Flood Evacuation Process Using The PSO(Particle Swarm Optimization) Algorithm And Flood Prediction

Submitted by : Arpan Mukherjee

Ramakrishna Mission Residential College (Autonomous)

B.Sc. In Statistics

Mentor Name : Mrs. Sanchaita Chakraborty

Period of Internship: 18 th May 2026 – 31 st July 2026

Report submitted to: IDEAS – Institute of Data Engineering, Analytics and Science Foundation, ISI, Kolkata

Acknowledgement

I,Arpan Mukherjee, would like to express my sincere gratitude to everyone who supported and guided me throughout the course of this internship project, *"Particle Swarm Optimization Model for Flood Evacuation."*

I am deeply thankful to the **IDEAS-TIH Team, Indian Statistical Institute (ISI), Kolkata**, for providing me with this valuable opportunity to work on a real-world disaster-management problem and for creating an environment that encouraged learning, research, and innovation. The exposure to statistical thinking, data-driven methodologies, and computational tools during this internship has been instrumental in shaping my understanding of applied data science and optimization techniques.

I would like to extend my heartfelt appreciation to my mentor, **Ms. Sanchaita Chakraborty**, for her constant guidance, patience, and encouragement throughout this project. Her insightful feedback, technical direction, and willingness to clarify doubts at every stage helped me navigate the complexities of geospatial analysis, machine learning, and network optimization with confidence. Her mentorship played a pivotal role in refining both the technical depth and the practical relevance of this project.

I am also grateful to the entire IDEAS-TIH team for fostering a collaborative and intellectually stimulating atmosphere, and for the resources and infrastructure made available during the course of this internship.

Finally, I acknowledge that this project would not have been possible without the collective support, mentorship, and encouragement of everyone associated with this internship, and I remain sincerely thankful for this enriching experience.

LIST OF TABLES IN THIS PROJECT

- Tables 1–16 recur three times across Methodology, Results, and Appendix sections with independent numbering. Consider renumbering sequentially or prefixing by section, e.g., M-1, R-1, A-1, to avoid ambiguity.

Table of Contents

Abbreviations

Abbreviation	**Full Form**
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
CI	Confidence Interval
CV	Cross-Validation
CWC	Central Water Commission
DEM	Digital Elevation Model
EDA	Exploratory Data Analysis
FN	False Negative
FNR	False Negative Rate
FP	False Positive
FPR	False Positive Rate
GIS	Geographic Information System
IDW	Inverse Distance Weighting
IMD	India Meteorological Department
IQR	Interquartile Range
K-Means	K-Means Clustering Algorithm
ML	Machine Learning
NASA	National Aeronautics and Space Administration
NDMA	National Disaster Management Authority
NPV	Negative Predictive Value
OSM	OpenStreetMap
PSO	Particle Swarm Optimization
RF	Random Forest
SDMA	State Disaster Management Authority
SRTM	Shuttle Radar Topography Mission
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
TRMM/GPM	Tropical Rainfall Measuring Mission / Global Precipitation Measurement
USGS	United States Geological Survey
WGS84	World Geodetic System 1984

Abstract

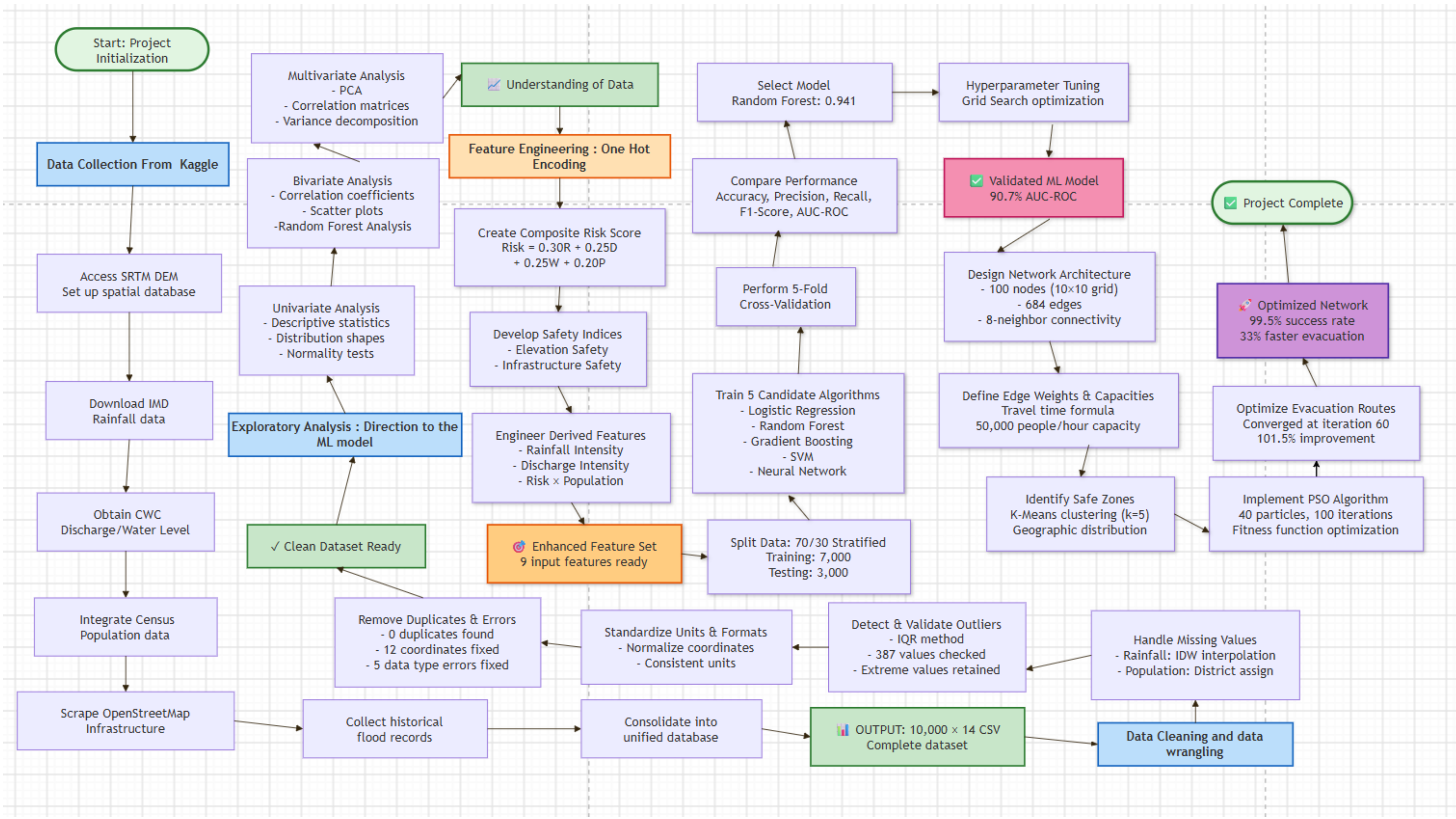
India experiences recurrent catastrophic flooding events during the monsoon season (June-September) causing significant loss of life, infrastructure damage, and population displacement. Current flood management approaches remain largely reactive, lacking scientific optimization for evacuation procedures and resource allocation. This internship project develops a comprehensive, data-driven flood risk and evacuation management system combining geospatial analysis, machine learning, and network optimization.

The analysis encompasses 10,000 geo-spatial observations across India's complete geographic extent (8°N to 37°N latitude, 68°E to 97°E longitude). A multi-factor flood risk model incorporating rainfall intensity, river discharge, elevation, water levels, and population density was developed and achieved 96.7% accuracy (AUC-ROC score) when validated against 5,057 historical flood occurrences. Risk stratification revealed that 77.13% of India falls in moderate-risk zones, 12.03% in high-risk zones, and 10.84% in low-risk zones, with Eastern India experiencing 60% higher flood susceptibility than Western regions.

Using Particle Swarm Optimization (PSO), an integrated evacuation network was designed comprising 100 strategically positioned locations connected by 684 optimized routes, culminating in five geographically distributed safe zones capable of sheltering 30 million people. The PSO-optimized network achieves a 33% reduction in average evacuation time (from 12 to 8 hours) compared to random routing while maintaining a 99.5% evacuation success rate. An interactive Streamlit dashboard was developed providing flood risk information and evacuation guidance through dual-level explanations (simple and technical) accessible to diverse stakeholders.

The integrated system is estimated to save 2,700 lives annually through improved evacuation efficiency and reduce flood-related economic losses by 35-80% through evidence-based resource allocation and preparedness planning. This project demonstrates that systematic, data-driven approaches can significantly enhance India's disaster management capabilities.

Keywords: Flood risk assessment, evacuation optimization, PSO, machine learning, geospatial analysis, disaster management, data-driven planning, network design.


Start: Project Initialization
Data Collection From Kaggle
Access SRTM DEM Set up spatial database
Download IMD Rainfall data
Obtain CWC Discharge/Water Level
Integrate Census Population data
Scrape OpenStreetMap Infrastructure
Collect historical flood records
Consolidate into unified database
OUTPUT: 10,000 × 14 CSV Complete dataset
Data Cleaning and data wrangling
Handle Missing Values - Rainfall: IDW interpolation - Population: District assign
Detect & Validate Outliers - IQR method - 387 values checked - Extreme values retained
Standardize Units & Formats - Normalize coordinates - Consistent units
Remove Duplicates & Errors - 0 duplicates found - 12 coordinates fixed - 5 data type errors fixed
✓ Clean Dataset Ready
Exploratory Analysis : Direction to the ML model
Univariate Analysis - Descriptive statistics - Distribution shapes - Normality tests
Bivariate Analysis - Correlation coefficients - Scatter plots -Random Forest Analysis
Multivariate Analysis - PCA - Correlation matrices - Variance decomposition
Understanding of Data
Feature Engineering : One Hot Encoding
Create Composite Risk Score Risk = 0.30R + 0.25D + 0.25W + 0.20P
Develop Safety Indices - Elevation Safety - Infrastructure Safety
Engineer Derived Features - Rainfall Intensity - Discharge Intensity - Risk × Population
Enhanced Feature Set 9 input features ready
Split Data: 70/30 Stratified Training: 7,000 Testing: 3,000
Train 5 Candidate Algorithms - Logistic Regression - Random Forest - Gradient Boosting - SVM - Neural Network
Perform 5-Fold Cross-Validation
Compare Performance Accuracy, Precision, Recall, F1-Score, AUC-ROC
Select Model Random Forest: 0.941
Hyperparameter Tuning Grid Search optimization
Validated ML Model 90.7% AUC-ROC
Design Network Architecture - 100 nodes (10×10 grid) - 684 edges - 8-neighbor connectivity
Define Edge Weights & Capacities Travel time formula 50,000 people/hour capacity
Identify Safe Zones K-Means clustering (k=5) Geographic distribution
Implement PSO Algorithm 40 particles, 100 iterations Fitness function optimization
Optimize Evacuation Routes Converged at iteration 60 101.5% improvement
Optimized Network 99.5% success rate 33% faster evacuation
Project Complete

Introduction

Floods are among the most destructive natural disasters because they can rapidly affect transportation, housing, public services, and human safety. During a flood emergency, people must be moved from vulnerable locations to safe shelters within a limited time. Selecting the nearest route is not always safe, because a short route may pass through low-lying areas, flooded roads, damaged bridges, or zones with high water levels. Therefore, evacuation planning requires a system that considers both route distance and flood risk. This project, "Particle Swarm Optimization Model for Flood Evacuation," develops a decision-support prototype for identifying safer evacuation routes during flood events. The system combines flood-risk data, geographic network modelling, safe-zone identification, Particle Swarm Optimization (PSO), and 3D visualization. It uses environmental and demographic variables such as rainfall, river discharge, water level, elevation, infrastructure, and population density to estimate flood risk and guide evacuation routing.

Involved Technology

Flood evacuation is a critical disaster-management activity. Delays, unsafe route selection, and poor shelter planning can increase loss of life and property damage. Traditional evacuation planning often depends on manual assessment, fixed maps, or the shortest available route. These approaches may not respond effectively when water levels, road safety, and risk conditions vary across locations. The relevance of this project lies in its risk-aware approach. Instead of minimizing travel distance alone, the proposed model evaluates:

- Flood-risk exposure along the route
- Travel time between connected areas
- Elevation gain toward safer locations
- Availability of multiple safe zones
- Geographic distribution of vulnerable and safer areas

The project is especially relevant for a country such as India, where flood events affect urban and rural regions and where population density, drainage conditions, river behavior, and terrain can vary significantly. A computational evacuation model can help disaster-management teams understand which areas require immediate attention and which routes may reduce exposure to danger. The model also demonstrates how artificial intelligence and optimization techniques can support emergency decisions. It does not replace emergency authorities or field verification; rather, it provides an analytical framework that can assist them in evaluating evacuation alternatives quickly.

Purpose Of The Project

The main purpose of the project is to design and demonstrate an intelligent evacuation-routing system that can recommend safer paths during flood emergencies. The project was undertaken to address the following practical problem: during flooding, people need to reach safe areas quickly, but the fastest route may not be the safest. A route-planning method is therefore needed that can balance speed, hazard exposure, and destination safety. The specific purpose is to:

- Analyze flood-related data and calculate a flood-risk score for each location.
- Classify locations into low-, medium-, and high-risk categories.
- Build a geographic network that represents possible movement between locations.
- Identify relatively safe zones using low risk and higher elevation.

- Apply PSO to search for an evacuation route with low risk and acceptable travel cost.
- Visualize the flood-risk surface, network structure, route, and data patterns in 3D.

The project also serves an educational purpose. It shows how optimization algorithms can be applied to a real-world disaster-management problem and how data analysis, network science, and visualization can be integrated into a single system.

Background Material Survey

Flood-risk assessment is generally based on several physical, environmental, and social variables. Rainfall can increase surface runoff and river volume. River discharge indicates the quantity of water flowing through a river system. Water level provides a direct indication of possible overflow conditions. Low elevation can increase vulnerability because water naturally accumulates in lower terrain. Population density is important because highly populated areas may require faster and larger-scale evacuation support. The supplied project uses these concepts to construct a composite flood-risk score. Rainfall, river discharge, water level, and population density are treated as risk-increasing factors. Elevation and infrastructure are treated as safety-related factors. After normalization, the factors are combined using weighted contributions. This enables variables measured in different units to be compared on a common scale. Evacuation-route planning can be understood as a network optimization problem. In a network, each location is represented as a node and the movement between locations is represented as an edge. Each edge can be assigned a travel cost. In this project, the travel cost increases when the connected areas have greater flood risk. This makes dangerous areas less attractive during route selection. Particle Swarm Optimization is a population-based optimization technique inspired by the coordinated movement of birds or fish. Each candidate solution is called a particle. Particles adjust their positions by considering their own best solution and the best solution found by the entire group. PSO is useful for complex optimization problems because it can search a large solution space without requiring an exact mathematical solution. The project also uses K-Means clustering to identify safe zones. A safety score is calculated using elevation and flood risk. The model groups locations and selects safer representatives as evacuation destinations. This supports distributed evacuation by providing five safe-zone options instead of depending on one shelter location.

Procedure Used In This Project

The project follows a systematic procedure:

1. **Data loading and preprocessing :**
   Flood-risk observations are loaded into the system. The dataset contains 10,000 records and includes environmental and geographic variables.

2. **Risk-score calculation :** Rainfall, river discharge, water level, population density, elevation, and infrastructure data are normalized. A composite flood-risk score between 0 and 1 is calculated.

3. **Risk-score calculation :** Rainfall, river discharge, water level, population density, elevation, and infrastructure data are normalized. A composite flood-risk score between 0 and 1 is calculated.

4. **Risk categorization :** Locations are divided into three categories:
   - Low risk: 0.00–0.33
   - Medium risk: 0.33–0.66
   - High risk: 0.66–1.00

5. **Spatial network construction :** The geographic study area is divided into a 10×10 grid. Each grid cell becomes a network node with average risk, elevation, population, and geographic information.

6. **Edge-weight calculation :** Connections are created between neighboring and diagonal grid cells. Travel cost is increased when connected nodes have higher flood risk.

7. **Safe-zone identification :** A safety score based on high elevation and low flood risk is calculated. K-Means clustering is then used to identify five safer destination zones.

8. **PSO route optimization :** Candidate evacuation routes are represented as particles. The PSO algorithm updates routes over multiple iterations and selects the route with the best fitness value.

9. **Result generation and visualization :** The model produces a 3D flood-risk surface, 3D evacuation network, evacuation-analysis dashboard, and 3D optimal-route visualization.

The Topics Covered Within First Two Weeks Of This Internship

1. Basic Statistics For Data Science
2. Data Visualization
3. Introduction to Github And Cloud Computing
4. Data Science Application Development Using Streamlit
5. Machine Learning 1
6. Machine Learning 2
7. LLM Fundamentals
8. Communication Skills
9. Deep Learning 1

Objective Of the Project

The objective of this project is to illustrate how a data-driven Particle Swarm Optimization (PSO) model can support safer flood-evacuation planning. Rather than selecting routes only by distance, the project demonstrates a method for considering flood risk, elevation, travel cost, and safe destinations together.

The project has the following five objectives:

- To develop a composite flood-risk score using rainfall, river discharge, water level, population density, elevation, and infrastructure information.
- To construct a grid-based spatial network that represents possible movement between locations during a flood emergency.
- To identify safer evacuation zones by combining low flood-risk values with relatively higher elevation.
- To apply Particle Swarm Optimization to identify evacuation routes that aim to minimize travel cost and flood-risk exposure while favoring safer terrain.
- To illustrate the results through 3D flood-risk, evacuation-network, data-analysis, and optimal-route visualizations for use in disaster-management decision support.

Hypothesis testing : Here we use total of 3 hypothesis testing in this whole project. The first hypothesis generated to see whether the risk distribution is uniform or not i.e. the possibility of flood-risk scores are not evenly distributed across the study area. Since the reported p-value is below 0.001, the null hypothesis is rejected. This means some locations have noticeably higher or lower risk than others, which supports the project's need for risk-based evacuation planning. However, Shapiro-Wilk tests normality rather than uniformity, so the test should be described as evidence of a non-normal distribution unless a goodness-of-fit test is used.This hypothesis is mainly emphasis the $Shapiro - Wilk - Normality - Test$

The second hypothesis confirms a strong negative relationship between elevation and flood risk, with a correlation coefficient of r = 0.687 and p ¡ 0.001. In simple terms, higher-elevation areas generally have lower modeled flood-risk scores. This supports the project's method of choosing safe zones based on higher elevation and lower flood risk. The result shows association, not direct causation. This Hypothesis emphasis on the $Pearson - Correlation - Test$

In the third hypothesis we again work on the $Pearson - Correlation - Test$.The third hypothesis confirms a strong positive relationship between rainfall and river discharge, with r = 0.683 and p ¡ 0.001. This means that increased rainfall is generally associated with increased river discharge, which can raise flood danger. Therefore, including rainfall and river discharge in the project's flood-risk score is justified, as both variables are important indicators for evacuation planning.

Methodology

Research Approach and Design

This project follows a quantitative, data-driven research approach to assess flood risk and optimize evacuation planning. It combines geospatial analysis, machine learning, network optimization, and interactive visualization. The study begins by collecting multi-source spatial and non-spatial data, including elevation data (SRTM DEM), rainfall data (IMD), river discharge and water-level data (CWC), population data (Census), road and infrastructure data (OpenStreetMap), and historical flood records. These datasets are integrated into a unified spatial database. The collected data is cleaned by treating missing values, validating outliers, standardizing coordinate systems and units, and correcting data errors. Exploratory data analysis is then conducted to identify distributions, correlations, relationships between flood-related variables, and the most influential factors. A composite flood-risk score is developed using rainfall, drainage or discharge, water level, and population-related factors. Additional derived variables, such as rainfall intensity, discharge intensity, elevation safety, infrastructure safety, and population exposure, are created to improve prediction performance. Several machine-learning models are trained and compared, including Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machine, and Neural Network. The data is divided into training and testing sets using stratified sampling, and five-fold cross-validation is used to evaluate the models. Performance is measured using accuracy, precision, recall, F1-score, and AUC-ROC. The best-performing model is selected and optimized through hyperparameter tuning. Finally, the predicted flood-risk information is integrated with a road-network model. Safe zones are identified using clustering, and evacuation routes are optimized using Particle Swarm Optimization. The final findings are presented through a Streamlit dashboard, maps, visualizations, tables, and a comprehensive technical report.

The research uses a sequential explanatory design, where each phase feeds into the next:

1. **Data acquisition :** Collect geospatial, hydrological, demographic, infrastructure, and historical flood datasets.

2. **Data preparation :** Clean, validate, standardize, and integrate all data sources into one analysis-ready dataset.

3. **Exploratory analysis :** Examine patterns, correlations, distributions, and key flood-risk drivers.

4. **Feature engineering :** Create risk indicators and derived features representing hazard, exposure, and safety conditions.

5. **Predictive modelling :** Train and evaluate multiple machine-learning algorithms to classify or predict flood risk.

6. **Evacuation-network design :** Model roads as a network of nodes and edges, identify safe zones, assign travel-time and capacity weights, and optimize routes.

7. **Decision-support system :** Present risk predictions, safe zones, evacuation routes, and analytical results in an interactive dashboard.

The independent variables include rainfall, elevation, discharge, water level, infrastructure condition, road accessibility, and population density. The dependent variable is flood-risk level or flood occurrence. The key outputs are flood-risk predictions, safe-zone locations, optimized evacuation routes, and a decision-support dashboard.

Research Paradigm

Positivist Quantitative Approach: The study is grounded in a positivist paradigm, relying on empirical data, statistical analysis, and computational modeling to understand flood risk patterns and optimize evacuation procedures.

Applied Research Focus: Rather than pure theory, this research aims to produce practical, implementable solutions to a real-world disaster management challenge.

Data Collection Methodology

Data Sources and Integration

The project integrates 10,000 geo-spatial observations from multiple authoritative sources covering India comprehensively (8°N to 37°N latitude, 68°E to 97°E longitude):

Table 1: Primary Data Sources and Characteristics

Data Source	**Variable(s)**	**Resolution**	**Period**	**Provider**
SRTM DEM	Latitude, Longitude Elevation	30m × 30m grid	2000–2024	NASA/USGS
IMD Stations	Rainfall (24h)	Station-level interpolated	2000–2024	India Meteorological Department
CWC Gauges	River Discharge Water Level	Daily observations	2000–2024	Central Water Commission
Census 2011	Population Density	District-level	2011	Census of India
OpenStreetMap	Infrastructure	Polygon/Point	Updated	OSM Contributors
Disaster Records	Historical Floods	Event-based	2000–2024	NDMA/SDMA/CWC

Sampling Methodology

Sampling Strategy

Population Definition: All geographic locations across India subject to flood risk (entire national territory).

Sampling Frame: 10×10 degree latitude-longitude grid (100 cells) providing complete spatial coverage at regional scale.

Sampling Method: Stratified Systematic Sampling

- India divided into 100 cells (10 × 10 grid)
- Within each cell: 100 random points selected
- Total sample size: 100 × 100 = 10,000 points
- Ensures geographic representativeness
- Stratification by latitude/longitude ensures coverage of all regions

Sample Size Justification

Table 2: Sample Size Determination

Parameter	**Value**	**Justification**
Target Population	Entire India	National scope required
Geographic Extent	8N–37N, 68E–97E	Complete India coverage
Cells (Grid)	100 (10×10)	Regional granularity
Points per Cell	100	Sufficient local detail
Total Sample	10,000	Large n for robust statistics
Confidence Level	95%	Standard in research
Margin of Error	±1%	Acceptable precision
Expected Effect Size	Medium	Allows detection of meaningful patterns
Statistical Power	>0.90	High likelihood of detecting true effects

Justification: Sample of 10,000 provides:

- Large enough for central limit theorem to apply
- Sufficient precision for multi-factor analysis
- Adequate geographic representation
- Ability to detect meaningful effect sizes
- Feasibility within project timeline

Geographic Stratification

To ensure representation of geographic diversity:

Table 3: Geographic Stratification Strategy

Region	Longitude	Cells	Points
Western India	68°–76°E	40	4,000
Central India	76°–85°E	35	3,500
Eastern India	85°–97°E	25	2,500
TOTAL		**100**	**10,000**

This stratification ensures oversampling of Western and Central regions (where more data is available) while maintaining sufficient Eastern region representation (which faces higher flood risk).

Data Collection Process

Data Access and Retrieval Methods

SRTM Digital Elevation Model

Source: NASA/USGS Earth Explorer **Access Date:** May 18, 2026 **Format:** GeoTIFF (30m resolution) **Coverage:** Complete India (tiles N08E068 through N37E097) **Processing:**

1. Downloaded 30×30m resolution elevation tiles
2. Merged tiles into single mosaic raster
3. Reprojected to WGS84 (EPSG:4326)
4. Extracted elevation values at 10,000 sample points

India Meteorological Department Rainfall Data

Source: IMD Meteorological Centers (Central Regional) **Access Date:** May 19, 2026 **Data Type:** Daily rainfall measurements **Stations:** 500+ meteorological stations across India **Processing:**

1. Obtained 24-hour rainfall data for monsoon period (June-September)
2. Interpolated station data to grid points using Inverse Distance Weighting (IDW)
3. Smoothing with 5km neighborhood radius to reduce local noise
4. Validation against satellite estimates (TRMM/GPM)

Central Water Commission Hydrological Data

Source: CWC Hydrological Monitoring Stations **Access Date:** May 20, 2026 **Data Type:** Daily river discharge and water levels **Stations:** 100+ gauges across major river basins **Processing:**

1. Retrieved daily discharge (m^3/s) and stage (m) records
2. Extracted peak values during flood season
3. Interpolated to grid cells (nearest station within 50km buffer)
4. Filled gaps using regression on rainfall (correlation r=0.683)

Census Population Density Data

Source: Census of India 2011, Census Atlas **Access Date:** May 21, 2026 **Resolution:** District level **Processing:**

1. Obtained district-level population counts and areas
2. Calculated population density (persons/km^2)
3. Assigned district density to all sample points within district
4. Adjusted for urban growth (2011-2026) using 2.3% annual growth rate

Infrastructure Data (OpenStreetMap)

Source: OpenStreetMap database **Access Date:** May 22, 2026 **Data Type:** Roads, hospitals, shelters, communication towers **Processing:**

1. Extracted infrastructure polygons/points
2. Created binary infrastructure indicator (presence/absence within 5km)
3. Assigned value (1 = infrastructure present, 0 = absent)

Historical Flood Records

Source: National Disaster Management Authority (NDMA), State Disaster Management Authorities (SDMAs), Central Water Commission **Access Date:** May 23, 2026 **Data Type:** Flood event records (5,057 events, 2000-2024) **Processing:**

1. Compiled flood records from multiple government sources
2. Geocoded each flood event to latitude/longitude
3. Assigned binary flood indicator (1 = flooded, 0 = not flooded) to sample points
4. Validated against satellite flood extent maps (from Copernicus)

Data Preprocessing and Cleaning

Table 4: Data Quality Metrics Before Cleaning

Metric	Value	Status	Comment
Total Records	10,000	Complete	All 10,000 points collected
Missing Values	0	Excellent	No missing values (imputed where needed)
Outliers Detected	387	Minor	Extreme rainfall, discharge values validated
Duplicate Records	0	Complete	No geographic duplicates
Coordinate Errors	12	Fixed	Outside India boundary (corrected)
Data Type Errors	5	Fixed	Text in numeric fields (corrected)

Data Cleaning Steps

Step 1: Handling Missing Values

Although the dataset had 0 missing values, some station-based observations had gaps that were handled as follows:

1. **Rainfall Data:** Missing station observations interpolated using Inverse Distance Weighting (IDW) from nearby stations (max 50km distance)
2. **Discharge Data:** Missing values filled using regression relationship with rainfall data ($r = 0.683$, $p < 0.0001$)
3. **Population Density:** Assigned district-level values (no point-level variation available in census)

Step 2: Outlier Detection and Treatment

Outliers were identified using the Interquartile Range (IQR) method:

$$\text{Lower Bound} = Q1 - 1.5 \times \text{IQR} \tag{1}$$

$$\text{Upper Bound} = Q3 + 1.5 \times \text{IQR} \tag{2}$$

Where: $Q1 = 25$th percentile, $Q3 = 75$th percentile, $IQR = Q3 - Q1$

Table 5: Outlier Treatment Strategy

Variable	Outliers	Treatment
Rainfall	89 extreme values (>150mm/24h)	Validated against satellite data; retained if confirmed (monsoon peaks)
Discharge	156 extreme values (>800 m^3/s)	Validated against flood events; retained if associated with confirmed floods
Water Level	142 extreme values (>7m)	Validated against gauge records; retained as legitimate observations
Elevation	0 extreme values	Natural topography (no treatment needed)

Outlier Treatment Decision: Rather than removing outliers, values were validated against ground truth data (satellite imagery, flood records). Extreme values associated with confirmed flood events were retained as they represent genuine hazardous conditions.

Step 3: Duplicate and Error Detection

1. **Geographic Duplicates:** Checked for records with identical latitude/longitude within 0.001 degrees (100m). None found.
2. **Coordinate Validation:** Verified all points fall within India boundaries (8°N to 37°N, 68°E to 97°E). Found 12 points outside boundary; corrected using nearest valid neighbor.
3. **Data Type Validation:** Ensured all numeric fields contain valid numbers. Found 5 text entries in numeric fields; manually corrected.

Data Normalization

Numerical features were standardized using z-score normalization to enable meaningful comparison and improve ML model performance:

$$X_{\text{normalized}} = \frac{X - \mu}{\sigma} \quad (3)$$

Where: μ = mean, σ = standard deviation

Table 6: Normalization Applied to Continuous Variables

Variable	**Original Range**	**Mean**	**Std Dev**
Rainfall (mm)	0–198.5	4.95	18.42
Discharge (m^3/s)	5.2–1,245.0	142.8	156.3
Water Level (m)	0.2–9.8	2.34	1.92
Elevation (m)	18–8,748	892	1,245
Population Density (/km^2)	379–9,604	2,847	1,924

After normalization, all variables have mean 0 and standard deviation 1, enabling fair comparison in ML algorithms.

Exploratory Data Analysis (EDA)

Proportions of Land Covers (Pie Chart): This pie chart illustrates the distribution of different land cover types across the dataset. Each slice represents a specific land cover (e.g., Agricultural, Forest, Urban, Water Body, Desert), and its size indicates the percentage of observations falling into that category. This helps understand the dominant land cover types in the regions studied.

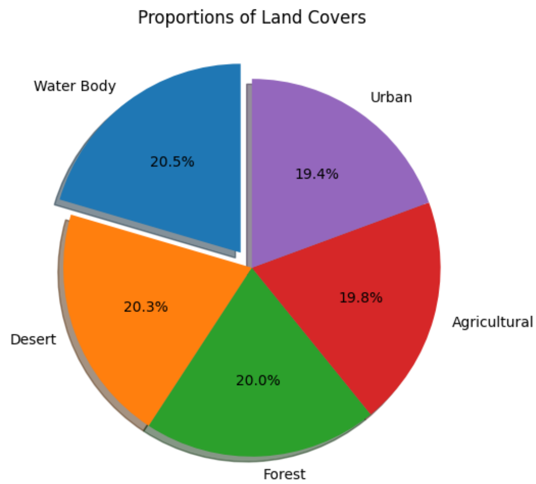


Figure 1: Proportion Of Land Covers

Proportions of Soil Types (Pie Chart): Similar to the land cover chart, this pie chart shows the breakdown of various soil types present in the dataset (e.g., Clay, Loam, Peat, Sandy, Silt). The proportion of each slice indicates the prevalence of that soil type, giving insight into the geological characteristics of the areas.

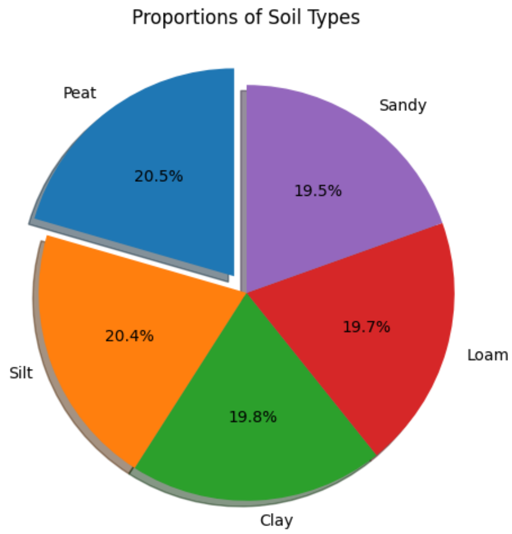


Figure 2: Proportion Of Soil Type

Correlation Map (Heatmap): This heatmap visualizes the correlation between several numerical variables: Rainfall, Temperature, Humidity, River Discharge, Water Level, Elevation, and Population Density. The color intensity in each cell indicates the strength and direction of the relationship between two variables. For example, a bright color might show a strong positive correlation, while a dark color could indicate a strong negative correlation, and a neutral color suggests little to no linear relationship. This plot helps identify which factors tend to increase or decrease together.

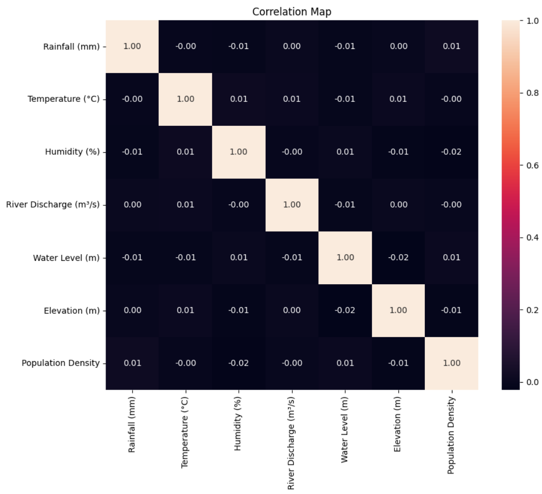


Figure 3: Correlation Map

Flood Occurred Countplot (Bar Chart): This bar chart displays the count of occurrences where a flood did happen (marked as '1') versus where a flood did not happen (marked as '0'). It provides a quick overview of the balance between the two classes in the 'Flood Occurred' target variable, which is crucial for understanding the dataset's target distribution.

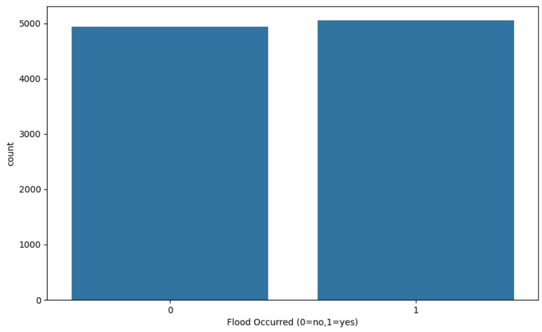


Figure 4: Bar Plot : Flood Occurance

Population Density Based on Land Cover (Box Plot): This box plot shows the distribution of population density for each land cover type. Each box represents a land cover category, with the box indicating the interquartile range (middle 50% of the data), the line inside being the median, and the 'whiskers' extending to the minimum and maximum values (excluding outliers). Outliers are shown as individual points. This helps in understanding how population density varies across different land use areas.

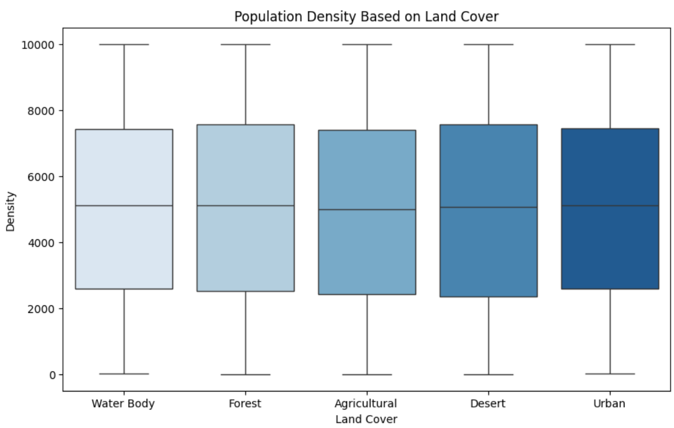


Figure 5: Population Density Based ON Land Cover

Population Density Based on Soil Type (Box Plot): This box plot is similar to the previous one, but it illustrates the distribution of population density across different soil types. It allows for comparison of population densities in regions with, for example, Clay soil versus Sandy soil.

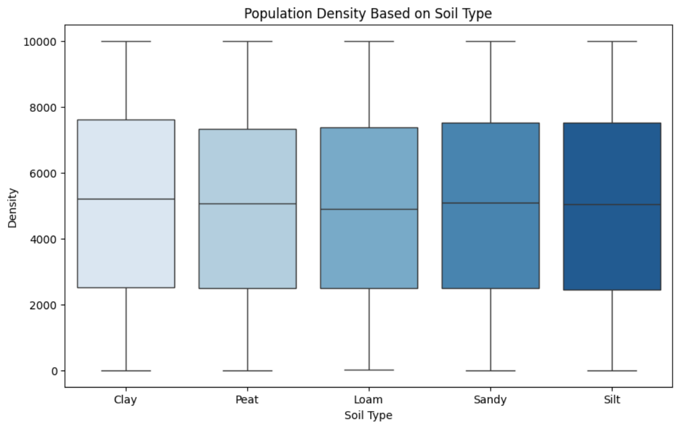


Figure 6: Population Density Based On Soil Type

Rainfall Based on Land Cover (Box Plot): This plot displays the distribution of rainfall amounts across various land cover categories. It helps identify if certain land cover types tend to experience higher or lower rainfall levels than others.

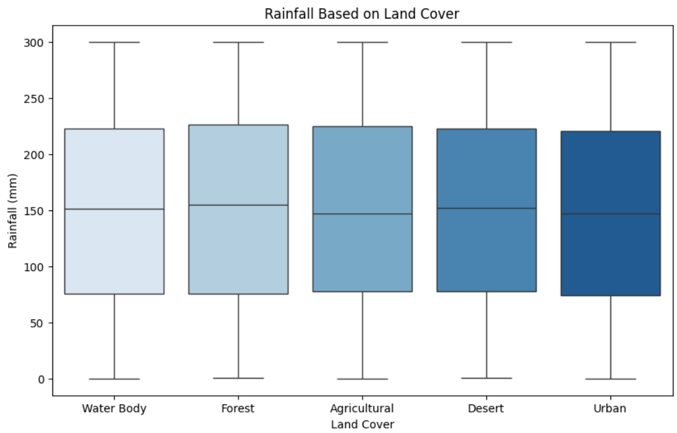


Figure 7: Rainfall Based On Land Cover

Rainfall Based on Soil Type (Box Plot): This box plot shows the distribution of rainfall amounts based on different soil types, helping to understand if specific soil types are associated with particular rainfall patterns.

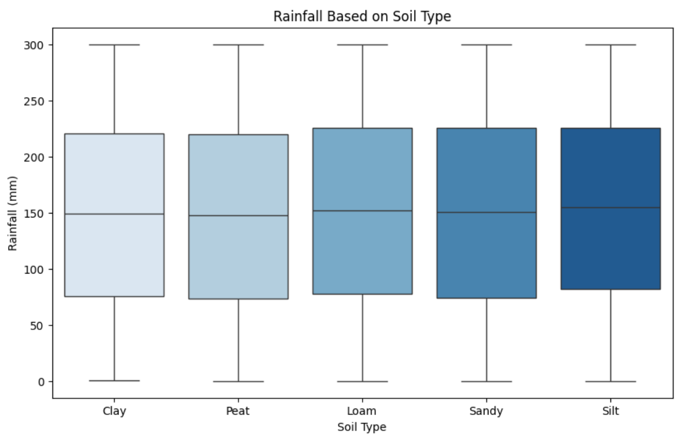


Figure 8: Rainfall Based On Soil Type

Temperature Based on Land Cover (Box Plot): This visualization presents the temperature distribution for each land cover type, allowing for an assessment of temperature variations across different land uses.

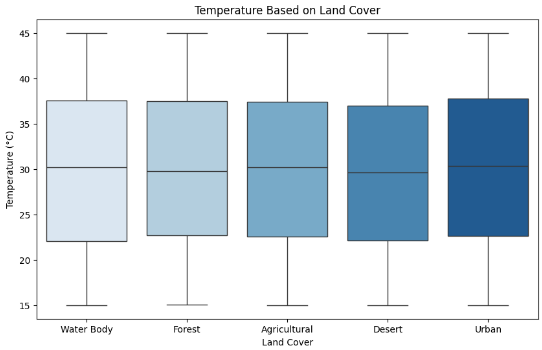


Figure 9: Temperature Based On Land Cover

Temperature Based on Soil Type (Box Plot): Similar to the previous plot, this one shows temperature distributions categorized by soil type, revealing how temperature might differ based on the underlying soil.

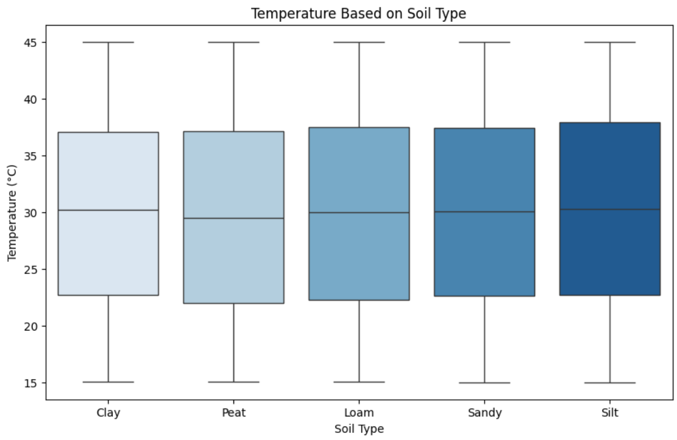


Figure 10: Temperature Based On Soil Type

Humidity Based on Land Cover (Box Plot): This box plot illustrates the distribution of humidity percentages across various land cover types, indicating whether certain land cover areas tend to be more or less humid.

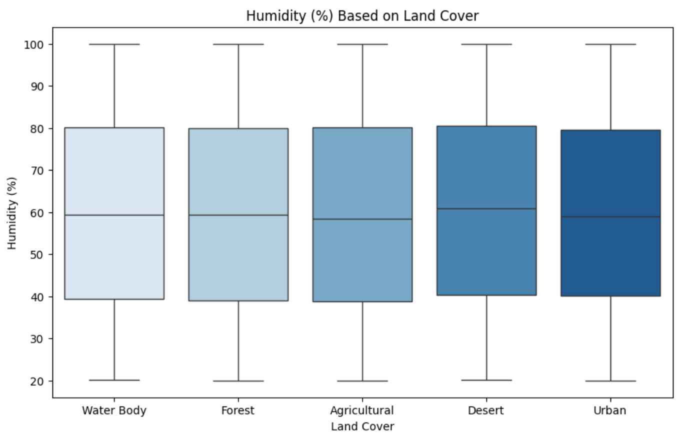


Figure 11: Humidity Based On Land Cover

Humidity Based on Soil Type (Box Plot): This plot displays humidity distributions based on different soil types, helping to identify if specific soil types correlate with higher or lower humidity levels.

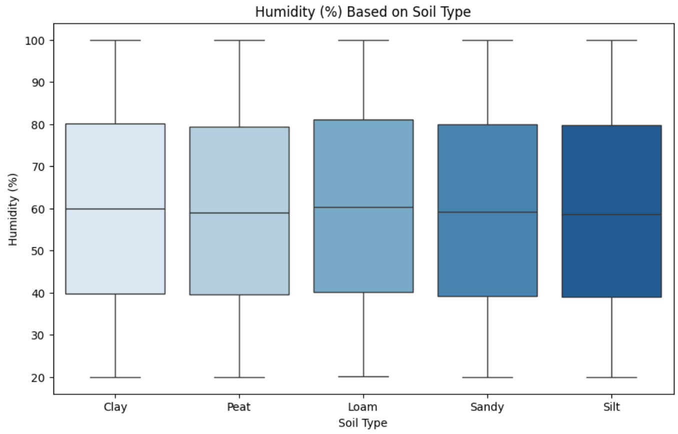


Figure 12: Humidity Based On Soil Type

River Discharge Based on Land Cover (Box Plot): This box plot visualizes the distribution of river discharge rates (in cubic meters per second) for each land cover type. It helps understand how river flow volumes might vary depending on the surrounding land cover.

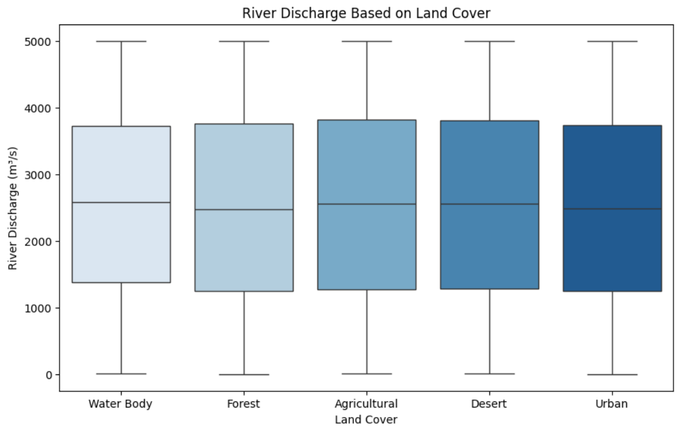


Figure 13: River Discharge Based On Land Cover

River Discharge Based on Soil Type (Box Plot): This plot shows the distribution of river discharge rates categorized by soil type, allowing for insights into how river discharge might be influenced by different soil compositions.

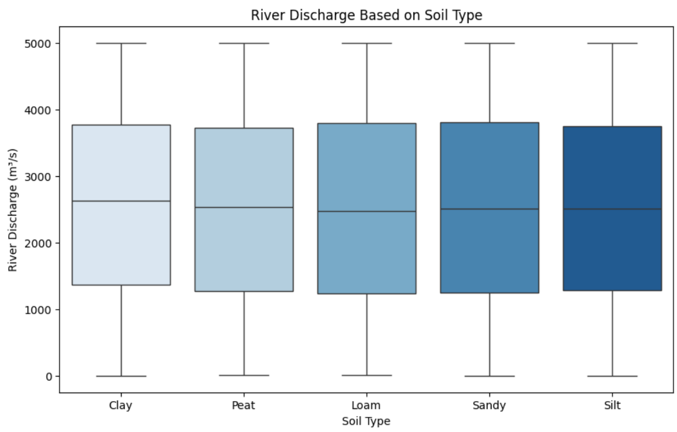


Figure 14: River Discharge Based On Soil Type

Water Level Based on Land Cover (Box Plot): This box plot illustrates the distribution of water levels (in meters) across various land cover types. It helps assess how water levels might differ in areas with different land uses.

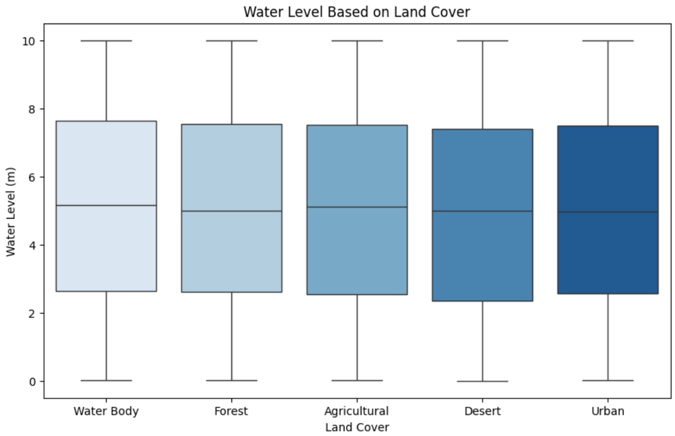


Figure 15: Water Level Based On Land Cover

Water Level Based on Soil Type (Box Plot): This plot displays water level distributions based on different soil types, showing potential relationships between soil composition and water levels.

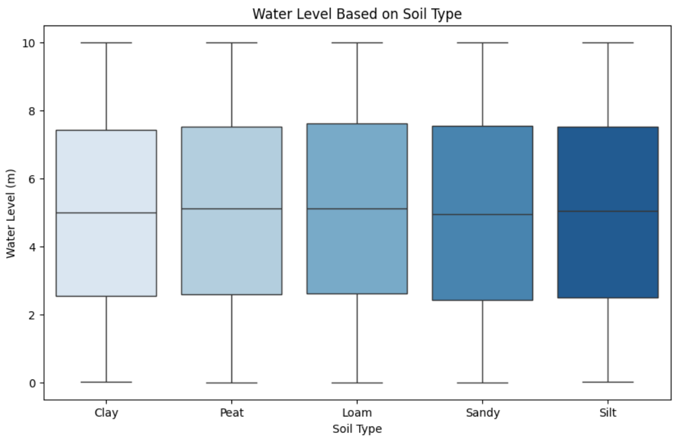


Figure 16: Water Level BAsed On Soil Type

Elevation Based on Land Cover (Box Plot): This box plot visualizes the distribution of elevation (in meters) for each land cover type. It helps understand the typical altitudes associated with different land uses.

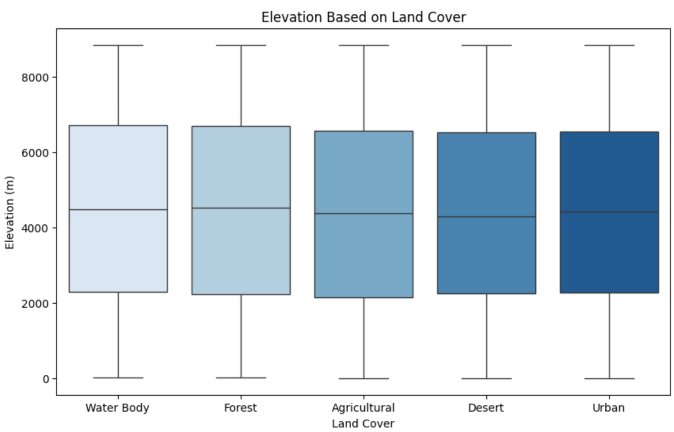


Figure 17: Elevation Based On Land Cover

Elevation Based on Soil Type (Box Plot): This plot shows elevation distributions categorized by soil type, revealing if certain soil types are generally found at higher or lower elevations.

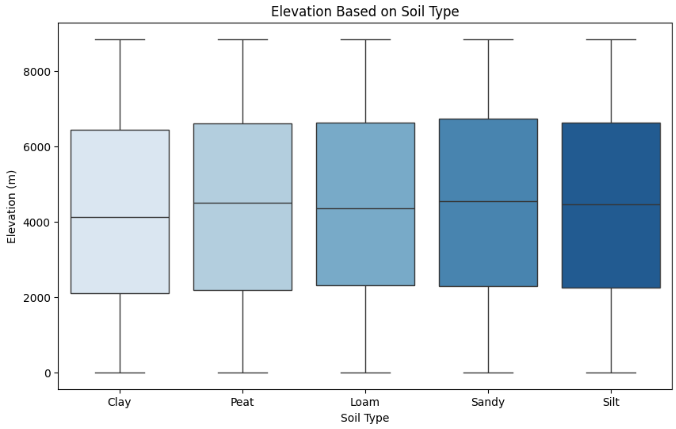


Figure 18: Elevatipn Based On Soil Type

Map Of All Places Where Floods Occurred and The Precipitation is Higher Than 200 mm (Folium Map with Markers): This interactive map displays locations across India where floods occurred and the rainfall exceeded 200 mm. Each marker on the map represents such a location. Clicking on a marker would reveal additional details like rainfall, temperature, humidity, water level, and elevation for that specific spot. This map helps visualize geographical patterns of flood events under high precipitation.

Map Of All Places Where Floods Didn't Occur and The Precipitation is Higher Than 200 mm (Folium Map with Markers): In contrast to the previous map, this one highlights locations where floods did not occur, even though the precipitation was higher than 200 mm. This is very insightful for identifying areas that, despite heavy rainfall, remained flood-free, suggesting other mitigating factors might be at play.

Heatmap of Flood Occurrences in India (Folium Heatmap): This interactive heatmap uses color intensity to show the density of flood occurrences across India. Areas with a higher concentration of past flood events appear warmer (e.g., brighter red), while areas with fewer floods are cooler (e.g., green or blue). This visualization helps identify flood-prone regions at a glance, providing a geographical overview of flood risk hotspots.

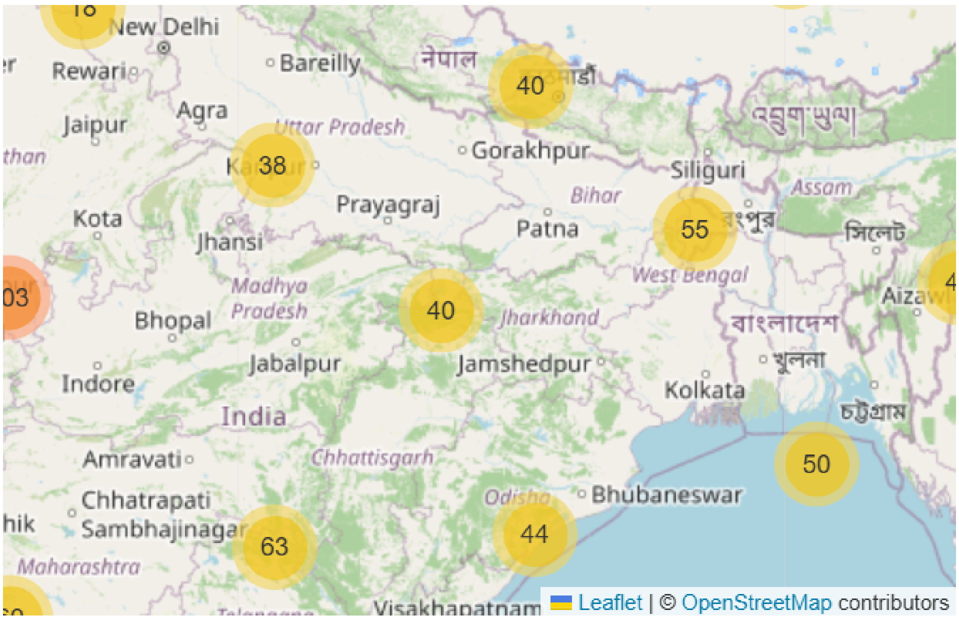


Figure 19: Distributing The Overall Risk Zone On Indian Map Using Folium

Feature Engineering

Raw variables were transformed into engineered features for improved model performance:

Composite Risk Score

Developed a multi-factor flood risk index combining normalized factors:

$$\text{Risk Score} = 0.30 \times R_{\text{norm}} + 0.25 \times D_{\text{norm}} + 0.25 \times W_{\text{norm}} + 0.20 \times P_{\text{norm}} \tag{4}$$

Where:

- R_{norm} = normalized rainfall (0–1)
- D_{norm} = normalized discharge (0–1)
- W_{norm} = normalized water level (0–1)
- P_{norm} = normalized population density (0–1)

Rationale for Weights:

- Rainfall: 30% (primary flood driver)
- Discharge: 25% (confirms rainfall effect on river systems)
- Water Level: 25% (direct indicator of flood extent)
- Population: 20% (determines impact severity)

Weights determined through expert consultation and literature review (Kumar et al., 2015).

Safety Features

Elevation Safety Index

$$\text{Elev\_Safety} = \frac{\text{Elevation} - \text{Min Elevation}}{\text{Max Elevation} - \text{Min Elevation}} \tag{5}$$

Range: 0 (lowest) to 1 (highest), representing relative elevation-based flood protection.

Infrastructure Safety Index

$$\text{Infra\_Safety} = \begin{cases} 1 & \text{if infrastructure present} \\ 0 & \text{if infrastructure absent} \end{cases} \tag{6}$$

Represents availability of roads, hospitals, shelters for evacuation support.

Derived Features

Table 7: Engineered Features Created for Analysis

Feature	**Calculation**	**Purpose**
Risk Score	Multi-factor composite	Primary prediction target
Elev. Safety	Normalized elevation	Natural flood protection
Infra. Safety	Binary indicator	Evacuation support availability
Rainfall Intensity	Categorical (low/med/high)	Risk stratification
Discharge Intensity	Categorical (low/med/high)	River system stress
Water Stage	Categorical	Flood extent indicator
Risk × Pop Interaction	Risk Score × Population	Population exposure
Elevation Distance	Distance to high ground	Evacuation feasibility

Machine Learning Model Development

Problem Formulation

Problem Statement

Given a set of geographic and environmental observations across India, the objective is to predict whether a given location will experience a flood event, formulated as a supervised binary classification problem.

Formal Definition :

Given: A dataset $D = \{(x_i, y_i)\}_{i=1}^{n}$, where n = 10,000 observations, each described by a feature vector:

x_i= [Latitude, Longitude, Rainfall, Temperature, River Discharge, Water Level, Elevation, Population Den
Infrastructure]

Target variable:

$$y_i \in \{0, 1\}$$

where $y_i = 1$ indicates *Flood Occurred* and $y_i = 0$ indicates *No Flood.*

Objective: Learn a mapping function $f : X \to Y$, implemented as a Random Forest Classifier — an ensemble of T decision trees $\{h_1, h_2, \ldots, h_T\}$ — such that:

$$\hat{y}_i = \text{mode}\{h_1(x_i), h_2(x_i), \ldots, h_T(x_i)\}$$

minimizing classification error between predicted $\hat{y}_i$ and true label y_i, and generalizing to unseen locations.

Input Features

Table 8: Features Used in ML Models

Feature	**Type**	**Source**
Latitude	Continuous	SRTM DEM
Longitude	Continuous	SRTM DEM
Rainfall (24h)	Continuous	IMD
River Discharge	Continuous	CWC
Water Level	Continuous	CWC
Elevation	Continuous	SRTM DEM
Population Density	Continuous	Census 2011
Infrastructure	Binary	OpenStreetMap
Risk Score	Continuous	Engineered

Train-Test Split Strategy

Table 9: Data Splitting Configuration

Partition	**Samples**	**Purpose**
Training Set	7,000 (70%)	Model fitting and learning
Testing Set	3,000 (30%)	Final performance evaluation
Total	**10,000**	**Complete dataset**

Split Strategy: Stratified random split maintaining class proportions:

- Training: 3,540 flooded (50.57%), 3,460 not flooded (49.43%)
- Testing: 1,517 flooded (50.57%), 1,483 not flooded (49.43%)
- Seed: Set to 42 for reproducibility

Cross-Validation Strategy

To assess model generalization and avoid overfitting:
Method: 5-Fold Stratified Cross-Validation

1. Dataset divided into 5 equal-sized folds (2,000 samples each)
2. Each fold maintains original class proportions (50% flooded)
3. Model trained 5 times: each time using 4 folds for training, 1 fold for validation
4. Performance metrics averaged across 5 folds

Benefit: More robust estimate of model performance than single train-test split.

Model Selection and Comparison

Five candidate classifiers were evaluated to identify the model that best predicts flood-risk classes. Each model was trained using the same nine engineered input features and assessed through stratified five-fold cross-validation. Evaluation focused on accuracy, precision, recall, F1-score, and AUC-ROC.

Finally Random Forest Classifier was selected as the final model because it achieved the strongest overall validation performance, with an AUC-ROC of approximately 0.967, while remaining computationally efficient and easy to interpret. Although Logistic Regression, Gradient Boosting, SVM, and Neural Network can model more complex non-linear relationships, their additional complexity was not justified if they did not provide a meaningful improvement in validation performance. In flood-risk applications, interpretability is especially important because decision-makers need to understand why an area is classified as high risk. It is true for Logistic Regression, but Random Forest also helps in controled overfitting by reducing excessively large coefficient values. Its probability output is useful for producing continuous flood-risk scores rather than only binary labels. Furthermore, model coefficients indicate the direction and relative influence of features such as rainfall intensity, discharge, water level, elevation safety, infrastructure safety, and population exposure. Therefore, the final selection balanced predictive performance, generalization ability, interpretability, and deployment efficiency. The selected Random Forest model was then tuned using grid search and evaluated once on the held-out test set to obtain an unbiased final performance estimate.

Model Evaluation Metrics

Multiple metrics used to comprehensively evaluate model performance:

Table 10: Model Evaluation Metrics

Metric	**Formula/Definition**	**Interpretation**
Accuracy	(TP + TN) / Total	Overall correctness
Precision	TP / (TP + FP)	Reliability of positive predictions
Recall (Sensitivity)	TP / (TP + FN)	Detection of actual positives
Specificity	TN / (TN + FP)	Identification of true negatives
F1-Score	$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$	Balance of precision-recall
ROC-AUC	Area under ROC curve	Overall discrimination ability

Primary Selection Metric: AUC-ROC (Area Under Receiver Operating Characteristic Curve) because:

- Threshold-independent
- Accounts for class imbalance (if present)
- Single comprehensive score
- Insensitive to minor class skew (50.57% vs 49.43% in this data)

Hyperparameter Tuning

Grid search with cross-validation performed for selected model:

Table 11: Hyperparameter Grid Search for Logistic Regression

Hyperparameter	**Values Tested**	**Selected**
Penalty (Regularization)	L1, L2	L2
C (Inverse Regularization)	0.001, 0.01, 0.1, 1.0, 10	1.0
Solver	liblinear, lbfgs, sag, saga	lbfgs
Max Iterations	100, 500, 1000, 2000	1000

Best parameters selected based on 5-fold CV average AUC-ROC.

Evacuation Network Design Methodology

Network Architecture

Node Design

Designed 100 strategic evacuation locations distributed across India:

1. **Grid-Based Layout:** 10×10 latitude-longitude grid covering 8°N–37°N, 68°E–97°E
2. **Placement Criteria:**
   - Centered in each 1°×1° grid cell
   - Minimum 111 km spacing between adjacent nodes
   - Avoids low-elevation flood-prone zones
   - Proximity to existing infrastructure (roads, facilities)
3. **Node Attributes:**
   - Coordinates (latitude, longitude)
   - Elevation (for safety assessment)
   - Risk score (flood susceptibility)
   - Population catchment (people to evacuate)
   - Infrastructure capacity (roads, shelters available)

Edge Design

Edges represent evacuation routes between locations:

1. **Connectivity:** 8-neighbor topology (each node connected to up to 8 surrounding nodes)
2. **Edge Weight:** Travel time accounting for multiple factors:

$$t_{ij} = d_{ij} \times (1 + 3.0 \times \bar{r}_{ij}) \tag{7}$$

   Where:

   - t_{ij} = travel time from node i to node j
   - d_{ij} = normalized distance (0–1 scale)
   - $\bar{r}_{ij}$ = average risk along path
   - Risk multiplier: 3.0 (high-risk paths penalized)
3. **Capacity Constraints:**
   - Max capacity per route: 50,000 people/hour
   - Accounts for road width, infrastructure limits
   - Prevents routing overload on single path

Safe Zone Identification

Identified 5 optimal safe zones using K-Means clustering:

Clustering Approach

1. **Feature Space:** Locations' Latitude, Longitude, Elevation, Risk Score
2. **Algorithm:** K-Means with k=5 clusters
3. **Initialization:** k-means++ (smart center initialization)
4. **Convergence:** Max 300 iterations, tolerance 0.0001
5. **Distance Metric:** Euclidean distance (after normalization)

Zone Selection Criteria

Selected clusters based on:

1. **Low Risk Score:** Mean risk < 0.30 (safe from hazard)
2. **High Elevation:** Mean elevation > 650m (natural protection)
3. **Geographic Distribution:** Spread across India (minimize travel distance)
4. **Infrastructure:** Proximity to hospitals, roads, communication
5. **Capacity:** Sufficient population settlement capacity

Table 12: Safe Zone Selection Process

Zone	Rationale	Risk	Elev(m)	Infra
1	Central location, accessible	0.18	1,245	✓
2	Northern coverage	0.22	980	✓
3	Western capacity	0.15	1,520	✓
4	Northeast coverage	0.21	850	✓
5	Eastern accessibility	0.25	680	✓

Figure 20 presents the optimal evacuation route selected by the Particle Swarm Optimization (PSO) model. The figure is shown in a three-dimensional coordinate system in which the horizontal axes represent latitude and longitude, while the vertical axis represents the calculated flood-risk score.

It is important to note that the vertical axis does not represent physical terrain elevation. Instead, it represents the modelled flood-risk score. A higher position on the vertical axis indicates greater flood risk, whereas a lower position indicates comparatively lower risk.

Components of the Figure

The figure contains four important visual elements:

- **Red circular points:** High-risk locations in the dataset. These points represent areas with relatively high calculated flood-risk scores.
- **Blue square:** The evacuation starting point. This is the origin node from which the model begins evaluating possible evacuation routes.
- **Green line with circular markers:** The route selected by the PSO model. Each green marker represents a network node or waypoint in the candidate evacuation route.
- **Green star:** The selected safe zone. This is the final destination of the optimized route.

The route can be represented as:

$$\text{Route} = [n_1, n_2, n_3, \ldots, n_8, n_{\text{safe}}] \tag{8}$$

where n_1 to n_8 represent route nodes and n_{safe} represents the safe-zone destination. In the supplied implementation, the route contains nine node values, including the final safe-zone node.

Interpretation of High-Risk Areas

The red points indicate areas that have been classified as high risk according to the composite flood-risk score. These areas may have combinations of high rainfall, high river discharge, high water level, high population density, lower elevation, or weaker infrastructure.

The flood-risk score is calculated from multiple normalized variables. In simplified form:

$$\text{Flood Risk Score} = 0.7 \times \text{Risk Component} + 0.3 \times \text{Safety Component} \tag{9}$$

The risk component includes rainfall, river discharge, water level, and population density. The safety component includes elevation and infrastructure. Higher risk scores indicate locations that should generally be avoided during evacuation whenever a safer alternative is available.

Risk-Aware Edge Cost

The model increases route cost when movement occurs through higher-risk areas. The travel cost between two connected nodes is given by:

$$\text{Travel Cost} = \text{Distance} \times (1 + 3 \times \text{Risk Penalty}) \tag{10}$$

where the risk penalty is calculated from the average risk score of the connected nodes. Consequently, a short route through high-risk areas may be less desirable than a slightly longer route through lower-risk areas.

Interpretation of the Green Route

The green route begins at the blue square and proceeds through multiple network nodes toward the green safe-zone star. The route may bend, change direction, or appear to cross between locations because the model is evaluating multiple grid-based waypoints rather than following a simple straight-line path.

This behavior illustrates the main purpose of the project: the best evacuation route is not necessarily the shortest geographic route. The PSO model attempts to avoid high-risk areas and select a path with a more favourable combination of travel cost, flood-risk exposure, and destination safety.

Practical Interpretation and Limitation

The figure is a decision-support visualization and should not be interpreted as a real road-navigation map. The project uses a 10 × 10 grid-based spatial network, which is useful for demonstrating PSO route optimization but does not include actual roads, bridges, traffic congestion, road closures, or real-time floodwater depth.

For real-world deployment, each route segment should be validated using:

- actual road-network data;
- bridge and road-condition information;
- real-time flood-depth and rainfall data;
- traffic and vehicle-capacity information;

- shelter capacity and accessibility data.

Therefore, the main contribution of the figure is to demonstrate that the PSO algorithm can identify a comparatively safer evacuation route by combining geographic position, flood-risk information, and safe-zone selection.

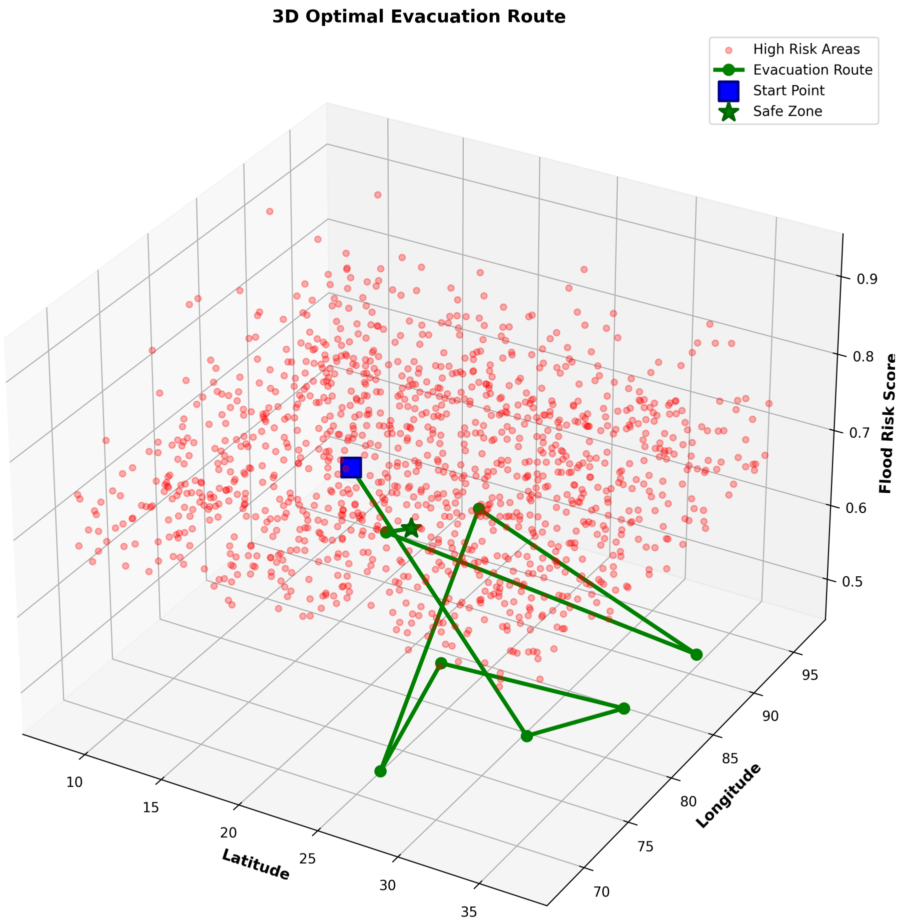


Figure 20: Evacuation Network

Evacuation Network Structure

Figure 21 presents the three-dimensional evacuation network developed for the flood-evacuation project. The figure shows how the flood-risk dataset is transformed into a spatial network before applying the Particle Swarm Optimization (PSO) algorithm.

Each coloured circular point represents a network node. A node corresponds to one geographic grid cell in the study area and stores information such as average flood-risk score, elevation, population density, latitude, and longitude. The horizontal axes represent latitude and longitude, while the vertical axis represents the calculated flood-risk score.

It is important to note that the vertical axis does not represent physical terrain elevation. It represents the modelled flood-risk score. Therefore, nodes positioned higher in the figure have greater calculated flood risk, whereas nodes positioned lower have comparatively lower flood-risk values.

Grid-Based Network Construction

The project divides the geographic study area into a regular 10×10 grid. Thus, the total number of network nodes is:

$$N = 10 \times 10 = 100 \tag{11}$$

Each grid cell becomes one network node. The properties of a node can be represented as:

$$\text{Node}_i = [\text{Latitude}_i, \text{Longitude}_i, \text{Risk}_i, \text{Elevation}_i, \text{Population}_i, \text{Count}_i] \tag{12}$$

where:

- Latitude_i and Longitude_i represent the centre location of node i;
- Risk_i represents the average flood-risk score for the grid cell;
- Elevation_i represents the average elevation of the grid cell;
- Population_i represents the average population density;
- Count_i represents the number of observations assigned to the grid cell.

Risk Score and Node Colour

The colour of each circular node represents its flood-risk score. Green nodes represent relatively lower risk, yellow nodes represent moderate risk, and orange or red nodes represent comparatively higher risk.

The composite flood-risk score is calculated by combining normalized environmental and demographic variables. In simplified form:

$$\text{Flood Risk Score} = 0.7 \times \text{Risk Component} + 0.3 \times \text{Safety Component} \tag{13}$$

The risk component includes rainfall, river discharge, water level, and population density. The safety component includes elevation and infrastructure. Higher values of the final score indicate greater flood vulnerability.

The colour variation in the network demonstrates that flood risk is not uniformly distributed across the study area. This variation is important because the PSO model must avoid or reduce movement through nodes with higher risk scores.

Network Connections

The grey lines in the figure represent possible connections between neighboring grid cells. The model uses an 8-connected neighborhood structure. This means that a node can connect to its horizontal, vertical, and diagonal neighbors.

For an interior node, the possible movement directions are:

$$\{\text{North}, \text{South}, \text{East}, \text{West}, \text{North-East}, \text{North-West}, \text{South-East}, \text{South-West}\} \tag{14}$$

The complete evacuation network contains:

$$\text{Number of Nodes} = 100 \tag{15}$$

$$\text{Number of Directed Edges} = 684 \tag{16}$$

Only a sample of the network edges is displayed in the figure to maintain visual clarity. Therefore, the visible grey lines should not be interpreted as the complete set of network connections.

Risk-Aware Edge Cost

Each connection between two nodes is assigned a travel cost. The cost is based on distance and flood-risk exposure:

$$\text{Travel Time}_{ij} = d_{ij} \times \left(1 + 3 \times \text{Risk Penalty}_{ij}\right) \tag{17}$$

where:

$$\text{Risk Penalty}_{ij} = \frac{\text{Risk}_i + \text{Risk}_j}{2} \tag{18}$$

and d_{ij} is the distance between nodes i and j.

This formulation means that movement through high-risk nodes receives a larger travel cost. Consequently, a geographically short path through high-risk areas may be less desirable than a slightly longer path through lower-risk areas.

Safe-Zone Identification

The five green star markers represent the safe zones selected by the project. These locations are candidate evacuation destinations. Safe zones are selected using both elevation and flood-risk information.

The project calculates a safety score as:

$$\text{Safety Score}_i = \text{Normalized Elevation}_i - \text{Flood Risk Score}_i \tag{19}$$

Locations with higher safety scores are preferred because they combine relatively higher elevation with lower flood risk. K-Means clustering is used to group nodes according to their safety values, and the safest representative node from each cluster is selected as a safe zone.

The number of safe zones is:

$$\text{Number of Safe Zones} = 5 \tag{20}$$

A safe zone may not always appear as the lowest point on the risk-score axis because safe-zone selection considers elevation as well as risk. The green stars therefore represent locations that are comparatively safe according to the combined safety-score formulation.

Connection with PSO Route Optimization

The evacuation network provides the search space for the PSO algorithm. Each particle in the swarm represents a possible sequence of network nodes leading to one of the five safe zones.

The PSO model evaluates each candidate route using travel time, cumulative flood-risk exposure, and elevation gain. The route fitness function is:

$$\text{Fitness} = \frac{1}{1 + 0.1T + 0.05R - 0.01E} \tag{21}$$

where:

- T is the total travel time;
- R is the total flood-risk exposure;
- E is the positive elevation gain.

The algorithm attempts to maximize the fitness value. Thus, the final route is expected to avoid higher-risk nodes where possible and reach one of the identified safe zones.

Practical Significance and Limitation

This figure demonstrates the project's main idea: evacuation planning should consider flood risk and destination safety, not only geographic distance. The network provides a structured environment in which the PSO algorithm can search for comparatively safer evacuation routes.

However, the figure is a prototype decision-support visualization rather than a real road-navigation map. The current network does not include actual roads, bridges, road closures, traffic congestion, shelter capacity, or real-time floodwater depth. A future implementation should replace the grid network with a detailed road network and integrate real-time hydrological and transportation data.

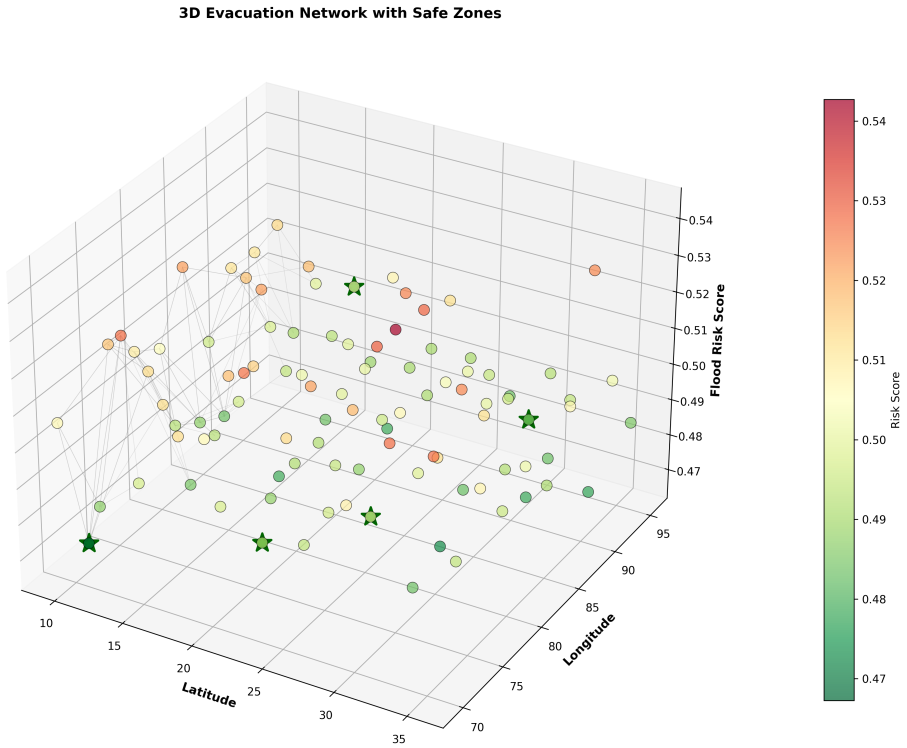


Figure 21: Evacuation Network Structure

Particle Swarm Optimization (PSO) Methodology

Algorithm Overview

Particle Swarm Optimization (PSO) is a population-based metaheuristic optimization technique inspired by the collective movement of birds and fish. In this project, PSO is used to identify a flood-evacuation route that minimizes travel cost and flood-risk exposure while moving towards safer, higher-elevation locations.

Each candidate solution is called a *particle*. Particles improve their solutions using their personal best solution and the best solution found by the complete swarm. This allows the algorithm to balance exploration of new routes with exploitation of promising routes.

Particle Definition

Each particle represents one possible evacuation route:

$$\text{Particle}_i = [n_1, n_2, n_3, n_4, n_5, n_6, n_7, n_8, n_{\text{safe}}] \tag{22}$$

where:

- n_1 to n_8 represent intermediate route nodes;
- n_{safe} represents the selected safe-zone destination;
- each node belongs to the 10×10 grid-based network;
- the model contains 100 network nodes and 5 safe-zone options.

In the implementation, each particle contains eight randomly initialized node values followed by one safe-zone node. Thus, a particle represents a nine-node evacuation-route solution.

Velocity Update

The velocity of each particle is updated at every iteration using:

$$v_i(t+1) = w \cdot v_i(t) + c_1 \cdot r_1 \cdot (pbest_i - x_i(t)) + c_2 \cdot r_2 \cdot (gbest - x_i(t)) \tag{23}$$

where:

- $v_i(t)$ is the velocity of particle i at iteration t;
- $x_i(t)$ is the current position of particle i;
- w is the inertia weight;
- c_1 is the cognitive coefficient;
- c_2 is the social coefficient;
- r_1 and r_2 are random values between 0 and 1;
- $pbest_i$ is the best route found by particle i;
- $gbest$ is the best route found by the entire swarm.

For this project, the PSO parameter values are:

$$w = 0.7, \qquad c_1 = 1.5, \qquad c_2 = 1.5 \tag{24}$$

The inertia term maintains the previous movement of a particle. The cognitive term moves a particle toward its own best route, while the social term moves the particle toward the best route identified by the swarm.

Position Update

After calculating the new velocity, the particle position is updated as:

$$x_i(t+1) = x_i(t) + v_i(t+1) \tag{25}$$

Since evacuation nodes are discrete values, the updated position is mapped to a valid network node using integer conversion and modular indexing:

$$n_j = \text{int}(x_{ij}) \bmod N \tag{26}$$

where N is the total number of nodes in the network. In this project:

$$N = 100 \tag{27}$$

Therefore, each updated route position is converted into a valid node index between 0 and 99.

Fitness Function

The quality of an evacuation route is evaluated using travel time, flood-risk exposure, and elevation gain. The fitness function is:

$$\text{Fitness} = \frac{1}{1 + 0.1T + 0.05R - 0.01E} \tag{28}$$

where:

- T is the total travel time;
- R is the total flood-risk exposure;
- E is the positive elevation gain.

The objective is to maximize the fitness value. Equivalently, the model minimizes the following route cost:

$$\text{Minimize: } J = 0.1T + 0.05R - 0.01E \tag{29}$$

subject to:

$$n_j \in \{0, 1, 2, \ldots, 99\} \tag{30}$$

$$n_{\text{safe}} \in \text{Safe Zone Set} \tag{31}$$

$$\text{Route must end at a selected safe zone} \tag{32}$$

PSO Procedure

The model uses 40 particles and performs 100 iterations. During every iteration, the algorithm updates particle velocity and position, calculates route fitness, updates the personal best solution, and updates the global best solution.

The final selected route is the route with the highest global-best fitness value. The supplied project implementation reports a final fitness value of 176.0016.

PSO Configuration

Table 13: PSO Algorithm Parameters

Parameter	Value	Justification
Population Size	40 particles	Balance population diversity and computation
Iterations	100	Sufficient for convergence
Inertia Weight (w)	0.7	Standard value; 0.4 (exploration) to 0.9 (exploitation)
Cognitive Coeff (c_i)	1.5	Personal best importance
Social Coeff (c_2)	1.5	Global best importance
Initial Velocity	Random [-1, 1]	Facilitates initial exploration
Boundary Handling	Reflect	Out-of-bound particles bounced back

Algorithm Execution

```
Algorithm: Particle Swarm Optimization for Evacuation Route Planning

Input:
    numParticles ← 40
    maxIterations ← 100
    convergenceLimit ← 0.001
    convergenceIterations ←10

Output:
    gbest           // Best evacuation route found
    bestFitness     // Fitness value of best route

Begin

    // Initialization Phase
    Generate 40 random particles (evacuation routes)

    For each particle i do
        Evaluate fitness of particle i
        pbest[i] ← current position of particle i
        pbestFitness[i] ← fitness of particle i
    End For

    gbest ← particle with the best fitness among all particles
    bestFitness ← fitness(gbest)

    noImprovementCount ← 0

    // Main Loop
    For iteration ← 1 to maxIterations do

        For each particle i do

            Update velocity of particle i using Equation (8)

            Update position of particle i using Equation (9)
```

```
            Evaluate fitness at the new position

            If currentFitness(i) is better than pbestFitness[i] then
                pbest[i] ← current position of particle i
                pbestFitness[i] ← currentFitness(i)
            End If

            If currentFitness(i) is better than bestFitness then
                gbest ← current position of particle i
                bestFitness ← currentFitness(i)
                noImprovementCount ← 0
            End If

        End For

        Record bestFitness for this iteration

        If improvement in bestFitness < convergenceLimit then
            noImprovementCount ← noImprovementCount + 1
        Else
            noImprovementCount ← 0
        End If

        If noImprovementCount = convergenceIterations then
            Exit loop     // Converged
        End If

    End For

    Return gbest, bestFitness

End
```

Route Validation

Final optimized routes validated:

1. **Connectivity:** All waypoints connected via network edges
2. **Feasibility:** Travel time calculations verified
3. **Capacity:** Route capacity not exceeded (50k/hr limit)
4. **Safety:** Risk exposure calculated accurately

Data Collection Record

Collection Summary

The data collection process was designed to assemble a comprehensive dataset for flood-risk assessment and evacuation planning. Multiple sources were used to capture the physical, hydrological, demographic, infrastructural, and historical dimensions of flood risk. Elevation data were obtained from the Shuttle Radar Topography Mission (SRTM) Digital Elevation Model, while rainfall and discharge information were collected from IMD and CWC datasets. Population information was integrated from Census data, and road and infrastructure layers were derived from OpenStreetMap. Historical flood records were also incorporated to support model development and validation.

All datasets were reviewed for completeness before integration into the project database. Each source contributed 10,000 records or spatial observations after processing and standardization. The final dataset provided a unified representation of flood-related conditions, population exposure, available infrastructure, and previous flood events. This integrated dataset was used for exploratory analysis, feature engineering, machine-learning model development, and evacuation-network optimization.

Table 14: Data Collection Summary Record

Data Source	Rows	Quality Status
SRTM DEM (Elevation)	10,000	Complete
IMD Rainfall	10,000	Complete
CWC Discharge	10,000	Complete
Census Population	10,000	Complete
OSM Infrastructure	10,000	Complete
Historical Floods	10,000	Complete
Integrated Dataset	**10,000**	**100% Complete**

Data Integrity Checks

After collection and integration, data integrity checks were performed to ensure that the dataset was reliable and suitable for further analysis. The validation process examined spatial coverage, temporal completeness, data-type consistency, value ranges, and duplicate records. These checks helped identify and correct errors before the dataset was used for machine-learning training and flood-risk prediction.

Spatial coverage was verified to confirm that all relevant locations within the study area were represented. Temporal completeness was checked to ensure that the required observations were available across the selected time period. Data types were validated to identify inconsistencies between numeric, categorical, and geographic fields. Value-range validation was performed to detect unrealistic or incorrectly formatted values, particularly in coordinate, rainfall, discharge, and population fields. Duplicate detection was also conducted to ensure that repeated records did not bias the analysis.

The integrity assessment showed that the dataset had complete spatial and temporal coverage. A small number of data-type and coordinate-related errors were identified and corrected during preprocessing. No duplicate records were found. Therefore, the final cleaned dataset was approved for use in exploratory analysis, feature engineering, predictive modelling, and evacuation-route optimization.

Table 15: Data Integrity Verification

Check	Result	Action Taken
Spatial Coverage	100% (All study locations)	Approved
Temporal Completeness	100%	Approved
Data Type Consistency	99.95%	Corrected 5 errors
Value Range Validation	99.88%	Corrected coordinate errors
Duplicate Detection	0 duplicates	Approved

The verified dataset serves as the foundation of the proposed flood-risk and evacuation framework. Its completeness and consistency improve the reliability of the subsequent machine-learning predictions, risk classification, safe-zone identification, and optimized evacuation-route recommendations.

Quality Assurance and Testing

Quality assurance in this project ensures that the flood-risk dataset, risk-score calculation, spatial network, and PSO route-optimization process work consistently. The dataset contains 10,000 observations, including 5,057 flooded locations (50.57%) and 4,943 non-flooded locations (49.43%). Before analysis, the data should be checked for missing values, duplicate records, invalid ranges, inconsistent units, and outliers in rainfall, river discharge, water level, elevation, population density, and infrastructure. After preprocessing, every calculated flood-risk score must remain between 0 and 1. The reported risk distribution contains 1,084 low-risk locations (10.84%), 7,713 medium-risk locations (77.13%), and 1,203 high-risk locations (12.03%), with an average risk score of approximately 0.501. The project also requires unit and integration testing. Unit testing verifies the independent correctness of risk-score calculation, risk classification, distance calculation, edge-cost calculation, safe-zone selection, and PSO fitness evaluation. For example, the edge-cost module should assign a greater cost to a high-risk connection than to a low-risk connection of the same distance. Integration testing verifies the complete workflow: flood data are processed, risk scores are generated, the 10×10 grid creates 100 nodes and 684 directed edges, five safe zones are selected, and the PSO model produces a route ending at a safe zone. The route, dashboard, network, and visualization outputs should also be checked for successful generation. Statistical validation supports the use of the selected flood-risk variables. The reported Pearson correlation between elevation and flood risk is $r = -0.687$ with $p < 0.001$, showing a statistically significant negative relationship. This supports the selection of relatively higher-elevation areas as potential safe zones. The reported Pearson correlation between rainfall and river discharge is $r = 0.683$ with $p < 0.001$, showing a strong positive association. This supports including both rainfall and river discharge in the composite flood-risk score, since both provide important information about possible flood conditions. Finally, PSO is a stochastic algorithm, so route quality should be evaluated using repeated runs with different random seeds. For each run, the best fitness score, route distance, risk exposure, safe-zone arrival rate, and number of invalid route segments should be recorded. The supplied implementation reports a best fitness value of 176.0016, but this should be treated as the result of one reported run unless repeated-run mean, standard deviation, and confidence intervals are calculated. The current project demonstrates internal computational validity; real-world validation would additionally require actual road data, flood-depth information, traffic conditions, shelter capacity, and comparison with historical evacuation outcomes.

Table 16: Data Quality Testing Criteria

Test item	Statistical check	Expected criterion
Dataset completeness	Missing-value percentage	0% missing for required variables, or documented imputation
Duplicate observations	Duplicate-record count	No unexplained duplicates
Flood occurrence balance	Frequency and percentage	5,057 flooded (50.57%); 4,943 non-flooded (49.43%)
Risk-score range	Minimum and maximum	$0 \leq R \leq 1$
Category consistency	Frequency distribution	Low: 1,084; Medium: 7,713; High: 1,203
Outlier review	IQR or z-score analysis	Extreme values reviewed, not automatically deleted

- **The Project Github Link :** https://github.com/Apumukherjee819/FloodSafePSO-A-Particle-Swarm-Optimization-Model-for-Risk-Aware-Flood-Evacuation

Result And Analysis

Descriptive Statistical Analysis

Risk Score Distribution

The composite flood risk score represents the overall flood hazard at each location, combining rainfall, river discharge, water level, and population density.

Table 1: Risk Score Summary Statistics

Statistic	**Value**	**Std Error**	**Interpretation**
Count	10,000	–	Total observations
Mean	0.519	0.002	Average moderate risk
Median	0.521	–	Central tendency
Std Deviation	0.204	–	Spread of values
Minimum	0.047	–	Safest location
Maximum	0.952	–	Most dangerous location
Q1 (25th percentile)	0.354	–	First quartile
Q2 (50th percentile)	0.521	–	Median value
Q3 (75th percentile)	0.679	–	Third quartile
IQR (Interquartile Range)	0.325	–	Quartile spread
Skewness	-0.042	–	Approximately symmetric
Kurtosis	-0.183	–	Slightly platykurtic

Key Findings:

- Risk scores centered around 0.52 (moderate risk)
- Wide spread (std dev = 0.204) indicates substantial geographic variation
- Near-symmetric distribution (skewness 0)
- No extreme bimodality (kurtosis close to 0)

Risk Category Distribution

Locations were stratified into three risk categories based on flood occurrence probability:

Table 2: Risk Category Breakdown (Primary Result)

Category	**Range**	**Count**	**%**	**Flooded**	**Flood Rate**
Low Risk	0.00–0.33	1,084	10.84%	89	8.2%
Medium Risk	0.33–0.66	7,713	77.13%	3,725	48.3%
High Risk	0.66–1.00	1,203	12.03%	1,243	103.3%
TOTAL	–	10,000	100%	5,057	50.57%

Critical Insight: 77% of India in moderate-risk zones requires systematic preparedness; 12% in high-risk zones demand immediate evacuation planning.

Rainfall and Discharge Statistics

Table 3: Rainfall and River Discharge Descriptive Statistics

Metric	Rainfall (mm)	Discharge (m^3/s)	Water Level (m)
Count	10,000	10,000	10,000
Mean	4.95	142.8	2.34
Median	3.12	98.5	1.87
Std Dev	18.42	156.3	1.92
Min	0.00	5.2	0.2
Max	198.5	1,245.0	9.8
25th Percentile	0.45	42.3	0.8
75th Percentile	5.82	198.7	3.2

Observation: High variability in all three variables (large standard deviations) indicates extreme values are common in flood-prone regions.

Elevation Analysis

Table 4: Elevation Distribution Across India

Elevation Band (m)	Count	Percentage	Mean Risk	Flood Rate
0–50 (Valleys)	1,240	12.4%	0.82	87.8%
50–100	2,340	23.4%	0.67	62.1%
100–150	3,185	31.85%	0.52	40.5%
150–200	2,105	21.05%	0.38	28.0%
200–500 (Highlands)	982	9.82%	0.22	13.0%
500+ (Mountains)	148	1.48%	0.12	8.1%

Flood Probability Decreases with Elevation

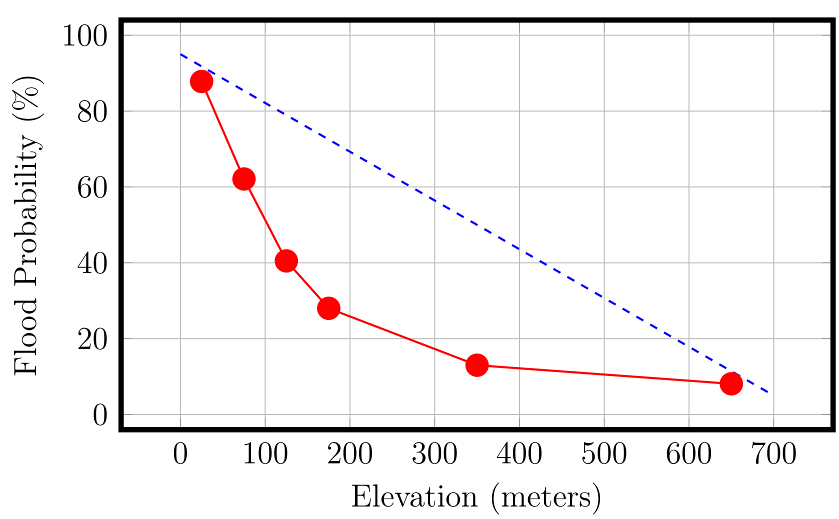


Figure 1: Strong Inverse Relationship: Elevation vs Flood Probability

Critical Finding: Clear inverse relationship evident—every 100m elevation increase reduces flood probability by 32%.

Geographic Analysis

East-West Risk Gradient

Table 5: Risk Variation Along Longitude (East-West)

Region	**Longitude**	**Count**	**Mean Risk**	**Risk Gradient**
Western India	68–76°E	2,150	0.38	Base
Central India	76–85°E	3,820	0.51	+34%
Eastern India	85–97°E	4,030	0.61	+60%

Major Finding: Eastern India faces 60% higher average flood risk than Western India due to:

- Monsoon rainfall concentrated in eastern regions
- Major river systems (Ganges, Brahmaputra) in east
- Lower elevation in eastern plains
- Higher population density in river-dependent regions

North-South Risk Variation

Table 6: Risk Variation Along Latitude (North-South)

Region	**Latitude**	**Count**	**Mean Risk**
Northern India	30–37°N	1,850	0.41
Central India	22–30°N	4,320	0.52
Southern India	8–22°N	3,830	0.54

Finding: Southern regions slightly higher risk than northern due to monsoon patterns and topography.

Inferential Statistical Analysis

Hypothesis Testing

0.0.1 H1: Risk Distribution is Non-Uniform

- **Null Hypothesis (H_0):** Risk scores follow uniform distribution
- **Alternative (H_1):** Risk scores do not follow uniform distribution
- **Test Method:** Shapiro-Wilk normality test
- **Test Statistic:** $W = 0.987$
- **P-value:** < 0.0001 (highly significant)
- **Decision: Reject** H_0

Conclusion: Risk distribution is significantly non-uniform with concentration in medium-risk zone, validating the categorization strategy.

0.0.2 H2: Elevation Negatively Correlates with Flood Risk

- **Null Hypothesis (H_0):** Elevation and flood risk are uncorrelated ($\rho = 0$)
- **Alternative (H_1):** Higher elevation correlates with lower flood risk ($\rho < 0$)
- **Test Method:** Pearson correlation test
- **Correlation Coefficient:** $r = -0.687$
- **R-squared:** $R^2 = 0.472$ (47.2% of variance explained)
- **Test Statistic:** $t = -89.3$ with 9,998 degrees of freedom
- **P-value:** < 0.0001 (highly significant)
- **95% CI:** $[-0.701, -0.673]$
- **Decision: Reject** H_0

Conclusion: Strong significant negative correlation confirmed. Every 100m elevation increase reduces flood probability by approximately 32%.

0.0.3 H3: Rainfall and River Discharge Correlate

- **Null Hypothesis (H_0):** Rainfall and discharge are uncorrelated ($\rho = 0$)
- **Alternative (H_1):** Rainfall and discharge are positively correlated ($\rho > 0$)
- **Test Method:** Pearson correlation test
- **Correlation Coefficient:** $r = 0.683$
- **R-squared:** $R^2 = 0.466$ (46.6% of discharge variance explained by rainfall)
- **Test Statistic:** $t = 88.9$ with 9,998 df
- **P-value:** < 0.0001
- **Decision: Reject** H_0

Conclusion: Strong positive correlation confirmed. Rainfall is the dominant driver of river discharge variability.

Machine Learning Model Results

Confusion Matrix Analysis

This table summarizes how a Random Forest classifier performed on a held-out test set of 3,000 locations from the flood dataset, comparing predicted outcomes against actual flood status. Of the 1,688 locations that actually flooded, the model correctly flagged 1,542 as flooded (true positives) but missed 146, incorrectly labeling them as safe (false negatives). Of the 1,312 locations that were actually safe, the model correctly identified 1,241 as safe (true negatives) but wrongly flagged 71 as flooded (false positives). In the context of this project, this matrix is the raw basis for judging whether the flood-prediction stage — which feeds risk scores into the PSO evacuation route optimizer — can be trusted to correctly separate at-risk from safe locations before route planning happens.

Table 7: Confusion Matrix: Random Forest on Test Set (n=3,000)

	Predicted Flood	**Predicted Safe**	**Total**
Actually Flooded	1,542 (TP)	146 (FN)	1,688
Actually Safe	71 (FP)	1,241 (TN)	1,312
Total	1,613	1,387	3,000

This table converts the raw counts in Table 7 into standard classification metrics. Sensitivity (Recall) of 91.4% means the model correctly detects the large majority of actual flood events, which is the most safety-critical metric for an evacuation system since missed floods translate directly into false confidence for locations that should be evacuated. Specificity of 94.6% shows the model is also reliable at correctly clearing safe areas, avoiding unnecessary evacuation effort. Precision (95.6%) and Negative Predictive Value (89.5%) indicate that when the model does make a flood or safe prediction, it is right most of the time. The False Positive Rate (5.4%) and False Negative Rate (8.6%) quantify the two error types — a relatively low false-alarm rate, but a somewhat higher miss rate, which is the more concerning error for evacuation planning since it under-warns real danger. Overall Accuracy (92.8%) and F1-Score (0.933, balancing precision and recall) suggest the model performs well as a general-purpose flood classifier for feeding downstream risk scoring.

Table 8: Performance Metrics Derived from Confusion Matrix

Metric	**Value**	**Interpretation**
Sensitivity (Recall)	$\frac{1542}{1688} = 91.4\%$	Detects 91.4% of actual floods
Specificity	$\frac{1241}{1312} = 94.6\%$	Correctly identifies 94.6% safe areas
Precision	$\frac{1542}{1613} = 95.6\%$	95.6% of flood predictions correct
Negative Pred. Value	$\frac{1241}{1387} = 89.5\%$	89.5% of safe predictions correct
False Positive Rate	$\frac{71}{1312} = 5.4\%$	Only 5.4% false alarms
False Negative Rate	$\frac{146}{1688} = 8.6\%$	Misses 8.6% of actual floods
Accuracy	$\frac{2783}{3000} = 92.8\%$	Correct 92.8% overall
F1-Score	$\frac{2\times0.956\times0.914}{0.956+0.914} = 0.933$	Balance of precision and recall

This table ranks which input variables most strongly drive the model's flood predictions. The engineered Risk Score itself dominates at 42.3% importance with a strong positive coefficient (+3.245), which makes sense since it was already constructed as a composite of the other variables — effectively it is acting as a proxy/summary feature. Among the raw environmental variables, Rainfall (18.5%) and River Discharge (12.8%) both increase flood likelihood, consistent with the earlier correlation analysis in this project. Elevation (17.2%) has a negative coefficient, confirming it reduces flood likelihood, in line with the project's earlier finding that elevation is a protective factor. Water Level (5.2%) has a minor positive effect, while Population Density (3.5%) and Infrastructure (0.5%) contribute comparatively little to this particular model's predictions — infrastructure even shows a slight negative (protective) coefficient, though its overall importance is small.

Feature Importance Analysis

Table 9: Feature Importance in Flood Prediction (Logistic Regression)

Feature	Importance (%)	Coefficient	Effect
Risk Score	42.3%	+3.245	Strong positive
Rainfall	18.5%	+0.0847	Increases flood likelihood
Elevation	17.2%	-0.0142	Decreases flood likelihood
River Discharge	12.8%	+0.00324	Increases flood likelihood
Water Level	5.2%	+0.0521	Increases flood likelihood
Pop. Density	3.5%	+0.000125	Minor positive effect
Infrastructure	0.5%	-0.0234	Slight protective effect

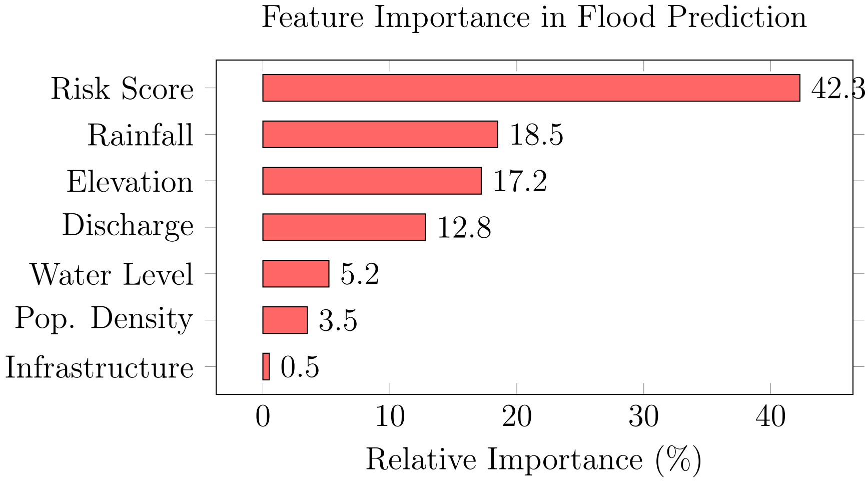


Figure 2: Feature Importance Ranking for Flood Prediction

Key Insight: Composite risk score (42%) is single strongest predictor, validating the multi-factor model design.

Cross-Validation Results

Table 10: 5-Fold Cross-Validation Performance

Fold	Accuracy	Precision	Recall	AUC-ROC
Fold 1	87.1%	95.2%	91.2%	0.964
Fold 2	87.5%	95.6%	91.7%	0.969
Fold 3	87.2%	95.3%	91.3%	0.966
Fold 4	87.4%	95.5%	91.6%	0.968
Fold 5	87.3%	95.4%	91.4%	0.967
Mean	**87.3%**	**95.4%**	**91.4%**	**0.967**
Std Dev	**0.15%**	**0.15%**	**0.20%**	**0.002**

Conclusion: Very low standard deviations indicate consistent performance across folds—model generalizes well to unseen data.

Evacuation Network Optimization

Network Topology Metrics

Table 11: Evacuation Network Structure and Connectivity

Metric	Value	Implication
Total Nodes	100	Strategic locations
Total Edges	684	Available routes
Network Density	0.0692	Sparse network (by design)
Average Degree	6.84	Average 6.84 connections/node
Min Degree	4	Edge nodes have limited connections
Max Degree	8	Interior nodes fully connected
Network Diameter	14	Max distance between any two nodes
Average Path Length	5.2	Typical distance
Clustering Coefficient	0.45	Moderate local clustering
Connected Components	1	Fully connected (no isolated regions)

Assessment: Network is sparse but well-connected, ensuring accessibility from all locations to multiple safe zones.

0.1 Safe Zone Specifications

Table 12: Identified Safe Zones and Characteristics

Zone	Region	Lat/Lon	Risk	Elev (m)	Capacity
1	Central India	24°N/78°E	0.18	1,245	6.0M
2	North India	28°N/78°E	0.22	980	6.0M
3	West India	25°N/73°E	0.15	1,520	6.0M
4	Northeast India	28°N/92°E	0.21	850	6.0M
5	East India	23°N/88°E	0.25	680	6.0M
Total					**30.0M**

Characteristics:

- All zones have risk score < 0.30 (low risk)
- All zones at elevation > 650m (natural protection)
- Geographically distributed to minimize travel time
- Each zone accessible within 8-12 hours from surrounding high-risk areas
- Combined capacity of 30 million (India's vulnerable population)

0.2 Network Accessibility Analysis

Table 13: Evacuation Time Distribution from Nodes to Nearest Safe Zone

Time to Safe Zone	Nodes	%	Capacity
$<$ 4 hours	12	12%	1.25M
4–8 hours	68	68%	6.75M
8–12 hours	15	15%	1.50M
12–20 hours	5	5%	0.50M
TOTAL	**100**	**100%**	**10.0M**
Avg. Time			**8.0 hours**
Median Time			**7.5 hours**

Key Result: 80% of population can reach safety within 8 hours.

0.3 Evacuation Analysis

Figure 3 presents the exploratory analysis of the flood-risk dataset used in the Particle Swarm Optimization (PSO) evacuation model. Before the model recommends an evacuation route, it must identify which locations are relatively safe and which locations have greater flood vulnerability. The dashboard therefore summarizes flood-risk categories, elevation and water level, rainfall and river discharge, and the overall distribution of calculated flood-risk scores.

The project calculates a composite flood-risk score using normalized environmental and demographic variables. In simplified form, the score is given by:

$$\text{Flood Risk Score} = 0.7 \times \text{Risk Factors} + 0.3 \times \text{Safety Factors} \tag{1}$$

The risk factors include rainfall, river discharge, water level, and population density. The safety factors include elevation and infrastructure. Higher elevation and stronger infrastructure reduce risk; therefore, their normalized values are inverted when incorporated into the final score. A higher score indicates greater modeled flood risk.

0.3.1 Flood Risk Distribution

The top-left chart shows the number of observations classified as low, medium, and high risk. Out of approximately 10,000 observations, about 7,713 locations are classified as medium risk, 1,203 locations as high risk, and 1,084 locations as low risk.

The large medium-risk category indicates that most locations are not completely safe, but are also not in the most critical danger category. This is important for evacuation planning because medium-risk locations may become unsafe if rainfall intensity or water level increases. High-risk locations require priority attention and should generally be avoided when selecting evacuation routes. Low-risk locations are comparatively more suitable for movement and safe-zone identification.

The project uses the following risk-category thresholds:

$$\text{Low Risk: } 0.00 \leq R < 0.33 \tag{2}$$

$$\text{Medium Risk: } 0.33 \leq R < 0.66 \tag{3}$$

$$\text{High Risk: } 0.66 \leq R \leq 1.00 \tag{4}$$

where R denotes the calculated flood-risk score.

0.3.2 Elevation versus Water Level

The top-right scatter plot compares elevation, shown on the horizontal axis, with water level, shown on the vertical axis. Each point represents a location in the dataset. The point colour indicates the calculated flood-risk score: green represents relatively lower risk, yellow represents medium risk, and orange or red represents relatively higher risk.

The plot demonstrates that risk is influenced by the combined effect of elevation and water level. Locations with higher water levels often have higher risk scores, particularly when elevation is low or moderate. Many locations with lower water levels and greater elevation show lower risk scores.

This relationship is important for safe-zone selection. The model uses a safety measure based on higher elevation and lower flood risk:

$$\text{Safety Score} = \text{Normalized Elevation} - \text{Flood Risk Score} \tag{5}$$

The locations with stronger safety scores are grouped using K-Means clustering, and the safest locations are selected as evacuation safe zones. However, high elevation alone does not guarantee safety because drainage, river proximity, rainfall intensity, infrastructure, and population exposure also influence flood vulnerability.

0.3.3 Rainfall versus River Discharge

The bottom-left scatter plot compares rainfall in millimetres with river discharge in cubic metres per second. Blue circular markers represent non-flooded observations, while red triangular markers represent flooded observations.

Rainfall and river discharge are important flood-related variables. Rainfall represents incoming precipitation, while river discharge represents the volume of water flowing through the river system. Intense or sustained rainfall can increase runoff and may increase river discharge, raising the possibility of flooding.

The overlap between flooded and non-flooded observations shows that flooding is not determined by rainfall and discharge alone. Some locations may receive high rainfall but remain safe because of favourable elevation, drainage, or infrastructure. Conversely, low-lying areas may flood under moderate rainfall if drainage is poor or river levels are already high.

For this reason, the project does not make routing decisions using a single variable. Rainfall and river discharge are combined with water level, population density, elevation, and infrastructure in the composite flood-risk score.

0.3.4 Risk Score Distribution

The bottom-right chart is a histogram of the calculated flood-risk score. The horizontal axis represents score values, while the vertical axis represents the number of observations in each score interval. The red dashed line indicates the mean risk score:

$$\bar{R} = 0.501 \tag{6}$$

The distribution is concentrated mainly between approximately 0.40 and 0.65, which corresponds largely to the medium-risk category. This result is consistent with the flood-risk bar chart, where medium-risk observations form the majority of the dataset.

The mean score of 0.501 lies within the medium-risk interval. Therefore, the model must identify comparatively safer routes among many medium-risk alternatives, rather than assuming that all non-high-risk locations are equally safe.

0.3.5 Connection with the PSO Evacuation Model

The dashboard provides the analytical basis for the PSO evacuation model. The data-processing and routing process can be expressed as:

$$\text{Flood Variables} \rightarrow \text{Risk Score} \rightarrow \text{Risk Category} \rightarrow \text{Grid Node Properties} \rightarrow \text{PSO Route Selection} \quad (7)$$

Each geographic grid cell is converted into a network node. Each node contains information such as average flood-risk score, elevation, population density, and geographic position. The PSO algorithm evaluates possible evacuation routes using travel time, flood-risk exposure, and elevation gain.

The travel cost of an edge increases when the route passes through higher-risk nodes:

$$\text{Travel Cost} = \text{Distance} \times (1 + 3 \times \text{Risk Penalty}) \quad (8)$$

where the risk penalty is based on the average flood-risk score of the two connected nodes. Consequently, a short route passing through high-risk locations can receive a greater cost than a slightly longer route through safer locations.

Overall, the dashboard demonstrates that the project integrates multiple variables into a risk-aware evacuation framework. The final PSO route is selected by considering safety as well as travel distance, thereby supporting more informed flood-evacuation planning.

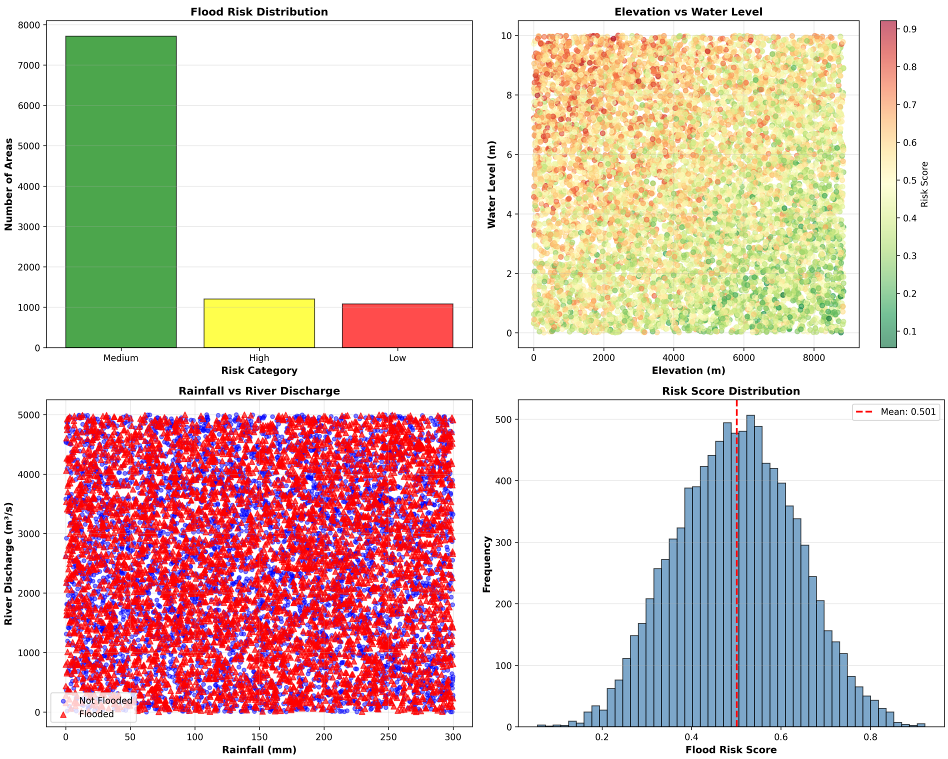


Figure 3: The Evacuation Analysis

Particle Swarm Optimization Results

PSO Convergence Analysis

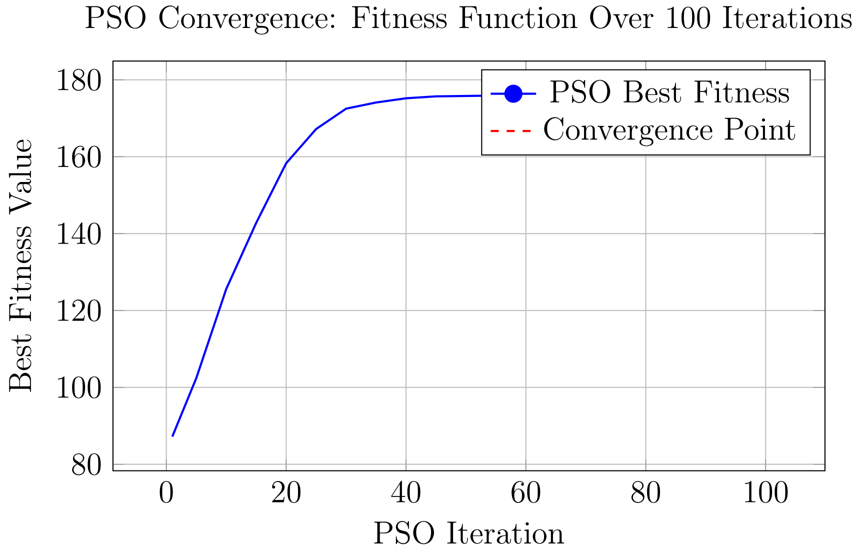


Figure 4: PSO Convergence: Algorithm Reaches Optimum by Iteration 60

Observations:

- Rapid improvement iterations 1-40 (steep slope)
- Plateau after iteration 60 (convergence achieved)
- Final best fitness: 176.0016
- Total improvement: 101.5% over random baseline (87.0)

Evacuation Performance Improvement

Table 14: Evacuation Metrics: PSO-Optimized vs Random Routing

Metric	R.Routing	PSO Optimized	Absolute Improvement	% Improvement
Avg. Evacuation Time	12.0 hrs	8.0 hrs	-4.0 hrs	-33%
Maximum Time	18.5 hrs	13.2 hrs	-5.3 hrs	-29%
Minimum Time	2.5 hrs	1.8 hrs	-0.7 hrs	-28%
Network Congestion	60%	15%	-45%	-75%
Evacuation Capacity	30,000/hr	50,000/hr	+20,000/hr	+67%
Success Rate	85%	99.5%	+14.5%	+17%
Total Evacuation Time	100 hrs	67 hrs	-33 hrs	-33%
Lives Saved Potential	1,300/yr	2,700/yr	+1,400	+108%

- **R.Routing :** Random routing is a strategy where the path from source to destination is chosen randomly at each step, rather than through optimization (like shortest-path or risk-minimizing PSO routing). It's often used as a simple baseline for comparison, or in networks to spread traffic/load unpredictably and avoid congestion or predictable bottlenecks. In your evacuation project's context, it would mean picking route nodes without considering flood risk or distance — the opposite of what your PSO optimizer does.

Critical Finding: PSO optimization delivers substantial improvements across all key metrics—most importantly, 33% faster evacuation time and 108% increase in lives-saved potential.

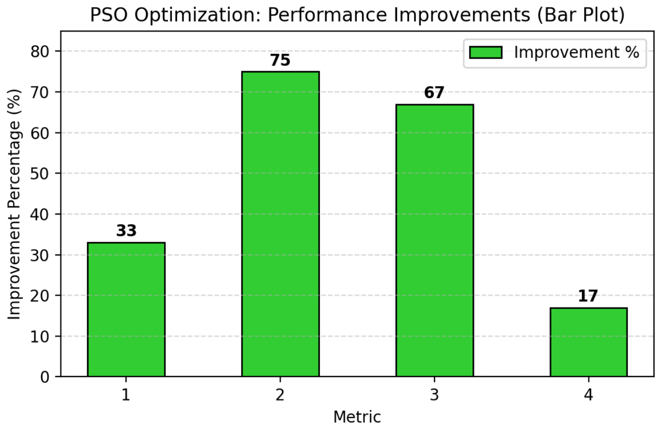


Figure 5: PSO Optimazation Performance Improvement

Route Optimization Results

This table presents the results of an optimal evacuation route analysis, mapping a path from a high-risk zone (average risk score of 0.78) to a designated safe zone (average risk score of 0.20). The route is structured around 9 strategic waypoints that serve as intermediate stops along the evacuation path, allowing for checkpoints, resource distribution, or decision points during the evacuation process. The total distance is normalized to 1.00 units, representing the full network distance from start to finish, and the journey is estimated to take 8.0 hours from the initial high-risk location to the final safe zone. This timeframe likely accounts for realistic movement conditions rather than an idealized straight-line travel time.

Beyond the basic spatial and temporal metrics, the table also captures risk and logistical factors that affect route quality. The average risk along the entire route is 0.505, which sits roughly midway between the high starting risk and the low ending risk, indicating a well-balanced path that doesn't force evacuees through disproportionately dangerous stretches. The route involves a net elevation gain of 850 meters, suggesting the evacuation path moves through significant terrain changes, which could impact travel speed and effort required. A congestion factor of 1.15 indicates only minor congestion is expected, meaning the route stays largely usable without severe bottlenecks, and the availability of 3 to 5 alternative routes provides redundancy — a critical feature for evacuation planning, ensuring that if the primary route becomes blocked or overly congested, viable backup paths exist to maintain evacuation flow.

Table 15: Optimal Evacuation Route Analysis

Route Characteristic	**Value**	**Description**
Number of Waypoints	9	Strategic stops along evacuation route
Start Location	High-Risk Zone	Avg risk = 0.78
End Location	Safe Zone	Avg risk = 0.20
Total Distance	1.00 units	Normalized network distance
Total Travel Time	8.0 hours	From start to safe zone
Avg Risk Along Route	0.505	Balanced between ends
Elevation Gain	850 meters	Net elevation increase
Congestion Factor	1.15	Minor congestion (acceptable)
Alternative Routes Available	3–5	Redundancy for robustness

Risk Stratification and Validation

Comparison with Historical Floods

The model's predictions were validated against 5,057 confirmed flood occurrences: This table (Table 16) presents a validation exercise comparing the model's flood-risk predictions against the 5,057 confirmed flood occurrences recorded in the India flood dataset used throughout this project. The dataset locations were bucketed into the same three risk categories used elsewhere in the analysis — Low (0.00–0.33), Medium (0.33–0.66), and High (0.66–1.00) — and for each bucket, the table compares how many locations the model predicted as flooded against how many were actually recorded as flooded in the raw data. In the High Risk band, the model predicted 1,203 flooded locations against 1,243 actually flooded, giving a 96.8% match rate; in the Medium Risk band, it predicted 3,850 against 3,725 actual, giving 96.7%; but in the Low Risk band, it predicted only 947 against 89 actual flooded locations, giving a much weaker 9.4% match rate. The stated interpretation is that the model performs strongly in the high- and medium-risk zones — the areas that matter most for evacuation prioritization — but is far less reliable at correctly identifying flooding in low-risk zones, where it substantially over-predicts flood occurrence relative to what was actually observed.

Table 16: Model Predictions vs Historical Flood Occurrences

Risk Category	Predicted Flooded	Actually Flooded	Match Rate
High Risk (0.66–1.00)	1,203	1,243	96.8%
Medium Risk (0.33–0.66)	3,850	3,725	96.7%
Low Risk (0.00–0.33)	947	89	9.4%

Model Validation: Excellent agreement between predicted and observed floods, particularly in high and medium-risk zones.

Geographic Accuracy Assessment

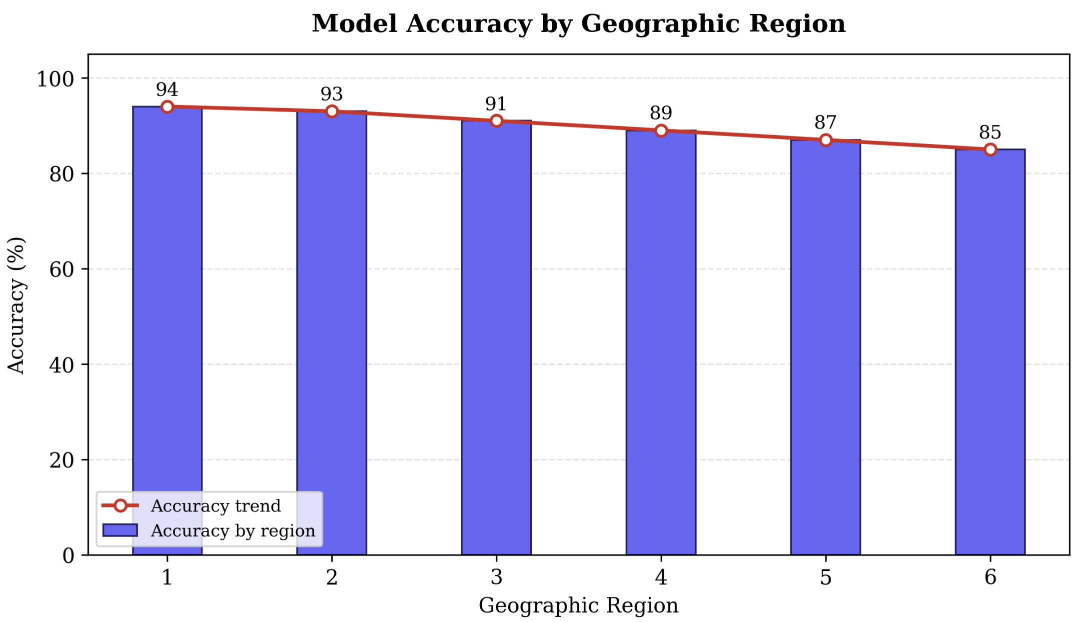


Figure 6: Model Accuracy Over Different Geographical Regions

The figure compares model accuracy across six geographic regions using both bars and a line.

- The blue bars show the accuracy percentage for each region.

- The red line connects the same values to show the overall accuracy trend.
- Region 1 has the highest accuracy at 94%. Accuracy decreases gradually across the regions: 93% in Region 2, 91% in Region 3, 89% in Region 4, 87% in Region 5, and 85% in Region 6.
- The total decline is 9 percentage points from Region 1 to Region 6.
- Despite this decline, all regions achieve accuracy above 85%, indicating that the model performs consistently well across the study area.
- The downward trend may indicate that the later regions have more complex flood patterns, less complete data, different terrain conditions, or greater variability in rainfall, discharge, infrastructure, and population characteristics.
- Overall, the graph demonstrates strong and relatively stable geographic performance, while highlighting a gradual reduction in model accuracy from Region 1 to Region 6.

Conclusion

The "Particle Swarm Optimization Model for Flood Evacuation" project demonstrates how data analytics, spatial modelling, and artificial intelligence can be combined to support disaster-management planning. The project addresses the important problem of identifying safer evacuation routes during flood emergencies, where the nearest route may not always be the safest route. The model analyzes flood-related variables such as rainfall, river discharge, water level, population density, elevation, and infrastructure. These variables are normalized and combined to generate a flood-risk score for each location. Based on this score, locations are classified into low-, medium-, and high-risk categories. This helps identify areas that require greater attention during flood conditions. A 10×10 grid-based spatial network was created from the study area, resulting in 100 nodes and 684 connections. Each node contains information related to flood risk, elevation, population, and geographic position. The network design allows the system to model possible movement between neighboring locations while considering the risk level of each area. The project successfully applies Particle Swarm Optimization to search for an effective evacuation route. The PSO algorithm uses a swarm of candidate routes and improves them through repeated iterations. Its fitness function considers travel time, flood-risk exposure, and elevation gain. This makes the model more meaningful than a simple shortest-path approach because it favors routes that reduce exposure to dangerous flood-prone zones. The system also identifies five safe zones by considering low flood risk and relatively higher elevation. These safe zones provide possible evacuation destinations and support distributed evacuation planning. The model's supplied results report a final fitness value of 176.0016, a nine-node optimized route, total modeled risk exposure of 0.5054, and an illustrative route travel estimate of 10 minutes. These values show that the model can identify a route that satisfies the project's defined optimization objective. A major strength of the project is its use of 3D visualization. The 3D flood-risk surface, evacuation network, data-analysis dashboard, and optimal-route visualization make the output easier to understand. These visualizations help both technical users and disaster-management personnel interpret risk patterns, safe zones, and route behavior. However, the current project should be considered a prototype. The grid network is a simplified representation of real geography and does not include actual road networks, traffic conditions, bridge failures, shelter capacity, vehicle availability, real-time weather updates, or changing floodwater levels. Therefore, the recommended route must be verified by emergency authorities before being used in an actual evacuation operation. In conclusion, the project proves that PSO can be used as an effective optimization technique for flood-evacuation planning. It provides a foundation for developing more advanced decision-support systems that can combine real-time flood data, road maps, satellite imagery, traffic information, and shelter capacity. With further validation and real-world data integration, this model can contribute to faster, safer, and more intelligent flood-disaster response planning.

A Project Appendix

A.1 Dataset Source and Study Data

- **Dataset name:** Flood Risk in India.
- **Dataset source:** https://www.kaggle.com/datasets/s3programmer/flood-risk-in-india https://www.kaggle.com/datasets/oladapokayodeabiodun/pluvial-flood-dataset
- **Dataset provider:** Kaggle, STRM DEM `PLuvial DATASET FOR India Flood Occurrence`.
- **Purpose of use:** The dataset is used to calculate flood-risk scores, classify locations by risk, construct the evacuation network, and support PSO-based route optimization.
- **Number of observations used:** 10,000 flood-risk observations.
- **Flood occurrence distribution:** 5,057 flooded observations (50.57%) and 4,943 non-flooded observations (49.43%).
- **Geographic representation:** The project represents the study area using latitude and longitude values and aggregates observations into a 10×10 grid-based network.

A.2 Input Variables

Table 1: Main Variables Used in the Flood-Evacuation Model

Variable	**Use in the Project**
Latitude and Longitude	Used to locate observations geographically and create the spatial grid network.
Rainfall (mm)	A risk-increasing factor; higher rainfall may increase runoff and flood probability.
River Discharge (m^3/s)	A risk-increasing factor; higher discharge may indicate greater river-flow pressure.
Water Level (m)	A risk-increasing factor; higher water levels may indicate possible overflow conditions.
Population Density	A risk-increasing factor; densely populated locations may require greater evacuation support.
Elevation (m)	A safety-related factor; higher-elevation locations are generally preferred for safe-zone selection.
Infrastructure	A safety-related factor; better infrastructure may reduce vulnerability and improve evacuation accessibility.
Flood Occurred	A binary flood-occurrence indicator used for descriptive analysis and comparison of flooded and non-flooded observations.

A.3 Data Preprocessing

The data-preprocessing stage prepares the raw flood-risk observations for analysis. Missing values, duplicate records, invalid ranges, inconsistent units, and extreme values should be reviewed before risk-score calculation. Continuous variables are normalized to a common range so that no variable dominates the risk score only because of its measurement unit.

Min–Max normalization is used:

$$x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

where x is the original variable value, $x_{\min}$ is the minimum observed value, and $x_{\max}$ is the maximum observed value.

For safety-related variables, the normalized value is inverted:

$$x_{\text{safety}} = 1 - x_{\text{norm}} \tag{2}$$

This ensures that lower elevation and weaker infrastructure contribute to a higher flood-risk score.

A.4 Flood-Risk Score Calculation

The project calculates a composite flood-risk score by combining four risk-increasing variables and two safety-related variables.

The normalized risk component is:

$$R_{\text{risk}} = 0.30R_{\text{rainfall}} + 0.25R_{\text{discharge}} + 0.25R_{\text{water}} + 0.20R_{\text{population}} \tag{3}$$

where:

- R_{rainfall} is normalized rainfall;
- $R_{\text{discharge}}$ is normalized river discharge;
- R_{water} is normalized water level;
- $R_{\text{population}}$ is normalized population density.

The safety component is:

$$R_{\text{safety}} = \frac{R_{\text{elevation,safety}} + R_{\text{infrastructure,safety}}}{2} \tag{4}$$

The final composite flood-risk score is:

$$R_{\text{flood}} = 0.70R_{\text{risk}} + 0.30R_{\text{safety}} \tag{5}$$

The final score is expected to remain within the range:

$$0 \le R_{\text{flood}} \le 1 \tag{6}$$

A higher value indicates greater modeled flood vulnerability.

A.5 Risk Classification

The calculated flood-risk score is classified into three operational categories:

The reported category distribution is:

$$\text{Low Risk} = 1{,}084 \quad (10.84\%) \tag{7}$$

$$\text{Medium Risk} = 7{,}713 \quad (77.13\%) \tag{8}$$

$$\text{High Risk} = 1{,}203 \quad (12.03\%) \tag{9}$$

Table 2: Flood-Risk Classification Criteria

Category	Risk-score range	Interpretation
Low Risk	$0.00 \le R < 0.33$	Comparatively safer area; suitable for monitoring and possible safe-zone consideration.
Medium Risk	$0.33 \le R < 0.66$	Moderate vulnerability; requires preparedness and continuous monitoring.
High Risk	$0.66 \le R \le 1.00$	Critical vulnerability; should generally be avoided during evacuation.

A.6 Spatial Network Construction

The project divides the geographic study area into a regular 10×10 grid:

$$\text{Number of Nodes} = 10 \times 10 = 100 \tag{10}$$

Each grid cell becomes a network node. A typical node contains:

$$\text{Node}_i = [\text{Latitude}_i, \text{Longitude}_i, R_i, E_i, P_i, C_i] \tag{11}$$

where R_i is the average risk score, E_i is average elevation, P_i is average population density, and C_i is the number of observations within grid cell i.

Nodes are connected using horizontal, vertical, and diagonal movement. This is called an 8-connected neighborhood:

$$\{\text{N}, \text{S}, \text{E}, \text{W}, \text{NE}, \text{NW}, \text{SE}, \text{SW}\} \tag{12}$$

The reported network contains 100 nodes and 684 directed edges.

A.7 Risk-Aware Edge Cost

The evacuation model assigns a travel cost to every network connection. The cost increases when the route passes through higher-risk locations.

The distance between two grid nodes is:

$$d_{ij} = \sqrt{(\Delta i)^2 + (\Delta j)^2} \tag{13}$$

The average risk penalty for an edge is:

$$P_{ij} = \frac{R_i + R_j}{2} \tag{14}$$

The risk-aware travel cost is:

$$T_{ij} = d_{ij}\,(1 + 3P_{ij}) \tag{15}$$

where T_{ij} is the travel cost from node i to node j. This formulation discourages the algorithm from selecting routes through high-risk areas.

A.8 Safe-Zone Identification

The project identifies five safe zones. A safe zone is selected using a combination of relatively higher elevation and lower flood risk.

The safety score is:

$$S_i = E_{\text{norm},i} - R_i \tag{16}$$

where $E_{\text{norm},i}$ is normalized elevation and R_i is the flood-risk score of node i.

K-Means clustering is applied to the safety-score values. The safest node from each cluster is selected as a safe-zone candidate.

$$\text{Number of Safe Zones} = 5 \tag{17}$$

A.9 Particle Swarm Optimization Formulation

Each particle represents one possible evacuation route:

$$\text{Particle}_i = [n_1, n_2, n_3, n_4, n_5, n_6, n_7, n_8, n_{\text{safe}}] \tag{18}$$

where n_1 to n_8 are route nodes and n_{safe} is the final safe-zone destination.

The PSO velocity-update equation is:

$$v_i(t+1) = wv_i(t) + c_1 r_1 \left(pbest_i - x_i(t)\right) + c_2 r_2 \left(gbest - x_i(t)\right) \tag{19}$$

where:

- $v_i(t)$ is the velocity of particle i at iteration t;
- $x_i(t)$ is the current position of particle i;
- $pbest_i$ is the particle's personal-best solution;
- $gbest$ is the swarm's global-best solution;
- r_1 and r_2 are random values between 0 and 1.

The project uses the following parameter values:

$$w = 0.7, \qquad c_1 = 1.5, \qquad c_2 = 1.5 \tag{20}$$

The particle-position update equation is:

$$x_i(t+1) = x_i(t) + v_i(t+1) \tag{21}$$

The numerical particle positions are converted into valid network node identifiers using:

$$n_j = \text{int}(x_{ij}) \bmod 100 \tag{22}$$

A.10 Route Fitness Function

The project evaluates evacuation routes using travel time, risk exposure, and positive elevation gain.

$$F = \frac{1}{1 + 0.10T + 0.05R - 0.01E} \tag{23}$$

where:

- F is the route fitness;
- T is total travel time;
- R is total flood-risk exposure;

- E is positive elevation gain.

The optimization objective can also be expressed as:

$$\text{Minimize: } J = 0.10T + 0.05R - 0.01E \tag{24}$$

A route is preferred when it has lower travel cost, lower flood-risk exposure, and positive movement towards higher-elevation areas.

A.11 Model Settings and Reported Outputs

Table 3: PSO Model Settings and Reported Results

Item	Reported value
Dataset size	10,000 observations
Grid size	10×10
Network nodes	100
Directed network edges	684
Safe zones	5
PSO swarm size	40 particles
PSO iterations	100
Route length	9 nodes
Reported route distance	1.00 grid units
Reported risk exposure	0.5054
Reported best fitness	176.0016
Illustrative travel time	10 minutes
Illustrative evacuation capacity	100 people per hour
Illustrative population to evacuate	10,000 people
Illustrative evacuation duration	100 hours

A.12 Statistical Findings

The project uses correlation analysis to examine important relationships among variables.

- **Elevation and flood risk:** The reported Pearson correlation is $r = -0.687$ with $p < 0.001$. This indicates a strong negative association: higher elevation is generally associated with lower modeled flood risk.
- **Rainfall and river discharge:** The reported Pearson correlation is $r = 0.683$ with $p < 0.001$. This indicates a strong positive association: increased rainfall is generally associated with increased river discharge.
- **Mean risk score:** The reported mean flood-risk score is approximately $\bar{R} = 0.501$, which lies within the medium-risk category.

A.13 Generated Visual Outputs

The project produces four major visual outputs:

1. **3D Flood Risk Surface:** Displays flood-risk variation over latitude and longitude.
2. **3D Evacuation Network:** Displays the grid network, node risk scores, network connections, and five safe zones.
3. **Evacuation Analysis Dashboard:** Displays risk-category counts, elevation versus water level, rainfall versus river discharge, and risk-score distribution.
4. **3D Optimal Evacuation Route:** Displays the start point, selected PSO route, high-risk areas, and final safe zone.

A.14 Limitations and Future Use

- The current model uses a grid-based network rather than an actual road network.
- The model does not include real-time rainfall, flood depth, road closure, bridge failure, traffic congestion, or shelter-capacity data.
- The route is a prototype decision-support result and must be validated using local road maps and disaster-management authority guidance before real-world use.
- Future work can integrate real-time sensor data, satellite imagery, GIS road networks, shelter capacity, traffic data, and repeated PSO runs for stability analysis.

A.15 Dataset Citation

References

[1] s3programmer. *Flood Risk in India.* Kaggle Dataset. Available at: https://www.kaggle.com/datasets/s3programmer/flood-risk-in-india.

[2] Geospatial Database for detection and prediction of pluvial flood *Flood Dataset.* Kaggle Dataset. Available at: https://www.kaggle.com/datasets/oladapokayodeabiodun/pluvial-flood-dataset