UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

HYDROLOGIC PEAK FLOW MODELLING USING MACHINE LEARNING

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE

By

AKHIL SANJAY POTDAR
Norman, Oklahoma
2020

HYDROLOGIC PEAK FLOW MODELLING USING
MACHINE LEARNING

A THESIS APPROVED FOR THE
GALLOGLY COLLEGE OF
ENGINEERING

BY THE COMMITTEE CONSISTING OF

Dr. Charles D. Nicholson,Chair

Dr. Pierre E. Kirstetter

Dr. Randa L. Shehab

Acknowledgements

Firstly, I would like to thank Dr. Pierre Kirstetter for his constant support and guidance throughout the past one and half year of my Masters' studies. Thank you for sharing your vision and giving me the opportunity to research into a domain that suited both our interests. Your support and the freedom to follow my own instincts made this journey a wonderful one. You have stood by me during my times of need and for that I am truly thankful. I couldn't have asked for a mentor more worthy than you.

I would also like to thank Dr. Charles Nicholson for his flexibility and his unconditional support as the committee head. Likewise, I thank Dr. Randa Shehab for her omnipresent help through this semester. Thank you both for agreeing to be in the committee in such short notice during the COVID-19 outbreak.

Also, I would love to thank Srushti for being by my side and appreciating all my efforts in my research work. Your unreserved support made my work all the more fulfilling. Thank you Shruti for your guidance through my journey in hydro metrology. And finally, I thank you mom, dad and Akshay, for all the efforts and sacrifices you have made to make this journey possible. Thank you!

This work was supported by the NOAA GOES-R Risk Reduction Science Program award NA16OAR4320115, and the NOAA Joint Technology Transfer Initiative award NA17OAR4590170.

Table of Contents

List of Tables

List of Figures

Abstract

The effect of rainfall spatial variability on catchment responses during floods remains poorly understood. The overall objective of this work is to develop a robust understanding of how rainfall spatial variability influences flood peak discharge, with a focus on its contribution relative to basin physiography. A machine learning approach is used on a high-resolution rainfall and flooding event dataset spanning 10 years and gathering rainfall events and basins of widely varying characteristics across the U.S. This approach overcomes a major limitation of prior studies based on limited observations or model simulations. This study explores the first-order dependencies in the relationships between peak discharge, rainfall variability, and basin physiography, and it disaggregates these complex interactions using a multi-dimensional statistical modeling approach. After selecting amongst the different regression methods (Lasso, Elastic Net, Multilinear Regression, Random Forest and Xgboost) we use Xgboost to generate regression models to predict peak discharge and perform predictor importance analysis. A parsimonious model is finally created that has low bias and variance and which can be deployed in the future for flash flood forecasting. The results confirm that the spatial organization of rainfall within a basin has a significant influence on the basin response, but the basin physiography is shown to be the primary driver of peak discharge. These findings have unprecedented representativeness in terms of flood characterization. An improved understanding of sub-basin scale rainfall spatial variability will aid in developing a robust flash flood characterization as well as identifying basins which could most benefit from distributed hydrologic modeling.

Key words: Hydrology, machine learning, flood prediction, hydrologic models, precipitation moments, basin physiography.

Chapter 1: Introduction & Literature Review

Flash Flood in the US

Floods hazards are ranked the third most frequent type of natural disaster behind severe storms and tropical cyclones. They have contributed to an estimated loss of $146.5 billion to the US economy in the last forty years and that number is steadily increasing (Smith, 2020). More flooding is expected along with more intense precipitation events globally under climate change (Sillmann et al., 2013). Fatalities under flash floods circumstances represent the major contribution of flood fatalities (Ashley & Ashley, 2008). Flash floods are rapid rises of water along a an existing waterway, that begins within 6 hours, and often within 3 hours, of the causative rainfall (NOAA, 2005). The ability to characterize and predict flash floods is increasingly important (Gourley et al. 2017).

The effect of rainfall on the discharge is shown through a *hydrograph* which shows precipitation rate and discharge as a function of time on the same graph. In figure 1, peak discharge is the maximum amount of water in a river after a rainfall event. If the peak discharge is more than the bank discharge capacity, then a flood will occur. As discharge is dependent on many factors of the basin response, geomorphological characteristics and the precipitation spatial distribution, we try to characterize the peak flow of a hydrograph.

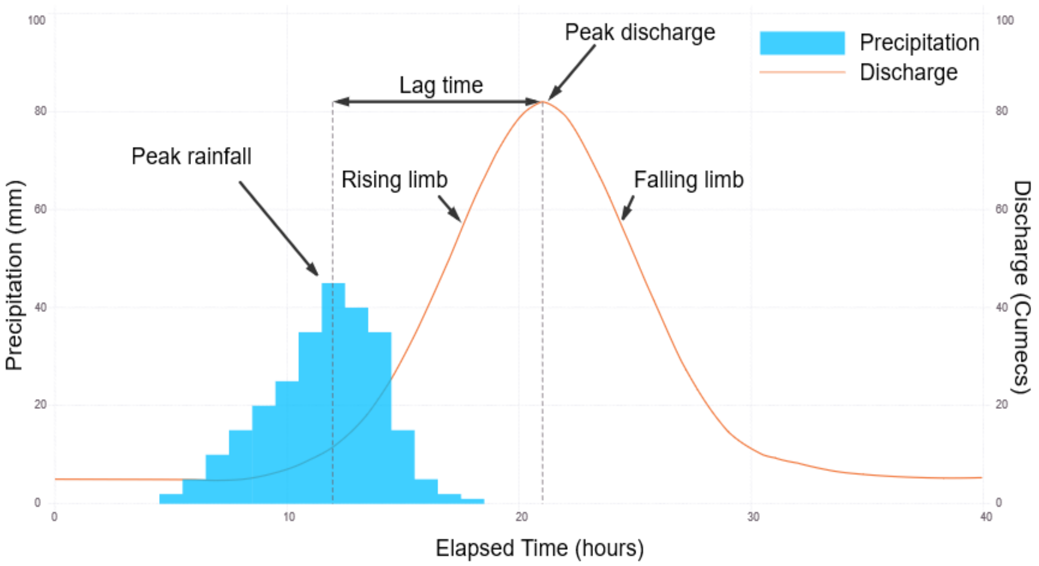


Figure 1. Flash Hydrograph (Jackson, 2014)

Flash flood monitoring systems include the Flash Flood Guidance (FFG) that is used worldwide and issues warnings based on the runoff generation (Sweeney & Baumgardner, 1999). However, FFG only represents parts of the floods characteristics as it is not focused on the propagation of water overland and along streams. Hence it misses any occurrence of flood downstream of the rainfall, especially delay, magnitude, and duration of the flood. A flood forecasting system needs to describe these characteristics to help predict events ahead of time, such that destruction of property and life can be mitigated by efficient warning systems.

Modelling in Hydrology

During a flash flood the discharge in the outlet increases suddenly under the integrated influences of specific hydrological processes which show variable effects under different basin geomorphology, climatology and spatiotemporal conditions (Saharia et al., 2017). Hydrological models are used to interpret and anticipate floods characteristics through simplified representations of the processes that take place in the watershed. Models can be classified in three categories based on how hydrological processes are described: empirical, conceptual, and physical models (Solomatine & Wagener, 2011).

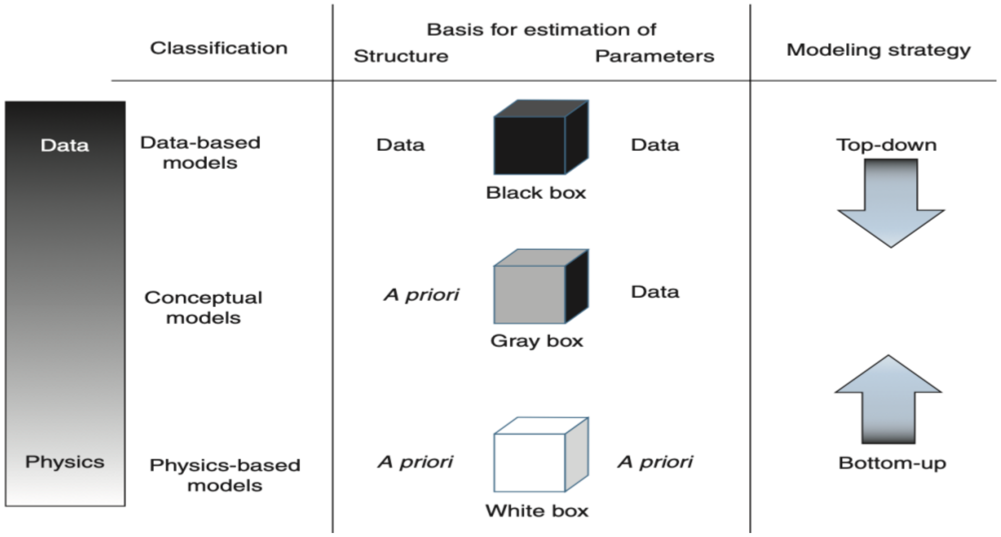


Figure 2. Classification of hydrological models (Solomatine & Wagener, 2011)

Traditional empirical (statistical) models are built from the joint analysis of precipitation (input) and discharge (response) time series data to derive statistical equations based on regression and correlation that represents the input-response behavior of a catchment. The unit hydrograph approach is an example of such empirical model. The data-based models do not consider catchment features (e.g. geomorphology) and hydrological processes, hence while they have high predictive power at a given location (basin outlet) they also have low explanatory efficiency and cannot be applied to a different basin (Devia et al., 2015).

Physically based models are mechanistic and designed to represent the physical processes of the system. The rationale expects a degree of physical realism to the extent that the laws of conservation of mass, momentum and energy are maintained. Such models use variables that are functions of space and time. The model's structure and parameters are designed *a priori* based on the understanding of the basin physics. As such the selected parameters are not calibrated, making diagnosis difficult. Such models initially rely on abundant geomorphological data on the catchment, in addition to hydrological and meteorological observations. However, this type of model also overcomes the limitation of versatility (for other basins) and interpretability that empirical models experience.

Conceptual models are parametric models with a structure that is decided *a priori*, while the parameters are calibrated using the observations of the catchment. A number of hydrologic processes are synthesized into single parameters, such that they are hard to interpret from the basin data-stream. As such they are imperfect representations of the physical processes. An example of a conceptual distributed model is the Ensemble Framework for Flash Flood Forecasting (EF5). It is the state of the art solution developed at the University of Oklahoma and the NOAA National Severe Storms Laboratory for flash flood prediction at the U.S. National Weather Service (Flamig et al., n.d.). It uses various conceptual models to simulate streamflow and soil saturation forecasts. With input of precipitation, temperature, evo-transpiration, discharge, the model parameters require large hydrological and meteorological data. These models identify processes which are important for flash floods.

The approach used here is a “physical- statistical” modeling approach that improves on the drawbacks of the above classification. Datasets are gathered that represent the geomorphology, the climatology and the spatiotemporal attributes of the precipitation forcing. While this approach primarily builds on observations like the empirically based modelling class, the physical depiction of the basin behavior is enriched through the integration of geomorphologic and climatological characteristics. As such, the contribution of hydrologic processes, that are driven by the basin features, are accounted for to represent their integrated response at the basin outlet. The curated dataset allows regression modelling using gradient boosted trees to identify the multivariate relationships that exist between the dependent (peak discharge) and independent input variables representing the geomorphology, the climatology, as well as the spatiotemporal attributes of precipitation. While empirical models are observed to overfit events in the training dataset (Devia et al., 2015), the addition of physical constraints along with the use of proper methodologies to mitigate overfitting, ease the generalization of the model for diverse catchments.

Unlike the conceptual and physically based modelling categories, this approach undertakes no prior assumptions (e.g. uniform depiction of basins) in the design of the model structure or the calibration of parameters. While biases arise from *a priori* structure design and parameter choices that impact the applicability over ungauged basins, this approach inherently calibrates the parameters to the data. Once calibrated by data training, it requires no further tuning. Unlike traditional empirical approaches, such a model can be used as a diagnostic tool to identify and interpret key hydrological processes. Through the study of feature importance, simpler and parsimonious models can be designed to represent the basin physics as mechanistic models do. By incorporating the central features of all the model types and eliminating its shortcomings, our modelling approach uniquely utilizes the best of all the approaches.

Solomatine & Wagener (2011) emphasized the advent of new data driven models through the integration of machine learning. The present approach is novel owing to a few reasons. First, it utilizes a diverse data set that comprehensively incorporates the physics of the hydrologic system. Second, it improves model validation by using a combination of evaluation metrics such as the “Mean Relative Error”, “Co-efficient of determination (R^2)” and the “Root Mean Squared Error”. Such implementation makes sure that the model explains the variance along with the systemic error. And finally, understanding of key parameters without an *a priori* basis gives us new insights on the science of basins response.

This new category of model can address important challenge in hydrologic sciences, i.e. characterizing of floods in ungauged basins. This novel approach seeks to provide a high predictive power interpretation with versatility over diverse basins.

Chapter 2: Data

2.1 Predictand: peak discharge

Times series information from USGS automated stream gauges are curated in the Unified Flash Flood Database (National Severe Storms Laboratory, n.d.) (https://blog.nssl.noaa.gov/flash/database/) to provide flooding peak discharge values at more than 10,000 locations across the U.S. A subset of 3,490 stream gauge locations is used, with stages corresponding to action, minor, moderate, and major flooding defined by the NWS in coordination with local stakeholders for modeling and diagnostics. This dataset covers diverse climatologist, hydrologic and weather conditions, which makes it a representative flash flood database over the U.S.

Gauges that are impacted by regulation or diversion are screened out using the regulation codes supplied by the USGS. In this database, a flood event is defined as the period when streamflow is above the defined action stage for that gauge. If there is a 24-hour period with discharge values below action stage, then the events are considered as separate. The database contains the start and end time when the flow first exceeded and dropped below the action stage threshold respectively, along with the time and magnitude of peak flow. The maximum basin area in this study is approximately 45,000 km^2 with a median area of 890 km^2 is suitable for analyzing the impact of rainfall spatial variability on floods

2.2 Predictor: geomorphology and climatology

A natural flood generally starts because of snowmelt or intense rainfall. But the physiography of the basin and sub-basin scale variability of rainfall will dictate the speed of conveyance of water through the channel network and the magnitude of the maximum discharge. Since the goal of this study is to understand the relative impact of rainfall variability and catchment features on flooding, the database is enhanced with attributes representing various landscape properties such as vegetation, topography, climatology, and soil. Several geomorphological parameters were derived from the Digital Elevation Model (DEM) data of the National Elevation Dataset (NED; http://ned.usgs.gov/) as potential explanatory variables of flash flooding. Flow accumulation and flow direction information was extracted by delineating basins with USGS stations. The National Hydrography Dataset (NHD; http://nhd.usgs.gov/) was used to resample the 30-m DEM to a 1-km grid to ensure compatibility between DEM-based flow accumulations and the actual river network across the Contiguous United States (CONUS). The geomorphologic parameters for delineated catchments were extracted from these grids using custom libraries developed using MATLAB. Variables representing soil properties such as mean depth-to-bedrock and K-factor (erodibility) were derived from the STATSGO database (Miller & White, 1998) while land cover and land use data from the National Land Cover Dataset (Fry et al., 2011) were used to estimate the runoff curve number. Lastly, the hydroclimatic variables of mean annual precipitation and temperature were extracted from the 30-year datasets (for period 1981-2010) prepared by the PRISM Climate Group of Oregon State University (http://www.prism.oregonstate.edu/normals/). The static spatially distributed basin attributes included in this study are provided in Table 1.

Table 1. Important Predictors for the study

Type	**Variable**	**Meaning**
Geomorphological	Area	Estimated Area (from Digital Elevation Model; flow grids)
	G1	First-order Moment of flow distance (Catchment averaged flow distance)
	G2	Second-order Moment of flow distance
	River Length	Length of the river systems
	Relief Ratio	R divided by Basin Length (highly correlated with drainage area)
	Ruggedness	Ruggedness expressed as drainage density multiplied by relief
	Slope to Outlet	Outlet Slope
	Rock Volume	Volume of rock; similar to rock depth
Precip moments	Activated Basin	Part of the basin where rainfall falls
	Rainfall Volume	Rainfall Volume

	Product Mean	Mean of the product of accumulated precipitation and flow distance of the activated basin
	Flow Distance (Mean)	Mean of flow distance of the activated basin
Climatological	bio_10	Mean Temperature of Warmest Quarter
	bio_15	Precipitation Seasonality
	Snow Percentage	Percentage of Snow in the Gauge
	Temp (Mean)	Climatological Average temperature

The dataset includes 21,143 rainfall events (observations) over 133 variables. These variables include morphological, bioclimatic, climatological, precipitation and gauge observations from across 902 different basins over the Contiguous United States (CONUS). Among these variables, the precipitation variability is described through precipitation moments (Zoccatelli et al., 2010). Flash flood are characterized by the observed *peak discharge* during a hydrological event at the basin outlet.

The provided dataset is devoid of missing values and is numerical in all its attributes. Out of the initial 133 variables, through prior domain knowledge and through previous studies on lag time studies (Duarte, 2019) a majority of variables were eliminated and only 50 predictors were selected. Eliminated variables were merely meant for quality control and deemed non-relevant for the model. The focus is primarily on the precipitation moments, the climatological and the morphological variables. Formulations for precipitation variability as moments are shown to provide a deeper understanding and representation of rainfall events (Duarte, 2019; Z. Zhang et al., 2012; Zoccatelli et al., 2010). They give an understanding on the spatial distribution of a rainfall event over a basin and how it impacts discharge at the basin outlet. Climatological and morphological attributes are equally important in describing the hydrological processes. Owing to such pruning measures 50 variables were selected.

This dataset repurposes existing data through a rigorous preprocessing that eases predicting, characterizing and understanding flash floods. Furthermore, its representativeness was demonstrated by Saharia et al. (2017) by mapping basin flashiness over the U.S. to predict flash flooding severity in ungauged regions with fair accuracy.

The dataset was created by (Saharia et al., 2017) by sourcing from multiple previous works (Gourley et al., 2013). The three primary sources for the database are: 1) the automated discharge observations from the U.S. Geological Survey, that have been reprocessed to describe individual flooding events, 2) flash-flooding reports collected by the National Weather Service from 2006 to 2013, the Multi-Radar/Multi-Sensor precipitation reanalysis (J. Zhang et al., 2016)(J. Zhang & Gourley, 2018).

Feature engineering is the process of transforming raw data into predictors using domain knowledge to better represent the objective. It improves the performance of machine learning algorithms. Feature engineering was performed retroactively once important predictors were identified through the modeling. As will be shown in next section, *Area* i.e., the Estimated Area (from Digital Elevation Model; flow grids) for the given basin is deemed an important variable. *Precipitation Mean* is the mean of precipitation averaged over the duration of the rain event on the activated basin (*Activated Basin;* part of the basin where rainfall falls). Knowing this, additional variables were engineered as follows:

$$Rainfall\ Proportion = \frac{Activated\ Basin}{Area}$$

$$Rainfall\ Volume = \ Activated\ Basin * Precipitation\ Mean$$

By expressing the active basin as a percentage, an additional predictor can be created that captures the volume of water collected from precipitation by the basin and that contributes to the peak discharge (m^3/s) which technically depends on the amount of rainfall accumulated in the basin.

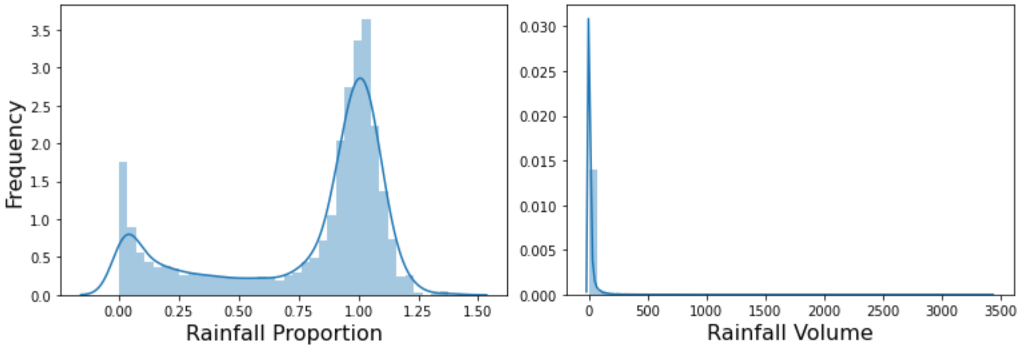


Figure 3. Histogram of engineered predictors

To understand the data and its distribution, univariate distribution plots, such as histograms with fitted and kernel density estimators, were plotted for all the predictors. The probability distribution function generated made clear that the predictors are not normally distributed as can be seen in the histograms in Figure 3 and 4.

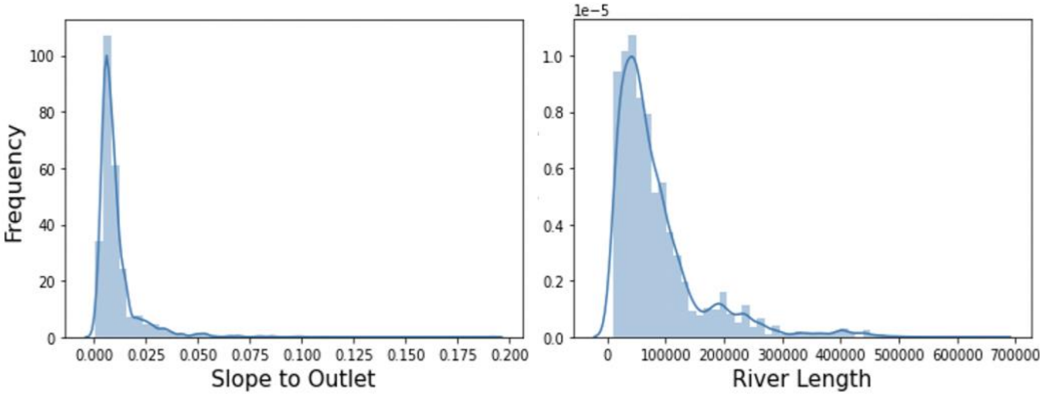


Figure 4. Histogram prior to transformation

The features in the dataset have different ranges. In order to make the data normally distributed and bring all attributes on a common scale, we need to normalize the data. This ensures that the algorithms sensitive to skewness and scaling will not be affected. Most of the attributes exhibited skewness in their distributions along with a wide range of value scales. Through the examples in figure 2 and 3 we can see that while *Slope to Outlet* has a very short range 0-0.2, while *Rainfall volume* has a range of 0-3500. Similarly, there are predictors having negative values as well. Owing to this, we needed to normalize the data.

Log and Box-Cox transformation could not be performed due to the presence of negative values in some of the attributes. Z-score standardization in the previous attempts (Duarte, 2019) did not produce great results. Hence Yeo-Johnson power transformation was selected, as it performs similar transformation to the log and Box-Cox while also managing negative values. Cubic transformations can also be tried in the future, as they too handle negative values and are used in rainfall datasets. The scipy package in python contains the Yeo-Johnsons transformation function which finds the optimal lambda (λ) parameter that maximizes the log-likelihood function and transforms the dataset. An example of the normalization of the target attribute peak discharge (*peakq*) is provided in figure 5.

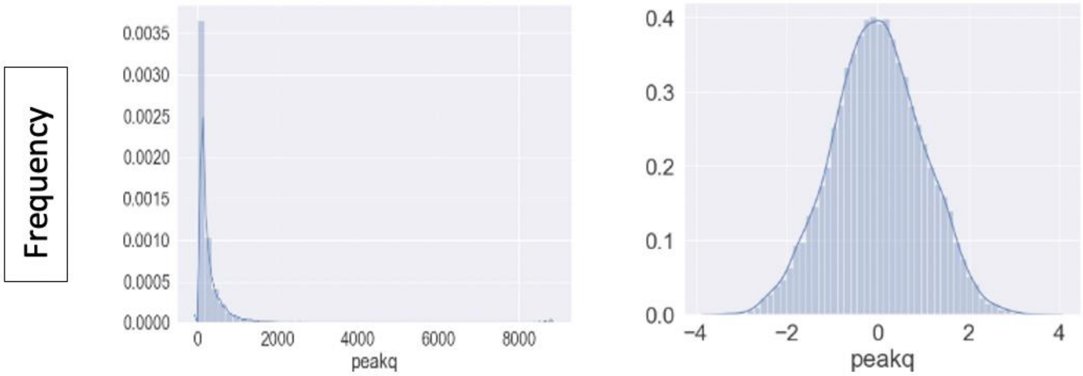


Figure 5. Histogram before (left) and after (right) transformation

These transformations were performed so that the regression methods such as multiple linear regression, lasso and elastic net could function as intended. Also, to keep a leveled comparison between the different regression methods, we used transformed data as inputs in all the 5 regression approaches that were compared. However, no transformation was used in the final implementation as we choose to perform a tree-based regression algorithm which is invariant to monotonic transformations of the independent variables. As we are using gradient boosting, a tree-based approach, to predict peak discharge no such preprocessing was performed.

Chapter 3: Methodology

3.1 Approach

A statistical (predictive) model is a mathematical representation of the problem statement concerning the data. Analysis of these models help understand and interpret the predictor relationships, make predictions on unseen data, and visualize that information. Our modeling approach involves the use of regression to perform the prediction of a continuous dependent variable "Peak Discharge" from several independent predictors.

3.2 Algorithm Selection

Algorithms selection (Lin & Li, n.d.-b, n.d.-a) is essential to identify the best performing model. We choose five main algorithms to compare with each other:

Multiple Linear Regression, Lasso, Elastic Net, Random Forests and XGBoost. Multiple Linear Regression identifies a linear relationship between multiple the predictor (explanatory) variables and target (response) variable using ordinary least-squares regression. Least Absolute Shrinkage and Selection Operator (Lasso) is a regression algorithm that performs both variable selection and regularization to increase the prediction accuracy of the target variable. Elastic Net is an embedded linear regression model trained with both l1 and l2 -norm regularization of the coefficients. This combination allows for learning a model where some weights are non-zero (as seen in Lasso) while also utilizing the regularization properties of Ridge. Random forests are an ensemble learning model that creates multiple decision trees at training time and generates a probability of the output in terms of mean prediction (for regression). They often overfit their training set and hence Extreme gradient boosting is also selected, as its more regularized model formalization is designed to mitigate this issue. Comparison of the models derived from these algorithms was done on default parameters.

The algorithm selection was based on accuracy metrics, and the domain understanding that the selected predictors imparted. The selected predictors should be able to explain basin physics and must adhere to the prior understanding from hydro metrology. Based on these constraints we finally end up using Extreme Gradient Boosting (XGBoost), a machine learning technique which produces a prediction model in the form of an ensemble of weak prediction decision trees. It is a supervised learning algorithm designed for fast computational time, especially on very large data sets. XGBoost is a form of gradient-boosted decision trees that can generate new models based on the prediction of the residuals' errors of prior models. The term "gradient boosting" refers to the utilization of a gradient descent to minimize the loss when adding additional models (Brownlee, 2016). XGBoost combines the benefits of the tree-based and gradient boosted models to overcome multi-collinearity. Its robustness towards correlated predictors is an advantage in the context of prediction with respect to its counterparts (e.g., Random Forests, Elastic Net, Lasso). This supervised learning algorithm builds models sequentially and generalizes them by allowing optimization of a differentiable loss function (root mean square error). XGBoost is well known for great performance in terms of speed and prediction accuracy and lower overfitting (Brownlee, 2016). The model predictions are evaluated through various error/performance metrics (see below).

3.3 Performance Metrics

Amongst many possible models, the best one should explain as much variance as possible (in the sense of R^2) and minimize the overall bias (in the sense of Mean Relative

Error) while minimizing over-fitting. Models were compared using performance metrics that target systematic discrepancies and random errors in the model predictions with respect to observations. The Mean Relative Error is used to quantify systematic error, while Root mean squared error (RMSE) is used to describe the random error. To quantify over-fitting amongst different models, an Accuracy Loss is introduced as the difference in R^2 values obtained when comparing the model predictions with the training and test data (see below).

3.4 Data Partitioning

Data partitioning is used to split the main data set before model creation, so that data are available to objectively assess the model. Such a testing approach is designed for reducing overfitting, bias and variance. The best practice is to split the data into three smaller data sets, i.e., training, validation and test sets. The training subset is used to create the model that relates the predictors and the predictand (flood peak values) and perform exploratory data analysis. The validation subset is unseen while model training. It is used to tune the model structure through hyperparameters (e.g., learning rate, depth of the tree) and compare performance between different models. The testing dataset is used to objectively assess the performance of the final model.

We performed data partitioning using stratified random sampling and by splitting data into training, validation and testing sets using a 70:15:15 ratio (cf. Figure 6). By dividing a population into distinct strata and then randomly sampling from each stratum, this sampling technique helps each of these datasets be representative of each other. We ensure this by comparing the mean peak discharge value (i.e., the predictand) in all the sets shown in Figure 7.

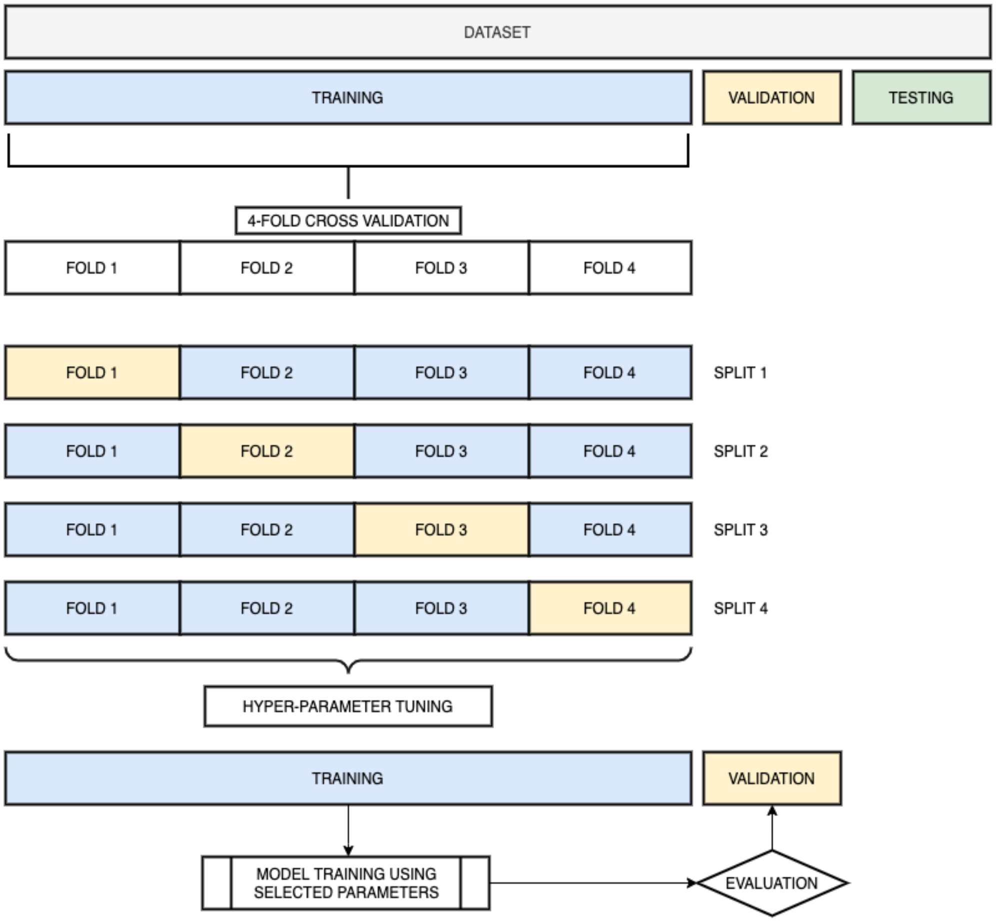


Figure 6. Data Partitioning using 70:15:15 ratio for training, validation and test dataset. K-fold cross-validation performed on training data.

3.5 Model Training

We build our predictive model on the training data set. An XGBoost model derives trees defined by varying depth and number of nodes according to user's specifications. The ensemble of trees that are generated (or learned) while training becomes the parameters for the predictive model. These trees are defined by hyper-parameters such as subsample ratios of predictors, learning rate, max depth, etc. Hyper-parameters cannot be estimated while training and are tuned manually to achieve the best model performance (hyper-parameter tuning).

In the present study, hyper-parameter tuning is performed on the training set to optimize performance (i.e., reducing bias and variance) and identify the best model. We do so through a testing technique known as 4-fold cross validation (cf. Fig. 6). The training data set (sample) is randomly partitioned into four equal sized subsamples. A single subsample is retained as the inner-fold validation set for checking the model performance, and the remaining three subsamples are used as inner-fold training sets. The cross-validation process is repeated four times, with each of the four subsamples used once as the inner-fold validation data. The four results are averaged to produce a single estimation for a single hyper-parameter combination.

Cross validation serves the goal improving the representativeness of the model by using all observations for both training and validation. Among the possible designs of k-fold cross validation, k = 4 was selected as a trade-off between model refinement and computational time. Searching for the best parameter combination in the hyper-parameter space occurs by random selection. Such combinations are selected 50 times and each of them under-go a 4-fold cross validation (cf. Fig. 6). Essentially, 200 (50*4) models are tested to find the best hyper parameters. Once the best hyper-parameter combination is identified, we use those settings to train the predictive model on the training data. The model performance is then checked on the validation dataset.

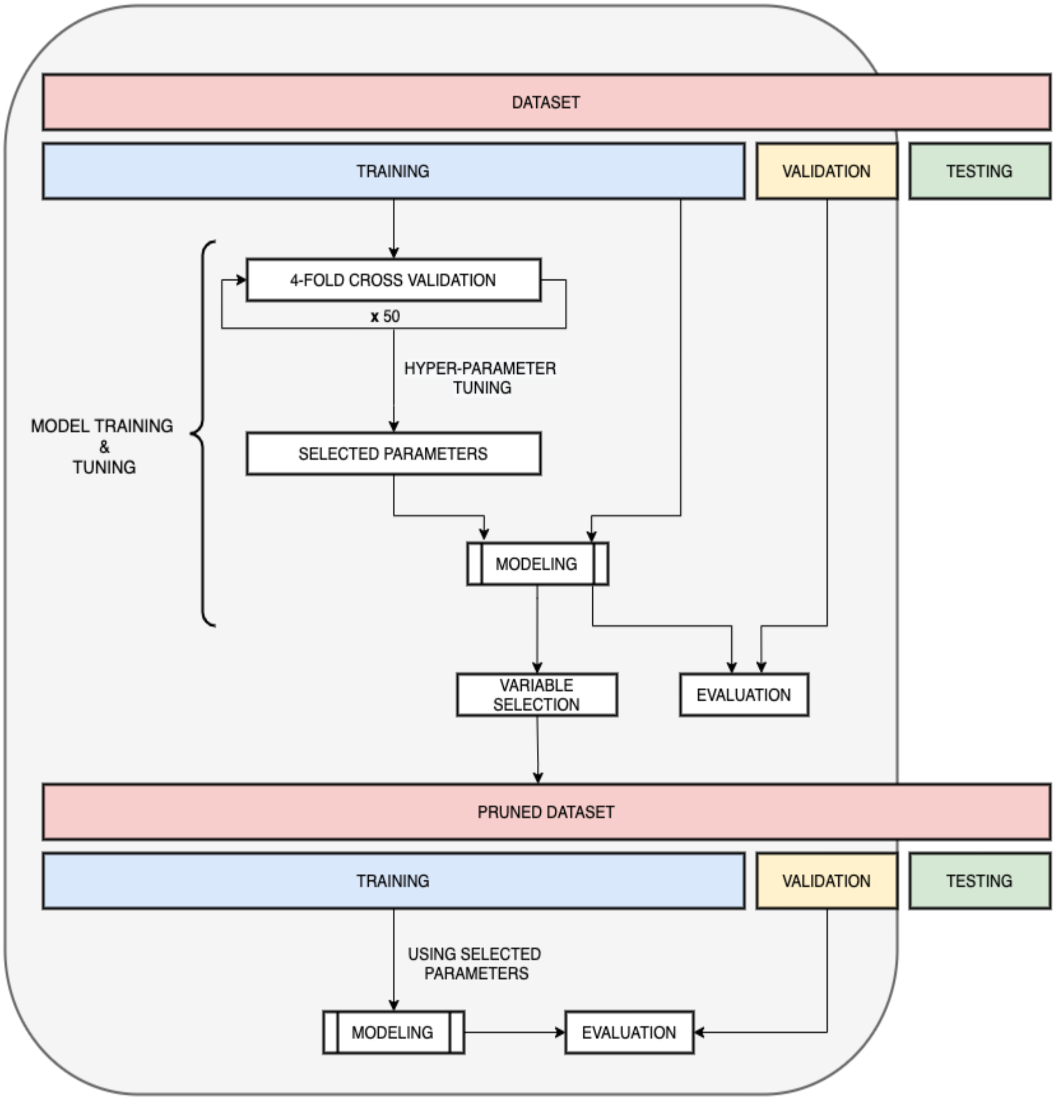


Figure 7. Modelling methodology

3.6 Predictor Selection and importance

Predictor selection is performed recursively after initial modelling of data (as mentioned above). Gradient boosted trees identify the predictor importance by measuring the mean decrease in impurity (variance). While training a tree, we can compute how much each predictor decreases the weighted impurity in a tree. Predictor Importance is a parameter that is used to rank predictors by averaging the impurity decrease from each predictor in a forest of trees.

‘Predictor importance’ is an absolute value, which implies how much reduction occurred in the standard deviation (at the leaf of the decision tree) when the said predictor was used. The more the reduction, the better the importance of the predictor. Hence, if the predictor keeps appearing as ‘important’ in the 40 modelling runs, its cumulative score will be higher, as is shown by the ‘Area (estimated area)’.

Upon receiving the importance scores for each predictor, we select the predictors whose importance is greater or equal to the mean of the all the predictor importance values. This selection technique is chosen after comparing with other techniques such as Recursive Feature Selection, Permutation Feature Importance and other embedded methods (LASSO and Elastic Net). By using the methodology of comparing means of the predictor importance values, we are able to run our large dataset with the Monte-Carlo sub-sampling experiments to get unbiased estimates.

The selected predictors are used to create a pruned training/validation/testing dataset, upon which we train a more parsimonious model. Performance metrics between the parsimonious model and the initial model are systematically compared and its observed that the difference is negligible , validating the fact that parsimonious models perform as good as the model with all the predictors.

3.7 Ranking of Variables

In order to avoid any bias associated with the dataset partitioning, the entire modelling methodology in Figure 2 (grey area) is performed 40 times with different subsets of training and validation splits and we obtain 40 different models. An ensemble of models gives more insights on the predictor importance and reduces bias induced (Beven & Binley, 1992). As such, an uncertainty analysis implemented as a Monte-Carlo experiment enables objective extraction of a set of empirical models to identify the most important associated predictors. Note that this approach is different from the Generalized Likelihood Uncertainty Estimator (GLUE) which applies on conceptual models to quantify the prediction uncertainty that results from their design and structure.

The “repeated random sub-sampling validation,” i.e., Monte Carlo cross-validation, is used to generate multiple random splits of the dataset into training and validation data. For each such split, a model is fit to the training data by identifying the best hyper-parameter combination and by assessing the predictive accuracy on the validation data. With 40 such unique random splits, a total of 8000 models are created. Monte Carlo cross-validation allows to keep the proportion of the training/validation split independent from the number of iterations (i.e., the number of partitions) to ensure proper representation of the training and the validation data.

Each of the 40 models provides its own unique predictor importance ranking (see Figure 3). The 40 rankings are aggregated to identify the most important predictors in terms of frequency of occurrence (i.e., how often they are selected across the 40 models) and importance (i.e., in each of the 40 models). Descriptive statistics of the predictor importance are derived, such as frequency of occurrence and importance sum for each predictor. The sum is used to rank the predictors overall. If a predictor was selected at least once in the 40 iterations, it is deemed as important. With this definition, 32 predictors out of the 50 predictors are identified as important.

Chapter 4: Results

4.1 Algorithm Selection

In Fig.8 we compare performance metrics on the validation dataset to see that the best R^2 is achieved by Xgboost with 83% explained variance followed by the Random Forests with 79% explained variance. While R^2 is indicative of the random error in the model, it does not give any information on the bias, hence we cannot yet choose Xgboost as the best performing model.

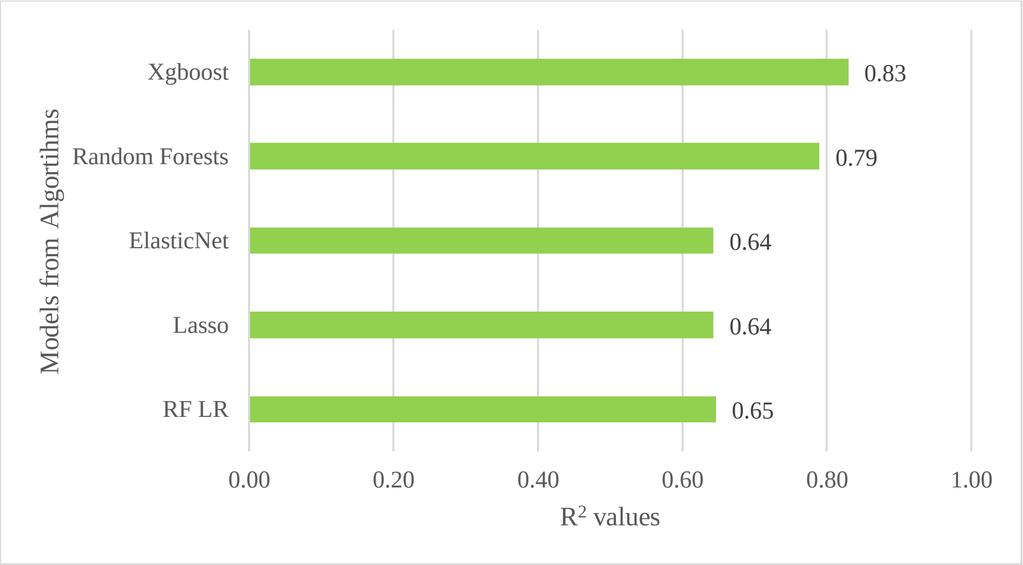


Figure 8. R^2 for models from different algorithms

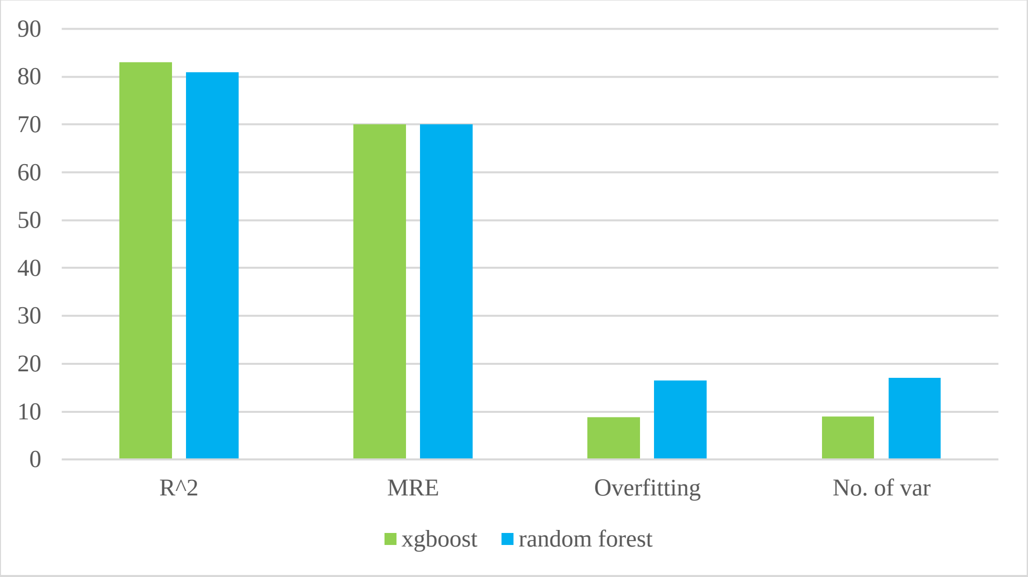


Figure 9. Evaluation metric comparison

As mentioned earlier, bias and variance must be taken into context through the introduction of mean relative error (MRE). Using R^2 and MRE, we can choose a comprehensive model. As shown in Figure 9, in comparison of the two best performing algorithms, for the validation dataset the R^2 and RMSE is better by 0.02 and 0.03, respectively, for the XGBoost. We see that the extreme gradient boosting algorithm and the random forests have equivalent MRE of 70%. Furthermore, overfitting and the number of variables for Xgboost is almost half of that of random forest. Also, we see an average of 18% and 10% accuracy loss across all the models of random forests and XGBoost respectively. Hence more overfitting is observed in the random forest models. Based on these inferences, we use XGBoost algorithm to create models and further analyze hydrological processes.

4.2 Model Selection

A model should be general enough such that it explains the variance of the entire dataset to a satisfactory level. This means that metrics such as the coefficient of determination (noted R^2 herein) is maximized while overfitting should be minimized. Amongst the 40 models we generate, we choose the model with the best R^2 on the validation dataset (cf. Figure 10), while also keeping the training and validation performance loss to less than 15% (cf. Figure 11). Furthermore, the number of selected predictors has to be below 20 (reducing more than 60% of predictors), such that we could generate a sufficiently parsimonious model, and these predictors have to explain more than or equal to 60% of predictor importance (cf. Figure 13). In figures 10, 11 and 12, the red line indicates the performance of selected model amongst the 40 runs while the green writing indicates the constraint that was subjected to select the best model.

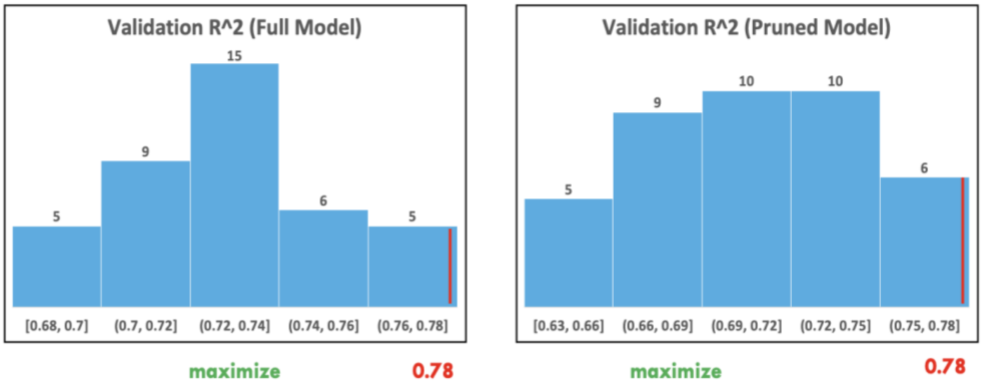


Figure 10. Selecting model based on performance on validation set.

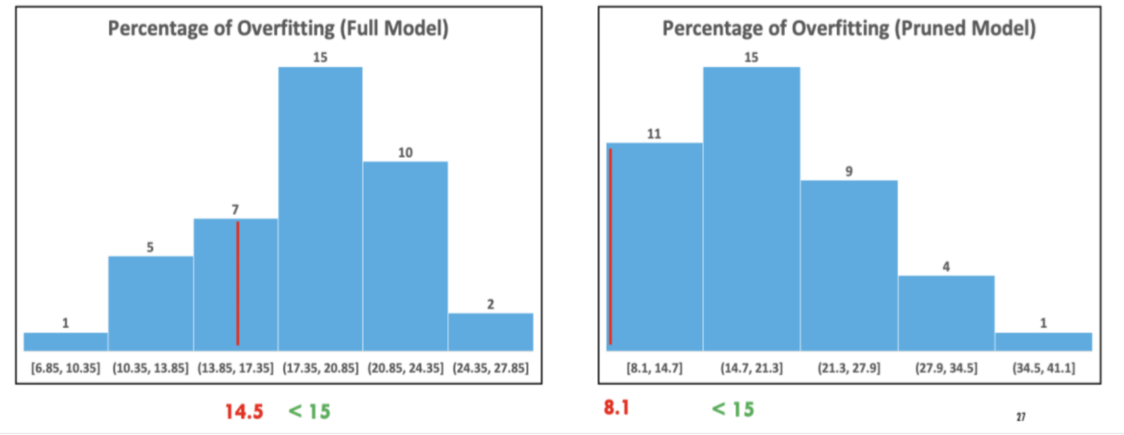


Figure 11. Selecting model based on overfitting

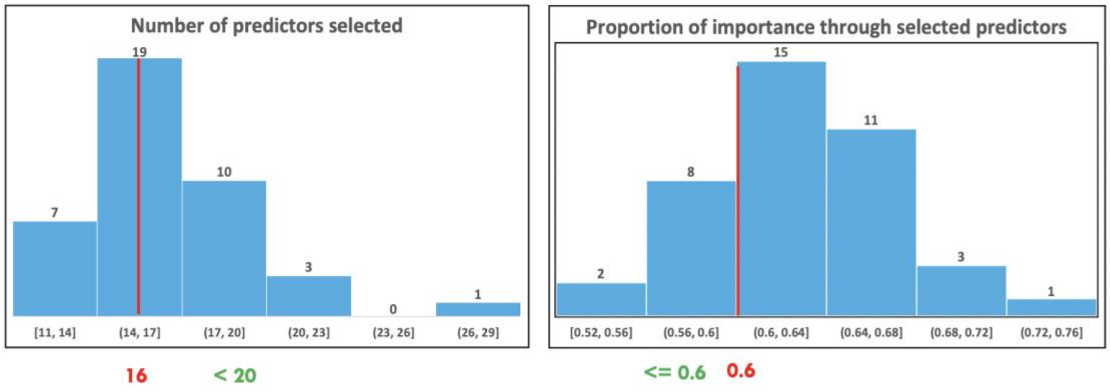


Figure 12. Selecting model based on number of predictors selected and proportion of importance explained.

Another substantiation is with regards to the physical realism of the empirical model. This is performed by checking the predictor importance and partial terms. Amongst the categories of processes that impact the hydrologic response of a basin, geomorphology is expected to have the greatest impact, followed by the spatial distribution of precipitation (precipitation moments), and finally the climatology. Hence a model should also reflect physical consistency in terms of predictor importance.

To assess such physical consistency globally, we consider the overall ranking of predictors across the 40 models. By combining the predictors by category of processes (geomorphology, precipitation spatial variability, climatology), a representative value of importance for each category is extracted with the importance median value. Upon doing so, we observe that the predictor importance map obtained from our chosen model is able to explain the physics of the basin in terms of hierarchy of importance of different predictor categories (mentioned above).

4.3 Model Performance & Feature selection

Through the model selection methodology, we select a model whose performance on the validation dataset has a R^2 of 0.78. The selected model is now tested on the testing dataset that was untouched in the all the selection and ranking procedures. On the testing dataset, the model predictions have a R^2 value of 0.77, mean relative error of 0.02 and the root mean squared error of 157.25 with a train vs test R^2 loss of 15%. With such metrics we can conclude to have an unbiased model.

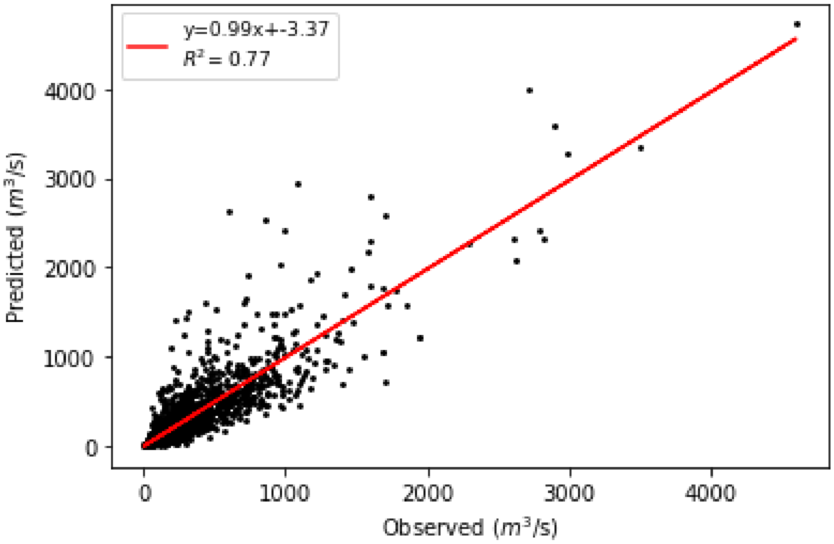


Figure 13. Predicted vs Observed peak discharge values (Test Dataset)

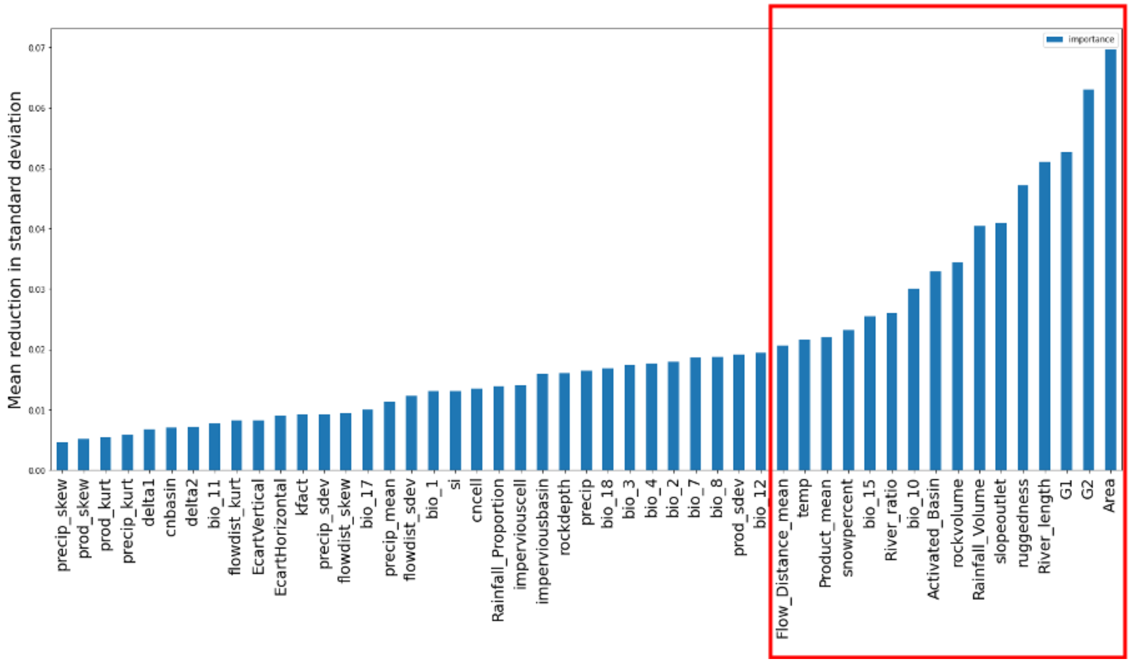


Figure 14. Predictor Importance Map.

Amongst the 50 predictors we choose 16 important predictors as seen in figure 14. Of these identified we have 9 geomorphological, 3 precipitation moments, 4 climatological variables as shown in table 1.

4.4 Accumulated Local Effects Plots

Additional insight is provided by Accumulated Local Effects (ALE) Plots. An ALE plot highlights the average impact of a given predictor on the model predictions (Molnar, 2019). ALE plots are unbiased and valid when predictors are correlated. They help reduce complex prediction functions to a newer function which solely depends on the predictor of interest. To understand the influence of a given predictor in multivariable functions, differences in prediction are calculated for the predictor, which are averaged (accumulated) to define the partial derivative of that predictor, before it is centered. The partial derivative for that predictor is computed by holding all the other predictors constant. ALE plots utilize this basic calculus and find partial derivatives conditional on the features' values. This derivative is further integrated to focus on the predictor and filter out the interaction with correlated predictors. The generated value is then centered by subtracting over a constant (e.g., mean value) to improve the interpretation.

To estimate the gradient with the Xgboost model, the predictor is binned into intervals and differences in predictions are computed. Bins are based on percentile values taken by the predictors to ensure uniformity across bins. The differences in the prediction relays the effect in terms of partial derivative of the predictor for each individual instance in a bin. These partial derivatives are conditionally averaged over each bin to estimate the local effects. These local effects are summed (accumulated) across all bins to derive ALE values, that are finally centered.

The ALE plots for geomorphological attributes, spatial distribution of precipitation, and climatology are provided below. To interpret the ALE values, one should consider the value on the y-axis as the conditional effect of the given predictor, when compared to the overall mean prediction for that bin. For instance, if the difference in the peak discharge is -65 for the 10th percentile of Area, then the prediction is lower by 50 cm^{3}s^{-1} in comparison with the mean prediction involving all predictors.

Figures 15, 16, 17 show the ALE plots for the geomorphological, precipitation moments, and the climatological predictors, respectively. To allow a comparison between variables, ALE values are computed at percentile bins of each variables. Consistent with the importance map (Fig. 3), in general the geomorphological ALE plots display larger ranges of variations than the precipitation and the climatological ALE plots, indicating that the geomorphological predictors have a higher impact on the model output (i.e., they generate higher differences in predicted peak discharge), while the climatological predictors have less impact.

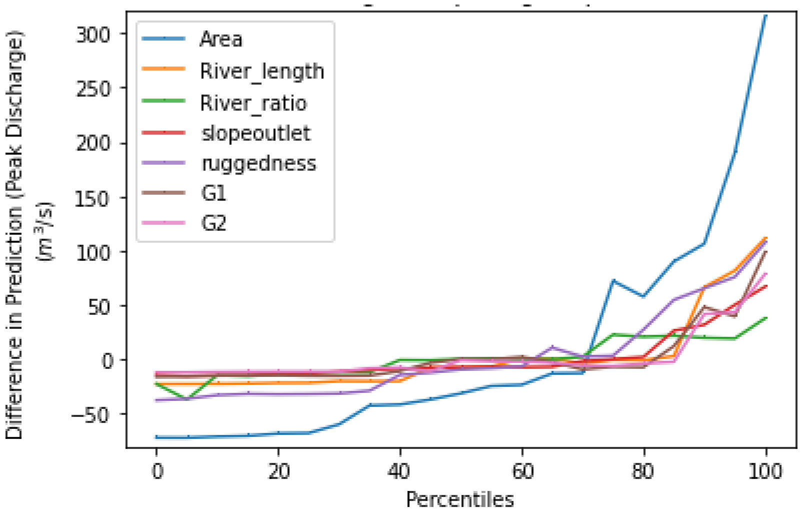
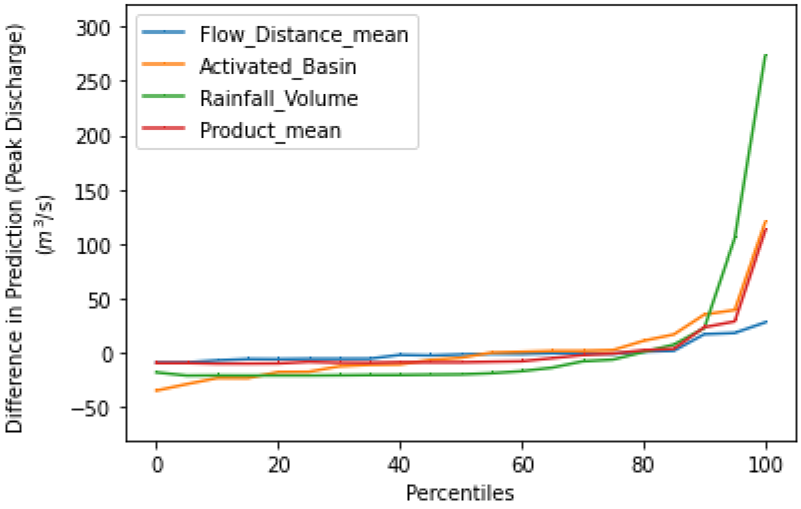


Figure 15. ALE analysis geomorphological predictors



.

Figure 16. ALE analysis Precipitation moments predictors

An estimation of volume of water, processed through hydrological processes in order to give is a peak value during the flood. Bio_10, identifies regions where we have specific atmospheric processes which generate high floods. High peak-discharge values are correlated with high precipitation, which are often associated with thunderstorms and convection. These occur in areas which are warm and moist, and hence it indirectly affects everything. The longer the flow distance, the longer the time for the flood peak to appear, important for showing the temporal delay, less important for the volume of water. Other plots are provided in the appendix.

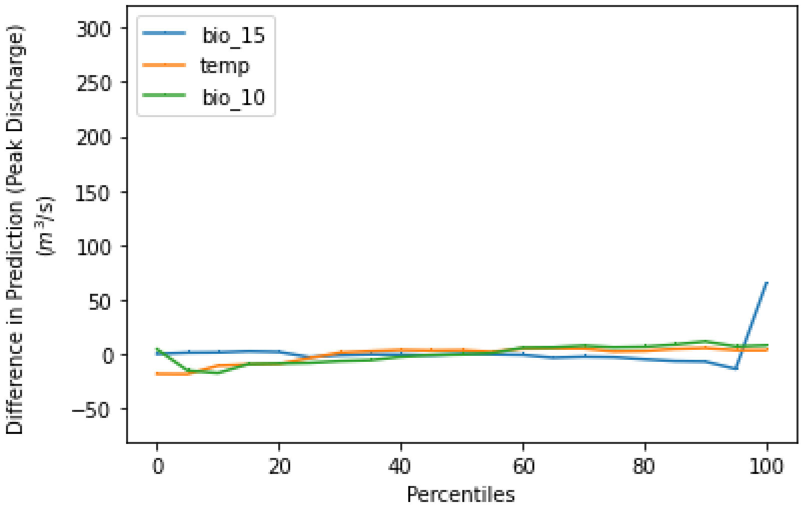


Figure 17. ALE analysis climatological predictors

4.5 Final Model Performance

The new model (parsimonious model), which is trained on the pruned dataset containing only the above selected variables, shows performance on the test dataset with R^2 value of 0.76, mean relative error of 0.02 and the root mean squared error of 159.62 with a train vs test R^2 loss of 10%. This model will be used for prediction if implemented for real time analysis.

Chapter 5: Conclusion & Future Work

5.1 Conclusion

Through this workflow, we successfully characterized flood peak discharge using machine learning. The dataset that was collected from precipitation and flood events across the US captures a large variety of precipitation spatial moments and basin geomorphological and climatological characteristics. Such a highly dimensional dataset helped train a statistical regression model for flash floods.

Among various regression algorithms and models considered, the selection was performed based on the coefficient of determination (R2), the mean relative error (MRE) and the root mean squared error (RMSE). In order to build a model with low bias and variance and with minimal overfitting, data partitioning was applied to create a training, validation and testing dataset. We then selected the XGBoost algorithm to fit the model on the training dataset. The best hyper-parameters for the XGBoost algorithm were identified prior to the training for best performance. The model was then tested on the validation dataset.

The entire process of data partitioning and model creation was performed 40 times in a Monte Carlo approach, and these results were aggregated to identify the model that reflects basin physiography the best through predictor significance. The predictor importance maps generated from the model helped quantify the importance of the basin characteristics.

The selected model was tested on the partitioned testing set to test its performance. The response of the peak discharge to the individual predictors is visualized using accumulated local effects plots. This measures the impact of each predictor and provides more insights on the response to various classes of predictors. A parsimonious model that can be used in future deployments was then built by pruning the dataset to contain only the most important predictors.

The identified key predictors were backed up with physical considerations, i.e., understanding of the hydrological processes. The methodology allows the modeler to use domain knowledge to select the models that conform with the base reality. Furthermore, these inferences also bolstered the idea of how a new approach to hydrological modeling using machine learning, that encompasses the best of the physical, conceptual and empirical models. Such an approach will potentially lead to new modeling techniques and contribute to analyze the hydrologic behavior of watersheds.

Likewise, the model shows promising performances in terms of predictions with great accuracy and low random error. It paves a way towards flood forecasting that can be considered in future work.

5.2 Future work

The current study provides a blueprint for creating models for real time flood prediction. It would require combining peak discharge with other flood characteristics such as lag time and flood threshold exceedance levels. Also, precipitation moments would need to be computed in real time.

Next, using the methodology outlined in this study, other efficient models can be created that minimize overfitting and are representative of the basin behavior to generate robust predictions. Further testing should be performed to check the visual and spatial consistency. This implementation could also be compared to existing flood forecasting systems like EF5. Finally, work should focus on making the information from this model simple for the forecaster to understand and ingest.

References

Ashley, S. T., & Ashley, W. S. (2008). Flood fatalities in the United States. *Journal of Applied Meteorology and Climatology*, *47*(3), 805–818. https://doi.org/10.1175/2007JAMC1611.1

Beven, K., & Binley, A. (1992). The future of distributed models: Model calibration and uncertainty prediction. *Hydrological Processes*, *6*(3), 279–298. https://doi.org/10.1002/hyp.3360060305

Brownlee, J. (2016). *XGBoost With Python: Gradient Boosted Trees with XGBoost and scikit-learn*. Machine Learning Mastery. https://www.goodreads.com/book/show/50621772-xgboost-with-python

Devia, G. K., Ganasri, B. P., & Dwarakish, G. S. (2015). A Review on Hydrological Models. *Aquatic Procedia*, *4*, 1001–1007. https://doi.org/10.1016/j.aqpro.2015.02.126

Duarte, J. (2019). *Probabilistic characterization of floods from catchment-scale precipitation moments*.

Flamig, Z. L., Vergara, H., & Gourley, J. J. (n.d.). *The Ensemble Framework For Flash Flood Forecasting (EF5) v1.2: Description and Case Study*. https://doi.org/10.5194/gmd-2020-46

Gourley, J. J., Hong, Y., Flamig, Z. L., Arthur, A., Clark, R., Calianno, M., Ruin, I., Ortel, T., Wieczorek, M. E., Kirstetter, P. E., Clark, E., & Krajewski, W. F. (2013). A unified flash flood database across the United States. *Bulletin of the American Meteorological Society*, *94*(6), 799–805. https://doi.org/10.1175/BAMS-D-12-00198.1

Jackson, A. (2014). *Discharge & Hydrographs*. https://geographyas.info/rivers/discharge-and-hydrographs/

Lin, H., & Li, M. (n.d.). *Chapter 11 Tree-Based Methods | Introduction to Data Science*. Retrieved April 16, 2020, from https://scientistcafe.com/ids/treemodel.html

Miller, D. A., & White, R. A. (1998). A Conterminous United States Multilayer Soil Characteristics Dataset for Regional Climate and Hydrology Modeling. *Earth Interactions*, *2*(2), 1–26. https://doi.org/10.1175/1087-3562(1998)002<0001:acusms>2.3.co;2

National Severe Storms Laboratory. (n.d.). *Database – FLASH*. Retrieved November 16, 2020, from https://inside.nssl.noaa.gov/flash/database/

NOAA. (2005). *The Awesome power of Floods*. http://www.redcross.org/http://www.fema.gov/

Saharia, M., Kirstetter, P.-E., Vergara, H., Gourley, J. J., Hong, Y., & Giroud, M. (2017). Mapping Flash Flood Severity in the United States. *Journal of Hydrometeorology*, *18*(2), 397–411. https://doi.org/10.1175/jhm-d-16-0082.1

Sillmann, J., Kharin, V. V., Zwiers, F. W., Zhang, X., & Bronaugh, D. (2013). Climate extremes indices in the CMIP5 multimodel ensemble: Part 2. Future climate projections. *Journal of Geophysical Research Atmospheres*, *118*(6), 2473–2493. https://doi.org/10.1002/jgrd.50188

Smith, A. B. (2020). *2010-2019: A landmark decade of U.S. billion-dollar weather and climate disasters | NOAA Climate.gov*. NOAA Climate.Gov. https://www.climate.gov/news-features/blogs/beyond-data/2010-2019-landmark-decade-us-billion-dollar-weather-and-climate

Solomatine, D. P., & Wagener, T. (2011). Hydrological Modeling. In *Treatise on Water Science* (Vol. 2, pp. 435–457). Elsevier. https://doi.org/10.1016/B978-0-444-53199-5.00044-0

Sweeney, T. L., & Baumgardner, T. F. (1999). *Modernized flash flood guidance*. NWS Hydrology Laboratory,Pg 11. https://www.nws.noaa.gov/oh/hrl/ffg/modflash.htm

Zhang, J., & Gourley, J. (2018). *Multi-Radar Multi-Sensor Precipitation Reanalysis*. Open Commons Consortium Environmental Data Commons. https://doi.org/https://doi.org/10.25638/EDC.PRECIP.0001

Zhang, J., Howard, K., Langston, C., Kaney, B., Qi, Y., Tang, L., Grams, H., Wang, Y., Cockcks, S., Martinaitis, S., Arthur, A., Cooper, K., Brogden, J., & Kitzmillller, D. (2016). Multi-Radar Multi-Sensor (MRMS) quantitative precipitation estimation: Initial operating capabilities. *Bulletin of the American Meteorological Society*, *97*(4), 621–638. https://doi.org/10.1175/BAMS-D-14-00174.1

Zhang, Z., Koren, V., Reed, S., Smith, M., Zhang, Y., Moreda, F., & Cosgrove, B. (2012). SAC-SMA a priori parameter differences and their impact on distributed hydrologic model simulations. *Journal of Hydrology*, *420–421*, 216–227. https://doi.org/10.1016/j.jhydrol.2011.12.004

Zoccatelli, D., Borga, M., Zanon, F., Antonescu, B., & Stancalie, G. (2010). Which rainfall spatial information for flash flood response modelling? A numerical investigation based on data from the Carpathian range, Romania. *Journal of Hydrology*, *394*(1–2), 148–161. https://doi.org/10.1016/j.jhydrol.2010.07.019