

Multidimensional Precipitation Index Prediction Based on CNN-LSTM Hybrid Framework

Yuchen Wang¹, Pengfei Jia^{1*}, Zhitao Shu², Keyan Liu³, Abdul Rashid Mohamed Shariff¹

¹Department of Biological and Agricultural Engineering, Universiti Putra Malaysia, Kuala Lumpur, Malaysia

²Vanderbilt University, Nashville, USA

³Arizona State University, Houston, USA

**Corresponding Author. Email: jiapengfei97@gmail.com*

Abstracts: Amid accelerating global climate change, accurate prediction of meteorological indicators is crucial for disaster prevention, agriculture, and transportation. Precipitation, a key factor, significantly impacts water resource management, agriculture, and urban flood control. This study proposes a CNN-LSTM hybrid model for multidimensional precipitation index prediction to enhance forecasting accuracy. Using monthly mean precipitation data from Pune, India (1972–2002), the model captures both local features and long-term dependencies in the time series. Experimental results demonstrate that the model achieves an RMSE of 6.752, outperforming traditional time series prediction methods. Although effective, the model demands high computational resources and its ability to predict multidimensional precipitation data remains limited. Future work should focus on optimizing computational efficiency and expanding the model's capacity for multidimensional prediction, contributing to more accurate and efficient meteorological forecasting technologies.

Keywords: CNN-LSTM, Precipitation Prediction, Hybrid Framework, Time Series, Meteorological Forecasting, Dataset

1. Introduction

As global climate change worsens, frequent meteorological disasters significantly impact society and the environment. Accurate meteorological data analysis and prediction are crucial for disaster prevention, agriculture, and transportation. Among various factors, precipitation is a critical indicator for water resource management, agricultural production, and urban flood control. However, due to the complexity and uncertainty of precipitation data, traditional prediction methods often face challenges such as high computational demands, limited accuracy, and slow response times, becoming increasingly inadequate in current applications.

Precipitation forecasting has long been a major focus in meteorology. Influenced by atmospheric circulation, temperature, and humidity, precipitation exhibits strong seasonal variations and short-term fluctuations, making prediction particularly difficult during extreme weather events. Traditional methods, including regression analysis, time series models, and numerical weather prediction (NWP) techniques [1], struggle to capture complex spatiotemporal relationships, especially for localized forecasts.

Recent advances in deep learning offer new opportunities for precipitation prediction. Deep learning models excel at capturing nonlinear relationships and spatiotemporal patterns in data, transforming precipitation forecasting into a more effective sequence prediction task. Models like Convolutional Neural Networks (CNN) and Long Short-Term Memory networks (LSTM) automatically learn complex patterns from historical data, significantly improving prediction accuracy and efficiency.

This study proposes a multidimensional precipitation index prediction model based on a hybrid CNN-LSTM framework, leveraging CNN's feature extraction capabilities and LSTM's strength in modeling long-term dependencies. The dataset, sourced from Pune, Maharashtra, India, spans 31 years (1972–2002) of monthly mean precipitation data, reflecting seasonal trends and long-term fluctuations. By analyzing this time series, the proposed model effectively captures both local temporal features and long-term dependencies, offering a novel and more reliable approach to multidimensional precipitation forecasting.

2. Related work

Time series forecasting is crucial in weather forecasting, financial - market analysis, and traffic flow prediction. In meteorology, precipitation forecasting is vital for water resource management, farming, and urban flood prevention. However, due to the complexity and nonlinearity of time series data, especially during extreme weather and with multidimensional meteorological data, traditional forecasting methods face challenges in accuracy and generalization.

In recent years, deep learning has advanced precipitation forecasting. Chen G et al. [2] developed a 3D CNN using 39 years of U.S. data, outperforming state-of-the-art models in daily and 5-day forecasts, especially for heavy precipitation, with millisecond-scale inference enabling large ensemble predictions. Jin W et al. [3] proposed a deep learning model for seasonal precipitation prediction in China, pre-trained on GCM hindcasts and fine-tuned with ERA5 data, improving PCCs by 0.10–0.13 over raw GCM outputs, and effectively guiding summer precipitation forecasts at China's National Climate Center.

Manna T et al. [4] integrated Rough Set on Fuzzy Approximation Space (RSFAS) with Deep Learning (DL) to predict precipitation in southern coastal India, enhancing accuracy and reducing errors. Rojas-Campos A et al. [5] developed DL models to generate precipitation maps using NWP variables, improving prediction resolution and bias correction. The CGAN model achieved the best results, though some biases remained, suggesting further research directions.

Huang et al. [6] proposed a wavelet-NAR neural network-based time series prediction model for meteorological elements, which combines the advantages of wavelet analysis and nonlinear autoregressive neural network, extracts the features and removes the noise through multi-scale decomposition, effectively captures the nonlinear dynamic relationship, and the prediction results are better than those of other models, which can provide a scientific basis for the development of weather derivatives. C.D. Tian et al. [7] proposed a combined kernel function least squares support vector machine (LSSVM) prediction model, which uses a combination of polynomial function and radial basis function to construct the kernel function, and uses an improved genetic algorithm to optimise the parameters, which has the advantages of both global and local kernel function and improves the model's generalisation ability, making it superior to the traditional LSSVM model with a single kernel function and other common prediction models. Jianxin Sun et al. [8] proposed a hybrid model for dynamic correlation time series, combining ARIMA model, multiple codecs and CNN, and weighted and summed by softmax function to get the final prediction results.

3. Data introduction

3.1. Data source

The dataset, from Pune, Maharashtra, India, covers monthly mean precipitation data from 1972 to 2002, reflecting long-term trends and seasonal variations. It captures regional precipitation patterns and extreme events, providing essential data for training weather prediction models. Missing values, marked as "-99" due to observational issues, were removed during data cleaning to ensure dataset integrity and reliability.

3.2. Descriptive statistical analysis

Figure 1 shows the monthly precipitation distribution from 1972 to 2002. The x-axis represents months, and the y-axis indicates precipitation in millimeters. A clear seasonal pattern is observed, with peaks from June to September due to the monsoon, and low, stable values from January to April. October to December marks the rainy season's end. The variability during the wet season reflects seasonal climatic influences, crucial for enhancing precipitation prediction models.

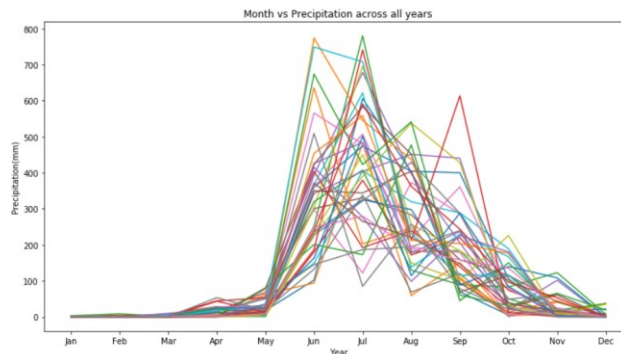


Figure 1: Yearly precipitation trends across different months

Figure 2 presents a boxplot showing the distribution of monthly precipitation across all years in the dataset. The x-axis represents the months from January to December, while the y-axis indicates precipitation in millimeters. From the plot, it is clear that the rainfall in November, December, January, February, March, and April is very low, with January and February experiencing almost zero precipitation, reflecting the dry season. In contrast, June, July, and August show significantly higher precipitation, with these months experiencing much higher rainfall compared to the rest of the year, indicating the influence of the monsoon season. The range of precipitation during these months is also broader, suggesting variability in rainfall even during the monsoon period, influenced by various climatic factors each year. This boxplot highlights the seasonal effect on precipitation, emphasizing the clear distinction between the wet and dry seasons, which is crucial for improving precipitation forecasting models.

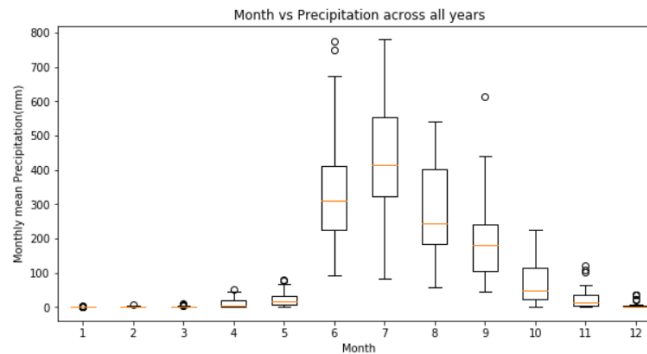


Figure 2: Boxplot of monthly precipitation distribution across all years in the pune region

4. CNN-LSTM fusion framework

4.1. CNN

Convolutional Neural Networks (CNN) are a type of feedforward neural network, particularly effective in processing data with grid-like structures, such as image data. The core advantages of CNN lie in its local perception and weight sharing. In the task of precipitation prediction, CNN automatically learns the local features of precipitation data through its convolutional layers, such as seasonal fluctuations and patterns of extreme precipitation events. By applying sliding convolution operations with different kernels on time-series data, CNN can extract various representative local patterns, which are crucial for recognizing the changing trends in precipitation. For example, specific patterns of precipitation distribution in certain seasons or months can serve as unique identifiers of particular precipitation types, and CNN can effectively capture these local features [9].

In this study, CNN is used to extract time-series features from precipitation data, leveraging the convolutional layers to identify patterns at different time scales. Specifically, CNN is able to recognize short-term fluctuations and seasonal characteristics in precipitation, which are essential for modeling long-term precipitation trends. Additionally, the pooling layer further reduces the dimensionality of the feature maps output by the convolutional layers, while retaining key features. This reduction in data size and computation enhances the training efficiency and generalization ability of the model. By incorporating CNN into the CNN-LSTM hybrid framework, this study effectively extracts local features from precipitation data and combines them with the LSTM module to capture long-term dependencies in precipitation.

4.2. LSTM

As a special type of recurrent neural network (RNN), LSTM excels in handling time - series data, making it highly effective for precipitation prediction. Its gate mechanisms help address the gradient vanishing and exploding problems in traditional RNNs, enabling it to capture long - term dependencies in sequences. This is particularly useful for precipitation data, which shows clear seasonality and long-term trends [10].

In the CNN-LSTM hybrid framework proposed in this study, LSTM processes time-series data. CNN first extracts spatial features from the input, and the features are then fed into the LSTM network. LSTM's cell states and gates allow it to selectively retain or discard information, effectively capturing temporal dependencies like seasonality and trends in precipitation data.

During training, LSTM adjusts its weights and biases via backpropagation to minimize prediction errors. In precipitation prediction, LSTM learns the relationships between different time steps,

enabling accurate future precipitation forecasting. Experiments show that LSTM can capture complex patterns in time-series data, enhancing the accuracy and reliability of precipitation predictions.

4.3. CNN-LSTM fusion framework

Model construction is an important part of the research investigation of weather indicator prediction models. In this paper, we constructed hybrid model Conv1D and LSTM. The main role is used to extract the local features of the time series data. The presence of a convolutional layer and pooling layer enables the model to automatically learn the spatial and in the original dataCNN temporal features, and to perform the dimensionality reduction and compression of the feature maps based on capturing the local features. In the model, the number of convolution kernels is 32, the kernel size is 5, the convolution step size is 1, and the Relu activation function used is to introduce nonlinearity, padding="causal".

LSTM, on the other hand, is primarily used to model long-term dependencies in time-series data by introducing three key gating structures; the forgetting gate determines the amount of information that needs to be discarded based on the current input and previous memory states. The input gate then determines what new information is worth being written to the state. Finally, the output gate combines the updated state and the current input to generate a prediction, which to some extent solves traditional gradient vanishing and gradient explosion problems when dealing with long sequential data RNNs. In this study, two layers of LSTM are by setting stacked the parameter values of units to LSTM 64 and 60 and return_sequences=True in the two models respectively. In addition, are also two fully connected layers used, the first containing 30 neurons and the second containing 10 neurons, and the activation functions are both ReLU, so that the last fully connected layer outputs only 1 value, which represents the predicted precipitation and the Lambda layer scales the output value to the original precipitation range. The period training epoch is set to 50, and the results are shown in Table 1.

Table 1: Model training parameters

mould	loss function	optimizer	learning rate strategy	batch size	window size	training cycle
CNN- LSTM model	Huber	Adam	Initial 1e-8; increases 10 times every 20 epochs.	256	64	50

5. Experiments

Figure 3 presents the performance of the CNN-LSTM model in predicting monthly precipitation on the test dataset from the Pune region. The graph plots actual future precipitation values against the model's forecasts. The lines for actual and predicted precipitation follow similar trends, especially in capturing the peaks and troughs of rainfall across the years. This shows the model's ability to recognize seasonal patterns and long-term trends in precipitation data. However, there are discrepancies in certain years where the forecast deviates from the actual values, indicating potential limitations in handling specific data fluctuations or extreme events. Overall, the model demonstrates reasonable accuracy in precipitation prediction, effectively capturing the general pattern of rainfall variation and providing a reliable reference for weather forecasting tasks.

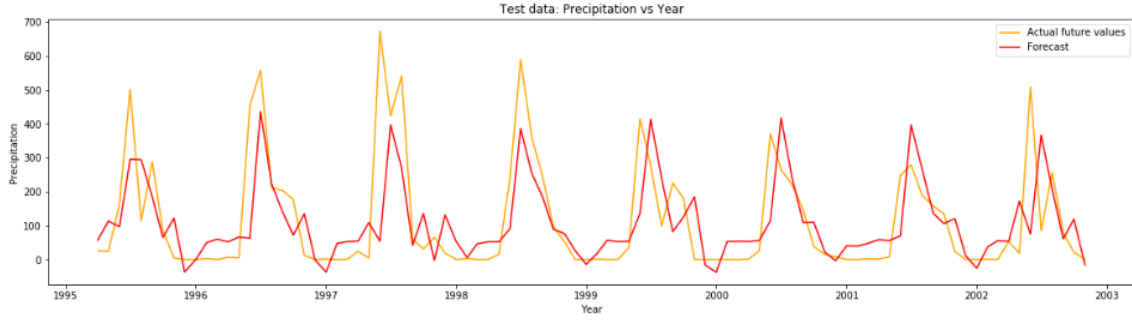


Figure 3: The predicted precipitation value of CNN-LSTM

The specific evaluation indicators for the CNN-LSTM model include mean square error and root mean square error. The formula is as follows:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (1)$$

$$RMSE = \sqrt{MSE} \quad (2)$$

These metrics measured the prediction error of the CNN-LSTM model on the test dataset of the precipitation data. The calculated Mean Squared Error (MSE) was approximately 45.59, and the Root Mean Squared Error (RMSE) was 6.752. Given that the monthly mean precipitation data in this experiment ranged from 0 to around 700 mm, the MSE and RMSE values indicate that the model's predictions were relatively close to the actual values. This suggests that the CNN-LSTM model has a decent performance in predicting precipitation, effectively capturing the general trend and seasonal variations of rainfall in the Pune region. However, there is still room for improvement, especially in reducing the prediction errors for years with significant precipitation fluctuations or extreme weather events.

6. Conclusions

This study proposes a multidimensional precipitation index prediction model based on the CNN-LSTM hybrid framework and conducts a systematic experimental investigation. By analyzing monthly precipitation data from Pune, Maharashtra, India, from 1972 to 2002, the experimental results show that the CNN-LSTM model performs well in precipitation prediction tasks. Specifically, the model's performance on the test dataset indicates that the Mean Squared Error (MSE) is approximately 45.59, and the Root Mean Squared Error (RMSE) is 6.752. Considering that the monthly mean precipitation data in the experiment range from 0 to around 700 mm, these MSE and RMSE values suggest that the CNN-LSTM model's predictions are relatively close to the actual values and are able to capture the overall trend and seasonal variations of precipitation in the Pune region effectively. However, for years with significant precipitation fluctuations or extreme weather events, the model's prediction errors still need to be further reduced.

Although the CNN-LSTM model can effectively capture the seasonal features and long-term trends of precipitation, its predictive ability remains limited when facing extreme weather events, especially in years with significant changes in precipitation. Future research can further optimize the model structure, such as incorporating additional data features or more complex network architectures, to improve prediction accuracy for extreme weather events. Moreover, given the high computational demands of large-scale dataset processing, future studies could explore lighter versions of the model to enhance its efficiency and applicability in practical scenarios.

References

- [1] Al-Yahyai, S., Charabi, Y., & Gastli, A. (2010). Review of the use of numerical weather prediction (NWP) models for wind energy assessment. *Renewable and Sustainable Energy Reviews*, 14(9), 3192-3198.
- [2] Chen, G., & Wang, W. C. (2022). Short-term precipitation prediction for contiguous United States using deep learning. *Geophysical Research Letters*, 49(8), e2022GL097904.
- [3] Jin, W., Luo, Y., Wu, T., Huang, X., Xue, W., & Yu, C. (2022). Deep learning for seasonal precipitation prediction over China. *Journal of Meteorological Research*, 36(2), 271-281.
- [4] Manna, T., & Anitha, A. (2023). Precipitation prediction by integrating rough set on fuzzy approximation space with deep learning techniques. *Applied Soft Computing*, 139, 110253.
- [5] Rojas-Campos, A., Langguth, M., Wittenbrink, M., & Pipa, G. (2023). Deep learning models for generation of precipitation maps based on numerical weather prediction. *Geoscientific Model Development*, 16(5), 1467-1480.
- [6] Huang JF, Lu WC. Wavelet - NAR neural network based time series prediction of meteorological elements and rainbow option valuation of weather index [J]. *Systems Engineering Theory and Practice*, 2016, 36(5): 1146-1155.
- [7] C.D. Tian, X.W. Gao, T. Shi. Combined kernel function least squares support vector machine for chaotic time series prediction [J]. *Physics Letters*, 2014, 63(16).
- [8] Sun Jianxin. Modelling for dynamic correlation time series forecasting [D]. Harbin Engineering University, 2021.
- [9] Sadeghi, M., Asanjan, A. A., Faridzad, M., Nguyen, P. H. U., Hsu, K., Sorooshian, S., & Braithwaite, D. A. N. (2019). PERSIANN-CNN: Precipitation estimation from remotely sensed information using artificial neural networks-convolutional neural networks. *Journal of Hydrometeorology*, 20(12), 2273-2289.
- [10] Shi, X., Chen, Z., Wang, H., Yeung, D. Y., Wong, W. K., & Woo, W. C. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28.