Advances in variable selection methods II: Effect of variable selection method on classification of hydrologically similar watersheds in three Mid-Atlantic ecoregions

H. Ssegane\textsuperscript{a}, E.W. Tollner\textsuperscript{a,b}, Y.M. Mohamoud\textsuperscript{b}, T.C. Rasmussen\textsuperscript{c}, J.F. Dowd\textsuperscript{d}

\textsuperscript{a}Department of Biological and Agricultural Engineering, University of Georgia, Athens, GA 30602, USA
\textsuperscript{b}Ecosystems Research Division, US Environmental Protection Agency, Athens, GA 30605, USA
\textsuperscript{c}Warnell School of Forestry and Natural Resources, University of Georgia, Athens, GA 30602, USA
\textsuperscript{d}Department of Geology, University of Georgia, Athens, GA 30602, USA

\begin{abstract}
Hydrological flow predictions in ungauged and sparsely gauged watersheds use regionalization or classification of hydrologically similar watersheds to develop empirical relationships between hydrologic, climatic, and watershed variables. The watershed classifications may be based on geographic proximity, regional frameworks such as ecoregions or classification using cluster analysis of watershed descriptors. General approaches used in classifying hydrologically similar watersheds use climatic and watershed variables or statistics of streamflow data. Use of climatic and watershed descriptors requires variable selection to minimize redundancy from a large pool of potential variables. This study compares classification performance of four variable groups to identify homogeneous watersheds in three Mid-Atlantic ecoregions (USA): Appalachian Plateau, Piedmont, and Ridge and Valley. The variable groups included: (1) variables that define watershed geographic proximity; (2) variables that define watershed hypsometry; (3) variables selected using causal selection algorithms; and (4) variables selected using principal component analysis (PCA) and stepwise regression. The classification results were compared to reference watersheds classified as homogeneous using three streamflow indices: Slope of flow duration curve; Baseflow index; and Streamflow elasticity using a similarity index (SI). Classification performance was highest using variables selected by causal algorithms (e.g., HITON-MB method, $SI = 0.71$ for Appalachian Plateau, $SI = 0.90$ for Piedmont, and $SI = 0.72$ for Ridge and Valley) compared to variables selected by stepwise regression ($SI = 0.72$ for Appalachian Plateau, $SI = 0.87$ for Piedmont, and $SI = 0.64$ for Ridge and Valley) and PCA ($SI = 0.71$ for Appalachian Plateau, $SI = 0.76$ for Piedmont, and $SI = 0.57$ for Ridge and Valley).
\end{abstract}

\begin{articleinfo}
Article history:
Available online 2 February 2012
This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of Attilio Castellarin, Associate Editor

Keywords:
Watershed classification
Hydrological similarity
Streamflow indices
Causal variable selection
Stepwise regression
Principal component analysis

\end{articleinfo}

1. Introduction

Development of regional frameworks such as hydrological landscape regions (Wolock et al., 2004) and ecoregions (Omernik and Bailey, 1997) has led to regionalization (Hall and Minns, 1999) of streamflow indices such that observed streamflow at gauged sites can be extrapolated to predict streamflow at ungauged sites in the same physiographic region. The concept of regionalization assumes that watersheds in the same physiographic region have similar hydrological signatures over a long period of time. Regionalization methods include: (1) statistical regionalization, where multiple regression is used to link hydrological responses to physical and climatic attributes (Kokkonen et al., 2003); (2) use of geospatial similarity (Merz and Blöschl, 2004); and (3) use of regional hydrological model parameters (Bastola et al., 2008). Irrespective of the approach used, observed data at gauged sites is used to model underlying hydrological processes at ungauged sites. Although previous studies have shown that geospatial similarity or geographical proximity does not always translate into hydrological similarity (Kokkonen et al., 2003; Acreman and Sinclair, 1986), geographic proximity may infer similarity in climatic conditions and watersheds. Commonly used approaches include those that infer similarity using climatic and watershed variables and those that use streamflow statistics or both.

Chiang et al. (2002) used cluster analysis and 16 streamflow statistics to generate six homogeneous regions from 94 watersheds in Alabama, Georgia, and Mississippi (USA). Kahya et al. (2008) used hierarchical clustering and streamflow patterns to classify 80 watersheds in Turkey. Acreman and Sinclair (1986) used 11 geographical similarities to classify 24 basins in the United Kingdom into 11 homogeneous regions.
watershed variables to classify 168 watersheds in Scotland into 5 homogeneous regions. And, Di Prinzio et al. (2011) used six streamflow statistics to establish reference homogeneous regions and compared results to four alternative classification methods using 12 watershed variables. The challenge with the above approaches is that there are no universally accepted similarity metrics (Wagener et al., 2007). Also, the watershed classification results depend on watershed descriptors used or the effectiveness of the variable selection methods.

On the choice of streamflow indices, Sawicz et al. (2011) suggest six streamflow metrics that define the different hydrologic functions of watersheds as possible universal metrics. The metrics include runoff ratio, flow duration curves, baseflow index, streamflow elasticity, ratio of snow days, and rising limb density. However, streamflow indices cannot be used to determine hydrological similarity of ungauged watersheds. On the choice of watershed descriptors, the most used variable selection methods are principal component analysis (PCA) (Salas et al., 2010; Ma et al., 2010; Álcazar and Palau, 2010) and stepwise regression analysis (SRA) (Barnett et al., 2010; Gong et al., 2010; Peña-Arangüebla et al., 2010). The conceptual basis of both approaches is not causality between response and explanatory variables. Stepwise regression analysis focuses on minimization of the predictive error while principal component analysis focuses on dimensional reduction (data extraction) by projecting high dimension data onto a low dimension space while maintaining the most relevant information.

Causal relationships between response and explanatory variables can be discovered by Bayesian networks. Bayesian networks consist of directed acyclic graphs whose nodes represent random variables and the edges conditional probabilities (Jensen and Nielsen, 2007; Karimi and Hamilton, 2009; Meganck et al., 2006). Therefore, the implied causation by this approach is probabilistic causation based on the theory that causes increase or change the probabilities of their effects such that the conditional probability of an effect given its cause is greater than the probability of the effect in absence of the cause (Hitchcock, 2010; Suppes, 1970; Cartwright, 1979). Thus, the possibility of event A occurring given that event B occurred is higher if event B causes event A and vice versa. Some of the algorithms that implement causal variable selection include: Grow-Shrink, GS (Margaritis and Thrun, 1999); interleaved Incremental Association Markov Boundary with PC algorithm, interIAMBnPC (Tsambardinos et al., 2003); Local Causal Discovery, LCD2 (Cooper, 1997); and HITON Markov Blanket, HITON-MB (Aliferis et al., 2003). For a brief description of the methods, the readers should refer to the first part of this study (reference for part I).

Therefore, the objective of the second part of the study is to compare the effectiveness of determining hydrologically similar watersheds using variables selected by causal algorithms (GS, interIAMBnPC, LCD2, and HITON-MB), stepwise regression analysis, principal component analysis, variables of geographical proximity, and watershed hypsometry in three Mid-Atlantic ecoregions: Appalachian Plateau, Piedmont, and Ridge and Valley (USA). The variable groups selected for comparison included: (1) variables that define watershed geographical proximity; (2) variables that define watershed hypsometry; (3) variables selected using causal selection algorithms; and (4) variables selected using principal component analysis (PCA) and stepwise regression. Hence, the focus of this study is on the effect of different variable selection methods on watershed classification while many previous studies have focused on different clustering or regionalization methods using the same set of variables.

We hypothesize that although hydrological similarity between watersheds in the same ecoregion is high when compared to watersheds from different ecoregions, all watersheds in the same ecoregion may not hydrologically behave in a similar manner. Therefore, the study used three streamflow indices: (1) slope of a flow duration curve (FDC); (2) the baseflow index (BFI); and (3) streamflow elasticity (SFE) with k-means clustering to classify reference homogeneous watersheds for each ecoregion. Watersheds classified using streamflow indices were considered to be the true hydrologically similar watersheds (reference watersheds) for each ecoregion. Then the ability of the four watershed variable groups to generate the exact homogeneous watersheds for the Appalachian Plateau, Piedmont, and Ridge and Valley were examined using a similarity index. The a priori assumption is that watershed classification using variables that typify the cause and effect relationship with the streamflow indices should give highest similarity when compared to reference watersheds.

The relevance of this approach was to emphasize the dependence and accuracy of watershed classification results on the variables used for classification. The interest in geographical proximity of watersheds is because proximity may infer similar climatic conditions and watershed form. While the interest in watershed hypsometry is based on the role of topography in hydrological processes. Stieglitz et al. (1997) highlighted the role of topography on soil moisture distribution, timing of discharge, and partitioning of streamflow into direct runoff and baseflow. Also, Vivoni et al. (2008) showed that total runoff reduced as the watershed hypsometric function changed from convex to concave. Therefore, this study also evaluates whether statistics of a hypsometric curve are adequate representatives of topography to differentiate hydrologic behavior across the three ecoregions.

2. Methods

2.1. Study area and data

Data used in this study covers three Mid-Atlantic physiographic regions (ecoregions) within USA (Fig. 1); the Appalachian Plateau (26 watersheds), the Piedmont (25 watersheds), and the Ridge and Valley (29 watersheds). Streamflow data used spanned the same 42 years of 1966–2007 epoch across all watersheds. Fig. 2 depicts topographic differences of headwaters of representative watersheds from each ecoregion. The watersheds were selected from Hydro-Climatic Data Network (HCDN) dataset (Slack and Landwehr, 1992) with emphasis on low extent of urbanization and minimum surface storage. For detailed description of the climatic and watershed descriptors used in this second part of the study, the reader is referred to part I of the study or Table 2.

2.2. Streamflow metrics

The common measures of watershed homogeneity or hydrological similarity analysis involve use of streamflow statistics (Kahya et al., 2008; Srinivas et al., 2008; Castellarin et al., 2008; Patil and Stieglitz, 2011). Three measures of watershed function signature were used to define hydrological similarity for watersheds in the same ecoregion. The measures included the slope of a flow duration curve (FDC), the baseflow index (BFI), and the streamflow elasticity. These three indices are a subset of six indices recommended by Sawicz et al. (2011). The choice of the three streamflow metrics was based on: (1) adequate representation of the watershed hydrologic response by the three metrics (refer to Sections 2.2.1, 2.2.2, 2.2.3); (2) use of fewer variables minimizes challenges of using high dimension data for unsupervised learning such as clustering (Fern and Brodley, 2003; Ding et al., 2002; Müller et al., 2009); and (3) the three metrics were easily extracted from readily available data compared to extracting all six indices. Watersheds classified as hydrologically homogeneous based on
these indices were considered to be the reference or true homogeneous watersheds for each ecoregion.

2.2.1. Flow duration curve

A flow duration curve (FDC) is a graphical representation of the percentage of time a streamflow is equaled or exceeded over a specified epoch (Vogel, 1994; Vogel and Fennessey, 1995). Therefore, the FDC depicts the integrated impacts of climate, geology, geomorphology, soils and vegetation on streamflow magnitudes. Flow duration curves for each watershed in the three ecoregions were generated using daily streamflows and a Weibull plotting position (Eq. (1)) for the 1966–2007 time period. The streamflows were standardized by the drainage area to minimize the effects of watershed size on slope of the flow duration curve. The slope of the curve between probabilities of exceedence of 20 % and 70 % was used as the overall slope.

\[ p_i(Q \geq q_i) = \frac{i}{N-1} \quad (1) \]

where \( p_i \) is probability of exceedence; \( Q \) is a random variable of \( q_i \); \( q_i \) is ordered streamflow; \( i \) is rank of \( q_i \); and \( N \) is total number of streamflow records.

2.2.2. Baseflow index

The baseflow index (BFI) describes the flow path and mean residence time of water through a watershed and therefore, quantifies the effects of watershed geology. The BFI for each watershed was estimated using the Eckhardt recursive digital filter (Eckhardt, 2005) in Eq. (2). A recession constant of 0.98 (\( \alpha = 0.98 \)) and a maximum baseflow index of 0.8 (\( BFImax = 0.8 \)) for humid areas such as the Mid-Atlantic ecoregions were used.

\[ b_t = \frac{(1 - BFImax)\alpha + b_{t-1}(1 - \alpha)BFImaxQ_t}{1 - BFImax} \quad (2) \]

where \( b_t \) is baseflow at time step \( t \) (daily); \( BFImax \) is maximum value of the baseflow index; \( \alpha \) is a recession constant; \( b_{t-1} \) is baseflow at a previous time step \( t = 1 \); and \( Q_t \) is streamflow at time step \( t \).

2.2.3. Streamflow elasticity

Streamflow (or climatic) elasticity of streamflow defines the sensitivity of streamflow to changes in precipitation (Sankarasubramanian et al., 2001). According to Zheng et al. (2009), streamflow is more sensitive to precipitation than to evapotranspiration. Therefore, this study used the precipitation based non-parametric estimator of streamflow elasticity (Eq. (3)) developed by Sankarasubramanian et al. (2001).

\[ SFE = \text{median} \left[ \frac{Q_t - \bar{Q}}{P_t - \bar{P}} \right] \quad (3) \]

where \( SFE \) is streamflow elasticity on annual basis; \( Q_t \) is annual total flow for year \( t \); \( \bar{Q} \) is average annual total flow; \( P_t \) is annual total precipitation for year \( t \); and \( \bar{P} \) is average annual total precipitation.

2.3. Watershed classification using streamflow metrics

Multivariate cluster analysis of \( k \)-means clustering was used to generate the reference set of hydrologically similar watersheds using streamflow metrics. Homogeneity measures developed by Hosking and Wallis (1997) were used to determine the homogeneity of the classified watersheds. For heterogeneous groups of watersheds, a discordancy index (Hosking and Wallis, 1997) was used to eliminate the non-group watersheds.
2.3.1. K-means clustering

The k-means clustering algorithm was used to classify hydrologically similar watersheds in each ecoregion using slope of the flow duration curve, baseflow index, and streamflow elasticity. The algorithm is an unsupervised iterative technique that groups multivariate data into clusters. According to (Wu et al., 2008) the k-means clustering is one of the top ten influential algorithms in data mining. Because the studied ecoregions are in close geographic proximity to each other, a k value of three was used such that each watershed could belong to any of the three ecoregions. Therefore, for each ecoregion, three clusters were generated. To improve the accuracy of the formed clusters, the algorithm was run 20 times using a squared euclidean distance as the metric for measuring within cluster and between cluster distance. Before clustering, all variables were standardized using min–max transformation (Eq. (4)).

\[
S_k = \frac{S_k - \min{S}}{\max(S) - \min(S)}
\]  

where \(S_k\) is transformed kth term of variable \(S\); and \(S_k\) is kth term of variable \(S\).

2.3.2. Homogeneity and discordancy tests

The Hosking and Wallis (1997) homogeneity tests were used to measure the degree of heterogeneity in a given cluster while the discordancy test was used to determine misclassified watersheds among the supposedly homogeneous watersheds. The underlying concept of the Hosking and Wallis (1997) \(H\)-statistics (\(H_1\), \(H_2\), and \(H_3\)) is to determine the variability of \(L\)-moment ratios (\(L\) – coefficient of variation, \(L – CV\); \(L\) – skewness; and \(L\) – kurtosis) and compare them to expected variability of a simulated homogeneous region using a four parameter kappa distribution. A group of watersheds is considered to be homogeneous if \(H < 1\), probably homogeneous if \(1 < H < 2\), and heterogeneous if \(H > 2\).

For each supposedly homogeneous set of watersheds, the homogeneity tests were computed on annual flows from 1966 to 2007. A non-supervised regional frequency analysis R-package (Viglione and Viglione, 2010) was used to calculate the \(H\)-statistics. For watershed groups whose \(H\)-statistics were greater than two, a discordancy test (Eq. (5)) was implemented to determine the misclassified watershed. The critical \(D\)-statistic is a function of the number of sites in a group. For example, if the number of sites in a homogeneous or heterogeneous region is equal to or greater than 15, the critical \(D\)-statistic is three such that sites with values of three or greater are eliminated from the group. For this study, sites with high \(D\)-statistic were eliminated until the \(H\)-statistics were about one or less than one.

\[
D_i = \frac{1}{N(N-1)} \sum_{j=1}^{N} (u_i - \bar{u}) S^{-1} (u_i - \bar{u})
\]

where \(D_i\) is discordancy measure for site \(i\) (watershed); \(N\) is number of sites in a group; \(u_i\) is vector containing the \(L – CV\), \(L\) – skewness, and \(L\) – kurtosis for site \(i\); \(\bar{u}\) is average of \(u_i\); and \(S\) is sample covariance matrix.

The cluster groups (set of homogeneous watersheds) with the maximum number of watersheds for each ecoregion were considered to be the characteristic (typical) watersheds for that ecoregion. These characteristic watersheds identified by the above approach were used to validate the accuracy of homogeneous watersheds identified using non-streamflow watershed characteristics.

To test the null hypothesis that selected homogeneous watersheds across the three ecoregions came from the same population distribution, nine pairwise comparisons of each streamflow metric (FDC, BFI, and SFE) were implemented using the Kruskal–Wallis test. Because the size of homogeneous watersheds is not the same across the three ecoregions, values of the first ten watersheds (Table 1) from each ecoregion were used for this analysis. Of the nine pairwise comparisons, only two were not significantly different (Baseflow index between Appalachian, and Ridge and Valley; and streamflow elasticity between Piedmont and Ridge and Valley). Therefore, the combined use of all three metrics provides adequate data structure to differentiate the three ecoregions.

2.4. Watershed classification using watershed descriptors

Four watershed variable groups that do not include streamflow statistics were analyzed for their suitability in watershed classification. This approach has implications for hydrological predictions in ungauged watersheds (Viola et al., 2011; Li et al., 2010; Ouarda and Shu, 2009). The suitability of a variable group is defined as its ability to identify the same hydrologically similar watersheds classified using streamflow metrics (reference watersheds). The four variable groups included (1) variables that define the geographical proximity of neighboring watersheds, (2) variables that define the
watershed hypsometry, (3) variables selected using causal variable selection algorithms, and (4) variables selected using principal component analysis (PCA) and stepwise regression. The third and fourth groups of variables seek to determine the dominant watershed variables that control streamflow in each ecoregion.

Three streamflow metrics were used and therefore, the top three variables from each variable group except for the watershed hypsometry were used to minimize misclassifications errors due to differences in data dimensions. The effect of data dimension on clustering results is documented in literature (Müller et al., 2009). For each variable group, variable transformation (Eq. (4)) was implemented prior to k-means clustering. The k-means algorithm was run 20 times using the squared euclidean distance to generate three clusters. The cluster group with the maximum number of watersheds was considered to be the set of representative homogeneous watersheds classified by the respective variable group. This procedure was implemented for each variable group on the three ecoregions. Details of the variable groups are discussed in the subsequent subsections.

2.4.1. Geographical proximity

Although Acreman and Sinclair (1986) showed that geographical proximity does not always infer hydrological similarity, geographical proximity may infer hydrological similarity because watershed neighborhood can translate into similarity in physical characteristics (e.g., watershed form) and climate conditions (e.g., similar rainfall and evapotranspiration). Watershed variables selected to represent geographical proximity included latitude, longitude, and elevation of the gauge station.

2.4.2. Watershed hypsometry

A hypsometric curve is a graphical representation of the distribution of area with elevation or the relative proportion of the watershed relief (Strahler, 1952). According to Luo (2000) a hypsometric curve can distinguish watersheds dominated by surface runoff (fluvial landforms defined by concave hypsometry) from watersheds dominated by subsurface runoff (terrestrial sapping landforms defined by convex hypsometry) using five statistical

Table 1
Hydrologically similar watersheds (reference watersheds) for each ecoregion.

<table>
<thead>
<tr>
<th>ID</th>
<th>Gauge name</th>
<th>USGS no.</th>
<th>DA</th>
<th>P</th>
<th>FDC</th>
<th>BFI</th>
<th>SFE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Tioga River, PA</td>
<td>01518000</td>
<td>730.4</td>
<td>897</td>
<td>-0.0191</td>
<td>0.574</td>
<td>1.227</td>
</tr>
<tr>
<td>2</td>
<td>Cowanesque River, PA</td>
<td>01520000</td>
<td>771.8</td>
<td>953</td>
<td>-0.0219</td>
<td>0.538</td>
<td>1.758</td>
</tr>
<tr>
<td>3</td>
<td>Towanda Creek, PA</td>
<td>01532000</td>
<td>556.8</td>
<td>1013</td>
<td>-0.0187</td>
<td>0.563</td>
<td>1.605</td>
</tr>
<tr>
<td>4</td>
<td>Tunkhannock Creek, PA</td>
<td>01534000</td>
<td>992.0</td>
<td>940</td>
<td>-0.0163</td>
<td>0.602</td>
<td>1.199</td>
</tr>
<tr>
<td>5</td>
<td>WB Susquehanna, PA</td>
<td>01541000</td>
<td>815.8</td>
<td>1113</td>
<td>-0.0146</td>
<td>0.599</td>
<td>1.208</td>
</tr>
<tr>
<td>6</td>
<td>Sinnemahoning Creek, PA</td>
<td>01543500</td>
<td>1774.1</td>
<td>1064</td>
<td>-0.0170</td>
<td>0.601</td>
<td>1.675</td>
</tr>
<tr>
<td>7</td>
<td>Pine Creek, PA</td>
<td>01548500</td>
<td>1564.4</td>
<td>1242</td>
<td>-0.0170</td>
<td>0.623</td>
<td>1.223</td>
</tr>
<tr>
<td>8</td>
<td>Blockhouse Creek, PA</td>
<td>01549500</td>
<td>97.6</td>
<td>993</td>
<td>-0.0170</td>
<td>0.598</td>
<td>1.234</td>
</tr>
<tr>
<td>9</td>
<td>Georges creek, MD</td>
<td>01599000</td>
<td>187.5</td>
<td>958</td>
<td>-0.0179</td>
<td>0.583</td>
<td>1.322</td>
</tr>
<tr>
<td>10</td>
<td>Buffalo Creek, PA</td>
<td>03049000</td>
<td>354.8</td>
<td>1006</td>
<td>-0.0172</td>
<td>0.579</td>
<td>0.836</td>
</tr>
<tr>
<td>11</td>
<td>Redstone Creek, PA</td>
<td>03074500</td>
<td>190.9</td>
<td>1069</td>
<td>-0.0126</td>
<td>0.641</td>
<td>2.250</td>
</tr>
<tr>
<td>12</td>
<td>Bluestone river, WV</td>
<td>03179000</td>
<td>1023.0</td>
<td>960</td>
<td>-0.0188</td>
<td>0.575</td>
<td>0.918</td>
</tr>
<tr>
<td>13</td>
<td>Big coal river, WV</td>
<td>03198500</td>
<td>1012.7</td>
<td>1176</td>
<td>-0.0184</td>
<td>0.556</td>
<td>1.162</td>
</tr>
</tbody>
</table>

Appalachian plateau (H1 = 0.52, H2 = 0.33, H3 = 0.067)

Piedmont (H1 = -0.46, H2 = 0.16, H3 = -0.33)

Ridge and Valley (H1 = -0.54, H2 = 0.54, H3 = 0.27)

a Drainage area (km²).
b Annual precipitation (mm).
c Slope of the flow duration curve (LS⁻¹ km⁻²).
d Baseflow index (−).
e Streamflow elasticity (−).
variables derived from the shape of a hypsometric curve. The five hypsometric variables selected include integral (HI), skewness (skew) and, kurtosis (kurtos) of the hypsometric curve, plus skewness (denSkew) and kurtosis (denKurtos) of the density function of the hypsometric curve. The first three hypsometric variables are defined by Eq. (6h) (Harlin, 1978; Pérez-Peña et al., 2009). The last two variables are estimated in a similar way by replacing \( f(x) \) with the density function \( g(x) = f'(x) \). The \( f(x) \) is the relative elevation corresponding to a relative watershed area \( x \).

\[
HI = \int_0^1 f(x)dx
\]

\[
\mu_1 = \frac{1}{T} \int_0^T x f(x)dx
\]

\[
\mu_2 = \frac{1}{T} \int_0^T (x - \mu_1)^2 f(x)dx
\]

\[
\sigma = \sqrt{\mu_2}
\]

\[
\mu_3 = \frac{1}{T} \int_0^T (x - \mu_1)^3 f(x)dx
\]

\[
\mu_4 = \frac{1}{T} \int_0^T (x - \mu_1)^4 f(x)dx
\]

\[
skew = \frac{\mu_3}{\sigma^3}
\]

\[
kurtos = \frac{\mu_4}{\sigma^4}
\]

The most common approach is to fit a continuous polynomial to the hypsometric data for each watershed. For the Mid-Atlantic watersheds, the polynomial fit gave high coefficients of determination \( R^2 \geq 0.9 \) in most cases, however, the graphical visual fit was not satisfactory. A combination of third order polynomial and a rational term (refer to Eq. (7)) gave high coefficients of determination and satisfactory graphical visual fits \( R^2 \geq 0.999 \) in over 90% of cases. For each watershed 200 points on a hypsometric curve were sampled using system for automated geoscientific analyses (SAGA) geographic information system (GIS) package (Olaya and Conrad, 2009).

\[
f(x) = a_1 + a_2x + a_3x^2 + a_4x^3 + \frac{1 - x^5}{(1 - x)^5 + a_5x^6}
\]

2.4.3. Variables selected by causal algorithms

For each ecoregion, four causal variable selection algorithms: GS, interAlMAnPC, LCD2, and HITON-MB were implemented to determine the dominant variables of 19 flow percentiles. The 19 flow percentiles were categorized as high flows (Q0.01, Q0.05, Q0.1, Q0.5, Q1, Q5, Q10); medium flows (Q20, Q30, Q40, Q50, Q60, Q70); and low flows (Q80, Q90, Q95, Q99, Q99.5, Q99.9) where Q10 represented the flow magnitude equaled or exceeded 10 percent of the flow record (1966–2007). Each algorithm was run 20 times by eliminating one data point (a watershed) on each run to improve the reliability of selected variables. The top three most selected variables across all flows were considered to be the dominant variables selected by each method.

2.4.4. Variables selected by principal component analysis and stepwise regression

Principal component analysis (PCA) is a common approach of reducing high dimension data in hydrological modeling (Salas et al., 2010; Ma et al., 2010; Alcázar and Palau, 2010; Gao et al., 2009). The PCA variable selection method implemented in this study is based on recommendations of Lu et al. (2007). The first five principal components of variables from each ecoregion were generated in the initial step. These components explained over 95% of the variability of initial variables. Five clusters were generated by \( k \)-means clustering of the five first principal components. The selected variables were the closest to each cluster centroid. The euclidean distance was used to determine the closest variables to each cluster centroid. This process was repeated 20 times by eliminating one watershed (data point) on each run. Again, the top three most selected variables after 20 runs were considered to be the dominant variables selected by PCA.

For stepwise regression, the method was implemented to select relevant variables for each of the 19 streamflow percentiles (refer to Section 2.4.3) on a single run. For each run a significance level of 0.1 was used to add a variable and a level of 0.2 to remove a variable. This process was repeated 20 times by eliminating a watershed on each run. The top three most selected variables across all flows were considered to be the dominant variables selected by stepwise regression.

2.5. Similarity between classifications based on streamflow indices and watershed descriptors

To assess the classification performance of the variable groups, three existing measures of similarity were initially used. These included the hamming distance (HD) by Dunne et al. (2002), a similarity index \( S_s \) by Kalousis et al. (2007), and a consistency index \( CI \) by Kuncheva (2007). Given two sets A and B such that set A consists of homogeneous watersheds classified by streamflow metrics (reference watersheds or hydrological similarity) and set B consists of homogeneous watersheds classified by a variable group (physical similarity), the similarities between sets A and B are computed as follows.

\[
HD = 1 - \frac{|A \cap B| + |B \setminus A|}{n}
\]

\[
S_s = 1 - \frac{|A| + |B| - 2|A \cap B|}{|A| + |B| - |A \cap B|}
\]

\[
CI = \frac{n|A \cap B| - k^2}{kn - k^2}
\]

where \( |A| \) is cardinality of set difference of A from B; \( |B| \) is cardinality of set difference of B from A; \( |A \cap B| \) is cardinality of set intersection of A and B; \( |A \cup B| \) is cardinality of set union of A and B; and \( n \) is the total number of features in the original dataset (e.g., 26 watersheds for Appalachian Plateau); \( k \) is the size of features to be compared in set A and B. This study used the minimum size for unequal set sizes. The hamming distance (HD) does not directly consider the cardinality of the set intersection while the consistency index (CI) does not directly consider the cardinality of the set differences. Both the hamming distance and the consistency index are greatly influenced by size of the original dataset \( n \) and are suited for equal set sizes. The Kalousis similarity index \( S_s \) just focuses on the two sets A and B without direct consideration for the set differences.

Therefore, this study proposed a fourth similarity index \( SI \) that accounts for the cardinality of the intersection of A and B, the cardinality of the set difference, and accounts for unequal number of features in the two sets A and B. The index is based on the assumption that the probability of a random algorithm (which does not consider correlation or causality) to generate two feature sets with similar features (intersection) is low while the probability of generating different features is high. Therefore, the cardinality of the set intersection is given a higher weight (rewarded) than the cardinality of the set differences (penalized). The index rewards the variable class for selecting the same homogeneous watersheds as the streamflow metrics (set intersection of A and B), however, penalizes it for selecting new watersheds not in set A (set differences of B from A) and for eliminating selected watersheds in A (set difference of A from B). This study used weighting factors of two for
Fig. 3. Comparison of similarity indices. The rand index, adjusted rand index, Jaccard index, and the Fowlkes-Mallows index are primarily used in analysis of cluster validity; while the hamming distance, Kalousis similarity, and the consistency index are used in analysis of stability (robustness) and consistency of variable selection methods. The similarity index developed in this study gives similar results as the Fowlkes-Mallows index and is comparable to the rand index and hamming distance.

Fig. 4. Clustering results for the Appalachian, Piedmont, and Ridge and Valley before homogeneity and discordancy tests. The three streamflow indices are standardized using minimum–maximum transformation (0–1 linear scale). The 2D plots depict the dominant clustering streamflow indices (where FDC = flow duration curve and BFI = baseflow index).
the set intersection and one for the set differences. This similarity index (SI) ranges from zero for totally different sets to one for exact sets.

\[
SI = \frac{1}{2} \left( 1 - \frac{|A \setminus B| + |B \setminus A| - 2|A \cap B|}{|A| + |B|} \right)
\]  

The performance of the above four similarity metrics was compared to four cluster validity indices of: (1) Rand index (Rand, 1971); (2) adjusted Rand index (Hubert and Arabie, 1985); (3) Jaccard index (Downton and Brennan, 1980); and (4) Fowlkes and Mallows index (Fowlkes and Mallows, 1983). The validity indices as defined by Steinley (2004) are expressed below.

\[
\text{RandIndex} = \frac{a + d}{a + b + c + d} - \frac{a + d}{N}
\]

(12)

\[
\text{AdjustedRandIndex} = \frac{N(a + d) - |(a + b)(a + c) + (b + d)(c + d)|}{N^2 - |(a + b)(a + c) + (b + d)(c + d)|}
\]

(13)

\[
\text{JaccardIndex} = \frac{a + d}{a + b + c + d} - \frac{a + d}{N}
\]

(14)

\[
\text{FowlkesMallows} = \frac{\sqrt{a + b} - (a + c)}{a}
\]

(15)

where \(a\) is the number of object pairs placed in the same cluster by both methods; \(b\) is the number of object pairs placed in the same cluster by method one but placed in a different cluster by method two; \(c\) is the number of object pairs placed in the same cluster by method two but placed in a different cluster by method one; and \(d\) is the number of object pairs that were not placed in the same cluster by either method. From the definitions of the similarity and cluster validity indices, it can be deduced that: (1) \(a = |A \cap B|\); (2) \(b = |A \setminus B|\); (3) \(c = |B \setminus A|\); and (4) \(d = n - |A \cup B|\).

Suppose a dataset consists of 10 catchments (\(n = 10\)). Based on hydrological similarity (use of streamflow metrics), all watersheds are classified as similar (\(|A| = 10\)), while using physical characteristics, only 8 are considered hydrologically similar (\(|B| = 8\)). Thus, the cardinality of the set intersection is 8 (\(|A \cap B| = 8\)), cardinality of set difference of B from A is 2 (\(|A \setminus B| = 2\)), and the cardinality of the set difference of B from A is zero (\(|B \setminus A| = 0\)). Therefore using Eqs. (8)–(15), the similarity indices are computed as below.

<table>
<thead>
<tr>
<th>Similarity Index (SI)</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hamming distance</td>
<td>0.80</td>
</tr>
<tr>
<td>Kalousis similarity</td>
<td>0.80</td>
</tr>
<tr>
<td>Consistency index</td>
<td>1.00</td>
</tr>
<tr>
<td>Rand index</td>
<td>0.80</td>
</tr>
<tr>
<td>Adjusted rand index</td>
<td>0.00</td>
</tr>
<tr>
<td>Jaccard index</td>
<td>0.80</td>
</tr>
<tr>
<td>Fowlkes-Mallows</td>
<td>0.8944</td>
</tr>
</tbody>
</table>

Fig. 3 depicts more results based on use of variable groups on the Appalachian data. All similarity metrics showed similar trend with causal variables for the Appalachian giving the highest similarity while the PCA selected variables for Piedmont giving the lowest similarity. The Fowlkes-Mallows index, and the developed similarity index (SI) gave similar results for all variable groups such that the trend lines are on top of each other. The Kalousis similarity index \(S_k\) and the Jaccard index gave similar results, and the hammering distance and the Rand index also, gave similar results. The adjusted Rand index was the most conservative with some values below zero followed by the consistency index. Work by Steinley (2004) showed that the minimum values of the adjusted Rand index may fall below zero. Since the performance of the developed similarity index (SI) is similar to that of Fowlkes-Mallows index and comparable to the Rand index and the hammering distance with values ranging from zero to one (better interpretability), subsequent assays in this study are based on SI.

3. Results and discussions

3.1. Watershed classification by streamflow metrics

3.1.1. Clustering Results of streamflow indices

Fig. 4 shows results of clustering three streamflow indices into three clusters for each ecoregion before the homogeneity and discordancy tests were implemented. The three dimension (3D) plots depict the three dimension spatial distribution of the clusters while the two dimension (2D) plots show the dominant clustering variable for each ecoregion. For the Appalachian plateau, the cluster centers are more distributed when projected onto streamflow elasticity axis (2D plot), flow duration curve (FDC) slope axis for the Piedmont, and streamflow elasticity axis for the Ridge and Valley. Therefore, streamflow elasticity is the dominant clustering variable followed by the slope of the flow duration curve. For the Appalachian plateau and the Piedmont ecoregions, the clusters with the most number of watersheds are visually obvious, however, for the Ridge and Valley, there are two clusters each with 12 watersheds. For each ecoregion, homogeneity and discordancy tests were implemented on clusters with the most watersheds to eliminate misclassified watersheds. Clusters with the most
watersheds after testing for homogeneity and discordancy were considered to be the reference homogeneous watersheds for each ecoregion.

3.1.2. Reference homogeneous watersheds

Table 1 shows the reference homogeneous watersheds (hydrologically similar) for each ecoregion after homogeneity and discordancy tests with corresponding streamflow metrics and Hosking and Wallis (1997) $H$-statistics, while Fig. 1 shows their map location. The results also depict the extent of heterogeneity in each ecoregion. For the Appalachian Plateau, 52% of the sampled watershed are homogeneous while 75% for the Piedmont, and 34.5% for the Ridge and Valley. This observation was supported by the $H$-statistics where the homogeneity of typical watersheds is highest for the Piedmont while lowest for the Ridge and Valley ($H_1$, $H_2$, and $H_3$ values in Table 1).

For the Appalachian Plateau, the North Central Appalachian and the Northern Allegheny Plateau sub-ecoregions have the highest concentration of reference homogeneous watersheds. The selected reference watersheds are dominated by first-order and second-order streams compared to the non-selected watersheds. This observation is explained by the average elongation ratios for the two groups, which is related to the watershed shape. The average elongation ratio (ER) of the reference watersheds is relatively low with small standard deviation while high with large standard deviation for the non-selected watersheds. Elongated watersheds (small elongation ratio) tend to be dominated by first-order and second-order streams compared to circular watersheds (high elongation ratios). Also, the average summer net precipitation, SNP (difference between summer precipitation and evapotranspiration) was negative for the reference watersheds while positive for the non-selected watersheds. For example, Georges (USGS 01599000) and NB Potomac (USGS 01595000) are close to each other and yet Georges is a reference watershed while NB Potomac is not. The difference is attributed to difference in shape (ER of 0.68 and 1.45 for Georges and NB Potomac, respectively) and summer net precipitation (SNP of $-117$ mm and $17.7$ mm for Georges and NB Potomac, respectively).

For the Piedmont ecoregion, both the Northern Piedmont and the Piedmont sub-ecoregions have a similar number of selected and non-selected watersheds. The selected reference watersheds for the Piedmont on average have a smaller drainage area ($DA = 478.4$ km$^2$) and a lower extent of urbanization ($Urban = 3.7\%$) compared to the non-selected watersheds ($DA = 888.9$ km$^2$ and $Urban = 7.9\%$). This explains the difference in classification of neighboring watersheds, for example, Big Pipe (USGS 01639500) is closer to Monocacy (USGS 01639000) yet Big Pipe ($DA = 264.2$ km$^2$ and $Urban = 1.8\%$) was classified while Monocacy ($DA = 448.1$ km$^2$ and $Urban = 4.6\%$) was not classified as a reference watershed. Note that some of the non-selected watersheds are close to neighboring ecoregions. For example, Stony (USGS 02046000) is close to the Piedmont-Southeastern plains border while West Conewago (USGS 01574000) is at the border between the Piedmont and Ridge and Valley.

![Fig. 6. Hypsometry and flow duration curves of representative watersheds from each ecoregion. The closest watershed to the cluster centroid of reference watersheds was chosen as a representative watershed.](image-url)
For the Ridge and Valley ecoregion, the reference watersheds are relatively smaller \( (DA = 644.9 \text{ km}^2) \) and have less surface storage (percent sum of open water and wetlands, \( SS = 0.21\% \)) compared to non-classified watersheds \( (DA = 756.3 \text{ km}^2 \text{ and } SS = 0.74\% \)). The non-selected watersheds are concentrated in the Northern part of the ecoregion \((44.4\% \text{ or } 8 \text{ of } 18 \text{ in Pennsylvania state})\) and are located near a neighboring ecoregion. For example, Cheat (USGS 03069500) and Youggheny (USGS 03075500) are located between two masses of the Central Appalachian. Wolf Creek (USGS 03175500), Walker (USGS 03173000), and Marsh Creek (USGS 01547700) are located near the Ridge and Valley and the Appalachian border. While Roanoke (USGS 02055000) and Marsh Run (USGS 01617800) are closer to the Blue Ridge ecoregion.

### 3.1.3. Hypsometry and flow duration curves of reference watersheds

Fig. 5 shows the general form of hypsometry (top) and flow duration curve (bottom) of the reference watersheds for each ecoregion. For each ecoregion, the general form was generated by computing the median of the reference watersheds. Fig. 6 shows hypsometry and flow duration curves of three representative watersheds.

As noted previously, the shape of a hypsometric curve can distinguish watersheds dominated by surface runoff (fluvial landforms-concave shape) from those dominated by sub-surface runoff (terrestrial sapping landforms-convex shape). The commonly used parameter to distinguish hypsometric shape is the hypsometric integral, where 0.5 is the threshold between concave \((HI < 0.5)\) and convex \((HI \geq 0.5)\). Vivoni et al. (2008) showed that total runoff was reduced as hypsometry changed from convex to concave if other watershed variables were held constant. Therefore, from Fig. 5, the flow that equaled or exceeded 50% \((Q_{50})\) of the record time \((1966–2007)\) should be highest for the Appalachian Plateau and lowest for the Ridge and Valley because \(HI_{Appa} > HI_{Pied} > HI_{Ridg}\). This is demonstrated by the corresponding \(Q_{50}\) in Fig. 5 in addition to hypsometric integrals and \(Q_{50}\) of representative watersheds in Fig. 6. The use of \(Q_{50}\) for comparison is because the main drivers of flood and drought streamflow conditions are at their minimum at \(Q_{50}\) and thus \(Q_{50}\) is the best streamflow percentile at which to compare the effects of topography.

Other watershed variables affect the shape of a flow duration curve. According to (Searcy, 1959), a steep curve in the flood region is representative of high flows over a short period, which is a characteristic of rain-caused floods compared to a relatively flat curve caused by prolonged travel time (a characteristic of snow-melt floods). These two scenarios are distinct between the Ridge and Valley (steep FDC in flood region; \(Q \leq Q_{10}\)) and the Appalachian Plateau (relatively flat FDC in flood region) in Figs. 5 and 6.

### 3.2. Watershed classification by selected variables

#### 3.2.1. Selected variable classes

The Table 2 defines the watershed variables while Table 3 depicts the three variables used to independently generate hydrologically similar watersheds by each selection method for each ecoregion. Variables of geographic proximity and watershed hypsometry are the same for the three ecoregions. However, variables selected using stepwise regression, principal component analysis (PCA), and causal algorithms differ for each ecoregion (refer to Table 3). The selected variables for classification of watersheds in the

---

### Table 2

<table>
<thead>
<tr>
<th>Variable</th>
<th>Units</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>RC</td>
<td>–</td>
<td>Circularity ratio (area to square of perimeter ratio)</td>
</tr>
<tr>
<td>RN</td>
<td>–</td>
<td>Ruggedness number (drainage density \times relief)</td>
</tr>
<tr>
<td>SLEN</td>
<td>m</td>
<td>Slope length</td>
</tr>
<tr>
<td>Cl</td>
<td>–</td>
<td>Convergence index</td>
</tr>
<tr>
<td>SMAX</td>
<td>m/km</td>
<td>Maximum slope</td>
</tr>
<tr>
<td>VDEP</td>
<td>m</td>
<td>Average valley depth</td>
</tr>
<tr>
<td>CPLAN</td>
<td>–</td>
<td>Plan curvature; rate of change of aspect along a contour</td>
</tr>
<tr>
<td>TWI</td>
<td>–</td>
<td>Topographic wetness index</td>
</tr>
<tr>
<td>MRVBF</td>
<td>–</td>
<td>Multi resolution index of valley bottom flatness</td>
</tr>
<tr>
<td>MRRTF</td>
<td>–</td>
<td>Multi resolution index of ridge top flatness</td>
</tr>
<tr>
<td>LDP</td>
<td>km</td>
<td>Longest drainage path (main channel length)</td>
</tr>
<tr>
<td>TSL</td>
<td>km</td>
<td>Total stream length</td>
</tr>
<tr>
<td>MCS</td>
<td>m/km</td>
<td>Main channel slope</td>
</tr>
<tr>
<td>AMEAN</td>
<td>deg</td>
<td>Average aspect</td>
</tr>
<tr>
<td>MAP</td>
<td>mm</td>
<td>Mean annual precipitation</td>
</tr>
<tr>
<td>MAET</td>
<td>mm</td>
<td>Mean annual evapotranspiration</td>
</tr>
<tr>
<td>NAP</td>
<td>mm</td>
<td>Net annual precipitation ((MAP-MAET))</td>
</tr>
<tr>
<td>Agric</td>
<td>%</td>
<td>Area under agriculture</td>
</tr>
<tr>
<td>Urban</td>
<td>%</td>
<td>Developed areas of low, medium, and high intensity</td>
</tr>
<tr>
<td>AWC</td>
<td>cm/</td>
<td>Available water content</td>
</tr>
<tr>
<td>Silt</td>
<td>cm/h</td>
<td>Saturated hydraulic conductivity</td>
</tr>
<tr>
<td>Porosity</td>
<td>–</td>
<td>Porosity</td>
</tr>
<tr>
<td>MSCL</td>
<td>cm</td>
<td>Macroscopic capillary length</td>
</tr>
<tr>
<td>T</td>
<td>cm²/h</td>
<td>Transmissivity</td>
</tr>
<tr>
<td>JUNP</td>
<td>mm</td>
<td>Average June precipitation</td>
</tr>
<tr>
<td>AUGP</td>
<td>mm</td>
<td>Average August precipitation</td>
</tr>
<tr>
<td>SepET</td>
<td>mm</td>
<td>Average September evapotranspiration</td>
</tr>
<tr>
<td>MMET</td>
<td>mm</td>
<td>Mean monthly potential evapotranspiration</td>
</tr>
<tr>
<td>ADI</td>
<td>mm</td>
<td>Annual dryness index; (ADI = \frac{Q_{50}}{Q_{60}})</td>
</tr>
<tr>
<td>RFx</td>
<td>mm</td>
<td>Rainfall equaled or exceeded (x%) of the record time</td>
</tr>
</tbody>
</table>

---

### Table 3

Top three variables\(^a\) selected by each method for watershed classification.

<table>
<thead>
<tr>
<th>Region</th>
<th>Variables</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appalachian</td>
<td>RF20, SEPP, RC</td>
<td>interAMBiPC</td>
</tr>
<tr>
<td>Piedmont</td>
<td>AWC, TWI, rockDep</td>
<td>Hydric, MRVBF, NAP</td>
</tr>
<tr>
<td>Ridge and Valley</td>
<td>RF10, TWI, MSCL</td>
<td>VDEP, TWI, LDP</td>
</tr>
<tr>
<td></td>
<td>LCD2</td>
<td>AMEAN, Agric, Urban</td>
</tr>
</tbody>
</table>

\(^a\) Refer to Table 2 for description of each variable.
Appalachian Plateau are dominated by climate using stepwise regression; climate and topography using PCA; and climate and topography using causal selection algorithms.

For the Piedmont ecoregion, the selected variables are dominated by soils, topography, and land use using stepwise regression; climate and topography using PCA; and topography and soils using causal selection algorithms. For the Ridge and Valley, the selected variables are dominated by soils and climate using stepwise regression; climate and topography using PCA; and topography, climate, and soils using causal selection algorithms.

3.2.2. Classification performance of selected variables

The results of Table 4 show that variables of geographical proximity performed best in the Appalachian Plateau. Most methods dominantly selected climatic variables for watershed classification in the Appalachian ecoregion (refer to Table 3). Accordingly, selected variables of geographical proximity performed better in the Appalachian (Table 4) because geographical proximity may infer the same climatic conditions. The hypsometric variables performed better in the Piedmont (Table 4) where topography was considered relevant by the causal algorithms (interIAMBnPC, LCD2, and HITON in Table 3).

The performance of the variable selection methods was based on two criteria. The first criterion was the similarity index (refer to Section 2.5): which is the similarity between watersheds classified as homogeneous using selected variables and the reference watersheds (watersheds classified using streamflow indices). Similarity index of zero meant that none of the watersheds classified as homogeneous using selected variables belonged to the reference watersheds while a value of one meant that all classified watersheds were exactly the same as the reference watersheds. The second criterion sought to assess the ability of variable selection methods to select variables that are unique to each ecoregion. For example, variables selected for the Appalachian Plateau should give the highest classification performance (highest similarity index) when applied to watersheds from the Appalachian Plateau and give relatively low classification performance for other ecoregions. Thus, the second criterion sought to emphasize the uniqueness of an ecoregion. Based on these two criteria, similarity indices of the main diagonal (3 × 3 matrix) should be higher than off-diagonal indices (Tables 4–6). Ideally, none of the off-diagonal indices should be equal or greater than the main diagonal similarity indices.

Table 4

<table>
<thead>
<tr>
<th>Variable class</th>
<th>Ecoregion data</th>
<th>Appalachian</th>
<th>Piedmont</th>
<th>Ridge and Valley</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proximity and hypsometry</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geographic proximity</td>
<td>0.64</td>
<td>0.52</td>
<td>0.48</td>
<td></td>
</tr>
<tr>
<td>Watershed hypsometry</td>
<td>0.40</td>
<td>0.71</td>
<td>0.29</td>
<td></td>
</tr>
<tr>
<td>PCA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appalachian</td>
<td>0.71</td>
<td>0.64</td>
<td>0.67</td>
<td></td>
</tr>
<tr>
<td>Piedmont</td>
<td>0.33</td>
<td>0.76</td>
<td>0.54</td>
<td></td>
</tr>
<tr>
<td>Ridge and Valley</td>
<td>0.50</td>
<td>0.71</td>
<td>0.57</td>
<td></td>
</tr>
<tr>
<td>Stepwise</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appalachian</td>
<td>0.72</td>
<td>0.58</td>
<td>0.64</td>
<td></td>
</tr>
<tr>
<td>Piedmont</td>
<td>0.67</td>
<td>0.87</td>
<td>0.50</td>
<td></td>
</tr>
<tr>
<td>Ridge and Valley</td>
<td>0.40</td>
<td>0.55</td>
<td>0.64</td>
<td></td>
</tr>
</tbody>
</table>

* Each cell represents classification performance (similarity index) of variables selected by the method (PCA or stepwise) for a specific ecoregion (corresponding row heading) when applied to data from a specific ecoregion (corresponding column heading). A value of one means watersheds classified as homogeneous using variables selected are exactly the same as the reference watersheds (classified using streamflow indices).

Based on results from Tables 4 and 5, only one method, the HITON Markov Boundary (HITON-MB) satisfied the two performance criteria. All other methods failed to meet the second criterion. For example, variables selected by the PCA for Ridge and Valley performed better when applied to data from the Piedmont (Table 4 row 9, column 3) than data from Ridge and Valley (Table 4 row 9, column 4). Also, variables selected by stepwise regression for the Appalachian gave the same performance as variables selected for the Ridge and Valley when applied to data from Ridge and Valley. Similar examples exist for the GS, interIAMBnPC, and LCD2 methods (Table 5). Table 6 shows improvement of watershed classification when variables selected by different methods are combined. However, combination of variables selected by stepwise regression and PCA still failed to meet the ecoregion uniqueness criterion.

On average, classification performance was higher for variable groups selected by causal algorithms compared to variable groups selected by stepwise regression and principal component analysis across all ecoregions. Higher classification performance by variables selected by causal algorithms was attributed to their intrinsic structure that seeks to establish causal associations between response and explanatory variables compared to stepwise regression that seeks to minimize the predictive error or the PCA that seeks to...
extract a subspace from high dimension data with the most information. Also, all variable groups performed best in the Piedmont and worst in the Ridge and Valley. This observation was attributed to the level of homogeneity of the reference watersheds in each ecoregion. The reference watersheds in the Piedmont were the most homogeneous whereas reference watersheds of the Ridge and Valley were the least homogeneous (refer to H-statistics of Table 1).

3.3. Hydrological implications of the results

Results show that ecoregion alone should not be a basis for regionalization because factors such as rate of urbanization, watershed shape, drainage area, and extent of surface storage introduce variability in hydrological functionality of watersheds in the same ecoregion. For this study, of the total sampled watersheds, 52% were classified as hydrologically similar for the Appalachian Plateau, 75% for the Piedmont and 34.5% for the Ridge and Valley. As shown in Table 2, this study presents a number of variables that were selected for watershed classification in each ecoregion by different methods. We hypothesize that these variables may have important hydrological implications and may contribute to watershed model parameterizations and for development of regional regression models. In this study, for the same ecoregion, different variable selection methods selected different variable groups which gave comparable classification results ($SI \geq 0.7$), however, only one method (HITON-MB) was able to identify variables that were unique to each ecoregion without compromising classification performance. This may imply that the robustness of regionalized flow indices and regionalized model parameters may greatly depend on robustness of the variable selection method.

4. Conclusions

This study evaluated the ability of variables selected using different methods to identify the same hydrologically similar reference watersheds classified using streamflow indices in three Mid-Atlantic physiographic provinces. Watersheds classified using three streamflow indices and $k$-means clustering were considered to be the reference (“typical”) watersheds for each ecoregion. We then evaluated the ability of four watershed variable classes to reproduce the exact homogeneous watersheds selected by the streamflow indices for the Appalachian Plateau, Piedmont, and Ridge and Valley using $k$-means clustering. A similarity index was used to compare classification results by streamflow indices and classification results by watershed variables. The four variable groups included: (1) geographical proximity; (2) watershed hypsometry; (3) variables selected using four causal selection algorithms; and (4) variables selected using principal component analysis (PCA) and stepwise regression.

On average, among variable groups, classification performance was higher for variables selected by causal algorithms (for GS method, $SI = 0.89$ for Appalachian, $SI = 0.86$ for Piedmont, and $SI = 0.67$ for Ridge and Valley) compared to variables selected by stepwise regression ($SI = 0.72$ for Appalachian, $SI = 0.87$ for Piedmont, and $SI = 0.64$ for Ridge and Valley) and principal component analysis ($SI = 0.71$ for Appalachian, $SI = 0.76$ for Piedmont, and $SI = 0.57$ for Ridge and Valley). Also, only one method (HITON-MB) was able to identify variables that were unique to each ecoregion without compromising classification performance (refer to Table 5; $SI = 0.71$ for Appalachian, $SI = 0.90$ for Piedmont, and $SI = 0.72$ for Ridge and Valley). Therefore, causal variable selection for watershed classification is recommended over stepwise regression and principal component analysis.

References


