SAN DIEGO STATE UNIVERSITY AND

UNIVERSITY OF CALIFORNIA

Santa Barbara

Flow duration curve prediction for ungauged basins: A data-driven study of the contiguous United States

A Dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Geography

by

Geoffrey George Fouad

Committee in charge:

Professor Allen Hope, Chair

Professor André Skupin

Professor Christina Tague

Professor Hugo Loáiciga

June 2016

The dissertation of Geoffrey George Fouad is approved.

Hugo Loáiciga

Christina Tague

André Skupin

Allen Hope, Committee Chair

March 2016

Flow duration curve prediction for ungauged basins: A data-driven study of the contiguous United States

by

Geoffrey George Fouad

ACKNOWLEDGEMENTS

Thanks are in order for my committee. They helped with this research each in their own unique way. Special thanks to my committee chair, Allen Hope, for sharing his vision. Allen, I will always try to think of how you would frame a study to distill the message in the clearest possible way. You have a special gift for that, and I thank you for trying to share it with me. I am also thankful for André Skupin's frankness and unrivaled expertise with the self-organizing map, a large part of this research. The infectious energy and insightful comments of Naomi Tague greatly improved this research. Finally, Hugo Loáiciga speculated early on in this research that the typical variables would inadequately represent the hydrologic processes necessary to predict the flow duration curve. This became a major finding of this research, and the exciting part of that is when something does not work in science, we now have more research to do. On we go.

I would like to acknowledge previous mentorship from Terrie Lee of the US Geological Survey. She inspired me to become a better scientist and continues to be a lifelong friend.

On a personal note, I would like to thank my unbelievable family. I would not have been able to accomplish this without the foundation they provide. My parents, Magdy and Marilyn Fouad, have provided ceaseless support throughout my life. Even when I was considering not attending college, they were supportive of me, or at least they have good poker faces. Thank you Mom and Dad, and thanks also to my brother, Daniel Fouad, for the amusement over the years.

To my wife, Chelsea Riela, I feel as though we accomplished this together. You were by my side the whole way, and never flinched through the thick and thin. It always amazes

me how much you have my back. You motivated me as you finished your own graduate studies while working full time, and I can't wait to start the next chapter of our lives together. Finally, to my baby, Boleyn, the British bulldog, who provided much needed diversions from this work.

ABBREVIATED VITA OF GEOFFREY GEORGE FOUAD
March 2016

EDUCATION

Doctor of Philosophy in Geography, San Diego State University and University of California, Santa Barbara, June 2016 (expected)
Master of Science in Environmental Science, University of South Florida, May 2009
Bachelor of Science in Environmental Science, Catawba College, May 2006

PROFESSIONAL EMPLOYMENT

2015: Adjunct Professor, Geography, San Diego Mesa College
2010-2015: Teaching Associate (Instructor of Record), San Diego State University, Earth's Physical Environment (GEOG 101) and Global Climate Change (GEOG 409)
2008-2010: Research Assistant, Florida Water Science Center, US Geological Survey
2007-2009: Graduate Assistant, University of South Florida

PUBLICATIONS

Hope, A., Fouad, G., Granovskaya, Y., 2014. Evaluating drought response of Southern Cape Indigenous Forests, South Africa, using MODIS data. International Journal of Remote Sensing 35, 4852-4864. doi: 10.1080/01431161.2014.930205

Alsharif, K.A., Fouad, G., 2012. Lake performance differences in response to land use and water quality: Data envelopment analysis. Lake and Reservoir Management 28, 130-141. doi: 10.1080/07438141.2012.667865

AWARDS

European Geosciences Union Early Career Scientist's Travel Award (2015)

San Diego State University Graduate Student Travel Fund Award (2013)

Jack and Laura Dangermond Geography Travel Scholarship (2012)

American Society for Photogrammetry and Remote Sensing Student Achievement Award (2011)

University of South Florida Fred and Helen Tharp Graduate Scholarship (2008)

FIELDS OF STUDY

Hydrology, environmental modeling, machine learning applications, remote sensing, geographic information systems

ABSTRACT

Flow duration curve prediction for ungauged basins: A data-driven study of the contiguous United States

by

Geoffrey George Fouad

The flow duration curve (FDC) is one of the most widely used tools for displaying streamflow data, and percentile flows derived from the FDC provide essential information for managing rivers. These statistics are generally not available since most basins are ungauged. Percentile flows are frequently predicted using regression models developed using streamflow and ancillary data from gauged basins. Many potential independent variables are now available to predict percentile flows due to the ready availability of spatially distributed physical and climatic data for basins. A subset of the variables is often selected using automated regression procedures, but these procedures only evaluate a portion of the possible variable combinations. Other approaches for exploiting the information from physical and climatic data may produce stronger models for predicting percentile flows. The overarching hypothesis guiding this dissertation research was that more extensive approaches for extracting information from large sets of independent variables may improve percentile flow predictions. The dissertation was organized into the following three linked studies: (1) a performance evaluation of various approaches for selecting the independent variables of percentile flow regression models, (2) a comparison of different sets of variables for percentile flow regression modeling with increasing amounts of information in terms of the number of variables and their description of the statistical distribution of the data, and (3) a proof-of-concept study using a neural network approach called the self-organizing map

(SOM) to account for the noise and non-linearity of predictive relations between the independent variables and percentile flows. Key findings from these studies were as follows: (1) random forests was the best approach for selecting the independent variables for regression models used to predict percentile flows, but variables selected based on a conceptual understanding of the FDC performed nearly as well, (2) a set of only three variables (mean annual precipitation, potential evapotranspiration, and baseflow index) performed as well as models with larger sets of variables representing more physical and climatic information, and (3) the SOM performed similarly to global regression models based on all the basins, but did not outperform regression models developed for regions composed of similar basins. This may be due to the SOM using all the independent variables, whereas the regression models discarded irrelevant variables that could increase the error in percentile flow predictions. All the studies of this dissertation were performed using 918 basins in the contiguous US, and the resulting predictive models provide a tool for local watershed managers to predict 13 percentile flows along with an estimate of the predictive error. These models could be improved through future research that (1) emphasizes the role of geology as this provided the most valuable information for predicting the percentile flows, (2) exploits new sources of remotely sensed information as classic topographic variables provided little predictive information, and (3) develops specialized models designed for high and low flows as these were the most difficult to predict.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1: Introduction

The flow duration curve (FDC) is one of the most important graphical representations of streamflow data that shows flow versus the percent of time it is equaled or exceeded (Smakhtin, 2001). The flow magnitude associated with a given percent of time is a percentile flow, and these statistics are a common diagnostic for water resource planning, such as hydropower feasibility, water use permitting, and wasteload allocation (Vogel and Fennessey, 1995). Percentile flows are calculated using long-term streamflow records. However, such records do not exist for most basins, and water resource planning for these ungauged basins must rely on predicted percentile flows.

The most common approach for predicting percentile flows uses information from gauged basins to infer values for ungauged basins (Hrachowitz et al., 2013). This process is called hydrologic regionalization, and can be used to predict percentile flows directly (Mohamoud, 2008) or parameters of statistical distributions (e.g. lognormal) or analytical equations (e.g. polynomial) for representing the FDC (Castellarin et al., 2004). Approaches based on statistical distributions or analytical equations assume the general shape of the FDC for a geographic region (see Castellarin et al., 2004; Mendicino and Senatore, 2013; Viola et al., 2011), and may not be suitable for large study areas, such as the US, with a wide variety of FDCs. Directly predicting percentile flows requires no assumptions on the shape of the FDC, and can be accomplished using percentile flows from gauged basins and independent variables describing physical and climatic basin characteristics associated with streamflow.

The simplest form of percentile flow predictions uses values from nearby gauged basins and rescales them for differences in drainage area (Smakhtin et al., 1997). However, this approach may not be reliable for ungauged basins far from gauged basins (Archfield and

Vogel, 2010) and heterogeneous regions with large spatial variability in streamflow (Patil and Stieglitz, 2012). An alternative approach that may be more robust to the distance between basins and different regional conditions is the development of empirical relations between percentile flows and independent variables. These relations are used to predict percentile flows based on the independent variables of ungauged basins. A common method for developing relations between percentile flows and independent variables is multivariate regression (see Holmes et al., 2002; Hope and Bart, 2012; Mohamoud, 2008).

The independent variables used for the multivariate regression are critical, yet few studies have investigated their effect on percentile flow predictions (Hope and Bart, 2012; Ssegane et al., 2012a). These studies have tested the use of remotely sensed vegetation variables (Hope and Bart, 2012) and a variety of variable selection methods for a small sample of 26 basins in the mid-Atlantic US (Ssegane et al., 2012a). Despite the recent attention, independent variables are still normally selected using stepwise regression procedures that evaluate a sequence of variable combinations using model performance criteria (see Boscarello et al., 2015; Mendicino and Senatore, 2013; Zhang et al., 2015 for recent examples). Stepwise regression is widely criticized in the statistical literature for only identifying locally optimum variable combinations (see Flom and Cassell, 2007; Harrell, 2001; Miller, 2002 for critiques). A global optimum could be identified by evaluating every variable combination in an all-models approach, but this is often not feasible given the number of available variables (> 300 in a national database for the US called GAGES-II). The growth of these variables is due to the proliferation of geographic information system and remote sensing data that can be used to create a variety of variables potentially associated with percentile flows. Access to this information and the potential to improve

predictions prompts the question of how to select the independent variables for percentile flow regression models.

A related question is how much information the initial set of independent variables should have in order to predict percentile flows. The data used to create the independent variables is distributed in space and time, and this information is typically aggregated (or lumped) as variables describing the average conditions for the basins. For example, precipitation time series may only be expressed as mean annual precipitation (see Archfield et al., 2009; Castellarin et al., 2004; Viola et al., 2011), or a digital elevation model is summarized using mean elevation and slope (see Hashmi and Shamseldin, 2014; Kim and Kaluarachchi, 2014; Zhang et al., 2015). The information from distributed data can be extended to include its statistical distribution and special features that may be related to percentile flows. The previous examples could be extended in the following manner to possibly improve percentile flow predictions: (1) the distribution of precipitation throughout the year and its peak could be quantified to capture the storms that contribute to high and average flows (Cheng et al., 2012) and (2) depressional storage may be represented as a proxy for groundwater recharge and associated low flows (Chiang et al., 2002b). These more detailed variables are now common in data-driven studies that use many independent variables (see Hashmi and Shamseldin, 2014; Mohamoud, 2008; Ssegane et al., 2012a).

Data-driven studies stand in contrast to a simple set of variables chosen based on a conceptual understanding of the processes that control the FDC. These processes were recently investigated by a series of studies. The first of these studies (Yokoo and Sivapalan, 2011) used simulations in hypothetical basins to deconstruct the FDC into two components: (1) high (fast) flows associated with precipitation and (2) average to low (slow) flows largely

contributed by groundwater with adjustments for evaporative losses. Follow-up studies were conducted using a relatively large sample of basins in the US, and found that (1) the fast and slow flow components largely explained the variability in the FDC (Cheng et al., 2012), (2) additional information may be needed to explain differences due to regional groundwater levels, snow, and vegetation (Ye et al., 2012), and (3) average flows are influenced by regional patterns in precipitation (Coopersmith et al., 2012). The knowledge developed from these studies could be converted into a simple set of variables for explaining the FDC. The number of independent variables used to predict percentile flows can therefore range from a simple set of hydrologically-based variables to the many variables of data-driven studies, and the amount of information for predicting percentile flows should be investigated to guide future modeling efforts.

It has been established that the a priori identification of regions improves percentile flow predictions (see Boscarello et al., 2015; Isik and Singh, 2008; Sauquet and Catalogne, 2011). These studies divide the basins into homogeneous regions with the goal of reducing the variance in percentile flows and improving their predictability. Homogeneous regions are identified using multivariate cluster analysis, but this involves decisions, such as the input variables, clustering method, and number of clusters, that can lead to predictive uncertainty. An alternative approach called the self-organizing map (SOM) can be used to cluster the data and generate predictions without the decisions of identifying regions a priori. The SOM is a neural network that iteratively adapts to the input data, revealing its cluster structure in an output layer (or grid) of neurons (Kohonen, 1998). The grid is a representation of the input data with neurons linked to the input data by a vector of values equal in length to the number of input variables. The neuron vectors can be used to predict percentile flows and show their

connection to the independent variables. The SOM has been used to cluster basins for prediction (Boscarello et al., 2015) and for exploratory analysis of controls on streamflow (Toth, 2012), but these two objectives (prediction and exploration) have not been pursued in the same study. Despite its ability to cluster data and avoid the decisions of a priori region identification, the SOM has not been used as a predictor for percentile flows.

A. Dissertation overview

The chapters of this dissertation address the uncertainty of percentile flow predictions stemming from (1) independent variable selection, (2) the content and quantity of information in the initial set of independent variables, and (3) a priori region identification. Each chapter presents research on predicting 13 percentile flows for 918 basins in the US. The large scale was chosen to produce more generally relevant results for future studies. Chapter 2 is titled "Independent variable selection for regression modeling of the flow duration curve for ungauged basins in the US", and evaluates different methods for selecting the independent variables of percentile flow regression models. An automated regression procedure for selecting the independent variables was used as a reference for assessing the performance of alternative variable selection methods including (1) knowledge-based variable selection according to the literature on controls of the FDC, (2) principal component analysis, (3) correlation analysis, (4) random forests, (5) symbolic regression driven by genetic programming, and (6) Bayesian networks. The methods were chosen from a review of the literature on variable (feature) selection, and represent different types of variable selection. Comparing the predictive performance of these methods addresses the following research question: How should independent variables be selected for the regression modeling of FDC percentile flows?

Chapter 3 investigates the research question in its title, "How much physical and climatic information is necessary for regional regression modeling of the flow duration curve?". Different sets of independent variables were used to perform a regional regression that first split the basins into regions and then developed percentile flow regression models. A regional regression was performed to improve the predictions from Chapter 2 that used all the basins to develop the regression models (rather than regions). The variable selection method that performed the best in Chapter 2 was applied in this study. The regional regression was repeated using three sets of variables with increasing amounts of information as follows: (1) three hydrologically-based variables for explaining the shape of the FDC, (2) 22 lumped variables describing average conditions in the basins, and (3) 37 distributed variables describing average conditions and the statistical distribution of the basin's data. The predictive performance of the three sets of variables was evaluated to assess the amount of information necessary for the regional regression of percentile flows.

Chapter 4 applies the SOM to both predict percentile flows and explore their associations with the independent variables, and is titled "Predicting and visualizing relations to the flow duration curve using the self-organizing map". The SOM was used to predict the percentile flows, and its output was used to conduct an exploratory analysis of variables related to the FDC. Data visualizations were created to compare the cluster structure and variation between the percentile flows and independent variables, and inferences were drawn on the factors controlling the FDC for future modelling efforts. The SOM clusters the data as it generates predictions, and can therefore be used to avoid the decisions of identifying a priori regions for percentile flow predictions. It was used to generate global percentile flow predictions based on all the basins as well as within the regions previously identified in

Chapter 3. The performance of the resulting predictions was assessed to test the hypothesis that the SOM could be used to predict percentile flows without the aid of a priori regions. The regional regression results from Chapter 3 were also included in this study as a reference for comparing the predictive performance of the SOM. Application of the SOM for both prediction and an exploratory analysis of percentile flows was conducted to answer the following two research questions: (1) How do global percentile flow predictions generated using the SOM compare to regional predictions? and (2) What can be learned from the SOM regarding the variables related to percentile flows?

The final chapter synthesizes the major findings of the dissertation, and concludes with future research recommendations.

Chapter 2: Independent variable selection for regression modeling of the flow duration curve for ungauged basins in the US

A. Abstract

The flow duration curve (FDC) is a widely used tool for hydrologic applications, and predictions of its percentile flows are frequently needed for ungauged basins. These predictions are traditionally produced using regression models with basin characteristics as independent variables. Due to the large number of potential independent variables, a subset must be selected for percentile flow regression models. An evaluation of all the possible regression models is impractical given the dimensionality of current basin databases with many independent variables. Instead, a portion of the possible variable combinations is typically evaluated using automated regression procedures based on model performance criteria. This represents the baseline approach for selecting the independent variables of percentile flow regression models, but alternative methods from the field of variable (feature) selection may identify a more optimum subset of variables. This study constructed regression models for predicting the FDC percentile flows of 918 basins in the United States, and tested a baseline regression procedure against alternative methods for selecting the independent variables of the regression models. The alternative methods either created latent variables without cross-correlation (principal component analysis) or selected variables based on their relation to the percentile flows (knowledge-based variable selection, correlation analysis, random forests, symbolic regression, and Bayesian networks). Performance of all the variable selection methods was evaluated using 184 validation basins excluded from any phase of regression model development. The predictive performance of the baseline regression procedure was only better than principal component analysis, which was the only method

that did not use the percentile flows to identify the independent variables. All other methods that used the percentile flows for variable selection performed better than the baseline regression procedure. The predictive error and complications with multicollinearity strongly suggest that baseline regression procedures should not be the first choice for selecting the independent variables of percentile flow regression models. Another notable result from the performance evaluation was that independent variables selected from subject matter knowledge performed nearly as well as the best data-based methods. Subject matter knowledge and data-based variable selection both emphasized the importance of geologic characteristics in shaping the FDC, and the geologic variable of baseflow index had the largest effect on predictive performance. All of the models suffered from unacceptable predictive error, and modeled percentile flows may be improved by (1) novel independent variables more representative of the processes that control streamflow, (2) regression models developed for regions with similar basins that constrain streamflow variability and increase predictability, and (3) more powerful predictive models, like artificial neural networks, capable of dealing with the noise and non-linearities in the relations between basin characteristics and percentile flows.

B. Introduction

The flow duration curve (FDC) is one of the most widely used tools for displaying streamflow data (Smakhtin, 2001). Flow magnitude is plotted against the probability it is equaled or exceeded (i.e. exceedance probability). This essentially gives a cumulative distribution function for daily streamflow, and is used for a variety of applications, such as hydropower, water quality, and water use assessments (Vogel and Fennessey, 1995). These

applications are often concerned with the probability of exceeding critical flows, which can readily be determined from gauged streamflow data.

A far more challenging problem lies in predicting statistics of the FDC where streamflow data is insufficient or unavailable. The ungauged basin problem is widely acknowledged as the ultimate challenge in hydrology (Seibert and Beven, 2009), and a large body of literature has developed around predicting streamflow variables for ungauged basins. Despite its widespread use, the FDC has garnered little attention compared to predicting flood and low flow statistics (Castellarin, 2014).

Predicting the FDC of ungauged basins can be accomplished by a variety of methods overviewed in Castellarin et al. (2004). A rainfall-runoff model can be parameterized for the ungauged basin, and the subsequent discharge estimates can be used to calculate the FDC. However, the parameterization of rainfall-runoff models for ungauged basins is open to great uncertainty (He et al., 2011), and simpler empirical methods have been shown to provide similar results (Zhang et al., 2014).

Empirical methods use data from surrounding gauged basins to infer values of the FDC for ungauged basins. The simplest of these methods transfers the nearest gauged FDC to the ungauged basin by scaling it according to the difference in drainage area (Stedinger et al., 1993). More advanced methods along these lines use gauged FDCs to produce predictions from distance-based weighting schemes (Ganora et al., 2009) or spatial interpolation methods (Castellarin, 2014). Methods that rely solely on gauged FDCs become less reliable in sparsely gauged regions, and alternative methods based on the FDC's physical foundation are needed.

The climate and physical setting of a basin influence its FDC (Yokoo and Sivapalan, 2011), and can subsequently be used as independent variables to predict the FDC. The advantage of this approach is that it is not dependent on the location of the gauges, and nearby gauges may not be the best way to predict the FDC in areas prone to high variability and non-linearities in streamflow, such as in dry climates (Patil and Stieglitz, 2012).

Basin characteristics can be used in regression models to predict either (1) the parameters of a statistical distribution for representing the FDC (Viola et al., 2011) or (2) the flows of specified exceedance probabilities called percentile flows (Mohamoud, 2008). The latter method is appealing because assumptions on the shape of the FDC are not required. Predicting percentile flows is also useful for evaluating the performance of regression models from low to high flows (Hope and Bart, 2012). Percentile flow regression models are fitted using basin characteristics thought to influence streamflow.

Developments in data collection facilitated by geographic information systems and remote sensing have greatly increased the number of basin characteristics that can serve as independent variables in percentile flow regression models, and over 300 of these variables are now available in a nationwide database for the United States (US) called GAGES-II (Falcone, 2011). A recent survey of the literature by Ssegane et al. (2012a) revealed that 251 different basin characteristics have been used to predict percentile flows and other streamflow statistics.

The growing number of basin characteristics readily available to investigators poses the question of which variables should be used in regression models for predicting the percentile flows of the FDC. Variable selection is a necessary step in building regression models in order to (1) eliminate irrelevant variables that introduce unnecessary variance in

model estimates, (2) reduce the number of model parameters for smaller sample sizes, and (3) shed redundant variables that can destabilize models applied to new data (Miller, 2002).

An optimum subset of independent variables could be identified if every combination of the proposed variables in a study were evaluated in an all possible models (all-models) regression. However, this approach can quickly become impractical considering that the number of possible models is given by $2^x - 1$, where x is the number of independent variables. Computational limits are reached even for a small fraction of the variables available in basin databases (> one trillion possible models for only 40 variables). Most FDC regression modeling studies avoid the all-models approach due to the computational cost. In fact, only two studies from the same authors (Hope and Bart, 2011; Hope and Bart, 2012) were discovered that used all-models regression to predict percentile flows, but these studies were obligated to use a manageable number of variables that do not reflect the dimensionality of current basin databases.

Due to the number of independent variables proposed in most studies, an all-models approach is not feasible for selecting the independent variables of percentile flow regression models. Instead, non-exhaustive methods are used to select a subset of variables. These methods either evaluate different combinations of variables for the regression model or they do not use regression to select the independent variables for the regression model. The former option is more common as it directly assesses how well each subset of variables performs for the regression model.

Different combinations of variables are typically evaluated using automated regression procedures that search a portion of the variable space for an optimum subset of variables. This represents a baseline approach widely applied to develop percentile flow

regression models. The baseline regression procedures limit the variable space by (1) evaluating a sequence of variables (stepwise regression) or (2) eliminating entire branches of variable combinations not expected to improve the model (branch-and-bound regression; Miller, 2002). These procedures have the advantage of directly evaluating the variables for the regression model, but are widely criticized in the field of statistics because they tend to produce biased models that underperform on new data used for model validation (see Copas, 1983; Flom and Cassell, 2007; Harrell, 2001 for critiques). The limitations of baseline regression procedures are further explained in the next section, and justify the need to test alternative methods for selecting the independent variables of percentile flow regression models.

1. Limitations of baseline regression procedures

Baseline regression procedures apply automated routines to test numerous models with the goal of identifying the best subset of variables for predicting the dependent variable. These procedures are used when an exhaustive search of all possible models is not feasible due to the number of variables. Instead, an automated search of the variable space is guided by the level of significance of the variables as in stepwise regression (Flom and Cassell, 2007) or performance criteria that quantify how well the model fits the data as in branch-and-bound searches (Miller, 2002).

Stepwise regression is by far the most common method for selecting the variables used to predict FDC statistics including percentile flows (see Archfield et al., 2009; Castellarin et al., 2004; Mohamoud, 2008; Ssegane et al., 2012a; Viola et al., 2011). This method adds or removes variables sequentially from the regression model according to their level of significance. In forward stepwise regression, the most significant variable is added to

the model at each step. The reverse occurs for backward stepwise regression, where the variable with the least significance is removed from the model at each step. Both of these processes continue until all the variables in the model are significant at a specified level.

Reliance on a variable's level of significance is the main criticism of stepwise regression because the significance of a variable can vary depending on the order it enters the regression model (Copas, 1983). This can lead to spurious regression models that may not contain the best variables for predicting percentile flows. In addition, the variables are selected from a limited search of the variable space that may only reach a poor local optimum of predictive performance (Miller, 2002).

The limitations of stepwise regression described above have prompted more extensive search procedures that evaluate different variable combinations based on their fit with the observed data rather than the significance level of individual variables. Subsets of variables are evaluated by branch-and-bound algorithms that approximate an all-models regression. The branch-and-bound approach is generally considered an improvement to stepwise regression since the broader search may provide a more global optimum for the variable space (Miller, 2002).

The basic principle of branch-and-bound algorithms is that the predictive performance of a regression declines as variables are removed from the model (Miller, 2002). Therefore, a larger set of variables can be abandoned if another smaller set of variables produces a better model. For example, if a set of only three variables outperformed a different set of five variables, then the larger set of variables can be dismissed since removing variables will not improve the resulting model.

All baseline regression procedures that evaluate multiple models are problematic for estimating model parameters because they do not account for the model selection process (Flom and Cassell, 2007). The distribution of the modeling results is not represented, and this introduces bias into the model parameters and associated statistics that cannot be corrected (Flom and Cassell, 2007). The consequences of this bias are summarized in Harrell (2001) as follows: (1) overinflated significance levels for the variables, (2) overestimated absolute values of the model parameters, and (3) underestimated errors for the model parameters resulting in narrow confidence intervals. The bias resulting from regression procedures can reduce the model's predictive performance on validation data withheld from the model selection process.

Predictive performance on validation data is also diminished by multicollinearity in the form of cross-correlated variables (Dormann et al., 2013). This is a common problem with basin characteristics describing climatic and physical conditions that have coevolved over geologic timescales (Wagener et al., 2010). Multicollinearity is exacerbated by regression procedures because cross-correlated variables are selected if they happen to improve the model (Flom and Cassell, 2007). This results in models that are unstable for validation data that may not have a similar correlation structure between the independent variables (Dormann et al., 2013).

Strategies for dealing with multicollinearity in regression procedures screen models according to diagnostics derived from the correlation matrix of the variables (Belsley et al., 2004). This process requires arbitrary thresholds that are typically drawn from the literature (Dormann et al., 2013). Such thresholds are proposed for general use, although multicollinearity diagnostics may be sensitive to the type of data, sample size, and model

specifications (Snee and Marquardt, 1984). Acceptable levels of multicollinearity may also depend on the variable targeted for prediction (Snee and Marquardt, 1984). All of this makes setting the multicollinearity threshold a dubious task.

In light of the problems with baseline regression procedures, alternative variable selection methods that do not use regression may be better suited to select the independent variables for percentile flow regression models.

2. Alternative variable selection methods

The following review covers the main categories of alternative variable selection methods that do not use regression. Alternative variable selection methods can be most broadly categorized as those that use subject matter knowledge of the given phenomenon or are strictly based on the data. Methods from either category are rarely used to select the independent variables of regression models in hydrology (Ssegane et al., 2012a), although they avoid the bias introduced by regression procedures and the need to set arbitrary multicollinearity thresholds. Despite these advantages, few studies have compared alternative variable selection methods to baseline regression procedures for predicting streamflow statistics (Wan Jaafar et al., 2011), and only one study from a review of FDC regression modeling research compared different variable selection methods (Ssegane et al., 2012a).

2.1 Variable selection based on subject matter knowledge

The knowledge-based approach is particularly suited for streamflow because it is controlled by physical properties that can be expressed as basin characteristics. Over a century of hydrologic research has investigated the relations between measurable basin characteristics and different flow magnitudes (Dawdy et al., 2012). This information can be

used to hypothesize the basin characteristics that would be most effective for predicting the percentile flows of the FDC.

Knowledge of the factors that shape the FDC is typically only exercised to identify the initial set of independent variables (Castellarin et al., 2004), and seldom used to select the final independent variables because of the subjective nature of knowledge-based variable selection. Despite its subjectivity, knowledge-based variable selection is recommended in widely cited regression modeling textbooks, which advise against automated regression procedures that select variables without substantive theory of the subject matter (see Harrell, 2001; Judd et al., 2009; Miller, 2002). This same concern has been voiced in hydrology because variables selected based on process controls are less likely to yield results that are an artifact of the dataset and more likely to maintain a connection with the targeted streamflow variable in ungauged basins (Bowden et al., 2005).

Physically relevant variables can be selected by reviewing the literature on the climatic and geomorphologic controls of long-term streamflow regimes. The earliest studies of this kind connected drainage area to the flow magnitude of a basin (O'Connell, 1868), and later studies have used more sophisticated statistical and simulation tools to elucidate the factors that influence a basin's response to precipitation (Hrachowitz et al., 2013), including how such factors shape the FDC (Yokoo and Sivapalan, 2011).

The physical underpinning of the FDC has long been recognized since its early use (Foster, 1934), and many studies have substantiated the link between FDC statistics and simple basin characteristics, such as mean annual precipitation, land surface slope, and geologic units (see Cheng et al. (2012), Searcy (1959), and Yaeger et al. (2012) among many others). Previous research has disaggregated the FDC into three main segments composed of

high, average, and low flows (Yokoo and Sivapalan, 2011). The processes affiliated with each of these segments and their controls were characterized by Yokoo and Sivapalan (2011) as follows: (1) high flows are contributed by surface runoff during storm events with a loss factor for infiltration, (2) average flows reflect long-term storage dictated by climatic and geologic conditions, and (3) low flows are groundwater contributions during the dry season influenced by evapotranspiration rates. This information could be interpreted to select the independent variables of regression models for predicting the percentile flows of the FDC, and at the very least, should be used to cross-check the physical relevance of selected variables.

2.2 Variable selection based on the data

Other alternative methods for selecting variables are strictly based on the data, and originate from the field of variable (or feature) selection. This field is dedicated to creating novel data-based variable selection methods for improving predictive models. Variable selection methods are rapidly evolving in response to the increasing dimensionality of modern datasets. Like baseline regression procedures, data-based variable selection methods automate the process of identifying variables to make predictions, but they operate independent of regression modeling.

The goal of data-based variable selection methods is either to (1) identify variables with a specified numerical relation (explanatory or probabilistic) to the targeted variable or (2) limit statistical redundancy (multicollinearity) in the dataset. The latter methods may suffer from the disadvantage of not using the variable targeted for prediction, but are advantageous for maximizing the information content of the dataset used to generate predictions (Abdi and Williams, 2010). Limiting multicollinearity is particularly critical for

handling basin databases prone to having statistically redundant variables. Such variables fail to provide additional information, and can be detrimental for regression modeling because the resulting multicollinearity may diminish the portability of the model (Dormann et al., 2013).

The risk of multicollinearity can be reduced through dimensionality reduction methods or correlation analysis. Dimensionality reduction methods transform a dataset into latent variables that are uncorrelated to each other. The most popular dimensionality reduction method is principal component analysis (PCA), which is the most common alternative to baseline regression procedures for hydrologic variable selection (Ssegane et al., 2012a). The original variables undergo PCA to produce a new set of uncorrelated variables known as principal components (PCs), which are a linear combination of the variables calculated as in Abdi and Williams (2010). Using the PCs as independent variables has the advantage of eliminating multicollinearity in the regression model, but does not incorporate information on the variable targeted for prediction.

Another method for minimizing the risk of multicollinearity is correlation analysis. This method is based on the correlation matrix of the variables. Cross-correlated sets of variables are identified, and a single variable from each set is selected on the basis of subject matter knowledge (Yadav et al., 2007) or correlation with the modeled variable (Povak et al., 2014). This is a hybrid of methods that attempts to limit multicollinearity while establishing a connection to the dependent variable of the regression. The weakness of correlation analysis is that it requires arbitrary correlation thresholds for screening variables. Another potential problem is that correlation analysis only compares pairs of variables, and the relations

between two variables is known to change in the presence of other variables (Dormann et al., 2013).

The other major category of data-based variable selection methods that do not use regression evaluate the predictive potential of the independent variables. Predictive potential is quantified as explanatory power or probabilistic associations with the targeted variable. Like baseline regression procedures, alternative variable selection methods concerned with explanatory power treat variable selection as an optimization problem. However, these approaches differ from baseline regression procedures in how they evaluate variables, and may select better variables through iterative processes that evaluate the predictive potential of the variables on different subsets of the data (i.e. bootstrapping; Breiman, 2001) or heuristic searches driven by an independent training algorithm (i.e. genetic programming; Koza, 1994). Both approaches focus on minimizing model residuals (error) through different iterative processes for evaluating the predictive potential of the variables.

Data partitioning is achieved by regression trees that split the modeled variable into more manageable groups. The average of the group then serves as an effective prediction. The partitions of the regression tree are determined by conditional statements regarding the values of affiliated independent variables. This process is repeated to produce random forests, which are an ensemble of regression trees.

Random forests were introduced by Breiman (2001) as an improvement to the use of single regression trees. The random part of the forest is the sample selected to build each regression tree. Data withheld from the tree is then used to generate predictions that are averaged over all the trees. The error associated with these predictions is called the out-of-bag error, and the independent variables can be ranked according to the out-of-bag error

produced by randomly permuting, or essentially removing, each independent variable from the regression trees. Variables that produce more out-of-bag error are considered more important.

An alternative form of optimization employs a heuristic search to test various models for predicting a quantity, such as percentile flows. This process is similar to baseline regression procedures except it employs a genetic program to solve the regression problem. This specialized form of genetic programming, called symbolic regression, evaluates a set of variables and mathematical operators using a training algorithm that mimics the evolution of a population. The variables and mathematical operators are the members of the population, and they are iteratively combined and evaluated using an objective function, such as root-mean-square error. Characteristics of the best models are passed on to the next generation via genetic operators, like mutation and crossover. The evolutionary process eventually converges on an optimum set of models, and little to no change in the objective function is observed. For a more detailed review of genetic programming and its application for symbolic regression, the reader is referred to Koza (1994).

The drawback of optimization methods is that they are susceptible to selecting irrelevant variables due to their focus on minimizing predictive error and limited attention to the conditional relation between variables (Ssegane et al., 2012a). Irrelevant variables may be selected as a result of Simpson's paradox (Simpson, 1951), a phenomenon in which the relation between two variables changes when a third variable is introduced. This problem can be mitigated by identifying causal (probabilistic) associations between variables (Pearl, 2014).

The term causal association refers to a variable's effect on the probability of a certain outcome (or dependent variable), and variables that exhibit this behavior are suited to predict the associated variable. Causal associations are discovered using Bayesian networks that evaluate the conditional probability of a variable changing in the presence of other variables (Meganck et al., 2006). Bayesian networks are composed of nodes (variables) and edges (arrows) that connect variables with causal associations. Edges are drawn when a variable is more likely to change in the presence of another variable rather than without it. Variables with edges toward the dependent variable are then selected as the final set of independent variables.

C. Study objectives

Regression modeling is the traditional approach for predicting streamflow statistics, including percentile flows, in ungauged basins. However, the appropriate method for selecting the independent variables of these models remains unclear. Most studies adopt baseline regression procedures that automatically test regression models to identify the independent variables for predicting percentile flows. Studies that compare commonly used baseline regression procedures against alternative variable selection methods may help to uncover the best methods for developing percentile flow regression models. An overview of the methods available to select the independent variables of percentile flow regression models is provided in Figure 1, and the methods applied in this study are specified.

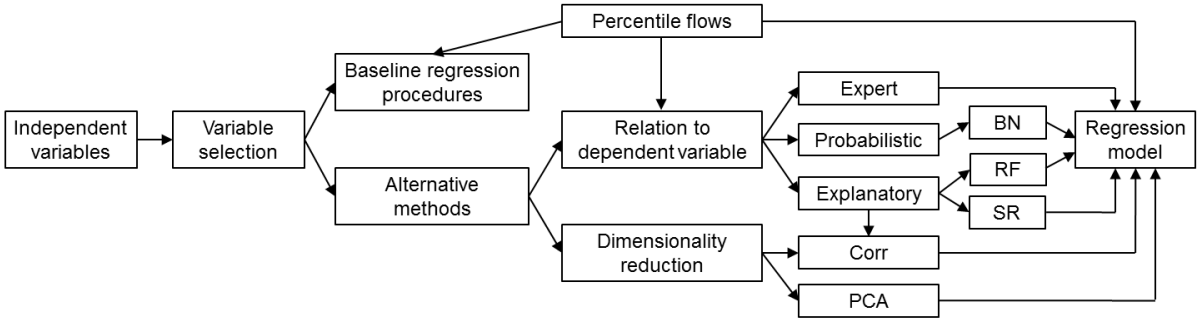

Figure 1. Overview of variable selection methods for percentile flow regression models with the methods applied in this study as follows: baseline regression, knowledge-based (expert), Bayesian network (BN), random forests (RF), symbolic regression (SR), correlation analysis (corr), and principal component analysis (PCA).

The overarching objective of this study was to develop regression models for predicting the percentile flows of ungauged basins in the US. The specific focus of the research was to compare variable selection methods for percentile flow regression models. A large number of independent variables are now available for the regression modeling of percentile flows in the US (see the GAGES-II database). These variables can be selected using a variety of variable selection methods. However, a recommended method has yet to emerge from the literature on FDC regression modeling. This study is designed to evaluate variable selection methods for predicting the percentile flows of the FDC using regression models, and addresses the following question:

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

A hypothesis on which variable selection method may perform the best could not be formulated from the literature, but the basis for selecting the independent variables provided some insight into which methods may perform better. Variable selection methods based on relations between the independent and dependent variable were expected to perform better than those that only evaluated the independent variables. Another expectation was that

methods concerned with the correlation between independent variables may reduce redundancy in the regression models and improve their stability for ungauged basins.

The rest of the paper is organized into the following sections. The methods section describes the overall research design, basins and variables used in this study, application of the variable selection methods, and the subsequent performance evaluation. The performance of the variable selection methods and interpretations are presented in the results and discussion section. Key findings and final recommendations are then provided in the conclusions section.

D. Methods

Competing variable selection methods were evaluated according to the predictive performance and multicollinearity of resulting regression models. Multicollinearity was a concern because correlated independent variables in regression models can hinder their transferability to ungauged basins. The performance of a baseline regression procedure was compared to alternative variable selection methods. The alternative methods operated independent of the regression to select independent variables using literature on the controls of the FDC and data-based methods, including PCA, correlation analysis, random forests, symbolic regression, and Bayesian networks.

The scope of this study was intended to produce generalizable results for future studies confronted with the question of how to select independent variables for FDC regression models. It has been recommended that generalizable results can be achieved by using many basins with an array of conditions (Andréassian et al., 2007). In light of this, 918 basins in the US were used to test the performance of variable selection methods. A larger variety of variable selection methods was tested than in previous studies, with a single

method chosen to represent each major category of variable selection (baseline regression procedures, knowledge-based, dimensionality reduction, optimization, and probabilistic). The performance of these methods was evaluated on a full range of FDC percentile flows to determine which methods performed better for low to high flows.

1. Study basins

Basins classified as "near-natural" in the GAGES-II database (Falcone, 2011) were used in this study. The near-natural designation is given to basins with minimal human impacts and water use. All of these basins were used provided they were in the contiguous US and had at least 30 years of continuous daily streamflow records. The minimum length of 30 years was used because it covers multidecadal shifts in climate and has been shown to produce stable streamflow statistics (Kennard et al., 2010). Streamflow records were not selected for a concurrent time period given the length of record and need for a large sample size. Nested basins were excluded from the analysis by removing any upstream basins. This ensured that the streamflow data was from hydrologically separate basins. The final number of basins was 918 after applying the above screening criteria.

The basins were then split into calibration and validation datasets (Figure 2). The calibration basins were used to select the independent variables and develop subsequent regression models, while the validation basins were withheld from these steps in order to assess how well the regression models resulting from the different variable selection methods were able to predict FDC percentile flows. The number of validation basins was set at 184 based on a review of 25 hydrologic prediction studies in which an average of about 20% of the basins were reserved for validation (see Heuvelmans et al., 2006; Holmes et al., 2002; Hope and Bart, 2011).

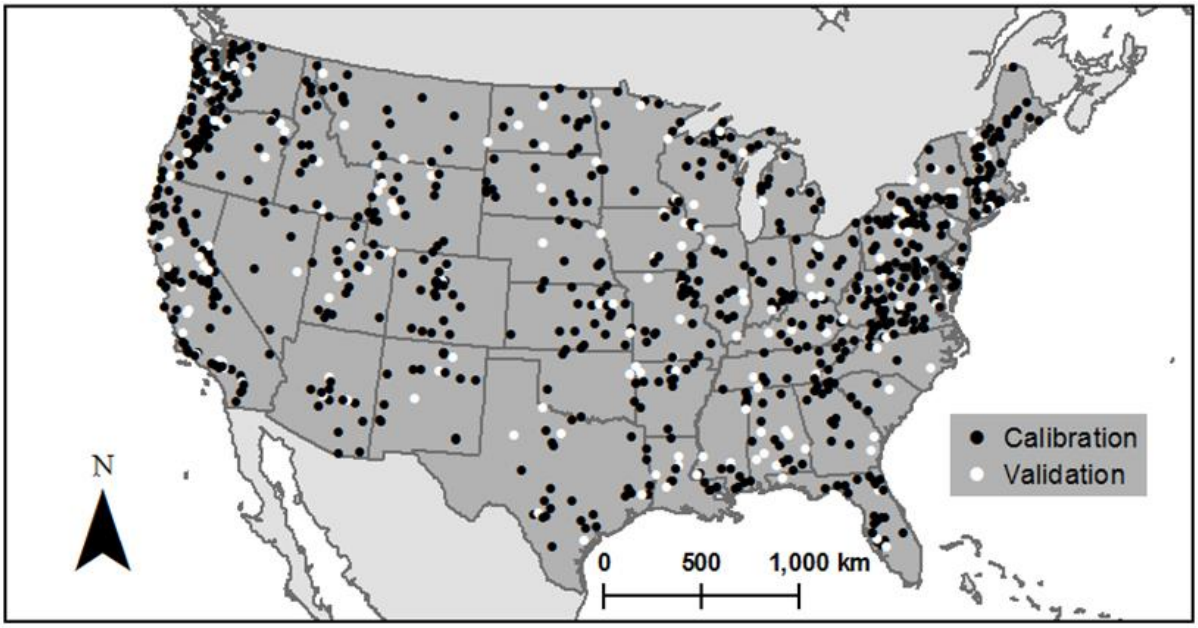

Figure 2. Location of the calibration and validation basins in the contiguous US.

A holdout validation approach was adopted given the large sample size. The diversity of the basins mimicked an ungauged situation, and a more complex cross-validation approach was deemed unnecessary. Following the recommendation of Klemeš (1986), a holdout validation should use basins that are representative of the entire sample in an approach called the "proxy-basin test". This validation approach can be accomplished using a sampling scheme that extracts a subset of basins which reflect the distribution of hydrologically critical variables in the entire sample.

A stratified random sampling scheme was used to select the validation basins for the proxy-basin test. Validation basins were extracted based on climate, geology, and drainage area because these are some of the most influential factors on the FDC (Yokoo and Sivapalan, 2011). First, the basins were geographically stratified by the broadest Köppen climate classes and major rock types as delineated by Peel et al. (2007) and Reed and Bush (2007), respectively. The geographic stratification was intentionally left as broad as possible

to limit any influence on the predictions, and this resulted in 11 groups for the contiguous US. A proportional number of basins were then extracted from each group. These basins were randomly selected within bins designed to sample across the group's drainage area distribution. The resulting validation basins have key characteristics that closely resemble the calibration basins (Table 1), and statistical tests (Kolmogorov-Smirnov and Mann-Whitney) were used to confirm that the calibration and validation datasets are from the same parent distribution.

Table 1. Distribution of key hydrologic characteristics for the calibration (C) and 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
25^{th} percentile	231	247	798	797	35	32	100	101	276	264
Median	409	412	1106	1100	48	46	292	303	498	470
75^{th} percentile	657	582	1308	1283	61	59	718	751	1194	1090
Maximum	3607	3507	4117	3965	85	82	25791	8265	3646	3435

2. Basin variables

A total of 13 percentile flows were predicted in this study. These included percentile flows at increments of ten from 10-90% and the extreme flows of 1%, 5%, 95%, and 99%. The percentile flows were calculated using 30 years of continuous daily streamflow records from potentially different time periods. The record length was more than adequate considering a previous study found that only five years are needed to reliably estimate the long-term FDC (Castellarin et al., 2007). Percentile flows were calculated using the Weibull plotting position as in Castellarin et al. (2004), and subsequently normalized to control for differences in magnitude between the basins. Normalization was achieved using the mean of

nonzero flows (Hope and Bart, 2011) since the mean and median flow equaled zero for some basins.

The percentile flows were predicted using 22 basin characteristics that served as the independent variables in regression models. A summary of these variables is provided in Table 2. Potential independent variables were chosen based on a review of the literature regarding FDC prediction and datasets covering the contiguous US. A representative variety of independent variables typically used to predict the FDC was then created from the available data. These variables are comprised of climatic, topographic, land cover, soil, and geologic variables.

Table 2. Dependent and independent variables created for the basins in this study.

Variable	Units	Description	Key reference	Data source
Dependent				
Q_p (e.g. Q_{01} for 1%)	-	Normalized percentile flows for 1, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 95, and 99%	Castellarin et al. (2004)	NWIS
Independent				
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. (2012a)	PRISM
Percent_Snow	%	Percent of precipitation as snow	Falcone (2011)	GAGES-II
Topography				
Area	km^2	Drainage area	Falcone (2011)	GAGES-II

Table continued on next page

Variable	Units	Description	Key reference	Data source
Density	km/km^2	Drainage density calculated as stream length divided by drainage area	Ssegane et al. (2012a)	NHDPlusV2, GAGES-II
Orientation	°N	Basin angle along main channel	Di Prinzio et al. (2011)	GAGES-II
Elev	m	Mean elevation	Ssegane et al. (2012a)	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. (2012a)	NED
Aspect	°N	Mean aspect	Ssegane et al. (2012a)	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. (2012a)	NED
Land cover				
Forest	%	Percent forest cover	Ssegane et al. (2012a)	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. (2012a)	CONUS-SOIL
Geology				
BFI	%	Mean baseflow index derived from a baseflow grid	Hope and Bart (2011)	BFI48GRD

Data sources: NWIS, National Water Information System (http://waterdata.usgs.gov/nwis); PRISM, Precipitation-elevation Regressions on Independent Slopes Model (http://prism.oregonstate.edu); GAGES-II, Geospatial Attributes of Gages for Evaluating Streamflow, version II (Falcone, 2011); NHDPlusV2, National Hydrography Dataset Plus Version 2 (http://www.nhdplus.com); NED, National Elevation Dataset (http://ned.usgs.gov); NLCD 1992, National Land Cover Dataset 1992 (Vogelmann et al., 2001); CONUS-SOIL, Conterminous US multilayer soil characteristics dataset (Miller and White, 1998); BFI48GRD, Base-flow index grid for the conterminous US (Wolock, 2003)

Climatic variables were generated from data temporally concurrent with the streamflow data except for variables requiring daily data (i.e. Precip_1D_Max and Precip_Intensity) and the only variable not produced specifically for this study (i.e. Percent_Snow). The daily climatic data did not span the period-of-record for the streamflow data, and the two climatic variables derived from daily data were generated for a 30-year

time period (1981-2010) with the most overlap in streamflow data. Percent_Snow is a GAGES-II variable for the average percent of precipitation delivered as snow from 1901-2000. The effect of snow was also accounted for using variables typically cross-correlated with snowfall, like elevation and temperature.

The only land cover variable was percent forest cover (Forest) since the amount of forest in a basin is known to affect its FDC (Brown et al., 2013b), and other vegetation classes have less of an effect on streamflow (see Bart and Hope, 2010; Brown et al., 2005; Wilcox and Huang, 2010).

Finally, the impact of geology on streamflow was represented using the baseflow index (BFI), which is the percent of streamflow contributed by groundwater. BFI is strongly linked to geologic conditions and was used instead of geologic units because it offers a way to quantify the effect of geology on streamflow (Bloomfield et al., 2009). A grid of BFI values has been produced for the contiguous US by spatially interpolating BFI values from gauged basins (Wolock, 2003). Although this product was derived from gauged streamflow data, it can be used for ungauged prediction since it is a pre-existing dataset with spatially contiguous coverage for the entire country. Similar products have been used for ungauged prediction in the past (see Hope and Bart, 2011; Wan Jaafar et al., 2011; Yadav et al., 2007).

The magnitude of the variables influenced some of the variable selection methods (i.e. PCA, symbolic regression, and Bayesian networks). For these methods, all the variables were converted to z-scores with a mean of zero and unit variance. This ensured that the magnitude of the variables did not influence the variable selection results.

3. Application of variable selection methods

The application of the variable selection methods in this study is described in the following sections (see the introduction for an overview of the different methods). Each variable selection method was limited to five variables. The number of variables was arbitrarily chosen in order to compare the performance of the different variable selection methods. Untransformed and natural log-transformed variables were assessed for the final regression models to accommodate linear and non-linear relations with the dependent variable. The performance of these models was evaluated in terms of their multicollinearity and accuracy for predicting the 13 percentile flows. All regression modeling and the accompanying performance evaluation were carried out in the R programming language (R Core Team, 2014).

3.1 Baseline regression procedure using a branch-and-bound search

A total of 44 independent variables were considered in this study including the untransformed and natural log-transformed variables listed in Table 2. This results in $2^{44} - 1$ independent variable combinations, which would take a typical computer over 5000 years to complete if it solved one regression equation every hundredth of a second. Computing technology may be able to solve this combinatorial problem in the future, but for now, reductionist methods are required to limit the variable space.

This study used a branch-and-bound search because it explores more of the variable space than stepwise regression and is the closest approximation to an all-models regression. The branch-and-bound search was conducted using the algorithm described in Miller (2002) and coded in the R leaps package (Lumley, 2009). This algorithm limits the variable space

based on the principle that removing variables from a model only increases the residuals. Larger sets of variables were completely eliminated if another smaller set performed better.

Model performance was assessed according to the residual sum of squares. A model averaging approach was adopted to identify the top five variables from a set of the best models. The top 40 models were returned by the branch-and-bound search for each level of complexity up to 20 independent variables. This produced a total of 800 candidate models. These models were then screened for multicollinearity using the condition number (CN) defined as

$$CN = \sqrt{\lambda_{max}/\lambda_{min}}, \quad (1)$$

where λ are the maximum and minimum eigenvalues of the cross-product matrix given by the selected independent variables (Belsley et al., 2004). Larger values signify greater multicollinearity between the independent variables of the regression model, and models with a CN .> 30 were discarded (Dormann et al., 2013). The remaining models were ranked according to their adjusted coefficient of determination (R^2) in order to compare models of varying complexity.

The top five variables were then discerned according to the number of times that each variable appeared in the top ten regression models. Ties between independent variables that appeared the same number of times in the top ten models were broken by the R^2 values resulting from univariate regressions with the percentile flow. This process was repeated to obtain the top five independent variables for each percentile flow.

3.2 Knowledge-based variable selection

Knowledge-based variable selection used previously developed understanding of the factors that control the FDC to identify the independent variables for the percentile flow

regression models. As previously discussed in the introduction, the FDC can be divided into three main segments: (1) high flows, (2) average flows, and (3) low flows. The high flows of the FDC are surface runoff during storms, and can be approximated by a precipitation duration curve adjusted for groundwater losses (Yokoo and Sivapalan, 2011). Average flows are associated with groundwater storage influenced by climate and geology (Coopersmith et al., 2012), and closely follow a groundwater (baseflow) duration curve (Yokoo and Sivapalan, 2011). Groundwater is also the source for the low tail of the FDC, except these flows are subject to evaporative demands during the dry season (Yaeger et al., 2012).

Based on the processes and related controls previously affiliated with the FDC, regression models for the percentile flows used the five independent variables listed below:

1. MAP – High flows are a function of rainfall, and average flows are related to climate.
2. PET – Low flows are suppressed by evapotranspiration, and PET is indicative of climate.
3. Slope – Surface runoff associated with high flows is affected by the slope of the land.
4. Soil_Porosity – Storage capacity and attendant subsurface drainage feed average and low flows.
5. BFI – Baseflow contributes to both average and low flows.

Although drainage area influences the overall magnitude of the FDC, it was omitted because normalizing the FDC by an index flow, such as the mean of nonzero flows, minimizes the effect of drainage area (Smakhtin, 2001). The final form of the variables was either untransformed or natural log-transformed depending on whichever had a higher R^2 in univariate regressions with each percentile flow.

3.3 Principal component analysis (PCA)

PCA was applied on the independent variables to produce uncorrelated PCs through orthogonal transformations of the data (Abdi and Williams, 2010). This was accomplished using the "prcomp" function in R (R Core Team, 2014). The same number of PCs as variables was produced, but only the first five PCs were used for the percentile flow regression modeling since all other variable selection methods were limited to five variables. Both untransformed and natural log-transformed variables were tested for the PCA, and natural log-transformed variables were used to compute the final PCs since they explained more variance than PCs based on the untransformed variables. The first five PCs of the natural log-transformed independent variables accounted for 76% of the variance in the data.

The PCs and independent variables representing the PCs were both tested for the percentile flow regression models. Representative variables for the PCs were selected using the method proposed in Lu et al. (2007). The first five PCs were used to generate five *k*-means clusters of the variable weights associated with each PC, and the independent variables with weights closest to the cluster centroids were selected to represent the PCs. The predictive performance of the models containing the PCs and independent variables representing the PCs was then compared to determine which variables should be used to predict the percentile flows.

3.4 Correlation analysis

Correlation analysis was used to identify groups of correlated independent variables, and select a representative variable from each group in order to reduce multicollinearity (Dormann et al., 2013). Correlation was quantified using Pearson's and Spearman's coefficients to account for linear and non-linear relations between variables. A correlation

threshold of 0.7 was used to identify groups of correlated independent variables. This threshold was used based on a recent review that reported 0.7 as the most commonly used correlation threshold (Dormann et al., 2013). The same review also found that the simple correlation threshold of 0.7 performed as well as other more sophisticated methods for dealing with multicollinearity. The correlation threshold identified groups of correlated independent variables and uncorrelated independent variables. These results were then used to select the top five independent variables for each percentile flow regression model as follows:

1. The variables from each correlated group were ranked by the R^2 values from univariate regressions with the percentile flow. The univariate regressions used both the untransformed and natural log-transformed variables to accommodate linear and non-linear relations with the dependent variable.
2. Correlated variables that did not have the strongest association with the percentile flow were excluded from further consideration.
3. The remaining variables were then ranked as in the first step, and the top five variables were used in the regression model for the percentile flow.
4. This process was then repeated for each percentile flow.

3.5 Random forests

Random forests were applied to rank the independent variables. This was accomplished by evaluating the variable's effect on the error associated with predicting the percentile flows (Breiman, 2001). Ensembles of regression trees were generated to recursively split the data into similar groups, which were averaged to produce percentile flow predictions. These predictions were then used to quantify the out-of-bag error for the random

sample of basins withheld from each regression tree. The out-of-bag error was used to rank the independent variables by randomly permuting one variable at a time. Variable rankings were quantified according to the change in out-of-bag error, with larger increases in out-of-bag error leading to higher rankings.

The variables were ranked by random forests generated using the randomForest package in R (Liaw and Wiener, 2002). Random forests split the data into ordinal groups. Therefore, natural log-transformed variables were not included because the monotonic transformation would yield the same results. These variables were considered after using random forests to rank the untransformed variables.

Random forests have three free parameters: (1) the number of variables used to recursively split the data (m_{try}), (2) the minimum size of the final groups in the regression trees (terminal nodes), and (3) the number of regression trees in the ensemble (n_{tree}). The first parameter is the only one that requires tuning because the last two can be set based on previously established guidelines (Svetnik et al., 2003). The default value for m_{try} is one-third of the number of independent variables rounded down to the nearest whole number (seven in this case). This tends to give reasonable results comparable to or better than the other possible values (Svetnik et al., 2003). Nonetheless, a full scale test of all possible m_{try} values was conducted using ten random forests for each percentile flow. An ensemble of random forests was employed because results can vary depending on the random samples used to build the regression trees, and numerous random forests may be required to obtain stable results (Saeys et al., 2008). The default value of seven was adopted because none of the other m_{try} values were consistently better than the default.

The last two parameters were set according to recommended guidelines. These guidelines suggest that the terminal nodes have little effect on predictions if they are a small fraction of the data (Svetnik et al., 2003). This parameter was set to only five basins, or about 1% of the basins used to generate the regression trees. The last parameter was determined using the traditional approach of plotting n_{tree} versus the out-of-bag error. The resulting graph usually shows a relation of exponential decay toward a limit where an increase in n_{tree} no longer reduces the out-of-bag error. The point where out-of-bag error stabilizes should be used to set n_{tree} (Svetnik et al., 2003). The random forests used for the m_{try} tuning were inspected to set n_{tree}, and the out-of-bag error stabilized after 100 regression trees for each percentile flow, which was then adopted as the n_{tree} value.

The independent variables were ranked according to 1000 random forests because an ensemble approach has been recommended to obtain stable variable rankings (Saeys et al., 2008). Average rankings from the 1000 random forests were used to select the top five independent variables for each percentile flow. Natural log-transformed independent variables were introduced at this stage, and the variable with a higher R^2 value in univariate regressions with the percentile flow was selected as the final independent variable. The above variable ranking process was repeated to formulate the regression model for each percentile flow.

3.6 Symbolic regression

Symbolic regression is similar to a baseline regression procedure in that it searches the variable space for an optimum model. However, the search also includes a set of mathematical operators (symbols) for testing different model structures. An independent algorithm called a genetic program drives the search for better models. The genetic program

attempts to mimic the evolution of a population (Koza, 1994). The symbols and independent variables are combined to produce a population of model formulations that evolve toward a set of optimum solutions. This process was implemented using the rgp package in R (Flasch et al., 2014). Mathematical operators were limited to addition, subtraction, and natural log to be consistent with the other variable selection methods. Given this restriction, the genetic program was essentially used as a means for selecting the best variable combinations. The resulting model solutions were evaluated according to their root-mean-square error, as is customary in symbolic regression (Flasch et al., 2014).

Characteristics of the models with less error were passed along to the next generation of models. This evolutionary process was influenced by four parameters: (1) the number of models considered in the first generation, (2) the number of models produced for subsequent generations, (3) the probability that models with less error will be combined to produce new models, and (4) the number of new models created for each generation. These parameters were optimized for each percentile flow using the sequential parameter optimization toolbox (SPOT) in R (Bartz-Beielstein and Zaefferer, 2012). The authors of the rgp package created SPOT as a companion package to automatically parameterize their genetic program.

Independent variables were ranked based on the final population of models for each percentile flow. At least 924 models were included in the final population, but few of these models were unique, indicating convergence toward an optimum set of variable combinations. The top ten models were then used to rank the independent variables as in the baseline regression procedure. The primary ranking criterion was the number of times that the independent variables appeared in the top ten models, and ties were broken according to

the R^2 values from univariate regressions with the percentile flow. The top five independent variables associated with each percentile flow were then used in the final regression model.

3.7 Bayesian networks

A variety of Bayesian networks have been developed for variable selection (Ssegane et al., 2012a). These methods differ in terms of how the Markov blanket is computed. Rather than computing the probabilistic relation between all variables, a Markov blanket identifies a subset of independent variables that makes the dependent variable probabilistically unrelated to all the other variables. The conditional probability of the dependent variable is explained by the subset of variables, and no further information is gained from the remaining variables. This process is detailed in Aliferis et al. (2010). Markov blankets can be produced by either constructing a portion of the Bayesian network or focusing directly on the probabilistic connections of the dependent variable (Aliferis et al., 2010).

Both approaches for computing the Markov blanket were tested for selecting the independent variables of the percentile flows. LCD2 is a revised version of the local causal discovery algorithm, which constructs an incomplete Bayesian network surrounding the dependent variable (Mani and Cooper, 1999). The other method used only evaluates connections to the dependent variable, and is called HITON-MB (Aliferis et al., 2003). These two methods were also chosen because they outperformed their Bayesian network counterparts in a previous study on selecting percentile flow independent variables (Ssegane et al., 2012a).

Bayesian network methods were implemented using the causal explorer toolkit in MATLAB. The input data consisted of individual percentile flows and the untransformed independent variables. Natural log-transformed independent variables were considered after

running the Bayesian networks. HITON-MB requires discrete data for the dependent variable. Percentile flow data was discretized using the minimum description length principle, a common method for preprocessing Bayesian network input data (Friedman et al., 1997). The HITON-MB algorithm has one more requirement of setting the maximum number of variables evaluated as the Markov blanket is assembled. All values of this parameter were tested to assess the sensitivity of variable selection results, and little change was observed, with an average of 98% agreement between parameter values. The default value of three was used given the lack of sensitivity to the parameter.

The output of the Bayesian networks was a list of selected independent variables for each percentile flow. These lists contained over five independent variables, so the following procedure was applied to rank the variables:

1. A random sample of 20% of the basins was removed from the calibration data.
2. The Bayesian network was then run on the remaining data, and selected variables were recorded.
3. A total of five runs were performed by replacing the basins and removing a different sample of basins on the next run.
4. The lists of selected variables were then tallied across the runs to rank the variables.
5. Ties between variables were broken by the R^2 values from univariate regressions with the percentile flow. This step considered the untransformed and natural log-transformed independent variables, and whichever explained more variance in the percentile flow was adopted to assign the final variable rankings.

The above steps were performed to identify the top five independent variables for the final regression model, and this process was repeated for each percentile flow.

4. Performance evaluation

The performance of the variable selection methods was tested on 13 percentile flow regression models, creating parallel experiments on flows ranging from low to high. Percentile flow regression models used the top five independent variables from each of the variable selection methods, and were evaluated in terms of their multicollinearity and predictive performance. The same regression model structure was used in order to compare the variable selection methods, and remove any effect from the regression modeling. The structure of the regression models was as follows

$$ln(Q_n) = \beta_0 + \beta_1 X_1 \ldots + \beta_5 X_5 , \quad (2)$$

where X_1-X_5 are the top five independent variables selected by the variable selection method either untransformed or natural log-transformed, and β_0-β_5 are the estimated parameters of the regression. Linear and non-linear relations to the percentile flows were accommodated by considering untransformed and natural log-transformed independent variables.

Percentile flows were predicted in log space because they were highly skewed, and this can lead to violating the basic regression assumption of homoscedastic (evenly varying) model residuals (Harrell, 2001). The natural log transformation has been recommended for the regression modeling of skewed flows (Archfield et al., 2009), and is widely used to predict percentile flows (Booker and Snelder, 2012; Over et al., 2014; Zhang et al., 2014). It is not mathematically possible to compute the natural log transformation on zero flows, so a constant of one was added to the percentile flows (Kilmartin and Peterson, 1972).

Regression models were assessed for multicollinearity since this condition can limit the predictive potential of the model on new data. Multicollinearity was quantified using the

CN specified in Equation 1. Higher values of the CN signify greater multicollinearity between the independent variables in the regression model.

Validation was then conducted to evaluate the predictive performance of the regression models and their associated variable selection method. Predictions were made for the validation basins not used to select the independent variables or calibrate subsequent regression models. Predictive performance was quantified using three performance metrics, namely, the R^2 of the observed and predicted percentile flows, the Nash-Sutcliffe coefficient of efficiency (Nash and Sutcliffe, 1970), and the relative error (RE). Each metric was calculated in log space to deemphasize the effect of large values in the skewed percentile flows (Sauquet and Catalogne, 2011). The Nash-Sutcliffe coefficient of efficiency (NSE) is calculated as

$$NSE = 1 - \frac{\sum_{i=1}^{n}\left(Q_i - \hat{Q}_i\right)^2}{\sum_{i=1}^{n}\left(Q_i - \bar{Q}\right)^2}, \quad (3)$$

where Q_i and $\hat{Q}_i$ are the observed and predicted percentile flow for basin *i*, $\bar{Q}$ is the mean of the percentile flow across all basins, and *n* is the number of basins. The metrics thus far have an upper bound of one indicating perfect performance, and lower values signify poorer performance.

The error of individual predictions was expressed as relative error (RE) obtained from

$$|RE| = \frac{\hat{Q}_i - Q_i}{Q_i + 1} \quad (4)$$

where Q_i and $\hat{Q}_i$ are as previously defined, one was added to the denominator to allow for zeros, and the absolute value of the relative error was used to calculate the sum of the error.

The performance metrics were calculated for each percentile flow, and summary statistics were taken to compare the predictive performance of the variable selection methods.

E. Results and discussion

Percentile flow regression modeling results focus on the model's (1) multicollinearity, (2) predictive performance, and (3) independent variables to elucidate the preferred variable selection method. Model multicollinearity is discussed first as it is a concern in the variable selection process and can subsequently degrade model performance in validation. The predictive performance of the models is then presented to show which variable selection methods chose better independent variables for predicting the percentile flows. Finally, the independent variables chosen by each variable selection method are summarized to reveal the variables associated with better percentile flow predictions. This information is interpreted to identify important variables for predicting the different percentile flows and make recommendations for future variable development. The figures and tables for the results use the following abbreviated names for the variable selection methods: baseline regression procedure (baseline), knowledge-based variable selection (expert), principal component analysis (PCA), correlation analysis (corr), random forests (RF), symbolic regression (SR), and Bayesian network (BN).

Regression models from the PCA either used the PCs as the independent variables or represented each PC in the model with an independent variable. The latter method is preferred for model interpretation. However, it does not ensure that multicollinearity will be eliminated in subsequent models, and some of the information contained in the PCs may be lost. PCs were represented using the independent variables closest to the cluster centroids of the PCs (Lu et al., 2007). As expected, the use of independent variables to represent the PCs

increased the multicollinearity of the models to a CN over 400, and some of the information content of the PCs was lost. Models that directly used the PCs explained an average of 29% of the variance in the percentile flows of the validation basins, while models that used independent variables to represent the PCs only explained 19%. In light of this, the PCs were used as the independent variables in the regression models, and these models are used to represent the PCA in the results.

Only one of the Bayesian networks are presented in the results since they both had very similar predictive performance. LCD2 was chosen to represent the Bayesian networks because it eliminated far more variables than HITON-MB. This was unexpected because HITON-MB has the potential to be more selective than LCD2 by only evaluating the probabilistic relations to the dependent variable. The poor selectivity of HITON-MB disagrees with previous findings on its ability to derive compact variable selections (Aliferis et al., 2010).

1. Multicollinearity

Multicollinearity was quantified as the CN. The average and range of the CN for the 13 percentile flow regression models resulting from each variable selection method are given in Table 3. There were no clear patterns in the variation of the CN for the percentile flows, and the discussion of multicollinearity focuses on the differences between the variable selection methods.

Table 3. The average and range of multicollinearity quantified as the CN for the 13 percentile flow regression models of each variable selection method.

	Baseline	Expert	PCA	Corr	RF	SR	BN
Minimum	32	1142	2	191	337	31	191
Average	68	8575	2	2124	22129	6502	6810
Maximum	158	31437	2	9892	68792	28011	21849

The baseline regression procedure had the second lowest multicollinearity next to PCA. However, the recommended CN threshold < 30 (Dormann et al., 2013) was still violated by the baseline regression procedure despite screening the models for multicollinearity. This occurred because the final models from the baseline regression procedure used a combination of independent variables from the top ten models identified by the branch-and-bound search. Subsequent variable combinations violated the recommended CN threshold. This approach was adopted because none of the top models from the branch-and-bound search at a complexity level of five independent variables passed the multicollinearity screening. Despite the high level of multicollinearity, the number of variables used for the regression models was not decreased below five given the large number of basins and variability in their percentile flows. The branch-and-bound method proceeded by selecting the top five variables according to the number of times that each variable appeared in the best models.

Only ten models were used to identify the final independent variables because the recommended CN threshold was highly restrictive. The CN threshold left just 13-27 eligible models out of the 800 best models for each percentile flow. Most of these models (82%) only contained a single independent variable, thereby eliminating any chance of multicollinearity. The most complex models contained three independent variables, but they only accounted for less than 1% of the models. These results indicate that the independent variables with predictive power were highly redundant, and few of them provided new information to explain the variance in the percentile flows.

The CN threshold in baseline regression procedures is an arbitrary value. Setting this value can be problematic since the CN may be sensitive to the data (Snee and Marquardt,

1984). This problem was briefly explored by increasing the CN threshold to 40. Loosening the threshold had the expected effect of increasing the number of eligible models (26% more models) and their complexity (78% more models with multiple independent variables). More interestingly, the average agreement between the observed and predicted percentile flows increased 7% in validation. The main difference in the resulting models was that some of them included BFI, which was completely excluded using the more restrictive CN threshold of 30. BFI is later identified as a critical variable for the predictive performance of the models (see the selected independent variables section). The slight increase in validation performance was attributable to the models that included BFI.

Increasing the CN threshold allowed an important independent variable to enter the models, and did not decrease model performance in validation. These results highlight the uncertainty of setting the CN threshold in baseline regression procedures. An additional source of uncertainty is that alternative multicollinearity diagnostics, such as the variance inflation factor or determinant of the correlation matrix (Belsley et al., 2004), may be more suited to screen percentile flow regression models. The question of how to screen for multicollinearity in baseline regression procedures used to predict percentile flows deserves further investigation in future studies. In this study, the widely cited CN threshold of 30 was retained because the goal was to compare a typical baseline regression procedure versus alternative variable selection methods.

Overall, the methods that addressed the correlation between independent variables (baseline regression procedure, PCA, and correlation analysis) limited multicollinearity more than other methods based on relations to the percentile flows. This confirmed the expectation of reducing the redundancy in regression models by using methods that account for the

correlation between the independent variables. The baseline regression procedure limited multicollinearity using the CN to reject models with redundant information. PCA virtually eliminated multicollinearity by producing uncorrelated PCs for the percentile flow regression models. The correlation analysis was the least successful of the methods that directly attempted to limit multicollinearity. This method limited multicollinearity by selecting one variable from a group of correlated variables, but the correlation threshold for identifying the groups needed to be lowered to further reduce the multicollinearity in resulting regression models. Pairwise correlation values may also fail to account for interactions between independent variables that can induce multicollinearity.

The other methods driven by relations to the percentile flows suffered from severe multicollinearity far greater than the acceptable threshold of 30 (Dormann et al., 2013). Only one model from these methods strayed from this trend with a CN of 31, while the rest had a CN of at least 172 and most models (73%) above 1000. This is a concern as multicollinearity can increase the divide between calibration and validation performance, which is discussed in the next section. The high multicollinearity of certain methods indicates that they selected redundant independent variables which contributed little information to the percentile flow regression models.

Random forests were the worst violator of selecting redundant independent variables, and might be because it is the only method that weighed the importance of one variable at a time rather than evaluating sets of variables. Knowledge-based variable selection had the second highest multicollinearity although an effort was made to select variables thought to be unrelated that would provide separate information to the regression models. This effort was thwarted by an unanticipated correlation between PET and BFI (Pearson = 0.58 and

Spearman = 0.63). The correlation between the rest of the knowledge-based independent variables was weak at < 0.5 and the majority of values < 0.3 for both correlation coefficients. This further demonstrates the sensitivity of the CN to the correlation among independent variables and the need to prescreen variables for cross-correlation.

Symbolic regression and Bayesian networks had the lowest multicollinearity of the methods based on relations to the percentile flows, but they were still far over any recommended CN thresholds. Both methods resulted in similar levels of multicollinearity, with the exception of a higher maximum for the symbolic regression. This was not expected because an advantage of Bayesian networks is that they evaluate the conditional probability between variable combinations, and therefore, should limit multicollinearity more than optimization methods like symbolic regression that only seek to maximize model fit (Sebastiani and Perls, 2008). Bayesian networks failed to produce such results in this application.

2. Predictive performance

Calibration and validation performance was compared to assess the stability and accuracy of the regression models on ungauged basins. As one would expect, overall model performance declined in validation (Table 4). However, the difference in R^2 from calibration to validation was small, and could be due to random error. Thus, the models resulting from each variable selection method appeared to be stable for ungauged basins. The stability of the models was not impaired by the elevated levels of multicollinearity discussed in the last section. This may be because the independent variables were similarly correlated in the calibration and validation data.

Table 4. Average R^2 in calibration (C) and validation (V).

	Baseline	Expert	PCA	Corr	RF	SR	BN
C	0.35	0.48	0.32	0.50	0.53	0.48	0.50
V	0.33	0.45	0.29	0.46	0.49	0.44	0.47

The baseline regression procedure had a similar decline in R^2 as the other methods, but its R^2 values were only higher than PCA. The baseline regression procedure and PCA both had little predictive value, only explaining about one third of the variance in percentile flows. Both of these methods limited multicollinearity more than the others, but it appears that they paid for this with reduced model fit.

The other methods all explained about half of the variance in the percentile flows. Random forests distinguished itself by having slightly higher R^2 values on average, but it can be concluded from the average R^2 values that none of the regression models formulated by the different variable selection methods performed at a high level. This indicates a deficiency in the predictive potential of the entire set of independent variables and alternative variables may be needed to explain more variance in the percentile flows.

The R^2 values in Table 4 should be evaluated in the context of the study area and results from previous studies. The contiguous US is a far larger study area than in previous FDC regression modeling studies that assume a regional scope (see Archfield et al., 2009; Castellarin et al., 2004; Mohamoud, 2008). The heterogeneity of this study area leads to much larger variance in the percentile flows. This can diminish the predictive power of the independent variables. It is expected that dividing the study area into more homogeneous regions would reduce the variance in percentile flows and improve resulting predictions as in previous studies (see Chiang et al., 2002; Mohamoud, 2008; Sauquet and Catalogne, 2011). Percentile flow regression models developed for the Mid-Atlantic US, a subregion of this

study, achieved R^2 values > 0.7 (Mohamoud, 2008). Another study reported R^2 values < 0.2 for FDC regression models covering all of France, but the R^2 of the models improved to > 0.5 upon using homogeneous regions (Sauquet and Catalogne, 2011). In light of these past results, the R^2 values in this study are more impressive.

The average R^2 values of the knowledge-based regression models were nearly equivalent to those of the other more complex data-based methods. The knowledge-based independent variables also produced models with similar R^2 values in calibration and validation. These results suggest that the knowledge-based models are at least as portable as the models derived from the data.

To test the portability of the knowledge-based models, a cross-validation experiment was conducted in which 20% of the basins were randomly removed, and the remaining data was used to calibrate regression models with the knowledge-based variables. This process was repeated five times, and revealed that the average R^2 values in calibration (0.49) and validation (0.50) were similar to the R^2 values from the original set of calibration and validation basins. Thus, the performance of the knowledge-based models appears to be stable when transported to different data, and remains competitive with the data-based methods. The portability of the knowledge-based variables indicates that they are physically meaningful.

Model performance is shown as the NSE for each percentile flow in Figure 3. Performance of the regression models produced by the variable selection methods typically peaked at Q_{20}, with the exception of the baseline regression procedure and symbolic regression peaking at Q_{10} and Q_{40}, respectively. A steady decline in model performance was the general trend for flows below Q_{20}. All methods had their lowest performance for the

highest flows at either Q_{01} or Q_{05}. These results differ from previous studies that did not have a marked decline in predictive performance for the high percentile flows (Hope and Bart, 2012) and obtained the best predictions for the average percentile flows around the middle of the FDC (Mohamoud, 2008).

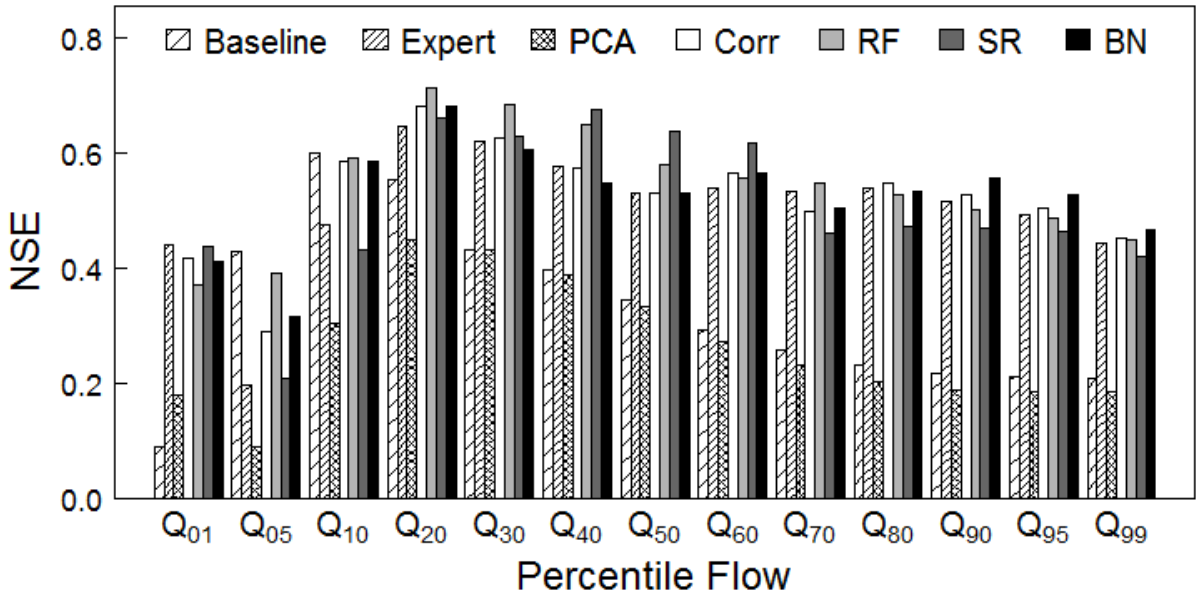

Figure 3. NSE of the models formulated by the variable selection methods for each percentile flow in validation.

Previous studies have also achieved better overall predictions for the percentile flows (Hope and Bart, 2012; Mohamoud, 2008; Ssegane et al., 2012a). However, those studies were conducted in homogeneous regions, whereas this study covered a much larger area with far more heterogeneous conditions. This likely introduced more variance in the percentile flows, and rendered them more difficult to predict because of weakened relations between the dependent and independent variables in the regressions.

Predictions at the extremes of the FDC suffered the most, which may be due to the increased variability and non-linearity of extreme flows (Salinas et al., 2013). The two

highest percentile flows were more difficult to predict than the low percentile flows. This may be because high flows are prone to more variance than low flows (Douglas et al., 2000).

Another reason why the extreme flows may have been poorly predicted is because of the independent variables, which may not have effectively represented the processes that control the high and low flows. This is especially evident considering that symbolic regression only selected two independent variables to predict percentile flows below Q_{70}, but still performed nearly as well as any other method. Independent variables for high flows should represent storm runoff, and additional variables may have been needed to describe the magnitude of storms, such as precipitation percentiles complementary to the percentile flows (Ssegane et al., 2012a). The number of rainy days is a surrogate for antecedent moisture conditions that may offer predictive information for high flows during storms (Jothityangkoon et al., 2001). The conversion of storms into runoff may have been better captured by land surface variables related to lateral flow, such as percent bare soil (Hashmi and Shamseldin, 2014). Both high and low flows are influenced by subsurface properties that may not have been adequately represented. Subsurface soil characteristics impact infiltration and lateral flow during storms (Merz and Blöschl, 2008), but the soil variables included in this study may have been affected by uncertainty or insufficient for representing the heterogeneity of soils at the basin scale. Low flows are contributed by groundwater. Although the independent variables included BFI, additional variables characterizing the subsurface drainage of a basin, such as a hydrologically sensitive geologic classification (Tague and Grant, 2004), may have improved low flow predictions.

In addition to the independent variables, the low percentile flows may have been poorly predicted because they contained zero flows. Approximately 12% of the flows below

Q_{80} were equal to zero. Predicting the FDC of intermittent streams is a special area of research with predictive models designed to accommodate zero flows (Hope and Bart, 2011). Such models would likely improve the predictions of the low percentile flows.

Relative performance of the different methods was variable across the percentile flows. No method clearly performed the best, but like the average R^2 values, the worst performance was typically associated with the baseline regression procedure and PCA. The baseline regression procedure offered the second poorest performance next to PCA for most of the percentile flows, but its performance reversed for the high percentile flows of Q_{05} and Q_{10}, where the baseline regression procedure was one of the better methods. This may be the case because there were not any strong independent variables for the high percentile flows. The other method that had particularly poor performance was PCA, which had the worst performance for every percentile flow except Q_{01}. PCA was the only method that did not use information on the percentile flows to select independent variables, and its predictive performance may have decreased as a result.

No method had consistently better performance across the percentile flows. The best performance was achieved by a variety of methods, and differences in the best performance for each percentile flow were mostly small with several methods near the top. The best method was typically random forests, symbolic regression, and Bayesian networks, which each had the best performance for three different percentile flows. These are all data-based methods that related the independent variables to the percentile flows via data partitioning, optimization, or causal associations.

Random forests performed the best at Q_{20} and Q_{30} where overall performance peaked and a variety of independent variables helped the models. This may be the case because

random forests were the most effective at ranking independent variables with predictive power. Random forests ranked the variables according to their effect on predictive error, whereas the other methods could result in ties between the best variables that had to be broken using univariate regressions with the percentile flows. Random forests also had the best performance at Q_{70}, but the difference was negligible and could be due to random error.

Symbolic regression achieved the best performance for the average percentile flows from Q_{40}-Q_{60}. This occurred although symbolic regression selected less than five independent variables for percentile flows below Q_{50}. The strong performance of symbolic regression with less than five independent variables highlights the redundancy (multicollinearity) of the independent variables. Few of the variables added information to the predictions, even for the average flows typically related to a wide variety of basin characteristics for climate, topography, and geology (Coopersmith et al., 2012). Symbolic regression employed an optimization routine that converged on the few independent variables with predictive relations to the average flows and successfully ruled out the other variables. The selectivity of symbolic regression was detrimental for the low percentile flows because it ruled out independent variables that contributed marginally to the predictions.

Bayesian networks were the best method for predicting the lowest percentile flows from Q_{90}-Q_{99}. Independent variables were selected based on their probabilistic relation with the percentile flows. The lowest percentile flows from Q_{90}-Q_{99} had a large fraction of zero flows (14%). The probabilistic relations developed by the Bayesian networks may have been more effective at handling the zero flows because they were treated as a single outcome linked to certain conditions (independent variables). This could account for the better performance of the Bayesian networks from Q_{90}-Q_{99}.

The simpler methods based on correlation and theoretical understanding of controls on the FDC performed nearly as well as the more complex data-based methods. Correlation and knowledge-based variable selection were typically among the best methods, and respectively scored the highest NSE values for Q_{80} and Q_{01}. Both methods performed similarly, but correlation typically outperformed knowledge-based variable selection because it selected the independent variables based on their predictive power in univariate regressions with the percentile flows. The performance of these methods peaked at the same percentile flows as the other methods from Q_{20}-Q_{40}, while their relative performance to the other methods was best for the highest percentile flow and those below Q_{70}. Predictions for these percentile flows were not improved by the more complex data-based methods due to an apparent lack of independent variables with predictive power.

The performance of the methods across the percentile flows was summarized using the average and range of the NSE for the regression models in validation (Table 5) and their cumulative error (Table 6). These results substantiated the findings on the predictive performance of the different methods discussed thus far. Nearly the same pattern in performance emerged using the mean NSE and sum of RE for all percentile flows. The lone exception was that Bayesian networks had a higher average NSE than the correlation analysis and symbolic regression, but all these methods had the same cumulative error. Besides this discrepancy, overall performance of the methods remained the same.

Table 5. The average and range of validation performance quantified as the NSE for the 13 percentile flow regression models of each variable selection method.

	Baseline	Expert	PCA	Corr	RF	SR	BN
Minimum	0.09	0.20	0.09	0.29	0.37	0.21	0.32
Average	0.33	0.50	0.27	0.52	0.54	0.51	0.53
Maximum	0.60	0.65	0.45	0.68	0.71	0.67	0.68

Table 6. The sum of the RE in validation for the 13 percentile flow regression models of each variable selection method. Lower values indicate better performance.

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

The baseline regression procedure and PCA were once again the only methods that stood out because of their poor performance. Both of these methods had the lowest ranges in NSE and the most RE. NSE values for predicting percentile flows were categorized in Castellarin et al. (2004) as poor (NSE < 0.50), fair (0.50 < NSE < 0.75), and good (NSE > 0.75). This categorized most of the models generated by the baseline regression procedure and all of the models from the PCA as poor.

The range of NSE values implied that neither the baseline regression procedure or PCA produced reliable models for predicting the percentile flows, and improvements are required to use these methods for applications in ungauged basins. The performance of the baseline regression procedure and PCA was comparable to one study covering all of France (Sauquet and Catalogne, 2011), which is a smaller study area with presumably less heterogeneity than the contiguous US. Predictive performance would likely be improved by dividing the basins into homogeneous regions as previously discussed for the poor R^2 values. Developing separate regression models for homogeneous regions typically improves predictions by reducing the variance in percentile flows (see Chiang et al., 2002; Mohamoud, 2008; Sauquet and Catalogne, 2011).

The poor performance of the baseline regression procedure and PCA is noteworthy since they are the two most common methods in hydrology for identifying the variables of flow-related regression models. Baseline regression procedures are particularly dominant in the realm of FDC regression modeling, yet these results suggest that other methods independent of regression may select better independent variables.

The other methods tested in this study had nearly the same performance. These methods had a higher range of NSE values and lower cumulative error than the baseline regression procedure and PCA, but they still only produced poor to fair models according to the categories in Castellarin et al. (2004). As discussed above, model performance is expected to increase if they are developed for homogeneous regions.

Random forests had the best overall performance with the highest range of NSE values and the least cumulative error. The correlation analysis and Bayesian networks closely followed random forests in both NSE and cumulative error. Symbolic regression was the least successful of the complex data-based methods, and had similar performance to knowledge-based variable selection. Previously developed knowledge of the FDC's physical controls proved useful in comparison to the more complex data-based methods, and the knowledge-based selection of variables may be preferable as a more parsimonious method with a physical foundation. A process-based understanding of the FDC is at least necessary to identify the initial set of variables for percentile flow regression models. Then, one of the data-based methods employing data partitioning, correlation, probability, or model optimization can be used to objectively select among physically meaningful variables for predicting FDC percentile flows.

3. Selected independent variables

The independent variables have been placed in categories (i.e. climate, topography, land cover, soil, and geology) to summarize the type of variables selected to predict high, average, and low flows (Table 7). These categories were devised from the major functional controls on streamflow identified in Wagener et al. (2004), and correspond to the categories in Table 2. The percentages in Table 7 have been normalized by the number of variables in

each category to show the relative importance of the different types of variables. Average flow is from Q_{30}-Q_{70}, and the outlying percentile flows are considered high and low flows. Discussion of the selected independent variables does not include PCA and knowledge-based variable selection since PCs are combinations of the variables and the knowledge-based variables are the same for each percentile flow (MAP, PET, Slope, Soil_Porosity, and BFI).

Table 7. Percent of selected independent variables from each variable category normalized by the number of variables in the category. Percentile flows are separated as high (Q_{01}-Q_{20}), average (Q_{30}-Q_{70}), and low (Q_{80}-Q_{99}).

	Baseline	Corr	RF	SR	BN
High					
Climate	46.7	16.0	15.1	11.1	17.8
Topography	8.9	2.6	4.5	3.7	4.3
Land cover	26.7	46.5	13.4	44.4	38.9
Soil	17.8	0.0	13.4	7.4	0.0
Geology	0.0	34.9	53.6	33.3	38.9
Average					
Climate	20.5	10.8	16.6	5.6	14.5
Topography	1.3	0.0	6.7	5.0	0.0
Land cover	58.6	39.3	0.0	44.7	41.4
Soil	19.5	10.5	16.1	0.0	2.8
Geology	0.0	39.3	60.5	44.7	41.4
Low					
Climate	27.3	9.7	5.1	0.0	20.0
Topography	0.0	0.0	3.4	1.6	0.0
Land cover	50.3	38.7	30.5	42.2	13.3
Soil	22.4	12.9	20.3	0.0	13.3
Geology	0.0	38.7	40.7	56.3	53.3

The most frequently selected types of variables for any flow were land cover and geology. Both of these variable categories were only comprised of a single variable (Forest and BFI). The recurring selection of these variables highlights their relations to the percentile flows and subsequent importance for modeling the FDC. The importance of Forest over climatic variables was unexpected, but it should be noted that Forest covaries with climate both in a general sense and as demonstrated by fairly large correlation coefficients in this

study (0.58 and -0.62 with MAP and Aridity, respectively). Forest was an important variable because it (1) moderates high flows during storm events through interception (Yokoo and Sivapalan, 2011), (2) is an indicator of long-term climatic factors that control average flows (Coopersmith et al., 2012), and (3) is related to evapotranspiration rates that affect low flows during the dry season (Yaeger et al., 2012). This indicates that additional vegetation information for canopy density and transpiration may be useful for predicting percentile flows. Such variables could be created using widely available remote sensing products, like the vegetation indices of the Moderate Resolution Imaging Spectroradiometer. Vegetation indices, such as leaf area index, may account for the effects of interception and transpiration on the FDC. Meanwhile, BFI was an important variable because the percent of streamflow delivered as groundwater is related to infiltration during storm events (high flows) and climatic and geologic conditions that influence the average and low flows of the FDC (Bloomfield et al., 2009).

Future work may benefit from including more land cover and geologic variables given their importance in this study. Land cover and geologic variables are typically categorical, but the inclusion of certain categories may be beneficial if they have a reasonable association to streamflow. For instance, the percent of sedimentary bedrock and water bodies in a basin may be useful as an indicator of storage conditions. Prior studies have also had success using baseflow recession statistics as a quantitative alternative to categorical geologic variables (Kroll et al., 2004). However, these statistics require streamflow data, and can only be used in an ungauged context if an interpolated product already exists for the study area. This is akin to the use of the BFI grid in this study, only the values of the grid represent other elements of baseflow recession related to geology.

The second most selected types of variables for any flow were climate and soil. Climatic variables were more consistently selected for high and average flows, which aligns with prior research linking the high tail and middle of the FDC to surface runoff generating processes and long-term climatic conditions (Yokoo and Sivapalan, 2011). Soil variables were more often selected for low flows. This is also consistent with prior research, which found that the low flows of the FDC are contributed by groundwater and subsequently influenced by soil storage properties (Yokoo and Sivapalan, 2011). The most surprising result of examining the variable categories was that topographic variables were the least frequently selected type of variable. This was unexpected given the widespread use of topographic variables to predict percentile flows and other streamflow statistics (Ssegane et al., 2012a). Results from this study downplay the importance of topographic variables.

The type of independent variables selected by the different methods was also compared. The baseline regression procedure was the only method not to select the lone geologic variable of BFI, whereas the other methods frequently used BFI to predict the percentile flows. The correlation analysis revealed that BFI was typically among the strongest independent variables in univariate regressions with the percentile flows, and influenced model performance as evidenced by the poor performance of methods that did not use BFI (i.e. baseline regression procedure and PCA).

BFI was excluded from the baseline regression procedure because it inflated multicollinearity in the regression models, and as a result, models containing BFI were rejected during the multicollinearity screening. It is not surprising that BFI heightened the multicollinearity of regression models since groundwater flows are related to other climatic, topographic, and soil variables (Santhi et al., 2008). Climate influences the amount of

precipitation available for groundwater flows, while topography and soil dictate groundwater infiltration.

The end result of excluding BFI was that the baseline regression procedure had poorer predictive performance than all other methods except PCA. BFI was clearly an important independent variable, yet its effect on multicollinearity led to its exclusion in the baseline regression procedure. This prompts the question of whether or not important independent variables should be sacrificed to guard against multicollinearity. This question may be resolved by using a more powerful predictive model, such as an artificial neural network, which is potentially robust to the adverse effects of multicollinearity (Dormann et al., 2013).

The remaining methods all heavily emphasized BFI alongside a mix of other independent variables. Correlation analysis, symbolic regression, and Bayesian networks all selected a similar distribution of independent variables for high, average, and low flows. In addition to BFI, these methods mainly selected land cover and to a lesser extent climatic variables. Similarity in the selected independent variables was reflected in predictive performance, where correlation analysis, symbolic regression, and Bayesian networks were nearly inseparable.

Random forests exhibited more variation in the independent variables selected to accompany BFI for the different types of flow. These changes in the independent variables selected for the various percentile flows resulted in random forests having the best overall predictive performance. Random forests made the best use of the independent variables, but its regression models still only achieved fair predictive performance at best.

The gap in predictive performance may be due to a shortage in useful independent variables. This calls for new variables more strongly associated with percentile flows and their governing processes. Current variables typically provide average values for the entire basin, but alternative variables that attempt to capture the temporal dynamics of climate and spatial distribution of physical features may improve percentile flow predictions. The FDC is influenced by subsurface properties that are underrepresented in current basin databases. Percentile flow predictions may benefit from new soil and geologic variables that quantify basin storage characteristics associated with the average and low flows of the FDC. Efforts are currently underway to better characterize subsurface properties that affect streamflow at Critical Zone Observatories in the US (see Takagi and Lin, 2010), and findings from these studies should be extrapolated to larger scales for predictions in ungauged basins.

A sample of the independent variables selected for one high, average, and low percentile flow is provided in Table 8. There is a strong level of agreement between the variables selected by each method. For instance, all the methods chose Aridity to predict Q_{10}, and four of the five methods predicted Q_{10} using BFI, Percent_Snow, and MAP. Similar degrees of overlap between the methods were present for the other percentile flows.

Table 8. Selected independent variables for a sample of percentile flows (see Table 2 for variable descriptions). Note that the baseline regression procedure contains Aridity twice as an untransformed and natural log-transformed variable.

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

The method that departed the most from the others was the baseline regression procedure because, as previously noted, BFI was rejected for all percentile flows. Another noteworthy result from the baseline regression procedure was that the branch-and-bound search returned models with Aridity as both an untransformed and natural log-transformed variable. Both Aridity variables were used in the same model during the branch-and-bound search since they improved model performance. These models also passed the multicollinearity screening, but including both forms of Aridity was a questionable result from the branch-and-bound search. The baseline regression procedure also selected independent variables that did not appear in any other method. This all contributed to the poor performance of the baseline regression procedure.

It should be noted that symbolic regression used less than five variables to predict all flows below Q_{50}. Despite this, symbolic regression remained among the best performing

methods. Only two independent variables were used for the percentile flow models below Q_{70}, yet these models were still among the best in predictive performance. This further confirms the lack of useful independent variables, particularly for the low percentile flows.

F. Conclusions

A variety of variable selection methods were tested to formulate regression models for predicting FDC percentile flows in ungauged US basins. Independent variables for percentile flows and other streamflow statistics are normally selected using stepwise regression procedures, but branch-and-bound regression procedures are an improvement to stepwise regression that search more of the variable space to find a more global optimum (Miller, 2002). Both of these procedures represent the baseline approach for identifying the independent variables of percentile flow regression models. Baseline regression procedures are used because an exhaustive search of all possible models is no longer feasible given the dimensionality of current basin databases. This problem, coupled with the model bias and multicollinearity introduced by baseline regression procedures (Harrell, 2001), motivated the use of alternative variable selection methods independent of the regression modeling.

A baseline regression procedure was compared to knowledge-based variables, PCA, correlation analysis, random forests, symbolic regression, and Bayesian networks. The variable selection methods were assessed according to resulting regression model multicollinearity and predictive performance. Regression models were developed to predict 13 percentile flows for 918 basins in the US, and the predictive performance of these models was evaluated using validation basins withheld from regression model development.

Regression model multicollinearity was only limited by the methods that assessed the correlation between independent variables. Multicollinearity was identified and removed in

the baseline regression procedure. However, this process diminished the predictive performance of the resulting models. The removal of multicollinearity using PCA also had the same effect on model performance. Thus, removing multicollinearity had the unanticipated effect of reducing the predictive performance of the regression models.

Multicollinearity removal was problematic in the baseline regression procedure because models had to be screened using an arbitrary threshold. A widely cited threshold was applied, but it restricted the number of eligible models along with their complexity. The arbitrary threshold was also responsible for rejecting one of the most important independent variables in BFI and reduced the predictive performance of resulting regression models. BFI greatly increased multicollinearity, but that did not hamper the predictive performance of the regression models.

Higher multicollinearity generally translated into better predictive performance in validation, which contradicts the assumption that multicollinearity degrades the transferability of regression models. Multicollinearity is a problem for transferring regression models to a dataset with a different correlation structure (Dormann et al., 2013). The validation data in this study likely had a similar correlation structure as the calibration data, and the predictive performance of the regression models was not negatively impacted by the presence of multicollinearity. Variable selection methods that resulted in elevated multicollinearity should be used with caution if the correlation structure may differ between the calibration and validation data.

The predictive performance of all the regression models was poor ($NSE < 0.50$) to fair ($0.50 < NSE < 0.75$), but these results may be improved by developing regression models for homogeneous regions in the US. Most of the regression models achieved similar

predictive performance despite using different variable selection methods. However, some of the variable selection methods performed better than the others, and the comparison of the variable selection methods revealed the following key findings:

- The baseline regression procedure only performed better than PCA, and both of these methods performed worse than the other methods. Baseline regression procedures and PCA are the two most common methods used to devise regression models in hydrology, but future studies may benefit from alternative variable selection methods that use knowledge of the controls on streamflow and numerical relations to the targeted streamflow variable established via correlation, regression trees, optimization algorithms, or causal associations.
- Methods other than the baseline regression procedure and PCA performed similarly, and consistently generated better regression models for predicting percentile flows. Random forests produced the best overall regression models. However, the best method for a given percentile flow varied, and the difference was often negligible.
- Knowledge-based variable selection performed similarly to the best data-based methods, and produced stable regression models for ungauged basins. This underscores the importance of using independent variables grounded in the physical understanding of runoff processes, which should at least be taken into consideration when formulating the initial set of independent variables.
- A small portion of independent variables was repeatedly selected by the best methods. These variables contributed redundant information as indicated by their high degree of multicollinearity, and had limited predictive power. Regression modeling of percentile flows requires new independent variables that better represent

the processes associated with streamflow generation. This was especially the case for the high and low flows with the most predictive error. These flows are both influenced by subsurface processes poorly characterized by current independent variables. New variables for the subsurface processes of infiltration and storage may improve regression models for high and low flows. Precipitation variables describing the statistical distribution of storm magnitude may also be critical for predicting high flows produced by storms.

- Widely used independent variables were mostly ineffective for predicting the percentile flows. Topographic variables are some of the most common independent variables, but they did not exhibit strong relations to the percentile flows, and, as a result, were not frequently selected by the variable selection methods.
- Geology and land cover were the most important independent variables, with BFI having the strongest influence on predictive performance. Given these results, future studies should more heavily emphasize geologic and land cover variables, and the development of such variables is needed to improve the representation of surface runoff and groundwater flows associated with the most difficult to predict percentile flows at the tails of the FDC.
- Even the best independent variable combinations had limited predictive potential, signifying that there was no underlying regression model solution for the given set of variables and basins. This could be due to missing information in the independent variables as previously discussed, but it could also stem from inadequacies in the regression modeling approach. For instance, regression models could be developed for hydrologically homogeneous regions in order to reduce the variability in

percentile flows and increase their predictability (see Chiang et al., 2002; Sauquet and Catalogne, 2011; Ssegane et al., 2012b). Another possibility is the use of a more powerful predictive model, like artificial neural networks, capable of assimilating the noise and non-linearities between percentile flows and basin characteristics. Both of these potential improvements to percentile flow predictions will be tested in future studies on clustering and neural network modeling of the basins.

Future research could exploit new datasets, such as the Gridded Soil Survey Geographic Database (Soil Survey Staff, 2014) and Soil Moisture Active Passive (Brown et al., 2013a), to account for subsurface processes that influence high and low flows prone to more predictive error. Modern datasets contain spatial and temporal information that is typically condensed into average values for the entire basin. Future studies could evaluate the utility of alternative variables that characterize the spatial distribution and temporal dynamics of the data. Such variables have corresponded to regional streamflow patterns (Toth, 2012), and may be effective independent variables for predicting percentile flows. An analysis of the factors related to the predictive error in the percentile flows was out of the scope of this study, but could lend insights into the information needed to improve predictions. A critical factor contributing to predictive error is the prediction of zero flows, and future studies should design predictive models specifically for intermittent streams.

The best method for the regression modeling of percentile flows remained unclear at the large scale of this study. A future study should be performed for a smaller, more homogeneous region to determine if one method is clearly better than the others. The percentile flow regression models developed for the US had poorer predictive performance than models developed for more homogeneous regions in previous studies (see Archfield et

al., 2009; Hope and Bart, 2012; Mohamoud, 2008). Identifying homogeneous regions with hydrologically similar basins may be a critical preliminary step in developing percentile flow regression models for the US. Subsequent regional regression models may achieve greater predictive performance than the global regression models developed in this study.

Percentile flow predictions may also be improved through the use of a more powerful predictive model, such as an artificial neural network. An advantage of artificial neural networks is that variable selection is not necessary since they are nonparametric models robust to noise and multicollinearity in the data (Coulibaly and Evora, 2007). Subsequent chapters of this dissertation will address the remaining uncertainty in the percentile flow regression models by employing a regional approach and artificial neural networks.

Chapter 3: How much physical and climatic information is necessary for regional regression modeling of the flow duration curve?

A. Abstract

The flow duration curve (FDC) expresses the percent of time a flow is exceeded, and its percentile flows are widely used in water resource applications. However, percentile flows are often needed for ungauged basins with insufficient flow data, and hydrologic regionalization approaches are used that pool data from gauged basins to predict percentile flows at ungauged sites. These approaches typically use regression models based on independent variables. The regression models are often developed for regions consisting of basins with similar independent variables in a process called regional regression modeling. Most regional regression studies have not focused on how the selection of independent variables can influence the predictive performance of subsequent models. This question was investigated in terms of the approach for selecting the initial set of variables and the amount of information necessary to develop regional regression models for 918 basins in the US. The regional regression modeling used three different sets of independent variables with varying levels of information as follows: (1) a simple set of three variables chosen based on hydrologic understanding of the FDC and subsequently called "hydrologic" variables, (2) a typical set of variables that summarize basin characteristics as average statistics called "lumped" variables, and (3) a more complex set of variables consisting of the typical variables and additional variables quantifying the statistical distribution of basin data. The different sets of variables were used to cluster the basins into regions and develop regional regression models for predicting 13 percentile flows. The regional regression approach achieved fair predictive performance based on validation results from 184 basins. Predictive

performance varied with the percentile flows and the different sets of independent variables. The approach performed best for percentile flows related to average conditions and worst for high and low flows subject to more regional variability. Predictive performance declined using the set of independent variables with the most information, and was similar for the hydrologic and lumped variables. This result indicates that variables typically used to predict the FDC offer little predictive information, and variables based on a physical understanding of the FDC are far more important. Future regional regression studies may consider developing new independent variables in light of the limited predictive potential of typical variables. Some of the predictive uncertainty detected in this study may be due to the use of regression, and the next study of this dissertation will evaluate a machine learning method for predicting the percentile flows.

B. Introduction

The flow duration curve (FDC) expresses flow as the percent of time it is equaled or exceeded. These values are called percentile flows, and they are widely used in water resource applications that depend on a minimum flow for a certain percent of time (Vogel and Fennessey, 1995), such as hydropower planning, water use permitting, and water quality management. The shape of the FDC is also strongly tied to the physical and climatic conditions of the contributing drainage basin (Yaeger et al., 2012; Ye et al., 2012; Yokoo and Sivapalan, 2011). This makes the FDC a valuable tool for investigating the basin characteristics associated with regional streamflow patterns.

Percentile flows are frequently needed for basins without flow data or insufficient data to construct a long-term FDC. For these ungauged basins, percentile flows must be predicted using information from surrounding gauged basins. This process is known as

hydrologic regionalization. The types of information used in a hydrologic regionalization procedure can vary depending on the situation (targeted streamflow variable, density of stream gauge network, and data availability). Since stream gauge networks are often too sparsely distributed to directly extrapolate percentile flows, the most common approach is to use basin characteristics to predict percentile flows at ungauged basins (Booker and Woods, 2014). The flow generated by a basin is related to its physical and climatic characteristics (e.g. drainage area, slope, and annual precipitation). These relations are then used in regionalization procedures to predict percentile flows. The performance of regionalization procedures largely depends on the region and independent variables used to develop the predictive relations.

1. Regional regression

In large study areas with potentially heterogeneous basins, regions composed of similar basins are often identified for the process of hydrologic regionalization (Sauquet and Catalogne, 2011). The purpose of these regions is to reduce the variance in percentile flows and enhance their predictability. A variety of prediction methods can be applied in each region in order to predict percentile flows. The most common approach calibrates a regression model to the region in a process called "regional regression". The term was popularized by Stedinger and Tasker (1985), and refers to the two-phase process of (1) designating hydrologic regions and (2) calibrating regression models to predict flow statistics for a region of interest. The region's basin characteristics are used as independent variables to calibrate the regression model. This approach has been adopted as a governmental standard for predicting flow statistics in the US (Ries, 2007) and UK (Robson and Reed, 1999). It has also been widely used to predict percentile flows (see Holmes et al., 2002; Hope

and Bart, 2012; Mohamoud, 2008). Critical components of regional regression are the regions and variables used to formulate the regression models. These two components are reviewed in the following sections, which outline the options for designating regions and different types of variables used for regional regression.

2. Hydrologic regions

Regional regression relies on hydrologic regions consisting of basins with similar percentile flows. Hydrologic regions are often identified geographically under the assumption that the basins of a geographic region have similar flow (see Archfield et al., 2009; Booker and Woods, 2014; Castellarin et al., 2004). However, that may not always be the case, particularly for drier climates where the variability in flow among basins increases (Patil and Stieglitz, 2012). In light of this, analytical methods are also applied to designate hydrologic regions (see Ganora et al., 2009; Holmes et al., 2002; Sauquet and Catalogne, 2011). A common method used in flood prediction is to establish a "region of influence" with a set number of nearby basins (Merz and Blöschl, 2008), but again, this method may be unreliable in drier climates and depends on the spatial density of the stream gauge network.

Other methods utilize the characteristics of basins to place them into clusters that may not be geographically contiguous. Specialized clusters for each basin can be assigned by ranking the most similar basins according to a set of characteristics (Oudin et al., 2008). However, this method requires an arbitrary limit to the number of basins included in the specialized clusters, and may inadvertently include dissimilar basins. Cluster analysis avoids this problem, and has become the preferred method for clustering basins based on a large number of hydrologically relevant characteristics (Sauquet and Catalogne, 2011).

3. Cluster analysis

The goal of cluster analysis in regional regression is to identify hydrologic regions with similar percentile flows. These "regions" are determined in the attribute space of basin characteristics, and are often formulated completely independent of basin location (see Laaha and Blöschl, 2006; Sanborn and Bledsoe, 2006; Srinivas et al., 2008). This means the resulting regions are clusters of basins with similar characteristics that may not be in the same geographic region.

Basin characteristics must be chosen for the cluster analysis. This involves subjective judgment of which basin characteristics should be used to create hydrologic regions. The obvious choice is to cluster the basins based directly on their percentile flows, but ungauged basins cannot be included in the clustering, and they may not be assigned to the appropriate cluster (Sanborn and Bledsoe, 2006). Percentile flows are usually excluded from the cluster analysis so that both gauged and ungauged basins can be included. In lieu of percentile flows, physical and climatic characteristics related to flow are used to identify hydrologic regions via cluster analysis (Olden et al., 2012).

Cluster analysis can take on many forms, but a common aspect of these methods is that they can be used to compute the distance between basins in multivariate attribute space. This measure of basin similarity is then used to assign clusters based on an objective function that attempts to maximize the similarity within clusters and dissimilarity between clusters. A number of these methods have been used for regional regression in hydrology (e.g. *k*-means, regression trees, and hierarchical), and some studies have compared their performance for predicting flow statistics (see Isik and Singh, 2008; Laaha and Blöschl, 2006; Lin et al., 2010). No method was clearly the best for clustering basins as their performance was likely

data specific. However, an important finding from this research is that clustering methods benefit from using derived input variables that reduce the dimensionality of the original basin data (see Di Prinzio et al., 2011; Farsadnia et al., 2014; Srinivas et al., 2008).

The purpose of creating derived input variables with reduced dimensionality is to treat the data for redundant information, noise (erroneous variation), and non-linearities, all of which can be problems for clustering methods. Derived input variables can be created using traditional statistical techniques that generate new variables based on the correlation structure of the data (e.g. principal component analysis) or machine learning techniques that produce new variables according to the underlying patterns in the data (e.g. the self-organizing map). The latter approach has gained ground in hydrology because machine learning accounts for the often complex and non-linear relations between hydrologic variables (Kalteh et al., 2008).

The self-organizing map (SOM) is a machine learning technique that transforms the input variables for the cluster analysis into a set of neuron vectors composed of generalized values for representing the data. The neuron vectors are derived through an iterative training process designed to reduce noise and capture non-linearities in the data (Kohonen et al., 1996). The trained neurons are arranged in a topologically preserving space that represents clusters in the data. Individual neurons could act as clusters in a small SOM. However, this does not allow the clusters to emerge in the SOM and requires the selection of an a priori number of clusters. Using a SOM with far more neurons than potential clusters creates a space in which the data can be organized according to its attributes. The resulting "map" of the attribute space can then be used to explore and define clusters. Exploration of the clusters is carried out through visualization and analysis of the neuron vectors, and clustering

methods are applied on the neuron vectors to cluster the input data. This process is becoming increasingly common for clustering the basins of regional regression studies (see Hall and Minns, 1999; Jingyi and Hall, 2004; Srinivas et al., 2008). The SOM is particularly appealing for clustering basins because of the noise and non-linearity often present in hydrologic data. Unlike other methods to create derived input variables for clustering basins, the SOM provides a platform for characterizing the clusters and their interconnections (Coleman, 2008).

4. Regional regression variables

Regional regression uses physical and climatic basin characteristics as variables for cluster analysis and regression. These variables are collectively referred to as independent variables. They are derived from ancillary data related to flow, and used to identify hydrologic regions and calibrate regression models for the resulting regions. Independent variables used in regional regression typically describe the central tendency of basin characteristics using mean or median statistics (Toth, 2012), and are called "lumped" variables because they use a single value to summarize the distributed data from a basin (e.g. .mean elevation, slope, and soil porosity).

A large variety of lumped variables have been used to predict the FDC, but an effective approach for selecting the variables may be to conceptualize the FDC as a gradient of flows consisting of two end members contributed by storms (highest flow) and groundwater (lowest flow). Flows between the end members are moderated by evaporative losses. Therefore, the factors that shape the FDC are storm and groundwater flows adjusted for evaporative losses. These factors could be represented for a simple, but hydrologically-based regional regression of the FDC using the following "hydrologic" variables:

1. Mean annual precipitation (MAP) accounts for the rainfall and antecedent moisture conditions associated with storm flows (Ye et al., 2012).
2. Baseflow index (BFI) quantifies groundwater flows as the percent of flow contributed by groundwater.
3. Potential evapotranspiration (PET) approximates evaporative losses that moderate the entire FDC (Yokoo and Sivapalan, 2011).

In contrast to the hydrologic variables, the complexity of the independent variables typically used for regional regression studies could be increased by including a variety of lumped variables along with statistics on the deviation about those variables. This set of variables would more closely account for the spatial and temporal distribution in basin data, and are therefore called "distributed" variables. These variables could be derived to characterize the spatial distribution of physiographic data (e.g. standard deviation in soil properties) and temporal components of climatic data (e.g. precipitation seasonality). Regional differences in the FDC may be better explained by distributed variables considering the factors that shape the FDC. This is illustrated by the following examples: (1) precipitation seasonality affects the variability in flow throughout the year and the slope of the FDC (Ye et al., 2012), (2) forest cover in the riparian corridor may reduce the low end of the FDC due to the transpiration from trees (Hope et al., 2009), and (3) variability in soil storage properties may reflect the middle of the FDC since these average flows are largely a function of the storage in a basin (Yokoo and Sivapalan, 2011). Despite the possible explanatory power of distributed variables, few regional regression studies have used them (Toth, 2012), and even fewer studies have investigated the effect of different independent variables on the flow predictions of a regional regression (Hope and Bart, 2012; Ilorme, 2011). More studies are

needed to resolve which type of variables (hydrologic, lumped, or distributed) and how much information is necessary to develop regional regression models. These questions are addressed for a regional regression using physical and climatic variables to cluster basins and model their percentile flows (Figure 4).

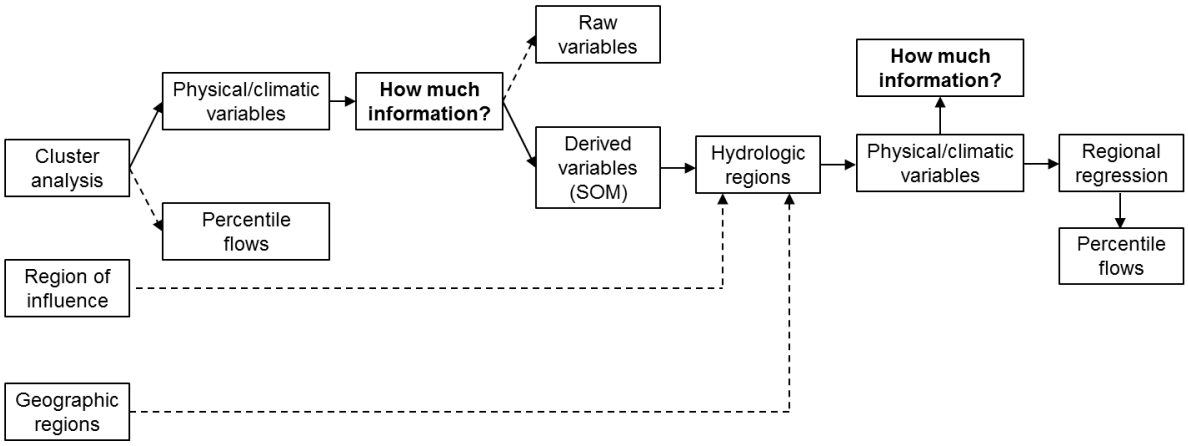

Figure 4. Regional regression approach applied to address the question of how much information is necessary to develop regional regression models. Solid lines indicate the methodological options chosen for this study.

C. Research design

Regional regression to predict percentile flows consists of two phases: (1) clustering basins into regions and (2) calibrating regression models for the resulting regions. Both phases involve the selection of variables used to approximate hydrologic conditions and predict percentile flows. Prior studies have primarily focused on how the basins should be clustered, but far less attention has been given to the variables used to formulate regional regressions. These variables may affect the regional regression results, and can be represented with varying amounts of information ranging from a simple set of hydrologically-based variables to variables describing many factors possibly related to the FDC. The amount of information refers to the number of variables and also how they aggregate the spatiotemporal data of the basins. The variables can either describe the average

of the data or additional information on the distribution of the data. Increasing the amount of information requires more computational effort to derive the independent variables. This regional regression study tests different sets of variables with varying amounts of information to answer the following question:

How much information is necessary for regional regression modeling of percentile flows?

The research question was evaluated using three sets of variables that can be viewed as a hierarchy with increasing numbers of variables and greater computational costs. The first level of the hierarchy only included three "hydrologic" variables chosen based on a conceptual understanding of the FDC gathered from the literature. "Lumped" variables were added to the second level of the hierarchy to describe the average for a variety of basin characteristics. This level of the hierarchy was intended to represent a typical set of variables used for regional regression modeling. The final level of the hierarchy included the lumped variables and additional "distributed" variables that characterize the statistical distribution of the basin data. Each level of the variable hierarchy was tested in a regional regression including a large sample of 918 basins in order to create more generalizable results for future studies (Andréassian et al., 2007). The study consisted of three steps in the evaluation of the different sets of variables: (1) designate hydrologic regions for the entire US according to physical and climatic basin characteristics, (2) develop regional regression models to predict the percentile flows of the FDC, and (3) compare the performance of the regional regression using the three different amounts of information on the basins. Hydrologic regions derived from the most parsimonious set of variables were then characterized to better understand regional patterns associated with percentile flows. This research was motivated by the hypothesis that the regional patterns derived from distributed variables would be more

closely tied to percentile flows due to the complexity of the processes that shape the FDC. A conceptual diagram of this research is given in Figure 5.

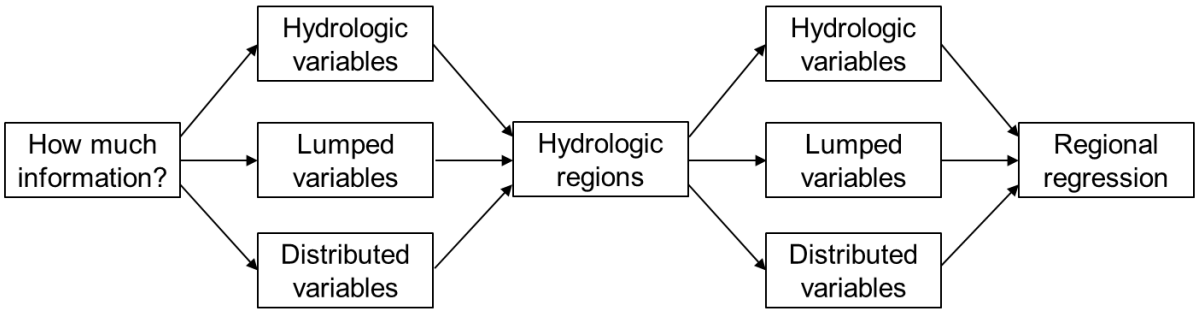

Figure 5. Research design of this study.

D. Methods

1. Overview

A regional regression routine was repeated using three different sets of independent variables (hydrologic, lumped, and distributed). Each set of variables was first fed into a SOM to create a new set of variables based on the neuron vectors generated by the SOM. This intermediary step is described further in the SOM section, and was performed to account for non-linearities and reduce noise in the data. The neuron vectors were then clustered using the *k*-means method as it has proven to be compatible with the SOM (Skupin, 2004). Neuron clusters were linked to the basins according to the neurons that best matched the basins (best-matching unit). Subsequent basin clusters were used as the hydrologic regions for regional regression models to predict 13 percentile flows including the high flow exceeded only 1% of the time (Q_{01}), low flow exceeded 99% of the time (Q_{99}), and flows between that range (Q_{05}, Q_{10}, Q_{20},…Q_{95}).

Regional regression models were calibrated using the independent variables. For the more complex sets of variables (lumped and distributed), a subset of variables was selected using random forests since it performed best in a prior study (see the first paper of this

dissertation) and offers a method for ranking variables based on their potential to predict percentile flows. Variables were selected for each hydrologic region, and used to calibrate the regional regression model. Calibration basins were used to develop the regression models, and their predictive performance was assessed using an independent set of validation basins. Results from the three sets of variables were then compared to determine which amount of information was most parsimonious for the regional regression of percentile flows. The preferred method was used to describe the physiographic and climatic characteristics of the US hydrologic regions. The entire regional regression study is summarized in Figure 6.

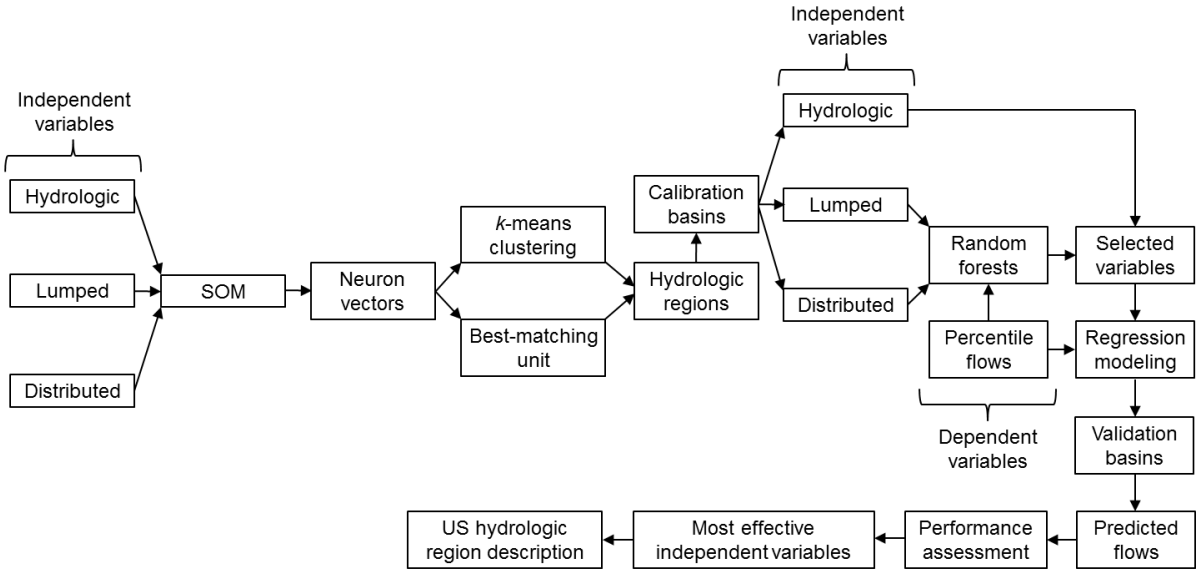

Figure 6. Summary of the regional regression study comparing different sets of independent variables.

2. Basins and variables

This study included 918 basins in the contiguous US classified as "near-natural" by the US Geological Survey's GAGES-II database (Falcone, 2011). The basins all had 30 years of continuous daily streamflow data, which was used to calculate 13 percentile flows ($Q_{01,}$ Q_{05}, Q_{10}, $Q_{20,}$…Q_{95}, Q_{99}). A streamflow record of 30 years is more than sufficient to compute

reliable percentile flows regardless of the record's starting date (Kennard et al., 2010). The percentile flows were computed using the Weibull plotting position (Castellarin et al., 2004), and normalized by the mean of nonzero flows in order to create dimensionless statistics less influenced by drainage area (Hope and Bart, 2011). The dimensionless percentile flows were used as the dependent variables in the regional regression models.

The dataset was split into calibration and validation basins for regression model development (Figure 7). Regression models were tested on 184 (20%) of the basins. A widely used split sampling technique called the "proxy-basin test" (Klemeš, 1986) was applied to split the basins. This technique tests the geographic transferability of the regression model using validation basins that are representative of the calibration basins. A representative sample of validation basins was selected using a stratified sampling approach that grouped the basins according to key factors related to the FDC (climate class, rock type, and drainage area). The validation therefore assessed the performance of the regression models for a variety of conditions with presumably different streamflow regimes. Percentile flows were not used to sample different streamflow regimes to avoid corrupting the validation.

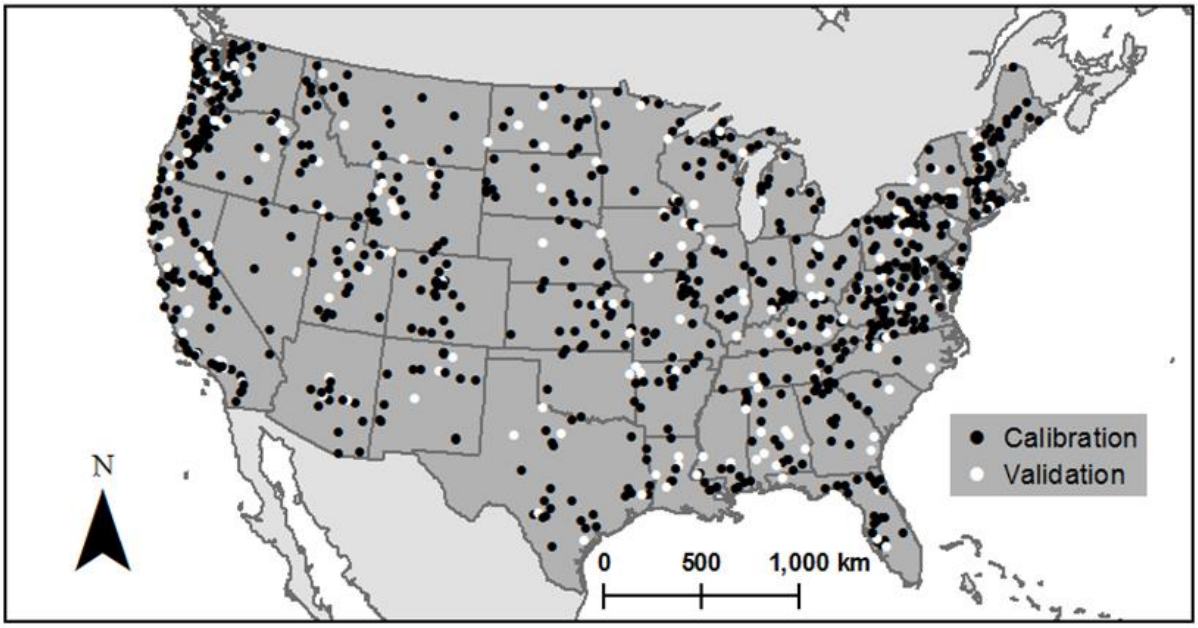

Figure 7. Location of the calibration and validation basins. The study included 918 basins, with 734 calibration and 184 validation basins.

Independent variables were used to cluster the basins into hydrologic regions and develop regional regression models to predict the percentile flows. These variables described the climate, topography, land cover, soil, and geology of the basins. Climatic variables were computed using 30 years of data. Most of them used monthly data concurrent with the streamflow data, but those concerned with storm intensity (Precip_1D_Max and Precip_Intensity) used a fixed timeframe (1981-2010) since the daily precipitation data did not cover all of the streamflow data. Land cover was represented using percent forest cover because it affects the FDC more than other major types of vegetation, such as shrub and grassland (Brown et al., 2005). Percent forest cover was quantified using the National Land Cover Dataset for 1992 as this year coincided with the most 30-year streamflow time periods. The influence of geology on streamflow was depicted using a spatially interpolated grid for BFI (percent of streamflow from groundwater). BFI values from gauged basins were spatially interpolated for the contiguous US (Wolock, 2003). It is acknowledged that the BFI

grid may have been derived using streamflow data from the validation basins. However, the BFI grid was used to create independent variables since it is a preexisting product that can be used by water resource managers to predict percentile flows (Wolock, 2003).

This study included three different sets of variables, and the variables in each set are given in Table 9 as hydrologic (H), lumped (L), and distributed (D). The different sets of variables are nested, with 37 distributed variables and the same, but fewer, variables in the 22 lumped and three hydrologic variables. The hydrologic variables included MAP, PET, and BFI based on a conceptual model of the factors that shape the FDC proposed in the introduction. The next set of variables included lumped variables that summarize physiographic and climatic data distributed in space and time using a single value. The most complex set of distributed variables used additional variables to describe the spatiotemporal distribution of basin data in more detail. The distribution of spatial physiographic data was quantified by its standard deviation in the basins, and percent forest cover was calculated within riparian corridors critical for groundwater discharge. The distributed variables also characterized the temporal dynamics of the climatic data. This was accomplished using (1) the standard deviation of annual storm intensity and aridity statistics, (2) the Fourier transform of the potential evapotranspiration time series to describe its amplitude and peak timing (Dalton, 2005), and (3) the autocorrelation function (Toth, 2012) and circular statistics (Dingman, 2001) to respectively describe the amplitude and peak timing of the precipitation time series. All the variables used in this study are described further in Table 9.

Table 9. Variables used in this study. The final column shows the variables included in the hydrologic (H), lumped (L), and distributed (D) sets of variables.

Variable	Units	Description	Data source	Set
Dependent				
Q_p (e.g. Q_{01} for 1%)	-	Normalized percentile flows for 1, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 95, and 99% exceedance	NWIS	-
Independent				
MAP	mm	Mean annual precipitation	PRISM	H, L, D
PET	mm	Mean annual potential evapotranspiration calculated using the Oudin et al. (2005) equation	PRISM	H, L, D
BFI	%	Mean baseflow index derived from a baseflow grid	BFI48GRD	H, 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	Precipitation per rainy day	PRISM	L, D
Spring_Temp	°C	Average temperature from April-June	PRISM	L, D
Aridity	-	Aridity index calculated as PET divided by 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/km^2	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

Table continued on next page

Variable	Units	Description	Data source	Set
Poorly_Drained	%	Percent poorly drained including hydrologic soil groups C and D	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 (2001)	PRISM	D
Precip_Seasonality	-	Distribution of monthly precipitation throughout the year calculated using circular statistics as in Dingman (2001)	PRISM	D
Precip_1D_Max_SD	mm	Standard deviation of Precip_1D_Max	PRISM	D
Precip_Intensity_SD	mm/d	Standard deviation of annual Precip_Intensity	PRISM	D
PET_Amp	mm	Amplitude of the first term of the Fourier transform as in Dalton (2005)	PRISM	D
PET_Ph	rad	Phase of the first term of the Fourier transform 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

Data sources: NWIS, National Water Information System (http://waterdata.usgs.gov/nwis); PRISM, Precipitation-elevation Regressions on Independent Slopes Model (http://prism.oregonstate.edu); GAGES-II, Geospatial Attributes of Gages for Evaluating Streamflow, version II (Falcone, 2011); NHDPlusV2, National Hydrography Dataset Plus Version 2 (http://www.nhdplus.com); NED, National Elevation Dataset (http://ned.usgs.gov); NLCD 1992, National Land Cover Dataset 1992 (Vogelmann et al., 2001); CONUS-SOIL, Conterminous US multilayer soil characteristics dataset (Miller and White, 1998); BFI48GRD, Base-flow index grid for the conterminous US (Wolock, 2003)

3. Identifying hydrologic regions

Hydrologic regions were identified using the independent variables. To deal with noise and non-linearities in the data, the SOM was applied as a preliminary step to cluster the

basins into hydrologic regions. The basins were clustered according to *k*-means clusters of the trained SOM neuron vectors. Finally, the appropriate number of clusters was determined based on the number of calibration basins per cluster and a variety of cluster validity indices for assessing the optimal number of clusters.

3.1 Self-organizing map (SOM)

The SOM was used to preprocess the data before clustering the basins. This step was performed to create SOM neuron vectors that represent the cluster structure of data containing non-linearities and noise. A separate SOM was produced for each set of independent variables using all the basins. In order to give the variables equal weight, they were normalized using their z-scores calculated as:

$$z = \frac{x_i - \bar{x}}{\sigma_x}, \quad (1)$$

where x_i is the value of the variable for basin i, $\bar{x}$ is the mean of the variable, and σ_x is the standard deviation of the variable. This rescales the variables to have a mean of zero and variance of one.

The rescaled variables were fed through a SOM, which is a grid of neurons with vectors equal in size to the number of variables. The layout of the neurons must be specified to create the SOM. Hexagonal neurons were used instead of squares so that all neighboring neurons share a side. The dimensions of the SOM had an equal number of neurons on either side (*x* and *y*) to limit the boundary effect problem in which neurons at the edge of the SOM fit the data less than internal neurons (Schmidt, 2008). The number of neurons was determined by testing two SOM sizes and assessing the number of neurons that did not represent any of the basins. Limiting the "empty" neurons helped ensure that the neuron vectors were representative of the basin data and subsequent neuron clusters could be

mapped back to the basins. The final SOM had 15 × 15 neurons based on the above conditions and the dimensions used in another study that evaluated various SOM sizes for clustering basins (Srinivas et al., 2008).

SOM neuron vectors were initialized with random values rather than using linear estimates for the initial neuron vectors. Random initialization is preferred for actual self-organization into clusters (Skupin and Hagelman, 2005). The self-organization process iteratively changed the neuron vectors as follows:

1. The data from a basin (b) was compared to every neuron vector (n_i), and the most similar neuron vector (n_c) was determined using the Euclidean metric:

$$\|b - n_c\| = \min_i\{\|b - n_i\|\} \tag{2}$$

2. The neuron vectors for n_c and its neighbors were updated to more closely match the incoming basin data as follows:

$$n_i(t+1) = n_i(t) + \alpha(t)h_{ci}(t)[b(t) - n_i(t)], \tag{3}$$

where t is the iteration number, α is the learning rate, and h_{ci} is the neighborhood function.

3. The learning rate controlled the magnitude of the updates to the neuron vectors and decreased monotonically with each iteration.

4. The neighborhood function decreased the effects on neurons farther from n_c using the following Gaussian equation:

$$h_{ci}(t) = \exp\left(-\frac{d_{ci}^2}{2\Theta^2(t)}\right), \tag{4}$$

where d_{ci} is the horizontal distance between n_c and the neighboring neuron i, and Θ is the width of the neighborhood, which decreased along with the learning rate.

5. The above steps were repeated over all the basins a set number of times to train the SOM.

SOM training was performed as recommended by Kohonen et al. (1996) in two stages: (1) a global training stage outlined the major clusters of the basin data, and (2) the SOM was then fine-tuned using a local training stage to reveal more detailed clusters. The different training stages were accomplished using three parameters: (1) training length (number of runs over all the basin data), (2) learning rate (magnitude of neuron vector updates), and (3) neighborhood radius (number of neurons updated around the central neuron). Training length was determined by plotting the quantization error for each training run. The quantization error summarized the agreement between the neuron vectors and basin data. This value initially decreased and then flattened out during training. The number of runs needed to flatten out the quantization error was then used as the training length. Global training required a much shorter training length than local training since it used a larger learning rate and neighborhood radius for broad-scale effects on the SOM. The learning rate and neighborhood radius both decreased monotonically during training. Initial values for these parameters were chosen based on the training stage and SOM size. The training parameters used for this study are listed in Table 10.

Table 10. SOM training parameters for the first stage of global training and second stage of local training.

Training stage	Training length (runs)	Learning rate (α)	Neighborhood radius (neurons)
Global	50	0.04	8
Local	4,000	0.03	5

3.2 Basin clustering

Basins were clustered into hydrologic regions using the trained SOMs. Each basin was assigned to its most similar neuron, or best-matching unit (BMU), using Equation 2. The

BMUs were then used to assign basins to clusters based on the neuron vectors. The neuron vectors were clustered using the *k*-means method since it creates similar-shaped clusters as the SOM (Skupin, 2004). This is because the *k*-means method uses a similar Euclidean metric to establish clusters:

$$SS = \sum_{j=1}^{k} \sum_{i=1}^{s} \left\| n_i - \bar{n}_j \right\|^2 \quad (5)$$

Neuron clusters were determined using the sum of squared distances (SS) within the clusters (k). Within-cluster distances were calculated for all the neurons (s) as the difference between the vector from neuron i and the average of neuron vectors in cluster j. The sum of squared distances was used as the objective function to find *k* clusters as follows:

1. Cluster centroids ($\bar{n}$) were randomly placed in the input data space of neuron vectors.
2. Each neuron was clustered according to the closest cluster centroid.
3. Cluster centroids were then recalculated to fit the clusters.
4. Steps 2 and 3 were repeated until the cluster centroids no longer changed, and the sum of squared distances within clusters was minimized.
5. The above process was repeated 1,000 times due to the random initialization of cluster centroids, and the cluster solution with the minimum sum of squared distances was used as the final neuron clusters.

Finally, the basins were clustered according to the cluster membership of their BMU.

3.3 Determining the number of basin clusters

The number of basin clusters (*k*) was determined using an approach that considered (1) the range of hydrologic conditions in the US, (2) the number of calibration basins per cluster for subsequent regression model development, and (3) various cluster validity indices to indicate the optimal number of clusters for the dataset. The number of clusters had to be

large enough to accommodate the wide range of hydrologic conditions in the study area. This served as a lower limit for identifying a reasonable number of clusters given the diversity of the basins. A reasonable number of clusters was gathered from previous work on splitting the US into hydrologic regions, which used 12 (Bailey, 1983), 15 (Commission for Environmental Cooperation, 1997), and 20 (Wolock et al., 2004) regions. The number of calibration basins per cluster served as an upper limit for the number of clusters since the clusters had to contain enough basins to calibrate regression models. At least 20 calibration basins have been recommended for regional streamflow predictions (Hosking and Wallis, 1997). This number permits simple regression models based on a small number of independent variables since the small sample size restricts the number of model parameters (Berger, 2004).

The optimal number of clusters was assessed using the following cluster validity indices: (1) silhouette, (2) Davies-Bouldin, (3) Xie-Beni, (4) Calinski-Harabasz, and (5) Dunn. All cluster validity indices were implemented as in Desgraupes (2013), and rescaled so that values closer to zero indicated more optimal cluster solutions. The different indices were chosen to represent the two main categories of methods for identifying the optimal number of clusters. The first category evaluated the number of clusters based on their compactness (Xie-Beni), and the second category considered both cluster compactness and the separation between clusters (remaining indices). Results from the different indices were then compared, and the appropriate number of clusters was identified.

4. Regression model development

Regression models were developed for each hydrologic region, and used to predict 13 percentile flows of the FDC. Model development included (1) independent variable selection

and (2) model calibration using the selected variables. Independent variables from the lumped and distributed datasets were selected using random forests to rank the variables. Random forests used a random sample from the calibration basins to create a regression tree for predicting percentile flows. Regression trees split the data into smaller and smaller bins until they approximated the percentile flow values. The entire process of randomly sampling and creating regression trees was repeated until the error in the withheld basins stabilized. The error in the withheld basins was also used to quantify the importance of the independent variables. Variable importance was quantified by removing the variable from the regression trees, and the error increased for more important variables. Variable rankings derived from random forests may change due to the random sampling. To improve the reliability of the variable rankings, 100 random forests were run, and the average error was used to rank the variables.

Regression models were calibrated using the ordinary least squares method. The number of independent variables used to calibrate the regression model depended on the sample size of the hydrologic region. An independent variable was added to the regression model for every ten calibration basins in the sample. This was an ample sample size considering regional regression studies often use less than ten basins per model parameter (e.g. Archfield et al., 2009; Hashmi and Shamseldin, 2014; Hope and Bart, 2011). The given number of variables was selected from the lumped and distributed variables ranked using random forests. All three hydrologic variables were used in the regression models (sample size > 30). The final phase of model development accounted for both linear and non-linear relations to the percentile flows. Untransformed and natural log-transformed variables were evaluated for the final regression model, and the variable that explained more variance (R^2)

in the percentile flow was adopted. The natural log of the percentile flow was used for all regression modeling due to the skew in flows and risk of violating the assumption that model errors vary evenly (Harrell, 2001). A constant of one was added to the percentile flows because the natural log cannot be calculated for zero values.

5. Regression model validation

The regional regression models were tested on validation basins withheld from all phases of regression model development. Model validation was conducted to assess the performance of the regional regression approach, and determine the effect of using the different sets of independent variables on the performance of the regional regression. Performance was quantified using the following metrics (or goodness-of-fit criteria) as defined in the accompanying references: (1) relative error (RE; Hope and Bart, 2012), (2) coefficient of determination (R^2; Sauquet and Catalogne, 2011), and (3) Nash-Sutcliffe efficiency (NSE; Nash and Sutcliffe, 1970). The performance metrics were calculated using the natural log of percentile flows to diminish the influence of large values in skewed flows (Sauquet and Catalogne, 2011). Due to zero flows, a constant of one was added to calculate RE, and its absolute value was used to calculate the sum of RE. The distribution of predictive performance was evaluated using the RE of individual predictions. These values were also mapped and regressed against the independent variables to investigate the factors contributing to predictive error.

Predictive performance was summarized for each percentile flow using the sum of absolute RE and the R^2 and NSE between observed and predicted values. Overall performance was quantified using the sum of absolute RE and average R^2 and NSE for all the percentile flows.

6. Describing the hydrologic regions of the US

The hydrologic regions used for the regional regression were described to identify regional characteristics associated with the percentile flows. A regional description was performed for the hydrologic regions defined using the preferred set of independent variables based on the results from the regression model validation. The regions were described using the mean z-score (Equation 1) of key independent variables. The key variables included the hydrologic variables conceptually associated with the FDC plus drainage area and mean elevation as they are commonly associated with streamflow. The z-scores were used along with geographic location to assign descriptive labels to the regions and short summaries of their conditions. Regional conditions were then descriptively linked to median FDCs for a sample of regions.

E. Results and discussion

1. SOM size

The SOM was used to map the basins according to their independent variables. A large SOM was initially used to evaluate the distribution of the basins in the SOM and assess the number of empty neurons without information for clustering the basins. The large SOM had 30 × 30 neurons to give nearly every basin a chance of belonging to their own neuron. The number of basins per neuron was mapped according to their BMU (Figure 8). The large number of empty neurons in the large SOM indicated that it should be reduced in size, and this task was accomplished based on the experiments of Srinivas et al. (2008), who evaluated a number of SOM sizes to cluster basins. The most suitable size from that study was used as a reference to scale down the SOM to 15 × 15 neurons. This adequately reduced the number of empty neurons. A larger size was subsequently ruled out since it could only increase the

number of empty neurons. A smaller size was also not considered because the neurons were beginning to accumulate more basins, and individual neurons were not intended to act as clusters. The neurons were instead used to map the basins in a new space according to their similarity. The distribution of the basins using the hydrologic variables is shown in Figure 8. This was the smallest dataset, yet it occupied the most space in the SOM. Larger datasets occupied less of the SOM possibly because the clusters became more well-defined as the basins were described by more variables.

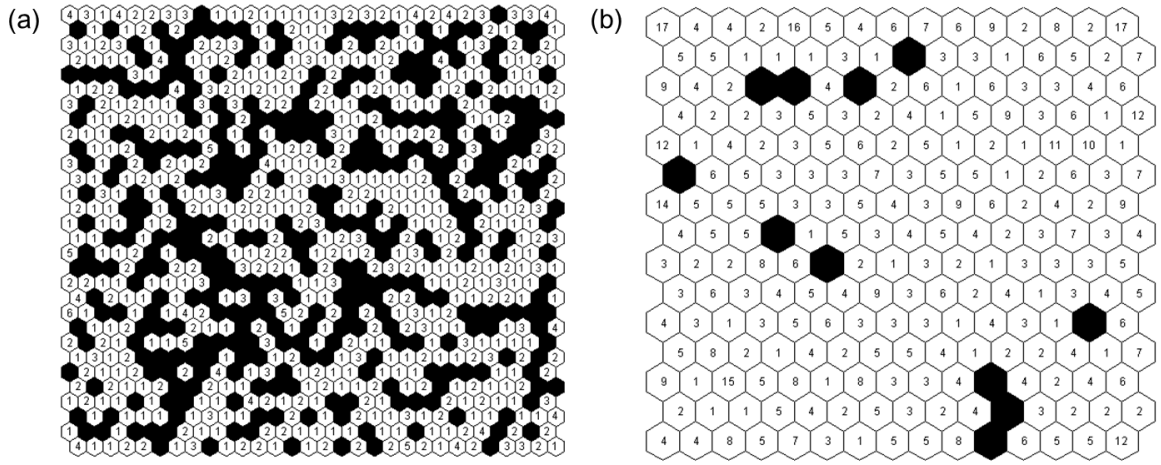

Figure 8. Number of basins assigned to each neuron for the (a) 30 × 30 and (b) 15 × 15 SOM trained using the hydrologic variables. Black neurons indicate empty neurons without basins.

2. Number of basin clusters

The number of basin clusters was determined based on (1) the number of calibration basins per cluster for subsequent model development and (2) the optimal number of clusters according to cluster validity indices. A minimum of 20 calibration basins per cluster was used as an upper limit for the number of clusters. This follows the recommendation that regional streamflow predictions should not use less than 20 calibration basins (Hosking and Wallis, 1997). The maximum number of clusters meeting this requirement is provided in Table 11 for each set of independent variables.

Table 11. Maximum number of clusters (*k*) with at least 20 calibration basins per cluster. Values provided for each set of independent variables.

	Hydrologic	Lumped	Distributed
k	15	17	16

The optimal number of clusters was evaluated using five different cluster validity indices. These values were calculated for each cluster solution of the SOM neuron vectors, and the results for each SOM trained using the different sets of independent variables are shown in Figure 9. Values closer to zero indicate better cluster solutions. The cluster validity indices either steadily decreased (Davies-Bouldin and Dunn) or increased (silhouette, Xie-Beni, and Calinski-Harabasz) with the number of clusters for each set of variables. This is not a useful trait for a cluster validity index because it will give a similar value regardless of the dataset. The cluster validity indices produced similar values (Table 12), and this resulted in an optimal number of clusters that was either greater than the upper limit for regional streamflow predictions (< 20 calibration basins per cluster) or unreasonably low for the 918 basins of the US used in this study.

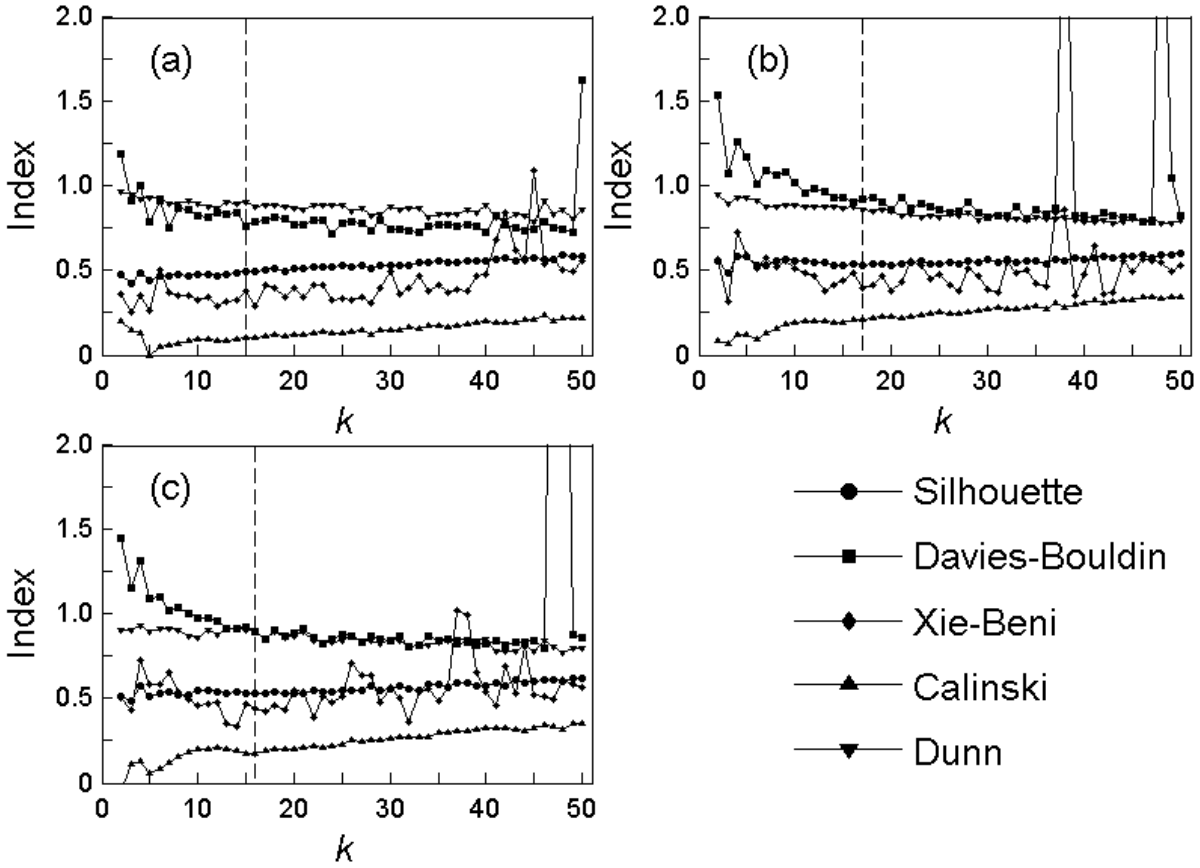

Figure 9. Cluster validity indices for each number of clusters (*k*) starting at two based on the (a) hydrologic, (b) lumped, and (c) distributed variables. Values closer to zero are more optimal cluster solutions, and the dashed line is the upper limit for *k* according to a minimum of 20 calibration basins per cluster.

Table 12. Optimal number of clusters for the different sets of independent variables and cluster validity indices. The upper limit for regional streamflow predictions is also shown for comparison.

	Silhouette	Davies-Bouldin	Xie-Beni	Calinski	Dunn	Upper limit
Hydrologic	3	24	3	5	45	15
Lumped	3	47	3	3	48	17
Distributed	3	46	14	2	48	16

The only optimal clustering solution that had enough calibration basins and a reasonable number of clusters for the US was produced by the Xie-Beni index for the distributed variables, which returned a value of 14 clusters (Table 12). This value was within the upper limit for each set of independent variables (Table 11), and it was reasonable given the number of hydrologic regions previously identified for the US ranging from 12-20 regions (Bailey, 1983; Commission for Environmental Cooperation, 1997; Wolock et al.,

2004). Fewer regions than this range may include more diverse basins in the same region. This could increase the variance in percentile flows and decrease the performance of regional models. A value of 14 was chosen for the final number of clusters based on the combined information of the number of calibration basins per cluster, optimal cluster solutions indicated by the cluster validity indices, and prior hydrologic regions for the US. The resulting regions for each set of independent variables are mapped in Figure 10.

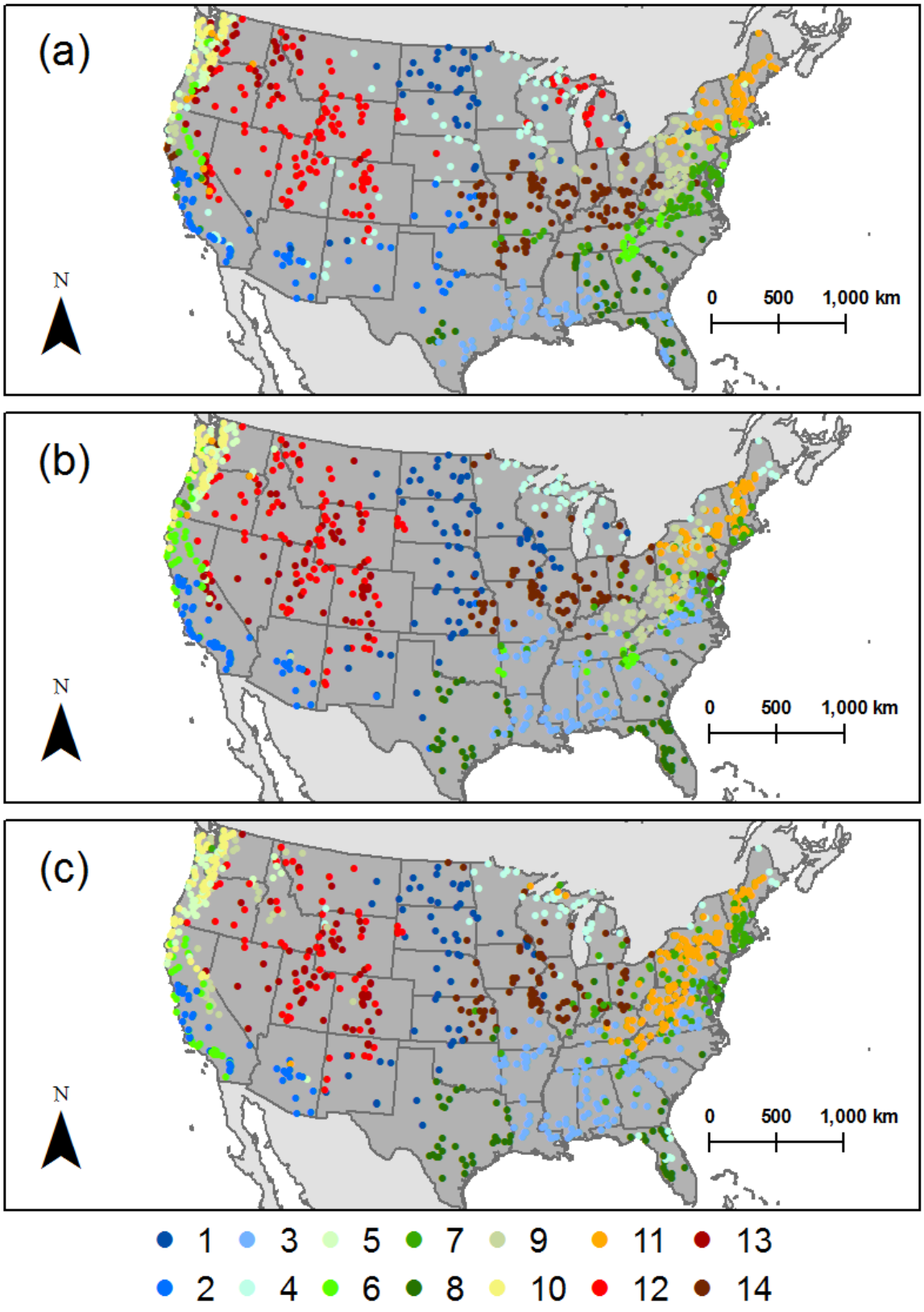

Figure 10. Regions for the (a) hydrologic, (b) lumped, and (c) distributed variables.

3. Regression models

Regression models were developed to predict 13 percentile flows for each region identified using the different sets of independent variables. There was a total of 546 regression models (3 sets of variables × 14 regions × 13 percentile flows). The independent variables for the regression models were selected based on their predictive potential, as measured by univariate regression for the three hydrologic variables and random forests for the more complex lumped and distributed variables. A sample of the regression models is provided in Table 13. The sample includes models from regions that clearly correspond between the separate cluster solutions and represent geographic regions with different hydrologic conditions (Figure 10). For each region, the models for predicting a high (Q_{05}), average (Q_{50}), and low (Q_{95}) flow are shown along with their adjusted R^2 and condition number (CN). The adjusted R^2 measures how well the model explains the flow's variance, and penalizes models with more independent variables. The CN was reported to show the degree of cross-correlation between the independent variables. A CN > 30 is often used to identify high cross-correlation in regression models (Belsley et al., 2004).

Table 13. Sample of regression models for predicting a high (Q_{05}), average (Q_{50}), and low (Q_{95}) flow of selected regions. All models formulated using the natural log of the percentile flows.

	Regression model	Adj. R^2	CN
Region 1 - Central Plains			
Hydrologic			
Q_{05}	$0.37 + 0.62\ln(BFI) - 0.16\ln(PET) + 2\times10^{-5}MAP$	0.11	37764
Q_{50}	$-0.37 + 2.8\times10^{-4}MAP + 4.2\times10^{-3}BFI + 3.4\times10^{-4}PET$	0.44	4057
Q_{95}	$-0.21 + 2.8\times10^{-4}PET + 2.1\times10^{-3}BFI - 6.2\times10^{-7}MAP$	0.28	4057
Lumped			
Q_{05}	$1.8 + 0.12\ln(Poorly\_Drained) + 0.32\ln(BFI) - 4.3\times10^{-4}Elev - 0.26\ln(MAP) - 0.19Aridity + 0.063\ln(Percent\_Snow)$	0.46	68762

Table continued on next page

	Regression model	Adj. R^2	CN
Q_{50}	-1.8 + 0.19ln(BFI) - 0.031ln(Poorly_Drained) + 0.0094Slope + 0.14ln(MAP) + 0.1ln(Precip_SD) + 3.5×10^{-4}Precip_1D_Max	0.69	2056
Q_{95}	0.19 - 0.025ln(Poorly_Drained) + 1.7×10^{-3}BFI - 0.065ln(Aridity) + 6.3×10^{-3}Slope + 4.3×10^{-5}MAP - 0.075ln(TWI)	0.57	56205
Distributed			
Q_{05}	0.058 + 0.34ln(BFI) + 0.21ln(Percent_Snow) - 1.6×10^{-3}Elev_SD - 0.23Aridity	0.58	2286
Q_{50}	-1.7 + 0.23ln(BFI) + 0.18ln(MAP) - 0.23Precip_Seasonality + 1.1×10^{-3}Precip_1D_Max	0.59	2753
Q_{95}	-0.12 + 1.5×10^{-3}BFI + 1.3×10^{-4}MAP + 3.8×10^{-3}Slope_SD - 9×10^{-3}ln(Aridity_SD)	0.35	2401
Region 2 - Southwest			
Hydrologic			
Q_{05}	-5.4 + 0.85ln(MAP) + 0.36ln(BFI) - 3.3×10^{-5}PET	0.52	18997
Q_{50}	-0.76 + 5.3×10^{-3}BFI + 0.13ln(MAP) - 1.2×10^{-4}PET	0.25	17883
Q_{95}	-0.025 + 1.2×10^{-3}BFI + 2.7×10^{-5}MAP - 1.2×10^{-5}PET	0.13	9248
Lumped			
Q_{05}	-1.6 + 0.25ln(Forest) + 0.29ln(MAP) - 0.049Aridity + 0.027ln(Precip_SD)	0.63	342
Q_{50}	-4.8 + 2.3ln(TWI) + 0.016Slope + 3.8×10^{-3}BFI + 1.8×10^{-3}Precip_Intensity	0.43	4663
Q_{95}	-1.4 + 0.74ln(TWI) + 4.9×10^{-3}Slope + 1.2×10^{-3}BFI - 0.045ln(Spring_Temp)	0.31	4976
Distributed			
Q_{05}	6.2 - 1.5ln(Aridity) - 0.65ln(MAP)	0.47	767
Q_{50}	-0.6 + 0.4TWI_SD - 0.084ln(Precip_Intensity)	0.27	74
Q_{95}	0.083 - 0.043ln(Precip_Intensity) + 4.2×10^{-5}Area	0.33	4113
Region 11 - Northeast			
Hydrologic			
Q_{05}	5.8 - 0.3ln(MAP) - 0.35ln(PET) + 9.2×10^{-4}BFI	0.21	4076
Q_{50}	-0.65 + 3.4×10^{-4}PET + 0.14ln(MAP) - 2.5×10^{-3}BFI	0.08	40143
Q_{95}	0.33 + 0.15ln(BFI) - 0.14ln(PET) + 8.1×10^{-5}MAP	0.12	90576
Lumped			
Q_{05}	3.5 - 0.32ln(MAP) + 0.075ln(Water_Capacity) + 0.068ln(Percent_Snow) - 7.8×10^{-3}Spring_Temp + 0.019ln(Density)	0.50	558

Table continued on next page

	Regression model	Adj. R^2	CN
Q_{50}	0.086 - 0.86Aridity - 2×10^{-3}Percent_Snow - 0.19ln(MAP) + 0.32ln(PET) + 0.011Spring_Temp	0.47	8128
Q_{95}	-0.047 - 0.034Density - 0.11ln(Aridity) + 1.7×10^{-4}Precip_SD + 0.019ln(Spring_Temp) - 1.9×10^{-6}MAP	0.28	77136
Distributed			
Q_{05}	6.6 - 0.26ln(BFI) - 0.21ln(Precip_1D_Max) - 0.064ln(Precip_Seasonality) - 0.062ln(Elev) - 0.45ln(PET) - 0.093ln(Percent_Snow) + 0.032ln(Poorly_Drained) + 9.7×10^{-3}ln(Elev_SD) - 0.022ln(Precip_1D_Max_SD) + 1.8×10^{-3}PET_Ph	0.61	1000
Q_{50}	0.1 + 0.21ln(BFI) + 0.065ln(Precip_Seasonality) + 0.2ln(Percent_Snow) - 0.04ln(Poorly_Drained) - 0.68Aridity + 1.3×10^{-3}PET + 0.035ln(Aridity_SD) - 0.028ln(Precip_1D_Max) - 0.14ln(MAP) - 5×10^{-3}Spring_Temp	0.56	184073
Q_{95}	0.059 + 4.2×10^{-3}BFI + 5.1×10^{-5}Elev + 0.019ln(Soil_Porosity_SD) + 0.28Aridity_SD - 0.025ln(Poorly_Drained) + 0.016Density - 0.045ln(TWI) - 9.3×10^{-3}ln(PET) - 7.6×10^{-6}Elev_SD + 2.8×10^{-3}ln(Percent_Snow)	0.63	149194

The adjusted R^2 values of all the models ranged from poor (< 0.2) to good (> 0.8), and averaged 0.48. Variation in the adjusted R^2 values for the different models depended on the region and percentile flow. Regions with more calibration basins tended to have models with larger adjusted R^2 values. Models using the lumped and distributed variables benefited from having additional independent variables in regions with more calibration basins. Adding more independent variables typically increased the adjusted R^2 values although this statistic penalizes more complex models. The number of independent variables did not change for the models that used the three hydrologic variables, and these models demonstrate how well the hydrologic variables explained the variance in flow for regions with different conditions.

The hydrologic variables produced regional models with average adjusted R^2 values ranging from 0.11-0.70. This indicates large variability in the predictive potential of the hydrologic variables depending on regional conditions. A sample of this variability is shown for three regions in Table 13. None of the three regions were adequately modeled using the

hydrologic variables. However, the relative fit of the models can provide some insight into the connection between the hydrologic variables and regional conditions. The hydrologic variables were most effective for the Central Plains and Southwest regions. Streamflow in the Central Plains region is associated with groundwater discharge from the northern Great Plains aquifer system (Downey and Dinwiddie, 1988), and this component of the percentile flows for this region may have been adequately represented using the BFI. The Southwest region is characterized by intermittent streams with percentile flows generated by storms (Yaeger et al., 2012). The hydrologic variables were related to storm flow generation (MAP) and losses (PET and BFI). The Northeast region had the weakest connection to the hydrologic variables because they did not account for the effect of snow on the percentile flows. This is evident from the inclusion of snowfall (Percent_Snow) and snowmelt (Spring_Temp) in the more complex models for the Northeast region, and previous efforts to model the FDC in the northeastern US have highlighted the importance of snowmelt (Ye et al., 2012).

The adjusted R^2 of the models created using the different sets of variables varied with the percentile flows. Regional models were summarized by high (Q_{01}-Q_{20}), average (Q_{30}-Q_{70}), and low (Q_{80}-Q_{99}) flows. Average flows were modeled the most effectively, with the largest average adjusted R^2 value (0.53). The average adjusted R^2 value decreased to 0.46 for the high flows, and was smallest for the low flows at 0.42. Regression models of percentile flows in previous studies have also explained more variance for average flows and less for extreme flows (Archfield et al., 2009; Hashmi and Shamseldin, 2014; Mohamoud, 2008). This indicates that the independent variables used in these models are less effective at representing the lateral transport of water during high flows and subsurface drainage

sustaining low flows. The models in this study likely explained the low flows the least because they included zero flows.

The cross-correlation between the independent variables was high according to the sample of CNs provided in Table 13. The minimum CN of all the regression models was greater than the threshold used to identify high cross-correlation (CN > 30). Thus, all the models had high cross-correlation. Models with high cross-correlation are a concern for two reasons: (1) the effects of individual independent variables can no longer be evaluated and (2) the model may generate less accurate predictions for a dataset with different cross-correlation between the independent variables (Baguley, 2012). The high cross-correlation in this study was not a concern since quantifying the effects of basin characteristics on percentile flows was not a goal and percentile flow predictions were generated for a representative sample of the basins.

Although cross-correlation was not a concern from a regression modeling standpoint, it is reported to highlight the redundancy in the independent variables. The lumped and distributed variables used in the regression models were selected based on their predictive potential. Therefore, it can be concluded that the variables with the most predictive potential provided redundant information to the regression models. The three hydrologic variables were chosen to represent different components of the FDC, but two of these components were related. The groundwater flows of the FDC were represented by BFI, while the evaporative losses that moderate the FDC were approximated using PET. Both of these variables were moderately correlated (Pearson = -0.58 and Spearman = -0.63), and this resulted in high multicollinearity for the regression models that used the hydrologic variables. However, neither variable was eliminated because screening variables for

multicollinearity impeded the performance of regression models in the first paper of this dissertation.

The independent variables for the regression models that used the lumped and distributed variables changed for the different percentile flows. The percent of models that included the variables can be viewed as a measure of variable importance, and is summarized for high (Q_{01}-Q_{20}), average (Q_{30}-Q_{70}), and low (Q_{80}-Q_{99}) flows in Table 14. BFI was included in the most models for each type of flow, and was the most important variable for predicting percentile flows from the lumped and distributed variables. BFI was expected to be a strong predictor of average and low flows because they are fed by groundwater (Cheng et al., 2012). High flows are strongly related to annual precipitation (Yokoo and Sivapalan, 2011), and these flows may have been associated with BFI since it varies with climatic conditions (Santhi et al., 2008).

Table 14. Importance of the independent variables as indicated by the percent of regression models that used the lumped and distributed variables to predict high (Q_{01}-Q_{20}), average (Q_{30}-Q_{70}), and low (Q_{80}-Q_{99}) flows.

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)

BFI was one of the three hydrologic variables proposed as the most important for predicting percentile flows. The other two hydrologic variables (MAP and PET) were used less frequently for the regression models based on the lumped and distributed variables

(Table 14). MAP was consistently included in the models that used the lumped variables, but was only in the top five of the distributed variables for low flows. PET was absent from the top five, except for models that used the distributed variables to predict high flows. The information from MAP and PET may have been adequately represented using Aridity (PET/MAP), which was commonly used in the models for different flows. Aridity was especially important for predicting high and average flows based on its place in the top five of both the lumped and distributed variables. High flows are generated by storms that transport water via shallow subsurface and surface runoff (Cheng et al., 2012). These processes are enhanced as antecedent moisture increases (Yokoo and Sivapalan, 2011), and Aridity quantifies excess water that may contribute to antecedent moisture. The amount of excess water approximated by Aridity may also indicate the groundwater storage of a basin that supplies average flows.

The distributed variables included the lumped variables and additional variables for characterizing the statistical distribution of the lumped variables. The additional variables did not consistently enter the regression models that used the distributed variables (Table 14). Only two of the additional variables (Precip_Seasonality and Aridity_SD) were among the top five of the distributed variables used most often to predict the different types of flow. This indicates that the additional variables had limited potential for predicting the percentile flows, and the typically used lumped variables were sufficient for the regional regression models. Although the additional variables were not commonly used in the regression models, the lumped variables that were selected for the models changed. This may have occurred because the importance of the lumped variables changed alongside the distributed variables. The distributed variables may have accounted for some of the information in the lumped

variables, and the importance of certain lumped variables may have been reduced. This could have changed the order in which the lumped variables entered the regression models.

Lumped variables that also ranked among the most used distributed variables for the same type of flow were considered important for predicting that component of the FDC. A few of these variables have not been discussed (Percent_Snow, Elev, and Poorly_Drained), but their importance should not be overlooked. Percent_Snow was important for predicting the high flows of snow-dominated regions (see Regions 1 and 11 in Table 13). This variable helped explain the high flows during the spring snowmelt season. Elev was used in over a third of the models for predicting average flows because it is related to climatic conditions and groundwater flows that influence the middle of the FDC. Elev is strongly associated with precipitation and snowfall (Grünewald et al., 2014), which in turn have an effect on average flows (Kult et al., 2014). Groundwater flows that contribute to average flows may diminish in higher elevation, headwater basins (Schaller and Fan, 2009). Poorly_Drained is an indicator of infiltration capacity, and was commonly used to predict low flows produced by groundwater.

4. Predictive performance

The regional regression approach was tested on 184 validation basins to (1) assess its predictive performance and (2) determine the amount of information required for the independent variables. Predictive performance was assessed using the relative error (RE), coefficient of determination (R^2), and Nash-Sutcliffe efficiency (NSE) between observed and predicted percentile flows. The distribution of predictive performance was examined using the RE of individual predictions, and the results are summarized by the box plots in Figure 11. Overall, the regional regression approach predicted the percentile flows with < 5% RE for

more than half of the basins and < 10% RE for more than three quarters of the basins. The whiskers of the box plots extend to 1.5 times the interquartile range, which is a common method for identifying outliers (Tukey, 1977, pp. 43-44). The upper range of RE (without outliers) was < 20%.

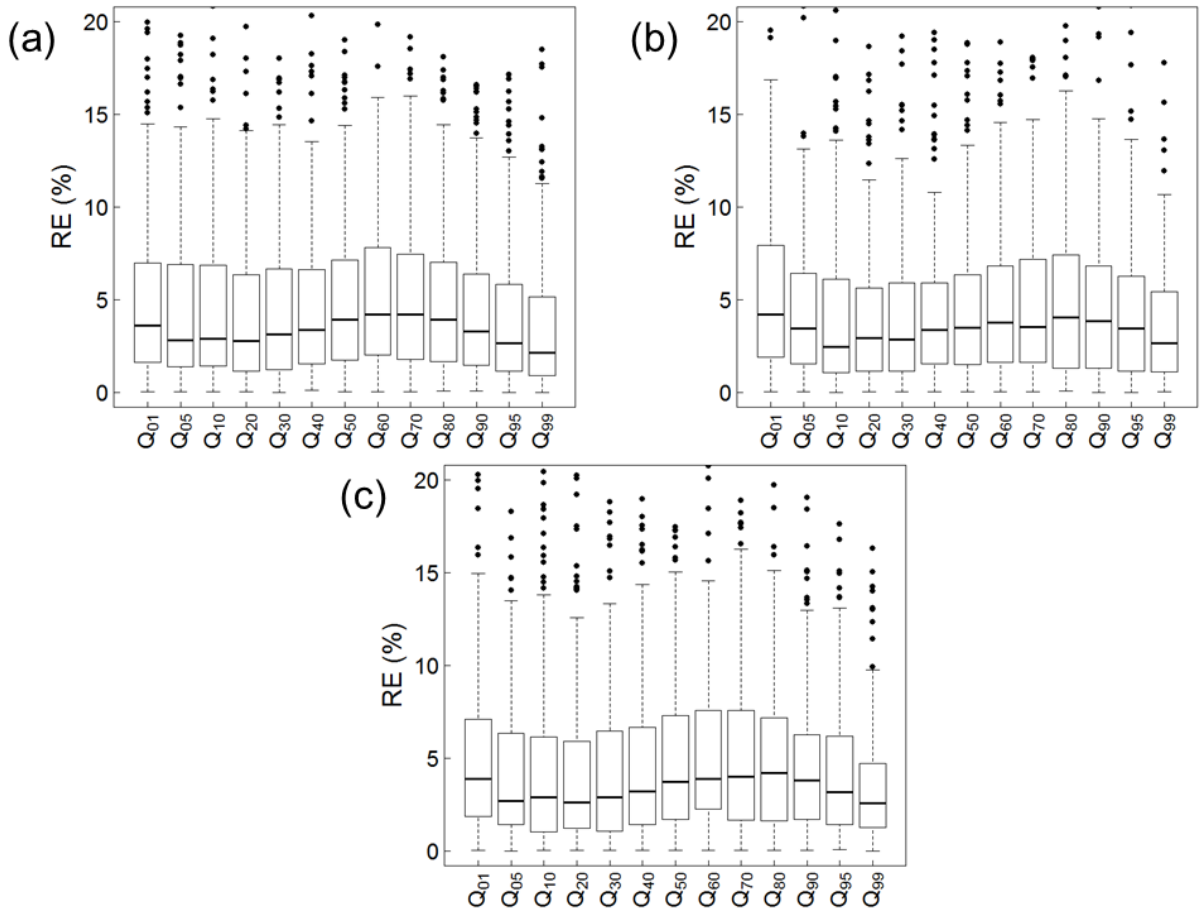

Figure 11. Box plots of absolute RE expressed as a percent for the percentile flows predicted using the (a) hydrologic, (b) lumped, and (c) distributed variables. The boxes show the median, first quartile, and third quartile, and the whiskers extend to 1.5 times the interquartile range. Points outside the whiskers are outliers.

The distribution of RE varied with the percentile flows (Figure 11), decreasing from Q_{01}-Q_{10}, increasing from Q_{20}-Q_{80}, and decreasing once more from Q_{90}-Q_{99}. The largest RE was often encountered for the highest flow (Q_{01}) and average to low flows (Q_{60}-Q_{80}). The highest flow represents flood events, which are notably difficult to predict due to potentially

large variability between basins (Salinas et al., 2013). The precipitation variables used in this study may have inadequately depicted extreme storms responsible for the highest flow. Average and low flows are mostly contributed by groundwater, and larger RE associated with these flows may reflect the uncertainty of BFI estimates or the need for variables to better represent subsurface storage, such as aquifer thickness or water table depth. The smallest RE was often achieved for high to average flows (Q_{10}-Q_{30}) and the lowest flow (Q_{99}). High to average flows integrate both precipitation and groundwater inputs (Yokoo and Sivapalan, 2011), which were sufficiently explained using MAP, Aridity, and BFI (see Table 14). Although RE is a normalized statistic, the small magnitude of the lowest flow may have given it the smallest range of RE.

The box plots in Figure 11 show the RE results for the different sets of independent variables, but the differences in predictive performance were minor at this level of detail. Summary statistics were used to further investigate the effect of the independent variables and overall performance of the regional regression approach.

Predictive performance was summarized for each percentile flow using the sum of absolute RE, R^2, and NSE (Table 15). The latter two performance metrics have an upper limit of 1 (perfect performance), and were used to evaluate the overall performance of the regional regression approach. Both metrics indicate similar performance, but NSE is consistently slightly lower. The regional regression approach had NSE values ranging from 0.39-0.76 depending on the independent variables and percentile flows. NSE increased for the percentile flows from Q_{01}-Q_{30}, and decreased thereafter for the lower percentile flows. This trend resulted in stronger predictive performance for the average flows in the middle of the FDC and weaker predictive performance for the extreme flows at the tails of the FDC. The

sum of absolute RE did not follow this trend as it was likely influenced by the magnitude of the percentile flows.

Table 15. Predictive performance of the different sets of independent variables for the percentile flows quantified as (a) the sum of absolute RE, (b) R^2, and (c) NSE. Bold numbers indicate the set of independent variables that performed the best for each percentile flow according to the given performance metric.

(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}
Hydrologic	**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}
Hydrologic	**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}
Hydrologic	**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

Predictive performance was strongest for the percentile flows in the middle of the FDC (Table 15b and c). This is typical of regional FDC studies (Booker and Woods, 2014; Hope and Bart, 2012; Sauquet and Catalogne, 2011), and may be due to the diminished variability of average flows for regions with similar physical and climatic conditions. Particularly useful variables for predicting the average flows were Aridity, Elev, and BFI (see Table 14), and their connection to average flows is described in the previous section on the regression models. Although the average flows were predicted best, the performance of the regional regression approach still typically only qualified as fair (NSE < 0.75) according to the NSE categories of Castellarin et al. (2004). The gap in predictive performance for the

average flows may be due to (1) uncertainty in the estimation of important variables like MAP, PET, and BFI, (2) large variance in the percentile flows of the regions used for the regression models, and (3) a limitation in the regression method for predicting the percentile flows. The regression method required a subset of the independent variables in order to reliably estimate the model parameters. A method that assimilates all the information from the variables may produce improved predictions. The variables of the regression models often had non-linear relations (see Table 13), and may be subject to noise (error in the independent variables and percentile flows). These complexities may be more effectively modeled using a machine learning method, such as the SOM. Machine learning methods may improve predictions because they can be used to develop non-parametric models that can account for non-linearities and give less weight to outliers that may be the product of noise (Maier et al., 2010).

The worst performance occurred for the extreme percentile flows at the tails of the FDC (Table 15b and c). High flows generated by storms were not well represented by the independent variables, and additional variables may be needed to quantify storms and the physical factors that affect storm flows. Storms were quantified using intensity statistics (Precip_1D_Max and Precip_Intensity) and their annual standard deviation (Precip_1D_Max_SD and Precip_Intensity_SD), but these variables were not frequently used in the regional regression models for high flows (see Table 14). Alternative variables for quantifying the magnitude of large storms may be more closely related to the high flows, and this could be accomplished using precipitation percentiles calculated like percentile flows. Another helpful variable for predicting high flows may be storm frequency as it relates to antecedent moisture conditions that can affect storm flows (Yokoo and Sivapalan, 2011).

Physical factors may also affect storm flows through rainfall interception and infiltration. These processes can be represented using land cover and soil variables, but land cover is preferred because it can be accurately measured using remote sensing. Forest cover was used in this study to account for the effect of land cover on storm flows, but it was not frequently used to predict high flows (see Table 14). Additional land cover variables, such as canopy density and bare soils, may have improved high flow predictions.

Predictive performance declined for the low percentile flows (Table 15b and c). This is commonly reported in the literature (Holmes et al., 2002; Hope and Bart, 2012; Ries, 2007), and may be due to increased variability of low flows between basins. Low flow variability is especially large in arid climates, where extremely low flows are prone to measurement error (Best et al., 2003) and zero flows are difficult to predict (Snelder et al., 2013). Arid climates are common in the western US (Peel et al., 2007), and may have decreased the performance of the regional regression models for the low flows. A possible solution to this problem is to screen the dataset for arid basins and develop predictive models specifically designed for intermittent streams (Croker et al., 2003; Hope and Bart, 2011; Pumo et al., 2014).

The decline in predictive performance for the low flows may also be due to a lack of useful independent variables. Low flows are supplied by groundwater during dry periods, and are therefore controlled by subsurface storage and evaporative losses when conditions are dry (Yokoo and Sivapalan, 2011). Subsurface storage was represented using BFI, which proved to be an important variable for predicting the low flows (see Table 14). BFI seems to have adequately represented subsurface storage because the regional regression models that all used BFI (hydrologic variables) performed better than the models that may not have included

BFI (lumped and distributed variables). However, additional variables describing subsurface drainage, such as a hydrogeologic classification (Tague and Grant, 2004), may improve low flow predictions. Low flows are also moderated by evaporative losses that may not have been adequately represented by any of the independent variables. Alternative variables describing evaporative losses during dry periods may be more useful, and this could be represented by PET during dry days or over the course of the dry season. Due to the error in calculating PET, another useful variable may be to simply quantify the duration of dry periods as a surrogate for the evaporative losses during low flows. Evaporative losses are influenced by the transpiration from vegetation, and the transpiration in riparian corridors can influence low flows by decreasing groundwater seepage into the stream (Smakhtin, 2001). Riparian vegetation conditions characterized using remotely sensed vegetation indices are strongly related to transpiration (Nagler et al., 2005), and can be used to produce variables that may help to predict low flows.

The different sets of independent variables (hydrologic, lumped, and distributed) used to perform the regional regression approach affected the predictive performance for the percentile flows (Table 15). The relative performance between the different sets of variables was similar for the various performance metrics. The simple set of hydrologic variables produced the best regional regression models for predicting the extreme percentile flows at the tails of the FDC, whereas the lumped variables typically used for regional regression slightly improved performance for the percentile flows in the middle of the FDC. The set of distributed variables with the most complexity consistently resulted in the worst performance for all the percentile flows. The additional variables that described the statistical distribution of the basin characteristics were not useful (see Table 14), and altered the percent of models

that used the more important variables. The use of MAP and Aridity declined for regression models developed with distributed variables, although MAP and Aridity helped the performance of the regional regression based on the lumped variables. This indicates that the distributed variables added an unnecessary layer of complexity that obscured the more important variables. The increased complexity of the distributed variables was not necessary for the regional regression approach, contradicting the hypothesis that the statistical distribution of basin characteristics are tied to percentile flows.

The hydrologic variables composed of only three variables resulted in similar predictive performance to the more complex set of lumped variables for the different percentile flows (Table 15). The additional information of the lumped variables only slightly improved predictive performance for the percentile flows in the middle of the FDC, while the simple set of three hydrologic variables produced the best performance for the more challenging to predict percentile flows at the tails of the FDC. These results demonstrate the importance of using independent variables with a physical connection to the streamflow variable targeted for prediction, and deemphasize the use of many independent variables as in data mining approaches. Such approaches may produce models with weak physical connections to the streamflow variable that are less reliable in ungauged basins. The use of many independent variables may be unnecessary in light of the redundancy and limited predictive potential of the variables used in this study. Only three carefully selected variables were necessary to perform the regional regression approach, and other typically used variables offered little, if any, improvement.

The effectiveness of the hydrologic variables was further confirmed by the overall performance of the regional regression approach summarized for all the percentile flows

(Table 16). The hydrologic variables produced the best overall performance for the regional regression approach according to the average R^2 and NSE, and only resulted in slightly more RE than the lumped variables. The success of the hydrologic variables hinged on representing the three major components of the FDC (storm flow, groundwater flow, and evaporative losses). These components were respectively represented using MAP, BFI, and PET. However, results from this study indicate that the hydrologic variables may be improved by combining the information from MAP and PET into an aridity index and using the third variable to account for evaporative losses during the low flows of dry periods. Alternatively, high and low flows may require currently unavailable variables for groundwater-surface water interactions, such as detailed soil moisture measurements for large areas, or national aquifer mapping efforts to characterize subsurface drainage. Despite these potential improvements, the hydrologic variables were still effective for executing the regional regression approach, and the regional regression results based on the hydrologic variables are used to investigate the factors related to RE and regional relations to the FDC in the following sections.

Table 16. Overall performance of the different sets of independent variables quantified as the sum of absolute RE and average R^2 and NSE for all the percentile flows. Bold numbers indicate the set of independent variables that performed the best overall according to the given performance metric.

	RE	R^2	NSE
Hydrologic	116	**0.67**	**0.66**
Lumped	**115**	0.65	0.65
Distributed	120	0.62	0.61

5. Factors related to relative error

The factors related to RE were investigated to identify the conditions associated with predictive error. This was done using the RE from the validation of the regional regression models based on the hydrologic variables since these simple models were highly effective for

predicting the percentile flows. Values of RE were categorized according to their percentile, and mapped to evaluate the geographic variation of RE for a high (Q_{05}), average (Q_{50}), and low (Q_{95}) flow of the validation basins (Figure 12). The maps illustrate that RE varied throughout the validation basins, and there were no obvious regional clusters of RE. This was a favorable result for the regional regression since it means that the approach did not produce large RE for certain parts of the country. The maps do, however, indicate that the RE of the high and average flow varied with aridity. This is evident from the larger RE in drier regions (see the Southwest and Central Plains) and smaller RE in wetter regions (see the Northwest and Northeast). The flow between basins may vary more in drier regions (Salinas et al., 2013), and this likely made the percentile flows more difficult to predict. The RE of the low flow was not clearly related to any factor based on the maps, and the factors related to RE were further examined using the independent variables.

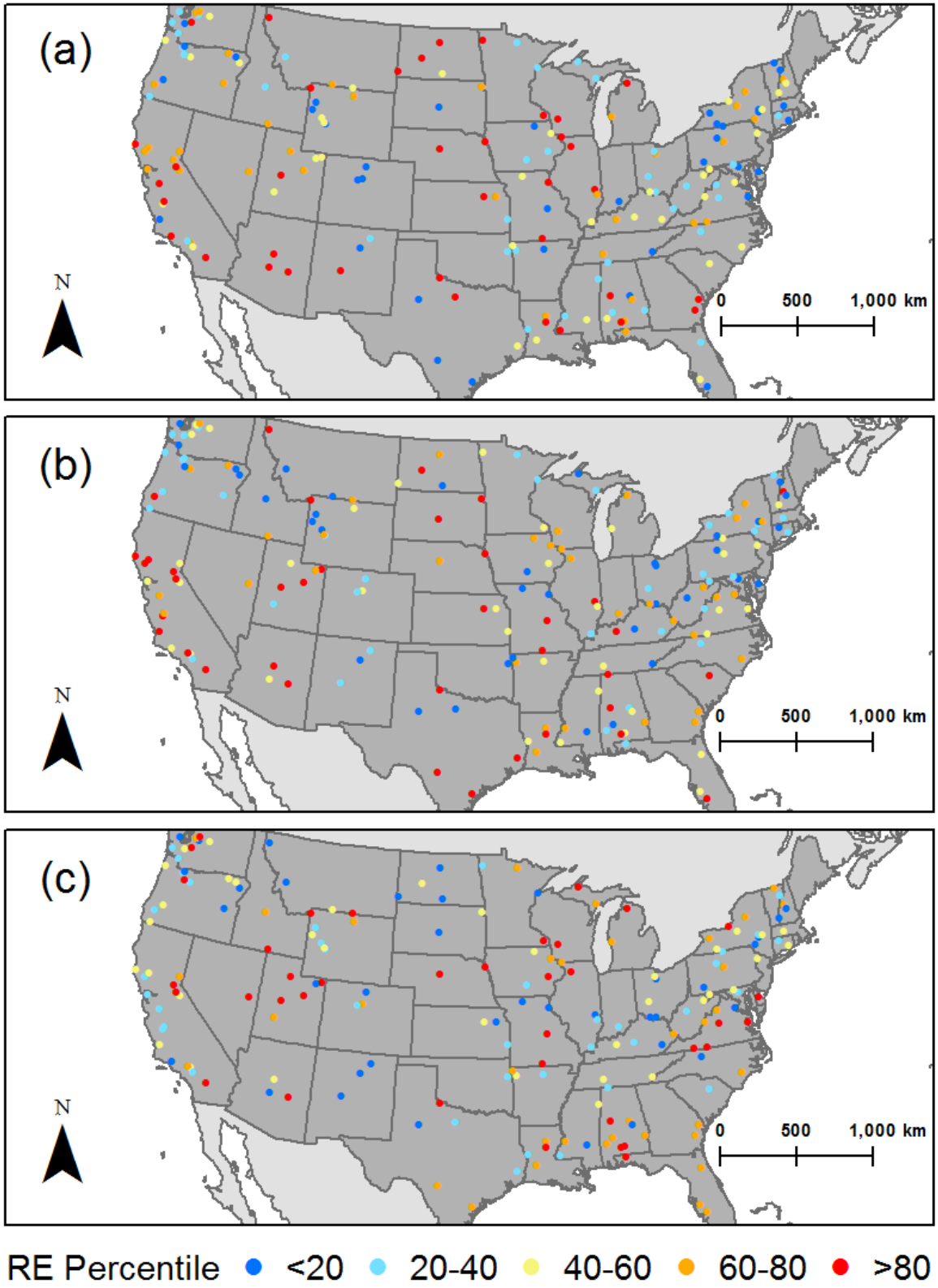

Figure 12. Geographic variation of RE for predicting one (a) high (Q_{05}), (b) average (Q_{50}), and (c) low (Q_{95}) flow of the validation basins. Values of RE are categorized according to their percentile, with lower percentiles indicating less RE.

The correlation between RE and the independent variables were evaluated using the Pearson correlation coefficient (Table 17). These statistics were calculated using the RE of a high (Q_{05}), average (Q_{50}), and low (Q_{95}) flow from the validation. Both the untransformed and semi-log transformed correlation coefficients were evaluated to account for linear and non-linear correlation to RE. The larger of the two correlation coefficients was then used to rank the independent variable. This process confirmed that Aridity was the dominant factor related to the RE of the high and average flow. The effect of Aridity on predictive error is well-documented in the literature (Salinas et al., 2013), and is attributed to the increased variability of streamflow generating processes in more arid regions. Precipitation can vary more between basins in arid regions due to complex terrain and local storm systems (Pilgrim et al., 1988), and this may be the reason why the RE of the high flow increased with Aridity in this study. The RE of the high flow also varied with forest cover, which is likely due to a correlation with Aridity, but worth noting because it further demonstrates the difficulty of predicting flows in drier environments with less forest cover.

Table 17. Relations between RE and the independent variables ranked according to the Pearson correlation coefficient (*r*). The largest correlation coefficient between the two untransformed or semi-log transformed variables was used to account for linear and non-linear relations to RE, and these values were generated for the RE of a high (Q_{05}), average (Q_{50}), and low (Q_{95}) flow in validation. All statistically significant relations to RE are shown (*p*-value < 0.05).

Q_{05}		Q_{50}		Q_{95}	
Variable	*r*	Variable	*r*	Variable	*r*
Aridity	0.49*	Aridity	0.34*	BFI	0.34*
Aridity_SD	0.45*	Aridity_SD	0.34*	Poorly_Drained	-0.25*
Forest	-0.36*	MAP	-0.22*	Slope_SD	-0.17*
Forest_Rip	-0.34*	PET	0.2*	Slope	-0.16*
MAP	-0.29*	Precip_Seasonality	0.19*	Soil_Porosity_SD	-0.15*
Precip_Seasonality	0.29*	Percent_Snow	-0.18*	PET	0.15*
Soil_Porosity	-0.22*	Soil_Porosity	-0.18*	Spring_Temp	0.14
BFI_SD	0.19*	Precip_Intensity_SD	0.17*	PET_Amp	0.13
Precip_Lag1	0.18*	Spring_Temp	0.16*	Precip_SD	-0.13
PET_Amp	0.16*	Forest_Rip	-0.16*	Precip_1D_Max_SD	0.12
Percent_Snow	-0.16*	Aspect_SD	-0.14	Elev_SD	-0.12
Precip_Intensity_SD	0.15*	Water_Capacity	-0.11	Density	-0.12
Precip_SD	-0.13	Relief_Ratio	0.11	TWI	0.12
Precip_1D_Max	-0.11	Area	0.11	Precip_Lag1	-0.11
Area	-0.11	PET_Amp	0.1	MAP	-0.11

*p-value < 0.05

The average flow encountered more RE as Aridity increased (Table 17), and this may be explained by the increased spatial variability of precipitation and resulting saturated areas contributing to average flows in arid regions (Morin et al., 2006). Unsaturated areas in arid regions can lead to considerable amounts of bank recharge (Pilgrim et al., 1988), which may increase the uncertainty of predicting percentile flows in arid regions. Error may also be greater in arid regions with smaller flows subject to increased gauging error (McMillan et al., 2012). The complexity of arid regions warrants the development of streamflow models specifically designed for such conditions (Pilgrim et al., 1988), and treating the arid basins separately may reduce the error in predicting their percentile flows.

The high and average flow shared an interesting relation with snowfall (Percent_Snow). As snowfall increased, the RE of the high and average flow decreased

(Table 17). The smaller RE associated with more snowfall may be due to more uniform runoff in snow-dominated regions (Saco and Kumar, 2000). In these regions, large storm systems deliver the winter snowpack, and subsequent runoff in the spring snowmelt season may be fairly consistent from basin to basin. The other side of this is that warmer regions with less snowfall may have more spatial variation in climatic patterns that affect flow, and this may increase the error in predicting high and average flows.

The main factor related to the RE of the low flow was BFI and other variables related to basin storage (see Poorly_Drained and Slope in Table 17). The correlation coefficients of these variables indicate that the RE of the low flow increased with basin storage. Larger groundwater contributions (BFI) were associated with more error. An increase in error also occurred for soils with more vertical drainage (Poorly_Drained) and flatter basins (Slope) with potentially more storage. The increase in RE with storage indicates that the low flows of basins connected to groundwater sources were more difficult to predict. More information on the subsurface drainage of these basins may improve the prediction of their low flows. Basins with more storage likely have larger low flows, and the dynamics affecting larger low flows, such as evaporative losses (Yokoo and Sivapalan, 2011), may have been a source of uncertainty.

6. US hydrologic regions

The regions derived from the hydrologic variables were adopted as the hydrologic regions of the US since they produced regional regression models with similar, if not better, performance than the other sets of variables for predicting the percentile flows. The US hydrologic regions were described to identify regional characteristics associated with the percentile flows. Regional characteristics were summarized using the mean z-score (Equation

1) of key independent variables (Figure 13). The combination of MAP and PET was used to describe the climate of the region, and other variables were used as an indicator of basin storage (Area, Elev, and BFI). Drainage area relates to storage, and elevation provides basic physiographic information that also often covaries with snowfall.

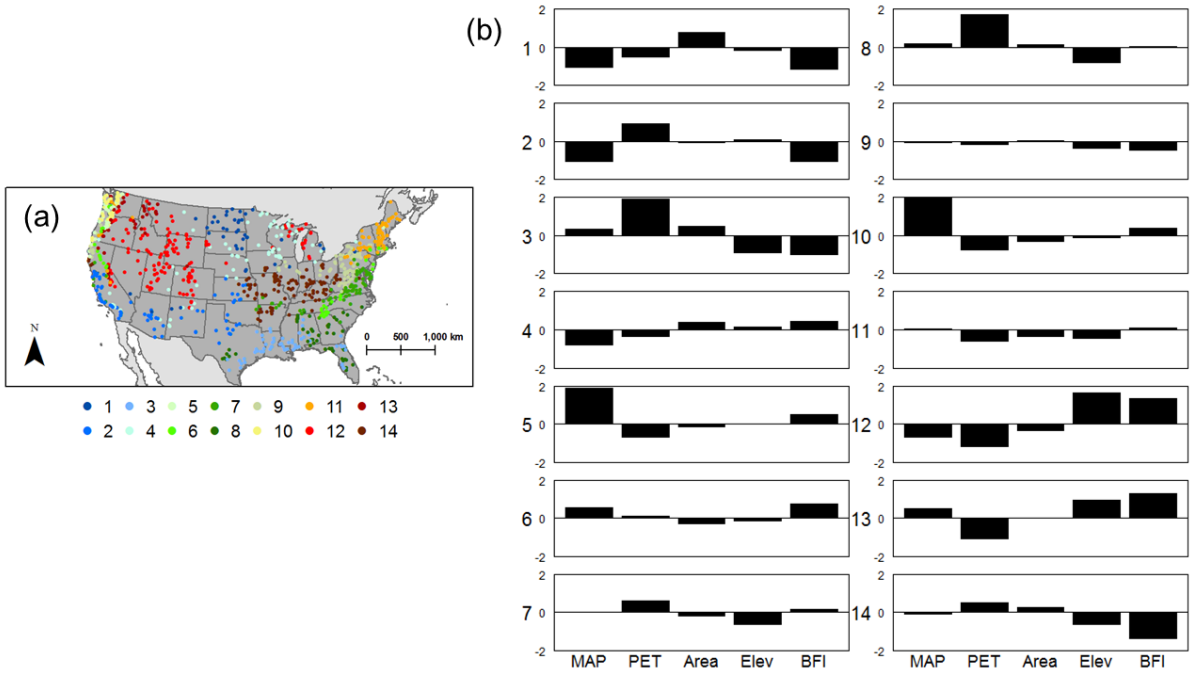

Figure 13. Regions derived from the hydrologic variables described by their (a) location and (b) mean z-score of key independent variables.

Descriptive labels were assigned to the regions based on their geographic location (Figure 13a) and key independent variables (Figure 13b), and the resulting classes are listed in Table 18. Geographic location was used to develop the classes since the hydrologic regions displayed geographic contiguity and spatial proximity was a first order indicator of basin similarity. The climate and storage of the regions were then qualitatively characterized using the key independent variables. The resulting classes are open to interpretation, but their value is briefly demonstrated by relating the characteristics from a sample of regions to the FDC. The same sample of regions from Table 13 (Regions 1, 2, and 11) are used to continue

the discussion of their characteristics and how they relate to the median FDC of the basins from each region (Figure 14).

Table 18. Descriptive classes assigned to the regions derived from the hydrologic variables. Classification developed based on geographic location and basin characteristics representing climate and storage.

Location	Climate	Storage	Description	Region
NW			Northwest	
	vw		Very wet	
		M	Moderate	5, 10
	w		Wet	
		S	Snow	13
RM			Rocky Mountains	
	c		Cold	
		S	Snow	12
MW			Midwest	
	sa		Semi-arid	
		LB	Large basins	1
	c		Cold	
		S	Snow	4
	t		Temperate	
		L	Low	14
NE			Northeast	
	t		Temperate	
		M	Moderate	7, 9, 11
	w		Wet	
		H	High	6
SE			Southeast	
	w		Wet	
		M	Moderate	8
		L	Low	3
SW			Southwest	
	a		Arid	
		L	Low	2

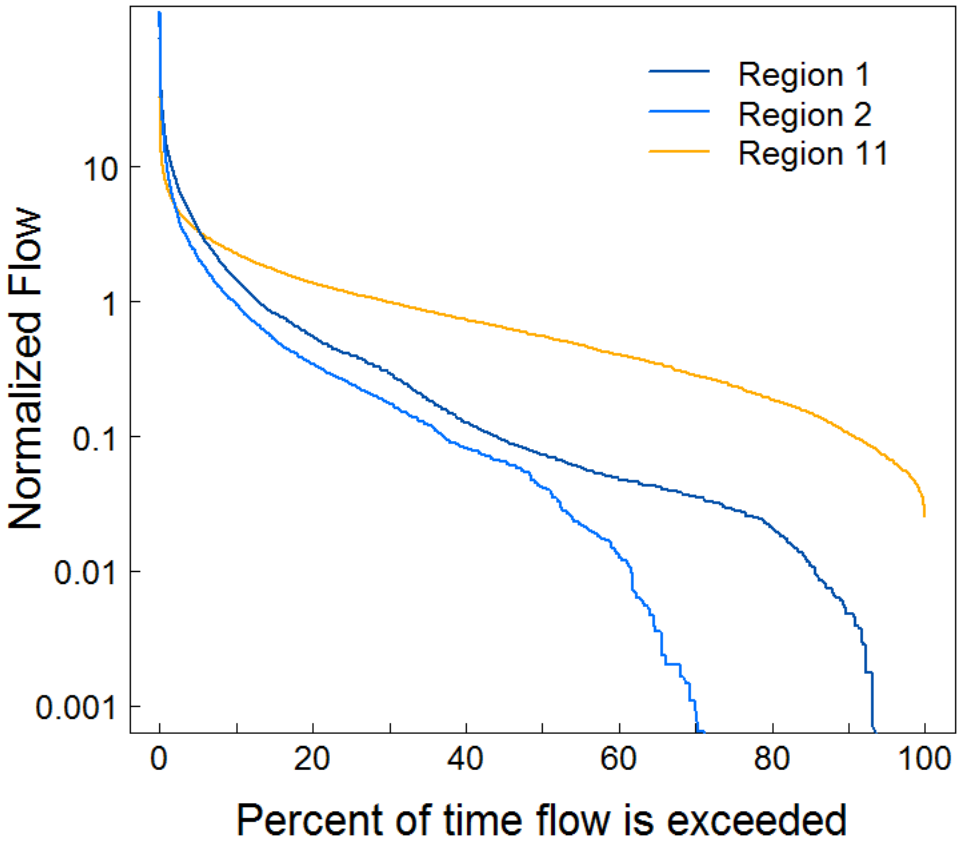

Figure 14. Median FDC of the basins from a sample of the hydrologic regions including regions 1, 2, and 11.

Region 1 was located in the Central Plains, and its semi-arid climate produced somewhat variable high flows and occasional periods of zero flow. The basins of the Central Plains were the largest of any region, and this added to their storage capacity and potential to sustain average flows. Region 2 occupied the Southwest, and was characterized by arid conditions with highly variable flows overall. The streams of the Southwest were disconnected from the water table, and received little flow from groundwater. These conditions produce streams that only flow in response to rainfall or during the wet season. Region 11 was mostly restricted to the New England portion of the Northeast, and its temperate climate produced perennial streams sustained by both rain and snowfall. High flows were the product of the summer storm season (Saco and Kumar, 2000), and average to

low flows are contributed by consistent year-round precipitation and local groundwater sources (Olcott, 1995).

F. Conclusions

A regional regression approach was applied on 918 basins in the US to predict 13 percentile flows of the FDC. The first phase of the approach split the basins into regions according to independent variables and a SOM-based cluster analysis to deal with noise and non-linearities in the data. The resulting regions were then used to develop regional regression models for predicting the percentile flows. The predictive performance of the regional regression models was assessed using 184 validation basins, and the entire regional regression approach was repeated using three different sets of independent variables to determine the necessary amount of information for predicting the percentile flows. The most efficient set of variables was then used to investigate the factors associated with predictive error and the regional conditions related to the percentile flows.

The regional regression approach achieved NSE values ranging from 0.39-0.76 and averaging 0.64. The predictive performance of the approach depended on the percentile flow and the set of independent variables used to formulate the regional regression models. The approach performed the best on the percentile flows in the middle of the FDC (average flows) and worst on percentile flows at the tails of the FDC (extreme flows). Average flows were modeled better than extreme flows likely because they have less variability (Salinas et al., 2013), and the extreme flows are controlled by processes, such as storm flows and subsurface drainage, that are more difficult to represent in the regression models. Additional variables may be needed to more closely represent the processes that control extreme flows,

such as the magnitude of extreme storms that produce high flows and the evaporative losses during dry periods that moderate low flows.

The performance of the regional regression approach was affected by the different sets of independent variables that were used to predict the percentile flows. The sets of independent variables represented the following three different amounts of information: (1) a simple set of hydrologic variables to represent the three components of the FDC, (2) a larger set of lumped variables typically used in regional regression studies to describe the average for a variety of basin characteristics, and (3) a more complex set of distributed variables including the lumped variables and additional variables to describe the statistical distribution of basin data. The distributed variables were used to test the hypothesis that variables describing the statistical distribution of basin conditions would improve percentile flow predictions by representing the variability of conditions that influence the FDC. However, the results do not support this hypothesis, and the distributed variables consistently produced the lowest performance for predicting the percentile flows. Variables describing the statistical distribution of basin conditions did not contribute useful information to the regional regression approach, and the simpler sets of variables were more effective for predicting the percentile flows.

The hydrologic variables consisted of only three variables to represent the dominant physical processes that control the FDC, and these variables (MAP, PET, and BFI) produced similar predictive performance to the more complex set of lumped variables. This means that additional variables typically used for regional regression added little information to the percentile flow predictions. The limited predictive potential of the typical variables was highlighted for the extreme flows, which were more effectively predicted using the three

process-oriented hydrologic variables. These results indicate that considerable time and effort could be saved by targeting independent variables to represent specific processes related to the FDC. The success of the hydrologic variables demonstrates the importance of using independent variables with a strong physical connection to the streamflow variable and the need to develop independent variables based on process understanding.

Results from the hydrologic variables were further investigated to understand the factors related to predictive error and regional conditions associated with the FDC. Predictive error increased with aridity and basin storage. This suggests that separate models should be developed to predict the percentile flows of arid basins and alternative variables may be needed to model the flows generated from basins with large storage components, such as snow or aquifers. The regions derived to predict the FDC showed strong geographic contiguity, and spatial proximity was a first order indicator of basin similarity. The regions distinguished different climatic and storage conditions, and these regional conditions were related to the FDC for a sample of regions.

The regional regression approach resulted in a range of predictive performance, and future research should target possible sources of uncertainty. The regions may have contained unacceptable levels of variability in the percentile flows, and regional homogeneity tests could be applied to identify problematic regions. This can be useful for adjusting the regions to make them more hydrologically homogeneous and improve subsequent predictions (Hosking and Wallis, 1997). The regions imposed discrete boundaries between the basins although their conditions varied along a continuum. A more appropriate way to identify the regions may be to give the basins partial membership in each region using fuzzy cluster analysis (Srinivas et al., 2008). Predictions could then be weighted according to the regional

membership of the basins, and this may improve predictions for basins that lie in the periphery of the regions. The regional regression models suffered from a lack of variables with predictive potential, and future work should focus on developing new variables for representing the processes that control the FDC. Finally, uncertainty in the regional regression approach may stem from the regression method used to generate the predictions, and a machine learning method, such as the SOM, may be better equipped to handle the noise and non-linearities in the data (Booker and Woods, 2014). The SOM is a particularly interesting option since its training routine clusters the data and may eliminate the need to identify hydrologic regions. This possibility is tested for the prediction of percentile flows in the final paper of this dissertation.

Chapter 4: Prediction and exploratory analysis of the flow duration curve using the self-organizing map

A. Abstract

Percentile flows of the flow duration curve (FDC) represent the flow magnitude exceeded for a given percent of time, and are widely used to manage water resources. These important statistics often need to be predicted for ungauged basins with insufficient streamflow data. A typical approach for predicting percentile flows uses independent variables consisting of physical and climatic characteristics to identify a priori regions and develop regional regression models. Identifying the regions can be a time-consuming process with uncertainties such as the appropriate clustering method and number of regions. A neural network approach, called the self-organizing map (SOM), is an alternative for clustering the basins and predicting percentile flows all in one step. It can also be used to inform future modeling efforts through an exploratory analysis of the factors related to the percentile flows. The SOM approach was used on 918 basins in the US for the prediction and exploratory analysis of 13 percentile flows. Global predictions using all the basins were generated using the SOM. A priori regions based on a cluster analysis of the independent variables for the subject basins were also used for SOM-based predictions to test the hypothesis that a priori regions do not improve predictions generated by the SOM. In addition, the predictive performance of the SOM was compared to a typical regional regression approach. Global and regional predictions of the SOM achieved similar performance, which confirms the hypothesis that a priori regions do not improve SOM predictions. Although the SOM did not benefit from the regions, it failed to outperform the regional regression. This may be because the regional regression used a subset of the independent variables based on their predictive

potential, and the SOM may have included irrelevant variables since it used all the independent variables. Future studies should pair the SOM with a variable selection method to discard irrelevant variables for predicting the percentile flows. The exploratory analysis using the SOM revealed that high flows were associated with the overall wetness of the basin and its snowfall possibly because of the spring snowmelt season or rain-on-snow events. Average and low flows were primarily associated with the groundwater contributions of baseflow. The overall relation between the percentile flows and independent variables was weak according to the discordancy between clusters derived from the two datasets, and future work should investigate new sets of variables to improve the connection to percentile flows. Such research would advance the identification of regions and independent variables for predicting percentile flows, and could be performed using the SOM.

B. Introduction

A widely used tool for representing streamflow data is the flow duration curve (FDC). This graphical representation of streamflow shows the flow magnitude equaled or exceeded for a given percent of time as percentile flows. These statistics are critical information for stream uses with flow requirements, such as hydropower, wasteload allocation, and habitat maintenance (Vogel and Fennessey, 1995). Percentile flows are readily calculated using sufficiently long streamflow records, but they must be predicted for most locations with insufficient or no streamflow data. This falls under the Predictions in Ungauged Basins problem (Sivapalan et al., 2003), and is addressed using information from gauged basins to predict percentile flows for ungauged basins.

Information from gauged basins is either used to establish functional relations between measurable basin characteristics and percentile flows or estimate the parameters of a

rainfall-runoff model. The latter approach is subject to both parameter and model uncertainty, whereas empirical approaches that relate basin characteristics to percentile flows require less effort and have achieved similar performance in comparative studies. (Booker and Woods, 2014; Müller and Thompson, 2015; Zhang et al., 2014). The basin characteristics used to predict percentile flows can include spatial proximity, but this may provide misleading information in drier climates with larger fluctuations in flow between neighboring basins (Patil and Stieglitz, 2012). Physical and climatic basin characteristics are preferred since they are related to the hydrologic processes that control percentile flows and may produce robust predictive models for a variety of environments (Sivapalan, 2005). Predictive models based on physical and climatic characteristics are often used to predict the parameters of statistical distributions for representing the FDC, but a suitable statistical distribution can vary depending on regional conditions (Castellarin et al., 2004). Predicting individual percentile flows is advantageous because assumptions on the statistical distribution of the FDC are not necessary.

A growing number of studies are using physical and climatic characteristics as independent variables to predict percentile flows (Hashmi and Shamseldin, 2014; Hope and Bart, 2012; Mohamoud, 2008). These studies have focused on particular geographic regions, such as the Auckland Region of New Zealand (Hashmi and Shamseldin, 2014), Cape Floristic Region of South Africa (Hope and Bart, 2012), and Mid-Atlantic Region of the US (Mohamoud, 2008). The success of these studies has varied possibly due to regional conditions affecting the heterogeneity of percentile flows and ability to model their relation to independent variables. Regional heterogeneity of percentile flows can diminish predictive performance. In order to control for such heterogeneity, the basins are often assigned to

regions (or groups) based on independent variables related to flow (Olden et al., 2012). Identifying a priori regions via independent variables is a longstanding research theme in flood prediction (Acreman and Sinclair, 1986), and this approach has recently been employed to predict the FDC (Boscarello et al., 2015). The first two papers of this dissertation demonstrated the importance of identifying a priori regions for predicting percentile flows. Global predictions based on all the basins were improved using a priori regions identified using cluster analysis. However, the use of cluster analysis is an additional step that involves decisions which may impact the predictive performance of subsequent models developed for each region (i.e. regional models).

A neural network approach, called the self-organizing map (SOM), is an alternative to identifying a priori regions for regional models. The SOM clusters features, such as basins, according to their attributes in a grid of neurons that can be used to generate predictions. This is an appealing alternative to identifying a priori regions because the SOM can cluster basins and generate percentile flow predictions in one step. Identifying a priori regions requires more effort and decisions that may influence predictive performance. The choice of a clustering method to identify the regions can affect the predictive performance of the regional models. This has been documented by studies that have compared the predictive performance of regional models developed using different clustering methods (see Boscarello et al., 2015; Di Prinzio et al., 2011; Sauquet and Catalogne, 2011).

Another decision that can affect predictive performance is the number of clusters (or regions) for the dataset. An optimal number of clusters can be identified using cluster validity indices, but these metrics can give conflicting results that either indicate a small or large number of clusters (Shim et al., 2005). This problem was observed using a variety of cluster

validity indices in the second paper of this dissertation, and an alternative strategy had to be developed to determine the number of clusters. Resulting clusters impose borders on continuously varying fields of data. This can be problematic for basins located along cluster borders with weak associations to individual clusters. The SOM avoids this issue by mapping the basins in a grid of neurons that represents the continuous variation in the data.

The neurons of the SOM are linked to the data through a vector of values equal in length to the number of input variables. These "neuron vectors" are incrementally adjusted according to the data during an iterative training process. Input data from a basin is presented to the SOM, and assigned to the most similar neuron, or best-matching unit (BMU). The BMU and its neighbors are then adjusted to more closely match the incoming input data. The resulting neurons can be thought of as representative samples of the basins (Kalteh et al., 2008), and the neuron vectors can serve as predictions if they include the percentile flows.

Application of the SOM for ungauged prediction has been an active area of research for nearly two decades (Hall and Minns, 1999), but using the SOM as a predictor for streamflow variables, like percentile flows, would be a new area of research. In the past, the SOM has been primarily used as a preliminary step towards identifying a priori regions (see the second paper of this dissertation and Boscarello et al., 2015; Di Prinzio et al., 2011; Hall and Minns, 1999). The purpose of using the SOM in this capacity is to organize the data for subsequent cluster analysis and account for non-linearities in the data. Output from the SOM has also been used for the exploratory analysis of the controls on streamflow to advance modeling efforts (see Farsadnia et al., 2014; Ley et al., 2011; Toth, 2012). These studies exploit the SOM's ability to visualize the structure of the data and connections between variables. The SOM has not been used extensively as a predictive method in hydrology.

These applications have been limited to estimating the parameters for a rainfall-runoff model (Wallner et al., 2013) and infilling missing data in hydroclimatic time series (see Kalteh and Berndtson, 2007; Mwale et al., 2012; Rustum and Adeloye, 2007). Applying the SOM for prediction and an exploratory analysis of the results would serve to illuminate future modeling needs for better predicting streamflow variables, such as percentile flows.

The SOM is a particularly suited approach to predict percentile flows for several reasons. Hydrologic data used to predict percentile flows are prone to noise (i.e. variation unrelated to the observed phenomenon). Streamflow data contains error due to gauging malfunctions and uncertainty in the rating curve used to estimate discharge (McMillan et al., 2012). Environmental data, like precipitation, have error introduced by spatial and temporal interpolation techniques (Daly et al., 2008). Noise in streamflow and environmental data may degrade the relations developed to predict percentile flows, but the SOM is resilient to reasonable levels of noise because the neuron vectors are computed using local neighborhoods less influenced by random variations in the data (Vesanto and Alhoniemi, 2000). Another reason why the SOM may be suited to predict percentile flows is that it is capable of modeling both linear and non-linear associations between variables (Kohonen, 1998). The relation between basin characteristics and percentile flows can take on a variety of functional forms, and the SOM may be more flexible than traditional statistical methods, like multivariate regression, since it adapts to the data based on an iterative training process (Kohonen, 2001). This is a common feature of artificial neural networks, but the output layer of the SOM allows the user to explore associations in the data that may otherwise be hidden in artificial neural networks (Kalteh et al., 2008). Associations to the percentile flows (or lack thereof) may be used to develop better predictive models.

The SOM's output layer consists of neuron vectors that reflect patterns in the input data. The neuron vectors can be used to conduct an exploratory analysis of the factors related to percentile flows, which could provide information for the future evolution of percentile flow modeling. The number of neurons for the exploratory analysis should be a fairly large fraction of the basins in order to use the SOM as a "spatial layout" tool for displaying the distribution of the data on a continuous surface (Skupin and Esperbé, 2011). Subsequent data visualizations are then capable of showing the cluster structure and connection between variables in the output layer of the SOM. The connection between percentile flows and independent variables is a critical aspect of identifying regions for ungauged predictions, and percentile flow predictions hinge on a strong connection to the independent variables. Data visualizations of the SOM can be used to examine the connection between percentile flows and independent variables. This can be accomplished by displaying how the data is ordered in the SOM. Data ordered in a similar fashion are related. This style of data visualization may reveal discordancy between the percentile flows and independent variables that can be interpreted to recommend future modeling directions.

The objective of this research was to predict percentile flows and evaluate their relations to independent variables using the SOM. This study highlights the SOM because of its (1) tolerance to noise (Vesanto and Alhoniemi, 2000), (2) modeling flexibility (Kohonen, 1998), and (3) exploratory component (Skupin and Esperbé, 2011). Another advantage of the SOM is that its clustered output may be a substitute for identifying a priori regions. The SOM was applied with and without a priori regions to test the hypothesis that a priori regions would not be needed to improve SOM predictions. Predictive performance of the SOM was compared to a typical regional approach that developed regression models for the individual

regions. This served as a baseline reference for assessing the applicability of the SOM for predicting percentile flows in ungauged basins. The study included three separate sets of percentile flow predictions produced using (1) the SOM without a priori regions, (2) the SOM with a priori regions, and (3) regional regression. The performance of these predictions was compared to test the hypothesized advantage of the SOM (i.e. a prior regions are not necessary) and its performance relative to a typical approach (i.e. regional regression). An exploratory analysis of the SOM predictions was conducted to reveal potential areas of improvement for modeling percentile flows. Results from the SOM were evaluated to answer the following two research questions:

How do percentile flow predictions generated using the SOM compare to regional predictions, and what potential improvements to percentile flow modeling can be identified using the SOM?

C. Methods

1. Overview

The SOM was used to predict 13 percentile flows ranging from a high flow exceeded 1% of the time (Q_{01}) to a low flow exceeded 99% of the time (Q_{99}) and 11 flows between that range (Q_{05}, Q_{10}, Q_{20},…Q_{95}). The percentile flows and a set of independent variables were used to train the SOM, and the output neuron vectors served as the percentile flow predictions. The performance of the predictions was assessed using validation basins treated as ungauged, and their percentile flows were withheld from any step to generate predictions using the SOM. Validation basins were assigned to SOM neurons according to their independent variables. The neuron with the most similar output vector to the independent

variables was designated the BMU, and its output vector was used to predict the percentile flows for the given validation basin.

A variety of approaches for training the SOM were tested since several options exist for generating ungauged predictions and these options may affect the predictive performance of the SOM. The SOM can be trained with missing data, and this meant that the SOM could be trained with the validation basins missing percentile flow data. This type of training originates from studies on infilling missing data (see Mwale et al., 2012 as an example). All the data was used to train the SOM including the missing percentile flow data from the validation basins. The SOM was also trained without the validation basins as in traditional neural network training (see Ssegane et al., 2012b as an example), and missing data was excluded from the SOM. Both types of training (with and without missing data) were used to generate SOMs including all 13 and individual percentile flows. This step was performed to determine if a single SOM could be used to predict all 13 percentile flows at once or individual SOMs had to be trained for each percentile flow.

The training approach with the best results was then applied in regions previously identified in the second paper of this dissertation. Global predictions based on all the basins and regional predictions were produced using the SOM to compare its performance with and without a priori regions. The utility of the SOM for predicting percentile flows was further evaluated relative to a typical regional regression from the second paper. Finally, the SOM was used to create data visualizations for an exploratory analysis of the percentile flows to illuminate potential modeling improvements. The methods of this study are summarized as a flow chart in Figure 15.

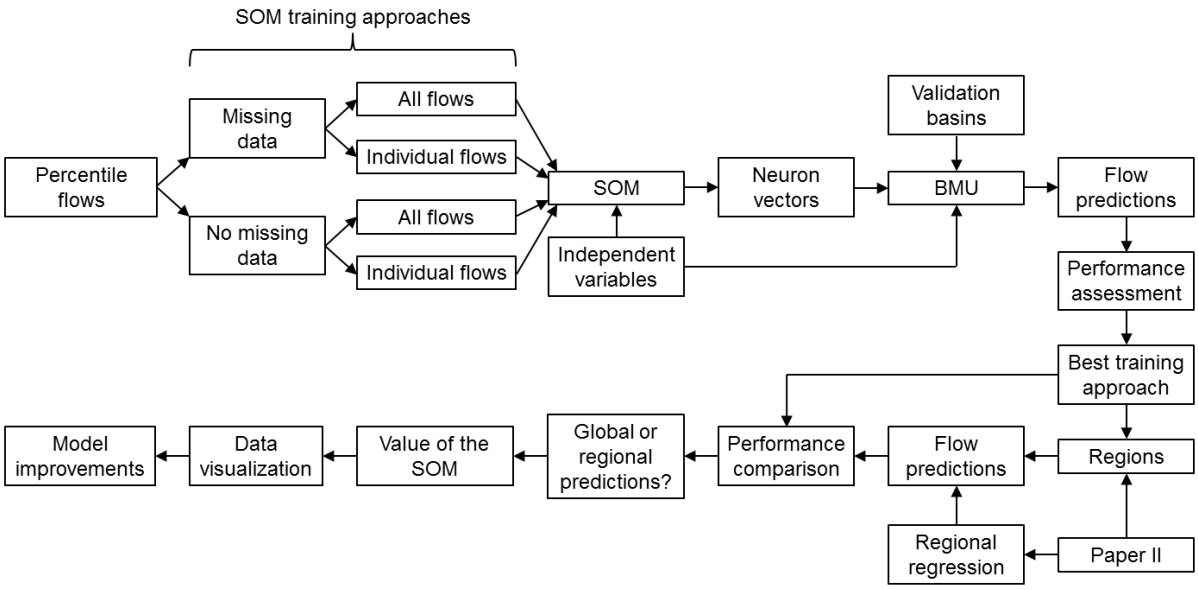

Figure 15. Flow chart of the methods for testing the SOM to predict percentile flows and improve future models.

2. Training data

The data used to train the SOM consisted of 13 percentile flows and 22 independent variables. These values were calculated for 918 basins in the contiguous US classified as "near-natural" (Falcone, 2011) and with at least 30 years of continuous daily streamflow data. The length of streamflow data was chosen to reliably calculate percentile flows for different time periods (Kennard et al., 2010). Normalized percentile flows were calculated using the Weibull plotting position and the mean of nonzero flows to control for differences in drainage area (Castellarin et al., 2004). The natural log transformation was applied on the percentile flows to minimize the potential influence of outliers (Allende et al., 2004). Percentile flow data was excluded from 184 validation basins (20% of the basins) to assess the performance of the SOM predictions. The validation basins were a representative sample of the basins selected according to climate class, rock type, and drainage area. Key hydrologic factors were used instead of the percentile flows to maintain the independence of

the validation. A map of the validation basins along with the other basins used in this study is provided in Figure 16.

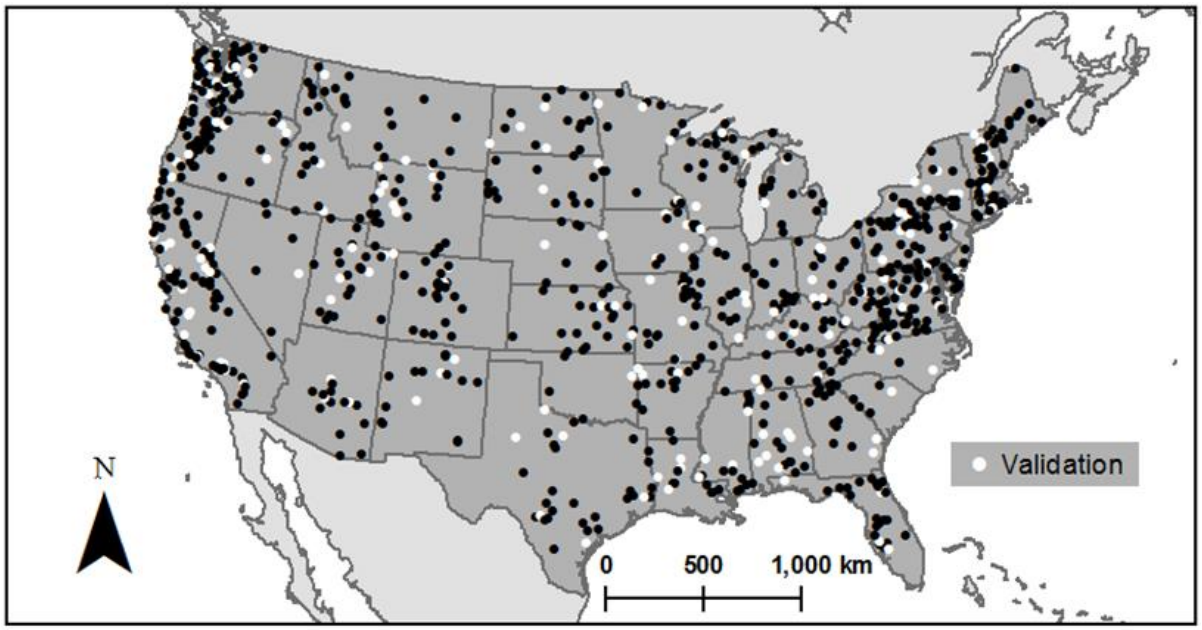

Figure 16. Map of the 918 basins used in this study, with the 184 validation basins highlighted in white.

The independent variables were used to indicate hydrologic similarity and assign the validation basins to SOM neurons. A typical set of independent variables was used to characterize the climate, topography, land cover, soil, and geology of the basins. Climatic variables were calculated using 30 years of data to effectively represent long-term conditions (Arguez et al., 2012). The only land cover variable was percent forest because the percent cover of the different land cover classes was related and differences in forest cover are strongly tied to the FDC (Brown et al., 2013). Geology was represented using a preexisting baseflow index (BFI) grid for the US expressing the percent of streamflow contributed by groundwater. This variable quantifies the effect of geology on streamflow and is very useful for predicting the FDC (see previous papers of this dissertation and Yokoo and Sivapalan, 2011). The rest of the independent variables are described in Table 19.

Table 19. Percentile flows and independent variables used to train the SOM.

Variable	Units	Description	Key reference	Data source
Percentile flows				
Q_p (e.g. Q_{01} for 1%)	-	Normalized percentile flows for 1, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 95, and 99%	Castellarin et al. (2004)	NWIS
Independent				
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
Spring_Temp	°C	Average temperature from April-June	Boscarello et al. (2015)	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. (2012b)	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. (2012b)	NHDPlusV2, GAGES-II
Orientation	°N	Basin angle along main channel	Di Prinzio et al. (2011)	GAGES-II
Elev	m	Mean elevation	Ssegane et al. (2012b)	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. (2012b)	NED
Aspect	°N	Mean aspect	Ssegane et al. (2012b)	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. (2012b)	NED
Land cover				

Table continued on next page

Variable	Units	Description	Key reference	Data source
Forest	%	Percent forest cover	Ssegane et al. (2012b)	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. (2012b)	CONUS-SOIL
Geology				
BFI	%	Mean baseflow index derived from a baseflow grid	Hope and Bart (2011)	BFI48GRD

Data sources: NWIS, National Water Information System (http://waterdata.usgs.gov/nwis); PRISM, Precipitation-elevation Regressions on Independent Slopes Model (http://prism.oregonstate.edu); GAGES-II, Geospatial Attributes of Gages for Evaluating Streamflow, version II (Falcone, 2011); NHDPlusV2, National Hydrography Dataset Plus Version 2 (http://www.nhdplus.com); NED, National Elevation Dataset (http://ned.usgs.gov); NLCD 1992, National Land Cover Dataset 1992 (Vogelmann et al., 2001); CONUS-SOIL, Conterminous US multilayer soil characteristics dataset (Miller and White, 1998); BFI48GRD, Base-flow index grid for the conterminous US (Wolock, 2003)

3. SOM training and predictions

The SOM was trained using the percentile flow and independent variable data. Prior to training the SOM, the data underwent z-score normalization in order to equally weight the input variables. This produced variables with a mean of zero and variance of one. The normalized variables were then used to train the SOM, which consisted of a set of output neurons arranged in a two-dimensional grid. The output neurons each had a vector equal in length to the number of input variables. The neuron vectors were first given random values, and then adjusted to more closely match the input data through an iterative training process. The input data was iteratively presented to the SOM, and assigned to the most similar neuron, or BMU, according to Euclidean distance. The neuron vectors of the BMU and its neighbors were adjusted to be more similar to the input data. This process was controlled by the learning rate and neighborhood function. The learning rate altered the magnitude of the

neuron vector adjustments, and decreased monotonically for each iteration of the training. The neighborhood function used the typical Gaussian equation to moderate the adjustments to neurons neighboring the BMU and decrease the radius of the neighborhood throughout training. Equations for the above SOM training process are supplied in the second paper of this dissertation.

As recommended by the creator of the SOM (Kohonen et al., 1996), training was conducted in two stages to first capture global structures in the data and then refine those structures with a local training stage. The global training used a larger learning rate (0.04) and neighborhood (half of the SOM) to make broad-scale adjustments to the SOM, while the local training was accomplished using a smaller learning rate (0.03) and neighborhood (one third of the SOM).

The only remaining parameters for training the SOM were the number of neurons and iterations. Both of these parameters were set relative to the number of input vectors (i.e. basins), and this was initially established using all the basins in experiments from the second paper of this dissertation. The number of neurons was determined by testing different sized SOMs and evaluating the number of "empty" neurons that did not serve as a BMU for any of the basins. The final SOM size of 15 × 15 neurons was selected to limit the number of empty neurons that were not linked to the data. The number of iterations was then selected to adequately train the neurons. This was determined based on the quantization error, which is a measure of how well the neurons match the data. The SOM with 15 × 15 neurons required 50 iterations for the global training and 4,000 iterations for the local training. The size of the SOM and number of training iterations established using all the basins were proportionally reduced to train the SOM using a subset of the basins.

Several approaches were used to train the SOM because the options for generating ungauged predictions may affect the SOM's predictive performance. The SOM was trained including and excluding the validation basins with missing percentile flow data. The approach that included the validation basins was akin to previous studies that have used the SOM to infill missing data (see Mwale et al., 2012). The SOM was also trained without the validation basins as in traditional neural network (NN) training (see Ssegane et al., 2012b). Both of the above training approaches were repeated to produce SOMs including all the percentile flows and individual percentile flows, and this resulted in four different training approaches for predicting percentile flows (Table 20). Each training approach used the same method for generating percentile flow predictions. Validation basins were assigned to the neurons using the independent variables. The BMU was identified based on Euclidean distance, and the output neuron vector of the BMU was used to predict the percentile flows.

Table 20. SOM training approaches used to predict percentile flows in this study.

SOM training approach	Validation basins	Percentile flows
Infill_All	Included	All
Infill_Individual	Included	Individual
NN_All	Excluded	All
NN_Individual	Excluded	Individual

4. SOM performance assessment

The performance of the SOM was assessed for predicting 13 percentile flows. Predictive performance was evaluated using 184 validation basins that were treated as ungauged. The difference between predicted and observed percentile flows was summarized using the sum of absolute relative error (RE), coefficient of determination (R^2), and Nash-Sutcliffe efficiency (NSE). These metrics were selected because they are widely used to assess the performance of percentile flow predictions (see Mendicino and Senatore, 2013 for

mathematical definitions). A normalized measure of error is provided by RE, which can be summed to assess the overall error of the predictions. The amount of variance in the observations explained by the predictions is given by R^2, with larger values signifying more accurate predictions. Like R^2, larger NSE values signify better predictive performance. The value of NSE also indicates if the predictions performed better than simply using the mean of observed values (NSE > 0). The performance assessment was conducted using the natural log of percentile flows to reduce the influence of outliers (Di Prinzio et al., 2011). Results from the performance assessment were compared for the four different approaches used to train the SOM, and the preferred approach was adopted to represent the global predictions from the SOM. Regional predictions using the SOM were generated using the preferred training approach and the regions from the second paper of this dissertation. Results from the regional regression of the second paper were also included as a reference for assessing the predictive performance of the SOM. Global predictions from the SOM were compared to the regional predictions to determine if the SOM could be used as an alternative to regional methods for predicting percentile flows.

5. Exploratory analysis using SOM data visualizations

An exploratory analysis of the percentile flows was conducted using data visualizations based on the SOM. The goal of the visualizations was to identify potential improvements for future predictive models by assessing the relation between the percentile flows and independent variables. The two sets of data were compared using the trained neuron vectors of the SOM. The neuron vectors were split in order to compare the percentile flows to the independent variables. Correspondence between the datasets was evaluated using the following visualization methods:

1. Pie charts were created to show the number of basins assigned to each neuron according to the Euclidean distance between the different input data and the neuron vectors.
2. The unified distance matrix (U-matrix) was used to visualize the cluster structure of the SOM by calculating the Euclidean distance between neighboring neurons. This calculation created a new representation of the SOM with map units equal to the distance between pairs of neighboring neurons and the average distance between all the neighboring neurons. Small distances between neurons represent clusters in the data, and large distances between neurons signify cluster borders.
3. Component planes were generated to map the value of individual variables in the SOM, and variables with similar component planes are related.
4. Independent variables related to the percentile flows were selected based on the component planes. The relative values of the selected independent variables were displayed as pie charts for each neuron, and overlaid on the component planes of individual percentile flows to show their relation to the independent variables.
5. Finally, the neuron vectors were clustered using the *k*-means method as described in Isik and Singh (2008). A comparison of the two cluster solutions was performed to assess the relation between the percentile flows and independent variables. The number of clusters (*k*) was determined using the "elbow" method. The sum of squared error (SSE) was plotted for consecutive clustering solutions with up to 50 clusters, and the point at which the decrease in SSE flattened out was chosen as the appropriate number of clusters. This approach was applied to cluster the neuron vectors for both the percentile flows and independent variables, and the resulting

clusters were mapped on the SOM. Neuron clusters were assigned to the basins according to their BMU. Basins were represented geographically using the Thiessen polygons of their gauge locations, and the clusters based on the SOM were mapped for the US. Cluster borders were mapped for the SOM and US to show the correspondence (or lack thereof) between the percentile flows and independent variables.

D. Results and discussion

1. Predictive performance of the different SOM training approaches

The performance of four different approaches for training the SOM to generate global percentile flow predictions was summarized for each percentile flow using the sum of absolute RE, R^2, and NSE (Table 21). The different training approaches affected the predictive performance of the SOM. Training approaches that included the validation basins with missing percentile flow values (Infill_All and Infill_Individual) consistently performed better than traditional NN training that excluded the validation basins (NN_All and NN_Individual). The data infilling approaches included the independent variables of the validation basins, and this may have improved how the validation basins were assigned to the SOM neurons. Training the SOM for individual percentile flows (Infill_Individual and NN_Individual) was not a clear advantage for generating predictions using the SOM. This may be the case because SOM training adjusts the values of the neuron vectors individually, and including all the percentile flows did not have a major effect on the neuron vectors. Based on these results, the preferred approach for training the SOM included the validation basins and all the percentile flows (Infill_All). This conclusion was confirmed by two of the

three metrics used to summarize the overall performance of the training approaches for all the percentile flows (Table 22).

Table 21. Predictive performance of the different SOM training approaches summarized for each percentile flow using (a) the sum of absolute RE, (b) R^2, and (c) NSE. Bold numbers indicate the SOM training approach that performed the best for each percentile flow according to the given performance metric.

(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}
Infill_All	**12.4**	**10.6**	**9.81**	**9.64**	**10.4**	**11.2**	**11.7**	**11.9**	**11.9**	**11.6**	10.5	9.49	8.06
Infill_ Individual	12.8	12.4	10.4	10.6	11.5	12.2	12.2	12.4	12.0	11.9	**10.1**	**9.44**	**7.93**
NN_All	13.0	11.5	10.1	10.3	11.5	12.4	12.9	13.1	12.9	12.3	11.1	10.1	8.52
NN_Individual	12.7	11.7	10.9	10.8	12.6	12.2	12.8	12.6	12.8	12.0	11.1	9.75	8.13

(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}
Infill_All	**0.44**	**0.50**	**0.58**	0.57	**0.58**	**0.55**	0.51	**0.46**	0.42	**0.39**	0.38	**0.38**	0.36
Infill_Individual	0.41	0.37	0.57	**0.59**	0.55	0.54	**0.53**	0.45	**0.50**	0.37	**0.41**	0.38	0.43
NN_All	0.38	0.41	0.55	0.54	0.52	0.49	0.44	0.40	0.37	0.35	0.34	0.34	0.32
NN_Individual	0.36	0.42	0.50	0.55	0.48	0.54	0.46	0.42	0.36	0.37	0.35	0.38	**0.45**

(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}
Infill_All	**0.42**	**0.49**	0.57	0.55	**0.56**	0.54	0.48	0.42	0.38	0.36	0.35	0.36	0.34
Infill_Individual	0.41	0.37	**0.57**	**0.59**	0.55	0.54	**0.53**	**0.44**	**0.49**	0.36	**0.41**	**0.37**	0.42
NN_All	0.36	0.41	0.54	0.53	0.50	0.46	0.41	0.37	0.34	0.32	0.30	0.31	0.30
NN_Individual	0.36	0.42	0.50	0.54	0.48	**0.54**	0.46	0.42	0.35	**0.37**	0.35	0.37	**0.43**

Table 22. Overall performance of the different SOM training approaches summarized as the sum of absolute RE and average R^2 and NSE for all the percentile flows. Bold numbers indicate the SOM training approach that performed the best overall according to the given performance metric.

	RE	R^2	NSE
Infill_All	**139**	**0.47**	0.45
Infill_Individual	146	0.47	**0.47**
NN_All	150	0.42	0.40
NN_Individual	150	0.43	0.43

The predictive performance of the SOM varied for the percentile flows (Table 21). The sum of absolute RE was influenced by the magnitude of the flow (larger error for higher

flows and smaller error for lower flows), but the other performance metrics indicated the relative performance of the SOM for the various percentile flows. The values of R^2 and NSE ranged from 0.32-0.59 and 0.30-0.59, respectively, and indicated a similar pattern of performance for predicting the percentile flows. Predictive performance plateaued for the percentile flows from Q_{10}-Q_{40}, and decreased for higher and lower flows. This has been a typical outcome in the other papers of this dissertation and prior studies on predicting percentile flows (Archfield et al., 2009; Hashmi and Shamseldin, 2014; Ssegane et al., 2012b). These studies have used a variety of methods to predict percentile flows, such as multivariate regression, symbolic regression, and neural networks, and this indicates that the independent variables used for these methods were not sufficiently representing the processes that control the higher and lower flows. The input variables were one source of uncertainty for the SOM predictions. Additional sources of uncertainty were (1) the parameters used to train the SOM, such as the number of neurons, learning rate, and the neuron neighborhood settings, (2) assigning the validation basins to the neurons of the SOM, and (3) the smoothing of the neuron vectors oversimplified the variability of the percentile flows.

2. Global versus regional percentile flow predictions

Global percentile flow predictions using all the basins were compared to regional predictions to determine if the SOM could be used to forego the process of identifying regions. The preferred SOM training approach from the comparison of different approaches (Global_SOM) was applied in previously identified regions from the second paper of this dissertation (Regional_SOM). The performance of the global and regional SOM predictions was assessed using the sum of absolute RE, R^2, and NSE for each percentile flow (Table 23). These metrics were also summarized for all the percentile flows to compare overall

performance of the global and regional SOM predictions (Table 24). The performance of the global and regional SOM predictions was similar for each percentile flow and overall. This result confirms the hypothesis of this study that a priori regions would not be needed to improve SOM predictions. The a priori regions were not needed presumably because of the clustering that occurs during SOM training.

Table 23. Performance of the global (Global_SOM) and regional predictions using the SOM (Regional_SOM) and regression (Regional_Reg) summarized for each percentile flow using (a) the sum of absolute RE, (b) R^2, and (c) NSE. Bold numbers indicate the global or regional method that produced the best predictions for each percentile flow according to the given performance metric.

(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}
Global_SOM	12.4	10.6	9.81	9.64	10.4	11.2	11.7	11.9	11.9	11.6	10.5	9.49	8.06
Regional_ SOM	12.7	11.0	9.77	9.61	10.5	11.3	11.9	12.1	12.0	11.3	10.4	9.45	8.00
Regional_Reg	**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**

(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}
Global_SOM	0.44	0.50	0.58	0.57	0.58	0.55	0.51	0.46	0.42	0.39	0.38	0.38	0.36
Regional_SOM	0.34	0.42	0.57	0.62	0.61	0.57	0.52	0.46	0.41	0.38	0.35	0.33	0.31
Regional_Reg	**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**

(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}
Global_SOM	0.42	0.49	0.57	0.55	0.56	0.54	0.48	0.42	0.38	0.36	0.35	0.36	0.34
Regional_SOM	0.32	0.41	0.56	0.62	0.61	0.57	0.51	0.45	0.40	0.36	0.33	0.32	0.30
Regional_Reg	**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**

Table 24. Overall performance of the global (Global_SOM) and regional predictions using the SOM (Regional_SOM) and regression (Regional_Reg) summarized as the sum of absolute RE and average R^2 and NSE for all the percentile flows. Bold numbers indicate the global or regional method that produced the best overall predictions according to the given performance metric.

	RE	R^2	NSE
Global_SOM	139	0.47	0.45
Regional_SOM	140	0.45	0.44
Regional_Reg	**115**	**0.65**	**0.65**

The predictive performance of the SOM was compared to a typical regional regression from the second paper of this dissertation (Regional_Reg). This served as a reference for assessing the SOM's predictive performance. The SOM did not perform as well as the regional regression for every percentile flow according to each performance metric of Table 23, and this was reflected in the overall performance of the predictions summarized for all the percentile flows (Table 24). An advantage of the SOM is that it adapts to the input data. However, this can be a detriment if the SOM is over-fit to the input data and possibly less transferable to data excluded from the training (i.e. ungauged basins). This may be the result of a limited training sample that inadequately represents the entire population of the data. For this research, more basins may have been needed to capture the full range of hydrologic conditions in the US. This may have produced a more robust SOM for a wider variety of ungauged basins.

The SOM may not have performed as well because it included all of the independent variables, whereas the regional regression applied a variable selection method to discard irrelevant variables unrelated to the percentile flows. These variables may have diminished the predictive performance of the SOM, and applying a variable selection method to discard irrelevant variables may improve SOM predictions.

The SOM may be a viable alternative to regional predictions based on a priori regions provided that the independent variables used to train the SOM are associated with the percentile flows. This could be accomplished by paring the SOM with a variable selection method for identifying a relevant subset of independent variables to predict the percentile flows. Input variable selection has improved the predictions of neural networks for streamflow forecasting by enhancing the connection between model inputs and observed

flows (Bowden et al., 2005), and may be similarly needed to improve SOM percentile flow predictions for ungauged basins. The SOM produces unsupervised predictions that are not adjusted to minimize the error of an objective function. Supervised predictors that minimize error while clustering the data may perform better than regional predictions. The potential of supervised predictors that cluster the data, such as random forests, has been demonstrated in a pair of studies from New Zealand (Booker and Snelder, 2012; Booker and Woods, 2014), and such methods may be a substitute for identifying a priori regions to predict percentile flows in the US.

The independent variables used in this study were not suited to predict the high and low percentile flows as indicated by the poorer performance of both the global and regional predictions (Table 23). Extreme flows are notably difficult to predict due to large variability between basins (Salinas et al., 2013), and independent variables were needed to better represent the processes that control the high and low flows. The high flows are essentially flood events driven by heavy storms and antecedent moisture conditions (Cheng et al., 2012). High flow predictions may be improved by additional variables on the magnitude of given precipitation percentiles (Ssegane et al., 2012b) and average soil moisture conditions related to the runoff generated by storms (Brown et al., 2013). Low flows can be approximated using baseflow adjusted for evaporative losses (Yokoo and Sivapalan, 2011). This study included baseflow as an independent variable, but additional information on subsurface drainage, such as a hydrogeologic classification (Tague and Grant, 2004) or the hydraulic characteristics of underlying aquifers (Smakhtin, 2001), may have improved low flow predictions. Evaporative losses to low flows may have been better represented using variables focused on inter-storm periods, such as the number or PET of dry days (Carrillo et al., 2011). The independent

variables of this study largely neglected the effect of vegetation on flow, but remotely sensed vegetation indices related to transpiration may improve percentile flow predictions (Troch et al., 2009).

3. SOM data visualizations of percentile flows versus independent variables

The SOM was used to create data visualizations for an exploratory analysis of the relations between the percentile flows and independent variables. The correspondence between the two datasets was examined to reveal potential improvements for future predictive models. The SOM trained without the validation basins was used for this analysis because the missing percentile flow data from the validation basins appeared in subsequent visualizations and detracted from the information that could be extracted from the SOM. The first visualization illustrates the number of basins assigned to each SOM neuron according to the two different sets of data for the percentile flows and independent variables (Figure 17). A backdrop of the U-matrix calculated using all the input data is provided to show the clusters of the SOM (blue colors with low dissimilarity) where there should be more basins assigned to the neurons. The visualization confirmed that there were more basins located in the clusters of the SOM, but the basins also accumulated along the edges of the SOM. This is a common problem known as the "edge effect" (Schmidt et al., 2011), and occurs because the neurons at the edge of the SOM have fewer neighbors. Basins located at the edge of the SOM therefore had fewer neurons to which they could be reassigned, and were less likely to be relocated away from the edge.

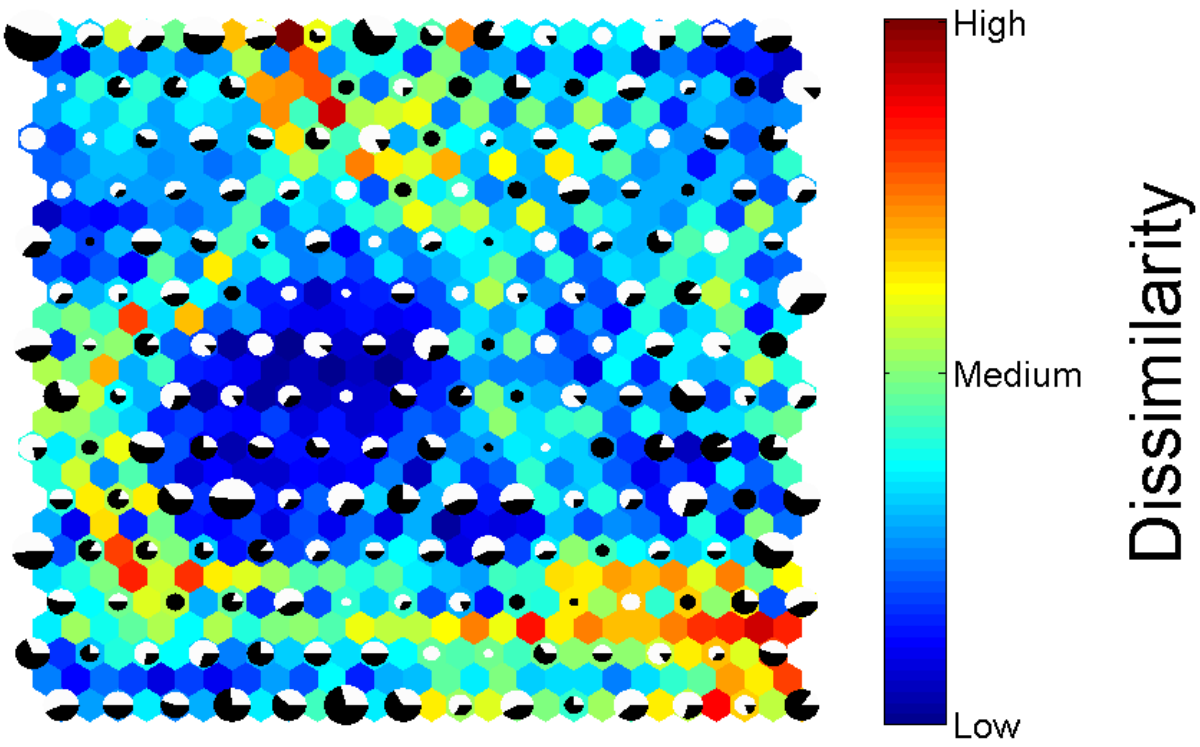

Figure 17. U-matrix of all the input variables with pie charts showing the number of basins assigned to each neuron according to the percentile flows (black) and independent variables (white). The size of the pie charts indicates the total number of basins assigned to each neuron based on both sets of input variables.

The pie charts of Figure 17 show the relative number of basins assigned to each SOM neuron based on the percentile flows and independent variables. Although this does not show the identity of the basins assigned to the neurons, it does give an indication of the difference in mapping the basins on the SOM using the two different sets of input data. Pie charts that are split in half indicate agreement between the two datasets, whereas disagreement between the datasets is shown as pie charts dominated by a single color. Agreement between the datasets was important for the SOM predictions since the independent variables were used to assign the basins to the neurons, and a possible source of uncertainty was that the independent variables assigned the basins to different neurons than the percentile flows. The pie charts of Figure 17 are representative of that uncertainty, and show that some basins were incorrectly assigned to the neurons (all black or white pie charts). Other pie charts show a

mix of correspondence between the percentile flows and independent variables, and the similarity of these two sets of data was further examined using continuous surfaces illustrating patterns in the SOM.

The first visualization comparing continuous surfaces of the SOM shows the U-matrix calculated using (a) the percentile flows and (b) independent variables (Figure 18). The U-matrix represents the clusters (low dissimilarity between neurons) and cluster borders (high dissimilarity between neurons), and was used to compare the cluster structure of the two datasets. The percentile flows have less well-defined clusters than the independent variables. These results indicate that it is difficult to distinguish clusters using the percentile flows of the FDC. A possible reason for this may be large variability in the percentile flows. Evidence for this is that a large number of basins (Figure 17) were assigned to areas of the SOM with high dissimilarity in percentile flows. Despite having dissimilar percentile flows, these basins were assigned to the same neurons. Clusters of the percentile flows and independent variables had some agreement (see the clusters toward the center and upper right corner of the SOM), but disagreements between the datasets mark differences that may have led to uncertainty in predicting the percentile flows. The sources of (dis)agreement between the percentile flows and independent variables were investigated by comparing the values of individual variables in the SOM.

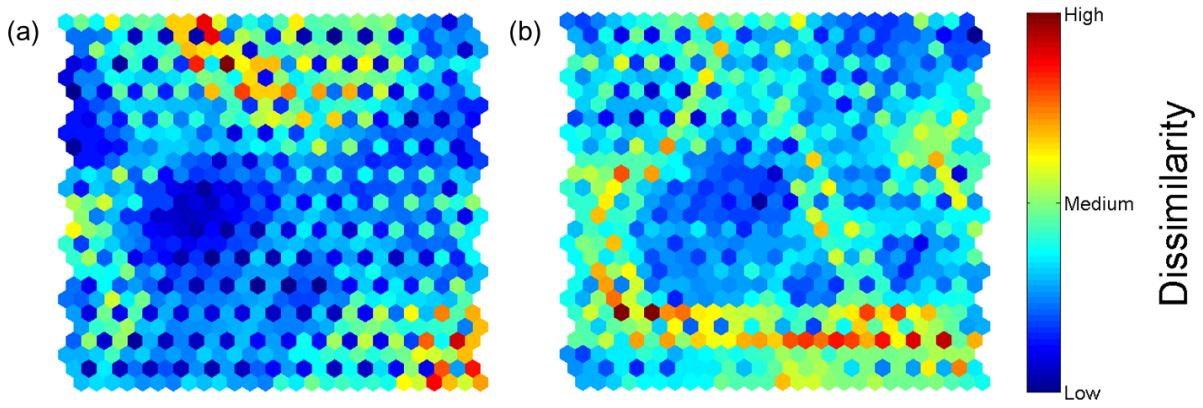

Figure 18. U-matrix of (a) percentile flows and (b) independent variables.

The values of individual variables in the SOM were visualized as component planes (Figure 19). A representative high (Q_{05}), average (Q_{50}), and low (Q_{95}) percentile flow was compared to all the independent variables. Overall, it is difficult to identify individual variables similar to the percentile flows, which is not surprising given the complexity of the processes and combination of variables that control percentile flows. Some parts of the variables were, however, related to the percentile flows as illustrated by the following examples: (1) the largest high flows (upper left corner of the SOM) corresponded with snow-dominated climates (Percent_Snow) perhaps due to the spring snowmelt season or rain-on-snow events known for generating floods (McCabe et al., 2007), (2) the smallest high flows (lower right corner of the SOM) were associated with arid regions (Aridity) characterized by low flows throughout the FDC (Pumo et al., 2014), (3) groundwater (BFI) is the main source of average flows (Yaeger et al., 2012), and tracked with the largest and smallest average flows (upper center and lower right corner of the SOM, respectively), (4) an additional area of large average flows (lower left corner of the SOM) was related to wetter climates (Aridity) able to generate more flow (Cheng et al., 2012), and (5) the largest low flows (upper center

of the SOM) overlapped with the largest groundwater contributions (BFI) as low flows are often a function of subsurface drainage (Tague and Grant, 2004).

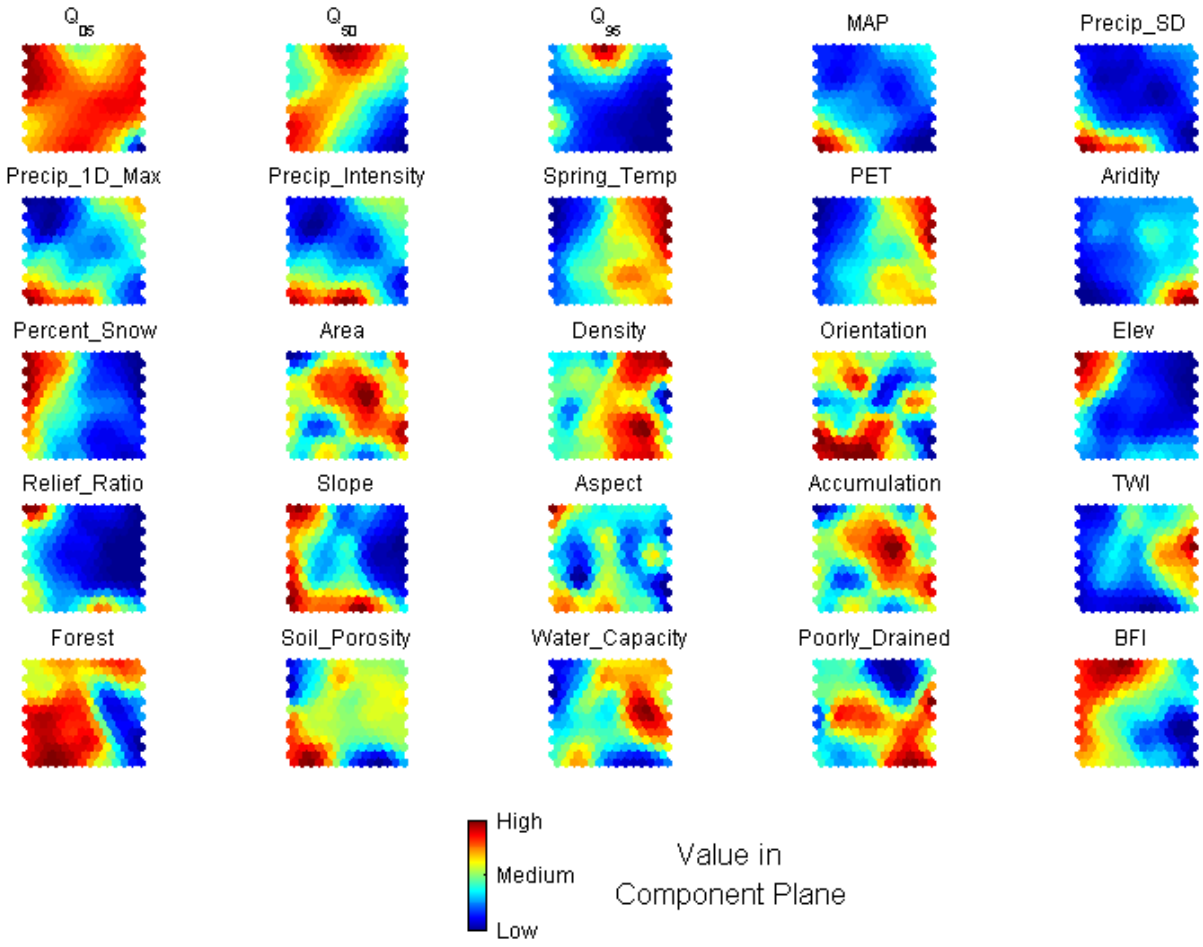

Figure 19. Component planes of representative flows (Q_{05}, Q_{50}, and Q_{95}) and all the independent variables.

The component planes of Figure 19 could be used to identify more connections between the percentile flows and independent variables, but explaining all the patterns in the percentile flows would be difficult using the given set of independent variables. Limitations in the explanatory power of the independent variables may be responsible for the uncertainty of predicting the percentile flows in this study. Additional variables may be needed to more fully capture the complexity of the percentile flows. For instance, the measures of storm magnitude (Precip_1D_Max and Precip_Intensity) were not well-related to the high flow

(Figure 19), and the connection between high flows and storms may be strengthened by variables that quantify the magnitude of large storms, such as precipitation percentiles (Ssegane et al., 2012b). The average flow was previously associated with BFI and Aridity, but other parts of the average flow's component plane are not clearly connected to the independent variables. This may be because the average flow is affected by the combined influence of small storms, groundwater discharge, and evaporative losses that cannot be encapsulated in a single variable. The component plane of the low flow consisted of mostly low values with subtle variations, and none of the variables reflected these subtle variations, which are likely the function of subsurface drainage properties (Tague and Grant, 2004). With the exception of BFI, the subsurface variables of this study (Soil_Porosity, Water_Capacity, and Poorly_Drained) were not strongly related to the low flow as they focused on soil properties and low flows may be the product of deeper groundwater flows (Schaller and Fan, 2009). A geologic classification could be developed to represent deep groundwater flows and the low flows that they produce (Tague and Grant, 2004).

Some other noteworthy observations were made using the component planes (Figure 19). The percentile flows were autocorrelated as order statistics derived from the same time series (see the similar structure of the component planes). Similar patterns in magnitude were observed for the different percentile flows, but the area with the largest magnitude changed location from the average flow (upper center of the SOM) to the high flow (upper left corner of the SOM). As previously mentioned, this was related to snow-dominated climates with the potential for large flows produced during the spring snowmelt season or rain-on-snow events. The pocket of snow-dominated climates persisted in the component planes of the percentile flows including those not pictured in Figure 19. This highlights the unique hydrology of

snow-dominated climates (Bales et al., 2006), and suggests that percentile flows may need to be predicted using models specifically designed for this special type of environment (Kim and Kaluarachchi, 2014).

Another special type of environment that emerged in the component planes of Figure 19 was the arid to semi-arid region located in the lower right quadrant of the SOM. This region contained a large number of zero flows for the low flow (dark blue color dominating the lower right quadrant of the SOM), and certain variables, such as MAP, Aridity, and BFI, only accounted for a portion of the zero flows. Specific variables may need to be developed to explain zero flows, such as the number of dry days or stream channel permeability to represent the possibility of bank recharge (Snelder et al., 2013). Finally, some independent variables were cross-correlated (see the component planes of Spring_Temp and PET for an example), and many of the widely used topographic variables, such as Density, Slope, and TWI, were not strongly associated with the percentile flows. The variables that stood out from the previous discussion of connections to the percentile flows were Aridity, Percent_Snow, and BFI, and these variables were related to the percentile flows for the next visualization.

The values of the three variables connected to the percentile flows (Aridity, Percent_Snow, and BFI) are shown as pie charts for each neuron in Figure 20, and the previously displayed component planes of the high (Q_{05}), average (Q_{50}), and low (Q_{95}) percentile flows are laid underneath to further examine their connection to the three variables. This visualization confirms the previous discussion of the connections between the selected variables and percentile flows, but it repackages the information in an integrated format (independent variables and percentile flows shown together). The result gives a

clearer view of the transition between environmental conditions and their relation to the percentile flows. This is illustrated by the following examples: (1) the high flow increased from the lower right corner of the SOM as Aridity decreased possibly due to more antecedent moisture and effective rainfall that can be transformed into flow during a storm (Ye et al., 2012), (2) average flows varied with BFI (see the transitions from the lower right corner and between the two high areas of the SOM) as groundwater is the main source of average flows (Yaeger et al., 2012), and (3) the low flow increased as Aridity decreased and BFI increased from the lower right corner of the SOM, which reflects the relation between groundwater flow and climate (Santhi et al., 2008) and their combined effect on low flows (Yokoo and Sivapalan, 2011). The three variables were of course not able to explain all the percentile flow patterns (see areas with changes in the variables but similar percentile flow values), and the relation between all the variables and percentile flows was investigated for the final visualization of clusters in the two datasets.

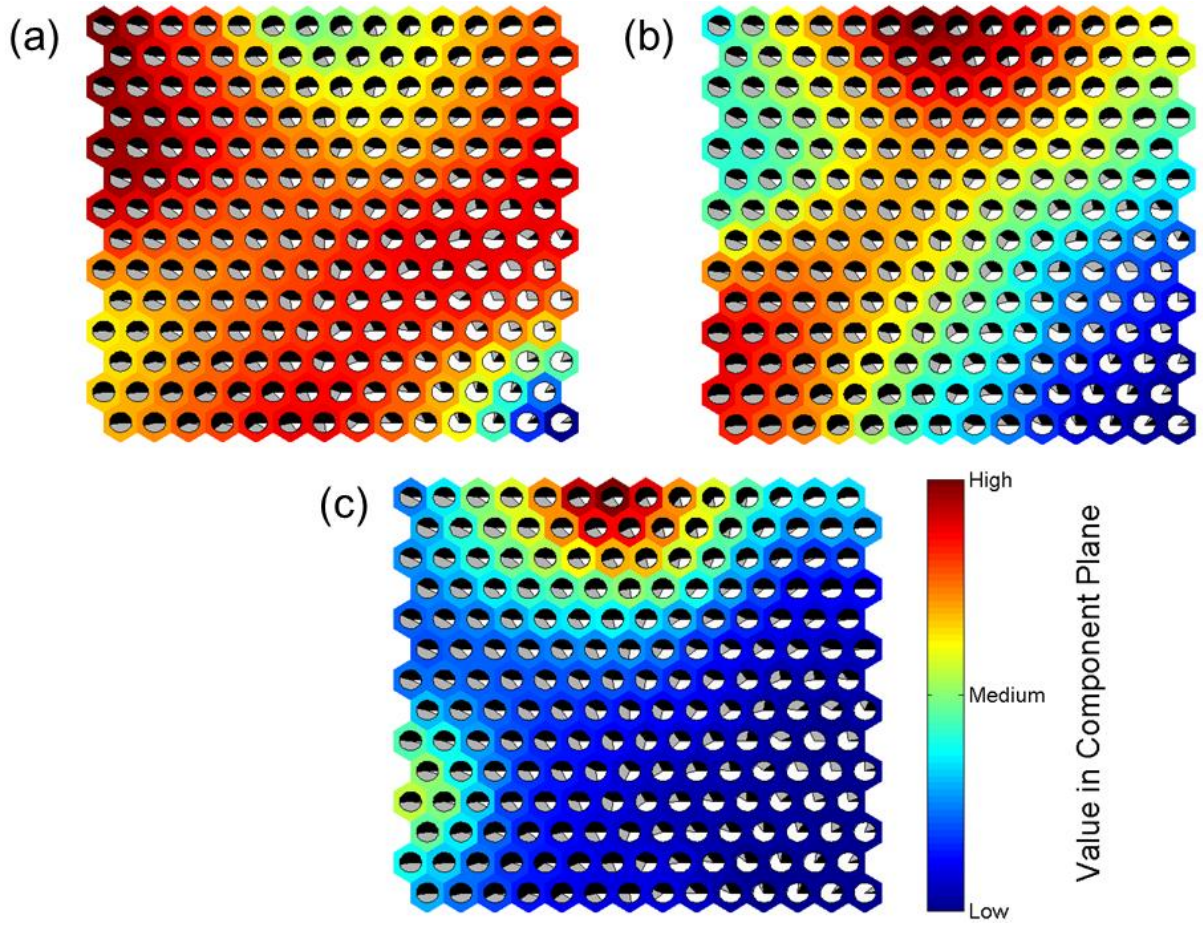

Figure 20. Component planes of (a) high (Q_{05}), (b) average (Q_{50}), and (c) low (Q_{95}) flows with pie charts showing the corresponding value of Aridity (white), Percent_Snow (gray), and BFI (black).

The final visualization compares *k*-means clusters of the SOM neuron vectors for the percentile flows versus the independent variables. The clustering was performed to view how the two datasets were organized in the SOM. Independent variables are widely used to identify regions for streamflow predictions as in the second paper of this dissertation. This approach assumes that the regions identified using independent variables follow the variation in streamflow. To test this assumption, two cluster solutions based on the independent variables and percentile flows were compared. An appropriate number of clusters for the two datasets was identified as ten using the elbow method of plotting the SSE for larger numbers of clusters. The agreement between the two cluster solutions was tabulated in a confusion

matrix (Table 25), and these results were visualized on the SOM and mapped for the US according to the BMU of the basins (Figure 21). Cluster borders were mapped for the US using Thiessen polygons of the gauge locations for the basins. The confusion matrix shows a wide range of agreement between the two cluster solutions, but the agreement was mostly poor (< 50%). The overall agreement of the two cluster solutions was assessed using a common metric called the adjusted Rand index (see Hubert and Arabie, 1985 for the mathematical definition), which ranges from ± 1 with zero indicating agreement due to chance and larger values signifying better agreement. The adjusted Rand index for the two cluster solutions was 0.19. This value is only slightly better than chance agreement, and confirms the poor agreement of the confusion matrix.

Table 25. Confusion matrix of the clusters based on the neuron vectors for the percentile flows (rows) and independent variables (columns) with bold numbers along the diagonal indicating agreement between the same cluster identified using the two different datasets. The percent agreement between the two cluster solutions is displayed in the last row and column.

Cluster	1	2	3	4	5	6	7	8	9	10	Agreement (%)
1	**3**	0	0	5	0	0	0	0	0	0	38
2	7	**6**	5	0	3	0	0	0	0	0	29
3	10	3	**2**	3	0	0	0	0	0	0	11
4	0	0	0	**3**	0	0	0	0	0	0	100
5	7	2	3	0	**7**	0	4	1	0	3	26
6	0	0	0	0	0	**17**	0	0	2	0	89
7	1	0	0	0	0	2	**6**	0	4	5	33
8	3	0	0	0	13	0	5	**16**	2	7	35
9	0	0	0	0	0	1	2	0	**8**	0	73
10	3	0	0	0	2	3	5	5	4	**3**	12
Agreement (%)	9	55	20	27	28	74	27	73	40	17	

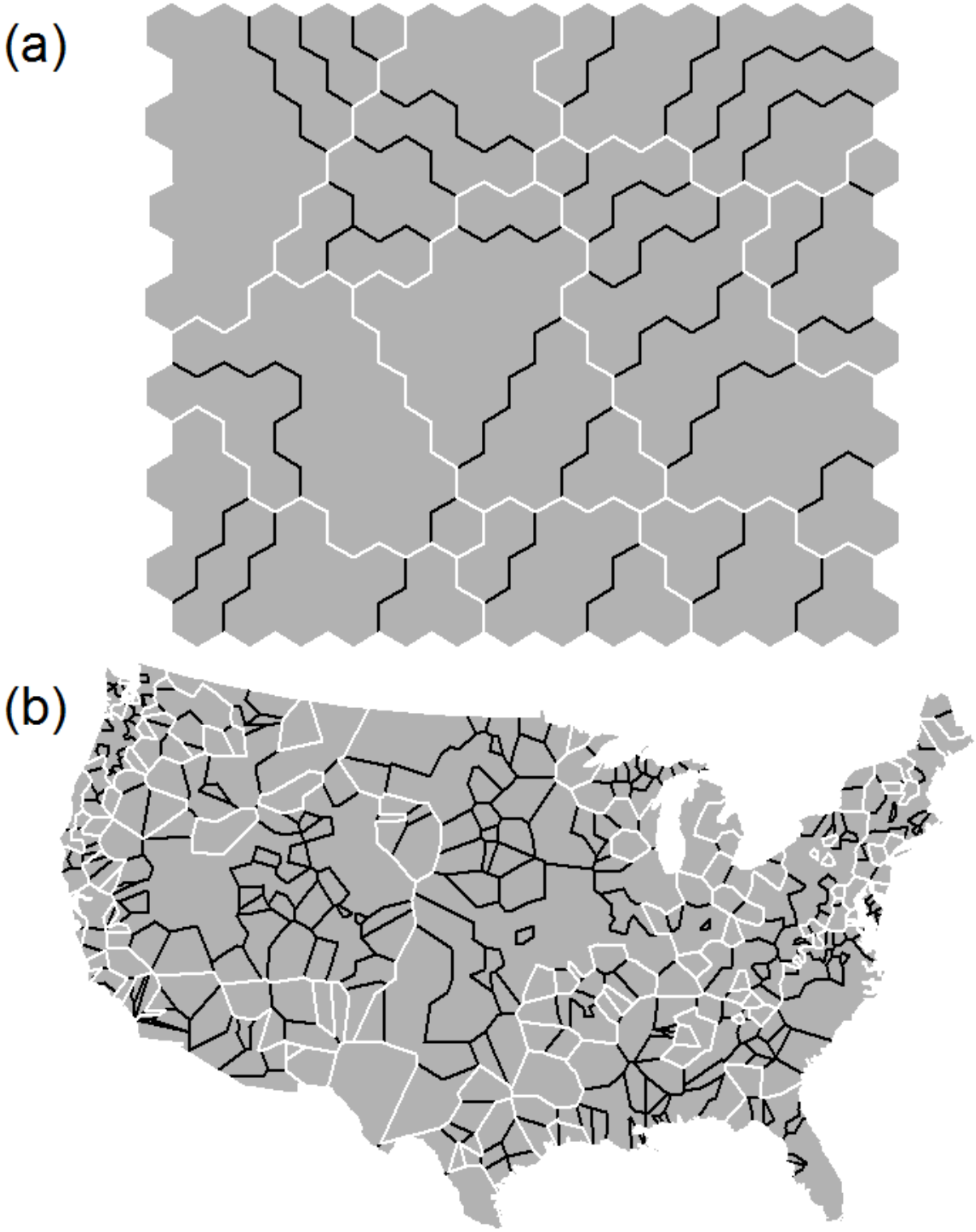

Figure 21. Clusters based on the neuron vectors for the percentile flows (black) and independent variables (white) shown on the (a) SOM and (b) mapped for the US using Thiessen polygons of the gauge locations.

The poor agreement between the percentile flows and independent variables was reflected in the maps of the cluster borders (Figure 21). The cluster borders frequently intersected, and the overall structure of the clusters was quite different. The clusters of the independent variables were cohesive in the attribute space of the SOM and the geographic space of the US. The geographic contiguity of clusters based on independent variables was previously observed in the second paper of this dissertation. Meanwhile, the percentile flow clusters were complex and fragmented, occurring as linear shapes spanning the entire SOM and geographically disconnected clusters. Some of the percentile flow clusters were clearly linked to the high and low flow areas previously discussed for the component planes (see the high flow cluster in the upper center and low flow cluster in the lower right of the SOM). The snow-dominated area in the upper left of the SOM was also demarcated, but the linear clusters between areas of high and low flow indicate a weak cluster structure dividing a continuous field.

The weak structure of the percentile flow clusters may be partially attributed to two factors: (1) the elbow method identified too many clusters and (2) the previously discussed autocorrelation of the percentile flows. The latter factor may have influenced the clusters because the autocorrelated percentile flows for a given basin were possibly larger or smaller than the other basins, and this resulted in neuron vectors ordered primarily by magnitude. The effects of autocorrelation on the SOM have previously been documented in a spatial context where closer features are more related (Bação et al., 2008), and this paper documents the effect of autocorrelation between the attributes used to train the SOM. The autocorrelated signal of the percentile flows presumably produced clusters arranged by the overall magnitude of the FDC, and these clusters were not geographically organized (Figure 21b).

This is noteworthy since many studies use spatial proximity to predict the FDC (see Booker and Snelder, 2012; Pugliese et al., 2013; Smakhtin et al., 1997), but spatial proximity was not strongly related to the FDC at the large scale of this study.

Clusters based on the percentile flows and independent variables largely disagreed (Figure 21). This is a concern because regions identified using independent variables are often used to develop models for predicting percentile flows. The discordancy between the two cluster solutions highlights the need to select independent variables that capture the variation in percentile flows. This could be performed using quantitative variable selection methods or knowledge of the factors that control percentile flows (see the first paper of this dissertation for examples).

The poor fit between the clusters in Figure 21 may be because the autocorrelated percentile flows simply represented the overall magnitude of flows for the basins, and independent variables selected to represent overall flow, such as Aridity and BFI, may have produced clusters more similar to the percentile flows. Alternatively, clusters based on the independent variables may have agreed more with streamflow variables representing different aspects of the hydrograph rather than only magnitude. This is the approach taken by hydrologic classification studies that use variables such as statistics on the rising and falling limb of the hydrograph or event runoff coefficients (see Sawicz et al., 2011 for an example). Clusters based on these types of variables have agreed more with independent variables in the past (Ley et al., 2011).

Relations between independent variables and percentile flows are critical for prediction. The SOM was used to display how the variables covaried in an ordered spatial layout of the data. This provided more detailed information than traditional statistical

measures of covariance, like correlation coefficients or statistics derived from a regression analysis. The SOM revealed several variables (Aridity, Percent_Snow, and BFI) that covaried with the percentile flows. These variables represent important factors that should be incorporated in future modeling efforts for the contiguous US. The SOM identified unique regions that may require specialized predictive models, such as snow-dominated and arid climates. Finally, discordancy between the percentile flows and independent variables in the SOM highlights the need for new variables more closely linked to the processes that shape the FDC.

E. Conclusions

The SOM was used to predict 13 percentile flows of the FDC and investigate their relation to independent variables consisting of measurable basin characteristics. The percentile flows were predicted for 184 validation basins in the US treated as ungauged. A typical procedure for predicting streamflow variables in a large study area splits the basins into regions with similar independent variables related to flow, but identifying a priori regions can be an uncertain and time-consuming process. The SOM was applied in this study since its training routine clusters the input data and may eliminate the need to identify a priori regions. SOM predictions were produced with and without a priori regions to test the hypothesis that a priori regions do not improve SOM predictions. Global predictions were generated using all the basins and four different approaches for training the SOM. The preferred training approach was then used to generate SOM predictions for a priori regions, and the performance of the global and regional predictions was compared. The results from a regional regression were also included as a reference for assessing the predictive performance of the SOM. Visualizations based on the SOM were produced for an

exploratory analysis of how the independent variables related to the percentile flows to potentially improve future modeling efforts.

The predictive performance of the SOM with and without a priori regions was similar. This confirmed the hypothesis that the SOM did not need a priori regions to predict percentile flows. Despite its success without a priori regions, the SOM did not perform as well as the regional regression. However, the regional regression likely benefited from using a variable selection method to discard irrelevant variables unrelated to the percentile flows. Performance of the SOM may be improved by applying a similar variable selection method to train the SOM without irrelevant variables. Uncertainty of the SOM predictions may also be related to the parameters used to train the SOM, such as the number of neurons. The SOM may also require more data to avoid over-fitting the predictions to a limited sample. Future studies should experiment with the training parameters to assess their effect on SOM predictions and expand the sample of training data to diminish the potential for over-fitting.

In light of the results from this study, machine learning methods that cluster data to generate predictions may be an alternative to regional percentile flow predictions if they are trained using a relevant set of independent variables. The SOM is an unsupervised learning method that does not use an objective function to minimize the error of its output. Percentile flow predictions may be improved using a supervised learning method, such as random forests, that applies an objective function to cluster the input data and minimize the error of the predictive model.

Information for future modeling efforts was extracted from SOM data visualizations of the relation between the percentile flows and independent variables. Individual percentile flows were related to several key variables. High flows were associated with the overall

wetness of the basin (Aridity) and the amount of snowfall (Percent_Snow) as it relates to the large flows of the spring snowmelt season or rain-on-snow events. Average and low flows were primarily related to groundwater contributions (BFI), which is interrelated with the wetness of the region (Aridity). The key variables (Aridity, Percent_Snow, and BFI) are essential factors for modeling percentile flows at the scale of the contiguous US. Some variables were not associated with the percentile flows, and the most noteworthy were topographic variables widely used for streamflow prediction. The independent variables were least connected to low flows fed by groundwater and subject to zero flows. Low flows may be better predicted using variables that represent subsurface drainage through a geologic classification and the likelihood of zero flows via the duration of dry periods.

Overall agreement between the percentile flows and independent variables was weak according to clusters derived from the two datasets. This is an important point since independent variables similar to those used in this study are widely used to identify regions for streamflow predictions. This approach operates on the assumption that regions derived from independent variables reflect variations in streamflow. The clusters in this study did not confirm this assumption, and independent variables should be more strategically selected to identify regions associated with percentile flows. This could include applying quantitative or knowledge-based variable selection methods prior to performing the cluster analysis for identifying regions. New variables may also need to be developed to strengthen the agreement between independent variables and percentile flows, and this could be investigated using the SOM.

Chapter 5: Conclusions

The goal of this dissertation was to investigate various sources of uncertainty for predicting percentile flows concerning (1) independent variable selection, (2) the amount of information for the initial set of independent variables, and (3) the identification of a priori regions to develop predictions. This goal was accomplished by evaluating various methods and sets of variables to predict 13 percentile flows for 918 basins in the US. The large sample of basins was used to improve the generality of the results for future studies and produce models that could be used to predict percentile flows for ungauged basins throughout the contiguous US. These models could be published to provide local watershed managers with a tool to predict percentile flows for ungauged basins with an estimate of the error.

The first study presented in Chapter 2 investigated how the independent variables should be selected for percentile flow regression models. An automated regression procedure for selecting the variables was evaluated against alternative methods from the field of variable (feature) selection. Common methods for selecting the independent variables of regression models in hydrology, including the automated regression procedure and principal component analysis, performed worse than the other variable selection methods. The other methods all performed similarly, but random forests produced the best overall results. Another notable result was that the variables selected based on hydrologic knowledge of the FDC performed nearly as well as the advanced machine learning methods, such as random forests. The other variables added little predictive information to the regression models, and widely used topographic variables were not useful for predicting the percentile flows. The most important variables for predicting the percentile flows at the scale of the US were groundwater flows expressed as the baseflow index (BFI) and percent forest cover in the

basin. These variables were likely important as they integrate the effects of climate, vegetation, and geology on the FDC. Overall, the regression models mostly explained less than half of the variance in the percentile flows ($R^2 < 0.5$). This was likely due to the large variability of percentile flows in the US, and a regional regression approach was adopted for the next study of the dissertation.

The amount of information for the independent variables of a regional regression was evaluated in Chapter 3. A regional regression was performed to predict percentile flows using different sets of variables ranging in complexity from three hydrologically-based variables to 37distributed variables describing average conditions and the statistical distribution of basin data. Only three hydrologically-based variables were necessary to perform the regional regression as they performed similarly to a typical set of 22 lumped variables describing average conditions and outperformed the more detailed set of 37 distributed variables. The result speaks to the importance of using variables with a strong hydrologic justification and downplays the use of data-driven approaches that include many variables with potentially weak connections to the percentile flows. The strong predictive performance of the three hydrologically-based variables once again highlights the limited predictive information provided by the additional variables. All sets of variables for the regional regression performed better than the global regression models from Chapter 2. The regions identified for the regression were related to the percentile flows, and regional differences in storage (BFI and snowfall) and climate (aridity) were associated with the FDC.

The final study of the dissertation in Chapter 4 used the SOM for prediction and an exploratory analysis of the percentile flows. The SOM was used to cluster the data and predict percentile flows in one step, which avoided the decisions of identifying a priori

regions, such as the number of regions. The approach was applied using all the basins to generate global predictions, and the SOM was also applied in the regions from Chapter 3. Performance of the global and regional predictions were compared to determine if the SOM could be used without the identification of a priori regions. The global and regional predictions generated using the SOM performed similarly, indicating that the SOM did not need regions to predict the percentile flows. The SOM achieved similar performance to the best global regression models from Chapter 2, but did not perform as well as the regional regression from Chapter 3. This is likely because the regional regression only used a subset of relevant independent variables, whereas the SOM used all the independent variables, potentially including irrelevant information that diminished the predictive performance of the SOM. The performance of the SOM may be improved by applying a variable selection method to exclude irrelevant variables. Output from the SOM was converted into data visualizations for an exploratory analysis of the variables related to the percentile flows. Overall agreement between the percentile flows and independent variables was weak, which confirmed that the SOM included irrelevant variables that may have reduced its predictive performance. Notable variables unrelated to the percentile flows were widely used topographic variables. The percentile flows were, however, related to several key variables including aridity, snowfall, and BFI. The importance of these three variables was previously highlighted by the regions from Chapter 3.

Overlying themes from the three studies of this dissertation are summarized as follows:

- Independent variable selection is a critical, but often overlooked, step in predicting percentile flows.

- A parsimonious set of hydrologically-based variables can be used to predict percentile flows as many widely used independent variables, such as topographic variables, offer little additional predictive information.
- Regions based on physical and climatic characteristics were related to the percentile flows and improved their prediction.
- The SOM may be an alternative to identifying a priori regions provided it is applied using a relevant set of independent variables.
- Several key independent variables for predicting percentile flows emerged from the studies, namely aridity, snowfall, and BFI. Snowfall was important at the scale of the US, but obviously may not be relevant for study areas without snow.
- BFI was an indispensable variable for predicting the percentile flows. However, it should be noted that BFI is derived from streamflow data. BFI was considered an independent variable in this study because a gridded product exists for the US (Wolock, 2003). This type of product should be developed for other study areas to predict percentile flows.
- Predictive performance decreased for the high and low percentile flows, and future research should address the uncertainty of predicting extreme flows with more variability between basins.

A. Future research

The independent variables used in this dissertation were representative of the variables typically used to predict percentile flows, and many of these variables offered little predictive information. Future research should develop new variables to predict percentile flows. These variables should focus on representing the processes that control the FDC. For

the high percentile flows, this would include variables that represent storm flows. Average to low percentile flows may be better explained using variables associated with subsurface drainage (e.g. water table depth and hydrologically-based geologic units) and evaporative losses (e.g. dry period duration and vegetation transpiration).

Baseflow was the most important independent variable for predicting percentile flows, but it must be derived using streamflow data. In countries with sufficient streamflow gauging networks, a gridded baseflow product like the one in the US should be produced to predict percentile flows along with other streamflow variables. For sparsely gauged regions, surrogate variables should be proposed to represent baseflow, such as a soil or geologic classification.

The exploratory components of this dissertation revealed that snow-dominated and arid regions have unique hydrologic properties. Separate predictive models may need to be developed for these unique regions. Such models would account for the effect of snow on the FDC, and attempt to predict zero flows in arid regions. None of the models in this dissertation predicted zero flows, and future research should develop models for intermittent streams given their widespread distribution throughout the US in both arid and non-arid regions.

The SOM showed that it could be used to predict percentile flows without regions. However, it was outperformed by a regional regression that discarded irrelevant independent variables. Future research should pair the SOM with a variable selection method to predict percentile flows. This would exclude irrelevant information, and may enhance the predictive performance of the SOM.

This dissertation investigated one alternative, the SOM, for identifying regions to predict percentile flows. Other alternatives exist to identify regions or avoid such a priori designations. A particularly intriguing approach for identifying regions is fuzzy clustering that represents the continuous gradient of similarity between the basins. Unlike discrete clustering, fuzzy clustering does not place potentially arbitrary divisions in the data, and the basins would be given partial membership to each region. This could improve predictions for basins with characteristics representative of multiple regions. A fuzzy clustering approach would be akin to ensemble modeling, with models averaged over regions. Finally, the SOM is not the only machine learning method that clusters data as it generates predictions. Other methods, such as random forests, should be evaluated for their potential to predict percentile flows and avoid identifying a prior regions.

References

Abdi, H., Williams, L.J., 2010. Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics 2, 433-459. doi: 10.1002/wics.101

Acreman, M.C., Sinclair, C.D., 1986. Classification of drainage basins according to their physical characteristics; an application for flood frequency analysis in Scotland. Journal of Hydrology 84, 365-380. doi: 10.1016/0022-1694(86)90134-4

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. Journal of Machine Learning Research 11, 171-234.

Aliferis, C.F., Tsamardinos, I., Statnikov, A., 2003. HITON: A Novel Markov Blanket Algorithm for Optimal Variable Selection. Proceedings of the American Medical Informatics Association Annual Symposium, 21-25.

Allende, H., Moreno, S., Rogel, C., Salas, R., 2004. Robust Self-organizing Maps, in: Sanfeliu, A., Trinidad, J.F.M, Ochoa, J.A.C. (Eds.), Progress in Pattern Recognition, Image Analysis and Applications. Springer, Berlin, Germany, pp. 179-186.

Andréassian, V., Lerat, J., Loumagne, C., Mathevet, T., Michel, C., Oudin, L., Perrin, C., 2007. What is really undermining hydrologic science today? Hydrological Processes 21, 2819-2822. doi: 10.1002/hyp.6854

Archfield, S.A., Vogel, R.M., 2010. Map correlation method: Selection of a reference streamgage to estimate daily streamflow at ungaged catchments. Water Resources Research 46. doi: 10.1029/2009WR008481

Archfield, S.A., Vogel, R.M., Steeves, P.A., Brandt, S.L., Weiskel, P.W., Garabedian, S.P., 2009. The Massachusetts Sustainable-Yield Estimator: A decision-support tool to assess water availability at ungaged sites in Massachusetts. US Geological Survey Scientific Investigations Report 2009-5227, 41 pp.

Arguez, A., Durre, I., Applequist, S., Vose, R.S., Squires, M.F., Yin, X., Heim, R.R., Owen, T.W., 2012. NOAA's 1981-2010 US Climate Normals: An Overview. Bulletin of the American Meteorological Society 93, 1687–1697. doi: 10.1175/BAMS-D-11-00197.1

Bação, F., Lobo, V., Painho, M., 2008. Applications of different self-organizing map variants to geographical information science problems, in: Agarwal, P., Skupin, A. (Eds.), Self-Organising Maps: Applications in Geographic Information Science. John Wiley and Sons, Chichester, UK, pp. 21-44.

Baguley, T., 2012. Multiple regression and the general linear model, in: Baguley, T. (Ed.), Serious Stats: A Guide to Advanced Statistics for the Behavioral Sciences. Palgrave Macmillan, New York City, USA, pp. 423-471.

Bailey, R.G., 1983. Delineation of Ecosystem Regions. Environmental Management 7, 365-373.

Bales, R.C., Molotch, N.P., Painter, T.H., Dettinger, M.D., Rice, R., Dozier, J., 2006. Mountain hydrology of the western United States. Water Resources Research 42. doi: 10.1029/2005WR004387

Bart, R., Hope, A., 2010. Streamflow response to fire in large catchments of a Mediterranean-climate region using paired-catchment experiments. Journal of Hydrology 388, 370-378. doi: 10.1016/j.jhydrol.2010.05.016

Bartz-Beielstein, T., Zaefferer, M., 2012. A Gentle Introduction to Sequential Parameter Optimization. CIplus 2.

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, USA, pp. 85-191.

Berger, D.E., 2004. Using Regression Analysis, in: Wholey, J.S., Hatry, H.P., Newcomer, K.E. (Eds.), Handbook of Practical Program Evaluation. Wiley, San Francisco, USA, pp.479-505.

Berger, K.P., Entekhabi, D., 2001. Basin hydrologic response relations to distributed physiographic descriptors and climate. Journal of Hydrology 247, 169-182.

Best, A.E., Zhang, L., McMahon, T.A., Western, A.W., 2004. Development of a Model for Predicting the Changes in Flow Duration Curves Due to Altered Land Use Conditions, in: Post, D. (Ed.), MODSIM 2003 International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand, Canberra, Australia, pp. 861-866.

Bloomfield, J.P., Allen, D.J., Griffiths, K.J., 2009. Examining geological controls on baseflow index (BFI) using regression analysis: An illustration from the Thames Basin, UK. Journal of Hydrology 373, 164-176. doi: 10.1016/j.jhydrol.2009.04.025

Booker, D.J., Snelder, T.H., 2012. Comparing methods for estimating flow duration curves at ungauged sites. Journal of Hydrology 434-435, 78-94. doi: 10.1016/j.jhydrol.2012.02.031

Booker, D.J., Woods, R.A., 2014. Comparing and combining physically-based and empirically-based approaches for estimating the hydrology of ungauged catchments. Journal of Hydrology 508, 227-239. doi: 10.1016/j.jhydrol.2013.11.007

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

Bowden, G.J., Dandy, G.C., Maier, H.R., 2005. Input determination for neural network models in water resources applications. Part 1-background and methodology. Journal of Hydrology 301, 75-92. doi: 10.1016/j.jhydrol.2004.06.021

Breiman, L., 2001. Random Forests. Machine Learning 45, 5-32.

Brown, M.E., Escobar, V., Moran, S., Entekhabi, D., O'Neill, P.E., Njoku, E.G., Doorn, B., Entin, J.K., 2013. NASA's soil moisture active passive (SMAP) mission and opportunities for applications users. Bulletin of the American Meteorological Society 94, 1125-1128. doi: 10.1175/BAMS-D-11-00049.1

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

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

Carrillo, G., Troch, P.A., Sivapalan, M., Wagener, T., Harman, C., Sawicz, K., 2011. Catchment classification: hydrological analysis of catchment behavior through process-based modeling along a climate gradient. Hydrology and Earth System Sciences Discussions 8, 4583-4640. doi: 10.5194/hessd-8-4583-2011

Castellarin, A., 2014. Regional prediction of flow-duration curves using a three-dimensional kriging. Journal of Hydrology 513, 179-191. doi: 10.1016/j.jhydrol.2014.03.050

Castellarin, A., Camorani, G., Brath, A., 2007. Predicting annual and long-term flow-duration curves in ungauged basins. Advances in Water Resources 30, 937-953. doi: 10.1016/j.advwatres.2006.08.006

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

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. Hydrology and Earth System Sciences Discussions 9, 7001-7034.
doi: 10.5194/hessd-9-7001-2012

Chiang, S., Tsay, T., Nix, S.J., 2002a. Hydrologic Regionalization of Watersheds. I: Methodology Development. Journal of Water Resources Planning and Management 128, 3-11. doi: 10.1061/(ASCE)0733-9496(2002)128:1(3)

Chiang, S., Tsay, T., Nix, S.J., 2002b. Hydrologic Regionalization of Watersheds. II: Applications. Journal of Water Resources Planning and Management 128, 12-20. doi: 10.1061/(ASCE)0733-9496(2002)128:1(12)

Coleman, A.M., 2008. Fundamental Basis of Artificial Neural Networks, in: Coleman, A.M. (Ed.), An Adaptive Landscape Classification Procedure Using Geoinformatics and Artificial Neural Networks. Vrije Universiteit, Amsterdam, The Netherlands, pp. 26-32.

Commission for Environmental Cooperation, 1997. Ecological Regions of North America: Toward a Common Perspective. Commission for Environmental Cooperation, Montréal, Canada.

Coopersmith, E., Yaeger, M.A., Ye, S., Cheng, L., Sivapalan, M., 2012. Exploring the physical controls of regional patterns of flow duration curves – Part 3: A catchment classification system based on regime curve indicators. Hydrology and Earth System Sciences 16, 4467-4482. doi: 10.5194/hess-16-4467-2012

Copas, J.B., 1983. Regression, Prediction and Shrinkage. Journal of the Royal Statistical Society. Series B (Methodological) 45, 311-354.

Coulibaly, P., Evora, N.D., 2007. Comparison of neural network methods for infilling missing daily weather records. Journal of Hydrology 341, 27-41. doi: 10.1016/j.jhydrol.2007.04.020

Croker, K.M., Young, A.R., Zaidman, M.D., Rees, H.G., 2003. Flow duration curve estimation in ephemeral catchments in Portugal. Hydrological Sciences Journal 48, 427-439. doi: 10.1623/hysj.48.3.427.45287

Dalton, K.L., 2005. Variation in timing of vegetation peak greenness on the north slope of Alaska, 1982-1999. San Diego State University, San Diego, USA.

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

Dawdy, D.R., Griffis, V.W., Gupta, V.K., 2012. Regional Flood-Frequency Analysis: How We Got Here and Where We Are Going. Journal of Hydrologic Engineering 17, 953-959. doi: 10.1061/(ASCE)HE.1943-5584.0000584

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

Dingman, S.L., 2001. Physical Hydrology, Second Edition. Pearson, Upper Saddle River, USA.

Di Prinzio, M., Castellarin, A., Toth, E., 2011. Data-driven catchment classification: application to the pub problem. Hydrology and Earth System Sciences 15, 1921-1935. doi: 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. doi: 10.1111/j.1600-0587.2012.07348.x

Douglas, E.M., Vogel, R.M., Kroll, C.N., 2000. Trends in floods and low flows in the United States: impact of spatial correlation. Journal of Hydrology 240, 90-105.

Downey, J.S., Dinwiddie, G.A., 1988. The Regional Aquifer System Underlying the Northern Great Plains in Parts of Montana, North Dakota, South Dakota, and Wyoming – Summary. US Geological Survey Professional Paper 1402-A, 73 pp.

Falcone, J.A., 2011. GAGES-II: Geospatial attributes of gages for evaluating streamflow. US Geological Survey Digital Spatial Dataset.

Farsadnia, F., Rostami Kamrood, M., Moghaddam Nia, A., Modarres, R., Bray, M.T., Han, D., Sadatinejad, J., 2014. Identification of homogeneous regions for regionalization of watersheds by two-level self-organizing feature maps. Journal of Hydrology 509, 387-397. doi: 10.1016/j.jhydrol.2013.11.050

Flasch, O., Mersmann, O., Bartz-Beielstein, T., Stork, J., Zaefferer, M., 2014. rgp: R genetic programming framework. R package version 0.4-1. http://CRAN.R-project.org/package=rgp.

Flom, P., 1999. Multicollinearity diagnostics for multiple regression: A Monte Carlo study. Fordham University, Doctoral Dissertation.

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 Annual Conference.

Foster, H.A., 1934. Duration curves. Transactions of the American Society of Civil Engineers 99, 1213-1267.

Friedman, N., Geiger, D., Goldszmidt, M., 1997. Bayesian Network Classifiers. Machine Learning 29, 131-163.

Ganora, D., Claps, P., Laio, F., Viglione, A., 2009. An approach to estimate nonparametric flow duration curves in ungauged basins. Water Resources Research 45. doi: 10.1029/2008WR007472

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

Hall, M.J., Minns, A.W., 1999. The classification of hydrologically homogenous regions. Hydrological Sciences Journal 44, 693-704.

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. Advances in Water Resources 68, 1-12. doi: 10.1016/j.advwatres.2014.02.009

He, Y., Bárdossy, A., Zehe, E., 2011. A review of regionalization for continuous streamflow simulation. Hydrology and Earth System Sciences 15, 3539-3553. doi: 10.5194/hess-15-3539-2011

Heuvelmans, G., Muys, B., Feyen, J., 2006. Regionalisation of the parameters of a hydrological model: Comparison of linear regression models with artificial neural nets. Journal of Hydrology 319, 245-265. doi: 10.1016/j.jhydrol.2005.07.030

Holmes, M.G.R., Young, A.R., Gustard, A., Grew, R., 2002. A region of influence approach to predicting flow duration curves within ungauged catchments. Hydrology and Earth System Sciences 6, 721-731.

Hope, A., Bart, R., 2011. Evaluation of a regionalization approach for daily flow duration curves in central and southern California watersheds. Journal of the American Water Resources Association 48, 123-133. doi: 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. doi: 10.4314/wsa.v38i2.4

Hope, A., Burvall, A., Germishuyse, T., Newby, T., 2009. River flow response to changes in vegetation cover in a South African fynbos catchment. Water SA 35, 55-60.

Hosking, J.R.M., Wallis, J.R., 1997. Regional Frequency Analysis: An Approach Based on L-Moments. Cambridge University Press, New York City, USA.

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. Hydrological Sciences Journal 58, 1198-1255. doi: 10.1080/02626667.2013.803183

Hubert, L., Arabie, P., 1985. Comparing Partitions. Journal of Classification 2, 193-218.

Ilorme, F., 2011. Development of a Physically-based Method for Delineation of Hydrologically Homogeneous Regions and Flood Quantile Estimation in Ungauged Basins Via the Index Flood Method, Michigan Technological University, Houghton, USA.

Isik, S., Singh, V.P., 2008. Hydrologic Regionalization of Watersheds in Turkey. Journal of Hydrologic Engineering 13, 824-834. doi: 10.1061/(ASCE)1084-0699(2008)13:9(824)

Jingyi, Z., Hall, M.J., 2004. Regional flood frequency analysis for the Gan-Ming River basin in China. Journal of Hydrology 296, 98-117. doi: 10.1016/j.jhydrol.2004.03.018

Jothityangkoon, C., Sivapalan, M., Farmer, D.L., 2001. Process controls of water balance variability in a large semi-arid catchment: downward approach to hydrological model development. Journal of Hydrology 254, 174-198.

Judd, C.M., McClelland, G.H., Ryan, C.S., 2009. Multiple Regression: Models with Multiple Continuous Predictors, in: Judd, C.M., McClelland, G.H., Ryan, C.S. (Eds.), Data Analysis: A Model Comparison Approach, Second Edition. Routledge, New York City, USA, pp. 99-128.

Kalteh, A.M., Berndtsson, R., 2007. Interpolating monthly precipitation by self-organizing map (SOM) and multilayer perceptron (MLP). Hydrological Sciences Journal 52, 305-317. doi: 10.1623/hysj.52.2.305

Kalteh, A.M., Hjorth, P., Berndtsson R., 2008. Review of the self-organizing map (SOM) approach in water resources: Analysis, modelling and application. Environmental Modelling and Software 23, 835-845. doi: 10.1016/j.envsoft.2007.10.001

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 Research and Applications 26, 137-156. doi: 10.1002/rra.1249

Kilmartin, R.F., Peterson, J.R., 1972. Rainfall-Runoff Regression with Logarithmic Transforms and Zeros in the Data. Water Resources Research 8, 1096-1099.

Kim, D., Kaluarachchi, J., 2014. Predicting streamflows in snowmelt-driven watersheds using the flow duration curve method. Hydrology and Earth System Sciences 18, 1679-1693. doi: 10.5194/hess-18-1679-2014

Klemeš, V., 1986. Operational testing of hydrological simulation models. Hydrological Sciences Journal 31, 13-24.

Kohonen, T., 1998. The self-organizing map. Neurocomputing 21, 1-6. doi: 10.1016/S0925-2312(98)00030-7

Kohonen, T., 2001. Self-Organizing Maps, Springer-Verlag, Berlin, Germany.

Kohonen, T., Hynninen, J., Kangas, J., Laaksonen, J., 1996. SOM_PAK: The Self-Organizing Map Program Package. Helsinki University of Technology, Helsinki, Finland.

Koza, J.R., 1994. Genetic programming as a means for programming computers by natural selection. Statistics and Computing 4, 87-112.

Kroll, C., Luz, J., Allen, B., Vogel, R.M., 2004. Developing a Watershed Characteristics Database to Improve Low Streamflow Prediction. Journal of Hydrologic Engineering 9, 116-125. doi: 10.1061/(ASCE)1084-0699(2004)9:2(116)

Kult, J.M., Fry, L.M., Gronewold, A.D., Choi, W., 2014. Regionalization of hydrologic response in the Great Lakes basin: Considerations of temperal scales of analysis. Journal of Hydrology 519, 2224-2237. doi: 10.1016/j.jhydrol.2014.09.083

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

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

Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest. R News 2, 18-22.

Lin, G., Wu, M., Chen, G., Liu, S., 2010. Construction of design hyetographs for locations without observed data. Hydrological Processes 24, 481-491. doi: 10.1002/hyp.7500

Lu, Y., Cohen, I., Zhou, X.S., Tian, Q., 2007. Feature Selection Using Principal Feature Analysis. ACM International Conference on Multimedia.

Lumley, T., 2009. leaps: regression subset selection. R package version 2.9. http://CRAN.R-project.org/package=leaps.

Maier, H.R., Jain, A., Dandy, G.C., Sudheer, K.P., 2010. Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions. Environmental Modelling and Software 25, 891-909. doi: 10.1016/j.envsoft.2010.02.003

Mani, S., Cooper, G.F., 1999. A Study in Causal Discovery from Population-Based Infant Birth and Death Records. Proceedings of the American Medical Informatics Association Annual Symposium, 315-319.

McCabe, G.J., Clark, M.P., Hay, L.E., 2007. Rain-on-snow events in the western United States. Bulletin of the American Meteorological Society 88, 319-328. doi: 10.1175/BAMS-88-3-319

McMillan, H., Krueger, T., Freer, J., 2012. Benchmarking observational uncertainties for hydrology: rainfall, river discharge and water quality. Hydrological Processes 26, 4078-4111. doi: 10.1002/hyp.9384

Meganck, S., Leray, P., Manderick, B., 2006. Learning Causal Bayesian Networks from Observations and Experiments: A Decision Theoretic Approach, in: Torra, V., Narukawa, Y., Valls, A., Domingo-Ferrer, J. (Eds.), Proceedings of the Third International Conference on Modeling Decisions for Artificial Intelligence. Springer, Berlin, Germany, pp. 58-69.

Mendicino, G., Senatore, A., 2013. Evaluation of parametric and statistical approaches for the regionalization of flow duration curves in intermittent regimes. Journal of Hydrology 480, 19-32. doi: 10.1016/j.jhydrol.2012.12.017

Merz, R., Blöschl, G., 2008. Flood frequency hydrology: 1. Temporal, spatial, and causal expansion of information. Water Resources Research 44. doi: 10.1029/2007WR006744

Miller, A., 2002. Finding subsets which fit well, in: Miller, A. (Ed.), Subset Selection in Regression, Second Edition. CRC Press, Boca Raton, 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 Interactions 2.

Mohamoud, Y.M., 2008. Prediction of daily flow duration curves and streamflow for ungauged catchments using regional flow duration curves. Hydrological Sciences Journal 53, 706-724. doi: 10.1623/hysj.53.4.706

Morin, E., Goodrich, D.C., Maddox, R.A., Gao, X., Gupta, H.V., Sorooshian, S., 2006. Spatial patterns in thunderstorm rainfall events and their coupling with watershed hydrological response. Advances in Water Resources 29, 843-860. doi: 10.1016/j.advwatres.2005.07.014

Müller, M.F., Thompson, S.E., 2015. Stochastic or statistic? Comparing flow duration curve models in ungauged basins and changing climates. Hydrology and Earth System Sciences Discussions 12, 9765-9811. doi: 10.5194/hessd-12-9765-2015

Mwale, F.D., Adeloye, A.J., Rustum, R., 2012. Infilling of missing rainfall and streamflow data in the Shire River basin, Malawi – A self organizing map approach. Physics and Chemistry of the Earth. doi:10.1016/j.pce.2012.09.006.

Nagler, P.L., Scott, R.L., Westenburg, C., Cleverly, J.R., Glenn, E.P., Huete, A.R., 2005. Evapotranspiration on western U.S. rivers estimated using the Enhanced Vegetation Index from MODIS and data from eddy covariance and Bowen ratio flux towers. Remote Sensing of Environment 97, 337-351. doi: 10.1016/j.rse.2005.05.011

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I – A discussion of principles. Journal of Hydrology 10, 282-290.

O'Connell, P.P.L., 1868. On the relation of the freshwater floods of rivers to the areas and physical features of their basins and on a method of classifying rivers and streams with reference to the magnitude of their floods. Minutes of the Proceedings of the Institution of Civil Engineers 27, 204-217.

Olcott, P.G., 1995. Ground Water Atlas of the United States: Connecticut, Maine, Massachusetts, New Hampshire, New York, Rhode Island, Vermont. US Geological Survey HA 730-M.

Olden, J.D., Kennard, M.J., Pusey, B.J., 2012. A framework for hydrologic classification with a review of methodologies and applications in ecohydrology. Ecohydrology 5, 503-518.

Oudin, L., Andréassian, V., Perrin, C., Michel, C., Le Moine, N., 2008. Spatial proximity, physical similarity, regression and ungaged catchments: A comparison of regionalization approaches based on 913 French catchments. Water Resources Research 44.
doi: 10.1029/2007WR006240

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. Journal of Hydrology 303, 290-306. doi: 10.1016/j.jhydrol.2004.08.026

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

Patil, S., Stieglitz, M., 2012. Controls on hydrologic similarity: role of nearby gauged catchments for prediction at an ungauged catchment. Hydrology and Earth System Sciences 16, 551-562. doi: 10.5194/hess-16-551-2012

Pearl, J., 2014. Understanding Simpson's Paradox. The American Statistician 68, 8-13. doi: 10.1080/00031305.2014.876829

Peel, M.C., Finlayson, B.L., McMahon, T.A., 2007. Updated world map of the Köppen-Geiger climate classification. Hydrology and Earth System Sciences 11, 1633-1644.

Pilgrim, D.H., Chapman, T.G., Doran, D.G., 1988. Problems of rainfall-runoff modelling in arid and semiarid regions. Hydrological Sciences Journal 33, 379-400.
doi: 10.1080/02626668809491261

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 Resources Research 50, 2798-2814. doi: 10.1002/2013WR014203

Pugliese, A., Castellarin, A., Brath, A., 2013. Geostatistical prediction of flow-duration curves. Hydrology and Earth System Sciences Discussions 10, 13053-13091. doi: 10.5194/hessd-10-13053-2013

Pumo, D., Viola, F., La Loggia, G., Noto, L.V., 2014. Annual flow duration curves assessment in ephemeral small basins. Journal of Hydrology 519, 258-270. doi: 10.1016/j.jhydrol.2014.07.024

R Core Team, 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Reed, J.C., Bush, C.A., 2007. Generalized Geologic Map of the United States, Puerto Rico, and the US Virgin Islands. US Geological Survey Digital Spatial Dataset.

Ries, K.G., 2007. 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.

Robson, A., Reed, D., 1999. Flood Estimation Handbook, Volume 3. Institute of Hydrology, Wallingford, UK.

Rustum, R., Adeloye, A.J., 2007. Replacing Outliers and Missing Values from Activated Sludge Data Using Kohonen Self-Organizing Map. Journal of Environmental Engineering 133, 909-916. doi: 10.1061/(ASCE)0733-9372(2007)133:9(909)

Saco, P., Kumar, P., 2000. Coherent modes in multiscale variability of streamflow over the United States. Water Resources Research 36, 1049-1067.

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.), Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Springer, Berlin, Germany, pp. 313-325.

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

Sanborn, S.C., Bledsoe, B.P., 2006. Predicting streamflow regime metrics for ungauged streams in Colorado, Washington, and Oregon. Journal of Hydrology 325, 241-261. doi: 10.1016/j.jhydrol.2005.10.018

Santhi, C., Allen, P.M., Muttiah, R.S., Arnold, J.G., Tuppad, P., 2008. Regional estimation of base flow for the conterminous United States by hydrologic landscape regions. Journal of Hydrology 351, 139-153. doi: 10.1016/j.jhydrol.2007.12.018

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

Sawicz, K., Wagener, T., Sivapalan, M., Troch, P.A., Carrillo, G., 2011. Catchment classification: empirical analysis of hydrologic similarity based on catchment function in the eastern USA. Hydrology and Earth System Sciences Discussions 8, 4495-4534.
doi: 10.5194/hessd-8-4495-2011

Schaller, M.F., Fan, Y., 2009. River basins as groundwater exporters and importers: Implications for water cycle and climate modeling. Journal of Geophysical Research 114.
doi: 10.1029/2008JD010636

Schmidt, C.R., 2008. Effects of irregular topology in spherical self-organizing maps. San Diego State University, San Diego, USA.

Searcy, J.K., 1959. Flow-Duration Curves. US Geological Survey Water-Supply Paper 1542-A, 33 pp.

Sebastiani, P., Perls, T.T., 2008. Complex genetic models, in: Pourret, O., Naïm, P., Marcot, B. (Eds.), Bayesian Networks: A Practical Guide to Applications. John Wiley and Sons, Chichester, UK, pp. 53-72.

Seibert, J., Beven, K., 2009. Gauging the ungauged basin: how many discharge measurements are needed? Hydrology and Earth System Sciences Discussions 6, 2275-2299.

Shim, Y., Chung, J., Choi, I., 2005. A Comparison Study of Cluster Validity Indices Using a Nonhierarchical Clustering Algorithm. Proceedings of the 2005 International Conference on Computational Intelligence for Modelling, Control and Automation. IEEE, New York City, USA.

Simpson, E.H., 1951. The Interpretation of Interaction in Contingency Tables. Journal of the Royal Statistical Society. Series B (Methodological) 13, 238-241.

Sivapalan, M., 2005. Pattern, Process and Function: Elements of a Unified Theory of Hydrology at the Catchment Scale, in: Anderson M.G. (Ed.), Encyclopedia of Hydrological Sciences. John Wiley and Sons, London, UK, pp. 193-219.

Sivapalan, M., Takeuchi, K., Franks, S.W., Gupta, V.K., Karambiri, H., Lakshmi, V., Liang, X., McDonnell, J.J., Mendiondo, E.M., O'Connell, P.E., Oki, T., Pomeroy, J.W., Schertzer, D., Uhlenbrook, S., Zehe, E., 2003. IAHS Decade on Predictions in Ungauged Basins (PUB), 2003-2012: Shaping an exciting future for the hydrological sciences. Hydrological Sciences Journal 48, 857-880. doi: 10.1623/hysj.48.6.857.51421

Skupin, A., 2004. The world of geography: Visualizing a knowledge domain with cartographic means. Proceedings of the National Academy of Sciences of the United States of America 101, 5274-5278. doi: 10.1073/pnas.0307654100

Skupin, A., Agarwal, P., 2008. Introduction: What is a Self-Organizing Map?, in: Agarwal, P., Skupin, A. (Eds.), Self-Organising Maps: Applications in Geographic Information Science. John Wiley and Sons, Chichester, UK, pp. 1-20.

Skupin, A., Esperbé, A., 2011. An alternative map of the United States based on an *n*-dimensional model of geographic space. Journal of Visual Languages and Computing 22, 290-304. doi: 10.1016/j.jvlc.2011.03.004

Skupin, A., Hagelman, R., 2005. Visualizing Demographic Trajectories with Self-Organizing Maps. GeoInformatica 9, 159-179.

Smakhtin, V.U., 2001. Low flow hydrology: a review. Journal of Hydrology 240, 147-186.

Smakhtin, V.Y., Hughes, D.A., Creuse-Naudin, E., 1997. Regionalization of daily flow characteristics in part of the Eastern Cape, South Africa. Hydrological Sciences Journal 42, 919-936. doi: 10.1080/02626669709492088

Snee, R.D., Marquardt, D.W., 1984. Collinearity Diagnostics Depend on the Domain of Prediction, the Model, and the Data. The American Statistician 38, 83-87.

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

Soil Survey Staff, 2014. Gridded Soil Survey Geographic Database. US Department of Agriculture Natural Resources Conservation Service Digital Spatial Dataset.

Srinivas, V.V., Tripathi, S., Rao, A.R., Govindaraju, R.S., 2008. Regional flood frequency analysis by combining self-organizing feature map and fuzzy clustering. Journal of Hydrology 348, 148-166. doi: 10.1016/j.jhydrol.2007.09.046

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

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

Stedinger, J.R., Tasker, G.D., 1985. Regional hydrologic analysis 1. Ordinary, weighted, and generalized least squares compared. Water Resources Research 21, 1421-1432.

Stedinger, J.R., Vogel, R.M., Foufoula-Georgiou, E., 1993. Frequency analysis of extreme events, in: Maidment, D. (Ed.), Handbook of Hydrology. McGraw-Hill, New York City, USA, pp. 18.1-18.66.

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. Journal of Chemical Information and Computer Sciences 43, 1947-1958.

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

Takagi, K., Lin, H.S., 2010. Temporal Dynamics of Soil Moisture Spatial Variability in the Shale Hills Critical Zone Observatory. Vadose Zone Journal 10, 832-842. doi: 10.2136/vzj2010.0134

Toth, E., 2012. Catchment classification based on characterization of streamflow and precipitation time-series. Hydrology and Earth System Sciences Discussions 9, 10805-10828. doi: 10.5194/hessd-9-10805-2012

Troch, P.A., Martinez, G.F., Pauwels, V.R.N., Durcik, M., Sivapalan, M., Harman, C., Brooks, P.D., Gupta, H., Huxman, T., 2009. Climate and vegetation water use efficiency at catchment scales. Hydrological Processes 23, 2409-2414. doi: 10.1002/hyp.7358

Tukey, J.W., 1977. Exploratory Data Analysis. Pearson, Upper Saddle River, USA.

Vesanto, J., Alhoniemi, E., 2000. Clustering of the Self-Organizing Map. IEEE Transactions on Neural Networks 11, 586-600. doi: 10.1109/72.846731

Viola, F., Noto, L.V., Cannarozzo, M., La Loggia, G., 2011. Regional flow duration curves for ungauged sites in Sicily. Hydrology and Earth System Sciences 15, 323-331. doi: 10.5194/hess-15-323-2011

Vogel, R.M., Fennessey, N.M., 1995. Flow Duration Curves II: A Review of Applications in Water Resources Planning. Journal of the American Water Resources Association 31, 1029-1039.

Vogelmann, J.E., Howard, S.M., Yang, L., Larson, C.R., Wylie, B.K., Van Driel, J.N., 2001. Completion of the 1990's National Land Cover Data Set for the conterminous United States. Photogrammetric Engineering and Remote Sensing 67, 650-662.

Wagener, T., Sivapalan, M., Troch, P.A., McGlynn, B.L., Harman, C.J., Gupta, H.V., Kumar, P., Rao, P.S.C., Basu, N.B., Wilson, J.S., 2010. The future of hydrology: An evolving science for a changing world. Water Resources Research 46. doi: 10.1029/2009WR008906

Wagener, T., Wheater, H.S., Gupta, H.V., 2004. Modelling Ungauged Catchments – Regional Procedures, in: Wei, T.K. (Ed.), Rainfall-Runoff Modelling in Gauged and Ungauged Catchments. Imperial College Press, London, UK, pp. 169-240.

Wallner, M., Haberlandt, U., Dietrich, J., 2013. A one-step similarity approach for the regionalization of hydrological model parameters based on Self-Organizing Maps. Journal of Hydrology 494, 59-71. doi: 10.1016/j.jhydrol.2013.04.022

Wan Jaafar, W.Z., Liu, J., Han, D., 2011. Input variable selection for median flood regionalization. Water Resources Research 47. doi: 10.1029/2011WR010436

Wilcox, B.P., Huang, Y., 2010. Woody plant encroachment paradox: Rivers rebound as degraded grasslands convert to woodlands. Geophysical Research Letters 37. doi: 10.1029/2009GL041929

Wolock, D.M., 2003. Base-Flow Index Grid for the Conterminous United States. US Geological Survey Open-File Report 03-263.

Wolock, D.M., Winter, T.C., McMahon, G., 2004. Delineation and Evaluation of Hydrologic-Landscape Regions in the United States Using Geographic Information System Tools and Multivariate Statistical Analyses. Environmental Management 34, S71-S88.
doi: 10.1007/s00267-003-5077-9

Yadav, M., Wagener, T., Gupta, H., 2007. Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins. Advances in Water Resources 30, 1756-1774. doi: 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. Hydrology and Earth System Sciences 16, 4483-4498. doi: 10.5194/hess-16-4483-2012

Ye, W., Bates, B.C., Viney, N.R., Sivapalan, M., Jakeman, A.J., 1997. Performance of conceptual rainfall-runoff models in low-yielding ephemeral catchments. Water Resources Research 33, 153-166. doi: 10.1029/96WR02840

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. Hydrology and Earth System Sciences 16, 4447-4465. doi: 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. Hydrology and Earth System Sciences 15, 2805-2819. doi: 10.5194/hess-15-2805-2011

Zhang, Y., Vaze, J., Chiew, H.S., Li, M., 2015. Comparing flow duration curve and rainfall-runoff modelling for predicting daily runoff in ungauged catchments. Journal of Hydrology 525, 72-86. doi: 10.1016/j.jhydrol.2015.03.043

Zhang, Y., Vaze, J., Chiew, F.H.S., Teng, J., Li, M., 2014. Predicting hydrological signatures in ungauged catchments using spatial interpolation, index model, and rainfall-runoff modelling. Journal of Hydrology 517, 936-948. doi: 10.1016/j.jhydrol.2014.06.032