Application of Machine Learning to Flood Prediction

Preprint · February 2022

DOI: 10.13140/RG.2.2.25882.72642

3 authors:

David Otoosakyi
University of Ibadan
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Ini Adinya
University of Ibadan
12 PUBLICATIONS **35** CITATIONS
SEE PROFILE

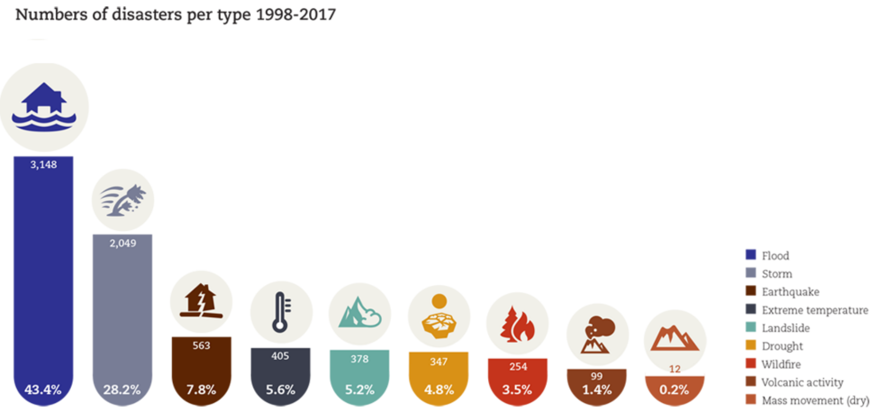
Enayon Taiwo
The University of Winnipeg
5 PUBLICATIONS **9** CITATIONS
SEE PROFILE

Application of Machine Learning to Flood Prediction

David Otoosakyi
Department of Mathematics, University of Ibadan - Nigeria

0.1 Introduction

Flood is an overflowing of a great body of water over land not usually submerged (Taiwo et al, 2019). It occurs when there is an excess of water which in turn covers a land that is usually dry. It is also regarded as the most frequent type of natural disaster which occur when an overflow of water submerges land that is usually dry (World Health Organisation – WHO). Floods are devastating natural disasters worldwide, it is a deleterious phenomenon that induces detrimental impacts on humans, properties and environment (Mind'je et al, 2019). Of all weather-related natural disasters, floods are the most common and widespread natural severe weather event which is estimated to affect 250 million people around the world every year, also costing billions of dollars in damages (Matias, 2018). Recent floods and consequences all over the world are becoming too frequent and threat to sustainable development in human settlements (Nwigwe and Emberga, 2014). According to The National Severe Storms Laboratory (NSSL), floods kill more people each year than tornadoes, hurricanes or lightning in the United States. Between 1998-2017, floods affected more than 2 billion people worldwide (WHO, 2021).



Source: CRED, UNISDR, 2018

Figure 1: Number of Disasters per type 1998-2017 [PreventionWeb, 2018]

Flooding is mostly experienced when water bodies, such as a river, lake, or ocean overflow. In such case, the water overtops or breaks the embankments built to prevent the overflow, which in turn result in some of that water escaping its normal borders. It as well occurs as a result of an accumulation of rainwater on saturated ground in an expanse of space or a region of land. Although, flooding is a natural occurrence, man-made changes to the land can also be a factor. Development does not cause flooding but can make it worse. In cities and suburbs, pavement and rooftops prevent some rainfall from being absorbed by the soil. Thus, can increase the amount of runoff flowing into low lying areas or storm drain system (ResearchClue, 2020). The occurrence of floods can be within minutes or over a long period, and may last for days, weeks, or longer.

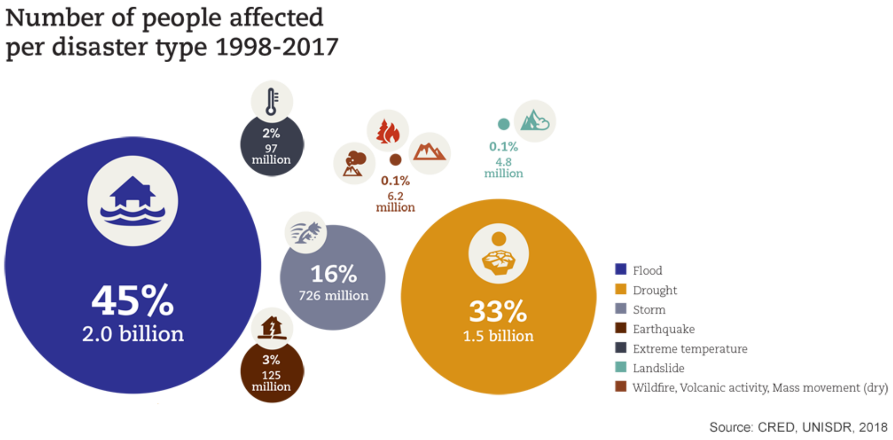


Figure 2: Number of people affected per disaster type 1998-2017
Source: [PreventionWeb, 2018]

0.1.1 Flood Types

Flood varies differently according to types, and this includes: river flood, coastal flood, storm surge, inland flooding and flash flood. From these types, we can infer that flooding can occur anywhere, both inland and coastal areas. Of these types, the common ones are flash floods, river floods and coastal floods.

Flash floods are termed to be the most dangerous kind of floods, because they combine the destructive power of a flood with incredible speed (NSSL, n.d). They are caused by rapid and excessive rainfall that raises water heights quickly, and rivers, streams, channels or roads may be overtaken. River floods are caused when consistent rain or snow melt forces a river to exceed its capacity, that is, water levels rise over the top of river banks. Coastal floods are caused by storm surges associated with tropical cyclones and tsunami. It is typically a result of a combination of sea tidal surges, high winds, and barometric pressure. When flash flood occurs – being the most common type of floods, some areas such as densely populated areas, areas near rivers and dam failures are very much at risk. Other areas include mountains and steep hill, which produce rapid runoff (as a result of high elevation/slope) and causes streams to rise quickly; rocks and shallow, clayey soils, which do not allow much water to infiltrate into the ground. In addition, Saturated soils and dry soils can also lead to rapid flash flooding as a result of not being able to absorb any more water due to very intense rainfall.

0.1.2 Problem Statement

Luo et al (2015) listed top 15 countries with greatest population that are exposed to river flood risk. These countries (spanning through three continents) are seen to be either developed or developing countries, and are the most vulnerable to natural disasters and climate change. In Africa, flood crisis has been one of the major disasters that befall Nigeria, claiming lives and properties and hindering man's social interaction with his environment and also food shortage.

One of the challenges posed by flood is to determine the next occurrence and the extent of such occurrence. It is therefore imperative that a prediction of such be carried out so as to forestall adequate preventive mechanisms and measures. Before now, several models (numerical and physical) have been developed for flood prediction of the areas prone to flood of which some are deterministic and others are stochastic models. Outcome of deterministic models is certain where the input parameters are fixed but less informative when compared to stochastic model which is suitable for forecasting based on the probability of outcome under varying circumstances. This research, therefore, focuses on implementing a data-driven model using machine learning algorithms with better predictive power that helps predict the location and extent of floods since data-driven method of prediction have been found to provide better insight through assimilation of measured climatic indices and hydro-meteorological parameters and are quicker to develop with minimal inputs (Mosavi et al, 2018).

0.1.3 Research Objective

As a means of mitigating the hardship caused by flood, this research is aimed at building a machine learning model from historical data that helps predict the next location susceptible to flooding and the extent of floods in such location. Since this approach is data-driven, the focus is on machine learning techniques in the analysis of the system's input and output variables rather than having a knowledge of the physical behaviour of the system.

0.2 Overview of Machine Learning and Review of Literature

We consider different works that are flood related by several authors in different continents of the world, specifically the review of the flood cases in countries that are mostly faced by flood hazards in major continents of the world and the models employed in predicting subsequent occurrences.

0.3 Overview of Machine Learning

Machine Learning (ML) is a subfield of Artificial Intelligence (AI) that employs intuitive training to understand patterns in a dataset as part of an algorithmic and heuristic approach, allowing for easier implementation with low computation costs, as well as fast training, validation, testing, and evaluation, with high performance compared to physical models and relatively less complexity. (Mosavi et al, 2018; Zehra, 2020). It is a collection of methods that enable computers to automate data-driven model building and programming through a systematic discovery of statistically significant patterns in the available data (Bhavsar et al, 2017). Machine learning, as defined by Arthur Samuel in 1959, is a ”field of study that offers computers the ability to learn without being expressly programmed.” Machine learning algorithms became conceivable with the advancement of computing and communications technologies, making it possible to find increasingly complex and hidden patterns in data. Recent trends in computational designs especially in machine learning have drastically increased the capability of empirical models (Vankatesan and Mahindrakar, 2019). Furthermore, models that can automatically adapt to larger and more complicated datasets are now possible, allowing decision-makers to evaluate the implications of numerous plausible scenarios in real time. Machine learning algorithms comb through several datasets and employ complex algorithms to find patterns, make choices, and/or forecast future trends.

0.3.1 Some Machine Learning Algorithms

Machine learning encompasses a variety of approaches and algorithms, some of which were developed long before the phrase "machine learning" was used, and academics are still refining and developing new and efficient ways today (Bhavsar et al, 2017). Machine learning methods are categorised based on how the algorithms learn patterns from data. These categories are majorly: Supervised Learning and Unsupervised Learning. The goal of a learning process is to find a function that minimises a risk of prediction error that is expressed as a difference between the real and computed output values when tested on a given dataset, and this is controlled by a predetermined threshold.

Supervised learning entails guiding the learning process with labelled data, that is, training a function (or algorithm) to compute output variables from provided data that has both input and output variables. The unsupervised learning depends only on the underlying unlabelled data to identify hidden patterns of data instead of inferring models for known input-output pairs.In this type of learning, clustering and association are the most common methods. The clustering technique involves grouping data in multiple clusters based on similarities between the data points of the given datasets. This approach relies on mathematical models to identify these similarities between the data points. A common approach is the Euclidean distance

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad \forall\, x, y \in \mathbb{R}^n$$

with $x = (x_1, x_2, \cdots, x_n)$ and $y = (y_1, y_2, \cdots, y_n)$ being input and output data points respectively.

Association method focuses on finding rules that connect data patterns with each other. The type of machine learning algorithms may vary from regression and classification to complex neuro-fuzzy systems (Bhavsar et al, 2017). Machine learning algorithms are successful at predicting floods and non-floods because they can take data from a variety of sources and categorize and regress it into flood and non-flood classes and has, therefore, emerged as a preferred instrument for delving into non-linear systems and exploring automatically generated predictions of flash floods, for example, thanks to advances in computing and algorithms (Zehra, 2020). Some selected machine learning algorithms that are found to be implemented in several open-source and commercial products are presented in the following subsections.

Gradient Boosting Algorithms

Gradient boosting is a ML technique that can be used for a variety of applications, including regression and classification. It returns a prediction model in the form of an ensemble of weak prediction models, most commonly decision trees. The resulting approach is called gradient-boosted trees when a decision tree is the weak learner; it usually outperforms random forest. A gradient-boosted trees model is constructed in the same stage-wise manner as other boosting approaches, but it differs in that it allows optimization of any differentiable loss function (Hastie et al, (2009); Piryonesi et al, (2020); Piryonesi et al, (2021))

Gradient boosting, like other boosting approaches, iteratively merges weak "learners" into a single strong learner. In least-squares regression setting, the gradient boosting method "teaches" a model F to predict values of the form $\hat{y} = F(x)$ by minimising the mean square error

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

where:

i iterates over n number of training set

$\hat{y}_i$ is the predicted value

y_i actual output

n number of samples in y

Support Vector Machine (SVM)

Support Vector Machine is a kernel-based function based on Vapnik-Chervonenkis (VC) theory - a learning theory that seeks to quantify the capability of a learning algorithm - whose output varies based on the parsed kernel functions used in the course of training a model. The goal of the kernel function is to make the primary inputs linearly separable in mapped high dimensional feature space. SVM has the potential of avoiding overfitting and can minimise estimation errors and model dimensions simultaneously with good generalisation (Vankatesan and Mahdrakar, 2019).

In developing a SVM model, the initial step entails the selection of support vectors and determination of weights. Our optimization objective is to maximize the margin - the distance between the separating hyperplane (decision boundary) and the training samples that are closest to this hyperplane, which are the so-called support vectors. By increasing the distance between the plane and the nearest input data points, support vector machines find a hyperplane (that classifies data). This is accomplished by minimizing the weight vector w that is used to define the hyperplane, relying on optimization theory and specific assumptions (Raschka and Mirjalili, 2017; Bhavsar et al, 2017)

Decision boundaries with big margins have a reduced generalization error, whereas models with short margins are more prone to overfitting. To understand the margin maximization, we consider the positive and negative hyperplanes which runs parallel to the decision boundary represented as linear equations below:

$$w_0 + w^T x_{pos} = 1 \tag{1}$$

$$w_0 + w^T x_{meg} = -1 \tag{2}$$

Subtracting the two linear equations (1) and (2) from each other, we get:

$$\Rightarrow \boldsymbol{w}^T \left(\boldsymbol{x}_{\text{pos}} - \boldsymbol{x}_{\text{neg}}\right) = 2$$

Normalising this equation by the length of the vector $\mathbf{w}$, defined as follows:

$$\|\boldsymbol{w}\| = \sqrt{\sum_{j=1}^{m} w_j^2}$$

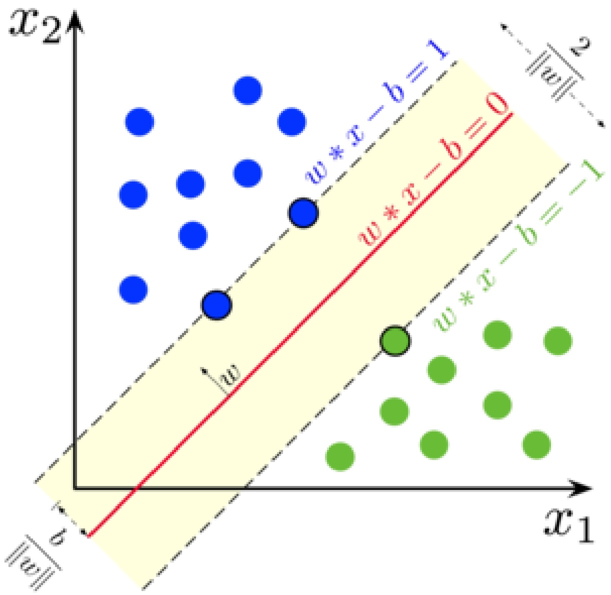


Figure 3: Concept of Support Vector Machine (SVM)
Source: Wikipedia, 2021

the following equation is arrived at:

$$\frac{w^T\left(x_{pas} - x_{neg}\right)}{\|w\|} = \frac{2}{\|w\|}$$

Given the set of labeled data points (input-output pairs) $S = \{(x_i, y_i)\}_{i=1}^n$, where $y_i \in \{-1, +1\}$ is the class label of high-dimensional point x_i, i.e., $(x_i, y_i) \in R^{d+1}$, and d and n are the numbers of features and labeled data points, respectively. In binary classification problem, points labeled with $+1$, and -1 belong to classes $\mathbf{C}^+$, and $\mathbf{C}^-$, respectively, i.e., $S = \mathbf{C}^+ \cup \mathbf{C}^-$.
The optimal classifier (also known as soft margin SVM) is determined by the parameters w and b through solving the convex optimization problem by quadratic programming:

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^n \xi_i \\ \text{subject to} & y_i\left(w^T x_i + b\right) \geq 1 - \xi_i, \quad i = 1, \ldots, n \\ & \xi_i \geq 0, \quad i = 1, \ldots, n \end{array}$$

with corresponding linear prediction function $f(x) = wx + b$, and can be used in estimating the linear regression of Support Vector Machine. The magnitude of penalization for misclassification is controlled by the parameter C and slack variables ξ_i.

Neural Networks

Neural Networks (NN) are statistical learning algorithms originally inspired by biological neural networks that can be used to estimate or approximate non-linear functions with an arbitrary number of inputs (Carvalho and Camelo, 2015). Neural networks have been widely utilized to assess flood in a threatened area of a river and its impact outside of that area.

The most common type of neural networks family used in flood prediction is the Artificial Neural Networks (ANN). The core unit of the ANN, like the nervous system, is a neuron,

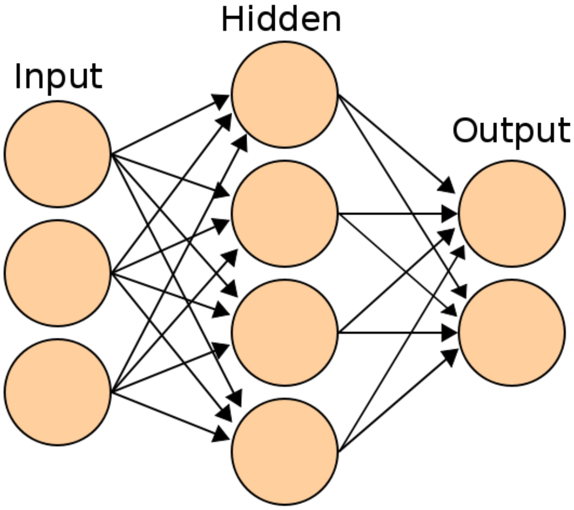


Figure 4: A neural network representation
Source: Wikipedia

which uses a "transfer function" to calculate output for a given input. These neurons are linked together to form a data-flowing network - input layer to other procession layers to output layer. The connections are weighted connections that scale the data flow, and each neuron's transfer functions map inputs to outputs. ANN has numerous benefits, including the ability to learn, manage extremely complicated nonlinear systems, and work in parallel. It is a versatile computation framework for nonlinear modeling that can be used in a variety of applications. The adjustable architecture allows the number of layers and neurons in each layer to be easily altered. The qualities of the data have a big impact on the network setup. (Bhavsar et al, 2017; Buyuksahin and Ertekin, 2019)

A feedforward artificial neural network with an input layer, one or more hidden layers, and an output layer is known as a multilayer perceptron. The predictor values are received by the input layer, and the prediction is provided by the output layer. To recognize features, hidden layers combine input predictors. (Ganguly et al, 2018). At each neuron, the general relationship between the x inputs and y output are expressed as:

$$y_k^{(i)} = \theta \times \left[\sum_{j=1}^{n} x_j^{(i)} \times w_{jk}^{(i)} - th_k \right]$$

Where:

$y_k^{(i)}$ is the output of neuron k in the i^{th} iteration.

$x_1, x_1, \cdots, x_n$ are the inputs from the previous layer.

$w_{jk}^{(i)}$ is the weight between input x_j and output neuron y_k

th_k is the threshold and θ is the activation function.

The activation (transfer) function is typically a non-linear function such as radial basis function (rbf)

$$\mathcal{K}\left(x^{(i)}, x^{(j)}\right) = \exp\left(-\frac{\left\|x^{(i)} - x^{(j)}\right\|^2}{2\sigma^2}\right) = \exp\left(-\gamma\left\|\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\right\|^2\right)$$

(where, $\gamma = \frac{1}{2\sigma^2}$ is a free parameter that is to be optimized);
or a sigmoid function

$$y_{sig} = \frac{1}{1+e^{-x}}$$

The ANN algorithm's learning (or training) process entails finding the weights using various methods. The backpropagation learning rule is a prominent method that is based on gradient descent error minimization. By adjusting the weights and biases, the backpropagation neural network (BPNN) propagates the error from the output layer to the input layer and improves accuracy. After each training epoch, the weights are usually changed to reduce the difference between the target and projected output.

$$e_o^{(i)} = y_{od}^{(i)} - y_o^{(i)}$$

Here, $y_{od}^{(i)}$ is the desired output and $y_o^{(i)}$ the output produced by the network in the output layer at iteration i. The weight at the output layer is updated using this error with Equation 2.3,2.4 and 2.5

$$w_{jo}^{(i+1)} = w_{jo}^{(i)} + \delta w_{jo}^{(i)} \tag{3}$$

$$\delta w_{jo}^{(i)} = \alpha \times y_j^{(i)} \times \delta_o^{(i)} \tag{4}$$

$$\delta_o^{(i)} = y_o^{(i)} \times \left(1 - y_o^{(i)}\right) \times e_o^{(i)} \tag{5}$$

Weights for the hidden layer, on the other hand, are updated as follows:

$$w_{sj}^{(i+1)} = w_{sj}^{(i)} + \delta w_{sj}^{(i)}$$

$$\delta w_{sj}^{(i)} = \alpha \times x_s^{(i)} \times \delta_j^{(i)}$$

$$\delta_j(i) = y_j^{(i)} \times \left(1 - y_j^{(i)}\right) \times \sum_{k=1}^{n} \delta_k^{(i)} \times w_{jk}^{(i)}$$

where, $x_s^{(i)}$ corresponds to the input of the s^{th} neuron in the i^{th} iteration and $y_j^{(i)}$ is the output of the neuron.Where n is the number of neurons in the output layer, the sum indicates that errors from the output layer are backpropagated.

Regression models can be "reduced" from neural networks. It's "reduced" in the sense that it can "pretend" to be any type of regression model. It can adapt to both linear and non-linear models, depending on how it was trained. It's a versatile model that can readily adjust to the data's shape. If the findings aren't precise enough, more hidden neuron layers can easily be added to improve the predictability and make the system more sophisticated.

0.3.2 Flood Vulnerability Assessment

Vulnerability assessment entails the process of identifying problems, quantifying it, and assessing the rate of risk involved in formulating developmental strategies to reduce the risk and

Countries	Population (in millions)	Continent
India	4.84	Asia
Bangladesh	3.48	Asia
China	3.28	Asia
Vietnam	0.93	Asia
Pakistan	0.71	Asia
Indonesia	0.64	Asia
Egypt	0.46	Africa
Myanmar	0.39	Asia
Afghanistan	0.33	Asia
Nigeria	0.29	Africa
Brazil	0.27	South America
Thailand	0.25	Asia
Congo D.R.	0.24	Africa
Iraq	0.19	Asia
Cambodia	0.19	Asia

Table 1: Annual Expected Population Affected by River floods
Source: World Resources Institute [WRI], 2015

vulnerabilities (Sudha Rani et al, 2015). Over the years, flood vulnerability has been as a root cause of human devastation, leading to food insecurity, outbreak of waterborne diseases, damages to crop productivity and infrastructure (Rehman et al, 2019). Luo et al (2015) ranked 164 countries of the world by the number of people affected by river flooding. It was discovered that the top 15 countries on the rank (**??**) accounted for about 80% of the total population that are affected by flood every year. These countries that made the top 15 rank were at least considered to be either developing or developed, and are the most vulnerable to natural disasters and climate change. The top 15 countries belong to 3 continents of the world – Africa, Asia and South-America – with most of these countries being Asian countries. Works by some authors around the leading countries in the continents listed are thus reviewed in the following subsections.

0.3.3 Asia - India and Bangladesh

Several studies in time past have shown that floods are a serious risk for hundreds of millions of people in India and Bangladesh. In terms of disasters caused by floods, Bangladesh is considered one of the highly flood-affected countries (Bhuiyan and Bak, 2014). At a point, about one-third of Bangladesh – a delta nation – was underwater, the country being endemic to monsoon flooding (Coca, 2020). Flood events over time indicate an increase in the intensity and frequency of flood risk and river erosion in the coming years in Bangladesh. Roy and Blaschke (2013) developed a methodology for spatial vulnerability assessment of floods in the coastal regions of Bangladesh using a GIS weighted overlay of 44 indicators. The resulting maps and figures reveal both the extent and levels of vulnerabilities. A similar study conducted by Hoque et al (2019) in Kalapara Upazila of Patuakhali district in Bangladesh, a coastal area, revealed that flooding is a frequent event in that area, with substantial rainfall experienced in most months of the year of an average annual rainfall of 2645mm, which further indicated that the vulnerability to flooding of human and all types of resources are high due to the area's lowland, geographic location, and dense population.

In many parts of India, floods have been found as the most occurring natural disaster, thereby causing massive wreck to lives and properties, and resulting in ecological and socioeconomic vulnerabilities. Floods in India are majorly caused by unexpected rainfall during the southwest monsoon, siltation in riverbeds, inefficiency of rivers to carry heavy discharge intensification of tropical storms and depressions (Rehman et al, 2019). In October 2021, Mukteshwar village in Nainital district in India experienced about 24 hours of rainfall estimated to be 341mm, and another in Pantnagar town in Udham Singh Nagar district which was estimated to be 404mm (Davies, 2021). Averagely each year, about 84% of India's estimated Gross Domestic Product (GDP) is affected by floods (India Today, 2015). Sowmya et al (2015) applied a multi-criteria evaluation approach in the analysis and assessment of urban flood vulnerability zoning of Cochin city, southwest coast of India using remote sensing and GIS and were able to map out three vulnerability zones with the very high and high zones constituting about 9% of the total area of the city. Their findings reveal that the major factors contributing to flooding in vulnerable areas are blockage of drainage channels and the nearness to coastal waters. A study by Ghosh and Kar (2018), applied the Analytical Hierarchy Process (AHP) for flood risk assessment in Malda district of West Bengal India, incorporating flood hazard elements and vulnerability indicators in GIS environment. The result of the study revealed that northern and western parts of the district are at the forefront of flood hazards, unlike the eastern part.

0.3.4 South America

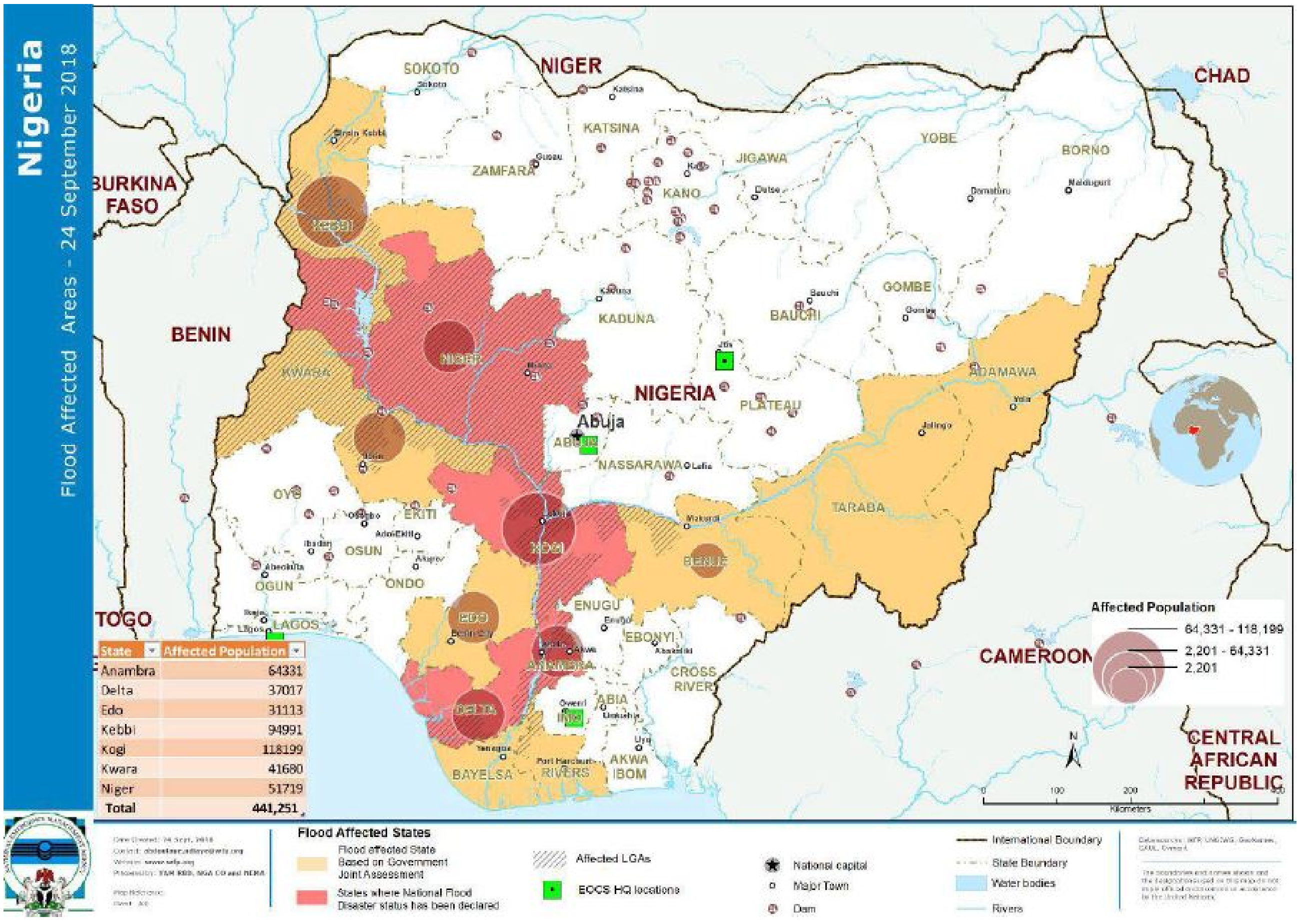
Several reports have it that South America is one of the continents that are not spared by hazardous floods. In 2008, persistent heavier-than-normal rain triggered deadly flooding in parts of South America, whose effect was enhanced by the cooler-than-normal ocean surface temperatures in the central and eastern Pacific that are associated with La Niña. To this end, the news reported that there were 52 fatalities in Bolivia, 19 in Peru, and 16 in Ecuador (National Aeronautics and Space Administration [NASA], 2008). The torrential rain of 2013 which caused flooding to hit Brazil and Mexico claimed at least 30 lives in two south-eastern states in Brazil and more than one hundred people were forced to evacuate their homes in Mexico (Ritorto, 2013). In 2015 through 2016, the severe weather brought by the El Niño Southern Oscillation (ENSO) including heavy downpours, floods, flash floods and landslides massively hit South America with countries like Venezuela, Columbia, Ecuador, Peru, Brazil, Bolivia, Paraguay, Uruguay, and Argentina being the most affected, accounting for the displacement of thousands of people in each country (Reliefweb, 2015). In 2017, Bolivia, Peru, and Chile were as well badly affected by floods, destroying home and properties. This was caused by torrential downpours which caused rivers to overflow in central Bolivia and southern Peru (Aljazeera, 2017). Paraguay, Peru, Ecuador and Bolivia were not left out of the 2019 South America floods triggered by heavy rain. About 70,000 people were affected by floods in several departments of Paraguay. The same caused a severe damage to dozens of homes in Peru, which prompted evacuations in the regions of Ancash, Amazons and Cusco. A state of emergency in several cantons was declared by the authorities in Ecuador due to flooding in Los Rios province. Over 2000 hectares of crops and 109 homes were destroyed by flooding which overwhelmed Parapeti River in Santa Cruz Department of Bolivia (FloodList, 2019). Martini (2020) reported that heavy downpours battered parts of South America, leaving hundreds of people to be evacuated from the northern parts of Argentina due to widespread flooding. In the same vein, Brazil and Paraguay was hit by flash flooding, which caused widespread destruction. Widespread flooding inundated Guyana – a small South American country – in 2021, swamping roads, homes and farmlands throughout the country with over 6900 households being severely affected by the devastating floods, reaching more than halfway up two-story houses. The overflow of a river in Madre de Dios Department in south-eastern Peru in February 2021 resulted in extensive flooding in which according to the United Nations Office for the Coordination of Humanitarian Affairs (UN OCHA) more than 6600 people were affected and about 1700 houses damaged and destroyed including 9 health centres. (Levenson, 2021; Reliefweb, 2021). Brazil's Amazon rainforest were struggling to cope with severe flooding after torrential downpour which caused nearby rivers to swell above normal level causing about 24,000 families homeless (Deutsch Welle [DW], 2021). Kinghorn (2017) outlined three major factors that make flooding in South America worse which are: global warming, rapid urbanisation and land use choices.

0.3.5 Africa - Nigeria

In recent decades, countries across Africa have experienced an increase in the frequency and severity of floods, Nigeria not excluded. Geopolitical zones like Northeast, Northcentral, South-south, and Southwest Nigeria are at the forefront of this disaster. States like Sokoto, Rivers, Plateau, Oyo, Osun, Ondo, Ogun, Niger, Yobe, Nasarawa, Lagos, Kwara, Kogi, and Kebbi are victims of flood in the year 2020 (National Emergency Management Agency [NEMA], 2020). In fact, no city or town of the southwestern Nigeria is absolutely free from floods in any year (Nwigwe and Emberga, 2014). Affected citizens have become homeless and internally displaced. The effect was as well significant on food production due to erosion, with ground saturated with water, hindering usual farming activities. Education was also affected as flooded schools are closed.

In 2012, Nigeria lost 363 people to flood while over 2.3 million people were displaced (Echendu, 2020). According to NEMA, 30 of Nigeria's 36 states were affected by the floods. The floods were termed the worst in 40 years and affected about seven million people. The significance of the year 2012 flood disaster in Nigeria lies in the fact that they were unprecedented in the past forty years. Most parts of the central states of Nigeria and other adjoining states along river Niger and Benue are devastated by these floods causing huge destruction to the rural and urban infrastructures (farmlands/crops, road, buildings, damages, bridges, powerlines etc) and socioeconomic lives of the areas. The estimated damages and losses caused by the floods were N2.6 trillion (Taiwo et al, 2019; ResearchClue, 2020). WHO categorised the 2012 flood disaster in Nigeria as the worst flood to have hit the country in the past 50 years (WHO, 2012). Flood impact analysis in Nigeria is not majorly categorized by gender but result shows that women and girls are mostly vulnerable to effects caused by natural hazards (Lucas, 2021)

A flood vulnerability assessment of communities in three of the Niger Delta states relative to the 2012 flood disaster in Nigeria using spatial data and GIS found out that a total of 1,110 towns were at risk of being inundated and about 7,120,028 people risk displacement (Amangabara and Obenade, 2015). A similar study by Okwu-Delunzu et al (2017) aimed at spatial assessment of flood vulnerability in Anambra East and environs using Remote Sensing (RS) and GIS to map flood prone area and modeling a digital elevation showed that 71% of the study area was liable to flooding.



State	Affected Population
Anambra	64331
Delta	37017
Edo	31113
Kebbi	94991
Kogi	118199
Kwara	41680
Niger	51719
Total	441,251

Figure 5: NEMA Situational report - Flood Affected States and flood risk
Source: European Commission Joint Research Centre

In 2016, 19% of respondents to a survey conducted by the Lagos State Government revealed that they had experienced flooding in years past (Lagos Bureau of Statistics, 2016), and most residents perceived flood, which occurs almost annually, as the second most hazardous disaster in the state. Another study by the World Bank in 2020 estimated the flood damage in some selected states in Nigeria to include Lagos, Delta and Cross River states, and the result shows that only in Lagos State, estimates of flood damage is about $3.992B annually, which is around 4% and 1% of the state's GDP and National GDP respectively (Croitoru et al, 2020). Flooding in Nigeria is not only limited to the natural course but also human induced (Echendu, 2020), poor or non-existent drainage systems, poor waste management system, unregulated urbanization, weak implementation of planning laws; and corruption – a political factor.

0.4 Scholarly Articles on Flood Prediction using Machine Learning

Recent researches on flood forecasting and prediction leverage on the predictive power of several machine learning algorithms that learn patterns from historical data. Over time, machine learning algorithms such as Decision Tree (DT), Random Forest (RF), Linear Regression (Linreg) Logistic Regression (LR), Extreme Gradient Boosting (XGBoost), K-Nearest Neighbour (KNN), Support Vector Machine (SVM) and Artificial Neural Network (ANN) have been implemented in flood prediction which in turn yields a reliable outcome. The choice and selection of suitable machine learning algorithm to apply in flood modelling and prediction is often restricted to data availability, quantity and quality, study objectives and to the scope of the study (Obarein et al, 2019).

Mind'je et al (2019) implemented a logistic regression model using Remote Sensing (RS) data and GIS over a flood inventory generated using 153 historical flood locations in Rwanda with a total of ten predictors (independent variables). 75% of the entire data was used in training the model while the rest was the validation set. The result indicated two features out of the ten – Normalised Difference Vegetation Index (NDVI) and Rainfall – to be the most influencing variables, while the validating set disclosed 79.8% accuracy of the prediction with Area Under Curve (AUC) as the evaluating metric. Talukdar and Pal (2020) modelled flood plain wetland transformation in consequences of flow alteration in Punargbhaba river in India and Bangladesh using Markov Chain Cellular Automata (MC-CA) and ANN with optimised parameters (Hidden layers =7, Activation function = 7, Learning algorithm = Back propagation; Learning rate = 0.2, Momentum = 0.22) that generated the best forecasting model. The accuracy of the model was assessed on wetland simulations of 2017 pre and post monsoon and a Receiver Operating Characteristics (ROC)-AUC curve values were 84.4% and 86.8% respectively. Talukdar et al (2020) as well employed four novel ensembles of machine learning algorithms with bagging technique – REPtree, Random Forest (RF), M5P, and Random tree – on twelve predictors, with the ROC curve as the evaluation metric. The models returned over 85% AUC, but the bagging with M5P model being the highest performing model. Rahman et al (2019) implemented Artificial Neural Network (ANN), Logistic Regression, Frequency Ratio (FR) and AHP, integrating statistical, machine learning and multi-criteria decision analysis in flood susceptibility assessment in Bangladesh on 475 data points with 9 predictors. The predictive power of the models were evaluated using AUROC with logistic regression model scoring the highest success rate (86%) and a prediction rate of 81.6%.

Cui and Cui (2020) implemented a linear regression model using four parameters - Minimum Temperature, Yesterday Temperature, Precipitation and Snow on Ground - to model spring flood in New Brunswick, Canada. The model was evaluated using R^2 which was 63.0%, with all the parameters being statistically significant.

0.5 Analysis of Flood Data and Probability Distributions

In this section, we analysed the precipitation data used for this research and examine the pattern of the probability distributions with histogram plot and kernel density estimation. Then we shall find the most suitable probability distribution for each of the precipitation features.

0.5.1 Data and Source

In predicting flood, the World Meteorological Organization (WMO) outlined necessary data requirements as features to be obtained in the modelling process depending on the particular nature of a flood warning system and its objectives. These features majorly include:

- **Hydrological data** e.g. measurement of river flow, etc
- **Meteorological data** e.g. precipitation amount, rainfall intensity and duration, etc.
- **Topographical data** e.g. elevation, slope, soil properties, land cover type, etc

The data for this study is a real-world secondary data of historical flood obtained from the repository of Zindi – the first data science competition platform in Africa. The dataset consists of major flooding that hit Southern Malawi with cyclone Idai in 2015 and 2019. The map of the location was broken up into approximately $1km^2$ rectangles, assigned with a target value which is a fraction (percentage) of that rectangle that was flooded in 2015, thus making the dataset to consist of 16466 rows (entries). For this research, we only made use of the 2015 flood extent data provided in the dataset. 80% of the dataset was used for training the machine learning model while the remaining 20% was used to measure the accuracy of the model. The following features (variables) are provided in the dataset:

- **Elevation** - the mean elevation over the rectangle, based on the NASA Shuttle Radar Topography Mission (SRTM) Digital Elevation 30m dataset in Google Earth Engine.
- **Dominant Land Cover Type**
- **Weekly precipitation** - Historical rainfall data for each rectangle, for 18 weeks beginning 2 months before the flooding based on the Tropical Rainfall Measuring Mission (TRMM) dataset in Google Earth Engine
- **Coordinates** - the longitude(Y) and the latitude (X) of the location, representing a rectangle 0.01 degrees on each side, centered on that $X - Y$ location.
- **Target** - the percentage of the given rectangle that was flooded, with a value between 0 and 1.

0.6 Data Analysis and Methods

0.6.1 Analysis of Precipitation Data

A major contributing factor to flood is rainfall (precipitation). The majority of existing flood prediction models are focused on severe rainfall and hurricanes (Cui and Cui, 2020). Hence it is important to consider this. In developing flood prediction models, rainfall data from weather stations or via remote sensing dataset are utilised. Some researchers are of the opinion that gridded and modeled rainfall data is often unable to effectively capture climate variability compared to station data which may cause some uncertainties when modeling flood susceptibility (Obarein et al,2019; Mind'je et al, 2019). Yet others are of the opinion that remotely sensed information about rainfall is the only source of reliable data despite that precipitation gauges are considered the standard for measuring precipitation. The precipitation data in the dataset used in this study is a remotely sensed data obtained from the Tropical Rainfall Measuring Mission (TRMM) dataset provided by NASA in Google Earth Engine. The assessment of the pattern is considered in the following subsection.

Function	**Notation**	**PDF**	**Mean (E[X])**	**Variance (Var[X])**
Normal	$\mathcal{N}\left(\mu, \sigma^2\right)$	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\mu - x)^2}{2\sigma^2}}$	μ	σ^2
Gamma	$\Gamma(k, \theta)$	$f(x; k, \theta) = \frac{x^{k-1} e^{-x/\theta}}{\theta^k \Gamma(k)} \quad \text{for } x > 0 \text{ and } k, \theta > 0$	$k\theta$	$k\theta^2$
Weibull	$Weibull(\lambda, k)$	$f(x; \lambda, k) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}, & x \geq 0 \\ 0, & x < 0 \end{cases}$	$\lambda\Gamma\left(1 \dashv \frac{1}{k}\right)$	$\lambda^2\left[\Gamma\left(1 + \frac{2}{k}\right) - \left(\Gamma\left(1 + \frac{1}{k}\right)\right)^2\right]$
Lognormal	$Lognormal(\mu, \sigma^2)$	$\frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln(x) - \mu)^2}{2\sigma^2}\right)$	$e^{\mu + \frac{1}{2}\sigma^2}$	$e^{2\mu + \sigma^2}\left(e^{\sigma^2} - 1\right)$

Table 2: Distribution table for common probability distributions

0.6.2 Distribution Pattern of the Precipitation Data

Probability distributions are a fundamental concept in probability theory that is used both theoretically and practically. Each possible outcome of a random experiment or event is represented by a probability distribution. In machine learning models, continuous probability distributions are frequently employed, particularly in the distribution of numerical input and output variables for models, as well as in the distribution of model errors.

Some commonly used probability distributions in rainfall analysis are Normal, Log-normal, Gamma, Gumbel and Weibull. A relatively common probability distribution is the Normal distribution. It's a widely used probability distribution. Because of the central limit theorem and its capacity to describe many natural occurrences, it is particularly essential in statistics. In probability theory and Statistics, the gamma distribution is a member of two-parameter family of continuous probability distribution. The common 'exponential' distribution and the 'chi-squared' distribution are special cases of the gamma distribution. The Weibull distribution is a continuous probability distribution. A positive random variable X is normally distributed if $\ln(X)$ is normally distributed.

The characteristics of some of these probability distributions are examined in the following sections.

0.7 Summary Statistics of the Precipitation Data

The precipitation data were analysed using Python programming language. The descriptive statistics for each week is shown below (in 2 decimal places):

Precipitation Week	**Min. Precip.**	**Max. Precip.**	**Mean (μ)**	**Std. Dev. (σ)**
Week 1	0.00	19.35	1.61	4.23
Week 2	0.00	41.02	2.50	8.63
Week 3	0.00	22.02	1.16	4.40
Week 4	1.41	18.87	8.27	4.27
Week 5	3.58	23.04	8.89	3.76
Week 6	1.25	21.75	9.57	4.52
Week 7	7.46	62.43	22.92	13.69
Week 8	15.65	51.20	28.11	7.79
Week 9	30.45	105.28	58.86	16.81
Week 10	0.00	11.10	1.25	1.97
Week 11	14.97	53.01	34.65	7.46
Week 12	13.26	44.34	28.32	8.05
Week 13	0.46	28.56	12.49	7.06
Week 14	0.28	15.72	3.80	2.67
Week 15	6.73	36.97	17.07	6.07
Week 16	3.28	25.71	9.11	4.57
Week 17	0.00	4.95	0.33	1.01

Table 3: Descriptive statistics of precipitation data

The mean precipitation for each week is computed using:

$$\mu = \frac{\Sigma x_i}{N}$$

and the standard deviation from the mean as:

$$\sigma = \sqrt{\frac{\Sigma(x_1 - \mu)^2}{N}}$$

?? shows that the maximum average precipitation was experienced in the middle of the weeks considered (Week 9) with an average precipitation amount of $58.86mm$. Following this is a sharp decline in ‘Week 10‘ with an average rainfall value of $1.25mm$, which tend to be an indicator that the amount of rainfall will be declining in the following weeks before the flood but this is not so in the following weeks (Week 11 and Week 12). The last week (Week 17) before experienced the least average precipitation for the entire period in the dataset with an average precipitation of $0.33mm$, which was barely captured in Figure 6.

Although, the summary statistics of the precipitation features gave a pointer about varying patterns for the precipitations, it is also of interest to understand the probability distribution of the rainfall for each week in all the locations. This will enable us determine if a change occurred in the distributional pattern of the precipitations over the weeks before the flood commenced.

A histogram plot for each week (Figure 7) showing the patterns of the precipitation was considered, and by facial inspection, it is obvious that the precipitation variables exhibited different probability distribution. Most of the precipitation data exhibited a positively skewed distribution. Weeks 1,2,3,10 and 17 seems to have similar shape of the distribution of precipitation data

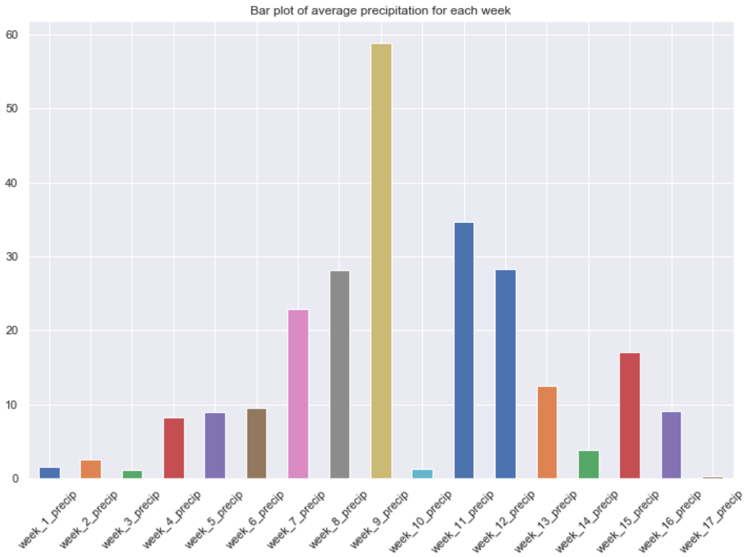


Figure 6: Bar plot of Average Precipitation for each week

which are highly positively skewed.

Since it is barely possible to get the exact probability distribution of a random variable with a histogram plot, it is therefore necessary to find the estimate of the probability distribution of each variable for a better insight of the behaviour of the data. One of the ways of achieving this is by estimating the distribution using a Kernel Density Estimator (KDE) function.

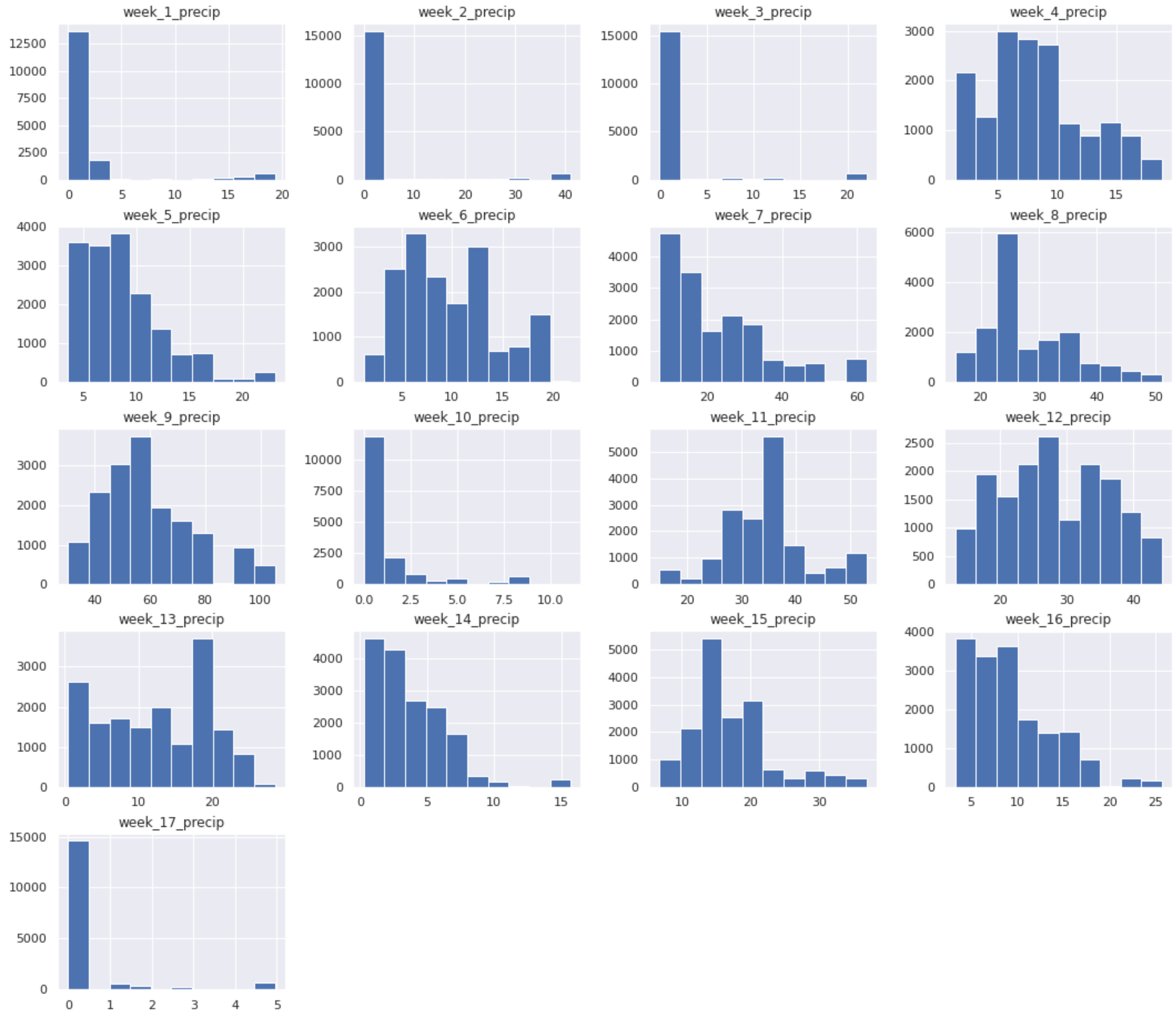


Figure 7: Histogram plot of precipitations for each week

0.7.1 Kernel Density Estimation (KDE)

Definition

Let $(x_1, x_2, \ldots, x_n)$ be independent and identically distributed (iid) samples drawn from some univariate distribution with an unknown density f at any given point x. We are interested in estimating the shape of this function f. Its kernel density estimator is:

$$\widehat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{K}_h\left(x - x_i\right) = \frac{1}{nh}\sum_{i=1}^{n} \mathcal{K}\left(\frac{x - x_i}{h}\right),$$

where $\mathcal{K}$ is the kernel - a non-negative function $-$ and $h > 0$ is a smoothing parameter called the bandwidth. A kernel with subscript h is called the scaled kernel and defined as:

$$\mathcal{K}_h(x) = \frac{1}{h}\mathcal{K}\left(\frac{x}{h}\right)$$

.

Kernel density estimation (KDE) is a non-parametric method for estimating a random variable's probability density function. It utilises a kernel function, $\mathcal{K}$, to estimate an unknown probability density function. Unlike histogram that counts the number of data points in random locations, KDE is the sum of a kernel function on each data point. It is a basic data smoothing problem in which population inferences are drawn from a small sample of data.

The following are typically properties of the kernel function:

1. It is symmetrical $\mathcal{K}(x) = \mathcal{K}(-x)$
2. It can be normalised such that $\int_{-\infty}^{\infty} \mathcal{K}(x)dx = 1$
3. It is monotonically decreases such that $\mathcal{K}'(x) < 0$ when $x > 0$
4. The expected value equals to zero $\mathrm{E}[\mathcal{K}] = 0$

Machine learning applications can make use of the Kernel Density Estimation approach. Because the kernel's scope is defined by parameters in the estimation function, a neural network can begin to train itself to correct its estimations and generate more accurate results. The bandwidth and amplitude estimations are continuously updated while the estimation process repeats itself, increasing the accuracy of the calculated probability density curve.

KDE was used in estimating each of the weekly precipitation data in the dataset in order to get an estimate of the type of probability distribution function for each variable. The result of the density plots (Figure 8) for the precipitation data shows that all the variables exhibit a non-normal distribution. This is in line with the outcome of the histogram plot for the data.

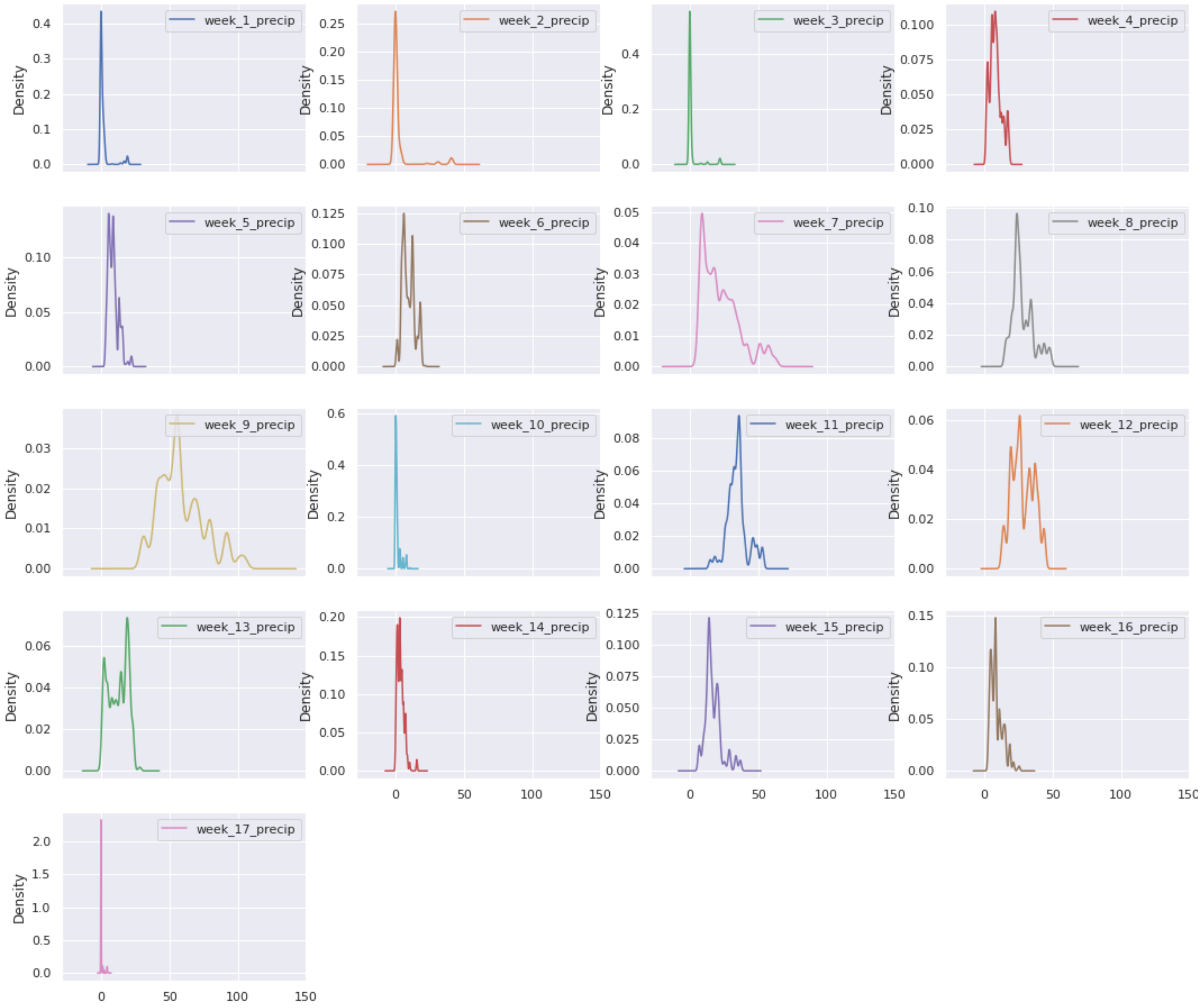


Figure 8: Kernel Density Estimation of the precipitation data

0.8 Fitting Probability Distributions on Precipitation Data

Comparing the histogram of the data with a PDF (probability distribution function) of a known distribution is a basic approach typically used to determine the underlying distribution that could have created a data set (e.g., normal). The distribution's parameters, however, are unknown, and there are many different distributions. As a result, an automatic method of fitting a large number of distributions to the data would be beneficial, which is what is implemented here.

The precipitation data were analysed to identify the best fit probability distribution for each period of study.This is done to find a distribution that suits the data well. The distribution giving a close fit is supposed to lead to good predictions. The best fit probability distribution was determined using the least square method and was identified based on the minimum deviation between actual and estimated values.

0.8.1 Finding Best-Fit PDF with Python *Fitter* Package

The *Fitter* package in Python was implemented on the precipitation data to find the best fit probability distribution for each of the weeks. A basic class in the *Fitter* package identifies the distribution from which a data sample is created. It employs 80 *SciPy* distributions and allows plotting of the results to see which distribution is the most likely and which parameters are the best.The best fit probability distribution was identified based on the minimum deviation between actual and estimated values, and the precipitation data assumed different probability distributions.

Table 4 contains the result of the fitted probability distribution functions for each weekly precipitation. Each table contains top 5 probability distributions out of the 80 fitted distributions. The best fit probability distribution is the in the first row of each table, having the least sum of square errors (SSE). The precipitation data for the period originates from one of Lomax, Wald, Double Gamma (dgamma) Skew Normal (skewnorm), Exponential (expon), Cauchy, Laplace, Half Cauchy, Semi Circular, Anglit, and Half Normal (halfnorm) distributions. The precipitation data of some weeks share the same distribution, such as Weeks 4,5 and 11 (Double Gamma), Weeks 7, 16 and 17 (Exponential), Weeks 1 and 2 (Lomax), and Weeks 8 and 15 (Cauchy). In addition, most of the tables have a special case of either Gamma or Weibull distributions,which are probability distributions that are commonly used to model rainfall data.

Each table as well consists of other columns that are used as measures in selecting the best fit probability distributions. The Akaike Information Criterion (AIC) evaluates how well each probability density function fits the precipitation data. Using the maximum likelihood estimate and the number of parameters (independent variables) in the probability density function, AIC calculates the relative information value of the model. The AIC is calculated using:

$$AIC = 2K - 2\ln(\hat{L})$$

where K is the number of parameters in the density function and $\hat{L}$ is the likelihood estimate. AIC scores with fewer parameters are better, whereas AIC penalizes models with more parameters. When two models explain the same amount of variation, the one with fewer parameters has a lower AIC score and is the better-fit model. To see the effect of AIC in selection of the best fit model in this study, we consider Week 5 table of Table 4. In Week 5, apart from the best-fit PDF (dgamma) with the least SSE (0.096229), others have the same SSE value (0.100198), but the erlang distribution which has the least AIC score (258.670917) is ranked the highest.

The Bayesian Information Criterion (BIC) is similar to the AIC in selection of best fit models. The BIC is obtained by evaluating:

$$\text{BIC} = K\ln(n) - 2\ln(\hat{L})$$

where: $\hat{L}$ is the maximized value of the likelihood function of the model, n is the number of data points in the sample size and k is the number of parameters estimated by the model. A lower BIC value is the preferred.

The Kullback-Leibler Divergence (KL_{div}) score, often known as the KL divergence score, measures how much one probability distribution differs from another.

$$\mathrm{KL}_{\mathrm{div}}(p,q) = \begin{cases} p\log(p/q) - p + q & p > 0, q > 0 \\ q & p = 0, q \geq 0 \\ \infty & \text{otherwise} \end{cases}$$

Thus, for distributions P and Q of a continuous random variable, KL_{div} is defined to be the integral:

$$KL_{div}(p,q) = D_{\mathrm{KL}}(P\|Q) = \int_{-\infty}^{\infty} p(x)\log\left(\frac{p(x)}{q(x)}\right)dx$$

0.8.2 Best-Fit Probability Density Function Plot

The Probability Density Function plot of each precipitation week data in Table 4 is shown in Figure **??**. The top legend in each plot indicates the best-fit probability density function.

0.8.3 Data Preprocessing and Normalization

The precipitation data were normalised using the z-score normalisation method. This was achieved by computing both the mean (μ) of the train data and the standard deviation (σ) of the train data. The standard score of a sample x is calculated as:

$$z = \frac{x - \mu}{\sigma}$$

With this done, the precipitation data are of the same range and normally distributed.

0.8.4 Training the Machine Learning Model

The preprocessed data was fed into six machine learning algorithms to train a model that will be suitable for flood prediction. Three of these algorithms are variants of the boosting algorithm, namely: CatBoost, Extreme Gradient Boosting Regressor (XGBoost) and Light Gradient Boosting Regressor (LGB). The other three are neural network (MultiLayerPerceptron), Support Vector Regressor and RandomForest

The Python package running on a Google Colaboratory cloud notebook was used to implement the algortithms on the data. The default parameters of each algorithm was used in training the model. At the completion of the learning process, the trained model was evaluated to see how well the algorithm was able to learn from the data.

0.9 Results and Discussion

0.9.1 Model Evaluation

Since the target variable of the dataset is a continuous, the performance metrics such as Mean Square Error (MSE), Root Mean Square Error (RMSE), the Mean Absolute Error (MAE) and the Co-efficient of Determination (R^2) were applied to evaluate the ability of the proposed model to predict flood. The statistical methods are defined as follows:

PDF	SSE	AIC	BIC	KL_div
lomax	0.056643	482.799173	−207113.667035	∞
gilbrat	0.083231	530.232699	−200786.518578	∞
beta	0.132962	918.551754	−193053.742790	∞
pearson3	0.195469	659.682700	−186718.510278	∞
burr	0.211742	565.465313	−185392.017518	∞

(a) Week 1

PDF	SSE	AIC	BIC	KL_div
lomax	0.070106	541.392375	−203602.636822	∞
wald	0.109301	741.502014	−196299.861689	∞
pearson3	0.119707	871.710054	−194792.609405	∞
exponweib	0.151789	545.013790	−190873.227922	∞
f	0.206320	375.470162	−185819.143322	∞

(b) Week 2

PDF	SSE	AIC	BIC	KL_div
wald	0.313566	796.083893	−178946.230783	∞
gengamma	0.578596	540.810597	−168839.819064	∞
halfgennorm	0.646049	1190.975805	−167033.793883	∞
exponnorm	0.666966	684.687662	−166509.130912	∞
genexpon	0.671923	683.757751	−166367.789418	∞

(c) Week 3

PDF	SSE	AIC	BIC	KL_div
dgamma	0.088844	236.580881	−199702.235317	∞
dweibull	0.089093	235.619610	−199656.153328	∞
foldcauchy	0.095206	235.903797	−198563.483725	∞
alpha	0.095703	226.889987	−198477.643324	∞
genextreme	0.095722	226.270348	−198474.496416	∞

(d) Week 4

PDF	SSE	AIC	BIC	KL_div
dgamma	0.096229	292.171451	−198387.427103	∞
erlang	0.100198	258.670917	−197721.948221	∞
pearson3	0.100198	258.672056	−197721.942564	∞
chi2	0.100198	258.671295	−197721.941973	∞
gamma	0.100198	258.671951	−197721.940992	∞

(e) Week 5

PDF	SSE	AIC	BIC	KL_div
skewnorm	0.088494	238.810197	−199767.227754	∞
moyal	0.088790	237.147174	−199721.885994	∞
kstwobign	0.089376	235.383197	−199613.665203	∞
gumbel_r	0.090281	237.586676	−199447.771977	∞
genlogistic	0.090978	240.133587	−199311.337092	∞

(f) Week 6

PDF	SSE	AIC	BIC	KL_div
expon	0.011823	320.118689	−232920.534648	∞
levy	0.012447	349.114618	−232074.773195	∞
halfcauchy	0.012655	331.001849	−231801.650943	∞
halflogistic	0.013159	317.810612	−231158.188933	∞
halfnorm	0.013555	314.731636	−230670.368647	∞

(g) Week 7

PDF	SSE	AIC	BIC	KL_div
cauchy	0.023629	313.926700	−221519.910570	∞
exponnorm	0.024444	287.036944	−220951.392876	∞
laplace	0.025271	300.390094	−220413.273279	∞
gumbel_r	0.025639	284.380700	−220175.628116	∞
moyal	0.025671	281.683165	−220154.730339	∞

(h) Week 8

PDF	SSE	AIC	BIC	KL_div
laplace	0.004684	344.419121	−248166.848221	∞
exponnorm	0.004924	334.834373	−247333.104187	∞
gumbel_r	0.004961	330.447946	−247221.344228	∞
erlang	0.005009	331.155969	−247051.073622	∞
gamma	0.005009	331.155894	−247051.072070	∞

(i) Week 9

PDF	SSE	AIC	BIC	KL_div
halfcauchy	0.101524	288.480399	−197515.154941	∞
gilbrat	0.107093	314.454867	−196635.894243	∞
wald	0.151924	318.536136	−190877.970873	∞
expon	0.267692	330.296612	−181550.716305	∞
cauchy	0.307349	334.111486	−179275.939013	∞

(j) Week 10

PDF	SSE	AIC	BIC	KL_div
dgamma	0.021283	295.603862	−223232.035078	∞
dweibull	0.021351	295.363435	−223178.913023	∞
gennorm	0.021503	295.445463	−223062.405693	∞
laplace	0.021566	293.934913	−223023.649583	∞
hypsecant	0.021576	293.733847	−223015.965548	∞

(k) Week 11

PDF	SSE	AIC	BIC	KL_div
semicircular	0.037257	247.636372	−214021.543681	∞
anglit	0.037645	251.513142	−213851.110196	∞
rayleigh	0.037695	255.733547	−213829.109212	∞
maxwell	0.037989	256.739313	−213701.360382	∞
gumbel_r	0.038059	260.094340	−213670.903432	∞

(l) Week 12

PDF	SSE	AIC	BIC	KL_div
anglit	0.040310	248.252330	−212724.628173	∞
gumbel_l	0.041968	264.956391	−212061.018749	∞
uniform	0.042450	237.506003	−211872.970057	∞
cosine	0.042641	253.711863	−211799.261309	∞
norm	0.044343	254.160055	−211154.663882	∞

(m) Week 13

PDF	SSE	AIC	BIC	KL_div
halfnorm	0.109094	265.923054	−196331.079234	∞
moyal	0.116474	263.189679	−195253.162883	∞
gumbel_r	0.117977	274.493330	−195042.100244	∞
rayleigh	0.120038	295.590630	−194756.832298	∞
maxwell	0.124257	302.296821	−194188.095566	∞

(n) Week 14

PDF	SSE	AIC	BIC	KL_div
cauchy	0.049247	303.591173	−209427.481522	∞
gumbel_r	0.049445	278.722427	−209361.417763	∞
hypsecant	0.049862	296.345384	−209223.054617	∞
laplace	0.049950	297.692542	−209194.157390	∞
logistic	0.050465	296.609682	−209025.259818	∞

(o) Week 15

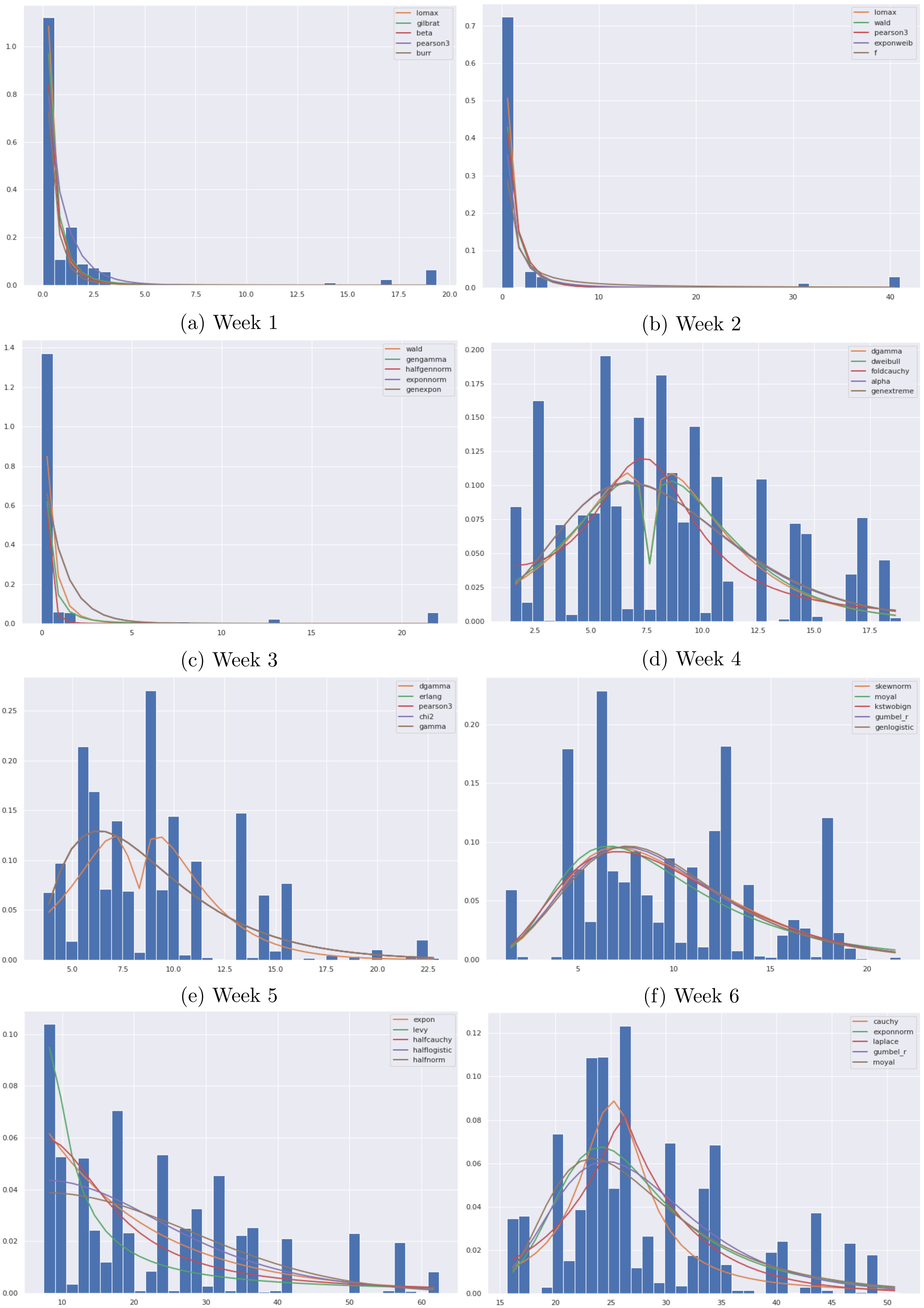
PDF	SSE	AIC	BIC	KL_div
expon	0.056182	262.084952	−207257.998509	∞
moyal	0.057234	271.622051	−206952.468397	∞
gumbel_r	0.059214	276.625015	−206392.558305	∞
rayleigh	0.062149	281.302021	−205596.146753	∞
laplace	0.063320	288.545496	−205288.690830	∞

(p) Week 16

PDF	SSE	AIC	BIC	KL_div
expon	16.282885	450.863241	−113907.821199	∞
gumbel_r	25.326163	455.649220	−106634.403485	∞
hypsecant	26.229316	478.194337	−106057.439121	∞
logistic	28.167568	456.758811	−104883.519408	∞
rayleigh	30.941705	309.590564	−103336.807339	∞

(q) Week 17

Table 4: Fitted probability distributions table for precipitation data



(a) Week 1

(b) Week 2

(c) Week 3

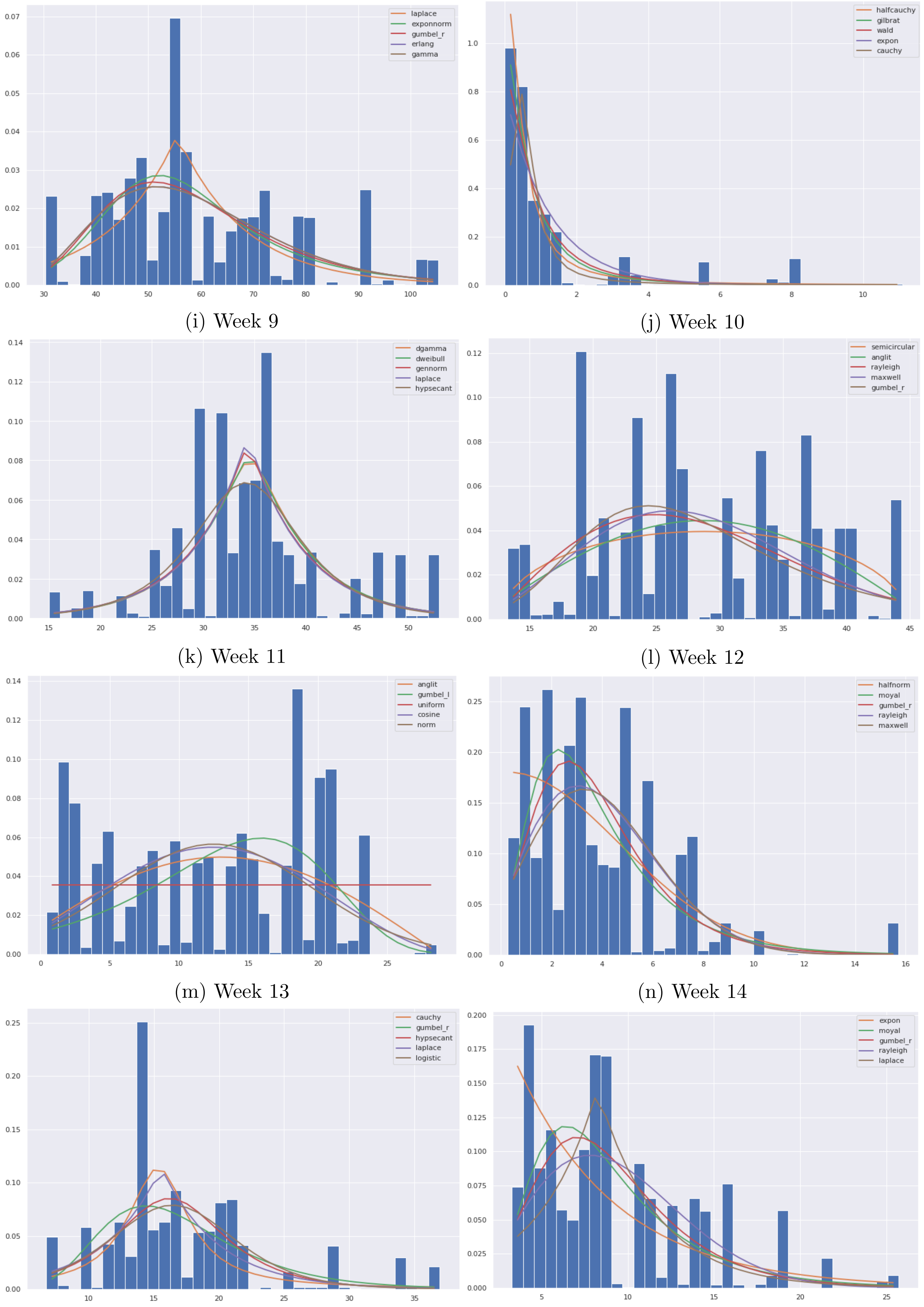
(d) Week 4

(e) Week 5

(f) Week 6

(g) Week 7

(h) Week 8



(i) Week 9

(j) Week 10

(k) Week 11

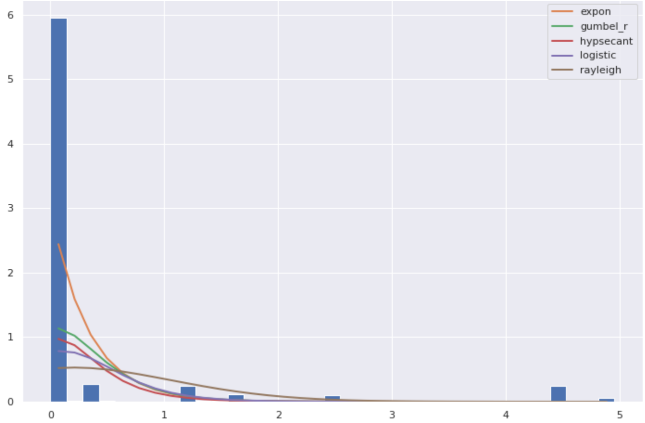
(l) Week 12

(m) Week 13

(n) Week 14

(o) Week 15

(p) Week 16



(q) Week 17

Figure 11: Plot of best-fitted probability density function

- **Mean Square Error (MSE)**

The mean squared error function calculates the expected value of the squared (quadratic) error or loss, which is a risk indicator. It is defined as:

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

Lower MSE indicates a better fit.

- **Root Mean Square Error (RMSE)**

This is the square root of the mean square error, and it is defined as:

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

- **Mean Absolute Error (MAE)**

This function calculates mean absolute error, which is a risk metric that represents the expected magnitude of an absolute error loss. It is defined as

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

- **Coefficient of Determination (R^2)**

It represents the fraction of the variation (of y) explained by the model's independent variables. Through the proportion of explained variance, it provides an indication of model goodness of fit and thus a measure of how well unseen samples are likely to be predicted by the model.

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$(\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\epsilon_i^2)$.

where n is the number of data points,y_i is the actual value , while $\hat{y}_i$ is the predicted value of the i-th data point

Table 5 is the performance of each model after it was being evaluated by the performance metrics From Table 5, the CatBoost model performs best with the lowest MSE and RMSE values, followed by RandomForest (having the least MAE) while SVR model perform worst. The implication of the CatBoost R^2 value is that the percentage of the flood (target variable) that is accounted for by the predictors (precipitation data, elevation and coordinates) is 56.5%, while the the remaining 43.5% are accounted for by other factors. The other factors may include: temperature data, wetlands, and other factors. RMSE implies that the prediction that a particular location will be flooded may seem to be 15.1% off the particular area.

Model	**MAE**	**MSE**	**RMSE**	R^2
SVR	0.140	0.048	0.220	0.075
XGBoost	0.073	0.027	0.166	0.472
Catboost	0.060	**0.022**	**0.151**	**0.565**
LGBoost	0.061	0.026	0.163	0.494
RandomForest	**0.058**	0.025	0.160	0.51
MultiLayerPerceptron	0.094	0.032	0.179	0.38

Table 5: Model performance metrics table

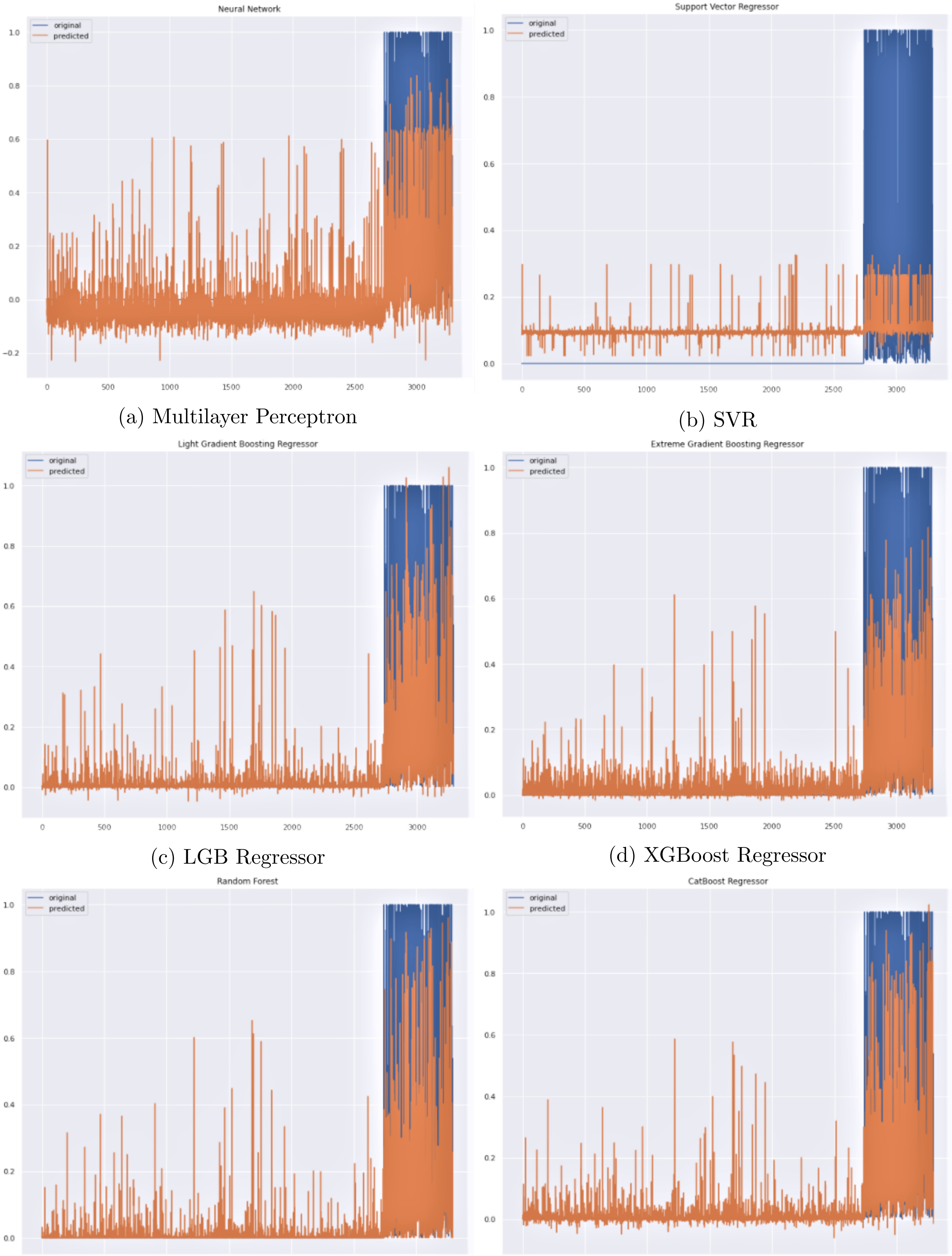
0.9.2 Prediction Plots

The prediction plot enables us to visualise and compare the actual values and the predicted values, in order to see how well the model perform. The prediction plot for each of the models is shown in Figure 4.1. The blue colour lines indicate the actual values of the target (y) while the orange colour lines ($\hat{y}$) indicate the predicted values From this plot, the SVR (first plot) did not learn well, and thus didn't capture the data, thereby underfitting the data.

0.9.3 Feature Importance Plot

In order to get a score of the predictors used in training the model based on their relevance, the feature importance technique is required. The ranking of the predictors based on the higher score from our best model (CatBoost) is obtained using a wrapper class that is contained in the algorithm. A higher score indicates that the specific feature will have a greater impact on the model used in predicting the flood extent.

Figure 13 is a visual representation of the important features from the CatBoost model. It shows the major contributing predictors in the model in descending order. The most influencing predictor in the model is "elevation", with a score greater than 50. Other important predictors are the location (X and Y coordinates), Week 17 (the week with the lowest average rainfall and the preceding week to the flooding period) and Week 9 (the week with the highest average rainfall) precipitation data, and Land Cover Type. Other predictors have a very low score



(a) Multilayer Perceptron

(b) SVR

(c) LGB Regressor

(d) XGBoost Regressor

(e) Random Forest

(f) Catboost Regressor

Figure 12: Prediction plots comparing actual values against predicted values

with Week 10 precipitation data being the least, indicating that the contributions of these low rank features (predictors) are not influencing the model significantly. Hence, dropping these variables may not necessarily affect the performance of the model in predicting flood extents in the targeted location. In fact, it may increase the efficiency of the model by reducing the error rate (bias) in the prediction.

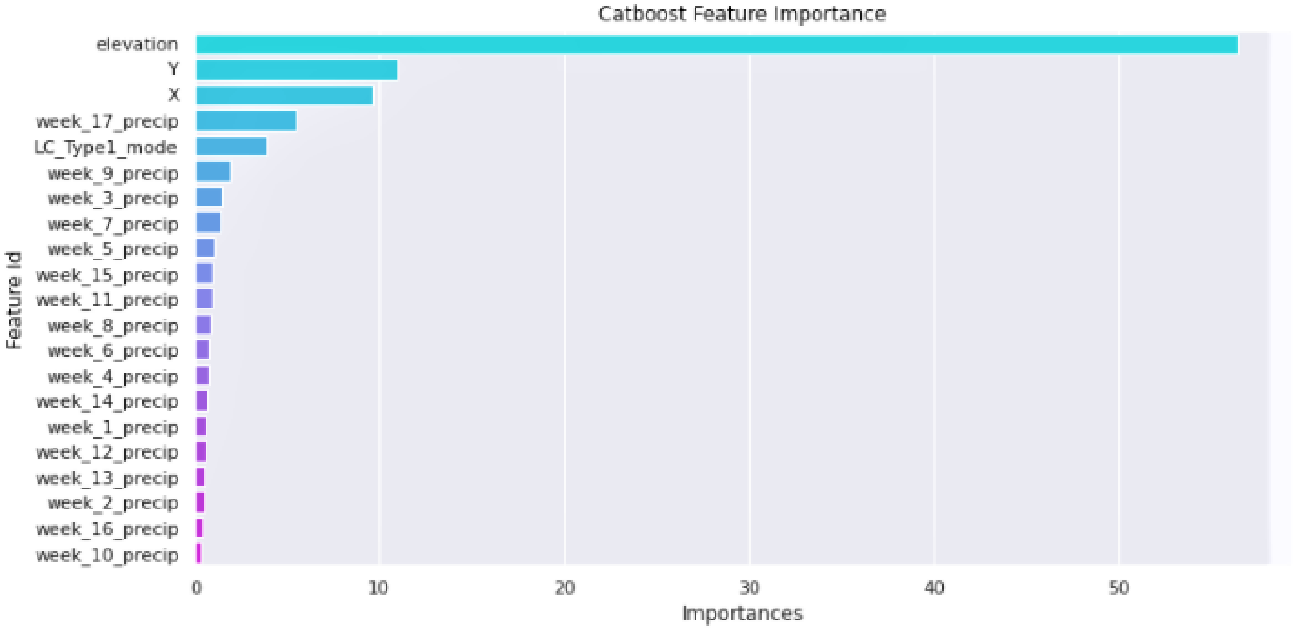


Figure 13: Rank of the important features

0.10 Conclusion

We considered mitigating flood occurrences by building a predictive machine learning model that learns historical flood patterns and is capable of making predictions of the location and the extent of flood.

In achieving the above, the probability distribution patterns of rainfall was first examined with a histogram plot and a kernel density estimation plot, and we found out that most of the precipitation features have different probability distributions, except for few which share the same distribution pattern. To determine the best fitted distribution of the data, the ”Fitter” package of Python programming language was used to fit and find the probability distributions, then the data was normalised to avoid varying range which may cause learning problems for the algorithms.

Six machine learning algorithms were used to train the data and the performance of the model was evaluated with four different metrics in which the best model was the Catboost with the ”elevation” feature as the most important feature in the model.

Hence, data-driven technique for flood prediction is more efficient if there is availability of sufficient data with relevant features. Thus, knowing this ahead, a cost effective and an optimal evacuation decision policy proposed by Taiwo et al (2019) can be adopted in mitigating huge loss of lives and properties during floods.

Further improvement can be made to the model for better prediction performance by including other relevant features such as climatic data (e.g Temperature) and topographical data (e.g distance to wetlands) could be in the course of training the model. In addition, predictors with low relative importance (score) may be dropped, this most times influences better prediction of machine learning models.

References

Aljazeera (2017, January 26). Flood lash Bolivia and Peru. Climate Crisis. https://www.aljazeera.com/news/2017/1/26/floods-lash-bolivia-and-peru

Amangabara, G.T, and Obenade M, (2015). Flood Vulnerability Assessment of Niger Delta States Relative to 2012 Flood Disaster in Nigeria. American Journal of Environmental Protection, vol. 3, no. 3: 76-83. DOI: 10.12691/env-3-3-3.

Bhavsar P., Safro I, Bouaynaya N, Polikar R and Dera D. (2017). Machine Learning in Transportation Data Analytics. Data Analytics for Intelligent Transport Systems. DOI: http://dx.doi.org/10.101
0-12-809715-1.00012-2

Bhuiyan, S.R. Baky, A.A (2014). Digital elevation based flood hazard and vulnerability study at various return periods in Sirajganj Sadar Upazila, Bangladesh. Int. J. Disas. Risk Reduct., 10, 48–58.

Buyuksahin U.C and Ertekin S. (2019). Improving forecasting accuracy of time series data using a new ARIMA-ANN hybrid method and empirical mode decomposition. Neurocomputing. https://doi.org/10.1016/j.neucom.2019.05.099; pp. 151-163

Carvalho J.P and Camelo F.V (2015). One Day Ahead of Stream Flow Forecasting. 16th World Congress of the International Fuzzy Systems Association (IFSA) Conference of the European Society for Fuzzy Logic and Technology (EUSFL). Atlantis Press; pp 1168-1175

CatBoost Regressor. https://catboost.ai/en/docs/concepts/python-reference_catboostregressor

Coca N. (2020). Flooded Asia: Climate change hits region the hardest. Nikkei Asia. https://asia.nikkei.com
Insight/Flooded-Asia-Climate-change-hits-
region-the-hardest

Cui C. and Cui L. (2020). An Innovative Flood Prediction System Using Improved Machine Learning Approach. The Canadian Science Fair Journal (CSFJ);Volume 2; Issue 2; DOI:10.18192/csfj.v2i32020115119

Davies R. (2021, October 20). India – Floods and Landslides Cause 27 Fatalities in Uttarakhand. FloodList. https://floodlist.com/asia/india-floods-landslides-uttarakhand-october-2021

DeepAI. Kernel Density Estimation. https://deepai.org/machine-learning-glossary-and-terms/kernel-density-estimation#: :text=The%20Kernel%20Density%20
Estimation%20is%20a%20mathematic%20process,a%20population%2C

%20based%20on%20a%20finite%20data%20set.

Deutsch Welle [DW] (2021, June 6). Brazil: Severe floods as Amazon rivers reach record highs. https://www.dw.com/en/brazil-severe-floods-as-amazon-rivers-reach-record-highs/a-57753700

Echendu, A.J (2020). “The impact of flooding on Nigeria’s sustainable development goals (SDGs):., Ecosystem Health and Sustainability, 6:1, 1791735, DOI: 10.1080/20964129.2020.1791735

Fitter Package Documentation. https://fitter.readthedocs.io/en/latest/

FloodList (2019, March 29). South America – Thousands Affected by Floods and Landslides in Paraguay, Peru, Ecuador and Bolivia. https://floodlist.com/america/floods-paraguay-peru-ecuador-bolivia-march-2019

Ganguly K.K, Nahar N. and Hossain, B.M.M (2018). A Machine Learning-Based Prediction and Analysis of Flood Affected Households: A Case Study of Floods in Bangladesh. International Journal of Disaster Risk Reduction, (), S221242091830311X–. doi:10.1016/j.ijdrr.2018.12.002

Ghosh, A., Kar, S.K. (2018) Application of analytical hierarchy process (AHP) for flood risk assessment: a case study in Malda district of West Bengal, India. Nat Hazards 94, 349–368. https://doi.org/10.1007/s11069-018-3392-y

Hastie T, Tibshirani R, and Friedman J. H. (2009). Boosting and Additive Trees. The Elements of Statistical Learning (2nd ed.). New York: Springer. pp. 337–384. ISBN 978-0-387-84857-0

Hoque M.A, Tasfia S., Ahmed N. and Pradhan B. (2019). Assessing Spatial Flood Vulnerability at Kalapara Upazila in Bangladesh Using an Analytical Hierarchy Process. Sensors 2019, 19, 1302, doi:10.3390/s19061302, https://www.mdpi.com/1424-8220/19/6/1302/pdf

India Today (2015, December 10). India is the most flood-prone country in the world. India Today Web Desk. https://www.indiatoday.in/education-today/gk-current-affairs/story/india-is-the-most-flood-prone-country-in-the-world-276553-2015-12-10

Kinghorn J. (2017, April 13). 3 Factors That Make Flooding in South America Worse. AIR. https://www.air-worldwide.com/blog/posts/2017/4/3-factors-that-make-flooding-in-south-america-worse/

Lagos Bureau of Statistics (2016). “Household Survey 2016”. Lagos State Government; retrieved online from http://mepb.lagosstate.gov.ng/wp-content/uploads/sites /29/2020/08/House-Hold-REPORT-Y2016.pdf

Levenson M. (2021, June 3). Severe Flooding in Guyana Prompts Extensive Relief Effort. New York Times. https://www.nytimes.com/2021/06/03/us/guyana-flooding-relief.html

Light Gradient Boosting Regresor (LGBoostRegressor). https://lightgbm.readthedocs.io/en/latest/Python-Intro.html

Lucas B. (2021). “Urban flood risks, impacts, and management in Nigeria”. K4D Helpdesk Report 948. Brighton, UK: Institute of Development Studies. DOI: 10.19088/K4D.2021.018

Luo T., Maddocks A., Iceland C., Ward P., Winsemius H. (2015). World's 15 Countries with the Most people exposed to river flood. World Resources Institute (WRI). Retrieved October 2021 from https://www.wri.org/insights/worlds-15-countries-most-people-exposed-river-floods

Martini A. (2020, February 26). Weatherwatch: floods across South America after heavy rain. The Guardian. https://www.theguardian.com/news/2020/feb/26/weatherwatch-floods-across-south-america-after-heavy-rain

Matias Y. (2018). Keeping people safe with AI-enabled flood forecasting. Google Blog, assessed online from https://www.blog.google/products/search/helping-keep-people-safe-ai-enabled-flood-forecasting/ Retrieved November 17th, 2020

Mind'je R, Li L, Amanambu A.C, Nahayo L, Nsengiyumva J.B, Gasirabo A, Mindje M. (2019). Flood susceptibility modelling and hazard perception in Rwanda. International Journal of Disaster Risk Reduction. https://doi.org/10.1016/j.ijdrr.2019.101211

Mosavi A, Ozturk P and Chau K. (2018). Flood Prediction Using Machine Learning Models: Literature Review. MDPI Journal. Water 2018, 10, 1536; doi:10.3390/w10111536

National Aeronautics and Space Administration [NASA] (2008, March 4). Heavy Rain Floods South America. NASA Earth Observatory.
https://earthobservatory.nasa.gov/images/19668/heavy-rain-floods-south-america

National Emergency Management Agency–NEMA (2020): "Flood Maps".
https://nema.gov.ng/docs-category/flood-maps/, retrieved 16th October 2020.

Neural Networks. https://scikit-learn.org/stable/modules/neural_networks_supervised.html

Neural Network MultiLayer Perceptron Regressor. https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html#sklearn.neural_network.MLPRegressor

Nwigwe, C. and Emberga T.T (2014). "An Assessment of cause and effects of flood in Nigeria". Standard Scientific Research and Essays Vol2 (7): 307-315, July 2014 (ISSN: 2310-7502)

Obarein O.A and Amanambu A.C (2019). Rainfall timing: variation, characteristics, coherence,and interrelationships in Nigeria, Theor. Appl. Climatol. pp.1–15.)

OCHA, (2016). "West Africa: Impacts of the floods", retrieved 13 April 2021, from https://www.humanitarianresponse.info/sites/www.humanitarianresponse.info/files/ documents/files/wca_a4_l_imp

Okwu-Delunzu V.U, Ogbonna C.E and Lamidi S. (2017). Spatial Assessment of Flood Vulnerability in Anambra East Local Government Area, Nigeria Using GIS and Remote Sensing. British Journal of Applied Science & Technology. 19(5): 1-11, 2017; Article no. BJAST.29378; DOI: 10.9734/BJAST/2017/29378

Piryonesi, S. Madeh; El-Diraby, and Tamer E. (2020). "Data Analytics in Asset Management: Cost-Effective Prediction of the Pavement Condition Index". Journal of Infrastructure Systems. 26 (1): 04019036. doi:10.1061/(ASCE)IS.1943-555X.0000512. ISSN 1943-555X. S2CID 213782055

Piryonesi, S. Madeh; El-Diraby, Tamer E. (2021). ”Using Machine Learning to Examine Impact of Type of Performance Indicator on Flexible Pavement Deterioration Modeling”. Journal of Infrastructure Systems. 27 (2): 04021005.

PreventionWeb (2018) “Disaster Data and Statistics”. Retrieved 13th April, 2021 from https://www.preven
statistics

Rahman M, Ningsheng C, Islam M.M, Dewan A, Iqbal J. Washakh R.M.A and Shufeng T. (2019). Flood Susceptibility Assessment in Bangladesh Using Machine Learning and Multi-criteria Decision Analysis. Earth Systems and Environment, (), –. doi:10.1007/s41748-019-00123-y

RandomForest Regressor. https://scikit-learn.org/stable/modules/generated/
sklearn.ensemble.RandomForestRegressor.html#: :text=A%20random%20forest
%20regressor.,accuracy%20and%20control%20over%2Dfitting.& text=
The%20number%20of%20trees%20in%20the%20forest.

Raschka S. and Mirjalili V. (2017). Python Machine Learning. 2nd Edition. Packt Publishing; Birmingham, UK.

Rehman S., Sahana M, Hong H, Sajjad H. and Ahmed B. (2019). A systematic review on approaches and methods used for flood vulnerability assessment: framework for future research. Natural Hazards. https://doi.org/10.1007/s11069-018-03567-z

Reliefweb (2015). South America: Floods and Landslides. Glide: FL-2015-000171-PRY. https://reliefweb.int/disaster/fl-2015-000171-pry

Reliefweb (2021). Peru: Floods – Feb 2021. Glide: FL-2021-000019-PER.
https://reliefweb.int/disaster/fl-2021-000019

ResearchClue. (2020). Flooding in Nigeria Causes, Effects and Solution. Available at: https://nairaproject.
[Accessed: 2021-4-13].

Ritorto D. (2013). South American floods: Dozens dead in Brazil as Mexico also hit. British Broadcasting Corporation (BBC). https://www.bbc.com/news/av/
world-latin-america-25514396

Roy D.C. and Blaschke T. (2015). Spatial vulnerability assessment of floods in the coastal regions of Bangladesh. Geomatics, Natural Hazards and Risk, 6:1, 21-44, DOI: 10.1080/19475705.2013.816785

Sowmya K, John C.M. and Shrivasthava, N.K. (2015). Urban flood vulnerability zoning of Cochin City, southwest coast of India, using remote sensing and GIS. Natural Hazards, 75(2), 1271–1286. doi:10.1007/s11069-014-1372-4

StandardScaler. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.
StandardScaler.html

StatsModels. Kernel Density Estimator (KDE). Statsmodel v0.13.1 documentation. https://www.statsmod

Sudha Rani, N.N.V., Satyanarayana A.N.V and Bhaskaran P.K (2015). Coastal vulnerability assessment studies over India: a review. Natural Hazards, 77(1), 405–428. doi:10.1007/s11069-015-1597-x

Support Vector Regressor. https://scikit-learn.org/stable/modules/generated/ sklearn.svm.SVR.html

Taiwo E.S, Adinya I., Edeki S.O (2019). Optimal evacuation decision policies for Benue flood disaster in Nigeria. Journal of Physics: Conference Series 1299 012137, doi: 10.1088/1742-6596/1299/1/012137

Talukdar S. and Pal S. (2020). Modeling flood plain wetland transformation in consequences of flow alteration in Punarbhaba river in India and Bangladesh. Journal of Cleaner Production. https://doi.org/10.1016/j.jclepro.2020.120767

Talukdar S; Ghose B; Shahfahad, Salam R, Mahato S, Pham, Q B, Linh N.T.T; Costache R, and Avand M. (2020). Flood susceptibility modeling in Teesta River basin, Bangladesh using novel ensembles of bagging algorithms. Stochastic Environmental Research and Risk Assessment, (), –. doi:10.1007/s00477-020-01862-5

The National Severe Storms Laboratory -NSSL(n.d). Severe Weather 101: Flood Basics. Online Learning Resource. Retrieved April 7th, 2021
from https://www.nssl.noaa.gov/education/svrwx101/floods/

Vankatesan E. and Mahdrakar A.B. (2019). Forecasting Floods using Extreme Gradient Boosting – A New Approach. International Journal of Civil Engineering and Technology (IJCIET). 10(2), 2019, pp. 1336-1346 http://www.iaeme.com/ijciet/issues .asp?JType=IJCIET&VType=10&IType=02

World Health Organisation–WHO (2012). “Public health risk assessment and interventions. Flooding disaster: Nigeria”. WHO/PEC/ERM/PHRA/2012.3; retrieved 13th April 2021, from http://www.who.int/hac/crises/nga/RA_ Nigeria_1Nov2012a.pdf.

World Health Organisation–WHO (2021). Floods. retrieved 7th April 2021 from https://www.who.int/heal topics/floods#tab=tab_1

World Meteorological Organization (WMO). Manual on Flood Forecasting and Warning. WMO-No. 1072; 2011 edition

XGBoost Documentation. https://xgboost.readthedocs.io/en/stable/

Zehra N. (2020). Prediction Analysis of Floods Using Machine Learning Algorithms (NARX & SVM). International Journal of Sciences: Basic and Applied Research (IJSBAR). Volume 49, No. 2, pp 24-34; http://gssrr.org/index.php?journal= JournalOfBasicAndApplied