CEW 613 Hydrological Modelling
Term Paper

Niharika Vadlamudi
2018122008

Survey : Methods in Flood Prediction & Risk Assessment

1 Introduction

Floods are amongst the most frequent and destructive types of disaster, causing significant damage and disrupting livelihoods throughout the world. There is a wide range of flood risk management methods available that can reduce this destruction, and managing flood risks requires the estimation of flood hazards and the impacts that they can cause. Proper estimation of risk is challenging and requires careful consideration of a number of factors, including watershed properties such as size, topography and land use, the types and characteristics of storms that produce rainfall and flooding in the region, and the number, location, and types of buildings and other assets that could be damaged. Poorly-conducted hazard and risk assessments can lead to poor risk management decisions, from insufficient protection to the wasting of scarce finances on unneeded protection. Well-conducted flood hazard and risk assessments, on the other hand, can provide valuable support for a range of decisions such as land-use master planning, design of infrastructure, and emergency response preparation.Before a hazard assessment is carried out, it is necessary to determine which types of floods are most common or destructive in the area, because in most cases the selection of hazard and risk modeling methods will vary depending on the type of flood.

1.1 Classification of Floods

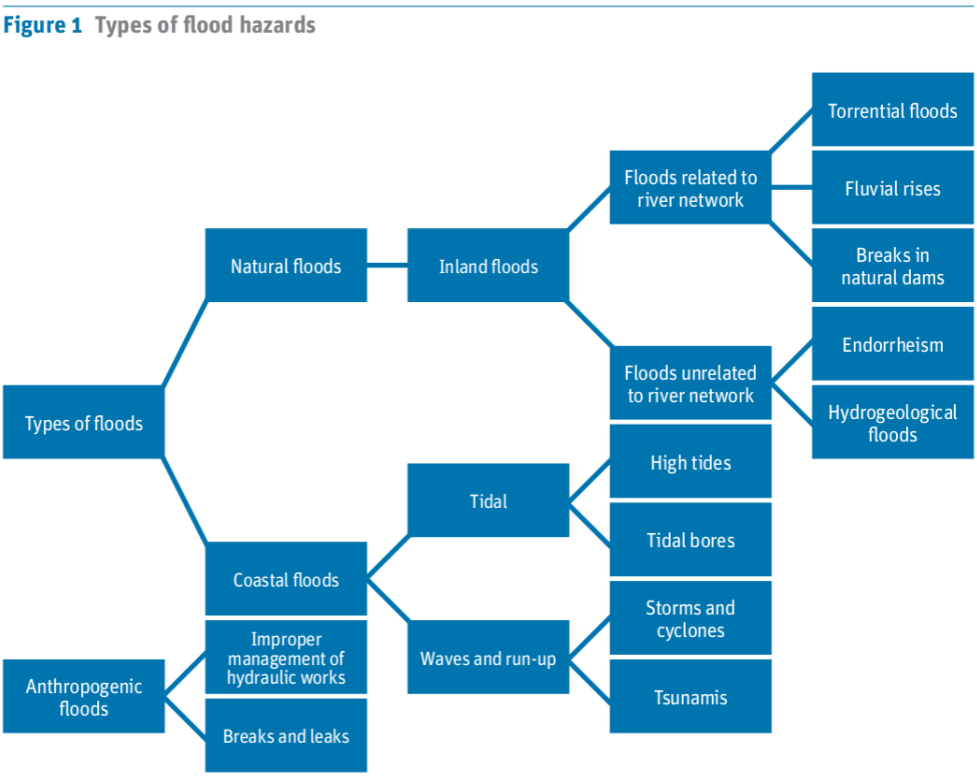


Figure 1: Hierarchy of Floods

2 Necessity for Hazard Assessments

A natural hazard is a potentially damaging physical event, phenomenon or human activity, which may cause the loss of life or injury, property damage, social and economic disruption, or environmental degradation (UNISDR-2004).

The goal of flood hazard assessment is to understand the probability that a flood of a particular intensity will occur over an extended period of time. Hazard assessment aims to estimate this probability over periods of years to decades to support risk management activities. Intensity usually refers to the combination of flood depth and horizontal flood extent; although other intensity measures such as flow velocity and flood duration can also be important depending on the situation.

3 Short Survey : Flood Prediction Techniques

This section focuses on the methods used to estimate design discharge, or the rate of the flow of water through the river or floodplain. The two traditional frequently-used approaches are outlined here are :

- **Traditional Methods**
    1. Discharge-Frequency Analysis
    2. Rainfall-Runoff Modelling
- **Machine Learning Techniques**
    1. Artificial Neural Networks(ANNs)
    2. Multi Layer Perceptrons
    3. Support Vector Machines
    4. Wavelet Neural Network
    5. Decision Trees

Let's understand the advantages of these systems , and the degree of accuracy in such models. An in-depth survey about the technique, their advantages,model complexity , etc .

3.1 Discharge Frequency Analysis

This approach relies on the existence of long records of accurate river discharge measurements. Usually, the highest recorded discharge record from each year is used in the analysis. After identification of annual discharge maxima and other data points,the analyst fits several statistical distributions (for example: log-normal, log-Pearson, or generalized extreme value) and selects the distribution that most accurately describes the data. Important considerations in discharge frequency analysis:

- **Quality of Discharge Records**: Proper measurement of discharge requires maintenance of equipment to provide for continual automatic monitoring of water levels, as well as verification using field measurements of flow and river cross sectional profiles at a range of flow conditions at least several times per year.
- **Length of Observations** : The discharge records must be sufficiently long to estimate the return periods that are required for the flood hazard analysis. There is no single guideline, but records should be at least 10 years in length to perform any sort of frequency analysis.
- **Time Resolution**: The temporal resolution of the discharge record needs to be fine enough to measure the important properties of floods in the river.
- **Change in Topography**: The upstream area cannot have undergone significant changes in terms of land use such as urbanization, agricultural development, or deforestation over the period of the discharge record. If significant changes have taken place, the results of standard statistical analyses will not be valid.

3.2 Rainfall-Runoff Modeling

In many situations, discharge measurements are either nonexistent or of insufficient quantity or quality to be able to conduct a discharge frequency analysis . In such situations, one of a broad class of tools known as rainfall-runoff models (also referred to as hydrologic models) can be used to convert estimates of extreme rainfall into design discharge estimates and design hydrographs.To do so, they must represent the movement of water across the landscape (a process known as runoff) and into the river channel.Many different rainfall-runoff models exist, each with certain advantages and disadvantages depending on a range of factors such as application, geographic setting, and data availability, and knowledge level of the user.

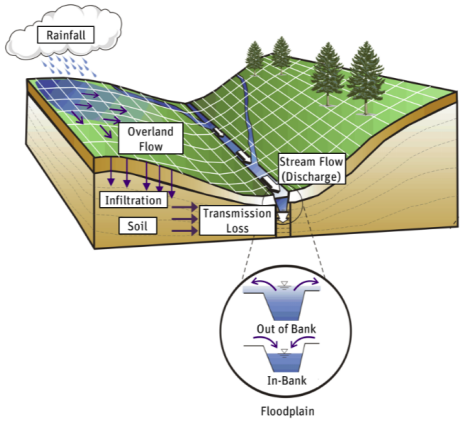


Figure 2: Rainfall-Runoff Diagram

The two main classes of rainfall-runoff models are:

- **Lumped Models**

  The watershed is treated as a single unit and the calculations are performed using simplified, spatially averaged processes.The resulting discharge estimate only applies to the watershed outlet (the most downstream modeled point of the river network). Well-known lumped models include TR-55 and other unit hydrograph-based methods.

- **Distributed Models**

  Distributed models use spatially varying input data for processes such as precipitation, infiltration, interception, interflow, infiltration, and base flow estimating discharge or other variables.

3.3 Machine Learning Techniques

To mimic the complex mathematical expressions of physical processes of floods, during the past two decades, machine learning (ML) methods have highly contributed to the advancement of prediction systems providing better performance and cost effective solutions. The main contribution is to demonstrate the state of the art of ML models in flood prediction and give an insight over the most suitable models.

3.3.1 Artificial Neural Networks

ANN as an efficient mathematical modeling system, through an efficient parallel processing, has the ability to mimic the biological neural network using interconnecting neuron units. Among all ML methods, ANNs, as the most popular learning algorithms,are known to be versatile and efficient in modeling the complex flood processes with a high fault tolerance and accurate approximation.Instead of catchment's physical characteristics, ANNs derive meaning from historical data. Thus, ANNs are considered as the reliable data-driven tools for constructing the black box models to model the complex and non-linear relationship of rainfall and flood as well as river flow and discharge forecasting. Modern ANNs systems include LSTMs and CNNs , Fast Forward Neural Networks (FCNNs) , etc .

3.3.2 Support Vector Machines

Support Vector (SV) as a nonlinear search algorithm using statistical learning theory. SVM classification is said to minimize the over-fitting and reduce the expected error of a learning machine. SVM, with a great popularity in flood modeling, is a supervised learning machine which works based on statistical learning theory and structural risk minimization rule. It has been applied in numerous flood prediction cases with promising results, excellent generalization ability and better performance, comparing to others. Unlike ANN, SVM is more suitable for nonlinear regression problems, to identify the global optimal solution in flood models. Although the high computation cost of using SVM and its unrealistic outputs might be demanding, due to its heuristic and semi-black box nature, the least square-support vector machine (LS-SVM), has highly improved the performance with acceptable computational efficiency.

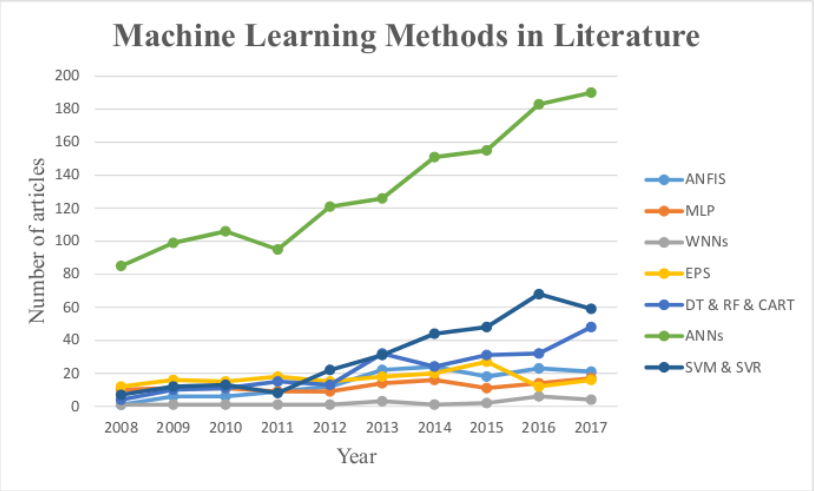


Figure 3. Major ML methods used for flood prediction in literature. Reference year 2008. Source: Scopus

Figure 3: Distributions of ML Techniques

3.3.3 Decision Trees

Decision Trees is one of the contributors in predictive modelling with a wide application in flood simulation. The technique uses a tree of decisions from branches to the target values of leaves. In classification trees (CT), the final variables in a DT contain a discrete set of values where leaves represent class labels and branches represent conjunctions of features labels. When the target variable in a DT has continuous values and an ensemble of trees are involved it would be regression trees (RT). As DTs are classified as fast algorithms, they have become very popular in ensemble forms to model and predict floods . Classification and regression tree (CART) which is a popular type of DT used in ML, have been applied successfully to flood modeling yet its applicability to flood prediction has not been fully investigated. The random forests (RF) is another popular DT method for flood prediction,it includes a number of tree predictors.

3.3.4 Wavelet Neural Networks

Wavelets transform (WT), as a mathematical tool, can be used to extract information from various data sources through analyzing local variation in time series . Wavelet transforms support reliable decomposition's of an original time series to improve data quality. The accuracy of prediction is improved through Discrete WT (DWT), which decomposes the original data into bands leading to improvement of flood prediction lead-times. DWTs due to their beneficial characteristics have been widely used in flood time series prediction.Furthermore, hybrid models of DWTs e.g. wavelet-based neural network, WNN, which combines the WT and FFNN, and wavelet-based regression model , which integrates WT and multiple linear regression (LR), have been used in time series prediction of flood . In fact, most recently, WNNs, due to their potential in enhancing the time series data, has gained popularity in flood modeling in applications like, daily flow , rainfall-runoff water level and flash flood.

4 Experiments

4.1 Aim

The main aim is to use exciting Machine Learning Algorithm's and modelling it for Flood prediction by regressing the river discharge volume.Using the river discharge values , we can have a dynamic threshold to determine the probability of flood occurrence on monthly/daily basis.

4.2 Dataset Analysis

The dataset used for this problem statement is ERA-5 Dataset , North America . This database consists of several parameters pertaining to climate and weather . The final features were decided based on several pre-processing techniques. The dataset comprises of 12941 samples , spanning from 1981-2017. The metrics and ranges of these parameters are widely varying. The dataset consists information regarding the soil profile, total precipitation , convectional precipitation , layer of snow , etc .

Parameter Name	Short Name	paramId	levels
Temperature	t	130	850, 700, 500
Specific humidity	q	133	850, 700, 500
Geopotential	z	129	850, 700, 500
Large-scale precipitation	lsp	142	-
Convective precipitation	cp	143	-
Volumetric soil water layer 1	swvl1	39	-
Volumetric soil water layer 2	swvl2	40	-
Runoff	ro	205	-
Total column water vapour	tcwv	137	-

Figure 4: Data Table

Let's look at few pre-processing and data-cleaning techniques:

Correlation Analysis Plot

Correlation analysis is a statistical method used to evaluate the strength of relationship between two quantitative variables. A high correlation means that two or more variables have a strong relationship with each other, while a weak correlation means that the variables are hardly related. This technique is strictly connected to the linear regression analysis that is a statistical approach for modeling the association between a dependent variable, called response, and one or more explanatory or independent variables.Using this technique,I eliminated 4 variables namely **lsp-4-11,lsp-55-128,sp-diff**,etc. The following diagram , is the proof :

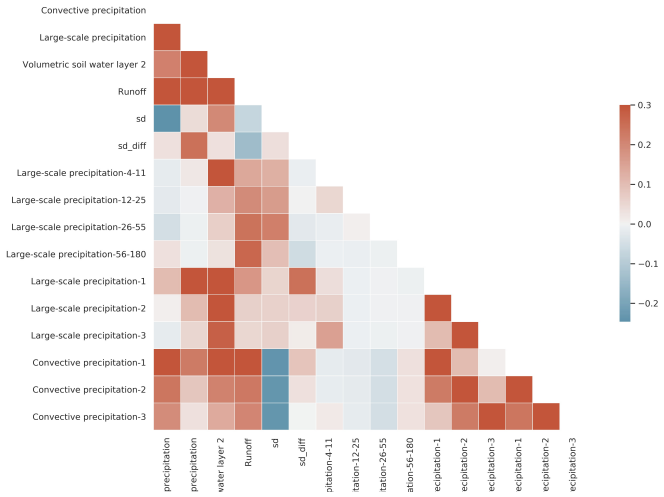


Figure 5: Correlation Plot

Box Plot Analysis

A box plot is a standardized way of displaying the distribution of data based on a five number summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum"). It picks out the outliers and gives us a notion of symmetry of data is symmetrical, how tightly your data is grouped, and degree of skewness.

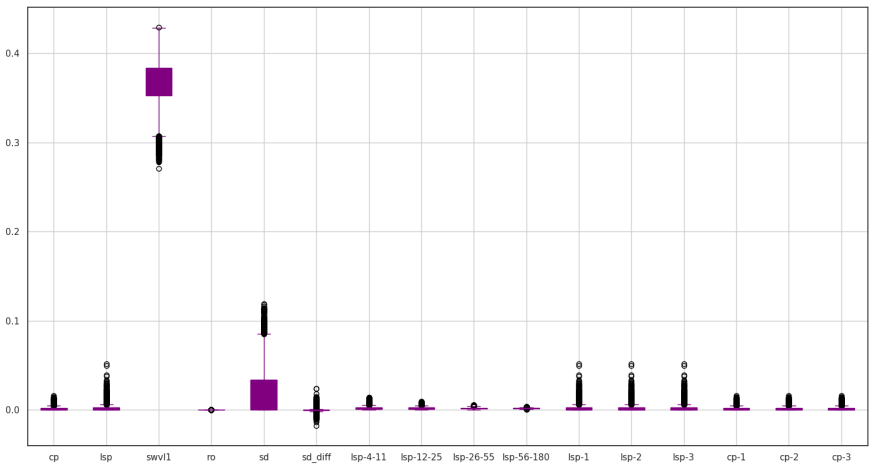
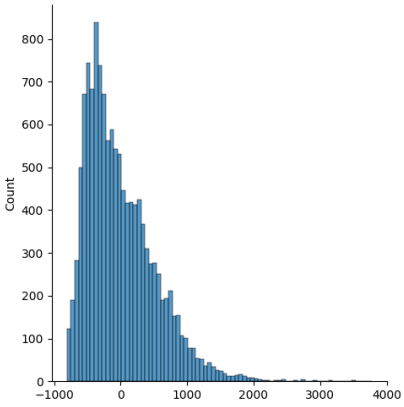


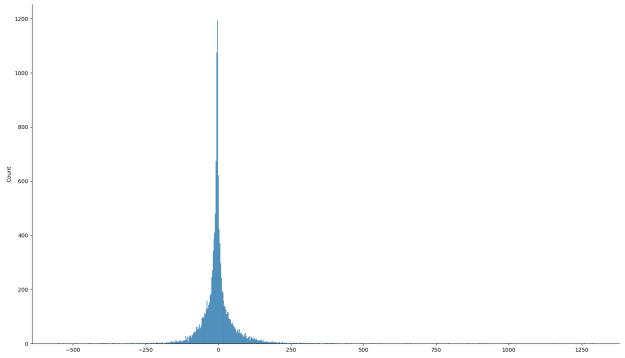
Figure 6: Coloured Box Plot

Gaussian Curve Fitting

Usually data is said to fit in various probability density functions .Here, we can observe that the discharge data is in the form of a Gaussian Curve , and the same trend is followed by the difference of discharge(across time).



(a) River Discharge - Skewed Gaussian Fitting

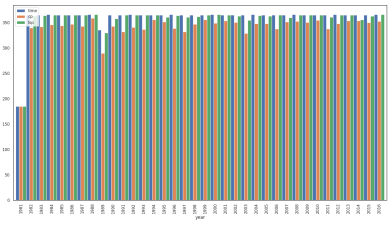


(b) Difference in River Discharge

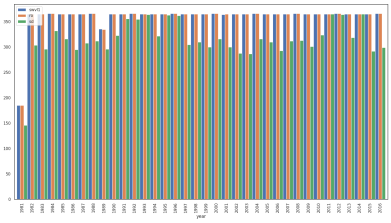
Figure 7: Gaussian Fitting

Bar Graph Analysis

A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally.A bar graph shows comparisons among discrete categories. One axis of the chart shows the specific categories being compared, and the other axis represents a measured value.This helps us in understanding , how our variables are varying across time.

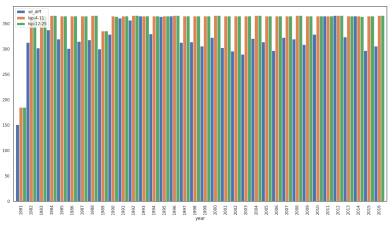


(a) Time vs Convectional Precipitation & Total Large Scale Precipitation

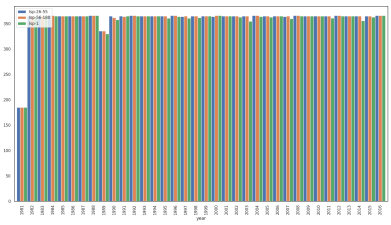


(b) Time vs Soil Layer Water & Runoff

Figure 8: Bar Plots of Parameters



(a) Time vs LP-4-11 vs LP-5-12



(b) Time vs LP-56-11

Figure 9: Bar Plots of Parameters

Quantile based Filtering

In this technique, we will do the flooring (e.g., the 10th percentile) for the lower values and capping (e.g., the 99th percentile) for the higher values.In our case , the values of high and low are 0.99 and 0.01 respectively.

4.3 Model Descriptions

For, assessing the performance of Machine Learning algorithms on this dataset, I've chosen 5 models , as follows :

- **Support Vector Machine (Regressors)** Here, I've experimented the traditional SVM with different kernels such as Radial Basis Function (RBF) , Linear and Polynomial Kernel.
- **Neural Network** A 3 layered convulational network was designed with (15-10-3-1) layers respectively .With increasing the depth in layers, the accuracy improved. All other parameters were set to default.
- **Decision Tree with Ada Boost Regressor** The regular Decision Tree , showed high loss despite changing the depth and the complexity of the algorithm.
- **Gradient Boost** Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiates loss function.

4.4 Quantitative Results-Comparative Analysis

Among the above ML models , we need to analyse how well the predictions align with the actual ground truth . For this experiment,I've taken a random sample from the **Test File** , across various years and observed the following graphs.

4.4.1 Comparison of ML Models

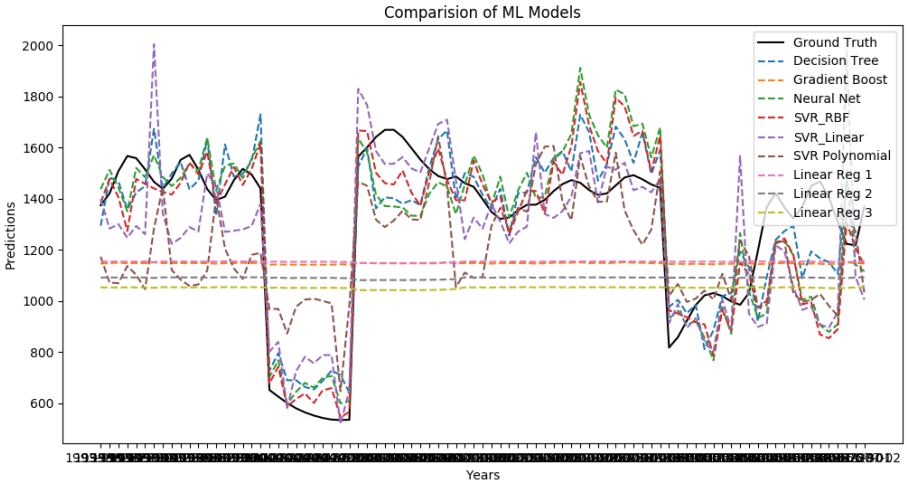


Figure 10: Comparison of all ML Models

We can observe that Linear Regression Models & Gradient Boost Algorithms perform poorly compared to other models .One way to reason is this is from the complexity of the data ,and that linear line fitting is not optimal in nature.

4.4.2 Comparison of SVR Models

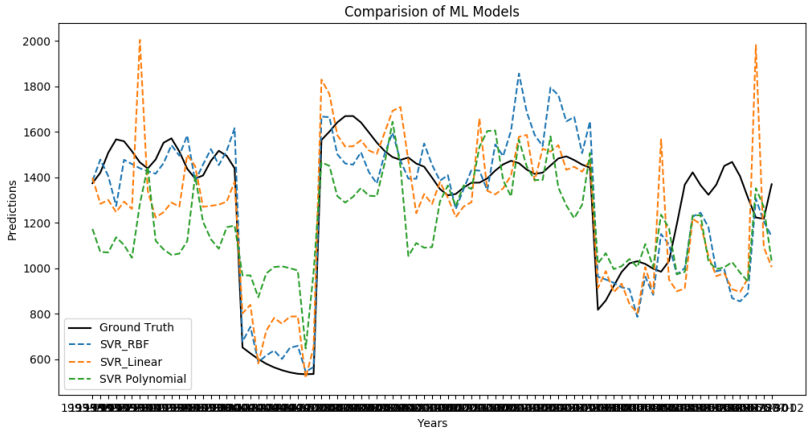


Figure 11: Comparison of all ML Models

Here's a comparison graph with the ground truth data and corresponding Support Vector Regressor models with various kernels - Radial Basis Function , Linear & Polynomial. We can see RBF kernel has outperformed the other 2 models due to advanced technique of univariate approximation.

4.4.3 Comparison of Other Models

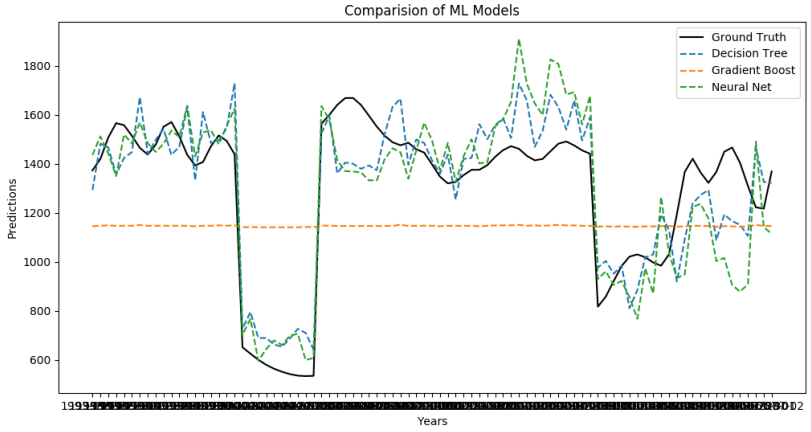


Figure 12: Other Models

A closer look at the remaining models - Neural Net , Gradient Boost & Decision Trees . Here , despite the Decision Tree depicting noisy observations we can see that it outperforms its group.The Neural Network which is Multi-layer Perceptron Model (MLP) , follows closely with the decision tree .

4.4.4 Linear Regression Models

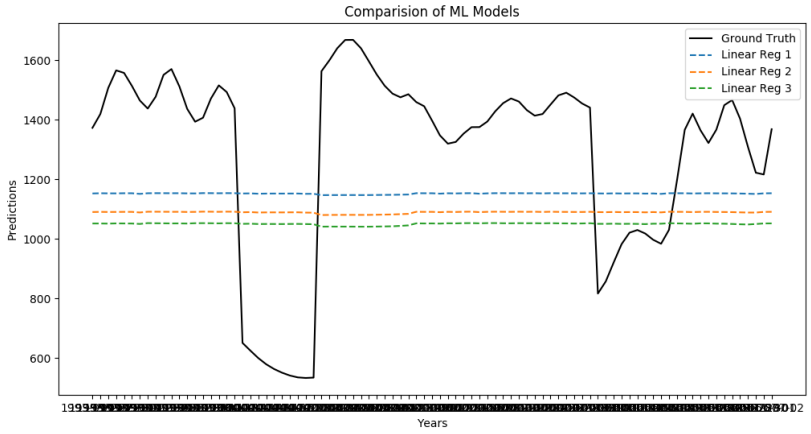


Figure 13: Linear Regression - ML Models

One of the plausible reasons for the failure of Linear Regression (Polynomial) to learn the dataset would be , that the data is to complex to fit curves on 2-dimension case.

5 Tabular Results

The following table was generated from Test File data , the loss taken here is Mean squared loss (L2-Loss). Note : Random Forest was also added here in the analysis .

Machine Learning Model	MSE Loss
SVR-Radial Basis Function Kernel	21.25
SVR - Linear Kernel	62.34
SVR - Polynomial	106.25
Linear Regression (Degree 1)	158.19
Linear Regression (Degree 2)	131.23
Linear Regression (Degree 3)	127.93
Decision Tree	37.087
Random Forest	48.68
Neural Network	32.81
Gradient Boost	188.24

Figure 14: L2 Loss on Test File

6 Tool : Quick Demo

Here, the picture of the GUI interface made for this tool . It gives you the final discharge in the terrain . Depending on the demography , and climatic conditions we can set thresholds for occurrence of floods . Each of the model has one dedicated page , here the user can give input via .csv file or manual input .

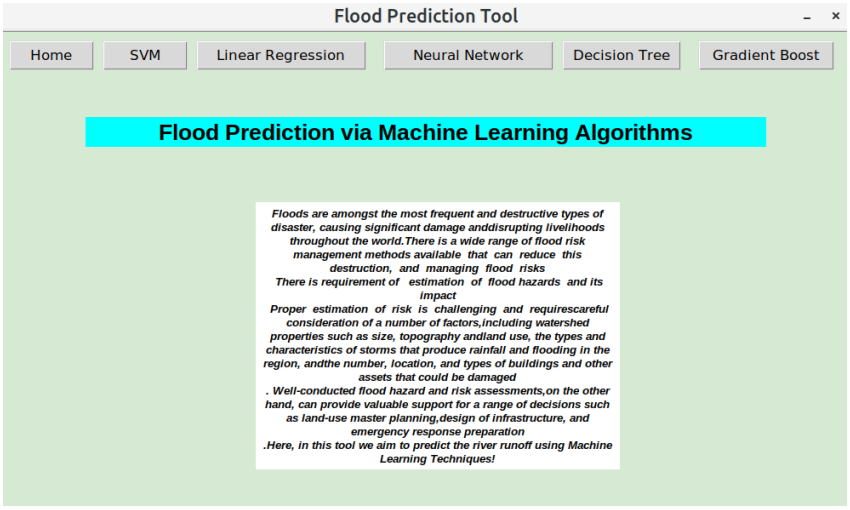


Figure 15: Home Page

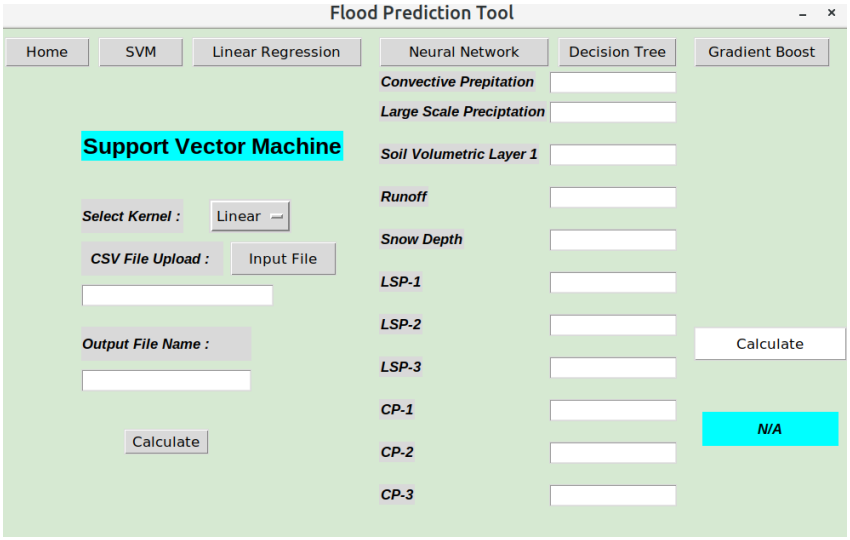


Figure 16: Support Vector Regressor Page

7 Flood Risk Prevention

Once a flood hazard and vulnerability assessment has been completed, it is relatively simple to arrive at an estimate of flood risk. It is clear that completely "correct" risk calculations can only be accomplished when the probability and the magnitude of the loss can be estimated with complete accuracy. This is impossible in practice due to the limited amount and accuracy of information. There are many situations in which flood hazard, vulnerability, or both are very difficult to quantify due to a lack of sufficient information. In such cases, one can either attempt to collect additional information to estimate hazard and vulnerability, or attempt to make risk management decisions without this information. Making flood risk management decisions with no or poor risk information can result either in too little or too much protection. Too little protection means that citizens or economic assets face continued exposure to flood impacts, while too much protection means that money has been unnecessarily spent on unneeded protection.

8 Uncertainty in Flood Assessments

A major challenge in flood hazard and risk assessment is to understand the uncertainties that exist at every stage of the process, and to decide how to incorporate these uncertainties into subsequent risk management decisions. For example, the estimation of design discharges, whether done using statistical methods or rainfall- runoff models, always depends on the use of multiple assumptions, incomplete

datasets, and imperfect models. This will lead to errors in design discharges, which will in turn lead to errors in the water levels estimated using a hydraulic model, which will combine with imperfections in the hydraulic model to affect the predicted spatial extent of flooding. Likewise, there are considerable uncertainties in vulnerability assessment. The inability to characterize building-level flood impacts will then translate into errors in estimated damage and economic losses. Flood hazard and risk experts put considerable efforts into understanding the various uncertainties and how they can affect risk estimates. These variations can translate into significant differences in estimated economic and other types of flood damage and loss. These differences imply significant uncertainty in the decision-making required to reduce these damages. Uncertainties should be carefully considered when evaluating risk management investment , provides a more detailed explanation of the various sources of uncertainty. We need to incorporate different uncertainties in flood hazards and risk assessment techniques .

9 Conclusion

After completing the term paper , I understood how advanced Machine Learning algorithms are applied to climate analysis.I understood various pre-processing techniques and important parameters relating to flood prediction modelling.I got a flavour of how traditional flood predictions function , their advantages & disadvantages.I realised the use-case of high-functioning flood prediction and warning systems .

10 References

- https://onlinelibrary.wiley.com/doi/10.1111/jfr3.12563
- https://www.researchgate.net/publication/330252952_A_simplified_approach_for_flood_vulnerability_assessment_of_historic_sites
- https://www.preventionweb.net/files/51114_capramethodsinfloodhazardandriskass.pdf
- https://s3.us-east-2.amazonaws.com/www.colmayor.edu.co/wp-content/uploads/2019/08/diseno-metodologico.pdf
- Prediction of Flood by Rainfall using MLP Classifier of Neural Network Model
- (GitHub) MATEHIW: MAchine learning TEchniques for High-Impact Weather
- (GitHub) https://github.com/NiharikaMessi/project-14012020