Posted on 24 Nov 2022 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.1002/essoar.10509264.1 — This a preprint and has not been peer reviewed. Data may be preliminary.

Regional flood frequency analysis using extreme gradient boosting based on Bayesian optimization.

Deva Jarajapu[1] and Rathinasamy Maheswaran[1]

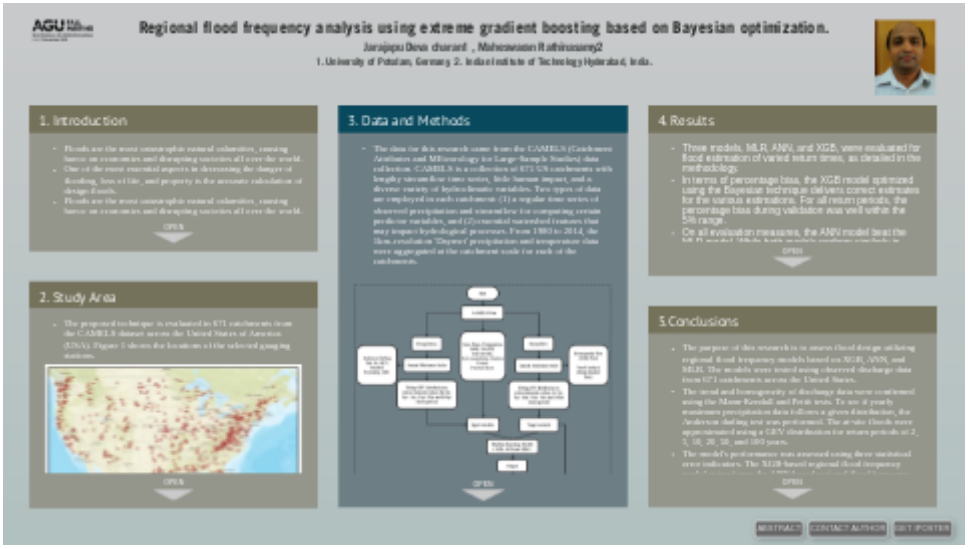
[1]MVGR College of Engineering

November 24, 2022

Abstract

Estimation of design flood is a crucial task in water resources engineering. Regional Flood Frequency Analysis is one of the widely used approaches for estimating design flood in ungauged basin. In the present research, we develop an eXtreme Gradient Boost based ML model for RFFA. The proposed approach relies on developing a regression model between flood quantiles and the commonly available catchment descriptors. In this study, the CAMELs data for 671 catchments from USA was used to study the efficiency of the approach. Further, the results were compared with the traditional methods such Multiple Linear Regression (MLR) and Artificial Neural Networks (ANN). XGB is a decision-tree-based ensemble machine learning algorithm that uses gradient boosting as a framework. The results revealed that the XGB based approach resulted in estimates with highest accuracy when using all the available catchment descriptors (i.e., mean annual rainfall(MAR), drainage area, fraction forest, mean annual potential evapotranspiration (MAPET), mean annual temperature, rainfall intensity, slope, fraction snow, soil porosity, and soil conductivity) both during training and validation. Four distinct models consisting of three to ten descriptors were examined for 2-, 5-, 10-, 25-, 50-, and 100-year return periods, all of the models exhibit smaller mean absolute error values and root mean square error values with percentage bias ranging from -10 to +10. A model with three predictor variables has comparable performance to other models. Drainage area, rainfall intensity, MAR, and fraction snow are the most efficient predictor variables, while MAPET, Slope, Temperature, Fraction Forest, Soil Porosity, and Soil Conductivity have low significance in predicting design flood for an ungauged catchment. The XGB modeling approach that has been proposed can be applied to different places throughout the world.

Regional flood frequency analysis using extreme gradient boosting based on Bayesian optimization.



Jarajapu Deva charan1 , Maheswaran Rathinasamy2

1. University of Potsdam, Germany, 2. Indian Institute of Technology Hyderabad, India.

PRESENTED AT:

1. INTRODUCTION

- Floods are the most catastrophic natural calamities, causing havoc on economies and disrupting societies all over the world.
- One of the most essential aspects in decreasing the danger of flooding, loss of life, and property is the accurate calculation of design floods.
- Floods are the most catastrophic natural calamities, causing havoc on economies and disrupting societies all over the world. One of the most essential aspects in decreasing the danger of flooding, loss of life, and property is the accurate calculation of design floods.
- Regional Flood Frequency Analysis (RFFA) has been a blessing for such regions with little to no flow data available.

2. STUDY AREA

- The proposed technique is evaluated in 671 catchments from the CAMELS dataset across the United States of America (USA). Figure 1 shows the locations of the selected gauging stations.

Figure.1 Locations of chosen gauging stations in the United States considered for model development in this research.

3. DATA AND METHODS

- The data for this research came from the CAMELS (Catchment Attributes and MEteorology for Large-Sample Studies) data collection. CAMELS is a collection of 671 US catchments with lengthy streamflow time series, little human impact, and a diverse variety of hydroclimatic variables. Two types of data are employed in each catchment: (1) a regular time series of observed precipitation and streamflow for computing certain predictor variables, and (2) essential watershed features that may impact hydrological processes. From 1980 to 2014, the 1km–resolution 'Daymet' precipitation and temperature data were aggregated at the catchment scale for each of the catchments.

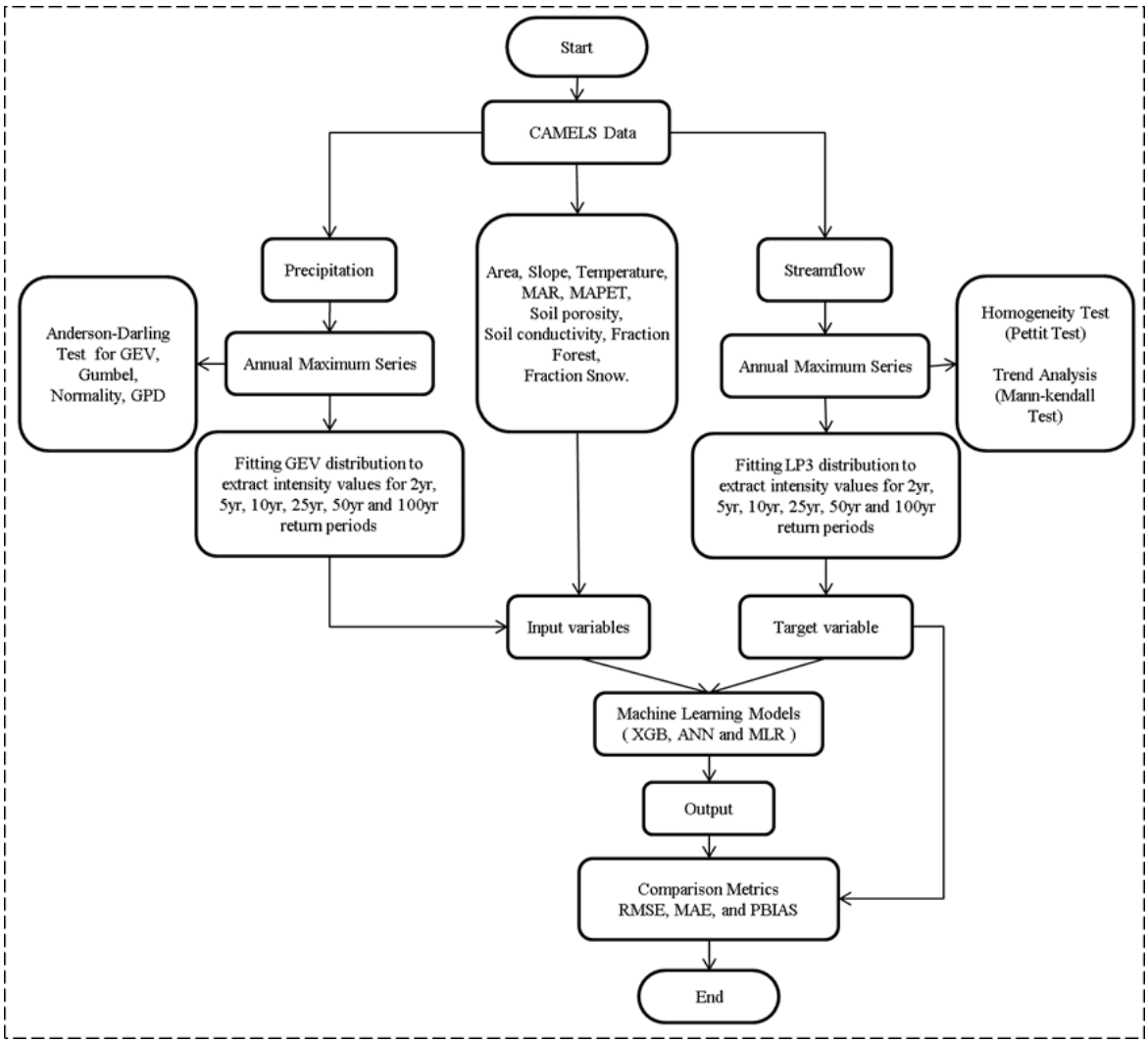


Figure 2. Flowchart of the proposed methodology

4. RESULTS

- Three models, MLR, ANN, and XGB, were evaluated for flood estimation of varied return times, as detailed in the methodology.
- In terms of percentage bias, the XGB model optimized using the Bayesian technique delivers correct estimates for the various estimations. For all return periods, the percentage bias during validation was well within the 5% range.
- On all evaluation measures, the ANN model beat the MLR-model. While both models perform similarly in terms of percentage bias in all return periods, the data demonstrate no difference in terms of performance. It's worth noting that none of the models is appropriate for all six return times when compared to XGB variants.

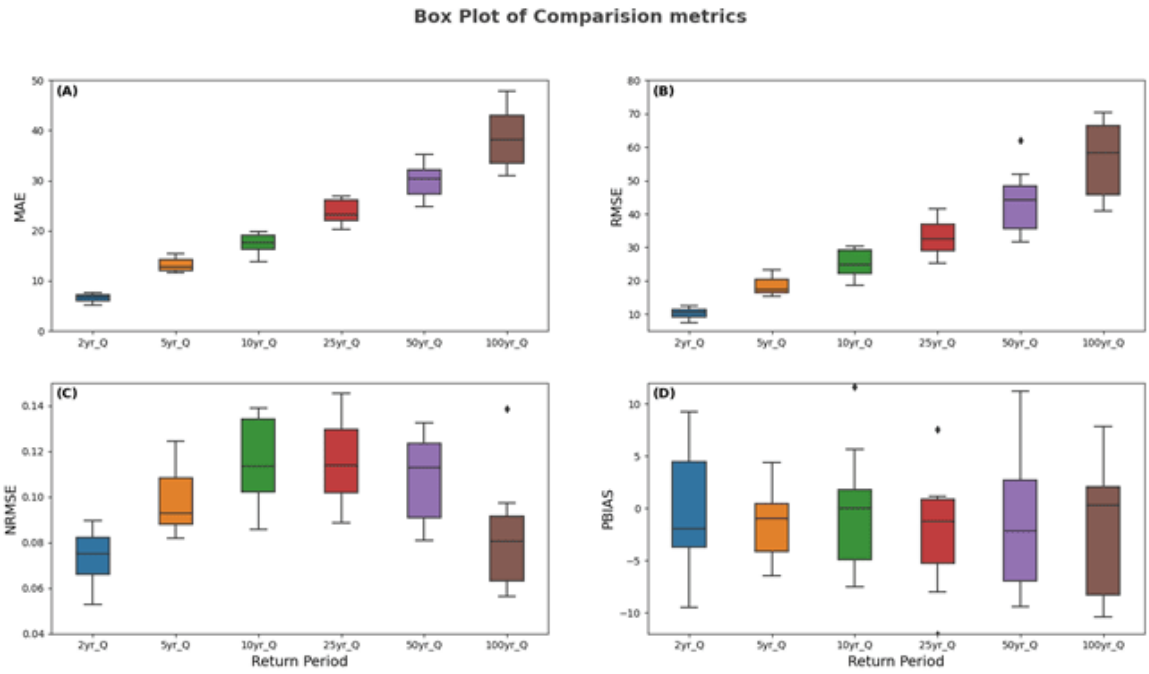
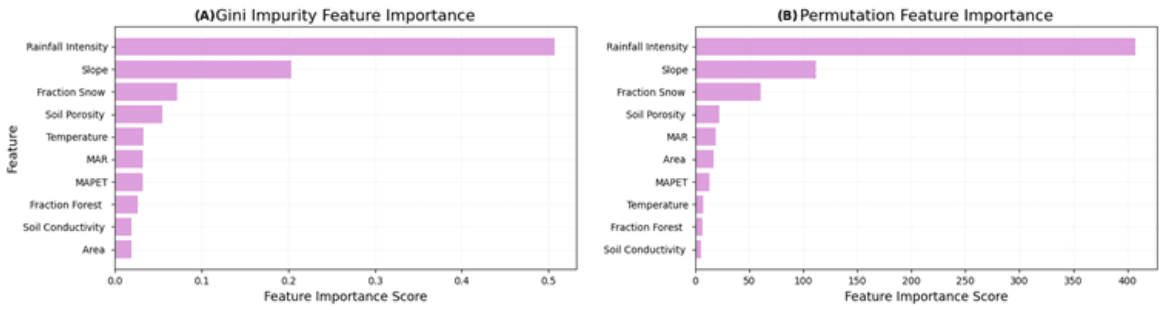
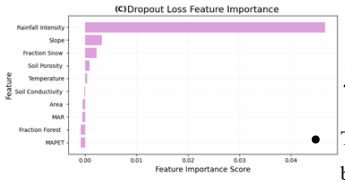


Figure 7. Displays the results of comparison metrics derived between observed and simulated streamflow. (A) and (B) depicts the mean absolute error (MAE) and root mean squared error (RMSE) values of different return periods. (C) and (D) shows the normalized root mean squared error (NRMSE) values and percentage bias for each return period.

- We employed three approaches to explore the effect of each feature used as predictor variables on target variable-design flow: Gini Impurity, Permutation, and Dropout. Methods of feature rating are used. The feature significance plots from the three techniques are shown in Figure 9. Rainfall intensity, slope, fraction snow, soil porosity, and temperature are the watershed variables that have a major influence on design flood, according to all three methodologies, whereas other catchment factors have a little impact. Rainfall Intensity, as predicted, is the most influential meteorological feature, scoring highest among the other attributes. In all three methodologies, soil conductivity, area, and fraction forest receive poor ratings.





5.CONCLUSIONS

- The purpose of this research is to assess flood design utilizing regional flood frequency models based on XGB, ANN, and MLR. The models were tested using observed discharge data from 671 catchments across the United States.
- The trend and homogeneity of discharge data were confirmed using the Mann-Kendall and Pettit tests. To see if yearly maximum precipitation data follows a given distribution, the Anderson darling test was performed. The at-site floods were approximated using a GEV distribution for return periods of 2, 5, 10, 20, 50, and 100 years.
- The model's performance was assessed using three statistical error indicators. The XGB-based regional flood frequency model outperforms the ANN-based regional flood frequency model and the MLR-based regional flood frequency model, according to the research.
- The Gini Impurity, Permutation, and Dropout Loss Feature Ranking techniques were used to highlight the feature relevance. In an ungauged watershed, the predictor variables MAPET, MAR, Drainage area, Temperature, Fraction Forest, and Soil Conductivity have little relevance in forecasting design flood. Rainfall Intensity, Slope, Soil Porosity, and Fraction Snow are the most important predictive factors. Because the bulk of the predictor variables are readily available for most of the catchments, the findings are extremely valuable. With PBIAS ranging from -10 to +10 for all design floods, the proposed xgboost technique holds a lot of potential for estimating severe flood magnitudes in the United States.

ABSTRACT

Estimation of design flood is a crucial task in water resources engineering. Regional Flood Frequency Analysis is one of the widely used approaches for estimating design flood in ungauged basin. In the present research, we develop an eXtreme Gradient Boost based ML model for RFFA. The proposed approach relies on developing a regression model between flood quantiles and the commonly available catchment descriptors. In this study, the CAMELs data for 671 catchments from USA was used to study the efficiency of the approach. Further, the results were compared with the traditional methods such Multiple Linear Regression (MLR) and Artificial Neural Networks (ANN). XGB is a decision-tree-based ensemble machine learning algorithm that uses gradient boosting as a framework. The results revealed that the XGB based approach resulted in estimates with highest accuracy when using all the available catchment descriptors (i.e., mean annual rainfall(MAR), drainage area, fraction forest, mean annual potential evapotranspiration (MAPET), mean annual temperature, rainfall intensity, slope, fraction snow, soil porosity, and soil conductivity) both during training and validation. Four distinct models consisting of three to ten descriptors were examined for 2-, 5-, 10-, 25-, 50-, and 100-year return periods, all of the models exhibit smaller mean absolute error values and root mean square error values with percentage bias ranging from -10 to +10. A model with three predictor variables has comparable performance to other models. Drainage area, rainfall intensity, MAR, and fraction snow are the most efficient predictor variables, while MAPET, Slope, Temperature, Fraction Forest, Soil Porosity, and Soil Conductivity have low significance in predicting design flood for an ungauged catchment. The XGB modeling approach that has been proposed can be applied to different places throughout the world.