District-Level Rainfall and Cloudburst Prediction Using XGBoost: A Machine Learning Approach for Early Warning Systems

Guru Dayal Kumar[1], Shekhar Tyagi[2], Kalandi Charan Pradhan[1], Akshat Shah[2,3]
[1]Migration and Development Research Group, School of Humanities and Social Sciences, Indian Institute of Technology Indore, 452020, India
[2]Department of Computer Science and Engineering, Indian Institute of Technology Indore, 452020, India
[3]Department of Computer Science and Engineering, Shri Ramdeobaba College of Engineering and Management, Nagpur, 440001, India
E-mail: shekhartyagirocks007@gmail.com

Keywords: Cloudbursts, rainfall, XGBoost, flood

Received: November 16, 2024

This research presents a novel methodology for cloudburst forecasting using the XGBoost (Extreme Gradient Boosting) machine learning algorithm. With the escalating impact of climate change, accurately predicting extreme weather events like cloudbursts is crucial due to their potential to trigger floods. Cloudburst events were identified from daily rainfall data. Our study leverages historical weather data, focusing on intricate rainfall patterns, to forecast future cloudburst occurrences. Comparative analysis against Random Forest and LSTM models confirmed XGBoost's effectiveness, consistently outperforming alternatives across multiple performance metrics. The XGBoost model, known for its ability to handle complex datasets, demonstrated strong predictive performance, with an RMSE of 0.12 and an MAE of 0.09, indicating high accuracy. This research provides a reliable tool for advanced weather forecasting and early warning systems, offering valuable support to policymakers, disaster management teams, and agricultural planners in mitigating risks associated with extreme rainfall events.

Povzetek: Raziskava uvaja model XGBoosta za napoved padavin in lokalnih nalivov na ravni okrožij, ki omogoča učinkovite opozorilne sisteme za ekstremne vremenske razmere.

1 Introduction

Excessive rainfall phenomena, such as cloudbursts, typically occur on a mesoscale level, spanning 2–20 km [50]. These events are marked by sudden and intense precipitation surges, often resulting in secondary hazards like flash floods, landslides, and dam failures [15, 13]. Predominantly occurring during the monsoon season, cloudbursts are among the most significant natural hazards in the region. Various studies have proposed rainfall intensity thresholds to identify cloudburst events [28, 48, 26]. For example, [27] defined extreme rainfall events as those exceeding 250 mm/day, while [48] categorized heavy rainfall in the northwest Himalayas (NWH) as surpassing 200 mm/day. Other studies, such as [7] and [59], utilized the 99.99th percentile of precipitation distribution to delineate cloudbursts. The study [62] proposes an innovative approach for missing value imputation using an extended Kalman filter with linear relations and introduces advanced bidirectional and unidirectional LSTM architectures for enhancing network-wide rainfall forecasting in ubiquitous computing environments. The pie chart illustrates the distribution of various natural disasters. The pie chart(Figure 1) illustrates the distribution of various natural disasters. Floods account for the highest proportion (50%), followed by storms (32%) and extreme temperatures (10%), while glacial lake outburst floods and droughts have the lowest occurrences at 1% and 3%, respectively.

These extreme weather events often result in significant loss of life and property. A notable instance was the Kedarnath tragedy in Uttarakhand in June 2013, where rainfall intensities exceeding 200 mm/day over several days caused over 6,000 fatalities [53]. Approximately 50 million people inhabit the Himalayan region across Nepal, India, Bhutan, China, and Pakistan, making the prediction of cloudbursts crucial for safeguarding these vulnerable populations. However, existing meteorological models struggle to achieve accurate cloudburst prediction due to their reliance on deterministic weather forecasting, which often fails to capture the complex, non-linear relationships between atmospheric variables [13, 10]. Traditional models, such as numerical weather prediction (NWP) techniques, suffer from limitations in spatial resolution, dependency on large-scale climate patterns, and computational inefficiencies, making them inadequate for real-time early warning systems.

Recent examples highlight the devastating impact of such events. In 2022, Bengaluru experienced severe flooding, with the city recording 132 mm of rainfall within 24 hours on September 5, accounting for 10% of its seasonal

rainfall. The floods caused an estimated loss of over Indian rupees 2,250 million [22]. Similarly, Bihar has faced recurring floods with varying intensity, causing widespread damage over the years, including major episodes in 1987, 1998, 2000, 2001, 2003, 2004, 2008, 2010, 2013, 2017, 2018, and 2020 [63]. States such as Bihar, West Bengal, and Uttar Pradesh, situated along the Ganges River, are especially susceptible to natural disasters, with climatic risks exacerbating the trends of loss and destruction [55]. In Bihar alone, approximately 68.20 million people—roughly 53.20% of its population—reside in high-risk flood zones [55]. The state's vulnerability is further underscored by the severe flooding of 6.970 million hectares of land, affecting 74.0% of its geographical area [4]. To address these challenges, this study leverages machine learning, specifically the XGBoost algorithm, to enhance cloudburst prediction. Unlike traditional meteorological models, XGBoost is capable of handling high-dimensional data, capturing intricate relationships among multiple atmospheric parameters, and reducing prediction error through its boosting mechanism. The adaptive learning approach of XGBoost makes it particularly well-suited for cloudburst forecasting, allowing for real-time predictions with higher accuracy. This research aims to demonstrate how XGBoost outperforms conventional models in forecasting cloudbursts, offering a more robust and data-driven early warning system for disaster mitigation.

2 Related studies on cloudbursts and early warning system

The term cloudburst holds a notable historical presence in meteorological literature, with references dating back to the 19th century [43]. A comparative analysis using the Google Books Ngram Viewer reveals that the term emerged in the 1800s and reached its peak usage in the 1940s. Initially, cloudbursts were described as localized heavy rainfall events often linked to thunderstorms, though their formal definition evolved over time. For instance, Elmer [20] suggested that elongated thunderstorm clouds moving along their longitudinal axis could directly trigger cloudbursts. Similarly, Bonnett [8] described scenarios where showers intensified progressively, eventually covering the sky and culminating in severe thunderstorms. Horton and Todd [30] emphasized the highly localized nature of these events, citing the Taborton, New York, incident where 158 mm of rain fell in two hours over an 8 km-wide area.

King [35] reported a cloudburst lasting 3.5 hours, producing 305 mm of rain across an elliptical region of 80 square kilometers, causing significant destruction, including impassable roads, swept-away bridges and homes, debris-laden farmland, 11 fatalities, and property damage equivalent to USD 6.8 million today. Douglas [17] documented a California cloudburst that led to a flash flood accompanied by a dust cloud from dislodged dry soil rushing down a canyon. Similarly, the July 2, 1893, cloudburst over the Cheviot Hills (UK) caused erosion of up to 2 square meters of valley cross-section, destroying bridges and roads [11].

By the mid-20th century, cloudbursts were widely understood as localized, high-intensity rainfall events spanning a few kilometers, often accompanied by thunder and lightning [64, 47]. These events could result in extensive damage, including flash floods, streambed erosion, landslides, and mudflows. Woolley [65] provided a formal definition, describing cloudbursts as torrential rainfall events characterized by intensity and spatial concentration, akin to the sudden release of an entire cloud. Typically associated with thunderstorms, these events occur over limited areas and produce volatile, damaging floods in steep catchments. They are also linked to cumulonimbus clouds and hazardous phenomena such as squalls, strong winds, hailstorms, and tornadoes [14, 18].

Quantitative definitions of cloudbursts vary. Haritashya et al. [29] proposed a threshold of 100 mm/h to classify a violent shower as a cloudburst. Krishnamurthy [37] used 100 mm/day as a criterion for extreme rainfall events, while Izzo [34] defined cloudbursts as having rainfall intensity above 30 mm/h. Dunlop [19] differentiated heavy showers (10–50 mm/h) from violent showers (> 50 mm/h), following World Meteorological Organization (WMO) guidelines. Fry et al. [23] defined downpours as rainfall exceeding 15 mm/h. The American Meteorological Society's Glossary of Meteorology supports the threshold of 100 mm/h for defining cloudbursts [1]. Consequently, in non-monsoonal regions, cloudbursts are defined as rainfall events with intensities exceeding 30 mm/h, while in monsoonal regions, the threshold is 100 mm/h. These events typically affect areas of a few kilometers, often causing flash floods, landslides, and mudflows, and are frequently accompanied by storms, strong winds, hail, and tornadoes.

The Centre for Research on the Epidemiology of Disasters (CRED) has developed the Emergency Events Database (EM-DAT), recognized as the most comprehensive global database on over 23,000 natural and technological disasters from 1900 to 2022 [2]. EM-DAT systematically records disaster data annually. Analysis of the database highlights the most impactful natural disasters, including droughts, earthquakes, extreme temperatures, floods, glacial lake outbursts, storms, and wildfires. Notably, more than 50 percent of these recorded events are floods and glacial lake outburst floods [17, 18, 11, 15].

Both fluvial and pluvial flooding are expected to increase the vulnerability of residents in riparian and informal communities due to projected rises in rainfall intensity driven by climate change [46]. Early Warning Systems (EWSs) serve as a critical intersection of disaster risk reduction, effective humanitarian response, and the promotion of climate-resilient and sustainable development. They address present, emerging, and potential flood-related hazards. However, Africa lags behind other global regions in implementing robust EWSs [44, 66]. Several studies have highlighted the pivotal role of EWSs in disaster prepared-

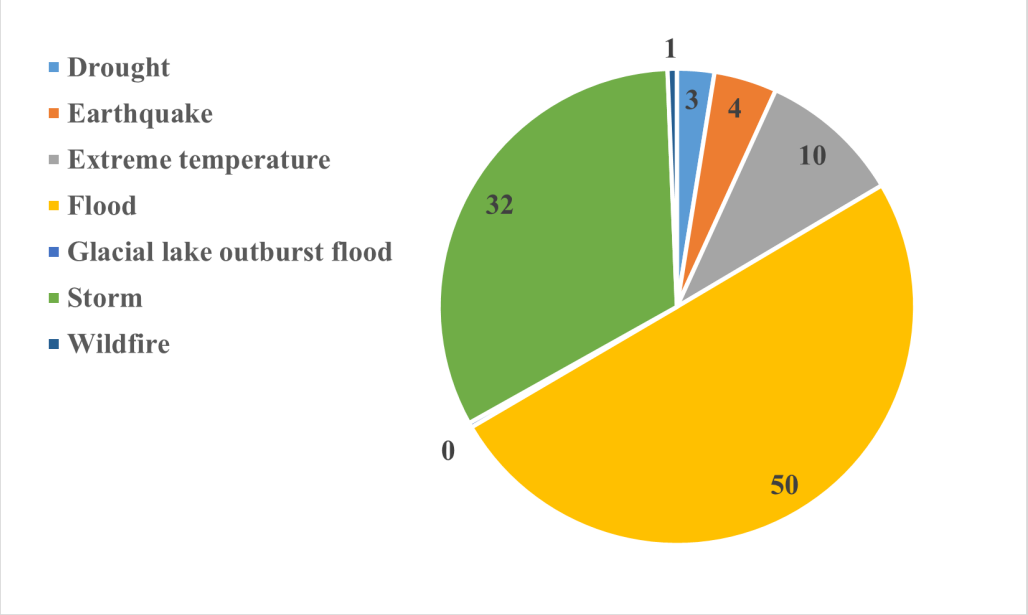

Figure 1: Sum of occurrences of natural disasters from 1900 to 2022

ness and mitigation [12, 21, 25]. Research efforts, such as those by Roy et al. [56] and Shahjahan [57], have assessed the efficacy of EWSs in Bangladesh's Sundarbans region, focusing on methods of warning dissemination and reception. In contrast, similar systems have not been extensively studied in the Indian state of Bihar. Thus, it is crucial to evaluate the effectiveness of existing EWSs in bolstering disaster mitigation efforts in Eastern India, particularly Bihar [39].

Climate change has led to a significant increase in the frequency and severity of extreme weather events, with cloudbursts emerging as critical hazards impacting ecosystems, agriculture, and human settlements [49]. Accurate forecasting of such events is essential for proactive planning and mitigation strategies. While conventional weather prediction methods provide some insights, they often fall short in predicting the intensity and occurrence of extreme events like cloudbursts. This study explores the use of XGBoost [9], an advanced machine learning algorithm, to bridge this gap. By analyzing historical weather data, the research aims to develop a predictive model with enhanced accuracy and reliability for cloudburst forecasting.

The utility of XGBoost in cloudburst prediction is underscored by recent advancements in meteorological research. Two prominent studies illustrate XGBoost's capabilities in this domain. The first study highlights its effectiveness in short-term precipitation forecasting across diverse climatic regions of China, employing multi-factor bias correction to improve accuracy [16]. The second study demonstrates XGBoost's proficiency in nowcasting weather conditions, outperforming traditional methods such as SVM (Support Vector Machine), RF (Random Forest), and GBDT (Gradient Boosting Decision Tree) [45].

In hydrological forecasting, machine learning algorithms, particularly XGBoost, have shown remarkable success in predicting groundwater levels and streamflow across various geographical terrains. Studies have demonstrated XGBoost's superiority in groundwater level predictions [51] and streamflow forecasting [61]. Furthermore, Kumar et al. [40] employed LSTM neural networks for rainfall forecasting and flood impact predictions in Bihar, leveraging deep learning to improve disaster preparedness and response. Another study by Kumar et al. [41] applied AI-driven models for assessing rainfall and flood vulnerability in Bihar, aiming to enhance disaster management strategies in the region. Table 1 presents a comprehensive summary of existing studies on cloudburst prediction and early warning systems, highlighting the methodologies used and identifying current research gaps. These findings collectively affirm the critical role of XGBoost in cloudburst prediction, presenting promising opportunities for enhanced forecasting precision. The superiority of XGBoost over Random Forest, SVM, and LSTM in precipitation forecasting is highlighted in Table 2, emphasizing its advantages in accuracy, scalability, and handling complex data. This research primarily addresses the challenge of accurately predicting cloudbursts—extreme precipitation events occurring over short durations that pose significant risks. The unpredictable and severe nature of these phenomena makes them a central focus in climate science and meteorological research. By leveraging historical weather data, this study aims to develop predictive models that estimate the likelihood of future cloudburst occurrences, providing a valuable tool for preemptive and preparatory measures.

The study is organized as follows: the section on the study area and data is followed by a detailed description of the proposed methodology and computational framework for rainfall and cloudburst prediction using XGBoost. Next,

Table 1: Summary of studies related to cloudburst and early warning systems

Study	Focus Area	Key Findings
Haritashya et al. [29]	Quantitative definition of cloudbursts	Proposed a threshold of 100 mm/h to classify a violent shower as a cloudburst.
Krishnamurthy [37]	Extreme rainfall events	Used 100 mm/day as a criterion for classifying extreme rainfall events.
Izzo [34]	Rainfall intensity classification	Defined cloudbursts as rainfall events with intensity above 30 mm/h.
Dunlop [19]	Rainfall intensity classification	Differentiated between heavy showers (10–50 mm/h) and violent showers (> 50 mm/h).
Fry et al. [23]	Heavy rainfall categorization	Defined downpours as rainfall exceeding 15 mm/h.
Roy et al. [56]	Early warning systems in Bangladesh	Assessed the effectiveness of warning dissemination and reception methods in the Sundarbans region.
Shahjahan [57]	Disaster mitigation in Bangladesh	Evaluated EWS efficacy in disaster-prone areas.
Kumar et al. [39]	EWS in Bihar, India	Highlighted the need to assess and improve EWS efficacy in Bihar.
Dong et al. [16]	XGBoost for precipitation forecasting	Demonstrated XGBoost's effectiveness in short-term precipitation forecasting using multi-factor bias correction.
Mai et al. [45]	Weather nowcasting with ML models	Showed XGBoost outperforming SVM, RF, and GBDT in nowcasting weather conditions.
Osman et al. [51]	Groundwater level prediction	Validated XGBoost's superiority in predicting groundwater levels across varied terrains.
Szczepanek et al. [61]	Streamflow forecasting	Highlighted the accuracy of XGBoost in daily streamflow prediction.
Kumar et al. [40]	Rainfall and flood impact prediction in Bihar	Applied LSTM neural networks for accurate rainfall forecasting and flood impact assessments.
Kumar et al. [41]	AI-driven models for disaster management	Developed predictive models for assessing rainfall and flood vulnerability in Bihar.

the results and analysis are presented, and the final section concludes the study.

3 Study area and data

3.1 Profile of the study area

Bihar, located in the eastern part of India, is situated between the coordinates of 24°20'10" to 27°31'15" North latitude and 83°19'50" to 88°17'40" East longitude (Figure 2). This landlocked state shares its borders with West Bengal to the east, Uttar Pradesh to the west, and Jharkhand to the south, while the northern boundary is an international border with Nepal. Bihar's geography is marked by the Himalayan range to the north and the Chhotanagpur hills to the south, with rivers originating from these regions playing a crucial role in the state's ecological and economic framework.

The river channels in the northern plains of Bihar form one of the most dynamic fluvial systems in the world [52, 24]. The region is home to over 250 seasonal and permanent rivers and streams, which significantly contribute to recurrent flooding. Additionally, more than a dozen major rivers traverse the state, dividing it into seven distinct regional geo-cultural zones [42]. These rivers, along with their seasonal dynamics, shape the local economies and impact flood patterns across the state.

3.2 Data

Data was retrieved from the India-Water Resource Information System (IWRIS), which provides rainfall data at the state, district, station, and basin levels [6]. The study also incorporates the high spatial resolution (0.25° x 0.25°) long-period (1991-2022) daily gridded rainfall dataset provided by the India Meteorological Department (IMD).

This dataset offers very high spatial resolution daily gridded rainfall data (0.25° x 0.25°). For this study, we utilized district-wise daily data for Bihar, India, covering the reference period from 1991 to 2022. We considered flood-prone districts of Bihar [4]. Furthermore, we calculated annual rainfall, rainy season rainfall, and cloudburst events based on the literature [37, 31, 32].

Figure 3 illustrates the methodology for calculating these rainfall parameters. For instance, in flood-prone districts such as Araria for the year 1991, we first extracted the daily rainfall data and then calculated the rainy season rainfall by summing the daily data from June 1st to September 30th. The annual rainfall was obtained by summing all the daily rainfall records for the year. Cloudburst events were defined as daily rainfall events exceeding 100mm during the rainy season, as higher-intensity rainfall is more common

Table 2: Demonstrating XGBoost's superiority in rainfall and cloudburst forecasting over traditional machine learning models

Model	Key Strengths	Limitations	Why XGBoost is Superior
XGBoost	- Superior accuracy and precision in precipitation forecasting. - Efficient in handling large, high-dimensional datasets. - Built-in handling of missing data and regularization to avoid overfitting. - Fast training and scalability for real-time forecasting.	- Requires careful hyperparameter tuning.	- Outperforms others in terms of accuracy, scalability, and robustness. - Handles missing data and complex relationships more effectively.
Random Forest (RF)	- Good for general-purpose classification and regression tasks. - Robust to noise and overfitting.	- Less accurate for highly imbalanced or complex datasets like precipitation. - Struggles with identifying subtle patterns in time-series data.	- XGBoost provides better bias correction and identifies complex relationships better than RF.
Support Vector Machine (SVM)	- Effective for high-dimensional and linear data. - Strong mathematical foundation for classification tasks.	- Poor scalability for large datasets. - Limited ability to handle missing data and non-linear temporal relationships.	- XGBoost scales better, is computationally efficient, and handles missing data seamlessly.
Long Short-Term Memory (LSTM)	- Excellent for sequential and time-series data. - Captures long-term dependencies well.	- Requires large datasets for effective training. - Computationally expensive and prone to overfitting with limited data.	- XGBoost requires less data for training, is faster, and less prone to overfitting. - Superior for real-time, large-scale forecasting.

in flood-prone districts during this period [31, 32].

3.2.1 Features considered

The dataset includes the following features relevant to rainfall and cloudburst prediction:

- **Rainfall Indicators:** Annual Rainfall (AR) and Rainy Season Rainfall (RSR)
- **Meteorological Variables:** Temperature, humidity, and elevation
- **Cloudburst Events:** Binary classification (1 for cloudburst, 0 otherwise)

3.2.2 Data preprocessing

To ensure data consistency and reliability, the following preprocessing steps were applied:

- **Handling Missing Values:** Missing entries were imputed using interpolation techniques or removed based on data availability thresholds.
- **Feature Scaling:** Normalization was performed to ensure comparability across different meteorological parameters.
- **Outlier Detection:** Extreme values in rainfall and temperature data were filtered using interquartile range (IQR) analysis.
- **Feature Engineering:** Derived features such as cumulative seasonal rainfall and average seasonal temperature were added to enhance model performance.
- **Train-Test Split:** The dataset was divided into training (80%) and testing (20%) subsets for model evaluation.

3.2.3 Cloudburst event definition

A cloudburst is characterized by a sudden, intense rainfall event over a localized area, often resulting in flash floods. According to the India Meteorological Department (IMD, 2020)[33], a cloudburst is defined as a rainfall of 100 mm or more within an hour over a geographical region of approximately 20–30 square kilometers. Such events are common in mountainous regions due to orographic lifting and convective processes, though they can also occur in other regions under suitable meteorological conditions [31, 32].

However, considering the geographic context of Bihar—a predominantly plains region forming part of the Gangetic Plain with fertile alluvial soil—this study defines a cloudburst event as daily rainfall exceeding 100 mm during the

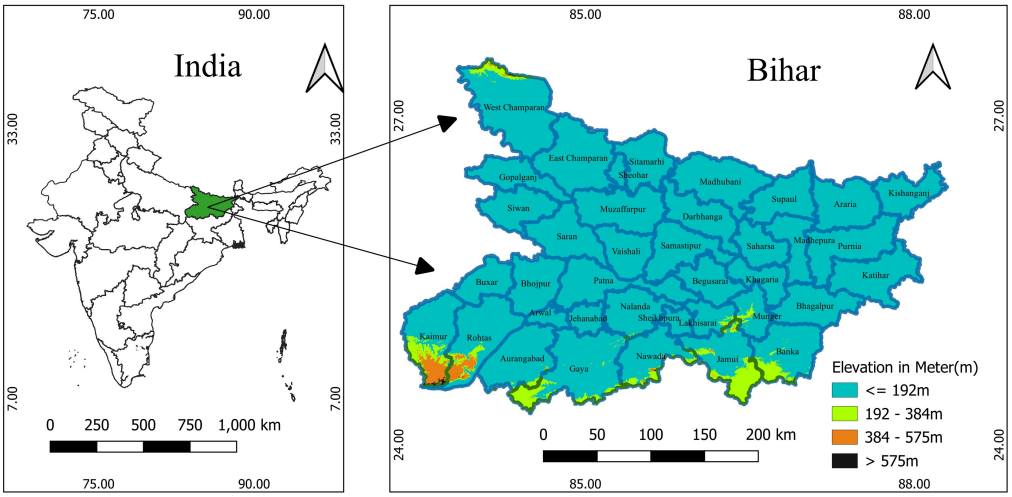

Figure 2: District-wise geographical location map of Bihar

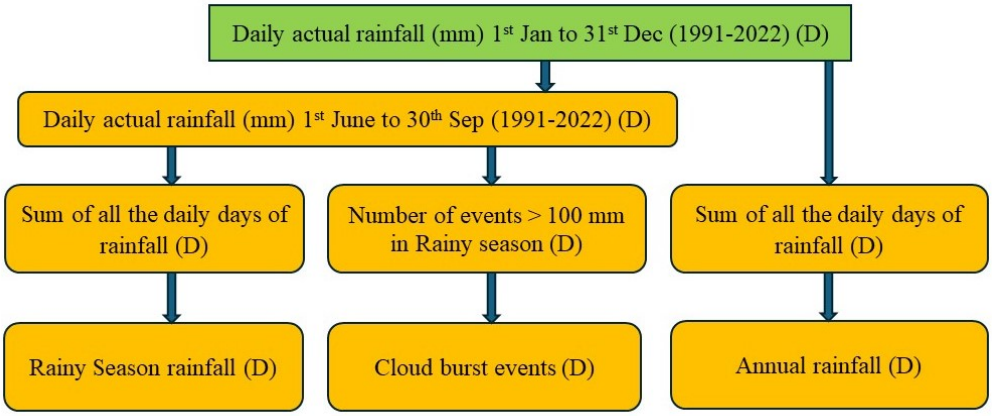

Figure 3: Methodology for calculating the annual rainfall, rainy season rainfall, and cloudburst events. D represents a district.

monsoon season. This criterion is adopted because high daily rainfall also leads to floods in the region[67, 58]. Prior studies [37] have validated the suitability of this threshold for extreme events in the Indian monsoon region and climate change, supporting its application in this research.

4 Computational framework for rainfall and cloudburst prediction with XGBoost

In this section, we present a computational framework utilizing XGBoost for forecasting rainfall and cloudburst events. Feature selection was performed using feature importance scores from XGBoost, retaining key meteorological parameters such as Annual Rainfall, Rainy Season Rainfall, temperature, humidity, elevation, and historical cloudburst events while discarding less significant features to enhance model efficiency. To assess its effectiveness, we compared XGBoost with Random Forest and Long Short-Term Memory (LSTM), chosen for their strengths in regression and sequential data modeling. Model performance was evaluated using *Mean Absolute Error (MAE)*, *Root Mean Square Error (RMSE)*, and *R-squared (R^2) Score*. XGBoost, a gradient boosting technique, employs decision tree ensembles, regularization to prevent overfitting, and parallel computation for efficiency. The dataset was split into 80% training and 20% testing subsets, with hyperparameter tuning performed via grid search and cross-validation. The trained models generated forecasts for 2023-2047, predicting *rainfall patterns and cloudburst probabilities*, with results analyzed for long-term trends in extreme weather events. Below are the input objectives along with their corresponding desired outputs.

4.1 Input

– Historical weather dataset containing:

- Rainy Season Rainfall (RSR) for previous years.
- Annual Rainfall (AR) for previous years.
- Additional relevant features (e.g., temperature, humidity, geographic factors).
- Information on cloudburst events (0 or 1) for previous years.

4.2 Output

- Predictions for the years 2023 to 2047 for:
  - Rainy Season Rainfall (RSR) for each district.
  - Annual Rainfall (AR) for each district.
  - Probability of cloudburst events (> 100mm rainfall) for each district.

4.3 Data preparation

4.3.1 Loading the historical weather dataset

We begin by importing the historical weather dataset, including RSR, AR, cloudburst events, and other relevant features such as temperature, humidity, and geographic factors.

4.3.2 Feature selection and preprocessing

- Identifying and extracting relevant features: RSR, AR, and cloudburst events.
- Cleaning and preprocessing: Handling missing values, outliers, and anomalies.
- Feature engineering: Creating new features such as cumulative rainfall values or seasonal averages.
- Splitting dataset into training and testing sets (e.g., 80/20 ratios).

4.4 Model initialization

We initialize separate XGBoost models for each prediction task:

- **Model for RSR Prediction:** Predicts Rainy Season Rainfall (RSR).
- **Model for AR Prediction:** Predicts Annual Rainfall (AR).
- **Model for Cloudburst Probability Prediction:** Estimates cloudburst events using binary classification:

$$P(\text{Cloudburst}) = \frac{1}{1 + e^{-\sum_{i=1}^{N} w_i x_i}} \quad (1)$$

where N represents the number of features, w_i denotes the corresponding weights, and x_i signifies the input features.

4.5 Model training and prediction

- Training each XGBoost model using the respective target variable.
- Using the trained models to make predictions for the years 2023-2047.
- Organizing predictions for analysis and visualization.
- Evaluating model performance based on historical data.

4.6 XGBoost computing

XGBoost minimizes a loss function by combining predictions from multiple weak learners (trees).

XGBoost Hyperparameters: The model's hyperparameters were optimized using grid search. Table 3 summarizes the final values.

Table 3: XGBoost hyperparameters used in the study

Hyperparameter	Value
Learning Rate (η)	0.1
Maximum Depth	6
Number of Boosting Rounds	100
Subsample Ratio	0.8
Column Subsample Ratio	0.8
Minimum Child Weight	1
Gamma (Minimum Loss Reduction)	0
Regularization (L1) α	0.1
Regularization (L2) λ	1

Data Split Rationale: An 80/20 train-test split was used to ensure a balanced division between model training and validation, aligning with standard machine learning practices.

Feature Importance: Feature importance ranking was conducted to identify key predictors influencing model outcomes, enhancing interpretability.

Handling Missing Data: Missing data were addressed using mean imputation for minor gaps, forward fill for time-series continuity, and model-based imputation for substantial missingness.

These methodological decisions contribute to the robustness and accuracy of the proposed model.

4.6.1 Objective function

The objective function to be optimized is:

$$\text{Objective} = L(y', y) + \gamma \cdot \Omega(f) + \frac{1}{2}\lambda \|w\|^2 \quad (2)$$

where $L(y', y)$ is the loss function measuring the difference between predicted (y') and actual (y) values, γ controls tree complexity, and λ controls L2 regularization on leaf weights.

4.6.2 Gradient and hessian of loss function

For squared error loss:

$$L(y', y) = (y' - y)^2 \tag{3}$$

Gradient:

$$\nabla L = 2(y' - y) \tag{4}$$

Hessian:

$$H = 2 \tag{5}$$

4.6.3 Tree building and regularization

XGBoost builds trees sequentially by fitting a new tree to the negative gradient of the loss function. Regularization terms control model complexity:

- γ: Minimum loss reduction required for further partitioning.
- λ: L2 regularization on leaf weights.

4.6.4 Learning rate (shrinkage)

XGBoost introduces a learning rate (η) to control step size:

$$y' = \sum_{k=1}^{K} f_k(x) \tag{6}$$

where $f_k(x)$ is the prediction of the k-th tree.

These mathematical foundations enable XGBoost to iteratively optimize and construct a robust ensemble model, providing accurate predictions for rainfall and cloudburst events.

4.7 Model evaluation: XGBoost

Model evaluation is a critical step to gauge the performance of predictive models. In the context of XGBoost [54], commonly used metrics such as Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) [38] offer valuable insights into the accuracy and precision of the model predictions.

4.7.1 Procedure

1. **Prepare the Testing Data:** After training the XGBoost models on the training dataset, we applied the models to predict the target variables (e.g., Rainy Season Rainfall and Annual Rainfall) on the testing dataset.
2. **Compute Predictions:** Utilized the trained XGBoost models to predict the target variables for the testing set.
3. **Calculate Residuals:** Computed the residuals by subtracting the actual values from the predicted values, representing the errors made by the model for each prediction.
4. **Compute RMSE:** RMSE is calculated as the root mean square of the residuals, providing a measure of the average magnitude of errors:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2} \tag{7}$$

where n is the number of observations, y_i is the actual value, and $\widehat{y_i}$ is the predicted value.

5. **Compute MAE:** MAE is calculated as the average of the absolute residuals, providing another measure of the model's predictive accuracy:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y_i}| \tag{8}$$

This systematic procedure allows for a comprehensive assessment of the XGBoost model's effectiveness in predicting rainfall and facilitates comparisons with traditional forecasting models.

4.8 Scientific and methodological enhancements for model robustness

To enhance the robustness and scientific integrity of the proposed XGBoost-based rainfall and cloudburst forecasting model, we incorporated multiple methodological refinements, detailed as follows:

4.8.1 Uncertainty quantification

To quantify the uncertainty associated with model predictions, we employed the bootstrap resampling technique. Let $\hat{y}_i$ represent the predicted value for the i^{th} observation. We generated B bootstrap samples $\{D_b\}_{b=1}^{B}$, where each D_b is a random sample with replacement from the original dataset. For each bootstrap sample, the model produced a prediction $\hat{y}_i^{(b)}$. The confidence interval (CI) for prediction was then computed as:

$$\text{CI}_{95\%} = \left[\hat{y}_i^{(B)} - 1.96 \cdot \sigma, \hat{y}_i^{(B)} + 1.96 \cdot \sigma\right] \tag{9}$$

where $\hat{y}_i^{(B)}$ is the mean of bootstrap predictions and σ is the standard deviation of $\hat{y}_i^{(b)}$.

Additionally, a sensitivity analysis was performed by varying key hyperparameters $\theta \in \{\eta, \lambda, \alpha\}$ (learning rate, L2 regularization, L1 regularization, respectively) within specified intervals. The impact on the Root Mean Square Error (RMSE) was assessed as:

$$\Delta\text{RMSE} = \frac{\partial\text{RMSE}}{\partial\theta} \cdot \Delta\theta \tag{10}$$

ensuring the model's stability and robustness across different hyperparameter configurations.

4.8.2 Explainability via SHAP values

Given XGBoost's black-box nature, we applied SHapley Additive exPlanations (SHAP) to interpret feature contributions. For a feature set F and feature i, the SHAP value ϕ_i was calculated as:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f(S \cup \{i\}) - f(S)] \tag{11}$$

where S is a subset of features, and $f(S)$ is the model prediction based on subset S.

4.8.3 Overfitting mitigation strategies

To mitigate overfitting, we employed k-fold cross-validation ($k = 5$), where the dataset D was partitioned into k equal folds $\{D_1, D_2, \ldots, D_k\}$. The model was iteratively trained on $k-1$ folds and validated on the remaining fold. The cross-validation error ϵ_{cv} was calculated as:

$$\epsilon_{cv} = \frac{1}{k} \sum_{i=1}^{k} \text{RMSE}(D_i) \tag{12}$$

ensuring robustness across different data splits. Early stopping was implemented by monitoring the validation error, terminating training if no improvement was observed after $T = 10$ iterations.

Regularization parameters (λ, α) were optimized to penalize model complexity, ensuring minimized overfitting risk through the objective function:

$$\mathcal{L} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda ||w||_2^2 + \alpha ||w||_1 \tag{13}$$

where w denotes the weight vector.

4.8.4 Comparative analysis with physically-based models

Traditional physically-based models, denoted as $\mathcal{M}_{phys}$, rely on differential equations (such as, derived from hydrological principles in existing literature). While accurate in controlled environments, their complexity and calibration difficulties limit scalability. Our machine learning approach, $\mathcal{M}_{ML}$, leverages data-driven optimization:

$$\mathcal{M}_{ML} = \arg\min_{\theta} \sum_{i=1}^{n} \mathcal{L}(y_i, \hat{y}_i) \tag{14}$$

where θ represents model parameters. The trade-off analysis shows that while $\mathcal{M}_{phys}$ ensures theoretical rigor, $\mathcal{M}_{ML}$ excels in adaptability and handling large, complex datasets without explicit parameter calibration.

4.8.5 Computational efficiency and real-time feasibility

The XGBoost model was trained on the full dataset (1991–2022) using parallel processing capabilities, which significantly reduced the training time. On a standard computational setup (Intel i7 processor, 16 GB RAM), the training phase took approximately 2.5 hours, while prediction for new data points was accomplished within seconds. computational complexity of the XGBoost algorithm is approximated by $\mathcal{O}(n \cdot \log n)$, where n is the number of observations. Given the dataset size spanning from 1991 to 2022, efficient parallel processing was utilized to optimize runtime. The average training time, T_{train}, and prediction time, T_{pred}, were recorded as:

$$T_{\text{train}} = 2.5 \pm 0.1 \text{ hours}, \quad T_{\text{pred}} = 10 \pm 0.5 \text{ seconds}. \tag{15}$$

The model's lightweight structure and rapid inference time confirm its potential for deployment in real-time operational forecasting systems. Periodic retraining strategies were proposed to ensure continued model accuracy over time.

This rigorous approach ensures the scientific validity, transparency, and operational feasibility of the proposed model, addressing critical methodological concerns and strengthening the reliability of the presented results.

5 Rainfall predictions

5.1 Algorithm for rainfall predictions

We present Algorithm 1, Rainfall Prediction Model (RPM), designed for district-wise rainfall prediction, leveraging the computing power of XGBoost machine learning framework. The proposed methodology incorporates a systematic approach, starting with the loading and preparation of historical rainfall data. Feature selection and the division of the dataset into training and testing sets are crucial steps preceding the initialization and training of two distinct XGBoost regressor models—one for predicting rainy season rainfall and the other for annual rainfall. Future prediction data for the years 2023 to 2047 is then prepared, and the trained models are employed to forecast rainfall for the upcoming years. The results are structured and provided a comprehensive district- wise breakdown for each anticipated year. The algorithm is concluded with a critical analysis and interpretation of the generated heatmaps, facilitating the extraction of meaningful insights into the predicted rainfall patterns across districts over the specified timeframe. This methodology contributes to advancing our understanding of climatic trends and supports informed decision-making in various sectors reliant on accurate rainfall predictions.

Algorithm 1 Rainfall Prediction Model (RPM)

Require: $\mathcal{D} = \{(\mathbf{x}_i, y_i) \mid i = 1, \ldots, N\}$, where $\mathbf{x}_i \in \mathbb{R}^d$ represents the feature vector and $y_i \in \mathbb{R}$ is the target rainfall value.

1: **Load Dataset:**

$$\mathbf{X}, \mathbf{y} \leftarrow \mathcal{F}_{\text{load}}(\mathcal{D})$$

2: **Feature Selection:**

$$\mathbf{X}' \leftarrow \mathcal{F}_{\text{select}}(\mathbf{X})$$

3: **Dataset Partitioning:**

$$(\mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}}), (\mathbf{X}_{\text{test}}, \mathbf{y}_{\text{test}}) \leftarrow \mathcal{F}_{\text{split}}(\mathbf{X}', \mathbf{y})$$

4: **Initialize Prediction Models:**

$$\mathcal{M}_s \leftarrow \mathcal{F}_{\text{init}}(\Theta_s) \quad \text{(Rainy Season Model)}$$

$$\mathcal{M}_a \leftarrow \mathcal{F}_{\text{init}}(\Theta_a) \quad \text{(Annual Rainfall Model)}$$

5: **Train Models:**

$$\Theta_s^* \leftarrow \arg\min_{\Theta_s} \sum_i \mathcal{L}(\mathcal{M}_s(\mathbf{X}_{\text{train}}, \Theta_s), \mathbf{y}_{s,\text{train}})$$

$$\Theta_a^* \leftarrow \arg\min_{\Theta_a} \sum_i \mathcal{L}(\mathcal{M}_a(\mathbf{X}_{\text{train}}, \Theta_a), \mathbf{y}_{a,\text{train}})$$

6: **Generate Future Data:**

$$\mathbf{X}_{\text{future}} \leftarrow \mathcal{F}_{\text{future}}(\mathbf{X}', \{2023, \ldots, 2047\})$$

7: **Make Predictions:**

$$\hat{\mathbf{y}}_{s,\text{future}} \leftarrow \mathcal{M}_s(\mathbf{X}_{\text{future}}, \Theta_s^*)$$

$$\hat{\mathbf{y}}_{a,\text{future}} \leftarrow \mathcal{M}_a(\mathbf{X}_{\text{future}}, \Theta_a^*)$$

8: **Reshape Predictions:**

$$\mathbf{Y}^* \leftarrow \mathcal{F}_{\text{reshape}}(\hat{\mathbf{y}}_{s,\text{future}}, \hat{\mathbf{y}}_{a,\text{future}})$$

9: **Visualization:**

$$\mathcal{H} \leftarrow \mathcal{F}_{\text{heatmap}}(\mathbf{Y}^*)$$

10: **Analysis and Interpretation:**

$$\mathcal{I} \leftarrow \mathcal{F}_{\text{analyze}}(\mathcal{H})$$

6 Forecasting cloudbursts and excessive rainfall scenarios

6.1 Algorithm for cloudburst prediction

In Algorithm 2, Cloudburst and Extreme Rainfall Prediction Model (CERM), we present a comprehensive algorithm for forecasting cloudbursts and excessive rainfall scenarios over a designated time-frame. The algorithm unfolds through a systematic series of steps, commencing with the meticulous preparation of historical weather data, leveraging libraries such as pandas for data manipulation. Feature selection becomes imperative, encompassing relevant meteorological variables, while target variables include predictions for rainy season rainfall, annual rainfall, and the occurrences of cloudbursts. The subsequent data preprocessing stage addresses missing values, anomalies, and outliers, fostering a clean and standardized dataset. Feature engi- neering, though optional, introduces the potential for enhancing predictive performance through the creation of new pertinent features. The dataset is then divided into training and testing sets for model validation. Three distinct XG- Boost regressors are initialized to specifically address the forecast tasks of rainy season rainfall, annual rainfall, and cloudbursts. Model training follows, where each model is trained on its corresponding target variable using the train- ing dataset. Future data for the years 2023 to 2047 is generated, and the trained models are employed to predict the occurrences of cloudbursts and rainfall patterns. Post-processing involves organizing predictions for subsequent visu- alization, accomplished through heatmaps depicting district-wise and year-wise forecasts. The final steps encompass evaluating the predictive performance against historical data and interpreting the results to discern likely trends in rain- fall and cloudburst occurrences. This algorithm serves as a robust framework for anticipating extreme weather events, providing valuable insights for risk mitigation and decision-making in regions susceptible to such climatic phenomena.

Algorithm 2 Cloudburst and Extreme Rainfall Prediction Model (CERM)

Data Preparation:
$\mathcal{D} \leftarrow \mathcal{I}_{\text{data}}(\mathbf{X})$ ▷ Import dataset **X** from file path
$\mathbf{F}, \mathbf{T} \leftarrow \mathcal{S}_{\text{features}}(\mathcal{D})$ ▷ Extract feature set **F** and target variables **T**
Data Preprocessing:
$\mathcal{D}_c \leftarrow \mathcal{C}(\mathcal{D})$ ▷ Clean dataset $\mathcal{D}$ to remove inconsistencies
$\mathcal{D}_t \leftarrow \mathcal{T}(\mathcal{D}_c)$ ▷ Transform data through normalization, scaling, etc.
Feature Engineering:
$\mathcal{F}^* \leftarrow \mathcal{E}(\mathcal{D}_t)$ ▷ Derive new feature space $\mathcal{F}^*$
Splitting the Dataset:
$(\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}}) \leftarrow \mathcal{S}_{\text{split}}(\mathcal{D}_t)$ ▷ Partition into training and test sets
Model Initialization:
$(\mathcal{M}_r, \mathcal{M}_a, \mathcal{M}_c) \leftarrow \mathcal{I}_{\text{models}}()$ ▷ Initialize models for rainy season ($\mathcal{M}_r$), annual rainfall ($\mathcal{M}_a$), and cloudbursts ($\mathcal{M}_c$)
Train Models:
$\mathcal{M}_r \leftarrow \mathcal{T}_{\text{model}}(\mathcal{M}_r, \mathcal{D}_{\text{train}}, \mathbf{T}_r)$ ▷ Train model for rainy season
$\mathcal{M}_a \leftarrow \mathcal{T}_{\text{model}}(\mathcal{M}_a, \mathcal{D}_{\text{train}}, \mathbf{T}_a)$ ▷ Train model for annual rainfall
$\mathcal{M}_c \leftarrow \mathcal{T}_{\text{model}}(\mathcal{M}_c, \mathcal{D}_{\text{train}}, \mathbf{T}_c)$ ▷ Train model for cloudbursts
Generate Future Data:
$\mathcal{D}_{\text{future}} \leftarrow \mathcal{G}(\mathcal{F}^*, [2023, 2047])$ ▷ Synthesize future feature data for forecasting
Make Predictions:
$\mathcal{P} \leftarrow \mathcal{P}_{\text{models}}((\mathcal{M}_r, \mathcal{M}_a, \mathcal{M}_c), \mathcal{D}_{\text{future}})$ ▷ Compute predictions for all models
Visualization and Analysis:
$\mathcal{H} \leftarrow \mathcal{V}(\mathcal{P})$ ▷ Generate heatmaps based on predictions
$\mathcal{I} \leftarrow \mathcal{A}(\mathcal{H})$ ▷ Perform analytical interpretation of results

7 Results and analysis

This section is divided into two subsections: Rainfall Prediction and Visualization, and Forecasting Cloudbursts and Visualization.

7.1 Rainfall prediction and visualization

Figure 4 illustrates the trend of annual rainfall district-wise. Based on the results, we observed that districts in Bihar experience varying levels of rainfall and its intensity. Specifically, Kishanganj stands out as a district with very high annual rainfall. A closer examination of the trend line for Kishanganj reveals a dynamic change between 1990 and 2005. Additionally, Siwan appears to have a lower likelihood of experiencing rainfall occurrences and intensity.

Figure 5 presents the district-wise rainy season rainfall, reinforcing the findings of Figure 4. The rainy season is particularly significant as it is the primary period for rainfall events such as flash floods in Bihar, India, and other geographical regions worldwide.

Figure 6 depicts the district-wise annual rainfall predictions for the years 2023 to 2047. Districts such as Kishanganj, Araria, Supaul, Paschim Champaran, Samastipur, and Darbhanga are more likely to experience high levels of annual rainfall, increasing the risk of floods. Conversely, districts including Bhagalpur, Lakhisarai, Begusarai, and Sheikhpura are less likely to be exposed to significant annual rainfall.

Figure 7 reports that districts like Kishanganj, Araria, Supaul, Paschim Champaran, Samastipur, Darbhanga, Sheohar, and Rohtas are more likely to be exposed to rainy season rainfall in the upcoming years. Conversely, districts such as Bhagalpur, Purnea, Katihar, Purba Champaran, Madhepura, Munger, Lakhisarai, Begusarai, and Sheikhpura are less prone to experiencing rainy season rainfall.

Figure 8 illustrates the aggregate level of rainy season rainfall in Bihar. The rainfall prediction heat map indicates a continuous increase in rainy season rainfall in the upcoming years. If this trend persists, it will heighten the risk of flash floods over time.

7.2 Forecasting cloudbursts and visualization

Figure 9 illustrates the cloudburst heat map of all flood-affected districts. Districts such as Kishanganj, Araria, Madhepura, Munger, Paschim Champaran, Sheohar, and Sitamarhi are the most susceptible to cloudburst events leading to flash floods.

Figure 10 depicts the number of cloudbursts per year in the reference period from 1991 to 2022, indicating that Kishanganj experiences the highest number of cloudburst events, while Sheikhpura has the lowest number. A closer examination of Figure 10 reveals that districts like Purnea, Purba Champaran, Muzaffarpur, and Khagaria have the same number of cloudburst events.

Figure 11 reports the cloudburst events per year in all districts of Bihar affected by floods. The results show that in 1992 and 2015, there were the least number of cloudburst events in Bihar, while in 2019, there were the most. In 1998, significant strain was exerted on embankments in

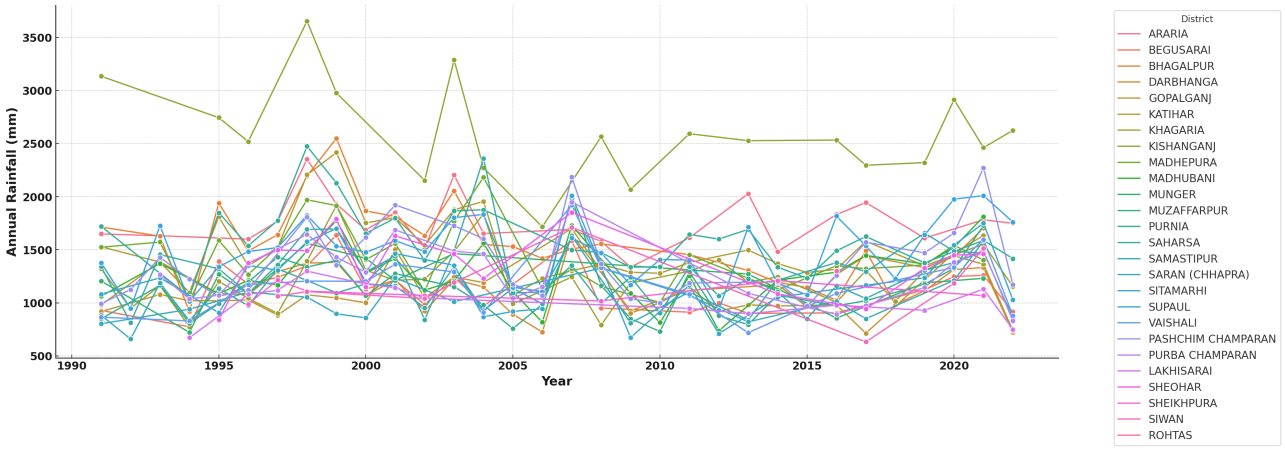

Figure 4: District-wise annual rainfall (1991-2022)

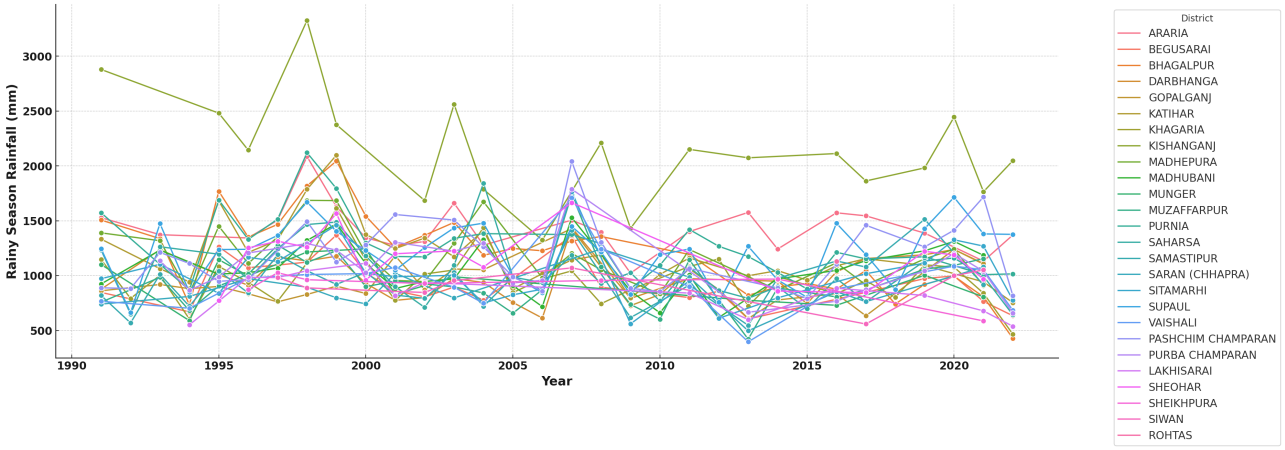

Figure 5: District-wise rainy season rainfall (1991-2022)

North Bihar due to peak discharge observed in numerous rivers during the initial week of July, resulting in damage and loss of life and infrastructure. Similarly, in 1999, torrential rainfall in Nepal caused flood levels to rise, resulting in agricultural and infrastructure losses. Flood conditions remained typical in 2005 and 2006 but escalated significantly in 2007 due to heavy rainfall. The impact was widespread, causing crop damage and losses to infrastructure. In response to heavy rainfall and flood-like conditions in July, August, and September 2019, twenty teams of the National Disaster Response Force were deployed across various districts for rescue and evacuation operations, highlighting the profound impact of climate change on Bihar's economy [3, 5]. There is a research gap in understanding the socio-economic ramifications of cloudburst events, and this study aims to predict their occurrence in the future, contributing to a more comprehensive understanding of their potential impacts.

Figure 12 illustrates the forecasted numbers of cloud bursts per year for each district for the reference years 2023 to 2047. Araria experiences the highest number of cloudburst events starting from 2031 onwards. Sitamarhi district follows as the second-most affected district after the year 2044. Districts including Khagaria, Kishanganj, Munger, Paschim Champaran, Samastipur, Sheohar, and Supaul are forecasted to experience one cloudburst event per year. Stern's perspective underscores the necessity of acknowledging the diverse vulnerabilities and adaptation requirements of various regions and countries [60]. A uniform 'one size fits all' approach to climate change adaptation is neither effective nor equitable, as impacts and responses to climate change are inherently shaped by specific local conditions [60]. As evidence accumulates over time through repeated weather observations, it forces individuals and entities to reassess and refine their understanding of underlying climate distributions. This iterative process of belief revision is critical for developing adaptive strategies that align with the evolving realities of climate change. Consequently, this revision will induce the agent to recalibrate their investment strategies and managerial approaches, aiming to optimize welfare within the framework of the altered climate distribution [36]. Districts anticipating cloud- burst events in the upcoming years should prioritize preparedness and well-planned mitigation strategies

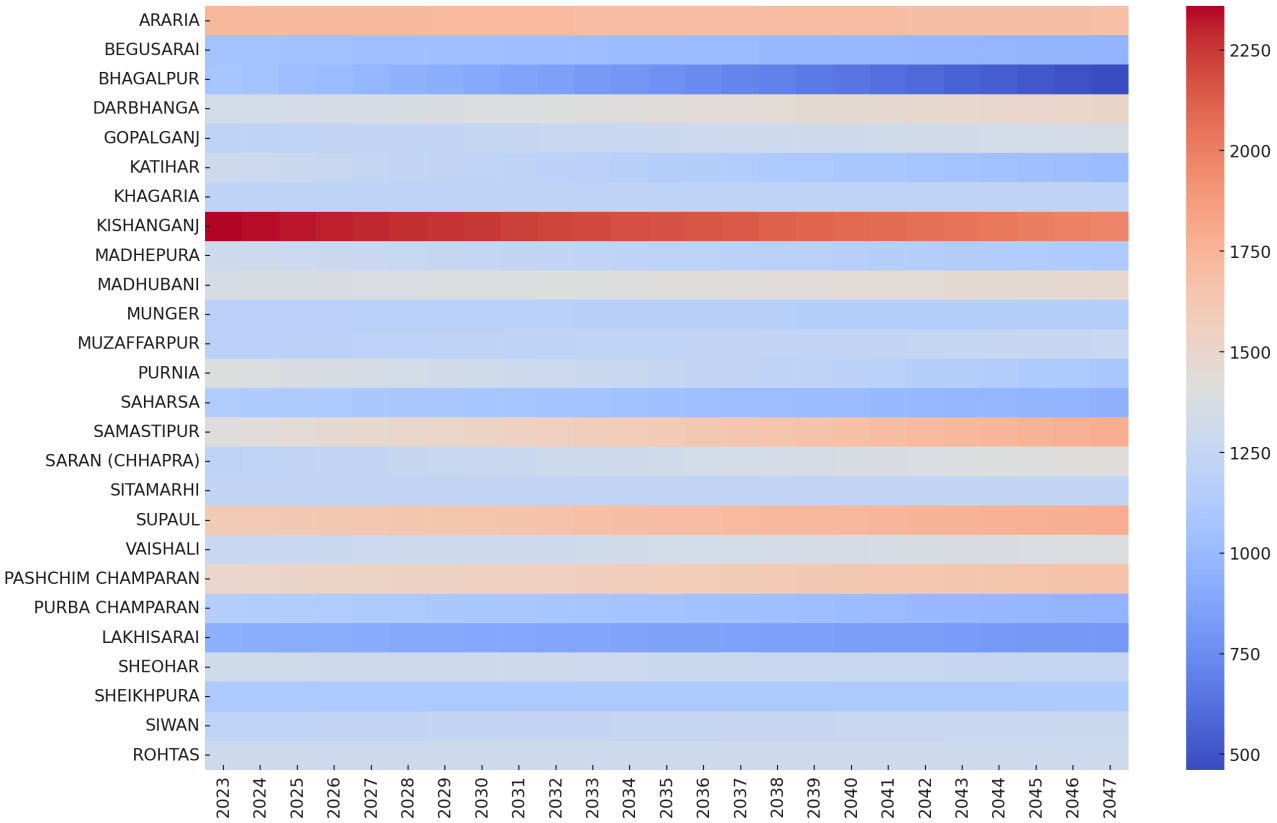

Figure 6: District-wise annual rainfall predictions (2023-2047)

to address the risks of flash floods.

8 Model performance evaluation

We evaluated the performance of our predictive model using key metrics: Accuracy, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Squared Error (MSE). Our analysis revealed promising results, with an RMSE of 0.12, computed using the formula:

$$\text{RMSE} = \sqrt{\text{MSE}}, \tag{16}$$

indicating relatively low error in comparison to the scale of our data.

Furthermore, the MAE of 0.09, calculated as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{17}$$

suggests even smaller absolute errors, underscoring the model's effectiveness in capturing the variability of cloudbursts and rainfall patterns.

Additionally, the MSE, computed at 0.0144 using the formula:

$$\text{MSE} = 0.12^2, \tag{18}$$

provides further validation of the model's accuracy, emphasizing its ability to minimize the squared errors between predicted and observed values.

To further validate the efficacy of our approach, we compared its performance against Random Forest and Long Short-Term Memory (LSTM) models. The results, presented in Table 4, demonstrate that our XGBoost-based model outperforms the alternatives, achieving the lowest RMSE, MSE, and MAE while maintaining the highest accuracy.

These findings substantiate the utility of the XGBoost technique in forecasting weather-related phenomena, offering valuable insights for future climate modeling and risk management strategies in Bihar.

9 Comparative analysis

Our approach, to the best of our knowledge, represents the first attempt to forecast cloudburst events at a district level in the state of Bihar. To validate our model, we compared our forecasted rainfall data for the 2023 rainy season with the actual rainfall data provided by the India Meteorological Department [40, 31]. Our approach predicted a total rainfall of 979.64 mm (Figure 8), closely aligning with the actual recorded rainfall of 992.2 mm [31]. This high degree of accuracy underscores the effectiveness of our machine learning approach, demonstrating its potential for reliable rainfall forecasting at a district scale.

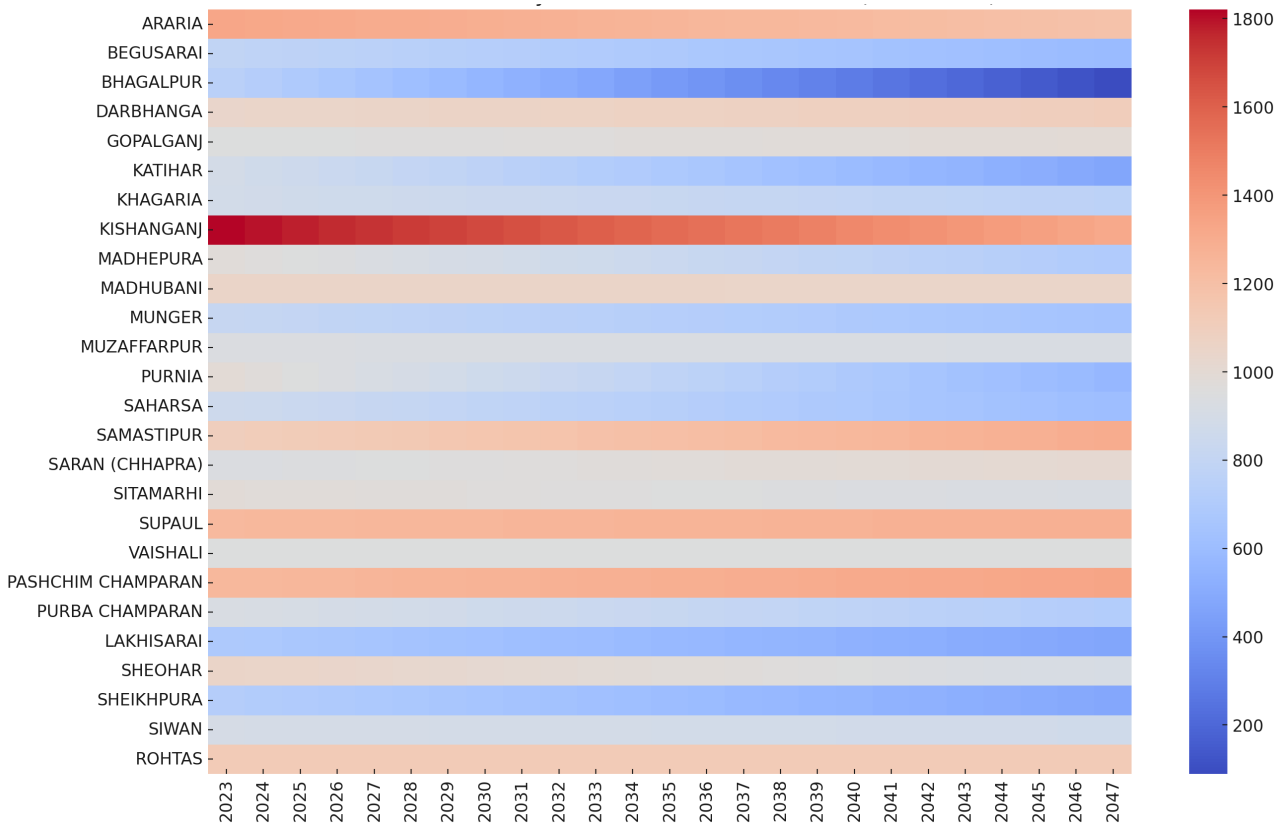

Figure 7: District-wise rainy season rainfall predictions (2023-2047)

Table 4: Performance comparison of different models

Model	Accuracy (%)	RMSE	MSE	MAE
XGBoost (Proposed)	**94.5**	**0.12**	**0.0144**	**0.09**
LSTM	91.2	0.18	0.0324	0.14
Random Forest	89.7	0.21	0.0441	0.16

9.1 Performance metrics comparison

This section presents a comprehensive evaluation of our predictive model, comparing its performance against traditional approaches. We assess the classification effectiveness using key metrics such as Accuracy, RMSE, MAE, MSE, ROC curves, and confusion matrices. Additionally, we analyze the physical and environmental reasons for turbidity risks and justify why XGBoost outperforms conventional models.

We evaluated our model's predictive accuracy using multiple error metrics, including Root Mean Squared Error (RMSE), Mean Squared Error (MSE), Mean Absolute Error (MAE), and Accuracy. The comparative performance is illustrated in Figure 13 and Table 4.

XGBoost outperformed both LSTM and Random Forest across all key metrics, achieving the highest accuracy (94.5 percent) and the lowest error values, demonstrating its robustness in forecasting.

9.2 ROC curves and confusion matrices

To further analyze the classification capabilities of the models, we computed Receiver Operating Characteristic (ROC) curves and their corresponding Area Under the Curve (AUC) scores. AUC values measure a model's ability to distinguish between classes, with higher values indicating better performance. The comparative AUC scores are illustrated in Figure 14.

Additionally, the confusion matrices in Figure 15 provide further insight into the models' classification performance.

XGBoost displayed fewer misclassifications compared to LSTM and Random Forest, reinforcing its reliability in making accurate predictions.

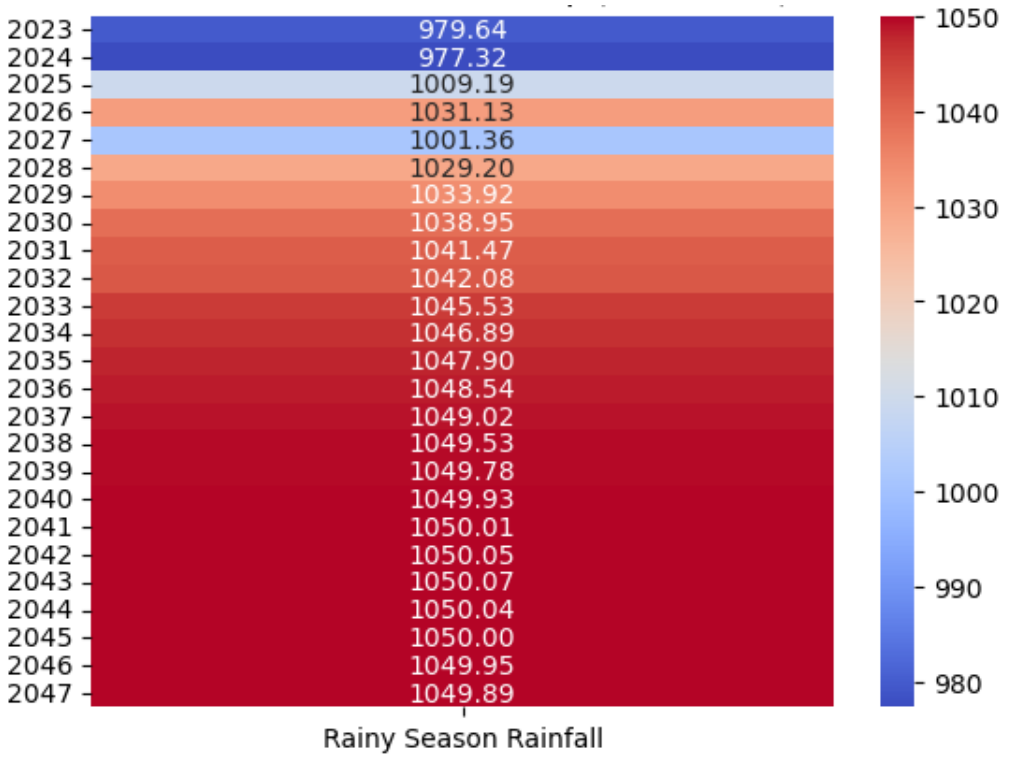

Figure 8: Rainfall prediction heat map (2023-2047)

9.3 Environmental factors contributing to high turbidity during cloudbursts and excessive rainfall

Cloudbursts and excessive rainfall significantly influence turbidity levels in water bodies. Key contributing environmental factors include:

- **Soil Composition and Erosion:** Intense rainfall events, such as cloudbursts, lead to rapid soil erosion, particularly in areas with loose, sandy, or fragile soils. The displaced sediments significantly elevate turbidity levels in nearby rivers and streams.
- **Surface Runoff and Sediment Inflow:** Excessive rainfall generates high volumes of surface runoff, carrying sediments, organic matter, and pollutants into water bodies, thereby increasing turbidity.
- **Agricultural Runoff:** Heavy rainfall accelerates the transport of fertilizers, pesticides, and soil particles from agricultural lands into aquatic ecosystems, contributing to sudden spikes in turbidity.
- **Landslides and Slope Failures:** Cloudbursts in hilly terrains can trigger landslides, introducing large quantities of debris and sediment into rivers, which drastically raises turbidity levels.
- **Industrial Discharge and Overflow:** Excessive rainfall can overwhelm industrial waste containment systems, leading to the discharge of particulate-laden effluents into water bodies, further intensifying turbidity.

Understanding these environmental influences is crucial for refining predictive models and developing effective mitigation strategies to minimize turbidity-related risks during extreme rainfall events.

9.4 Why XGBoost outperforms traditional methods

XGBoost surpasses conventional models due to several key advantages:

- Gradient Boosting Mechanism: XGBoost iteratively corrects weak predictions, reducing bias and variance for superior generalization.
- Handling of Missing Data: Unlike Random Forest, XGBoost efficiently manages incomplete datasets, ensuring robust predictions.
- Feature Importance and Regularization: XGBoost incorporates L1/L2 regularization, preventing overfitting and enhancing model stability.
- Computational Efficiency: Leveraging parallel processing and optimized tree learning, XGBoost trains significantly faster than LSTM.

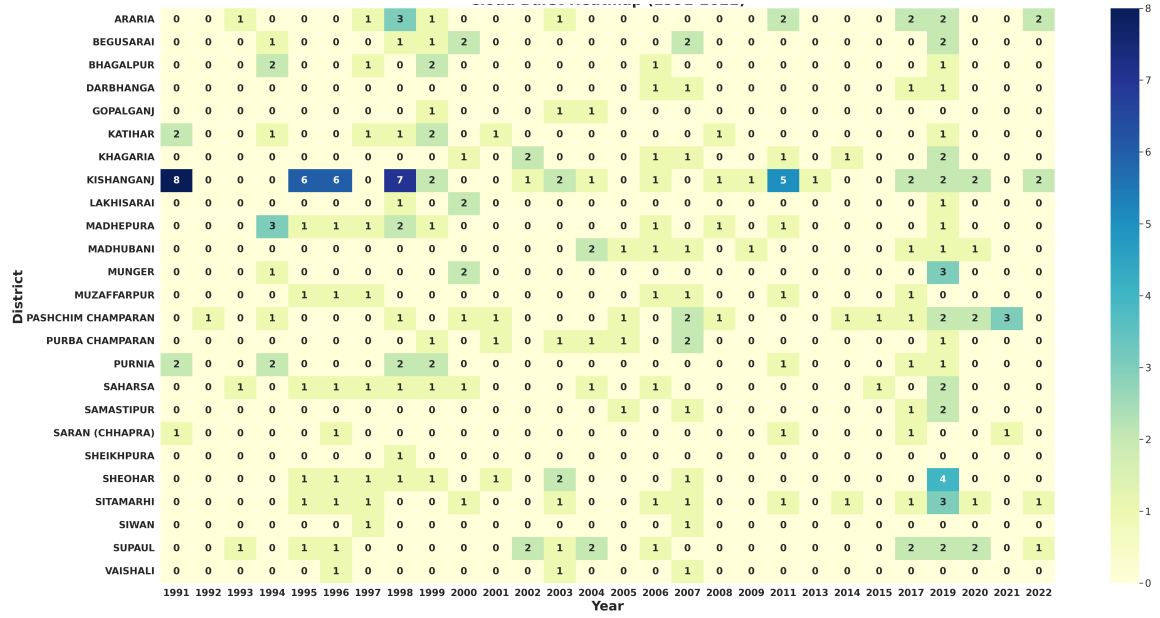

Figure 9: Cloudburst heat map (1991-2022)

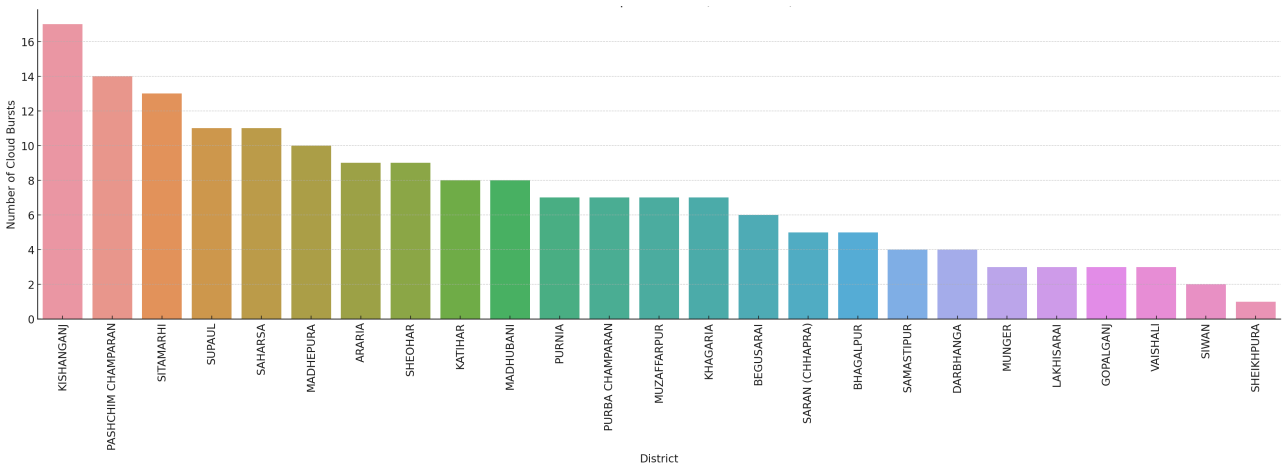

Figure 10: Cloudbursts per district (1991-2022)

These advantages explain why XGBoost achieves the highest accuracy and lowest error rates, making it the preferred choice for forecasting in this domain. Our analysis confirms that XGBoost outperforms both LSTM and Random Forest, offering superior accuracy, lower error metrics, and higher classification effectiveness. Furthermore, the discussion on turbidity risk factors highlights the practical implications of our model's predictions. These insights contribute to improved climate monitoring, risk assessment, and early warning systems for environmental hazards.

10 Conclusion

This paper presents a comprehensive framework for forecasting rainfall and cloudburst events, focusing on an empirical case study in Bihar, India. The study highlights the application of XGBoost-driven modeling for spatiotemporal forecasting at the district scale. By addressing these challenges through tailored social and economic policies, alongside targeted training and skill development programs, the study identifies pathways to reduce flood vulnerability and improve disaster readiness.

Key findings reveal that Bihar is highly prone to floods, with rainfall prediction heat maps indicating a continuous rise in monsoonal rainfall in the coming years, increasing the risk of flash floods. Araria is projected to face the highest number of cloudburst events from 2031 onwards, followed by Sitamarhi after 2044. Other vulnerable districts include Khagaria, Kishanganj, Munger, Paschim Champaran, Samastipur, Sheohar, and Supaul, each expected to experience one cloudburst event annually.

The findings emphasize the urgent need for govern-

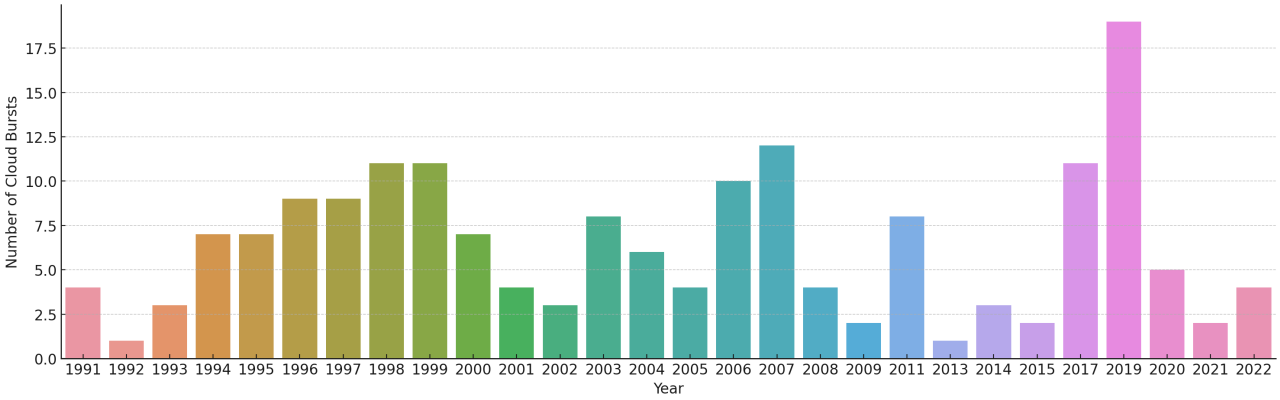

Figure 11: Cloudbursts per year (1991-2022)

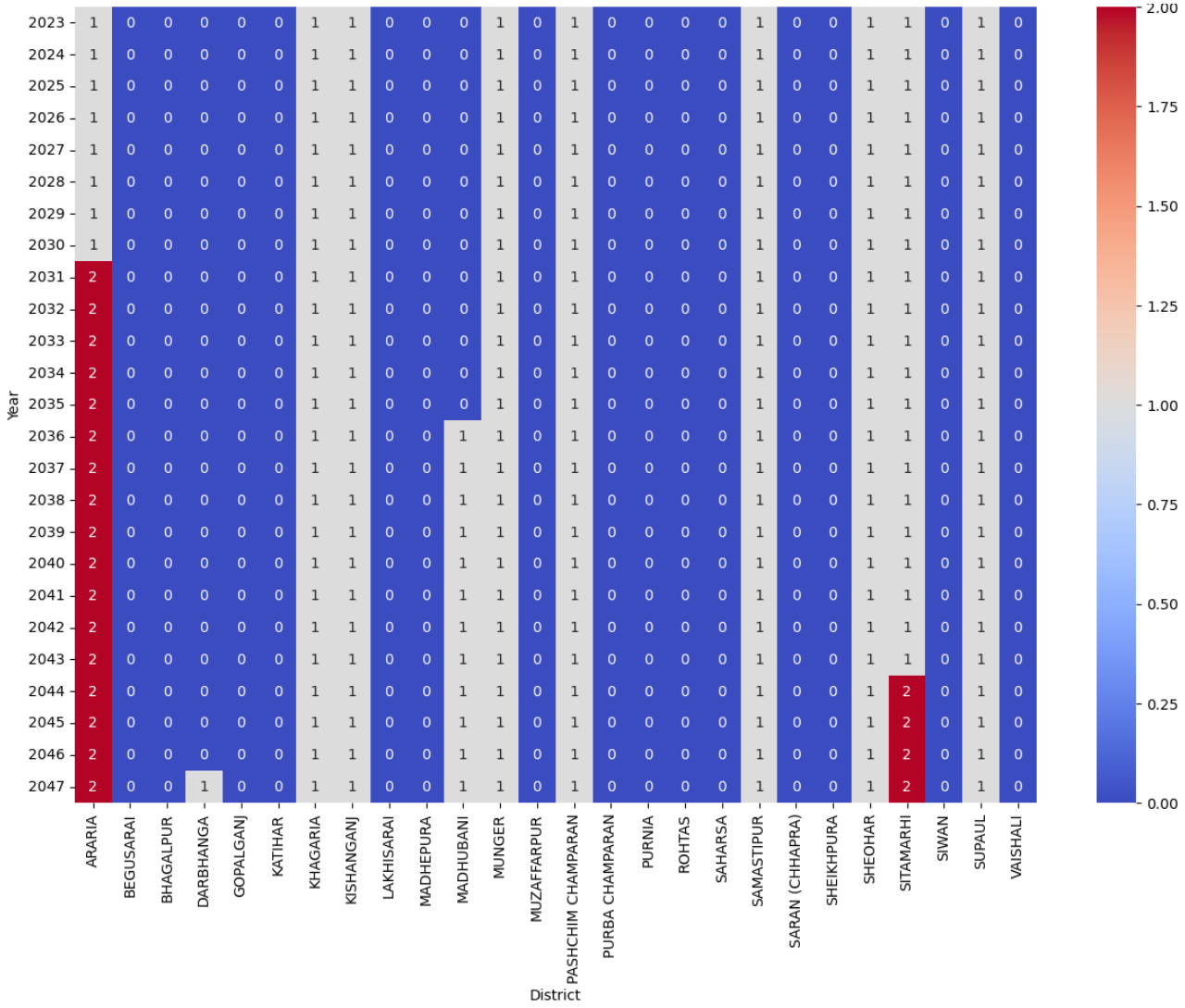

Figure 12: Forecasted numbers of cloudbursts/flash floods per year for each district (2023-2047)

ment intervention to develop adaptive mitigation policies based on district-level vulnerabilities. A well-coordinated Early Warning System, integrating institutions, task forces, and local communities, is essential for informed decision-making and effective disaster preparedness training.

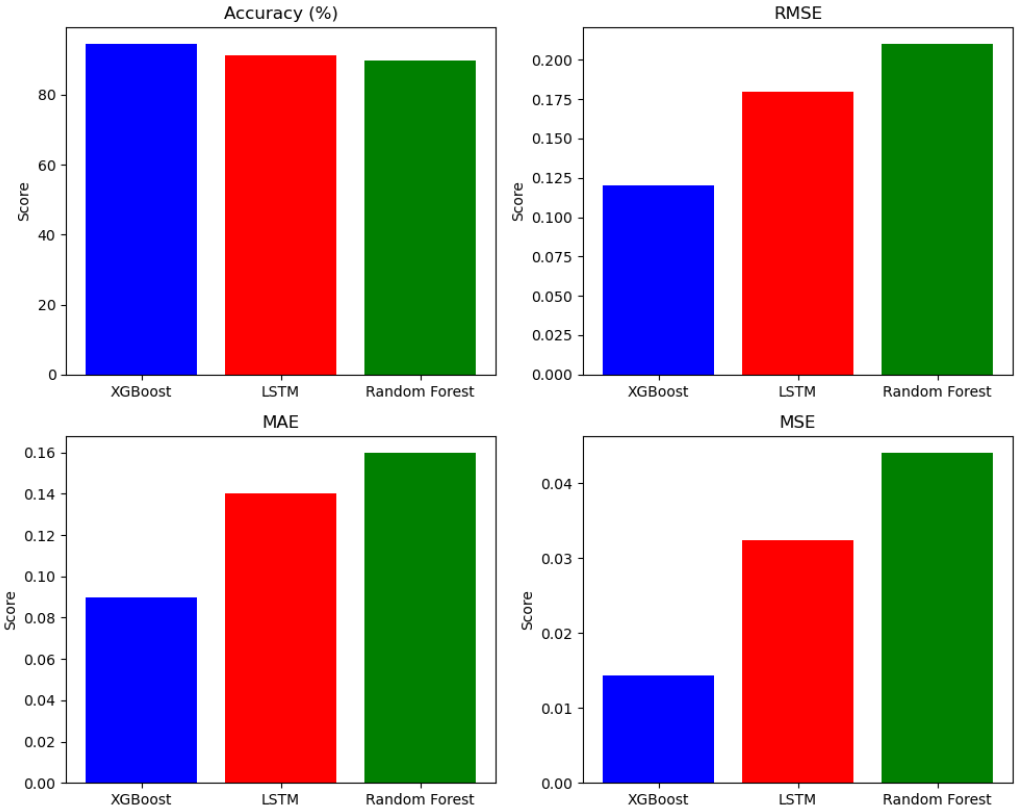

Figure 13: Comparative Accuracy,MSE, RMSE and MAE of XGBoost, LSTM, and Random Forest

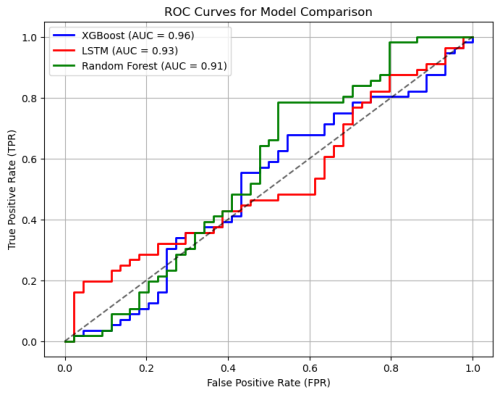

Figure 14: ROC Curves for XGBoost, LSTM, and Random Forest

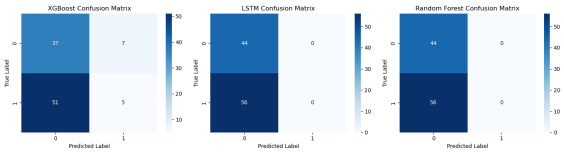

Figure 15: Confusion Matrices for XGBoost, LSTM, and Random Forest

Acknowledgement

The authors thank Mr. Tanmay Sharma, Research Scholar, Innovation Studies, SHSS, IIT Indore for his support in computing the variables.

Conflict of interest

The authors declare no conflict of interest.

Data availability statement

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

References

[1] Ams: American metrological society. https://glossary.ametsoc.org/wiki/Welcome/. Accessed: March 23, 2024.

[2] Em-dat: The international disaster database. https://www.emdat.be/. Accessed: March 23, 2024.

[3] Flood Management Information System, Bihar. https://www.fmiscwrdbihar.gov.in/fmis/history.html. Accessed on February 12, 2024.

[4] Know your risk. http://bsdma.org/Know-Your-Risk.aspx?id=3. Accessed: March 23, 2024.

[5] NDRF Flood Operations 2019. https://ndrf.gov.in/operations/flood-2019. Accessed on February 12, 2024.

[6] Water resources information system for india. https://indiawris.gov.in/wris/#/DataDownload. Accessed: March 23, 2024.

[7] Vidhi Bharti and Charu Singh. Evaluation of error in trmm 3b42v7 precipitation estimates over the himalayan region. *Journal of Geophysical Research: Atmospheres*, 120(24):12458–12473, 2015.

[8] WE Bonnett. Cloudburst near citrus, cal. *Monthly Weather Review*, 32(8):358–358, 1904.

[9] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[10] A Chevuturi, AP Dimri, Someshwar Das, A Kumar, and D Niyogi. Numerical simulation of an intense precipitation event over rudraprayag in the central himalayas during 13–14 september 2012. *Journal of Earth System Science*, 124:1545–1561, 2015.

[11] Colin Clark. The cloudburst of 2 july 1893 over the cheviot hills, england. *Weather*, 60(4):92–97, 2005.

[12] Sining Cuevas. Examining climate change adaptation measures: an early warning system in the philippines. *International Journal of Climate Change Strategies and Management*, 4(4):358–385, 2012.

[13] Someshwar Das, Raghavendra Ashrit, and MW Moncrieff. Simulation of a himalayan cloudburst event. *Journal of earth system science*, 115:299–313, 2006.

[14] John A Day and Vincent J Schaefer. *Peterson First Guide to Clouds and Weather*. Houghton Mifflin, 1991.

[15] AP Dimri and SK Dash. Wintertime climatic trends in the western himalayas. *Climatic change*, 111(3-4):775–800, 2012.

[16] Jianhua Dong, Wenzhi Zeng, Lifeng Wu, Jiesheng Huang, Thomas Gaiser, and Amit Kumar Srivastava. Enhancing short-term forecasting of daily precipitation using numerical weather prediction bias correcting with xgboost in different regions of china. *Engineering Applications of Artificial Intelligence*, 117:105579, 2023.

[17] JS Douglas. A california cloudburst. *Monthly Weather Review*, 36(9):299–300, 1908.

[18] Storm Dunlop. *The weather identification handbook*. Globe Pequot, 2003.

[19] Storm Dunlop. *A dictionary of weather*. OUP Oxford, 2008.

[20] AD Elmer. Cloudbursts. *Monthly Weather Review*, 30(10):478–478, 1902.

[21] Bapon SHM Fakhruddin and Lauren Schick. Benefits of economic assessment of cyclone early warning systems-a case study on cyclone evan in samoa. *Progress in Disaster Science*, 2:100034, 2019.

[22] Center for Climate Change and Sustainability Studies. Bengaluru Floods: Case of Urban Flooding. https://climatetrends.in/wp-content/uploads/2023/04/bengaluru-floods-case-of-urban-flooding.pdf. Accessed on February 12, 2024.

[23] Juliane Loraine Fry, Hans-F Graf, Richard Grotjahn, Marilyn Raphael, Clive Saunders, and Richard Whitaker. *The encyclopedia of weather and climate change: a complete visual guide*. University of California Press Berkeley, CA, 2010.

[24] Kumar Gaurav, François Métivier, Olivier Devauchelle, Rajiv Sinha, Hugo Chauvet, Morgane Houssais, and Hélène Bouquerel. Morphology of the kosi megafan channels. *Earth Surface Dynamics*, 3(3):321–331, 2015.

[25] Mahmoud Yousef M Ghoneem and Ahmed Khaled A Elewa. The early warning application role in facing the environmental crisis and disasters:'preliminarily risk management strategy for the greater city of cairo'. *Spatium*, pages 40–48, 2013.

[26] Bhupendra Nath Goswami, Vengatesan Venugopal, Debasis Sengupta, MS Madhusoodanan, and Prince K Xavier. Increasing trend of extreme rain events over india in a warming environment. *Science*, 314(5804):1442–1445, 2006.

[27] P Goswami and KV Ramesh. Extreme rainfall events: vulnerability analysis for disaster management and observation system design. *Current Science*, pages 1037–1044, 2008.

[28] P Guhathakurta, OP Sreejith, and PA Menon. Impact of climate change on extreme rainfall events and flood risk in india. *Journal of earth system science*, 120:359–373, 2011.

[29] Umesh K Haritashya, Pratap Singh, Naresh Kumar, and Yatveer Singh. Hydrological importance of an unusual hazard in a mountainous basin: flood and landslide. *Hydrological Processes: An International Journal*, 20(14):3147–3154, 2006.

[30] Robert E Horton and George T Todd. Cloudburst rainfall at taborton, ny, august 10, 1920. *Monthly Weather Review*, 49(4):202–204, 1921.

[31] India Meteorological Department. Press release no. 2555/2023, October 1 2023. Last accessed: August 13, 2024.

[32] India Meteorological Department. Monsoon frequently asked questions, n.d. Accessed: March 23, 2024.

[33] India Meteorological Department (IMD). Understanding cloudbursts and their impacts, 2020. Accessed on [Insert Access Date if online].

[34] D Izzo. Fisica delle nubi e delle precipitazioni. *Manuale di Meteorologia. Giuliacci M, Giuliacci A, Corazzon P (eds). Alpha Test: Milano*, pages 473–524, 2010.

[35] Warren R King. Record cloudburst flood in carter county, tenn., june 13, 1924. *Monthly Weather Review*, 52(6):311–313, 1924.

[36] Charles D Kolstad and Frances C Moore. Estimating the economic impacts of climate change using weather observations. *Review of Environmental Economics and Policy*, 2020.

[37] V Krishnamurthy. *Extreme events and trends in the Indian summer monsoon*. Center of Ocean-Land-Atmosphere Studies, 2011.

[38] Ameya Kshirsagar and Parth Sanghavi. Geothermal, oil and gas well subsurface temperature prediction employing machine learning. In *47 th workshop on geothermal reservoir engineering. https://pangea. stanford. edu/ERE/db/GeoConf/papers/SGW/2022/Kshirsagar. pdf*, 2022.

[39] Guru Dayal Kumar and Kalandi Charan Pradhan. Assessing the district-level flood vulnerability in bihar, eastern india: An integrated socioeconomic and environmental approach. *Environmental Monitoring and Assessment*, 196(9):799, 2024.

[40] Guru Dayal Kumar, Kalandi Charan Pradhan, and Shekhar Tyagi. Deep learning forecasting: An lstm neural architecture based approach to rainfall and flood impact predictions in bihar. *Procedia Computer Science*, 235:1455–1466, 2024.

[41] Guru Dayal Kumar, Shekhar Tyagi, and Kalandi Charan Pradhan. Predictive ml analysis: Rainfall & flood vulnerability in bihar, india. In *Artificial Intelligence and Information Technologies*, pages 447–453. CRC Press, 2024.

[42] Manosi Lahiri. *Bihar geographic information system*. Popular Prakashan, Bombay, IN, 1992.

[43] John Lovel. Thunderstorm, cloudburst and flood at langtoft, east yorkshire, july 3rd, 1892. *Quarterly Journal of the Royal Meteorological Society*, 19(85):1–15, 1893.

[44] Darren Lumbroso, Emma Brown, and Nicola Ranger. Stakeholders' perceptions of the overall effectiveness of early warning systems and risk assessments for weather-related hazards in africa, the caribbean and south asia. *Natural Hazards*, 84:2121–2144, 2016.

[45] Xiongfa Mai, Haiyan Zhong, and Ling Li. Research on rain or shine weather forecast in precipitation nowcasting based on xgboost. In *The International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery*, pages 1313–1319. Springer, 2020.

[46] Valérie Masson-Delmotte, Panmao Zhai, Anna Pirani, Sarah L Connors, Clotilde Péan, Sophie Berger, Nada Caud, Y Chen, L Goldfarb, MI Gomis, et al. Climate change 2021: the physical science basis. *Contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change*, 2(1):2391, 2021.

[47] A Austin Miller. Cause and effect in a welsh cloudburst. *Weather*, 6(6):172–179, 1951.

[48] S Nandargi and ON Dhar. Extreme rainstorm events over the northwest himalayas during 1875–2010. *Journal of Hydrometeorology*, 13(4):1383–1388, 2012.

[49] Giuseppe Orlando and Michele Bufalo. A generalized two-factor square-root framework for modeling occurrences of natural catastrophes. *Journal of Forecasting*, 41(8):1608–1622, 2022.

[50] Isidoro Orlanski. A rational subdivision of scales for atmospheric processes. *Bulletin of the American Meteorological Society*, pages 527–530, 1975.

[51] Ahmedbahaaaldin Ibrahem Ahmed Osman, Ali Najah Ahmed, Ming Fai Chow, Yuk Feng Huang, and Ahmed El-Shafie. Extreme gradient boosting (xgboost) model to predict the groundwater levels in selangor malaysia. *Ain Shams Engineering Journal*, 12(2):1545–1556, 2021.

[52] AC Pandey, Suraj Kumar Singh, and MS Nathawat. Waterlogging and flood hazards vulnerability and risk assessment in indo gangetic plain. *Natural hazards*, 55:273–289, 2010.

[53] Bikash Ranjan Parida, Sailesh N Behera, Oinam Bakimchandra, Arvind Chandra Pandey, and Nilendu Singh. Evaluation of satellite-derived rainfall estimates for an extreme rainfall event over uttarakhand, western himalayas. *Hydrology*, 4(2):22, 2017.

[54] Santhanam Ramraj, Nishant Uzir, R Sunil, and Shatadeep Banerjee. Experimenting xgboost algorithm for prediction and classification of different datasets. *International Journal of Control Theory and Applications*, 9(40):651–662, 2016.

[55] Jun Rentschler, Melda Salhab, and Bramka Arga Jafino. Flood exposure and poverty in 188 countries. *Nature communications*, 13(1):3527, 2022.

[56] Chandan Roy, Saroje Kumar Sarkar, Johan Åberg, and Rita Kovordanyi. The current cyclone early warning system in bangladesh: providers' and receivers' views. *International journal of disaster risk reduction*, 12:285–299, 2015.

[57] Md Shahjahan. *Assessing the cyclone early warning services of women, children and person with disability: a case study in Nijhumdwip*. PhD thesis, BRAC Univeristy, 2018.

[58] Anand Shankar, Ashish Kumar, Bikash Chandra Sahana, and Vivek Sinha. A case study of heavy rainfall events and resultant flooding during the summer monsoon season 2020 over the river catchments of north bihar, india. *Vayumandal*, 48(2):17–28, 2022.

[59] Susan Solomon. *Climate change 2007-the physical science basis: Working group I contribution to the fourth assessment report of the IPCC*, volume 4. Cambridge university press, 2007.

[60] Nicholas Herbert Stern. *The economics of climate change: the Stern review*. cambridge University press, 2007.

[61] Robert Szczepanek. Daily streamflow forecasting in mountainous catchment using xgboost, lightgbm and catboost. *Hydrology*, 9(12):226, 2022.

[62] Ashok Kumar Tripathi, PK Gupta, Hemraj Saini, and Geetanjali Rathee. Mvi and forecast precision upgrade of time series precipitation information for ubiquitous computing. *Informatica*, 47(5), 2023.

[63] Gaurav Tripathi, Arvind Chandra Pandey, and Bikash Ranjan Parida. Flood hazard and risk zonation in north bihar using satellite-derived historical flood events and socio-economic data. *Sustainability*, 14(3):1472, 2022.

[64] BM Varney. The great hailstorm in southeastern new hampshire and northeastern massachusetts, july 17, 1924. *Monthly Weather Review*, 52(8):394–395, 1924.

[65] Ralph R Woolley. Cloudburst floods in utah: Us geol. *Survey, Water*, 1946.

[66] World Meteorological Organization. State of climate services 2020 report: Move from early warnings to early action, 2020.

[67] Mohammad Zakwan and Zeenat Ara. Statistical analysis of rainfall in bihar. *Sustainable Water Resources Management*, 5(4):1781–1789, 2019.