Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-639, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 19 December 2016

Hydrology and Earth System Sciences
Open Access
Discussions

Regional regression models of percentile flows for the contiguous US: Expert versus data-driven independent variable selection

Geoffrey Fouad[1], André Skupin[2], Christina L. Tague[3]

[1]Geography Program, Monmouth University, West Long Branch, NJ, USA
[2]Department of Geography, San Diego State University, San Diego, CA, USA
[3]Bren School of Environmental Science and Management, University of California, Santa Barbara, CA, USA

Correspondence to: Geoffrey Fouad (gfouad@monmouth.edu)

Abstract. Percentile flows are statistics derived from the flow duration curve (FDC) that describe the flow equaled or exceeded for a given percent of time. These statistics provide important information for managing rivers, but are often unavailable since most basins are ungauged. A common approach for predicting percentile flows is to deploy regional regression models based on gauged percentile flows and related independent variables derived from physical and climatic data. The first step of this process identifies groups of basins through a cluster analysis of the independent variables, followed by the development of a regression model for each group. This entire process hinges on the independent variables selected to summarize the physical and climatic state of basins. Distributed physical and climatic datasets now exist for the contiguous United States (US). However, it remains unclear how to best represent these data for the development of regional regression models. The study presented here developed regional regression models for the contiguous US, and evaluated the effect of different approaches for selecting the initial set of independent variables on the predictive performance of the regional regression models. An expert assessment of the dominant controls on the FDC was used to identify a small set of independent variables likely related to percentile flows. A data-driven approach was also applied to evaluate two larger sets of variables that consist of either (1) the averages of data for each basin or (2) both the averages and statistical distribution of basin data distributed in space and time. The small set of variables from the expert assessment of the FDC and two larger sets of variables for the data-driven approach were each applied for a regional regression procedure. Differences in predictive performance were evaluated using 184 validation basins withheld from regression model development. The small set of independent variables selected through expert assessment produced similar, if not better, performance than the two larger sets of variables. A parsimonious set of variables only consisted of mean annual precipitation, potential evapotranspiration, and baseflow index. Additional variables in the two larger sets of variables added little to no predictive information. Regional regression models based on the parsimonious set of variables were developed using 734 calibration basins, and were converted into a tool for predicting 13 percentile flows in the contiguous US. Supplementary Material for this paper includes an R graphical user interface for predicting the percentile flows of basins within the range of conditions used to calibrate the regression models. The equations and performance statistics of the models are also supplied in tabular form.

1 Introduction

The flow duration curve (FDC) is composed of percentile flows that identify the flow equaled or exceeded for a given percent of time. Percentile flows are used to make decisions for streamflow applications, such as hydropower, wastewater dilution, and water abstractions (Vogel and Fennessey, 1995). These applications are often conducted without observed percentile flows as most basins are ungauged. In this case, regionalization procedures are typically adopted to predict percentile flows based on information from gauged basins.

A common type of regionalization procedure develops regression models that relate observed percentile flows to independent variables derived from physical and climatic basin data (see Hope and Bart, 2011; Mohamoud, 2008; Over et al., 2014). Hydrologic models based on predicted parameters and climatic forcing data are an alternative to derive percentile flows (Westerberg et al., 2014). However, the simplicity of using regression models presents an opportunity to provide a tool to predict percentile flows for ungauged basins. Regression models are known to perform poorly for study areas with a large variance in percentile flows, such as the contiguous United States (US). To reduce the variance in percentile flows, separate regression models are developed for groups of basins in a process called *regional regression modeling* (Sauquet and Catalogne, 2011). A typical regional regression first splits the basins into groups using cluster analysis and then develops regression models for each group. Independent variables, such as mean elevation and precipitation, are used to identify the groups and parameters of the regression models.

Regional regression modeling has been used to predict percentile flows in the US (see Archfield et al., 2007; Hope and Bart, 2011; Mohamoud, 2008). These studies have focused on particular geographic regions of the US, such as southern New England (Archfield et al., 2007), southern and central California (Hope and Bart, 2011), and the mid-Atlantic (Mohamoud, 2008). The National Streamflow Statistics Program of the US Geological Survey has published regional regression equations for individual states (see Ries (2007) for a summary of the program). Although previous studies have been performed for parts of the US, physical and climatic data now exist to develop regional regression models for the contiguous US. Independent variables could be derived to cluster basins in the contiguous US and develop regression models for subsequent groups of basins. The resulting models could be used as a tool to predict percentile flows for ungauged basins in the contiguous US.

1.1 Independent variables for regional regression modeling

Independent variables summarize physical and climatic basin data, and serve as the foundation for regional regression modeling. Both steps of regional regression modeling (basin grouping and model development) depend on the independent variables chosen to represent the basins. Despite the importance of independent variables, prior studies have primarily evaluated the use of different cluster analyses (see Di Prinzio et al., 2011; Laaha and Blöschl, 2006; Sauquet and Catalogne, 2011) and modeling methods (see Archfield et al., 2007; Holmes et al., 2002; Over et al., 2014), rather than the input information of the overall regional regression approach.

A limited number of studies have examined the input information for regional regression modeling of percentile flows. These studies have either investigated how to select independent variables from a large number of possible variables (Ssegane et al., 2012a,b) or experimented with the initial set of variables to assess the predictive potential of a certain type of variable (Hope and Bart, 2012; Ilorme, 2011). A two-part study compared the performance of different variable selection methods for clustering basins (Ssegane et al., 2012a) and modeling percentile flows (Ssegane et al., 2012b). These studies revealed the importance of different variables through the variable selection process (i.e. more important variables were selected more often). The importance of different variables can also be evaluated by changing the initial set of variables and assessing the difference in model performance. This approach has been used to evaluate the importance of variables describing vegetation cover (Hope and Bart, 2012) and the spatial distribution of land surface data (Ilorme, 2011). Both of these studies reported only minor differences to model performance after changing the initial set of variables. However, the study involving the spatial distribution of land surface data could be expanded to include information on climate and geology as these are dominant controls on the FDC (Yokoo and Sivapalan, 2011). Studies that evaluate different types of variables, such as Hope and Bart (2012) and Ilorme (2011), highlight the uncertainty of selecting the initial set of variables for a regional regression approach.

The approach for selecting the initial set of variables has evolved over the long history of regional regression studies. Early studies used a small number of variables due to the scarcity of spatially distributed data (see Dingman, 1978; Mimikou and Kaemaki, 1985; Singh, 1971). These studies included all of the variables in the regression models, and attempted to target variables with a strong physical connection to the FDC. This early approach to selecting variables could be implemented using contemporary data sources that provide more physical and climatic information. Variables derived from such data could be selected through an expert assessment of the dominant controls on the FDC.

An *expert assessment* of the FDC would select a small number of variables according to a physical understanding of the curve, which can be summarized as follows: the highest and lowest flows are respectively generated by storms and subsurface drainage, and flows in between are a mixture of these sources moderated by evapotranspiration losses (summary based on the work of Yokoo and Sivapalan, 2011). With this understanding, the climatic controls of the FDC could be summarized by *mean annual precipitation* (MAP) and *potential evapotranspiration* (PET), while subsurface drainage could be represented by *baseflow index* (BFI) values that describe the percent of streamflow attributed to groundwater discharge. Although BFI values are derived at gauged points, they can be interpolated to produce spatially distributed data, such as a gridded product for the contiguous US (Wolock, 2003). This data can then be used to create independent variables for regional regression models.

The growth of spatially distributed data has prompted recent regional regression studies to use a large number of independent variables (e.g. Over et al. (2014) evaluated 21 variables) in a *data-driven* approach that attempts to account for more complex and nuanced relations to percentile flows than is presumably provided by the limited number of variables identified through expert assessment. All of the variables are first used to cluster the basins, and then a subset of the variables is selected to model the percentile flows for each group of basins (see Di Prinzio et al. (2011), Laaha and Blöschl (2006), and

Sauquet and Catalogne (2011) for examples). The variables used in these studies may describe the average of the basin data via a single, *lumped* value or the statistical distribution of the basin data via multiple, *distributed* values. The latter set of variables describes the spatial distribution of physical data, such as topography and geology, and the temporal distribution of climatic data. This information describes factors potentially associated with streamflow generation, such as the variability of subsurface drainage conditions (Tague and Grant, 2004) or dispersion of precipitation throughout the year (i.e. seasonality; Ye et al., 2012). Distributed variables produce the largest set of variables for regional regression modeling, and are thought to be advantageous for accommodating a large variety of relations to the percentile flows.

2 Research objective and question

The objective of this research was to create regional regression models for predicting percentile flows in the contiguous US. The steps to complete this objective included (1) grouping basins and (2) developing regression models for each group of basins. Both of these steps were based on independent variables that summarized the physical and climatic data of the basins. The approach used to select the independent variables may influence the performance of the regression models. A small number of variables could be selected according to an expert assessment of the dominant controls on the FDC, or a data-driven approach could be adopted to account for many possible relations to the percentile flows using a large number of variables. Both of these approaches were applied to create the regional regression models. The difference in performance was then evaluated to answer the following research question:

How does the performance of regional regression models for predicting percentile flows differ when using an expert assessment to select a small number of variables versus a data-driven approach involving a large number of variables?

The hypothesis investigated in this study was that the data-driven approach would produce better regression models because the large number of variables may account for nuanced relations to the percentile flows not anticipated by the expert assessment. A performance evaluation was conducted to test this hypothesis and identify a parsimonious approach for creating the regional regression models. These models were then used to develop a tool for predicting the percentile flows of ungauged basins in the contiguous US.

3 Methods

3.1 Overview

Regional regression models were created using three different sets of independent variables:

(1) A limited number of variables identified through e*xpert assessment*,

(2) an expanded number of *lumped* variables, and

(3) a larger number of *distributed* variables.

The first set of variables was selected based on the expert assessment outlined in the Introduction, and included MAP, PET, and BFI (herein referred to as *expert variables*). All of the expert variables were used in the regional regression models. Larger sets of variables were used in a data-driven approach to identify the regional regression models. A set of lumped variables was used to describe the averages of data for each basin, while distributed variables described both the average and distribution of the basin data in space and time. A subset of the lumped and distributed variables was selected for the regional regression models using a regression tree method called random forests (Breiman, 2001) to rank the predictive potential of the variables.

The different sets of variables were used in a cluster analysis to split the basins into groups. As a precursor to cluster analysis, the variables were fed through a neural network called the self-organizing map (SOM). This is an increasingly popular step to reduce noise in hydrologic data (i.e. variance unrelated to the actual value) and account for non-linearities in the cluster analysis (see Boscarello et al., 2015; Di Prinzio et al., 2011; Toth, 2013). With each basin characterized by *n* variables, SOM neural network training transforms *n*-dimensional basin vectors into *n*-dimensional neuron vectors. Those neuron vectors then become the subject of multivariate clustering, which ultimately leads to the grouping of basins. The *k*-means clustering method was applied to the neuron vectors as it identifies clusters similar to how the data is organized in the SOM (Skupin, 2004). The basins were then assigned to the neuron clusters using the neuron with the vector that best-matched the basin data (i.e. best-matching unit).

Groups of basins identified based on the SOM were used to develop regression models for predicting percentile flows. This was accomplished using a set of calibration basins, and an independent set of validation basins was used to evaluate the performance of the regression models. The entire process of (1) identifying groups of basins, (2) developing regression models, and (3) evaluating their performance was repeated using the three different sets of independent variables (expert, lumped, and distributed). The performance evaluation was used to identify a parsimonious set of variables for creating the regional regression models, and a tool based on these models was developed to predict percentile flows for the contiguous US. The entire regional regression study is summarized as a flow chart in Fig. 1.

3.2 Basins and percentile flows

All basins used in this study were located in the contiguous US, and were selected based on being classified as "near-natural" by the US Geological Survey's GAGES-II database. The near-natural class consists of basins with little human influences to the flow of water (see Falcone (2011) for more details). Near-natural basins with at least 30 years of continuous daily streamflow data were used to calculate 13 percentile flows including the high flows exceeded 1 and 5 % of the time (Q_{01} and Q_{05}), low flows exceeded 95 and 99 % of the time (Q_{95} and Q_{99}), and decile values between those flows (Q_{10}, Q_{20},…Q_{90}). Streamflow data was downloaded from the National Water Information System (http://waterdata.usgs.gov/nwis). Daily data for 30 years was used to calculate percentile flows reasonably stable for different time periods (Kennard et al., 2010). Percentile flows were calculated for 918 near-natural basins using the Weibull plotting position to identify the percent of time that a given flow was equaled or exceeded (p) as follows:

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

where r is the rank of the daily flow according to its magnitude, n is the total number of daily flow values, and the flows were normalized using the mean of nonzero values as in Hope and Bart (2011) to control for differences in magnitude between the basins.

A subset of validation basins was used to evaluate the performance of the regression models. The validation included 184 (20 %) of the basins, which meets the recommendation that the validation should have at least 100 samples for a continuous dependent variable, such as percentile flows (Harrell, 2001). The sample of validation basins was selected using a "proxy-basin" approach to identify a sample of basins representative of the remaining calibration basins used to develop the regression models (Klemeš, 1986). The representative sample was identified using a stratified random sample based on independent variables to avoid corrupting the validation. The independent variables used to stratify the basins were thought to be indicative of major controls on the FDC, and consisted of the broadest Köppen climate classes (Peel et al., 2007), the three major rock types (Reed and Bush, 2007), and drainage area categories. The validation basins were then randomly selected within the strata, and the remaining calibration basins had a similar distribution of independent variables according to descriptive statistics and statistical tests (Kolmogorov-Smirnov and Mann-Whitney). A map of the calibration and validation basins is provided in Fig. 2.

3.3 Independent variables

Independent variables describing topography, land cover, soil, geology, and climate were used to develop regional regression models for predicting the percentile flows. Topographic variables were derived from the National Elevation Dataset (NED) 1 arc-second (~ 30-m) grid (http://ned.usgs.gov), and the stream channels for calculating additional topographic variables were acquired from GAGES-II or the National Hydrography Dataset Plus Version 2 (NHDPlusV2) 1:100,000-scale product (http://www.nhdplus.com). Land cover was assessed using the 30-m National Land Cover Dataset (NLCD) for the year 1992 (Vogelmann et al., 2001), as this was the year of the NLCD that coincided with streamflow data from the most basins. The NLCD was used to calculate percent forest cover since synthesis of paired catchment experiments identifies a strong relation between forest cover and annual flow (Brown et al., 2005). Soil variables were calculated using a multilayer soil characteristics dataset for the contiguous US (CONUS-SOIL), which provides the State Soil Geographic Database (STATSGO) as 1-km grids (Miller and White, 1998). Geology was summarized using the BFI as the two are known to be strongly correlated (Price, 2011). Estimates of BFI were previously generated by Wolock (2003) for the contiguous US (BFI48GRD). The 1-km grid was spatially interpolated using BFI values at 8,249 gauges with at least ten years of daily streamflow data (see Wolock (2003) for more information on the methodology). Although the BFI grid was produced using gauged data, it is a pre-existing dataset that can be used to create independent variables for predicting the percentile flows of ungauged basins, as previously demonstrated (Dudley, 2015; Hope and Bart, 2011; Yuan, 2013).

Climatic variables were calculated using 30 years of the *Precipitation-elevation Regressions on Independent Slopes Model* (PRISM) 4-km grids (http://prism.oregonstate.edu). The only exception was a GAGES-II variable for the average percent of precipitation delivered as snow from 1901-2000 (Percent_Snow). The other climatic variables used monthly PRISM data concurrent with the streamflow data for each basin or daily PRISM data from 1981-2010 as the daily data was unavailable for all of the streamflow data and the chosen time period overlapped with the most streamflow data. Precipitation depths (mm) were weighted according to the fraction of the grid cell located within the basin boundary.

The independent variables were organized into three different sets of variables listed in Table 1 and named *expert* (E), *lumped* (L), and *distributed* (D). These different sets of variables may include some of the same variables, but the number of variables increased for each successive dataset. The expert variables included MAP, PET, and BFI based on expert assessment of the FDC, as discussed in the Introduction. The larger sets of 22 lumped and 37 distributed variables were used for a data-driven approach to identify the regional regression models. The lumped variables mainly described the average of the basin data, while the distributed variables expanded on this information using the following types of variables: (1) the standard deviation of gridded physical data, such as elevation, and annual climatic statistics, such as precipitation intensity (mm d^{-1}), (2) the percent forest cover in riparian corridors since they are critical areas for groundwater discharge (Hope et al., 2009), and (3) the amplitude and peak timing of monthly precipitation and PET data described using the lag-1 autocorrelation coefficient (Toth, 2013), circular statistics (Dingman, 2002), and first term of the Fourier transform (Dalton, 2005).

3.4 Cluster analysis

Individual basins are assigned to groups through cluster analysis of independent variables converted into z-scores with a mean of zero and variance of one in order to give variables on different scales comparable weight. The z-scores were used as the weights of input vectors for training the SOM, which was composed of hexagonal neurons (i.e. each neuron has six neighbors) arranged in a two-dimensional grid. The number of neurons in the grid was chosen to be significantly larger than the anticipated number of clusters in order to avoid individual neurons acting as cluster centroids. Neurons are later linked to basins through computation of the similarity of basin input vectors and neuron output vectors. “Empty” neurons not linked to any of the basins through strong similarity were deemed unrepresentative of the input data, and limited for the cluster analysis. Preliminary experiments were performed on the total number of neurons in order to limit the occurrence of empty neurons, and this led to the choice of a 15x15-neuron SOM for all training.

Prior to training, the neurons were given a random vector of values equal in length to the number of input variables. The neuron vectors were adjusted through an iterative training process that presented the basin data to the SOM and assigned it to the most similar neuron according to the Euclidean distance metric. The receiving neuron, or best-matching unit (BMU), and its neighbors were modified to more closely match the incoming data using a Gaussian neighborhood function and a learning rate that decreased the magnitude of the modifications as the training proceeded. Neural network training was performed with the SOM Toolbox (http://www.cis.hut.fi/projects/somtoolbox), using techniques described by Vesanto et al. (2000).

The SOM was trained using a global and local stage as recommended by Kohonen (1990). The global stage used a large neighborhood size (8 neurons), relative to the size of the SOM (15x15 neurons). A large learning rate (0.04), which decreased over a short number of runs (50), was used during global training to reveal large structures in the data. Smaller clusters were then distinguished using a smaller neighborhood (5 neurons) and learning rate (0.03) for a longer number of runs (4,000) during local training. Both training stages were run until the neuron vectors converged on the basin data (i.e. the difference between the neurons and basins no longer decreased).

The trained neuron vectors were then clustered using the *k*-means method. Cluster centroids were initially given a random vector of values, and the neuron vectors were assigned to the cluster centroids according to the Euclidean distance metric. The cluster centroids were then recalculated using the mean of the neuron vectors assigned to each cluster. This process continued until the cluster centroids no longer changed, and was repeated 1,000 times to prevent the randomly initiated cluster centroids from influencing the performance of the clustering. The final cluster solution had the minimum sum of squared Euclidean distances between the cluster centroids and neuron vectors. The basins were assigned to the neuron clusters according to their BMU (i.e. neuron with the shortest Euclidean distance to the basin).

The number of clusters was evaluated for the cluster solutions with 2-50 clusters. The criteria for evaluating the number of clusters were (1) the number of calibration basins per cluster available to develop subsequent regression models and (2) the validity of the clusters in terms of their compactness and separation. The minimum number of calibration basins per cluster was 20 as recommended by Hosking and Wallis (1997) for regional streamflow predictions. This served as a limit on the number of clusters. The validity of the different number of clusters was evaluated using the following indices defined as in Desgraupes (2013): silhouette, Davies-Bouldin, Xie-Beni, Calinski-Harabasz, and Dunn. The number of clusters identified using the indices was limited to solutions that returned at least 20 basins per cluster. From these solutions, the largest number of clusters was chosen to contend with the large variability of basins in the contiguous US. The different sets of variables (Table 1) resulted in different numbers of clusters, and the largest number was chosen so that each set of variables had the same number of clusters. The cluster analysis for each set of variables split the basins into 14 groups.

3.5 Regression model development

Regression models were developed for each group of basins. The number of calibration basins in the group dictated the number of independent variables used for the model. An independent variable was used for every ten calibration basins as this provides an adequate sample size for estimating regression model parameters (Austin and Steyerberg, 2015). The regression models were able to use all three of the variables from the expert assessment of the FDC (> 30 calibration basins per group), but a subset of the larger sets of variables (lumped and distributed) had to be selected for the regression models.

The lumped and distributed variables were selected using random forests because this approach can be used to estimate the importance of each variable. Random forests were generated for each group using the calibration basins and their percentile flows. The calibration basins were randomly sampled to grow regression trees until the error of the percentile flow predictions stabilized for the out-of-bag sample (i.e. calibration basins excluded from the tree). The regression trees recursively split the calibration basins into smaller groups using a series of rules based on the independent variables. The percentile flows of the smallest groups were averaged to generate predictions for the out-of-bag sample. The mean squared error (MSE) of out-of-bag predictions was used as an estimate of variable importance. Each variable was randomly permuted (or essentially removed) to grow the regression trees, and the increase in MSE signified the importance of the variable. The variable rankings derived from this process may change due to the random samples used to grow the regression trees. As a result, the entire process was repeated using 100 random forests, and the mean increase in MSE was used to rank the lumped and distributed variables for each group of basins. The lowest ranked variable included in the regression model was determined by the number of calibration basins in the group as previously described.

The independent variables selected for each group of basins were used in a regression model of the form:

$$ln(Q_i) = \beta_0 + \beta_1 I_1 \ldots + \beta_j I_j \qquad (2)$$

where the dependent variable (Q_i) is the percentile flow i transformed using the natural log to reduce the skew of the flows and their potential to violate the assumption of homoscedasticity (i.e. evenly varying model residuals), I_1-I_j are the untransformed or natural log-transformed independent variables whether they had a non-linear or linear relation to the percentile flow i, and β_0-β_j are the parameters of the regression model estimated using the ordinary least squares method. Separate regression models were developed to predict each of the 13 percentile flows ($Q_{01,}$ Q_{05}, Q_{10}, Q_{20},…Q_{95}, Q_{99}).

3.6 Performance evaluation

The performance of the regional regression models was evaluated using the percentile flows of the validation basins excluded from regression model development. The natural log of the percentile flows was applied to reduce the potential influence of large flows. The predictive performance for each percentile flow was quantified using the relative error (E), coefficient of determination (R^2), and Nash and Sutcliffe (1970) efficiency (N) calculated as:

$$E = \left|\frac{P_b - O_b}{O_b + 1}\right| \qquad (3)$$

$$R^2 = \left(\frac{\sum_{b=1}^{n}(O_b - \bar{O})(P_b - \bar{P})}{\sqrt{\sum_{b=1}^{n}(O_b - \bar{O})^2}\sqrt{\sum_{b=1}^{n}(P_b - \bar{P})^2}}\right)^2 \qquad (4)$$

$$N = 1 - \frac{\sum_{b=1}^{n}(O_b - P_b)^2}{\sum_{b=1}^{n}(O_b - \bar{O})^2} \qquad (5)$$

where O_b and P_b are respectively the observed and predicted percentile flow for basin b, $\bar{O}$ and $\bar{P}$ are the mean of the observed and predicted percentile flows, respectively, and n is the number of validation basins. A constant of one was added to the denominator of E to accommodate zero flows, and the absolute value was used to calculate the sum of E for each percentile flow.

4 Results

The study basins were split into 14 groups using three different sets of variables (Fig. 3). The groups were largely geographically contiguous, although the independent variables did not describe the location of the basins. The geographic contiguity of the groups signifies that spatial proximity is a strong indicator of similarity between the independent variables. Notable exceptions were distant areas with similar physical and climatic conditions, such as southern Appalachia and northern California (group 6 of Figs. 3a and b), and mountainous areas with sharp changes in elevation, such as the Pacific Northwest (groups 5 and 10 of Fig. 3). The groups derived using the different sets of variables had some major differences, such as an additional group for the Rocky Mountains when information on snow (Percent_Snow) was included (group 13 of Figs. 3b and c) and the loss of a group in the Appalachian Mountains (group 9 of Figs. 3a and b) when using the distributed variables. The differences between the groups could be further characterized by an exploratory analysis like that of Ley et al. (2011), but the present study was concerned with the effect of using the different sets of variables on regional regression models for predicting percentile flows. The rest of the results therefore examine the independent variables used for the regression models, and compare the performance of the regression models developed using the different sets of variables to evaluate the hypothesis that a data-driven approach with a large number of variables is more effective than an expert assessment of the FDC that uses a small number of variables.

4.1 Independent variables selected for the regression models using random forests

The data-driven approach used random forests to select a subset of the most important lumped or distributed variables for the regression models, and the percent of the models that included each variable was used to rank the importance of the variables for predicting high (Q_{01}-Q_{20}), average (Q_{30}-Q_{70}), and low (Q_{80}-Q_{99}) percentile flows (Table 2). The most frequently selected variable was BFI. The expert assessment of the FDC identified BFI as an important variable for flows other than the highest flows generated by storms, but BFI was the most important variable for predicting the entire FDC including the highest flows. The other variables selected by the expert assessment of the FDC (MAP and PET) were among the top five most selected variables. These variables were also frequently used when combined as Aridity (PET/MAP).

The remaining frequently selected variables in Table 2 described snow accumulation and melt (Percent_Snow and Spring_Temp), subsurface drainage (Poorly_Drained), and mean elevation (Elev). The snow-related variables were frequently used for snow-dominated groups. For instance, the Rocky Mountains (groups 12 and 13) included the snow-related variables in 62 % of the models. The importance of subsurface drainage was previously highlighted by BFI, and is further demonstrated by the frequent use of a variable describing the percent of the basin covered in poorly drained (NRCS (2007) groups C and D) soils (Poorly_Drained). Mean elevation was frequently used for the regression models, but other topographic variables, such as mean slope and drainage density, were not frequently used, which is noteworthy as these variables have been widely used to predict percentile flows.

The distributed variables included additional variables for describing the within-basin statistical distribution of independent variables, but these variables were largely absent from the regression models (Table 2). Only two of these variables (Precip_Seasonality and Aridity_SD) were among the top five most selected variables. This indicates that using the distributed variables had little effect on the performance of the regression models (see the next section for those results). The more frequently used lumped variables (i.e. basin averages) were stronger predictors of the percentile flows, and the additional variables on the distribution of the basin data may have been statistically redundant (i.e. cross-correlated with the lumped variables).

The statistical redundancy of the variables was evaluated using the condition number (CN) of the regression models (Belsley et al., 2004), with a CN > 30 signifying the presence of redundant variables. All of the regression models included redundant variables as indicated by the minimum CN > 30 (Table 3). A large CN may be a problem for transferring the regression model to new (validation) data (Kroll and Song, 2013), but this was not a concern here since the validation used a representative sample of the basins. The CN was used here to evaluate the redundancy of the variables with the greatest predictive potential selected for the regression models. These variables were highly redundant, and the degree of redundancy increased for the distributed variables according to the mean CN. This once again indicates that the distributed variables added little predictive information to the regression models (i.e. the extra variables were statistically redundant), and the contribution of the distributed variables to the performance of the regression models is evaluated in the following section.

4.2 Regression model performance

The performance of the regression models was evaluated to identify a parsimonious set of variables (expert, lumped, or distributed) and create a tool for predicting percentile flows in the contiguous US. Regression model performance in calibration (goodness-of-fit) was assessed using adjusted R^2 to compare models with a different number of independent variables. The distribution of adjusted R^2 values is shown in Fig. 4. The smallest adjusted R^2 values of the regression models were produced using the smallest set of variables (expert), while the larger sets of variables (lumped and distributed) had similar adjusted R^2 values. Although adjusted R^2 considers the number of variables in a model, it can still favor larger models (Greene, 2003). Calibration performance as measured by adjusted R^2 was influenced by the number of variables in the models. An adjusted $R^2 \geq 0.75$ ("good" according to Castellarin et al., 2004) was obtained by 3.3, 11.0, and 13.7 % of the models developed using the expert, lumped, and distributed variables, respectively. The same order of variables explained over half of the variance in the percentile flows for 25.8, 62.1, and 61.0 % of the models. The median adjusted R^2 of the models developed using the expert, lumped, and distributed variables was respectively 0.34, 0.57, and 0.58.

The predictive performance of the regression models was evaluated using validation basins and summarized for each percentile flow using the sum of absolute relative error (E), R^2, and Nash-Sutcliffe efficiency (N) (Table 4). Smaller E values signify better predictive performance, whereas R^2 and N have a maximum of 1 indicating perfect performance. The sum of absolute E was largest for the highest percentile flow (Q_{01}) and smallest for the lowest percentile flow (Q_{99}), and likely influenced by the magnitude of the flow. The other performance metrics (R^2 and N) summarized the predictive performance for the percentile flows independent of their magnitude. The N values were slightly less than the R^2 values, and ranged from 0.39-0.76. Predictive performance peaked toward the middle of the FDC (average flows), and declined at the tails (high and low flows).

The different approaches for selecting the independent variables (expert assessment or data-driven) influenced the performance of the regression models. The data-driven approach performed better using the lumped variables, as opposed to the distributed variables, for almost all of the percentile flows (Table 4). Therefore, the additional information of the distributed variables (i.e. the statistical distribution of the basin data) did not contribute to the performance of the regression models. This may be because the additional information was statistically redundant (Table 3) or produced variables that were not strong predictors of the percentile flows (Table 2).

The hypothesis that the data-driven approach would outperform the expert assessment of the FDC was contradicted by the predictive performance for the percentile flows. Although the data-driven approach performed better in calibration (Fig. 4), expert assessment of the FDC achieved similar performance to the data-driven approach in validation (Table 4). The percentile flows at the tails of the FDC (high and low flows) were predicted better using the three variables of the expert assessment, whereas the larger set of lumped variables slightly improved the predictions for the percentile flows in the middle of the FDC. The three variables of the expert assessment (MAP, PET, and BFI) explained most of the variance in the percentile flows, and little to no additional variance was explained using the larger sets of variables for the data-driven approach.

The expert assessment required less computational effort (i.e. calculating and selecting variables), and produced simple regression models that only need three values to predict the percentile flows for ungauged basins. These models had similar overall performance to the more complex data-driven models according to the sum of absolute E and mean R^2 and N for all the percentile flows (Table 5). Due to the simplicity and overall performance of the expert assessment, it was identified as a parsimonious approach for creating the regional regression models, and these models were used to develop a tool for predicting the percentile flows in the contiguous US (see the Supplementary Material section).

5 Discussion

5.1 Important variables for predicting the FDC

The most important variable for predicting the entire FDC was BFI (Table 2). The expert assessment of the FDC linked BFI to average and low flows at least partially supplied by subsurface drainage (Yokoo and Sivapalan, 2011), but BFI may also be related to the excess precipitation of a basin (MAP-PET), explaining its ability to predict high flows. Larger BFI values and high flows may be expected for basins with more excess precipitation, and excess precipitation had a statistically significant (p-value < 0.01) correlation with BFI and the representative high flow of Q_{10} (Spearman's rho of 0.39 and 0.49, respectively). The baseflow of a basin may therefore be indirectly related to its potential to produce high flows. This is in line with previous findings that the process of infiltration, a major control of baseflow, also plays a part in controlling floods (see Gioia et al. (2012) among others).

The top five most selected variables of Table 2 included the other variables from the expert assessment of the FDC (MAP and PET). A substitute for these variables may be Aridity (PET/MAP) since it was also frequently selected. Aridity is a measure of the long-term water balance that generally represents the proportion of incoming precipitation lost to evapotranspiration. High to low flows of the FDC were related to Aridity because it influences antecedent moisture and subsurface drainage. Antecedent moisture moderates the higher flows generated by storms (Muneepeerakul et al., 2010), whereas subsurface drainage provides the lower flows between storms (Botter et al., 2008). The long-term water balance expressed as Aridity has been linked to the variation of the FDC in the contiguous US (Cheng et al., 2012), and may be a more effective means of representing MAP and PET.

Other important variables for predicting the FDC (Table 2) included variables representing snow accumulation and melt (Percent_Snow and Spring_Temp), subsurface drainage (Poorly_Drained), and mean elevation (Elev). Snow accumulation and melt were important for snow-dominated basins, and closely associated with high flows generated by spring snowmelt (Rosenberg et al., 2013). Subsurface drainage represented via poorly drained soils was most strongly related to the low flows. A larger percent of poorly drained soils with less subsurface drainage reduced the low flows as illustrated by the statistically significant (*p*-value < 0.01) negative correlation between Poorly_Drained and Q_{90} (Pearson's *r* of -0.38). Elevation is an integrative variable that was likely important because it covaries with other important factors, such as precipitation (Daly et al., 2008), snow accumulation and melt (Grünewald et al., 2014), and subsurface drainage (Schaller and Fan, 2009).

Largely absent from the important variables listed in Table 2 are the distributed variables that describe the statistical distribution of spatial and temporal basin data. This is either because basin averages provide better information for predicting percentile flows (Table 2) or distributed variables mainly add redundant information to the regression models (Table 3).

5.2 Improving performance of percentile flow predictions

The performance of the regression models was strongest for percentile flows in the middle of the FDC (average flows) and poorest for percentile flows toward the tails of the FDC (high and low flows). This is a commonly noted pattern in other FDC regionalization studies (see Hope and Bart, 2012; Mohamoud, 2008; Sauquet and Catalogne, 2011). Poorer performance toward the tails may be attributed to the large variability of basin responses to storms that generate high flows and the challenge of representing the contribution of subsurface drainage to low flows (Salinas et al., 2013).

The predictive performance for the high and low flows may have been improved by using additional independent variables. High flows are the product of storms, which were represented by intensity (mm d^{-1}) and maximum events (largest 1-day totals). However, these variables (Precip_Intensity and Precip_1D_Max) were not frequently used for the regression models (Table 2), and additional information on the frequency and magnitude of storms may have been useful. This information could be represented using a precipitation duration curve (PDC) derived like a FDC. The PDC has previously been effective at reconstructing the high end of the FDC (Yokoo and Sivapalan, 2011), and may be equally effective at yielding independent variables (i.e. precipitation percentiles) for predicting high flows. The frequency of rainy days (e.g. rainy days y^{-1}) may also be informative as an indicator of average antecedent moisture conditions, which mediate the high flows generated by storms. Physical factors also play a role in mediating high flows through interception and infiltration. Land cover provides readily available information on these physical factors, but may have been underrepresented in this study. Percent forest cover (Forest) was the only land cover variable, and it was not frequently used to predict high flows (Table 2). Additional land cover variables specifically targeting the processes of interception (e.g. tree canopy density) and infiltration (e.g. natural impervious area) may have improved high flow predictions. Such variables could be developed using remote sensing technology and evaluated for their potential to predict high flows.

Additional independent variables may also have benefited the prediction of low flows controlled by subsurface drainage. Variables describing subsurface drainage (BFI and Poorly_Drained) were among the most strongly associated variables to low flows (Table 2). Subsurface drainage could be further characterized using a hydrologically relevant geologic (hydrogeologic) classification. A hydrogeologic classification of the Pacific Northwest was previously linked to summer low flows (Tague and Grant, 2004), and this prompts the hypothesis that groups of basins identified using a hydrogeologic classification may improve subsequent low flow predictions. Subsurface drainage may also be characterized by the storage properties of regional aquifers. These properties (e.g. aquifer thickness) have been mapped for regional aquifers using spatial interpolation methods (see Williams and Dixon (2015) among others), and may indicate the contribution of aquifers to low flows. The low flows of this study included zero flows, which are notably difficult to predict (Snelder et al., 2013) and may require modeling schemes designed to accommodate intermittent streams (Hope and Bart, 2011).

The uncertainty of the percentile flow predictions may be attributed to the use of regression. Regression models are sensitive to outliers and measurement noise (i.e. errors) in the data (Harrell, 2001). Anomalous values, such as outliers or noise, may be given less weight using machine learning methods, such as neural networks, capable of smoothing the data (Dawson and Wilby, 2001). Neural networks are also known for their ability to capture the non-linear relations of hydrologic data (Abrahart and See, 2007), and should be evaluated for the modeling of percentile flows. Regression was applied in this study to produce simple models, which could then be converted into a tool for predicting percentile flows in the contiguous US (see the next section on Supplementary Material).

The expert assessment may have been improved by combining MAP and PET as Aridity since this variable was an effective substitute frequently used in the data-driven regression models (Table 2). The third variable of the expert assessment could then target the more challenging to predict high or low flows. For instance, the high flows may be explained using a variable describing the variability of large storms (e.g. slope of the PDC above the 20th percentile), and the low flows may be associated with the percent of the basin underlain by a general hydrogeologic class, such as unconsolidated material.

6 Supplementary Material – CONUS Percentile Flow Predictor

The Supplementary Material for this paper provides the contiguous US (CONUS) Percentile Flow Predictor, an open source R graphical user interface for predicting 13 percentile flows ($Q_{01,}$ Q_{05}, Q_{10}, Q_{20},…Q_{95}, Q_{99}) of ungauged basins. The tool uses regression models developed based on 734 calibration basins in the contiguous US and a set of parsimonious independent variables identified through expert assessment of the FDC. Input data includes MAP and PET calculated using long-term PRISM data (http://prism.oregonstate.edu) and BFI based on a grid for the contiguous US (http://water.usgs.gov/lookup/getspatial?bfi48grd). For convenience, MAP and PET grids have been calculated using data from 1981-2010, and are provided with the aforementioned BFI grid in the Supplementary Material. Input data should be within the range of the data used to create the CONUS Percentile Flow Predictor (Table 6), and a warning is generated if the input data is outside of this range.

The percentile flows of the ungauged basin are predicted by first assigning the basin to a group (Fig. 3a) and then solving the regression equations for the assigned group. The groups were identified using the SOM, and the ungauged basin is assigned to the SOM neuron with the shortest Euclidean distance between the neuron vector and input data. The group membership of the neuron is then used for the ungauged basin. The regression models of that group are used to predict the percentile flows. The percentile flows were normalized using the mean of nonzero daily flows (i.e. the index flow) and natural log-transformed to develop the regression models. The predictions are converted into cubic meters per second using the index flow and the Duan (1983) smearing estimate to back transform the percentile flows. The index flow is predicted using regression models developed the same way as the percentile flows. Output of the tool can be generated for multiple ungauged basins, and the predictions are accompanied by statistics on the performance of the regression models (adjusted R^2, CN, and standard error (SE) of the model). The predicted percentile flows can be used to reconstruct the complete FDC through an interpolation method such as the one in Mohamoud (2008). The CONUS Percentile Flow Predictor is provided with a metadata file (Readme.txt) including instructions on how to use the tool. Regression models used to create the CONUS Percentile Flow Predictor are provided with associated statistics in tabular form for use in a variety of software packages.

7 Conclusions

Regional regression models were developed to predict percentile flows for the contiguous US. The two steps of a regional regression (i.e. identifying groups of basins and developing models) depend on the independent variables used to summarize the physical and climatic data of the basins. This study compared the following two approaches for selecting the independent variables: (1) an expert assessment of the factors that control the FDC to identify a small number of variables and (2) a data-driven approach with many variables possibly linked to the FDC. The data-driven approach was performed using lumped variables (i.e. basin averages) and a larger set of distributed variables (i.e. basin averages and statistical distribution). The predictive performance of the regression models was evaluated to identify the most parsimonious approach for selecting independent variables.

An underlying hypothesis of this study was that the data-driven approach would produce better regression models for predicting the percentile flows. This was contradicted by the results of the performance evaluation. The predictive performance of the expert assessment (mean N = 0.66) was similar to the data-driven approach using the lumped variables (mean N = 0.65) and slightly better than the data-driven models derived from the distributed variables (mean N = 0.61). The additional information of the distributed variables (i.e. the statistical distribution of the basin data) did not contribute to the regression models because it was redundant and less important than the lumped variables. The expert assessment included three variables (MAP, PET, and BFI), and performed similarly to the 22 lumped variables. This signifies that many of the lumped variables were either redundant or otherwise not useful. With the exception of mean elevation, topographic variables widely used in FDC regionalization studies were not useful predictors of percentile flows. An important predictor of percentile flows was Aridity (PET/MAP), which could have been used as a substitute for MAP and PET in the expert assessment. Another variable alongside Aridity and BFI could be evaluated to possibly improve performance of the expert assessment.

The expert assessment produced simple models that did not decrease predictive performance, and was deemed the parsimonious approach for selecting the independent variables of the regional regression. The regression models can be easily used to predict percentile flows for ungauged basins based on MAP, PET, and BFI, and were used to create a tool for predicting percentile flows in the contiguous US (see the Supplementary Material for this paper). The CONUS Percentile Flow Predictor generates predictions for the 13 percentile flows of this study, along with estimates of the predictive uncertainty.

The regional regression was used to predict high to low percentile flows (Q_{01}- Q_{99}). Predictive performance was the worst for the percentile flows at the tails of the FDC (high and low flows). The highest predictive performance for these flows was obtained using the three variables of the expert assessment, and the other variables used in this study did not improve the high and low flow predictions. Additional variables may have been needed to characterize the magnitude of storms that drive high flows (e.g. precipitation percentiles) and the potential contribution of subsurface drainage to low flows (e.g. aquifer thickness). The development of new variables may also be in order to characterize factors such as the spatial variability of storms (Zoccatelli et al., 2011) and groundwater levels (Costelloe et al., 2015). The low flow predictions may also have been improved using a modeling scheme that takes the probability of zero flows into account (Hope and Bart, 2011).

The percentile flow predictions may have been improved by modeling methods other than regression. The strongest predictor of the percentile flows was BFI, and this was derived from a gridded product for the contiguous US that was spatially interpolated between stream gauges. A similar spatial interpolation approach could be used to predict percentile flows, and has outperformed regression in a previous study (Archfield et al., 2013). The output of the spatial interpolation could be served as a data product for predicting the percentile flows of ungauged basins. Regression may also be outperformed by neural networks that are more resilient to the noise and non-linearity of hydrologic data (Hall et al., 2002).

Neural networks, such as the SOM, could be used to cluster the basins and generate percentile flow predictions in one step. This would eliminate the need to identify groups of basins for percentile flow modeling, and should be evaluated in future studies.

Author contribution. All authors contributed to the design of the study. G. Fouad conducted the research, developed the CONUS Percentile Flow Predictor, and prepared the manuscript with contributions from all co-authors.

Acknowledgements. This work was performed for a doctoral study, and could not have been completed without the advising of Allen Hope at San Diego State University (SDSU). Hugo Loáiciga of the University of California, Santa Barbara (UCSB) also provided helpful comments as a member of the doctoral committee. The doctoral study was completed for the SDSU-UCSB Joint Doctoral Program in Geography. The SDSU Department of Geography provided computational support and helpful computing staff. Barbara Bailey of the SDSU Department of Mathematics and Statistics offered insights on matters of the validation.

Disclaimer. Predictions from the CONUS Percentile Flow Predictor are meant as a first approximation of the percentile flows, and not intended for engineering design of any kind. Users should refer to the governmental standards for predicting percentile flows in the given jurisdiction.

References

Abrahart, R. J. and See, L. M.: Neural network modelling of non-linear hydrological relationships, Hydrol. Earth Syst. Sci., 11, 1563–1579, doi:10.5194/hess-11-1563-2007, 2007.

Archfield, S. A., Pugliese, A., Castellarin, A., Skøien, J. O., and Kiang, J. E.: Topological and canonical kriging for design flood prediction in ungauged catchments: an improvement over a traditional regional regression approach?, Hydrol. Earth Syst. Sci., 17, 1575–1588, doi:10.5194/hess-17-1575-2013, 2013.

Archfield, S. A., Vogel, R. M., and Brandt, S. L.: Estimation of flow-duration curves at ungaged sites in southern New England, in: World Environmental and Water Resources Congress 2007: Restoring Our Natural Habitat, Tampa, FL, 15-19 May 2007, doi:10.1061/40927(243)407, 2007.

Austin, P. C. and Steyerberg, E. W.: The number of subjects per variable required in linear regression analyses, J. Clin. Epidemiol., 68, 627–636, doi:10.1016/j.jclinepi.2014.12.014, 2015.

Belsley, D. A., Kuh, E., and Welsch, R. E.: Detecting and Assessing Collinearity, in: Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, John Wiley and Sons, Hoboken, NJ, 85–191, 2004.

Boscarello, L., Ravazzani, G., Cislaghi, A., and Mancini, M.: Regionalization of Flow-Duration Curves through Catchment Classification with Streamflow Signatures and Physiographic-Climate Indices, J. Hydrol. Eng., 21, 05015027, doi:10.1061/(ASCE)HE.1943-5584.0001307, 2015.

Botter, G., Zanardo, S., Porporato, A., Rodriguez-Iturbe, I., and Rinaldo, A.: Ecohydrological model of flow duration curves and annual minima, Water Resour. Res., 44, W08418, doi:10.1029/2008WR006814, 2008.

Breiman, L.: Random Forests, Mach. Learn., 45, 5-32, doi: 10.1023/A:1010933404324, 2001.

Brown, A. E., Zhang, L., McMahon, T. A., Western, A. W., and Vertessy, R. A.: A review of paired catchment studies for determining changes in water yield resulting from alterations in vegetation, J. Hydrol., 310, 28–61, doi:10.1016/j.jhydrol.2004.12.010, 2005.

Castellarin, A., Galeati, G., Brandimarte, L., Montanari, A., and Brath, A.: Regional flow-duration curves: reliability for ungauged basins, Adv. Water Resour., 27, 953–965, doi:10.1016/j.advwatres.2004.08.005, 2004.

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

Costelloe, J. F., Peterson, T. J., Halbert, K., Western, A. W., and McDonnell, J. J.: Groundwater surface mapping informs sources of catchment baseflow, Hydrol. Earth Syst. Sci., 19, 1599–1613, doi:10.5194/hess-19-1599-2015, 2015.

Dalton, K. L.: Variation in timing of vegetation peak greenness on the north slope of Alaska, 1982-1999, M.S. thesis, Department of Geography, San Diego State University, USA, 75 pp., 2005.

Daly, C., Halbleib, M., Smith, J. I., Gibson, W. P., Doggett, M. K., Taylor, G. H., Curtis, J., and Pasteris, P. P.: Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States, Int. J. Climatol., 28, 2031–2064, doi:10.1002/joc.1688, 2008.

Dawson, C. W. and Wilby, R. L.: Hydrological modelling using artificial neural networks, Prog. Phys. Geog., 25, 80–108, doi:10.1177/030913330102500104, 2001.

Desgraupes, B.: Clustering Indices, University Paris Ouest, Nanterre, France, 2013.

Dingman, S. L.: Synthesis of flow-duration curves for unregulated streams in New Hampshire, Water Resour. Bull., 14, 1481–1502, doi:10.1111/j.1752-1688.1978.tb02298.x, 1978.

Dingman, S. L.: Precipitation, in: Physical hydrology, Prentice Hall, Upper Saddle River, NJ, 94–165, 2002.

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

Duan, N.: Smearing Estimate: A Nonparametric Retransformation Method, J. Am. Stat. Assoc., 78, 605–610, doi:10.1080/01621459.1983.10478017, 1983.

Dudley, R. W.: Regression Equations for Monthly and Annual Mean and Selected Percentile Streamflows for Ungaged Rivers in Maine: US Geological Survey, Scientific Investigations Report 2015–5151, 35 p., doi:10.3133/sir20155151, 2015.

Falcone, J. A.: GAGES-II: Geospatial Attributes of Gages for Evaluating Streamflow, Digital Spatial Dataset, US Geological Survey, available at: http://water.usgs.gov/GIS/metadata/usgswrd/XML/gagesII_Sept2011.xml, 2011.

Gioia, A., Iacobellis, V., Manfreda, S., and Fiorentino, M.: Influence of infiltration and soil storage capacity on the skewness of the annual maximum flood peaks in a theoretically derived distribution, Hydrol. Earth Syst. Sci., 16, 937–951, doi:10.5194/hess-16-937-2012, 2012.

Greene, W.H.: Least Squares, in: Econometric Analysis, Prentice Hall, Upper Saddle River, NJ, 19-40, 2003.

Grünewald, T., Bühler, Y., and Lehning, M.: Elevation dependency of mountain snow depth. Cryosphere, 8, 2381-2394, doi: 10.5194/tc-8-2381-2014, 2014.

Hall, M. J., Minns, A. W., and Ashrafuzzaman, A. K. M.: The application of data mining techniques for the regionalisation of hydrological variables, Hydrol. Earth Syst. Sci., 6, 685–694, doi:10.5194/hess-6-685-2002, 2002.

Harrell, F. E.: Multivariable Modeling Strategies, in: Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis, Springer, New York, NY, 53–86, 2001.

Holmes, M. G. R., Young, A. R., Gustard, A., and Grew, R.: A region of influence approach to predicting flow duration curves within ungauged catchments, Hydrol. Earth Syst. Sci., 6, 721–731, doi:10.5194/hess-6-721-2002, 2002.

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

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

Hope, A., Burvall, A., Germishuyse, T., and Newby, T.: River flow response to changes in vegetation cover in a South African fynbos catchment, Water SA, 35, 55–60, doi:10.4314/wsa.v35i1.76652, 2009.

Hosking, J. R. M. and Wallis, J. R.: Identification of homogeneous regions, in: Regional Frequency Analysis: An Approach Based on L-Moments, Cambridge University Press, Cambridge, UK, 54–72, 1997.

Ilorme, F.: Delineation of Hydrologically Homogeneous Regions Using Spatially Distributed Data, in: Development of a Physically-based Method for Delineation of Hydrologically Homogeneous Regions and Flood Quantile Estimation in Ungauged Basins Via the Index Flood Method, Ph.D. thesis, Department of Civil and Environmental Engineering, Michigan Technological University, USA, 97–119, 2011.

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

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

Kohonen, T.: The Self-Organizing Map, P. IEEE, 78, 1464–1480, doi:10.1109/5.58325, 1990.

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

Laaha, G. and Blöschl, G.: A comparison of low flow regionalisation methods – catchment grouping, J. Hydrol., 323, 193–214, doi:10.1016/j.jhydrol.2005.09.001, 2006.

Ley, R., Casper, M. C., Hellebrand, H., and Merz, R.: Catchment classification by runoff behavior with self-organizing maps (SOM), Hydrol. Earth Syst. Sci., 15, 2947–2962, doi:10.5194/hess-15-2947-2011, 2011.

Miller, D. A. and White, R. A.: A Conterminous United States Multilayer Soil Characteristics Dataset for Regional Climate and Hydrology Modeling, Earth Interact., 2, 2-002, doi:10.1175/1087-3562(1998)002<0001:ACUSMS>2.3.CO;2, 1998.

Mimikou, M. and Kaemaki, S.: Regionalization of flow duration characteristics, J. Hydrol., 82, 77–91, doi:10.1016/0022-1694(85)90048-4, 1985.

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

Muneepeerakul, R., Azaele, S., Botter, G., Rinaldo, A., and Rodriguez-Iturbe, I.: Daily streamflow analysis based on a two-scaled gamma pulse model, Water Resour. Res., 46, W11546, doi:10.1029/2010WR009286, 2010.

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

Natural Resources Conservation Service (NRCS): Hydrologic Soil Groups, in: Part 630 Hydrology National Engineering Handbook, NRCS, Washington, DC, 630.0700–630.0703, 2007.

Nyeko-Ogiramoi, P., Willems, P., Mutua, F. M., and Moges, S. A.: An elusive search for regional flood frequency estimates in the River Nile basin, Hydrol. Earth Syst. Sci., 16, 3149–3163, doi:10.5194/hess-16-3149-2012, 2012.

Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andréassian, V., Anctil, F., and Loumagne, C.: 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, doi:10.1016/j.jhydrol.2004.08.026, 2005.

Over, T. M., Riley, J. D., Sharpe, J. B., and Arvin, D.: Estimation of Regional Flow-Duration Curves for Indiana and Illinois: US Geological Survey, Scientific Investigations Report 2014–5177, 24 pp., doi:10.3133/sir20145177, 2014.

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

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

Reed, J. C. and Bush, C. A.: Generalized Geologic Map of the United States, Puerto Rico, and the US Virgin Islands, Digital Spatial Dataset, US Geological Survey, available at: https://pubs.usgs.gov/atlas/geologic/, 2007.

Ries, K. G.: The National Streamflow Statistics Program: A Computer Program for Estimating Streamflow Statistics for Ungaged Sites: US Geological Survey, Techniques and Methods 4–A6, 48 pp., 2007.

Rosenberg, E. A., Clark, E. A., Steinemann, A. C., and Lettenmaier, D. P.: On the contribution of groundwater storage to interannual streamflow anomalies in the Colorado River basin, Hydrol. Earth Syst. Sci., 17, 1475–1491, doi:10.5194/hess-17-1475-2013, 2013.

Salinas, J. L., Laaha, G., Rogger, M., Parajka, J., Viglione, A., Sivapalan, M., and Blöschl, G.: Comparative assessment of predictions in ungauged basins – Part 2: Flood and low flow studies, Hydrol. Earth Syst. Sci., 17, 2637–2652, doi:10.5194/hess-17-2637-2013, 2013.

Sauquet, E. and Catalogne, C.: Comparison of catchment grouping methods for flow duration curve estimation at ungauged sites in France, Hydrol. Earth Syst. Sci., 15, 2421–2435, doi:10.5194/hess-15-2421-2011, 2011.

Schaller, M. F. and Fan, Y.: River basins as groundwater exporters and importers: Implications for water cycle and climate modeling, J. Geophys. Res., 114, D04103, doi:10.1029/2008JD010636, 2009.

Singh, K. P.: Model Flow Duration and Streamflow Variability, Water Resour. Res., 7, 1031–1036, doi:10.1029/WR007i004p01031, 1971.

Skupin, A.: Visualizing a knowledge domain with cartographic means, P. Natl. Acad. Sci. USA, 101, 5274–5278, doi:10.1073/pnas.0307654100, 2004.

Snelder, T. H., Datry, T., Lamouroux, N., Larned, S. T., Sauquet, E., Pella, H., and Catalogne, C.: Regionalization of patterns of flow intermittence from gauging station records, Hydrol. Earth Syst. Sci., 17, 2685–2699, doi:10.5194/hess-17-2685-2013, 2013.

Ssegane, H., Tollner, E. W., Mohamoud, Y. M., Rasmussen, T. C., and Dowd, J. F.: Advances in variable selection methods II: Effect of variable selection method on classification of hydrologically similar watersheds in three Mid-Atlantic ecoregions, J. Hydrol., 438–439, 26–38, doi:10.1016/j.jhydrol.2012.01.035, 2012a.

Ssegane, H., Tollner, E. W., Mohamoud, Y. M., Rasmussen, T. C., and Dowd, J. F.: 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, doi:10.1016/j.jhydrol.2012.01.008, 2012b.

Tague, C. and Grant, G. E.: A geological framework for interpreting the low-flow regimes of Cascade streams, Willamette River Basin, Oregon, Water Resour. Res., 40, W04303, doi:10.1029/2003WR002629, 2004.

Toth, E.: Catchment classification based on characterisation of streamflow and precipitation time series, Hydrol. Earth Syst. Sci., 17, 1149–1159, doi:10.5194/hess-17-1149-2013, 2013.

Vesanto, J., Himberg, J., Alhoniemi, E., and Parhankangas, J.: Self-Organizing Map (SOM), in: SOM Toolbox for Matlab 5, Helsinki University of Technology, Helsinki, Finland, 7–11, 2000.

Vogel, R. M. and Fennessey, N. M.: Flow Duration Curves II: A Review of Applications in Water Resources Planning, J. Am. Water Resour. As., 31, 1029–1039, doi:10.1111/j.1752-1688.1995.tb03419.x, 1995.

Vogelmann, J. E., Howard, S. M., Yang, L., Larson, C. R., Wylie, B. K., and Van Driel, J. N.: Completion of the 1990's National Land Cover Data Set for the conterminous United States, Photogramm. Eng. Rem. S., 67, 650–662, 2001.

Westerberg, I.K., Gong, L., Beven, K.J., Seibert, J., Semedo, A., Xu, C.Y., and Halldin, S.: Regional water balance modelling using flow-duration curves with observational uncertainties, Hydrol. Earth Syst. Sci., 18, 2993-3013, doi: 10.5194/hess-18-2993-2014, 2014.

Williams, L. J. and Dixon, J. F.: Digital Surfaces and Thicknesses of Selected Hydrogeologic Units of the Floridan Aquifer System in Florida and Parts of George, Alabama, and South Carolina, Data Series 926, US Geological Survey, available at: http://pubs.usgs.gov/ds/0926/, 2015.

Wolock, D. M.: Base-Flow Index Grid for the Conterminous United States, Open-File Report 03–263, US Geological Survey, available at: http://water.usgs.gov/lookup/getspatial?bfi48grd, 2003.

Ye, S., Yaeger, M., Coopersmith, E., Cheng, L., and Sivapalan, M.: 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, doi:10.5194/hess-16-4447-2012, 2012.

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

Yuan, L. L.: Using correlation of daily flows to identify index gauges for ungauged streams, Water Resour. Res., 49, 604–613, doi:10.1002/wrcr.20070, 2013.

Zoccatelli, D., Borga, M., Viglione, A., Chirico, G. B., and Blöschl, G.: Spatial moments of catchment rainfall: rainfall spatial organisation, basin morphology, and flood response, Hydrol. Earth Syst. Sci., 15, 3767–3783, doi:10.5194/hess-15-3767-2011, 2011.

1 **Table 1.** Independent variables used in this study, with the different sets of variables used to develop regional regression
2 models identified as expert (E), lumped (L), and distributed (D) in the last column.

Independent variable	Units	Description	Data source	Set
MAP	mm	Mean annual precipitation	PRISM	E, L, D
PET	mm	Mean annual potential evapotranspiration calculated using the Oudin et al. (2005) equation	PRISM	E, L, D
BFI	%	Mean baseflow index derived from a spatially interpolated grid	BFI48GRD	E, L, D
Precip_SD	mm	Standard deviation of annual precipitation	PRISM	L
Forest	%	Percent forest cover	NLCD 1992	L
Precip_1D_Max	mm	Median of annual 1-day maximum precipitation	PRISM	L, D
Precip_Intensity	mm d^{-1}	Precipitation per rainy day	PRISM	L, D
Spring_Temp	°C	Average temperature from April-June	PRISM	L, D
Aridity	-	Aridity index calculated as PET/MAP	PRISM	L, D
Percent_Snow	%	Mean annual percent of precipitation as snow	GAGES-II	L, D
Area	km^2	Drainage area	GAGES-II	L, D
Density	km^{-1}	Drainage density calculated as stream length divided by drainage area	NHDPlusV2, GAGES-II	L, D
Orientation	°N	Basin angle along main channel	GAGES-II	L, D
Elev	m	Mean elevation	NED	L, D
Relief_Ratio	%	Relief ratio calculated as elevation range divided by basin length along main channel	NED, GAGES-II	L, D
Slope	%	Mean slope	NED	L, D
Aspect	°N	Mean aspect	NED	L, D
Accumulation	km^2	Mean flow accumulation expressed as upslope area	NED	L, D
TWI	-	Mean topographic wetness index calculated as ln(Accumulation/tan(Slope))	NED	L, D
Soil_Porosity	%	Mean soil porosity expressed as percent pore volume	CONUS-SOIL	L, D
Water_Capacity	%	Mean water capacity expressed as percent volume at field capacity	CONUS-SOIL	L, D
Poorly_Drained	%	Percent poorly drained including hydrologic soil groups C and D (NRCS, 2007)	CONUS-SOIL	L, D
Precip_Lag1	-	Lag-1 autocorrelation coefficient of monthly precipitation data	PRISM	D
Wet_Season	-	Binary variables indicating season with peak precipitation calculated using circular statistics as in Dingman (2002)	PRISM	D

Table continued on next page

Independent variable	Units	Description	Data source	Set
Precip_Seasonality	-	Distribution of monthly precipitation throughout the year calculated using circular statistics as in Dingman (2002)	PRISM	D
Precip_1D_Max_SD	mm	Standard deviation of Precip_1D_Max	PRISM	D
Precip_Intensity_SD	mm d^{-1}	Standard deviation of annual Precip_Intensity	PRISM	D
PET_Amp	mm	Amplitude of the first term of the Fourier transform for monthly PET data as in Dalton (2005)	PRISM	D
PET_Ph	rad	Phase of the first term of the Fourier transform for monthly PET data as in Dalton (2005)	PRISM	D
Aridity_SD	-	Standard deviation of annual Aridity	PRISM	D
Elev_SD	m	Standard deviation of elevation	NED	D
Slope_SD	%	Standard deviation of slope	NED	D
Aspect_SD	°N	Standard deviation of aspect	NED	D
Accumulation_SD	km^2	Standard deviation of flow accumulation	NED	D
TWI_SD	-	Standard deviation of topographic wetness index	NED	D
Forest_Rip	%	Percent forest cover within 800 m of a stream channel	GAGES-II	D
Soil_Porosity_SD	%	Standard deviation of soil porosity	CONUS-SOIL	D
Water_Capacity_SD	%	Standard deviation of water capacity	CONUS-SOIL	D
BFI_SD	%	Standard deviation of baseflow index	BFI48GRD	D

1

Table 2. Top five lumped and distributed variables for predicting high (Q_{01}-Q_{20}), average (Q_{30}-Q_{70}), and low (Q_{80}-Q_{99}) percentile flows based on the percent of the regression models that included each variable (in parentheses).

Lumped			Distributed		
High	Average	Low	High	Average	Low
BFI (62.5)	BFI (81.4)	BFI (76.8)	BFI (60.7)	BFI (75.7)	BFI (80.4)
Aridity (55.4)	MAP (52.9)	MAP (44.6)	Aridity (41.1)	Aridity_SD (40.0)	Poorly_Drained (44.6)
MAP (48.2)	Aridity (45.7)	Aridity (44.6)	Percent_Snow (32.1)	Elev (37.1)	Percent_Snow (28.6)
Percent_Snow (39.3)	Elev (37.1)	Poorly_Drained (42.9)	Precip_Seasonality (30.4)	Poorly_Drained (34.3)	Elev (25.0)
Spring_Temp (37.5)	Percent_Snow (34.3)	Spring_Temp (35.7)	PET (30.4)	Aridity (30.0)	MAP (21.4)

Table 3. The mean and range of the CN for the regression models developed using the three different sets of variables.

	Expert	Lumped	Distributed
Minimum	522	161	49
Mean	4.8×10^{4}	3.2×10^{5}	7.1×10^{7}
Maximum	3.4×10^{5}	4.9×10^{7}	1.2×10^{10}

Table 4. Predictive performance of the regression models developed using expert, lumped, and distributed variables and summarized as **(a)** the sum of absolute E, **(b)** R^2, and **(c)** N. Bold numbers indicate the set of variables that produced the best regression models for each percentile flow.

(a)

	Q_{01}	Q_{05}	Q_{10}	Q_{20}	Q_{30}	Q_{40}	Q_{50}	Q_{60}	Q_{70}	Q_{80}	Q_{90}	Q_{95}	Q_{99}
Expert	**9.89**	**9.00**	8.93	8.83	9.01	9.46	9.80	9.89	9.64	9.23	**8.31**	**7.61**	**6.69**
Lumped	11.5	10.1	**8.51**	**8.24**	**8.36**	**8.70**	**8.83**	**8.92**	**8.97**	**9.03**	8.76	8.14	7.15
Distributed	11.5	9.32	9.05	9.01	9.14	9.31	9.45	10.0	9.54	9.56	8.81	8.11	6.97

(b)

	Q_{01}	Q_{05}	Q_{10}	Q_{20}	Q_{30}	Q_{40}	Q_{50}	Q_{60}	Q_{70}	Q_{80}	Q_{90}	Q_{95}	Q_{99}
Expert	**0.60**	**0.67**	0.71	0.71	0.72	0.72	0.69	0.66	0.64	0.63	**0.64**	**0.64**	**0.63**
Lumped	0.47	0.58	**0.71**	**0.77**	**0.77**	**0.75**	**0.74**	**0.71**	**0.68**	**0.64**	0.58	0.56	0.52
Distributed	0.42	0.60	0.67	0.71	0.72	0.71	0.69	0.64	0.63	0.60	0.58	0.55	0.52

(c)

	Q_{01}	Q_{05}	Q_{10}	Q_{20}	Q_{30}	Q_{40}	Q_{50}	Q_{60}	Q_{70}	Q_{80}	Q_{90}	Q_{95}	Q_{99}
Expert	**0.59**	**0.67**	**0.70**	0.71	0.72	0.71	0.69	0.66	0.63	0.63	**0.63**	**0.62**	**0.60**
Lumped	0.45	0.58	0.70	**0.75**	**0.76**	**0.74**	**0.74**	**0.71**	**0.68**	**0.63**	0.58	0.56	0.51
Distributed	0.39	0.60	0.66	0.69	0.70	0.70	0.68	0.63	0.62	0.59	0.58	0.54	0.51

Table 5. The overall performance of the regression models developed using the expert assessment (expert) and data-driven approach (lumped and distributed) quantified as the sum of absolute E and mean R^2 and N for all the percentile flows. Bold numbers indicate the approach that produced the best overall regression models.

	Expert	Lumped	Distributed
E	116	**115**	120
R^2	**0.67**	0.65	0.62
N	**0.66**	0.65	0.61

Table 6. Range of the data used to create the CONUS Percentile Flow Predictor. Input data for the tool should be within this range.

	MAP	PET	BFI
Minimum	234	292	3
Maximum	4117	1390	85

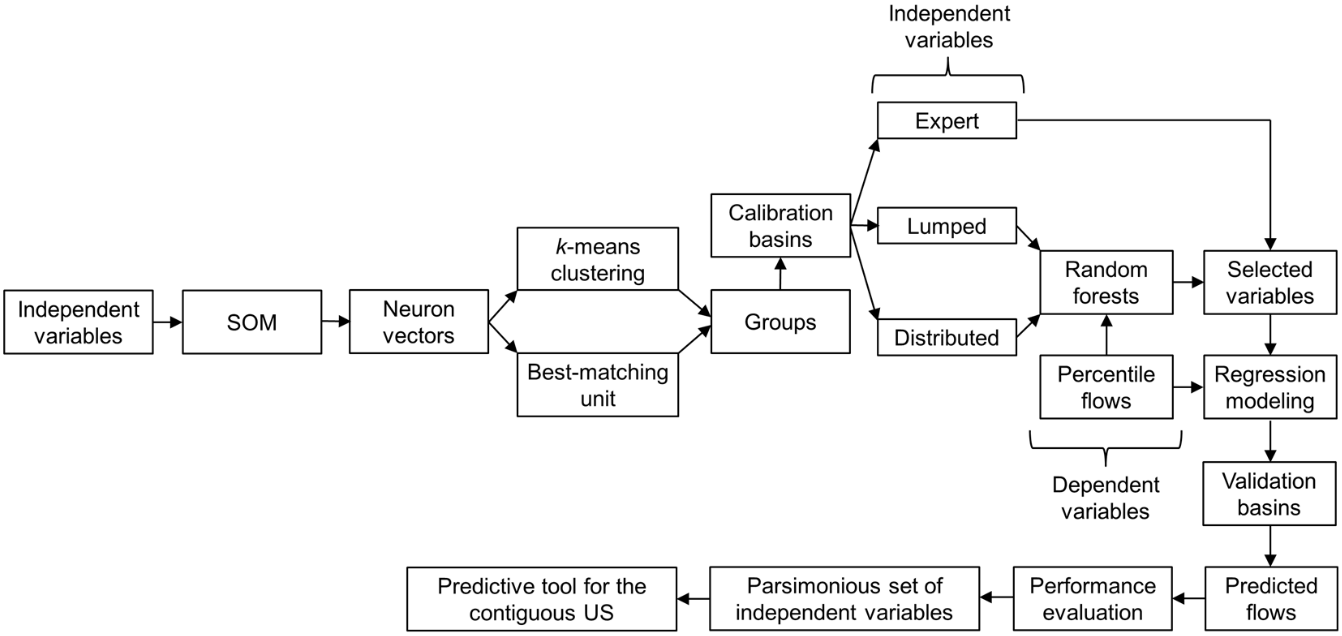
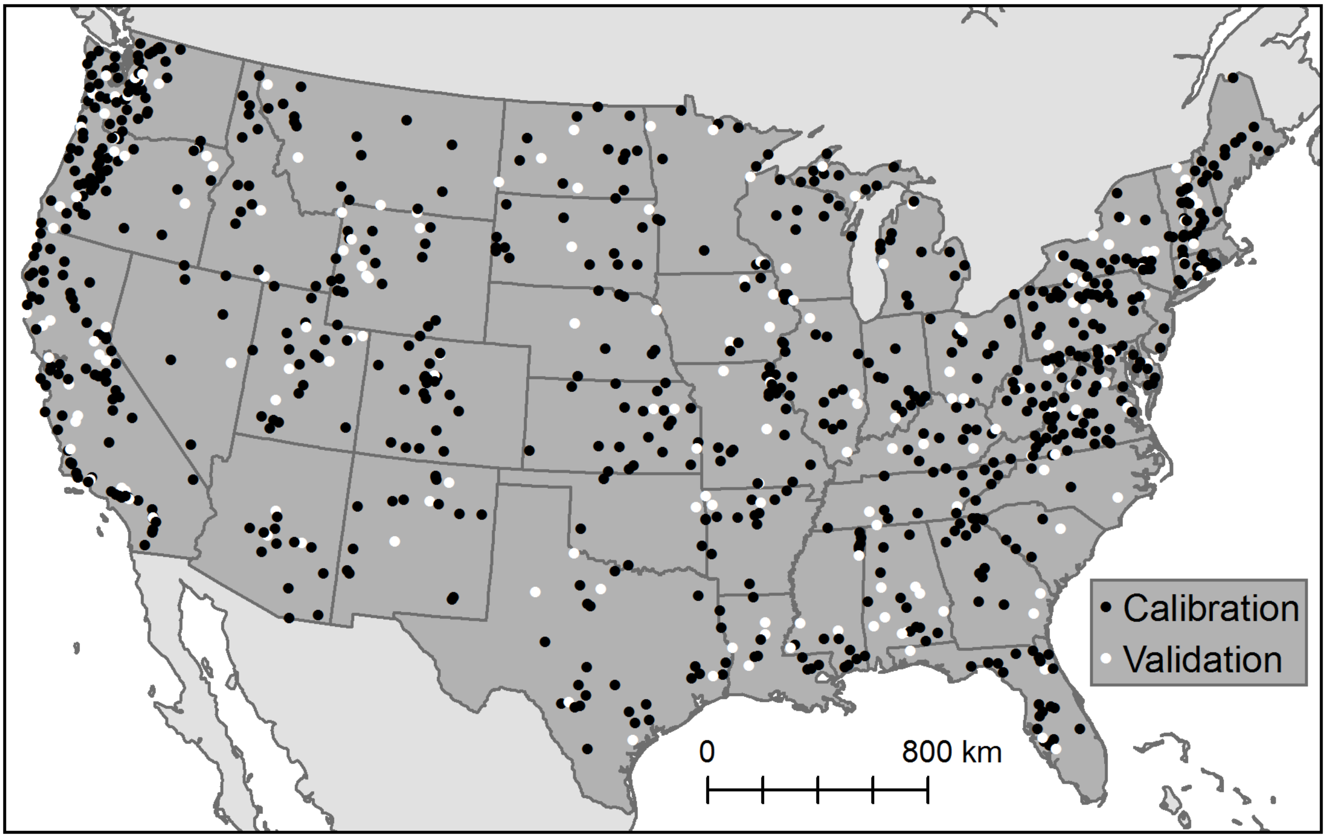


Fig. 1. Steps for developing regional regression models using expert, lumped, and distributed variables in order to identify a parsimonious set of variables and introduce a tool for predicting percentile flows in the contiguous US.

1 

2 **Fig. 2.** Map of the 734 calibration and 184 validation basins in the contiguous US represented by the location of their stream
3 gauges.

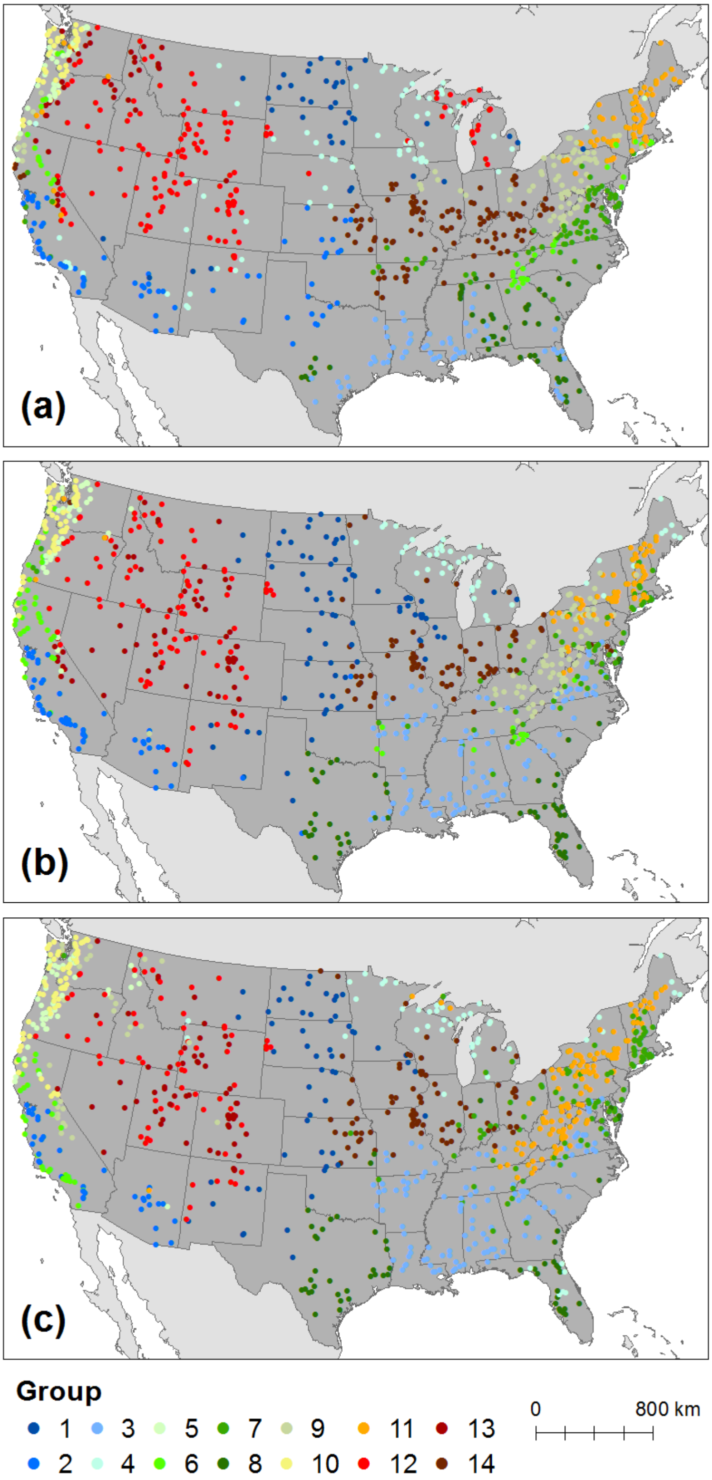


Fig. 3. Maps of the study basins split into 14 groups using the **(a)** expert, **(b)** lumped, and **(c)** distributed variables.

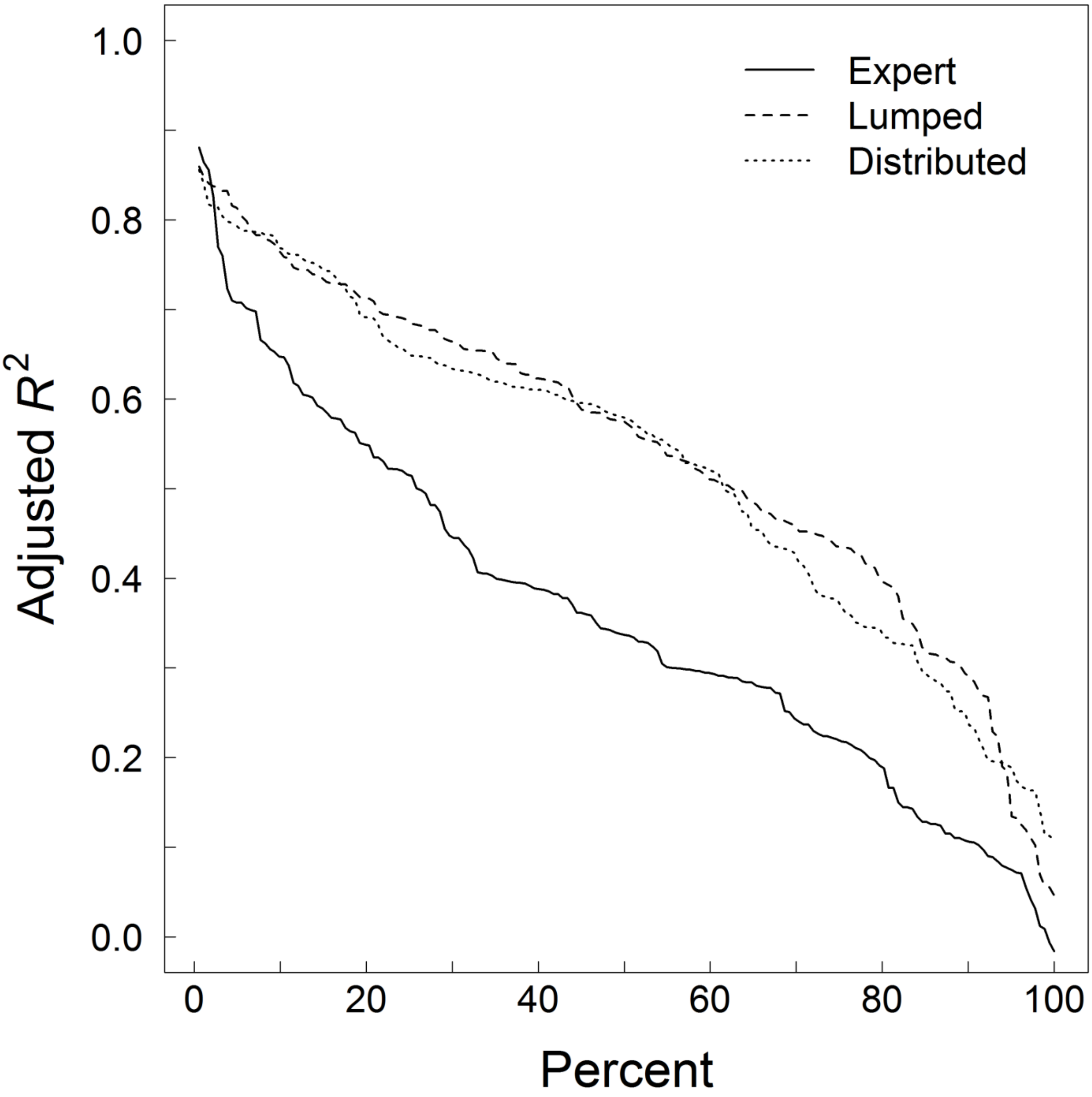


Fig. 4. Percent of the regression models with an adjusted $R^2 \geq$ the given value.