# Study and Prediction of Extreme Rainfall Events On Indian Region

*A dissertation submitted in*
*partial fulfilment for the degree of*

**Master of Technology**

in

**Computer Science**

*by*

**Nayan Giri**
Roll no. - [**CS2218**]

*under the supervision of*

**Dr. Sarbani Palit**
Compute Vision and Pattern Recognition Unit


Indian Statistical Institute, Kolkata
**June, 2024**

# CERTIFICATE

This is to certify that the dissertation entitled "**Study and Prediction of Extreme Rainfall Events On Indian Region**" submitted by **Nayan Giri** to the Indian Statistical Institute, Kolkata, in partial fulfillment for the award of the degree of Master of Technology in Computer Science is an authentic and genuine record of the research work conducted by the candidate under my supervision and guidance. I affirm that the dissertation has met all the necessary requirements in accordance with the regulations of this institute.

June,2024

Sarbani Palit

**Dr. Sarbani Palit**
Compute Vision and Pattern Recognition Unit,
Indian Statistical Institute,
Kolkata-700108

# Acknowledgement

I would like to express my utmost gratitude to my esteemed advisor, Dr. Sarbani Palit, from the Computer Vision and Pattern Recognition Unit at the Indian Statistical Institute, Kolkata. Her unwavering guidance, continuous support, and encouragement have been invaluable to me throughout this research journey. Under her mentorship, I have learned the art of conducting thorough and impactful research, and her insightful ideas have constantly motivated and inspired me.

I extend my deepest thanks to all the research scholars at the Indian Statistical Institute for their invaluable suggestions and engaging discussions, which have significantly enriched the depth and quality of my research work.

I would also like to express my sincere appreciation to all my friends for their unwavering assistance and support. I am grateful to everyone who has contributed to my growth and success, even if I have inadvertently missed mentioning them in the above list.

# Declaration

I, **Nayan Giri**, with Roll No. **CS2218** hereby declare that the material presented in the dissertation titled **Study and Prediction of Extreme Rainfall Events On Indian Region** represents original work carried out by me for the degree of **Master of Technology, Computer Science** at **Indian Statistical Institute, Kolkata**.

Furthermore, I affirm that no sections of this report have been sourced or copied from external references without proper attribution. I am aware that any instances of plagiarism or the utilization of unacknowledged materials from third parties will be treated with utmost seriousness and consequences.

**Nayan Giri**
Roll no. - CS2218

# Abstract

Extreme rainfall events, commonly known as cloudbursts, are significant weather phenomena characterized by exceptionally high intensity of precipitation within a short period. These events can result in devastating flash floods, landslides, and avalanches, causing extensive damage to infrastructure, property, and significant loss of life. Despite various measures and advancements in meteorological science to mitigate their impact, predicting these events with high accuracy remains a formidable challenge. This study focuses on applying cutting-edge machine learning techniques to anticipate heavy rainfall events in the Indian subcontinent, specifically leveraging ConvLSTM neural networks. By integrating diverse meteorological datasets, including potential vorticity, relative humidity, cloud cover, temperature, and surface pressure, this research aims to develop a robust predictive model. Leveraging the historical data, the ConvLSTM model is trained to discern intricate patterns and correlations between the input variables and cloudburst incidence, thus enabling accurate predictions of cloudburst probabilities within future timeframes. The empirical findings of this study reveal the ConvLSTM-based prediction model's remarkable accuracy and its capacity to furnish valuable insights into cloudburst event occurrences. To summarize, this project encompasses the development of an advanced ConvLSTM-based prediction model for cloudburst events, effectively harnessing historical meteorological data.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

This report investigates and predicts extreme rainfall events, commonly known as cloudbursts, on an particular region in India. Cloudbursts are intense weather phenomena that occur when a large amount of moisture in the air condenses rapidly into water droplets, resulting in a sudden and heavy downpour over a small area. In India, cloudbursts are most frequently observed in the Himalayan region, which is particularly susceptible to heavy rainfall due to its proximity to the monsoon winds. This region includes states such as Uttarakhand, Himachal Pradesh, and Jammu Kashmir, where steep terrain exacerbates the consequences of these events [3]. The high intensity of rainfall during cloudbursts can lead to flash floods, landslides, and avalanches, causing widespread damage and loss of life. [4]

To reduce the impact of extreme rainfall in India, various measures have been undertaken. One such measure is the construction of check dams—small, temporary barriers built across rivers and streams to reduce water flow velocity and prevent erosion. Additionally, early warning systems utilizing meteorological data have been established to predict the likelihood of cloudbursts and alert communities to take necessary precautions. These systems help authorities evacuate people from danger zones to safer locations before a disaster strikes.

Despite these efforts, preventing the impact of cloudbursts entirely remains a significant challenge due to their unpredictable nature. Factors such as deforestation, illegal mining, and uncontrolled urbanization have exacerbated the effects of cloudbursts in some regions. Deforestation, for example, has led to soil erosion, increasing the likelihood of landslides during cloudbursts.

This study aims to advance the prediction of extreme rainfall events in India using advanced machine learning techniques, specifically ConvLSTM neural networks. By integrating diverse meteorological datasets, this research seeks to develop a robust predictive model.

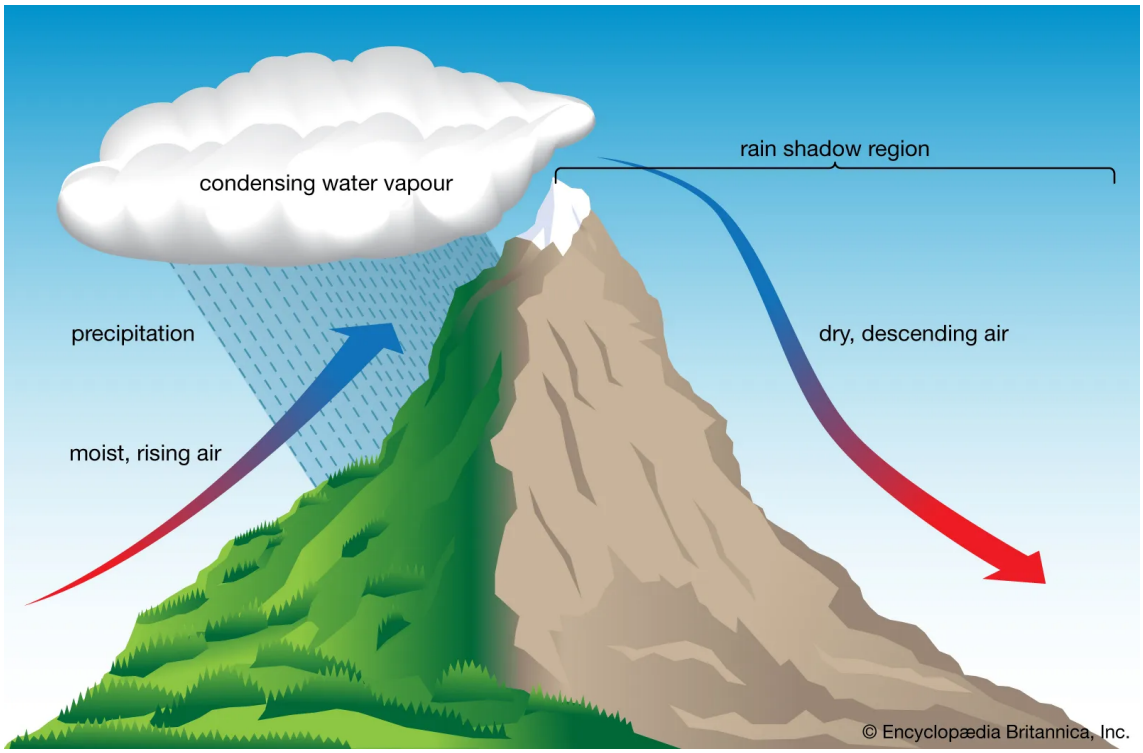


Figure 1.1: Condensation-precipitation-rain shadow effect [1]

## 1.1 Some major Atmospheric variables influencing Rainfall

Extreme rainfall events in the Indian region are influenced by a variety of atmospheric variables. Understanding these can help in predicting and mitigating the impacts of such events. Here are some major atmospheric variables that can cause extreme rainfall:

### 1.1.1 Monsoon Winds:

#### Southwest Monsoon

The primary driver of rainfall during the summer months (June to September). The strength and onset of the southwest monsoon can significantly influence rainfall patterns.

#### Northeast Monsoon

Affects the southern parts of India, particularly Tamil Nadu, during the winter months (October to December).

### 1.1.2 Temperature:

Surface temperatures can affect the convective activity. Higher temperatures can lead to increased evaporation and subsequently more moisture in the atmosphere.

### 1.1.3 Relative Humidity:

High humidity levels in the lower and mid-troposphere provide the necessary moisture for rainfall. Heavy precipitation may occur due to the presence of winds carrying moisture from the Bay of Bengal and the Arabian Sea.

### 1.1.4 Potential Vorticity:

Potential vorticity (PV) is a crucial atmospheric dynamic quantity that helps in understanding and predicting heavy rainfall events. It combines the effects of rotation (vorticity), stratification (static stability), and thickness (depth of the atmospheric layer). High PV values often indicate the presence of cyclonic activity, which can enhance upward motion and lead to heavy rainfall. When a region of high PV interacts with moist air, it can trigger strong convection, leading to intense and localized rainfall. This interaction is particularly significant in regions where atmospheric conditions are already conducive to storm development, such as during the monsoon season. Therefore, monitoring PV can provide valuable insights into the development and intensification of heavy rainfall events.

It is important to note that cloudbursts can be unpredictable and difficult to forecast, making them a significant natural hazard. Proper disaster management and preparedness measures are crucial to minimize the impact of cloudbursts on communities and infrastructure.

## 1.2 Some Extreme Rainfall Events

### 1.2.1 Recent Floods

During the summer monsoon of 2016, the central Indian region faced severe and extreme rainfall events, which resulted in extensive flooding. These floods caused significant damage to both life and property along the west coast, including the city of Mumbai, and led to the submergence of Kaziranga National Park in Assam, located in northeastern India. Analysis of satellite rainfall data from this period reveals a major heavy rainfall event that persisted for three days. The regional extent of this event mirrors the growing trends in extreme rainfall incidents. Similarly, the floods in Mumbai in 2017 exhibited a comparable pattern

### 1.2.2 Historical Floods

We investigated several historical floods that led to significant loss of life and property. Notable examples include the central Indian floods of 1989 and 2000, the Mumbai floods of 2005, and the South Asian floods of 2007. These events were marked by intense rainfall over a span of three days, with moisture sourced from the westerly flow of the Arabian Sea.

## 1.3 Problem Definition

The objective of this project is to study and predict extreme rainfall events in the Indian region, with a particular focus on Mumbai. Mumbai, being a coastal city, is highly susceptible to severe weather conditions, including severe rainfall occurrences that have the potential to cause serious socioeconomic problems like flooding, traffic jams, and property and human casualties. Precise and prompt forecasting of these occurrences is essential for efficient catastrophe management and reduction tactics.

To address this problem, we employ advanced machine learning techniques, specifically Convolutional Long Short-Term Memory (ConvLSTM) networks and ConvLSTM networks augmented with Attention mechanisms. These methods are chosen for their ability to capture both spatial and temporal dependencies in the data, which are essential for accurately modeling complex atmospheric processes leading to extreme rainfall.

# Chapter 2

# Related Work

Recent advancements in deep learning algorithms have greatly improved performance in tasks such as image classification, pattern recognition, and computer vision. Convolutional Neural Networks (CNNs) have been particularly successful in applications like object detection and tracking.

- **Tiwari and Verma (2015)** [5] proposed a method using Arduino, an open-source electronics platform, as a central component for cloudburst prediction. The system integrates various sensors and data processing capabilities to perform real-time calculations of rainfall intensity. This setup allows for continuous monitoring and immediate assessment of rainfall data, aiding in the prediction of potential cloudburst events.

- **Dimri, Thayyen, et al. (2016)** [6] focused on predicting cloudburst events in the Indian Himalayan region using an Artificial Neural Network (ANN) model. By collecting meteorological data and applying the ANN model, they provided valuable insights into forecasting cloudburst occurrences in this area.

- **Goswami (2017)** [7] introduced a cloudburst prediction mechanism based on detecting cumulonimbus clouds by utilizing the brightness temperature (TB) difference between 19 and 91 GHz values in horizontal polarization from NASA's SSMI satellite. This method relies on a threshold for the TB difference to identify cumulonimbus clouds, often associated with cloudbursts. Tests conducted across multiple locations in India from 2013 to 2016 indicated that this approach could provide cloudburst predictions with a lead time of 1-4 days.

- **Pabreja and Datta (2016)** [8] employed data mining techniques to analyze weather forecasts, aiming to extract valuable insights about cloudburst events. By exploring and analyzing large datasets, they sought to uncover patterns and trends that could signal the likelihood of cloudbursts. Their objective was

to enhance the interpretation of weather forecasts, improving the accuracy and timeliness of cloudburst predictions through the identification of specific weather patterns or indicators.

- **Sivagami et al. (2021)** developed a novel cloudburst prediction model leveraging deep learning techniques. Their model is based on a comprehensive dataset of cloudburst events in Uttarakhand over the past decade. They evaluated the model using Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) time series sequence models, demonstrating its effectiveness in forecasting cloudbursts in specific geographical regions.

- In their 2016 study, **Gope et al.** [9] introduced a novel deep-learning approach utilizing a Stacked AutoEncoder (SAE) model for predicting heavy rainfall in Mumbai and Kolkata with a lead time ranging from 6 to 48 hours. Their model incorporated various climatic parameters including temperature, relative humidity, and also other atmospheric variables to enhance rainfall forecasts. Despite these advancements, the study fell short in addressing extreme weather events, and there was room for improvement in capturing spatial variability.

# Chapter 3

# Proposed Architecture

In our study on predicting rainfall, we employed two main models: the random forest regressor and the convLSTM2d model. We tested both with and without attention mechanisms to see which performed better.

## 3.1 Baseline Models

### 3.1.1 Random forest regressor

The random forest regressor is a type of machine learning algorithm used for prediction tasks, like forecasting rainfall. It works by creating multiple decision trees during training and then averaging their predictions to improve accuracy. In rainfall prediction, the random forest regressor considers various input factors such as temperature, humidity, cloud coverage percentage, and historical rainfall data to make predictions about future rainfall amounts. It's favored for its ability to handle complex relationships between input variables and output predictions.

### 3.1.2 Vanilla stacked LSTM Model

Long Short-Term Memory (LSTM) networks were introduced to address the limitations of traditional Recurrent Neural Networks (RNNs) in learning long-term dependencies in sequential data. Traditional RNNs struggle with vanishing and exploding gradient problems, making them ineffective for capturing relationships over long time intervals. LSTMs, with their unique cell structure, including forget, input, and output gates, can selectively remember and forget information, making them well-suited for handling long-term dependencies.

In time series forecasting, LSTMs are highly effective as they can learn patterns and trends over time. They are used to model sequences of data points, such as weather data, stock prices, or any other sequential dataset. By capturing temporal

dependencies, LSTMs can predict future values based on past observations, making them a powerful tool for forecasting tasks. Their ability to handle varying lengths of sequences and maintain relevant information over long periods makes LSTMs ideal for time series analysis and prediction.

## 3.2 ConvLSTM 2D model

The ConvLSTM2d model, introduced by Shi et al. in 2015 [10], combines convolutional and LSTM layers. Convolutional layers extract spatial features, while LSTM layers capture temporal dependencies. This architecture is valuable for analyzing spatio-temporal data like video sequences or time-series data. It finds application in video prediction, action recognition, and anomaly detection. Specifically, it excels in predicting future frames of a video by understanding temporal relationships between frames. During training, backpropagation through time (BPTT), a variant of backpropagation accounting for temporal dependencies, optimizes the model. The objective is to minimize the mean squared error (MSE) between predicted and actual outputs, as outlined by Kato and Hotta in 2021.

The input to the ConvLSTM2d model is a 4D tensor, defined by dimensions (batch size, time steps, channels, height, width). Here, the batch size signifies the number of samples in a batch, time steps indicate the frames in a video sequence or time-series data, channels represent the input feature maps, while height and width denote the spatial dimensions of the input.

The ConvLSTM model leverages convolutional layers to extract spatial features from the input data, enabling the capture of significant patterns and relationships. To improve upon this model, we can consider extending or modifying the basic time series ConvLSTM structure by integrating various layers or experimenting with different combinations of weights for the input data. These adjustments could potentially boost the model's overall performance and accuracy.

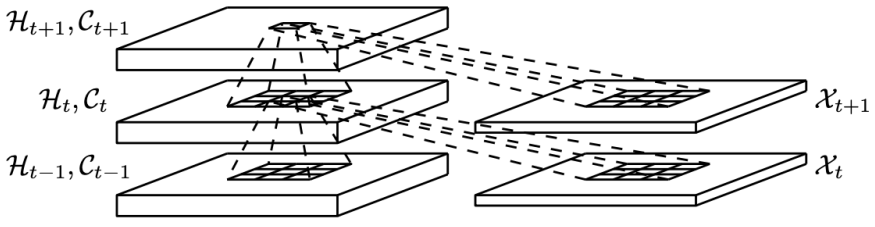


Figure 2: Inner structure of ConvLSTM

The methodological flowchart of the model outlines the process steps, including selecting input variables for 2D spatial grids, preparing and normalizing data, defin-

ing the ConvLSTM2D model layers, and generating model outputs for the selected grids. Figure 2 [11] then delves into the inner architecture of the ConvLSTM model. This model controls data flow within the cell through the forget gate ($f_t$), input gate ($i_t$), and output gate ($o_t$). The forget gate decides how much information the model should retain or discard, determining whether to remove or keep relevant data at each iteration's end. In ConvLSTM2D models, the memory cell ($c_{t-1}$) serves as an information accumulator, helping to gather information at every state. The LSTM uses multiple self-parametrized gates to manage access to the cell. The input gate gathers information into the cell, allowing new data to be incorporated into long-term memory. The forget gate helps to remove outdated information from the cell state ($c_{t-1}$). And, finally that information will passes to next LSTM cell through the output gate. The updated cell information is multiplied by the output gate values, and the result is the hidden state $h_t$), which is then calculated using an activation function (tanh, in this case).

The expression (mathematical) of the ConvLSTM layer is as follows (for each time step $t$):

1. Input gate

$$i_t = \sigma(w_{xi} * x_t + w_{hi} * h_{t-1} + w_{ci} \odot c_{t-1} + b_i)$$

2. Forget gate

$$f_t = \sigma(w_{xf} * x_t + w_{hf} * h_{t-1} + w_{cf} \odot c_{t-1} + b_f)$$

3. Cell state

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(w_{xc} * x_t + w_{hc} * h_{t-1} + b_c)$$

4. Output gate

$$o_t = \sigma(w_{xo} * x_t + w_{ho} * h_{t-1} + w_{co} \odot c_t + b_o)$$

5. Hidden state

$$h_t = o_t \odot \tanh(c_t)$$

where $*$ denotes the convolution operation and $\odot$ is the elementwise Hadamard product.

## 3.3 Data Details

### 3.3.1 Area of interest

Our area of interest is Mumbai, which is situated in the Western Ghats and has a tropical monsoon climate.The city, which has a 437.79 $km^2$ area, is situated between 18° and 19.2° N latitude and 72° and 73° E longitude (see Figure 3.1).

Mumbai experiences heavy rainfall primarily due to the southwest monsoon winds, which bring moisture-laden air from the Arabian Sea. When these winds hit the Western Ghats, they rise and cool, leading to intense precipitation. The city's coastal location and proximity to the Western Ghats contribute to orographic rainfall, where moist air is forced to ascend the mountains, cooling and condensing to form heavy rains. The urban heat island effect, created by Mumbai's dense infrastructure, can enhance local convection, leading to increased rainfall. Additionally, cyclonic activity in the Arabian Sea can bring heavy rains and strong winds, contributing to extreme weather events.

Climate change also plays a role, with increased sea surface temperatures and altered atmospheric circulation patterns leading to more frequent and intense rainfall events. Other meteorological phenomena, such as the Madden-Julian Oscillation (MJO), can influence rainfall patterns, resulting in periods of heavy rain. These factors, individually or in combination, contribute to the heavy rainfall events that Mumbai frequently experiences.

### 3.3.2 Used Data Details

For this project purpose we collected data from 'ERA-5' (re-analysis data). In this study we mainly focused on 6 predictor variables to forecast rainfall.These variables are temperature (t), potential vorticity(pv), relative humidity (rh), total cloud coverage (tcc), cloud coverage at high level (hcc), and surface pressure (sp). We collected hourly data for these variables from 2012 to 2023, which created a large dataset for our purpose. For the ConvLSTM2D model, we chose "total precipitation (tp)" as the target variable.

In this study, I used data in the NetCDF file format. NetCDF (Network Common Data Form) is a widely used format for storing array-oriented scientific data. The basic structure of a NetCDF file includes dimensions, variables, and attributes. Dimensions define the axes of the data, such as time, latitude, and longitude. Variables contain the data values and are associated with these dimensions. Attributes provide additional metadata, such as units and descriptions, to help interpret the data.

Figure 3.1: Interested region of study in India

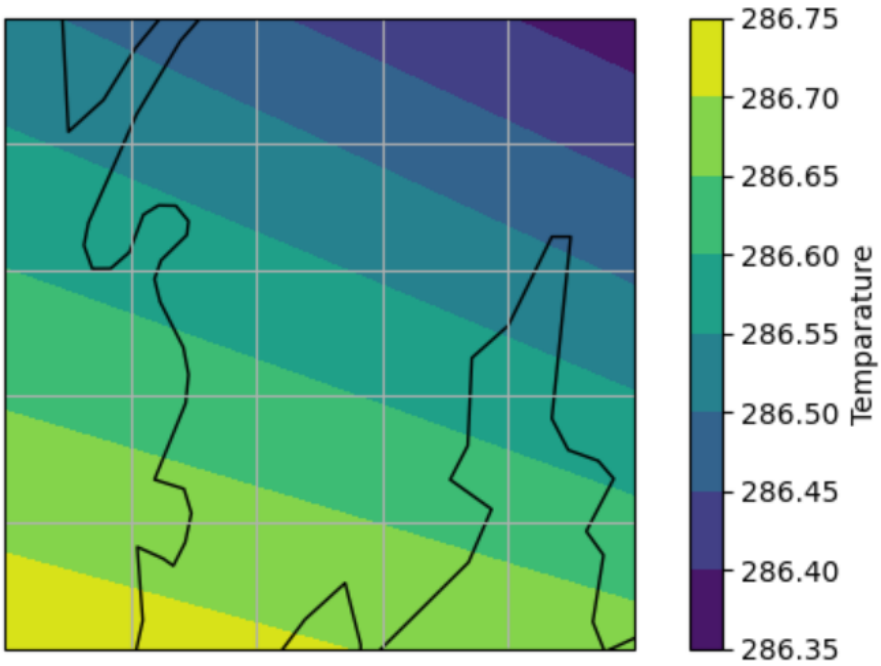


Figure 3.2: A netCDF data (temparature) on map.

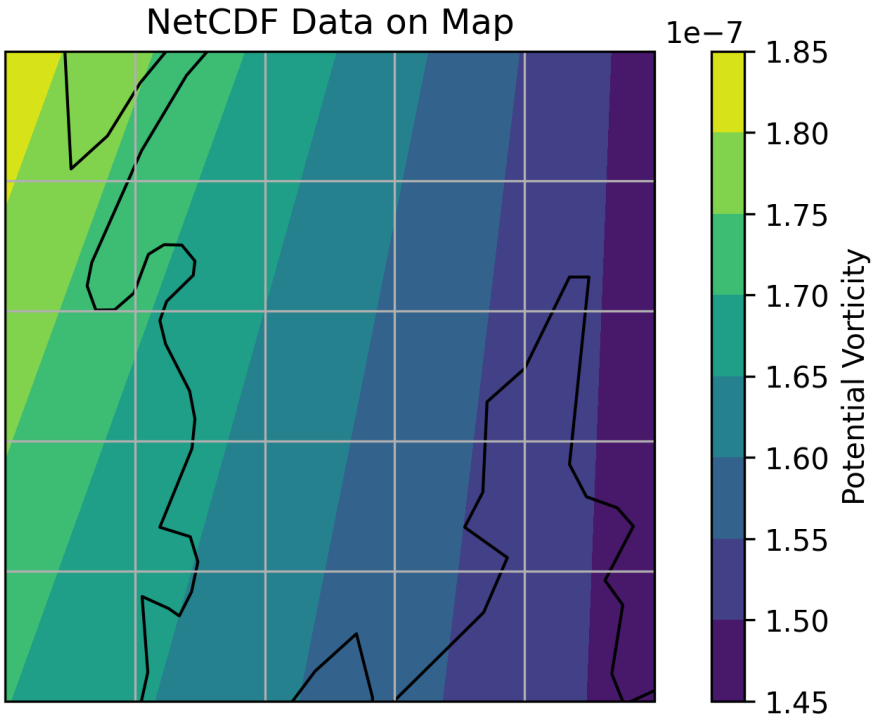


Figure 3.3: A netCDF data (potential vorticity) on map.

# Chapter 4

# Methodology

In this chapter, we outline the approach taken to carry out the study. This includes the selection of data, preprocessing steps, model development, and evaluation metrics used to assess the performance of our predictive models. We start by detailing the data sources and the variables selected for analysis, followed by the methods employed to preprocess and normalize the data. Next, we describe the architecture and training process of the ConvLSTM2D model used for precipitation forecasting. Finally, we explain the criteria and techniques used to evaluate the accuracy and reliability of the predictions.

## 4.1 Selection of Data

The research examines four grid areas covering Mumbai. We used the ERA-5 reanalysis (global) dataset [12]. The data was specifically tailored to Mumbai's geographical coordinates for the period from 2012 to 2023, resulting in 105,194 hourly observations for the selected six predictor variables.

## 4.2 Data Preprocessing

The data preprocessing involved loading and normalizing the dataset to ensure consistency and facilitate effective model training. The key steps are as follows:

**Loading Data:** The data was read from NetCDF files using the xarray library, which provides a convenient way to handle multi-dimensional arrays.

**Normalization:** Each variable's data was normalized to a range between 0 and 1. This was done by subtracting the minimum value and dividing by the range (maximum value minus minimum value). This step ensures that all variables are on a comparable scale, improving the performance and convergence of the model.

Next, we examined the relationship between the predictor and predicted variables

using a correlation matrix heat map. The matrix is displayed below.

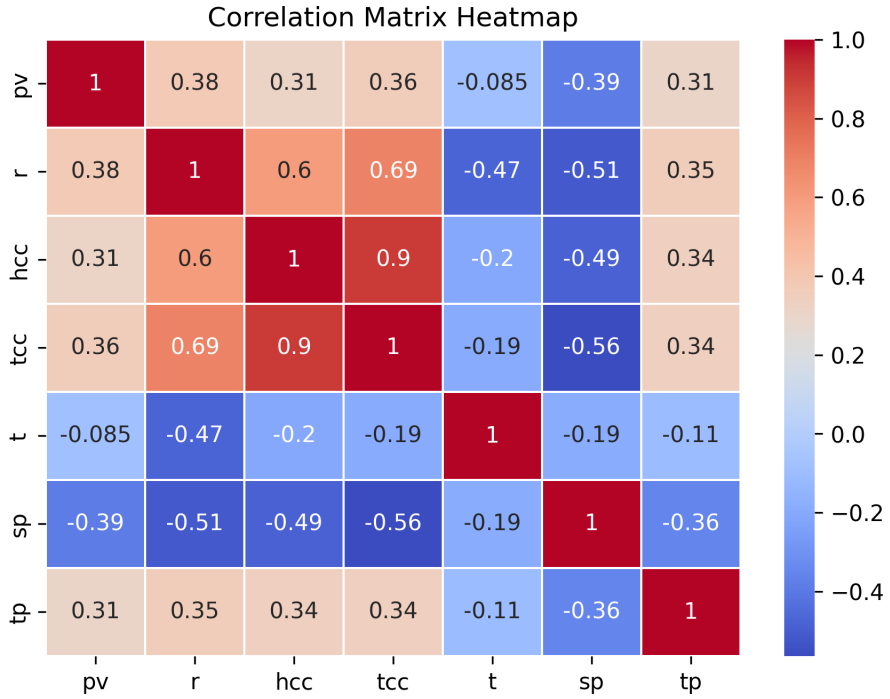


Figure 4.1: The correlation matrix of the atmospheric variables.

Out of the 6 predictors analyzed, surface pressure and temperature at 1000 hPa are the only ones showing a negative correlation with 'tp' (total precipitation).

The correlation heat map clearly shows that the predictor variables are not correlated with each other. Therefore, we cannot ignore any of the six predictor variables when predicting the amount of precipitation.

## 4.3 Model Development

Our proposed model is designed to predict precipitation using a ConvLSTM2D with attention architecture. Below are the details of the model:

### 4.3.1 Architecture:

**ConvLSTM2D Layers:**

- The first ConvLSTM2D layer has 256 filters, a kernel size of 2x2, and uses the ReLU activation function. It takes the input shape and returns sequences, maintaining the same padding.
- The second ConvLSTM2D layer has 128 filters, a kernel size of 2x2, and also uses the ReLU activation function. It continues to return sequences with the same padding.

**Batch Normalization:** Batch normalization is applied after each ConvLSTM2D layer to improve training stability and performance.

**Attention Layer:** An attention mechanism is added to focus on the most relevant parts of the input sequences.

**Flatten Layer:** This layer flattens the 3D output from the previous layers into a 1D array.

**Dense Layers:** The model includes two dense layers:

- The first dense layer has 256 units with ReLU activation and includes a dropout rate of 0.4 to prevent overfitting.
- The second dense layer has 4 units with a linear activation function, corresponding to the 2x2 grid output.

### 4.3.2 Compilation:

- The model is compiled using the Adam optimizer with a specified learning rate of 0.0001.
- The loss function used is Mean Squared Error (MSE), suitable for regression tasks such as precipitation prediction.

### 4.3.3 Training details:

The ConvLSTM2D model was trained using sequences created for 12-hour forecasts. We set the sequence length to 24 time steps and split the dataset into training and testing sets with an 85%-15% ratio. The data was reshaped to fit the input shape required by the model.

The model architecture includes ConvLSTM2D layers, batch normalization, an attention mechanism, and dense layers. The model was compiled with the Adam optimizer and a learning rate schedule that reduces the learning rate if the validation loss does not improve for 5 epochs.

Training was performed with a validation split of 15%, for 75 epochs, and a batch size of 32. After training, predictions were made on both the training and testing sets, reshaped to match the 2x2 grid format.

## 4.4 Experimental Details

In the context of rainfall prediction, it is crucial to use appropriate evaluation metrics to assess the performance of different models. Here are brief definitions of the evaluation metrics used:

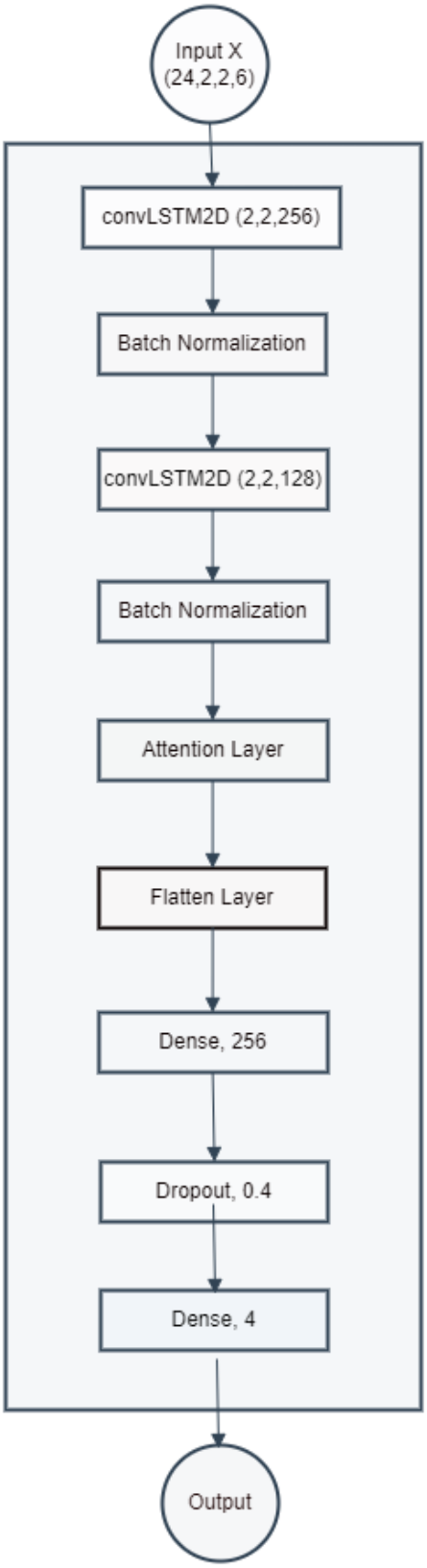


Figure 4.2: An overview of the used convLSTM with attention model.

**Correlation Coefficient:** The correlation coefficient measures the strength and direction of the linear relationship between the predicted and actual values. It ranges from -1 to 1, where 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no correlation.

**Normalized Root Mean Square Error (NRMSE):** NRMSE is a measure of the differences between predicted and actual values, normalized by the range of the observed data. It is calculated by taking the square root of the average squared differences between predictions and actual observations, divided by the range of the observed values. A lower NRMSE indicates better model performance, as it means the model's predictions are closer to the actual values relative to the variability of the data.

**Precision:** Precision, in the context of precipitation prediction, refers to the accuracy of the positive predictions. It is calculated as the ratio of true positive predictions (correctly predicted rainfall events) to the total number of positive predictions (true positives plus false positives). Higher precision means fewer false alarms.

**Recall:** Recall measures the model's ability to identify all relevant instances. In precipitation prediction, it is the ratio of true positive predictions to the total number of actual positive instances (true positives plus false negatives). Higher recall indicates the model is good at capturing all rainfall events, though it may include some false positives.

**Specificity:** Specificity, also known as the true negative rate, is a metric used in classification problems to measure the proportion of actual negatives that are correctly identified as such by the model. In other words, it tells us how well the model is at identifying negative cases.

### 4.4.1 Using Random Forest Regressor:

We used 100 estimators, which corresponds to the number of trees in the forest. This number was chosen to balance computational efficiency and model performance.

Followings are the results obtained from training set using Random Forest Regressor:

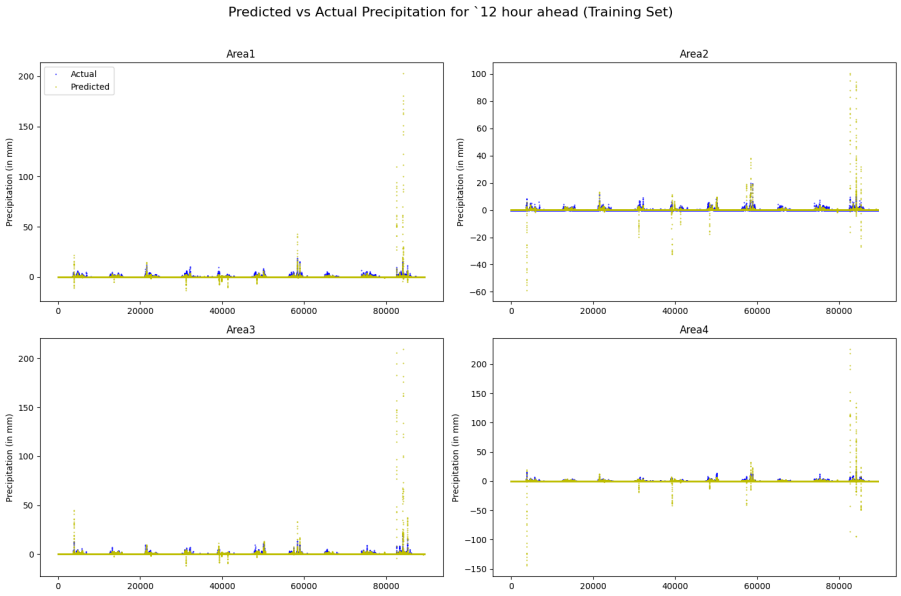


Figure 4.3: Actual vs Predicted rainfall using Random Forest regressor.

Followings are the results obtained from testing set using Random Forest Regressor:

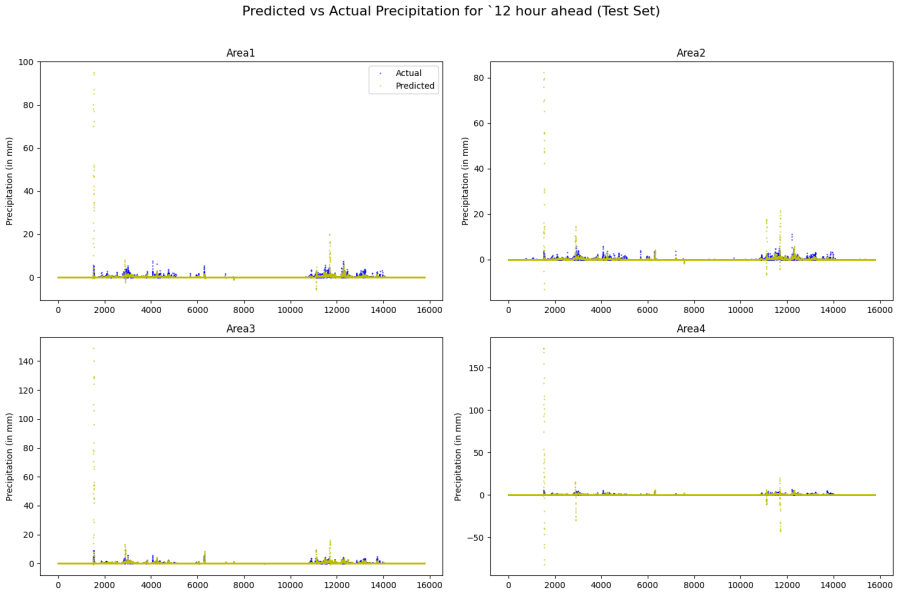


Figure 4.4: Actual vs Predicted rainfall using Random Forest regressor.

### 4.4.2 Using vanila stacked LSTM model:

In this section, we present the experimental results of vanila stacked LSTM model designed to predict rainfall 12 hours into the future.

Followings are the results obtained from training set using LSTM model:

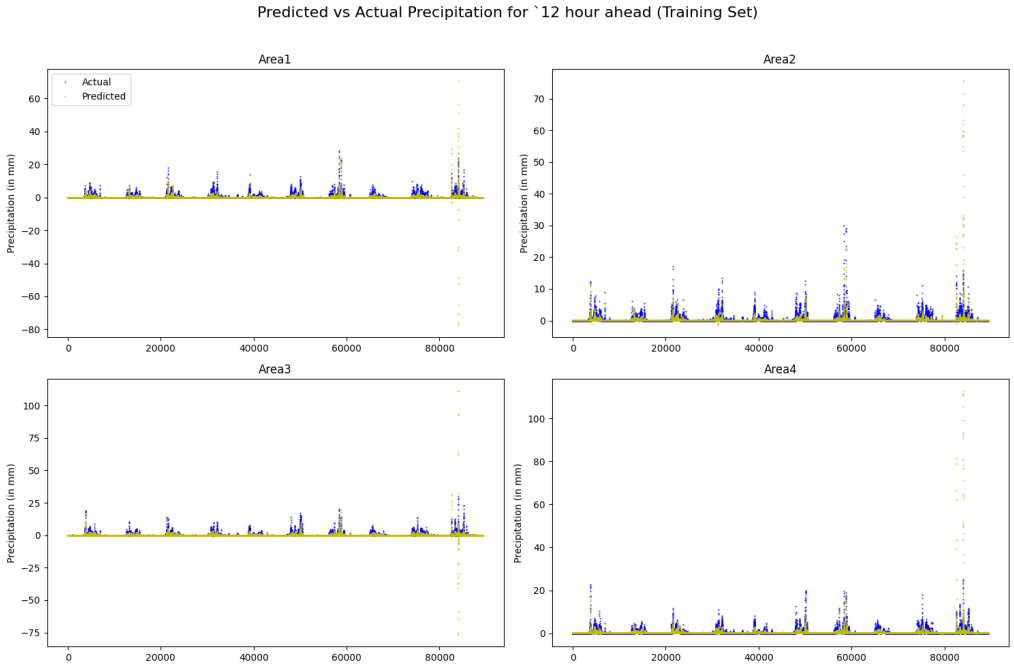


Figure 4.5: Actual vs Predicted rainfall using LSTM model.

Followings are the results obtained from testing set using LSTM model:

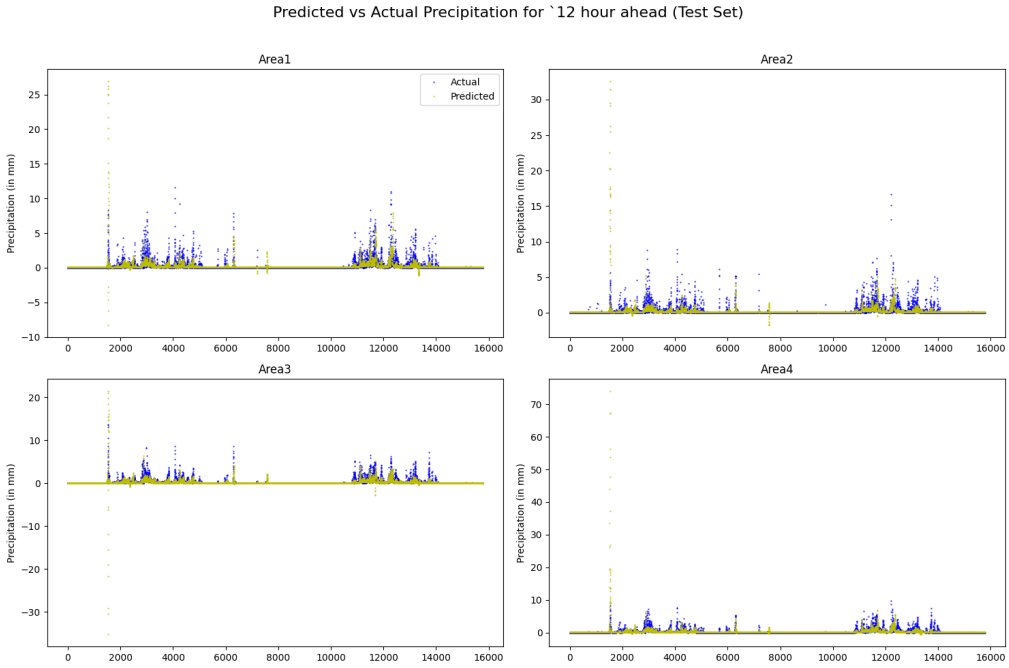


Figure 4.6: Actual vs Predicted rainfall using LSTM model.

### 4.4.3 Using ConvLSTM model:

In this section, we present the experimental results of our ConvLSTM model designed to predict rainfall using 24-hour time steps. The model was trained and evaluated on a dataset derived from ERA-5 reanalysis data, which includes a variety of meteorological variables. Sequences of 24-hour time steps were used to forecast rainfall 12 hours into the future.

Followings are the results obtained from training set using convLSTM model:

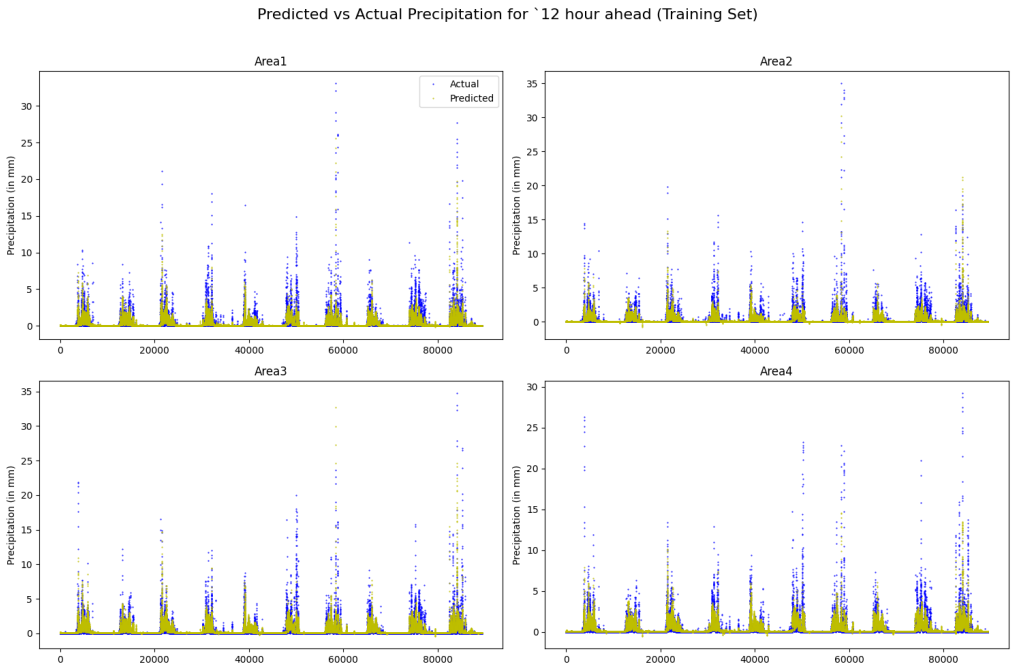


Figure 4.7: Actual vs Predicted rainfall using convLSTM model.

Followings are the results obtained from testing set using convLSTM model:

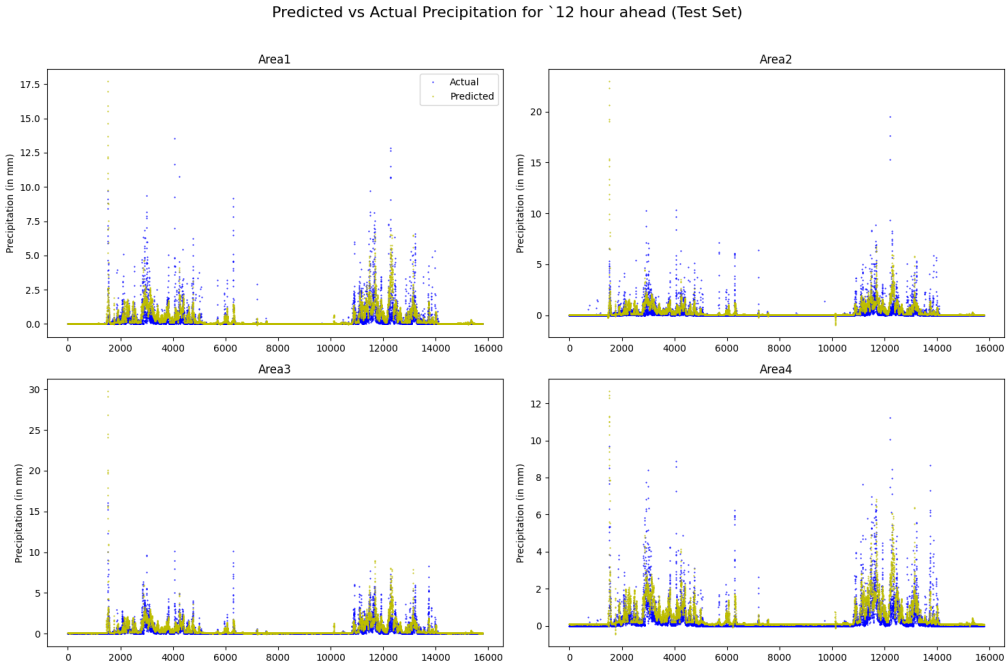


Figure 4.8: Actual vs Predicted rainfall using convLSTM model.

### 4.4.4 Using ConvLSTM model with Attention:

In this section, we present the experimental results of our enhanced ConvLSTM model, which includes an additional attention layer to predict rainfall using 24-hour time steps.The inclusion of an attention layer in our model allows it to focus on the most relevant parts of the input sequences, thereby improving its predictive capabilities.

Followings are the results obtained from training set using convLSTM model with attention:

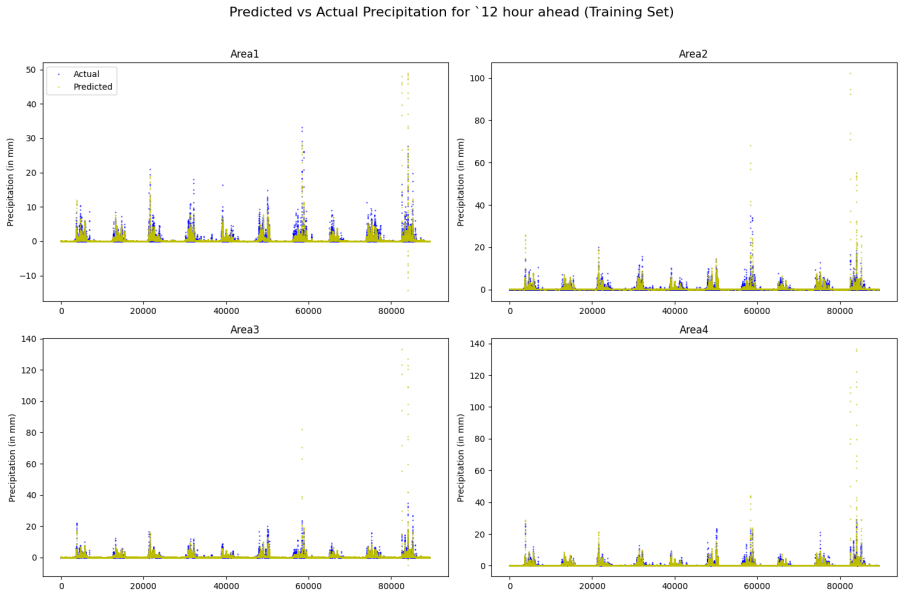


Figure 4.9: Actual vs Predicted rainfall using convLSTM with attention model.

Followings are the results obtained from testing set using convLSTM model with attention:

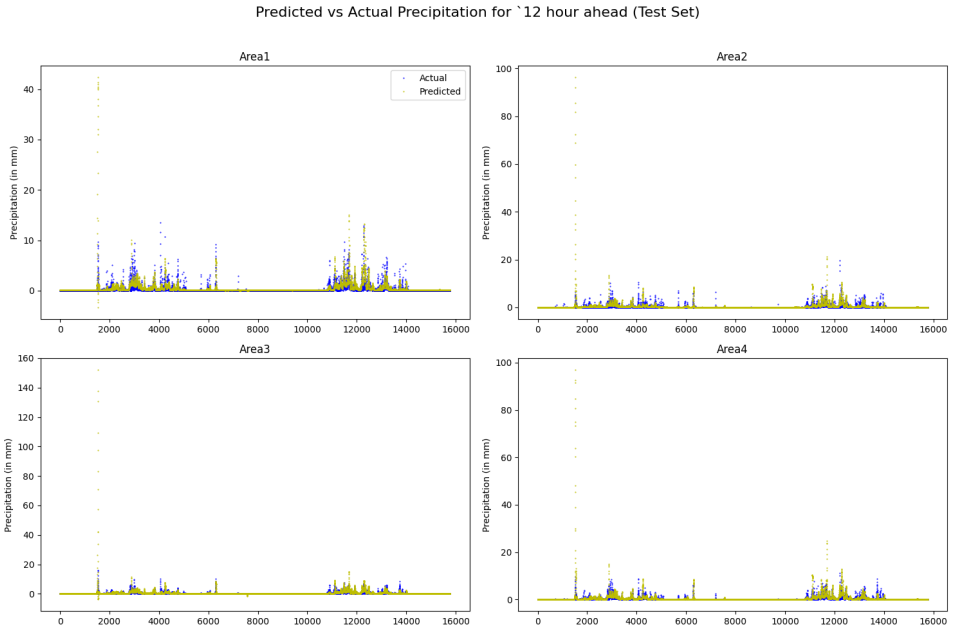


Figure 4.10: Actual vs Predicted rainfall using convLSTM with attention model.

## 4.5 Thresholding for Extreme Rainfall Classification

The model is predicting rainfall values at 12-hour intervals, not hourly. This means that it forecasts the amount of rainfall expected to occur over a specific area every 12 hours. So, the model provides predictions for two distinct time periods within a day, typically covering daytime and nighttime intervals.

Therefore, when determining the threshold for classifying an event as extreme, we should consider the cumulative rainfall amount predicted for each 12-hour interval rather than hourly predictions.

The following table depicted the rainfall intensity classifiaction provided by IMD 2015:

Table 4.1: Rainfall intensity classification (IMD 2015) [2].

| **No.** | **Terminology** | **Rainfall range (mm)** |
|---|---|---|
| 01 | Very light rainfall | Trace–2.4 |
| 02 | Light rainfall | 2.5–15.5 |
| 03 | Moderate rainfall | 15.6–64.4 |
| 04 | Heavy rainfall | 64.5–115.5 |
| 05 | Very heavy rainfall | 115.6–204.4 |
| 06 | Extremely heavy rainfall | $\geq$204.5 |

Based on the observed data, the city experiences extreme rainfall events (ERF) of 204.5 mm per day approximately once every other year during the summer monsoon season. To classify an event as extreme rainfall, I have set a threshold value of 180 mm per day. This threshold is used to identify extreme rainfall events if the predicted rainfall amount exceeds 180 mm within a 24-hour period. Additionally, since our model provides predictions at 12-hour intervals rather than on an hourly basis, the threshold value for each 12-hour interval is set to 90 mm. This ensures consistency and accuracy in classifying extreme events based on the 12-hour predictions [13].

The following table depicted some extreme rainfall days in Mumbai:

Table 4.2: Extreme rainfall days.

| **No.** | **Extreme rainfall days** | **Rainfall amount (mm)** |
|---|---|---|
| 08 | 13 Jul 2000 | 351.5 |
| 09 | 26 Jul 2005 | 944.4 |
| 10 | 24 Jun 2007 | 279.4 |
| 11 | 15 Jul 2009 | 274.1 |
| 12 | 19 Jun 2015 | 283.4 |
| 13 | 29 Aug 2017 | 331 |
| 14 | 20 Sep 2017 | 303.7 |
| 15 | 1 Jul 2019 | 375 |

For classification using your threshold of 180 mm per day (or 90 mm per 12-hour interval), these can be identified as follows:

- **TP:** Predicted rainfall $\geq$ 180 mm (or 90 mm per 12-hour interval) and actual rainfall $\geq$ 180 mm (or 90 mm per 12-hour interval).
- **TN:** Predicted rainfall $<$ 180 mm (or 90 mm per 12-hour interval) and actual rainfall $<$ 180 mm (or 90 mm per 12-hour interval).
- **FP:** Predicted rainfall $\geq$ 180 mm (or 90 mm per 12-hour interval) but actual rainfall $<$ 180 mm (or 90 mm per 12-hour interval).
- **FN:** Predicted rainfall $<$ 180 mm (or 90 mm per 12-hour interval) but actual rainfall $\geq$ 180 mm (or 90 mm per 12-hour interval).

### 4.5.1 Comparison of results

Table 4.3: Results using Random Forest Regressor

| | CC | NRMSE |
|---|---|---|
| Area 1(Train) | 0.56 | 0.0231 |
| Area 2(Train) | 0.58 | 0.0200 |
| Area 3(Train) | 0.57 | 0.0228 |
| Area 4(Train) | 0.60 | 0.0274 |
| Area 1(Test) | 0.52 | 0.0370 |
| Area 2(Test) | 0.49 | 0.0325 |
| Area 3(Test) | 0.56 | 0.0301 |
| Area 4(Test) | 0.53 | 0.0511 |

Table 4.4: Results using vanilla stacked LSTM model

| | CC | NRMSE |
|---|---|---|
| Area 1(Train) | 0.73 | 0.0193 |
| Area 2(Train) | 0.70 | 0.0202 |
| Area 3(Train) | 0.73 | 0.0281 |
| Area 4(Train) | 0.72 | 0.0266 |
| Area 1(Test) | 0.67 | 0.0378 |
| Area 2(Test) | 0.62 | 0.0312 |
| Area 3(Test) | 0.69 | 0.0332 |
| Area 4(Test) | 0.65 | 0.0372 |

Table 4.5: Results using convLSTM model

| | CC | NRMSE |
|---|---|---|
| Area 1(Train) | 0.72 | 0.0202 |
| Area 2(Train) | 0.71 | 0.0223 |
| Area 3(Train) | 0.76 | 0.0278 |
| Area 4(Train) | 0.74 | 0.0251 |
| Area 1(Test) | 0.65 | 0.0391 |
| Area 2(Test) | 0.64 | 0.0301 |
| Area 3(Test) | 0.71 | 0.0311 |
| Area 4(Test) | 0.68 | 0.0361 |

Table 4.6: Results using convLSTM with attention

| | CC | NRMSE |
|---|---|---|
| Area 1(Train) | 0.75 | 0.0137 |
| Area 2(Train) | 0.74 | 0.0213 |
| Area 3(Train) | 0.78 | 0.0235 |
| Area 4(Train) | 0.79 | 0.0240 |
| Area 1(Test) | 0.67 | 0.0344 |
| Area 2(Test) | 0.69 | 0.0312 |
| Area 3(Test) | 0.75 | 0.0337 |
| Area 4(Test) | 0.71 | 0.0381 |

The following table describes the precision,recall and specificity values obtained from the classification task of extreme rainfall events based on predictions from different models:

Table 4.7: Comparison of Precision, Recall and Specificity

| | Precision | Recall | Specificity |
|---|---|---|---|
| Random Forest Regressor | 0.722 | 0.713 | 0.701 |
| vanilla stacked LSTM | 0.877 | 0.881 | 0.823 |
| convLSTM | 0.873 | 0.891 | 0.843 |
| convLSTM with attention | 0.893 | 0.914 | 0.883 |

# Chapter 5

# Conclusions and Future Work

In this study, we compared the performance of four different models for predicting extreme rainfall events: the random forest regressor, vanilla stacked LSTM, ConvLSTM without attention and ConvLSTM with attention. Our results clearly demonstrate that the ConvLSTM model with attention mechanisms outperforms the other three models in terms of accuracy and predictive capability. The inclusion of attention mechanisms allows the model to focus on the most relevant parts of the input data, leading to improved performance in capturing complex temporal and spatial dependencies inherent in rainfall prediction. This enhanced performance underscores the importance of incorporating attention mechanisms in deep learning models for more accurate and reliable weather forecasting.

Moving forward, there are several ways to improve our model. One approach is to incorporate enhanced attention layers, which could help the model focus even more effectively on the most important parts of the data, potentially leading to better results.

Additionally, we could analyze the main atmospheric factors that cause extreme rainfall and use this information to improve our predictions. By identifying and incorporating these key factors into our model, we can make our rainfall predictions more accurate and reliable.

Another area for improvement is the integration of real-time data, such as live satellite and radar images, which could enhance the model's ability to predict sudden changes in weather patterns. We could also explore the use of more advanced machine learning techniques, such as hybrid models that combine different approaches, to leverage the strengths of multiple methods.

# Bibliography


[1] "orographic precipitation,"

[2] IMD, "India meteorological department doc.," 2015.

[3] Kumar, Amit, and Asthana, "Assesment and review of hydrometeorological aspects for cloudburst and flash flood events in the third pole region (indian himalaya)," 2018.

[4] Sherstinsky and Alex, ""fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network," in *Physica D: Nonlinear Phenomena 404*, p. 132306, 2020.

[5] Tiwari, Arpit, Verma, and S. K, "Cloudburst predetermination system," 2015.

[6] Dimri, A.P., and Thayyen, "A review of atmospheric and land surface processes with emphasis on flood generation in the southern himalayan rivers," in *Science of The Total Environment 556*, p. 98–115, 2016.

[7] Goswami and Bikramjit, "Prediction of cloudburst using passive microwave remote sensing," 2017.

[8] Pabreja and Kavita, "Clustering technique for interpretation of cloudburst over uttarakhand.," 2016.

[9] S. Gope, S. Sarkar, P. Mitra, and S. Ghosh, "Early prediction of extreme rainfall events: A deep learning approach," 2016.

[10] Sivagami, Radha, Balasundaram, and Ananthakrishnan, "Sequence model based cloudburst prediction for the indian state of uttarakhand," pp. 1–9, 2021.

[11] Shi, Xingjian, Chen, Zhourong, Wong, and Wangchun, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," vol. arXiv: 1506.04214, 2015.

[12] Hersbach and H. Bell, "The era5 global reanalysis.," *Quarterly Journal of the Royal Meteorological Society*, vol. abs/2110.04596, 2020.

[13] Mohanty, Swain, Nadimpalli, and Osuri, "Meteorological conditions of extreme heavy rains over coastal city mumbai. journal of applied meteorology and climatology," pp. 191–208, 2023.

[14] Ali, H., Mishra, V., Pai, and D. S., "Observed and projected urban extreme rainfall events in india," in *Journal of Geophysical Research: Atmospheres,*, p. 119(22), 2014.

[15] Sønderby, C. K., Espeholt, L. Heek, Dehghani, and Oliver, "Metnet: A neural weather model for precipitation forecasting.," 2020.

[16] Espeholt, L., Agrawal, Sønderby, and Kumar, "Deep learning for twelve hour precipitation forecasts. nature communications," 2022.

[17] Rawat, K. Singh, Sahu, S. Ranjan, Singha, S. Kumar, Mishra, and A. Kumar, "Cloudburst analysis in the nainital district, himalayan region," 2022.

[18] Sinha and Amitabh, "Explained: What are cloudburst incidents and are they rising across india?," 2022.

[19] Staudemeyer, R. C, Morris, and E. Rothstein, "Understanding lstm – a tutorial into long short-term memory recurrent neural networks," *arXiv: 1909.09586 [cs.NE]*, 2019.

[20] Vakili, Meysam, Ghamsari, Mohammad, Rezaei, and Masoumeh, "Performance analysis and comparison of machine and deep learning algorithms for iot data classification.," 2020.

[21] Sherstinsky and Alex, "Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network," 2020.

[22] "Observed and projected urban extreme rainfall events in india," *Journal of Geophysical Research: Atmospheres*, 2014.

[23] Bi, Xie, L. Zhang, Chen, and Tian, "Accurate medium-range global weather forecasting with 3d neural networks," *Nature*, 2023.

[24] V. Bajpaia, A. Bansala, and S. A. Kshitiz Vermab, "Prediction of rainfall in rajasthan, india using deep and wide neural networks," 2020.