Risk Factors for 100-Year Flood Events in the Mid-Atlantic Region of the U.S.

Gina Tonn[1*], Seth Guikema[2]

[1] Delaware Department of Natural Resources and Environmental Control (DNREC), Dover, DE, USA

[2] Department of Industrial & Operations Engineering, University of Michigan, Ann Arbor, MI, USA

*Address correspondence to Gina Tonn, Delaware DNREC, 285 Beiser Blvd., Suite 102, Dover, DE 19904, USA, gina.tonn@delaware.gov.

ABSTRACT:

Anecdotal information indicates that streams in the Mid-Atlantic region of the U.S. experience more extreme flood events than might be expected. This leads to the question of whether this is an unfounded perception or if these extreme events are actually occurring more than should be expected. If the latter is true, is this due solely to randomness, or alternately to characteristics that make certain watersheds more prone to repeated events that may be defined as 100-year or greater floods? These questions are investigated through analysis of flood events based on standard flood frequency analysis. 100-year streamflow rates for stream gages were estimated using Bulletin 17B flood frequency analysis methods, and the probability of the annual peak flow record for each gage was calculated. These probabilities were compared to a set of synthetic probabilities to evaluate their distribution. This comparison indicates that for the Mid-Atlantic region as a whole, the Bulletin 17B method does not systematically over or underestimate flood frequency. A Random Forest model of probability of actual flood record (PAFR) versus watershed and stream gage characteristics was developed and used to understand if certain characteristics are associated with PAFR. This analysis indicated that unexpected numbers of large flood events in a stream gage period of record can be attributed primarily to randomness, but there is some correlation with watershed and gage characteristics including weighted skew, drainage area, and mean annual peak discharge. The results indicate that watersheds with high values of these characteristics may warrant advanced flood frequency methods.

KEY WORDS: Flood frequency; Random Forest; 100-year flood

1. INTRODUCTION

Anecdotal information indicates that along some rivers and streams in the U.S. Mid-Atlantic region, 100-year flood events occur more frequently than might be expected. News headlines such as "100-Year Flood, for the Second Straight Year" (Clines, 2002) and "Potomac '100-year flood' hits twice in eight months" (Roylance, 1996) reinforce the perception that the occurrence of these floods may not follow expectations. The terminology used – 100-year flood – creates confusion amongst the

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/risa.13659.

general public who may assume that only one 100-year flood can occur in a 100-year period. The term 1% annual flood is used alternatively to address this concern (USGS, 2016). In either case, some rivers and streams experience repeat occurrences of these extreme events, which leads to the question of whether this can be attributed solely to randomness, or alternately to some characteristics of these watersheds and streams that make them more prone to repeated events that may be defined as 100-year or greater floods based on standard flood frequency analysis (FFA) methods.

FFA is a commonly used tool for quantifying flood risk. There are different types of FFA, and widely used methods include statistical analysis of local flood or regional flood records and rainfall-runoff modeling (Merz & Bloschl, 2008). Within the categories of statistical analysis of records and rainfall-runoff modeling, many different methods exist (Villarini & Smith, 2010, Villarini et al., 2011, Paquet et al., 2013). The focus of this paper is FFA involving statistical analysis of local stream gage records in order to estimate peak discharge for specified recurrence intervals. While FFA is useful and widely used, it is based on data sets with limited records and uncertainty in the methods is considerable (Merz & Thieken, 2005). Sources of uncertainty include the magnitude of future hydrologic events, use of simplified models, economic and social uncertainty that influence land use change, performance of water-control measures like levees, dams, and stormwater management features, accuracy and length of observations, other flood-influencing variables, and climate non-stationarity (USACE, 1996; Morss et al., 2005). Recent advances in FFA have employed probabilistic and synthetic hydrographs (Ahmadisharaf et al., 2018; Brunner et al. 2017, 2018a, 2018b), bivariate or multivariate return periods (Brunner et al., 2016; Graler et al., 2013), and watershed characteristics (Rogger et al., 2012).

Additional statistical analysis may serve to better evaluate flood risk and identify conditions for which standard FFA may misestimate flood risk. This project uses statistical analysis to evaluate whether or not stream gages in the Mid-Atlantic region experience unexpected numbers of 100-year or greater flood events based on standard FFA. It also uses statistical analysis to investigate whether the likelihood of an observed flood frequency record is at least partially explained by watershed characteristics and stream gage statistical characteristics. The focus of this research is extreme streamflow events (100-year or greater) in the Mid-Atlantic region, and the results are intended for use in identifying watersheds where advanced FFA methods may be warranted.

One of the first steps typically completed in assessing and managing flood risk is FFA. FFA is often followed by hydraulic modeling to estimate flood elevations at specific locations and to generate floodplain maps. Floodplain maps are used by communities as tools to regulate development in floodplains and are developed by FEMA for setting flood insurance rates. Many flood risk management decisions are based on FFA. Use of the 100-year event, which is common in floodplain maps, was meant to be a preliminary approach, but has become a de facto standard for flood risk management in the United States (Galloway, 2011). Quite often the 100-year flood (the flow rate with a 1% probability of being exceeded in a given year) is used for design, analysis, and decision-making with little regard for how uncertainty factors into this figure (Christian et al., 2013). Research is underway to improve standard FFA methods (Stedinger, 2008). However, flood frequency results in the form of Federal Emergency Management Agency (FEMA) flood insurance rate maps (FIRMs) and studies are in wide use, and even with improved FFA methods, uncertainty is still considerable.

Bulletin 17B is a standard FFA method used in the U.S. (IACWD, 1982), and is the method employed in this study. The Bulletin 17B methods are not necessarily efficient, but they are consistent. Their adoption seeks to have all 20,000+ floodplains in the U.S. demarcated by the same methods and practices by engineers (Merz & Thieken, 2005). This method and the associated narrow decision-making process is deeply uncertain, as the analysis is primarily based on available stream gage data and a single flood frequency distribution in practice (Merz & Thieken, 2005).

Some common issues with FFA methods including Bulletin 17B are the lack of a physical basis for determining the underlying flood frequency distribution, and the need to look at flood risk for return periods longer than the period of stream gage record (Lettenmaier et al., 1987). Flood frequency results at stream gages vary, and a single type of distribution for flood frequency may not work equally well at different gage locations (Benson, 1962a). Villarini and Smith (2010) noted that spatial heterogeneity was apparent in flood peaks at stream gages in the eastern U.S. and should be addressed. Villarini et al. (2011) observed a heavier-tailed flood frequency distribution in the Eastern U.S. than in the Midwest and identified relationships between watershed characteristics and distribution parameters. Some gages experience more 100-year events than would be expected based on FFA given the period of record, while others experience fewer. Additionally, 100-year streamflow may significantly increase in some regions of the U.S. due to climate change and land use change driven by population growth (Kollat et al., 2012). From a risk analysis perspective, it would be useful to understand whether standard FFA is regularly over or underestimating the frequency of low likelihood events in the Mid-Atlantic region, and to have some estimate of which stream gage locations may experience records that standard FFA results suggest would be unlikely. This would aid in risk-based decisions around flood risk (Rosner et al., 2014) in applications such as siting of detention basins (Ahmadisharaf et al., 2016), reservoir management (Naz et al., 2018), managing land use, and protecting life and property.

The purpose of this study is to address these issues through the use of statistical learning methods. We evaluate the likelihood of the flood frequency outcomes at stream gages in the Mid-Atlantic region as a whole, and then identify watershed characteristics that are associated with conditions in which observed records would be judged to have higher or lower likelihoods based on Bulletin 17B results. That is, we seek to identify watershed and gage record characteristics that are associated with the probability of record.

The rest of this paper is organized as follows. Section 2 provides background information on FFA, watershed characteristics, and uncertainty. Section 3 describes the data and methods. Section 4 includes a presentation and discussion of results, and Section 5 presents conclusions from the study.

2. BACKGROUND

Various types of FFA that are widely used include statistical analysis of local flood records, statistical analysis of regional flood records, and rainfall-runoff modeling (Merz & Bloschl, 2008). The FFA method used in this study is the Log Pearson Type III method as implemented in Bulletin 17B, developed by the Interagency Advisory Committee on Water Data (IACWD, 1982). This method was selected due to its wide usage and acceptance in the U.S., including regulatory requirements to use the method for certain applications, such as FEMA flood insurance rate mapping. The Bulletin 17B method is an evolution of previous methods developed by the U.S. Water Resources

Council and was developed in an effort to provide an accurate and standard method to estimate flood frequency based on stream gage data. Bulletin 17B estimates are based primarily on stream gage records for the stream being studied and use the method-of-moments approach with a log-Pearson Type III distribution to determine the statistical parameters for a given gage station. Bulletin 17B includes methods to incorporate the systematic record, as well as historic data, regional data, and flood estimates based on precipitation records (IACWD, 1982). The method is reasonable and performs well compared to other potential methods (Stedinger, 2008).

An update of 17B was released in 2018 and incorporates proposed improvements such as the use of historical and interval data, regional skew computation and precision, and confidence intervals. Generally, it is still unclear what the contribution of nonstationarity is to uncertainty and whether estimates would be improved by including it, and difficulties in resolving the skew may still remain (診Stediner & Griffis, 2011, Ouarda & El Adlouni, 2011). Some other suggested approaches to improve the accuracy of FFA include more substantial use of historic or paleoflood data (Kirby & Moss, 1987). However, historic data are often limited, and there is no certainty that historic data can be found or will improve flood frequency estimates (Payrastre et al., 2011). “A simple model with well-understood flaws may be preferable to a sophisticated model whose correspondence to reality is uncertain” (Lins & Cohn, 2011). Because these FFA methods add complexity, it would be useful to identify watershed or stream gages characteristics for which advanced methods are warranted due to poor performance of the standard method.

Studies have been performed to explain how flood magnitudes vary based on physical and climatic characteristics of a watershed. A study by Benson (1962b) found that drainage area, main channel slope, and surface area of lakes and ponds were important variables. Watershed characteristics have also been widely used in developing regional regression equations and in estimating peak discharge at ungaged watersheds (Lettenmaier et al., 1987; Pandey & Nguyen, 1999; Wiltshire, 1985). Statistical characteristics of gage records have also been used in the development of regional models (Lettenmaier et al., 1987; Burn, 1988). A study by Kidson and Richards (2005) suggests that it is impossible to determine which FFA tool is best for a given watershed and that a multi-disciplinary approach employing physical modeling supplemented with regional, historic, and paleoflood information may be best. Studies correlating Bulletin 17B performance with watershed or gage record characteristics seem to be lacking, but one study found that Bulletin 17B had poor performance for watersheds with negative skew values (Wallis & Wood, 1985).

Even with proposed improvements to FFA methods, uncertainty is still considerable, and the flood record at some stream gages may be considered a low probability outcome (e.g., three 100-year floods in 50 years of record where a 100-year flood is estimated by Bulletin 17B methods), with “outcome” defined as the number of flood events over the period of record. A low-probability outcome could be considered an indication that the FFA method is less accurate for a particular watershed. It could be a signal that more uncertainty exists at a gage location, or that flood risk is either greater or smaller in and around that 100-year floodplain than Bulletin 17B suggests. Conversely, it could be the result of random meteorological events. Given the extensive use of the flood frequency results for flood risk management, it would be useful to understand which gages have low probability outcomes and to identify watershed and stream gage record characteristics that are associated with probability of outcome. This would allow risk managers to identify study locations where they might want to consider more advanced flood frequency and risk analysis methods versus those where they might be more comfortable using simpler flood frequency methods. This study applies statistical learning methods to this problem to generate a model of probability of outcome versus watershed characteristics. The use of probability of outcome as a measure of flood frequency model accuracy is a novel approach.

3. METHODS AND DATA

3.1. Data

Stream gage data for this project were obtained from the United States Geological Survey (USGS) National Water Inventory System (NWIS) website (2018). Annual peak streamflow data were retrieved for the stream gages with at least 40 years of record in the states of Delaware, Maryland, Pennsylvania, West Virginia, Virginia, and North Carolina. Only stream gages with 40 years or more of non-regulated flow were included in the analysis, resulting in a total of 515 gages. The record lengths for the gages used in this analysis ranged from 40 to 123 years, with an average record length of 70 years. Fig. 1 provides a histogram of the gage record lengths. Additionally, a subset of 128 gages with 80 or more years of record was identified. These gages were used to develop train and test datasets. The train dataset consisted of peak flows from the start of the gage record up to the last 40 years of record. The test dataset consisted of the last 40 years of record for each of these gages.

FFA was performed for each stream gage using the PeakFQ software, which implements the Bulletin 17B methods. FFA was performed on the train and test dataset as well. Streamflow qualification codes were evaluated and peaks were disqualified based on the specifications in the PeakFQ manual (Flynn et al., 2006). This included peaks affected by dam failure and known effects of regulation, urbanization, or other watershed change. Adjustments were made for low outliers, while high outliers were retained without adjustment per the Bulletin 17B guidance for analysis where useful historic information is not available to adjust for high outliers (IACWD, 1982). Weighted skew values based on the station skew and generalized regional skew were used. Sources of generalized skew values for each state are presented in Table I. No historic or other adjustments (e.g. two-station comparisons) were included in order to maintain consistency with the simplest implementation of the Bulletin 17B methods.

Table I: Regional Skew Value Data

State	**Source of Regional Skew Values**
Delaware	U.S. Geological Survey (USGS) (Ries & Dillow, 2006)
Maryland	Maryland Hydrology Panel (2010)
North Carolina	USGS (Weaver et al., 2009)
Pennsylvania	US Army Corps of Engineers (USACE) Delaware River Basin (Goldman et al., 2009) and statewide (Roland & Stuckey, 2008)

Virginia	Generalized skew coefficients map in Bulletin 17B, with values generated in PeakFQ based on station location (Austin et al., 2011)
West Virginia	USGS (Wiley & Atkins, 2010)

Once the 100-year discharge was estimated for each gage, this value was compared to the annual peak discharge time-series for each gage to determine the actual number of years in the period of record in which the annual peak discharge met or exceeded the estimated 100-year discharge. This actual number of years for each gage that include a 100-year or greater discharge event is termed the "number of floods" for purposes of this study. The probability of this specific number of floods occurring considering the length of gage record was calculated and is termed the "probability of actual flood record" (PAFR). Higher values of PAFR indicate a gage that experienced a number of floods over the period of record that is more probable while lower values of PAFR represent a number of floods that is less probable. Particularly low values of PAFR indicate gages with rare outcomes. We explore these gages to understand whether these very low, potentially unexpected, values may be attributed to the stochastic nature of floods or indicates a poor fit of the FFA method for certain types of watersheds.

The PAFR for each gage was calculated using the binomial equation presented as equation 1. In this equation, n is the number of years of record for the gage, k is the number of years in which the annual peak discharge met or exceeded the 100-year discharge, p is 0.01, which is the probability of experiencing at least one 100-year or greater discharge event in any year, and X is the observed number of years that the peak annual discharge at the gage met or exceeded the 100-year discharge. In this calculation, the likelihood of a 100-year flood event occurring in any given year remains constant over the entire period of record and is independent of events occurring in other years. For example, a stream gage that had one annual peak that met or exceeded the 100-year discharge in 50 years of record would have a probability of 0.31.

$$P(X=k)=\left(\frac{n!}{k!(n-k)!}\right)p^{k}(1-p)^{(n-k)} \tag{1}$$

PAFR was plotted and evaluated geospatially to identify any potential spatial trends. A density plot of PAFR was also generated. In order to determine whether the distribution of the PAFR values for the set of gages as a whole was as should be expected given the number of stream gages and the length of record for each of the gages, a synthetic record analysis was completed. For each stream gage, a synthetic record of number of 100-year events (events that meet or exceed the 100-year discharge) was randomly generated using the actual number of years of record for each gage and a probability 0.01 of a 100-year event occurring in a given year. The probability of this record was then calculated, yielding one replication for that gage. One hundred thousand replications were performed for each gage and the set of synthetic probabilities for all gages was used to generate a synthetic distribution of probability of record. The density of this synthetic distribution was plotted along with the density of the probabilities based on the actual data set to evaluate how the actual probabilities compare to theoretical expected probabilities, given the length of record at each of the gages.

Watershed data were obtained from the USGS Geospatial Attributes of Gages Evaluating Streamflow II (2011) data set. This data set contains watershed characteristic data for USGS stream gages. Covariates were chosen to reflect commonly used watershed characteristics that could conceivably be related to either the accuracy of the FFA method, the runoff generating mechanisms for the watershed, or the meteorological conditions at the watershed location. The covariates fall into the following categories: basin identification, basin classification, basin morphology, climate, geology, hydrology, hydrologic modifications, landscape patterns, land use, population and infrastructure, soil, and topography. The full list of the 60 covariates used in the analysis is included as Appendix A.

In addition to the watershed characteristics, the mean annual peak discharge (log), standard deviation of mean annual peak discharge, and weighted skew were calculated for each stream gage. These values are critical elements of the 17B analysis, as they are used to fit the station data to the log-Pearson type III using the method of moments. Considering them was important not only in completing the 17B analysis but also in understanding whether certain ranges of these key values are more likely to result in low probability outcomes. The weighted skew is of particular interest given the role of the skew in characterizing the tails of the distribution.

3.2. Statistical Modeling

Statistical analysis was performed using the R software (R Development Core Team, 2008). For the watershed characteristics analysis, several models appropriate for the response variable, probability of outcome, constrained to the 0 to 1 interval were selected and run, including beta regression, Classification and Regression Trees (CART), and Random Forest. Random Forest is a non-parametric ensemble decision tree method. In the method, a large number of regression trees are developed, with each tree based on a bootstrapped sample of the data set. The prediction is averaged from the set of trees. Random Forest models are good for data sets with non-linear relationships, outliers, and noise (Hastie et al., 2001). Two sets of models were generated – one using $p(x=k)$ as the response variable and one using $p(x \leq k)$. For each set of models, multiple versions were run, including models with the full set of watershed characteristics as covariates, models with the full set of watershed characteristics plus the gage mean, standard deviation, and weighted skew, and models with a reduced set of watershed characteristics selected to reduce redundancies from a physical perspective.

These models were tested in two ways. First, train and test datasets based on the subset of 128 gages with at least 80 years of record were used. The model was fitted using the train dataset and tested using the test dataset. Second, holdout analysis was run on the full set of 515 gages with 50 repeated, random holdouts with a randomly selected 20% of the data held out each time. The predictive accuracy of the models was compared amongst the models and to a mean-only model where predictions were made using only the mean probability for all gages.

The models tested provided improvements in predictive accuracy as compared to a mean-only model. The models using $p(x=k)$ as the response variable provided better accuracy improvement over the mean only model than those using $p(x \leq k)$. Results presented herein reflect the $p(x=k)$ models.

3.3. Clustering

To further explore the relationship between key covariates and the PAFR, k-means clustering was performed. K-means clustering is a method to partition a data set into a specified number of non-overlapping clusters based on data values (James et al., 2013). The purpose of this analysis was to determine whether certain ranges of covariate values might be associated with low PAFR values. Based on preliminary evaluation of the most functional number of clusters for purposes of this analysis, four clusters were chosen for the PAFR k-means analysis, and the stream gages were separated into four clusters based solely on PAFR. Empirical Cumulative Distribution Function (ECDF) plots were generated for each cluster, for each of the four most important covariates from the Random Forest model and are presented alongside the partial dependence plots for each covariate.

3.4. Bootstrapped Data Analysis

In order to partially address the limitations of our study pertaining to the variation in years of record for stream gages, we generated ten bootstrapped samples of 40 years of record for each gage. Because each bootstrapped sample had exactly 40 randomly drawn years of record, we eliminated the effect of differing stream gage record lengths. This yielded ten separate bootstrapped data sets. Results from the bootstrapped analysis were compared to the results of the analysis of the full data set to determine if the same variables had high importance, and how the variable importance differed amongst the data sets.

4. RESULTS AND DISCUSSION

4.1. PAFR

To visualize how PAFR varied geographically, a map of the study area and the PAFR (p=k) for each stream gage was generated and is included as Fig. 2. The red dots on the map represent the stream gages with the lowest PAFR values. Because of the nature of precipitation and flooding events, some grouping of low PAFR stream gages was expected. However, visual analysis of Fig. 2 fails to show any spatial grouping of similar probability gages. This indicates that low PAFR values are not confined to a certain geographic portion of the study area and are not generally grouped geospatially. While watershed boundaries are not displayed on Fig. 2, review of these results along with HUC-4 watershed boundaries indicates no obvious grouping by watershed. Fig. 2 also illustrates that the lower PAFR gages include some with long record lengths (80+ years) and some with shorter record lengths.

We see from Fig. 3 that the actual and expected PAFR align reasonably well for the set of 515 stream gages that was analyzed. However, there are fewer low-PAFR gages and slightly more very high-PAFR gages than would be expected. For the gages evaluated in the Mid-Atlantic region, the 17B method does not appear to result in systematic over or under estimation of flood frequency, but there are some differences between the actual and expected set of probabilities.

While evaluation of Fig. 3 reveals no systematic over or underestimation of flood frequency, the underrepresentation of low-PAFR gages does *not* mean that these are not problematic from a risk perspective. These gages have flood risk that is either higher or lower than Bulletin 17B would suggest. Understanding if watershed and stream gage characteristics are associated with these low-PAFR estimates, and how so, is a major goal of this paper.

4.2. Statistical Model and Clustering Analysis

In order to evaluate the accuracy of the models and to choose the model with the best predictive accuracy, two types of testing were performed as described in Section 3.2. The results of the testing are presented in Table II. While additional variations of the models were run, including models with the full set of 60 watershed characteristics, only a few of the more accurate models are presented here. The "A" models include a subset of 33 watershed covariates selected to reduce physical redundancy while the "B" models include the same 33 watershed covariates plus the mean, standard deviation, and weighted skew of the gage record. In generating these models, covariates representing similar physical characteristics were removed. For instance, the average basin temperature was retained, while the maximum basin temperature was removed.

In comparing the models, Beta regression B had the lowest average mean absolute error (MAE) and average mean squared error (MSE) across the test/train analysis, while Random Forest B had the lowest MAE and MSE in the holdout testing. A comparison of the models indicated that the same three covariates were of highest importance in both Beta regression B and Random Forest B. For simplicity in presentation of results, only the Random Forest B model was utilized for further analysis.

Table II: Comparison of model predictive accuracy based on average Mean Average Error (MAE) and Mean Square Error (MSE)

		Test/train analysis		**Holdout analysis**	
Model	**Covariates included**	**Avg. MAE (std.dev.)**	**Avg. MSE (std.dev.)**	**Avg. MAE (std.dev.)**	**Avg. MSE (std.dev.)**
Beta regression A	Subset of watershed covariates	0.2136 (0.0912)	0.0539 (0.0406)	0.1199 (0.0073)	0.0235 (0.0025)
Beta regression B	Subset of watershed covariates plus gage characteristics	0.1764 (0.1023)	0.0415 (0.0405)	0.1108 (0.0047)	0.0198 (0.0019)
CART A	Subset of watershed covariates	0.2130 (0.0906)	0.0535 (0.0402)	0.1189 (0.0074)	0.0230 (0.0026)
CART B	Subset of watershed covariates plus gage characteristics	0.1968 (0.0864)	0.0461 (0.0376)	0.1146 (0.0055)	0.0209 (0.0017)
Random Forest A	Subset of watershed covariates	0.2058 (0.1006)	0.0524 (0.0446)	0.1149 (0.0073)	0.0215 (0.0023)
Random	Subset of watershed	0.1993	0.0481	0.1091	0.0190

Forest B	covariates plus gage characteristics	(0.0921)	(0.0386)	(0.0057)	(0.0016)
Mean Only	Mean of PAFRfor all stream gages in the training set used as prediction for the holdout set	0.2582 (0.1976)	0.1054 (0.1229)	0.1151 (0.0078)	0.0223 (0.0021)

To determine whether model accuracy could be improved by using a subset of the most important covariates from the selected model, recursive feature elimination was performed using the Classification and Regression Training (CARET) package in R with 200 bootstrap samples. In recursive feature elimination, backwards selection of covariates is performed based on importance ranking. Less important covariates are sequentially removed to identify the subset of predictors that provides the most accurate model. The output indicated that a reduction in covariates from the selected model would not result in a more accurate model. Thus, the selected model (Random Forest B) was used for the remainder of the analysis.

In considering Table II with regard to the question of whether watershed and stream gage characteristics are correlated with PAFR, it is apparent that there is some correlation since the model provides an improvement in fit over the mean only model. We believe that this model provides useful information about watersheds that might be at higher risk for low PAFR values. However, the accuracy of this model is limited, and it is clear that randomness plays a significant role in flood frequency outcomes. Variable importance is calculated as the percent increase in MSE resulting from permuting each covariate and recording the out-of-bag prediction error (James et al., 2013). This is a measure of the contribution of each variable to the out of sample predictive accuracy of the model. Fig. 4 shows the top fifteen most important covariates, based on the Random Forest variable importance. As shown in Fig. 4, weighted skew was the most important covariate, followed by drainage area, mean annual peak log discharge, and watershed slope. Weighted skew, drainage area, and mean annual peak log discharge were also the most important covariates in the Beta Regression B model, which reinforces the importance of these covariates. A correlation matrix for the top fifteen covariates is provided in Appendix B. While there is some correlation between certain covariates, most notably percent developed and population density, Random Forest models are generally robust to correlation in covariates.

In order to further analyze the influence of the covariates, partial dependence plots and ECDF plots were generated for each of the four covariates. Partial dependence plots show the marginal influence of a covariate on the response variable after integrating out the other covariates (James et al., 2013). In each of the partial dependence plots, the influence of the covariate changes with the covariate values.

We partitioned the response variable into four groups using k-means clustering and then plotted the ECDFs for each of the four groups in order to better understand how different or similar the covariates are across different ranges of the response variable. The objective was to identify differences in covariate values that are associated with low PAFR values. The four clusters are described in Table III, including the value of the center (centroid or median) PAFR of the cluster and the range of PAFR in the cluster. The partial dependence and ECDF plots for each of the six covariates are shown in Figs. 5 through 8. The results of these ECDFs and the partial dependence plots are discussed for each of the top four covariates. The partial dependence plots indicate the marginal influence of the covariate on the response variable (probability). For our top covariate, the partial dependence is in the range of 0.31 to 0.46. Covariates of lesser importance have a narrower partial dependence range. The ECDF plots show the distribution of the covariate values for each of the four clusters described above.

Table III: k-means clustering

	Number of stations	Cluster PAFR center	Cluster PAFR range
Cluster 1	83	0.14	0.03-0.22
Cluster 2	112	0.31	0.27-0.35
Cluster 3	180	0.38	0.35-0.48
Cluster 4	140	0.57	0.48-0.67

4.2.1. Weighted Skew

The skew of the stream gage record (station skew) represents the asymmetry of the values about the mean and the regional skew represents an average skew value for gages within a geographic area. The weighted skew is a weighted average of the station and regional skew. Weighted skew is included as an input in the 17B FFA method. The partial dependence plot in Fig. 5 shows that predicted probability tends to decrease with increasing weighted skew. In the ECDF plot, the stream gages in the lowest PAFR cluster (cluster 1) tend to have higher weighted skew values than the stream gages in the other clusters. Weighted skew is used in the Bulletin 17B method to fit the stream gage record to the log Pearson type III distribution. A higher skew value would indicate that the shape of the distribution is wider on the right side than on the left side, that is, it is right-skewed. This indicates that gages with a greater number of annual peak discharge values at the high end of the distribution are more likely to have low PAFR values, which is likely due to a poor fit with the thin-tailed log Pearson type III distribution.

4.2.2. Drainage Area

Drainage area is defined as the watershed area that drains to the stream gage location, and in our study has units of square kilometers. The partial dependence plot in Fig. 6 indicates that lower probability predictions generally tend to have higher drainage areas. The partial dependence plot is flat for very high values of drainage area where there are very few data points. The ECDF plot shows that stream gages in the lowest PAFR cluster generally tend to have larger drainage areas. This may be due to greater spatial variation in storm events and resulting runoff generation in larger watersheds than in smaller watersheds where precipitation events are more likely to impact the entire watershed concurrently and land use may be more consistent. Another possible reason for this outcome is that large drainage basins in the Mid-Atlantic region tend to experience extremely high/tail events due to extreme storms like tropical cyclones. Smaller watersheds may experience high discharge events due to different types of storms, including thunderstorms, extra-tropical cyclones, and tropical systems. A relatively high discharge event can occur more frequently in the smaller watersheds (Gamble, 1997), with the statistical characteristics of the stream gage record (mean, skew, standard deviation), the 100-year discharge estimate, and the PAFR all reflecting this difference.

4.2.3. Mean Annual Peak Discharge

The mean of the log of the stream gage annual peak discharge serves as an indicator of the magnitude of the annual peak discharge time series at each gage. The partial dependence plot in Fig. 7 shows that the lowest probability predictions tend to have higher mean annual peak discharge values, generally above 3.5. The ECDF plot also indicates that the lowest PAFR cluster tends to have slightly higher mean annual peak discharge values than the other clusters. This result is consistent with the Drainage Area result. While other watershed characteristics influence flow generation, watersheds with higher mean annual peak discharge would generally tend to come from watersheds with larger drainage areas.

4.2.4. Mean Watershed Slope

This covariate represents the mean watershed slope as a percent, which is a key factor in driving watershed runoff. Watersheds with a higher mean slope will potentially generate higher peak runoff values than similar watersheds with a lower mean slope. The partial dependence plot in Fig. 8 indicates that watershed slope values in the 3 to 13 percent range correspond to higher PAFR. As watershed slope increases in the 13 to 25 percent range, PAFR decreases. The ECDF plot shows that clusters 1 and 2 tend to have steeper watershed slopes than the higher PAFR clusters. This indicates that watersheds with steeper average slopes may be more prone to low PAFR values.

4.2.5. Comparison

In reviewing the partial dependence plots, the marginal influence of the weighted skew covariate spans a range of 0.15. The marginal influence for each the other covariates is less, spanning ranges of about 0.06 for drainage area and 0.03 for mean annual peak discharge, the next most important covariates, to 0.007 for mean watershed slope. While the marginal influence of these covariates is somewhat small, summing these influences could result in significant influence on the response variable. Higher weighted skew values, larger drainage areas, higher mean annual peak discharge, and higher mean watershed slope are watershed characteristics associated with lower PAFR. The ECDF plots corroborate the findings of the partial dependence plots. This points to the conclusion that the accuracy of standard FFA results may not be equivalent for all watersheds.

4.3. Bootstrapped Data Analysis

Using the covariates included in the Random Forest B model, a Random Forest model was created for each of the ten bootstrapped samples. The purpose of this analysis was to determine whether the same covariates had high variable importance in randomly selected sets of equal record length. Because each bootstrapped gage data set contained exactly 40 years of record, only a limited number of discrete PAFR values were possible. Therefore, the response variable was treated as categorical for this analysis. In each bootstrapped model, weighted skew was the most important covariate. The importance of the other covariates varied. In addition to the four covariates of higher importance in our original model, covariates with high importance in the models for some of the bootstrapped data sets included fragmentation index, standard deviation of gage record, mean watershed elevation, watershed percent forested, and watershed percent developed.

Additionally, the range of variable importance for each covariate in the bootstrapped models was evaluated, and boxplots of the relative importance of the covariates are displayed on Fig. 9. Because the magnitude of variable importance values differs for each Random Forest run, the variable importance is plotted as a percent of total variable importance, so that the different runs can be compared (Tonn et al., 2016). Review of Fig. 9 indicates that the variable importance values for weighted skew and mean watershed slope in the original model are near the median percent importance for these covariates in the bootstrapped analysis. Variable importance for drainage area and mean annual peak discharge for the original model are at the upper end of the range of importance for these covariates in the bootstrapped analysis. The bootstrapped analysis reinforced the finding of the importance of the top four covariates from the original analysis.

5. CONCLUSIONS

The synthetic streamflow record analysis indicates that for the Mid-Atlantic region as a whole, the Bulletin 17B method is generally producing results in line with expectations. While the results generally match expectations, the method indicates fewer low PAFR and slightly more very high PAFR gages than should theoretically be expected. Given the extensive use of FFA results for flood risk management, it would be useful to be able to identify stream gages that are likely to have low PAFR values when judged relative to a Bulletin 17B analysis and to identify watershed characteristics that may be correlated with PAFR. This would allow risk managers to identify stream gages where they might want to consider more advanced flood frequency and flood risk analysis methods versus those where they might be more comfortable using basic FFA results. This study is an effort to apply statistical learning methods to this problem to generate a model of PAFR versus watershed characteristics.

Choosing a response variable for this analysis was challenging, and PAFR was selected as the most feasible option. Using PAFR as a response variable allows for stream gages with differing record lengths to be analyzed as a set. It provides a value for analysis that gives an indication of the likelihood or expectedness of a discharge estimate. However, there are limitations associated with the use of this response variable. The definition of the response variable is somewhat convoluted, and the value does not give an indication of whether a low or high probability value is due to an excess or a deficit of 100-year events. Other potential response variables, such as ratio of actual to expected years with 100-year events, or deviation from the expected number of years with 100-year events, have limitations associated with disparate periods of record.

Beta regression, CART, and Random Forest models with different covariate selections were compared, and a Random Forest model was selected for further analysis. Variable importance and partial dependence plots were generated and analyzed to interpret model output. Clustered data analysis was performed to further analyze the relationship between the covariates and probability of outcome. Covariates that are associated with lower PAFR in the Mid-Atlantic region included higher weighted skew, larger drainage area, higher mean annual peak discharge, and higher watershed slope. The clustering analysis reinforced the findings of the Random Forest model, and showed that cumulatively, gages with a low probability of outcome had values for several covariates that were generally higher or lower than most other gages. In both the original analysis and an analysis of ten bootstrapped data sets, weighted skew was the most important covariate. In evaluating the ECDF plot for weighted skew, there was clear separation in skew values for the lowest probability cluster as compared to the other clusters. Higher weighted skew values are clearly correlated with lower PAFR values, which makes sense given that gages with high skew values have distributions that are right-skewed and given the significance of the skew value in the Bulletin 17B calculations.

The results of this study identify the covariates that are most important in modeling PAFR at stream gages in the Mid-Atlantic region. The key finding is that certain watershed characteristics are correlated with PAFR, indicating that the results of standard FFA may not be equivalent across differing watersheds. Analysts may want to consider enhanced flood frequency methods for watersheds with these characteristics. These results can be used in evaluating floodplain maps generated using the Bulletin 17B methods, such as FEMA FIRMs. Watershed characteristics could be compared to those found to be important in this study to determine if a watershed area is more likely to experience an unexpected outcome (i.e. the floodplain map is less reliable).

While our model provided an improvement in predictive accuracy over the other evaluated models including a mean-only model, the model accuracy is limited. Flood frequency is highly dependent on random weather events and other meteorological conditions that could not be captured by this study. This study was limited to stream gages in the Mid-Atlantic region of the U.S., and the findings may not apply to other regions. The study included only the 100-year return period, and results might differ for other return periods. This study focused on the Bulletin 17B flood frequency method, and the new 17C method (England et al., 2019) could be evaluated in a similar manner. The main methodological improvements associated with 17C apply to low outliers, historic and paleoflood data, regional skew approximation, and confidence interval calculation. Any changes to our study results associated with a switch from 17B to 17C would likely be most pronounced for Virginia gages, for which we used the Bulletin 17B generalized skew coefficients map. Furthermore, some of the covariates included in our study such as the percent forested and percent developed land are subject to change over time, but are modeled as stationary values due to data limitations.

Acknowledgements:

This work was supported by a fellowship to Gina Tonn by the Environment, Energy, Sustainability and Health Institute (E2SHI) and the NSF IGERT Water, Climate, and Health program at Johns

Hopkins University. The views in this paper are those of the authors and do not necessarily reflect those of the sponsors.

APPENDIX A: LIST OF MODEL COVARIATES

Covariate	**Abbreviation**	**Units**	**Included in RFB**
Drainage Area	DRAIN_SQKM	square kilometers	Yes
Hydrologic Disturbance Index	HYDRO_DISTURB_INDX	unitless	Yes
Watershed compactness ratio	BAS_COMPACTNESS	unitless	Yes
Mean annual Precipitation for the basin	PPTAVG_BASIN	centimeters	Yes
Average annual air temperature for the watershed	T_AVG_BASIN	degrees C	Yes
Average monthly maximum air temperature	T_MAX_BASIN	degrees C	No
Watershed average number of days of measurable precipitation (based on 30 year average)	WD_BASIN	days	Yes
Site average number of days of measurable precipitation (based on 30 year average)	WD_SITE	days	No
Watershed average of monthly maximum number of days of measureable precipitation	WDMAX_BASIN	days	No
Watershed average of monthly minimum number of days of measureable precipitation	WDMIN_BASIN	days	No
Site average of monthly maximum number of days of measureable precipitation	WDMAX_SITE	days	No
Site average of monthly minimum number of days of measureable precipitation	WDMIN_SITE	days	No
Maximum Strahler stream order in watershed	STRAHLER_MAX	unitless	Yes
Sinuosity of main stream line	MAINSTEM_SINUOUSITY	unitless	Yes
Percent of mainstem stream(s) coded as artificial path in NHDPlus	ARTIFPATH_MAINSTEM_PCT	Percent	Yes
Percent of watershed area covered by lakes/ponds and reservoirs	HIRES_LENTIC_PCT	percent	Yes
Base flow index	BFI_AVE	percent	Yes
Dunne overland flow	PERDUN	Percentage of total streamflow	Yes
Horton overland flow	PERHOR	Percentage of total streamflow	Yes
Topographic wetness index	TOPWET	ln(meters)	Yes

Estimated average annual watershed runoff	RUNAVE7100	mm/year	Yes
Percent of watershed stream lengths which are first order streams	PCT_1ST_ORDER	percent	Yes
Percent of watershed stream lengths which are second order streams	PCT_2ND_ORDER	percent	Yes
Dam density (2009)	DDENS_2009	number of dams/100 km sq	Yes
Major dam density (2009)	MAJ_DDENS_2009	Number of major dams/100 km sq	No
Fragmentation Index of undeveloped land in the watershed	FRAGUN_BASIN	unitless	Yes
Watershed percent developed, 2006	DEVNLCD06	Percent	Yes
Watershed percent forest, 2006	FORESTNLCD06	Percent	Yes
Watershed percent agriculture, 2006	PLANTNLCD06	Percent	Yes
Watershed percent open water, 2006	WATERNLCD06	Percent	Yes
Mainstem 100 m buffer developed	MAINS100_DEV	Percent	No
Mainstem 100 m buffer forest area	MAINS100_FOREST	Percent	No
Mainstem 100 m buffer planted/cultivated (agricultural) area	MAINS100_PLANT	Percent	No
Mainstem 100 m buffer open water area	MAINS100_11	Percent	No
Mainstem 800 m buffer developed area	MAINS800_DEV	Percent	No
Mainstem 800 m buffer forest area	MAINS800_FOREST	Percent	No
Mainstem 800 m buffer agricultural area	MAINS800_PLANT	Percent	No
Mainstem 800 m buffer open water area	MAINS800_11	Percent	No
Riparian 100 m buffer developed area	RIP100_DEV	Percent	No
Riparian 100 m buffer forested area	RIP100_FOREST	Percent	No
Riparian 100 m buffer agricultural area	RIP100_PLANT	Percent	No
Riparian 100 m buffer open water area	RIP100_11	Percent	No
Riparian 800 m buffer developed area	RIP800_DEV	Percent	No
Riparian 800 m buffer forested area	RIP800_FOREST	Percent	No
Riparian 800 m buffer agricultural area	RIP800_PLANT	Percent	No
Riparian 800 m buffer open water area	RIP800_11	Percent	No
Population density in the watershed (2000)	PDEN_2000_BLOCK	Persons/sq km	Yes
Road density	ROADS_KM_SQ_KM	km/sq km	No
Number of road/stream intersections	RD_STR_INTERS	Number of intersections/km of stream length	Yes
Watershed percent impervious	IMPNLCD06	Percent	Yes
Percentage of soils in hydrologic group A	HGA	Percent	Yes
Percentage of soils in group A/D	HGAD	Percent	No
Percentage of soils in group D	HGD	Percent	Yes
Percentage of soils in group C/D	HGCD	Percent	No
Mean watershed elevation	ELEV_MEAN_M_BASI	Meters	No

	N		
Elevation at gage location	ELEV_SITE_M	Meters	Yes
Elevation-relief ratio	RRMEAN	unitless	Yes
Mean watershed slope	SLOPE_PCT	percent	Yes
Aspect northness (range -1 to 1 with 1 meaning watershed faces/drains due north and -1 means due south)	ASPECT_NORTHNESS	Unitless	Yes
Aspect eastness (range -1 to 1 with 1 meaning watershed faces/drains due east and -1 means due west)	ASPECT_EASTNESS	unitless	Yes

APPENDIX B: MATRIX OF CORRELATION*

	GageSkew	DRAIN_SQKM	GageMean	SLOPE_PCT	STRAHLER_MAX	T_AVG_BASIN	PDEN_2000_BLOCK	RUNAVE7100	GageSD	RRMEAN	PLANTNLCD06	DEVNLCD06	FORESTNLCD06	WD_BASIN	WATERNLCD06
GageSkew	1.000	0.004	-0.058	-0.108	-0.006	-0.165	0.042	0.013	0.153	0.068	0.070	0.029	-0.061	-0.034	-0.021
DRAIN_SQKM	0.004	1.000	0.577	0.050	0.166	-0.124	-0.075	-0.044	-0.265	-0.092	0.012	-0.091	0.064	0.035	0.202
GageMean	-0.058	0.577	1.000	0.338	0.325	-0.211	-0.164	0.081	-0.524	-0.318	-0.139	-0.192	0.317	0.224	0.221
SLOPE_PCT	-0.108	0.050	0.338	1.000	0.099	-0.428	-0.275	0.417	-0.230	-0.216	-0.513	-0.347	0.768	0.565	-0.225
STRAHLER_MAX	-0.006	0.166	0.325	0.099	1.000	-0.081	-0.050	0.030	-0.085	-0.091	0.020	-0.044	0.105	0.037	0.119

T_AVG_ BASIN	-0.165	-0.124	-0.211	-0.428	-0.081	1.000	0.133	-0.535	0.302	-0.077	0.204	0.193	-0.442	-0.682	-0.001
PDEN_2000_BLOCK	0.042	-0.075	-0.164	-0.275	-0.050	0.133	1.000	-0.116	0.055	0.137	-0.121	0.894	-0.477	-0.205	-0.077
RUNAVE 7100	0.013	-0.044	0.081	0.417	0.030	-0.535	-0.116	1.000	-0.399	0.008	-0.296	-0.120	0.363	0.696	-0.024
GageSD	0.153	-0.265	-0.524	-0.230	-0.085	0.302	0.055	-0.399	1.000	0.053	0.231	0.073	-0.239	-0.524	-0.165
RRMEAN	0.068	-0.092	-0.318	-0.216	-0.091	-0.077	0.137	0.008	0.053	1.000	-0.119	0.129	-0.093	0.117	-0.011
PLANT NLCD06	0.070	0.012	-0.139	-0.513	0.020	0.204	-0.121	-0.296	0.231	-0.119	1.000	-0.078	-0.687	-0.409	0.003
DEV NLCD06	0.029	-0.091	-0.192	-0.347	-0.044	0.193	0.894	-0.120	0.073	0.129	-0.078	1.000	-0.575	-0.256	-0.024
FOREST NLCD06	-0.061	0.064	0.317	0.768	0.105	-0.442	-0.477	0.363	-0.239	-0.093	-0.687	-0.575	1.000	0.556	-0.059
WD_ BASIN	-0.034	0.035	0.224	0.565	0.037	-0.682	-0.205	0.696	-0.524	0.117	-0.409	-0.256	0.556	1.000	-0.034
WATER NLCD06	-0.021	0.202	0.221	-0.225	0.119	-0.001	-0.077	-0.024	-0.165	-0.011	0.003	-0.024	-0.059	-0.034	1.000

*Refer to Appendix A for descriptions associated with covariate abbreviations

REFERENCES

Ahmadisharaf, E., Tajrishy, M., & Alamdari, N. (2016). Integrating flood hazard into site selection of detention basins using spatial multi-criteria decision-making, *Journal of Environmental Planning and Management*, 59:8, 1397-1417, DOI: 10.1080/09640568.2015.1077104

Ahmadisharaf, E., Kalyanapu, A., Lillywhite, J., & Tonn, G. (2018). A probabilistic framework to evaluate the uncertainty of design hydrograph case study of Swannanoa River watershed. Hydrological Sciences Journal 63(12), 1776–1790.

Austin, S.H., Krstolic, J.L., & Wiegand, U. (2011). Peak-Flow Characteristics of Virginia Streams. U.S. Geological Survey Scientific Investigation Report 2011-5144. 106 p. http://pubs.usgs.gov/sir/2011/5144/.

Benson, M.A. (1962a). Evolution of methods for evaluating the occurrence of floods (No. 1580-A). USGPO.

Benson, M.A. (1962b). Factors influencing the occurrence of floods in a humid region of diverse terrain (No. 1580-B). USGPO.

Brunner, M.I., Seibert, J., & Favre, A.C. (2016). Bivariate return periods and their importance for flood peak and volume estimation. *Wiley Interdisciplinary Reviews: Water*, *3*(6), 819-833.

Brunner, M.I., Viviroli, D., Sikorska, A.E., Vannier, O., Favre, A.C., & Seibert, J. (2017). Flood type specific construction of synthetic design hydrographs. *Water Resources Research*, *53*(2), 1390-1406.

Brunner, M.I., Seibert, J., & Favre, A.C. (2018a). Representative sets of design hydrographs for ungauged catchments: A regional approach using probabilistic region memberships. *Advances in water resources*, *112*, 235-244.

Brunner, M.I., Furrer, R., Sikorska, A.E., Viviroli, D., Seibert, J., & Favre, A.C. (2018b). Synthetic design hydrographs for ungauged catchments: a comparison of regionalization methods. *Stochastic Environmental Research and Risk Assessment*, *32*(7), 1993-2023.

Burn, D.H. (1988). Delineation of groups for regional flood frequency analysis. *Journal of Hydrology*, 104(1), 345-361. DOI:10.1016/0022-1694(88)90174-6.

Christian, J., Duenas-Osorio, L., Teague, A., Fang, Z., & Bedient, P. (2013). Uncertainty in floodplain delineation: expression of flood hazard and risk in a Gulf Coast watershed. *Hydrological Processes*, *27*(19), 2774-2784. DOI: 10.1002/hyp.9360.

Clines, F.X. (2002). 100-year Flood for the Second Straight Year. New York Times [internet] 2002 May 9. [cited 2016 Sept 22] Available from http://www.nytimes.com/2002/05/09/us/100-year-flood-for-the-second-straight-year.html

England, J.F., Cohn, T.A., Faber, B.A., Stedinger, J.R., Thomas Jr, W.O., Veilleux, A.G., Kiang, J.E., & Mason Jr, R. R. (2019). *Guidelines for determining flood flow frequency—Bulletin 17C, Version 1.1*. US Geological Survey.

Flynn, K.M., Kirby, W.H., & Hummel, P.R. (2006). User's Manual for Program PeakFQ, Annual Flood-Frequency Analysis Using Bulletin 17B Guidelines. USGS.

Galloway, G.E. (2011). If Stationarity Is Dead, What Do We Do Now? *Journal of the American Water Resources Association,* 47(3): 563–570. DOI: 10.1111/j.1752-1688.2011.00550.x.

Gamble, D.W. (1997). The relationship between drainage basin area and annual peak-flood seasonality in the southeastern United States. Southeastern Geographer, 37(1), 61-75.

Gräler, B., van den Berg, M., Vandenberghe, S., Petroselli, A., Grimaldi, S., De Baets, B., & Verhoest, N. (2013). Multivariate return periods in hydrology: a critical and practical review focusing on synthetic design hydrograph estimation. *Hydrology and Earth System Sciences*, *17*(4), 1281-1296.

Goldman, D., Konieczki, M., & Faber, B. (2009). Delaware River Basin Regional Skew Analysis. US Army Corps of Engineers Project Report 70 (PR-70).

Hastie, T, Tibshirani, R, & Friedman, J. (2001). The Elements of Statistical Learning; Data Mining. Inference and Prediction, 1st ed., New York: Springer. ISBN-13: 978-0387848570.

Interagency Advisory Committee on Water Data. (1982). Guidelines for Determining Flood Flow Frequency, Bulletin #17B of the Hydrology Subcommittee.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. Springer, New York. ISBN-13: 978- 1461471370. http://pubs.usgs.gov/sir/2008/5102/pdf/sir2008-5102.pdf.

Kidson, R. & Richards, K.S. (2005). Flood frequency analysis: assumptions and alternatives. *Progress in Physical Geography*, *29*(3), 392-410. DOI: 10.1191/0309133305pp454ra.

Kirby, W.H. & Moss, M.E. (1987). Summary of flood-frequency analysis in the United States. *Journal of Hydrology*, *96*(1), 5-14. DOI: 10.1016/0022-1694(87)90139-9

Kollat, J.B., Kasprzyk, J.R., Thomas Jr., W.O., Miller, A.C., & Divoky, D. (2012). Estimating the impacts of climate change and population growth on flood discharges in the United States. *Journal of Water Resources Planning and Management*, 138(5), 442-452. DOI: 10.1061/(ASCE)WR.1943-5452.0000233.

Lettenmaier, D.P., Wallis, J.R., & Wood, E.F. (1987). Effect of regional heterogeneity on flood frequency estimation. *Water Resources Research*, *23*(2), pp.313-323. DOI: 10.1029/WR023i002p00313.

Lins, H.F., & Cohn, T.A. (2011). Stationarity: Wanted dead or alive? *Journal of the American Water Resources Association*, 47(3), 475-480. DOI: 10.1111/j.1752-1688.2011.00542.x

Maryland Hydrology Panel. (2010). *Applications of Hydrologic Methods in Maryland*.

Merz, R., & Bloschl, G. (2008). Flood Frequency Hydrology: 2. Combining Data Evidence. *Water Resources Research*, 44.8. DOI:10.1029/2007WR006745.

Merz, B., & Thieken, A.H. (2005). Separating Natural and Epistemic Uncertainty in Flood Frequency Analysis. *Journal of Hydrology,* 309.1-4, 114–132. DOI:10.1016/j.jhydrol.2004.11.015.

Morss, R.E., Wilhelmi, O.V., Downton, M.W., & Gruntfest, E. (2005). Flood Risk, Uncertainty, and Scientific Information for Decision Making: Lessons from an Interdisciplinary Project. *Bulletin of the American Meteorological Society*, 86.11, 1593–1601.

Naz, B.S., Kao, S.C., Ashfaq, M., Gao, H., Rastogi, D., & Gangrade, S. (2018). Effects of climate change on streamflow extremes and implications for reservoir inflow in the United States. *Journal of Hydrology*, 556, 359-370.

Ouarda, T., & El Adlouni, S. (2011). Bayesian Nonstationary Frequency Analysis of Hydrological Variables. *Journal of the American Water Resources Association*, 47(3), 496-505. DOI: 10.1111/j.1752- 1688.2011.00544.x

Pandey, G.R., & Nguyen, V.T.V. (1999). A comparative study of regression based methods in regional flood frequency analysis. *Journal of Hydrology*, *225*(1), 92-101.

Paquet, E., Garavaglia, F., Garçon, R., & Gailhard, J. (2013). The SCHADEX method: A semi-continuous rainfall–runoff simulation for extreme flood estimation. *Journal of Hydrology*, 495, 23-37.

Payrastre, O., Gaume, E., & Andrieu, H. (2011). Usefulness of historical information for flood frequency analyses: Developments based on a case study. *Water Resources Research*, 47(8). DOI: 10.1029/2010WR009812.

R Development Core Team. (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Ries, K.G., & Dillow, J. (2006). Magnitude and frequency of floods on nontidal streams in Delaware. U.S. Geological Survey Scientific Investigations Report 2006-5146. 59 p. http://pubs.usgs.gov/sir/2006/5146/pdf/sir2006-5146.pdf.

Rogger, M., Kohl, B., Pirkl, H., Viglione, A., Komma, J., Kirnbauer, R., Merz, R., & Blöschl, G. (2012). Runoff models and flood frequency statistics for design flood estimation in Austria–Do they tell a consistent story? *Journal of Hydrology*, *456*, 30-43.

Roland, M.A., & Stuckey, M.H. (2008). Regression Equations for Estimating Flood Flows at Selected Recurrence Intervals for Ungaged Streams in Pennsylvania. U.S. Geological Survey Scientific Investigations Report 2008-5102.

Rosner, A., Vogel, R. M., & Kirshen, P. H. (2014). A risk-based approach to flood management decisions in a nonstationary world. *Water Resources Research*, 50(3), 1928-1942.

Roylance, F.D. (1996). Potomac '100-year flood' hits twice in eight months. Baltimore Sun [internet] 1996 Sept 11. Accessed at: http://articles.baltimoresun.com/1996-09-11/news/1996255144_1_potomac-hurricane-fran-flow on September 22, 2016.

Stedinger, J.R. (2008). Flood Frequency Analysis in the United States: Time to Update. *Journal of Hydrologic Engineering*, 13(4), 199–204. DOI: 10.1061/(ASCE)1084-0699(2008)13:4(199).

Stedinger, J.R., & Griffis, V.W. (2011). Getting From Here to Where? Flood Frequency Analysis and Climate. *Journal of the American Water Resources Association*, 47.3, 506–513. DOI: 10.1111/j.1752-1688.2011.00545.x

Tonn, G.L., Guikema, S.D., Ferreira, C.M., & Quiring, S.M. (2016). Hurricane Isaac: a longitudinal analysis of storm characteristics and power outage risk. *Risk Analysis* 36(1), 1936-1947. DOI:10.1111/risa.12552

USACE (1996). Risk-Based Analysis for Flood Damage Reduction Studies - Manual No. 1110-2-1619.

USGS (2014). Geospatial Attributes of Gages for Evaluating Streamflow (GAGES) II data set accessed at: http://water.usgs.gov/lookup/getspatial?gagesII_Sept2011, on November 19, 2014.

USGS (2016). Floods: Recurrence Intervals and 100-Year Floods. Accessed at: http://water.usgs.gov/edu/100yearflood.html.

USGS (2018). National Water Inventory System (NWIS) accessed at: http://nwis.waterdata..usgs.gov/nwis on June 13, 2018.

Villarini, G., & Smith, J.A. (2010). Flood peak distributions for the eastern United States. *Water Resources Research*, 46(6).

Villarini, G., Smith, J.A., Baeck, M.L., & Krajewski, W.F. (2011). Examining Flood Frequency Distributions in the Midwest US. *Journal of the American Water Resources Association*, 47(3), 447-463. DOI: 10.1029/2009WR008395.

Wallis, J.R., & Wood, E.F. (1985). Relative accuracy of log Pearson III procedures. *Journal of Hydraulic Engineering*, *111*(7), 1043-1056. DOI: 10.1061/(ASCE)0733-9429(1985)111:7(1043).

Weaver, J., Feaster, T.D., & Gotvald, A.J. (2009). Magnitude and frequency of rural floods in the Southeastern United States, through 2006: Volume 2, North Carolina. U.S. Geological Survey Scientific Investigations Report 2009-5158.

Wiley, J.B., & Atkins, J.T. (2010). Estimation of Flood-Frequency Discharges for Rural, Unregulated Streams in West Virginia. U.S. Geological Survey Scientific Investigations Report 2010-5033. http://pubs.usgs.gov/sir/2010/5033/pdf/sir2010-5033.pdf.

Wiltshire, S.E. (1985). Grouping basins for regional flood frequency analysis. *Hydrological Sciences Journal*, 30(1): 151-159. DOI:10.1080/02626668509490976.

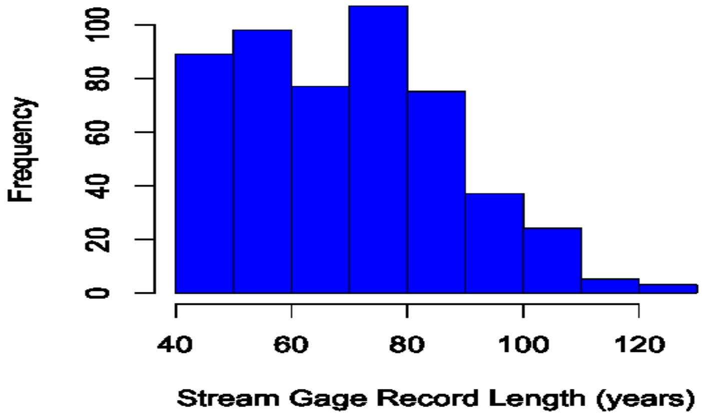

Fig. 1: Histogram of stream gage record length

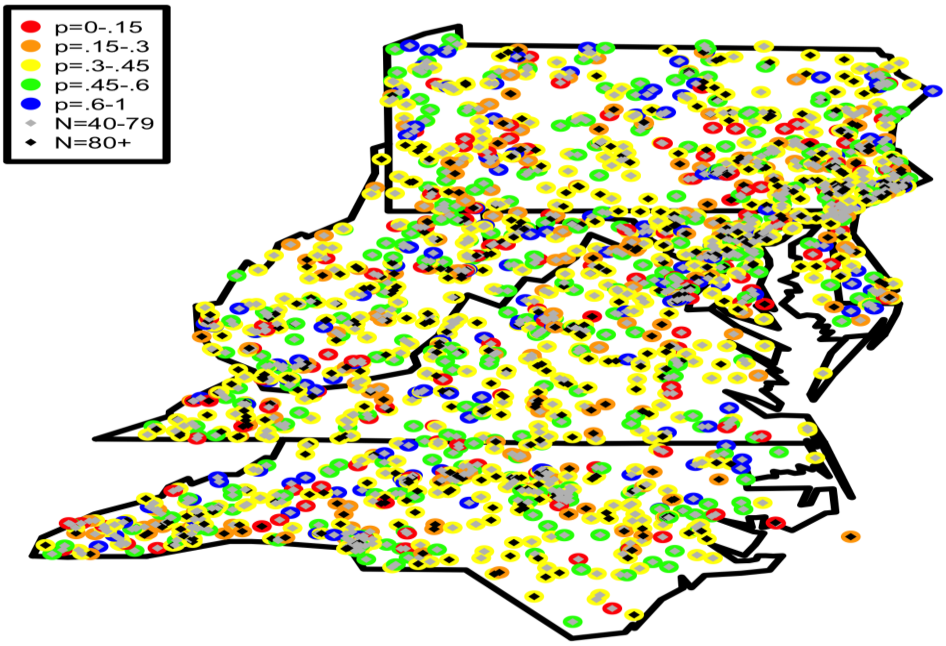

Fig. 2: Map of PAFR (p) and record length (N) in the stream gages analyzed

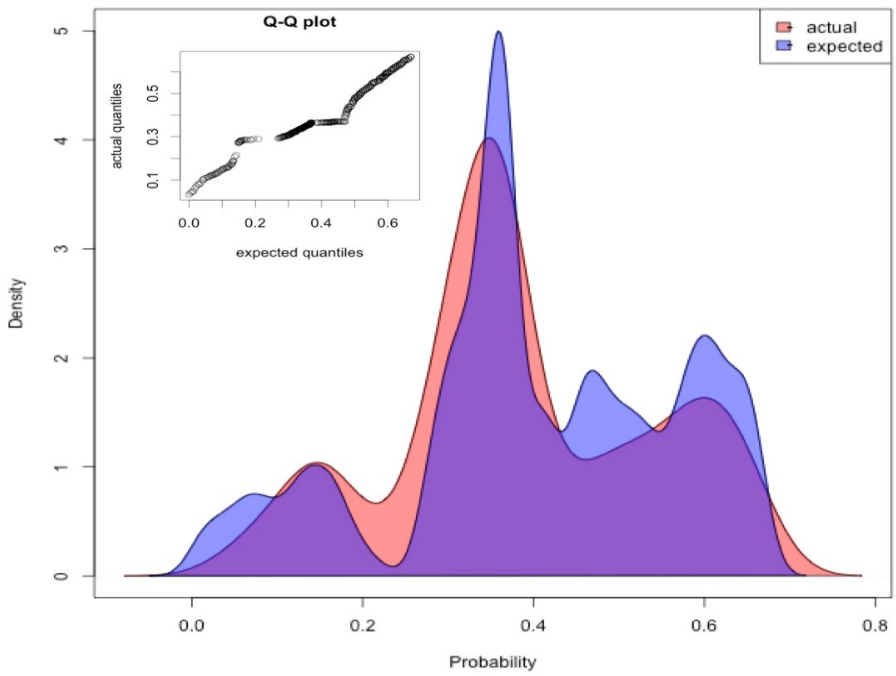

Fig. 3: Density of PAFR and Q-Q plot (inset). Red represents density of PAFR for the 515 studied stream gages. Blue represents density of PAFR from the synthetic probability analysis.

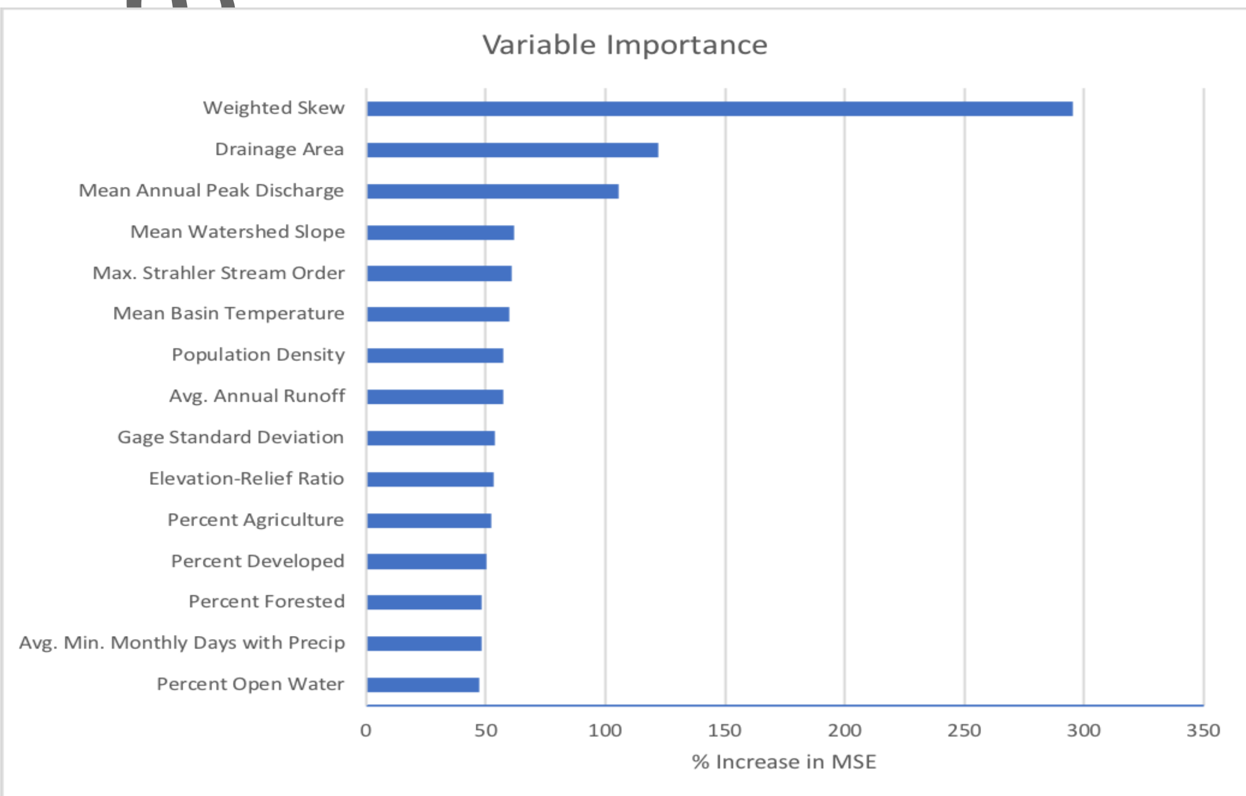

Fig. 4: Random Forest Variable Importance

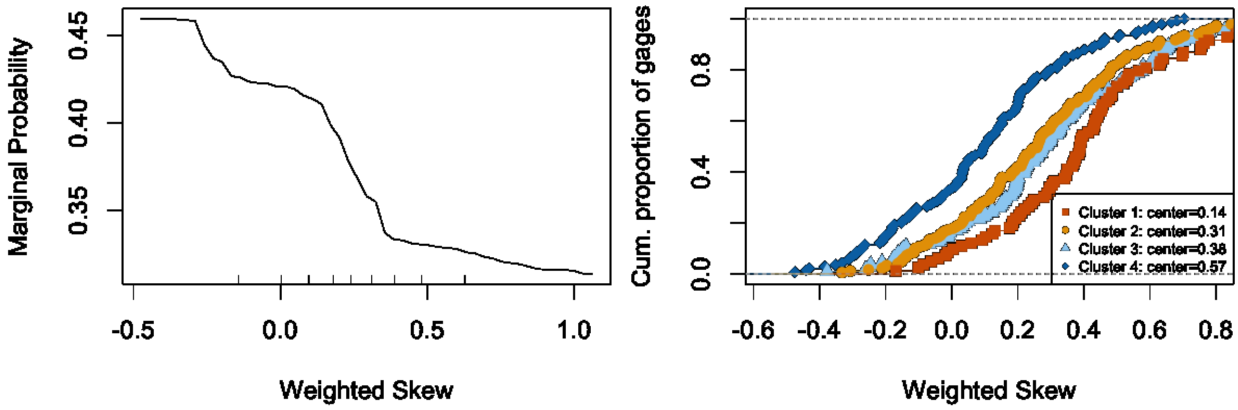

Fig. 5: Partial dependence and Empirical Cumulative Distribution Function (ECDF) plots for Weighted Skew

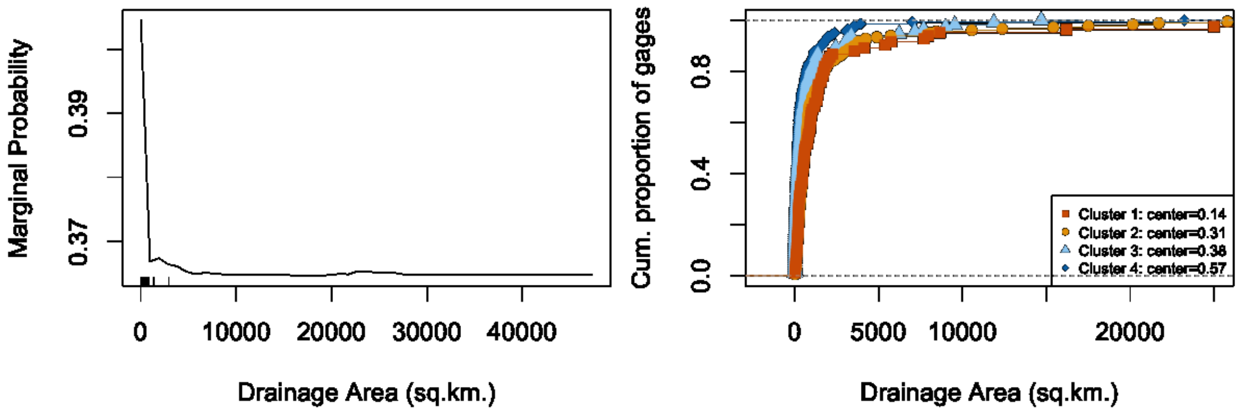

Fig. 6: Partial Dependence and Empirical Cumulative Distribution Function (ECDF) plots for Drainage Area

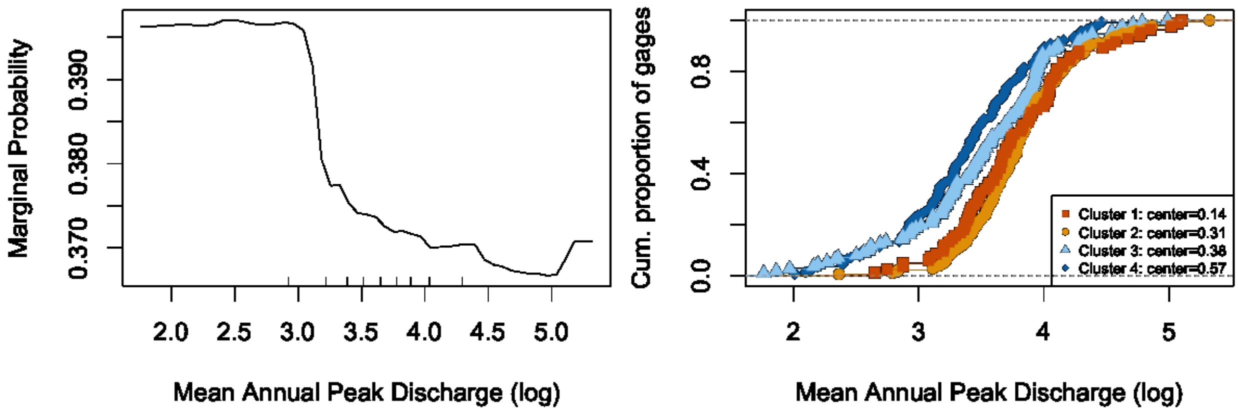

Fig. 7: Partial Dependence and Empirical Cumulative Distribution Function (ECDF) plots for Mean Annual Peak Discharge

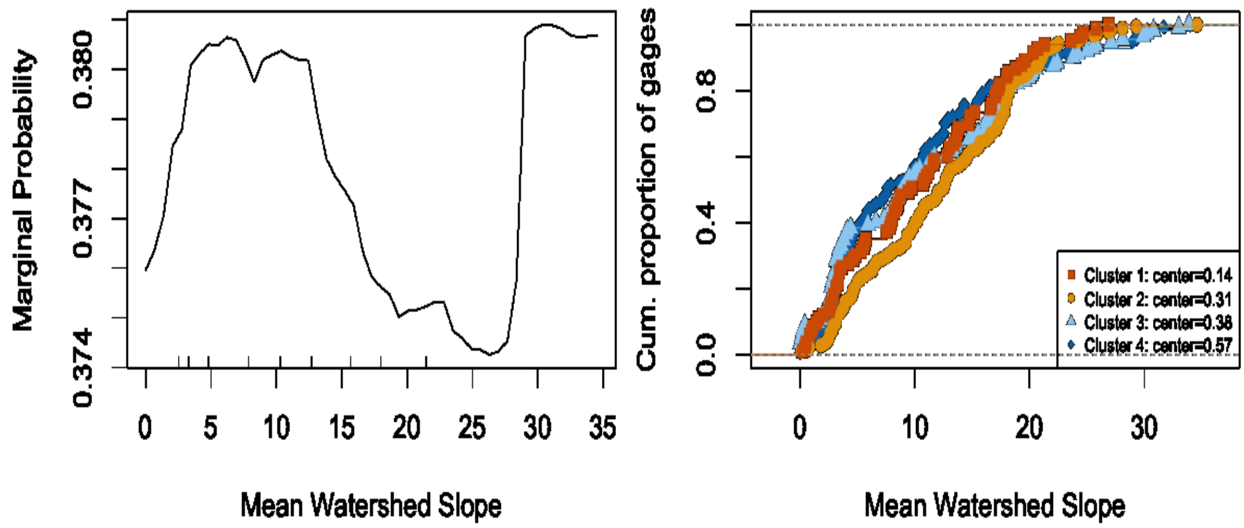

Fig. 8: Partial Dependence and Empirical Cumulative Distribution Function (ECDF) plots for Mean Watershed Slope

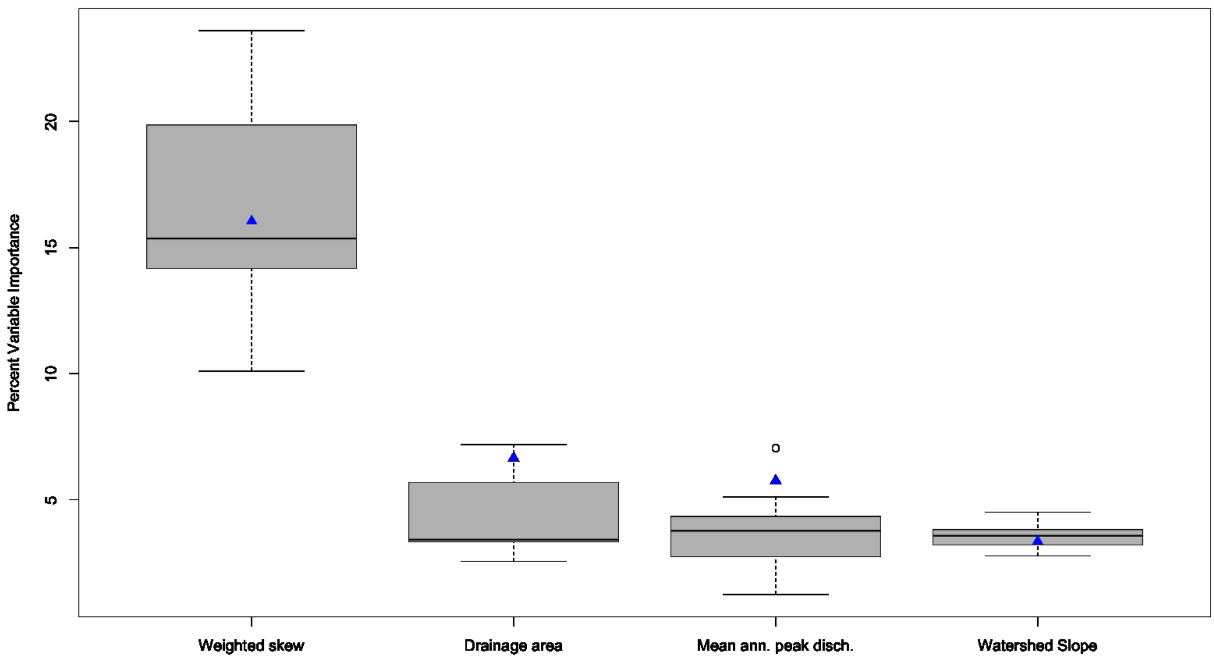

Fig. 9: Percent variable importance for bootstrapped data analysis (box plots) and percent variable importance from original model (triangular points)