

Literature Review

F22-4202 \_ How Climate Change Effects Flood

Final Year Project

Ahmed Aqil (K19-1334)

Shayan Malik (K19-1375)

Umer Murtaza (K19-1349)

Department of Computer Science BS(CS)

FAST-NUCES Karachi

# How Climate Change affects Flood

***Abstract – Flooding is a natural disaster that can have devastating effects on people, their homes, and their community's infrastructure. It can cause significant damage and result in economic losses, including destruction of property and disruption of daily life. Flooding can occur anywhere in the world and can have severe consequences for those affected by it. There has been a recent increase in the use of machine learning techniques for flood assessment, as these methods are effective at identifying patterns and relationships in data. Thus we are doing research on this, which involves collecting and analysing data about variables that influence flooding over time and using Machine learning techniques which allow us to efficiently analyse and train our model with large amounts of data, making them a valuable tool for understanding floods. The goal of this research is to identify the key factors that contribute to flood prediction and analysis. By the end of the study, we hope to have a clear understanding of the parameters that are most important in predicting and analysing flood events.***

1. INTRODUCTION

Floods are the most damaging of all natural disasters, seriously harming infrastructure, agriculture, society's economy, and human lives. Therefore, there is a demand on governments to create trustworthy and accurate maps of flood risk areas and to further plan for sustainable flood risk management that focuses on prevention, protection, and preparedness. Models for flood prediction are crucial for managing extreme events and assessing hazards. The development of future evacuation modelling, policy ideas, and analysis all benefit greatly from robust and accurate prediction. Therefore, it is crucial to emphasise the value of cutting-edge technologies for both short- and long-term flood forecasting and other hydrological events in order to reduce damage. Research on the improvement of flood prediction models has helped to lower risks, suggest policy changes, minimise the loss of life from floods, and lessen property damage. Machine learning (ML) techniques have greatly advanced prediction systems over the past two decades, offering higher performance and more affordable solutions by simulating the intricate mathematical expressions of the physical processes of floods. The potential and extensive advantages of ML led to a sharp rise in its acceptance among hydrologists. Researchers hope to find more precise and effective prediction models by incorporating innovative machine learning techniques and hybridising already existing ones.

1. LITERATURE REVIEW

According to the paper [1] they tried to prove that every real world problem can be mapped to a function. In order to achieve the results, the input (data) and the mapping processes should be correct. They discussed how two of the machine learning algorithms namely NARX (Non-linear Autogenous Exogenous Model) and SVM (Support Vector Machine) helped in forecasting floods early so damage could be minimised.

The major findings of the research were:

* The parameters used to predict flood are majorly non-linear thus many linear practices for time series predictions were used for flood predictions such as AR, ARMAX, Kalman Filtration. But in most cases the linear approach has not provided appropriate results. Therefore, complex structures of real world parameters that are used to predict floods could be handled much more efficiently using non-linear models.
* There were some comparisons between models such as NARX with AAN which showed that NARX outperformed the conventional neural network when it came to prediction of water flow firm rain data.

The aim of this study [2] was to propose a reliable solution for a model which could predict floods ahead of time as static neural networks with back propagation were not as efficient. Flood forecasting requires dynamic nonlinear properties as the data used is always complexly structured therefore, NARX is used in this study for the prediction of dynamic characteristics of water flood level.

The major finding of the paper were:

* The proposed NARX model was able to give great results as it predicted flood way ahead of time that is 10.83 hours using the Klang River data but the basin information was excluded from the input parameters as it did not provide any significant effects to prediction results.
* It was observed that the prediction performances were showing great results with Best Fit of 87% and almost no errors.

The purpose of this study [3] was to select better attributes, and discharge level of Rawal Lake, Islamabad so better perdition can be made. This paper was divided into two parts. In Part I, they present the month on month trends for hydrological parameters and in Part II, they chose the most suitable hydrological parameters for classification. SVM was one of the comparing algorithms for the above mentioned study alongside MLP, ARIMA and RBF and the results could be seen in the picture, where SVM proved to be the best among them.

In this paper [4] multiple neural network models were used to predict flood water level from meteorological data sets. The main models were ENN and BPN both of them were used with the integration of EKF also. As the data related to fold is mostly complex only non-linear techniques were used. Both models were compared in terms of their tracking errors. ENN was determined to be better in terms of RMSE result only but BPN model took the lead in terms of FPE with error less than 1. BPN and ENN performed extremely better when integrated with EKF which showed that EKF is the best nonlinear estimator for the BPN and ENN model because it is able to track the dynamics of the nonlinear system itself. Hence this was observed that nonlinear neural network models for flood water level prediction are reliable network structures.

**Methodology in Papers:**

A flood warning system was created for which data [8] has been collected from remote sensing satellites and ground application. The parameters considered are the amount of rainfall and water-level of nearby water bodies. Gradient Boost Algorithm is used to obtain a non-linear relationship between the total sum of rainfall and runoff, thus reducing the mean-squared error. For the datasets which were not a part of the training dataset, the Decision Tree Algorithm is used to predict floods on datasets which were not used in the training dataset. Random Forest Method was used to make different decision trees and get the prediction from each tree.

The Decision Tree Algorithm gave an accuracy of 94.4%, the Gradient Boost Algorithm gave an accuracy of 87.9% whereas the Random Forest Algorithm gave an accuracy of 92.4%. Hence, the Decision Tree Algorithm was chosen for the model.

In this paper [7], they propose the convolutional LSTM (ConvLSTM) to extract spatiotemporal aspects of hydrological information by integrating a convolutional neural network (CNN) and a long short term memory network (LSTM). In this study, the region of interest is segmented into grids based on latitude and longitude, and the rainfall and discharge data gathered at stations are merged into tensors based on station coordinates. Our input feature is a two-dimensional time series with spatial information, as opposed to one-dimensional time series. The trials also demonstrate that the ConvLSTM performs better than the current models in terms of peak discharge and flood arrival time, demonstrating that it is a viable option.

Different techniques, including Linear Regression, Logistic Regression, and Decision Tree, are utilised in this research [6]to classify flood predictions. The classification and differentiation of algorithms are based on some key characteristics like accuracy, time, and speed. The government-provided dataset on the 2016 Kerala floods will be presented in this study. The decision tree method produces findings that are more accurate and easy to grasp when compared to the other algorithms. For nonlinear datasets, the decision tree also generates a model. This nonlinearity can be used to determine the precision of a dataset that is logistic or linear. As evidenced by the results of the comparison, the decision tree offers greater accuracy when compared to other straightforward machine learning techniques.

This paper's [5] main contribution is to show the current state of ML models for flood prediction and to provide insight into the most appropriate models. In order to give a thorough overview of the many ML algorithms utilised in the area, the literature where ML models were benchmarked through a qualitative examination of robustness, accuracy, efficacy, and speed is specifically explored in this work. The performance comparison of ML models offers a thorough grasp of the various methodologies within the context of a thorough assessment and discussion. Traditional rainfall and water level measurements made with ground rain gauges or more recent remote sensing methods using satellites, multi-sensor systems, and/or radars are the sources of the dataset.

The historical dataset of hourly, daily, and/or monthly values is divided into individual sets to construct and evaluate the learning models, whether using a radar-based dataset or ground gauges to create a prediction model. The ANN technique was employed for prediction with higher accuracy compared to conventional statistical models. Since they were originally used to model flood prediction in the 1990s, ANN algorithms have been the most used. ANNs interpret historical data rather than the physical qualities of a catchment. ANNs are therefore regarded as trustworthy data-driven tools for building black-box models of intricate and nonlinear interactions between rainfall and flooding. Backward Propagation, Recurrent Neural Networks, and Feed Forward Neural Networks were also used.

This study also used the MLP (Multilayer Perceptron), a kind of FFNN that uses supervised learning of BP to train a network of interconnected nodes with numerous layers. Time-series flood forecasts were made using DWT (Discrete Wavelet Transform) hybrid models, such as wavelet-based neural networks (WNNs), which integrate WT and FFNNs, and wavelet-based regression models, which incorporate WT and multiple linear regression (MLR). As ML alternatives to ANNs, SVM (Support Vector Machine) and SVR (Support Vector Regression) have gained appeal among hydrologists for flood prediction. One of the contributors to predictive modelling with widespread use in flood simulation is the ML approach of DT. DT employs a decision tree with branches leading to target values for leaves. The most promising flood prediction techniques are thus presented in this work for both long-term and short-term floods. Investigated are also the key trends in raising the standard of flood prediction models. The most successful methods for enhancing ML techniques are said to be hybridization, data decomposition, algorithm ensemble, and model optimization.

**Models and Algorithms Used:**

Decision Tree:A non-parametric supervised learning technique for classification and regression is called a decision tree (DT). The objective is to learn straightforward decision rules derived from the data characteristics in order to build a model that predicts the value of a target variable. A tree can be seen as a piecewise constant approximation. It can be used in decision analysis to formally and graphically reflect decisions and decision-making. It employs a decision-tree-like approach, as the name suggests. Though it is primarily used in data mining to develop a plan for achieving a certain objective, it is also frequently used in machine learning.

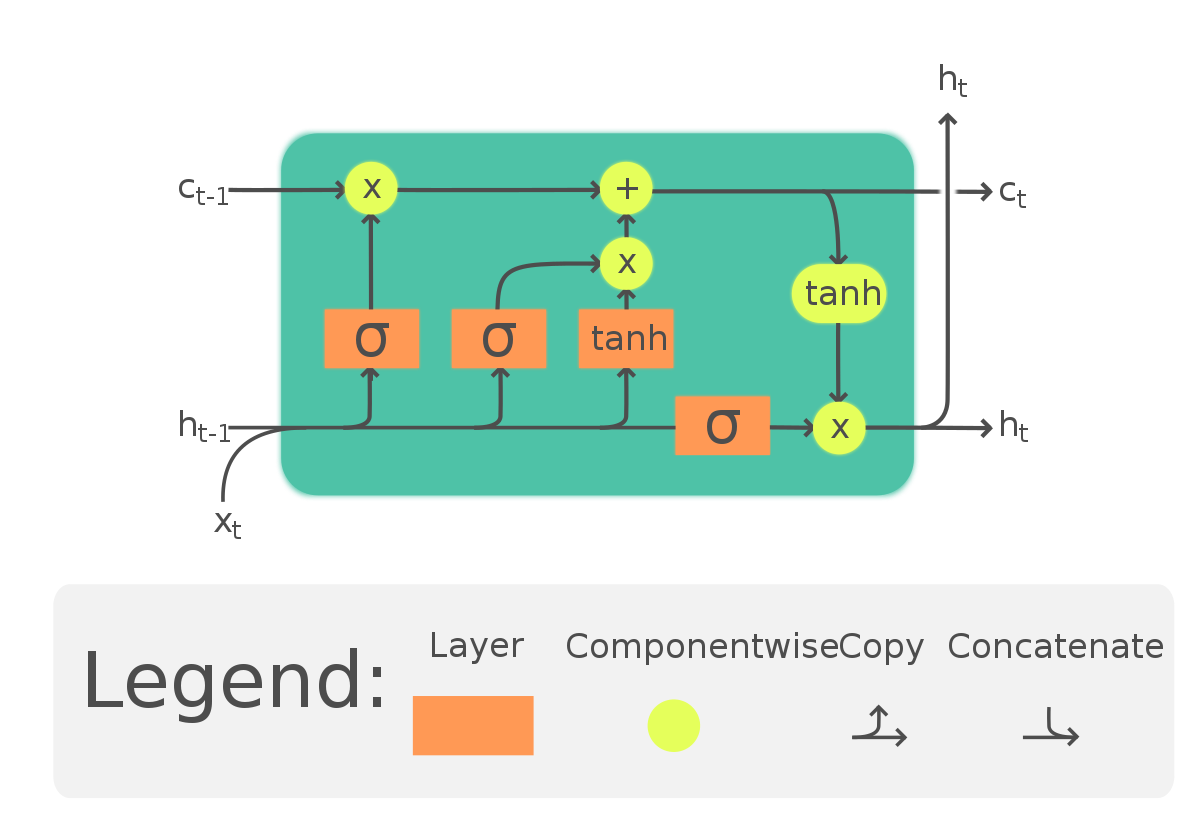
Gradient Boosting:One of the most well-liked machine learning techniques for tabular datasets is gradient boosting. It has excellent usability, can deal with missing values, outliers, and large cardinality categorical values on your features, and is strong enough to detect any nonlinear relationship between your model target and features. Gradient boosting is one of the variants of ensemble methods where you create multiple weak models and combine them to get better performance as a whole.

Random Forest:Like its name suggests, a random forest is made up of numerous independent decision trees that work together as an ensemble. Every tree in the random forest spits out a class forecast, and the classification that receives the most votes becomes the prediction made by our model. The fundamental concept behind random forest is a simple but powerful one: *“A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.”*

**

Multi Layer Perceptron: While BPNNs are currently often employed in this field, the MLP, a more sophisticated version of ANNs, has recently become more well-known. The MLP is a subclass of FFNN that trains its multiple-layer network of interconnected nodes using supervised learning from BP. The MLP is characterised by simplicity, nonlinear activation, and a large number of layers. These qualities led to the model's widespread application in complicated hydrogeological models and flood prediction. MLP models were shown to be more effective and more generalizable in a review of ANN classes used in flood modelling. However, it is typically discovered that the MLP is more challenging to optimise.

Long Short Term Memory Networks: Long short-term memory networks is a special type of RNN (Recurrent Neural Network) and is capable of processing the entire series of data. LSTMs are designed specifically to avoid long-term dependency problem. LSTM has a chain-like structure, but every repeating module has a different structure. It consists of four neural network layers which are interacting in a special manner and uses output from the previous state as an input. Multivariate LSTMs are trained to predict the next time step in a sequence given a series of previous time steps. The model does this by learning to recognize patterns in the data and using this information to make predictions about future time steps. The LSTM model is able to learn long-term dependencies in the data by using a special type of memory cell, called a "memory gate," which can selectively retain or forget information from the cell's memory. The LSTM model also uses input, output, and forget gates to control the flow of information into and out of the memory cell. To train a multivariate LSTM model, we have defined the number of time steps (also known as the "window size") to use as input, as well as the number of input and output variables temperature, amount of rainfall, soil moisture, humidity , water runoff, and vapour pressure. The model is then trained using a sequence of input data and the corresponding output data, using an optimization algorithm. Once the model is trained, you can use it to make predictions by providing it with a series of input time steps and asking it to predict the next time step in the sequence. When a value is zero, "let nothing through" and when a value is one "let everything through" respectively.

****

ARIMA: Autoregressive integrated moving average (ARIMA) is a statistical model for forecasting time series data. It is a generalisation of the simpler autoregressive (AR) and moving average (MA) models, and it allows for the incorporation of both past values and past errors in the model. Here's how ARIMA works: Autoregression (AR): The AR component of the model captures the relationship between the time series and its own lagged (i.e past) values. It models the time series as a linear function of its past values, with coefficients (called "AR coefficients") that determine the strength and lag of the relationships. Integration (I): The I component of the model accounts for non-stationarity in the time series data, such as trend or seasonality. It does this by taking the difference between the time series and its own lagged values, which removes the trend and leaves a stationary series. The degree of differencing is called the "differencing order." Moving average (MA): The MA component of the model captures the relationship between the time series and residual errors from past predictions. It models the time series as a linear function of the past errors, with coefficients (called "MA coefficients") that determine the strength and lag of the relationships. To fit an ARIMA model, you need a time series dataset with timestamps and the corresponding values. You can then use statistical techniques to estimate the values of the AR, I, and MA coefficients and test the statistical significance of the relationships between the time series and its lagged values and errors. Once you have fit the model, you can use it to make forecasts for future time points. ARIMA models are widely used for forecasting time series data, and they can be effective in a variety of situations. However, they can be difficult to fit and interpret, especially when dealing with long-term dependencies or complex seasonality in the data.

Prophet: Prophet is a machine learning model for forecasting time series data. It is based on a decomposable time series model with three main components: trend, seasonality, and holidays.

Here's how Prophet works:

Trend: The trend component models the long-term growth or decline of the time series. It uses a non-linear model, such as a logistic curve, to capture changes in the trend over time.

Seasonality: The seasonality component models the periodic fluctuations in the time series data, such as soil moisture. It uses Fourier series to model the periodic patterns and allows the strength and frequency of the seasonality to vary over time.

Holidays: The holiday component allows you to include additional regressors that capture the effects of holidays and other special events on the time series.

To fit the Prophet model, use a dataset with timestamps and the corresponding values. You can then use the Prophet API to specify the model components and fit the model to the data. Once the model is fit, it can be used to make forecasts for future time points by specifying a forecast horizon. One of the main advantages of Prophet is that it is relatively easy to use and can be implemented with just a few lines of code. It also has good performance on a wide range of time series forecasting tasks and can handle missing data and outliers in the data.

VAR Model: Vector autoregression (VAR) is a statistical model used to capture the linear interdependencies among multiple time series variables. It assumes that the current value of each time series can be explained by a linear combination of its past values and the past values of all other time series in the model. Then statistical techniques can be used to estimate the values of the coefficients and test the statistical significance of the relationships between the time series. Once you have fit the model, you can use it to make forecasts or to identify the sources of changes in the time series. VAR models can be extended to include more than one lag of each time series, which allows you to capture longer-term relationships between the time series. You can also include additional time series in the model to capture the relationships between them.

**References**

| **[1]** | N. Zehra, *“Prediction Analysis of Floods Using Machine Learning Algorithms (NARX & SVM)”*, *IJSBAR*, vol. 49, no. 2, pp. 24–34, Jan. 2020 |
| --- | --- |
| **[2]** | F. A. Ruslan, A. M. Samad, Z. M. Zain and R. Adnan, "Flood prediction using NARX neural network and EKF prediction technique: A comparative study," *2013 IEEE 3rd International Conference on System Engineering and Technology*, 2013, pp. 203-208, doi: 10.1109/ICSEngT.2013.6650171 |
| **[3]** | M. Ali, A. M. Qamar and B. Ali, "Data Analysis, Discharge Classifications, and Predictions of Hydrological Parameters for the Management of Rawal Dam in Pakistan," *2013 12th International Conference on Machine Learning and Applications*, 2013, pp. 382-385, doi: 10.1109/ICMLA.2013.78 |
| **[4]** | R. Adnan, A. Samad, F. Ruslan, and Z. Zain, *“Flood prediction modelling using hybrid BPN-EKF and Hybrid Enn-EKF : A Comparative Study”* |
| **[5]** | A. Mosavi, P. Ozturk, and K. Chau, “Flood Prediction Using Machine Learning Models: Literature Review,” *Water*, vol. 10, no. 11, p. 1536, Oct. 2018, doi: 10.3390/w10111536 |
| **[6]** | Flood prediction forecasting using machine Learning Algorithms - IJSER Journal Publication |
| **[7]** | C. Chen, J. Jiang, Z. Liao, Y. Zhou, H. Wang, and Q. Pei, “A short-term flood prediction based on Spatial Deep Learning Network: A case study for Xi County, China,” Journal of Hydrology, vol. 607, p. 127535, 2022 |
| **[8]** | Kunverji, Kruti and Shah, Krupa and Shah, Nasim, A Flood Prediction System Developed Using Various Machine Learning Algorithms. Proceedings of the 4th International Conference on Advances in Science & Technology (ICAST2021) |