Flood extent prediction using Machine models with Google Earth Engine.

Eshan Jairath — eshanjairath@outlook.com — Northumbria University
Second Author — secondauthor@email.com — organisation
third Author — thirdauthor@email.com — organisation

Abstract

Natural disasters like floods are one of the most dangerous and are the ones that leave a heavy aftereffect on people and properties. People living near the flood-prone area live in constant terror of evacuating their homes and property whenever a flood strikes. Therefore, a requirement of a prediction model that can assist in estimating the magnitude of flood along with the location of impact becomes necessary. This study focuses on an automated and approachable technique to predict the extent of the flood by leveraging Google Earth Engine and applying machine learning to learn the flood extent patterns. A total of 10 machine learning algorithms have been evaluated in this study. Many factors and techniques that can help predict flood extent after their arrival to prevent huge losses of life and property by evacuating the majorly flooded areas are discussed in this study. All the models used in this study are evaluated using RMSE, NSE, Standard deviation of the features, and flood extent prediction graphs to correctly identify the best performing model. The result of this study will be a set of unique features that can be useful in predicting the flood extent along with a model cable of making strong and reliable predictions.

1 Introduction

Floods are the most common and lethal natural catastrophes in the world, inflicting havoc on both people and property. Floods are usually produced by natural reasons, but they may also be caused by man-made causes, such as infrastructure failures, the development of infrastructure in flood-prone areas, and deforestation, to name a few (Greentumble, 2016). The Organisation for Economic Cooperation and Development estimated that disasters like floods cause damage worth more than $40 billion annually (OECD, 2016). The need for a continual supply of water to maintain life, rich fertile grounds, access to water transport, and the luxury of living near rivers all led to civilization in flood-prone places. In many countries, civilization is established near flood plains. For example, in Japan, 49 percent of the population is established in coastal floodplains (Sato, 2006), 60 percent of the land in the Netherlands is below or at sea level, with two-thirds of the entire population established in flood prone areas, and 70 percent of the Dutch gross domestic product is made on flood-prone lands (Jones-Bos, 2011). The benefits of living near a river or a sustainable waterbody have led to an exponential rise in civilization near flood-prone locations, increasing the amount of damage to life and property with each disaster. This frequently results in a decrease in rainfall absorption rates since the water output surpasses what would be expected on natural terrain (OECD, 2016). Aside from civilization, unexpected climatic changes such as increased precipitation levels, change in NDVI - normalised difference vegetation index (GISGeography, 2022b), land elevation, soil organic carbon, and soil moisture levels all contribute to the likelihood of flooding in a region. When it comes to flood prediction using machine learning there are two key categories to consider. The first is flood prediction using classification approaches, which merely indicates if there will be a flood or not based on prior training data. The disadvantage of this domain is that it does not give enough information to make an informed choice because it does not indicate the magnitude of the flood. If an area is just 1% flooded, it will still be considered flooded, leaving that region unfilled owing to the risk of flooding, and a threshold number of the extent is always necessary for this classification. Whereas the other category is the flood extent prediction which forecast the degree of flooding in an area. This approach uses regression techniques to predict the flood percentage P of the region R (*R is the 1 km sq. polygon of the area of interest)* being flooded where ($0 \leq P \leq 1$) based on K features on time T. This study focusses on the second category in which a list of features K will be used to predict fraction of flood P of region R in an area of interest on time T with the help of machine learning techniques. Following that, an attempt would be made to extract required data from Google Earth Engine (Appendix 1) in order to create the final dataset for training the models. The result will be a flood prediction model capable of forecasting the percentage of the area that will be covered by flood water as well as the precise coordinates that will be flooded with that percentage. This approach will benefit disaster management teams as well as those living in flood-prone regions by allowing them to vacate the area which is most flooded, saving both lives and property.

1.1 Research Approach

This study will employ the experimental research technique, which entails a series of experiments conducted with analysis, planning, and execution. The model in this study will focus on the flood events of the year 2015 and 2019 in the southern region of Malawi. The dataset supplied by the ZINDI (ZINDI, 2020) for the Malawi region will be utilised as a base dataset in this study, with external characteristics obtained from Google Earth Engine and MASDAP (See Appendix) on top to improve flood extent forecasts. As the project entails a significant amount of study on prior studies, resolving flaws with the models used, gathering and processing data from Google Earth engine, and using it to develop a strong and dependable model. The strategy used to construct a flood extent prediction model that tries to identify new available ideas and approaches in flood extent prediction models using GEE data.

2 Background

2.1 Literature Review

Flood forecasting is an area of study that has received a lot of attention. However, machine learning has just recently been applied to this sector (Mosavi et al., 2018). Many types of research have been conducted using various technologies such as IoT sensors, machine learning techniques, GIS (geographical information system), GPS (global positioning system), and satellite imagery to prevent the loss of life and property due to floods, but most of the results are unreliable, because data isn't always sufficient, and system accuracy isn't always robust. This Section will describe the approaches that were previously used in the area of flood prediction and will also give a glimpse of the results achieved.

The flood monitoring system that was first introduced by Adams and Pagano (2016) was capable of notifying when water levels increase over time, and this technology also provided exact data on dangerous weather conditions throughout the world. Han et al. (2017) used Bayesian forecasting System (BFS) which provides an ideal framework for quantifying uncertainty theoretically, that may be used to construct probabilistic flood forecasting systems using any deterministic hydrologic model. There were certain limitations in this technique of flood forecasting since there were only a few types of river basins studied in this work to train and construct a Bayesian forecasting system. It was still unclear if these ideas were applicable to various scenarios and climate conditions.

Ghorbani et al. (2018) The use of time series data mining was also proven to be a good approach in the direction of flood prediction as used by Damle and Yalcin (2007) in river flood prediction. In this technique, the model was able to effectively and precisely anticipate all floods in the testing stage, even though the relationship between earliness of prediction and correct prediction varied. The application of time series demands considerable amount of training time along with accurate and precise threshold selections to attain the highest possible accuracy. Most of the studies that are mentioned above work on the classification basis that tells only the occurrence of floods in yes or no format. Opella et al. (2019) proposed a solution that uses GIS to produce the probability map of the area that could be susceptible to flood. This method uses CNN (Convolution Neural Networks) along with SVM classification techniques to produce a map image output of the area which could be emersed into the water in the case of floods and this study also justifies that why producing a flood map is necessary factor in handling Disaster management planning. Initially, most flood prediction studies relied on hydrological models to analyse and forecast changes in several characteristics that cause floods, but as technology and machine learning advanced, extremely good values were achieved in order to construct a system for predicting floods. In a flood forecasting model, the main source of dataset are the rainfall and water levels as floods occur across the world, a wide range of datasets are accessible. Depending on the weather conditions and water levels of a given place, each of the research described employs a different dataset for their model. As a result, it is impossible to say which solution works best in every situation because they were all trained and created using numerous factors.

3 Data Extraction

3.1 Data Collection

Because the size of the flood is determined by the number of features, the major reason why a prediction system fails is that sometimes the information available about the beginning state is insufficient, and there is a lack of data coverage at the global levels (Wetterhall, 2017). As a result, data collecting is the most important aspect of this study, as the basic Malawi dataset only gives limited information. External features to improve flood extent predictions are taken from Google Earth Engine (See Appendix 1) and MASDAP (See Appendix 2).

3.2 Google Earth Engine (GEE)

GEE is a geospatial image data viewer that provides access to a set of global and regional datasets available in the Earth Engine Data Catalogue(Kumar and Mutanga, 2018). Google earth engine can be accessed with the help of an internet-based IDE (interactive development environment) or web-based python API (Application programming Interface). A relatively sizable collection of publicly accessible geographic datasets may be found in the earth engine data catalogue. Images in both optical and non-optical wavelengths and across a range of spectral bands are included in these files from a number of satellite and aerial imaging systems (Gorelick et al., 2017).

Types of data that will be used in this study from GEE.

1. Vector Data – This uses combinations of the pairs of coordinates as well as longitude and latitude to depict things on the surface of the world.
2. Raster Data – In this the earth's surface is represented as a matrix of values in the form of pixels, cells, or grids. Raster is a picture that contains a matrix of values that correspond to the values of an attribute that was observed. raster bands that represent several factors (Datacarpentry.org, 2022).

The only way a geographic raster differs from a digital image is if spatial information is included to link the data to a specific location. The extent, cell size, number of rows and columns, and spatial reference system of the raster are all included in this. Every piece of information of this kind has a CRS (coordinate reference system) value that identifies the sort of 2D projection used on the map. The term ”map projection” simply refers to the process of attempting to represent the Earth's surface, or a portion of it, on a flat piece of paper or computer screen. The choice of map projection and CRS relies on the geographical coverage of the study area as well as the data's availability (QGIS-Documentation, 2022). For this study all the available data, resources and the data extracted form GEE are projected on “EPSG:4326” CRS.

3.3 Important Features for consideration

Since unexpected changes in the environment are the main cause of floods, it is important to consider the unique characteristics of each flood event that occurs in a particular location. According to the features of the research area, different factors were chosen for different study areas. While one factor may have a major influence on floods in one place, it may not have any impact in another (Kia et al., 2012). The following list factors that will be used in this study and these are considered on the basis of previous research conducted by (Shafapour Tehrany et al., 2017) and (Muckley and Garforth, 2021).

1. **Elevation** - The height of the provided place above a given level. This feature has a pixel size of 30 metres and a measurement range of -10 to 6500 metres.
2. **Slope** – It is the angle between a high elevated and low elevated land. This feature is calculated in degrees from a terrain DEM (Digital Elevation model).
3. **Aspect** – It is the compass direction that a slope faces (GISGeography, 2022a). Units are degrees where 0=N, 90=E, 180=S, 270=W.
4. **Soil Surface Moisture** - The relative water content of the top few centimetres of soil, measured in % saturation, is called surface soil moisture (SSM), and it indicates how moist or dry the soil is in that layer.
5. **Soil Organic Carbon** – Higher crop yields and better soil moisture retention are typically linked to increased soil organic matter. Having a high count of this feature in the soil means the less risk of soil erosion, nutrient leaching and more stable carbon cycle which will result in greater agriculture production.
6. **Clay Content** – Percentage of Clay content present in soil (kg / kg) at 6 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250 m resolution (Hengl, 2018). Clay content in soil can influence the rainfall when it reaches down, as the soils with high clay content does not allow rainwater to infiltrate and forcing it to run off while reducing the river lag periods and increasing the flood risk.
7. **Weekly Precipitation levels of before and after the flood event** – Total weekly precipitation levels of the region based on the coordinates.
8. **Distance to water** – This is calculated in degrees for each datapoint a.k.a. coordinates in our dataset to nearest area of water using the country level free spatial data of inland waters from DIVA-GIS.
9. **Distance to River** – Distance of every point (X, Y) to nearest river in in meters.
10. **Land Cover Type** – This feature provides the spatial information on different types of physical coverage of the earth's surface such as forests, grasslands, croplands, lakes, and wetlands.

11. **NDVI (Normalised Difference Vegetation Index)** – By measuring the difference between near-infrared i.e. which vegetation strongly reflects and red light, the Normalized Difference Vegetation Index (NDVI) measures vegetation (GISGeography, 2022b).

    Formula to calculate NDVI -

    $$NDVI = \frac{(NIR–RED)}{(NIR + RED)} \tag{3.1}$$

    Where NIR is the light reflected in the near-infrared spectrum and RED is the light reflected by the red range of the spectrum.

3.4 Method of extracting features

The majority of features can be extracted from Google Earth Engine (GEE) using the local coordinates, since GEE delivers point values; consequently, having the coordinates of the area of interest simplifies the task exponentially.

3.5 Calculating Flood Fraction for Testing

The target variable that needs to be predicted is called ”flood fraction,” which refers to the percentage of the study area that is drowned in flood water. In this case, the study area is divided into rectangles of 1 km sq., and flood fraction will be calculated for each of these rectangles. Without the proper techniques and knowledge, this operation can take a very long time to finish manually, but satellite technology can assist with this enormous effort. In order to obtain the flood extent data for this study, a recent method developed by (Huang and Jin, 2020) is employed to efficiently map the flood zones and estimate the flood's extent. This method compares SAR (Synthetic aperture radar) images taken before and after flood occurrences with a predetermined threshold function in order to determine the extent of the flood. The threshold function utilised in this study has a value of 1.25, which was selected using the hit-and-miss technique. The primary benefit of employing SAR images was their capacity to see through dense cloud cover and rain, which is crucial for flood mapping. To make this technique more robust, even more information about the locations that are submerged in water for more than 10 months of the year was utilised. The end results achieved with the help of this technique are demonstrated in Figure 1.

The black polygon in Figure 1 indicates the area of Malawi where the flood portion needs to be estimated, and the blue colour in the Figure shows the flood water for the year 2019 in Malawi. These results are possible to download from the Google Earth engine in raster or vector format; however, in order to simplify the study, vector format was chosen. Now that each square in the dataset needs to have its flood percentage calculated, the entire map of Malawi is divided into small squares of area 1 km sq. using the minimum and maximum values of the X and Y coordinates that are available for Malawi. This method creates a grid with 23023 rows

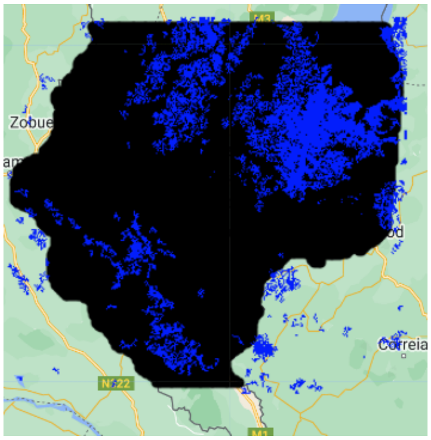

Figure 1: Flood extent map of Malawi

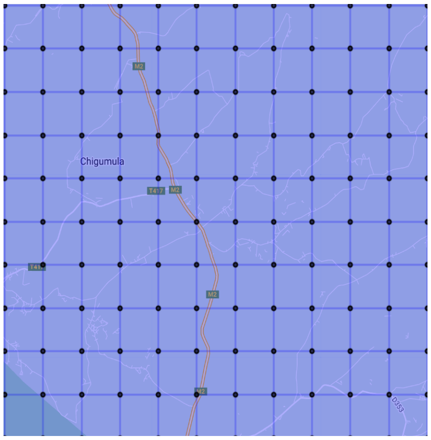

Figure 2: Grid Generated for Malawi Zoomed In

and one geometry column with the coordinates of a single square polygon, in this case a square whose flood % needs to be calculated. Figure 2 shows a Zoomed in version of the grid generated. The area of the flood present in one square is calculated and then divided by the overall area of that square, this process is repeated for every square, in order to determine the flood percentage of each square.

With the help of grid generated and flood extent map vector extracted from GEE the flood percentage of each square can be calculated (Figure. 3). Once each square's flood percentage has been calculated, the results can be added to the dataset as the new target using the coordinates X and Y.

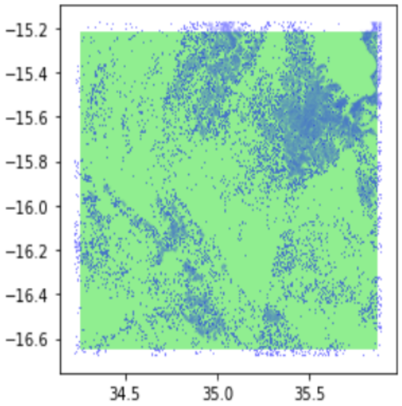

Figure 3: Plotting the grid over flood image.

4 Methodologies

Building a machine learning model involves numerous processes, all of which must be carried out precisely and meticulously to produce accurate predictions. This section includes all the data pre-processing steps carried out in this study. All the significant machine learning models used in this study are also covered in detail later in this section.

4.1 Data used

The dataset used for training the models is the dataset based on the flood events of 2015 and 2019 in Malawi, where external features from google earth engine are gathered. The final data consists of 16466 rows and 27 columns in respect to year 2015 for training and respect to year 2019 in the test dataset. The final goal is to train all the models on 2015 data and find the optimal model which gives lowest RMSE score on 2019 data. Because both datasets are of the same region but from different years, they contain some values that are same, such as X and Y coordinates, elevation, distance from river, and soil organic carbon, because these attributes do not vary greatly over time, the only varying variable is the precipitation values.

4.2 Data Pre-processing

Data pre-processing is the considered to be the first step that is taken before building any machine learning model, this step usually includes data cleaning, data engineering, feature selection, outlier detection, dimensionality reduction etc (Kotsiantis et al., 2006). The training set and test set for this study's data are based on two severe floods that occurred in Malawi in the years 2015 and 2019, respectively. The training set contains all the 2015 precipitation values, the target as well as all of the features

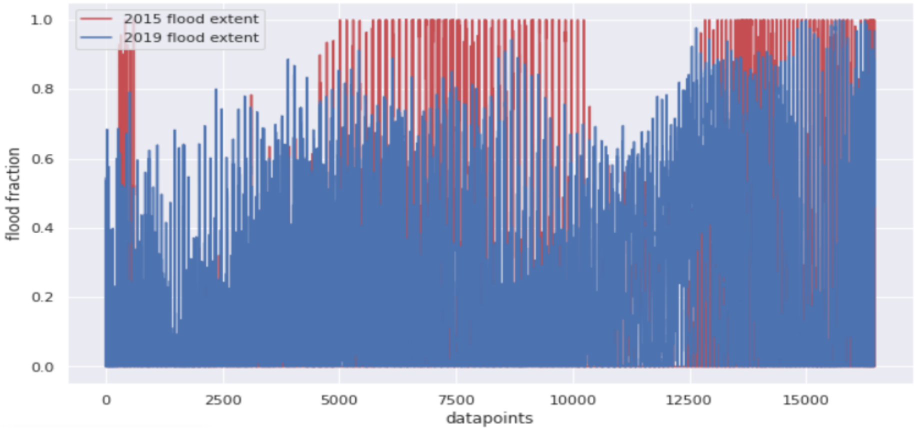

Figure 4: Data Analysis of Flood Extent in 2015 and 2019

Table 1: Metric comparison of target between train and test set

Target	Train set (2015)	Test set (2019)
Mean	0.076609	0.222080
Standard Deviation	0.228734	0.244667

listed in Section 3.3 with respect to the year 2015, and the testing set contains all of the 2019 precipitation, feature values and target for that year. To make the names of all the columns same in both the datasets the precipitation column names were reduced to name “intensity week” + ‘number of the week’ after that all the columns were aligned in both the datasets. Unwanted columns like the index and square Id were dropped from both the dataset to reduce the complexity in training the machine learning models. Lastly, 2015 data was normalized using MinMaxScaler() and the same normalizations were applied in 2019 data.

4.3 Data Analysis

The datasets utilised in this study underwent some analysis to provide a clearer understanding of the data. The initial comparison of the summaries of the two datasets revealed that there was a difference of 0.145 between the mean of the 2015 and 2019 dataset target, which indicates that the area experienced 14.5% more flooding in 2019 than in 2015. Even the standard deviation of the 2019 test set is higher compared to 2015 train set (as shown in Table 1) indicating more spreading out of data distribution in test set.

As most of the features were static in both the sets because certain features like distance to water, elevation, slope, soil carbon does not change with time, apart from precipitation, SSM and NDVI being dynamic in respect of year 2019. The metric comparison between NDVI and ssm values

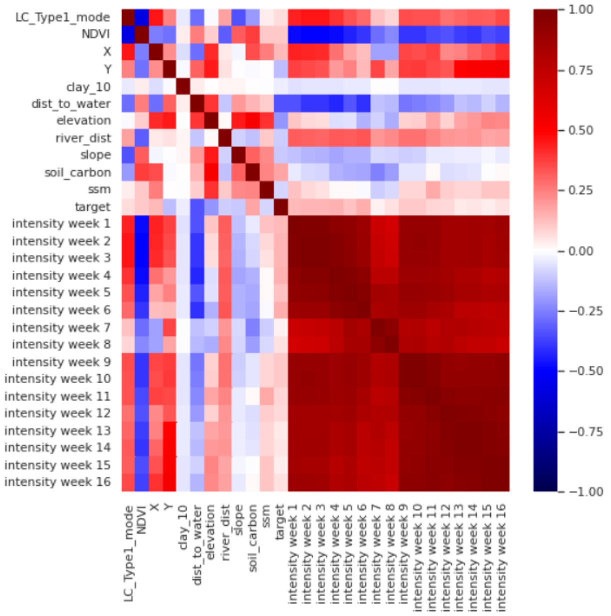

Figure 5: Correlation matrix of 2015 dataset

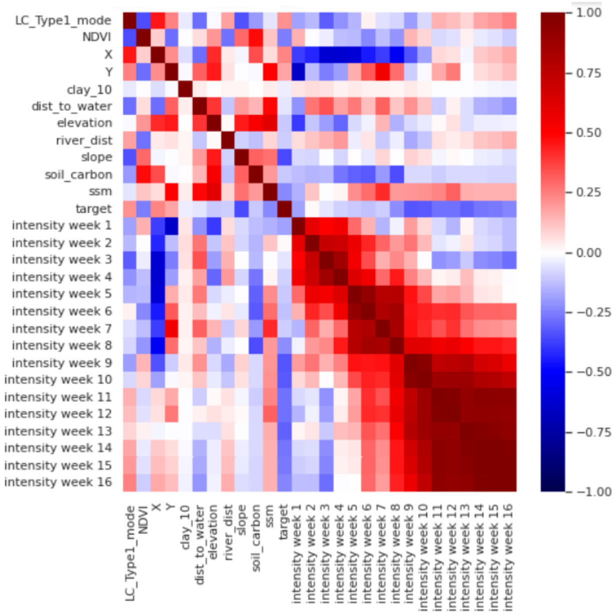

Figure 6: Correlation matrix of 2019 dataset

Table 2: Metric comparison of SSM between train and test set

NDVI	Train set (2015)	Test set (2019)
Mean	10.081447	22.734636
Standard Deviation	2.656261	0.909326

Table 3: Metric comparison of NDVI between train and test set

NDVI	Train set (2015)	Test set (2019)
Mean	0.331285	0.313523
Standard Deviation	0.118333	0.115540

Table 4: Correlation Values of Train and Test set.

Correlation	Train set (2015)	Test set (2019)
Max. Positive	Soil carbon = 0.173711	LC Type1 mode = 0.212423
Min. Positive	aspect = 0.0023	intensity week 2 = 0.009651
Max. Negative	Clay at band 200 = -0.023501	aspect = -0.006156
Min. Negative	Dist.to water = -0.339443	Slope = -0.355545

is shown in Tables 2 and 3, where it is noticeable that in 2019 there has been a 12.65% increase in surface soil moisture (ssm) and a little change in NDVI. After that, a correlation matrix is plotted to figure out the features with lowest positive and negative correlation. According to the results in Table 4, the aspect and clay content at band 200 has very little correlation with the 2015 and 2019 targets. On the other hand, precipitation especially intensity week 2 has a very low correlation with 2019 target data because other than rainfall the main factor of the 2019 Malawi floods was the cyclone Idai (Charrua et al., 2021).

4.4 Machine Learning Models

Data analysis is followed by the creation of distinct train and testing datasets that are scaled using the maximum and minimum values of the training set. A list of several Machine learning models was introduced, and each model was trained using the 2015 data before being tested on the 2019 data.

Machine Learning models that are used in this study -

1. **Linear Regression**

   This algorithm is used to predict continuous values based on the values of other variables in the dataset, the value that needs to be predicted is known as dependent variable and the other variables used to predict dependent variable are called independent variable. This algorithm is also known as the basic regression model which fits a linear or a straight line of the predicted values $Y_{pred} = (Y_1, \ldots, Y_j)$ and calculates the sum of squared error using the above equation of SSE.

2. **SVR (Support Vector Regression)**

   Support Vector Machines are supervised learning models with associated learning algorithms used in machine learning that examine data used for regression and classification analysis. The straight line needed to fit the data is referred to as the hyperplane in Support Vector Regression. This method, which uses the same fundamentals as the SVM's (support vector machines) to predict discrete values, (Noble, 2006).

3. **K Nearest Neighbour Regressor (KNN)**

   This algorithm is mostly used in classification problems due to its simplicity and efficiency. But this machine learning algorithm can be used for both regression and classification problems. Based on the value of *(K)*, the number of nearest neighbours, this algorithm determines the distance between a given datapoint and its nearest neighbours. There are various methods for calculating the distance between the datapoints and their nearest neighbours. The three most common measures are, **Euclidean distance, Manhattan distance, Minkowski distance.**

4. **Decision Tree Regressor**

   This algorithm makes decisions based on a flow-chart like tree structure which consist of a root, node and leaves. Each node in this tree is a testcase for an attribute (such as whether a coin flip results in heads or tails), and each leaf that is associated with that node is the result of that testcase. Both continuous and discrete decisions can be made using this technique. In regression problems Decision tree trains a model in the form of the tree to predict the continuous values. These trees are not made of a single observation as they are more adopted towards the dataset with many features i.e., high dimensional datasets.

5. **Random Forest Regressor**

   Random Forest is a supervised learning algorithm which follows an ensemble learning (Dietterich, 2002) approach. Random forest constructs number of Decision trees and outputs the mean of the all the predictions made by decision trees as the result.

6. **AdaBoost Regressor**

   AdaBoost, also known as Adaptive Boosting, is a Machine Learning technique that is an Ensemble Method. As primary idea underlying boosting is to combine a lot of weak learners or classifiers to create a strong model capable of producing good results. The most frequent AdaBoost algorithm is decision trees with one level, which is decision trees with only one split. These trees are often referred to as Decision Stumps. As in Random Forest vote of each tree has an equal importance on the outcome, but in the forest made with Adaboost some stumps have more importance than others.

7. **Gradient Boosting Regressor**

   This is a part of the family of potent machine learning approaches that have been extremely successful in a variety of specific real-world applications. Unlike Adaboost trees are bigger in Gradient Boost, it begins with a single leaf, rather than a tree or stump, and this leaf represents an initial guess for the weights of all the samples. This method deals with the problem of low bias and high variance by introducing a learning rate to scale the contribution.

8. **XGB Regressor (Xtreme Gradient Boost)**

   XGB stands for extreme gradient boost which is an open-source library that provides an efficient and effective implementation of the gradient boosting method(Chen et al., 2015). In addition to gradient boost, XG Boost has a separate approach for building trees in which Similarly score, and Gain identify the best node split. The formula for similarity Score is given by -

$$S.S \;=\; \frac{\left(\sum_{i=1}^{n} Res_i\right)^2}{\sum_{i=1}^{n}\left[Prev\ Prob_i * (1 -\ \ Prev\ Prob_i)\right] + \lambda} \tag{4.1}$$

   Where the residual is the difference between the actual and predicted values, the previous probability is the probability of split calculated at the previous step while keeping the initial probability at 0.5, and lambda(λ) is the regularisation parameter. The best split is chosen as the node with the highest Gain, after the Similarity score Gain is calculated using the formula –

$$Gain \;=\; Leaf\ Leaf_{sim} + Right\ Leaf_{sim} - Root_{sim} \tag{4.2}$$

9. **Catboost Regressor**

   CatBoost is a free and open-source library developed by the company Yandex (Yandex, 2020), for gradient boosting that was created specifically to handle categorical features, which are discrete features that can take on a limited range of values. For regression tasks where the objective is to predict a continuous target variable, the CatBoostRegressor is a type of model that can be used in the CatBoost library. The handling of categorical features, which are frequently present in real-world datasets, is one of CatBoost's key features. CatBoost uses specialised techniques like permutation feature importance, which measures the difference in the model's performance when a feature's values are randomly permuted, to automatically convert categorical features into numerical ones.

CatBoost also includes missing data handling functions that enables the model to continue learning even when some features have missing values. This is done by using an algorithm which calculates the missing values in a way that has least negative effect on the model performance. RMSE, MAE, Logloss, and other loss functions that can be specified during model building can all be used with CatBoost.

10. **Ridge Regression**

    This is a parameter estimation method that is commonly used to solve the multicollinearity problem that is when two or more feature are highly correlated in multiple linear regression. (McDonald, 2009). This linear regression technique aims to reduce the overall model complexity and reduce overfitting by introducing a penalty term known as the L2 regularization term to the cost function. This term is the sum of the squares of the model coefficients, multiplied by the scalar parameter lambda (λ). By making the model less complex, this term attempts to reduce the coefficients toward zero, which helps to prevent overfitting. This model can make good predictions even when the features are correlated to each other because of its ability to shrink the coefficients.

5 Experiments and Results

This section of the study outlines all the results obtained by different models used in this study along with their hyperparameter Configuration. Both the 2014 dataset and 2019 dataset were utilized in the experiment conducted. First K-fold cross validation was done on both the datasets where the entire datasets were divided into 10 folds, in which in every iteration randomly 9 folds were used in training and 1-fold was used in testing. Later the datasets were divided into training and testing to check the performance of the models on same year.

5.1 Experiments Conducted

The metrics chosen for evaluation of the models is RMSE (Root mean square error) which is basically the square root of the MSE, where RMSE = 0 means a perfect fit. To check the performance of the models splitting the dataset of one year, say 2014 data into training data and test data would simply mean dividing coordinates of the dataset leaving important information of certain region in the area being unavailable to the model. Although this method produces positive findings when evaluated with K-fold cross validation (Table 5) and when the 2015 and 2019 data is divided separately into 70% of the data used for training and 30% used for testing. After applying K-Fold cross validation (where $\boldsymbol{K}$ = 10) to all the models mentioned in Section 4.3 the RMSE (Root Mean Square Error) and K-Fold scores obtained are displayed in Table 5 and 6. According to Table 6, when 2015 and 2019 data is analysed using K-fold cross validation with

Table 5: Mean RMSE of K-Fold Cross validation

Model	Train set (2015)	Test set (2019)
Catboost	0.102716	0.109994
XGB	0.120654	0.131949
Gradient Boosting	0.115145	0.125763
Random Forest	0.101452	0.109587
Decision Tree	0.138380	0.144700
KNN	0.128324	0.123893
Adaboost	0.175628	0.169232
SVR	0.164650	0.140567
Ridge	0.190167	0.179237
Linear	0.188968	0.176805

Table 6: RMSE scores when 30 % data used for testing

Model Name	RMSE on 2015	RMSE on 2019 data
Linear	0.188	0.179
SVR	0.161	0.142
Random Forest	0.098	0.106
Gradient Boosting	0.109	0.115
AdaBoost	0.176	0.171
Decision Tree	0.131	0.145
XGB	0.101	0.111
K Neighbours	0.133	0.127
Cat boost	0.102	0.109
Ridge	0.188	0.182

k = 10, the Random Forest regressor performs best with the lowest mean RMSE score on 10 folds. When each one of the models were trained on 70% of the data and tested on 30% of the data chosen randomly from both datasets, the random forest regressor had the lowest RMSE scores.

Because flooding is a natural and random process, there is relatively very low similarity between the targets of 2015 and 2019. Apart from that, gathering data for a year and predicting flood extents for the same year will render the model unusable because availability of the data can be a problem, the predictions would be faulty and training data will be insufficient. In actuality, no one will gain from this model since, to successfully estimate future flood extent levels, training data must be resilient and contain all the relevant information about the region. Therefore, to produce reliable predictions a separate dataset based on the flood events 2019 has be selected in this study for testing. This dataset contains the target flood fraction calculated in Section 3.5 which needs to be predicted by the model along with all the features in respect to year 2019. The end goal is to achieve the lowest RMSE on the 2019 data with training done on 2015 data.

5.2 Evaluation of the Models

After training the models on entire 2015 dataset, all the models were tested using the entire 2019 data to predict the flood extent in Malawi for the year 2019. When the testing dataset is as large as the training dataset, as in this study, it is challenging to achieve very low RMSE scores since the datapoints tend to scatter significantly on a residual plot due to their massive numbers. Therefore, only depending upon RMSE scores for model performance evaluation is not the best option in this case, as these scores can be affected by model bias, outliers, or repeated data. That's why this study accommodates two other techniques to evaluate hydrological model performance which are suggest by (Ritter and Munoz-Carpena, 2013). In this technique NSE (Nash–Sutcliffe Efficiency coefficient) and $\mathrm{SD_F}$ are used, where $\mathrm{SD_F}$ is the standard deviation of all the features and then the models are classified into 4 categories which are very good, good, acceptable, and unsatisfactory. The Equation to calculate the Standard deviation of features is given by –

$$SD_F \ = \sqrt{\ \sum_{i=1}^{n} (V_{fi} + V_{f1} \ldots V_{fn})} \tag{5.1}$$

Where $\mathrm{SD_F}$ is the standard deviation of all the features, $\mathrm{V_{fi}}$ is the summation of the variance of each individual feature and n is the total number of features present in the dataset. This $\mathrm{SD_F}$ is then used to calculate the NSE whose equation is given by –

$$NSE = 1 - \ \frac{\sum_{i=1}^{N} (O_i - P_i)^2}{\sum_{i=-1}^{N} \left(O_i - \overline{O}_i\right)^2} \tag{5.2}$$

$\mathrm{O_i}$ and $\mathrm{P_i}$ in Equation 5.2 are the observed and predicted observations. $\overline{O}$ is the mean of observed values, where NSE = 1 means a perfect fit the equation 5.2 can also be written as -

$$NSE \ = 1 - \left(\frac{RMSE}{SD_o}\right)^2 \tag{5.3}$$

here $\mathrm{SD_O}$ is the same standard deviation of output calculated in above equation.

Where models with very good performance rating have **NSE $\geq$ 0.9** and **RMSE $\div$ SD $\leq$ 0.31**, Good performance rating has **NSE $\geq$ 0.8** and **RMSE $\div$ SD $\leq$ 0.45**, Acceptable performance rating has the **NSE $\geq$ 0.65** and **RMSE $\div$ SD $\leq$ 0.83**, else the model performance is rated unsatisfactory.(Ritter and Munoz-Carpena, 2013).

When training on 2015 data is followed by testing on 2019 data Table 7 clearly illustrates that SVR outperforms the random forest regressor with RMSE of 0.255, which performed best while training and testing were done on the same year of data.

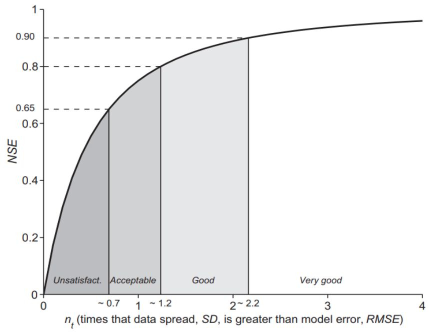

Figure 7: Performance indicator graph.

Table 7: Metric scores on test dataset

Model Name	RMSE	NSE	RMSE/$\mathrm{SD_O}$
Linear	0.358	0.752	0.497
SVR	0.255	0.874	0.354
Random Forest,	0.319	0.802	0.443
Gradient Boosting	0.349	0.764	0.484
AdaBoost	0.301	0.824	0.419
Decision Tree	0.357	0.753	0.496
XGB	0.404	0.683	0.562
KNN	0.339	0.777	0.471
Cat boost	0.320	0.802	0.444
Ridge	0.312	0.812	0.433

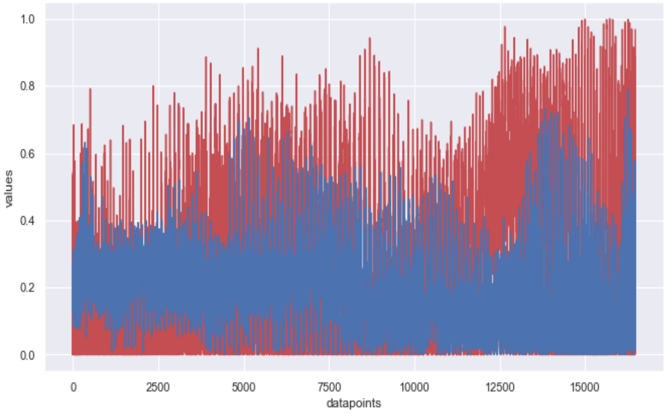

Figure 8: Predicted test values by Catboos Regressor

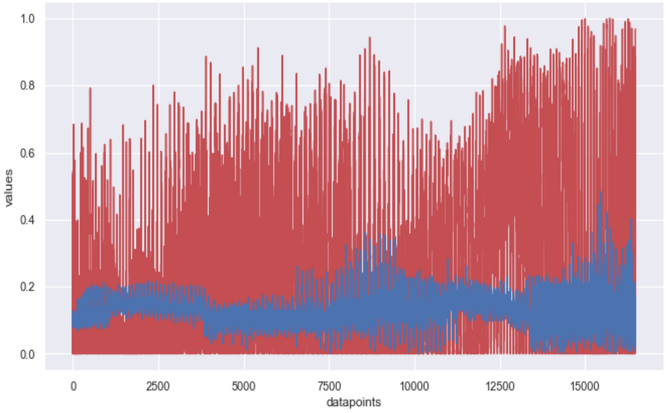

Figure 9: Predicted test values by SVR Regressor

5.3 Results

On the 2019 data, the Support Vector Regressor with default ‘rbf’ kernel and degree = 3 has the lowest RMSE score, 0.255 (Table 7), and the SD_F of the features is 0.73. When eqns. (5.1), (5.2) & (5.3) are used to test the model's performance, the result obtained is 0.874 for the NSE and 0.354 when the RMSE is divided by SD_F. According to the performance criteria provided above in Section 5.2 and the performance indicator graph in Figure 7, the SVR's performance is close to very good.

Figure 11 and 12 illustrates all the flood extent graphs created with different models where the predictions provided by Catboost regressor are very close to the actual values which is target_2019, despite having a larger RMSE score than SVR. Evaluating equations eqns. (5.1), (5.2) & (5.3) in respect to RMSE of Catboost regressor the outcomes are 0.802

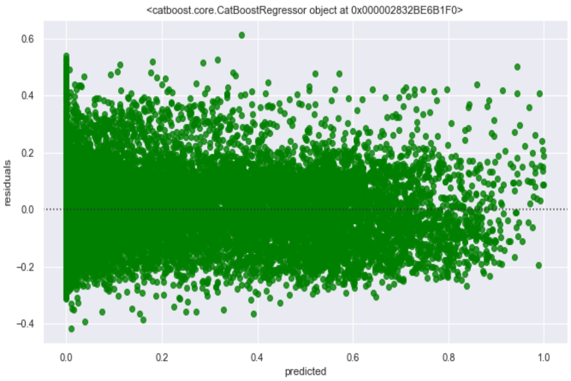

Figure 10: Residual Plot of 2019 Data on Catboost regressor

of **NSE** and 0.444 when **RMSE** divided by $\mathbf{SD_F}$. According to criteria the performance of the Catboost regressor better than good as the predicted flood extent is close to the original values making the predictions more robust. As evidenced by the Figure 8 and 9 the values predicted by Catboost are closer to the actual values of the target values of 2019 flood extent. Therefore, it can be stated that Catboost performs better than the other models that were used in this study for flood extent prediction in Malawi. As in the case of SVR, it predicts the values within the decision boundaries of the hyperplane which results in some predicted values that are far away from actual values making the model predictions unreliable. This study makes use of additional features that provide useful information about the geography of the area. The reason for using more than one performance indicator in this study is to assists in verification of the best performing model.

5.4 Future Scope

Flood prediction is a broad subject with several features to evaluate, models to consider and configure; this study focuses on some of the elements and models. There are numerous unknown features that may be retrieved from GEE from different satellites and time stamps relating the flood occurrence; even the area of grid size can be changed depending on the availability of the coordinates. Alternative deep learning techniques such as stacking can be deployed to produce enhanced results, as well as automated (optimal) hyperparameter tuning of the models utilised in this study. Flood extent prediction of another location outside Malawi can be done simply by extracting the relevant coordinates of the area and then extracting all the required information from those coordinates by repeating the methods accomplished in this study. The coordinates of an area can be generated by Collect Earth grid generator tool which is an earth engine app made by open fortis (fortis, 2022). After selecting an area of interest and a CRS system, a csv file with the coordinates can be downloaded and uploaded to GEE to extract more information.

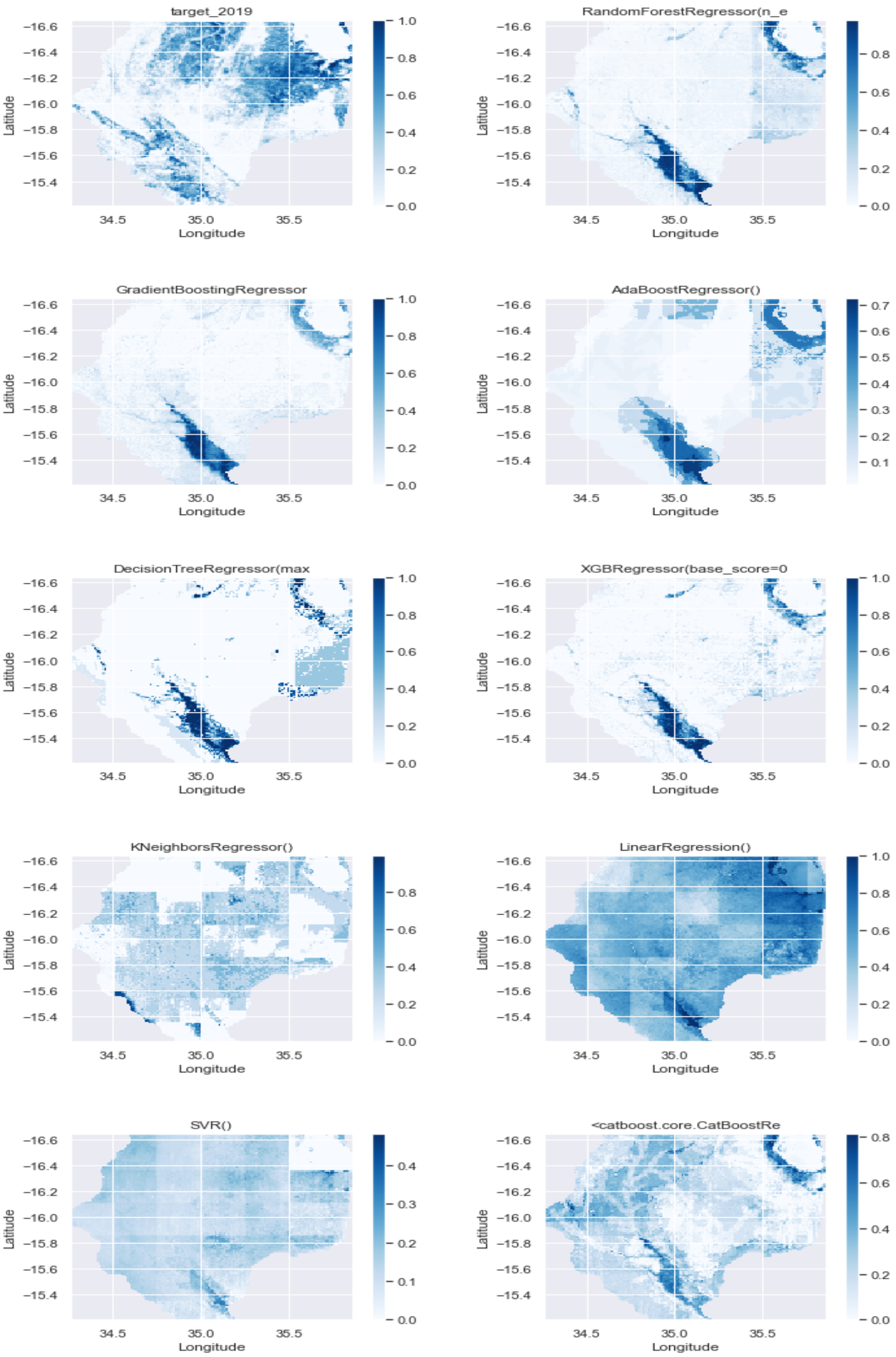

Figure 11: Flood Extent prediction Graphs

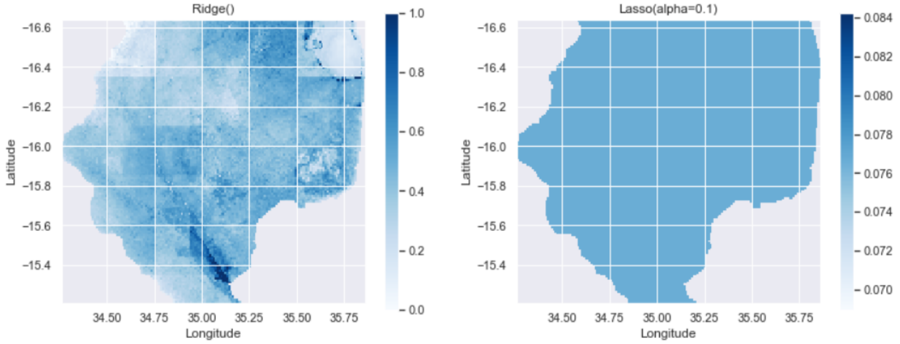

Figure 12: Flood Extent prediction Graphs

6 Conclusions

Considering floods being the most dangerous natural disasters, there are many factors involved in an occurrence of floods and there is no exact answer to main reason behind a region being flooded. Where rainfall plays a minor part when compared to other features like the elevation, slope, land cover type, soil moisture etc which gives an enormous amount of information about the region’s geography and terrain. When predicting natural disasters like floods, having a solid understanding of the region's geography can be very helpful. This information aids in a better understanding of the minor details in the data of the area where specific feature engineering techniques may be used. In this study, various regression models were applied in search of best performing model that can be utilized for flood extent prediction in the region of Malawi. The experiments conducted in this study revealed that in the case of disaster prediction modelling depending on one evaluation metric can often lead to incorrect model selection. The Random Forest Regressor had the lowest RMSE score when the model training and testing were conducted in the same year, whereas the SVR had the lowest RMSE of when the complete 2014 data set was used to train the models and the entire 2019 data set was used to test the models. Since the SVR predicts values within the hyperplane's decision boundaries, critical information may be left outside those boundaries during model fitting, therefore the predictions made by SVR are often incorrect and not close to the actual values. The predictions provided by Catboost were the closest to the actual values, according to the flood extent graphs in Figure 10 and 11, except for having a greater RMSE score than SVR but a good NSE score. These results claim catboost as the best performing model for flood extent prediction in Malawi when taken along other features mentioned in this study. With the aid of GIS and predictive modelling, the work provided in this paper reveals all the traits and methodologies that can help to anticipate the extent of a flood in a particular area. This work is not limited to specific area or disaster and will also act as a foundation for future research conducting in the area of natural disaster prediction.

REFERENCES

ADAMS, T. E. & PAGANO, T. 2016. *Flood Forecasting: A Global Perspective*, Elsevier Science.

CHARRUA, A. B., PADMANABAN, R., CABRAL, P., BANDEIRA, S. & ROMEIRAS, M. M. 2021. Impacts of the tropical cyclone idai in mozambique: A multi-temporal landsat satellite imagery analysis. *Remote Sensing,* 13**,** 201.

CHEN, T., HE, T., BENESTY, M., KHOTILOVICH, V., TANG, Y., CHO, H. & CHEN, K. 2015. Xgboost: extreme gradient boosting. *R package version 0.4-2,* 1**,** 1-4.

DAMLE, C. & YALCIN, A. 2007. Flood prediction using Time Series Data Mining. *Journal of Hydrology,* 333**,** 305-316.

DATACARPENTRY.ORG. 2022. *Introduction to Geospatial Concepts - Introduction to Raster Data* [Online]. Available: `https://datacarpentry.org/organization-geospatial/01-intro-raster-data/#:~:text=Raster%20data%20is%20any%20pixelated,we%20represent%20any%20digital%20image`. [Accessed 13/08/ 2022].

DIETTERICH, T. G. 2002. Ensemble learning. *The handbook of brain theory and neural networks,* 2**,** 110-125.

FORTIS, O. 2022. Collect earth grid generator. `https://collectearth.users.earthengine.app/view/collect-earth-grid-generator`

GHORBANI, M. A., KAZEMPOUR, R., CHAU, K. W., SHAMSHIRBAND, S., TAHEREI GHAZVINEI & P. 2018. Forecasting pan evaporation with an integrated artificial neural network quantum behaved particle swarm optimization model: a case study in Talesh, northern Iran. *Eng. Appl. Comput. Fluid Mech.,* 12 (1),**,** 724–737.

GISGEOGRAPHY. 2022a. *What is an Aspect Map?* [Online]. Available: `https://gisgeography.com/aspect-map/` [Accessed 13/08/ 2022].

GISGEOGRAPHY. 2022b. *What is NDVI (Normalized Difference Vegetation Index)?* [Online]. Available: `https://gisgeography.com/ndvi-normalized-difference-vegetation-index/` [Accessed 23/07/2022].

GORELICK, N., HANCHER, M., DIXON, M., ILYUSHCHENKO, S., THAU, D. & MOORE, R. 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment,* 202**,** 18-27.

GREENTUMBLE. 2016. *What are the Human Causes of floods* [Online]. [Accessed 23/07/2022].

HAN, S. & COULIBALY, P. 2017. Bayesian flood Forecasting methods: a Review. *J. Hydrol.,* 551**,** 340-351.

HENGL, T. 2018. *Clay content in % (kg / kg) at 6 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250 m resolution* [Online]. Zenodo. Available: `https://zenodo.org/record/2525663#.YvqQ_nbMK3A` [Accessed 15/08/2022].

HUANG, M. & JIN, S. 2020. Rapid flood mapping and evaluation with a supervised classifier and change detection in Shouguang using Sentinel-1 SAR and Sentinel-2 optical data. *Remote Sensing,* 12**,** 2073.

JONES-BOS, R. 2011. As the Mississippi floods, follow the Dutch model. Available from: `https://www.washingtonpost.com/opinions/as-the-mississippi-floods-follow-the-dutch-model/2011/05/23/AGP9kICH_story.html` [Accessed 23/07/2022.]

KIA, M. B., PIRASTEH, S., PRADHAN, B., MAHMUD, A. R., SULAIMAN, W. N. A. & MORADI, A. 2012. An artificial neural network model for flood simulation using GIS: Johor River Basin, Malaysia. *Environmental earth sciences,* 67**,** 251-264.

KOTSIANTIS, S. B., KANELLOPOULOS, D. & PINTELAS, P. E. 2006. Data preprocessing for supervised leaning. *International journal of computer science,* 1**,** 111-117.

KUMAR, L. & MUTANGA, O. 2018. Google Earth Engine Applications Since Inception: Usage, Trends, and Potential. *Remote Sensing,* 10**,** 1509.

MCDONALD, G. C. 2009. Ridge regression. *Wiley Interdisciplinary Reviews: Computational Statistics,* 1**,** 93-100.

MOSAVI, A., OZTURK, P. & CHAU, K.-W. 2018. Flood Prediction Using Machine Learning Models: Literature Review. *Water,* 10.

MUCKLEY, L. & GARFORTH, J. Multi-Input ConvLSTM for Flood Extent Prediction. International Conference on Pattern Recognition, 2021. Springer, 75-85.

NOBLE, W. S. 2006. What is a support vector machine? *Nature biotechnology,* 24**,** 1565-1567.

OECD 2016. *Financial Management of Flood Risk.*

OPELLA, J. M. A., HERNANDEZ & A., A. 2019. Developing a flood risk assessment using support vector machine and convolutional neural network: a conceptual framework. *In: 2019 IEEE 15th International Colloquium on Signal Processing & Its Applications*, 260 - 265.

QGIS-DOCUMENTATION. 2022. *Coordinate Reference Systems* [Online]. Available: `https://docs.qgis.org/3.22/en/docs/gentle_gis_introduction/coordinate_reference_systems.html` [Accessed 13/08/ 2022].

RITTER, A. & MUNOZ-CARPENA, R. 2013. Performance evaluation of hydrological models: Statistical significance for reducing subjectivity in goodness-of-fit assessments. *Journal of Hydrology,* 480, 33-45.

SATO, T. 2006. Fundamental characteristics of flood risk in Japan's urban areas. *A better integrated management of disaster risks: Toward resilient society to emerging disaster risks in mega-cities, Tokyo: TERRAPUB and NIED*, 23-40.

SHAFAPOUR TEHRANY, M., SHABANI, F., NEAMAH JEBUR, M., HONG, H., CHEN, W. & XIE, X. 2017. GIS-based spatial prediction of flood prone areas using standalone frequency ratio, logistic regression, weight of evidence and their ensemble techniques. *Geomatics, Natural Hazards and Risk,* 8, 1538-1561.

WETTERHALL, F. 2017. Flood forecasting. *Advanced School and Workshop on Subseasonal to Seasonal (S2S) Prediction and Application to Drought Prediction* Online: YouTube.

YANDEX. 2020. *Comapny description* Online. Available: `https://yandex.com/company/` [Accessed 27/08/ 2022].

ZINDI. 2020. *UNICEF Arm 2030 Vision #1: Flood Prediction in Malawi* [Online]. Available: `https://zindi.africa/competitions/2030-vision-flood-prediction-in-malawi/data` [Accessed 13/06/2022].

Appendix

1. Introduction to Google Earth Engine -
   The data utilised in this study was obtained from the Google Earth Engine, which is a planetary scale platform for Earth science data and analysis. Google Earth Engine blends planetary-scale analytical capabilities with a multi-petabyte database of satellite photos and geographical information.With the provided coordinates of the region, the user may extract essential information about any region of interest and then download the multi-spectral images or .SHP vector files, which can then be converted into csv files for future use.

Google Earth Engine can be accessed by -
`https://earthengine.google.com//`

2. MASDAP (Malawi Spatial Data Platform) - A public platform for GIS Data to support development in Malawi.
   A web-based data sharing facility called the Malawi Spatial Data Platform (MASDAP) was introduced in November 2012 and is run by the National Spatial Data Centre in the Department of Surveys in association with the National Statistics Office and other technical Ministries. Users can upload, manage, and browse spatial data, create, and share maps, search for and download pertinent documents, and have access at multiple levels (viewing, editing, uploading) depending on their needs and the nature of their work.

   MASDAP can be accessed by -
   `https://www.masdap.mw/`