WATER RESOURCES RESEARCH, VOL. 49, 3756–3769, doi:10.1002/wrcr.20315, 2013

Impact of multicollinearity on small sample hydrologic regression models

Charles N. Kroll[1] and Peter Song[1,2]

Received 4 June 2012; revised 10 May 2013; accepted 15 May 2013; published 21 June 2013.

[1] Often hydrologic regression models are developed with ordinary least squares (OLS) procedures. The use of OLS with highly correlated explanatory variables produces multicollinearity, which creates highly sensitive parameter estimators with inflated variances and improper model selection. It is not clear how to best address multicollinearity in hydrologic regression models. Here a Monte Carlo simulation is developed to compare four techniques to address multicollinearity: OLS, OLS with variance inflation factor screening (VIF), principal component regression (PCR), and partial least squares regression (PLS). The performance of these four techniques was observed for varying sample sizes, correlation coefficients between the explanatory variables, and model error variances consistent with hydrologic regional regression models. The negative effects of multicollinearity are magnified at smaller sample sizes, higher correlations between the variables, and larger model error variances (smaller R^2). The Monte Carlo simulation indicates that if the true model is known, multicollinearity is present, and the estimation and statistical testing of regression parameters are of interest, then PCR or PLS should be employed. If the model is unknown, or if the interest is solely on model predictions, is it recommended that OLS be employed since using more complicated techniques did not produce any improvement in model performance. A leave-one-out cross-validation case study was also performed using low-streamflow data sets from the eastern United States. Results indicate that OLS with stepwise selection generally produces models across study regions with varying levels of multicollinearity that are as good as biased regression techniques such as PCR and PLS.

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

1. Introduction

[2] Applications of regression analyses in the field of water resources are extensive. One of the widest applications of regression in hydrology is its use in developing regional models that relate hydrologic characteristics, such as low-flow and flood-flow statistics, to watershed characteristics. Such models have been developed for most states in the United States [*Riggs*, 1972; *U.S. Geological Survey*, 2010] and have been incorporated into StreamStats, an interactive web-based application to estimate streamflow statistics at ungauged river sites [*Reis et al*., 2008]. Regression has been used to estimate both the mean and variance of annual watershed runoff for all regions of the United States as a function of climate and basin characteristics [*Vogel et al*., 1999]. In addition, regression has been used to estimate sediment loads [*Syvitski and Milliman*, 2007; *Roman et al*., 2012], water quality constituents [*Tasker and Driver*, 1988; *Driver and Troutman*, 1989] such as fecal coliform [*Kelsey et al*., 2004] and pesticides [*Kreuger and Tornqvist*, 1998], urban runoff magnitudes and chemistry [*Gallo et al*., 2012], groundwater levels [*Thomas and Vogel*, 2012], groundwater quality [*Gardner and Vogel*, 2005], and a host of other applications.

[3] A linear regression model can be represented by

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{E} \tag{1}$$

where $\mathbf{Y}$ is a vector of dependent variables, $\mathbf{X}$ is a matrix of explanatory variables (typically augmented by a column of ones), $\boldsymbol{\beta}$ is a vector of model parameters, and $\mathbf{E}$ is a vector of model residuals. Often ordinary least squares (OLS) regression procedures are used to estimate the model parameters by minimizing the sum of squared residual terms. Via the Gauss Markov theorem, for OLS estimators to be the best linear unbiased estimators of the parameters, the residuals have a mean of zero, a constant variance (homoscedastic), and be independent from the explanatory variables and each other [*McElroy*, 1967]. In hydrologic regression analyses, weighted least squares procedures have been applied to address heteroscedasticity of the

[1]Baker Lab, Environmental Resources Engineering, SUNY ESF, Syracuse, New York, USA.
[2]Anchor QEA, LLC, Montvale, New Jersey, USA.

Corresponding author: C. N. Kroll, 402 Baker Lab, Environmental Resources Engineering, SUNY ESF, Syracuse, NY 13210, USA. (cnkroll@esf.edu)

0043-1397/13/10.1002/wrcr.20315

residuals [*Tasker*, 1980], while generalized least squares procedures address the lack of independence [*Stedinger and Tasker*, 1985; *Kroll and Stedinger*, 1998].

[4] To create unique model parameter estimators, the explanatory variables should be linearly independent. A violation of this condition is referred to as multicollinearity. Multicollinearity can yield highly variable parameter estimators, erroneous selection of explanatory variables from a data set, and the inability to understand the precise effects of certain explanatory variables [*Johnston*, 1972; *Greene*, 1990]. This linear dependence exists in many hydrologic regression models because correlated watershed characteristics representing topography, geology, meteorology, and geomorphology are often employed as explanatory variables. For example, watershed area and stream length are typically correlated throughout a region, as are different estimators of watershed slope. This problem is exacerbated by the availability of large geographic information system (GIS)-derived databases of watershed characteristics for model development [*Kroll et al*., 2004; *Falcone et al*., 2010]. The use of standard model selection criteria, such as stepwise selection, may also be adversely impacted by multicollinearity. *Graham* [2003] has shown that with increasing multicollinearity, significant explanatory variables are more vulnerable to erroneous variable exclusion.

[5] It is currently unclear as to the best method to employ when confronted with multicollinearity in hydrologic regression modeling. Common techniques include using variance inflation factors (VIFs) to screen for multicollinearity [*Mansfield and Helms*, 1982; *Kroll et al*., 2004; *O'brien*, 2007], transforming the original explanatory variables into new uncorrelated variables such as principal component regression (PCR) [*Haan and Allen*, 1972; *Jolliffe*, 1986], adding a constant to the diagonal of the cross product matrix ($\mathbf{X}'\mathbf{X}$) such as ridge regression (RR) [*Hoerl and Kennard*, 2000], using an alternative model selection criterion such as Mallow's Cp [*Laaha and Bloschl*, 2007], or completely ignoring the problem and performing OLS. Outside of the hydrologic literature other techniques have been explored that take advantage of the correlation between the dependent variables and the response variable, such as partial least squares regression (PLS) [*Wold et al*., 2001; *Tootle et al*., 2007].

[6] While some researchers have examined the impact of multicollinearity when sample sizes are large [*Frank and Friedman*, 1993; *Grewal et al*., 2004], only limited studies have addressed the issue of multicollinearity on small sample regression [*Mason and Perreault*, 1991; *Kiers and Smilde*, 2007], a common problem in hydrology. *Mason and Perreault* [1991] found that the adverse impact of multicollinearity is magnified at a sample size of 30 observations and poor overall model fit. When comparing the model development techniques with highly correlated explanatory variables, *Kiers and Smilde* [2007] found that PCR and PLS produce the most accurate parameter estimators at a sample size of 50 observations, while OLS produces the best model predictions for sample sizes of 10, 20, and 50 observations. For asymptotic calculations where the number of observations is large, *Mela and Kopalle* [2002] show that the prediction differences between PCR and PLS is small. The results from these previous studies are not directly applicable to the hydrologic regional regression problem because they assume that the true model is known. The true model is seldom known in practice, and even when the true model is known there is often uncertainty regarding how best to estimate the model parameters.

[7] *Frank and Friedman* [1993] conducted a Monte Carlo simulation with 50 observations to compare the prediction performances of various regression model building techniques. Their study found that RR performs the best, closely followed by PCR and PLS, while OLS performed worst when comparing the average squared prediction error. While *Frank and Friedman* [1993] provide a detailed description of these techniques, their simulation was for a very limited case, and the trade-offs between these estimation techniques were not fully explored. Note that we do not compare RR in this analysis for a variety of reasons. There is no definitive technique for selecting the proper constant to add to the $\mathbf{X}'\mathbf{X}$ matrix in RR. As the RR constant increases, the variance of the parameter estimators decreases, and thus, one avoids the inflated variances of multicollinearity. Subsequently, as the RR constant increases, the bias of the parameter estimators also increases. The RR constant is often chosen to balance the variance and bias of the parameter estimators, or a cross-validation is employed. Since PCR and PLS performed similarly to RR in *Frank and Friedman*'s [1993] limited analysis, RR was not considered here.

[8] Employing a Monte Carlo simulation, we extend Frank and Friedman's analysis by comparing OLS, VIF, PCR, and PLS model development techniques for various sample sizes and at higher degrees of collinearity. We also examine how these techniques are impacted by stepwise (forward and backward) variable selection. Techniques are compared based on their ability to estimate model parameters, the predictions from the developed model, and the probability of selecting the correct model. An example for low-streamflow prediction in the eastern United States is also presented which supports the results of the Monte Carlo simulation.

2. Model Development Techniques

[9] The four model development techniques employed in this experiment are briefly described later and are summarized in Table 1.

2.1. Ordinary Least Squares (OLS)

[10] For OLS, the model parameters are determined by minimizing the sum of squared residual terms. In this analysis, models are developed using standard stepwise variable selection procedures based on an F test with an $\alpha = 0.05$ [*Draper and Smith*, 1981]. Use of stepwise variable selection (both forward and backward) with an F test to determine the model explanatory variables has been criticized for a number of reasons, including biased parameter estimators, incorrect p values, and an upwardly biased coefficient of determination [*Pope and Webster*, 1972; *Rencher and Pun*, 1980; *Arditi*, 1989]. In spite of this, the F test was employed in this analysis due to its common use in practice.

2.2. Variance Inflation Factor With OLS (VIF)

[11] The VIF statistic is commonly employed to screen for multicollinearity [*Greene*, 1990; *Johnston*, 1972; *Kroll*

Table 1. Summary of Model Development Techniques

Model Development Technique	Description
OLS	Standard OLS regression with original watershed characteristics as explanatory variables. Employ stepwise variable selection to develop model.
OLS with variance inflation factor screening (VIF)	Same as OLS except add a screening procedure to sequentially remove highly correlated watershed variables from model.
PCR	Transform watershed characteristics into independent components. Use a two-step stepwise variable selection procedure to develop model.
PLS	Same as PCR except transform watershed characteristics and response variable into independent components.

et al., 2004]. Each explanatory variable is regressed against the other remaining explanatory variables, and the VIF is calculated as

$$\mathrm{VIF} = \frac{1}{1 - R^2} \tag{2}$$

where R^2 is the regression model coefficient of determination [*Rawlings et al*., 1988]. A VIF greater than 10 is a common threshold for detecting severe multicollinearity [*Chatterjee and Price*, 1990; *O'brien*, 2007]. In our analysis, highly correlated explanatory variables are sequentially removed if the VIF > 10, and the model with the lowest sum of squared errors is kept as the best model. Similar to OLS, standard stepwise regression procedures are employed with this method, which will be referred to as VIF throughout this paper.

2.3. Principal Component Regression (PCR)

[12] In PCR, the original correlated explanatory variables are linearly transformed into a new set of uncorrelated variables known as principal components (PCs). This linear transformation involves weighting each explanatory variable by the eigenvectors of the explanatory variable correlation or covariance matrix. Each PC explains a fraction of the total variance within the data set. In this study, weights based on the correlation matrix were used because the covariance matrix can be sensitive to the units of measurement, which vary across different watershed characteristics [*Jolliffe*, 1986]. The PCs are defined as

$$\mathbf{T} = \mathbf{XW} + \mathbf{E}_2 \tag{3}$$

where $\mathbf{T}$ is a matrix of PCs (scores), $\mathbf{X}$ is the original explanatory variable matrix, $\mathbf{W}$ are the weights (sometimes referred to as the loadings) found as eigenvectors of the correlation matrix, and $\mathbf{E}_2$ is a vector of residuals. The PCs ($\mathbf{T}$) are then used to model the response variable $\mathbf{Y}$

$$\mathbf{Y} = \mathbf{T}\beta' + \mathbf{E}_{\mathrm{pc}} \tag{4}$$

where $\boldsymbol{\beta}'$ are the transformed parameter estimators and $\mathbf{E}_{\mathrm{pc}}$ are the model residuals.

[13] A two-step, stepwise variable selection technique is employed to determine the explanatory variables in the PCR model. In the first step, a new potential explanatory variable is either entered into or removed from the current model, and a set of PCs are calculated for the new explanatory variable data set. In the second step stepwise variable selection procedures, based on an F test, are performed using the PCs. The selected PCs are then used to develop the regression model, and this model is then transformed back to the original explanatory variable space ($\boldsymbol{\beta} = \mathbf{W}\boldsymbol{\beta}'$) to create the final model. If all of the PCs are retained, then the final model is identical to OLS. One common use of PCR is to reduce the dimensionality of the regression model by including fewer PCs than the original explanatory variables.

2.4. Partial Least Squares Regression (PLS)

[14] OLS is based upon minimizing the sum of squared differences between the observed $\mathbf{Y}$ and predicted $\mathbf{Y}$, calculated from $\mathbf{X}\boldsymbol{\beta}$. PCR is based upon maximizing the variance of the linear combinations of $\mathbf{X}$. PLS is based upon maximizing the covariance between the $\mathbf{Y}$ and the linear combinations of $\mathbf{X}$ [*Helland and Almoy*, 1994].

[15] Similar to PCR, in PLS the score matrix is found by multiplying the $\mathbf{X}$ matrix by a weight matrix (equation (3)). While the $\mathbf{W}$ matrix in PCR is computed to reflect the correlation structure between the explanatory variables, the $\mathbf{W}$ matrix in PLS is computed to represent the covariance structure of the response and explanatory variables. One method for computing the $\mathbf{W}$ matrix is using the Non-linear Iterative PArtial Least Squares (NIPALS) algorithm [*Geladi and Kowlaski*, 1986]. $\mathbf{Y}$ is decomposed to compute $\mathbf{Q}$ by performing OLS of $\mathbf{Y}$ on $\mathbf{T}$:

$$\mathbf{Y} = \mathbf{TQ} + \mathbf{E}_{\mathrm{pls}}, \tag{5}$$

where $\mathbf{T}$ is the same score matrix previously found for $\mathbf{X}$ (equation (3)), $\mathbf{Y}$ is the original response (dependent) variable, $\mathbf{E}_{\mathrm{pls}}$ is the residual vector, and $\mathbf{Q}$ are the loadings. Once $\mathbf{Q}$ has been found, PLS employs the following prediction model:

$$\mathbf{Y} = \mathbf{XWQ} + \mathbf{E}_{\mathrm{pls}}, \tag{6}$$

where the regression coefficients for PLS are computed as $\mathbf{WQ}$. As with PLS, a two-step stepwise variable selection technique is employed to determine the explanatory variables in the model. PLS can be geometrically represented as a plane bounded by components and projected on a slope with respect to the original coordinate axes [*Wold et al*., 2001]. The components describe the data set variability, while the slope defines the best correlation of $\mathbf{X}$ with $\mathbf{Y}$.

3. Monte Carlo Simulation

3.1. Experimental Design

[16] For the Monte Carlo simulation, the following model is examined:

$$\mathbf{y} = b_0 + b_1\mathbf{x}_1 + b_2\mathbf{x}_2 + b_3\mathbf{x}_3 + \varepsilon, \tag{7}$$

where $\mathbf{y}$ is a vector of the dependent variable, $\mathbf{x}_i$ is a vector of the ith explanatory variable, b_i is the ith model parameter, and ε is the model residual. The explanatory variables were randomly generated following a standard normal distribution ($N(0,1)$) where $\mathbf{x}_2$ and $\mathbf{x}_3$ are correlated with a Pearson

correlation coefficient ρ, and $\mathbf{x}_1$ is independent of $\mathbf{x}_2$ and $\mathbf{x}_3$. While a different model could be used to generate $\mathbf{x}_i$, in hydrologic regression models these variables often represent the log of watershed characteristics (such as in a log-linear regression model), which often appear normally distributed in a region. In addition a fourth random independent potential explanatory ($\mathbf{x}_4$) is also generated and is available for variable selection. The true value for all parameters (b_i) is set to 1. The model residual, ε, was randomly generated from a normal distribution with a mean of zero and constant variance of σ_ε^2. Changing σ_ε^2 varies the model's coefficient of determination (R^2). The dependent variable, $\mathbf{y}$, was generated using the true values of the parameters, and random realizations of the explanatory variables and model error.

[17] In the Monte Carlo simulation, the performance of the model development techniques was observed for changing simulation parameters. The sample size was set to values of $n = 20$, 50, 100, and 1000 observations, the Pearson correlation coefficient between $\mathbf{x}_2$ and $\mathbf{x}_3$ was set to values of $\rho = 0.90$, 0.95, and 0.99, and the model error variance was set to value of $\sigma_\varepsilon^2 = 0.15$ and 1.5. The model error variances produced models with average R^2 values of 0.75 $\left(\sigma_\varepsilon^2 = 1.5\right)$ and 0.95 $\left(\sigma_\varepsilon^2 = 0.15\right)$ when $n = 1000$. For the Monte Carlo simulation, the number of simulation replications was set to $m = 100{,}000$.

[18] In this experiment two simulations are performed: one with the true model known, and one where the true model is unknown and stepwise procedures are employed to develop the model. When the true model is known (i.e., $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$ are in the model), only the first two of the three PCs for PCR and PLS are retained in the model. This reduces the dimensionality of the problem, as keeping all three PCs would yield the same result as OLS. For the situation where the true model is unknown, it is important to note that this assumes that the form of the true model is known, and that the true model is included as a subset of the potential explanatory variables. In practice, one would expect that neither situation might actually be true.

3.2. Performance Metrics

[19] A number of statistics were calculated to analyze the performance of the model development techniques. Model performance was assessed based on three results: (1) properties of the parameter estimators, (2) model predictions, and (3) probability of selecting the correct model. Properties of the parameter estimators were based on an estimator's bias, mean-square error (MSE), and variance, which were calculated as

$$\text{Bias}(\hat{b}_j) = \sum_{i=1}^{m} \frac{\hat{b}_{ji} - 1}{m} \tag{8}$$

$$\text{MSE}\left(\hat{b}_j\right) = \sum_{i=1}^{m} \frac{\left(\hat{b}_{ij} - 1\right)^2}{m} \tag{9}$$

$$\text{Var}\left(\hat{b}_j\right) = \sum_{i=1}^{m} \frac{\left(\hat{b}_{ij} - \overline{b_j}\right)^2}{m-1}, \tag{10}$$

where $\hat{b}_{ji}$ is the ith estimate of jth parameter estimator, and $\overline{b_j}$ is the average of $\overline{b_{ji}}$.

[20] To determine the predictive ability of the model development techniques, an analysis was performed to examine how well each technique predicted the median of the data set as well as an extrapolated value outside the range of the data set. To examine the median, a leave-one-out cross-validation was performed where the median value of $\mathbf{y}$ was removed from the data set, the remaining data set of size $n - 1$ was employed to develop regression models with each of the model development techniques, and then the resulting model was used to estimate the removed value of $\mathbf{y}$. To examine how well each technique estimates an extrapolated value of $\mathbf{y}$, a new value of $\mathbf{y}$ was generated for each data set using equation (7) by generating random realizations of the explanatory variables ($\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$) from a $N(3,1)$ distribution and the model residuals from a $N(0, \sigma_\varepsilon^2)$ distribution, and setting $b_i = 1$ for all i. It is important to note that this assumes the true model is valid outside the range of measurements, which may not be true in practice. Also note that predicting the largest value of original data set via a leave-one-out cross-validation would create a bias for all methods since the largest value is typically associated with an observation generated with a large value of ε. To examine model predictions, three performance metrics were calculated:

$$\text{Bias}(\hat{y}) = \sum_{i=1}^{m} \frac{\hat{y}_i - y_i}{m} \tag{11}$$

$$\text{MSE}(\hat{y}) = \sum_{i=1}^{m} \frac{(\hat{y}_i - y_i)^2}{m} \tag{12}$$

$$\text{Var}(\hat{y}) = \sum_{i=1}^{m} \frac{(\hat{y}_i - \overline{y})^2}{m-1}, \tag{13}$$

where $\hat{y}_i$ is the ith prediction of the observed medium or extrapolated value of $\mathbf{y}$.

[21] The frequency of accepting or rejecting explanatory variables is examined to determine how well the model development techniques select the correct model when multicollinearity is present. One would expect that when ρ is high, the information in $\mathbf{x}_2$ and $\mathbf{x}_3$ is nearly identical, and thus, the model performance would be similar if $\mathbf{x}_2$, $\mathbf{x}_3$, or both $\mathbf{x}_2$ and $\mathbf{x}_3$ were selected, resulting in a situation where the incorrect model is more frequently chosen.

3.3. Results

[22] A Monte Carlo simulation was performed with the four model development techniques with varying correlation coefficients, sample sizes, and model error variances. The results are broken into three subsections: (1) properties of the parameter estimators, (2) model predictions, and (3) probability of selecting the correct models. For subsections (1) and (2), we examine results for each of the model development techniques when the true model is known and unknown. When the true model is known, equation (7) is employed for OLS, PCR, and PLS, and if the VIF > 10, then either $\mathbf{x}_2$ or $\mathbf{x}_3$ is included for VIF. For PCR and PLS, only the first two components are retained in the regression equations. When the true model is unknown, each model

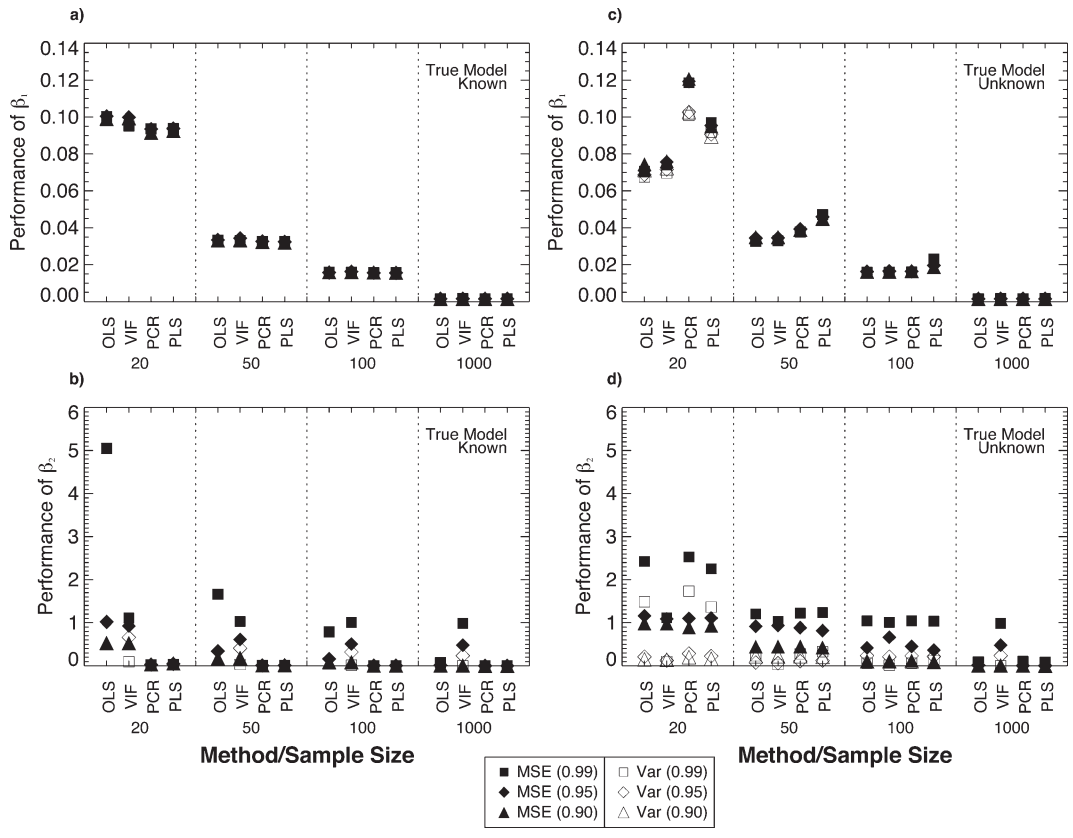

Figure 1. MSE and variance (Var) of $\boldsymbol{\beta}_1$ when the true model is (a) known and (c) unknown and $\boldsymbol{\beta}_2$ when the true model is (b) known and (d) unknown. Correlation coefficient is indicated in parentheses.

development technique selects the explanatory variables using stepwise procedures, with VIF screening for situations with a VIF > 10.

3.3.1. Properties of the Parameter Estimators

3.3.1.1. True Model Known

[23] Figure 1a presents the MSE and variance of the uncorrelated parameter estimator b_1, when $\sigma_\varepsilon^2 = 1.5$ and the true model is known. Higher values of MSE and variance indicate poorer performance of the model development technique parameter estimators. Since the MSE $\approx$ Variance + (Bias)^2, differences in the MSE and variance are an indirect measure of bias. All four model development techniques perform well when the true model is known, with MSE values of b_1 close to zero (<0.1). Slightly lower MSE values of PCR and PLS were observed for $n = 20$; however, the difference in the MSE compared to the other model development techniques was relatively small. The MSE values for all techniques decreases as the sample size increases. Increasing the correlation coefficient between $\mathbf{x}_2$ and $\mathbf{x}_3$ generally did not affect the performance of the b_1 parameter estimator. Results for $\sigma_\varepsilon^2 = 0.15$ (not shown here) followed similar patterns but were smaller in magnitude.

[24] Figure 1b presents the MSE and variance of the correlated parameter estimator b_2 when $\sigma_\varepsilon^2 = 1.5$ and the true model is known. The MSE and variance of the OLS b_2 estimators increases as the sample size decreases, the correlation coefficient between $\mathbf{x}_2$ and $\mathbf{x}_3$ increases, and the model error variance increases. At the highest correlation, VIF b_2 estimators consistently produces MSE values close to 1 because the b_2 estimator is compensating for the absence of b_3 in the model (which is screened since the VIF > 10), thus doubling b_2 from a value of 1 to $\sim$2. Only VIF produced biased parameter estimators. PCR and PLS produce the best b_2 estimators for all correlations and samples sizes, with MSE values close to zero. PCR and PLS estimators produced nearly identical results. Performance metrics for b_3 were identical to b_2, while the performance for b_1 was better than for b_2. These results are consistent with reported results that when multicollinearity is present in a model, OLS produces parameter estimators with inflated variances, and that the use of PCR or PLS can reduce these parameter estimator variances [*Kiers and Smilde*, 2007; *Mela and Kopalle*, 2002; *Frank and Friedman*, 1993].

3.3.1.2. True Model Unknown (*Stepwise Selection*)

[25] Figure 1c shows that when the true model is unknown, OLS and VIF b_1 estimators have smaller MSE values than PCR and PLS estimators. This is especially true when the sample size is small, where PCR b_1 estimators have the largest MSE. When the sample size increases, the difference between the OLS, VIF, PCR, and PLS b_1 estimators narrows. Interestingly, OLS and VIF b_1 estimators produce smaller MSE values when the true model is unknown than when the true model is known. When the true model is unknown, OLS and VIF can choose the explanatory variables that can best minimize the variance of the model residuals and thus minimize the variance of the parameter estimators. The variance of the b_1 estimators for

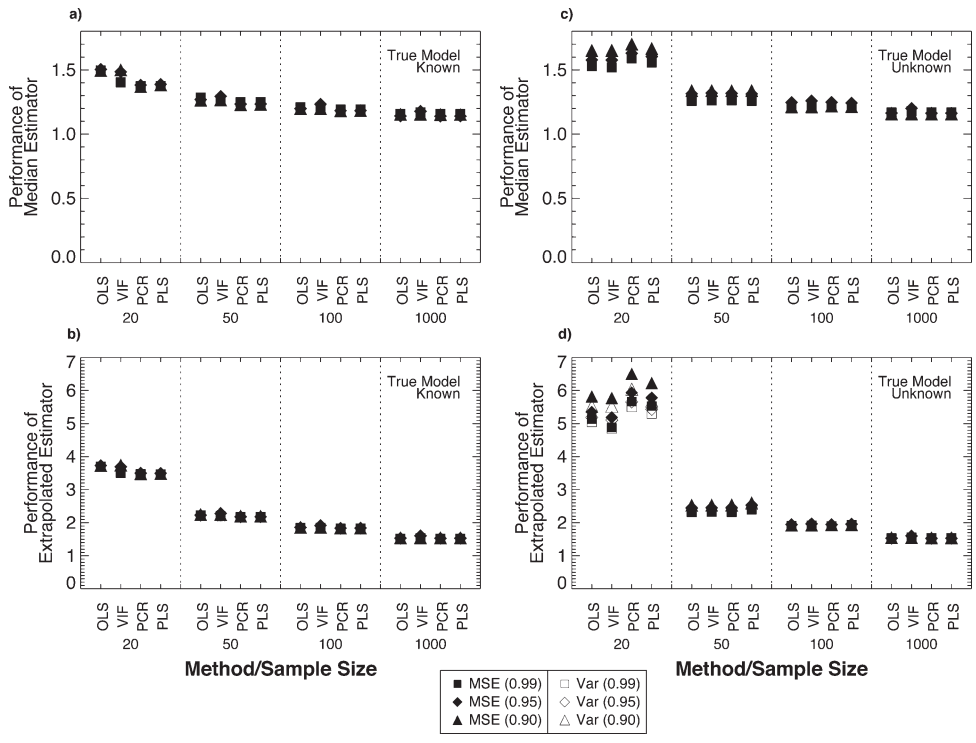

Figure 2. MSE and variance (Var) of median estimator when the true model is (a) known and (c) unknown and extrapolated estimator when the true model is (b) known and (d) unknown. Correlation coefficient is indicated in parentheses.

PCR and PLS increases when the techniques choose the model using stepwise selection. Regardless, all four model development techniques generally perform well when estimating uncorrelated explanatory variables using stepwise regression.

[26] The results of the Monte Carlo simulation with stepwise variable selection for b_2 when $\sigma_\varepsilon^2 = 1.5$ is summarized in Figure 1d. At the highest correlation, the MSE of the VIF b_2 estimators was again approximately one for all sample sizes. It can be seen from Figure 1d that VIF b_2 estimators perform better than the other techniques when the correlation is high and the sample size is low. The adverse impacts of multicollinearity decrease as the sample size increases for OLS, PCR, and PLS. OLS, PCR, and PLS all had similar results. Interestingly, when OLS is allowed to choose the model using stepwise selection (i.e., Figure 1d versus Figure 1b), the variance of the parameter estimator decreases, indicating that only one of the two correlated variables is entering the model. This observation is supported in section 3.3.3.

3.3.2. Model Predictions

3.3.2.1. True Model Known

[27] Figure 2a presents the MSE and variance of estimators of the median observation when $\sigma_\varepsilon^2 = 1.5$ and the true model is known. Although PCR and PLS produced better parameter estimators when the true model is known, these model development techniques produced only slight improvements over OLS in predicting the median observation at 20 observations, and nearly no improvement at higher sample sizes. As sample size increased, MSE and variance decreased, though the magnitude of this decrease was relatively small (<10%) when moving from 50 to 1000 observations. Except for VIF when the sample size is 20, increasing the correlation between the variables had no effect on the MSE and variance of the median estimators. The MSE and variance of the estimators were nearly identical, indicating no bias from any of the techniques when estimating the median. While not shown here, the MSE and variance for when $\sigma_\varepsilon^2 = 0.15$ are similar to when $\sigma_\varepsilon^2 = 1.5$, though smaller in magnitude. This is also true for other results presented in this section.

[28] As the presence of multicollinearity inflates the variance of the parameter estimators and the model predictions, of concern is the impact of multicollinearity on predicting observations where one must extrapolate outside of the range of observations used to develop the model. Figure 2b presents the MSE and variance of the extrapolated observation when $\sigma_\varepsilon^2 = 1.5$ and the true model is known. While the pattern of MSE and variance of the extrapolated estimators is similar to that of the median estimators, the MSE and variance are over twice a big for the extrapolated estimators when the sample size is 20. This difference decreases substantially as the sample size increases. Again the OLS estimators have a MSE and variance that is slightly larger than PCR and PLS when the sample size is 20, but for all other sample sizes the MSE and variance of the OLS estimators are similar to those from PCR and PLS. As seen with the median estimators, the extrapolated estimators exhibited no bias, and changing the correlation between $\mathbf{x}_2$ and $\mathbf{x}_3$ had no impact on the MSE and variance.

[29] It is important to note that while the variance of the extrapolated estimators converges to 1.5 as the sample size increases (Figures 2b and 2d), the variance of the median

estimators converges to a value less than 1.5 (Figures 2a and 2c). The reason is that when large residuals are used to generate observations in equation (7), those observations generally are not the median of the data set. Thus, the variance of the residuals for the median observations and the variance of the median estimators are less than 1.5. For the extrapolated estimators, the observations and residuals are generated independently of the data set used to fit the model, and thus the variance of these estimators converges to 1.5, the variance of the residuals.

3.3.2.2. True Model Unknown

[30] Figure 2c presents the MSE and variance of estimators of the median observation when $\sigma_\varepsilon^2 = 1.5$, and stepwise selection is employed to develop the models. Unlike when the true model is known, when one builds a model with stepwise regression and a sample size of 20, OLS and VIF perform slightly better than PCR and PLS, though these differences diminish as the sample size increases. At a sample size of 20, the MSE and variance of the median estimators were slightly larger when the true model was unknown compared to when it was known, though at larger sample sizes these differences are small. As sample size increased, MSE and variance decreased, though the magnitude of this decrease was relatively small for sample sizes of 50, 100, and 1000. When the sample size was 20 or 50, increasing the correlation between the variables produced a decrease in the MSE and variance of the median estimators, a result not observed when the true model was known. The reason for this result is that at lower correlations, the incorrect model (with just one explanatory variable) is chosen nearly as often as at higher correlations, and the model performs worse because more model information is missing due to the variable being left out of the model. The MSE and variance of the median estimators were nearly identical, indicating no bias from any of the techniques when estimating the median.

[31] Figure 2d presents the MSE and variance of the extrapolated observation when $\sigma_\varepsilon^2 = 1.5$ and the true model is unknown. While the pattern of MSE and variance of the extrapolated estimators is again similar to that of the median estimators, the MSE and variance are much larger for the extrapolated estimators when the sample size is 20. This difference decreases substantially as the sample size increases. Again the OLS and VIF estimators have a MSE and variance that are slightly smaller than PCR and PLS when the sample size is 20, but for all other sample sizes the MSE and variance of the OLS estimators are similar to those from PCR and PLS. As seen with the median estimators, the extrapolated estimators exhibited no bias except for when the sample size is 20, where a slight bias was observed for all methods. Changing the correlation between $\mathbf{x}_2$ and $\mathbf{x}_3$ had no impact on the MSE and variance of the extrapolated estimators unless the sample size was 20.

3.3.3. Probability of *S*electing the *C*orrect *M*odel

[32] Also of interest is how often the correct set of explanatory variables is chosen when the true model is unknown. Table 2 contains the frequency at which the true model is chosen by each model development technique for different sample sizes and $\rho = 0.90$ and 0.99. The correct model is more frequently selected when the sample size increases, especially when $\rho = 0.90$, which is just below the threshold indicating a high level of multicollinearity. At higher levels of multicollinearity, all model development techniques have a lower frequency of choosing the correct model for all sample sizes, choosing $\mathbf{x}_2$ or $\mathbf{x}_3$ but not both variables. As expected, VIF, which screen for multicollinearity, always chooses the incorrect model when $\rho = 0.99$. The other methods also choose the incorrect model frequently because the first variable entered into the model explains most of the information in the second variable, and thus adding the second variable does not provide a large decrease in the sum of squared errors so that the second variable is rejected from the model. A higher frequency of choosing the correct model, though, did not translate into better model predictions, as shown in section 3.3.2. At higher levels of multicollinearity, exclusion of one variable did not have large impact on model predictions since much of the information in the excluded variable was contained within the highly correlated variable that was included in the model.

Table 2. Probability of Selecting the Correct Model

Sample Size	Frequencies of Selected Variables	Model Development Techniques			
n	x_1, x_2, x_3	OLS	VIF	PCR	PLS
Correlation Coefficient = 0.90					
20		1.2%	1.2%	2.6%	1.8%
50		28.1%	28.0%	28.3%	28.6%
100		80.1%	80.1%	71.1%	75.7%
1000		95.0%	95.0%	95.0%	95.5%
Correlation Coefficient = 0.99					
20		1.3%	0.0%	1.8%	1.2%
50		0.6%	0.0%	0.9%	1.1%
100		0.2%	0.0%	0.2%	1.3%
1000		85.7%	0.0%	75.8%	80.0%

4. Low-Streamflow Regional Regression Analysis

[33] Low-streamflow statistics are needed for a variety of water quality and water quantity management purposes. At ungauged river sites, a common technique to estimate low-streamflow statistics is to employ a regional regression model which has been developed between low-streamflow statistics and watershed characteristics at sites in the region of interest [*Riggs*, 1972; *Vogel and Kroll*, 1992]. Here an analysis is performed to compare the four regression model development techniques at 130 gauged river sites in the Blue Ridge, Piedmont, and Valley and Ridge physiographic provinces in the eastern United States (Figure 3). These three physiographic provinces were the focus of a U.S. Geological Survey (USGS) Regional Aquifer-System Analysis study, and thus, information is available regarding the hydrogeologic (and thus, low flow) characteristics within this region [*Sun and Weeks*, 1991]. This study area was chosen because low-streamflow regression models have been shown to perform poorly in this region of the United States [*Kroll et al*., 2004].

4.1. Study Sites

[34] Using the guidance provided by *Falcone et al*. [2010], the sites employed in this study were identified as having minimal anthropogenic disturbances based on three

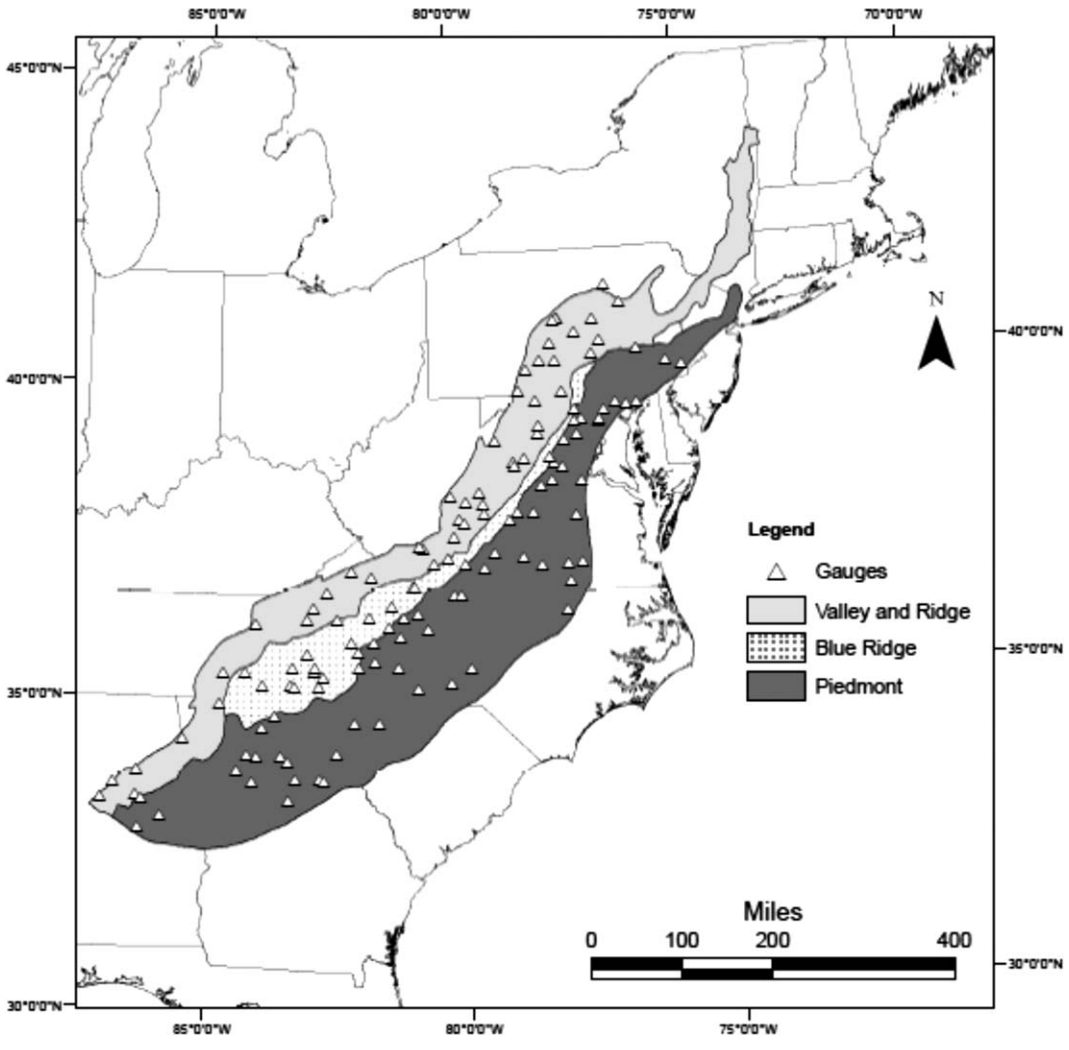

Figure 3. Gauging stations employed in low-streamflow regional regression analysis in the Blue Ridge, Piedmont, and Valley and Ridge physiographic regions.

criteria: a GIS-based index to quantify anthropogenic modification in a watershed, visual inspection of every site and its watershed using recent high-resolution imagery, and anthropogenic influences as described in the State USGS Annual Water Data Reports. The drainage area of these sites ranged from 25 to 3000 km^2. The low-streamflow statistic of interest in this analysis was the 7 day, 10 year low streamflow ($Q_{7,10}$), a common design statistic [*Smakhtin*, 2001]. At site, estimates of the $Q_{7,10}$ were obtained by a frequency analysis with a log-Pearson type 3 distribution whose parameters were estimated by method of moments [*Stedinger et al*., 1992]. Since a log-linear regression model was employed in this study, sites where the $Q_{7,10}$ was estimated as zero were removed from the analysis. A Tobit model could be used to include sites where $Q_{7,10}$ is estimated as zero [*Kroll and Stedinger*, 1999]. In addition, only sites that were included in the watershed characteristic databases developed by *Kroll et al*. [2004] and *Falcone et al*. [2010] were included so that information from both databases could be utilized in this study. Both of these databases were developed using spatially explicit raster data sets and automated GIS processing and contain watershed characteristics that are highly correlated. It is hypothesized that addressing multicollinearity within these regions may improve the low-streamflow regression models. Table 3 presents the 78 watershed characteristics considered in this study.

4.2. Low-Flow Regression Model

[35] The regional regression model employed in this analysis had the form:

$$Q_{7,10} = e^{b_o} x_1^{b_1} x_2^{b_2} x_3^{b_3} \ldots e^{\varepsilon} \tag{14}$$

where $\mathbf{x}_i$ are the model's explanatory variables (watershed characteristics), b_i are model parameters to be estimated, and ε is the model residual. By taking the logarithm of both sides of equation (14), a log-linear model is obtained. In this study, sites were regionalized based on both state boundaries and physiographic regions. State boundaries are sometimes used as regional boundaries because many watershed characteristics are independently developed and stored in state-based GIS clearinghouses [*Kroll et al*., 2004], and often regression models are developed by and

Table 3. Watershed Characteristics Included in Low Streamflow Analysis

Data Type/Source	Name	Variable Description
USGS	DRN_SQKM	Watershed drainage area, sq km
	SNOW_PCT_PRECIP	Snow percent of total precipitation estimate, mean for period 1901–2000
	STREAMS_KM_SQ_KM	Stream density, km of streams per watershed sq km, from NHD 100k streams
	BFI	Base Flow Index (BFI), The BFI is a ratio of base flow to total streamflow
	ELEV_MEAN_M_BASIN	Mean watershed elevation (meters) from 100 m National Elevation Dataset
	ELEV_MAX_M_BASIN	Maximum watershed elevation (meters) from 100 m National Elevation Dataset
	ELEV_MIN_M_BASIN	Minimum watershed elevation (meters) from 100 m National Elevation Dataset (may include sinks)
	ELEV_MEDIAN_M_BASIN	Median watershed elevation (meters) from 100 m National Elevation Dataset
	ELEV_STD_M_BASIN	Standard deviation of elevation (meters) across the watershed from 100m National Elevation Dataset
	RRMEAN	Dimensionless elevation – relief ratio, calculated as (ELEV_MEAN – ELEV_MIN)/(ELEV_MAX – ELEV_MIN)
	RRMEDIAN	Dimensionless elevation – relief ratio, calculated as (ELEV_MEDIAN – ELEV_MIN)/(ELEV_MAX – ELEV_MIN)
	SLOPE_PCT	Mean watershed slope, percent. Derived from 100 m resolution National Elevation Dataset
Precipitation-elevation Regressions on Independent Slopes Model (PRISM)	PPTAVG_BASIN	Mean annual precip (cm) for the watershed, from 800 m PRISM data
	PPTAVG_SITE	Mean annual precip (cm) at the gauge location, from 800 m PRISM data. 30 years period of record 1971–2000
	PPTMAX_BASIN	Watershed average of maximum monthly precipitation (cm) from 2 km PRISM, derived from 30 years of record (1961–1990)
	PPTMIN_BASIN	Watershed average of minimum monthly precipitation (cm) from 2 km PRISM, derived from 30 years of record (1961–1990)
	PPTMAX_SITE	Site average of maximum monthly precipitation (cm) from 2 km PRISM, derived from 30 years of record (1961–1990)
	PPTMIN_SITE	Site average of minimum monthly precipitation (cm) from 2 km PRISM, derived from 30 years of record (1961–1990)
	T_AVG_BASIN	Average annual air temperature for the watershed (°C), from 800 m PRISM data. 30 years period of record 1971–2000
	T_AVG_SITE	Average annual air temperature at the gauge location(°C), from 2 km PRISM data. 30 years period of record 1971–2000
	T_MAX_BASIN	Watershed average of maximum monthly air temperature (°C) from 800 m PRISM, derived from 30 years of record (1971–2000)
	T_MAXSTD_BASIN	Standard deviation of maximum monthly air temperature (°C) from 800 m PRISM, derived from 30 years of record (1971–2000)
	T_MIN_BASIN	Watershed average of minimum monthly air temperature (°C) from 800 m PRISM, derived from 30 years of record (1971–2000)
	T_MINSTD_BASIN	Standard deviation of minimum monthly air temperature (°C) from 800 m PRISM, derived from 30 years of record (1971–2000)
	T_MAX_SITE	Gauge location maximum monthly air temperature (°C) from 800 m PRISM, derived from 30 years of record (1971–2000)
	T_MIN_SITE	Gauge location minimum monthly air temperature (°C) from 800 m PRISM, derived from 30 years of record (1971–2000)
	RH_BASIN	Watershed average relative humidity (percent), from 2 km PRISM, derived from 30 years of record (1961–1990)
	RH_SITE	Site average relative humidity (percent), from 2 km PRISM, derived from 30 years of record (1961–1990)
	FST32F_BASIN	Watershed average of mean day of the year (1–365) of first freeze, derived from 30 years of record (1961–1990), 2 km PRISM
	LST32F_BASIN	Watershed average of mean day of the year (1–365) of last freeze, derived from 30 years of record (1961–1990), 2 km PRISM
	FST32F_SITE	Site average of mean day of the year (1–365) of first freeze, derived from 30 years of record (1961–1990), 2 km PRISM
	LST32F_SITE	Site average of mean day of the year (1–365) of last freeze, derived from 30 years of record (1961–1990), 2 km PRISM
	WD_BASIN	Watershed average of annual number of days (days) of measurable precipitation, derived from (1961–1990) 2 km PRISM
	WDMAX_BASIN	Watershed average of monthly maximum number of days of measurable precipitation, derived from(1961–1990) 2 m PRISM
	WDMIN_BASIN	Watershed average of monthly minimum number of days (days) of measurable precipitation, derived from (1961–1990) 2 m PRISM
	WD_SITE	Site average of annual number of days (days) of measurable precipitation, derived from (1961–1990) 2 km PRISM
	WDMAX_SITE	Site average of monthly maximum number of days (days) of measurable precipitation, derived from (1961–1990) 2 km PRISM
	WDMIN_SITE	Site average of monthly minimum number of days (days) of measurable precipitation, derived from (1961–1990) 2 km PRISM
	PET	Mean-annual potential evapotranspiration (PET), estimated using the Hamon (1961) equation

Table 3. (continued)

Data Type/Source	Name	Variable Description
	PRECIP_SEAS_IND	Index of annual precipitation falling seasonally (1) or spread over the year (0). Based on monthly precip values from (1971–2000) PRISM
	JAN_PPT7100_CM	Mean January precip (cm) for the watershed, from 800 m PRISM data. 30 years period of record 1971–2000
	FEB_PPT7100_CM	Mean February precip (cm) for the watershed, from 800 m PRISM data. 30 years period of record 1971–2000
	MAR_PPT7100_CM	Mean March precip (cm) for the watershed, from 800 m PRISM data. 30 years period of record 1971–2000
	APR_PPT7100_CM	Mean April precip (cm) for the watershed, from 800 m PRISM data. 30 years period of record 1971–2000
	MAY_PPT7100_CM	Mean May precip (cm) for the watershed, from 800 m PRISM data. 30 years period of record 1971–2000
	JUN_PPT7100_CM	Mean June precip (cm) for the watershed, from 800 m PRISM data. 30 years period of record 1971–2000
	JUL_PPT7100_CM	Mean July precip (cm) for the watershed, from 800 m PRISM data. 30 years period of record 1971–2000
	AUG_PPT7100_CM	Mean August precip (cm) for the watershed, from 800 m PRISM data. 30 years period of record 1971–2000
	SEP_PPT7100_CM	Mean September precip (cm) for the watershed, from 800 m PRISM data. 30 years period of record 1971–2000
	OCT_PPT7100_CM	Mean October precip (cm) for the watershed, from 800 m PRISM data. 30 years period of record 1971–2000
	NOV_PPT7100_CM	Mean November precip (cm) for the watershed, from 800 m PRISM data. 30 years period of record 1971–2000
	DEC_PPT7100_CM	Mean December precip (cm) for the watershed, from 800 m PRISM data. 30 years period of record 1971–2000
	JAN_TMP7100_DEGC	Average January air temperature for the watershed (°C), from 800 m PRISM data. 30 years period of record 1971–2000
	FEB_TMP7100_DEGC	Average February air temperature for the watershed (°C), from 800 m PRISM data. 30 years period of record 1971–2000
	MAR_TMP7100_DEGC	Average March air temperature for the watershed (°C), from 800 m PRISM data. 30 years period of record 1971–2000
	APR_TMP7100_DEGC	Average April air temperature for the watershed (°C), from 800 m PRISM data. 30 years period of record 1971–2000
	MAY_TMP7100_DEGC	Average May air temperature for the watershed (°C), from 800 m PRISM data. 30 years period of record 1971–2000
	JUN_TMP7100_DEGC	Average June air temperature for the watershed (°C), from 800 m PRISM data. 30 years period of record 1971–2000
	JUL_TMP7100_DEGC	Average July air temperature for the watershed (°C), from 800 m PRISM data. 30 years period of record 1971–2000
	AUG_TMP7100_DEGC	Average August air temperature for the watershed (°C), from 800 m PRISM data. 30 years period of record 1971–2000
	SEP_TMP7100_DEGC	Average September air temperature for the watershed (°C), from 800 m PRISM data. 30 years period of record 1971–2000
	OCT_TMP7100_DEGC	Average October air temperature for the watershed (°C), from 800 m PRISM data. 30 years period of record 1971–2000
	NOV_TMP7100_DEGC	Average November air temperature for the watershed (°C), from 800 m PRISM data. 30 years period of record 1971–2000
	DEC_TMP7100_DEGC	Average December air temperature for the watershed (°C), from 800 m PRISM data. 30 years period of record 1971–2000
	RFACT	Rainfall and runoff factor (1R factor1 of universal soil loss equation); average annual value for period 1971–2000
State Soil Geographic (STATSGO) Data Base	PERMAVE	Average permeability (inches/hour)
	AWCAVE	Average value for the range of available water capacity for the soil layer or horizon (inches of water per inches of soil depth)
	BDAVE	Average value of bulk density (grams per cubic centimeter)
	OMAVE	Average value of organic matter content (percent by weight)
	WTDEPAVE	Average value of depth to seasonally high water table (feet)
	ROCKDEPAVE	Average value of total soil thickness examined (inches)
	NO4AVE	Average value of percent by weight of soil material less than 3 inches in size and passing a no. 4 sieve (5 mm)
	NO200AVE	Average value of percent by weight of soil material less than 3 inches in size and passing a no. 200 sieve (.074 mm)
	NO10AVE	Average value of percent by weight of soil material less than 3 inches in size and passing a no. 10 sieve (2 mm)
	CLAYAVE	Average value of clay content (percentage)
	SILTAVE	Average value of silt content (percentage)
	SANDAVE	Average value of sand content (percentage)
	KFACT_UP	Erodibility factor which quantifies the susceptibility of soil particles to detachment and movement by water.

for state agencies. If a state did not have at least 10 sites, it was combined with an adjacent state with the fewest sites. Each physiographic region was divided in half to create regions with a smaller number of sites. A total of 11 subregions were developed for this study, and the number of sites within each subregion ranged from 14 to 48. Due to the limited number of sites within each subregion and to prevent overfitting, the stepwise selection procedure stopped if three explanatory variables had been entered in the model. Standard assumptions of regression analyses (homoscedastic, independent, and normally distributed residuals) were verified for each model.

[36] To compare the performance of the model development techniques, the adjusted coefficient of determination (Adj-R^2), the predicted coefficient of determination (Pred-R^2), and the percent standard error of prediction (SE) were calculated. The Pred-R^2 is calculated by sequentially removing one observation from the data set, developing the model with all observations except the removed observation, and evaluating how well the model predicts the removed observation (a leave-one-out cross-validation). SE is calculated using *Hardison*'s [1971] unbiased estimator of the variance of the residuals

4.3. Results

[37] Regression models were developed in 12 different regions across the eastern United States (1 for entire region, 5 subregions based on state boundaries, and 6 subregions based on physiographic provinces). The parameter estimators were derived by the four model development techniques: OLS, VIF, PCR, and PLS. Model selection was based on the same stepwise selection procedures employed in the Monte Carlo analysis. Table 4 lists the regional regression models across the 12 regions. Most regions contained at least one topographic watershed characteristic. Drainage area (DRN_SQKM) entered the models most frequently and was generally the first watershed characteristics to enter all model development techniques. Drainage area is often an important variable in low-streamflow regional regression models [*Kroll et al*., 2004; *Vogel et al*., 1999] given that the size of the watershed has a direct impact on the magnitude of streamflows. Maximum (ELEV_MAX_M_BASIN) or minimum (ELEV_MIN_M_BASIN) basin elevations were also commonly entered variables. These watershed characteristics can be surrogates for temperature and thus evapotranspiration, which can impact low-streamflow conditions.

[38] Watershed characteristics derived from the STATSGO soils data set frequently entered the models. No particular soil characteristic entered the models most frequently. Soil characteristics generally present were average total soil thickness (ROCKDEPAVE) and the average depth to the seasonally high water table (WTDEPAVE). These watershed characteristics have an impact on groundwater storage characteristics; hydrogeologic characteristics of a watershed have been shown to have an influence on low streamflows [*Kroll et al*., 2004]. The watershed characteristic from the PRISM temperature and precipitation data set that most often entered models was the average April basin precipitation (APR_PPT7100_CM), which may represent groundwater recharge characteristics in this region.

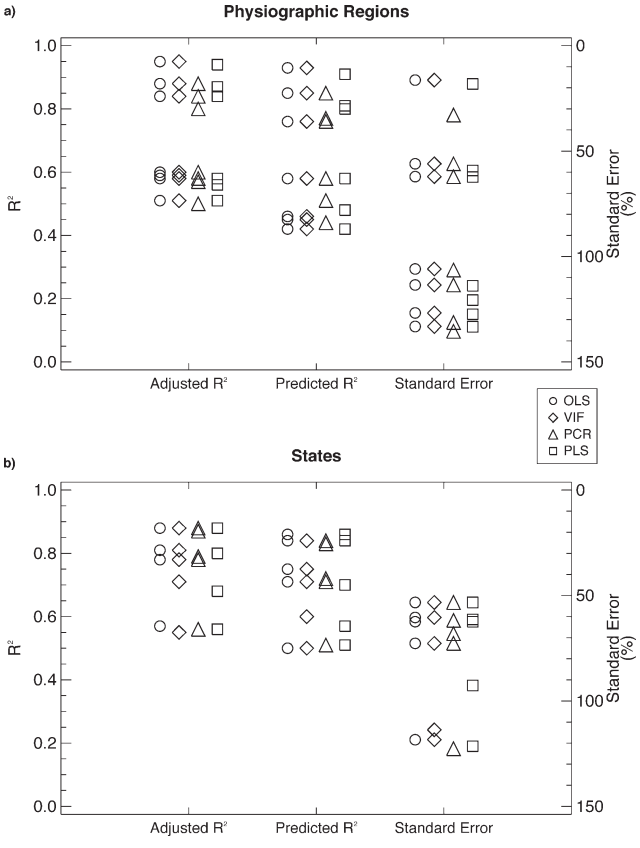

Figure 4. Adjusted R^2, predicted R^2, and standard error for OLS, VIF, PCR, and PLS for each subregion based on (a) physiographic regions and (b) States.

[39] While many of the potential watershed characteristics were highly correlated, multicollinearity was only detected (with a VIF >10) for one regression model developed, the Georgia and South Carolina region. Here a basin's monthly average number of measurable precipitation events (WDMAX_BASIN) was highly correlated with soil erosion potential following rainfall (RFACT). In this case, OLS, PCR, and PLS all developed models with the same explanatory variables, while VIF screened for multicollinearity and developed a model with only DRN_SQKM and WDMAX_BASIN.

[40] Figures 4a and 4b present the Adj-R^2, Pred-R^2, and the SE for each model development technique for subregions developed based on physiographic regions and states, respectively. In general, the Valley and Ridge and the southern half of the Blue Ridge physiographic regions produced the best regression models. These regions yielded Adj-R^2 and Pred-R^2 greater than 75%. Strong performances within these regions are attributed to watersheds of similar sizes. However, with the exception of the South Blue Ridge, all regional regression models produced relatively large SE% ($>60\%$). Surprisingly, regionalization by states tended to produce better regression models than regionalization by physiographic regions. This result may be due to having sites in closer proximity when state boundaries were employed.

[41] Across all of the subregions, results indicate that all techniques perform similarly. VIF mimics OLS, as the

Table 4. Number of Sites in Each Region and the Explanatory Variables Selected for the Low Streamflow Regional Regression Models

	Number of	Model Development Techniques			
Regions/States	Sites	OLS	VIF	PCR	PLS
All sites	130	DRN_SQKM, WDMAX_SITE, BFI	DRN_SQKM, WDMAX_SITE, BFI	DRN_SQKM, WDMAX_SITE, BFI	DRN_SQKM, WDMAX_SITE, BFI (2)
North Piedmont	27	T_MAXSTD_BASIN, RH_BASIN, APR_PPT7100_CM	T_MAXSTD_BASIN, RH_BASIN, APR_PPT7100_CM	T_MAXSTD_BASIN, RH_BASIN, APR_PPT7100_CM(2)	T_MAXSTD_BASIN, RH_BASIN
South Piedmont	25	DRN_SQKM, ELEV_MAX_M_BASIN	DRN_SQKM, ELEV_MAX_M_BASIN	DRN_SQKM, ELEV_MAX_M_BASIN (1)	DRN_SQKM, ELEV_MAX_M_BASIN (1)
North Blue Ridge	17	DRN_SQKM, ELEV_MIN_M_BASIN	DRN_SQKM, ELEV_MIN_M_BASIN	DRN_SQKM, ELEV_MIN_M_BASIN (1)	DRN_SQKM, ELEV_MIN_M_BASIN (1)
South Blue Ridge	16	DRN_SQKM, WTDEPAVE, ELEV_MIN_M_BASIN	DRN_SQKM, WTDEPAVE, ELEV_MIN_M_BASIN	DRN_SQKM, WTDEPAVE, ELEV_MIN_M_BASIN	DRN_SQKM, WTDEPAVE, ELEV_MIN_M_BASIN(2)
North Valley and Ridge	23	DRN_SQKM, APR_PPT7100_CM, ROCKDEPAVE	DRN_SQKM, APR_PPT7100_CM, ROCKDEPAVE	DRN_SQKM, APR_PPT7100_CM, ROCKDEPAVE	DRN_SQKM, APR_PPT7100_CM, ROCKDEPAVE
South Valley and Ridge	22	DRN_SQKM, WTDEPAVE, ELEV_MIN_M_BASIN	DRN_SQKM, WTDEPAVE, ELEV_MIN_M_BASIN	DRN_SQKM, WTDEPAVE, ELEV_MIN_M_BASIN	DRN_SQKM, PERMAVE, WTDEPAVE (2)
Pennsylvania and Maryland	26	DRN_SQKM, FST32F_BASIN, AWCAVE	DRN_SQKM, FST32F_BASIN, AWCAVE	DRN_SQKM, FST32F_BASIN, AWCAVE	DRN_SQKM, FST32F_BASIN, AWCAVE
North Carolina	26	DRN_SQKM, T_MAXSTD_BASIN, APR_PPT7100_CM	DRN_SQKM, T_MAXSTD_BASIN, APR_PPT7100_CM	DRN_SQKM, T_MAXSTD_BASIN, BFI (2)	DRN_SQKM, T_MAXSTD_BASIN, BFI (2)
Georgia and South Carolina	17	DRN_SQKM, WDMAX_BASIN, RFACT[a]	DRN_SQKM, WDMAX_BASIN	DRN_SQKM, WDMAX_BASIN, RFACT (2)	DRN_SQKM, WDMAX_BASIN, RFACT
Virginia and West Virginia	48	DRN_SQKM, AWCAVE, ELEV_MIN_M_BASIN	DRN_SQKM, AWCAVE, ELEV_MIN_M_BASIN	DRN_SQKM, ROCKDEPAVE, ELEV_MIN_M_BASIN(2)	DRN_SQKM, BFI, ELEV_MIN_M_BASIN (1)
Alabama and Tennessee	14	AWCAVE, ROCKDEPAVE, ELEV_STD_M_BASIN	AWCAVE, ROCKDEPAVE, ELEV_STD_M_BASIN	AWCAVE, ROCKDEPAVE, ELEV_STD_M_BASIN	ROCKDEPAVE, ELEV_STD_M_BASIN (1)

Number of components indicated in parentheses.
[a]Indicates regional regression model with VIF>10. "(2)" indicates the number of components added to the final model if less than 3.

calculated VIF was greater than 10 for only one subregion. For this case of high multicollinearity (Georgia and South Carolina region), all techniques performed similarly except for VIF, which performed much worse when a highly correlated explanatory variable was omitted from the model. Even though many watershed characteristics considered were highly correlated, OLS performed as well as PCR and PLS across these regions. This is because models with highly correlated variables were not selected for all but one region. This supports the results of the Monte Carlo simulation that when the true model is unknown OLS generally performs as well as other more complicated techniques which have been developed to address multicollinearity.

5. Conclusions

[42] This study used a Monte Carlo simulation to compare the performance of regional regression techniques when multicollinearity is present. The four model development techniques include OLS, OLS with variance inflation factor screening (VIF), PCR, and PLS. The purpose of this study was to understand the performance of model development techniques with respect to the (a) properties of the parameter estimators, (b) model predictions, and (c) probability of selecting the correct model. Performances of these techniques were observed with changing sample sizes, correlations between variables, and model error variances. Of particular interest was how these techniques compare when the true model is both known and unknown.

[43] When the true model is known:

The presence of multicollinearity greatly inflates the variance of the correlated OLS parameter estimators but does not influence the variance of the uncorrelated parameter estimators.

[44] Correlated parameter estimators for PCR and PLS are less affected by multicollinearity and consistently yield the parameter estimators with smaller MSE and variances than OLS parameter estimators, especially as smaller sample sizes.

[45] The MSE and variance of the parameter estimators decreases with increasing sample sizes.

[46] The predictive ability of the four techniques improves with increasing sample sizes.

[47] Multicollinearity has little effect on the performance of the predictions from the techniques within the data set range exampled ($n = 20$, 50, 100, and 1000).

[48] PCR and PLS, which produced the best parameter estimators, produced prediction with properties similar to OLS.

[49] When the true model is unknown:

[50] The presence of multicollinearity inflates the variance of the correlated parameters for all models, including PCR and PLS.

[51] At 20 observations, incorporating stepwise selection impairs model predictions compared to predictions when the true model is known.

[52] At 50 or more observations, the performance of model predictions is similar between models when the true model is known and or unknown.

[53] A higher presence of multicollinearity leads to incorrect selection of the true model; however, choosing the correct model more frequently did not improve model predictions at small sample sizes.

[54] The model's predictive ability and its ability to select the correct model are primarily influenced by the number of observations as opposed to the magnitude of the multicollinearity.

[55] A case study developing low-streamflow regional regression models in the eastern United States was also performed. Even though explanatory variables were highly correlated for this case study, only 1 of the 12 regional models has a high level of multicollinearity. In all regions, OLS performed as well as any of the other techniques in terms of model adjusted coefficient of determination, predicted coefficient of determination, and percent standard error of prediction. This result supports those from the Monte Carlo analysis: (1) when the true model is unknown, standard selection techniques rarely select a model with a high level of multicollinearity and (2) in the presence of highly correlated potential explanatory variables, OLS performs as well as more complicated techniques which have been proposed to address multicollinearity.

[56] The results of this study show that multicollinearity should not be viewed in isolation; instead, one should consider the sample size and overall model fit which provides a model framework for the correlated explanatory variables. If one is only interested in model predictions within the data set range of the model development techniques, the use of OLS for hydrologic regional regression analyses appears warranted; employing complex, biased regression techniques, such as PCR and PLS, to address multicollinearity does little to improve model predictions. In practice, models can exhibit a myriad of possible conditions, and no one technique exists to address the combination of problems that may be present in the data set [Wallis, 1965].

[57] **Acknowledgments.** The authors would like to acknowledge Richard Vogel, Gregor Laaha, and two anonymous reviewers who provided insightful comments that helped improve this manuscript, as well as Joana Luz who previously worked on a similar experiment.

References

Arditi, R. (1989), Avoiding fallacious significance tests in stepwise regression: A Monte Carlo method applied to a meteorological theory for the Canadian lynx cycle, *Int. J. Biometeorol.*, *33*(1), 24–26.

Chatterjee, S., and B. Price (1990), *Regression Diagnostics*, Wiley, New York.

Draper, N. R., and H. Smith (1981), *Applied Regression Analysis*, Wiley, New York.

Driver, N. E., and B. M. Troutman (1989), Regression models for estimating urban storm-runoff quality and quantity in the United States, *J. Hydrol.*, *109*(3–4), 221–236.

Falcone, J. A., D. M. Carlisle, D. M. Wolock, and M. R. Meador (2010), GAGES: A stream gage database for evaluating natural and altered flow conditions in the conterminous United States, *Ecology*, *91*(2), 621–621.

Frank, I. E., and J. H. Friedman (1993), A statistical view of some chemometrics regression tools, *Technometrics*, *35*(2), 109–135.

Gallo, E. L., P. D. Brooks, K. A. Lohse, and J. E. T. McLain (2012), Temporal patterns and controls on runoff magnitude and solution chemistry of urban catchments in the semiarid southwestern United States, *Hydrol. Processes*, *27*(7), 995–1010, doi:10.1002/hyp.9199.

Gardner, K. K., and R. M. Vogel (2005), Predicting groundwater nitrate concentration from land use in Nantucket, Massachusetts, *Groundwater*, *43*(3), 343–352.

Geladi, P., and B. R. Kowalski (1986), Partial least-squares regression: A tutorial, *Anal. Chim. Acta*, *185*, 1–17.

Graham, M. H. (2003), Confronting multicollinearity in ecological multiple regression, *Ecology*, *84*(11), 2809–2815.

Greene, W. (1990), *Econometric Analysis*, Macmillan, New York.
Grewal, R., J. A. Cote, and H. Baumgartner (2004), Multicollinearity and measurement error in structural equation models: implications for theory testing, *Marketing Sci.*, *23*(4), 519–529.
Haan, C. T., and D. M. Allen (1972), Comparison of multiple regression and principal component regression for predicting water yields in Kentucky, *Water Resour. Res.*, *8*(6), 1593–1596.
Hardison, C. H. (1971), Prediction error of regression estimates of streamflow characteristics at ungauged sites, *U.S. Geol. Surv. Prof. Pap. 750-C*, C228–C236.
Helland, I. S., and T. Almoy (1994), Comparison of prediction methods when only a few components are relevant, *J. Am. Stat. Assoc.*, *89*(426), 583–591.
Hoerl, A. E., and R. W. Kennard (2000), Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics*, *42*(1, special 40th Anniversary Issue), 80–86.
Johnston, J. (1972), *Econometric Methods*, McGraw-Hill, New York.
Jolliffe, I. T. (1986), *Principal Component Analysis*, Wiley, New York.
Kelsey, H., D. E. Porter, G. Scott, M. Neet, and N. White (2004), Using geographic information systems and regression analysis to evaluate relationships between land use and fecal coliform bacterial pollution, *J. Exp. Mar. Biol. Ecol.*, *298*(2), 197–209.
Kiers, H., and A. Smilde (2007), A comparison of various methods for multivariate regression with highly collinear variables, *Stat. Methods Appl.*, *16*(2), 193–228.
Kreuger, J., and L. Tornqvist (1998), Multiple regression analysis of pesticide occurrence in streamflow related to pesticide properties and quantities applied, *Chemosphere*, *37*(2), 189–207.
Kroll, C., J. Luz, B. Allen, and R. M. Vogel (2004), Developing a watershed characteristics database to improve low streamflow prediction, *J. Hydrol. Eng.*, *9*(2), 116–125.
Kroll, C. N., and J. R. Stedinger (1998), Regional hydrologic analysis: Ordinary and generalized least squares revisited, *Water Resour. Res.*, *34*(1), 121–128.
Kroll, C. N., and J. R. Stedinger (1999), Development of regional regression relationships with censored data, *Water Resour. Res.*, *35*(3), 775–784.
Laaha, G., and G. Bloschl (2007), A national low flow estimation procedure for Austria, *Hydrol. Sci. J.*, *52*(4), 625–644.
Mansfield, E. R., and B. P. Helms (1982), Detecting multicollinearity, *Am. Stat.*, *36*(3), 158–160.
Mason, C. H., and W. D. Perreault Jr. (1991), Collinearity, power, and interpretation of multiple regression analysis, *J. Market. Res.*, *28*(3), 268–280.
McElroy, F. W. (1967), A necessary and sufficient condition that ordinary least-squares estimators be best linear unbiased, *J. Am. Stat. Assoc.*, *62*(320), 1302–1304.
Mela, C. F., and P. K. Kopalle (2002), The impact of collinearity on regression analysis: The asymmetric effect of negative and positive correlations, *Appl. Econ.*, *34*(6), 667–677.
O'brien, R. (2007), A caution regarding rules of thumb for variance inflation factors, *Qual. Quant.*, *41*(5), 673–690.
Pope, P. T., and J. T. Webster (1972), The use of an F-statistic in stepwise regression procedures, *Technometrics*, *14*(2), 327–340.
Rawlings, J. O., S. G. Pantula, and D. A. Dickey (1988), *Applied Regression Analysis: A Research Tool*, 2nd ed., Springer, New York.
Reis, K.G. III, J. G. Guthrie, A. H. Rea, P. A. Steeves, and D. W. Stewart (2008), *StreamStats: A water resources web application*, U.S. Geol. Surv. Fact Sheet 2008–3067, 6 pp.
Rencher, A. C., and F. C. Pun (1980), Inflation of R2 in Best Subset Regression, *Technometrics*, *22*(1), 49–53.
Riggs, H. C. (1972), Low Streamflow Investigations, Techniques of Water-Resources Investigations of the United States Geological Survey, Book 4, chap. B1.
Roman, D., R. M. Vogel, and G. E. Schwarz (2012), Regional regression models of watershed suspended-sediment discharge for the eastern United States, *J. Hydrol.*, *472–473*, 53–62.
Smakhtin, V. U. (2001), Low flow hydrology: A review. *J. Hydrol.*, *240*(3–4), 147–186.
Stedinger, J. R., and G. D. Tasker (1985), Regional hydrologic analysis: 1. Ordinary, weighted, and generalized least squares compared, *Water Resour. Res.*, *21*(9), 1421–1432.
Stedinger, J. R., R. M. Vogel, and E. Foufoula-Georgiou (1992), Frequency analysis of extreme events, in *Handbook of Hydrology*, edited by D. R. Maidment, chap. 18, pp. 18.19–18.21, McGraw-Hill, New York.
Sun, R. J., and J. B. Weeks (1991), Bibliography of regional aquifer-systems analysis program of the U.S. Geological Survey, 1978–91, U.S. Geol. Surv. Water-Resour. Invest. Rep. 91–4122, 92 pp.
Syvitski, J. P. M., and J. D. Milliman (2007), Geology, geography, and humans battle for dominance over the delivery of fluvial sediment to the coastal ocean, *J. Geol.*, *115*(1), 1–19.
Tasker, G. D. (1980), Hydrologic regression with weighted least squares, *Water Resour. Res.*, *16*(6), 1107–1113.
Tasker, G.D., and N. E. Driver (1988), Nationalwide regression models for predicting urban runoff water quality at unmonitored sites, *J. Am. Water Resour. Assoc.*, *24*(5), 1091–1101.
Thomas, B., and R. M. Vogel (2012), The impact of stormwater recharge practices on Boston groundwater levels, *J. Hydrol. Eng.*, *17*(8), 923–932.
Tootle, G. A., A. K. Singh, T. C. Piechota, and I. Farnham (2007), Long lead-time forecasting of U.S. streamflow using partial least squares regression, *J. Hydrol. Eng.*, *12*(5), 442–451.
U.S. Geological Survey (2010), National streamflow statistics program (NSS), January. [Available at http://water.usgs.gov/osw/programs/nss/index.html.].
Vogel, R. M., and C. N. Kroll (1992), Regional geohydrologic-geomorphic relationships for the estimation of low-flow statistics, *Water Resour. Res.*, *28*(9), 2451–2458.
Vogel, R. M., I. Wilson, and C. Daly (1999), Regional regression models of annual streamflow for the United States, *J. Irrig. Drain. Eng.*, *125*(3), 148–157.
Wallis, J. R. (1965), Multivariate statistical methods in hydrology—A comparison using data of known functional relationship, *Water Resour. Res.*, *1*(4), 447–461.
Wold, S., M. Sjöström, and L. Eriksson (2001), PLS-regression: A basic tool of chemometrics, *Chemom. Intell. Lab. Syst.*, *58*(2), 109–130.