Title

Independent variable selection for regression modeling of the flow duration curve for ungauged basins in the United States

Authors

Fouad, Geoffrey

Loáiciga, Hugo A

Publication Date

2020-08-01

DOI

10.1016/j.jhydrol.2020.124975

Peer reviewed

Journal of Hydrology 587 (2020) 124975

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

Independent variable selection for regression modeling of the flow duration curve for ungauged basins in the United States

Geoffrey Fouad[a,*], Hugo A. Loáiciga[b]

[a] Geographic Information Systems Program, Monmouth University, West Long Branch, NJ, USA
[b] Department of Geography, University of California, Santa Barbara, CA, USA

ARTICLE INFO

This manuscript was handled by A. Bardossy, Editor-in-Chief, with the assistance of Jie Chen, Associate Editor

Keywords:
Flow duration curve
Ungauged basin
Regression
Variable selection
United States

ABSTRACT

The flow duration curve (FDC) is widely used for river management operations, such as hydropower. Percentile flows of the FDC express the percent of time a flow is equaled or exceeded, and often need to be predicted for ungauged basins. Regression models are commonly used to predict percentile flows. However, a major challenge of this approach is selecting basin characteristics to serve as independent variables. The number of basin characteristics precludes an analysis of all possible models. Thus, a subset of models are typically evaluated using an automated regression procedure, like stepwise regression or the more exhaustive branch-and-bound approach. The latter was used as a baseline approach in this study, and was compared to six other less commonly used methods from the field of variable (feature) selection. The performance of all seven approaches was evaluated based on percentile flow regression models developed for a large sample of 918 basins in the United States. The baseline regression procedure only outperformed principal component analysis, the only method that did not use the percentile flows to select variables. Of the variable selection methods that used the percentile flows, the regression procedure performed the worst. This suggests that regression procedures should not be the first choice among variable selection methods for developing percentile flow regression models. Variables selected based on knowledge of the FDC performed nearly as well as the best overall data-based method (i.e. random forests). Random forests, and other methods that performed well, emphasized the influence of geology on percentile flows. A geologic variable (i.e. baseflow index) had the largest effect on predictive performance. All of the models suffered from large predictive error, and future work should apply a regional approach that groups similar basins to predict percentile flows.

1. Introduction

The flow duration curve (FDC) is a widely used tool for managing rivers. The percent of time a flow is equaled or exceeded (i.e. percentile flow) is a component of the FDC that supplies essential information for river applications, such as hydropower, wastewater treatment, and water abstractions (Vogel and Fennessey, 1995). However, gauged flow data is normally not available to calculate percentile flows. In this case, methods are used to transfer percentile flows from gauged basins to ungauged basins.

Regression models are a simple and common method to predict percentile flows (see the United States (US) Geological Survey StreamStats Application summarized in Ries et al., 2017). The method is founded on the idea that climatic and physical basin characteristics influence the FDC and therefore can be used as independent variables in percentile flow regression models. A major challenge of this method is selecting independent variables among many possible basin characteristics. The number of basin characteristics has grown due to advances in geographic information systems and remote sensing. A database in the US called geospatial attributes of gages for evaluating streamflow version II (GAGES-II) has over 300 basin characteristics (Falcone, 2011). Studies to predict percentile flows and other streamflow statistics have used 251 different basin characteristics as independent variables (Ssegane et al., 2012).

Given the number of independent variables, it is not possible to evaluate all models formed from every combination of variables. A study using only 40 independent variables surpasses a trillion possible models (i.e. 2^{40}–1). For this reason, many studies evaluate a subset of models using automated regression procedures (e.g. Eash and Barnes, 2017; Hsu and Huang, 2017; Painter et al., 2017). Despite their frequent use, automated regression procedures are widely criticized in the field of statistics (see Copas, 1983; Flom and Cassell, 2007; Harrell,

* Corresponding author at: Monmouth University, 400 Cedar Avenue, West Long Branch, NJ 07764, USA.
E-mail address: gfouad@monmouth.edu (G. Fouad).

https://doi.org/10.1016/j.jhydrol.2020.124975
Received 19 September 2019; Received in revised form 11 January 2020; Accepted 14 April 2020
Available online 18 April 2020
0022-1694/

2001 for criticisms). Automated regression procedures are known to produce biased model parameters and, as a result, often underperform on new data in model validation. Alternative approaches do not use regression, and select variables according to (1) knowledge of the dependent variable or (2) data-based methods from the field of feature selection.

1.1. Knowledge-based variable selection

Knowledge-based variable selection uses a conceptual understanding of the dependent variable to select independent variables. A conceptual understanding can be developed based on the literature. Although this can be subjective, a small number of carefully selected independent variables is recommended for regression modeling (see Harrell, 2001 among others). A similar recommendation for ungauged basin model development suggests that independent variables should be carefully selected to reflect the processes that influence streamflow (Castellarin et al., 2013). In doing so, models may be less specific to a particular dataset and more transferable to ungauged basins.

The FDC has a strong physical underpinning (Vogel and Fennessey, 1995). Geographic variation of the FDC is tied to the water balance (Yaeger et al., 2012), topographic setting (Ye et al., 2012), and geology (Cheng et al., 2012) of a region. A modeling study (Yokoo and Sivapalan, 2011) decomposed the FDC into: (1) high flows contributed by surface runoff during storm events with a loss factor for infiltration, (2) average flows reflecting long-term storage conditions influenced by climate and geology, and (3) low flows supplied by groundwater in the dry season subject to evapotranspiration losses. This information can be used to select independent variables associated with the FDC.

1.2. Data-based variable selection

Data-based variable selection methods are developed in the field of feature selection to reduce large datasets for modeling purposes. Feature selection can be broadly categorized into methods that (1) control for redundant (cross-correlated) independent variables and (2) select a subset of variables based on a specified relation to the dependent variable. The former treats cross-correlated independent variables (multicollinearity) to address possible model instability on new data (Dormann et al., 2013). Treating multicollinearity can involve deriving new latent variables that reduce cross-correlation or screening variables based on cross-correlation. A disadvantage of these methods is that they do not use information from the dependent variable. In addition, multicollinearity may not be a concern for predictions (Harrell, 2001) if the model data represents systematic relations between variables that also occur in new data (e.g. snowfall varies with elevation).

A subset of independent variables can be selected based on a mathematical relation to the dependent variable. The mathematical relation may express the independent variable's explanatory power or probabilistic association with the dependent variable. A variable's explanatory power is evaluated according to its effect on model performance. Models are evaluated by excluding one variable at a time (Breiman, 2001), or a training algorithm can be used to screen variables that do not contribute to model performance (Koza, 1994). These methods are however prone to including irrelevant variables due to Simpson's (1951) paradox, a condition in which a relevant variable becomes irrelevant in the presence of another variable. Methods that evaluate probabilistic relations to the dependent variable are effective at filtering irrelevant independent variables (Pearl, 2014). Probabilistic relations are evaluated based on the conditional probability that an observed outcome (flow) changes in the presence of another set of variables (e.g. the balance between precipitation and evapotranspiration changes the probability of observing a particular flow).

The present study compares a typical automated regression procedure to six other less widely used variable selection methods on a large sample of 918 basins for more generalizable results. The objectives of this study were to (1) develop regression models to predict percentile flows for ungauged basins in the US and (2) evaluate the performance of different variable selection methods for developing the regression models. Other modeling approaches, such as artificial neural networks, may be applied to predict percentile flows. However, these approaches also require variable selection. The choice of a modeling approach is unlikely to change the relative performance of different variable selection methods. Therefore, regression models were applied here as a simple and widely used approach. The study addresses the following research question:

How should independent variables be selected for the regression modeling of FDC percentile flows?

A hypothesis is that a small number of carefully selected independent variables may perform similar to evaluating many independent variables in a data-based approach because multicollinearity is common among independent variables (Kroll and Song, 2013) and the FDC can be decomposed into only three distinct physical components (Yokoo and Sivapalan, 2011).

2. Methods

2.1. Basins

The study used 918 basins in the GAGES-II database (Falcone, 2011) classified as "near-natural" with little development. Each basin had to have 30 years of continuous, daily streamflow record to calculate percentile flows (see the next section). The basins were split into 734 calibration basins and 184 validation basins (Fig. 1), meeting a recommendation that model validation should be conducted with at least 100 samples (Harrell, 2001). Validation basins were selected at random within groups based on Köppen climate classes (Peel et al., 2007), three major rock types (Reed and Bush, 2005), and drainage area. The result facilitated a "proxy-basin test" (Klemeš, 1986) in which the calibration and validation basins had similar physiographic conditions (Table 1). A cross validation approach in which small samples of the data are iteratively excluded from model development could have been applied, but was not required given the large calibration and validation sample sizes, which generally produce stable results (Harrell, 2001). The calibration and validation samples share similar statistical distributions, which should suppress the effect of multicollinearity on predictions (Harrell, 2001).

2.2. Percentile flows

Percentile flows were calculated using 30 years of daily streamflow data to generate stable streamflow statistics for different time periods (Kennard et al., 2010). The Weibull plotting position was used to calculate the probability of a flow being equaled or exceeded as follows:

$$p = \frac{r}{(n+1)} \times 100 \quad (1)$$

where p is the exceedance probability of a flow with rank r (in decreasing order) among n flow observations. Other equations may be used to calculate the exceedance probability, but are known to generate similar results for large records of $n > 100$ (Sadegh et al., 2016). A total of 13 percentile flows were calculated at intervals of ten percent from Q_{10}-Q_{90}, extreme high flows (Q_1 and Q_5), and extreme low flows (Q_{95} and Q_{99}). Each percentile flow was normalized (divided) by the mean of nonzero flows to compare basins with different flow magnitudes (Hope and Bart, 2011).

2.3. Independent variables

The independent variables are described in Table 2. A review of the literature on FDC prediction and data covering the contiguous US was conducted to select a representative array of independent variables

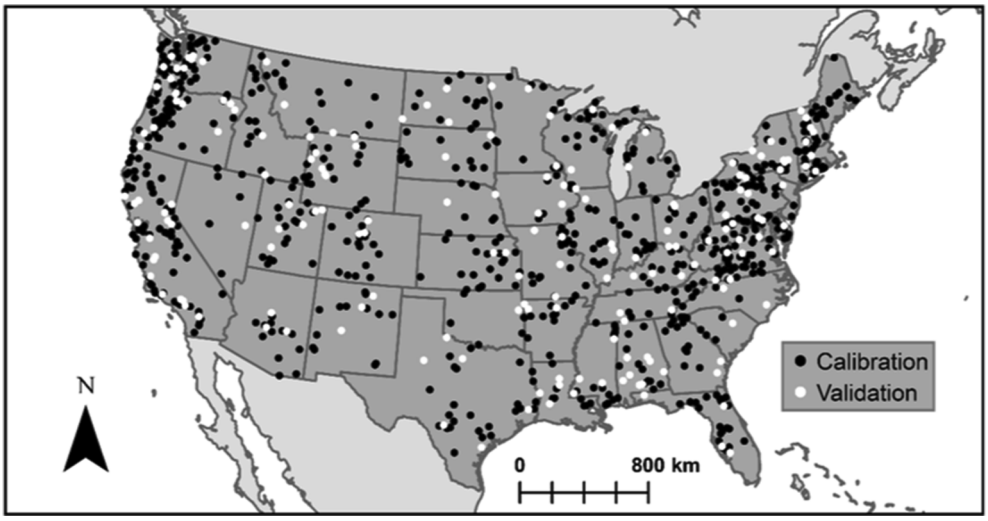

Fig. 1. Map of 734 calibration and 184 validation basins represented by the location of the stream gauge at the basin outlet.

including climatic, topographic, land cover, soil, and geologic basin characteristics. Climatic characteristics were derived from monthly data during the same time periods as the streamflow data, except for daily precipitation variables, which used a 30-year time period from 1981 to 2010 because daily data was not available for all streamflow time periods. The time period of 1981–2010 was used because it overlapped with streamflow data from the most basins. The percent of precipitation delivered as snow is a GAGES-II variable calculated using long-term data from 1901 to 2000. Forest cover was used to summarize land cover due to cross-correlation among land cover classes and the noted effect of forest cover on the FDC (Brown et al., 2013). Geology was represented using the baseflow index (i.e. the percent of streamflow from groundwater), a strong indicator of basin geology (Price, 2011) useful for predicting the FDC in prior studies (see Castellarin et al., 2013; Eash and Barnes, 2017; Yokoo and Sivapalan, 2011). The baseflow index was calculated using streamflow data at gauged locations and interpolated into a grid for ungauged streamflow predictions (Wolock, 2003). Both non-transformed and natural log-transformed independent variables were used in variable selection methods. Independent variables were converted to z-scores (i.e. number of standard deviations from the mean) for variable selection methods sensitive to variables on different scales (i.e. principal component analysis, symbolic regression, and Bayesian networks).

2.4. Variable selection methods

The independent variables were selected for percentile flow regression models using a number of methods. A summary of each method follows with references that cover the method in more depth. Because methods may select a different number of variables, each method was applied in a way that supplied the top five variables. The different methods could then be compared for constructing percentile flow regression models. Again, in the interest of comparing variable selection methods, regression models were not developed for sub-regions as in some other studies (see Tsakiris et al., 2011 for example).

2.4.1. Automated regression procedure

An automated regression procedure served as the "baseline" approach commonly used in percentile flow regression modeling studies. A branch-and-bound procedure was used in lieu of stepwise regression because it conducts a more exhaustive search of all possible models. To do so, the branch-and-bound procedure applies the logic that adding an independent variable should improve the fit of the model. If a variable is added and model fit is not improved, then the branch of models including that variable is eliminated (see Miller, 2002 for a more in depth review). Model fit was evaluated using the sum of squared residuals (error). A number of other objective functions could be used, but the sum of squared residuals was applied as it forms an unbiased estimate of error variance and reflects the overall fit of the model. An increase in the sum of squared residuals signaled that a branch of models could be eliminated. Models were also eliminated based on multicollinearity measured using the condition number (Belsley et al., 2004). A condition number > 30 was used to eliminate models (Dormann et al., 2013). Candidate models were ranked using the adjusted coefficient of determination (adjusted R^2) to compare models with a different number of independent variables. Other metrics, like the Akaike information criterion, can be used to compare models of different complexity, but are not based solely on how well the model fits the calibration data. Therefore, models ranked here based on adjusted R^2 were in order from best to worst fit. The five most frequently used independent variables in the top ten regression models were used in the final percentile flow regression model.

2.4.2. Hydrologic understanding

Hydrologic understanding of the FDC was used to select a small number of independent variables based on the conceptual model that

Table 1
Characteristics of 734 calibration (C) and 184 validation (V) basins.

	Mean annual flow (mm)		Mean annual precipitation (mm)		Baseflow index (%)		Drainage area (km^2)		Mean elevation (m)	
	C	V	C	V	C	V	C	V	C	V
Minimum	1	4	234	287	5	3	2	4	9	16
25th percentile	231	247	798	797	35	32	100	101	276	264
Median	409	412	1106	1100	48	46	292	303	498	470
75th percentile	657	582	1308	1283	61	59	718	751	1194	1090
Maximum	3607	3507	4117	3965	85	82	25,791	8265	3646	3435

Table 2
Independent variables used in percentile flow regression models, with key references that provide a more in depth definition of the variable. Data source abbreviations defined in table footnote.

Variable	Units	Description	Key reference	Data source
Climate				
MAP	mm	Mean annual precipitation	Hope and Bart (2011)	PRISM
Precip_SD	mm	Standard deviation of annual precipitation	Hope and Bart (2011)	PRISM
Precip_1D_Max	mm	Median of annual 1-day maximum precipitation	Yadav et al. (2007)	PRISM
Precip_Intensity	mm/d	Precipitation per rainy day	Kroll et al. (2004)	PRISM
Mean_Temp	°C	Average daily mean temperature	Hope and Bart (2011)	PRISM
PET	mm	Mean annual potential evapotranspiration calculated using the Oudin et al. (2005) equation	Oudin et al. (2005)	PRISM
Aridity	–	Aridity index calculated as PET divided by MAP	Ssegane et al. (2012)	PRISM
Percent_Snow	%	Percent of precipitation as snow	Falcone (2011)	GAGES-II
Topography				
Area	km^2	Drainage area	Falcone (2011)	GAGES-II
Density	km/km^2	Drainage density calculated as stream length divided by drainage area	Ssegane et al. (2012)	NHDPlusV2, GAGES-II
Orientation	°N	Basin angle along main channel	Di Prinzio et al. (2011)	GAGES-II
Elev	m	Mean elevation	Ssegane et al. (2012)	NED
Relief_Ratio	%	Relief ratio calculated as elevation range divided by basin length along main channel	Berger and Entekhabi (2001)	NED, GAGES-II
Slope	%	Mean slope	Ssegane et al. (2012)	NED
Aspect	°N	Mean aspect	Ssegane et al. (2012)	NED
Accumulation	km^2	Mean flow accumulation expressed as upslope area	Povak et al. (2014)	NED
TWI	–	Mean topographic wetness index calculated as ln(accumulation/tan(slope))	Ssegane et al. (2012)	NED
Land cover				
Forest	%	Percent forest cover	Ssegane et al. (2012)	NLCD 1992
Soil				
Soil_Porosity	%	Mean soil porosity expressed as percent pore volume	Hope and Bart (2011)	CONUS-SOIL
Water_Capacity	%	Mean water capacity expressed as percent volume at field capacity	Mohamoud (2008)	CONUS-SOIL
Poorly_Drained	%	Percent poorly drained including hydrologic soil groups C and D	Ssegane et al. (2012)	CONUS-SOIL
Geology				
BFI	%	Mean baseflow index derived from a baseflow grid	Hope and Bart (2011)	BFI48GRD

BFI48GRD: base-flow index 1-km grid for the conterminous United States (Wolock, 2003); CONUS-SOIL: conterminous United States soil characteristics dataset 1-km grid (Miller and White, 1998); GAGES-II: geospatial attributes of gages for evaluating streamflow version II database (Falcone, 2011); NED: national elevation dataset 30-m grid (https://ned.usgs.gov); NHDPlusV2: national hydrography dataset plus version 2 (https://www.nhdplus.com); NLCD 1992: national land cover database 30-m grid of 1992 (https://www.mrlc.gov), the product year in most streamflow time periods; PRISM: parameter-elevation regressions on independent slopes model climate data, long-term model (LT81m) 4-km grid product (http://prism.oregonstate.edu).

the FDC can be decomposed into high, average, and low flows (Yokoo and Sivapalan, 2011). The selection process is of course subjective, and arguments may be made for different independent variables. However, the variables chosen here are based on readily accessible data for the contiguous US and a common understanding of the FDC. High flows are the product of storms related to total precipitation over the course of a year (MAP). Storm runoff is moderated by infiltration (Soil_Porosity). The slope of the basin influences the magnitude and duration of high flows. Average flows are associated with basin storage conditions influenced by the balance between MAP and PET and geology (BFI). Low flows in the dry season are generated by baseflow subject to evaporative losses. Together, the conceptual model of different flows was used to select the following independent variables from Table 2: MAP, PET, Slope, Soil_Porosity, and BFI.

2.4.3. Principal component analysis

Principal component analysis (PCA) transforms the independent variables into a new set of uncorrelated variables called principal components (PCs; see Kroll and Song, 2013 for a more in depth explanation of PCA). Each PC explains a percent of the variance in the original data. The natural log-transformed independent variables produced PCs that explained more variance in the data than PCs based on the non-transformed variables. The first five PCs based on the log-transformed variables explained 76% of the variance in the data, and were used for the percentile flow regression models.

2.4.4. Correlation analysis

A correlation analysis was performed to screen cross-correlated variables. Cross-correlation was assessed using Pearson's and Spearman's coefficients. A coefficient > 0.7 was used to identify cross-correlated groups of variables (Dormann et al., 2013). A univariate regression with each percentile flow was conducted to rank both the non-transformed and natural log-transformed independent variables. The variable with the largest R^2 in a cross-correlated group was ranked alongside the uncorrelated variables. The top five variables were then used for the percentile flow regression models.

2.4.5. Random forests

Random forests are regression trees, which apply rules (e.g. MAP > 1000 mm) to split the data samples (basins) into progressively smaller groups. The smallest groups (terminal nodes) are averaged to generate predictions (see Breiman, 2001 for a more in depth review). The regression trees were developed using random subsets of the basins, and evaluated for predictive performance using the "out-of-bag" samples not used to develop the regression trees. The out-of-bag error decreased until about 100 regression trees, which was the number of trees used in the analysis. Terminal nodes of the regression trees included five basins, or less than 1% of the basins, because the size of the terminal nodes has little effect on predictions if a small percentage of the data is used (Svetnik et al., 2003). The number of independent variables evaluated at each split in the regression tree was determined based on an analysis of all values (i.e. 1–22) on all percentile flows (i.e. Q_1–Q_{99}). The analysis revealed little difference between the values and a commonly used rule of thumb (i.e. one-third of the total number of independent variables rounded down to the nearest whole number). The number of variables evaluated at each split was seven.

Independent variables were ranked according to the out-of-bag error. Each variable was randomly permuted, effectively removing that variable from the predictions. A larger increase in the out-of-bag error quantified as mean squared error indicated a more important variable, which received a higher ranking. The top five independent variables were based on average rankings from 1000 random forests because rankings can change from one random forest run to the next and an ensemble approach like this is advised (Saeys et al., 2008). Natural log-

transformed independent variables were evaluated for the top five variables. The form of the variable that generated a larger R^2 was used in the final percentile flow regression model.

2.4.6. Symbolic regression

Symbolic regression evaluates regression models based on a genetic program, which simulates the evolution of a population (see Koza, 1994 for a review of the genetic programming process). An initial set of independent variables and mathematical operators were presented to the program. The operators were limited to addition, subtraction, and natural log for consistency with the other variable selection methods. The variables and operators were combined to generate an initial population of regression models. The root mean squared error was used to evaluate the models. A model with less error was more likely to pass on attributes to the next generation of models. The evolution of the models was influenced by four parameters: (1) the number of models in the first generation, (2) the number of models in each subsequent generation, (3) the number of completely new models generated for each generation, and (4) the probability that models with less error were combined to form new models. The parameters were set using a sequential parameter optimization procedure defined in Bartz-Beielstein and Zaefferer (2012). The final generation of models was ranked according to adjusted R^2. The top ten models were used as in the automated regression procedure to identify the five most frequently used independent variables, which were then used in the final percentile flow regression model.

2.4.7. Bayesian networks

Bayesian networks assess probabilistic relations to the dependent variable. To do so, a Markov blanket is computed to identify the independent variables that make the dependent variable statistically independent of other variables (see Aliferis et al., 2010 for an explanation of this process). A subset of variables explained the probability of observing a particular flow, and other variables that had no effect on the flow's probability were eliminated. Variables can be eliminated either by developing part of the Bayesian network around the dependent variable or directly evaluating the conditional probability of the dependent variable. The latter approach was applied using HITON Markov blanket (Aliferis et al., 2003), but ultimately discarded because it eliminated few independent variables.

The Bayesian network was developed around the dependent variable using a local causal discovery algorithm described in Mani and Cooper (1999). Because the approach selected more than five independent variables, the Bayesian network was developed five times, excluding 20% of the basins each time as in a *k*-fold cross-validation. The independent variables were ranked based on the number of times that they were selected by the Bayesian networks. The form of the top five variables (i.e. non-transformed or natural log-transformed) that generated a larger R^2 was used in the final percentile flow regression model.

2.5. Regression model evaluation

Regression models used the top five independent variables as follows:

$$ln(Q_i + 1) = \beta_0 + \beta_1 X_1 \cdots + \beta_5 X_5 \quad (2)$$

where the natural log transformation was used to model skewed percentile flows (Q_i), a constant of one was added to the percentile flows to calculate the natural log of zero flows, β_0, β_1, …, β_5 are the model coefficients, and X_1, X_2, …, X_5 are the non-transformed or natural log-transformed independent variables.

The regression models were assessed for multicollinearity and predictive performance. Multicollinearity was assessed using the condition number (Belsley et al., 2004) to evaluate the redundancy of independent variables selected by the different variable selection methods. Predictive performance was assessed using the 184 validation basins withheld from regression model development. The difference between observed and predicted percentile flows was evaluated using R^2, Nash and Sutcliffe (1970) efficiency, and relative error calculated as the absolute difference between observed and predicted values divided by the observed value plus one to accommodate zero flows.

3. Results

The results compare the following variable selection methods (abbreviations in parentheses): automated regression procedure (baseline), knowledge-based variable selection (expert), principal component analysis (PCA), correlation analysis (corr), random forests (RF), symbolic regression (SR), and Bayesian networks (BN).

3.1. Multicollinearity

Multicollinearity was not assessed to screen the final percentile flow regression models because it is generally not a concern given large, representative samples (Harrell, 2001) and did not impair predictive performance as shown in the next section. Instead, multicollinearity is evaluated here to assess the variable selection method's ability to select non-redundant independent variables. A larger condition number indicates that the variable selection method chose more redundant variables (Table 3). The three methods that controlled for multicollinearity (baseline, PCA, and corr) had the smallest condition numbers. PCA effectively eliminated multicollinearity by using uncorrelated PCs derived from the independent variables. The baseline regression procedure had the second least multicollinearity because it rejected models with a condition number > 30. This did however result in few models (18%) having multiple independent variables, and only three independent variables were used in the largest models, accounting for less than 1% of the models. The final percentile flow regression models could have a condition number > 30 because, rather than use fewer variables than other variable selection methods, the baseline regression procedure selected the five most frequently used independent variables in high-performing models. The correlation analysis (corr) was the least effective of methods that treated multicollinearity because pairwise correlation values failed to account for multicollinearity in the regression models.

The methods that did not control for multicollinearity (expert, RF, SR, and BN) had the largest condition numbers (Table 3). Symbolic regression (SR) and Bayesian networks (BN) had similar levels of multicollinearity, although Bayesian networks can exclude variables with similar information (Aliferis et al., 2010). Variables selected based on hydrologic understanding of the FDC (expert) experienced multicollinearity due to correlation between PET and BFI (Pearson correlation coefficient = 0.58 and Spearman correlation coefficient = 0.63), but neither variable was screened to evaluate a purely knowledge-based variable selection method. Random forests had the greatest multicollinearity possibly because the process of evaluating one variable at a time can assign similar rankings to cross-correlated variables (Breiman, 2001) and highly ranked variables with the greatest predictive power were cross-correlated.

Table 3
The average and range of multicollinearity expressed as the condition number for the 13 percentile flow regression models developed using different variable selection methods.

	Baseline	Expert	PCA	Corr	RF	SR	BN
Minimum	32	1142	2	191	337	31	191
Average	68	8575	2	2124	22,129	6502	6810
Maximum	158	31,437	2	9892	68,792	28,011	21,849

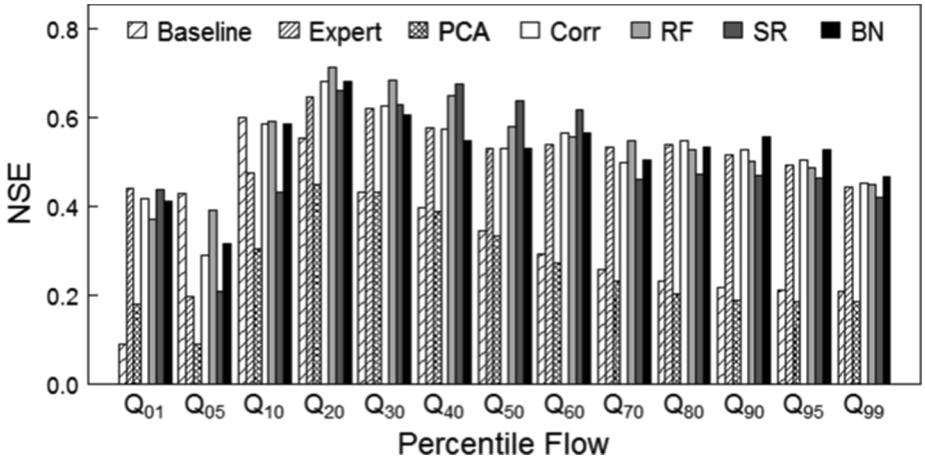

Fig. 2. Nash-Sutcliffe efficiency (NSE) for percentile flow regression models developed using different variable selection methods.

3.2. Predictive performance

Predictive performance was quantified using the Nash-Sutcliffe efficiency (NSE) for each percentile flow (Fig. 2). Similar patterns in performance were observed for R^2 and relative error values, and are not shown here. Performance (i.e. NSE) peaked at the flow equaled or exceeded 20% of the time (Q_{20}), and steadily declined for lower flows from Q_{30}-Q_{99}. Extreme high flows (Q_{01} and Q_{05}) had the lowest performance.

The variable selection methods had different levels of predictive performance (Fig. 2). The baseline regression procedure consistently had the second worst performance (see flows from Q_{20}-Q_{99}). PCA had the worst performance (except for Q_{01}) likely because it was the only method that did not use information from the dependent variables. The baseline regression procedure and PCA performed the worst despite controlling for multicollinearity. This indicates the calibration basins were representative of the validation basins, and multicollinearity did not impair predictions as in a proxy-basin test (Klemeš, 1986).

The best performance was achieved mostly by methods other than the baseline regression procedure (Fig. 2). Random forests (RF), symbolic regression (SR), and Bayesian networks (BN) each had the best performance on three different percentile flows. Random forests performed best where performance peaked at Q_{20} and Q_{30}, indicating it was most effective at distinguishing between more influential independent variables. Average flows from Q_{40}-Q_{60} were predicted best by symbolic regression, which notably used fewer than five variables to predict Q_{50} and Q_{60}, showing that many of the independent variables were redundant or did not have useful information. The same can be said for lower flows from Q_{70}-Q_{99} for which symbolic regression used only two independent variables, but had only slightly lower performance than the best methods (i.e. less than 16% difference from the maximum NSE values). The lowest flows from Q_{90}-Q_{99} had a large fraction of zero flows (14%). Bayesian networks predicted these flows best perhaps because it was most effective at predicting zero flows and able to identify a certain set of conditions (e.g. low baseflow and high evapotranspiration) associated with the probability of observing a zero flow. Finally, variables selected based on hydrologic understanding of the FDC (expert) performed best for Q_{01}, and had NSE values within 12% of the best methods on average. Therefore, five carefully selected independent variables were almost as effective as more complex data-based methods that had access to 22 independent variables.

The overall performance of the different variable selection methods is summarized as the sum of relative error in Table 4, and largely confirms prior results. The worst performance (i.e. largest relative error) belonged to PCA. Other than PCA, the baseline regression procedure did not perform better than any other method. The simple approach of selecting variables based on hydrologic understanding (expert) performed similar to three other data-based methods (corr, SR, and BN), and had 8% more relative error than random forests (RF), which minimized relative error the most. These results indicate hydrologic understanding is a critical step in predicting the FDC as suggested in Castellarin et al. (2013).

Table 4
Sum of the relative error expressed as a percent for 13 percentile flow regression models developed using different variable selection methods.

	Baseline	Expert	PCA	Corr	RF	SR	BN
Sum (%)	170	147	181	145	139	145	145

3.3. Selected independent variables

Independent variables selected for a sample of high (Q_{10}), average (Q_{50}), and low (Q_{90}) percentile flow regression models are shown in Table 5. Similar variables were chosen for other high, average, and low flows not shown. Some variables were chosen frequently regardless of flow (see Aridity, BFI, and Forest). These variables were related to the overall shape of the FDC, whereas other variables were only related to certain flows (see MAP, Percent_Snow, and Poorly_Drained). Mean annual precipitation (MAP) was frequently used for high and average percentile flow regression models, as these flows are fed by rainfall events and accumulated precipitation over the course of a year (Yokoo and Sivapalan, 2011). The percent of precipitation delivered as snow (Percent_Snow) was often used for high percentile flow regression models, possibly due to the contribution of spring snowmelt to high flows in alpine rivers (Ye et al., 2012). The percent of a basin covered in poorly drained soils (Poorly_Drained) was a common independent variable in low percentile flow regression models. Poorly drained soils may reduce groundwater supplies for low flows in the dry season (Cheng et al., 2012). The repeated use of independent variables, both for different flows and across different variable selection methods, indicates once again that many independent variables were either redundant or not useful. Topographic variables, like mean slope, were absent from many regression models and not generally useful for predicting percentile flows.

The variable selection methods had some notable differences in the variables that they chose for the percentile flow regression models (Table 5). Most notably, BFI was not used in the baseline regression models, but was used in at least 12 of 13 percentile flow regression models formulated by the other variable selection methods. This helps explain the poor performance of the baseline regression procedure for many of the percentile flows (see Fig. 2). The baseline regression procedure rejected models that used BFI due to multicollinearity, whereas

Table 5
Independent variables selected by different variable selection methods for sample high (Q_{10}), average (Q_{50}), and low (Q_{90}) percentile flow regression models. Variable selection methods do not include principal component analysis (PCA) and variables selected based on hydrologic understanding of the FDC (expert) because these methods used the same five independent variables for each percentile flow regression model. The baseline regression procedure (baseline) has Aridity twice as non-transformed and natural log-transformed independent variables.

Flow	Baseline	Corr	RF	SR	BN
Q_{10}	Aridity	Aridity	Aridity	Aridity	Aridity
	Aridity	BFI	BFI	BFI	BFI
	Percent_Snow	Forest	MAP	Forest	Forest
	Precip_Intensity	MAP	Percent_Snow	MAP	MAP
	Water_Capacity	Percent_Snow	PET	Orientation	Percent_Snow
Q_{50}	Aridity	Aridity	Aridity	Aspect	Aridity
	Aridity	BFI	BFI	BFI	BFI
	Forest	Forest	Elev	Elev	Forest
	Percent_Snow	MAP	MAP	Forest	MAP
	Poorly_Drained	Soil_Porosity	Soil_Porosity	–	Soil_Porosity
Q_{90}	Aridity	Aridity	Aridity	BFI	BFI
	Aridity	BFI	BFI	Forest	Mean_Temp
	Forest	Forest	Forest	–	Percent_Snow
	Mean_Temp	Percent_Snow	Poorly_Drained	–	PET
	Poorly_Drained	Poorly_Drained	TWI	–	Poorly_Drained

other variable selection methods did not screen for multicollinearity. BFI elevated multicollinearity because it is related to other climatic, topographic, land cover, and soil variables (Price, 2011). Despite this, BFI was a critical variable for raising predictive performance in validation. Although topographic variables were not generally useful for predicting percentile flows, they were useful in specific instances. Symbolic regression (SR) was the only method that used mean aspect (Aspect) and mean elevation (Elev) to predict Q_{50} and Q_{60}, and performed the best on these percentile flows (see Fig. 2). Both of these variables (Aspect and Elev) are likely related to other long-term water balance and climatic conditions associated with average flows (Ye et al., 2012). Bayesian networks (BN) performed best on low flows from Q_{90}-Q_{99} (see Fig. 2), and unlike other methods, used mean daily temperature (Mean_Temp) and mean annual potential evapotranspiration (PET), both of which have been tied to low flows in previous work (see Kroll et al., 2004; Pumo et al., 2013; Yokoo and Sivapalan, 2011).

4. Discussion

Multicollinearity was prominent among independent variables (see Table 3). This is a well-documented problem in studies that use a variety of climatic and physical basin characteristics to predict flows (see Eash and Barnes, 2017; Kroll et al., 2004; Kroll and Song, 2013). The problem occurs because climatic and physical characteristics co-evolve over time (e.g. poorly developed soils in arid basins) and are therefore interdependent (Hrachowitz et al., 2013). Multicollinearity is less of a problem for model predictions given a large, diverse sample (Kroll and Song, 2013) as in this study. However, if interpreting model coefficients is a priority, then an alternative form of regression, like partial least squares, may be preferred. Regression procedures screen models for multicollinearity using arbitrary thresholds, like a condition number > 30 (Dormann et al., 2013). This is an uncertain process since different statistics, like the determinant of the correlation matrix or variance inflation factor, can be used to screen models, and these statistics may be sensitive to the dataset (Snee and Marquardt, 1984). In this case, the condition number was relaxed to 40 to illustrate the uncertainty in setting the multicollinearity threshold. After relaxing the threshold, baseline regression models began to use the critical variable of BFI, and model performance improved. The question of how to treat multicollinearity in regression models for predicting flow is an active area of research (Kroll and Song, 2013), and should be investigated for percentile flows using a wide variety of models and multicollinearity statistics as in Dormann et al. (2013).

Predictive performance for the different percentile flows (see Fig. 2) followed patterns from previous studies, except for the decline in performance for high flows (see Hashmi and Shamseldin, 2014; Hope and Bart, 2012; Hsu and Huang, 2017). Like these studies, predictive performance decreased for lower percentile flows. This is common due to the complex, non-linear processes that govern low flows (Hope and Bart, 2011). Furthermore, independent variables may not adequately represent subsurface properties that influence low flows (Kroll et al., 2004). This study used BFI and several soil variables to represent subsurface properties. Both BFI and poorly drained soils (Poorly_Drained) were important variables for predicting low flows (see Table 5). Similar variables related to the storage of a basin should be emphasized in future low flow studies. The low flows in this study had zero flows, which may require specialized modeling approaches as in Hope and Bart (2011). The decline in performance for high flows (Q_{01} and Q_{05}) was uncharacteristic of previous studies, and may be due to the large variance in floods across the contiguous US. Prior studies, like Ssegane et al. (2012), have had more success in predicting high flows at a regional scale. Dividing the contiguous US into flood regions, as in regional flood frequency analysis (Burn, 1990), may improve the prediction of high flows. Identifying homogenous regions (i.e. similar basins) reduces the variability of high flows and generally improves the accuracy of predictions (Burn, 1990). However, this approach was not applied here because the focus of the study was on comparing variable selection methods and not on grouping basins for regional predictions. Overall, the regression models in this study explained about half of the variance in percentile flows, whereas prior studies have explained about three-quarters of the variance (see Hope and Bart, 2012; Hsu and Huang, 2017; Ssegane et al., 2012). Again, the difference may be due to the scale of this study.

Variable selection method comparisons revealed that conventional approaches, like a baseline regression procedure and PCA, did not perform as well as alternative methods (see Table 4). PCA has not improved predictions in previous studies (see Kroll and Song, 2013; Ssegane et al., 2012), and should only be used if controlling for multicollinearity is a priority. A baseline regression procedure has performed similarly to Bayesian networks in a previous study (Ssegane et al., 2012), but the same study found that Bayesian networks more consistently chose the correct independent variables of a known model and therefore more accurately represent a system. This can lead to improved percentile flow predictions, which was the case when comparing Bayesian networks to the baseline regression procedure in this study. The best method varied from percentile flow to percentile flow (see Fig. 2), a result that substantiates the recommendation to use a combination of variable selection methods (Ssegane et al., 2012). Percentile flow predictions may also be improved through the use of an ensemble (e.g. model averaging) approach (Waseem et al., 2015).

A small number of independent variables were repeatedly selected for percentile flow regression models (see Aridity, BFI, and Forest in Table 5). Aridity (i.e. the ratio of mean annual potential evapotranspiration to mean annual precipitation) is a measure of the long-term water balance generally related to storm runoff (i.e. high percentile flows) (Rossi et al., 2016) and groundwater storage conditions (Istanbulluoglu et al., 2012) that influence average and low percentile flows (Yokoo and Sivapalan, 2011). The percent of streamflow from groundwater (BFI) is indirectly related to high percentile flows as an indicator of infiltration (i.e. a loss factor for storm runoff) (Yokoo and Sivapalan, 2011), and directly affects average and low percentile flows (Cheng et al., 2012). BFI describes subsurface drainage, and other variables that describe groundwater storage (e.g. aquifer thickness) and groundwater discharge (e.g. a baseflow recession constant) available as interpolated grids, like BFI, may help predict percentile flows. Forest coverage (Forest) moderates high to low percentile flows via interception and evapotranspiration (Pumo et al., 2013; Yaeger et al., 2012). Vegetation indices that convey similar information, like leaf area index, may improve percentile flow predictions in future studies.

Independent variables selected based on hydrologic understanding of the FDC could be improved based on independent variables frequently used in percentile flow regression models (see previous paragraph) and predictive performance across the percentile flows (see Fig. 2). Frequently used variables (i.e. Aridity, BFI, and Forest) could be used to replace MAP and PET (combined in Aridity) and Slope (not a frequently used variable). The other two variables could target high and low percentile flows that had lower predictive performance. Variables for low percentile flows, like poorly drained soils, aquifer thickness, and a baseflow recession constant, have been previously discussed. High percentile flows correlate strongly to variables that describe maximum precipitation, such as maximum daily precipitation times the fraction of days without precipitation (Cheng et al., 2012). A revised set of independent variables based on hydrologic understanding of the FDC is Aridity, BFI, Forest, a variable that describes subsurface drainage (e.g. poorly drained soils), and a variable that describes maximum precipitation (e.g. maximum daily precipitation times the fraction of days without precipitation).

5. Conclusions

Variable selection methods to develop percentile flow regression models were compared for a large sample of 918 basins in the United States. Because of the large sample, multicollinearity (i.e. cross-correlation between independent variables) was not a problem for predictive performance on the 184 validation basins withheld from model development. Instead, high levels of multicollinearity indicate that many commonly used independent variables, such as an array of different climatic variables, are redundant. Treating multicollinearity using a regression diagnostic (i.e. the condition number) and principal components was problematic for an automated regression procedure and PCA, both of which had the worst overall performance of variable selection methods. Other variable selection methods performed better because they used BFI, a highly important, albeit cross-correlated, independent variable. The best overall method (i.e. random forests) only performed marginally better than variables selected based on hydrologic understanding of the FDC, which indicates at the very least the initial set of independent variables should be explicitly linked to hydrologic processes that influence the FDC. The best predictive performance did not belong to any one method for the various percentile flows, suggesting that using a combination of variable selection methods, either to rank variables or develop ensemble models, could enhance percentile flow regression models. Predictive performance declined for high and low percentile flows. Independent variables specifically targeting these flows, such as variables that characterize storm runoff and basin storage, should be developed in future work. Homogenous regions that reduce the variance in percentile flows should also be developed to further advance percentile flow regression models for the contiguous US. The study was limited to the context of developing regression models. However, future work may leverage large datasets such as in this study to use more adaptable model forms from the field of soft computing. Emphasis of this research thus far has been on streamflow forecasting (see Yaseen et al., 2019 for example), but now with access to large basin databases, machine learning models should be applied to predict percentile flows at ungauged basins.

Conflict of interest

None.

CRediT authorship contribution statement

Geoffrey Fouad: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing - original draft, Writing - review & editing. **Hugo A. Loáiciga:** Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank Allen Hope for guidance on this research. We also thank André Skupin and Christina Tague for helpful comments.

References

Aliferis, C.F., Statnikov, A., Tsamardinos, I., Mani, S., Koutsoukos, X.D., 2010. Local causal and Markov blanket induction for causal discovery and feature selection for classification part I: algorithms and empirical evaluation. J. Mach. Learn. Res. 11, 171–234.

Aliferis, C.F., Tsamardinos, I., Statnikov, A., 2003. HITON: a novel Markov blanket algorithm for optimal variable selection. In: American Medical Informatics Association Annual Symposium Proceedings. American Medical Informatics Association, Bethesda, MD, USA, pp. 21–25.

Bartz-Beielstein, T., Zaefferer, M., 2012. A gentle introduction to sequential parameter optimization. CIplus, Band 1/2012.

Belsley, D.A., Kuh, E., Welsch, R.E., 2004. Detecting and assessing collinearity. In: Belsley, D.A., Kuh, E., Welsch, R.E. (Eds.), Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. John Wiley and Sons, Hoboken, NJ, USA, pp. 85–191.

Berger, K.P., Entekhabi, D., 2001. Basin hydrologic response relations to distributed physiographic descriptors and climate. J. Hydrol. 247, 169–182. https://doi.org/10.1016/S0022-1694(01)00383-3.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Brown, A.E., Western, A.W., McMahon, T.A., Zhang, L., 2013. Impact of forest cover changes on annual streamflow and flow duration curves. J. Hydrol. 483, 39–50. https://doi.org/10.1016/j.jhydrol.2012.12.031.

Burn, D.H., 1990. Evaluation of regional flood frequency analysis with a region of influence approach. Water Resour. Res. 26, 2257–2265. https://doi.org/10.1029/WR026i010p02257.

Castellarin, A., Botter, G., Hughes, D.A., Liu, S., Ouarda, T.B.M.J., Parajka, J., Post, D.A., Sivapalan, M., Spence, C., Viglione, A., Vogel, R.M., 2013. Prediction of flow duration curves in ungauged basins. In: Blöschl, G., Sivapalan, M., Wagener, T., Viglione, A., Savenije, H. (Eds.), Runoff Prediction in Ungauged Basins: Synthesis Across Processes, Places and Scales. Cambridge University Press, Cambridge, United Kingdom, pp. 135–162.

Cheng, L., Yaeger, M., Viglione, A., Coopersmith, E., Ye, S., Sivapalan, M., 2012. Exploring the physical controls of regional patterns of flow duration curves – Part 1: insights from statistical analyses. Hydrol. Earth Syst. Sci. 16, 4435–4446. https://doi.org/10.5194/hess-16-4435-2012.

Copas, J.B., 1983. Regression, prediction and shrinkage. J. Roy. Stat. Soc. B Met. 45, 311–354. https://doi.org/10.1111/j.2517-6161.1983.tb01258.x.

Di Prinzio, M., Castellarin, A., Toth, E., 2011. Data-driven catchment classification: application to the pub problem. Hydrol. Earth Syst. Sci. 15, 1921–1935. https://doi.org/10.5194/hess-15-1921-2011.

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., García Marquéz, J.R., Gruber, B., Lafourcade, B., Leitão, P.J., Münkemüller, T., McClean, C., Osborne,

P.E., Reineking, B., Schröder, B., Skidmore, A.K., Zurell, D., Lautenbach, S., 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36, 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x.

Eash, D.A., Barnes, K.K., 2017. Methods for estimating selected low-flow frequency statistics and harmonic mean flows for streams in Iowa. US Geological Survey Scientific Investigations Report 2012–5171. 99 p. https://doi.org/10.3133/sir20125171.

Falcone, J.A., 2011. GAGES-II: Geospatial attributes of gages for evaluating streamflow. US Geological Survey Dataset. https://doi.org/10.3133/70046617.

Flom, P.L., Cassell, D.L., 2007. Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use. NorthEast SAS Users Group (NESUG): Statistics and Data Analysis, Baltimore, MD, USA.

Harrell, F.E., 2001. Multivariable modeling strategies. In: Harrell, F.E. (Ed.), Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. Springer, Berlin, Germany, pp. 53–85.

Hashmi, M.Z., Shamseldin, A.Y., 2014. Use of gene expression programming in regionalization of flow duration curve. Adv. Water Resour. 68, 1–12. https://doi.org/10.1016/j.advwatres. 2014.02.009.

Hope, A., Bart, R., 2011. Evaluation of a regionalization approach for daily flow duration curves in central and southern California watersheds. J. Am. Water Res. Assoc. 48, 123–133. https://doi.org/10.1111/j.1752-1688.2011.00597.x.

Hope, A., Bart, R., 2012. Synthetic monthly flow duration curves for the Cape Floristic Region, South Africa. Water SA 38, 191–200. https://doi.org/10.4314/wsa.v38i2.4.

Hrachowitz, M., Savenije, H.H.G., Blöschl, G., McDonnell, J.J., Sivapalan, M., Pomeroy, J.W., Arheimer, B., Blume, T., Clark, M.P., Ehret, U., Fenicia, F., Freer, J.E., Gelfan, A., Gupta, H.V., Hughes, D.A., Hut, R.W., Montanari, A., Pande, S., Tetzlaff, D., Troch, P.A., Uhlenbrook, S., Wagener, T., Winsemius, H.C., Woods, R.A., Zehe, E., Cudennec, C., 2013. A decade of predictions in ungauged basins (PUB) – a review. Hydrol. Sci. J. 58, 1198–1255. https://doi.org/10.1080/02626667.2013.803183.

Hsu, N.S., Huang, C.J., 2017. Estimation of flow duration curve at ungauged locations in Taiwan. J. Hydrol. Eng. 22. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001511.

Istanbulluoglu, E., Wang, T., Wright, O.M., Lenters, J.D., 2012. Interpretation of hydrologic trends from a water balance perspective: the role of groundwater storage in the Budyko hypothesis. Water Resour. Res. 48. https://doi.org/10.1029/2010WR010100.

Kennard, M.J., Mackay, S.J., Pusey, B.J., Olden, J.D., Marsh, N., 2010. Quantifying uncertainty in estimation of hydrologic metrics for ecohydrological studies. River Res. Appl. 26, 137–156. https://doi.org/10.1002/rra.1249.

Klemeš, V., 1986. Operational testing of hydrological simulation models. Hydrol. Sci. J. 31, 13–24. https://doi.org/10.1080/02626668609491024.

Koza, J.R., 1994. Genetic programming as a means for programming computers by natural selection. Stat. Comput. 4, 87–112. https://doi.org/10.1007/BF00175355.

Kroll, C., Luz, J., Allen, B., Vogel, R.M., 2004. Developing a watershed characteristics database to improve low streamflow prediction. J. Hydrol. Eng. 9, 116–125. https://doi.org/10.1061/(ASCE)1084-0699(2004)9:2(116).

Kroll, C.N., Song, P., 2013. Impact of multicollinearity on small sample hydrologic regression models. Water Resour. Res. 49, 3756–3769. https://doi.org/10.1002/wrcr.20315.

Mani, S., Cooper, G.F., 1999. A study in causal discovery from population-based infant birth and death records. In: American Medical Informatics Association Annual Symposium Proceedings. American Medical Informatics Association, Bethesda, MD, USA, pp. 315–319.

Miller, A., 2002. Finding subsets which fit well. In: Miller, A. (Ed.), Subset Selection in Regression. CRC Press, Boca Raton, FL, USA, pp. 37–88.

Miller, D.A., White, R.A., 1998. A conterminous United States multilayer soil characteristics dataset for regional climate and hydrology modeling. Earth Interact. 2. https://doi.org/10.1175/1087-3562(1998)002<0001:ACUSMS>2.3.CO;2.

Mohamoud, Y.M., 2008. Prediction of daily flow duration curves and streamflow for ungauged catchments using regional flow duration curves. Hydrol. Sci. J. 53, 706–724. https://doi.org/10.1623/hysj.53.4.706.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I – a discussion of principles. J. Hydrol. 10, 282–290. https://doi.org/10.1016/0022-1694(70)90255-6.

Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andréassian, V., Anctil, F., Loumagne, C., 2005. Which potential evapotranspiration input for a lumped rainfall-runoff model? Part 2 – towards a simple and efficient potential evapotranspiration model for rainfall-runoff modelling. J. Hydrol. 303, 290–306. https://doi.org/10.1016/j.jhydrol.2004.08.026.

Painter, C.C., Heimann, D.C., Lanning-Rush, J.L., 2017. Methods for estimating annual exceedance-probability streamflows for streams in Kansas based on data through water year 2015. US Geological Survey Scientific Investigations Report 2017–5063. 20 p. https://doi.org/10.3133/sir20175063.

Pearl, J., 2014. Understanding Simpson's paradox. Am. Stat. 68, 8–13. https://doi.org/10.1080/00031305.2014.876829.

Peel, M.C., Finlayson, B.L., McMahon, T.A., 2007. Updated world map of the Köppen-Geiger climate classification. Hydrol. Earth Syst. Sci. 11, 1633–1644. https://doi.org/10.5194/hess-11-1633-2007.

Povak, N.A., Hessburg, P.F., McDonnell, T.C., Reynolds, K.M., Sullivan, T.J., Salter, R.B., Cosby, B.J., 2014. Machine learning and linear regression models to predict catchment-level base cation weathering rates across the southern Appalachian Mountain region, USA. Water Resour. Res. 50, 2798–2814. https://doi.org/10.1002/2013WR014203.

Price, K., 2011. Effects of watershed topography, soils, land use, and climate on baseflow hydrology in humid regions: a review. Prog. Phys. Geog. 35, 465–492. https://doi.org/10.1177/0309133311402714.

Pumo, D., Noto, L.V., Viola, F., 2013. Ecohydrological modelling of flow duration curve in Mediterranean river basins. Adv. Water Resour. 52, 314–327. https://doi.org/10.1016/j.advwatres.2012.05.010.

Reed, J.C., Bush, C.A., 2005. Generalized geologic map of the United States, Puerto Rico, and the US Virgin Islands. US Geological Survey Dataset. https://pubs.usgs.gov/atlas/geologic.

Ries, K.G., Newsom, J.K., Smith, M.J., Guthrie, J.D., Steeves, P.A., Haluska, T.L., Kolb, K.R., Thompson, R.F., Santoro, R.D., Vraga, H.W., 2017. StreamStats, version 4. US Geological Survey Fact Sheet 2017–3046. 4 p. https://doi.org/10.3133/fs20173046.

Rossi, M.W., Whipple, K.X., Vivoni, E.R., 2016. Precipitation and evapotranspiration controls on daily runoff variability in the contiguous United States and Puerto Rico. J. Geophys. Res. Earth 121, 128–145. https://doi.org/10.1002/2015JF003446.

Sadegh, M., Vrugt, J.A., Gupta, H.V., Xu, C., 2016. The soil water characteristic as new class of closed-form parametric expressions for the flow duration curve. J. Hydrol. 438–456. https://doi.org/10.1016/j.jhydrol.2016.01.027.

Saeys, Y., Abeel, T., Van de Peer, Y., 2008. Robust feature selection using ensemble feature selection techniques. In: Daelemans, W., Goethals, B., Morik, K. (Eds.), Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Springer, Berlin, Germany, pp. 313–325.

Simpson, E.H., 1951. The interpretation of interaction in contingency tables. J. Roy. Stat. Soc. B Met. 13, 238–241. https://doi.org/10.1111/j.2517-6161.1951.tb00088.x.

Snee, R.D., Marquardt, D.W., 1984. Collinearity diagnostics depend on the domain of prediction, the model, and the data. Am. Stat. 38, 83–87. https://doi.org/10.2307/2683239.

Ssegane, H., Tollner, E.W., Mohamoud, Y.M., Rasmussen, T.C., Dowd, J.F., 2012. Advances in variable selection methods I: causal selection methods versus stepwise regression and principal component analysis on data of known and unknown functional relationships. J. Hydrol. 438–439, 16–25. https://doi.org/10.1016/j.jhydrol.2012.01.008.

Svetnik, V., Liaw, A., Tong, C., Culberson, C., Sheridan, R.P., Feuston, B.P., 2003. Random forest: a classification and regression tool for compound classification and QSAR modeling. J. Chem. Inf. Comp. Sci. 43, 1947–1958. https://doi.org/10.1021/ci034160g.

Tsakiris, G., Nalbantis, I., Cavadias, G., 2011. Regionalization of low flows based on canonical correlation analysis. Adv. Water Resour. 34, 865–872. https://doi.org/10.1016/j.advwatres. 2011.04.007.

Vogel, R.M., Fennessey, N.M., 1995. Flow duration curves II: a review of applications in water resources planning. J. Am. Water Res. Assoc. 31, 1029–1039. https://doi.org/10.1111/j.1752-1688.1995.tb03419.x.

Waseem, M., Ajmal, M., Kim, T.W., 2015. Ensemble hydrological prediction of streamflow percentile at ungauged basins in Pakistan. J. Hydrol. 525, 130–137. https://doi.org/10.1016/j.jhydrol.2015.03.042.

Wolock, D.M., 2003. Base-flow index grid for the conterminous United States. US Geological Survey Open-File Report 03–263. http://water.usgs.gov/lookup/getspatial?bfi48grd.

Yadav, M., Wagener, T., Gupta, H., 2007. Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins. Adv. Water Resour. 30, 1756–1774. https://doi.org/10.1016/j.advwatres.2007.01.005.

Yaeger, M., Coopersmith, E., Ye, S., Cheng, L., Viglione, A., Sivapalan, M., 2012. Exploring the physical controls of regional patterns of flow duration curves – Part 4: a synthesis of empirical analysis, process modeling and catchment classification. Hydrol. Earth Syst. Sci. 16, 4483–4498. https://doi.org/10.5194/hess-16-4483-2012.

Yaseen, Z.M., Sulaiman, S.O., Deo, R.C., Chau, K., 2019. An enhanced extreme learning machine model for river flow forecasting: state-of-the-art, practical applications in water resource engineering area and future research direction. J. Hydrol. 569, 387–408. https://doi.org/10.1016/j.jhydrol.2018.11.069.

Ye, S., Yaeger, M., Coopersmith, E., Cheng, L., Sivapalan, M., 2012. Exploring the physical controls of regional patterns of flow duration curves – Part 2: role of seasonality, the regime curve, and associated process controls. Hydrol. Earth Syst. Sci. 16, 4447–4465. https://doi.org/10.5194/hess-16-4447-2012.

Yokoo, Y., Sivapalan, M., 2011. Towards reconstruction of the flow duration curve: development of a conceptual framework with a physical basis. Hydrol. Earth Syst. Sci. 15, 2805–2819. https://doi.org/10.5194/hess-15-2805-2011.