International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 02 | Feb - 2025 SJIF Rating: 8.448 ISSN: 2582-3930

Flood Prediction using Machine Learning

Dr. M. Sengaliappan[1], Muthu Sahin S H [2]

[1]Head of the Department, Department of Computer Applications, Nehru College of Management, Coimbatore, Tamilnadu, India, ncmdrsengaliappan@nehrucolleges.com

[2] II MCA, Department of Computer Applications, Nehru College of Management, Coimbatore, Tamilnadu, India, muthusahin123@gmail.com

Abstract: *Due to urbanization and climate change, flooding has increased in frequency and severity, upsetting lives and seriously damaging property. Flood Susceptibility Modeling (FSM), which employs sophisticated machine learning approaches, helps identify flood-prone locations and the elements that contribute to these risks in order to solve this problem. This study explores hybrid FSM models that integrate the Index of Entropy (IOE) with Decision Tree (DT), Support Vector Machine (SVM), and Random Forest (RF) to offer a dependable approach for flood prediction and prevention. To assess the predictive power and correlations between influencing elements, the study started with feature selection and multicollinearity analysis. The relationship between several flood-causing components and their total effect on flooding was measured by IOE. Weighted inputs from these findings were used to train the hybrid models. Metrics like the Area Under the Curve (AUC) and other statistical indicators were used to evaluate the models in order to ensure correctness and reliability. The standalone DT model performed the worst (77.0%), while the hybrid DT-IOE model had the best prediction accuracy (87.1%), followed by SVM-IOE and RF-IOE. These findings show that prediction accuracy is increased when machine learning and statistical techniques are combined. 21% of the study region is extremely sensitive to floods, according to the final susceptibility maps, underscoring the major impact of human-induced factors including land-use changes and urban growth. By enhancing feature analysis and prediction accuracy, generative AI significantly enhanced model performance. The significance of hybrid machine learning approaches in developing efficient flood risk management plans is highlighted by this study, which also supports disaster resilience and sustainable urban design.*

Keywords: *Human-induced factors, Flood occurrences, Flood susceptibility modeling (FSM), and hybrid models Artificial Intelligence (ML), Natural Causes Remote Observation.*

1. Introduction:

One of the most destructive natural disasters, floods have a devastating effect on people, infrastructure, agriculture, and the economy. There is increasing demand on governments to provide precise maps of flood risk and long-term management strategies that emphasize readiness and prevention. Flood prediction is difficult because of the changing climate, and

physical models limit short-term forecasts by requiring large amounts of data and computer power. An alternative that is quicker and data-driven is machine learning (ML), which uses previous data to efficiently identify flood patterns with little input. When compared to conventional techniques, machines (SVMs), and neuro-fuzzy systems offer superior accuracy and lower complexity. Combining machine learning with other methods or models improves their resilience and flexibility. But in order to guarantee accuracy and prevent generalization problems, machine learning models rely on high-quality data, necessitation a variety of training datasets. How well they function depends on the kind of prediction (e.g., short-term vs. long-term) and the available data. Despite obstacles, machine learning (ML) is still a crucial tool risk management and flood prediction, especially in locations with intelligent sensors or rain gauges.

1.1 Flood Prediction Techniques of ML

One of the most damaging natural calamities, floods seriously harm both property and human life. Flood prediction has become increasingly important due to climate change induced increases in rainfall. By learning form past data, machine learning (ML) is essential for forecasting floods and other natural disasters. Three types of machine learning can be distinguished: supervised learning, in which models are trained on labelled data to predict outcomes (e.g., classification for categorical variables, regression for continuous variables); unsupervised learning, in which models identify patterns in unlabeled data (e.g., association for identifying relationships between variables, clustering for grouping similar items); and reinforcement learning. By aiding in decision-making via the examination of past data, machine learning (ML) promotes preparedness for catastrophes.

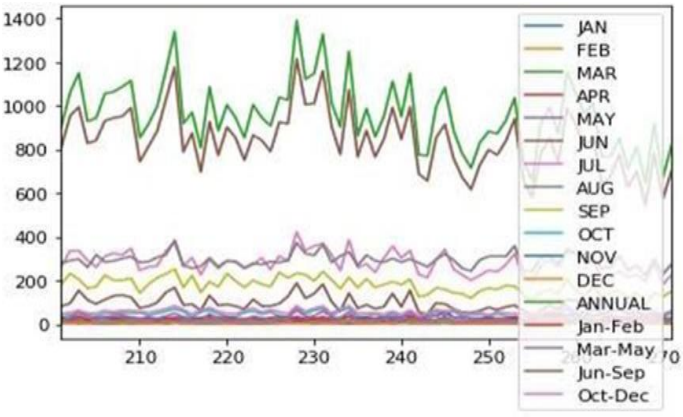


Fig 1: *Flood Prediction of Each Month*

1.2 GenAI and Deep Learning

- Artificial neural networks are used in deep learning, a subfield of machine learning, which draws inspiration from the structure and function of the human brain. Deep learning models, particularly deep neural networks, are composed of several layers of interconnected neurons that can learn hierarchical data representations. They have demonstrated great effectiveness in a variety of fields, like speech and picture recognition, and are adept at picking up intricate patterns.

- Generative AI (GenAI) aims to produce fresh content, as was previously said. It falls under the subgroup of Deep Learning. It uses generative models, such as generative adversarial networks (GANs) and variational autoencoders (VAEs), to learn from data and generate new samples with similar trails. Generative models must first understand the underlying patterns and structures of the data in order to produce outputs that resemble the training data.

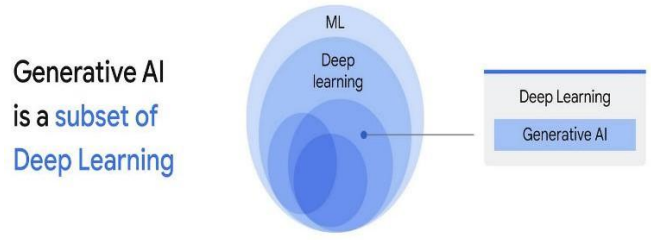


Fig 2: *Understanding with GenAI and Deep Learning*

2. Proposed Method:

The Indian Meteorological Department's rainfall data for Kerala (1901-2018), which used machine learning methods like KNN, LR, SVM, DT, and RF to analyze the millimeter-scale data. Performance evaluation and quality indicator (SNIP, Cite Score, SJR, h-index) were used to rank peer-reviewed studies on flood prediction using machine learning. In order to determine which models performed best for particular application kinds, and results. Applications for flood prediction depend on important factors such as streamflow, rainfall, water level, river flow, and soil moisture. Rainfall has a major impact on flood modelling and runoff, particularly for short-term forecasts and flash floods. Rainfall by itself, however, is not enough to accurately anticipate floods, especially in long-term scenarios when catchment conditions and soil moisture are crucial. The most important flood resource variable is covered in this paper, along with how they can be integrated with machine learning (ML) techniques. It highlights how various machine learning approaches depend on the dataset, use case, and kind of forecast (e.g., short-term water level projections or long-term streamflow modeling).

2.1 Working of Modules

Using data validation, the prediction accuracy of various models is assessed, and the accuracy is obtained by comparing the results. By comparing algorithms, the false-positive rate specification precision, recall, training dataset accuracy, and testing dataset correctness are determined.

The steps involved:

- Define a problem
- Preparing data
- Evaluating algorithms
- Prediction result

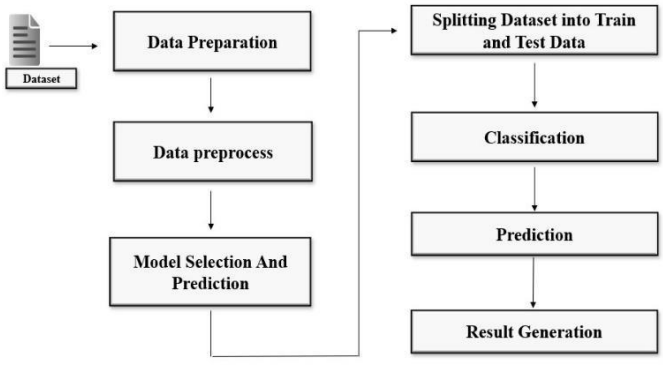


Fig 3: *System Architecture*

2.2 Rainfall Data Set

A dataset that includes historical flood data and rainfall information for particular regions, such as Kerala, is employed. With average rainfall estimated every ten days and displayed on a graph, the dataset covers around three months. Flood occurrences are employed as output labels in a machine learning model trained on Kerala's yearly rainfall data. The model can be used to predict floods in any Indian state with comparable data because it is trained and stored using daily rainfall criteria.

	SUBDIVISION	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	ANNUAL RAINFALL	FLOODS
0	KERALA	1901	28.7	44.7	51.6	160.0	174.7	824.6	743.0	357.5	197.7	266.9	350.8	48.4	3248.6	YES
1	KERALA	1902	6.7	2.6	57.3	83.9	134.5	390.9	1205.0	315.8	491.6	358.4	158.3	121.5	3326.6	YES
2	KERALA	1903	3.2	18.6	3.1	83.6	249.7	558.6	1022.5	420.2	341.8	354.1	157.0	59.0	3271.2	YES
3	KERALA	1904	23.7	3.0	32.2	71.5	235.7	1098.2	725.5	351.8	222.7	328.1	33.9	3.3	3129.7	YES
4	KERALA	1905	1.2	22.3	9.4	105.9	263.3	850.2	520.5	293.6	217.2	383.5	74.4	0.2	2741.6	NO
..	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
113	KERALA	2014	4.6	10.3	17.9	95.7	251.0	454.4	677.8	733.9	298.8	355.5	99.5	47.2	3046.4	YES
114	KERALA	2015	3.1	5.8	50.1	214.1	201.8	563.6	406.0	252.2	292.9	308.1	223.6	79.4	2600.6	NO
115	KERALA	2016	2.4	3.8	35.9	143.0	186.4	522.2	412.3	325.5	173.2	225.9	125.4	23.6	2176.6	NO
116	KERALA	2017	1.9	6.8	8.9	43.6	173.5	498.5	319.6	531.8	209.5	192.4	92.5	38.1	2117.1	NO
117	KERALA	2018	29.1	52.1	48.6	116.4	183.8	625.4	1048.5	1398.9	423.6	356.1	125.4	65.1	4473.0	YES

[118 rows x 16 columns]

Fig 4: *Rainfall Data Set*

2.1.1 Data Preparation

This study aims to collect and organize rainfall data from Kerala and other regions of India in order to investigate flood-prone areas. Data needs to be collected, aggregated, profiled, verified, and converted before it can be used for analytics and visualization. It entails gathering data from internal and external sources and incorporating it into warehouses, data lakes, and NoSQL databases. This procedure, which is frequently referred to as "data prep" or "data wrangling," is crucial for creating analytics applications and is completed by analysts, data scientists, and IT teams via self-service technologies.

2.1.2 Data Pre-Processing

The first stage in getting raw data ready for machine learning models is data pre-processing, which makes sure the data is clean and formatted. This procedure eliminates missing, null, or duplicate values that are frequently present in raw datasets. A "flood" column is added based on meteorological data, and labels are transformed into numericrepresentations for machine readability. The process entails loading the dataset, handling null values effectively, importing libraries like Numpy and Pandas, and saving the cleaned data as a CSV file for further use.

Fig 5: *Data Pre-Processing*

	SUBDIVISION	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	ANNUAL RAINFALL	FLOODS
0	KERALA	1901	28.7	44.7	51.6	160.0	174.7	824.6	743.0	357.5	197.7	266.9	350.8	48.4	3248.6	1
1	KERALA	1902	6.7	2.6	57.3	83.9	134.5	390.9	1205.0	315.8	491.6	358.4	158.3	121.5	3326.6	1
2	KERALA	1903	3.2	18.6	3.1	83.6	249.7	558.6	1022.5	420.2	341.8	354.1	157.0	59.0	3271.2	1
3	KERALA	1904	23.7	3.0	32.2	71.5	235.7	1098.2	725.5	351.8	222.7	328.1	33.9	3.3	3129.7	1
4	KERALA	1905	1.2	22.3	9.4	105.9	263.3	850.2	520.5	293.6	217.2	383.5	74.4	0.2	2741.6	0

2.1.3 Model Selection and Prediction

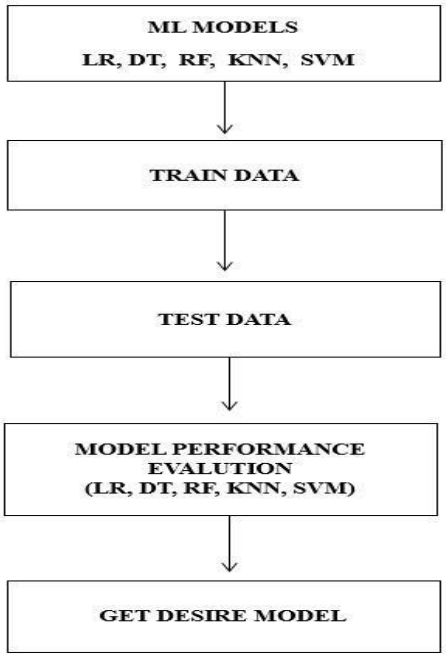


Fig 6: *ML Model Selection Process*

Selecting the optimal model for a job is known as model selection. Floods are predicted using methods such as SVM, Random Forest, KNN, Decision Tree, and Linear Regression. These

models examine input data, spot trends, and forecast outcomes. To Find patterns and learn, they are trained on datasets. The decision is influenced by variables such as the complexity, characteristics, and quantity of the dataset. It is advised to begin with a basic model, progressively add complexity, and use cross-validation and parameter adjustment to maximize accuracy.

2.1.4 Splitting Dataset into Train and Test Data

Data pre-processing involves splitting the dataset into training and test sets in order to improve model performance in machine learning.

Training Set: A subset of the dataset with known results that is used to train the model.

Test Set: A subset of the dataset used to access the model's forecasts.

Action to take: Divide the dataset into tests and train subsets at random.

- X_train: Training features
- X_test: Test-related feature
- Y_train: Training dependent variables
- Y_test: Testing dependent variables

To guarantee consistent results, use train_test_split() with: Data arrays test_size for train-test ratio random_state.

2.1.5 Classification

A supervised learning method for classifying fresh observations using training data is the classification algorithm. With classes standing in for objectives or labels, it learns from labelled datasets and divides data into categories like Yes/No or 0/1. Finding the input data's category is the main objective, especially for categorical outputs. Effective algorithms for classifying rainfall data include SVM, Random Forest, KNN, LR, and Decision Tree.

Classification Task Types:

- One of two classes is predicted by binary classification, such as spam or non-spam.
- Predicting one of more than two classes (such as flower types) is known as multi-class classification.
- Several classes are predicted for every instance using multi-label classification (e.g., tagging photographs with several labels).
- Imbalanced Classification: Addresses unequal distribution of classes (fraud detection, for example).

2.1.6 Tasks including email filtering, fraud detection, medical diagnosis, and weather forecasting all make extensive use of classification. Certain problem types are better suited for different algorithms, and in order to attain the best results, model parameters must be how well the data is classified, hence data pre-processing is essential**. Prediction Result Generation**

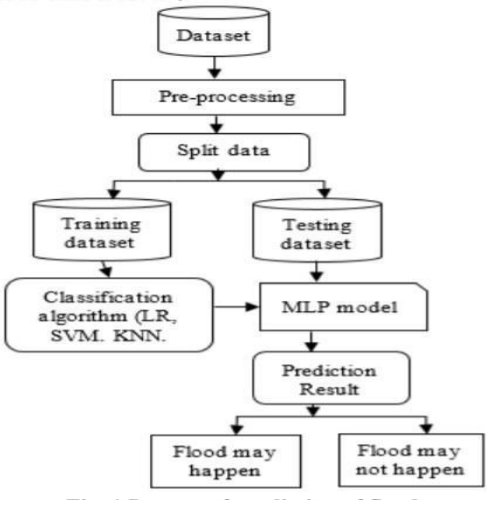


Fig 7: *Prediction Process of Flood*

LASS 1: "FLAH FLOOD MAY OCCUR":

This class is set to 1 if the input is greater than 2400mm.

CLASS 2: "FLASH FLOOD MAY NOT OCCUR": The flash flood class is set to 0 it the input is less than 2400mm.

2.3 Materials and Methods

The dataset used for the analysis was rainfall data from 1901 to 2018. The dataset is arranged by month, state, and district and is generated as a CSV file. The unit used to measure rainfall is the millimeter (mm). The dataset was collected monthly from 36 locations in the metrological department. By examining historical data, such as flood events in the past, machine learning technology can be used to forecast the future. LR, SVM, KNN, and MLP are some of the methods used to measure ML performance.

3. Result:

Using machine learning techniques, the proposed work assesses a rainfall dataset to make highly accurate prediction about flash flood. Using training and testing datasets, flood models' performance is assessed using ROC curves and statistical markers. The study's main goals are to create maps of flood-prone areas and construct machine-learning-based flood susceptibility models by examining a variety of influencing factors, such as natural and of human origin causes. The predictive power, importance, and correlations between these variables—which were found to have a major influence on flood occurrences—were examined using feature engineering. The dearth of earlier research on the region's vulnerability to flooding is addressed in this paper.

3.1 The actions listed below demonstrated that the suggested model offers a very simple and effective way to forecast floods:

Step 1: Pre-processing is done on the rainfall collection dataset.

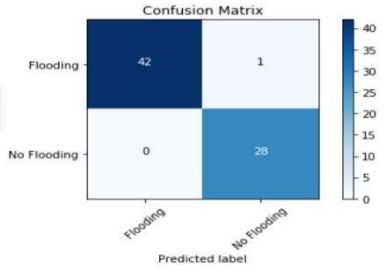
Step 2: The rainfall dataset is divided into training and testing groups at random.

Step 3: LR, DT, KNN, RF, and MLP algorithms were used to train the dataset.

Step 4: The most accurate SVM method is used to build the model, and it is verified using metrics like accuracy, sensitivity, specificity, recall, precision, and f1-score.

***Step 5*:** Feed the prediction model test data and confirm the outcomes. The algorithm's calculated accuracy.

***Step 6*:** The flood warning system also appears.



	Name	Score
0	KNN	0.833333
1	LR	0.875000
2	DT	0.666667
3	RF	0.791667
4	SVM	0.958333

Fig 8*: Prediction Label and Accuracy*

3.2 Performance Analysis of Algorithm:

The procedures listed below, along with the suggested model, offer a relatively simple and effective way to forecast flooding:

Step 1: Pre-processing is done on the rainfall collection dataset.
Step 2: The rainfall dataset is divided into training and testing groups at random.
Step 3: The dataset was trained using the LR, SVM, KNN, RF, and DT algorithms in step three.
Step 4: The model is built with the highest accuracy using the SVM method and verified using characteristics like accuracy, precision, confusion matrix, and fi-score.
Step 5: Feed the prediction model test data and confirm the outcomes.

Accuracy of the algorithm calculated from the F1 score and Confusion Matrix.

From (1) and (2), the precision measured.

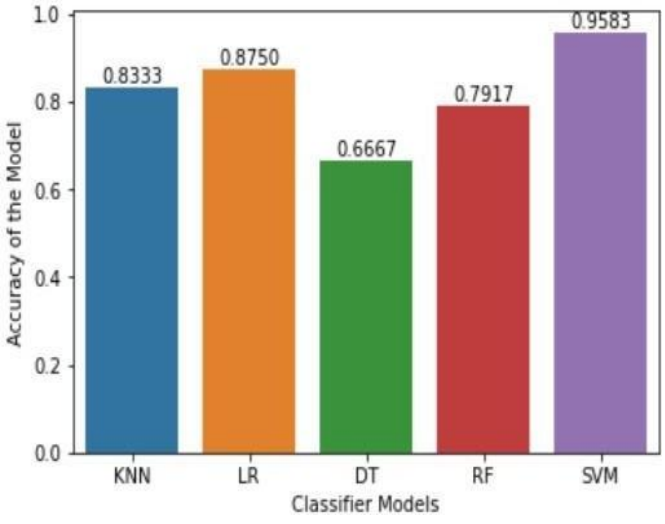


Fig 9*: Performance Analysis of Classification Algorithm*

Discussion

By examining a variety of parameters in the study area, this research paper aims to create machine learning-based flood susceptibility models that will provide maps that are prone to flooding. Both natural and of human origin element that have been show to affect flood events were taken into account. The predictive power, importance, and correlations between these characteristics were evaluated through feature engineering, and the results were utilized for modelling. Finding and assessing region-specific characteristics was crucial because there is no previous research in this field. Determining contributing elements, confirming previous flood disasters, and identifying regions that are prone to hazard were all made possible by remote sensing. Effective machine learning methods including DT, SVM, and RF were used both alone and in conjunction with IOE to create accurate flood susceptibility maps. The goal of the project is to assess rainfall data in order to make accurate flash flood forecasts.

4. Conclusion

In addition to encouraging collaboration between public institutions for better flood prediction and warning system, this initiative offers insights into the needs and readiness of local communities. By connecting stakeholders, public officials, and citizens, a cooperative platform will enable prompt flood warnings to lessen negative effects. Flash floods seriously harm people and property. A model was created to predict flash floods using rainfall data from 1901 to 2018. The pre-processed dataset was split between 80% training and 20% testing, and it was

analyzed using the SVM, Logistic Regression, KNN, Decision Tree, and Random Forest algorithms. Calculations were made for performance criteria such as sensitivity, specificity, F1 score, recall, and accuracy. SVM successfully forecasts floods based on rainfall data and reached the best classification accuracy, surpassing 90%. Disaster management agencies can use this model to help predict floods. The goal of future research is to automate prediction outputs using desktop or online applications and integrate cutting-edge AI approaches.

Reference:

1. Y. Wang, Z. fang, H. Hong, and L. Peng, "Flood susceptibility mapping using convolutional neural network frameworks," J. Hydrol., vol. 582, no. March, p. 124482, 2020, doi: 10.1016/j.jhydrol.2019.124482.

2. R. Mind'je et al., "Flood susceptibility modeling and hazard perception in Rwanda," Int. J. Disaster Risk Reduct., vol. 38, no. April 2018, p. 101211, 2019, doi: 10.1016/j.ijdrr.2019.101211.

3. W. Chen et al., "Modeling flood susceptibility using data-driven approaches of naïve Bayes tree, alternating decision tree, and random forest methods," Sci. Total Environ., vol. 701, 2020, doi: 10.1016/j.scitotenv.2019.134979.

4. R. Costache et al., "Spatial predicting of flood potential areas using novel hybridizations of fuzzy decision-making, bivariate statistics, and machine learning," J. Hydrol., vol. 585, no. December 2019, p. 124808, 2020, doi: 10.1016/j.jhydrol.2020.124808.

I. E. Olorunfemi, A. A. Komolafe, J. T. Fasinmirin, A. A. Olufayo, and S. O. Akande, "A GISbased assessment of the potential soil erosion and flood hazard zones in Ekiti State, Southwestern Nigeria using integrated RUSLE and HAND models," Catena, vol. 194, no. January, p. 104725, 2020, doi: 10.1016/j.catena.2020.104725.

5. P. T. Padi, G. Di Baldassarre, and A. Castellarin, "Floodplain management in Africa: Large scale analysis of flood data," Phys. Chem. Earth, vol. 36, no. 7–8, pp. 292– 298, 2011, doi: 10.1016/j.pce.2011.02.002.

6. I. Ajibade, G. McBean, and R. Bezner-Kerr, "Urban flooding in Lagos, Nigeria: Patterns of vulnerability and resilience among women," Glob. Environ. Chang., vol. 23, no. 6, pp. 1714–1725, 2013, doi: 10.1016/j.gloenvcha.2013.08.009

7. J. Ntajal, B. L. Lamptey, I. B. Mahamadou, and B. K. Nyarko, "Flood disaster risk mapping in the Lower Mono River Basin in Togo, West Africa," Int. J. Disaster Risk Reduct., vol. 23, no. October 2016, pp. 93–103, 2017, doi: 10.1016/j.ijdrr.2017.03.015.

8. I. Douglas, "Flooding in African cities, scales of causes, teleconnections, risks, vulnerability and impacts," Int. J. Disaster Risk Reduct., vol. 26, no. September, pp. 34–42, 2017, doi: 10.1016/j.ijdrr.2017.09.024

9. C. C. Olanrewaju, M. Chitakira, O. A. Olanrewaju, and E. Louw, “Impacts of flood disasters in Nigeria: A critical evaluation of health implications and management,” 52 Jamba J. Disaster Risk Stud., vol. 11, no. 1, pp. 1–9, 2019, doi: 10.4102/jamba.v11i1.557.