# Hybrid Deep Learning Approach for Monthly Rainfall Prediction Using Endogenous property


Submitted by

**Mandala Venkata Surendra**

**17CE33003**

Under the supervision of

**Prof. Rajib Maity**




**Department of Civil Engineering**
**Indian Institute of Technology Kharagpur**
**Kharagpur-721302, India**


**2020**

# CERTIFICATE

This is to certify that the Dissertation Report entitled, "**Hybrid Deep Learning Approach for Monthly Rainfall Prediction Using Endogenous property**" submitted by **Mr. Mandala Venkata Surendra** to Indian Institute of Technology, Kharagpur, India, is a record of bonafide Project work carried out by him under my supervision and guidance and is worthy of consideration for the award of the degree of Bachelor of Technology in Civil Engineering of the Institute. The Dissertation Report has fulfilled all the requirements as per the regulations of the institute and in my opinion reached the standard for submission.

# Abstract:

Among various spatio-temporal scales, seasonal or monthly prediction of rainfall over a subdivision (~ 1-2×$10^5$ sq. km) prediction is one of the most important tasks for agricultural and water resources management for a region. In general, climate and rainfall are highly non-linear and complicated phenomena that require advanced modelling strategies and improved simulation for an acceptable prediction accuracy. In recent studies, a variety of statistical and other modelling approaches have been developed to capture their endogenous properties of hydrological time series with respect to their temporal evolution. Deep Learning (DL) is an effective technique for dealing with many complex systems. This study proposes a hybrid DL approach, a combination of one-dimensional Convolutional Neural Network (Conv1D) and Multi-Layer Perceptron (MLP) (hereinafter referred to as hybrid Conv1D-MLP model), to capture the endogenous properties of sub-divisional monthly rainfall series with a goal for an improved prediction performance. The developed hybrid model is applied to three different subdivisions in India that are climatologically different from each other. The model performance is assessed through mean absolute error, root mean square error (RMSE), Pearson correlation (r), and Nash Sutcliffe coefficient of efficiency (NSE). Overall, this study establishes the potential of the proposed hybrid Conv1D-MLP model in capturing the hidden complex endogenous characteristics of regional monthly rainfall that is effective for a reliable prediction performance.



# Keywords

Rainfall prediction; Deep learning; Multi-Layer Perceptron; Convolutional Neural Network; Hybrid deep learning models.

# 1. INTRODUCTION

Precipitation is one of the important components in the hydrologic system. It is also one of the six intrinsic parts of weather prediction. In the last few decades, the spatiotemporal distribution of precipitation is getting modified as an impact of changing climate. This leads to simultaneous drought and flood like situation within a spatial distance of a few hundred kilometers. Considering the Indian mainland, many cases of extreme rainfall events had been recorded in the southern, east-central, northern and north-western parts of the country. For instance, Gujarat and Maharashtra in 2005, Ladakh in 2010, Uttarakhand in 2013, Tamil Nadu and Puducherry in 2015 and Kerala in 2018 received unusually heavy rainfall that led to a huge loss of life and property.

There are several methods available with their own merits and demerits. Many of them require detailed information on the physical processes responsible for the occurrence of rainfall, and also their simulations are computationally challenging. Recent developments in Artificial Intelligence (AI)/Machine Learning (ML) are proven to be highly potential approaches in understanding many such complex phenomena with a wide range of computational challenges, such as image processing, sequence learning, speech recognition, communication network etc. The application of AI/ML in hydrology has a long history since the 1990s. In these approaches, inherent physical processes are implicitly considered through AI without any explicit requirement as it is in physical and conceptual models. Such approaches are widely used in hydroclimatic modelling viz. rainfall and drought prediction, extreme climate event identification and water demand forecast. Methods like Artificial Neural Networks (ANNs), Support Vector Regression (SVR), Gene Expression Programming (GEP) etc. are being successfully used recently. Charaniya and Dudul [13] proposed two different ANN models for forecasting consecutive rainfall values based on previous day lagged rainfall values. In this study, a pattern recognition approach of the neural network was adopted to extract the relevant spatiotemporal feature of historical rainfall data Lee et al. [1] presented an ANN

algorithm for forecasting early summer rainfall for the Geum River Basin in South Korea. In this analysis, the observed rainfall data was assumed to follow a normal distribution, and it was classified into three categories viz. below, near and above normal. The model gave very poor performance in classifying the above and below the normal category of rainfall. Several studies have effectively carried out rainfall forecasting and reported the applicability of Machine Learning (ML) algorithms.

In contrast to all the aforementioned computing techniques, Deep Learning (DL) is one of the recently popularized AI approaches that is beneficial in many multi-dimensional aspects. It has the ability to utilize raw data and automatically extracts the data features using successive layer representation. These layers contribute information to the model and provide the flexibility to use multiple hierarchical layers and learn from exposure to data without any human expertise. Several DL techniques viz. Multi-Layer Perceptron (MLP), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Long-Short Term Memory (LSTM) etc., so far has been applied successfully in various fields such as in image recognition [3], [4], speech recognition [2], [5], medical science [6], language understanding [7], rainfall forecasting [8] etc., and it has outperformed the existing AI/ML algorithms. These results suggest that DL may have potential applicability in many different domains, which has motivated researchers to explore and apply it for forecasting hydrologic variables at various spatiotemporal scales.

This report provides a literature review on rainfall prediction using different DL approaches used by different researchers. The report also discusses briefly about the concept of one dimensional convolution neural network architecture. The objective of this study is to develop a hybrid DL approach, a combination of one-dimensional Convolutional Neural Network (Conv1D) and Multi-Layer Perceptron (MLP) (hereinafter referred to as hybrid Conv1D-MLP model), to capture the endogenous properties of sub-divisional monthly rainfall series with a goal for an improved prediction performance. The rest of the paper is organized as follows. In Section 2, we discuss the existing research work in the rainfall prediction system. In Section 3, the proposed

system implemented on the dataset is discussed. Section 4 describes the experimental results and analysis. Finally, in the last Section 5 the work is concluded along with a discussion of some future possibilities.

# 2. LITERATURE REVIEW

Prediction methods have come a long way, from relying on an individual's experience to simple numeric methods to complex atmospheric models. Although machine learning algorithms like Artificial Neural Network (ANN) have been utilized by researchers to forecast rainfall. But studies on deep learning shows that DL may have potential applicability in many different domains and forecasting hydrologic variables at various spatiotemporal scales. For instance, Liu et al. [9] presented a DL based deep MLP approach to process a huge weather dataset which was used to forecast the weather for the next 24 hour. It was the first DL based study performed for detecting climate extremes. Zhang et al. [10] presented a DL based deep belief network algorithm for forecasting next day precipitation using seven environmental factors from the previous day. They found a better accuracy in the forecast as compared with various ML and statistical algorithms. However, there were several days for which the forecast was not reasonably good. Aswin et al. [11] presented DL based architectures, namely LSTM and CNN for prediction of rainfall magnitude. In this study, rainfall dataset for January month (1979-2018), from the Global Precipitation Climatology Project (GPCP) was used. Both the DL architectures were trained and optimized on this Global Average Monthly (GAM) dataset. The proposed architecture predicted the GAM rainfall value. However, both the architectures have a similar root mean squared error (RMSE) indicating a scope of improvement in both the architectures. Haidar and Verma [8] used a DL based one-dimensional deep CNN approach (Conv 1D) to forecast the monthly rainfall at Innisfail, Australia. In this study, eleven climate indices and sunspot values were used as the predictors. The obtained result was compared with MLP and the forecasting model of the Bureau of Meteorology, Australia. The analysis revealed that Conv 1D model

performance was better for the months having higher annual mean whereas it was not good for months having lower annual mean of rainfall. These studies form the motivation of this study, i.e. to explore the potential of the DL approaches in hydrometeorological studies. Hydrometeorological prediction of daily rainfall prediction is considered in this study. Objective of this study is to extract the hidden sequential information in the sub-divisional monthly rainfall series in order to develop a monthly rainfall prediction model. Following specific contributions are made in this study:

- This study proposes a hybrid DL approach, a combination of one-dimensional Convolutional Neural Network (Conv1D) and Multi-Layer Perceptron (MLP) (hereinafter referred to as hybrid Conv1D-MLP model), for monthly rainfall prediction of different subdivisions of India.
- The hybrid model is trained on a dataset of rainfall data of India's different subdivisions from year 1951-2000 in such a way that the model uses previous months' data as input with different lag periods to obtain the consecutive month rainfall data.
- Next, the trained model is tested on a dataset containing data from 2001-2015 and predicted the monthly data in time series form for some of the subdivisions.

# 3. PROPOSED METHODOLOGY

## A. MODEL ARCHITECTURE

The proposed hybrid Conv1D-MLP model is developed in the Jupyter notebook using Keras, which is a powerful library for large scale DL algorithms. The developed model is a sequential type. A sample schematic diagram of the model is shown in Fig 1. The first part consists of a Conv1D network and the second part consists of an MLP network. Conv1D is a type of CNN used in various applications including sequence prediction problems like time series analysis and forecasting. It comprises the input layer, a fully connected output layer with activation function,and between them, there is an arbitrary number of hidden layer(s) along with the activation functions. The function of the input

layer is to receive the signal (input data) and transfer it to the hidden layer. Hidden layers are the computational engine of the model. These may have one or more layers of the Conv1D layer, flatten layer and fully connected layers.

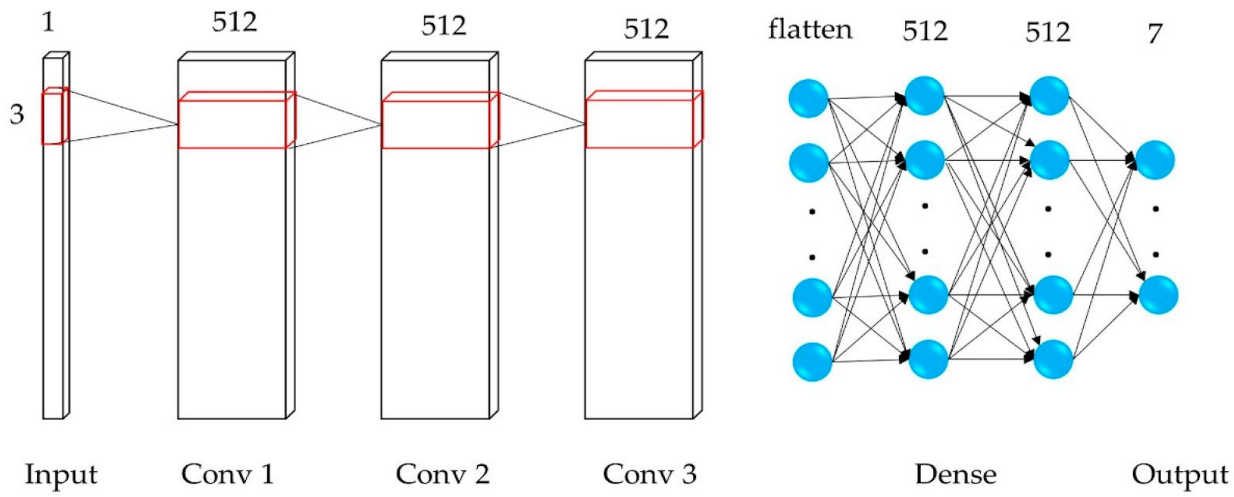


**Figure 1**.(Chunhua Liao et.al[14],2020,p.8)A sample 1D CNN configuration with 3 CNN and 2 MLP layers

Basic functioning and process of Convolution Neural networks is described below:

Basically for every model there are 2 sets i.e. Training dataset and Validation/test set which are given as input and obtain trained CNN model. Training starts with the initialization of network weights and bias. After initialization the process runs for the number of epochs required. In each epoch, the model processes the records of the training data cases, compares the actual values to the predicted values and calculates the loss function, all this process is known as forward propagation calculation. Now the model backpropagates the error through the layers and adjust the network weights.After updating the weights the model is validated on the validation dataset and check if better loss value is obtained and if so it saves the network weights and runs for epochs the model provided with. And finally after completion of all the iterations the trained model is obtained. Now we test this model on a test set.

The Conv1D layer is the main building block of CNN. It consists of filters to extract features from the input signal and kernels to specify the height of the filter. The model is

trained on the defined dimension and extracts the hidden information of the sequence.MLP network receives the input from the Conv1D model (Fig. 1). It is also a fully connected ANN that receives the data in a one-dimensional vector form. Therefore, after the Conv1D network a flattened layer is added. Next, a fully connected dense layer along with the activation functions are added. The fully connected layers have a more number of neurons than output layers. In this way, neural networks are allowed to think wider before they converge to the output layer.There are several parameters to be specified to fix the model architecture. This is problem specific and details are provided in the results and discussion section of this report. After configuring the layers, the model is trained on a set of input and output data to learn the relationship between them. Training involves adjustments of weights and biases (parameters) to minimize the error. It is achieved through backpropagation, which adjusts the parameters considering the error (loss) in the predictions. There are several error based metrics, e.g., mean squared error (MSE) and log loss etc. Once the model is properly trained, it is ready for further use.

## B. DATA DESCRIPTION

Rainfall prediction is clearly of great importance for India. India Meteorological Department (IMD) provides monthly rainfall data that is used in this study. Main focus is to predict the amount of rainfall in a particular division or state well in advance using the temporal evolution of time series hidden in the past data.

The dataset consists of the monthly rainfall for the period 1901-2015 for each state in India. The data for this study were obtained from IMD. (URL: https://data.gov.in/resources/subdivision-wise-Rainfall-and-its-departure-1901-2015 accessed in November 2020). The selected dataset contains 19 attributes (individual months, annual, and combinations of three consecutive months) for 36 sub divisions from 1901 to 2015. The data from 1951–2000 were used to train the different models, and the data from 2001–2015 were used for testing. Previous months rainfall data is used as input to predict the consecutive month rainfall, some of these monthly data had either zero or

very few missing values that were handled during data preprocessing. Three different lags, i.e., 3, 6 and 9 months are considered. Lag indicates the gap (number of months) between input i.e. past data and the starting month of prediction. For example, if we consider 3 months lag we use rainfall data of January, February and March to predict the rainfall of April of 2001 and it goes on predicting the rainfall of all the consecutive months in the time series model until 2015. In the similar way 6, 9 months lag is also considered in this study. Variation of monthly rainfall, seasonal rainfall of India is distributed over years from 1951-2000 is shown in Figures 2 and 3. The correlation coefficients between rainfall in different months and seasons and the same with annual rainfall are shown in the heat maps (Figures 4 and 5).

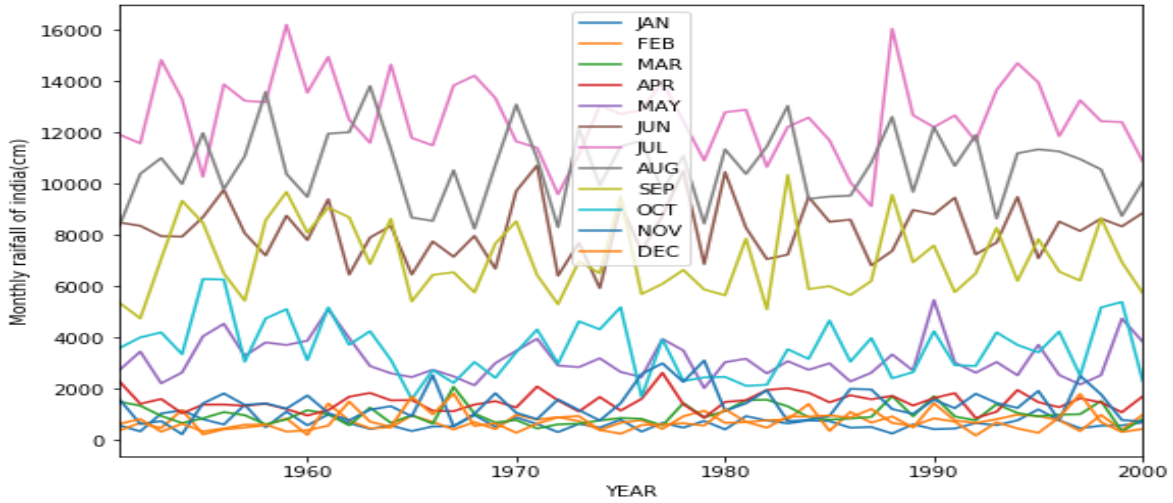


**Figure 2**: Variation of monthly rainfall of India

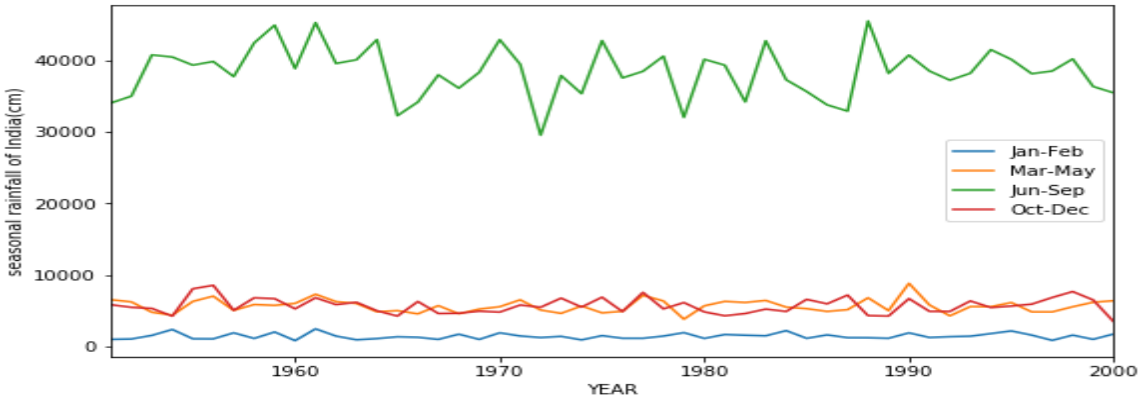


**Figure 3**: Variation of seasonal rainfall of India

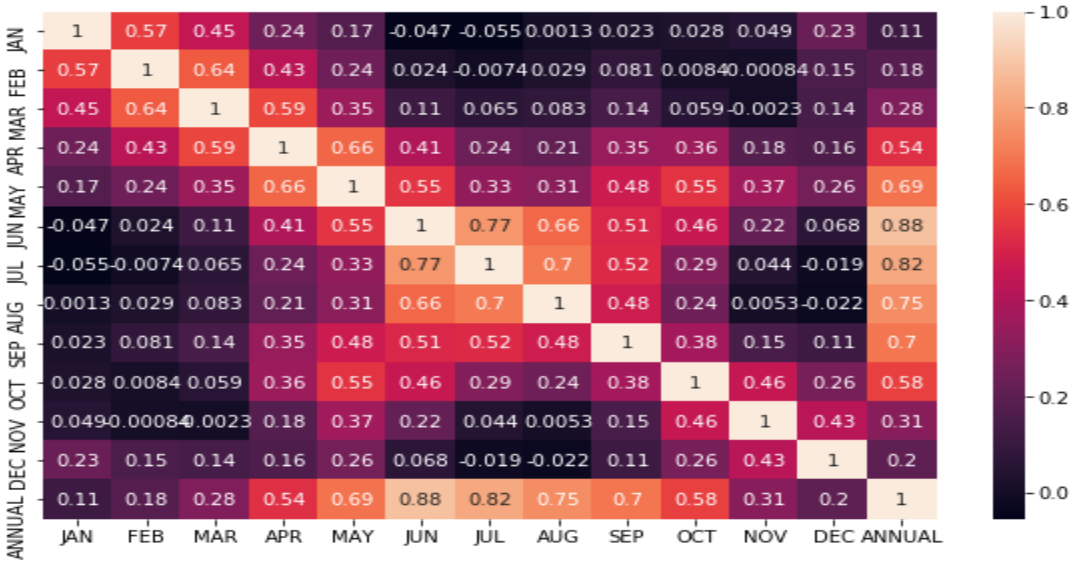


**Figure 4.** Heat map showing the correlation coefficients between rainfall in different months and the same with annual rainfall


**Figure 5.** Heat map showing the correlation coefficients between rainfall in different seasons and the same with annual rainfall

## C. PERFORMANCE EVALUATION CRITERIA

The performance of the hybrid Conv1D-MLP model is evaluated through three statistical measures viz. Root Mean Squared Error (RMSE), coefficient of correlation (r), Nash–Sutcliffe Efficiency (NSE). The coefficient of correlation is a measure of linear association between two variables. The value of r is computed as:

$$r = \frac{\sum_{t=1}^{n}(Y_t - \overline{Y})(Y'_t - \overline{Y'})}{\sum_{t=1}^{n}(Y_t - \overline{Y}) \sum_{t=1}^{n}(Y'_t - \overline{Y'})}$$

The value of r ranges between [−1,1], where −1 represents a perfect negative linear association, 0 denotes no linear association, and 1 represents a perfect positive linear association. Higher the value of r, the better the model performs.

RMSE is used frequently to measure the difference between observed and predicted values. It is always positive and a lower RMSE indicates a better model performance. RMSE is expressed as:

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{n}(Y_t - Y'_t)^2}{n}}$$

NSE is used to measure the efficiency of the proposed model. The value NSE is computed by the equation:

$$\text{NSE} = 1 - \frac{\sum_{t=1}^{n}(Y_t - Y'_t)^2}{\sum_{t=1}^{n}(Y_t - \overline{Y})^2}$$

NSE ranges between (−∞, 1]. A value greater than 0 indicates a better efficiency of the model as compared to a value equal to 0 that signifies the predicted values are as good as the mean of the observed values. NSE values less than zero indicates an unacceptable model performance.

# 4. RESULTS AND DISCUSSIONS

The models were created in python on the Jupyter notebook using Keras (https://github.com/surendra093/Rainfall_prediction/blob/main/Rainfall%20prediction.ipynb) deep learning. Several network architectures were evaluated by varying model parameters (viz. number of hidden layers, number of filters, kernel size) and optimizing several hyperparameters (viz. learning rate, batch size, number of epochs, loss functions, and activation functions) in order to ascertain the best possible architectural configuration. All the experiments were run for 10 epochs, but by using callbacks in Keras only the best weight for each test run was saved. The finalized architecture of the proposed hybrid model comprises seven layers. Details of the finalized configurations are shown in Table 1.

**TABLE 1**. Configurations of the proposed hybrid Conv1D-MLP model

| Layer no. | Layer | Type | Parameters of layers | | | |
|---|---|---|---|---|---|---|
| | | | Activation func. | Kernel Size | No.of filters | Neurons |
| 1 | Conv1D | Convolution Layer | ReLU | 1 | 64 | - |
| 2 | Conv1D | Convolution Layer | ReLU | 2 | 128 | - |
| 3 | Flatten | Flatten Layer | - | - | - | - |
| 4 | dense | Fully connected Layer | ReLU | - | - | 128 |
| 5 | dense | Fully connected Layer | ReLU | - | - | 64 |
| 6 | dense | Fully connected Layer | ReLU | - | - | 32 |
| 7 | dense | Fully connected Layer (output layer) | Linear | - | - | 1 |

Next, the trained model with the aforementioned architecture is tested on three subdivisions of India i.e Telangana, Orissa and Jharkhand and observed the predicted rainfall for all the months from 2001-2015 in time series fashion and the model performance is evaluated using RMSE, coefficient of correlation and NSE values. Model can be tested on any of the subdivisions given in the test set.

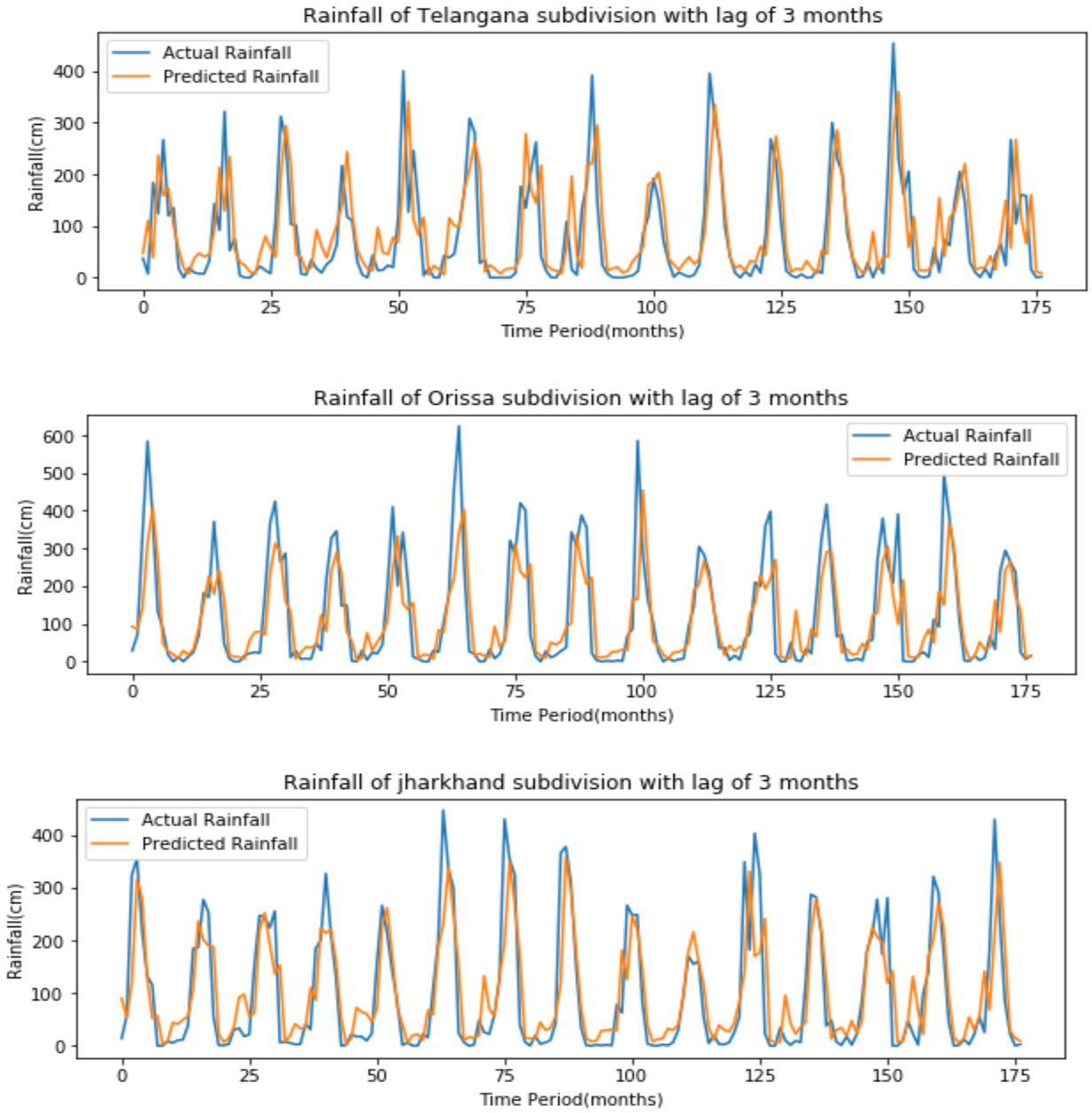


**Figure 6.** Comparison plots between actual and predicted monthly using inputs from 3 previous months for 3 different subdivisions. X-axis shows the months from the year 2001 to 2015.

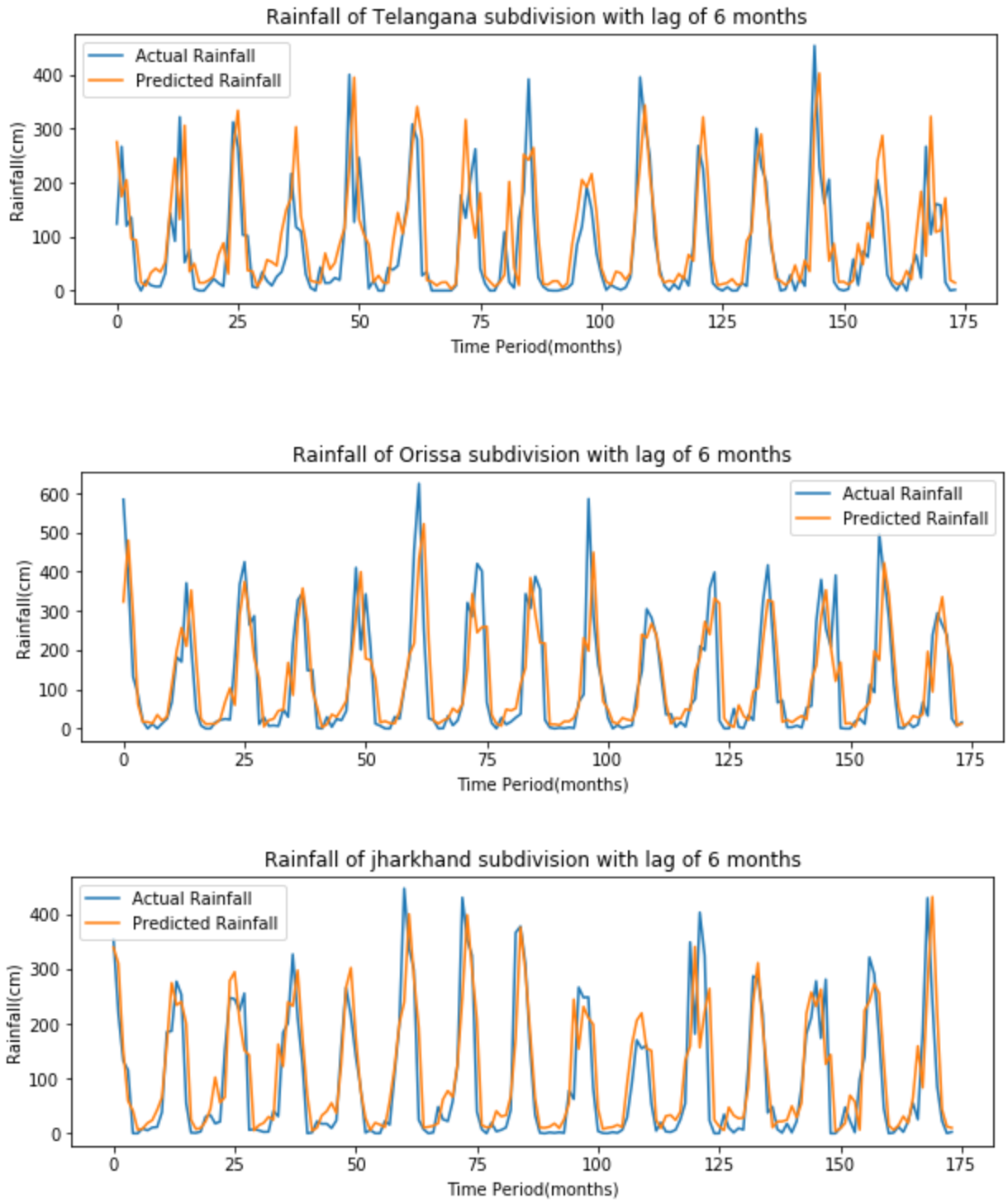


**Figure 7** Comparison plots between actual and predicted monthly using inputs from 6 previous months for 3 different subdivisions. X-axis shows the months from the year 2001 to 2015.

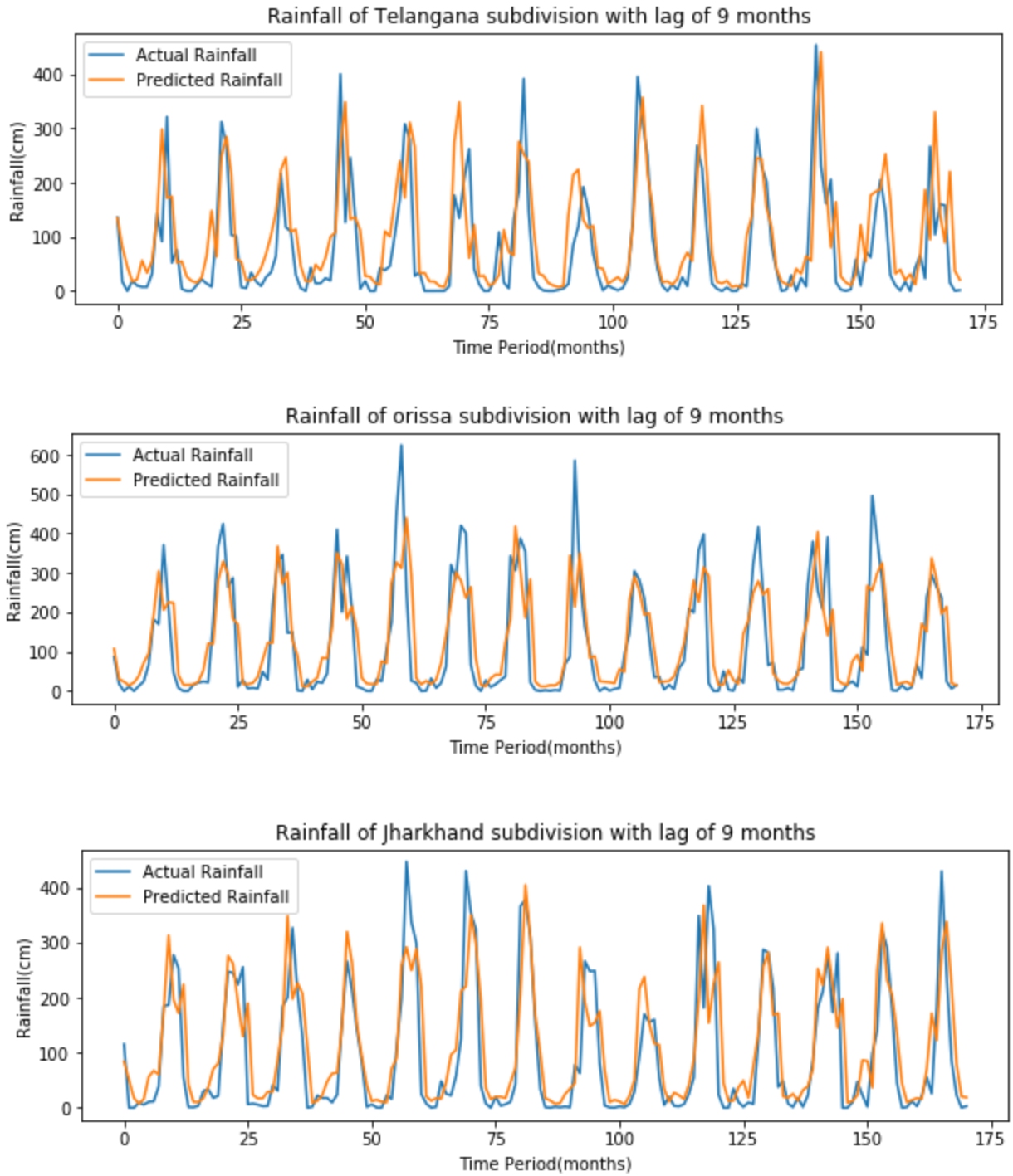


**Figure 8** Comparison plots between actual and predicted monthly using inputs from 9 previous months for 3 different subdivisions. X-axis shows the months from the year 2001 to 2015.

**TABLE 2:** Performance statistics viz. r, RMSE and NSE for different lag periods at different subdivisions.

| Subdivision | No.of inputs/ Lag months | Performance Statistics | | |
|---|---|---|---|---|
| | | RMSE | Correlation coefficient(r) | NSE coefficient |
| Telangana | 3 | 74.25 | 0.686 | 0.445 |
| | 6 | 74.10 | 0.708 | 0.399 |
| | 9 | 73.36 | 0.728 | 0.459 |
| Orissa | 3 | 100.74 | 0.747 | 0.545 |
| | 6 | 95.68 | 0.770 | 0.590 |
| | 9 | 92.2 | 0.773 | 0.597 |
| Jharkhand | 3 | 71.87 | 0.803 | 0.642 |
| | 6 | 68.8 | 0.824 | 0.670 |
| | 9 | 67.65 | 0.806 | 0.676 |

The performance statistics shown in Table 2, indicates that there is an increase in the performance with the increase in no. of inputs. Considering the 9 months lag to predict the monthly rainfall NSE coefficient which indicates the better performance of the model with 9 inputs. NSE values lie between 0.44 - 0.46 for Telangana subdivision, 0.54 - 0.6 for Orissa subdivision and 0.64 - 0.68 for Jharkhand subdivision, all of them have NSE values greater than 0 which indicates better efficiency of model. Moreover, the root mean square error of the model decreases with the increase in no. of inputs. Coefficients of correlation considering all the tested subdivisions lie between 0.6 - 0.8 which indicates all the models with different input sets gives good performance but with 9-month lag input best results are obtained. Also, the model yields better results for some of the

subdivisions that is reflected through the above performance statistics table. We observe that the model performance is best in case of Jharkhand subdivision when compared to other two subdivisions, i.e. Telangana and Orissa. Therefore, spatial variation of predictability exists in some cases in which the model is sensitive towards different subdivisions.

# 5. CONCLUSIONS

This study presents the potential of a proposed hybrid DL approach, namely hybrid Conv1D-MLP model, for monthly rainfall prediction of daily rainfall using past data as the input variables. From the performance of the proposed model, it is concluded that deep learning has the potential to capture the non-linear relationship between the past data and monthly rainfall variability. So in general, the predictions obtained from the proposed hybrid DL model can be helpful in agriculture, irrigation scheduling, and even flooding due to heavy rainfall.

Future plan of work includes i) comparison with other machine learning models, such as SVR, ii) application to other sub-divisions and iii) month-wise analysis of predictability. Moreover, there is a huge scope to apply the hybrid DL approach for spatial variation in long-lead predictability of monthly rainfall using global climatic indices which is also kept as future scope of study.

.

# REFERENCES

[1] J. Lee, C.-G. Kim, J. Lee, N. Kim, and H. Kim, ‘‘Application of artificial neural networks to rainfall forecasting in the geum river basin, korea,’’ Water, vol. 10, no. 10, p. 1448, 2018.

[2] T. Mikolov, A. Deoras, D. Povey, L. Burget, and J. Cernocky, ‘‘Strategies for training large scale neural network language models,’’ in Proc. IEEE Workshop Autom. Speech Recognit. Understand., Dec. 2011, pp. 196–201

[3] R. V. Ramana, B. Krishna, S. R. Kumar, and N. G. Pandey, ‘‘Monthly rainfall prediction using wavelet neural network analysis,’’ Water Resour. Manage., vol. 27, no. 10, pp. 3697–3711, Aug. 2013.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ‘‘ImageNet classification with deep convolutional neural networks,’’ in Proc. 25th Int. Conf. Neural Inf. Process. Syst., 2012, vol. 60, no. 6, pp. 1097–1105.

[5] T. N. Sainath, A.-R. Mohamed, B. Kingsbury, and B. Ramabhadran, ‘‘Deep convolutional neural networks for LVCSR,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., May 2013, pp. 8614–8618.

[6] J. Ma, R. P. Sheridan, A. Liaw, G. E. Dahl, and V. Svetnik, ‘‘Deep neural nets as a method for quantitative structure-activity relationships,’’ J. Chem. Inf. Model., vol. 55, no. 2, pp. 263–274, 2015.

[7] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, ‘‘Natural language processing (Almost) from scratch,’’ J. Mach. Learn. Res., vol. 12, pp. 2493–2537, Aug. 2011.

[8] A. Haidar and B. Verma, ‘‘Monthly rainfall forecasting using one dimensional deep convolutional neural network,’’ IEEE Access, vol. 6, pp. 69053–69063, 2018.

[9] J. N. K. Liu, Y. Hu, J. J. You, and P. W. Chan, ‘‘Deep neural network based feature representation for weather forecasting,’’ in Proc. Int. Conf. Artif. Intell. (ICAI), 2014, p.1.

[10] P. Zhang, L. Zhang, H. Leung, and J. Wang, ‘‘A deep-learning based precipitation forecasting approach using multiple environmental factors,’’ in Proc. IEEE Int. Congr. Big Data (BigData Congress), Jun. 2017, pp. 193–200

[11] S. Aswin, P. Geetha, and R. Vinayakumar, ‘‘Deep learning models for the prediction of rainfall,’’ in Proc. Int. Conf. Commun. Signal Process. (ICCSP), Apr. 2018, pp. 657–661.

[12] M. I. Khan and R. Maity, ”Hybrid Deep Learning Approach for Multi-Step-Ahead Daily Rainfall Prediction Using GCM Simulations” IEEE Access March 16 ,2020.

[13] N. A. Charaniya and S. V. Dudul, ‘‘Design of neural network models for daily rainfall prediction,’’ Int. J. Comput. Appl., vol. 61, no. 14, pp. 23–27,2013.

[14] Chunhua Liao et. al,”Synergistic Use of Multi-Temporal RADARSAT-2 and VENμS Data for Crop Classification Based on 1D Convolutional Neural Network”,p.9,2020.