https://doi.org/10.48047/AFJBS.6.Si3.2024.3054-3065

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper Open Access

Temporal-Spatial Clustering Analysis for Precipitation Prediction: Entropy-TCN-GRU Technique in Environmental Time Series Analysis

E.Valli[1], S.Nirmala Sugirtha Rajini[2]

[1]Research Scholar, Department of Computer Applications, Dr.MGR. Educational and Research Institute, Chennai, India
Email: [1]vallielumalai2795@gmail.com
[2]Professor, Department of Computer Applications, Dr.MGR. Educational and Research Institute, Chennai, India
Email:[2]nirmalasugirtharajini.mca@drmgrdu.ac.in

ArticleInfo

Volume6, Issue Si3,July2024

Received: 24May2024

Accepted:21June 2024

Published:10July2024

doi:

10.48047/AFJBS.6.Si3.2024.3054-3065

ABSTRACT:

Accurate precipitation prediction is crucial due to its potential to trigger various disasters, profoundly impacting communities worldwide. Reliable forecasts enable proactive measures to mitigate risks, such as floods and droughts, which regularly afflict populations globally. This precision is particularly significant for countries like India, where agriculture plays a pivotal role in the economy. Effective prediction relies on addressing challenges such as seasonal variability and transient patterns. This study employs a concise spatial analysis method to tackle these variations, integrating advanced deep learning techniques like the Multivariate Transient Convolutional Network (TCN) and Gated Recurrent Unit (GRU). Referred to as the e-TCN-GRU model, it achieved impressive metrics with a coefficient of determination (R2) and explained variance score (EVar) reaching 98.42% and 98.49%, respectively. Additionally, mean absolute error (MAE) and root mean square error (RMSE) were significantly lower compared to alternative models. These findings underscore the model's robustness and reliability in enhancing precipitation forecasting, crucial for environmental sustainability and disaster preparedness efforts globally.

Keywords: Precipitation Prediction, Accuracy, Tempo-Spatial, Entropy-TCN-GRU, mix-max scaler and Pytorch.

1. Introduction

Rainfall continues to be one of the weather factors that has the biggest impact on many aspects of our daily lives. The effects of the rainy season on the socioeconomic system are significant, and they range from disruptions in transportation to infrastructure damage in the event of a flood. As a result, numerous studies have looked into and recommended temperature and precipitation forecasting methods as a mitigation strategy in advance of any disaster or eventuality. However, these strategies must be effective and timed to improve human mobility, agriculture, and industrial development. Building a prediction model based on the historical data of bike-sharing demand caneffectively explain the time series characteristics of this phenomenon, but the influenceof other elements in the bike-sharing system is not considered; thus, there is a certain one-sidedness, and a limit to the ability to explain and predict the fluctuation mechanismof bicycle travel demand [1]. Pointed out that climatic conditions such as temperature,wind, and precipitation are the main factors affecting the demand for bike-sharing, and Faghih et al.[2]Established an LSTM linear regression model considering the distancevariable of users' rides, and the results of the study show that the prediction accuracy was improved compared with the existing time series prediction models. Li et al.[3]

From the perspective of forecasting model development, statistical methods such as theAutoregressive Integrated Moving Average model (ARIMA) were first applied to solve the bike-sharing cycling demand forecasting problem.Statistical inferential forecastingmethods based on statistics include traditional models such as ARIMA models, regression analysis and Markov chains[4] .Data driven is an expression for an activity influenced by data and did not impacted by instinct or individual observation. The most common example is clustering. Clustering allows one to search for the group with similar characteristics in a partition considering them belongs to unsupervised learning style[5]. In DT, entropy and information gain play the most important role in splitting the attributes. Entropy is computing the vulnerability within the set of training because of the possibility of more than one possible splitting solution[6].GRU solves the problem with its two gates named update and reset to manage the succession of data in the time of learning them for predicting through the information organized in a long chain[7].

In order to forecast hourly rainfall using time-series data from five significant UK cities, three models based on the LSTM and Stacked-LSTM Networks were modified from the reviewed literature. Utilizing time-series data from five significant UK cities, a model based on bidirectional LSTM networks was suggested for the task of hourly rainfall forecasting. In order to forecast the amount of rainfall per hour using time-series data from five significant UK cities, models based on LSTMNetworks, Stacked-LSTM Networks, Bidirectional-LSTM, Radial Basis Function Networks (RBFNs) Networks, XGBoost, and the final model from AutoML were tested. However, forecasting precipitation is a much trickier task because regional rainfalls frequently vary in space and time. A. Barrera-Animas. Y. Oyedele,et.al.,(2022)[1].

2. Literature Survey

One of the most significant meteorological variables continues to be rainfall. In a number of areas of our daily lives.The performance of models based on LSTM, Stacked-LSTM, and Bidirectional-LSTM Networks is specifically compared in this study to that of an XGBoost decision tree model and a proposed model that will be created using an Automated Machine

Learning (AutoML) tool Barrera-Animas, et.al.,2022 [1] .Their ARIMA model was successfully able to predict the COVID-19 trends in countries such as Iran, Italy, Spain and France. Ardabili et al.have used the machine learning techniques for predicting the COVID-19 outbreak, they found that the multi-layer perceptron model and adaptive network-based fuzzy interface system are found to give promising results[8]. Deep learning models are superior over the statistical machine learning models for forecasting the non-linear applica- tions such as prediction of weather, stock prices, electro car- diogram (ECG) recordings and crude oil prices etc[9][10]

However, there is no clear trends for the COVID-19 cases throughout main land China. In a study, Is- mail et al[11] conducted a comparative study based on ARIMA, LSTM, Non-linear Autoregressive Neural Networks (NARNN) mod- elstoforecast the COVID-19 cases in Denmark, Belgium, Germany, France, United Kingdom, Finland, Switzerland, and Turkey. They found that the LSTM offers lowest Root Mean Square Error (RMSE) compared to other models. Shawni Dutta et al. [12] have used LSTM, Grated Recurrent Unit (GRU), Recurrent Neural Networks (RNN) to predict the confirmed, released, negative, death cases of COVID-19 pandemic. Their study has recommended that individual machine learning models are needed for each country due to the presence of fundamental differences between the countries.K-Nearest Neighbor(KNN), Support vector machine(SVM), and Neural Network(NN).These have been applied on the rainfall data of North Carolina from 2007 – 2017 and also the performance is calculated by applying different metrics F-score, precision, accuracy, recall. Finally, eight hybrid models have been proposed and Gradient boosting-Ada boost has been the superior which exhibited good results. Kar, Kaveri, Neelima Thakur, and Prerika Sanghvi [6] has used the fuzzy logic approach for the prediction of rainfall on the data of temperature in a geographic location.

The fuzzy model has been applied Due to other climatic factors the prediction is not accurate so they have considered other influencing factors like humidity also analyzed the advantages of fuzzy system over other techniques[20]. Radial Basis Function Neural Network (RBFNN) and generalized regression (GRNN) on the rainfall data of India mainly Nanded district, Maharashtra was considered and the data is normalized between 0 to 1 and the algorithms are applied and the performance of those was calculated and compared. BPNN and RBFNN has given good results compared to GRNN. Chen, Binghong, et al. [21] focuses on the non-linear machine learning approaches like gradient boosting decision tree model and deep neural networks for a short term prediction of rainfall and these algorithms were built on Alibaba cloud and data was collected from different sites and effectiveness is calculated by using classification metrics AUC, F1 score, precision and accuracy and by Regression metric RMSE, correlation. It has been observed that DNN showed better result than ECData.

The largest SI value is expected to change in the future, according to this paper. Regarding anticipated water savings from rainwater tanks, more details are provided in the paper. Climate change simulation uses ETank, a daily water balance model that was previously created. For the daily rainfall forecast and simulations of potential water savings from rainwater tanks.

In order to recreate the long-term seasonal rainfall patterns in Western Australia, this paper compares the effectiveness of linear and non-linear modeling techniques. Commonly used multiple linear regression (MLR) and artificial neural network (ANN) modeling approaches were used to create the linear and non-linear models. El Nio Southern Oscillation (ENSO) and Indian Ocean Dipole (IOD) lagged (past) values were thought to be the most promising predictorsof seasonal rainfall. The Lavenberg-Marquardt algorithm was used in conjunction with the Multilayer Perceptron training rule to build the non-linear ANN models. Iqbal &Rasel,et.al.,(2020).

This review paper presents an introduction tothe implementation of machine learning on geographical datasets.The main area covered in this paper are meteorological data, Ensemble models, regression, classification including Neural Networks, Support vector machine (SVM), Decision trees, Naïve Bayes, J48, CART, and ID3. However, the main thrust area has been the implementation of various neural network models which includes BPNN, FFNN, GWLM-NARX, RNN, and TDNN in geographical data sciences.Sheikh Amir Fayaz .,et.al.,[6]

3. Proposed System

The objective of this investigation is to investigate rainfall volume data using time-series data from five major UK cities. The Epsilon-TCN-GRU created with the PyTorch Tool and TCN-GRUwere found to outperform all other models tested. This indicates that models based on Epsilon TCN-GRN with fewer hidden layers perform better for this approach and control the temporal-spatio problem datasets in difficult features like topographic and stations .GRU and TCN, with the principle of the least-squared error sum. In so doing we aimed to reduce the possibility of overfitting and to take advantage of the fitting of deep learning models on nonlinear and non-stationary data, in order to improve the prediction ability of the models.In particular, this study aims to manage the temporal-spatio problem analyst from difficult features from topographic area to improve the prediction accuracy robust.

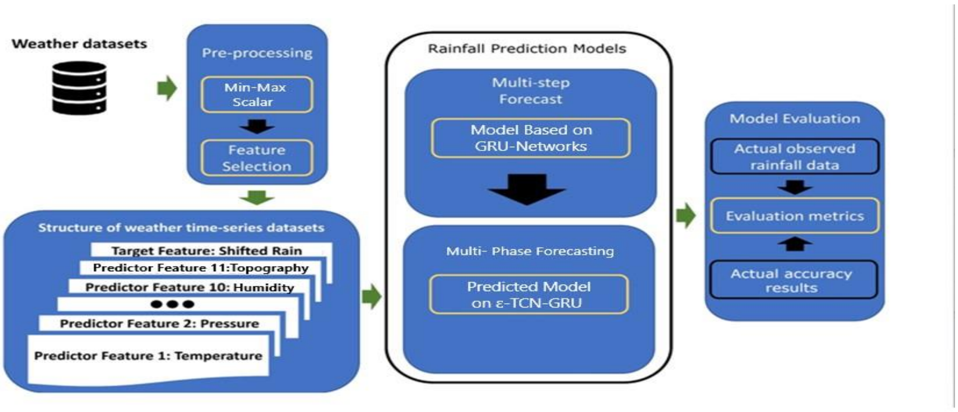


Fig1: Model Evaluation Methodology

Temporal Convolutional Network:
TCN is a clever design in view of a Convolutional Brain Organization (CNN). Dissimilar to general CNNs, TCNs use designs, for example, extended causal convolution and remaining blocks [38-40]. This enables them to remove includes and accomplish forecast from enormous example time series, and TCNs can actually address the presentation corruption of profound organizations during network preparing. TCN comprises of enlarged, causal 1D completely convolutional layers with similar information and result lengths. The convolution in the TCN model is causal convolution, wherein the layers are causally connected with one another, in this way guaranteeing that no verifiable data or future information will be remembered fondly. Also, TCN can plan arrangements of inconsistent length to yield

successions of similar length, utilizing leftover modules and widening convolution to more readily control the memory length of the model and work on the prescient power.

Gated Recurrent Unit (GRU)

LSTM and GRU show solid expected appropriateness in the information expectation issue concentrated on in this paper, with GRU performing somewhat better. Contrasted and the TCN strategy, GRU requires less preparation boundaries, is more straightforward to combine and can decrease the gamble of model overfitting on account of restricted time series information. GRU improves the three entryway elements of TCN, turning the arrangement of neglecting doors and info entryways into a solitary update door, and blending the neuron states with the secret states. This can actually ease the issue of "slope vanishing" in RNN organizations and diminish the quantity of boundaries of TCN network units, shortening the preparation season of the model. The fundamental design is displayed in, and the numerical depiction is shown ,

$rt = \sigma(Wr \cdot [ht-1, xt])$ (1)

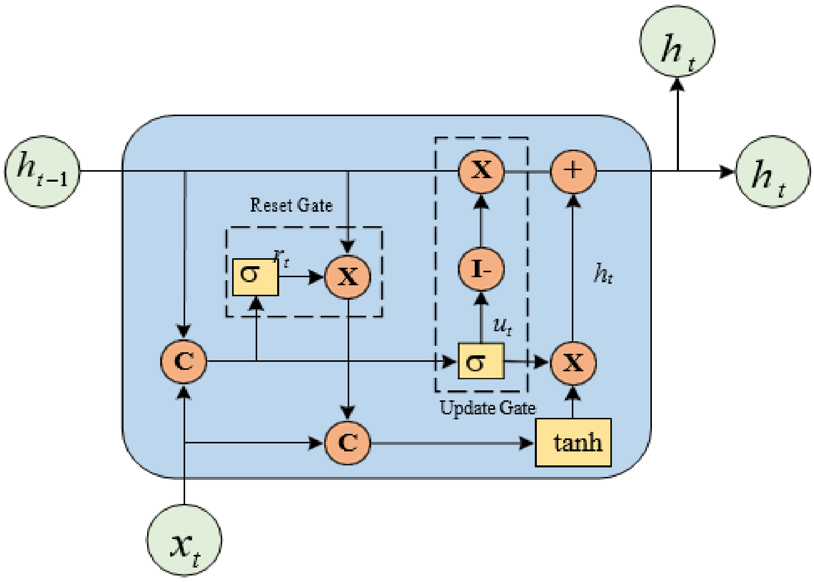
$ut = \sigma(Wu \cdot [ht-1, xt])$ (2)

$ht = \tanh(Wh \cdot [rt * ht-1, xt])$ (3)

$ht = (1 - ut) * ht-1 + ut*ht$ (4)

$yt = \sigma(Wo \cdot ht)$ (5)

LSTM and GRU show solid possible materiality in the information expectation issue concentrated on in this paper, with GRU performing somewhat better. Contrasted and the TCN technique, GRU requires fewer preparation boundaries, is more straightforward to merge and can diminish the gamble of model overfitting on account of restricted time series information. GRU upgrades the three entryway elements of TCN, turning the arrangement of neglecting doors and information doors into a solitary update door, and blending the neuron states with the secret states. This can really lighten the issue of "slope vanishing" in RNN organizations and decrease the quantity of boundaries of TCN network units, shortening the preparation season of the model. The fundamental design is displayed in, and the numerical portrayal is shown,



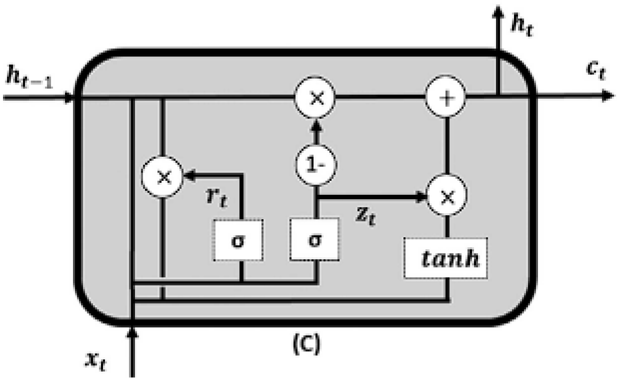


Fig 2: Architecture of GRU Transformation to E-TCN-GRU

Entropy Computing:

Entropy is processing the weakness inside the arrangement of preparing as a result of the chance of more than one potential parting arrangement [22]. The ideal arrangement is having the most reduced entropy that will put the likelihood, p either in 0 or 1 as displayed in Fig 3.

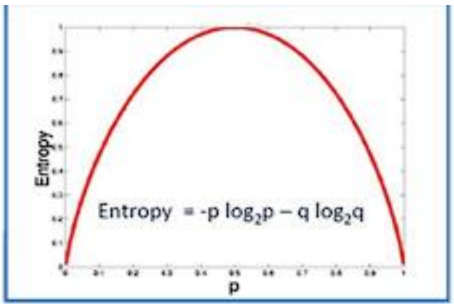


Fig 3: Graph of Entropy ft probability

$$\text{Entropy} = 0.5\log_2 0.5 - 0.5\log_2 0.5 = 1 \qquad (6)$$

Entropy is connected with data gain as an exact dt possesses lower entropy and the most elevated data gain. data Gain as displayed in (6) is the distinctions in the entropy because of the segment [22]

$$\text{Gain}_{\text{spit}} = \text{Entropy(p)} - \left(\sum_{i=1}^{k} \frac{n_i}{n}\,\text{Entropy(i)}\right) \quad (7)$$

4. Data Overview and Preprocessing

Data Overview

This dataset contains meteorological and geographical data collected from various weather stations across a specific region. The data is collected over several years and includes features relevant to rainfall prediction**about 8 features and 39K rows.** The month of the observation, recorded as a numeric value from seasonal month January to December ..

- Data consists of 8 attributes (individual months, annual, and combinations of 3 consecutive months).
- The data is available only from 1940s to 2022 for some of the stations
- The attributes are the amount of rainfall measured in mm

The identifier for the weather station where the observation was made to be relative humidity measured at the weather station, recorded as a percentage. The amount of rainfall recorded during the observation period, measured in millimetres. Information about the local topography surrounding the weather station, including elevation, terrain type, and geographical features.

Table 1: Data Fields and Descriptions

Field Name	Description	Example
timestamp	Timestamp for grouping data together	4 January 1980, 12:00
Year	Gathering Year of sesaonal datasets	32
Max(t 1&t2)	Actual temperature (°C)	3.0
Min(t 1&t2)	Subjective perception of temperature (°C)	2.0
hum	Humidity percentage (%)	93.0
wind_speed	Wind speed value (km/h)	6.0
Weather_stations	station observed main city poles	25
season	Spring: 0; Summer: 1; Autumn: 2; Winter: 3	3
hour	24 h per day	12
day_of_month	Natural days per month	1
Topography types	elevation, terrain and geographical feature	3
month	January: 1, . . . , December: 12	6

Data Preprocessing:

Before building predictive models, the dataset undergoes preprocessing steps to ensure data quality and compatibility for analysis. Identify and handle missing values in the dataset, either through imputation or removal of incomplete records. Scale numerical features such as temperature, humidity, and rainfall amount using techniques like Min-Max scaling to normalize the data within a specified range (e.g., between 0 and 1). Encode categorical features like station, year, and month into numerical representations using techniques such as one-hot encoding or label encoding. Create additional features or transform existing ones to capture relevant information, such as aggregating temperature data over different time intervals or extracting seasonal patterns from the date features. Split the dataset into training, validation, and test sets to evaluate model performance effectively. Normalize topographic features to ensure consistency and comparability across different stationsand regions, taking into account elevation differences and terrain variations. The specific preprocessing steps may vary depending on the characteristics of the dataset and the requirements of the predictive modelling task.

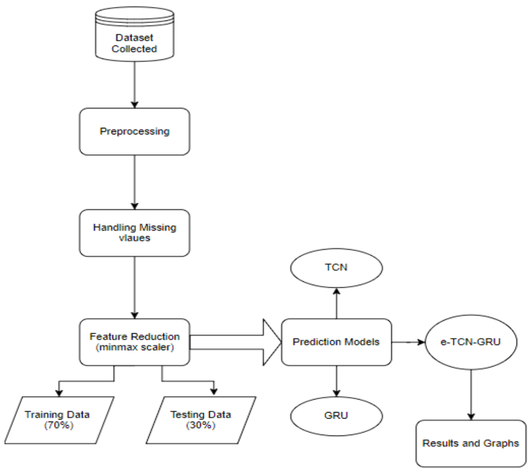


Fig 4: Preprocessing Data Collection

It's fundamental to painstakingly break down the information and pick fitting preprocessing strategies to set up the dataset for precise precipitation expectation models. This includes a direct change of the first information that maps the information values to the [0, 1] span. The change capability is displayed in Condition.

$$x^* = \frac{x - \min}{\max - \min} \quad (8)$$

where max is the maximum value of the data, and min is the minimum value.

E-TCN-GRU Technique Algorithm:
Rainfall prediction Input: Rainfall data set
Output: Accuracy/error of the prediction
Step1: Import the rainfall data set csv file.
Step2: Fill the missing values with mean value of the data.
Step3: Scaling the features- scaling the data to a fixed scale.
Step4: Feature Reduction- minmax scaler is used to minimize the data.
Step5: The data is divided into training set (70%) and testing set (30%).
Step6: Temporal convolutional Network, Gated Recurrent Unit and E-TCN-GRU is applied and the Mean Absolute Error, r2 score is calculated.
Step7: The scatter plots are plotted between predicted and testing data for the applied models and the errors are compared and best model among them is selected.
Step8: Display the results.

Model Evaluation Methods
Model Evaluation Methods
To validate and compare the accuracy as well as the robustness of the models, we used R2, EVar, MAE, MedAE, and RMSE as evaluation metrics, respectively.

(1) Coefficient of determination (R2)
The coefficient of determination characterizes the extent to which the regression model explains the variation in the dependent variable, or the goodness of fit of the model to the observations.

$$R^2 = 1 - \frac{\sum_{1}^{N} (y_1 - y_2)^2}{\sum_{i=1}^{N} ((y_i - \bar{y}_i)^2} \quad (9)$$

Here, yi is the actual value of the ith data point; yˆi is the corresponding predicted value; and y¯i is the mean value of the time series. In general, the value of the coefficient of determination R2 ranges from 0 to 1, where an R2 equal to 0 means that the model cannot predict the target variable at all, and an R2 equal to 1 means that the model can make aperfect prediction. R2 can also have negative values, in which case the model's prediction ability is not as good as calculating the mean of the target variable directly.

(2) Explainable Variance Score (EVar)
The explainable variance score measures the degree to which the dispersion of errors between all predicted and actual values is similar to the dispersion of the true values themselves.

$$\mathrm{EVar} = 1 - \frac{\mathrm{Var}(y-\hat{y})}{\mathrm{Var}(y)} \qquad (10)$$

A larger value of EVar indicates the better prediction ability of the model, and the best possible value is 1.

(3) Mean Absolute Error (MAE)

The mean absolute error is the expectation of the absolute value of the error between the predicted and actual values at each moment in time.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \qquad (11)$$

(4) Median Absolute Error (MedAE)

$$\mathrm{MedAE}(y,\hat{y}) = \mathrm{median}(|y_1 - \hat{y}_1|, \ldots, |y_n - \hat{y}_n|) \qquad (12)$$

The median absolute error is the median of the absolute error of the predicted and actual values for all data points. The metric is robust to outliers.

(5) Root Mean Square Error (RMSE)

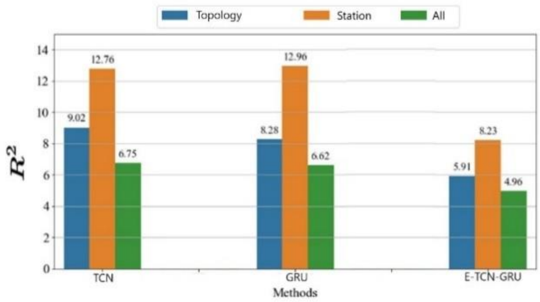
The mean square error calculates the mean of the square of the error between the predicted and true values. The root mean square error, on the other hand, is the open square of the mean square error, which is consistent with the target variable in terms of magnitude.

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \qquad (13)$$

5. Result and Discussion

Table 2:Error Loss of Prediction model Environmental Station results

Prediction Model	MAE	R2score	Evar	MedAE	RMSE
Temporal ConvolutionaNetwork	12.76	12.76	19.44	12.76	19.44
Gated Recurrent Unit	10.26	12.96	20.28	12.96	20.28
E-TCN-GRU	8.10	8.23	15.43	8.23	15.43



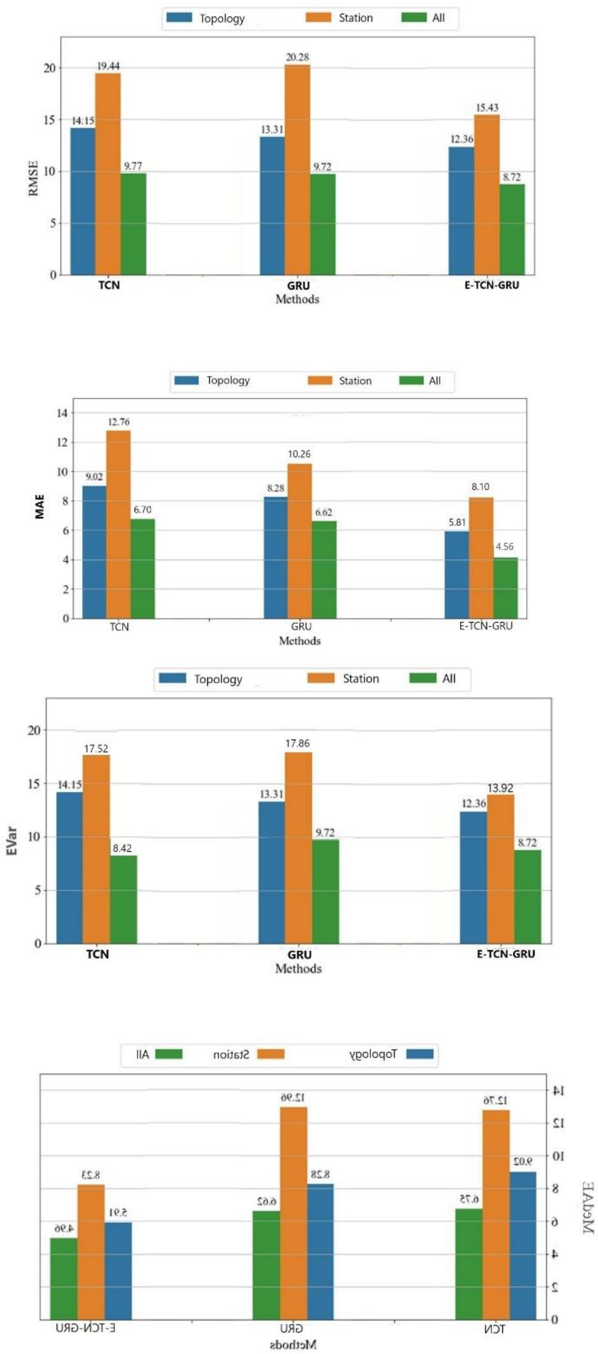


Fig 5: Error Loss Comparison Charts Among Topology and Stations of Methods

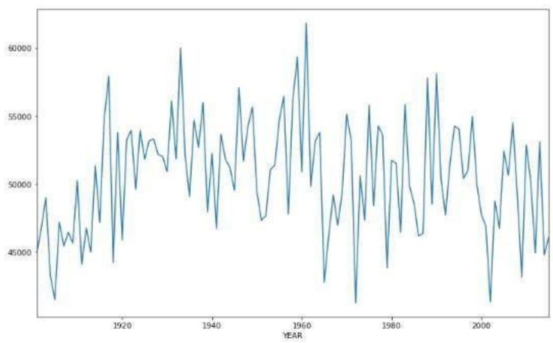


Figure 6: line graph for distribution of rainfall from the year 1940-2022.

6. Conclusion

This study aims to evaluate the effectiveness of three models in addressing the temporal-spatial challenges of rainfall prediction: Temporal Convolutional Networks (TCN), Gated Recurrent Units (GRU), and the innovative Epsilon TCN-GRU (E-TCN-GRU) method. TCN excels at capturing long-range dependencies, while GRU is adept at handling sequential data through its gating mechanism. The E-TCN-GRU model introduces an epsilon mechanism to integrate the strengths of both TCN and GRU architectures, offering a novel approach to enhancing the prediction of complex temporal and spatial patterns in rainfall data. Through comparative analysis, E-TCN-GRU demonstrates superior performance over TCN and GRU models alone, making it a promising choice for improving accuracy in rainfall predictions. This research addresses critical environmental challenges by enhancing our ability to forecast precipitation accurately, crucial for mitigating risks associated with floods and droughts and supporting sustainable environmental management practices.

7. References:

1. Barrera-Animas, A. Y., Oyedele, L. O., Bilal, M., Akinosho, T. D., Delgado, J. M. D., &Akanbi, L. A. (2022). “Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting”. Machine Learning with Applications, 7(August 2021), 100204. https://doi.org/10.1016/j.mlwa.2021.100204
2. Imteaz, M. A., & Hossain, I. (2022). “Climate Change Impacts on ‘Seasonality Index’ and its Potential Implications on Rainwater Savings”. Water Resources Management, 0123456789. https://doi.org/10.1007/s11269-022-03320-z
3. Hossain, Iqbal & Rasel, H. M. &Imteaz, Monzur&Mekanik, Fatemeh. (2020). “Long-term seasonal rainfall forecasting using linear and non-linear modeling approaches: a case study for Western Australia”. Meteorology and Atmospheric Physics. 132. 10.1007/s00703-019-00679-4.
4. Zou, Mingzhe& Fang, Duo & Harrison, Gareth &Djokic, Sasa. (2019). “Weather-Based Day-Ahead and Week-Ahead Load Forecasting using Deep Recurrent Neural Network”. 341-346. 10.1109/RTSI.2019.8895580.
5. Rebecca Emerton., QingyunDuan ., Andrew W. Wood, Fredrik&Wetterhall, David E.Robertson(2021). “Ensemble Flood Forecasting: Current Status and Future Opportunities”,
https://doi.org/10.1002/wat2.1432.
6. Sheikh Amir Fayaz, Majid Zaman, Muheet Ahmed Butt(2022). “Knowledge Discovery in GeographicalSciences—A Systematic Survey of Various Machine Learning Algorithms for Rainfall Prediction”. https://www.researchgate.net/publication/354289092.
7. Qiao, S., Chen, D., Wang, B., Cheung, H. N., Liu, F., Cheng, J., Tang, S., Zhang, Z., Feng, G., & Dong, W. (2021). “The Longest 2020 Meiyu Season Over the Past 60 Years: Subseasonal Perspective and Its Predictions”. Geophysical Research Letters, 48(9), 1–11. https://doi.org/10.1029/2021GL093596.
8. Ardabili SF, et al. “Covid-19 outbreak prediction with machine learning” 2020. Available at SSRN 3580188.
9. Chawla M, Verma H, Kumar V. “RETRACTED: A new statistical PCA–ICA algorithm for location of R-peaks in ECG”. Elsevier; 2008.
10. Li X, Shang W, Wang S. “Text-based crude oil price forecasting: A deep learning approach”. Int J Forecast 2019;35(4):1548–60.
11. Kırbas¸ I˙, et al. “Comparative analysis and forecasting of COVID-19 cases in

var-ious European countries with ARIMA, NARNN and LSTM approaches". Chaos, Solitons & Fractals; 2020.
12. Bandyopadhyay SK, Dutta S. "Machine learning approach for confirmation of covid-19 cases: Positive, negative, death and release", 2020. medRxiv.
13. Chen, X.Y.; Jiang, Y.K.; Li, M.Y.; Ke, X.W. "A study of public bicycle single-site scheduling demand based on BP neural network". Transp. Res. 2016, 3, 30–35.
14. Mattson, J.; Godavarthy, R. "Bike share in Fargo, North Dakota: Keys to success and factors affecting ridership". Sustain. Cities Soc.2017, 34, 174–182.
15. Li, Y.H.; Ma, Y. "LSTM-based bicycle sharing demand predictionSmart City", 2019, 5, 1–4.
16. Liu, X.N.; Wang, J.J.; Zhang, T.F. "A Method of Bike Sharing Demand Forecasting". In Applied Mechanics and Materials; Trans TechPublications Ltd.: Bach, Switzerland, 2014; Volume 587–589, pp. 1813–1816.
17. A. Azran and Z. Ghahramani, "A new approach to data driven clustering". Proceedings of the 23rd International conference on Machine Learning, pp. 57-64, 2006.
18. M. Singh, S. Sharma, and A. Kaur, "Performance analysis of decision trees," International Journal of Computer Applications, vol. 71, no. 19, pp. 10-14, 2013.
19. J. Yuan and Y. Tian, "An intelligent fault diagnosis method using GRU neural network towards sequential data in dynamic processes", vol. 7, no. 3, pp. 152-172, 2019.
20. Kar, Kaveri, Neelima Thakur, and Prerika Sanghvi. "Prediction of Rainfall Using Fuzzy Dataset." (2019).
21. Chen, Binghong, et al. "Non-Linear Machine Learning Approach to Short-Term Precipitation Forecasting." (2018).