SUMMER RESEARCH FELLOWSHIP (SRF) 2021

USING DEEP LEARNING METHODS TO STATISTICALLY DOWNSCALE PRECIPITATION REANALYSIS DATA

INDIAN ACADEMY OF SCIENCES, BENGALURU
INDIAN NATIONAL SCIENCE ACADEMY, NEW DELHI
THE NATIONAL ACADEMY OF SCIENCES INDIA, PRAYAGRAJ

SUMMER FELLOWSHIP REPORT
SUBMITTED BY,

Hamsaa Sayeekrishnan *(EPSS565),*
420AS2144 (Roll No.),
M.Sc. Atmospheric Sciences,
NIT Rourkela.

UNDER THE SUPERVISION OF,

DR. K.V. RAMESH,
SENIOR PRINCIPAL SCIENTIST,
CSIR FOURTH PARADIGM INSTITUTE, BENGALURU

MAY 2021

ACKNOWLEDGEMENT

I would like to express my sincere gratitude and regard to my summer research fellowship supervisor Dr. K.V. Ramesh, for the continuous support, advice, and guidance throughout this project. Without his invaluable time, guidance and encouragement, the task would not be completed. The vast knowledge and extraordinary guidance of my supervisor helped me with each phase of the summer project. I am overwhelmed by all the support and enthusiasm; it has helped me shape my career.

I take this opportunity to express my sincere gratitude to Kiran Kumar (Ph.D. scholar) for his support and guidance to complete the task.

I am also thankful to the Indian Academy of Sciences, Bengaluru, Indian National Science Academy, New Delhi, and The National Academy of Sciences India, Prayagraj for giving me the opportunity to carry out the summer research in the prestigious CSIR fourth paradigm institute.

Last, but not least, I would like to express a heartfelt thanks to my family and friends for their unconditional support and encouragement while carrying out the research work.

February 2022 *Hamsaa Sayeekrishnan*

CONTENTS

ABSTRACT

The assessment of precipitation under the changing climate is important as the spatial and temporal variation of precipitation is highly influenced by changes in climate In this study, Statistical downscaling is performed using a deep learning technique on low-resolution reanalysis data**.** Statistical downscaling is a technique in which the data over a larger grid and time period is used to determine the data on a smaller grid scale. Studies revealed that downscaling models created using machine learning approaches outperform downscaling models created using classic statistical regression techniques. The National Centers for Environmental Prediction (NCEP) reanalysis precipitation data is statistically downscaled for the period of 1951 to 2020 using Long Short Term Memory Deep Learning Model, by using High-resolution IMD observation data and the coarser-resolution NCEP reanalysis as the predictor variables and the higher resolution reanalysis data as the predictand variable.

Keywords: Statistical Downscaling , Machine Learning

INTRODUCTION

The assessment of precipitation under the changing climate is important as the spatial and temporal variation of precipitation is highly influenced by changes of climate. The NCEP reanalysis data is one of the best sources for gridded precipitation reanalysis data, but likely of a coarser resolution of 2.5 ˚by 2.5˚ when compared to observation data, like, 0.25˚ by 0.25˚ for the Indian Meteorological Department's (IMD) gridded precipitation rainfall data. In order to bridge the spatial gap between the coarser resolution and finer resolution data, statistical and dynamical downscaling methods have been developed. In statistical downscaling, empirical statistical relationships between coarser and finer resolution precipitation data are developed to bridge the spatial scale gap between NCEP reanalysis precipitation data and IMD observation data. In dynamic downscaling, physics based equations are used for the same purpose Statistical downscaling has gained wide popularity due to its low computational cost and simplicity.

Statistical downscaling is a technique in which the data over a larger grid and time period is used to determine the data on a smaller grid scale. This model is usually of linear regression type between the Predictand and the Predictor Variables but in cases where the relationship is not linear, non-linear and non-parametric regression methods are needed. Deep learning models, like Support Vector Machines (SVM), Artificial Neural Network (ANN), Principal Component Regression Long short-term memory (LSTM) model, in such cases can be employed to statistically downscale the data (Tahir et.al, 2018).

Statistical downscaling approaches can be split into three categories, according to Wilby et al (2004): regression-based approaches, weather classification-based approaches, and approaches based on weather generators. Because of their ease of use, regression-based statistical downscaling procedures have gained favor among the three groups above. Multi Linear Regression (MLR), Generalized Linear Models (GLMs), Artificial Neural Networks (ANNs), Support Vector Machine (SVM), Relevance Vector Machine (RVM), Genetic Programming (GP), and Gene Expression Programming are some of the regression approaches commonly utilized in statistical downscaling (GEP). Techniques like ANN, SVM, RVM, and GP are generally referred to as machine learning techniques because of their ability to learn from data and their use in computer algorithms.

Studies comparing the performance of different downscaling algorithms produced with machine learning techniques and classical statistical techniques have been recorded in the historical literature. Coulibaly (2004) discovered that GP-based downscaling models were able to better mimic both daily minimum and maximum temperature than MLR-based downscaling models in a downscaling experiment. Sachindra et al (2013) discovered that a Least Square Support Vector Machine (LSSVM)-based downscaling model was able to better represent the observed streamflow than an MLR-based model in a streamflow downscaling investigation. In comparison to ANN and MLR-based models, Duhan and Pandey (2015) discovered that SVM-based downscaling models are superior at reproducing observed monthly maximum and minimum temperatures.

For downscaling large scale atmospheric variables to monthly precipitation, Goly et al (2014) used MLR, positive coefficient regression (PCR), stepwise regression (SR), and SVM, and found that SVM-based downscaling models outperform models developed with all other techniques in simulating statistics of monthly observed precipitation. According to the experiments mentioned above, downscaling models created using machine learning approaches outperform downscaling models created using classic statistical regression techniques.

OBJECTIVE

The aim of this study is to use deep learning to convert low-resolution NCEP precipitation reanalysis data to high-resolution data. In this study, the Predictand variable is the high-resolution rainfall reanalysis data with the predictor being the observed rainfall data and the low-resolution rainfall reanalysis data.

DATA AND METHODOLOGY

Data

India is taken as the study region located between latitudes 8˚4'N to 37˚6'N andlongitudes 68˚7'E and 97˚25'E. The data consists of 70 years (1951 to 2020) daily precipitation from IMD (0.25˚by 0.25˚) and NCEP reanalysis data (2.5 ˚by 2.5˚). The IMD gridded data forIndian region is of 25kms (0.25 degree) resolution. This observation data is considered to be the true data. *(2022, https://www.imdpune.gov.in/Clim_Pred_LRF_New/Grided_Data_Download.ht ml).* The NCEP/NCAR Reanalysis data set is a continually updated (1948– present) globally gridded data set that represents the state of the Earth's atmosphere, incorporating observations and numerical weather prediction (NWP) model output from 1948 to present. It is a joint product from the National Centre for Environmental Prediction (NCEP) and the National Centre for Atmospheric Research (NCAR). This low-resolution reanalysis data will be re-gridded and statistically downscaled *(https://psl.noaa.gov/data/gridded/data.ncep.reanalysis.surface.html).*

Methodology

Error statistics was compared between the IMD and NCEP reanalysis data for the year 2020. The root mean square error (RMSE), Mean absolute error (MAE)and Pearson's correlation score, between the true data (IMD) and the interpolated data (NCEP Reanalysis) are evaluated to produce the error statistics. The years from 1951 to 2020 are classified then, based on normal, excess, drought, El-Nino and La- Nina years to prepare train and test data. An LSTM model is developed to establish the non-linear relationship between the true and the interpolated NCEP data and necessary corrections are applied.

LSTM Network:

Deep learning, also known as deep structured learning, is a type of neural network with numerous layers. These networks outperform regular neural networks when it comes to retaining information from prior events. A recurrent neural network (RNN) is a system that loops a number of different networks. The information is preserved thanks to the looped networks. Each network in the loop receives data and input from the previous network, conducts the necessary action, and produces output while also sending the data to the next network. Some apps just require recent data, while others may demand more information from the past. As the distance between required previous information and the point of necessity widens, the common recurrent neural networks lag in learning.

However, Long Short Term Memory (LSTM) Networks [18], a type of RNN capable of learning such instances, are available. These networks are specifically designed to avoid the recurrent networks' long-term dependency issue. LSTMs are excellent in remembering information for a lengthy period of time. Because more prior information can affect the model's accuracy, LSTMs are a natural choice. As illustrated in Fig. 1, a typical LSTM module called a repeating module contains four neural network layers interacting in a unique way. As shown in Fig. 1, the module contains three gate activation functions 1, 2, and 3 as well as two output activation functions 1 and 2. Multiplication and addition of elements are represented by the symbols and. The operation of concatenation is denoted by the symbol (•) bullet. Cell state, a line extending from Memory from Previous Block (S_{t1}) to Memory from Current Block (S_{t2}), is the most basic component of LSTMs (S_t). It permits data to travel down the wire in a straight line. The network can decide how many past data to send. It is managed by the first layer (1).

The operation that this layer performs is described in (1) Two network layers are used to compute the new information to be stored in the cell state. A sigmoid layer (2), which determines which values to update (It) (see (2)), and a Tanh layer (1, which evolves a vector of new candidate values (St), as illustrated in (3). In the state, a combination of both will be added. Finally, the status of the cells is updated using (4).

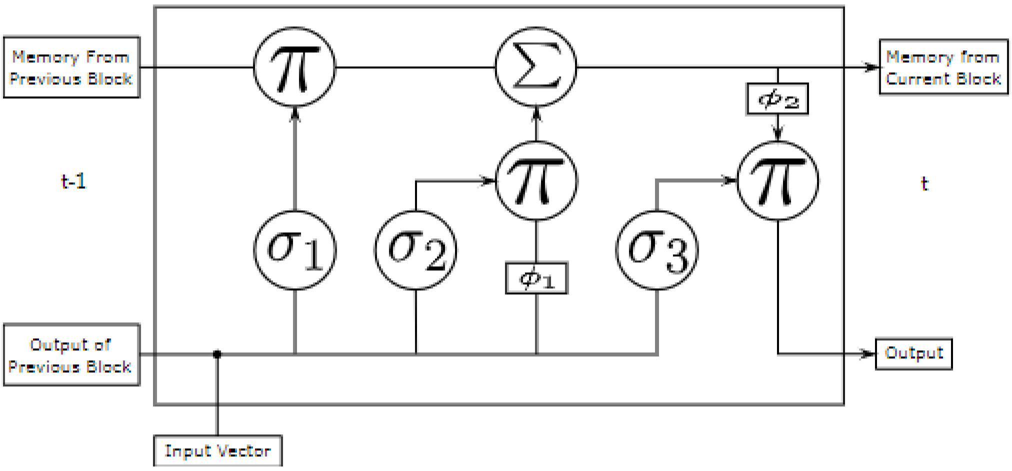

Fig: LSTM network

$$cf_1 = \sigma 1(W_{cf} \cdot [O_{t-1}, x_t] + b_{cf}) \tag{1}$$

$$I_t = \sigma 2(W_1 \cdot [O_{t-1}, x_t] + b_I) \tag{2}$$

$$S_t = \tanh\,(W_S\,[O_{t-1}, x_t] + b_S) \tag{3}$$

$$S_t = cf_1 \times S_{t1} + I_t \times S_{t-1} \tag{4}$$

RESULTS AND DISCUSSION

1. Error Statistics of IMD and NCEP data for 2020

The NCEP data is linearly interpolated from 250km to 25km resolution and data is extracted for the Indian region. Both the datasets are set up with the same dimensions, in order to proceed further with the error statistics. The histogram plots of raw and interpolated data (Fig.1) clearly distinguish raw data from the interpolated data. The mean rainfall plots over India (Fig 2) using both the datasets are also plotted

.

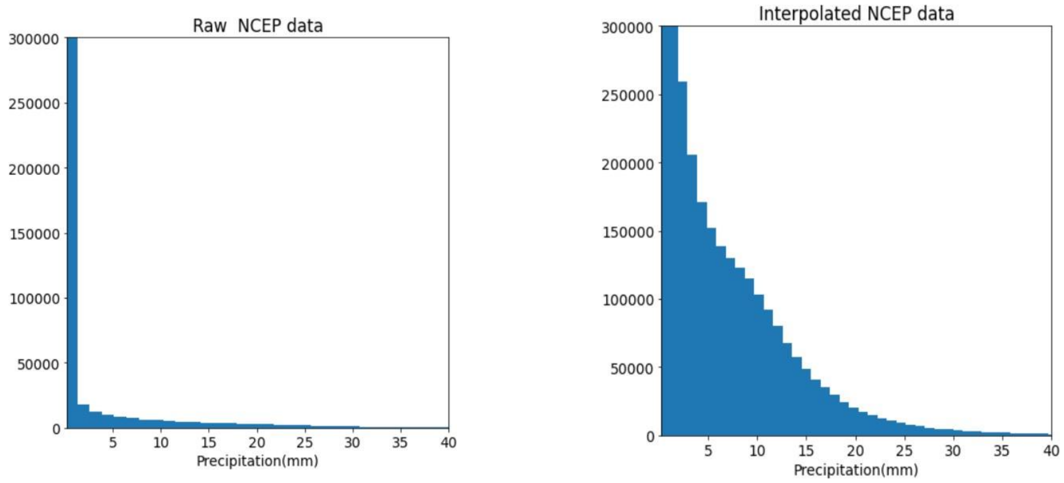

Fig 1 – Histogram plots of raw and interpolated data

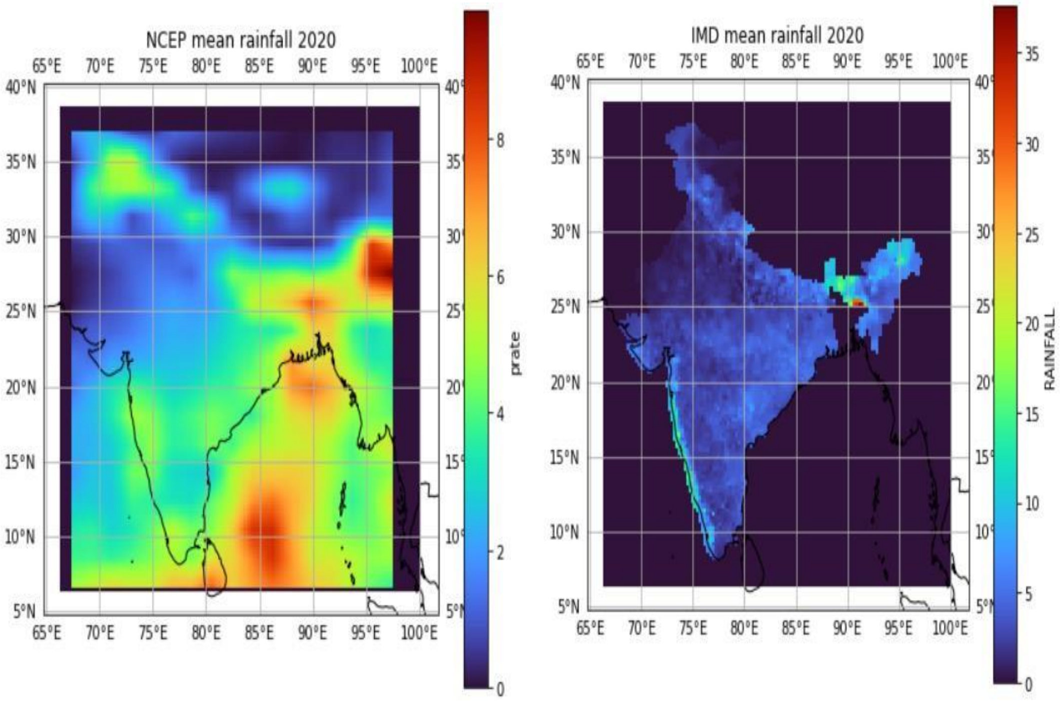

Fig 2 – Mean rainfall plots for the year 2020 using NCEP and IMD data

The Spatial and distribution plots of RMSE, MAE and Pearson's correlation are plotted to analyze the error between the two datasets. The plots reveal that there is a significant RMSE of around 5 – 10 mm and a significant MAE ofaround 2- 6 mm. There is also a 0.4 positive correlation between the two datasets.

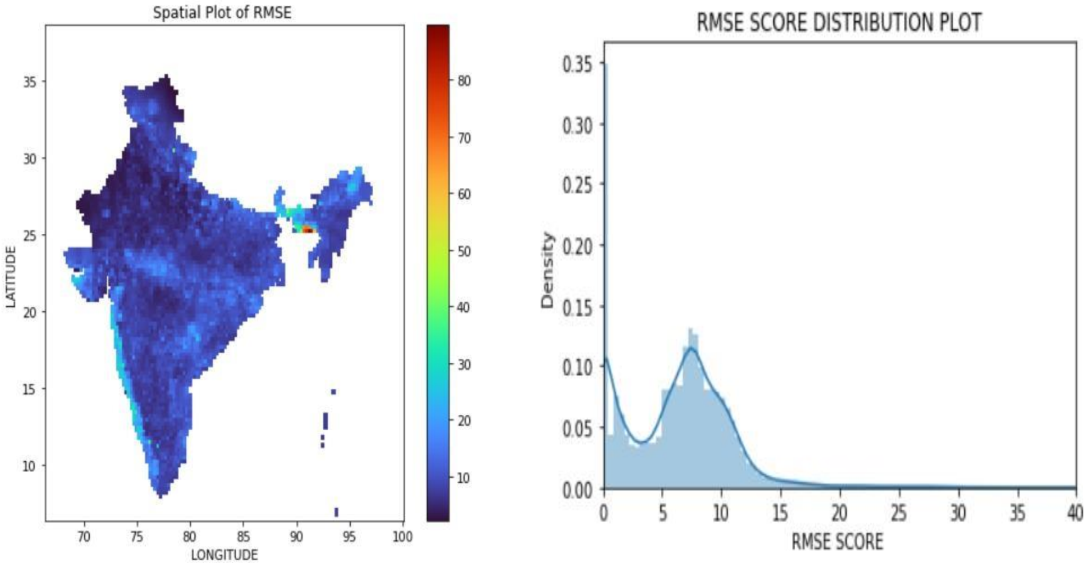

Fig 3 – RMSE spatial and distribution plots

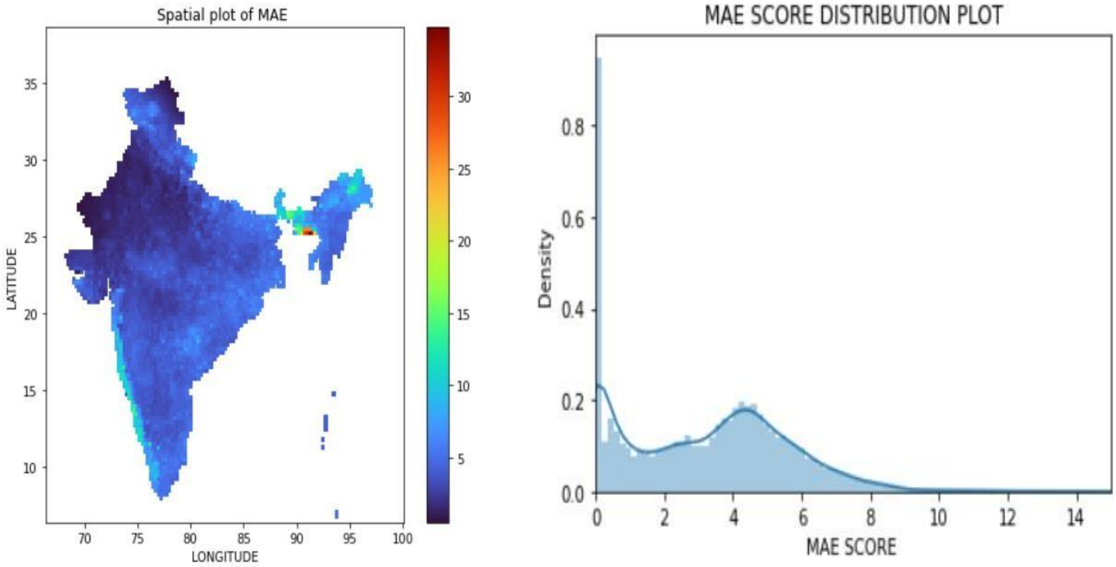

Fig 4- MAE – spatial and distribution plots

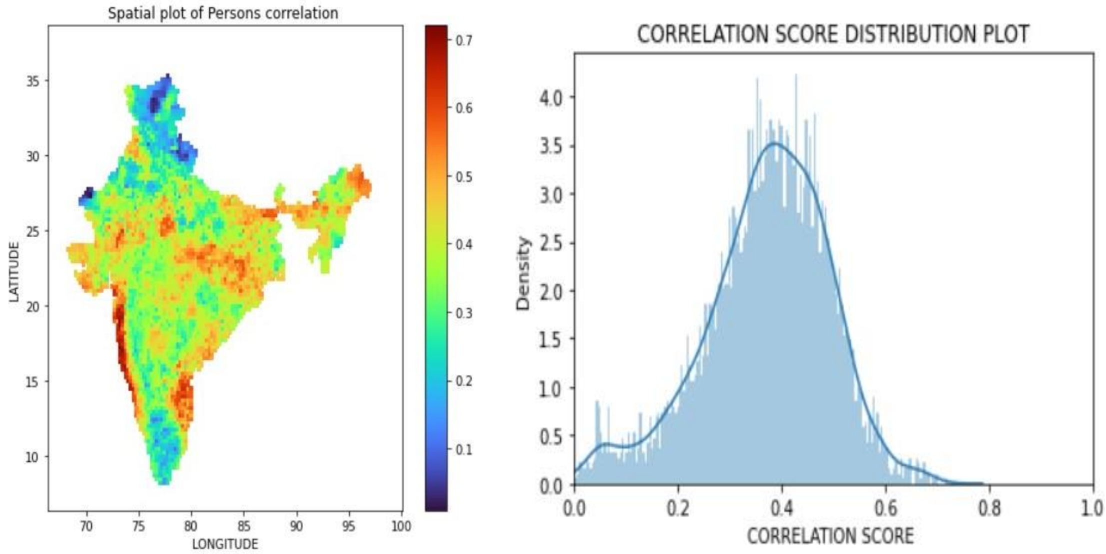

Fig 5 – Pearson's correlation spatial and distribution plots

The RMSE dispersion plots will give us an idea of the error present in theinterpolated NCEP reanalysis. These are plotted as time series plots, establishing the trend of RMSE score over each year.

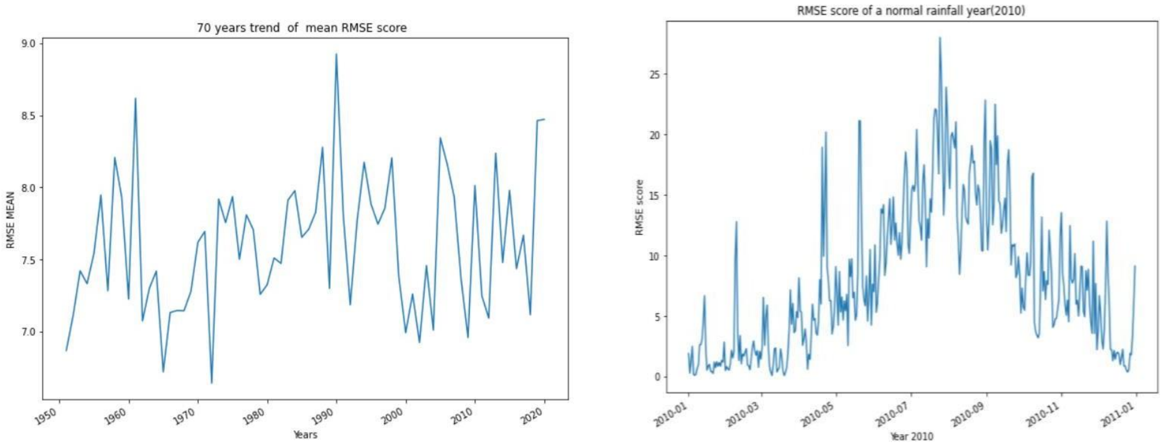

Fig 6 – 70 years trend of mean RMSE score over India and RMSE time series for a normal rainfall year (2010)

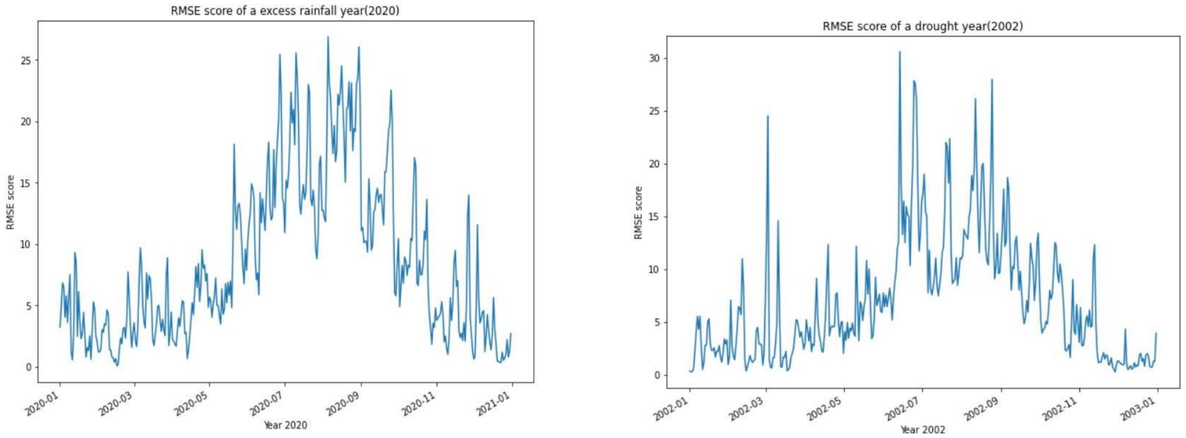

Fig 7 – RMSE time series of an excess rainfall year (2020) and RMSE time series of a drought year(2002)

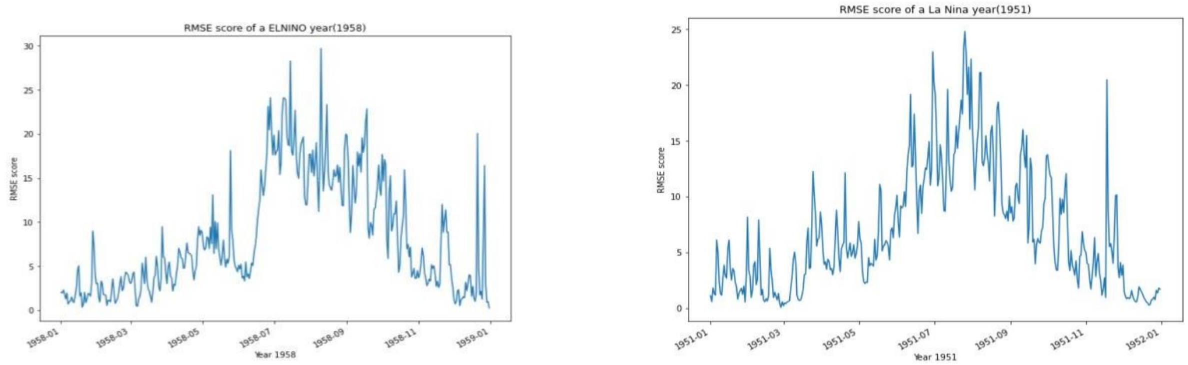

Fig 8- RMSE score of an ELNINO year (1958) and RMSE time series of a La Nina year

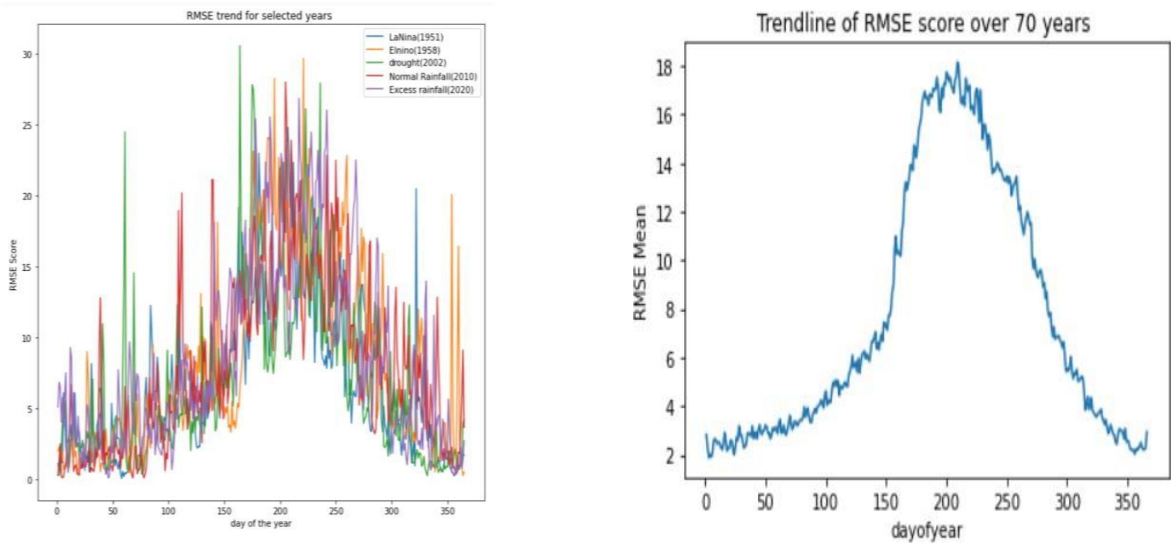

Fig 9- RMSE trend of selected years and RMSE trend line of 70 years

From the time series for each of the selected years, we see that the trendline peak is during the JJAS season. This is significant error we want to reduce byusing deep learning to prepare a model that will convert our low resolution NCEP data into a high-resolution data and hence can be used for further application.

2. Preparation of Train and Test data

We classify the 1951 to 2020 precipitation data based on two factors- the net annual rainfall and NINO3.4 index. We interpolate the NCEP reanalysis data asdone for the year 2020 above. The Indian Summer monsoon rainfall (ISMR) isextracted and resampled for year-wise, where the data is summed over time andspatially averaged. The year-wise net rainfall for JJAS (June July August September) is extracted for both the data sets. We can find the percentage deviation from mean rainfall and use a deviation of +10% for excess rainfall and -10% for drought years and the rest being normal rainfall years. The table showing the final classification is attached.

Table 1 – Classification of years from 1951 to 2020 based on ISMR deviation

IMD - Drought years	1951,1965,1966,1968,1972,1979,1982, ,1987,2002,2004,2009,2014,2015
IMD - Excess rainfall years	1959,1961,1970,1975,1983,1988,1994, 2019,2020
NCEP interpolated – Droughtyears	1951, 1963,1964,1965,1966,1972, 1974,1986,1987,1992,1996,2000,2002, 2014
NCEP interpolated – Excess rainfall years	1955,1956,1958,1959,1961,1970,1971, 1977,1978,1980,1983,2006,2011,2013, 2019,2020

The NINO3.4 SST index is used to classify years as, El Niño and La Niña. The NINO3.4 anomaly data is obtained from PSL using the HadlSST1 dataset. Sincethe ISMR rainfall is considered, and a six-month lag exists between ISMR and NINO3.4, December January February (DJF) mean is calculated for each year
and deviations greater than 0.5 ˚ C is considered as a El Niño year and lesser than -0.5 ˚ C is considered as La Niña year. The classification table is attached.

Table 2 – classification based on NINO3.4 Index

El Niño	1958,1959,1964,1966,1969,1970,1973,1977, 1978,1980,1983,1987,1988,1992,1995,1998, 2003,2005,2007,2010,2015,2016,2019,2020
La Nina	1951,1955,1956,1965,1968,1971,1972, 1974,1975,1976,1984,1985,1986,1989, 1996,1999,2000,2001,2006,2008,2009,2011, 2012,2018

2. Deep learning – LSTM model

The flowchart of the entire process of statistical downscaling of the NCEP reanalysis data is discussed below. The LSTM model was run for the JJAS season. After various combinations of hyper-parameters, the final set of Hyper parameters that produced optimal results are as shown in Table 3.

Table 3: LSTM model summary

HYPERPARAMETERS	**OPTIMUM VALUE**
Nodes	50
Dropout	0,20
Optimizer	Adam(Learning rate=0.01)
Activation Function	Tanh
Epoch	50
Batch Size	64
Predictor variables	IMD observations, NCEP reanalysis data
Predictand	High-Resolution Precipitation Reanalysis data

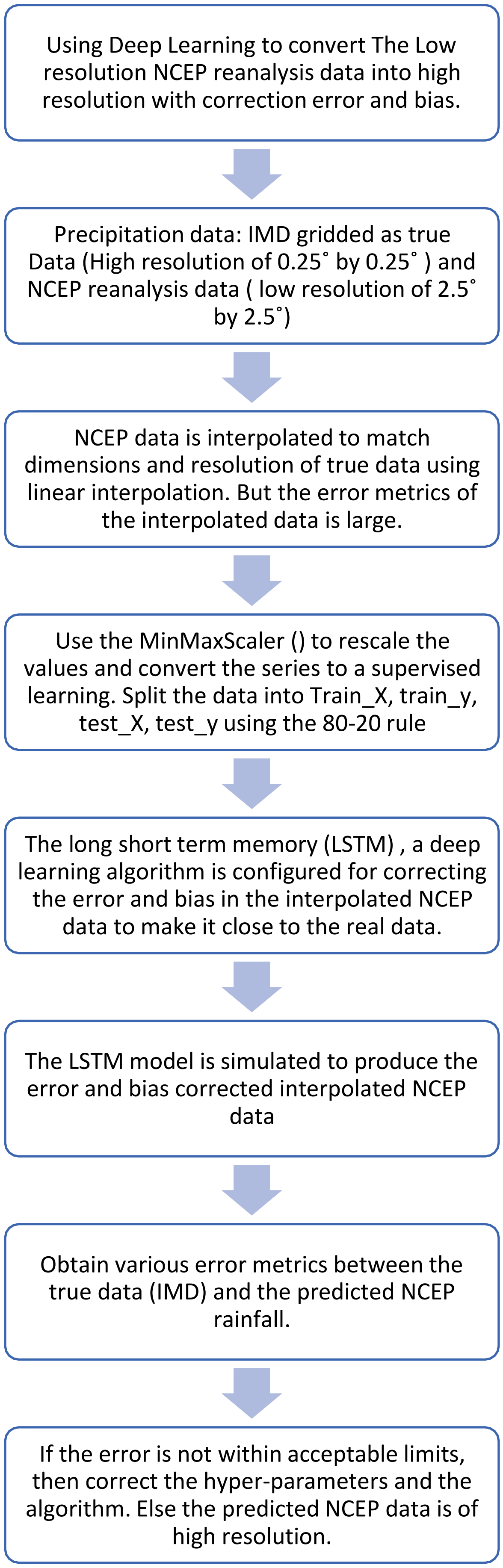
Using Deep Learning to convert The Low resolution NCEP reanalysis data into high resolution with correction error and bias.
Precipitation data: IMD gridded as true Data (High resolution of 0.25˚ by 0.25˚) and NCEP reanalysis data (low resolution of 2.5˚ by 2.5˚)
NCEP data is interpolated to match dimensions and resolution of true data using linear interpolation. But the error metrics of the interpolated data is large.
Use the MinMaxScaler () to rescale the values and convert the series to a supervised learning. Split the data into Train_X, train_y, test_X, test_y using the 80-20 rule
The long short term memory (LSTM) , a deep learning algorithm is configured for correcting the error and bias in the interpolated NCEP data to make it close to the real data.
The LSTM model is simulated to produce the error and bias corrected interpolated NCEP data
Obtain various error metrics between the true data (IMD) and the predicted NCEP rainfall.
If the error is not within acceptable limits, then correct the hyper-parameters and the algorithm. Else the predicted NCEP data is of high resolution.

Fig 10- LSTM flowchart

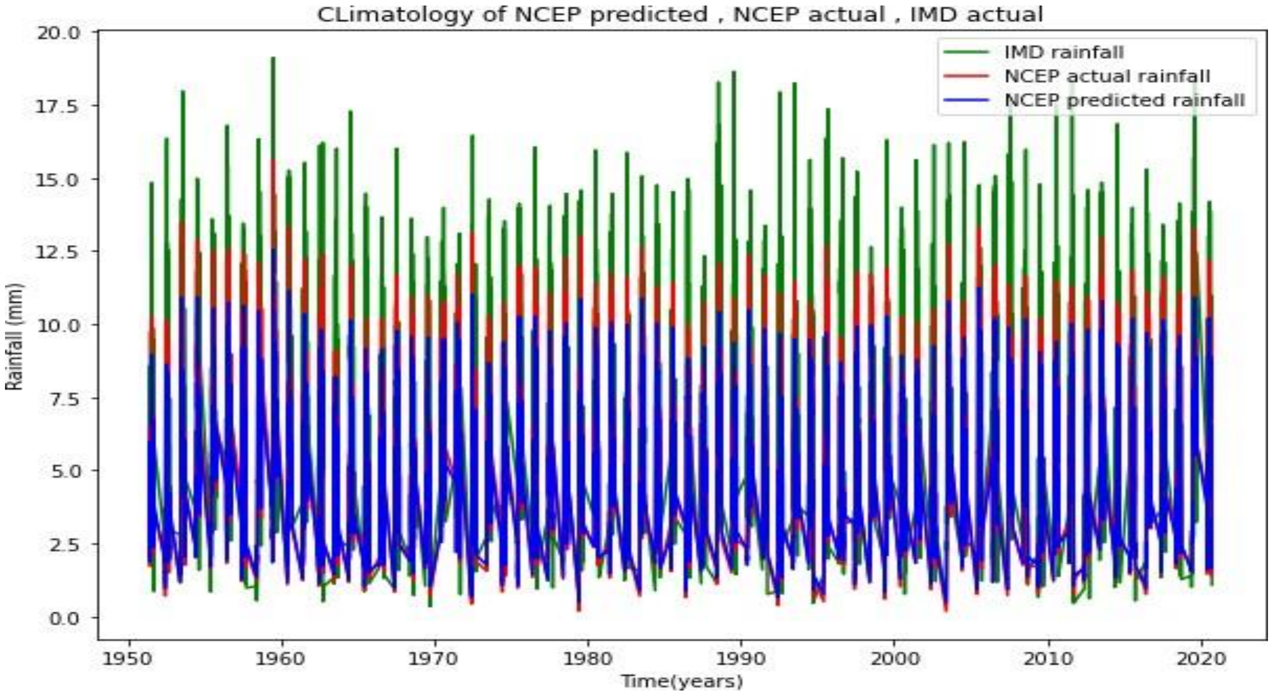

Fig 11-Climatology plots of predicted, actual and IMD datasets.

The climatology plots of the downscaled NCEP rainfall data, the low resolutionNCEP data and the IMD data are plotted. The LSTM model has statistically downscaled the reanalysis rainfall, which can be deduced from the above plot.

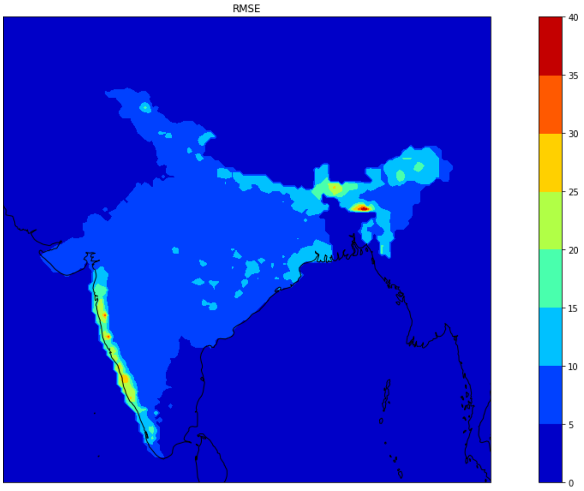

Fig 12 -Mean RMSE Spatial plot (mm)

Upon plotting the RMSE spatial mean plot for all the timesteps, though the overall error has reduced significantly, there is still bias error that is not accounted for, which can be improved by using further advanced deep learningmodels.

CONCLUSION

The assessment of precipitation in a changing climate is critical because climate change has a significant impact on the geographical and temporal variance of precipitation. On low-resolution reanalysis data, statistical downscaling is accomplished using a deep learning technique in this study. Statistical downscaling is a technique for determining data on a smaller grid-scale using data from a bigger grid and time period. Downscaling models built using machine learning methodologies beat downscaling models created using traditional statistical regression techniques, according to studies. The precipitation data from the National Centers for Environmental Prediction (NCEP) reanalysis is statistically downscaled for the period 1951 to 2020 using the Long Short Term Memory Deep Learning Model, with high-resolution IMD observation data and coarser-resolution NCEP reanalysis as predictor variables and higher resolution reanalysis data as predictand variables.The error statistics were analyzed and the data was prepared for training the model. The LSTM model was trained with 50 nodes, Adam optimizer with 0.01 learning rate, and Tanh activation function. The error and climatology plots reveal the downscaling but with bias error. The model can be further developed to eliminate the bias error.

REFERENCES

1. *Guhathakurta, Pulak & Rajeevan, M. & Thapliyal, V.. (1999). LongRange Forecasting Indian Summer Monsoon Rainfallby a Hybrid Principal Component Neural Network Model. Meteorology and Atmospheric Physics. 71. 255-266. 10007/s007030050059.*

2. *Singh, Pritpal & Borah, Bhogeswar. (2013). Indian summer monsoon rainfall prediction using artificial neural network. Stochastic Environmental Research and Risk Assessment. 27. 10.1007/s00477-013-0695-0.*

3. *Sahai, A., Soman, M. & Satyan, V. All India summer monsoon rainfall prediction using an artificial neural network. Climate Dynamics 16, 291– 302 (2000). https://doi.org/10.1007/s003820050328*

4. *Tahir, T., A. M. Hashim, and K. W. Yusof. "Statistical downscaling ofrainfall under transitional climate in Limbang River Basin by using SDSM." IOP conference series: earth and environmental science. Vol 140*

5. *Osman, Yassin Z., and Mawada E. Abdellatif. "Improving Accuracy of Downscaling Rainfall by Combining Predictions of Different Statistical Downscale Models." Water Science, vol. 30, no. 2, Informa UK Limited, 1 Oct. 2016, pp. 61–75. Crossref, doi: 10.1016/j.wsj.2016.10.002.*

6. *Vu, M.T., Aribarg, T., Supratid, S. et al. Statistical downscaling rainfall using artificial neural network: significantly wetter Bangkok?. Theor Appl Climatol 126, 453–467 (2016). https://doi.org/10.1007/s00704-015-1580-1*

7. *Ahmed, K., Shahid, S., Haroon, S.B. et al. Multilayer perceptron neural network for downscaling rainfall in arid region: A case study of Baluchistan, Pakistan. J Earth Syst Sci 124, 1325–1341 (2015). https://doi.org/10.1007/s12040-015-0602-9*

8. *Tran Anh, Duong, et al. "Downscaling Rainfall Using Deep Learning Long Short-term Memory and Feedforward Neural Network." International Journal of Climatology, vol. 39, no. 10, Wiley, Apr. 2019, pp. 4170–4188. Crossref, doi:10.1002/joc.6066.*