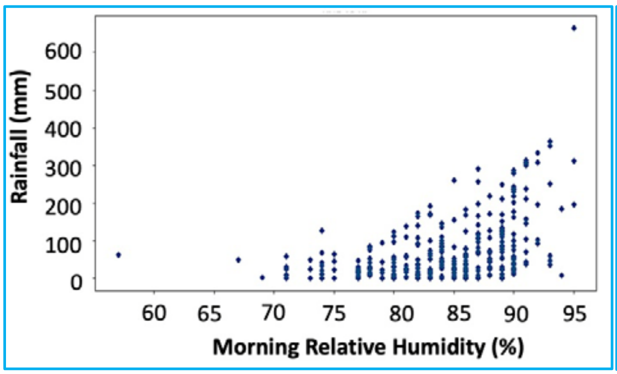


Fig. 3(c). Morning Relative Humidity *vs* rainfall

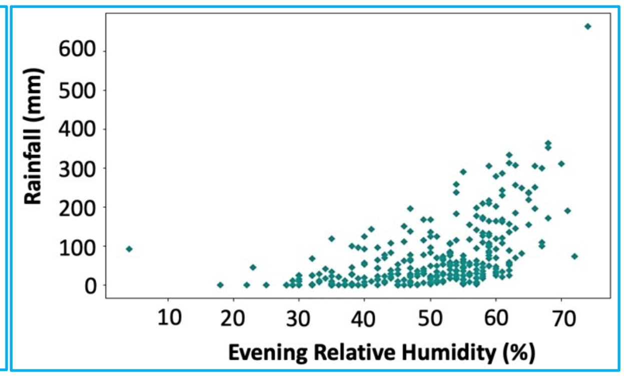


Fig. 3(d). Evening Relative Humidity *vs* rainfall

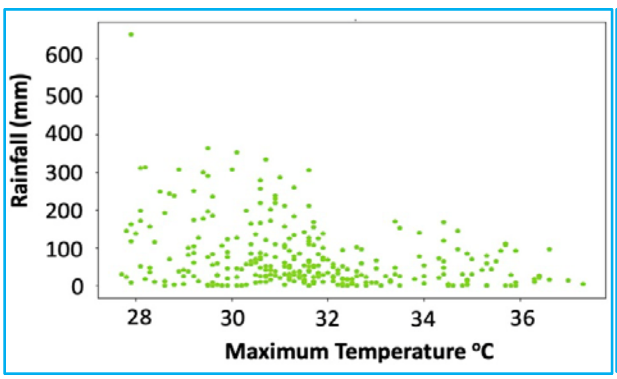


Fig. 3(e). Maximum Temperature *vs* rainfall

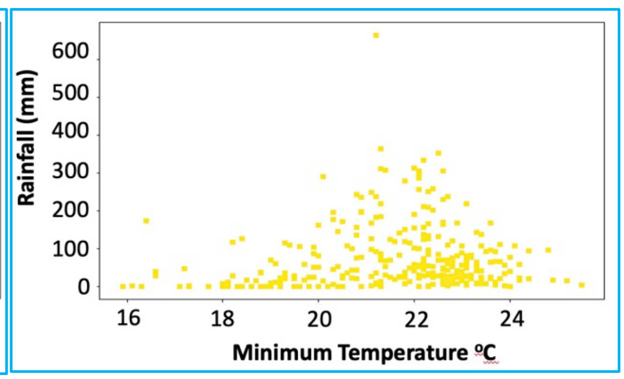


Fig. 3(f). Minimum Temperature *vs* rainfall

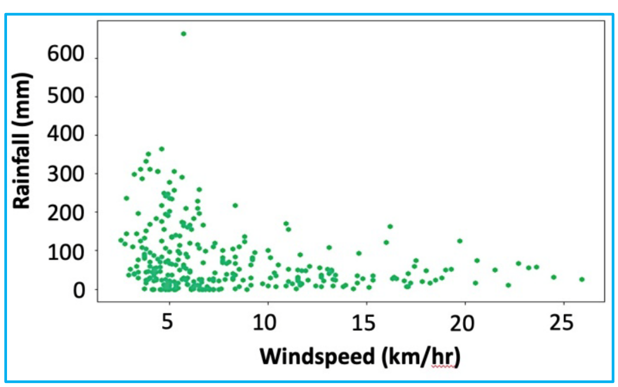


Fig. 3(g). Wind speed *vs* rainfall

machine learning algorithms [Table 1, Figs. 2(a-d)]. PCA method is potentially very well suited for neural networks training methods (Mandziuk, 2002)

Using PCA, auto-generated equations are developed for the components and are given below:

Component 0 : 0.002 × DMI + -0.693 × SOI + 0.099 × Max_Temp + -0.015 × Min_Temp + -0.208 × RelativeHumidity_1 + -0.683 × RelativeHumidity_2 + 0.021 × Wind speed

Component 1 : -0.007 × DMI + 0.709 × SOI + 0.079 × Max_Temp + -0.049 × Min_Temp + -0.073 × RelativeHumidity_1 + -0.687 × RelativeHumidity_2 + -0.105 × Wind speed

Component 2 : -0.006 × DMI + 0.132 × SOI + 0.086 × Max_Temp + 0.180 × Min_Temp + -0.749 × RelativeHumidity_1 + 0.121 × RelativeHumidity_2 + 0.606 × Wind speed

Component 3 : 0.010 × DMI + -0.006 × SOI + 0.559 × Max_Temp + 0.541 × Min_Temp + -0.243 × RelativeHumidity_1 + 0.132 × RelativeHumidity_2 + -0.565 × Wind speed

Component 4 : -0.004 × DMI + -0.010 × SOI + -0.422 × Max_Temp + -0.422 × Min_Temp + -0.576 × RelativeHumidity_1 + 0.116 × RelativeHumidity_2 + -0.547 × Wind speed

Irrespective of the different algorithm studied, the weightage assigned to component 0 was found to be higher. Higher the values, the component influence have a profound relationship with the targeted test data. These generated components were taken for training and testing data input for the machine learning algorithms.

3.3. *Relationship of rainfall with the selected features*

By employing, K- Means centroid algorithm, 90 per cent of the original dataset from the data period (42 years) with unique characteristics was taken into consideration. Out of 406 rows, 452 rows of data were selected for the training period which eliminates the redundancy nature of the data which were used as training set and are presented from Figs. 3(a-g).

TABLE 2

Accuracy level and Algorithms

Algorithm	Accuracy level (%)
Decision Tree Regression	60.0
Gradient Boosting	40.6
Ada Boost	68.7
Random Forest Regression	89.2

The relationship of rainfall was established only with the selected features of dataset and the rainfall distribution was found to be more than 300 mm during Southern Oscillation Index (SOI) values was less than -15. The relationship during NEM (October - December) for Tamil Nadu was found to be positive, *i.e*., rainfall is more than normal during El Niño and *vice-versa* during La Niña years (Geethalakshmi *et al*., 2005; Kokilavani *et al*., 2015). Similarly, whenever positive Indian Ocean Dipole or Dipole Mode Index values were recorded, the probability to receive more quantity of rainfall was witnessed in the selected features of dataset for Coimbatore.

During good showers, the morning relative humidity recorded more than 90 per cent and evening relative humidity was found to be more than 60 per cent. The maximum temperature was found to be around 28 to 30 °C and minimum temperature was found to be 20 to 22 °C in the selected features of dataset. There exists a negative relationship between rainfall and temperature (Cong and Brady, 2012). Interestingly, when the wind speed ranged from 3 to 6 Kilometer per hour, good rainfall was reported in the selected dataset. Similar findings were reported by Guo *et al*. (2011).

3.4. *Selection of best algorithm*

Based on the training data set obtained from the PCA, the machine learning algorithm creates its own intelligence in terms of numerous nodes with different weightage and assembling the mean in arriving the best accuracy level by employing scikit learn. Random Forest is a tree based ensemble method where all the trees depend on a collection of random variables. The final decision was obtained by aggregation over the ensemble and thus improved the stability and accuracy of the model with 89.2 per cent(Table 2).

3.5. *Monthly rainfall hindcast using RFR*

After the selection of best algorithm, RFR was employed to predict the monthly rainfall values for three

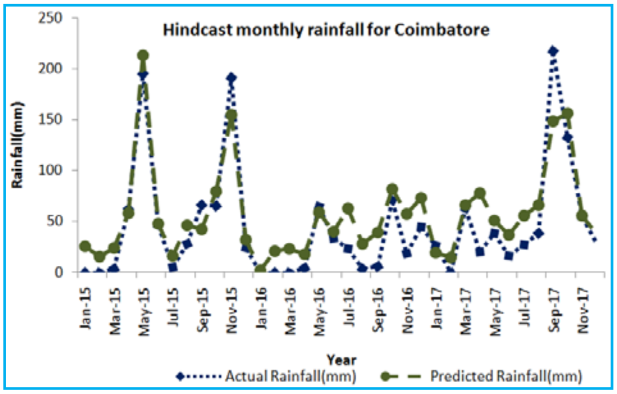


Fig. 4. Hindcast monthly rainfall for Coimbatore

Years from 2015 to 2017 (Fig. 4). The co-efficient of determination for the predicted and the observed values was run using python and found to be 0.8 which indicated the better performance of the algorithm in simulating the rainfall. By using Random Forest for three-rainfall-states classification was improved in the extreme rainfall downscaling (Pham *et al*., 2019).

4. Conclusions

In this study, it has been attempted to evaluate the different machine learning algorithms, *viz*., Decision Tree Regression (DTR), Gradient Boosting (GB), Ada Boost (AB) and Random Forest Regression (RFR) to hindcast the monthly rainfall from 2015 to 2017. By employing PCA, the feature has been reduced from seven to five numbers. K-means algorithm was employed for centroid identification (which selects the rows with unique distinguished features) at 90 per cent of the original data by eliminating the redundancy nature of the data which were used as training set. Among the four algorithms, the accuracy level was high in RFR which was found to be 89.2 per cent. The Co-efficient of Determination (R^2) for the predicted and observed values was found to be 0.8 for the monthly rainfall from 2015 to 2017 which indicated the better performance of the algorithm. For future work, focus will be emphazied to reduce the difference in the error percentage using neural network techniques.

Acknowledgement

The contents and views expressed in this research paper/article are the views of the authors and do not necessarily reflect the views of the organizations they belong to.

Disclaimer : The contents and views expressed in this study are the views of the authors and do not necessarily reflect the views of the organizations they belong to.

References

Abdul Nazeer, K. A. and Sebastian, M. P., 2009, "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm", Proceeding of the World Congress on Engineering,

Abhisek, K., Singh, M. P., Ghosh, S. and Anand, A., 2012, "Weather forecasting model using artificial Neural Network", *Procedia Technology*, **4**, 311-318.

Anyanwu, M. N. and Shiva, S. G., 2009, "Comparative analysis of serial decision tree classification algorithms", *International Journal of Computer Science and Security*, **3**, 3, 230-240

Breiman, L., 2001, "Random forests", *Mach. Learn*., **45**, 5-32.

Colomo, R. A., Dagoberto Castellanos-Nieves and Máximo Méndez,2019, "Comparative Analysis of Rainfall Prediction Models Using Machine Learning in Islands with Complex Orography : Tenerife Island", *Applied Sciences*, **9**, 4931, 1-17.

Cong, R. G. and Brady, M. V., 2012, "The Interdependence between Rainfall and Temperature : Copula Analyses", *The Scientific World Journal*, **3**, 1-11.

Fahim, M., A., Salem, M. and Torkey, F. A., 2006, "An efficient enhanced k-means clustering algorithm", *Journal of Zhejiang University Science*, **10**, 1626-1633.

Geethalakshmi, V., Bride J. M. and Huda, A. K. S., 2005, "Impact of ENSO on Tamil Nadu Rainfall", *Vatavaran*, **29**, 2, 9-16.

Guo, H., Xu, M. and Hu, Q., 2011, "Changes in near-surface wind speed in China: 1969-2005", *International J. of Climatology*, **31**, 349-358.

Jadhav, S. D. and Channe, H. P., 2016, "Efficient recommendation system using decision tree classifier and collaborative filtering", *Int. Res. J. Eng. Technol*., **3**, 2113-8.

Jehad, Ali, Khan, Rehanullah, Ahmad, Nasir and Maqsood, Imran, 2012, "Random Forests and Decision Trees", *International Journal of Computer Science*, **9**, 5, 3.

Jigui, S., Jie, L. and Lianyu, Z., 2008, "Clustering algorithms Research", *Journal of Software*, **19**, 1, 48-61.

Johnson, R. and Zhang, T., 2012, "Learning Nonlinear Functions Using Regularized Greedy Forest", Technical Report.

Karthick, S., Malathi, D. and Arun, C., 2018 "Weather Prediction Analysis Using Random Forest Algorithm", *International Journal of Pure and Applied Mathematics*, **118**, 20, 255-262.

Kokilavani, S., Ramaraj A. P. and Panneerselvam, S., 2015, "Exploring the relationship of Enso and rainfall variability over southern zone of Tamil Nadu", *International Journal of Environmental Science and Technology*, **4**, 955-965.

Mańdziuk, J., 2002. "Application of PCA method to weather prediction task", Conference : Neural Information Processing, ICONIP '02. Proceedings of the 9th International Conference.

Marsland, S., 2014, "Machine Learning" , CRC Press, Taylor & Francis Inc., Boca Raton, FL.

Narayanan, B. and Govindarajan, M., 2016, "Rainfall Prediction based on Ensemble Model", *International Journal of Innovative Research in Science, Engineering and Technology*, **5**, 5, 8237-8243.

Neha, D. and Vidyavathi, B. M., 2015, "A Survey on Applications of Data Mining using Clustering Techniques", *International Journal of Computer Applications*, **126**, 2, 7-12.

Neto, F. A. de A. and Castro, A. 2017, “A reference architecture for educational data mining”, in 2017 IEEE Frontiers in Education Conference (FIE), 1-8.

Pham, Q. B., Tao-Chang, Yang, Chen-Min Kuo, Hung-Wei Tseng and Pao-Shan Yu., 2019, “Combing Random Forest and Least Square Support Vector Regression for Improving Extreme Rainfall Downscaling”, *Water*, **11**, 451, 1-17.

Purnomo, H. D., Hartomo, K. D. and Prasetyo, S. Y. J., 2017, “Artificial Neural Network for Monthly Rainfall Rate Prediction”, IOP Conference Series : Materials Science and Engineering, 1-9.

Shrivastava, G., Karmakar, S., Guhathakurta, P. and Kowar, M. K., 2012, “Application of Artificial Neural Networks in Weather Forecasting”, A Comprehensive Literature review, *International Journal of Computer Applications*, **51**, 18, 17-29.

Surajit, C. and Manojit. 2007, “A soft computing technique in rainfall forecasting”, International Conference on IT, HIT, March 19-21.

Wasik, E., Chmielowski, K., Operacz, A. , 2017, “PCA as a data mining tools characterizing the work of nitrification reactors in the sewage treatment plant in Trepcza”,*Acta Sci. Pol. Formatio Circumiectus*, **16**, 209-222. (In Polish)

Zakir, Hossain Md., Nasim, Akhtar Md., Ahmad, R. B. and Mostafijur Rahman, 2019, “A dynamic K-means clustering for data mining”, *Indonesian Journal of Electrical Engineering and Computer Science*, **13**, 2, 5.