In Search of Causal Watershed Variables for Watershed Classification and Daily Streamflow Prediction in Ungauged Watersheds

by

Herbert Ssegane

(Under the direction of E.W. Tollner)

Abstract

Hydrological predictions at a watershed scale are generally made by extrapolating and upscaling hydrological behavior at point and hillslope scales. However, some dominant hydrological drivers at a hillslope may not be as relevant at the watershed scale because of watershed heterogeneities. Quantifiable watershed data in the form of watershed descriptors and streamflow indices are becoming readily available such that appropriate variable selection provides new insight into the watershed descriptors that dominate different streamflow regimes at the watershed scale. Stepwise regression and principal components analysis are commonly used to select descriptive variables for relating runoff to climate and watershed descriptors. These methods do not derive causal associations between response and explanatory variables. Therefore, this study compares the accuracy, stability, and predictive power of variables selected by stepwise regression and principal components analysis with causal selection methods(e.g. HITON Markov Blanket) and their relevance in watershed hydrologic modeling. The results demonstrate that causal variable selection methods (especially HITON Markov Blanket) have a high probability of selecting true variables compared to stepwise regression and principal component analysis. Also, variables selected by causal methods give high classification accuracy of hydrologically similar watersheds and improve the predictive power for regionalized flow duration curves. Classification of hydrologically similar

watersheds in three Mid–Atlantic regions of Appalachian Plateau (28 basins; 98–1779 km^2), Piedmont (19 basins; 34.8–620 km^2), and Ridge and Valley (25 basins; 48–1857 km^2) are highest for variables selected by causal algorithms using a similarity index (SI) which quantifies agreement between hydrological similarity (based on streamflow indices) and physical similarity (based on selected variables). For the HITON-MB method, SI=0.71 for Appalachian, 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, 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).

INDEX WORDS: Causal variable selection, Stepwise regression, Principal component analysis (PCA), Probabilistic causation, Watershed variables, Hypsometric curve, Geographic proximity, Streamflow separation, Temporal sequence, Flow Duration Curve, Regional streamflow equations, HITON–MB, Markov Blanket, Feature selection, Feature relevance, Grow–Shrink, Local Causal Discovery

In Search of Causal Watershed Variables for Watershed Classification and Daily Streamflow Prediction in Ungauged Watersheds

by

Herbert Ssegane

B.S., Makerere University (Uganda), 2002

M.S., University of Georgia, Athens, 2007

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2011

In Search of Causal Watershed Variables for Watershed Classification and Daily Streamflow Prediction in Ungauged Watersheds

by

Herbert Ssegane

Approved:

Major Professor: E.W. Tollner

Committee: Yusuf M. Mohamoud
Todd C. Rasmussen
John F. Dowd

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2011

In Search of Causal Watershed Variables for Watershed Classification and Daily Streamflow Prediction in Ungauged Watersheds

HERBERT SSEGANE

December 12, 2011

DEDICATION

To my wife, Christine Kyosimire, my two daughters, Petra Mirembe and Michelle L. Patra, and my son, Preston M. Ssegane, who provide meaning to everyday life. To my Auntie, Christine Sanyu Kyobe, who has provided countless opportunities for my personal growth after the passing away of my Father and Mother. To my Auncle and all Aunties who continuously provided a place I would call home.

ACKNOWLEDGEMENTS

Let me take this opportunity to express my deep and sincere gratitude to Prof. E.W. Tollner for providing me with an opportunity to pursue graduate studies. His guidance, support, encouragement, and accessibility has been invaluable to the successfully completion of my Ph.D. He created an environment that allowed me to grow academically, professionally, and as an independent thinker. I am greatly indebted to him.

To the members of my advisory committee: Dr. Yusuf M. Mohamoud, Prof. Todd C. Rasmussen, and Prof. John F. Dowd, for your guidance, insightful comments and suggestions, commitment, and time. I am very appreciative of your support because your doors were always open for questions and discussions. You allowed me to freely express and explore different suppositions and concepts. To this effect, I am fortunate to have been introduced to Dr. Mohamoud by Dr. Chittaranjan, who helped me formulate and focus the research questions. I am grateful to Dr. Dowd and Dr. Rasmussen, who motivated me to improve my computer programming knowledge.

My fellow graduate students: Xianben Zhu, Umakanta Jena, Weilin Wang, Leopold Matamba, and their families for demonstrating genuine and true friendship and for exemplifying hard work with a niche for excellence.

I also want to thank Brenda Valeska Ortiz, Lina Wang, Praveen Kolar and his wife, Kaushlendra Singh and his wife, Yeshanth Rao, Rajan, Ke Cheng and the graduate class of 2005 for their support.

Finally and most importantly, I want to thank my wife, Christine Kyosimire who allowed me to be both a father and a student and for enduring lonely nights while I was at the department. Your patience, love, and care are greatly appreciated

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background and justification

Hydrologic modeling and ungauged watersheds

Use of hydrologic models is common practice as a scientific and technical basis for improving decision making in water resource planning, flood forecasts, prediction of hydrologic responses at ungauged stream watersheds (Harmel et al., 2008; Kim et al., 2009; Marshall and Randhir, 2008; Sun et al., 2008), management of surface runoff, sediment, nutrient leaching, and pollutant transport processes. Watershed models are conceptualizations of physical hydrological processes at watershed scales based on experimental data and field observations. Therefore, accuracy of model outputs is based on how realistic a model represents watershed and environmental processes. Apart from model structure, accuracy of data inputs derived from spatial and temporal sampling frequencies influence accuracy of model outputs. Therefore, the modeling process is complicated by limited understanding of how physical processes scale from point observations to integrated complex watershed interactions. Even with limited understanding of physical processes that drive hydrological processes, there are numerous watershed hydrological models with different model structures, process conceptualization, data input and resolution, state variables, and model parameters. Singh and Frevert (2002, 2006), USEPA (2008), and Donigian et al. (1991) discuss over 40 watershed

models commonly used by water resource managers, engineers, and hydrologists. The challenge of limited understanding of dominant physical processes at watershed scale can be minimized at gauged watersheds by calibration of models.

Model calibration is a process of determining model parameters using historical observations at a specified watershed location such that future predictions of system response can be inferred. The concept of model calibration is based on the assumption that past observations and watershed responses are predictors of future system response under different management practices. However, the modeling challenges are compounded in ungauged and poorly gauged watersheds, in watersheds where monitoring has been discontinued, and watersheds with few observations. Some of the current methodologies for simulating stream flow time series at ungauged and poorly gauged watersheds include; use of neighboring gauged watershed response characteristics; use of remote–sensing data; and use of physically based models (Sivapalan et al., 2003). Irrespective of the method used, the concepts utilize data and data derived relationships at gauged watersheds.

Development of regional frameworks such as hydrologic landscape regions (Wolock et al., 2004) and Eco-regions (Omernik and Bailey, 1997) that aggregate hydrological, geological, biotic, and abiotic factors have led to regionalization (Hall and Minns, 1999) of streamflow characteristics such that observed responses in gauged watersheds can be extrapolated to predict responses of ungauged watersheds in the same physiographic or hydrometric region. The flow duration curves (FDC) is a commonly used tool for hydrologic predictions in ungauged watershed. Flow duration curves estimate percentage of time specific stream flows are equaled or exceeded based on historical flow records of a watershed. The flow duration for such analyses ranges from hourly, to daily, to monthly, and to annual time steps. Vogel and Fennessey (1995) review applications of flow duration curves in water resource planning and management. Applications of FDC include water–use planning, waste–load allocations, frequency of suspended–sediment loads, and examination of streamflow suitability for stream habitat. Mohamoud (2008) illustrates that use of FDC provides more detailed information on watershed functionality and behavior than indices such as mean annual flow and the base flow index.

Regionalized flow duration curves

The regionalization of flow duration curves (FDC) involves development of empirical relationships between major hydrologic, climatic, and watershed characteristics for a region. The region may be based on geographic proximity or other regional frameworks. The concept is based on the assumption of watersheds in the same physiographic or hydrometric region have similar hydrological behavior over time. However, caution must be taken because geographic neighborhood of two watersheds may not always produce similar hydrological signatures (Acreman and Sinclair, 1986). Castellarin et al. (2004) classified methods of regionalization of FDC into statistical, parametric, and graphical methods. Statistical approaches include use of log–normal (LeBoutillier and Waylen, 1993) and normal frequency distributions (Singh et al., 2001). The parametric approaches fit data to an exponential function (Quimpo et al., 1983), third order polynomial function (Mimikou and Kaemaki, 1985) and a power function (Franchini et al., 2005). The most commonly used approach utilizes tools of regression analysis to develop regionalized flow equations linking watershed hydrologic response to climatic and geophysical characteristics (Castellarin et al., 2007; Chalise et al., 2003; Sanborn and Bledsoe, 2006). Hydrologic response measurements used include lowest consecutive seven–day, ten–year streamflow, mean annual flow, base flow index, and stream frequency. Watershed characteristics used include climatic conditions, drainage area, geology, geomorphology, soils, and land cover and land use.

1.2 Statement of the problem

Most previous studies use multiple stepwise regression and a limited pool of climatic and physiographic data to explore regional physical hydrological drivers (variables) that are operational under low, medium, and high stream flow regime. Examination of over 42 published papers (Alcázar et al., 2008; Castellarin et al., 2007; Eng et al., 2007; Johnston and Shmagin, 2008; Kroll et al., 2004; Laaha and Blöschl, 2006; Mohamoud, 2008; Sanborn and Bledsoe, 2006; Sando et al., 2009, see, e.g.) determined 72 topographic variables, 66 climatic variables, 98 soil variables, and 15 lan-

duse and landcover variables used by different researchers. The deductions from these studies vary depending on: (1) the region; (2) the initial watershed variables used; (3) the conceptualization by different researchers of what constitutes relevant variables; and (4)the variable selection method.

Although the majority of the variables are statistically redundant, the challenge is to devise an approach that identifies relevant variables that characterize different flow regimes on a regional basis. The current approaches include stepwise regression (Barnett et al., 2010; Brandes et al., 2005; Gong et al., 2010; Heuvelmans et al., 2006; Peña-Arancibia et al., 2010) and principal component analysis (Alcazar and Palau, 2010; Ma et al., 2010; Morris et al., 2009; Salas et al., 2010). Both approaches provide useful results but are susceptible to elimination of relevant variables because they are not based on the principle of causality between dependent and independent variables. Also, use of a limited pool of independent variables may result in selection of irrelevant variables in absence of other relevant variables. This phenomenon is referred to as Simpsons paradox (Ma et al., 2010).

For example, regional flow equations developed by: Verdin and Worstell (2008) for conterminous U. S.; Yu et al. (2002) for Taiwan; and Zhu and Day (2009) for Pennsylvania (U.S.) show that an increase in the watershed mean elevation increases flow while regional flow equations developed by Mohamoud (2008) for Mid-Atlantic ecoregions (U.S.) show that an increase in median elevation decreases flow. Also, flow equations developed by Mohamoud (2008) show that minimum elevation has a positive effect while equations by Castellarin et al. (2007) in Italy show that maximum elevation has a negative effect. The simple justification for the inconsistency of the impact of elevation on flow (in above studies) can be attributed to difference in regions; however, these observations can also be attributed to absence of dominant independent variables in the initial pool of variables such that elevation becomes a surrogate variable. Therefore, the inconsistencies can broadly be attributed to: (1) different number of initial variables used in the respective studies; (2) different region of study; (3) the variable selection method used. The above probable sources of errors in current regional flow equations can be minimized by use of a large initial pool of variables and use of causal variable selection methods.

1.3 Objectives

The main research objective of this study is to develop a hydrologic predictive system for ungauged watersheds based on variable selection methods that infer causal association between response (dependent) and explanatory (independent) variables. Specific objectives include

1 To assess the accuracy, consistency, and predictive potential of five causal variable selection methods in comparison to stepwise regression and principal component analysis

   a) For accuracy, all methods were evaluated for their ability to select true variables of two known functional relationships

   b) Regarding consistency (reliability), all algorithms were implemented on datasets with a known functional relationship and watershed data to quantify their ability to select same variables when data was slightly perturbed item[c)]Selected variable classes for high, medium, and low flows for each method were classified to quantify the dominant variable class for Piedmont physiographic province

2 To compare the effectiveness of determining hydrologically similar watersheds using different variable selection methods in three Mid-Atlantic ecoregions. The overall variable groups for comparison included variables

   a) that define watershed geographical proximity

   b) that define watershed hypsometry

   c) selected using causal selection algorithms

   d) selected using principal component analysis (PCA) and stepwise regression

3 Daily streamflow prediction for ungauged watersheds by independent estimation of magnitude and sequence

   a) Prediction of streamflow magnitude using regionalized flow duration curves developed from variables selected by a causal variable selection method

b) Examination of the effect of the relative distance and drainage area of the donor (gauged) and target (ungauged) watersheds on the accuracy of the predicted sequence

c) Improvement of the accuracy of predicted daily streamflow by generating a sequence from an ensemble of streamflow data of more than one donor watershed

BIBLIOGRAPHY

Acreman, M., Sinclair, C., 1986. Classification of drainage basins according to their physical characteristics; an application for flood frequency analysis in Scotland. Journal of Hydrology 84 (3-4), 365–380. 3

Alcazar, J., Palau, A., 2010. Establishing environmental flow regimes in a Mediterranean watershed based on a regional classification. Journal of Hydrology 388 (1-2), 41–51. 4

Alcázar, J., Palau, A., et al., 2008. A neural net model for environmental flow estimation at the Ebro River Basin, Spain. Journal of Hydrology 349 (1-2), 44–55. 3

Barnett, F., Gray, S., Tootle, G., et al., 2010. Upper Green River Basin (United States) Streamflow Reconstructions. Journal of Hydrologic Engineering 15, 567. 4

Brandes, D., Hoffmann, J. G., Mangarillo, J. T., 2005. Base flow recession rates, low flows, and hydrologic features of small watersheds in Pennsylvania, USA. Journal of the American Water Resources Association 41 (5), 1177–1186. 4

Castellarin, A., Camorani, G., Brath, A., 2007. Predicting annual and long-term flow-duration curves in ungauged basins. Advances in water resources 30 (4), 937–953. 3, 4

Castellarin, A., Galeati, G., Brandimarte, L., Montanari, A., Brath, A., 2004. Regional flow-duration curves: reliability for ungauged basins. Advances in Water Resources 27 (10), 953–965. 3

Chalise, S., Kansakar, S., Rees, G., Croker, K., Zaidman, M., 2003. Management of water resources and low flow estimation for the Himalayan basins of Nepal. Journal of Hydrology 282 (1-4), 25–35. 3

Donigian, A., Consultants, A., Huber, W., Barnwell Jr, T., Branch, A., 1991. Modeling of Nonpoint Source Water Quality in Urban and Non-urban Areas. 1

Eng, K., Milly, P., Tasker, G., 2007. Flood Regionalization: A Hybrid Geographic and Predictor-Variable Region-of-Influence Regression Method. Journal of Hydrologic Engineering 12, 585. 3

Franchini, M., Galeati, G., Lolli, M., 2005. Analytical derivation of the flood frequency curve through partial duration series analysis and a probabilistic representation of the runoff coefficient. Journal of Hydrology 303 (1-4), 1–15. 3

Gong, G., Wang, L., Condon, L., Shearman, A., Lall, U., 2010. A Simple Framework for Incorporating Seasonal Streamflow Forecasts into Existing Water Resource Management Practices1. JAWRA Journal of the American Water Resources Association 46 (3), 574–585. 4

Hall, M., Minns, A., 1999. The classification of hydrologically homogeneous regions. Hydrological sciences journal 44 (5), 693–704. 2

Harmel, R., Rossi, C., Dybala, T., Arnold, J., Potter, K., Wolfe, J., Hoffman, D., 2008. Conservation effects assessment project research in the Leon River and Riesel watersheds. Journal of Soil and Water Conservation 63 (6), 453. 1

Heuvelmans, G., Muys, B., Feyen, J., 2006. Regionalisation of the parameters of a hydrological model: Comparison of linear regression models with artificial neural nets. Journal of Hydrology 319 (1-4), 245–265. 4

Johnston, C., Shmagin, B., 2008. Regionalization, seasonality, and trends of streamflow in the US Great Lakes Basin. Journal of Hydrology 362 (1-2), 69–88. 3

Kim, S., Tachikawa, Y., Sayama, T., Takara, K., 2009. Ensemble flood forecasting with stochastic radar image extrapolation and a distributed hydrologic model. Hydrological Processes 23 (4), 597–611. 1

Kroll, C., Luz, J., Allen, B., Vogel, R., 2004. Developing a watershed characteristics database to improve low streamflow prediction. Journal of Hydrologic Engineering 9, 116. 3

Laaha, G., Blöschl, G., 2006. Seasonality indices for regionalizing low flows. Hydrological processes 20 (18), 3851–3878. 3

LeBoutillier, D., Waylen, P., 1993. A stochastic model of flow duration curves. Water resources research 29 (10), 3535–3541. 3

Ma, H., Liu, L., Chen, T., 2010. Water security assessment in Haihe River Basin using principal component analysis based on Kendall τ. Environmental monitoring and assessment 163 (1), 539–544. 4

Marshall, E., Randhir, T., 2008. Effect of climate change on watershed system: a regional analysis. Climatic Change 89 (3), 263–280. 1

Mimikou, M., Kaemaki, S., 1985. Regionalization of flow duration characteristics. Journal of Hydrology 82 (1-2), 77–91. 3

Mohamoud, Y., 2008. Prediction of daily flow duration curves and streamflow for ungauged catchments using regional flow duration curves/Prevision de courbes de debits classes et de debit fluviatile pour des bassins. Hydrological Sciences Journal-Journal des Sciences Hydrologiques 53 (4), 706–724. 2, 3, 4

Morris, A. J., Donovan, J. J., Strager, M., 2009. Geospatial Analysis of Climatic and Geomorphic Interactions Influencing Stream Discharge, Appalachian Mountains, USA. Environmental Modeling and Assessment 14 (1), 73–84. 4

Omernik, J., Bailey, R., 1997. DISTINGUISHING BETWEEN WATERSHEDS AND ECOREGIONS1. JAWRA Journal of the American Water Resources Association 33 (5), 935–949. 2

Peña-Arancibia, J., van Dijk, A., Mulligan, M., Bruijnzeel, L., Gebrehiwot, S., Ilstedt, U., Gärdenas, A., Bishop, K., Brocca, L., Melone, F., et al., 2010. The role of climatic and terrain attributes in estimating baseflow recession in tropical catchments. Hydrology and Earth System Sciences Discussions 7, 4059–4087. 4

Quimpo, R., Alejandrino, A., McNally, T., 1983. Regionalized flow duration for Philippines. Journal of Water Resources Planning and Management 109 (4), 320–330. 3

Salas, J., Fu, C., Rajagopalan, B., 2010. Long Range Forecasting of Colorado Streamflows Based on Hydrologic, Atmospheric, and Oceanic Data. Journal of Hydrologic Engineering 1, 210. 4

Sanborn, S., Bledsoe, B., 2006. Predicting streamflow regime metrics for ungauged streamsin Colorado, Washington, and Oregon. Journal of Hydrology 325 (1-4), 241–261. 3

Sando, S., Fish, U., (US), G. S., 2009. Estimation of Streamflow Characteristics for Charles M. Russell National Wildlife Refuge, Northeastern Montana. US Dept. of the Interior, US Geological Survey. 3

Singh, R., Mishra, S., Chowdhary, H., 2001. Regional flow-duration models for large number of ungauged Himalayan catchments for planning microhydro projects. Journal of hydrologic engineering 6, 310. 3

Singh, V., Frevert, D., 2002. Mathematical modeling of watershed hydrology. Mathematical models of large watershed hydrology, 1–22. 1

Singh, V., Frevert, D., 2006. Watershed models. CRC Press. 1

Sivapalan, M., Takeuchi, K., Franks, S., Gupta, V., Karambiri, H., Lakshmi, V., Liang, X., McDonnell, J., Mendiondo, E., O'connell, P., et al., 2003. IAHS Decade on Predictions in Ungauged

Basins (PUB), 2003–2012: Shaping an exciting future for the hydrological sciences. Hydrological Sciences Journal 48 (6), 857–880. 2

Sun, G., McNulty, S., Moore Myers, J., Cohen, E., 2008. Impacts of Multiple Stresses on Water Demand and Supply Across the Southeastern United States1. JAWRA Journal of the American Water Resources Association 44 (6), 1441–1457. 1

USEPA, 2008. Handbook for developing watershed plans to restore and protect our waters. US Environmental Protection Agency, Office of Water, Nonpoint Source Control Branch. 1

Verdin, K., Worstell, B., 2008. A Fully Distributed Implementation of Mean Annual StreamFlow Regional Regression Equations1. JAWRA Journal of the American Water Resources Association 44 (6), 1537–1547. 4

Vogel, R., Fennessey, N., 1995. Flow Duration Curves II: A Review of Applications in Water Resources Planning. JAWRA Journal of the American Water Resources Association 31 (6), 1029–1039. 2

Wolock, D., Winter, T., McMahon, G., 2004. Delineation and evaluation of hydrologic-landscape regions in the United States using geographic information system tools and multivariate statistical analyses. Environmental Management 34, 71–88. 2

Yu, P., Yang, T., Wang, Y., 2002. Uncertainty analysis of regional flow duration curves. Journal of Water Resources Planning and Management 128, 424. 4

Zhu, Y., Day, R., 2009. Regression modeling of streamflow, baseflow, and runoff using geographic information systems. Journal of environmental management 90 (2), 946–953. 4

CHAPTER 2

LITERATURE REVIEW

2.1 Watershed hydrologic modeling

Overview

Watersheds (or drainage basins apart from aquifers) are the basic units where interactions between surface, subsurface, groundwater and the respective hydrologic processes are related. Therefore, hydrologic modeling is a process by which our understanding of interactions between the climatic, pedologic, lithologic, and hydrospheric are represented at a watershed scale (Singh and Frevert, 2002, 2006). This is because streamflow can be gauged at point outlets. The watershed representations vary from conceptual to mathematical models and vary in complexity from simple to detailed mathematical descriptions of interactions between dominant hydrological processes. Hydrologic models are useful in water resource planning and management to simulate and predict responses of hydrologic systems due to changes in system inputs (climatic conditions) and system characteristics (changes in landuse and landcover). Specific uses of hydrologic models include; simulation of fate and movement of solutes and contaminants (Jacques et al., 2008; Poole et al., 2008) and design of hydrologic structures (Afshar et al., 2009; Taguas et al., 2008).

Model types

Model types differ in modeling approach and model structure. Based on modeling approach, some models are physically based (deterministic) models and some models are stochastic (Obropta and Kardos, 2007). Deterministic models use physics (e.g. conservation of mass and momentum) while stochastic models are based on statistical relations from observed data. Deterministic models produce the same result for given model inputs and do not take into account the uncertainties in input data and due to model structure. Examples include the Antecedent Precipitation Index (Pan et al., 2003) and the Sacramento Soil Moisture Accounting (Anderson et al., 2006) models. Stochastic models consider inputs as random variables with probabilities and therefore, account for uncertainties in data inputs. Examples include the Markov chain models (Smith and Marshall, 2008; Zhang et al., 2009). Because stochastic models are data driven, some may be specific to a watershed or region. Based on model structure, some models are lumped and some are distributed models (Carpenter and Georgakakos, 2006; Reed et al., 2004). Lumped hydrologic models ignore spatial variability of watershed parameters and climatic inputs by lumping them into a single value while distributed models account for the spatial variability. However, there is some level of lumping in distributed models since watershed characteristics cannot be represented at every point.

Other model types include scenario based deterministic models (Koutsoyiannis et al., 2007). Scenario based deterministic models use some aspects of stochastic and deterministic models. Instead of using random variable inputs, the models produce multiple simulations for predicting probabilistic values using a deterministic model. Model states such as current soil moisture are used as stationary conditions for multiple runs. The runs are based on historical records or short term forecasts. Examples of scenario based deterministic models include the Ensemble Streamflow Predictions.

2.2 Historic perspectives of rainfall runoff process

Horton (Horton, 1935a,b) conceptualized runoff generation process as a result of rain falling or snow melting at intensity in excess of the rate of infiltration into the soil. This conceptual model is known as the infiltration excess or Hortonian overland flow model. Horton assumed that at the start of a storm event, all rainfall infiltrates and as the storm event proceeds, the instantaneous infiltration rate decreases exponentially through time. The infiltration excess runoff generation is more prevalent in areas where soils have low initial moisture and low infiltration rate like bare soils in arid and semi-arid environments. Research by Burman (1969) concluded that the Hortonian model fitted experimental data because of the many parameters in the model rather than the models representation of the infiltration process.

Forested watersheds have soils with high infiltration rates due to presence of above ground and decomposing vegetation, therefore infiltration excess runoff process is not as important. Research (Beven and Freer, 2001a; Steenhuis and Muck, 1988) showed that soils in the Northeast U.S have infiltration rates that are rarely exceeded by rainfall rate. For such watersheds, runoff generation is due to precipitation falling on saturated areas or local regions. This process is called the saturation excess runoff generation. Betson and Marius (1969) deduct that saturation excess runoff is likely to be generated only on some parts of the hillslope, a concept known as partial area contribution. Research by Dunne and Black (1970a,b) showed that saturation excess runoff may vary between two similar storm events, a concept known as variable source area (VSA). The locations of areas (VSA) generating saturation excess runoff depend on the topography and soil transmissivity of the watershed, and expand and contract in size as the water table rises and falls respectively.

Earlier conceptualization of rainfall runoff process treated a watershed as a lumped, linear, and time invariant system Singh (1988). Such models included the Zoch model Zoch (1934), the Clark model Clark (1945), the Nash model Nash (1960); Nash et al. (1958) and Dooge model Dooge (1959). The Zoch model represents watershed system behavior with time area concentration (TAC) curves and routes the watershed through a linear reservoir to form a unit hydrograph. A unit hydrograph is a runoff hydrograph in response to a uniformly distributed unit rainfall excess

for a given time duration. The Clark model replaced the unit hydrograph with an instantaneous unit hydrograph (IUH) by developing a unit hydrograph of a watershed due to instantaneous rainfall and routing the TAC curves through a linear reservoir. The Nash model assumes that the watershed behavior is similar to a cascade of linear reservoirs each having a time lag with the instantaneous rainfall imposed on the uppermost reservoir. The watershed outflow based on the Nash model is expressed by a gamma distribution with shape and scale parameters (Singh, 1988). The Dooge model is a generalized instantaneous unit hydrograph that uses the Nash model but accounts for the effect of flow transition. The concept of a geomorphologic unit hydrograph or a geomorphologic instantaneous unit hydrograph (Moussa, 2008) augments the Dooge model by relating the watershed IUH to the geometry of the watershed stream network.

2.3 Advances in hydrologic modeling

The technological advancements in remote sensing, satellite sensor technology, geographic information systems (GIS) and database management have greatly improved representation of watershed spatial and temporal heterogeneity compared to improved understanding of the underlying physical processes. Compared to the historical perspectives of Hortonian overland flow and saturation excess runoff generation, studies have shown that pre–event soil moisture constitutes the bulk of the observed streamflow. Additional understandings of the hydrologic process have included temporal and spatial scaling effects in hydrologic modeling. However, most of recent developments are in the areas of remote sensing and data acquisition, use of GIS with improved computational capabilities, scaling effects in time and space, use of data mining and optimization techniques, use of chemical and isotopic tracers, and improved statistical methods to explore error propagation from model inputs and model structure. Therefore, the discussion of recent developments in hydrologic modeling will focus on the above mentioned areas.

Data acquisition, visualization, and processing

Advancements in satellite sensor technology and remote–sensing have included measurement of near surface soil moisture. From 1978 to 1987, the scanning multichannel microwave radiometer (SMMR) on a space–borne platform was used to measure the vertical and horizontal polarized radiations (from which near surface soil moisture is derived) at spatial resolutions of 27 km to 147 km (Yu and Gloersen, 2005) . The SMMR was replaced by the special sensor microwave imager (SSM/I) in 1987 (Ferraro et al., 1996). In 2002, an advanced microwave scanning radiometer (AMSR-E) was launched to improve the spatial resolution to 38 km to 56 km (Njoku et al., 2003) . Data assimilation for the above radiometers was based on the C-band (frequency, f = 4–8 GHz). Recent studies (Joseph et al., 2008; Wigneron et al., 2008) have shown that best soil moisture retrieval could be achieved using the L-band (f = 1–2 GHz). Therefore, in 2009, the soil moisture and ocean salinity mission (SMOS) was launched to observe moisture over land surfaces and salinity over the oceans. Other sensor technologies include the space borne onboard synthetic aperture radar (SAR) imaging radar designed to achieve images of spatial resolution less than 30 m and the European remote sensing scatterometers (ERS).

Developments in precipitation estimates include Next Generation Radar (NEXRAD) and the Tropical Rainfall Measuring Mission (TRMM) data. The NEXRAD data is a high resolution spatial (4 km) and temporal (1 hr) precipitation, however, it is vulnerable to errors such as the complexity of the relationship between radar reflectivity and rainfall rate (Z–R relationship) at the land surface (Dinku et al., 2002).

Recent developments in GIS applications for hydrologic modeling include integration of GIS with expert systems and analytic hierarchy process for multi-criteria site analysis (Nekhay et al., 2009). Most GIS developments related to hydrologic modeling are in the area of watershed representation. The availability of GIS datasets in form of digital terrain models (DTM) and digital elevation models (DEM) has enabled representation of watershed spatial heterogeneity.

Currently available DEM data (Sanders, 2007) include airborne light ranging and detection (LiDAR) data at one-ninth second (1/9 s) resolution, airborne interferometric synthetic aperture

radar (IfSAR) data at 1/9 s resolution, United States Geologic Survey (USGS) national elevation data (NED) at 1/9, 1/3 and 1 second (s) resolution, and shuttle radar topography mission (SRTM) data at 1 and 3 s resolutions. A DEM resolution of 1 s corresponds to about 30 m. However, most of the current digital terrain analysis (DTA; techniques used to derive terrain parameters) algorithms, assume the earth's spherical surface as a flat surface (Weber et al., 2007). Problems with such techniques include geometric distortions and spatial data overlapping or spatial data breaking of derived terrain or topographic parameters. Advancements in DTA have included use of quaternary triangular mesh (QTM) method to overcome the above mentioned problems.

Use of chemicals and isotopic tracers

The use of chemical tracers (Botter et al., 2008; Flury and Wai, 2003) and isotopic tracers (Jones et al., 2006) has been fundamental in exploring questions of sources of water contributing to streamflow, age of water, and sources of solutes and contaminants in surface and ground water. Chlorine and bromide ions are some of the chemical tracers used in geochemical processes of groundwater recharge estimation while the water isotope ($\delta^{18}O$ and $\delta^{2}H$) are some of the isotopes used to determine different sources of streamflow water. Tracers have been used to separate pre-event and event based (Kvaerner and Klove, 2006) runoff (Buttle, 1994; Winston and Criss, 2002). Other uses of tracers have included separation of streamflow based on whether the source is groundwater or event water or soil water. The traditional techniques of tracer studies include digging of trenches and multiple sampling to generate a representative spatial variation of flow paths. The recent developments include use of imaging techniques as alternatives.

Computational intelligence and hydrologic modeling

Artificial neural networks are black box modeling systems that relate complex physical phenomena using weights to link inputs and outputs. The modeling approach conceptualizes behavior of synaptic strength between the neurons of a biological nervous system. Artificial neural networks (ANN) are used to heuristically estimate response without deterministically defining the under-

lying nonlinear dynamics of the physical process (Rabunal and Dorado, 2006). Each neuron is connected to other neurons by means of direct linkages that weigh the transferred information. Specifically, the data transferred through the links between neurons are multiplied by weights that define the strength of a transient signal between neutrons. The process of determining the network of weights for information transfer between neutrons is called learning or training, similar to calibration of mathematical models. The most used learning procedure for artificial neural networks uses the feed–forward, error back-propagation algorithm (De Vos and Rientjes, 2008). The initial randomly estimated weights are corrected during training by comparing the ANN outputs to target outputs (such as measurements or field data). The errors are then backward propagated to determine the optimal weight adjustments necessary to minimize the errors.

The capability of ANN to represent nonlinear relationships has led to their application in the fields of finance and manufacturing (Kamruzzaman et al., 2006), image processing (Mas and Flores, 2008), aircraft design (Jules et al., 2002), cancer diagnosis (Naguib and Sherbet, 2000), and hydrology (Chen et al., 2008; Lee et al., 2008; Prabha and Hoogenboom, 2008). Recent applications of ANN have demonstrated their effectiveness in determining soil water content (Jana et al., 2008), aquifer parameters (Karahan and Ayvaz, 2006), daily evaporation (Tabari et al., 2010) , and flow forecasting (Lee et al., 2008; Ochoa-Rivera, 2008). Other applications of artificial intelligence techniques in hydrologic modeling include use of fuzzy logic (Jia and Culver, 2008), support vector machines (Kaheil et al., 2008), and genetic algorithms (Ines and Mohanty, 2008; Kamp and Savenije, 2006).

Scaling effects in hydrology

Scaling in hydrologic modeling refers to a process of predicting responses for a longer time scale (e.g. landuse change in 25 years) based on observations at shorter time scales or the use of large scale models (models developed based on regional or global observations) to make predictions at catchment or watershed scale and vice-versa. The scaling processes interpolate or extrapolate hydrologic responses in time and space. Scaling effects in hydrologic modeling refer to propagation

of errors or the uncertainty of the predicted hydrologic responses due to interpolation or extrapolation. Scaling effects may refer to impact of using different spatial representations of watershed.

2.4 Hydrologic predictions in ungauged watersheds

Hydrologic predictions in ungauged watersheds refers to reconstruction of past hydrologic responses (river flow or water level) of watersheds; 1) that have no flow measuring instruments (ungauged); 2) that are poorly gauged (less gauges compared to watershed size); and 3) that are gauged with few years of data records; using climatic inputs, landuse and landcover data, and watershed topography (Sivapalan et al., 2003). Current and future hydrologic data (e.g. water level in a river or lake, river discharge, sediment and water quality) for an ungauged watershed is critical to water supply planning and water engineering works (construction of dams, reservoirs, and spillways). Also, hydrologic data is relevant in analyzing the impact of significant modifications of landuse such as deforestation and urbanization, off-stream withdraws, an operation of dams on magnitude and frequency of downstream flows.

Several methods have been used to model hydrological responses of ungauged watersheds. The methods include; 1) statistical regionalization, where multiple regression is used to link hydrological responses of watersheds to their respective physical and climatic attributes (Kokkonen et al., 2003) ; 2) use of geospatial similarity (Merz and Bloschl, 2004); and 3) use of regional hydrological model parameters (Bastola et al., 2008). Irrespective of the approach used, all methods directly or indirectly interpolate or extrapolate observed data at gauged watersheds and use some form of mathematical representation of our understanding of the underlying physical processes. Previous studies have shown that geospatial similarity or geographical proximity does not always translate into hydrological similarity (Kokkonen et al., 2003). Therefore, the folowing literature review focuses on regionalization of watershed model parameters and regional flow equations.

2.5 Regionalization of watershed model parameters

The availability of high resolution spatial and temporal watershed and climatic data is the driving force behind regional calibration of watershed models. The method uses streamflow data and corresponding climatic and watershed characteristics at gauged watersheds to calibrate watershed models. The optimized model parameters at gauged watersheds are related to watershed characteristics of ungauged watersheds with similar hydrological response (Singh and Frevert, 2006). Singh and Frevert (2006) summarizes different methods for regional calibration of watershed models as; 1) multiple regression, 2) cluster analysis, 3) kriging, 4) artificial neural networks, and 5) hydrological homogeneity.

Various studies have applied and compared different methods of regionalizing watershed parameters. Bastola et al. (2008) used regionalization schemes of multiple regressions, artificial neural network, and partial least squares regression to generate regionalized parameters of TOPMODEL (Beven, 1997) for watersheds across the World. The results showed that the above regionalization schemes did not account for uncertainties in the input data. Heuvelmans et al. (2006a) compared linear regression and artificial neural network regionalization schemes to generate SWAT (Brown and Hollis, 1996) model parameters. The artificial neural network scheme overall performed better within the data range while the linear regression scheme depicted better extrapolation results. (Gotzinger and Bárdossy, 2007) modified the Lipschitz and monotony conditions to determine regionalized parameters of HBV (GRAHAM and Jacob, 2000) model. The main challenge of regionalization of watershed model parameters is the existence of multiple parameter sets with equally good model outputs, a scenario coined as equifinality (Beven and Freer, 2001b) and thus introducing more uncertainties in the model results.

2.6 Regional flow equations

Regional flow equations or regionalized flow indices utilize statistical techniques (e.g. multiple regressions) to relate watershed characteristics and climatic conditions to hydrological responses of

hydrologically similar watersheds or watersheds in the same geographical proximity (Heuvelmans et al., 2006a; Kokkonen et al., 2003; Mwakalila, 2003). Watershed characteristics, climatic conditions, and hydrological responses are measured at gauged watersheds and the derived relationships are transferred to ungauged watersheds. Acreman and Sinclair (1986) showed that geographical proximity of watersheds does not translate into hydrological similarity. Therefore, statistical techniques such as cluster analysis have been applied to classify watersheds of similar hydrological response (Nathan and McMahon, 1990). Broadly, the explanatory variables used include; climatic conditions, geological data, geomorphologic data, soils, landuse and landcover data. Different studies have used different climatic data and watershed characteristics. Garcia-Martinó et al. (1996) used 53 parameters to describe watershed characteristics while (Mohamoud, 2008) used 42 parameters. Some of the used hydrological flow characteristics include; mean annual flow, mean monthly flow, mean daily flow, flow duration curves, and baseflow indices.

The various approaches to regionalization of watersheds try to group watersheds with similar hydrologic responses based on physiographic characteristics, geographical location, climatic conditions, and hydrological response. According to Rao and Srinivas (2008) methods of regionalization of watersheds include; 1)method of residuals, 2) canonical correlation analysis, 3) region of influence, and 4) cluster analysis. A detailed analysis of each method is contained in the above analysis. Mazvimavi (2003) uses ordination techniques to select watershed characteristics that better hydrologic response for regionalization of flow prediction. Ordinate techniques used include principal component analysis, redundancy analysis, correspondence analysis, detrended correspondence analysis, and canonical correlation. Discriminant analysis tests the significance of the cluster difference; thus, each cluster represents one hydrologic region. Principal component analysis interprets the regional differences and similarities (Chiang et al., 2002).

Attributes that have been used for watershed regionalization include; physiographic characteristics, soil, landuse and landcover, drainage, geographical location, meteorology, geology, watershed response time, flood seasonality, and watershed shape indicators. The main measure of watershed homogeneity is use of flood statistics. Rao and Srinivas (2008) recommend use of flood seasonality to determine watershed homogeneity compared to other flood statistics of magnitude.

2.7 Flow duration curves

Flow duration curves (FDC) are graphical representation of the frequency of time a streamflow is equaled or exceeded over a specified historical period for a given watershed. Vogel and Fennessey (1994) defined FDC as a complement of the cumulative distribution of hourly or daily or weekly or monthly or annual streamflow that relate streamflow magnitude and frequency. Therefore, FDC represents the combined effects of climate, geology, geomorphology, soils and vegetation. The flow duration curves are used in water resource projects like estimation of streamflow at ungauged watersheds (Mohamoud, 2008), water quality management (Pomeroy et al., 2008), analysis of hydrological flow regimes (Castellarin et al., 2007), and sediment studies (Schmidt and Morche, 2006) . Additional applications of FDC can be found in Vogel and Fennessey (1995).

Construction of FDC from streamflow data is achieved using two methods. The first method is the traditional method (Castellarin et al., 2004) which ranks the streamflow data in a descending order and computes the exceedence probability based on the Weibull plotting position of the Gumbel distribution (Equation 2.1).

$$p_i = P(Q \geq q_i) = \frac{i}{N+1} \tag{2.1}$$

Where p_i is probability of exceedence, q_i is ordered streamflow, i is rank of q_i, N is total number of streamflow records, and Q is random variable of q_i.

The second approach was called the annual interpretation of flow duration curves (AFDC) by Vogel and Fennessey (1994). The method determines annual FDC for each year of the historical

record using the previous procedure. The mean or median for each quartile for the historical period is then used to generate the watershed mean or median annual FDC. Vogel and Fennessey (1994) showed that AFDC are less sensitive to the period of record compared to the traditional FDC and are effective in estimating low and flood streamflow. Some studies have normalized the streamflow data used to generate FDC and AFDC.

A common method for determining streamflow predictions at ungauged watersheds is the use regionalized flow duration curves. Castellarin et al. (2004) reviews and classifies procedures of regionalization of FDC into three groups; 1) statistical, 2) parametric, and 3) graphical procedures. The statistical approaches include use of log–normal frequency distribution (LeBoutillier and Waylen, 1993) and use of normal frequency distributions. The parametric approaches fit data to an exponential function (Quimpo et al., 1983), third–order polynomial function (Mimikou and Kaemaki, 1985) and a power function. While graphical methods use standardized curves (Gustard et al., 1992).

2.8 Causal variable selection

The need for variable selection

The availability of multiple landscape and geomorphic variables in observed hydrologic data presents a challenge of identifying patterns and relationships between causal predictor (independent) and response (dependent) variables for predictive purposes. The presence of variables with marginal relevance to the response variable may result in data over–fitting, and thus poor predictive accuracy when the resultant model is presented to new observations. Variable selection (also known as feature selection or feature extraction) addresses the above challenge. The concept of variable selection involves dimension reduction by transforming the high–dimension variable space to a subset with the same information content as the original high–dimension variable space.

The most commonly used variable selection methods include stepwise regression (Barnett et al., 2010; Brandes et al., 2005; Gong et al., 2010; Heuvelmans et al., 2006b; Peña-Arancibia

et al., 2010) and principal component analysis (Alcazar and Palau, 2010; Ma et al., 2010; Morris et al., 2009; Salas et al., 2010). Both approaches give high performance results (high coefficient of determination; $R^2 \geq 0.8$) but are susceptible to elimination of relevant variables. Stepwise regression seeks to minimize the prediction error while principal component analysis focuses on dimension reduction which may not utilize information from the response variable. Therefore, they are not structured to derive causal associations between dependent and independent variables. Also, use of a limited pool of independent variables may result in selection of irrelevant variables as relevant in absence of other relevant variables; a concept referred to as Simpson's paradox (Whittaker, 1990, pg. 24), such that two variables are marginally independent in absence of a third variable, but are dependent when conditioned on third variable.

Causation, Bayesian Networks, and Markov Blanket

Advancements in the fields of artificial intelligence, machine learning, and data mining, in addition to increased computational speed and capabilities of computers, have led to the development of algorithms that seek to infer causal associations between explanatory and response variables. Causal relationships between explanatory and response 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). 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 (Cartwright, 1979; Hitchcock, 2010; Suppes, 1970).

Causal variables for a given response variable can be inferred from a Bayesian network by constructing a Markov Blanket for the response variable. A Bayesian network is a graphical representation of a joint probability distribution over a set of random variables using a directed acyclic graph (Fu and Desmarais, 2010). The nodes of the directed acyclic graph are the random variables while the edges are the direct relationships between the variables. Given a Bayesian network, the

Markov Blanket M (Equation 2.2) of a response variable is the minimal set of explanatory (predictor) variables conditioned on which all other variables are independent of the response variable (Fu and Desmarais, 2010). Other statistical concepts considered in construction of a Markov Blanket include: i) assumptions of Markov condition property; ii) definition of faithfulness; iii) Bayes feature relevance; and iv) feature irrelevance. For detailed definitions of these concepts, the reader is referred to Fu and Desmarais (2010), Han et al. (2010) and Aliferis et al. (2010).

$$(Y \perp (X - M) \mid M) \tag{2.2}$$

Such that

$$P(Y, (X - M) \mid M) = P(Y \mid M)P((X - M) \mid M) \tag{2.3}$$

Or for

$$P((X - M) \mid M) > 0 \tag{2.4}$$

$$P(Y \mid (X - M), M) = P(Y \mid M) \tag{2.5}$$

Where X is feature vector or a vector of random variables, Y is response variable, M is subset of X also called the Markov boundary; $P(Y \mid M)$ is probability of Y given M; and $P((X - M) \mid M)$ is probability of the set difference between X and M given M.

Causal selection algorithms

The causal algorithms (discussed below) seek to select causal variables by reconstructing a Markov blanket of the response variable based on probabilistic definition of causation and variable relevance. The algorithms (defined below) differ on how explanatory variables are added to the Markov blanket and in the implementation of conditional independence tests.

Grow–Shrink, GS

The Grow–Shrink (GS) algorithm (Margaritis and Thrun, 1999) induces a Bayesian network by first identifying each node's Markov blanket and then connecting the nodes in a consistent way. The algorithm depends on two assumptions: (1) faithfulness and (2) correct conditional independence test. The algorithm first implements the growing phase where variables that form a Markov boundary of the target and some false positives are added, then the shrinking phase where the false positives are removed. The algorithm statically orders the variables based on their association with the target (T) given the empty Markov Blanket, ($MB(T)$). It then admits into $MB(T)$ the variable in the ordering that is not conditionally independent with T given the current $MB(T)$. One problem with this approach occurs when spouses form part of the $MB(T)$, they will be picked last because the association between spouses and T are weaker than associations between descendants and T. This means that more false positives will be included in the $MB(T)$ and thus the conditional independence tests will become more unreliable. The algorithm requires manually defined parameters to limit the number of conditional independence tests, and therefore, cannot always give the correct $MB(T)$.

Incremental Association Markov Blanket, IAMB and its variants

The Incremental Association Markov Blanket, IAMB (Tsamardinos et al., 2003) is similar to the GS algorithm in that it is based on the same two assumptions and has both the growing and the shrinking phases. However, instead of statically ordering the associations between variables and T in the growing phase, each time a new variable enters a candidate $MB(T)$, the algorithm reorders the variables based on the updated conditional independence test. This theoretically, allows the IAMB to outperform the GS. The IAMB is not data efficient since the conditional independence tests are conditioned on the entire $MB(T)$ which may include false positives. Therefore, several variations of the algorithm have been developed (Tsamardinos et al., 2003). The interleaved IAMB, interIAMB interleaves the growing phase of IAMB with the shrinking phase in an attempt to keep the size of the $MB(T)$ small during all the steps of the algorithm execution. This seeks to improve

the reliability of the conditional independence tests. The IAMBnPC replaces the shrinking phase of IAMB with the Peter-Clark algorithm.

Local Causal Discovery, LCD2

The Local Causal Discovery, LCD2 is an extension of the algorithm developed by Cooper (1997). The method is based on four assumptions (Mani and Cooper, 1999): 1) The Markov condition property; 2) Faithfulness between directed acyclic graph and a probability distribution; 3) the statistical test of independence on a finite dataset is approximates results on an infinite dataset; and 4) there exists an instrumental variable (feature) that is not caused by any other measured variable in the dataset. The method implements five tests of dependence and one test of independence between the instrumental variable (W), the response variable (Y), and the variable of interest (X). For details the reader is referred to Mani and Cooper (1999). If there exists a causal relationship between X and Y, then the above six tests hold.

HITON and its variants

The HITON algorithm (Aliferis et al., 2003) first induces the Markov Blanket of the variable to be predicted or classified, and then seeks to eliminate unneeded variables by using a wrapper technique (classification). The accuracy of the wrapper (classifier) is evaluated on smaller subsets of the Markov Blanket and all variables that do not affect classification are removed. A wrapper is an algorithm that solves the feature selection problem by searching in the space of feature subsets and evaluates each one with a user–specified classifier and loss function estimator. HITON accelerates the search with a number of heuristics, including limiting conditioning sets to sizes permitting the sound estimation of conditional probabilities and prioritizing candidate variables. For detailed description of the algorithm and its variants of HITON–PC and HITON–MB, refer to Aliferis et al. (2003).

Algorithm applications

The GS and interIAMBnPC algorithms have been successfully tested on their ability to recapture a Bayesian network of a medical monitoring system and hailfinder (Choi and Jun, 2010; Tsamardinos et al., 2003) while the HITON algorithms have been implemented in areas of drug discovery, clinical diagnosis, gene expression, infant mortality, Ovarian cancer, ecology, and text categorization with a variable to sample size ratio ranging between 0.67 and 60 (Aliferis et al., 2010, 2003).

BIBLIOGRAPHY

Acreman, M., Sinclair, C., 1986. Classification of drainage basins according to their physical characteristics; an application for flood frequency analysis in Scotland. Journal of Hydrology 84 (3-4), 365–380. 21

Afshar, A., Rasekh, A., Afshar, M., 2009. Risk-based optimization of large flood-diversion systems using genetic algorithms. Engineering Optimization 41 (3), 259–273. 12

Alcazar, J., Palau, A., 2010. Establishing environmental flow regimes in a Mediterranean watershed based on a regional classification. Journal of Hydrology 388 (1-2), 41–51. 24

Aliferis, C., Statnikov, A., Tsamardinos, I., Mani, S., Koutsoukos, X., 2010. Local causal and Markov blanket induction for causal discovery and feature selection for classification part I: algorithms and empirical evaluation. The Journal of Machine Learning Research 11, 171–234. 25, 28

Aliferis, C. F., Tsamardinos, I., Statnikov, A., 2003. HITON: a novel Markov Blanket algorithm for optimal variable selection. AMIA Annu Symp Proc, 21–5. 27, 28

Anderson, R., Koren, V., Reed, S., 2006. Using SSURGO data to improve Sacramento Model a priori parameter estimates. Journal of Hydrology 320 (1-2), 103–116. 13

Barnett, F., Gray, S., Tootle, G., et al., 2010. Upper Green River Basin (United States) Streamflow Reconstructions. Journal of Hydrologic Engineering 15, 567. 23

Bastola, S., Ishidaira, H., Takeuchi, K., 2008. Regionalisation of hydrological model parameters under parameter uncertainty: A case study involving TOPMODEL and basins across the globe. Journal of Hydrology 357 (3-4), 188–206. 19, 20

Betson, R., Marius, J., 1969. Source areas of storm runoff. Water Resources Research 5 (3), 574–582. 14

Beven, K., 1997. TOPMODEL: a critique. Hydrological Processes 11 (9), 1069–1085. 20

Beven, K., Freer, J., 2001a. A dynamic topmodel. Hydrological Processes 15 (10), 1993–2011. 14

Beven, K., Freer, J., 2001b. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. Journal of Hydrology 249 (1-4), 11–29. 20

Botter, G., Peratoner, F., Putti, M., Zuliani, A., Zonta, R., Rinaldo, A., Marani, M., 2008. Observation and modeling of catchment-scale solute transport in the hydrologic response: A tracer study. Water Resources Research 44 (5), W05409. 17

Brandes, D., Hoffmann, J. G., Mangarillo, J. T., 2005. Base flow recession rates, low flows, and hydrologic features of small watersheds in Pennsylvania, USA. Journal of the American Water Resources Association 41 (5), 1177–1186. 23

Brown, C., Hollis, J., 1996. SWAT: A Semi-empirical Model to Predict Concentrations of Pesticides Entering Surface Waters from Agricultural Land. Pesticide science 47 (1), 41–50. 20

Burman, R., 1969. Plot runoff using kinematic wave theory and parameter optimization. Cornell Univ. 14

Buttle, J., 1994. Isotope hydrograph separations and rapid delivery of pre-event water from drainage basins. Progress in Physical Geography 18 (1), 16. 17

Carpenter, T., Georgakakos, K., 2006. Intercomparison of lumped versus distributed hydrologic model ensemble simulations on operational forecast scales. Journal of Hydrology 329 (1-2), 174–185. 13

Cartwright, N., 1979. Causal laws and effective strategies. Nous 13 (4), 419–437. 24

Castellarin, A., Camorani, G., Brath, A., 2007. Predicting annual and long-term flow-duration curves in ungauged basins. Advances in water resources 30 (4), 937–953. 22

Castellarin, A., Galeati, G., Brandimarte, L., Montanari, A., Brath, A., 2004. Regional flow-duration curves: reliability for ungauged basins. Advances in Water Resources 27 (10), 953–965. 22, 23

Chen, J., Ning, S., Chen, H., Shu, C., 2008. Flooding probability of urban area estimated by decision tree and artificial neural networks. Journal of Hydroinformatics 10 (1), 57–67. 18

Chiang, S., Tsay, T., Nix, S., 2002. Hydrologic regionalization of watersheds. I: methodology development. Journal of water resources planning and management 128, 3. 21

Choi, Y., Jun, C., 2010. A causal discovery algorithm using multiple regressions. Pattern Recognition Letters 31 (13), 1924–1934. 28

Clark, C., 1945. Storage and the unit hydrograph. Transactions of the American Society of Civil Engineers 110, 1419–1446. 14

Cooper, G. F., 1997. A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. Data Mining and Knowledge Discovery 1, 203–224. 27

De Vos, N., Rientjes, T., 2008. Multi-objective training of artificial neural networks for rainfall–runoff modeling. Water Resources Research 44 (8), W08434. 18

Dinku, T., Anagnostou, E., Borga, M., 2002. Improving radar-based estimation of rainfall over complex terrain. Journal of Applied Meteorology 41 (12), 1163–1178. 16

Dooge, J., 1959. A general theory of the unit hydrograph. Journal of Geophysical Research 64 (2), 241–256. 14

Dunne, T., Black, R., 1970a. An experimental investigation of runoff production in permeable soils. Water Resources Research 6 (2), 478–490. 14

Dunne, T., Black, R., 1970b. Partial area contributions to storm runoff in a small New England watershed. Water Resources Research 6 (5), 1296–1311. 14

Ferraro, R., Weng, F., Grody, N., 1996. An eight-year (1987-1994) time series of rainfall, clouds, water vapor, snow cover, and sea ice derived from SSM/I measurements. Bulletin of the American Meteorological Society 77 (5). 16

Flury, M., Wai, N., 2003. Dyes as tracers for vadose zone hydrology. Rev. Geophys 41 (1), 1002. 17

Fu, S., Desmarais, M. C., 2010. Markov Blanket based Feature Selection: A Review of Past Decade. In: Proceedings of the World Congress on Engineering 2010. Vol. I. 24, 25

Garcia-Martinó, A., Scatena, F., Warner, G., Civco, D., 1996. Statistical low flow estimation using GIS analysis in humid montane regions in Puerto Rico. Water resources bulletin 32 (6), 1259–1271. 21

Gong, G., Wang, L., Condon, L., Shearman, A., Lall, U., 2010. A Simple Framework for Incorporating Seasonal Streamflow Forecasts into Existing Water Resource Management Practices1. JAWRA Journal of the American Water Resources Association 46 (3), 574–585. 23

Gotzinger, J., Bárdossy, A., 2007. Comparison of four regionalisation methods for a distributed hydrological model. Journal of Hydrology 333 (2-4), 374–384. 20

GRAHAM, L., Jacob, D., 2000. Using large-scale hydrologic modeling to review runoff generation processes in GCM climate models. Meteorologische Zeitschrift 9 (1), 49–57. 20

Gustard, A., Bullock, A., Dixon, J., 1992. Low flow estimation in the United Kingdom. Institute of Hydrology. 23

Han, B., Park, M., Chen, X., 2010. A Markov blanket-based method for detecting causal SNPs in GWAS. BMC bioinformatics 11 (Suppl 3), S5. 25

Heuvelmans, G., Muys, B., Feyen, J., 2006a. Regionalisation of the parameters of a hydrological model: Comparison of linear regression models with artificial neural nets. Journal of Hydrology 319 (1-4), 245–265. 20, 21

Heuvelmans, G., Muys, B., Feyen, J., 2006b. Regionalisation of the parameters of a hydrological model: Comparison of linear regression models with artificial neural nets. Journal of Hydrology 319 (1-4), 245–265. 23

Hitchcock, C., 2010. Probabilistic causation. Stanford Encyclopedia of Philosophy. 24

Horton, R., 1935a. Surface runoff phenomena. Vol. 1. Edwards Brothers, Inc. 14

Horton, R., 1935b. Surface runoff phenomena. Part 1. Analysis of the hydrograph. Horton Hydrological Laboratory, Publication 101. Edward Bros. Ann Arbor, Michigan. 14

Ines, A., Mohanty, B., 2008. Near-surface soil moisture assimilation for quantifying effective soil hydraulic properties using genetic algorithm: 1. Conceptual modeling. Water Resources Research 44 (6), W06422. 18

Jacques, D., Simunek, J., Mallants, D., Van Genuchten, M., 2008. Modeling coupled hydrologic and chemical processes: Long-term uranium transport following phosphorus fertilization. Vadose Zone Journal 7 (2), 698. 12

Jana, R., Mohanty, B., Springer, E., 2008. Multiscale Bayesian neural networks for soil water content estimation. Water Resources Research 44 (8), W08408. 18

Jensen, F., Nielsen, T., 2007. Bayesian networks and decision graphs. Springer Verlag. 24

Jia, Y., Culver, T., 2008. Uncertainty analysis for watershed modeling using generalized likelihood uncertainty estimation with multiple calibration measures. Journal of Water Resources Planning and Management 134, 97. 18

Jones, J., Sudicky, E., Brookfield, A., Park, Y., 2006. An assessment of the tracer-based approach to quantifying groundwater contributions to streamflow. Water Resources Research 42 (2), W02407. 17

Joseph, A., van der Velde, R., O'Neill, P., Lang, R., Gish, T., 2008. Soil moisture retrieval during a corn growth cycle using L-band (1.6 GHz) radar observations. Geoscience and Remote Sensing, IEEE Transactions on 46 (8), 2365–2374. 16

Jules, K., Lin, P., Center, N. G. R., 2002. Artificial neural networks applications: from aircraft design optimization to orbiting spacecraft on-board environment monitoring. Citeseer. 18

Kaheil, Y., Gill, M., McKee, M., Bastidas, L., Rosero, E., 2008. Downscaling and assimilation of surface soil moisture using ground truth measurements. Geoscience and Remote Sensing, IEEE Transactions on 46 (5), 1375–1384. 18

Kamp, R., Savenije, H., 2006. Optimising training data for ANNs with Genetic Algorithms. Hydrology and Earth System Sciences 10 (4), 603–608. 18

Kamruzzaman, J., Begg, R., Sarker, R., 2006. Artificial neural networks in finance and manufacturing. IGI Global. 18

Karahan, H., Ayvaz, M., 2006. Forecasting aquifer parameters using artificial neural networks. Journal of Porous Media 9 (5). 18

Karimi, K., Hamilton, H., 2009. Finding temporal relations: Causal bayesian networks vs. C4. 5. Foundations of Intelligent Systems, 266–273. 24

Kokkonen, T., Jakeman, A., Young, P., Koivusalo, H., 2003. Predicting daily flows in ungauged

catchments: model regionalization from catchment descriptors at the Coweeta Hydrologic Laboratory, North Carolina. Hydrological Processes 17 (11), 2219–2238. 19, 21

Koutsoyiannis, D., Efstratiadis, A., Georgakakos, K., 2007. Uncertainty assessment of future hydroclimatic predictions: A comparison of probabilistic and scenario-based approaches. Journal of Hydrometeorology 8 (3), 261–281. 13

Kvaerner, J., Klove, B., 2006. Tracing sources of summer streamflow in boreal headwaters using isotopic signatures and water geochemical components. Journal of Hydrology 331 (1-2), 186–204. 17

LeBoutillier, D., Waylen, P., 1993. Regional variations in flow–duration curves for rivers in British Columbia, Canada. Physical Geography 14 (4), 359–378. 23

Lee, K., Hung, W., Meng, C., 2008. Deterministic insight into ANN model performance for storm runoff simulation. Water resources management 22 (1), 67–82. 18

Ma, H., Liu, L., Chen, T., 2010. Water security assessment in Haihe River Basin using principal component analysis based on Kendall τ. Environmental monitoring and assessment 163 (1), 539–544. 24

Mani, S., Cooper, G. F., 1999. A study in causal discovery from population-based infant birth and death records. Journal of the American Medical Informatics Association, 315–319. 27

Margaritis, D., Thrun, S., 1999. Bayesian network induction via local networks. [Research paper] / Carnegie Mellon University. School of Computer Science,. School of Computer Science Carnegie Mellon University, Pittsburgh, Pa. 26

Mas, J., Flores, J., 2008. The application of artificial neural networks to the analysis of remotely sensed data. International Journal of Remote Sensing 29 (3), 617–663. 18

Mazvimavi, D., 2003. Estimation of flow characteristics of ungauged catchments. Unpublished

PhD Thesis, Wageningen University and International Institute for Geo-Information and Earth Observation, ITC, Enschede, The Netherlands. 21

Meganck, S., Leray, P., Manderick, B., 2006. Learning causal bayesian networks from observations and experiments: A decision theoretic approach. Modeling Decisions for Artificial Intelligence, 58–69. 24

Merz, R., Bloschl, G., 2004. Regionalisation of catchment model parameters. Journal of Hydrology 287 (1-4), 95–123. 19

Mimikou, M., Kaemaki, S., 1985. Regionalization of flow duration characteristics. Journal of Hydrology 82 (1-2), 77–91. 23

Mohamoud, Y., 2008. Prediction of daily flow duration curves and streamflow for ungauged catchments using regional flow duration curves/Prevision de courbes de debits classes et de debit fluviatile pour des bassins. Hydrological Sciences Journal-Journal des Sciences Hydrologiques 53 (4), 706–724. 21, 22

Morris, A. J., Donovan, J. J., Strager, M., 2009. Geospatial Analysis of Climatic and Geomorphic Interactions Influencing Stream Discharge, Appalachian Mountains, USA. Environmental Modeling and Assessment 14 (1), 73–84. 24

Moussa, R., 2008. Effect of channel network topology, basin segmentation and rainfall spatial distribution on the geomorphologic instantaneous unit hydrograph transfer function. Hydrological Processes 22 (3), 395–419. 15

Mwakalila, S., 2003. Estimation of stream flows of ungauged catchments for river basin management. Physics and Chemistry of the Earth, Parts A/B/C 28 (20-27), 935–942. 21

Naguib, R., Sherbet, G., 2000. Artificial neural networks in cancer diagnosis, prognosis, and patient management. CRC Press, Inc. 18

Nash, J., 1960. A unit hydrograph study with particular reference to British catchments. Proc. Instn Civ. Engrs 17 (5). 14

Nash, J., et al., 1958. Determining run-off from rainfall. In: ICE Proceedings. Vol. 10. Ice Virtual Library, pp. 163–184. 14

Nathan, R., McMahon, T., 1990. Evaluation of automated techniques for base flow and recession analyses. Water Resources Research 26 (7), 1465–1473. 21

Nekhay, O., Arriaza, M., Guzmán-Álvarez, J., 2009. Spatial analysis of the suitability of olive plantations for wildlife habitat restoration. Computers and Electronics in Agriculture 65 (1), 49–64. 16

Njoku, E., Jackson, T., Lakshmi, V., Chan, T., Nghiem, S., 2003. Soil moisture retrieval from AMSR-E. Geoscience and Remote Sensing, IEEE Transactions on 41 (2), 215–229. 16

Obropta, C., Kardos, J., 2007. Review of Urban Stormwater Quality Models: Deterministic, Stochastic, and Hybrid Approaches1. JAWRA Journal of the American Water Resources Association 43 (6), 1508–1523. 13

Ochoa-Rivera, J., 2008. Prospecting droughts with stochastic artificial neural networks. Journal of Hydrology 352 (1-2), 174–180. 18

Pan, F., Peters-Lidard, C., Sale, M., 2003. An analytical method for predicting surface soil moisture from rainfall observations. Water resources research 39 (11), 1314. 13

Peña-Arancibia, J., van Dijk, A., Mulligan, M., Bruijnzeel, L., Gebrehiwot, S., Ilstedt, U., Gärdenas, A., Bishop, K., Brocca, L., Melone, F., et al., 2010. The role of climatic and terrain attributes in estimating baseflow recession in tropical catchments. Hydrology and Earth System Sciences Discussions 7, 4059–4087. 23

Pomeroy, C., Postel, N., ONeill, P., Roesner, L., 2008. Development of storm-water management

design criteria to maintain geomorphic stability in Kansas City metropolitan area streams. Journal of Irrigation and Drainage Engineering 134, 562. 22

Poole, G., O'Daniel, S., Jones, K., Woessner, W., Bernhardt, E., Helton, A., Stanford, J., Boer, B., Beechie, T., 2008. Hydrologic spiralling: the role of multiple interactive flow paths in stream ecosystems. River research and applications 24 (7), 1018–1031. 12

Prabha, T., Hoogenboom, G., 2008. Evaluation of the Weather Research and Forecasting model for two frost events. Computers and Electronics in Agriculture 64 (2), 234–247. 18

Quimpo, R., Alejandrino, A., McNally, T., 1983. Regionalized flow duration for Philippines. Journal of Water Resources Planning and Management 109 (4), 320–330. 23

Rabunal, J., Dorado, J., 2006. Artificial neural networks in real-life applications. Idea Group Pub. 18

Rao, A., Srinivas, V., 2008. Regionalization of watersheds: an approach based on cluster analysis. Vol. 58. Springer Verlag. 21, 22

Reed, S., Koren, V., Smith, M., Zhang, Z., Moreda, F., Seo, D., DMIP, P., 2004. Overall distributed model intercomparison project results. Journal of Hydrology 298 (1-4), 27–60. 13

Salas, J., Fu, C., Rajagopalan, B., 2010. Long Range Forecasting of Colorado Streamflows Based on Hydrologic, Atmospheric, and Oceanic Data. Journal of Hydrologic Engineering 1, 210. 24

Sanders, B., 2007. Evaluation of on-line DEMs for flood inundation modeling. Advances in Water Resources 30 (8), 1831–1843. 16

Schmidt, K., Morche, D., 2006. Sediment output and effective discharge in two small high mountain catchments in the Bavarian Alps, Germany. Geomorphology 80 (1-2), 131–145. 22

Singh, V., 1988. Hydrologic systems. Prentice Hall. 14, 15

Singh, V., Frevert, D., 2002. Mathematical modeling of watershed hydrology. Mathematical models of large watershed hydrology, 1–22. 12

Singh, V., Frevert, D., 2006. Watershed models. CRC Press. 12, 20

Sivapalan, M., Takeuchi, K., Franks, S., Gupta, V., Karambiri, H., Lakshmi, V., Liang, X., McDonnell, J., Mendiondo, E., O'connell, P., et al., 2003. IAHS Decade on Predictions in Ungauged Basins (PUB), 2003–2012: Shaping an exciting future for the hydrological sciences. Hydrological Sciences Journal 48 (6), 857–880. 19

Smith, T., Marshall, L., 2008. Bayesian methods in hydrologic modeling: A study of recent advancements in Markov chain Monte Carlo techniques. Water Resources Research 44 (12), W00B05. 13

Spirtes, P., Glymour, C. N., Scheines, R., 2000. Causation, prediction, and search, 2nd Edition. Adaptive computation and machine learning. MIT Press, Cambridge, Mass.

Steenhuis, T., Muck, R., 1988. Preferred movement of nonadsorbed chemicals on wet, shallow, sloping soils. Journal of environmental quality (USA). 14

Suppes, P., 1970. A probabilistic theory of causality. North-Holland. 24

Tabari, H., Marofi, S., Sabziparvar, A., 2010. Estimation of daily pan evaporation using artificial neural network and multivariate non-linear regression. Irrigation Science 28 (5), 399–406. 18

Taguas, E., Ayuso, J., Pena, A., Yuan, Y., Sanchez, M., Giraldez, J., Pérez, R., 2008. Testing the relationship between instantaneous peak flow and mean daily flow in a Mediterranean Area Southeast Spain. Catena 75 (2), 129–137. 12

Tsamardinos, I., Aliferis, C., Statnikov, A., 2003. Algorithms for large scale markov blanket discovery. In: The 16th International FLAIRS Conference. Vol. 103. 26, 28

Vogel, R., Fennessey, N., 1994. Flow-duration curves. I: New interpretation and confidence intervals. Management 120 (4). 22, 23

Vogel, R., Fennessey, N., 1995. Flow Duration Curves II: A Review of Applications in Water Resources Planning. JAWRA Journal of the American Water Resources Association 31 (6), 1029–1039. 22

Weber, G., Bremer, P., Pascucci, V., 2007. Topological landscapes: A terrain metaphor for scientific data. Visualization and Computer Graphics, IEEE Transactions on 13 (6), 1416–1423. 17

Whittaker, J., 1990. Graphical models in applied multivariate statistics. Wiley series in probability and mathematical statistics: Probability and mathematical statistics. Wiley.
URL http://books.google.com/books?id=MAfvAAAAMAAJ 24

Wigneron, J., Chanzy, A., de Rosnay, P., Rudiger, C., Calvet, J., 2008. Estimating the effective soil temperature at L-band as a function of soil properties. Geoscience and Remote Sensing, IEEE Transactions on 46 (3), 797–807. 16

Winston, W., Criss, R., 2002. Geochemical variations during flash flooding, Meramec River basin, May 2000. Journal of hydrology 265 (1-4), 149–163. 17

Yu, J., Gloersen, P., 2005. Interannual variations in global SST deviations through SMMR from 1978 to 1987. International journal of remote sensing 26 (24), 5419–5431. 16

Zhang, X., Liang, F., Srinivasan, R., Van Liew, M., 2009. Estimating uncertainty of streamflow simulation using Bayesian neural networks. Water Resources Research 45 (2), W02403. 13

Zoch, R., 1934. On the relation between rainfall and stream flow. Monthly Weather Review 62, 315. 14

CHAPTER 3

ADVANCES IN VARIABLE SELECTION METHODS I: CAUSAL SELECTION METHODS VERSUS STEPWISE REGRESSION AND PRINCIPAL COMPONENT ANALYSIS

[1]Ssegane, H., Tollner E. W., Mohamoud Y. M., Rasmussen T. C., and Dowd J. F. Submitted to *Journal of Hydrology*, 06/24/2011.

Abstract

Hydrological predictions at a watershed scale are commonly based on extrapolation and upscaling of hydrological behavior at plot and hillslope scales. Yet, dominant hydrological drivers at a hillslope may not be as dominant at the watershed scale because of the heterogeneity of watershed characteristics. With the availability of quantifiable watershed data (watershed descriptors and streamflow indices), variable selection can provide insight into the dominant watershed descriptors that drive different streamflow regimes. Stepwise regression and principal components analysis have long been used to select descriptive variables for relating runoff to climate and watershed descriptors. Questions have remained regarding the robustness of the selected descriptors. This paper evaluates five new approaches: Grow-Shrink, GS; a variant of Incremental Association Markov Boundary, interIAMBnPC; Local Causal Discovery, LCD2; HITON Markov Blanket, HITON–MB; and First–Order Utility, FOU. We demonstrate their performance by quantifying their accuracy, consistency and predictive potential compared to stepwise regression and principle component analysis on two known functional relationships. The results show that the variables selected by HITON–MB and the first–order utility are the most accurate while variables selected by Stepwise regression, although not accurate have a high predictive potential. Therefore, a model with high predictive power may not necessary represent the underlying hydrological processes of a watershed system.

3.1 Introduction

Hydrological predictions in ungauged watersheds include estimation of hydrological responses in watersheds that have no flow measuring instruments, watersheds with fewer gauges compared to watershed size, and watersheds that are gauged but have few years of data (Sivapalan et al., 2003). These predictions are based on climatic inputs, land use and land cover, soil and physical descriptors, and watershed topography. Hydrological predictions are relevant to analysis of changes due to deforestation, urbanization, stream withdrawals, and installation and operation of reservoirs. Several methods are used to model hydrological behavior in ungauged watersheds. The methods include statistical regionalization (Kokkonen et al., 2003) and the use of regional hydrological model parameters (Bastola et al., 2008). Both approaches use observed data at gauged sites to conceptualize and derive underlying hydrological processes for predictions at ungauged sites.

Examination of 42 published papers (e.g. Alcazar et al., 2008; Johnston and Shmagin, 2008; Mohamoud, 2008; Sando et al., 2009) identified 72 unique topographic variables, 66 climatic variables, 98 soil variables, and 15 land use and land cover variables used by different researchers. The selection of relevant variables and whether their effect is positive or negative vary: 1) from region to region; 2) depending on the initial watershed variables used; 3) depending on the conceptualization by different researchers of what constitutes relevant variables; and 4) depending on the variable selection method. Figure 3.1 summarizes some of the watershed descriptors used in the above studies. These studies, though not extensive, provide a basis for *a priori* assumptions on the role of topography, climate, land use, and soil descriptors at different flows.

[Figure 3.1 about here]

Although the majority of the variables are statistically redundant, the challenge is to devise approaches that minimize variable redundancy and identify relevant variables that characterize the full behavior of flow regimes on a regional basis. Commonly used approaches include stepwise regression (Barnett et al., 2010; Brandes et al., 2005; Gong et al., 2010; Heuvelmans et al., 2006; Peña-Arancibia et al., 2010) and principal component analysis (Alcazar and Palau, 2010; Ma et al.,

2010; Morris et al., 2009; Salas et al., 2010). Stepwise regression seeks to minimize the prediction error while principal component analysis focuses on dimension reduction which may not utilize information from the response variable. Both approaches perform well (high coefficient of determination; $R^2 \geq 0.8$) but are susceptible to the elimination of relevant variables. Also, neither method is structured to derive causal associations between dependent and independent variables. Also, use of a limited pool of independent variables may result in selection of irrelevant variables as relevant in absence of other relevant variables; a concept referred to as Simpson's paradox (Whittaker, 1990, pg. 24). This paradox states that two variables may be marginally independent in absence of a third variable but become dependent when conditioned on the third variable.

Advancements in the fields of artificial intelligence, machine learning, and data mining, in addition to increased computational speed and capabilities of computers have led to development of methods that seek to infer causal associations between explanatory and response variables. Causal relationships between response and explanatory variables can be discovered using 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 found using Bayesian networks is a probabilistic causation based on the precept 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 (Cartwright, 1979; Hitchcock, 2010; Suppes, 1970).

Some of the causal methods include: Grow-Shrink, GS (Margaritis and Thrun, 1999); a variant of Incremental Association Markov Boundary (IAMB), interIAMBnPC (Tsamardinos et al., 2003); Local Causal Discovery, LCD2 (Cooper, 1997); HITON Markov Blanket, HITON–MB (Aliferis et al., 2003); and First Order Utility, FOU (Brown, 2009). The first four methods seek to select causal variables by reconstructing a Markov blanket of the response variable based on probabilistic definition of causation and variable relevance, while the fifth method (FOU) uses mutual information to derive variable relevance, redundancy, and conditional redundancy. The four causal

selection algorithms have two major phases; the growing phase where variables are added to a Markov blanket (MB) and a shrinking phase where false positives are removed. The GS statically orders the variables based on their association with the response variable given the empty Markov blanket (MB) and then admits into MB the variable in the ordering that is not conditionally independent with response given the current MB. The IAMB is similar to GS; however, each time a new variable enters a candidate MB, the algorithm reorders the variables based on the updated conditional independence test. The interIAMBnPC interleaves the growing phase of IAMB with the shrinking phase; however it replaces the shrinking phase of IAMB with the Peter-Clark algorithm (Spirtes et al., 2000). The LCD2 implements five tests of dependence and one test of independence between an instrumental variable, the response variable, and the variable of interest. The HITON algorithms first induce the Markov Blanket of response variable and then eliminate false positives using a wrapper. A wrapper is an algorithm that solves the variable selection problem by searching in the space of variable subsets and evaluating each one with a user specified classifier and loss function estimator (Zheng and Zhang, 2008).

A Markov blanket of a response variable is the minimal set of explanatory variables conditioned on which all other variables are independent of the response variable, while a variable is strongly relevant to the response if and only if the joint probability of the response given that variable and the remaining variables is not equal to the conditional probability of the response given the remaining variables. For details readers may refer to Fu and Desmarais (2010) and Aliferis et al. (2010). The GS and interIAMBnPC methods have been successfully tested on their ability to recapture a Bayesian network of a medical monitoring system and hailfinder (Tsamardinos et al., 2003) while the HITON methods have been implemented in areas of drug discovery, clinical diagnosis, gene expression, and text categorization with a ratio of variable to sample size ranging between 0.67 and 60 (Aliferis et al., 2003).

Therefore, the objective of this study was to assess the accuracy, consistency, and predictive potential of the five variable selection methods in comparison to stepwise regression and principal component analysis. For accuracy, all methods were evaluated for their ability to select true

variables of two known functional relationships. Regarding consistency (method reliability) and predictive potential, all methods were implemented on the two datasets of a known functional relationship in addition to watershed and streamflow data from Mid-Atlantic Piedmont ecoregion (USA). For consistence of the methods, variables selected by each method on subsequent runs when the original data sample is slightly changed were analyzed to check for ability of methods to select the same variables.

3.2 Methods

Performance on a known relationship

The methods were initially tested on two datasets with known functional relationships to assess their ability to select the true variables from a pool of variables. The first dataset consisted of data for predicting the weight of a hollow cylinder originally generated by (Wallis, 1965) and the second dataset was generated by the authors from a known functional relationship for predicting pressure drop of a fluid flowing through a circular pipe.

Weight of a hollow cylinder

The data consisted of 75 synthetic samples of 14 explanatory variables and one response variable of weight of a hollow cylinder (Wallis, 1965). Only four (radius of inside cylinder, RI; radius of outside cylinder, RO; density, D; and height, H) of the 14 explanatory variables accurately define the known functional relationship of weight of a hollow cylinder. Other variables included diagonals and surface areas of the inner (DIAGI and 2KRIH) and outer (DIAGO and 2KROH) cylinders in addition to second order powers (RI^2, RO^2, H^2, and D^2) and combinations of the above variables (DDIAGI and DDIAGO). Prior to implementation of the methods, primary variable reduction (refer to section 3.2) was carried out to minimize the effects of variable redundancy. This evaluation retained 9 of the 14 explanatory variables. Sample sizes of 10, 20, 30, 40, 50, and 60 were randomly drawn from the original data. Variable selection by each of the methods was

implemented on each sample size and method reliability or accuracy was estimated as the ratio of selected true positives to the total number of true positives. The procedure of randomly drawing different sample sizes and variable selection was repeated 40 times; from which average reliability for each method and sample size were reported (six sample sizes with 40 replications each gives 240 randomly selected samples with replacement).

Pressure drop in a circular pipe

The pressure drop (Δp) in a circular pipe can be estimated by equation 3.1 as a function of flow velocity (v), pipe length (L), fluid density (ρ), pipe diameter (d), and pipe friction factor (f). The pipe friction factor is a function of the Reynold's number (Re), while the Reynold's number is a function of fluid kinematic viscosity (ν), pipe diameter (d), and flow velocity. Seventy five samples for each of the five explanatory variables were randomly generated from a uniform distribution and pressure drop estimated using equation 3.1. Other generated explanatory variables included the flow cross section area ($A = \pi d^2/4$), volumetric flow rate ($V = v \times A$), mass flow rate ($M = \rho V$), dynamic viscosity (μ), total flow contact area (πdL).

$$\Delta p = \frac{v_2 \times f \times L \times \rho}{2d} \tag{3.1}$$

Therefore, from the relevant five variables, a total of 14 explanatory variables were generated to form a pool of potential variables that may drive the dynamics of pressure drop in a circular pipe. An initial synthetic data of 75 samples was generated. Similar variable selection procedures undertaken in the preceding subsection were implemented for this dataset.

Watershed data and representative watershed descriptors

The data used in this study consisted of 26 watersheds in the Piedmont physiographic province of the Mid-Atlantic hydrological region, USA (Figure 3.2). Streamflow data used spanned the same 42 years (1965 to 2007) across all watersheds. The selected watersheds were predominantly

forested (greater than 60 % forest cover) with lower levels of urbanization and surface storage (areal coverage of open water surfaces and wetlands). Data sources included the U.S. Geological Survey (USGS) for streamflow, the National Weather Service (NWS) for climatic data, the Natural Resources Conservation Service (NRCS) for STATSGO soil data, and the National Hydrology Dataset (NHD) compiled by USGS for sample watersheds with minimum level of urbanization and surface storage. Data preparation was achieved using readily available geographical information service (GIS) tools such as MicroDEM (U.S. Navy – public domain), ArcGIS (ESRI Inc. – proprietary), BASINS 4.0 (USEPA – public domain), and Systems for Automated Geoscientific Analyses (SAGA-GIS – public domain).

[Figure 3.2 about here]

Watershed characteristics were selected based on their likely contribution to the hydrological response as supported by information from the literature (See, e.g. Alcazar et al., 2008; Castellarin et al., 2007; Eng et al., 2007; Johnston and Shmagin, 2008; Mohamoud, 2008; Sanborn and Bledsoe, 2006; Sando et al., 2009; Srinivas et al., 2008). Tables 3.1 to 3.4 present lists of geomorphological descriptors, land use and land cover descriptors, soils and physical descriptors, and climatic descriptors used in this study to generate the original data.

[Table 3.1 about here]

[Table 3.2 about here]

[Table 3.3 about here]

[Table 3.4 about here]

Watershed data preprocessing

The initial set of variables constituted 111 parameters (41 topographic, 39 climatic, 6 land use and land cover, and 25 soil and physical parameters) for 26 piedmont watersheds (Tables 3.1 to

3.4). Only a few land use and land cover (LULC) variables because the selected watersheds were predominantly forested. A correlation matrix of the variables was generated, from which pairwise variables with a correlation coefficient greater than 0.9 were identified for primary dimension reduction. Given two highly correlated variables, the variable which provided the highest incremental gain (information gain) about the response variable was retained. The incremental gain (Schroedl, 2010) was computed as a function of: 1) mutual information between the variable and the response variable (variable relevance); 2) mutual information of different variables (variable redundancy); and 3) the increase of mutual information between previously selected variables and the response variable conditioned on a selected variable (conditional redundancy). The incremental gain of highly correlated variables was computed for 19 flow percentiles and the average value was used as the representative information gain. 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, as an example Q10 represents the flow magnitude equaled or exceeded 10 percent of the flow record (1965 to 2007). This process reduced the 111 original variables to 92 variables.

Watershed data transformation for variable selection

The usefulness of data normalization is to rescale variables of different scales of magnitude onto a similar scale such that the underlying data structure and not the magnitudes are comparable. The streamflow percentiles were normalized using drainage area to minimize its effect on variable selection and were subsequently logarithmic transformed. A minimum–maximum standardization method (equation 3.2) was then implemented on the transformed streamflow percentiles and explanatory variables.

$$F(S_k) = \frac{S_k - \min\{S\}}{\max\{S\} - \min\{S\}} \tag{3.2}$$

where $F(S_k)$ is the transformed k^{th} term of variable *S*; and S_k is the k^{th} term of variable *S* resulting in $0.0 \leq F(S_k) \leq 1.0$.

Variable Selection of watershed descriptors

The causal explorer toolkit was used to implement the variable selection methods of GS, interIAMBnPC, LCD2, and HITON–MB (Aliferis et al., 2003). The principal component analysis method (PCA) implemented in this study is based on recommendations of Lu et al. (2007). The first five principal components of the covariance matrix between transformed variables from Piedmont physiographic province were generated in the initial step. These components explained over 99 % 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 variables to the cluster centroids. The euclidean distance was used to determine the closest variables to each cluster centroid. For stepwise regression, the method was implemented to select relevant variables for each of the 19 streamflows 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 (these values were used because significance level of 0.05 did not select any variable for most streamflows).

Overall, variable selection by each method was implemented by: 1) randomly deleting a single watershed; 2) running the variable selection methods on the remaining watersheds; 3) summarizing variables selected by each method for each of the 19 flow percentiles; and 4) repeating Steps 1 to 3 twenty six times to improve the reliability of the results. The slight data perturbation was achieved by excluding a different watershed with replacement on each run. The top five variables selected by each method for each streamflow percentile were chosen based on the aggregate number of times they were selected after 26 runs.

Consistency of methods

Consistency of method (reliability) in this study refers to the ability of an method to select the same set of variables on subsequent runs when the initial sample data is slightly changed. Method reliability provides confidence on the robustness and stability of both the method and the selected variables. For example, if a method selects the same top three variables when a different watershed is removed on multiple runs, that method is assumed to be more reliable and increases the

confidence in the selected variables. Reliability of methods was estimated by computing similarity indices for variable subsets generated by the same method on subsequent runs. The sample dataset was changed by randomly excluding a single watershed from the original sample and implementing the variable selection procedure. This process was repeated 26 times and thus provided 25 variable subset pairs for comparison. Three existing measures of method reliability initially used included; hamming distance (Dunne et al., 2002, equation 3.3), similarity index (Kalousis et al., 2007, equation 3.4), and the consistency index (Kuncheva, 2007, equation 3.5).

$$S_h = 1 - \frac{|A \setminus B| + |B \setminus A|}{n} \tag{3.3}$$

$$SI = 1 - \frac{|A| + |B| - 2\,|A \cap B|}{|A| + |B| - |A \cap B|} = \frac{|A \cap B|}{|A \cup B|} \tag{3.4}$$

$$CI = \frac{n\,|A \cap B| - k^2}{nk - k^2} \tag{3.5}$$

Where $|A \setminus B|$ is cardinality of set difference of A from B; $|B \setminus A|$ is cardinality of set difference of B from A; $|A|$ is cardinality of set A; $|B|$ is cardinality of set B; $|A \cap B|$ is cardinality of set intersection of A and B; $|A \cup B|$ is cardinality of set union of A and B; n is total number of variables in the original dataset; and k is the size of features to be compared (the minimum of sizes of set A and B).

The hamming distance always yielded a high index while the similarity index was more conservative than the consistency index. The consistency index overestimated method reliability when two variable subsets had different number of variables. For example, when the first subset has four variables, the second subset has six, and the intersection between the two subsets has four, then the computed consistency index is one (1.0) yet the two variable subsets do not exactly match. Therefore, this study proposed a fourth index that accounts for the cardinality of the intersection of two variable sets, the cardinality of the set differences, and unequal number of variables.

The proposed index assumes that the probability of a random method (which does not consider causality or correlation) to generate two variable subsets with similar features (cardinality of intersection) is low while the probability of generating different variable subsets is high. Therefore, the cardinality of the set intersection was given a higher weight than the cardinality of the set differences. This study used weighting factors of two for the cardinality of set intersection and one for the set differences. Because these preliminary results showed comparable performance between SI, CI, and RI, subsequent analysis was based on RI.

$$RI = \frac{1}{2}\left(1 - \frac{|A \setminus B| + |B \setminus A| - 2\,|A \cap B|}{|A| + |B|}\right) \tag{3.6}$$

Predictive potential of selected variables

To examine the predictive performance of the selected variables by each method, a two–layer feed forward back propagation artificial neural network was used to generate a predictive model. The choice was based on the increased use of artificial neural networks for prediction and forecasting in water resources (Chiang and Chang, 2009; He et al., 2011; Maier et al., 2010; Tiwari and Chatterjee, 2010) and the approach tries to capture linearities and non–linearities of a system without knowing the functional relationship.

Each dataset was split into training (70 %) and validation (30 %) subsets. The same training and validation data was used for variables selected by different methods. The performance of each trained model on the validation data was used as the predictive potential of selected variables. The performance metrics used included the coefficient of determination (R^2), Nash–Sutcliffe coefficient of efficiency (NSE), mean absolute error (MAE), and root mean square error (RMSE). The R^2 varies from zero to one, the NSE varies from negative infinity to one, while the MAE and $RMSE$ depend on the magnitude of the response variable. For both R^2 and NSE, a value of one is optimum while for MAE and $RMSE$ greater prediction power of the selected variables is indicated by smaller values.

3.3 Results

Data of known functional relationships

Accuracy and predictive potential of selected variables

Tables 3.5 and 3.6 depict the top–four and top–five most selected variables by each method after 240 runs (40 runs for each of the six sample sizes) for the hollow cylinder and pressure drop data, respectively. For the 240 runs of the hollow cylinder data, the first order utility (FOU) method selected at least three of the four true variables on 111 runs while the HITON–MB selected these on 37 runs, and the stepwise regression on 9 runs. Aggregation of the top–four selected variables by each method gave three true variables by FOU, two by HITON–MB and LCD2, and none by all other methods. The prediction performance using a feed–forward back–propagation neural network was highest for variables selected by GS, interIAMBnPC, and stepwise regression while lowest for true variables (refer to Table 3.7).

[Table 3.5 about here]

[Table 3.6 about here]

[Table 3.7 about here]

With respect to the 240 runs of the pressure drop data, the first order utility (FOU) method selected at least four of the five true variables on 132 runs while the HITON–MB on 50 runs, and the stepwise regression on 9 runs. Aggregation of the top five selected variables by each method gave four true variables by FOU, four by HITON–MB, three by LCD2, PCA, Stepwise, and interIAMBnPC, and two by GS. The prediction performance using a feed–forward back–propagation neural network was highest for variables selected by HITON–MB, true variable, GS, interIAMBnPC, and PCA while lowest for variables selected by FOU (refer to Table 3.8).

[Table 3.8 about here]

Consistency of the selection methods

Tables 3.9 and 3.10 show reliability indices of variable selection methods as a function of the sample size for the hollow cylinder and pressure drop data, respectively. The reliability index of stepwise regression, HITON–MB, FOU, and LCD2 increased with increasing sample size while the reliability index of GS, interIAMBnPC, and PCA were not significantly affected by sample size. On average, the LCD2, PCA, and the HITON–MB were the most consistent methods while the GS and interIAMBnPC were the least consistent.

[Table 3.9 about here]

[Table 3.10 about here]

Piedmont streamflow percentiles; unknown functional relationship

Consistency and prediction potential

Figure 3.3 compares consistency of the methods using the reliability index (RI) for selected watershed descriptors across 19 streamflow percentiles. Each data point is an average of 25 pairwise comparisons for a single flow percentile. Values below 0.6 depicted low method reliability (robustness) while values equal or greater than 0.6 depicted high reliability. A reliability of 0.6 showed that, on average, the method selected the same three variables of the top five variables on each of the 25 pairwise comparisons. There was no single method that was consistently more reliable than others across all flows. However, on average, LCD2, PCA, and HITON–MB were the most consistent at high, medium, and low flows (more points with $RI \geq 0.8$). The first order utility (FOU) was the least consistent.

[Figure 3.3 about here]

Figure 3.4 shows the average coefficients of determination (R^2) for each method and for each streamflow percentile. The simulated flow percentiles for comparison with observed were based

on flow percentile prediction using selected variables by each method and a feed forward back propagation neural network. Based on data in the Figure 3.4, one can observe that there was no single method that consistently outperformed other methods across all flows. However, on average the variables selected by GS, interIAMBnPC, and stepwise regression gave the best streamflow predictions across high, medium, and low flows (more points with $R^2 \geq 0.8$).

[Figure 3.4 about here]

Selected watershed descriptors

Table 3.11 shows the percent proportion of variable classes of the most selected variables across high flows (Q0.01 to Q10; 7 flows), medium flows (Q20 to Q70; 6 flows), and low flows (Q80 to Q99.9; 6 flows). For high flows, topographic variables were the most commonly selected variables by all methods while topographic and soil variables were the most commonly selected variables for medium and low flows.

[Table 3.11 about here]

Table 3.12 shows top–five most selected variables (watershed descriptors) by each method for five sample streamflow percentiles. There was some level of consistence with respect to variables selected by different methods for the same flow percentile. For example four of the six methods selected monthly February precipitation (FEBP) for Q10, and three methods selected topographic wetness index (TWI), Porosity, percent urban coverage (Urban), and convergence index (CI) for various flows.

[Table 3.12 about here]

3.4 Discussion

Accuracy, consistency and predictive potential of methods

The HITON-MB and first order utility (FOU) methods selected the most true variables based on results of the most selected variables for data of known functional relationships. Since the reliability index of HITON–MB was relatively better and consistent for data of known and unknown relationships compared to the FOU, variables selected by HITON–MB were assumed to have a higher probability of being causal compared to those selected by FOU. The low reliability indices by FOU on watershed data depict failure of the method to capture uncertainties introduced by the statistical nature of the data (watershed average values).

The relatively high reliability values for the LCD2 were attributed to its ability to consistently identify at least one relevant variable on multiple runs while the GS and the interIAMBnPC did not identify relevant variables on most runs. The poor reliability results by the GS and interIAMB-nPC are attributed to more false positives the methods identify during the growing phase such that independent tests are likely to be unreliable during the shrinking phase. And thus, the final selected variables have more false positives. The consistently high reliability of variables selected by HITON–MB for all flow ranges was attributed to its intrinsic structure which does not require conditioning on the entire Markov blanket to determine conditional independence. Any false positives can be removed by the wrapper (Fu and Desmarais, 2010).

The most selected variables were determined by aggregating selected variables by all methods across high, medium, and low streamflow percentiles. For high, medium, and low flows, the most selected topographic variables were related to the control of rate of flow accumulation and transport. Also, the most selected soil variables for high flows were connected to water accumulation and movement. For medium and low flows, the most selected soil variables control subsurface flow and flow under unsaturated conditions and emphasize the relevance of soil pore connectivity and soil structure with regard to water movement.

The coefficient of determination (R^2) and Nash–Suticliffe efficiency (NSE) values for data of a known functional relationship show that surrogate variables provide higher predictive performance than true variables. For example, the most selected variables by GS, interIAMBnPC, and stepwise regression for the hollow cylinder showed greater predictive potential (Table 3.7; $R^2 = 0.82$ and $NSE = 0.81$) than the true variables ($R^2 = 0.71$ and $NSE = 0.67$), while none of the methods selected any true variable. This observation is attributed to the inability of the neural network to capture the underlying functional relationship. However, the neural networked gave high prediction for pressure drop data (Table 3.8;$R^2 = 0.97$ and $NSE = 0.96$) using the true variables. Even for the pressure drop data the highest predictive variables included a surrogate variable ($R^2 = 0.99$ and $NSE = 0.96$ for variables selected by HITON–MB).

Two major challenges are exhibited by these results; 1) the need to select the true system variables; and 2) the need to determine the underlying functional relationship. Caution should be used when evaluating hydrological predictive models with a high coefficient of determination or Nash–Suticliffe efficiency because results demonstrate that high predictive potential does not infer system representation. A key observation is that the two most accurate methods (HITON–MB and FOU) use discrete data compared to other methods that use continuous data.

3.5 Conclusions

This paper compares the accuracy, consistency, and predictive potential of variables selected by stepwise regression and principal component analysis to variables selected by five methods that seek to infer causal associations between explanatory and response variables. For accuracy, data of two known functional relationships: weight of a hollow cylinder and pressure drop of a fluid within a circular pipe were used. For consistency and predictive potential, data of known and unknown functional relationships were used. The unknown functional relationship consisted of 26 Mid–Atlantic Piedmont watersheds with 111 watershed descriptors and 19 streamflow percentiles.

The accuracy of some causal selection methods is greater than others. Overall, the HITON–MB and first order utility (FOU) methods are the most accurate followed by principal component analysis (PCA). The accuracy of the Grow–Shrink (GS) and a variant of the incremental association Markov boundary (interIAMBnPC) were not better than the accuracy of the stepwise regression. Because of the high accuracy of the HITON–MB and its high consistency on data of known and unknown functional relationship, variables selected by this method have a high probability of being causal compared to stepwise regression. The authors recommend use of more than one selection method to improve the reliability of the selected variables. Future efforts should focus on quantifying the probability that a selected variable for a specific response variable is causal based on selection accuracy of various methods. Data of known functional relationships with varying system complexities should be used.

BIBLIOGRAPHY

Alcazar, J., Palau, A., 2010. Establishing environmental flow regimes in a Mediterranean watershed based on a regional classification. Journal of Hydrology 388 (1-2), 41–51. 43

Alcazar, J., Palau, A., Vega-Garcia, C., 2008. A neural net model for environmental flow estimation at the Ebro River Basin, Spain. Journal of hydrology 349 (1-2), 44–55. 43, 48

Aliferis, C., Statnikov, A., Tsamardinos, I., Mani, S., Koutsoukos, X., 2010. Local causal and Markov blanket induction for causal discovery and feature selection for classification part I: algorithms and empirical evaluation. The Journal of Machine Learning Research 11, 171–234. 45

Aliferis, C., Tsamardinos, I., Statnikov, A., Brown, L., 2003. Causal explorer: A causal probabilistic network learning toolkit for biomedical discovery. In: International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS03). Citeseer, pp. 371–376. 44, 45, 50

Barnett, F., Gray, S., Tootle, G., et al., 2010. Upper Green River Basin (United States) Streamflow Reconstructions. Journal of Hydrologic Engineering 15, 567. 43

Bastola, S., Ishidaira, H., Takeuchi, K., 2008. Regionalisation of hydrological model parameters under parameter uncertainty: A case study involving TOPMODEL and basins across the globe. Journal of Hydrology 357 (3-4), 188–206. 43

Brandes, D., Hoffmann, J. G., Mangarillo, J. T., 2005. Base flow recession rates, low flows, and

hydrologic features of small watersheds in Pennsylvania, USA. Journal of the American Water Resources Association 41 (5), 1177–1186. 43

Brown, G., 2009. A new perspective for information theoretic feature selection. In: 12th International Conference on Artificial Intelligence and Statistics. Vol. 5. Citeseer, pp. 49–56. 44

Cartwright, N., 1979. Causal Laws and Effective Strategies. Nous 13 (4), pp. 419–437. URL http://www.jstor.org/stable/2215337 44

Castellarin, A., Camorani, G., Brath, A., 2007. Predicting annual and long-term flow-duration curves in ungauged basins. Advances in Water Resources 30 (4), 937–953. 48

Chiang, Y., Chang, F., 2009. Integrating hydrometeorological information for rainfall-runoff modelling by artificial neural networks. Hydrological Processes 23 (11), 1650–1659. 52

Cooper, G. F., 1997. A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. Data Mining and Knowledge Discovery 1, 203–224. 44

Dunne, K., Cunningham, P., Azuaje., F., 2002. Solution to instability problems with sequential wrapper-based approaches to feature selection. Tech. rep., Department of Computer Science, Trinity College. 51

Eng, K., Milly, P. C. D., Tasker, G. D., 2007. Flood regionalization: A hybrid geographic and predictor-variable region-of-influence regression method. Journal of Hydrologic Engineering 12 (6), 585–591. 48

Fu, S., Desmarais, M. C., 2010. Markov Blanket based Feature Selection: A Review of Past Decade. In: Proceedings of the World Congress on Engineering 2010. Vol. I. 45, 56

Gong, G., Wang, L., Condon, L., Shearman, A., Lall, U., 2010. A Simple Framework for Incorporating Seasonal Streamflow Forecasts into Existing Water Resource Management Practices1. JAWRA Journal of the American Water Resources Association 46 (3), 574–585. 43

He, J., Valeo, C., Chu, A., Neumann, N., 2011. Prediction of event-based stormwater runoff quantity and quality by ANNs developed using PMI-based input selection. Journal of Hydrology. 52

Heuvelmans, G., Muys, B., Feyen, J., 2006. Regionalisation of the parameters of a hydrological model: Comparison of linear regression models with artificial neural nets. Journal of Hydrology 319 (1-4), 245–265. 43

Hitchcock, C., 2010. Probabilistic causation. Stanford Encyclopedia of Philosophy. 44

Jensen, F., Nielsen, T., 2007. Bayesian networks and decision graphs. Springer Verlag. 44

Johnston, C. A., Shmagin, B. A., 2008. Regionalization, seasonality, and trends of streamflow in the US Great Lakes Basin. Journal of Hydrology 362 (1-2), 69–88. 43, 48

Kalousis, A., Prados, J., Hilario, M., 2007. Stability of feature selection algorithms: a study on high-dimensional spaces. Knowl. Inf. Syst. 12 (1), 95–116. 51

Karimi, K., Hamilton, H., 2009. Finding temporal relations: Causal bayesian networks vs. C4. 5. Foundations of Intelligent Systems, 266–273. 44

Kokkonen, T. S., Jakeman, A. J., Young, P. C., Koivusalo, H. J., 2003. Predicting daily flows in ungauged catchments: model regionalization from catchment descriptors at the Coweeta Hydrologic Laboratory, North Carolina. Hydrological Processes 17 (11), 2219–2238. 43

Kuncheva, L. I., 2007. A stability index for feature selection. ACTA Press, pp. 390–395. 51

Lu, Y., Cohen, I., Zhou, X., Tian, Q., 2007. Feature selection using principal feature analysis. In: Proceedings of the 15th international conference on Multimedia. ACM, pp. 301–304. 50

Lyon, J. G., 2003. GIS for water resources and watershed management. CRC Press, Boca Raton, Fla. 67

Ma, H., Liu, L., Chen, T., 2010. Water security assessment in Haihe River Basin using principal component analysis based on Kendall τ. Environmental monitoring and assessment 163 (1), 539–544. 43

Maier, H., Jain, A., Dandy, G., Sudheer, K., 2010. Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions. Environmental Modelling & Software 25 (8), 891–909. 52

Margaritis, D., Thrun, S., 1999. Bayesian network induction via local networks. [Research paper] / Carnegie Mellon University. School of Computer Science,. School of Computer Science Carnegie Mellon University, Pittsburgh, Pa. 44

Meganck, S., Leray, P., Manderick, B., 2006. Learning causal bayesian networks from observations and experiments: A decision theoretic approach. Modeling Decisions for Artificial Intelligence, 58–69. 44

Mohamoud, Y. M., 2008. Prediction of daily flow duration curves and streamflow for ungauged catchments using regional flow duration curves. Hydrological Sciences Journal-Journal Des Sciences Hydrologiques 53 (4), 706–724. 43, 48

Morris, A. J., Donovan, J. J., Strager, M., 2009. Geospatial Analysis of Climatic and Geomorphic Interactions Influencing Stream Discharge, Appalachian Mountains, USA. Environmental Modeling and Assessment 14 (1), 73–84. 44

Peña-Arancibia, J., van Dijk, A., Mulligan, M., Bruijnzeel, L., Gebrehiwot, S., Ilstedt, U., Gärdenas, A., Bishop, K., Brocca, L., Melone, F., et al., 2010. The role of climatic and terrain attributes in estimating baseflow recession in tropical catchments. Hydrology and Earth System Sciences Discussions 7, 4059–4087. 43

Rawls, W. J., Brakensiek, D. L., Saxton, K. E., 1982. Estimation of Soil-Water Properties. Transactions of the ASAE 25 (5), 1316–1328. 70

Salas, J., Fu, C., Rajagopalan, B., 2010. Long Range Forecasting of Colorado Streamflows Based on Hydrologic, Atmospheric, and Oceanic Data. Journal of Hydrologic Engineering 1, 210. 44

Sanborn, S. C., Bledsoe, B. P., 2006. Predicting streamflow regime metrics for ungauged streams in Colorado, Washington, and Oregon. Journal of Hydrology 325 (1-4), 241–261. 48

Sando, S. K., Fish, U., Service., W., (U.S.), G. S., 2009. Estimation of streamflow characteristics for Charles M. Russell National Wildlife Refuge, northeastern Montana. Scientific investigations report. U.S. Geological Survey, Reston, Va. 43, 48

Saxton, K., Romberger, W., Papendick, J., et al., 1986. Estimating Generalized Soil-water Characteristics from Texture1. Soil Science Society of America Journal 50 (4), 1031. 69

Schroedl, S., 2010. Feature Selection Based on Interaction Information 2010.
URL http://www.mathworks.com/matlabcentral/fileexchange/26981-feature-selection-based-on-interaction-information/content/select_features/html/demo_feature_select.html 49

Sivapalan, M., Takeuchi, K., Franks, S. W., Gupta, V. K., Karambiri, H., Lakshmi, V., Liang, X., McDonnell, J. J., Mendiondo, E. M., O'Connell, P. E., Oki, T., Pomeroy, J. W., Schertzer, D., Uhlenbrook, S., Zehe, E., 2003. IAHS decade on Predictions in Ungauged Basins (PUB), 2003-2012: Shaping an exciting future for the hydrological sciences. Hydrological Sciences Journal-Journal Des Sciences Hydrologiques 48 (6), 857–880. 43

Spirtes, P., Glymour, C., Scheines, R., 2000. Causation, prediction, and search. 45

Srinivas, V. V., Tripathi, S., Rao, A. R., Govindaraju, R. S., 2008. Regional flood frequency analysis by combining self-organizing feature map and fuzzy clustering. Journal of Hydrology 348 (1-2), 148–166. 48

Suppes, P., 1970. A probabilistic theory of causality. North-Holland. 44

Tiwari, M., Chatterjee, C., 2010. Uncertainty assessment and ensemble flood forecasting using bootstrap based artificial neural networks (BANNs). Journal of Hydrology 382 (1-4), 20–33. 52

Tsamardinos, I., Aliferis, C., Statnikov, A., 2003. Algorithms for large scale markov blanket discovery. In: The 16th International FLAIRS Conference. Vol. 103. 44, 45

Wallis, J., 1965. Multivariate statistical methods in hydrologya comparison using data of known functional relationship. Water Resources Research 1 (4), 447–461. 46

Whittaker, J., 1990. Graphical models in applied multivariate statistics. Wiley series in probability and mathematical statistics: Probability and mathematical statistics. Wiley.
URL http://books.google.com/books?id=MAfvAAAAMAAJ 44

Zavoianu, I., 1985. Morphometry of drainage basins. Developments in water science ;. Elsevier, Amsterdam ; Oxford ; New York [etc.]. 67

Zheng, H., Zhang, Y., 2008. Feature selection for high-dimensional data in astronomy. Advances in Space Research 41 (12), 1960–1964. 45

Table 3.1: Topographic descriptors

Variable	Units	Description
DA	km^2	drainage area
EMEAN	*m*	mean elevation
EMAX	*m*	maximum elevation
EMIN	*m*	minimum elevation
EMED	*m*	median elevation
ESTD	*m*	standard deviation of elevation
RLF	*m*	relief; $RLF = EMAX - EMIN$
SMEAN	*m/km*	mean slope
SMAX	*m/km*	maximum slope
SSTD	*m/km*	standard deviation of slope
CPLAN	–	plan curvature; rate of change of aspect along a contour. It measures the propensity of water to converge as it flows across the land
TWI	–	topographic wetness index
OLFD	*m*	overland flow distance = overland flow distance to the stream network
SLEN	*m*	slope length; distance from the point of origin of overland flow to the point where either the slope gradient decreases enough that deposition begins, or the runoff water enters a well-defined channel
MRVBF	–	multi resolution index of valley bottom flatness; measures the extent of watershed valley bottoms at multiple DEM resolutions
MRRTF	–	multi resolution index of ridge top flatness; measures the extent of watershed ridge tops at multiple DEM resolutions.

Table 3.1: continued on next page

Variable[a]	Units	Description
SHGT	*m*	slope height = average watershed peak heights.
VDEP	*m*	valley depth = average watershed valley depths.
TSL	*km*	total stream length
MCL	*km*	main channel length; longest drainage path from watershed divide to outlet
MCS	*m/km*	main channel slope; $MCS = (E_{divide} - E_{outlet})/MCL$
AMEAN	*degree*	average aspect
HPC10	*m*	hypsometric curve elevation corresponding to relative watershed area of 0.1
HPC50	*m*	hypsometric curve elevation corresponding to relative watershed area of 0.5
HPC90	*m*	hypsometric curve elevation corresponding to relative watershed area of 0.9
MCSR	*km/km*	Main channel sinuosity; $MCSR = MCL/BL$; BL = watershed length
CI	–	convergence index
BP	*km*	watershed perimeter
BL	*km*	watershed length; distance of straight line from outlet to intersection of watershed divide and longest drainage path
BW	*km*	watershed width; $BW = DA/BL$
SF	*km/km*	shape factor; $SF = BL/BW$
ER	*km/km*	elongation ratio; $ER = \left[\frac{4DA}{\pi BL^2}\right]^{\frac{1}{2}}$
RB	km^2/km^2	watershed rotundity ratio (Lemniscate index); $RB = \frac{\pi BL^2}{4DA}$
CR	*km/km*	compactness ratio (Gravelius shape index); $CR = 0.282\frac{BP}{\sqrt{DA}}$

Table 3.1: continued on next page

Variable[a]	Units	Description
RR	*m/km*	relief ratio; $RR = \frac{RLF}{BP}$
DD	km/km^2	drainage density; $DD = \frac{TSL}{DA}$
RN	$m.km/km^2$	ruggedness number; $RN = DD \times RLF$
SR	–	slope ratio; $SR = \frac{MCS}{SMEAN}$
HI	*m/m*	hypsometric integral; $HI = \frac{EMEAN-EMIN}{EMAX-EMIN}$
HFF	km^2/km^2	Hortons form factor; $HFF = \frac{DA}{BL^2}$
RC	km^2/km^2	circularity ratio; $RC = \frac{4\pi DA}{BP^2}$

[a] Definitions of watershed shape and channel parameters (variables BL to RC) are given in Lyon (2003) pages 99 to 111 and Zavoianu (1985) pages 39 to 41.

Table 3.2: Land use and land cover descriptors

Variable	Units	Description
Water	%	open water, generally with less than 25% cover of vegetation or soil
Urban	%	developed areas of low, medium, and high intensity
Barren	%	barren areas of bedrock and unconsolidated shores
Forest	%	deciduous forest, evergreen forest, mixed forest, shrub/scrub, and grassland/herbaceous
Agric	%	pasture or hay and cultivated crops
Wetland	%	all types of wetlands

Table 3.3: Soil and physical descriptors

Variable	Units	Description
WTD	*cm*	water table depth
rockDep	*cm*	depth to bedrock
hsgA	%	hydrological soil group A
hsgB	%	hydrological soil group B
hsgC	%	hydrological soil group C
hsgD	%	hydrological soil group D
Hydric	%	hydric soils
Sdepth	*cm*	soil depth
KFACT	–	soil erodibility factor without rocks
KFFACT	–	soil erodibility factor with rocks
Sand	%	sand
Silt	%	silt
Clay	%	clay
AWC	*cm/cm*	available water content
BulkD	g/cm^3	bulk density
OM	%	percent organic matter
PERM	cm/hr	permeability from STATSGO data
KSAT	cm/hr	saturated hydraulic conductivity by pedotransfer function (Saxton et al., 1986)
SAT	$cm^3 cm^3$	saturation

Table 3.3: continued on next page

Variable	Units	Description
Porosity	–	porosity
Void	–	void ratio
PSDI[a]	–	pore size distribution index
MSCL[a]	*cm*	macroscopic capillary length
T	cm^2/hr	transmissivity
STORG	*cm*	storage; $STORG = void \times Sdepth$

[a] The PSDI and MSCL are functions of the soil texture based on work by Rawls et al. (1982).

Table 3.4: Climatic descriptors

Variable	Units	Description
mP	*mm*	monthly precipitation (January to December; 12 variables)
mET	*mm*	monthly evapotranspiration (January to December; 12 variables)
MAP	*mm*	mean annual precipitation
MMP	*mm*	mean monthly precipitation (MAP/12)
MAET	*mm*	mean annual potential evapotranspiration
MMET	*mm*	mean monthly potential evapotranspiration
NAP	*mm*	net annual precipitation ; $NAP = MAP - MAET$
NMP	*mm*	net monthly precipitation ; $NMP = MMP - MMET$
ADI	*mm*	annual dryness index ; $ADI = \frac{MAP}{MAET}$
PI	–	Prescott index; $PI = \frac{0.445MAP}{MAET^{0.75}}$
RFx	*mm*	Rainfall amount equaled or exceeded x % of the record time (x=[0.01, 0.05, 0.1, 0.5, 1, 5, 10, 20]; 7 variables)

Table 3.5: Selection results for data of a hollow cylinder

Method	Most selected variables[a]			
GS	2KROH	DDIAGO	DDIAGI	2KRIH
interIAMBnPC	2KROH	DDIAGO	2KRIH	DDIAGI
LCD2	DDIAGO	2KROH	**D**	DDIAGI
HITON-PC	DDIAGO	2KROH	**D**	**H**
FOU	**RO**	DDIAGO	**RI**	**H**
PCA	DDIAGO	2KROH	DIAGI	**D**
STEPWISE	DDIAGO	2KROH	2KRIH	DDIAGI
True variables	**D**	**H**	**RI**	**RO**

[a] The bold variables are the true variables

Table 3.6: Selection results for pressure drop data

Method	Most selected variables				
GS	**v**	Re	Fr	**f**	Q
interIAMBnPC	**v**	Fr	Q	**L**	**D**
LCD2	**f**	**v**	**d**	Fr	Re
HITON-PC	**v**	**d**	**f**	Fr	**L**
FOU	**v**	**D**	Q	**f**	**d**
PCA	Re	**L**	**v**	KV	**D**
STEPWISE	Q	Fr	**d**	**D**	**f**
True variables	**v**	**d**	**f**	**D**	**L**

Table 3.7: Prediction performance for hollow cylinder data

Method	R^2	NSE	MAE (kg)	RMSE (kg)
GS	0.82	0.81	35	125
interIAMBnPC	0.82	0.81	35	125
LCD2	0.77	0.75	66	140
HITON-MB	0.77	0.74	82	144
FOU	0.76	0.73	97	147
PCA	0.8	0.77	67	137
STEPWISE	0.82	0.81	35	125
True variables	**0.71**	**0.67**	**100**	**163**

Table 3.8: Prediction performance for pressure drop data

Method	R^2	NSE	MAE (kPa)	RMSE (kPa)
GS	0.94	0.92	474	721
interIAMBnPC	0.96	0.89	468	787
LCD2	0.85	0.81	573	1062
HITON-MB	0.99	0.96	320	465
FOU	0.46	0.42	1242	1827
PCA	0.86	0.84	629	960
STEPWISE	0.85	0.83	640	1001
True variables	**0.97**	**0.96**	**198**	**482**

Table 3.9: Reliability index[a] for hollow cylinder data

Sample[b]	GS	IAMB[c]	LCD2	HITON[d]	FOU	PCA	STEPWISE
10	0.14	0.07	0.11	0.17	0.49	0.64	0.27
20	0.43	0.19	0.36	0.41	0.48	0.51	0.65
30	0.52	0.19	0.59	0.6	0.47	0.61	0.62
40	0.52	0.19	0.62	0.67	0.52	0.6	0.63
50	0.52	0.19	0.8	0.74	0.63	0.73	0.59
60	0.52	0.19	0.97	0.8	0.71	0.73	0.74
Average[e]	0.49	0.19	0.71	0.67	0.59	0.66	0.64

[a] Each value is an average of 40 values

[b] Randomly selected sample size from initial data

[c] interIAMBnPC

[d] HITON–MB

[e] Weighted average by sample size

Table 3.10: Reliability index for pressure drop data

Sample	GS	IAMB	LCD2	HITON	FOU	PCA	STEPWISE
10	0.49	0.38	0.47	0.41	0.47	0.83	0.45
20	0.49	0.5	0.65	0.51	0.46	0.84	0.37
30	0.49	0.48	0.73	0.58	0.53	0.82	0.4
40	0.49	0.48	0.78	0.6	0.53	0.83	0.42
50	0.49	0.47	0.91	0.66	0.6	0.86	0.56
60	0.49	0.48	0.99	0.78	0.69	0.86	0.72
Average	0.49	0.48	0.83	0.64	0.58	0.84	0.53

Table 3.11: Percent proportion of variable classes of most selected variables

	GS	IAMB	LCD2	HITON	FOU	Stepwise
			High flows			
Climatic	16.7	14.7	25	34.6	40.4	7.5
LULC	18.3	21.2	15.7	5.5	11.2	6.1
Soil	26.4	12	28.4	19	6	15.5
Topography	38.7	52.1	30.8	40.9	42.4	70.8
			Medium flows			
Climatic	10.6	36.9	4	19.5	40.5	28.9
LULC	0	0	6.4	11.7	4.4	0
Soil	59.5	22.5	51.9	20.8	28.4	21.9
Topography	29.8	40.5	37.8	48	26.7	49.2
			Low flows			
Climatic	17	21	34.8	15.4	44.4	11.4
LULC	0	0	0	23	14.5	0
Soil	40.1	30.1	2.1	37.6	19	27.9
Topography	42.9	48.9	63.1	24	22.2	60.7

Table 3.12: Selected variables for sample streamflow percentiles

Flow	GS	IAMB	LCD2	HITON	FOU	Stepwise
	TWI	VDEP	rockDep	VDEP	Wetland	MRVBF
	Urban	Urban	MRVBF	RF0.05	MCSR	MCL
Q1	rockDep	MCL	Wetland	TWI	AUGP	CPLAN
	NOVP	CPLAN	MCSR	APRP	RF0.05	Urban
	hsgC	TWI	TWI	EMIN	RF0.01	VDEP
	FEBP	MAYP	Forest	EMIN	JANP	FEBP
	rockDep	Water	DecET	OLFD	SLEN	Wetland
Q10	PSDI	Wetland	MarET	hsgB	APRP	Forest
	OM	FEBP	OM	JANP	OLFD	EMIN
	JUNP	OM	NovET	FEBP	SR	MMET
	Porosity	JUNP	TWI	Wetland	OM	TWI
	CI	MSCL	MRRTF	CI	SMEAN	hsgC
Q50	KSAT	Porosity	Clay	APRP	CPROF	JUNP
	Sand	hsgC	KSAT	hsgD	RF20	CI
	AWC	MAYP	Porosity	HCP90	AUGP	MAYP
	MSCL	MSCL	EMED	Rock	T	VDEP
	EMED	VDEP	MRRTF	RF20	AugET	MSCL
Q90	CI	AUGP	TWI	KSAT	Rock	RF0.1
	Porosity	CI	CI	T	RF20	HI
	MRRTF	Porosity	RF20	BW	SEPP	PERM
	Porosity	MSCL	EMED	Rock	MCL	MSCL
	MSCL	AMEAN	AugET	MCL	OLFD	HI
Q99	EMED	VDEP	RF20	RF20	SMEAN	VDEP
	AugET	AUGP	MMET	T	Wetland	RF10
	CI	EMIN	Porosity	AWC	FEBP	NOVP

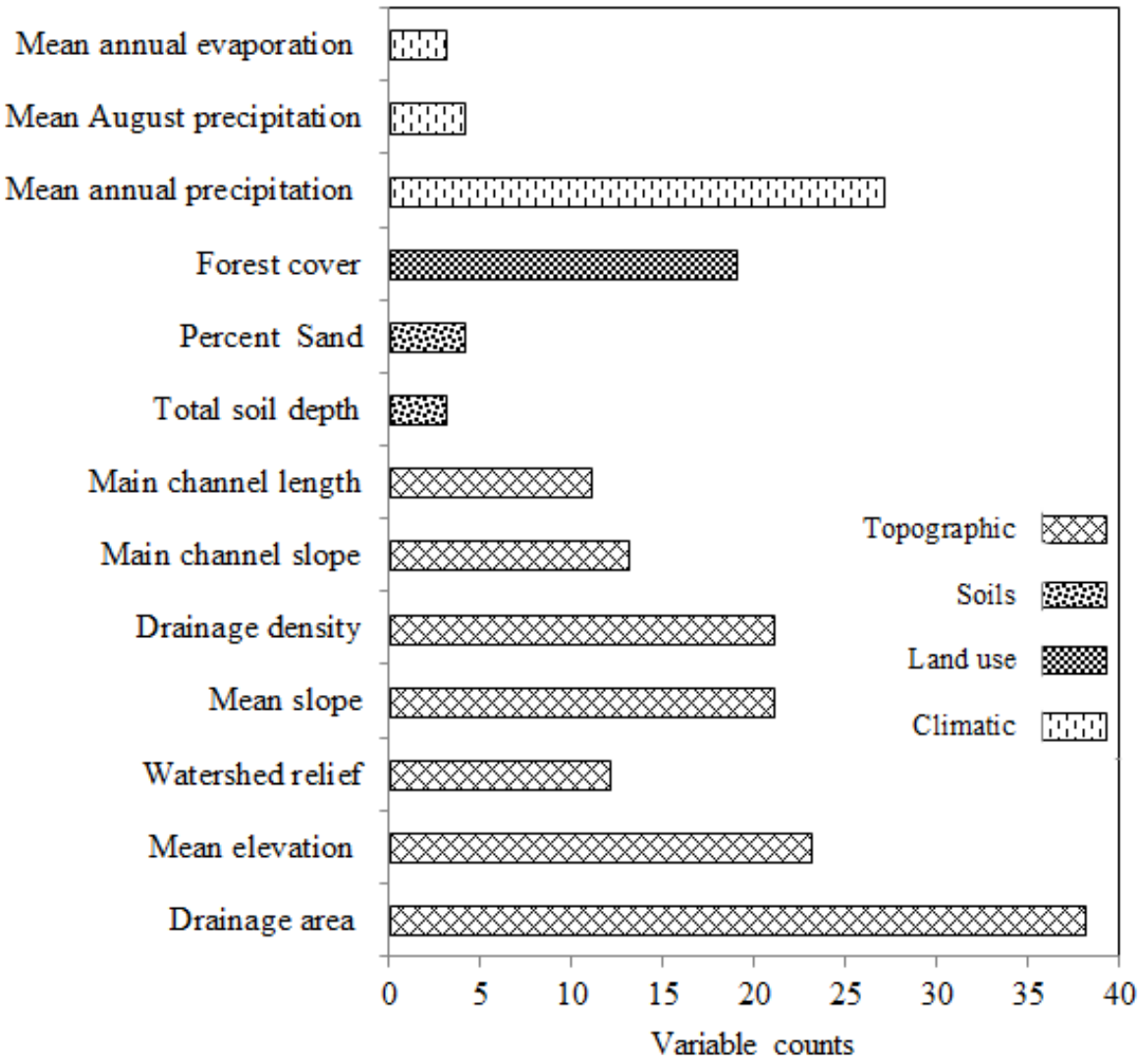

Figure 3.1: Frequency counts of some of the variables found in literature review of 42 studies (1989 - 2009)

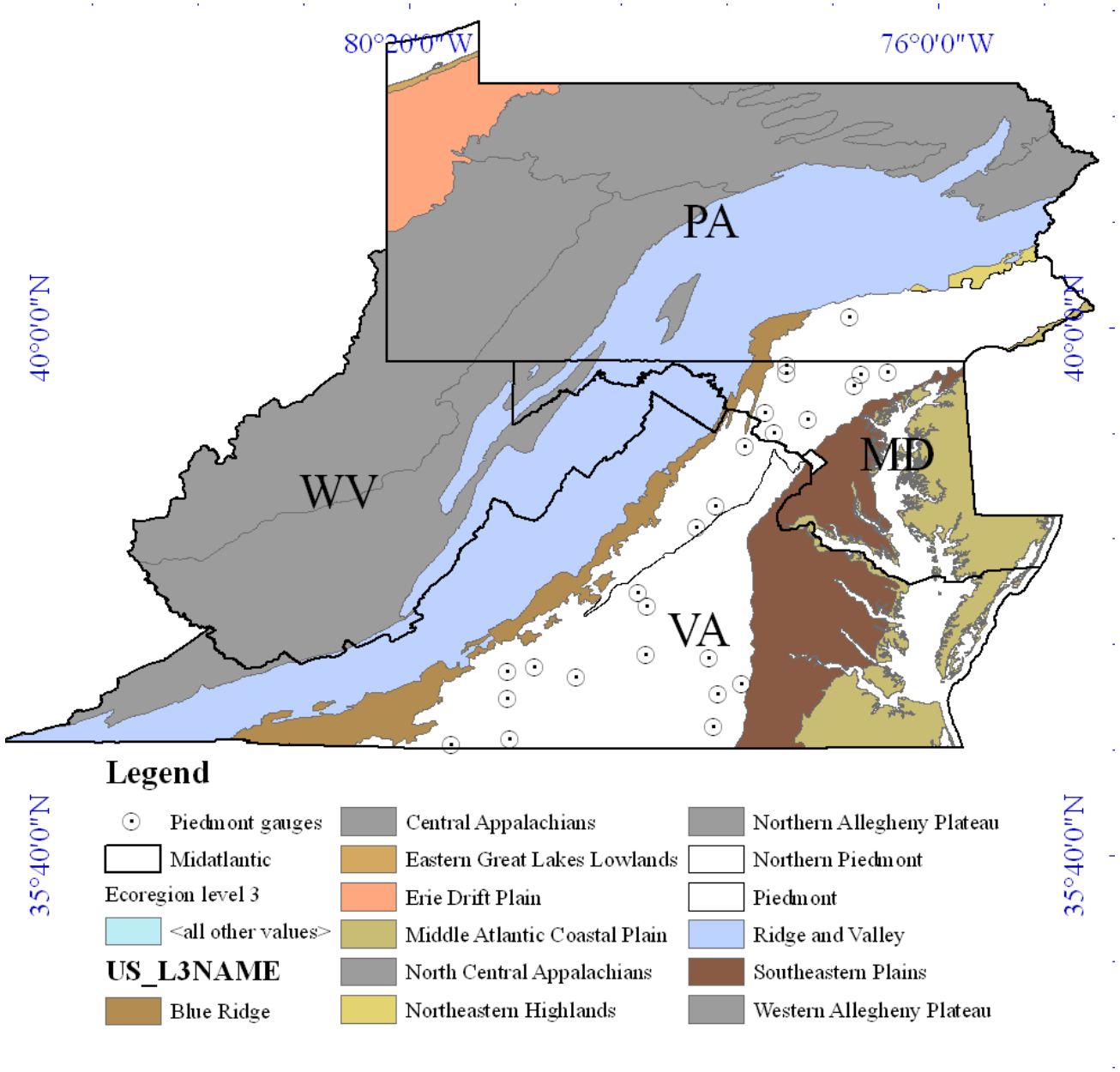

Figure 3.2: Location of watershed gauges in the Mid–Atlantic Piedmont, USA

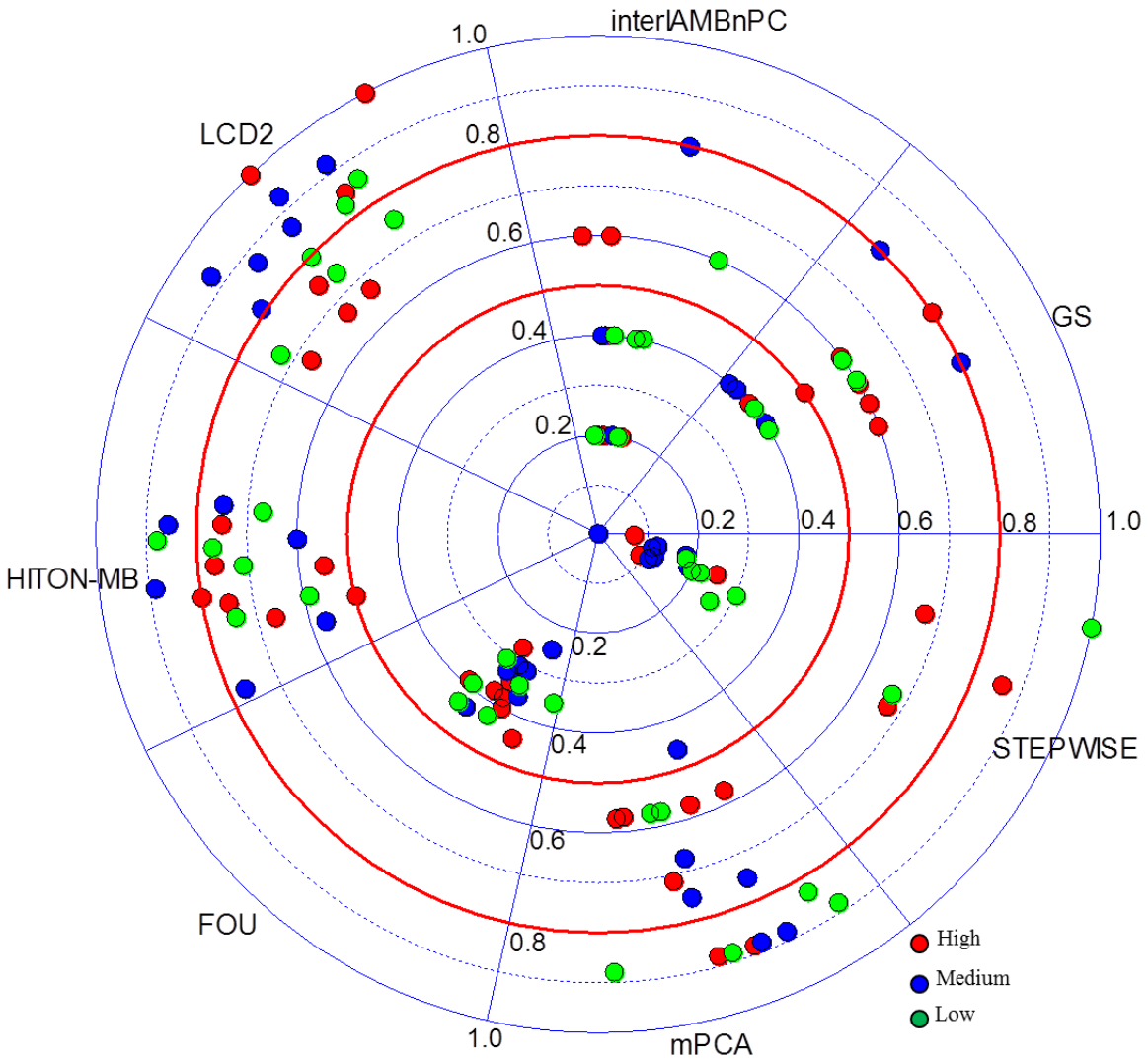

Figure 3.3: Bull's eye plot of method reliability indices (RI). Each data point is an average of 25 values for a single flow percentile. There are 19 points (flow percentiles) in each sector (method). Points closer to the center (RI=0) depict poor method reliability while points on the outermost circumference (RI=1) depict high method reliability. The red filled circles are for flows classified as high (Q0.01 to Q10), the blue filled circles for flows classified as medium (Q20 to Q70), and the green filled circles for flows classified as low (Q80 to Q99.9).

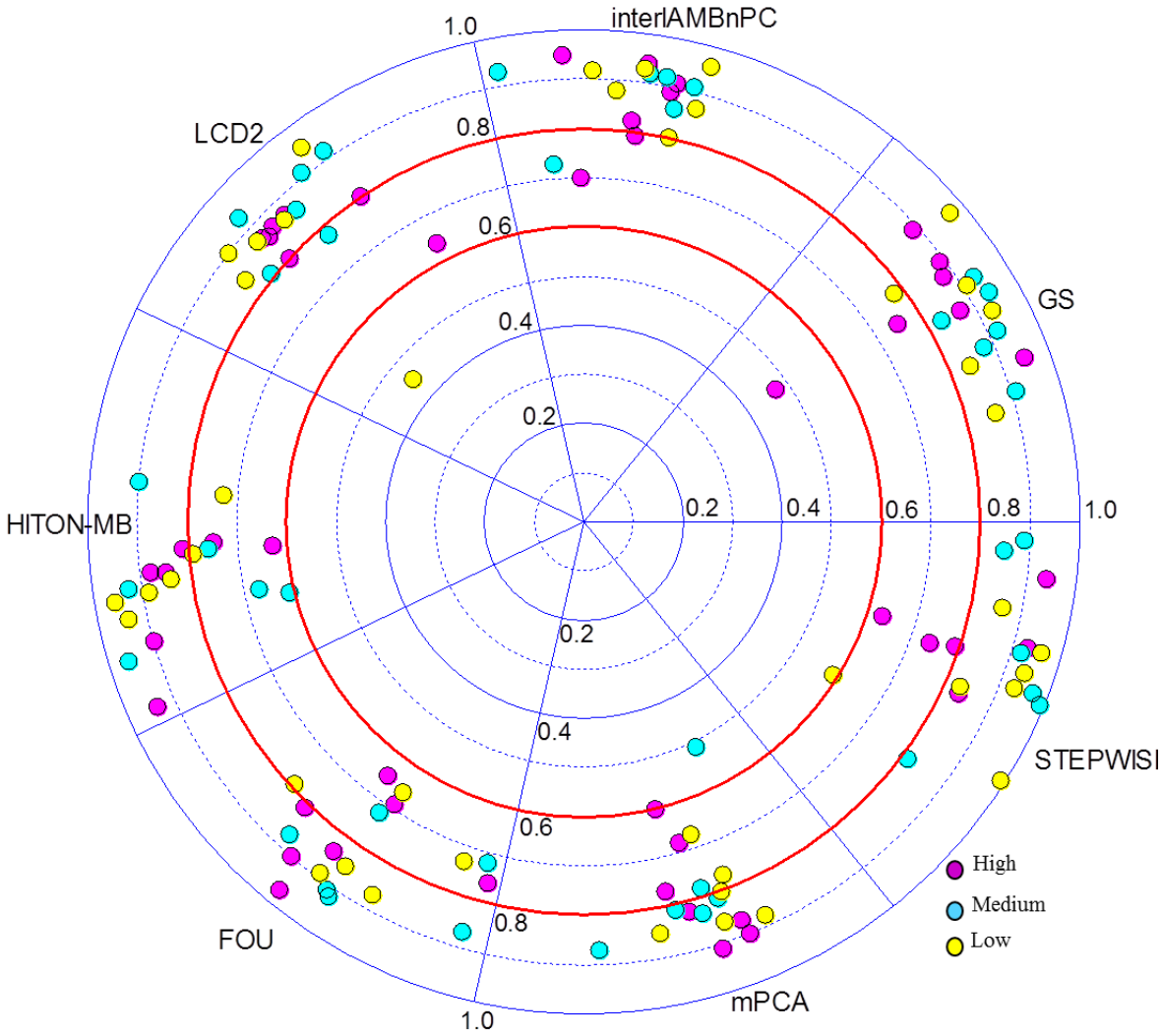

Figure 3.4: Bulls eye plot of prediction performance (R^2) of variables selected by each method. The prediction performance was based on feed forward back propagation neural network. Each data point is an average of 26 R^2 values for a single flow percentile. There are 19 points (flow percentiles) in each sector (method). Points closer to the center ($R^2 = 0$) depict poor method prediction performance while points on the outermost circumference ($R^2 = 1$) depict high method prediction performance

CHAPTER 4

ADVANCES IN VARIABLE SELECTION METHODS II: CLASSIFICATION OF HYDROLOGICALLY SIMILAR WATERSHEDS

[1]Ssegane, H., Tollner E. W., Mohamoud Y. M., Rasmussen T. C., and Dowd J. F. Submitted to *Journal of Hydrology*, 06/24/2011.

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).

4.1 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 (Acreman and Sinclair, 1986; Kokkonen et al., 2003), geographic proximity may infer similarity in climatic conditions and watershed form. 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 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 met-

rics 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) (Alcázar and Palau, 2010; Ma et al., 2010; Salas et al., 2010) and stepwise regression analysis (SRA) (Barnett et al., 2010; Gong et al., 2010; Peña-Arancibia 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 (Cartwright, 1979; Hitchcock, 2010; Suppes, 1970). 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 (Tsamardinos 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 form 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.

4.2 Methods

Study area and data

Data used in this study covers three Mid-Atlantic physiographic regions (ecoregions), USA (Figure 4.1); 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 to 2007 epoch across all watersheds. Figure 4.2 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 4.2.

[Figure 4.1 about here]

[Figure 4.2 about here]

Streamflow metrics

The common measures of watershed homogeneity or hydrological similarity analysis involve use of streamflow statistics (Castellarin et al., 2008; Kahya et al., 2008; Patil and Stieglitz, 2010; Srinivas et al., 2008). 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 subsequent subsections); 2) use of fewer variables minimizes challenges of

using high dimension data for unsupervised learning such as clustering (Ding et al., 2002; Fern and Brodley, 2003; M"uller 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.

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 (equation 4.1) for the 1966 to 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.

$$p_i \left(Q \geq q_i\right) = \frac{i}{N+1} \tag{4.1}$$

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.

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 equation 4.2. A recession constant of 0.98 ($\alpha = 0.98$) and a maximum baseflow index of 0.8 ($BFI_{max} = 0.8$) for humid

areas such as the Mid-Atlantic ecoregions were used.

$$b_t = \frac{(1 - BFI_{max})\,\alpha + b_{t-1}\,(1 - \alpha)\,BFI_{max} Q_t}{1 - \alpha BFI_{max}} \tag{4.2}$$

where b_t is baseflow at time step t (daily); BFI_{max} is maximum value of the baseflow index; α 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.

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 (equation 4.3) developed by Sankarasubramanian et al. (2001).

$$SFE = median\left[\frac{Q_t - \bar{Q}}{P_t - \bar{P}} \cdot \frac{\bar{P}}{\bar{Q}}\right] \tag{4.3}$$

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.

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.

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 k 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 (equation 4.4).

$$S_k^{'} = \frac{S_k - min\left\{S\right\}}{max\left\{S\right\} - min\left\{S\right\}} \tag{4.4}$$

where $S_k^{'}$ is transformed k^{th} term of variable S; and S_k is k^{th} term of variable S.

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 ($H1$, $H2$, and $H3$) 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 $1 \leq H \leq 2$, 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 (equation 4.5) 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.

$$D_i = \frac{1}{3} N (u_i - \bar{u})^T S^{-1} (u_i - \bar{u}) \tag{4.5}$$

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 4.1) 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.

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 (Li et al., 2010; Ouarda and Shu, 2009; Viola et al., 2011). 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 misclassification errors due to differences in data dimensions. The effect of data dimension on clustering results is documented in literature (M"uller et al., 2009). For each variable group, variable transformation (equation 4.4) 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 subsection*s.

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.

Watershed hypsometry

A hypsometric curve is a graphical representation of the distribution of area with elevation or the relative proportion of the watershed area that lies at or above a given height relative to total 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 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 equation 4.6 (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.

$$HI = \int_0^1 f(x)dx \tag{4.6a}$$

$$\mu_{01} = \frac{1}{I}\left[\int_0^1 xf(x)dx\right] \tag{4.6b}$$

$$\mu_2 = \frac{1}{I}\left[\int_0^1 (x-\mu_{01})^2 f(x)dx\right] \tag{4.6c}$$

$$\sigma = \sqrt{\mu_2} \tag{4.6d}$$

$$\mu_3 = \frac{1}{I}\left[\int_0^1 (x-\mu_{01})^3 f(x)dx\right] \tag{4.6e}$$

$$\mu_4 = \frac{1}{I}\left[\int_0^1 (x-\mu_{01})^4 f(x)dx\right] \tag{4.6f}$$

$$skew = \frac{\mu_3}{\sigma^3} \tag{4.6g}$$

$$kurtos = \frac{\mu_4}{\sigma^4} \tag{4.6h}$$

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 equation 4.7) 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).

$$f(x) = a_1 + a_2 x + a_3 x^2 + a_4 x^3 + \frac{(1-x)^{a_5}}{(1-x)^{a_5} + a_6 x^{a_7}} \tag{4.7}$$

Variables selected by causal algorithms

For each ecoregion, four causal variable selection algorithms: GS, interIAMBnPC, 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 to 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.

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 (Alcázar and Palau, 2010; Gao et al., 2009; Ma et al., 2010; Salas et al., 2010). 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 99 % 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 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.

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.

$$HD = 1 - \frac{|A \setminus B| + |B \setminus A|}{n} \tag{4.8}$$

$$S_s = 1 - \frac{|A| + |B| - 2\,|A \cap B|}{|A| + |B| - |A \cap B|} = \frac{|A \cap B|}{|A \cup B|} \tag{4.9}$$

$$CI = \frac{n\,|A \cap B| - k^2}{kn - k^2} \tag{4.10}$$

Where $|A \setminus B|$ is cardinality of set difference of A from B; $|B \setminus A|$ is cardinality of set difference of B from A; $|A|$ is cardinality of set A; $|B|$ is cardinality of set B; $|A \cap B|$ is cardinality of set intersection of A and B; and $|A \cup B|$ is cardinality of set union of A and B; 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 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.

$$SI = \frac{1}{2}\left(1 - \frac{|A \setminus B| + |B \setminus A| - 2\,|A \cap B|}{|A| + |B|}\right) \tag{4.11}$$

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.

$$RandIndex = \frac{a+d}{a+b+c+d} = \frac{a+d}{N} \tag{4.12}$$

$$AdjustedRandIndex = \frac{N(a+d) - [(a+b)(a+c)+(b+d)(c+d)]}{N^2 - [(a+b)(a+c)+(b+d)(c+d)]} \tag{4.13}$$

$$JaccardIndex = \frac{a}{a+b+c} \tag{4.14}$$

$$FowlkesMallows = \frac{a}{sqrt(a+b)(a+c)} \tag{4.15}$$

Where a is number of object pairs placed in the same cluster by two 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 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 A from B 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 Equations 4.8 to 4.15, the similarity indices are computed as below.

Figure 4.3 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

Hamming distance = 0.80	Rand index = 0.80
Kalousis similarity = 0.80	Adjusted Rand index = 0.00
Consistency index = 1.00	Jaccard index = 0.80
Similarity index (SI) = 0.8889	**Fowlkes-Mallows =0.8944**

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_s) and the Jaccard index gave similar results, and the hamming 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 hamming distance with values ranging from zero to one (better interpretation), subsequent assays in this study are based on SI.

[Figure 4.3 about here]

4.3 Results and discussions

Watershed classification by streamflow metrics

Clustering Results of streamflow indices

Figure 4.4 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.

Reference homogeneous watersheds

Table 4.1 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 Figure 4.1 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 (H1, H2, and H3 values in Table 4.1).

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.

For the Ridge and Valley ecoregion, the reference watersheds are relatively smaller(DA=644.9 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 km^2 and SS=0.74 %). The non–selected watersheds are concentrated in the Northern part of the ecoregion (44.4 % or 8 of 18 in Pennsylvania state) and are located near a neighboring ecoregion. For example, Cheat (USGS 03069500) and Youghiogheny (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.

Hypsometry and flow duration curves of reference watersheds

Figure 4.5 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. Figure 4.6 shows hypsometry and flow duration curves of three representative watersheds.

[Figure 4.5 about here]

[Figure 4.6 about here]

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 Figure 4.5, the flow that equaled or exceeded 50 % ($Q50$) of the record time (1966 to 2007) should be highest for the Appalachian Plateau and lowest for the Ridge and Valley because $HI_{Appa} > HI_{Pied} > HI_{RnV}$. This is demonstrated by the corresponding $Q50$ in Figure 4.5 in addition to hypsometric integrals and $Q50$ of representative watersheds in Figure 4.6. The use of $Q50$ for comparison is because the main drivers of flood and drought streamflow conditions are at their minimum at $Q50$ and thus $Q50$ 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; $Q \leq Q10$) and the Appalachian Plateau (relatively flat FDC in flood region) in Figure 4.5 and Figure 4.6.

Watershed classification by selected variables

Selected variable classes

Table 4.2 defines the watershed variables while Table 4.3 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 4.3). The selected variables for classification of watersheds in the 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.

[Table 4.2 about here]

[Table 4.3 about here]

Classification performance of selected variables

The results of Table 4.4 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 4.3). Accordingly, selected variables of geographical proximity performed better in the Appalachian (Table 4.4) because geographical proximity may infer the same climatic conditions. The hypsometric variables performed better in the Piedmont (Table 4.4) where topography was considered relevant by the causal algorithms (interIAMBnPC, LCD2, and HITON in Table 4.3).

[Table 4.4 about here]

The performance of the variable selection methods was based on two criteria. The first criterion was the similarity index (refer to section 4.2): 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 4.4, 4.5, and 4.6). Ideally, none of the off-diagonal indices should be equal or greater than the main diagonal similarity indices.

[Table 4.5 about here]

[Table 4.6 about here]

Based on results from Table 4.4 and Table 4.5, 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 4.4 row 9, column 3) than data from Ridge and Valley (Table 4.4 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 4.5). Table 4.6 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 4.1).

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 4.2, 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.

4.4 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 4.5; 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.

BIBLIOGRAPHY

Acreman, M., Sinclair, C., 1986. Classification of drainage basins according to their physical characteristics; an application for flood frequency analysis in Scotland. Journal of Hydrology 84 (3-4), 365–380. 83, 91

Alcázar, J., Palau, A., 2010. Establishing environmental flow regimes in a Mediterranean watershed based on a regional classification. Journal of Hydrology 388 (1-2), 41–51. 84, 93

Aliferis, C., Tsamardinos, I., Statnikov, A., 2003. HITON: A Novel Markov Blanket Algorithm for Optimal Variable Selection. In: AMIA Annual Symposium Proceedings. Vol. 2003. American Medical Informatics Association, p. 21. 84

Barnett, F., Gray, S., Tootle, G., et al., 2010. Upper Green River Basin (United States) Streamflow Reconstructions. Journal of Hydrologic Engineering 15, 567. 84

Bastola, S., Ishidaira, H., Takeuchi, K., 2008. Regionalisation of hydrological model parameters under parameter uncertainty: A case study involving TOPMODEL and basins across the globe. Journal of Hydrology 357 (3-4), 188–206. 83

Cartwright, N., 1979. Causal laws and effective strategies. Nous 13 (4), 419–437. 84

Castellarin, A., Burn, D., Brath, A., 2008. Homogeneity testing: How homogeneous do heterogeneous cross-correlated regions seem? Journal of Hydrology 360 (1-4), 67–76. 86

Chiang, S. M., Tsay, T. K., Nix, S. J., 2002. Hydrologic regionalization of watersheds. II: Applications. Journal of Water Resources Planning and Management–ASCE 128 (1), 12–20. 83

Cooper, G. F., 1997. A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. Data Mining and Knowledge Discovery 1, 203–224. 84

Di Prinzio, M., Castellarin, A., Toth, E., 2011. Data-driven catchment classification: application to the PUB problem. Hydrology and Earth System Sciences Discussions 8, 391–427. 83

Ding, C., He, X., Zha, H., Simon, H., 2002. Adaptive dimension reduction for clustering high dimensional data. In: Data Mining, 2002. ICDM 2002. Proceedings. 2002 IEEE International Conference on. IEEE, pp. 147–154. 87

Downton, M., Brennan, T., 1980. Comparing classifications: an evaluation of several coefficients of partition agreement. Class. Soc. Bull 4 (4), 53–54. 96

Dunne, K., Cunningham, P., Azuaje, F., 2002. Solutions to instability problems with sequential wrapper-based approaches to feature selection. Journal of Machine Learning Research. 94

Eckhardt, K., 2005. How to construct recursive digital filters for baseflow separation. Hydrological Processes 19 (2), 507–515. 87

Fern, X., Brodley, C., 2003. Random projection for high dimensional data clustering: A cluster ensemble approach. In: MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-. Vol. 20. p. 186. 87

Fowlkes, E., Mallows, C., 1983. A method for comparing two hierarchical clusterings. Journal of the American Statistical Association, 553–569. 96

Gao, Y., Vogel, R., Kroll, C., Poff, N., Olden, J., 2009. Development of representative indicators of hydrologic alteration. Journal of Hydrology 374 (1-2), 136–147. 93

Gong, G., Wang, L., Condon, L., Shearman, A., Lall, U., 2010. A Simple Framework for Incorporating Seasonal Streamflow Forecasts into Existing Water Resource Management Practices1. JAWRA Journal of the American Water Resources Association 46 (3), 574–585. 84

Hall, M., Minns, A., 1999. The classification of hydrologically homogeneous regions/Classification en régions hydrologiques homogènes. Hydrological Sciences Journal 44 (5), 693–704. 83

Harlin, J., 1978. Statistical moments of the hypsometric curve and its density function. Mathematical Geology 10 (1), 59–72. 92

Hitchcock, C., 2010. Probabilistic causation. Stanford Encyclopedia of Philosophy. 84

Hosking, J., Wallis, J., 1997. Regional frequency analysis: An approach based on L-moments. Cambridge Univ Pr. 88, 89, 98

Hubert, L., Arabie, P., 1985. Comparing partitions. Journal of classification 2 (1), 193–218. 95

Jensen, F., Nielsen, T., 2007. Bayesian networks and decision graphs. Springer Verlag. 84

Kahya, E., Demirel, M., Bég, O., 2008. Hydrologic homogeneous regions using monthly Streamflow in Turkey. Earth Sciences Research Journal 12 (2), 181–193. 83, 86

Kalousis, A., Prados, J., Hilario, M., 2007. Stability of feature selection algorithms: a study on high-dimensional spaces. Knowledge and Information Systems 12 (1), 95–116. 94

Karimi, K., Hamilton, H., 2009. Finding temporal relations: Causal bayesian networks vs. C4. 5. Foundations of Intelligent Systems, 266–273. 84

Kokkonen, T., Jakeman, A., Young, P., Koivusalo, H., 2003. Predicting daily flows in ungauged catchments: model regionalization from catchment descriptors at the Coweeta Hydrologic Laboratory, North Carolina. Hydrological Processes 17 (11), 2219–2238. 83

Kuncheva, L., 2007. A stability index for feature selection. In: Proceedings of the 25th conference on Proceedings of the 25th IASTED International Multi–Conference: artificial intelligence and applications. ACTA Press, pp. 390–395. 94

Li, M., Shao, Q., Zhang, L., Chiew, F., 2010. A new regionalization approach and its application to predict flow duration curve in ungauged basins. Journal of Hydrology 389, 137–145. 91

Lu, Y., Cohen, I., Zhou, X., Tian, Q., 2007. Feature selection using principal feature analysis. In: Proceedings of the 15th international conference on Multimedia. ACM, pp. 301–304. 93

Luo, W., 2000. Quantifying groundwater-sapping landforms with a hypsometric technique. Journal of Geophysical Research 105, 1685–1694. 92

Ma, H., Liu, L., Chen, T., 2010. Water security assessment in Haihe River Basin using principal component analysis based on Kendall τ. Environmental monitoring and assessment 163 (1), 539–544. 84, 93

Margaritis, D., Thrun, S., 1999. Bayesian network induction via local networks. [Research paper] / Carnegie Mellon University. School of Computer Science,. School of Computer Science Carnegie Mellon University, Pittsburgh, Pa. 84

Meganck, S., Leray, P., Manderick, B., 2006. Learning causal bayesian networks from observations and experiments: A decision theoretic approach. Modeling Decisions for Artificial Intelligence, 58–69. 84

Merz, R., Blöschl, G., 2004. Regionalisation of catchment model parameters. Journal of Hydrology 287 (1-4), 95–123. 83

M"uller, E., G"unnemann, S., Assent, I., Seidl, T., 2009. Evaluating clustering in subspace projections of high dimensional data. Proceedings of the VLDB Endowment 2 (1), 1270–1281. 87, 91

Olaya, V., Conrad, O., 2009. Geomorphometry in SAGA. Developments in Soil Science 33, 293–308. 93

Omernik, J., Bailey, R., 1997. Distinguishing between watersheds and ecoregions. Water resources bulletin 33 (5), 935–949. 83

Ouarda, T., Shu, C., 2009. Regional low-flow frequency analysis using single and ensemble artificial neural networks. Water Resources Research 45 (11), W11428. 91

Patil, S., Stieglitz, M., 2010. Hydrologic similarity among catchments under variable flow conditions. Hydrology and Earth System Sciences Discussions 7, 8607–8630. 86

Peña-Arancibia, J., van Dijk, A., Mulligan, M., Bruijnzeel, L., Gebrehiwot, S., Ilstedt, U., Gärdenas, A., Bishop, K., Brocca, L., Melone, F., et al., 2010. The role of climatic and terrain attributes in estimating baseflow recession in tropical catchments. Hydrology and Earth System Sciences Discussions 7, 4059–4087. 84

Pérez-Peŋa, J., Azaŋkŋ, J., Azor, A., 2009. CalHypso: An ArcGIS extension to calculate hypsometric curves and their statistical moments. Applications to drainage basin analysis in SE Spain. Computers & Geosciences 35 (6), 1214–1223. 92

Rand, W., 1971. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical association, 846–850. 95

Salas, J., Fu, C., Rajagopalan, B., 2010. Long Range Forecasting of Colorado Streamflows Based on Hydrologic, Atmospheric, and Oceanic Data. Journal of Hydrologic Engineering 1, 210. 84, 93

Sankarasubramanian, A., Vogel, R., Limbrunner, J., 2001. Climate elasticity of streamflow in the United States. Water Resources Research 37 (6), 1771–1781. 88

Sawicz, K., Wagener, T., Sivapalan, M., Troch, P., Carrillo, G., 2011. Catchment classification: empirical analysis of hydrologic similarity based on catchment function in the eastern USA. Hydrology and Earth System Sciences Discussions 8, 4495–4534. 83, 86

Searcy, J., 1959. Flow-duration curves: US Geol. Survey Water-Supply Paper 1542. 100

Slack, J., Landwehr, J., 1992. Hydro-climatic data network (HCDN): A US Geological Survey

streamflow data set for the United States for the study of climate variations, 1874-1988. US Geological Survey. 86

Srinivas, V., Tripathi, S., Rao, A., Govindaraju, R., 2008. Regional flood frequency analysis by combining self-organizing feature map and fuzzy clustering. Journal of Hydrology 348 (1-2), 148–166. 86

Steinley, D., 2004. Properties of the Hubert-Arable Adjusted Rand Index. Psychological Methods 9 (3), 386. 96, 97

Stieglitz, M., Rind, D., Famiglietti, J., Rosenzweig, C., 1997. An Efficient Approach to Modeling the Topographic Control of Surface Hydrology for Regional and Global Climate Modeling. Journal of Climate 10, 118–137. 85

Strahler, A., 1952. Hypsometric (area-altitude) analysis of erosional topography. Geological Society of America Bulletin 63 (11), 1117. 92

Suppes, P., 1970. A probabilistic theory of causality. North-Holland. 84

Tsamardinos, I., Aliferis, C., Statnikov, A., 2003. Algorithms for large scale markov blanket discovery. In: The 16th International FLAIRS Conference. Vol. 103. 84

Viglione, A., Viglione, M., 2010. Package nsRFA. 90

Viola, F., Noto, L., Cannarozzo, M., La Loggia, G., 2011. Regional flow duration curves for ungauged sites in Sicily. Hydrol. Earth Syst. Sci 15, 323–331. 91

Vivoni, E., Di Benedetto, F., Grimaldi, S., Eltahir, E., 2008. Hypsometric control on surface and subsurface runoff. Water Resources Research 44 (12), W12502. 85, 100

Vogel, R., 1994. Flow-duration curves. I: New interpretation and confidence intervals. Management 120 (4). 87

Vogel, R., Fennessey, N., 1995. Flow Duration Curves II: A Review of Applications In Water Resources Planning. JAWRA Journal of the American Water Resources Association 31 (6), 1029–1039. 87

Wagener, T., Sivapalan, M., Troch, P., Woods, R., 2007. Catchment classification and hydrologic similarity. Geography Compass 1 (4), 901–931. 83

Wolock, D., Winter, T., McMahon, G., 2004. Delineation and evaluation of hydrologic-landscape regions in the United States using geographic information system tools and multivariate statistical analyses. Environmental Management 34, 71–88. 83

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G., Ng, A., Liu, B., Yu, P., et al., 2008. Top 10 algorithms in data mining. Knowledge and Information Systems 14 (1), 1–37. 89

Zheng, H., Zhang, L., Zhu, R., Liu, C., Sato, Y., Fukushima, Y., 2009. Responses of streamflow to climate and land surface change in the headwaters of the Yellow River Basin. Water Resources Research 45 (7). 88

Table 4.1: Hydrologically similar watersheds (Reference watersheds) for each ecoregion

ID	Gauge Name	USGS No.	DA^a	P^b	FDC^c	BFI^d	SFE^e
	Appalachian plateaus (H1=1.06, H2=-0.37, H3=0.07)^f						
1	Tioga River, PA	01518000	730.4	897	-0.0191	0.574	1.227
2	Cowanesque River, PA	01520000	771.8	953	-0.0219	0.538	1.758
3	Towanda Creek, PA	01532000	556.8	1013	-0.0187	0.563	1.605
4	Tunkhannock Creek, PA	01534000	992.0	940	-0.0163	0.602	1.199
5	WB Susquehanna, PA	01541000	815.8	1113	-0.0146	0.599	1.208
6	Sinnemahoning Creek, PA	01543500	1774.1	1064	-0.0170	0.601	1.675
7	Pine Creek, PA	01548500	1564.4	1242	-0.0170	0.623	1.223
8	Blockhouse Creek, PA	01549500	97.6	993	-0.0170	0.598	1.234
9	Georges creek, MD	01599000	187.5	958	-0.0179	0.583	1.322
10	Buffalo Creek, PA	03049000	354.8	1006	-0.0172	0.579	0.836
11	Redstone Creek, PA	03074500	190.9	1069	-0.0126	0.641	2.250
12	Bluestone river, WV	03179000	1023.0	960	-0.0188	0.575	0.918
13	Big coal river, WV	03198500	1012.7	1176	-0.0184	0.556	1.162
	Piedmont (H1=-0.46, H2=0.16, H3=-0.33)						
1	Deer Creek at Rocks, MD	01580000	244.5	1135	-0.0088	0.737	1.943
2	Little Falls, MD	01582000	137.0	1064	-0.0090	0.755	2.052
3	Western Run, MD	01583500	154.9	1074	-0.0088	0.745	1.600
4	Patuxent River, MD	01591000	90.1	1080	-0.0116	0.693	2.205
5	Big pipe Creek, MD	01639500	264.2	1095	-0.0121	0.644	1.536
6	Seneca Creek, MD	01645000	261.6	996	-0.0103	0.658	1.795
7	Rapidan River, VA	01667500	1212.1	1072	-0.0129	0.659	1.390
8	Appomattox River, VA	02039500	782.2	1118	-0.0113	0.608	1.893
9	Nottoway River, VA	02044500	821.0	1189	-0.0131	0.604	1.736
10	Pigg River, VA	02058400	909.1	1146	-0.0082	0.669	1.462
11	Goose Creek, VA	02059500	486.9	914	-0.0102	0.654	2.192
12	Big otter River, VA	02061500	815.8	1113	-0.0113	0.670	2.294
13	Falling River, VA	02064000	427.3	1123	-0.0109	0.607	1.657
14	North Mayo River, VA	02070000	279.7	1278	-0.0072	0.691	1.332
15	Sandy River, VA	02074500	287.5	1151	-0.0077	0.649	1.904

Table 4.1: continued on next page

Table 4.1: continued

ID	Gauge Name	USGS No.	DA^a	P^b	FDC^c	BFI^d	SFE^e
	Ridge and Valley (H1=-0.54, H2=0.94, H3=0.27)						
1	Dunning Creek, PA	01560000	445.5	1003	-0.0189	0.582	1.058
2	Aughwick Creek, PA	01564500	530.9	993	-0.0208	0.574	1.482
3	Wills Creek, MD	01601500	639.7	1077	-0.0187	0.586	1.275
4	Patterson Creek, WV	01604500	572.4	899	-0.0223	0.584	2.399
5	Cacapon River, WV	01611500	1748.2	884	-0.0177	0.593	1.705
6	Back Creek, WV	01614000	608.6	970	-0.0195	0.505	1.334
7	N F Shenandoah, VA	01632000	543.9	899	-0.0227	0.517	1.795
8	Cedar Creek, VA	01634500	264.2	874	-0.0173	0.592	2.264
9	Dunlap Creek, VA	02013000	419.6	1034	-0.0185	0.567	1.404
10	Calfpasture River, VA	02020500	365.2	1085	-0.0208	0.533	1.683

[a] Drainage area (km^2)

[b] Annual precipitation (mm)

[c] Slope of the flow duration curve ($LS^{-1}km^{-2}$)

[d] Baseflow index (–)

[e] Streamflow elasticity (–)

[f] The regional Hosking and Wallis (1997) $H1$, $H2$, and $H3$ statistics

Table 4.2: Table of watershed descriptors

Variable	Units	Description
RC	–	circularity ratio (area to square of perimeter ratio)
RN	–	ruggedness number (drainage density×relief)
SLEN	m	slope length
CI	–	convergence index
SMAX	m/km	maximum slope
VDEP	m	average valley depth
CPLAN	–	plan curvature; rate of change of aspect along a contour
TWI	–	topographic wetness index
MRVBF	–	multi resolution index of valley bottom flatness
MRRTF	–	multi resolution index of ridge top flatness
LDP	km	longest drainage path (main channel length)
TSL	km	total stream length
MCS	m/km	main channel slope
AMEAN	deg	average aspect
MAP	mm	mean annual precipitation
MAET	mm	mean annual evapotranspiration
NAP	mm	net annual precipitation (MAP – MAET)
Agric	%	area under agriculture
Urban	%	developed areas of low, medium, and high intensity
AWC	cm/cm	available water content
rockDep	cm	depth to bedrock
hsgC	%	hydrological soil group C
Hydric	%	hydric soils
Silt	%	silt
KSAT	cm/hr	saturated hydraulic conductivity
Porosity	–	porosity
MSCL	cm	macroscopic capillary length
T	cm^2/hr	transmissivity
JUNP	mm	average June precipitation
AUGP	mm	average August precipitation
SEPP	mm	average September precipitation
SepET	mm	average September evapotranspiration
MMET	mm	mean monthly potential evapotranspiration
ADI	mm	annual dryness index ; $ADI = \frac{MAP}{MAET}$
RFx	mm	Rainfall equaled or exceeded x % of the record time

Table 4.3: Top three variables[a] selected by each method for watershed classification

	GS	**interIAMBnPC**	**PCA & Stepwise**
Appalachian	RF20, SEPP, RC	Hydric, MRVBF, NAP	SepET, RF20, JUNP
Piedmont	AWC , TWI, rockDep	VDEP, TWI, LDP	hsgC, CPLAN, MMET
Ridge & Valley	RF10, TWI, MSCL	AMEAN, Agric, Urban	KSAT, AUGP, Porosity
	LCD2	**HITON–MB**	**HITON & IAMB**
Appalachian	SEPP, ADI, MRVBF	Silt, MRVBF, MSCL	Hydric, MRVBF, NAP
Piedmont	MRRTF, MRVBF, CI	SLEN, MRRTF, T	SLEN, MRRTF, T
Ridge & Valley	ADI, RF10, RF20	SEPP, AMEAN, KSAT	SEPP, AMEAN, KSAT
	Stepwise	**PCA**	**HITON & GS**
Appalachian	NAP, SepET, RF20	MAP, JUNP, RN	RF20, SEPP, RC
Piedmont	hsgC, CPLAN, Urban	VDEP, RN, MMET	SLEN, MRRTF, rockDep
Ridge & Valley	KSAT, AUGP, Porosity	RN, SMAX, TSL	SEPP, AMEAN, KSAT

[a] Refer to Table 4.2 for description of each variable

Table 4.4: Classification performance of variables for geographic proximity, watershed hypsometry, and variables selected by principal component analysis (PCA) and stepwise regression

Variable class	**Ecoregion data**[a]		
	Appalachian	**Piedmont**	**Ridge & Valley**
	Proximity & Hypsometry		
Geographic proximity	**0.64**	0.52	0.48
watershed hypsometry	0.40	**0.71**	0.29
	PCA		
Appalachian	**0.71**	0.64	**0.67**
Piedmont	0.33	**0.76**	0.54
Ridge & Valley	0.50	**0.71**	**0.57**
	Stepwise		
Appalachian	**0.72**	0.58	**0.64**
Piedmont	0.67	**0.87**	0.50
Ridge & Valley	0.40	0.55	**0.64**

[a] 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 selected variables are exactly the same as the reference watersheds (classified using streamflow indices).

Table 4.5: Classification performance of variables selected by causal variable selection algorithms of Grow-Shrink (GS), interleaved Incremental Association Markov Blanket with Parents and Children (interIAMBnPC), Local Causal Discovery (LCD), and the HITON Markov Blanket (HITON–MB).

Variable class	**Ecoregion data**		
	Appalachian	**Piedmont**	**Ridge & Valley**
	GS		
Appalachian	**0.89**	0.43	0.40
Piedmont	0.67	**0.86**	0.25
Ridge & Valley	0.38	**0.72**	**0.67**
	interIAMBnPC		
Appalachian	**0.81**	**0.64**	**0.62**
Piedmont	0.50	**0.81**	0.46
Ridge & Valley	0.59	**0.67**	**0.62**
	LCD2		
Appalachian	**0.77**	**0.90**	0.67
Piedmont	0.50	**0.93**	0.31
Ridge & Valley	0.52	0.42	**0.73**
	HITON-MB		
Appalachian	**0.71**	0.50	0.46
Piedmont	0.35	**0.90**	0.48
Ridge & Valley	0.67	0.58	**0.72**

Table 4.6: Classification performance of variables selected by combining selected variables from two variable selection methods.

Variable class	Ecoregion data		
	Appalachian	**Piedmont**	**Ridge & Valley**
	PCA & Stepwise		
Appalachian	**0.81**	0.54	0.26
Piedmont	0.67	**0.90**	0.50
Ridge & Valley	**0.64**	0.36	**0.64**
	HITON-MB & interIAMBnPC		
Appalachian	**0.81**	0.64	0.62
Piedmont	0.35	**0.90**	0.48
Ridge & Valley	0.67	0.58	**0.72**
	HITON-MB & GS		
Appalachian	**0.89**	0.43	0.40
Piedmont	0.38	**0.93**	0.48
Ridge & Valley	0.67	0.58	**0.72**

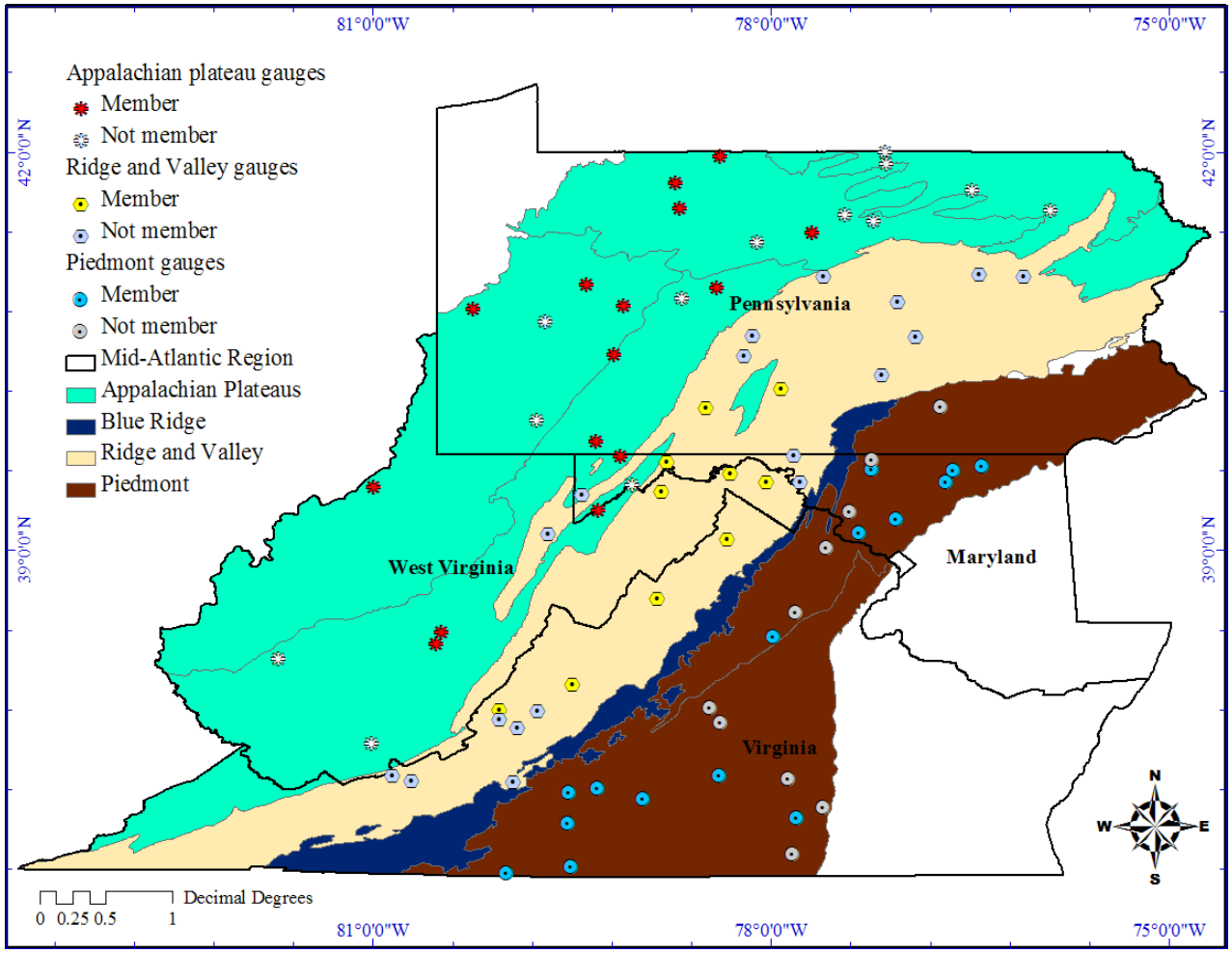

Figure 4.1: Location of watersheds in the three physiographic provinces of Appalachian Plateaus, Piedmont, and Ridge and Valley (Mid-Atlantic region, USA). For each ecoregion, homogeneous (reference) and non–homogeneous (non–reference) watersheds are differentiated

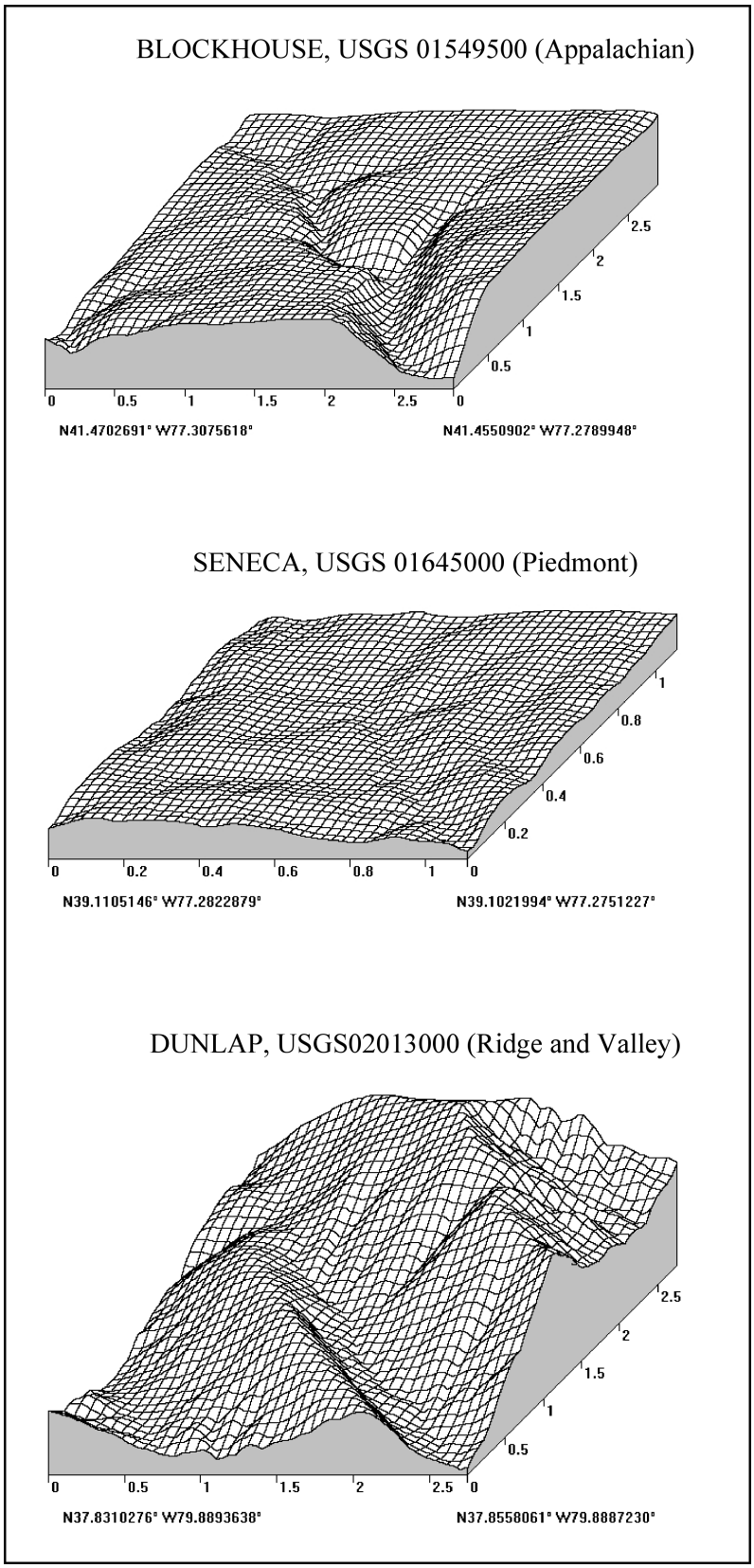

Figure 4.2: Fishnet plots of topography of headwaters of representative watersheds for each ecoregion.

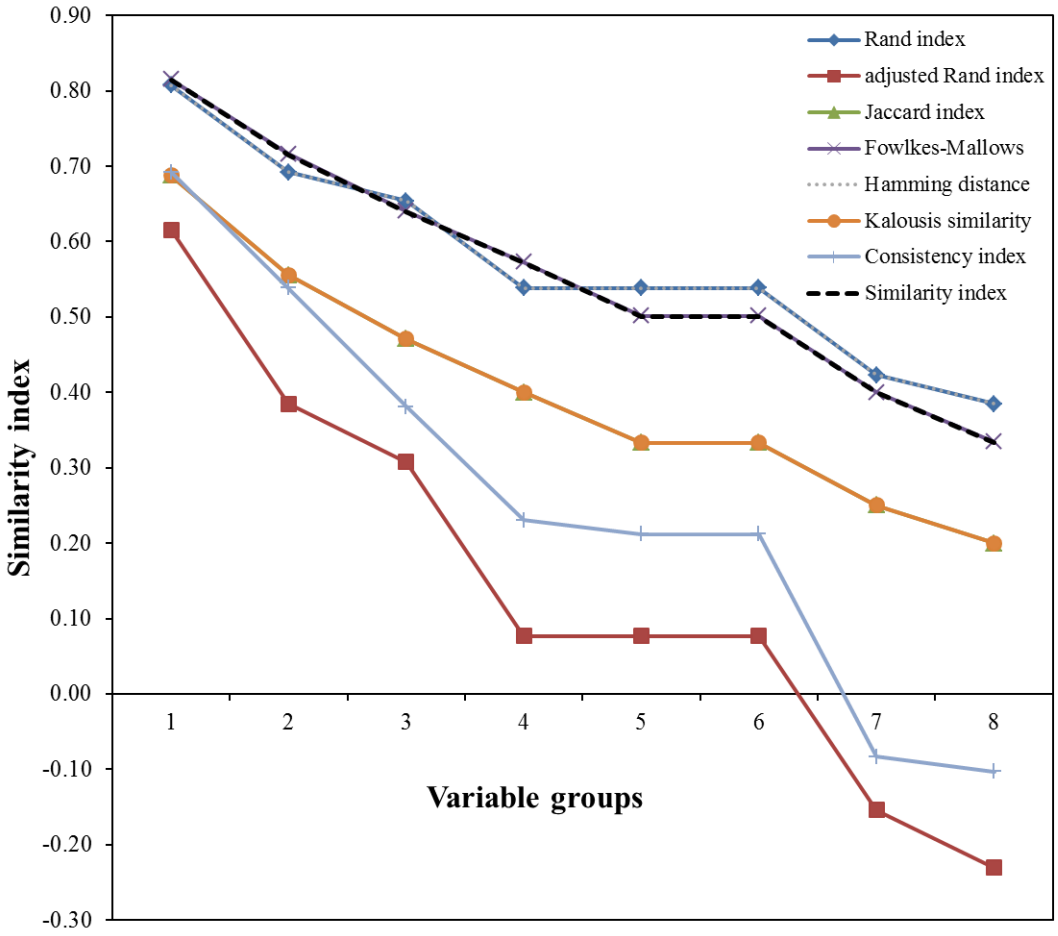

Figure 4.3: Comparison of similarity indices. The Rand Index, adjusted Rand Index, Jaccard Index, and the Fowlkes–Mallows Index are primarily used in analysis of cluster validity; while the Hamming Distance, Kalousis Similarity, and the Consistency Index are used in analysis of stability (robustness) 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.

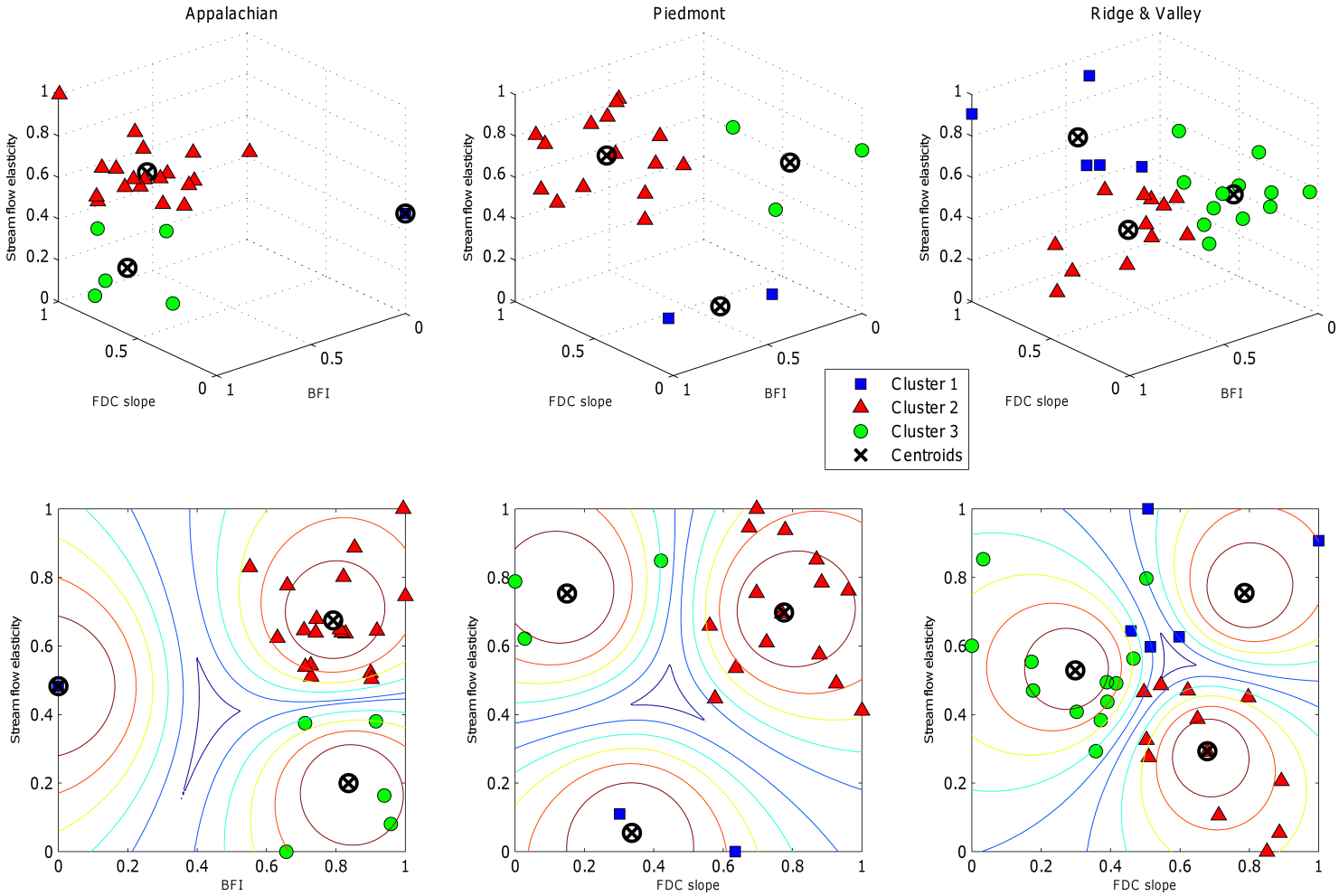

Figure 4.4: Clustering results for the Appalachian, Piedmont, and Ridge & Valley before homogeneity and discordancy tests. The three streamflow indices are standardized using minimum–maximum transformation (0 – 1 linear scale). The 2D plots depict the dominant clustering streamflow indices (where FDC=Flow duration curve and BFI=Baseflow index).

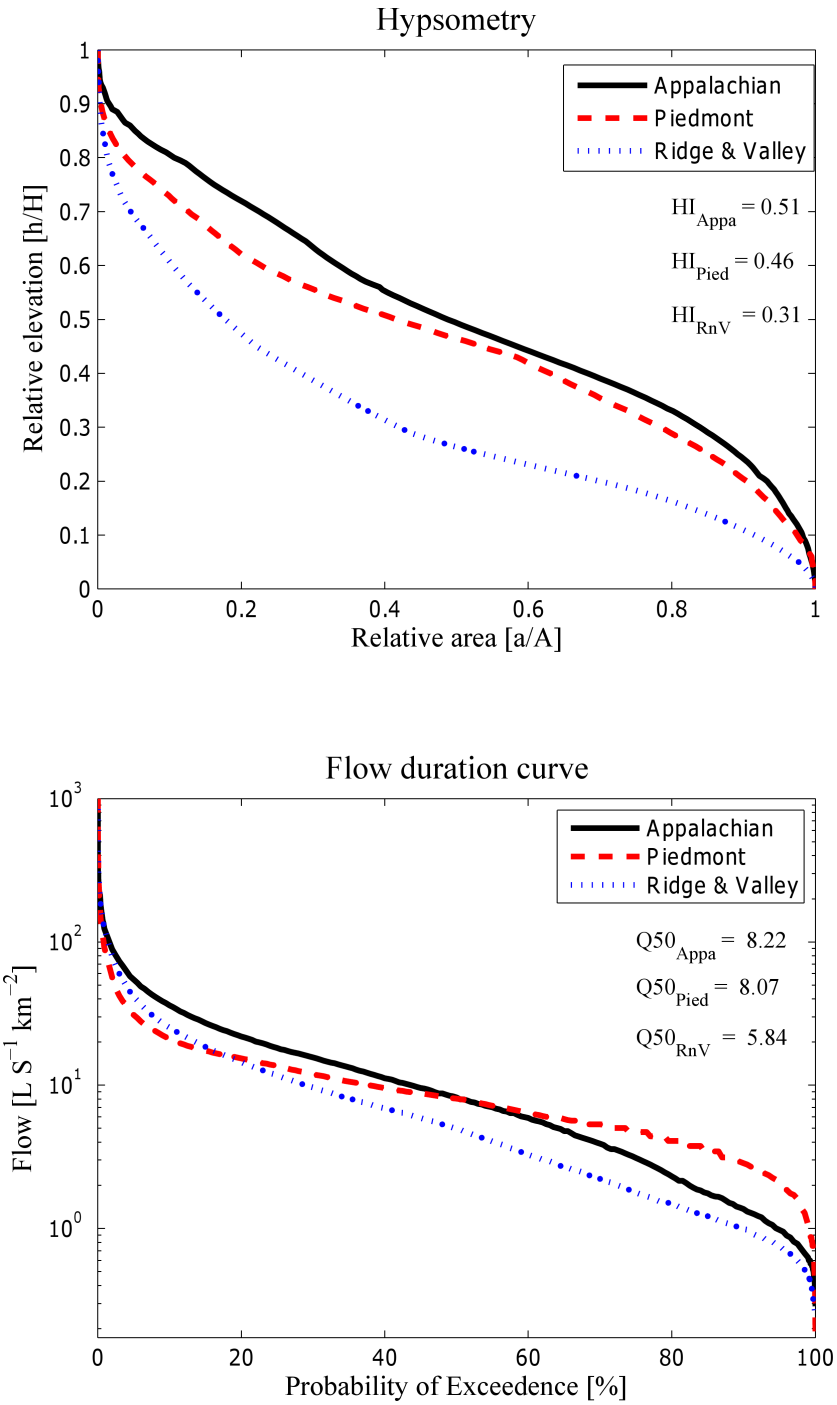

Figure 4.5: Characteristic hypsometry (top) and flow duration curve for each ecoregion. Characteristic curves were computed as medians of reference watersheds. Appalachian plateaus are dominated by convex shape ($HI_{Appa} > 0.5$) while Piedmont and Ridge and Valley are dominated by concave hypsometry (HI_{Pied} and $HI_{RnV} < 0.5$). Corresponding flow decreases as hypsometry changes from convex to concave ($Q50_{Appa} > Q50_{Pied} > Q50_{RnV}$).

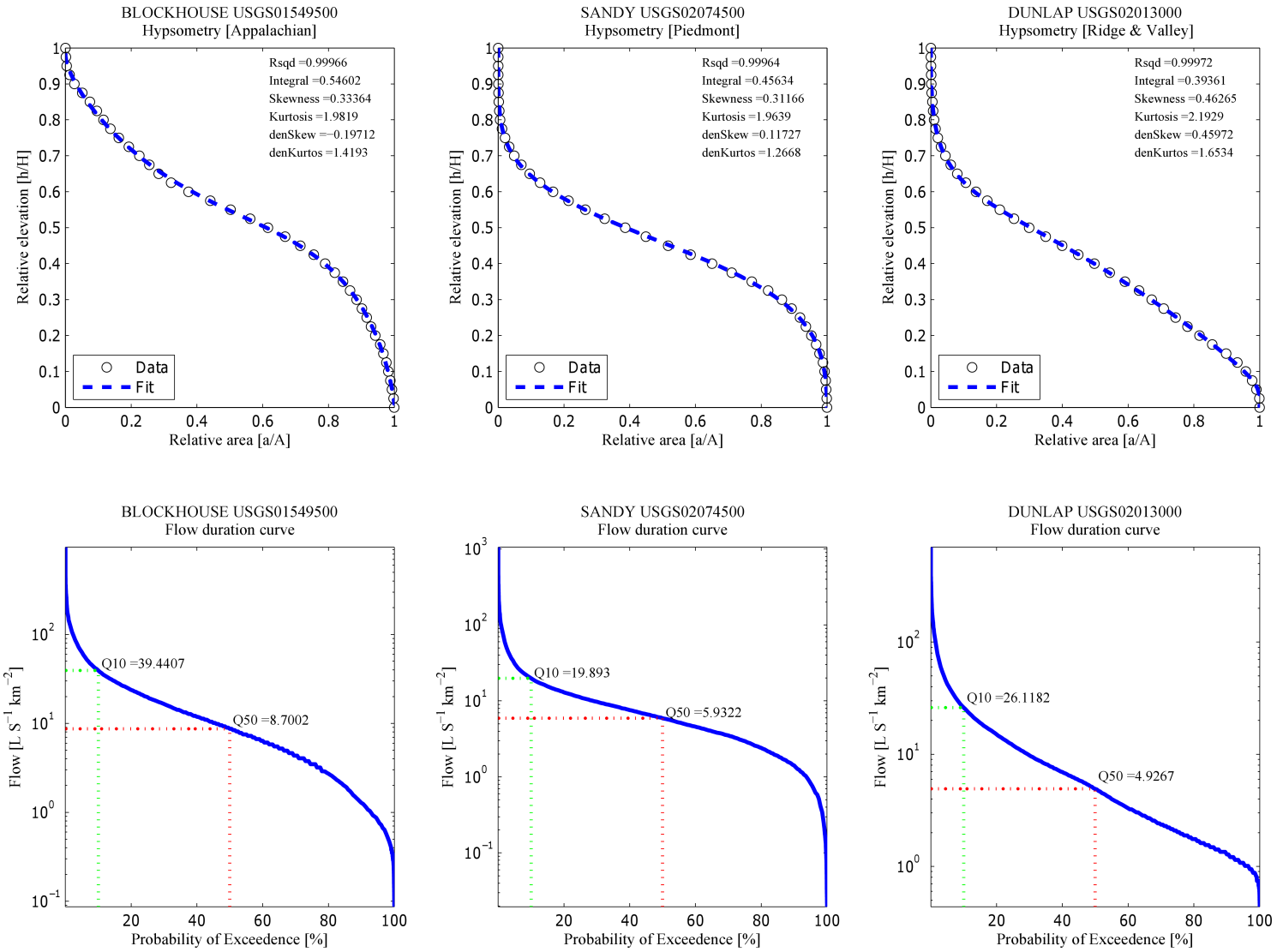

Figure 4.6: Hypsometry and flow duration curves of representative watersheds from each ecoregion. The closest watershed to the cluster centroid of reference watersheds was chosen as a representative watershed.

CHAPTER 5

DAILY STREAMFLOW PREDICTION FOR UNGAUGED WATERSHEDS BY INDEPENDENT ESTIMATION OF MAGNITUDE AND TEMPORAL SEQUENCE

[1]Ssegane, H., Tollner E. W., Mohamoud Y. M., Rasmussen T. C., and Dowd J. F. To be submitted to *Journal of Environmental Modelling and Software*

Abstract

The process of predicting daily streamflow using hydrologic models implicitly predicts streamflow magnitude and temporal sequence concurrently. However, if one conceptualizes streamflow as a composite of two separable components of magnitude and sequence, then each component can be predicted separately and then combined. This study independently predicted streamflow magnitude using regionalized flow duration curves and streamflow sequence for watersheds (basins) in three Mid–Atlantic regions of Appalachian Plateaus (28 basins; 98–1779 km^2), Piedmont (19 basins; 34.8–620 km^2), and Ridge and Valley (25 basins; 48–1857 km^2). The two components were combined using a shuffling technique (sorting) to generate predicted daily streamflow. The study also assessed the effect of relative drainage area of gauged (donor) and ungauged (target) watersheds, distance between gauged and ungauged watersheds, and use of ensemble techniques, on accuracy of predicted sequence. The results show that this approach better predicted daily streamflow (Nash–Sutcliffe Efficiency, $NSE = 0.88$ for Cheat River near Parsons, WV) than Hydrologic Simulation Program–Fortran (HSPF; $NSE = 0.33$), even after HSPF calibration using parameter estimation software (PEST; $NSE = 0.43$). The drainage area of the donor watershed had no effect on accuracy of predicted sequence while distance had. Ensemble methods of geometric mean, Monte–Carlo, bootstrap aggregation (bagging), and modified bagging (boosting) provided the greatest improvements in accuracy of predicted sequence. For Cheat River (USGS 03069500), boosting improved NSE of predicted daily streamflow from 0.76 to 0.88.

5.1 Introduction

Water resources planning and management requires long–term streamflow data to assess feasibility of establishing engineering structures such as dams and bridges, assessment of environmental and ecological integrity of rivers, allocation of water resources among competing uses, establishment of water quality standards, watershed management, and disaster preparedness in case of flood and drought conditions. Such data may be available at gauged sites; however many watersheds (drainage basins) in developed and developing countries lack long–term streamflow data and thus the need to predict streamflow in ungauged watersheds (Mazvimavi et al., 2005; Pavizham et al., 2009; Sivapalan, 2003). For example, according to Besaw et al. (2010), less than 10 % of rivers in the U.S are gauged by the U.S. Geological Survey (USGS) on a daily basis.

Two methods for predicting streamflow in ungauged watersheds include: 1) statistical regionalization (Engeland and Hisdal, 2009; Garcia-Martinó et al., 1996; Kokkonen et al., 2003; Laaha and Bloschl, 2006; Mohamoud, 2008; Zhu and Day, 2009), where multiple regression analysis is used to correlate hydrological responses of watersheds to physical and climatic attributes; and 2) use of regionalized hydrological model parameters (Bárdossy, 2007; Bastola et al., 2008; Gotzinger and Bárdossy, 2007; Hughes et al., 2010), where watershed characteristics of ungauged watersheds are related to optimized hydrologic model parameters at gauged watersheds. Both approaches use regionalization to infer hydrological similarity between watersheds in the same hydrological region. The regionalization process involves grouping watersheds with similar hydrological response such that relationships between flow regimes, climatic and watershed characteristics derived at gauged watersheds are used to predict flow at ungauged watersheds (Srinivas et al., 2008).

Regarding statistical regionalization, Garcia-Martinó et al. (1996) used 53 watershed characteristics and multiple regression analysis to derive equations for low–flow indices for 19 watersheds (0.19–38.50 km^2) of humid Montane regions of Puerto Rico (U.S) while Mohamoud (2008) used 42 variables and regression analysis to derive regional equations of 15 percentiles of a flow duration curve for 29 watersheds (23.30–4,250.94 km^2) in Mid-Atlantic ecoregions (U.S). For regionalized hydrological model parameters, Bastola et al. (2008) used regionalization schemes of

multiple regressions, artificial neural network, and partial least–squares regression on 26 watersheds (25.5–8,936 km^2) to generate regionalized parameters of TOPMODEL (Beven et al., 1995). Heuvelmans et al. (2006) compared linear regression and artificial neural network regionalization schemes to generate SWAT (Neitsch et al., 2001) model parameters based on data of 25 watersheds (2.24–209.93 km^2).

Some of the approaches above, predict streamflow time series and thus implicitly predict streamflow magnitude and temporal sequence concurrently. An alternative approach that has not been fully explored is the conceptualization of streamflow as a composite of magnitude and sequence, such that each component (magnitude and sequence) can be modeled independently (Mohamoud, 2008, 2011). The magnitude can be modeled using the flow duration curve (FDC), which estimates percentage of time specific stream flows are equaled or exceeded based on historical flow records of a watershed (Ganora et al., 2009; Vogel and Fennessey, 1994, 1995). The FDC, based on daily streamflow, may be viewed as the hydrological system memory because it represents all flow magnitudes experienced by the watershed for the epoch under consideration. For this study the terms magnitude and FDC are used interchangeably. The streamflow sequence is defined as the timing or the temporal occurrence of streamflow magnitudes and therefore determines the date or the Julian day number when a specific magnitude occurred.

Many studies are replete with methods of estimating flow duration curves for ungauged watersheds.Commonly used approaches utilize regression analysis to develop regionalized flow equations linking watershed hydrologic response to climatic and geophysical characteristics (Castellarin et al., 2007; Chalise et al., 2003; Li et al., 2010; Mohamoud, 2008; Sanborn and Bledsoe, 2006). Also, the use of Monte Carlo resampling, bootstrap aggregating (bagging), and boosting (modified bagging) have become common place in hydrology (Anctil and Lauzon, 2004; Barnett et al., 2008; Boucher et al., 2010; Ebtehaj et al., 2010; Rustomji and Wilkinson, 2008; Selle and Hannah, 2010; Tiwari and Chatterjee, 2010). For Monte Carlo resampling, given mean and standard deviation of a sample, and an assumed distribution of the population, a new dataset can be generated. Bootstrapping involves randomly drawing values from a sample with replacement a

predefined number of times, while aggregation is taking the arithmetic mean of bootstrap sample. The concept of bagging assumes equal weight for each instance of the sample in the original sample such that each instance has an equal probability of selection during resampling. Boosting starts off with equal weight for each instance, however, the weights are increased for sample instances with poor classification or prediction results during training such that the probability of selecting such instances is higher during training. Boucher et al. (2010) and Anctil and Lauzon (2004) have shown that the performance of boosting in hydrology is comparable to performance of bagging and sometimes even better.

Studies by Mohamoud (2008) and Mohamoud (2011) detail the concept of streamflow separation. The streamflow magnitude was predicted using regionalized flow duration curve while streamflow sequence was predicted using sequence of a neighboring gauged watershed. Daily streamflow was then predicted using a reshuffling technique that sorted predicted streamflow magnitude with respect to the predicted sequence. Earlier literature that introduce the concept include Hughes and Smakhtin (1996) and Smakhtin et al. (1997) where the approach is used to patch and extend daily streamflow in South African watersheds. The two papers refer to the approach as a spatial interpolation method. Smakhtin et al. (1997) cites the ability to accurately predict daily FDC and use of appropriate source of sequence as the major challenges of method. None of the above literature quantifies the effect of drainage area of the donor watershed (gauged) and the effect of the distance between donor and target (ungauged) watersheds on the accuracy of the predicted daily streamflow.

The first objective of this study is to develop regionalized flow duration curves for three Mid-Atlantic physiographic provinces (Appalachian Plateau, Piedmont, and Ridge and Valley) using variables selected by a causal variable selection method, the HITON Markov boundary (HITON–MB) algorithm (Aliferis et al., 2010, 2003a; Fu and Desmarais, 2010) instead of the commonly used stepwise regression analysis and principal component analysis. Previous work by the authors (Ssegane et al., 2011a,b) demonstrated that HITON–MB better selected process driven variables than stepwise regression and principal component analysis. The HITON–MB algorithm seeks to

select causal variables by reconstructing the Markov Blanket of the response variable, therefore, the implied causation is probabilistic causation. The second objective examines the effect of the relative distance between the gauged (donor) and ungauged (target) watersheds, and their relative drainage areas, on the accuracy of the predicted daily streamflow using the streamflow separation (SFS) method. The third objective improves the accuracy of predicted daily streamflow by enhancing streamflow sequence of ungauged watershed from an ensemble of streamflow data of more than one gauged watershed. The explored ensemble techniques include: Pythagorean means (arithmetic, geometric, and harmonic); quadratic mean; weighted means (distance and drainage area); and use of Monte Carlo resampling from a normal distribution, bagging, and boosting. The fourth objective compares prediction performance of SFS method to simulations by Hydrological Simulation Program–Fortran (HSPF). HSPF is a semi–distributed hydrologic and water quality model developed by U.S. Environmental Protection Agency (Bicknell et al., 2001). The model has been applied on watersheds to estimate flow, sediment and nutrient transport (Anne and Uchrin, 2007; Chung and Lee, 2009; Diaz-Ramirez et al., 2008).

5.2 Methods

Study area and watershed data

Three Mid-Atlantic (U.S.) physiographic provinces (Figure 5.1; Appalachian Plateau, Piedmont, and the Ridge and Valley) are examined in this study. According to Haering and Evanylo (2006), the long term annual precipitation of the Mid–Atlantic region varies between 889 mm and 1270 mm. The landscape and geology vary between provinces (see Haering and Evanylo, 2006, chap. 2). The Appalachian Plateau are steeply sloping with forest coverage and are dominated by sandstones, siltstones, and shales; the Piedmont is characterized by gentle slopes underlain by igneous and metamorphic rocks; and the ridges of the Ridge and Valley are covered by rocky soils on steep slopes while the valleys are covered by limestones and shales.

[Figure 5.1 about here]

The streamflow data used with all watersheds spanned the same 42 year epoch (1965 to 2007). The sample data consisted of 28 watersheds (98–1,779 km^2) for the Appalachian Plateau, 19 watersheds (34.8–620 km^2) for the Piedmont, and 25 watersheds (48–1,857 km^2) for the Ridge and Valley. Data sources included the U.S. Geological Survey (USGS) for streamflow, the National Weather Service (NWS) for climatic data, the Natural Resources Conservation Service (NRCS) for STATSGO soil data, and the National Hydrology Dataset (NHD) compiled by USGS for sample watersheds with minimum level of urbanization and surface storage. Data extraction, preprocessing, and management used readily available geographical information service (GIS) tools such as ArcGIS (ESRI Inc. – proprietary), BASINS 4.0 (USEPA – public domain), and Systems for Automated Geoscientific Analyses (SAGA-GIS – public domain). Watershed characteristics were selected based on their likely contribution to the hydrological response as supported by information from the literature (e.g. Alcazar et al., 2008; Castellarin et al., 2007; Eng et al., 2007; Johnston and Shmagin, 2008; Mohamoud, 2008; Sanborn and Bledsoe, 2006; Sando et al., 2009; Srinivas et al., 2008).

Data preprocessing

The initial set of variables constituted 111 parameters (41 topographic, 39 climatic, 6 land use and land cover, and 25 soil and physical parameters) for each watershed. The use of few land use and land cover (LULC) variables was based on selection of watersheds that were predominantly forested. A correlation matrix of the variables was generated, from which pairwise variables with a correlation coefficient greater than 0.9 were identified for primary dimension reduction. Given two highly correlated variables, the variable which provided the highest incremental gain about the response variable was retained. The incremental gain (Schroedl, 2010) was computed as a function of: (1) mutual information between the variable and the response variable (variable relevance); (2) mutual information of different variables (variable redundancy); and (3) the increase of mutual information between previously selected variables and the response variable conditioned on a selected variable (conditional redundancy).

The incremental gain of highly correlated variables was computed for 19 flow percentiles and the average value was used as the representative information gain. The 19 flow percentiles included: high flows ($Q_{0.01}$, $Q_{0.05}$, $Q_{0.1}$, $Q_{0.5}$, Q_1, Q_5, Q_{10}); medium flows (Q_{20}, Q_{30}, Q_{40}, Q_{50}, Q_{60}, Q_{70}); and low–flows (Q_{80}, Q_{90}, Q_{95}, Q_{99}, $Q_{99.5}$, $Q_{99.9}$); where, Q_p represents the flow magnitude equaled or exceeded p percent of the flow record (1965 to 2007). This process reduced the 111 original variables to 92 variables (Refer to Ssegane et al. (2011a,b)).

Flow duration curves for each watershed in the three physiographic provinces were generated using daily streamflow and a Weibull plotting position for the 1965 to 2007 time period. The streamflow percentiles were normalized by dividing them with respective drainage areas to minimize the effect of drainage area on variable selection. A minimum–maximum standardization (Equation 5.1) was, then implemented on the logarithmic transformed streamflow percentiles and the explanatory variables.

$$F(S_k) = \frac{S_k - \min\{S\}}{\max\{S\} - \min\{S\}} \tag{5.1}$$

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; N is total number of streamflow records; $F(S_k)$ is the transformed k^{th} term of variable *S*; and S_k is the k^{th} term of variable *S*.

Variable Selection of watershed descriptors

The causal explorer toolkit was used to implement the causal variable selection method of HITON–MB (Aliferis et al., 2003b). The variable selection process entailed: 1) randomly deleting a single watershed; 2) running the variable selection on the remaining watersheds; 3) summarizing variables selected by each method for each of the 19 flow percentiles; 4) repeating steps 1 to 3 by excluding a different watershed with replacement on each run. The top five most selected variables for each flow percentile were used to generate the regional equations. Details for the Piedmont physiographic province are contained in Ssegane et al. (2011a).

Estimation of streamflow magnitude

Our conceptualization of daily streamflow separates streamflow magnitudes from their sequence. Therefore, given accurate characterization of streamflow magnitude (FDC) at ungauged watershed, one can reconstruct the daily streamflow using a surrogate sequence. This two step approach is referred to as streamflow separation (SFS). The fist step independently determines streamflow magnitude and streamflow sequence while the second step combines magnitude and sequence to generate streamflow time series.

The streamflow magnitudes were estimated using a regionalized flow duration curve (RFDC). The development of RFDC for each province entailed: 1) calculation of the 19 flow percentiles for each watershed; 2) variable selection for each flow percentile; 3) and determination of the regional equation for each flow percentile. The regional equations were determined by curve fitting of selected variables to predetermined functional forms. Several functional forms were explored, however, the two optimum forms of Equations 5.2 and 5.3 gave the highest predictive power across all 19 streamflow percentiles. For each physiographic province and streamflow percentile, the functional form and the two or three variables (out of top five variables) that best minimized the sum of square errors during calibration were selected as the optimum predictor equations.

$$Q_p = \frac{aDA}{1 + X_1}\left(b + c\frac{X_2}{X_3}\right) \tag{5.2}$$

$$Q_p = a10^b DA^c X_1^d X_2^e + f \tag{5.3}$$

where Q_p is the p^{th} flow percentile; DA is watershed drainage area ; X_i is a selected variable; and a,b, c, d, e, and f are optimized regional coefficients that vary across physiographic provinces.

The above described procedure only generates 19 points on a flow duration curve with probabilities ranging from 0.01 to 99.9 and corresponding flow percentiles of $Q_{0.01}$ to $Q_{99.9}$. However, if the duration of interest is 11 years between 01 January 2000 to 31 December 2010, one needs a total of 4017 points on the FDC with corresponding probabilities ranging from 0.025 to 99.975. This study used linear interpolation and extrapolation to generate all points of the FDC for the period under consideration.

Estimation of streamflow sequence

The streamflow sequence was estimated using the sequence obtained from neighboring gauged watershed, also referred to as the donor watershed. Generation of a sequence involves chronicling the Julian day number and rank for the specific streamflow magnitude. For example, given daily streamflow data for a non leap year, the Julian day numbers range from 1 to 365 and the sequential value on each day is the rank of the magnitude for that day. Thus the sequence will take on numbers between 1 and 365. For 11 years between the first of January 2000 to the thirty first of December 2010, the sequence will take on numbers between 1 and 4017. Therefore, the sequential values are determined by the period under consideration. For more information on streamflow sequence estimation, readers may refer to Hughes and Smakhtin (1996); Mohamoud (2008, 2011); Smakhtin et al. (1997).

Effect of distance and drainage area

We quantified the effect of distance and drainage area of donor watersheds by using the sequences of four closest donor watersheds. The use of the term closest is limited to watersheds in our database for each province and therefore does not characterize the entire true geographical neighborhood. The geographical proximity between gaged and ungauged watersheds was calculated by computing for the Euclidean distance between the centroids of target and all remaining watersheds. The effect of drainage area was determined using the ratio of the donor to target drainage area (DAR). The effect of distance and drainage area were tested using the student t–statistic at 5

% level of statistical significance. The performance of predicted daily streamflow was used as the response variable, while predicted magnitude (FDC) performance, DAR, and Euclidean distance used as the explanatory variables.

Use of Pythagorean, quadratic, and weighted means

This study explored ways of improving the predicted sequence by using Pythagorean, quadratic, and weighted means of daily streamflow values of the two closest donor watersheds. The schematic representation of the process is depicted by Figure 5.2. Preliminary analysis showed that use of two closest donor watersheds gave better results than use of three or four. This was attributed to introduction of more errors as the distance from the ungauged watershed increases because effects of watershed heterogeneity become more apparent. The Pythagorean means (Eves, 2003) included arithmetic mean(Equation 5.4), geometric mean (Equation 5.5), and the harmonic mean (Equation 5.7). The weighted means used the euclidean distance between the centroids of target and donor watersheds, and the drainage area as the weighting factors. The use of Julian day means generated six additional sets of new sequences.

$$A_i = \frac{1}{n}\sum_{j=1}^{n} x_i^j \tag{5.4}$$

$$G_i = \sqrt[n]{\prod_{j=1}^{n} x_i^j} \tag{5.5}$$

$$\frac{1}{H_i} = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{x_i^j} \tag{5.6}$$

$$Q_i = \left[\frac{1}{n}\sum_{j=1}^{n} \left(x_i^j\right)^2\right]^{\frac{1}{2}} \tag{5.7}$$

where A_i, G_i, H_i, and Q_i are the arithmetic, geometric, harmonic, and quadratic means on the i_{th} Julian day number, respectively; x_i^j is the streamflow magnitude on i^{th} Julian day number of the j^{th} donor watershed; and n is the total number of donor watersheds. For this study only two donor watersheds were used.

Monte Carlo resampling, bagging, and boosting

For this study both resampling techniques were performed 100 times on each Julian day number and the arithmetic mean of the new sample was the new estimated value for sequence prediction. Sampling more than 100 times did not improve the results. This might be attributed to a small original sample (17 values) for each Julian day, such that multiple sampling beyond 100 just replicates the same distribution as 100 samples. The Julian day number sample consisted of values from the four donor watersheds and values from Pythagorean, quadratic, and weighted means, in addition to the minimum and maximum of the two closest donor watersheds. Regarding Monte Carlo resampling, assumptions of normal and log–normal distributions were tested. The preliminary results for log–normal distribution were relatively poor and therefore subsequent analysis assumed only normal distribution.

For this study the concept of boosting was implemented by increasing the probability of selecting values from better performing ensemble techniques. For example, preliminary results showed that distance weighted and the geometric means performed better than quadratic and arithmetic means, therefore, the new sample for bootstrapping was stacked with more values from these means.

Combination of predicted magnitude and sequence

The procedure of combining predicted streamflow magnitude and sequence to generate daily streamflow is described as follows

Practically, Steps 2 and 3 can be combined into one step by directly sorting the streamflow of the donor watershed in a descending order, however, the description above tries to emphasize the rank as the main focus compared to the magnitude of the donor watershed. For details refer to (Mohamoud, 2011).

Step	Action
1	Start with daily streamflow data of a donor watershed covering the same period as the predicted FDC. By default the data is sorted in an ascending order of the Julian day number. Therefore, you should have two columns. Column 1 is the Julian day number and column 2 the corresponding streamflow magnitude of the donor watershed.
2	Generate column 3 as the rank of the streamflow magnitude corresponding to the respective Julian day number.
3	Sort the rank (column 3) in a descending order while maintaining the corresponding Julian day number and donor streamflow magnitude
4	Copy and paste the predicted magnitudes (FDC flow percentiles) into a new column (column 4). Recall, by default the predicted magnitudes are already sorted in a descending order.
5	Re–sort the Julian day number (column 1) in an ascending order while maintaining the corresponding values in columns two to four.
6	The rearranged magnitudes in column 4 are the predicted streamflow for the corresponding Julian day number.

Jackknife cross–validation

To assess the predictive performance of the method, a process of jackknife cross–validation was implemented. The process consisted of the following steps

For both R^2 ($0 \leq R^2 \leq 1.0$) and NSE ($-\infty \leq NSE \leq 1.0$), a value of one is optimum while for MAE and $RMSE$, better prediction power coincides with smaller values.

Step	Action
1	Elimination of the watershed of interest (target) from the sample data
2	Generation of regionalized flow duration curves (RFDC) based on the remaining data
3	Prediction of magnitude, sequence, and daily streamflow of the target watershed based on methods described in sections 5.2, 5.2, and 5.2
4	Given predicted and observed daily streamflow, evaluation of predictive performance using coefficient of determination (R^2), Nash–Sutcliffe coefficient of efficiency (NSE), mean absolute error (MAE), and root mean square error ($RMSE$)

Comparison of streamflow separation to HSPF simulations

One watershed from each province was selected for comparing HSPF simulations to results of the method described above. West Branch Susquehanna River, PA (01541000) was selected for the Appalachian, Nottoway River near Rawlings, VA (02044500) for the Piedmont, and Cheat River near Parsons, WV (03069500) for the Ridge and Valley. Digital elevation model (DEM) resolution of 30 m was used to generate watershed boundary and sub–watersheds for each watershed. U.S. EPA - BASINS 4.0 program was used to download the respective data sets and preparing the input files for WinHSPF simulations. The following weather stations were used for input climatic data: PA367167 (1982–2007) for Susquehanna, VA441322 (1972–2006) for Nottoway, and WV463464 (1973–2006) for Cheat River. The HSPF simulations were implemented using WinHSPF for the 1985 to 2005 epoch. Two simulation scenarios were implemented: one focused on forward simulation without calibration (HSPF) while the second focused on automated calibration using parameter estimation software, PEST (HSPF_PEST).

5.3 Results and discussions

Streamflow magnitude

Tables 5.1 and 5.2 define watershed and climatic variables while tables 5.3 to 5.5, depict equations of regionalized flow duration curve (RFDC) for the three provinces. The equations were developed using all data samples for each ecoregion. The coefficients of determinations (R^2) corresponding to predicted 19 streamflow percentiles show better predictions ($R^2 \geq 0.8$) for high ($Q_{0.01}$–Q_{10}) and medium (Q_{20}–Q_{70}) flows compared to low–flows (Q_{80}–$Q_{99.9}$). The same trend was observed during the Jackknife cross–validation (section 5.2 and Figure 5.3). The most selected variables for magnitude prediction are dominated by: topography and climate for Appalachian Plateau; topography and soils for Piedmont; and soils and climate for Ridge and Valley.

[Table 5.1 about here]

[Table 5.2 about here]

[Table 5.3 about here]

[Table 5.4 about here]

[Table 5.5 about here]

Figure 3.2 compares predicted and observed streamflow magnitudes for sample watersheds from each province. The graphs compare prediction results of the RFDC to predictions by HSPF hydrologic model before calibration (HSPF) and after calibration using PEST (HSPF_PEST). For all the three sample cases, the RFDC gave better predictions than HSPF, even after PEST calibration.

[Figure 5.3 about here]

For some instances (Nottoway River near Rawlings, VA; Figure 5.3), there is better visual agreement between HSPF predictions after calibration and observed data than with RFDC prediction, yet the Nash–Sutcliffe efficiency (NSE) is greater for RFDC predictions. This observation is attributed to better predictions at high streamflow magnitudes by RFDC compared to medium and low magnitudes. This skews the overall performance metric because of the larger errors for high flows compared to medium and low flows. The results also show that watershed and climatic variables selected by a causal variable selection algorithm (HITON–MB) have high predictive power for daily streamflow magnitudes. The high predictive power does not mean the regionalized flow duration curve better represents the underlying hydrological process than HSPF since predictive power may depend on measurement noise than the hydrological processes. However, use of causal variable selection algorithm seeks to minimize the effect of measurement noise on selected features compared to stepwise regression.

Effect of drainage area, distance, and land cover on predicted sequence

Figure 5.4 illustrates the orientation, size, and euclidean distance between centroids of the four closest donor watersheds in the neighborhood of the target watershed (Cheat River 03069500). Statistical analysis of the effect of drainage area ratio (ratio of donor to target drainage area) showed no significant effect by drainage area of the donor watershed on accuracy of predicted daily streamflow at 5 % level of significance. However, the distance between donor and target watersheds had a significant effect. On average, the accuracy (NSE) of the predicted magnitude and distance explained 89 % of the total variability of accuracy of the predicted streamflow.

Figure 5.5 quantifies the relationship between accuracy of predicted magnitude, distance, and accuracy of predicted streamflow. The figure seeks to identify an upper limit of the radius of influence from which candidate donor watersheds can be selected. On average, a distance of 30 km can be considered appropriate given the accuracy of the predicted magnitude is greater or equal to 0.8 ($NSE \geq 0.8$). The figure also shows high prediction accuracy using sequences of donor watersheds whose distance is greater than 30 km and vice–versa. This observation is attributed

to other factors such as similarity in the level of urbanization and surface storage (sum of percent surface water and wetlands).

[Figure 5.4 about here]

[Figure 5.5 about here]

Additional statistical assays of watersheds in the Appalachian Plateau showed that the level of urbanization and surface storage between donor and target watersheds had a significant effect on the accuracy of the predicted daily streamflow. Inclusion of the square of the variable differences between donor and target watersheds for the above variables increased the explained variability from 89 % to 92 %. The assays further showed that a 1 km increase in the euclidean distance between donor and target watersheds, decreased the NSE of the predicted daily streamflow by 0.005 from the NSE of the predicted magnitude. Also, 1 % increase or decrease between the surface storage and level of urbanization of donor and target watersheds decreased the streamflow NSE by 0.05 and 0.0006, respectively.

The above analysis is supported by streamflow predictions at Blockhouse Creek near English center (01549500). The closest donor watershed in our sample data is Tioga (01518000) at 29.7 km while the second closest is Pine Creek (01548500) at 38.1 km. However, use of sequence from Tioga gave NSE of 0.52 compared to NSE of 0.76 for Pine Creek. Assessment of surface storage for both watersheds showed comparable storage at 0.7 %, but the level of urbanization at Pine Creek (4.8 %) was more comparable to that at Blockhouse (5.2 %) than at Tioga (2.1 %).

Improvement of sequence prediction

Tables 5.6 to 5.8 depict accuracy of predicted daily streamflow using sequence of neighboring four watersheds and sequences generated by ensemble methods. The first row demonstrates that use of true sequence yields the exact accuracy as the predicted magnitude. With a few exceptions, the accuracy then decreases as the distance between centroids of donor and target watersheds increases. The results of row 6 to row 13 are based on sequence derived from respective operations on each

Julian day number (JDN) for the closest two donor watersheds. Results for row 14 are based on a sequence of the arithmetic mean on each JDN for the two watersheds whose flow duration curves are most similar to the predicted flow duration curve. Rows 15 to 17 are results of Monte–Carlo resampling, bagging, and boosting (refer to section 5.2). The improvements were more prevalent in watersheds where the distance between the first and second closest donor watersheds was less than 20 km.

[Table 5.6 about here]

[Table 5.7 about here]

[Table 5.8 about here]

[Figure 5.6 about here]

The sequence generated from geometric mean of two closest donor watersheds gave better predictions than any other mean. This is explained by the observation that for a sample with large and small values, the geometric mean provides the best measure of central tendency while the arithmetic and the harmonic means are biased toward the large and small values, respectively. The results derived from geometric mean, Monte–Carlo resampling, bagging, and boosting gave comparable prediction performance, however, bagging and boosting better improved sequence prediction than any other method. Therefore, subsequent results are based on boosting (Figure 5.6). Tables 5.9 to 5.11 show prediction accuracy of magnitude using regionalized flow duration curve (RFDC) and corresponding accuracy of daily streamflow based on sequence prediction by boosting. Some cases of high accuracy of predicted magnitude (NSE_{fdc}) with low accuracy of predicted daily streamflow (NSE_{ts}) are related to location of the closest donor watershed. For example, the closest watershed in the sample data for Bluestone (03179000; Table 5.9) is 67 km ($NSE_{fdc} = 0.94$; $NSE_{ts} = 0.46$) while for Middle Island (03114500; Table 5.9) is 108.2 km ($NSE_{fdc} = 0.95$; $NSE_{ts} = 0.24$).

[Table 5.9 about here]

[Table 5.10 about here]

[Table 5.11 about here]

Analysis of the results showed that the procedure improved the predicted daily streamflow on 57 % of the watersheds in Appalachian Plateau, 81 % of Piedmont watersheds, and 55 % of Ridge and Valley watersheds. One plausible explanation for difference in levels of improvement, is the level of homogeneity of sampled watersheds for each physiographic province. Earlier work by Ssegane et al. (2011b) showed that for the Appalachian Plateau, 52 % of the sampled watershed were hydrologically similar while 75 % for the Piedmont, and 34.5 % for the Ridge and Valley. The level of hydrologic homogeneity among watersheds in each province was computed on annual flows from 1966 to 2007 using Hosking and Wallis (1997) homogeneity tests. The level of hydrologic homogeneity and the level of sequence improvement follow a similar trend across physiographic provinces. Therefore, although distance has primary effect on sequence estimation, other factors that are specific to each province are relevant. Ssegane et al. (2011b) showed that climatic and topographic variables were major hydrological drivers for watersheds in Appalachian Plateau; topographic and soil variables for Piedmont; and, topographic, climatic, and soil variables for Ridge and Valley.

Implications for hydrologic modeling

Figure 5.3 illustrated that streamflow magnitude predictions by regionalized flow duration curve gave better predictions for high flows compared to calibrated HSPF hydrological model. However, the PEST calibrated HSPF model gave better predictions for medium and low flow magnitudes. The two approaches can be combined to utilize their strength and thus provide better characterization of entire flow duration curve (streamflow magnitude). Figure 5.6 shows a time shot of predicted daily streamflow using streamflow separation (SFS method) in contrast to observed data and predictions by HSPF. The results show that even after PEST calibration, the HSPF (HSPF_PEST) explained less than 50 % of the total variability of observed streamflow for 1985 to 2005 epoch

compared to over 85 % by the SFS method yet the predicted magnitudes were comparable (Figure 5.3). This observation is indicative of the limitation of HSPF to concurrently capture both magnitude and sequence. The HSPF prediction accuracy greatly depends on quality of the climatic data (precipitation in particular) and therefore, the poor sequence prediction by HSPF may guide a modeler to revisit input data and examine whether the data captures most of the temporal and spatial variation of precipitation in the watershed. In absence of better input data, the independent prediction of streamflow sequence can be implemented following the procedure outlined in this study.

[Figure 5.6 about here]

5.4 Conclusions

The study set out to predict daily streamflow for ungauged watersheds by independently predicting streamflow magnitude and sequence. The streamflow magnitude was estimated using regionalized flow duration curves while sequence was estimated by transferring donor (gauged) watershed sequence to the target (ungauged) watershed. The effects of drainage area and distance between donor and target watersheds on the accuracy of predicted sequence were assessed. The relative drainage area of the donor watershed was not statistically significant during sequence prediction, however, the distance was. Therefore, the Euclidean distance between centroids of donor and target watersheds had primary effects on accuracy of predicted sequence, with better results for a distance less than 40 km. Other factors such as surface storage and level of urbanization had secondary effects. The geometric mean of streamflow of two closest donor watersheds gave better sequence prediction than arithmetic, harmonic, and quadratic means.

Ensemble methods of bagging and boosting better improved sequence prediction than use of just the closest donor watershed given the distance between the first and second closest gauged watersheds was less than 20 km. Across all watersheds in each province, the boosting method compared to use of the closest donor watershed, increased NSE of predicted daily streamflow by

0.040 for the Appalachian watersheds, by 0.042 for the Piedmont, and by 0.100 for the Ridge and Valley. Specifically, boosting improved the the accuracy (NSE) of predicted daily streamflow for West Branch Susquehanna River from 0.88 to 0.94; from 0.77 to 0.83 for Nottoway River; and from 0.76 to 0.88 for Cheat River.

The performance of the streamflow separation method and use of the regionalized flow duration curves may be limited to watersheds with drainage areas ranging between 98–1,779 km^2 for the Appalachian Plateau, 34.8–620 km^2 for the Piedmont, and 48–1,857 km^2 for the Ridge and Valley. The approach also poorly predicted low flows.

BIBLIOGRAPHY

Alcazar, J., Palau, A., Vega-Garcia, C., 2008. A neural net model for environmental flow estimation at the Ebro River Basin, Spain [electronic resource]. Journal of hydrology 349 (1-2), 44–55. 131

Aliferis, C., Statnikov, A., Tsamardinos, I., Mani, S., Koutsoukos, X., 2010. Local causal and Markov blanket induction for causal discovery and feature selection for classification part I: algorithms and empirical evaluation. The Journal of Machine Learning Research 11, 171–234. 129

Aliferis, C., Tsamardinos, I., Statnikov, A., 2003a. HITON: a novel Markov Blanket algorithm for optimal variable selection. In: American Medical Informatics Association Annual Symposium Proceedings. Vol. 2003. American Medical Informatics Association, p. 21. 129

Aliferis, C., Tsamardinos, I., Statnikov, A., Brown, L., 2003b. Causal explorer: A causal probabilistic network learning toolkit for biomedical discovery. In: International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS03). Citeseer, pp. 371–376. 132

Anctil, F., Lauzon, N., 2004. Generalisation for neural networks through data sampling and training procedures, with applications to streamflow predictions. Hydrology and Earth System Sciences 8 (5), 940–958. 128, 129

Anne, A., Uchrin, C., 2007. Modeling the hydrology and water quality using BASINS/HSPF for the upper Maurice River watershed, New Jersey. Journal of Environmental Science and Health Part A 42 (3), 289–303. 130

Bárdossy, A., 2007. Calibration of hydrological model parameters for ungauged catchments. Hydrology and Earth System Sciences 11 (2), 703–710. 127

Barnett, T., Pierce, D., Hidalgo, H., Bonfils, C., Santer, B., Das, T., Bala, G., Wood, A., Nozawa, T., Mirin, A., et al., 2008. Human-induced changes in the hydrology of the western United States. science 319 (5866), 1080. 128

Bastola, S., Ishidaira, H., Takeuchi, K., 2008. Regionalisation of hydrological model parameters under parameter uncertainty: A case study involving TOPMODEL and basins across the globe. Journal of Hydrology 357 (3-4), 188–206. 127

Besaw, L., Rizzo, D., Bierman, P., Hackett, W., 2010. Advances in ungauged streamflow prediction using artificial neural networks. Journal of Hydrology 386 (1-4), 27–37. 127

Beven, K., Lamb, R., Quinn, P., Romanowicz, R., Freer, J., Singh, V., 1995. TOPMODEL. Computer models of watershed hydrology., 627–668. 128

Bicknell, B., Imhoff, J., Kittle Jr, J., Jobes, T., Donigian Jr, A., 2001. Hydrological Simulation Program-Fortran (HSPF). User's Manual for Release 12. US EPA National Exposure Research Laboratory, Athens, GA, in cooperation with US Geological Survey. Water Resources Division, Reston, VA. 130

Boucher, M., Laliberté, J., Anctil, F., 2010. An experiment on the evolution of an ensemble of neural networks for streamflow forecasting. Hydrology and Earth System Sciences 14 (3), 603–612. 128, 129

Castellarin, A., Camorani, G., Brath, A., 2007. Predicting annual and long–term flow duration curves in ungauged basins. Advances in Water Resources 30 (4), 937–953. 128, 131

Chalise, S., Kansakar, S., Rees, G., Croker, K., Zaidman, M., 2003. Management of water resources and low flow estimation for the Himalayan basins of Nepal. Journal of Hydrology 282 (1-4), 25–35. 128

Chung, E., Lee, K., 2009. Prioritization of water management for sustainability using hydrologic simulation model and multicriteria decision making techniques. Journal of Environmental Management 90 (3), 1502–1511. 130

Diaz-Ramirez, J., Alarcon, V., Duan, Z., Tagert, M., McAnally, W., Martin, J., O'Hara, C., 2008. Impacts of land use characterization in modeling hydrology and sediments for the Luxapallila Creek Watershed, Alabama and Mississippi. Transactions of the ASABE 51 (1), 139–151. 130

Ebtehaj, M., Moradkhani, H., Gupta, H., 2010. Improving robustness of hydrologic parameter estimation by the use of moving block bootstrap resampling. Water Resources Research 46 (7), W07515. 128

Eng, K., Milly, P. C. D., Tasker, G. D., 2007. Flood regionalization: A hybrid geographic and predictor variable region of influence regression method. Journal of Hydrologic Engineering 12 (6), 585–591. 131

Engeland, K., Hisdal, H., 2009. A comparison of low flow estimates in ungauged catchments using regional regression and the HBV–model. Water resources management 23 (12), 2567–2586. 127

Eves, H., 2003. Means Appearing in Geometric Figures. Mathematics magazine 76 (4), 292. 135

Fu, S., Desmarais, M. C., 2010. Markov Blanket based Feature Selection: A Review of Past Decade. In: Proceedings of the World Congress on Engineering 2010. Vol. I. 129

Ganora, D., Claps, P., Laio, F., Viglione, A., 2009. An approach to estimate nonparametric flow duration curves in ungauged basins. Water Resources Research 45, 10418. 128

Garcia-Martinó, A., Scatena, F., Warner, G., Civco, D., 1996. Statistical low flow estimation using GIS analysis in humid Montane regions in Puerto Rico. Water Resources Bulletin 32 (6), 1259–1271. 127

Gotzinger, J., Bárdossy, A., 2007. Comparison of four regionalisation methods for a distributed hydrological model. Journal of Hydrology 333 (2-4), 374–384. 127

Haering, K., Evanylo, G., 2006. The Mid-Atlantic nutrient management handbook. The Mid-Atlantic regional Water Program. 130

Heuvelmans, G., Muys, B., Feyen, J., 2006. Regionalisation of the parameters of a hydrological model: Comparison of linear regression models with artificial neural nets. Journal of Hydrology 319 (1-4), 245–265. 128

Hosking, J., Wallis, J., 1997. Regional frequency analysis: an approach based on L-moments. Cambridge Univ Pr. 143

Hughes, D., Kapangaziwiri, E., Sawunyama, T., 2010. Hydrological model uncertainty assessment in southern Africa. Journal of Hydrology 387 (3-4), 221–232. 127

Hughes, D., Smakhtin, V., 1996. Daily flow time series patching or extension: a spatial interpolation approach based on flow duration curves. Hydrological Sciences Journal-Journal des Sciences Hydrologiques 41 (6), 851–872. 129, 134

Johnston, C. A., Shmagin, B. A., 2008. Regionalization, seasonality, and trends of streamflow in the US Great Lakes Basin. Journal of Hydrology 362 (1-2), 69–88. 131

Kokkonen, T., Jakeman, A., Young, P., Koivusalo, H., 2003. Predicting daily flows in ungauged catchments: model regionalization from catchment descriptors at the Coweeta Hydrologic Laboratory, North Carolina. Hydrological Processes 17 (11), 2219–2238. 127

Laaha, G., Bloschl, G., 2006. A comparison of low flow regionalisation methods–catchment grouping. Journal of Hydrology 323 (1-4), 193–214. 127

Li, M., Shao, Q., Zhang, L., Chiew, F., 2010. A new regionalization approach and its application to predict flow duration curve in ungauged basins. Journal of Hydrology 389 (1-2), 137–145. 128

Mazvimavi, D., Meijerink, A., Savenije, H., Stein, A., 2005. Prediction of flow characteristics using multiple regression and neural networks: A case study in Zimbabwe. Physics and Chemistry of the Earth, Parts A/B/C 30 (11-16), 639–647. 127

Mohamoud, Y., 2008. Prediction of daily flow duration curves and streamflow for ungauged catchments using regional flow duration curves. Hydrological Sciences Journal 53 (4), 706–724. 127, 128, 129, 131, 134

Mohamoud, Y., 2011. Streamflow Separation Method for Ungauged Watersheds. *(under review)*. Journal of Hydrology. 128, 129, 134, 136

Neitsch, S., Arnold, J., Kiniry, J., Williams, J., King, K., 2001. Soil and water assessment tool theoretical documentation version 2000. Grassland, Soil and Water Research Laboratory, Temple, Texas. 128

Pavizham, A., Sudheer, K., Chaubey, I., 2009. Predictions in Ungauged Basins using Distributed Hydrologic Models: Regionalization of Parameters and Quantification of Predictive Uncertainty. In: AGU Spring Meeting Abstracts. Vol. 1. p. 01. 127

Rustomji, P., Wilkinson, S., 2008. Applying bootstrap resampling to quantify uncertainty in fluvial suspended sediment loads estimated using rating curves. Water Resources Research 44 (9), W09435. 128

Sanborn, S., Bledsoe, B., 2006. Predicting streamflow regime metrics for ungauged streamsin Colorado, Washington, and Oregon. Journal of Hydrology 325 (1-4), 241–261. 128, 131

Sando, S. K., Fish, U., Service., W., (U.S.), G. S., 2009. Estimation of streamflow characteristics for Charles M. Russell National Wildlife Refuge, northeastern Montana. Scientific investigations report. U.S. Geological Survey, Reston, Va. 131

Saxton, K., Romberger, W., Papendick, J., et al., 1986. Estimating Generalized Soil-water Characteristics from Texture1. Soil Science Society of America Journal 50 (4), 1031. 160

Schroedl, S., August 2010. Feature Selection Based on Interaction Information. 131

Selle, B., Hannah, M., 2010. A bootstrap approach to assess parameter uncertainty in simple catchment models. Environmental Modelling & Software 25 (8), 919–926. 128

Sivapalan, M., 2003. Prediction in ungauged basins: a grand challenge for theoretical hydrology. Hydrological Processes 17 (15), 3163–3170. 127

Smakhtin, V., Hughes, D., Creuse-Naudin, E., 1997. Regionalization of daily flow characteristics in part of the Eastern Cape, South Africa. Hydrological sciences journal 42 (6), 919–936. 129, 134

Srinivas, V., Tripathi, S., Rao, A., Govindaraju, R., 2008. Regional flood frequency analysis by combining self-organizing feature map and fuzzy clustering. Journal of Hydrology 348 (1-2), 148–166. 127, 131

Ssegane, H., Tollner, E. W., Mohamoud, Y. M., Rasmussen, T. C., Dowd, J. F., 2011a. Advances in Variable Selection Methods I: Causal Selection Methods versus Stepwise Regression and Principle Components on Data of Known and Unknown Functional Relationships. *(under review)*. Journal of Hydrology. 129, 132

Ssegane, H., Tollner, E. W., Mohamoud, Y. M., Rasmussen, T. C., Dowd, J. F., 2011b. Advances in Variable Selection Methods II: Effect of Variable Selection Method on Classification of hydrologically similar watersheds in three Mid–Atlantic Ecoregions *(under review)*. Journal of Hydrology. 129, 132, 143

Tiwari, M., Chatterjee, C., 2010. Development of an accurate and reliable hourly flood forecasting model using wavelet-bootstrap–ANN (WBANN) hybrid approach. Journal of Hydrology. 128

Vogel, R., Fennessey, N., 1994. Flow-Duration Curves I: New Interpretation and Confidence Intervals. Journal of Water Resources Planning and Management 120 (4), 485–504. 128

Vogel, R., Fennessey, N., 1995. Flow Duration Curves II: A Review of Applications in Water Resources Planning 1. JAWRA Journal of the American Water Resources Association 31 (6), 1029–1039. 128

Zhu, Y. H., Day, R. L., 2009. Regression modeling of streamflow, baseflow, and runoff using geographic information systems. Journal of Environmental Management 90 (2), 946–953. 127

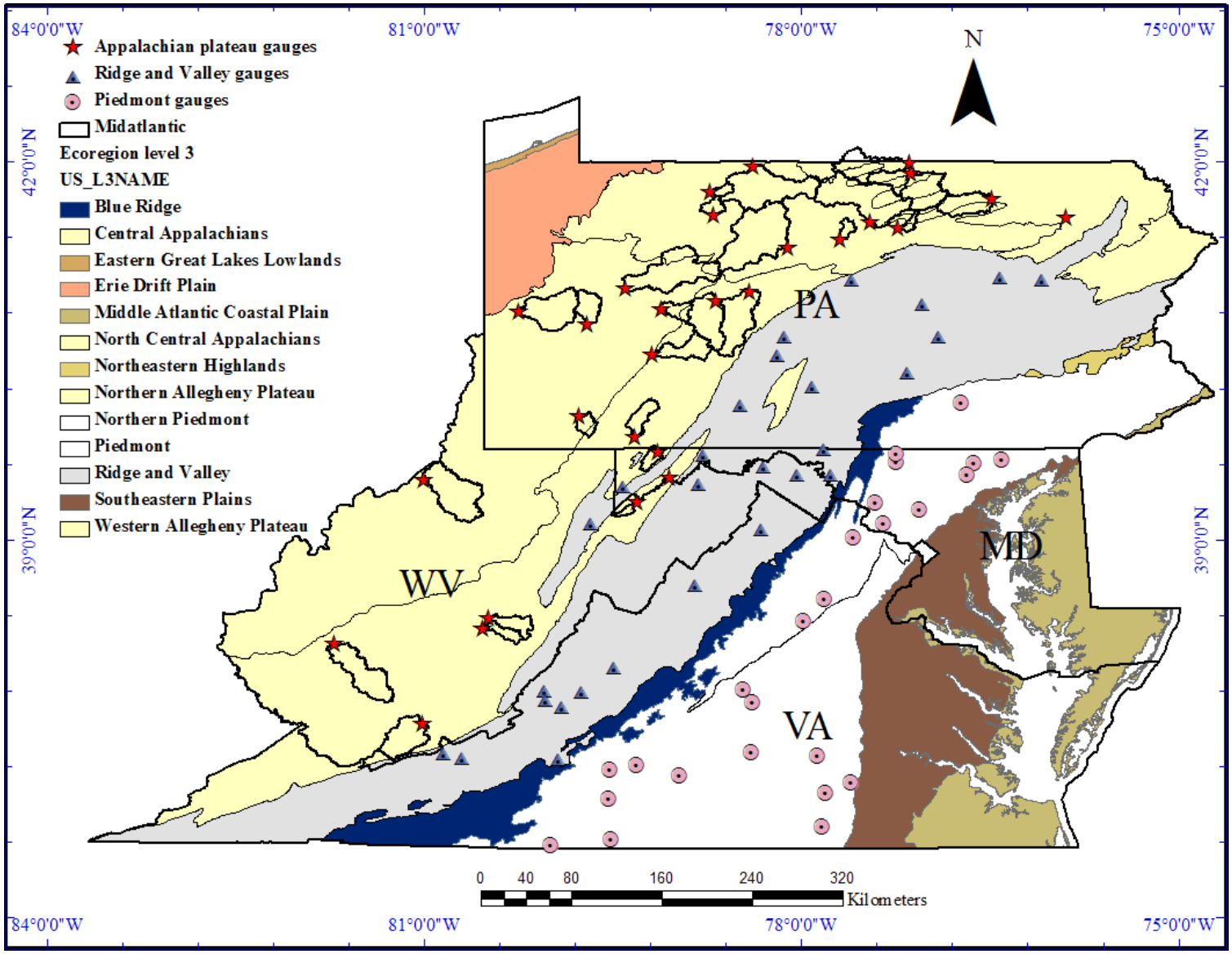

Figure 5.1: Location of watersheds in the Appalachian Plateau (98 – 1779 km^2), Piedmont (34.8 – 620 km^2), and Ridge and Valley (48 – 1857 km^2), Mid-Atlantic, USA.

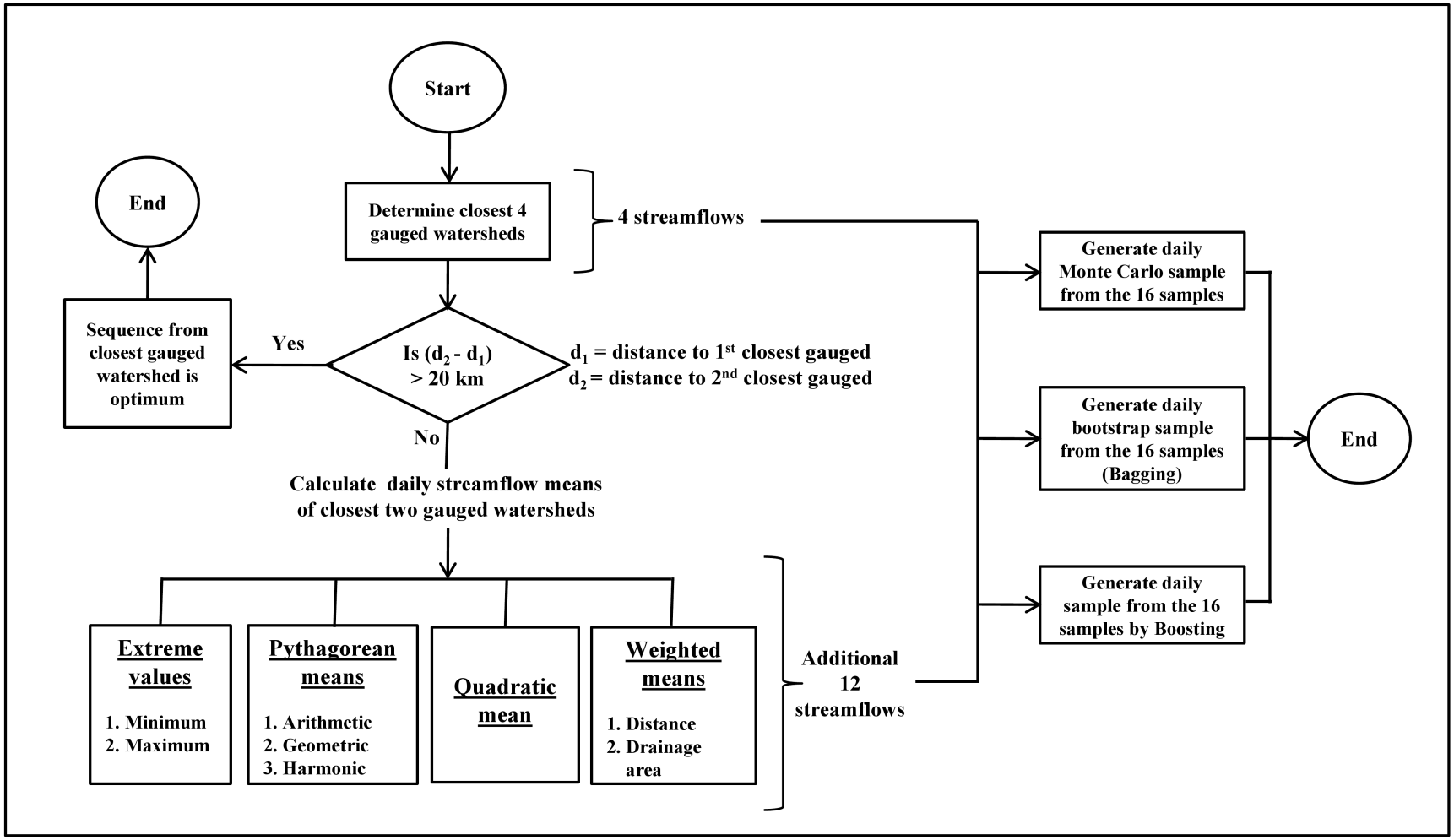

Figure 5.2: Procedure for sequence improvement. This is achieved when Euclidean distance between 1^{st} and 2^{nd} closest gauged watersheds is less than 20 km. If the distance is greater, then sequence from 1^{st} closest is the best estimate.

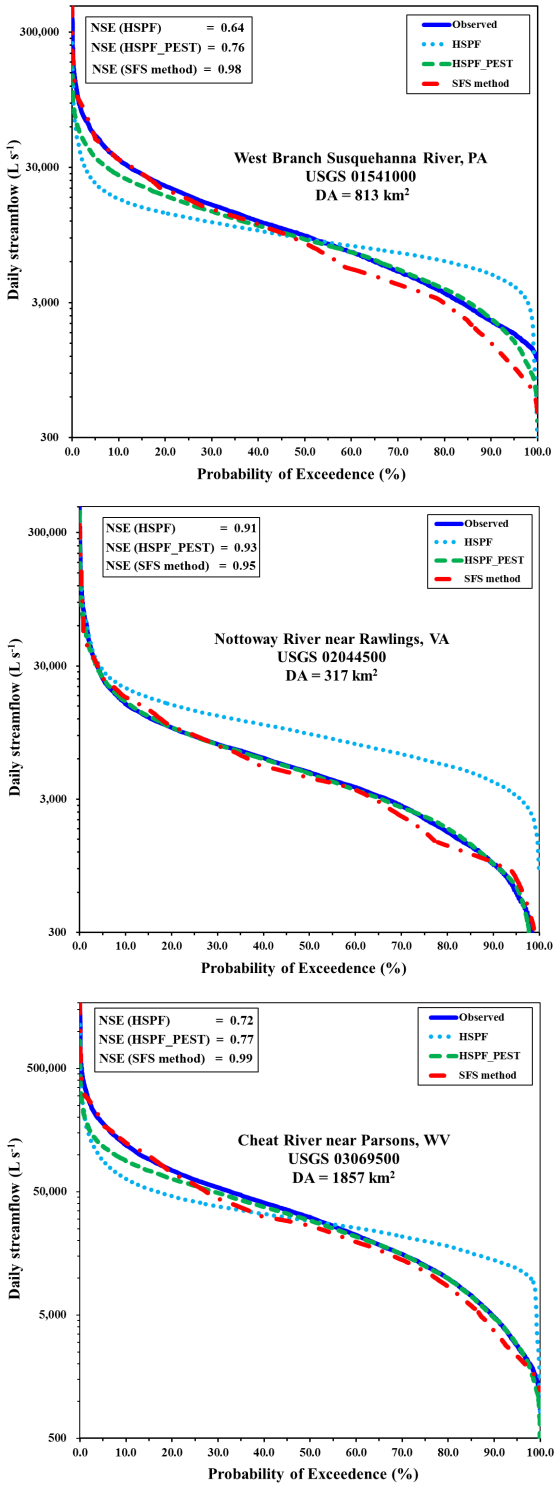

Figure 5.3: Observed and predicted flow duration curves for sample watersheds from Appalachian(top), Piedmont (middle), and Ridge and Valley (bottom). The results include HSPF prediction without (HSPF) and with PEST calibration (HSPF_PEST); and use of streamflow separation method (SFS method) for the 1985 to 2005 epoch

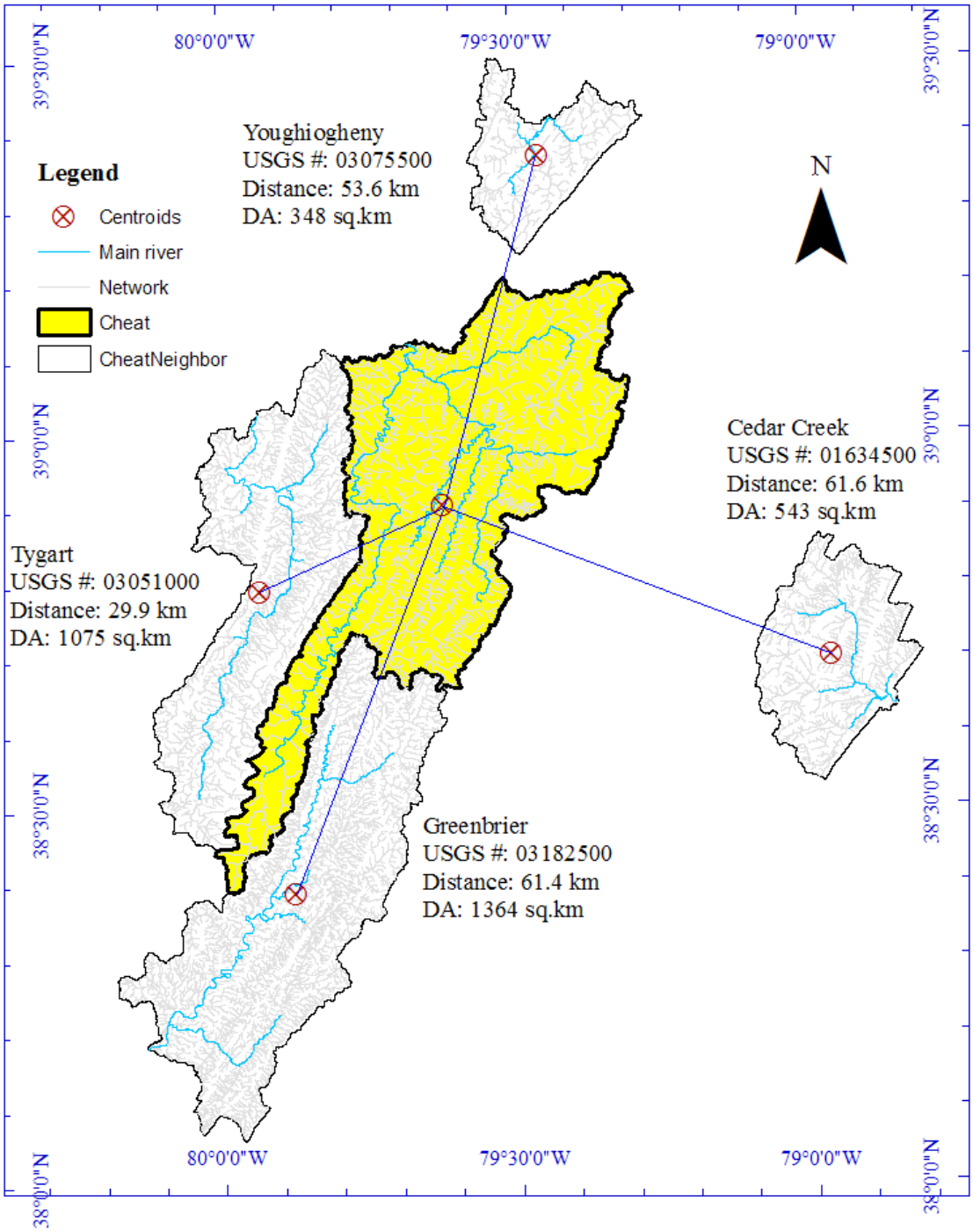

Figure 5.4: Proximity and orientation of four neighboring watersheds to Cheat river near Parsons, WV (USGS 03069500; Ridge and Valley ecoregion). Neighborhood is based on only the sample data. The highlighted watershed is the Cheat river watershed.

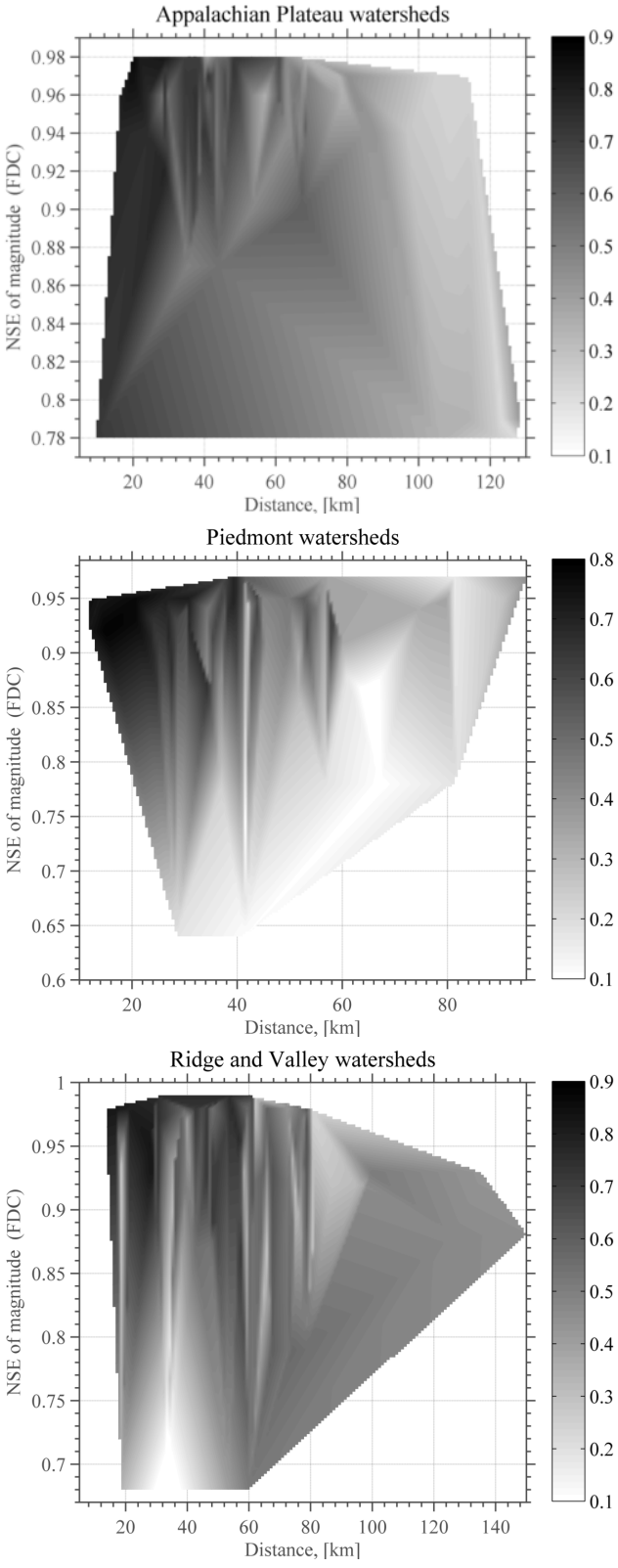

Figure 5.5: Relationship between accuracy of predicted daily streamflow (colorbar), the accuracy of predicted magnitude (y-axis), and the distance between centroids of target and donor watershed (x-axis) for Appalachian(top), Piedmont (middle), and Ridge and Valley (bottom).

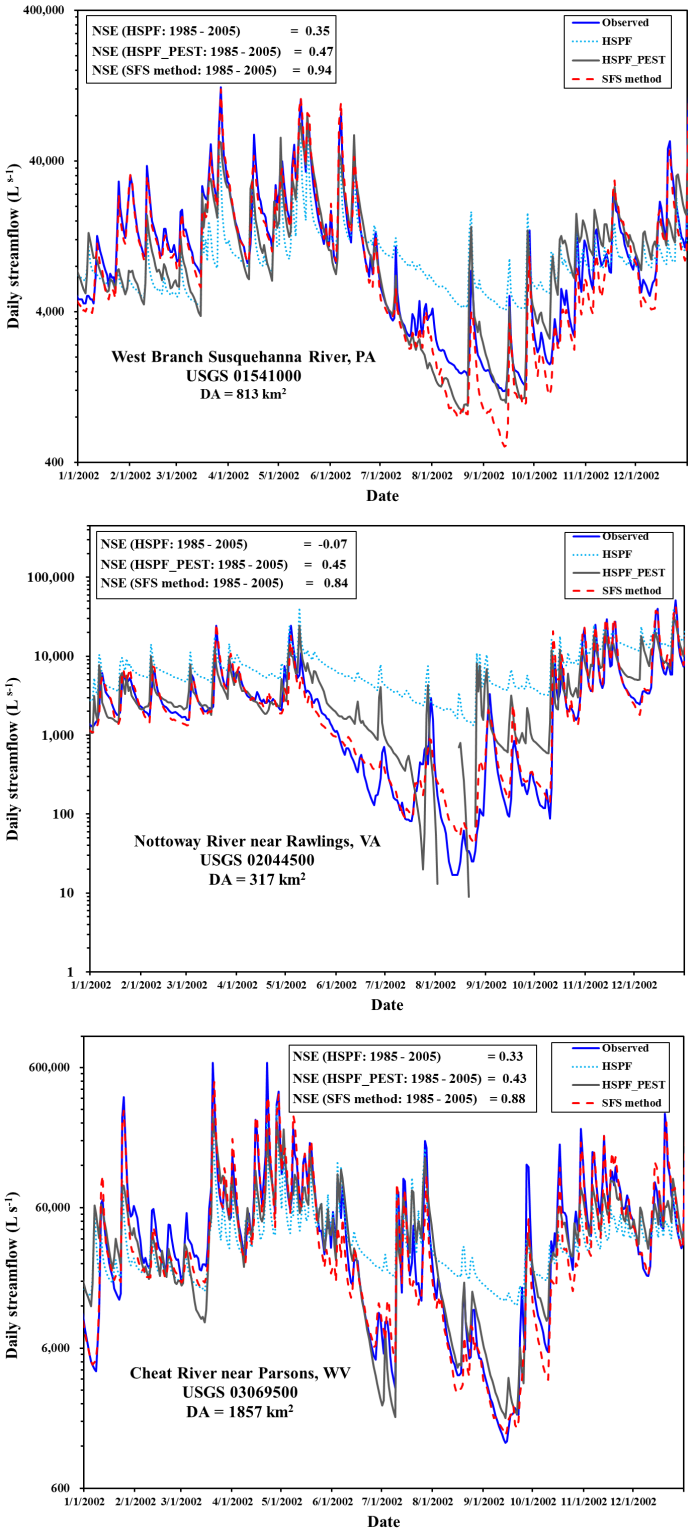

Figure 5.6: Observed and predicted daily streamflow for sample watersheds from each province. The corresponding predicted flow duration curves (magnitudes) are given in Figure 5.3. The above figures display daily streamflow for 1/1/2002 to 12/31/2002, however, the displayed accuracies (NSE) cover the period 1/1/1985 to 12/31/2005.

Table 5.1: Topographic and land cover descriptors

Variable	Units	Description
DA	km^2	drainage area
EMIN	*m*	minimum elevation
ESTD	*m*	standard deviation of elevation
RLF	*m*	watershed relief
SMAX	*m/km*	maximum slope
SSTD	*m/km*	standard deviation of slope
CPLAN	–	plan curvature; rate of change of aspect along a contour
WI	–	topographic wetness index
MRVBF	–	multi resolution index of valley bottom flatness
MRRTF	–	multi resolution index of ridge top flatness
TSL	*km*	total stream length
MCS	*m/km*	main channel slope
AMEAN	*deg*	average aspect
HPC90	*m*	hypsometric curve elevation corresponding to relative watershed area of 0.9
BW	*km*	watershed width
DD	km/km^2	drainage density; $DD = TSL \div DA$
Barren	%	barren areas of bedrock and unconsolidated shores
Urban	%	developed areas of low, medium, and high intensity

Table 5.2: Soil, physical, and climatic descriptors

Variable	Units	Description
WTD	*cm*	water table depth
rockDep	*cm*	depth to bedrock
hsgB	%	hydrological soil group B
Sdepth	*cm*	soil depth
KFFACT	–	soil erodibility factor with rocks
Silt	%	silt
OM	%	percent organic matter
PERM	cm/hr	permeability from STATSGO data
KSAT	cm/hr	estimated hydraulic conductivity using pedotransfer function (Saxton et al., 1986)
Porosity	–	porosity
PSDI[a]	–	pore size distribution index
MSCL[a]	*cm*	macroscopic capillary length
T	cm^2/hr	transmissivity
STORG	*cm*	storage; $STORG = void \times Sdepth$
mP	*mm*	monthly precipitation (JANP; average January precipitation)
mET	*mm*	monthly evapotranspiration
MMET	*mm*	mean monthly potential evapotranspiration
ADI	*mm*	annual dryness index ; $ADI = \frac{MAP}{MAET}$
RFx	*mm*	Rainfall amount equaled or exceeded x % of the record time

Table 5.3: Regional equations for Appalachian Plateau

Equation	R^2
$Q_1 = 1.79 \times 10^{3.66} DA^{0.98} DecET^{-1.05} SMAX^{0.026} + 0.13$	0.86
$Q_2 = 0.81 \times 10^{0.53} DA^{1.18} ESTD^{0.66} SMAX^{0.066} + 0.16$	0.83
$Q_3 = 1.49 \times 10^{2.49} DA^{0.88} SAT^{0.35} MAYP^{0.12} + 0.16$	0.92
$Q_4 = 0.96 \times 10^{1.09} DA^{0.99} SMAX^{0.034} Silt^{0.685} + 0.13$	0.96
$Q_5 = 1.48 \times 10^{2.02} DA^{0.93} Urban^{-0.04} RF0.5^{0.668} + 0.34$	0.96
$Q_6 = 1.36 \times 10^{1.89} DA^{0.99} Urban^{-0.088} SSTD^{-0.073} + 0.21$	0. 97
$Q_7 = 0.56 \times 10^{0.42} DA^{1.04} SMAX^{-0.02} OCTP^{0.712} + 0.12$	0.95
$Q_8 = 0.66 \times 10^{0.35} DA^{1.02} OCTP^{0.706} RF20^{0.272} + 0.19$	0.93
$Q_9 = 0.74 \times 10^{0.62} DA^{1.20} SSTD^{-0.05} KFFACT^{-0.423} + 0.16$	0.91
$Q_{10} = 0.55 \times 10^{0.37} DA^{0.98} SMAX^{-0.047} OCTP^{0.629} + 0.12$	0.91
$Q_{11} = 0.45 \times 10^{0.21} DA^{0.98} OCTP^{1.0} JunET^{-0.387} + 0.14$	0.89
$Q_{12} = 0.76 \times 10^{0.74} DA^{1.16} SSTD^{-0.085} SMAX^{-0.033} + 0.11$	0.83
$Q_{13} = 0.73 \times 10^{0.55} DA^{1.15} ADI^{0.70} SSTD^{-0.167} + 0.19$	0.83
$Q_{14} = 0.74 \times 10^{0.64} DA^{1.01} SSTD^{-0.265} ADI^{2.15} + 0.16$	0.86
$Q_{15} = 0.74 \times 10^{0.67} DA^{1.09} MRVBF^{0.709} SMAX^{-0.112} + 0.14$	0.87
$Q_{16} = 0.96 \times 10^{0.60} DA^{1.03} MRVBF^{0.619} MSCL^{-0.323} + 0.33$	0.78
$Q_{17} = 0.65 \times 10^{0.35} DA^{1.17} BW^{-0.415} MRVBF^{0.788} + 0.24$	0.72
$Q_{18} = 0.57 \times 10^{0.30} DA^{1.22} MRVBF^{0.843} BW^{-0.476} + 0.22$	0.70
$Q_{19} = 0.61 \times 10^{0.34} DA^{1.09} MRVBF^{0.891} RLF^{-0.166} + 0.19$	0.63

Table 5.4: Regional equations for Piedmont

Equation	R^2
$Q_1 = 26.2DA\left(24.61DD + 9.34RF5^2 - \frac{0.65}{SR}\right)$	0.76
$Q_2 = \frac{25.04DA}{1+Porosity}\left(25.15 - 2.72\frac{MarET}{MSCL}\right)$	0.86
$Q_3 = \frac{147.65DA}{1+MarET}\left(96.70 - 64.73\frac{STORG}{MSCL}\right)$	0.91
$Q_4 = \frac{69.09DA}{1+MarET}\left(76.48 - 0.923\frac{WI}{Urban}\right)$	0.92
$Q_5 = 1.103DA\left(0.03EMIN + 0.347WI^2 - \frac{0.998}{RF0.05}\right)$	0.89
$Q_6 = \frac{87.37DA}{1+DECP}\left(83.8 - 31.35\frac{rockDep}{FEBP}\right)$	0. 93
$Q_7 = 4.59DA(3.72 + 0.31EMIN)$	0.85
$Q_8 = 1.1DA\left(0.12JANP + 1.48OM^2 + \frac{1}{MarET}\right)$	0.90
$Q_9 = 1.05DA\left(0.051AMEAN + 1.13OM^2 - \frac{0.098}{CI}\right)$	0.94
$Q_{10} = 1.14DA\left(0.85MRRTF + 0.168DD^2 - \frac{1.614}{MRRTF}\right)$	0.94
$Q_{11} = 3.27 \times 10^{-1.87}DA^{0.90}HPC90^{0.23}APRP^{0.94} - 1.9$	0.94
$Q_{12} = 3.45 \times 10^{-1.57}DA^{0.92}APRP^{0.985}MRRTF^{-0.44} + 2$	0.96
$Q_{13} = 0.617 \times 10^{-1.20}DA^{0.86}MRRTF^{-0.47}APRP^{1.203} + 1.1$	0.93
$Q_{14} = 8.98 \times 10^{-14.2}DA^{0.734}MRRTF^{-1.78}WI^{13.5} + 3.1$	0.88
$Q_{15} = 3.14 \times 10^{3.88}DA^{0.71}T^{-2.34}KSAT^{1.71} + 8.2$	0.65
$Q_{16} = 3.40 \times 10^{3.72}DA^{0.65}KSAT^{1.39}T^{-2.21} + 3.4$	0.52
$Q_{17} = 0.26 \times 10^{5.56}DA^{-0.98}RF20^{4.57}LDP^{2.61} - 10$	0.57
$Q_{18} = 1526 \times 10^{73.4}DA^{72.2}T^{-73}TSL^{-60} + 218.5$	0.66
$Q_{19} = 136.7 \times 10^{-49.4}DA^{96.2}BW^{-153.9}T^{-32.5} + 79$	0.70

Table 5.5: Regional equations for Ridge and Valley

Equation	R^2
$Q_1 = \frac{15.62DA}{1+MRVBF}\left(-3.08 + 15.84\frac{NOVP}{MSCL}\right)$	0.82
$Q_2 = \frac{12.89DA}{1+MRVBF}\left(10.53 + 7.53\frac{SEPP}{MSCL}\right)$	0.88
$Q_3 = \frac{34.81DA}{1+MRVBF}\left(19.98 - 30.97\frac{MSCL}{NOVP}\right)$	0.92
$Q_4 = \frac{53.30DA}{1+MSCL}\left(48.43 + 41.11\frac{RF20}{MRVBF}\right)$	0.89
$Q_5 = \frac{56.44DA}{1+MSCL}\left(57.13 - 1.49\frac{hsgB}{MCS}\right)$	0.88
$Q_6 = \frac{20.87DA}{1+PERM}\left(20.90 + 0.07\frac{KFFACT}{CPLAN}\right)$	0. 96
$Q_7 = \frac{5.94DA}{1+PSDI}\left(6.03 + 1.43\frac{Rock}{KFFACT}\right)$	0.96
$Q_8 = \frac{33.08DA}{1+MMET}\left(29.27 + 5.32\frac{Rock}{PSDI}\right)$	0.96
$Q_9 = \frac{16.91DA}{1+WTD}\left(16.94 - 0.06\frac{PERM}{RF20}\right)$	0.84
$Q_{10} = \frac{1.29DA}{1+MMET}\left(1.0 - 1.29\frac{JUNP}{PSDI20}\right)$	0.95
$Q_{11} = \frac{1.11DA}{1+MMET}\left(1.0 + 1.11\frac{JUNP}{PSDI}\right)$	0.95
$Q_{12} = \frac{9.75DA}{1+WTD}\left(-6.0 + 7.75\frac{JUNP}{AprET}\right)$	0.92
$Q_{13} = \frac{5.74DA}{1+WTD}\left(-4.85 + 3.10\frac{JUNP}{MSCL}\right)$	0.94
$Q_{14} = 2.13 \times 10^{-3} DA^{1.3}(NOVP^{1.8}WTD^{-1.03} + MSCL^{-5.68}PERM^{0.52})$	0.88
$Q_{15} = 1.78 \times 10^{-5} DA^{1.1}(DECP^{0.97}WI^{2.82} + PSDI^{79.26}NOVP^{-100.35})$	0.75
$Q_{16} = 6.2 \times 10^{-5} DA^{0.93}(WI^{6.57}MSCL^{-1.85} + Sdepth^{0.87}Porosity^{-10.6})$	0.70
$Q_{17} = 3.0 \times 10^{-7} DA^{0.79}(Porosity^{1.57}WI^{7.29} + AMEAN^{-11.5}JANP^{16.47})$	0.53
$Q_{18} = 0.14DA^{1.67}(AMEAN^{-1.21}Sdepth^{1.32} + OM^{11.35}JANP^{-3.52})$	0.46
$Q_{19} = 34.66DA^{0.81}(SEPP^{0.19}Porosity^{5.21} + JANP^{11.89}AMEAN^{-11.16})$	0.41

Table 5.6: Prediction of sequence for Susquehanna (USGS 01541000; Appalachian)

ID	Source of sequence	R^2	NSE	MAE	$RMSE$
1	SUSQUEHANNA (True sequence)	0.98	0.98	1236	3182
2	CLEARFIELD (19.8 km)	0.88	0.88	3639	8251
3	LITTLE MAHONING (21.8 km)	0.81	0.81	4502	10395
4	BLACKLICK (29.1 km)	0.82	0.82	4225	10065
5	REDBANK (48.4 km)	0.76	0.76	5207	11692
6	Arithmetic mean	0.93	0.92	2895	6500
7	Minimum	0.83	0.83	4377	9737
8	Maximum	0.88	0.88	3617	8169
9	Geometric mean	0.93	0.93	2616	6279
10	Harmonic mean	0.89	0.89	3394	7930
11	Distance weighted mean	0.92	0.92	3020	6753
12	Area weighted mean	0.91	0.91	3216	7191
13	Quadratic mean	0.90	0.90	3335	7431
14	similarity of FDCs	0.88	0.88	3708	8312
15	Monte-Carlo sampling	0.93	0.93	2801	6156
16	Bagging	0.94	0.94	2672	5818
17	Boosting	0.94	0.94	2574	5894

- R^2 is the coefficient of determination ($0 \leq R^2 \leq 1.0$) and NSE is the Nash–Suticliffe coefficient of efficiency ($-\infty \leq R^2 \leq 1.0$).
- MAE and $RMSE$ are the mean absolute error and root mean square error in liters per second. Smaller values indicate high prediction.
- Results depict performance of closest donor watersheds and different ensemble methods for sequence improvement (refer to section 5.2)

Table 5.7: Prediction of sequence for Nottoway River (USGS 02044500; Piedmont)

ID	Source of sequence	R^2	NSE	MAE	$RMSE$
1	NOTTOWAY (True sequence)	0.94	0.93	941	5212
2	DEEP CREEK (20.7 km)	0.77	0.77	2406	9298
3	MEHERRIN (23.9 km)	0.66	0.66	2421	11341
4	STONY (30.7 km)	0.69	0.69	2748	10859
5	APPOMATTOX (54.5 km)	0.50	0.48	3954	14081
6	Arithmetic mean	0.80	0.80	1989	8728
7	Minimum	0.79	0.78	2353	9083
8	Maximum	0.68	0.68	2398	11018
9	Geometric mean	0.84	0.84	1811	7915
10	Harmonic mean	0.82	0.82	2034	8304
11	Distance weighted mean	0.82	0.81	1932	8420
12	Area weighted mean	0.71	0.70	2296	10611
13	Quadratic mean	0.75	0.75	2209	9785
14	similarity of FDCs	0.73	0.73	2407	10187
15	Monte-Carlo sampling	0.83	0.82	1931	8166
16	Bagging	0.83	0.83	1878	8031
17	Boosting	0.84	0.83	1830	7993

Table 5.8: Prediction of sequence for Cheat River (USGS 03069500; Ridge and Valley)

ID	Source of sequence	R^2	NSE	MAE	$RMSE$
1	CHEAT (True sequence)	0.99	0.99	4132	7559
2	TYGART (29.9 km)	0.78	0.76	14777	35240
3	YOUGHIOGHENY (53.6 km)	0.77	0.76	16973	35779
4	GREENBRIER (61.4 km)	0.71	0.69	18572	40587
5	CEDAR CREEK (61.6 km)	0.37	0.22	31653	64103
6	Arithmetic mean	0.86	0.85	12491	28007
7	Minimum	0.80	0.79	16134	33591
8	Maximum	0.79	0.77	14612	34576
9	Geometric mean	0.87	0.87	12339	26435
10	Harmonic mean	0.85	0.84	13706	29043
11	Distance weighted mean	0.85	0.84	12628	28806
12	Area weighted mean	0.81	0.80	13857	32596
13	Quadratic mean	0.82	0.81	13462	31649
14	similarity of FDCs	0.83	0.82	13974	31068
15	Monte-Carlo sampling	0.88	0.88	11774	25374
16	Bagging	0.88	0.88	11634	25234
17	Boosting	0.88	0.88	11752	25506

Table 5.9: Prediction performance on sample watersheds of Appalachian Plateau

USGS #	Name	Area[a]	Dist[b]	NSE^c_{fdc}	NSE^d_{near}	NSE^e_{boost}
03010500	Allegheny	1423	28.6	0.94	0.47	0.56
03042000	Blacklick	491	29.1	0.92	0.78	0.79
01549500	Blockhouse	98	29.7	0.96	0.52	0.75
03179000	Bluestone	1023	66.7	0.94	0.46	0.46
03049000	Buffalo	355	24.2	0.96	0.76	0.76
03078000	Casselman	164	35.1	0.96	0.71	0.84
01541500	Clearfield	961	19.8	0.97	0.85	0.85
03106000	Connoquene	922	24.2	0.94	0.75	0.75
01520000	Cowanesque	774	28.6	0.87	0.36	0.52
03187500	Cranberry	207	9.7	0.78	0.72	0.72
03011800	Kinzua	101	16.0	0.96	0.79	0.66
03080000	Laurel	312	37.3	0.9	0.69	0.76
03034500	Little Mahoning	222	21.8	0.97	0.81	0.81
03114500	Middle Island	1186	108.2	0.95	0.24	0.30
01595000	N B Potomac	235	45.8	0.97	0.66	0.66
01548500	Pine Creek	1558	28.6	0.97	0.48	0.67
03032500	Redbank	1375	34.2	0.98	0.7	0.76
01543500	Sinnemahoning	1779	38.9	0.98	0.7	0.67
01541000	Susquehanna	813	19.8	0.98	0.88	0.94
01518000	Tioga	723	29.7	0.96	0.56	0.61
03028000	W B Clarion	164	16.0	0.94	0.77	0.78
03186500	Williams	330	9.7	0.79	0.73	0.73
01545600	Young	119	29.1	0.96	0.85	0.85

[a] Watershed drainage area (km^2) derived from BASINS 4.0
[b] Distance (km) between ungauged and nearest gauged watershed
[c] The NSE of predicted flow duration curve (FDC)
[d] The NSE of predicted daily streamflow using nearest gauged watershed
[e] The NSE of predicted daily streamflow after sequence improvement

Table 5.10: Prediction performance on sample watersheds of Piedmont

USGS #	Name	Area[a]	Dist[b]	NSE^{c}_{fdc}	NSE^{d}_{near}	NSE^{e}_{boost}
01574000	West Conewago	512	25.1	0.91	0.63	0.70
01580000	Deer Creek	94.6	10.6	0.74	0.59	0.52
01583500	Western Rn	60.5	17.8	0.87	0.78	0.78
01591000	Patuxent	34.8	13.8	0.95	0.70	0.83
01639000	Monocacy	173	24.7	0.85	0.64	0.68
01639500	Big Pipe	103	24.7	0.87	0.65	0.73
01643500	Bennett	63.2	11.8	0.95	0.76	0.79
01645000	Seneca	102	11.8	0.92	0.73	0.80
01664000	Rappahannock	620	39.5	0.95	0.71	0.64
01667500	Rapidan	467	39.5	0.97	0.71	0.64
02030500	Slate	226	30.4	0.85	0.51	0.56
02039500	Appomattox	303	30.4	0.94	0.58	0.65
02041000	Deep Creek	205	20.7	0.95	0.77	0.80
02044500	Nottoway	317	20.7	0.93	0.77	0.83
02046000	Stony	112	26.3	0.96	0.73	0.73
02051500	Meherrin	552	23.9	0.91	0.69	0.69
02058400	Pigg River	343	28.3	0.93	0.60	0.73
02064000	Falling	165	34.4	0.88	0.40	0.62
02070000	North Mayo	104	37.9	0.98	0.62	0.64

[a] Watershed drainage area (km^2) derived from BASINS 4.0
[b] Distance (km) between ungauged and nearest gauged watershed
[c] The NSE of predicted flow duration curve (FDC)
[d] The NSE of predicted daily streamflow using nearest gauged watershed
[e] The NSE of predicted daily streamflow after sequence improvement

Table 5.11: Prediction performance on sample watersheds of Ridge and Valley

USGS #	Name	Area[a]	Dist[b]	NSE^{c}_{fdc}	NSE^{d}_{near}	NSE^{e}_{boost}
01564500	Aughwick	447	31.0	0.89	0.69	0.79
02020500	Calfpasture	366	21.0	0.87	0.72	0.72
03069500	Cheat	1857	29.9	0.99	0.76	0.76
01614500	Conococheague	1301	31.0	0.97	0.73	0.76
02016000	Cowpasture	1195	21.0	0.98	0.80	0.77
02018000	Craig	852	13.0	0.98	0.82	0.80
02013000	Dunlap	419	17.1	0.98	0.82	0.75
01560000	Dunning	444	29.0	0.93	0.82	0.83
01555500	East Mahantango	421	67.5	0.79	0.60	0.66
01539000	Fishing	707	35.6	0.93	0.62	0.62
01556000	Frankstown	748	28.1	0.95	0.83	0.90
03182500	Greenbrier	1364	38.5	0.98	0.60	0.77
01558000	Little Juniata	576	28.1	0.93	0.82	0.82
01555000	Penns Creek	793	61.3	0.92	0.54	0.63
02014000	Potts	397	13.0	0.98	0.83	0.91
02055000	Roanoke	995	30.0	0.87	0.59	0.61
01568000	Sherman	535	52.8	0.95	0.67	0.77
03051000	Tygart	1075	29.9	0.98	0.77	0.76
03173000	Walker	773	21.4	0.88	0.66	0.66
01538000	Wapwallopen	103	35.6	0.88	0.63	0.63
01601500	Wills	639	46.3	0.98	0.72	0.84
03175500	Wolf Creek	578	21.4	0.92	0.72	0.72
03075500	Youghiogheny	348	39.5	0.92	0.49	0.75

[a] Watershed drainage area (km^2) derived from BASINS 4.0
[b] Distance (km) between ungauged and nearest gauged watershed
[c] The NSE of predicted flow duration curve (FDC)
[d] The NSE of predicted daily streamflow using nearest gauged watershed
[e] The NSE of predicted daily streamflow after sequence improvement

CHAPTER 6

SUMMARY AND CONCLUSION

This study set out to examine the effect of using different variable selection methods on watershed hydrologic modeling. The first part of the study compared the accuracy, consistency, and predictive potential of variables selected by stepwise regression and principal component analysis to five causal variable selection methods (Grow–Shrink, interleaved Incremental Association Markov Blanket, Local Causal Discovery, First Order Utility, and HITON Markov Blanket). The causal variable selection methods seek to infer causal associations between explanatory and response variables by reconstructing a local Markov blanket of the response variable given explanatory variables. The second part evaluated the ability of the selected variables to classify the same hydrologically similar reference watersheds classified using streamflow indices in three Mid-Atlantic physiographic provinces of Appalachian Plateaus, Piedmont, and Ridge and Valley (watershed classification). This enabled quantification of which variable selection method better predicted agreement between physical and hydrological similarity of watersheds.

Based on accuracy, consistency, and predictive potential of causal variable methods, the third part predicted daily streamflow for ungauged watersheds by independently predicting streamflow magnitude using variables selected by HITON Markov Blanket method and independently predicting streamflow temporal sequence for the three Mid-Atlantic physiographic provinces. Several ensemble techniques were used to enhance prediction of streamflow temporal sequence using data from multiple neighboring gauged watersheds.

To examine the ability of the methods to select the true variables, data of two known functional relationships: weight of a hollow cylinder and pressure drop of a fluid in a circular pipe were used. To examine the ability of the methods to select the same variable, even when the data is slightly perturbed (method consistency); and predictive potential, data of known and unknown functional relationships were used. Data of unknown functional relationship consisted of 26 Mid–Atlantic Piedmont watersheds with 111 watershed descriptors and 19 streamflow percentiles. This data is considered to have an unknown functional relationship because, there is no universal agreement on the functional form and the true variables that drive the different streamflow percentiles are not known. The accuracy of some causal selection methods was greater than others. Overall, the HITON–MB and first order utility (FOU) methods were the most accurate followed by principal component analysis (PCA). The accuracy of the Grow–Shrink (GS) and a variant of the incremental association Markov boundary (interIAMBnPC) were not better than the accuracy of the stepwise regression. Because of the high accuracy of the HITON–MB and its high consistency on data of known and unknown functional relationship, variables selected by this method have a high probability of being causal compared to stepwise regression.

For the watershed classification objective, we evaluated the ability of watershed variables selected by different methods to classify the exact hydrologically similar watersheds selected by the streamflow indices for the Appalachian Plateaus, Piedmont, and Ridge and Valley using k–means clustering. A similarity index (SI) was used to compare classification results by streamflow indices and classification results by watershed variables. On average, 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). Only one method (HITON–MB) was able to identify variables that were unique to each ecoregion without compromising classification performance (SI=0.71 for Appalachian, SI=0.90 for Piedmont, and SI=0.72 for Ridge and Valley).

Regarding the third part, the streamflow magnitude was estimated using regionalized flow duration curves while sequence was estimated by transferring neighboring gauged watershed sequence to the ungauged watershed. This approach is referred to as streamflow separation technique. Also, the effects of drainage area and distance between gauged and ungauged watersheds on the accuracy of predicted daily streamflow were assessed. The relative drainage area of the donor watershed was not statistically significant during sequence prediction, however, the distance was. Other factors such as surface storage and level of urbanization had secondary effects. The geometric mean of streamflow of two closest donor watersheds gave better sequence prediction than arithmetic, harmonic, and quadratic means. Ensemble methods of bagging and boosting better improved sequence prediction than use of just the closest donor watershed given the distance between the first and second closest gauged watersheds was less than 20 km.

This study demonstrated that use of causal variable selection methods, especially HITON Markov Blanket has a high probability of selecting true variables that drive the different hydrologic regimes (High–flows, medium–flows, and low–flows). The study also showed that variables selected by causal methods, have a high predictive potential given a true functional form of response to explanatory variables. Use of more than one selection method to improve the reliability of the selected variables is recommended.

Future work on variable selection should focus on quantifying the probability that a selected variable for a specific response variable is causal based on selection accuracy of various methods on data of known functional relationships with varying system complexities. One recommendation is use of multiple variable selection methods from which a simple or a weighted rank may be used to determine probable causal variables. Daily streamflow prediction using the streamflow separation method showed high predictive power of the method using sequence of neighboring gauged watersheds, however most ungauged watersheds are not located in neighborhoods of gauged watersheds. Therefore, future work on sequence prediction should concentrate on use of only the rainfall data of the ungauged watershed or use rainfall data in combination with sequence of gauged watershed at a considerable distance from the ungauged watershed.

In conclusion, replication of similar performance of predicted daily streamflow by the streamflow separation method and use of the regionalized flow duration curves may be limited to watersheds with drainage areas ranging between 98–1,779 km^2 for the Appalachian Plateau, 34.8–620 km^2 for the Piedmont, and 48–1,857 km^2 for the Ridge and Valley. The approach also poorly predicted low flows.

APPENDIX A

AVERAGE MONTHLY VARIATION OF WATER BALANCE COMPONENTS

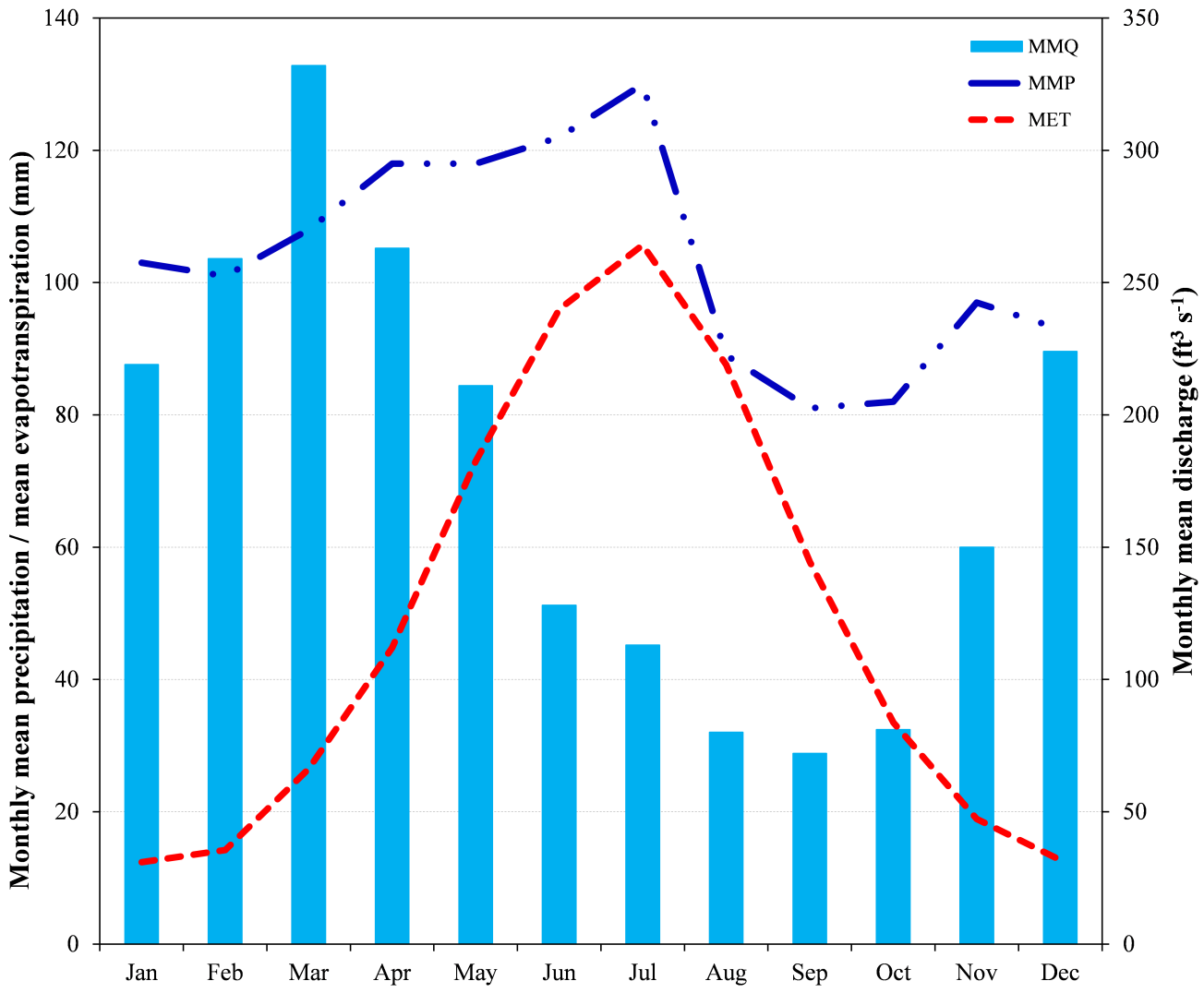

Figure A.1: Long–term water balance for Appalachian Plateau watersheds (1965–2007). The net monthly precipitation [Precipitation (MMP) — Evapotranspiration (MET)] is positive throughout the year and highest during the winter months. The highest monthly streamflow (MMQ) corresponds to months with highest snowmelt.

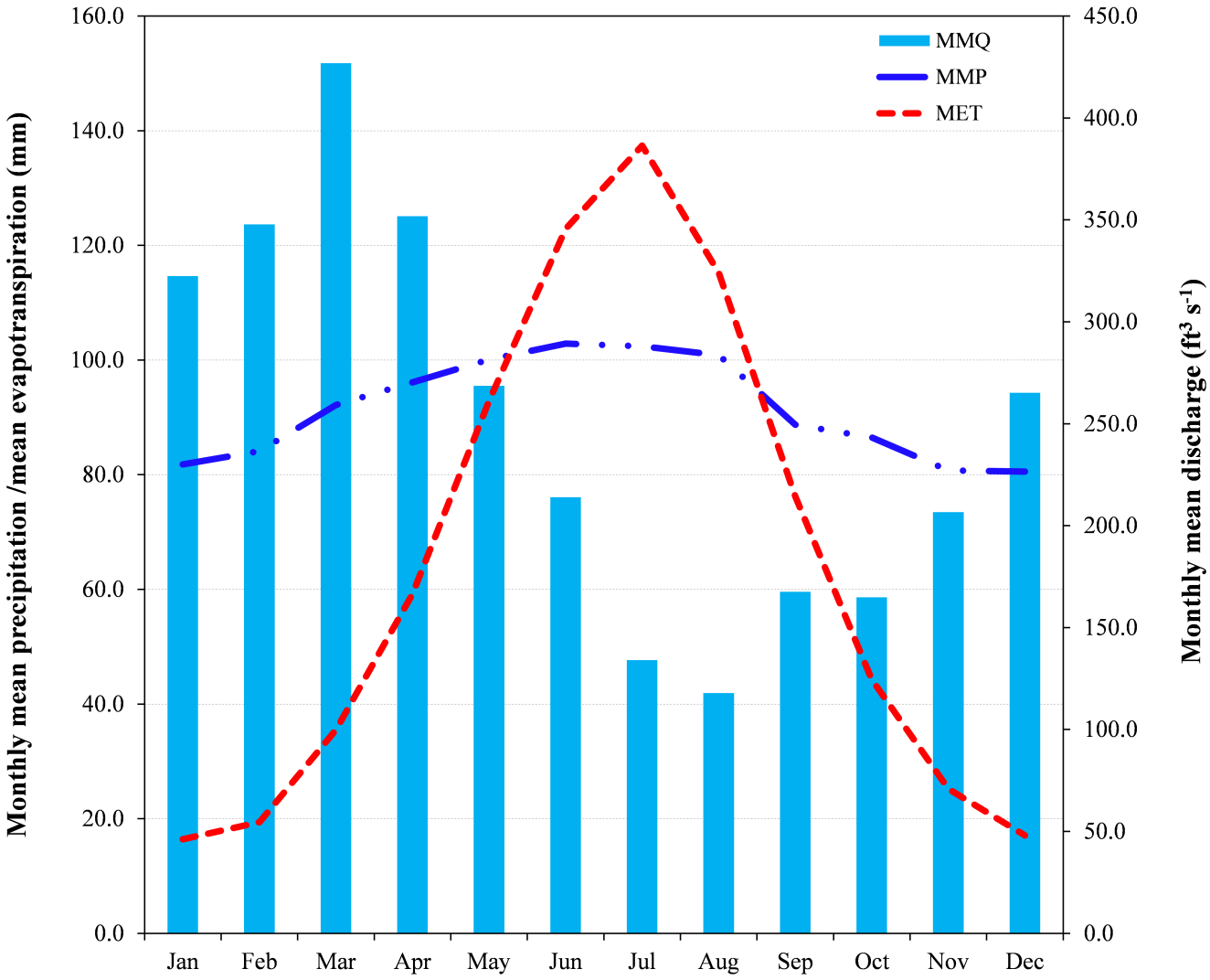

Figure A.2: Long–term water balance for Piedmont watersheds (1965–2007). The net monthly precipitation [Precipitation (MMP) — Evapotranspiration (MET)] is negative during the summer months due to high levels of evapotranspiration. The highest monthly streamflow (MMQ) behaviour is similar to that of Appalachian Plateaus although more in magnitude.

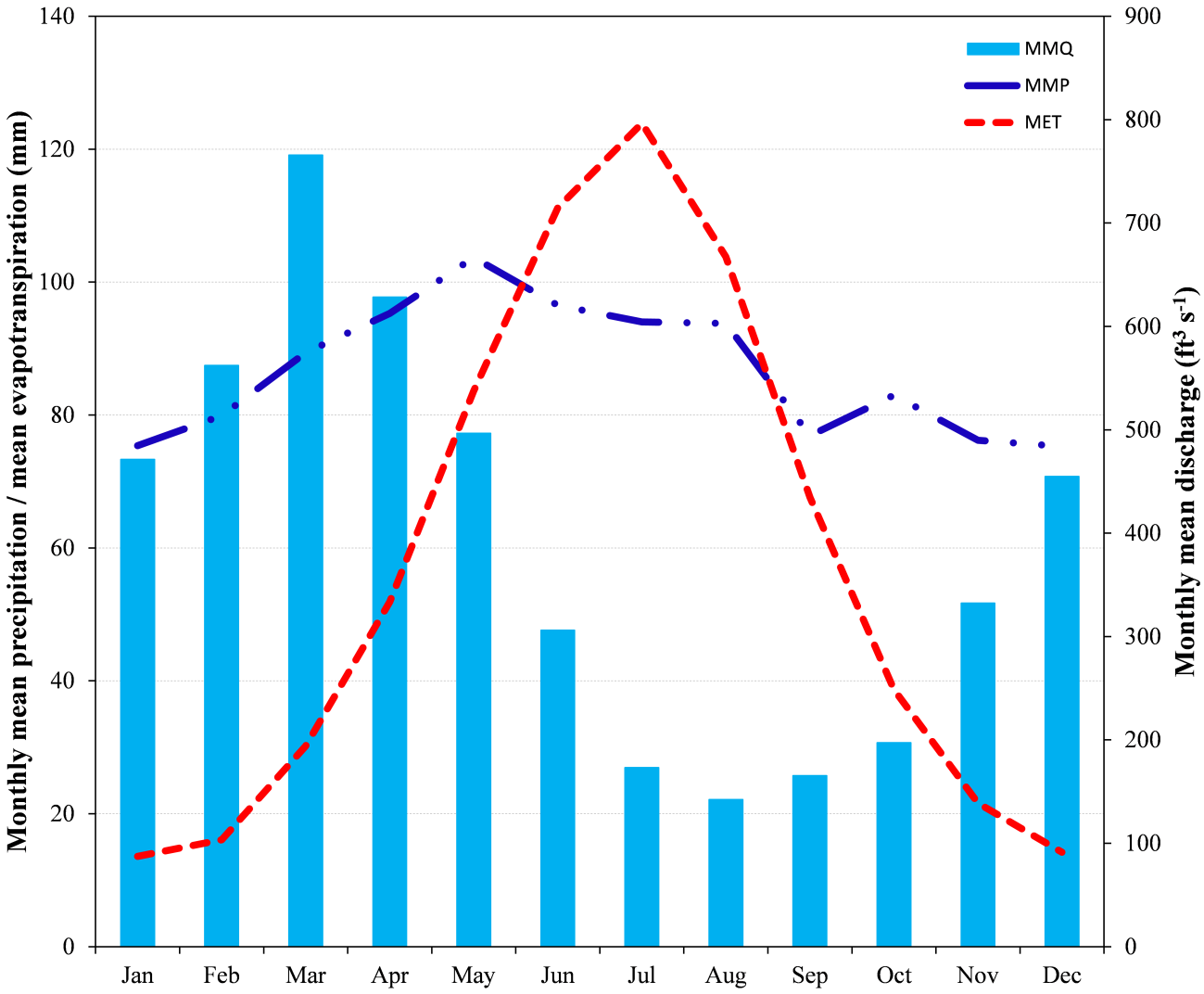

Figure A.3: Long–term water balance for ridge and Valley watersheds (1965–2007). The net monthly precipitation [Precipitation (MMP) — Evapotranspiration (MET)] and the streamflow behaviours are similar to those of the Piedmont, however, they are less in magnitude.

APPENDIX B

ANNUAL VARIATION OF PRECIPITATION, EVAPOTRANSPIRATION, AND MEDIUM STREAMFLOW WITH ELEVATION

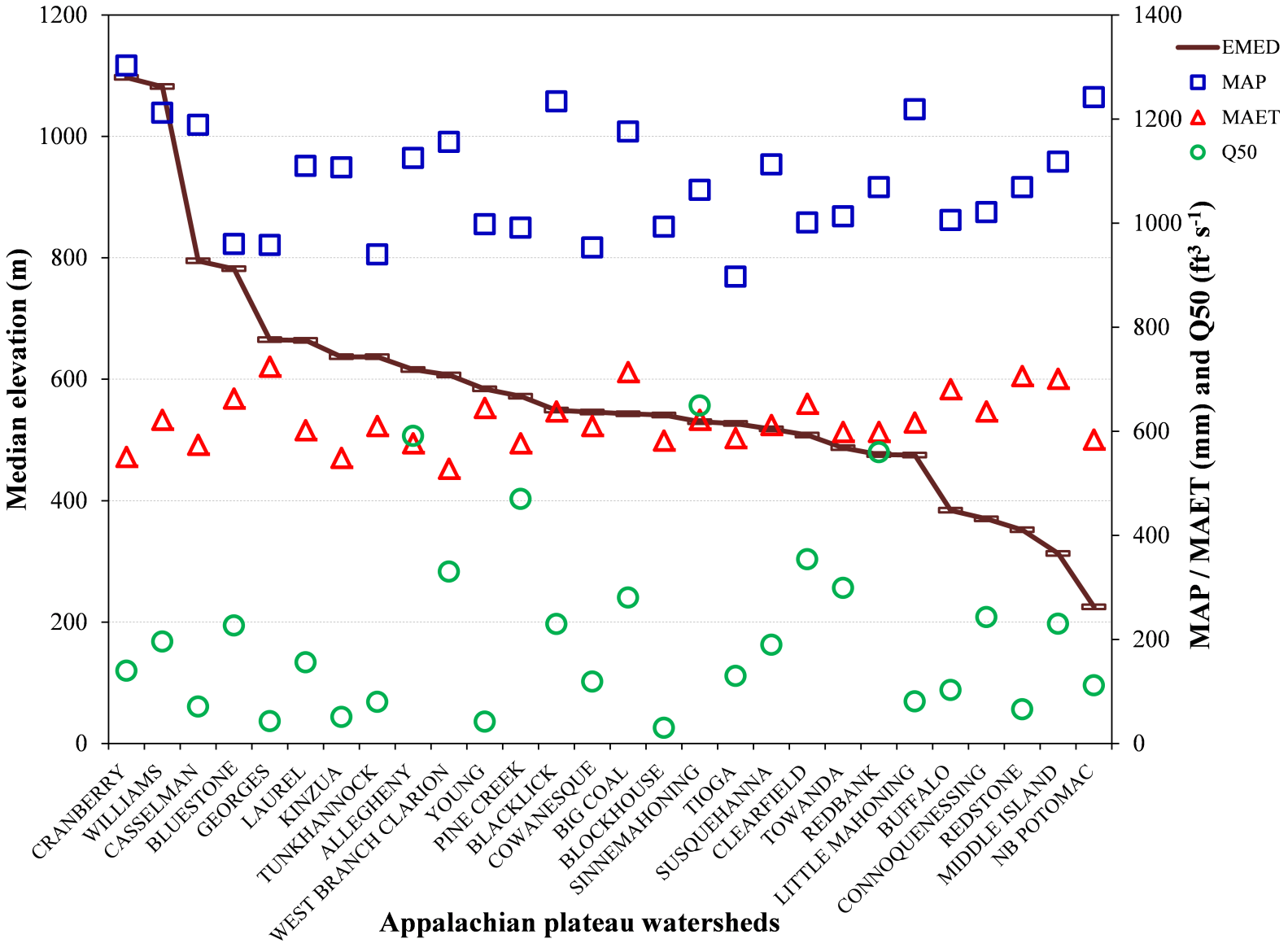

Figure B.1: Long–term (1965–2007) mean annual variation of precipitation (MAP), evapotranspiration(MAET), and medium streamflow (Q50) with median elevation (EMED) for Appalachian Plateau watersheds. There are no noticeable trends between each water balance component with decreasing elevation.

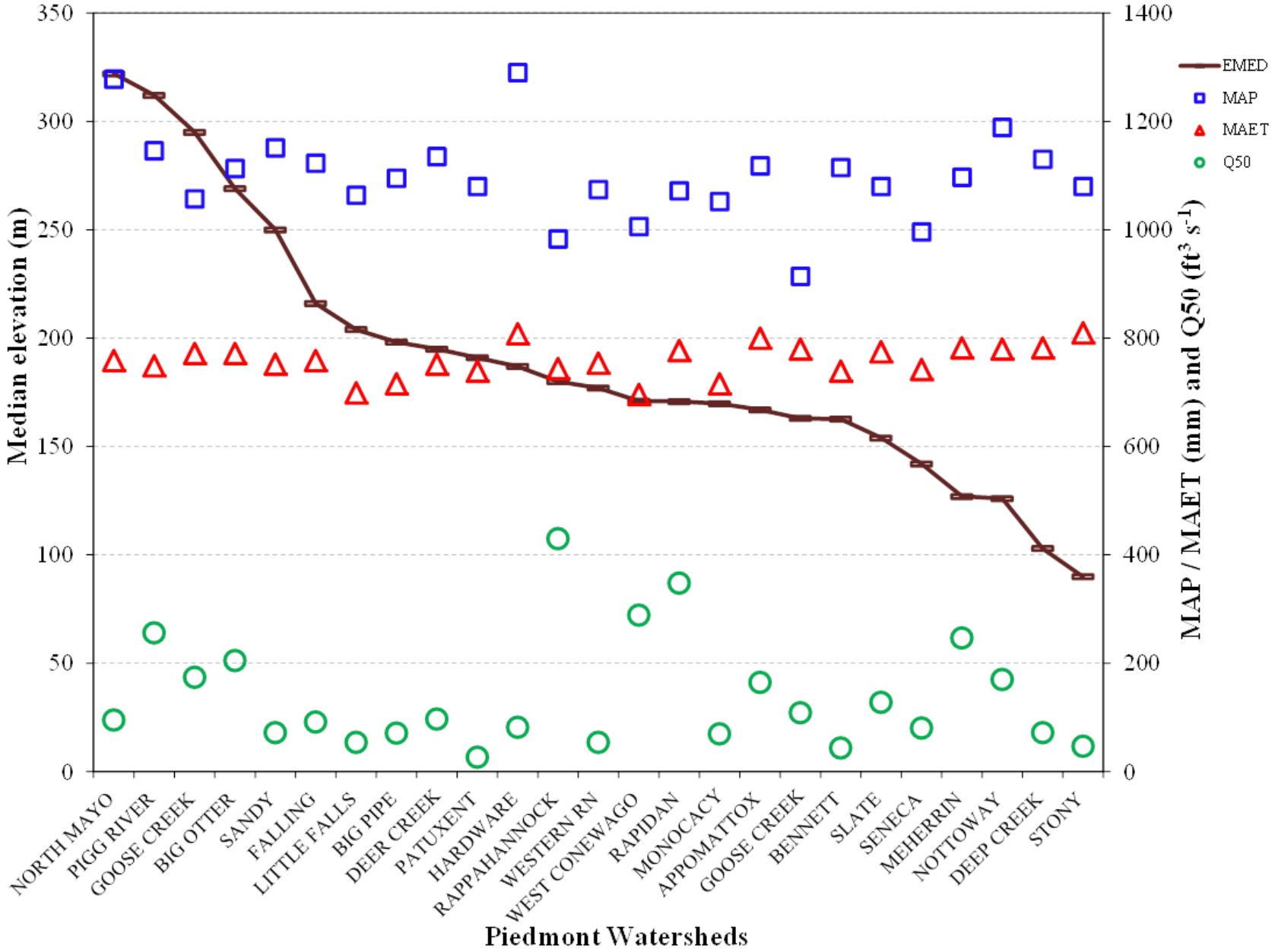

Figure B.2: Long–term (1965–2007) mean annual variation of precipitation (MAP), evapotranspiration(MAET), and medium streamflow (Q50) with median elevation (EMED) for Piedmont watersheds. There are no noticeable trends between each water balance component with decreasing elevation.

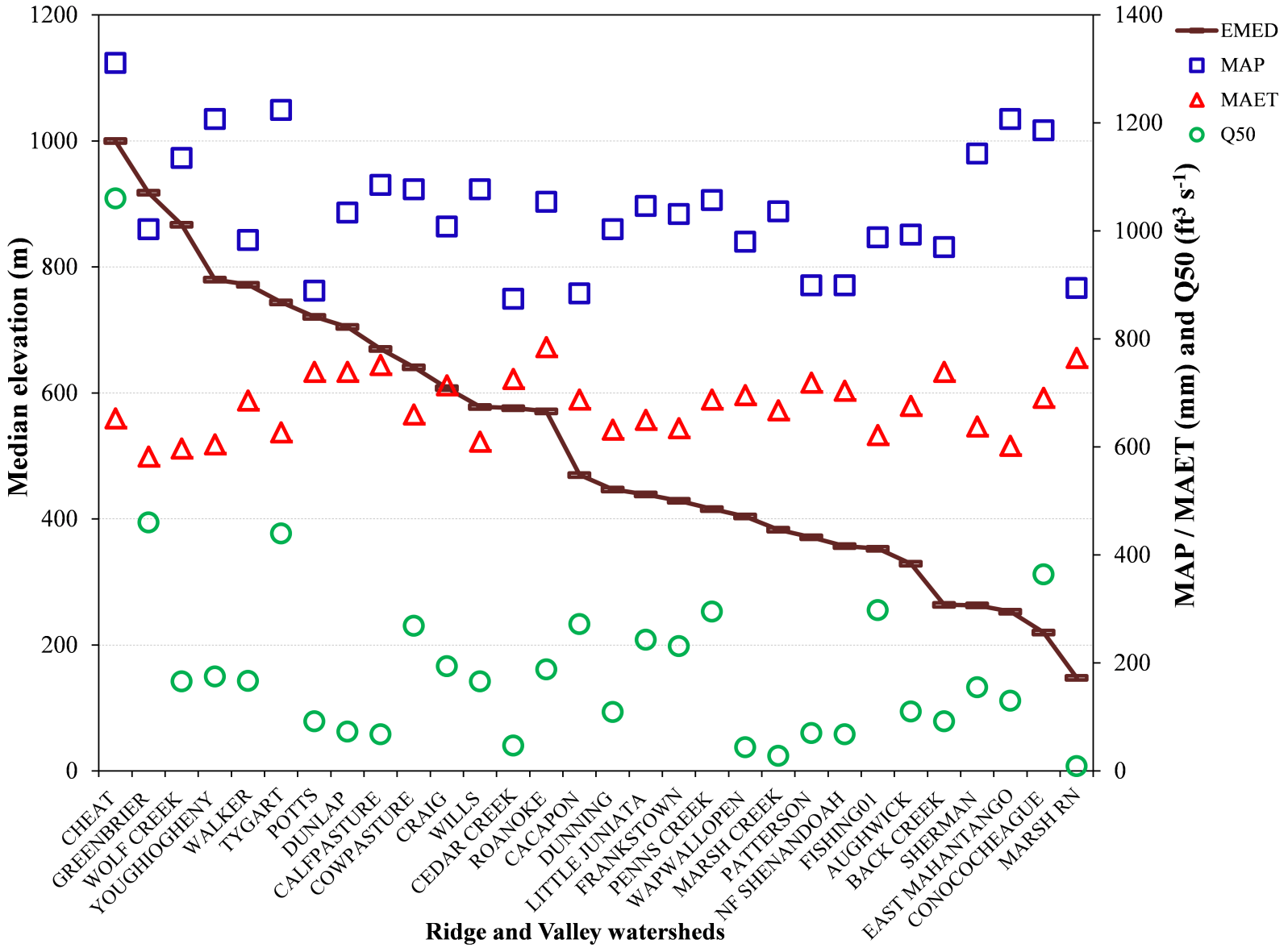

Figure B.3: Long–term (1965–2007) mean annual variation of precipitation (MAP), evapotranspiration(MAET), and medium streamflow (Q50) with median elevation (EMED) for Ridge and Valley watersheds. There are no noticeable trends between each water balance component with decreasing elevation.

APPENDIX C

MOST SELECTED VARIABLES FOR EACH PHYSIOGRAPHIC PROVINCE

Table C.1: Appalachian Plateaus: Percent proportion of most selected variable classes

Variable class†	GS	interIAMBnPC	LCD2	HITON–MB
		High flows		
Climatic	62.0	39.8	97.8	30.3
LULC	4.7	4.7	0.0	25.7
Soil	31.7	35.7	0.0	6.2
Topography	1.6	19.8	2.2	37.8
		Medium flows		
Climatic	98.0	82.4	100.0	39.4
LULC	0	0	0.0	19.8
Soil	0.0	10.4	0.0	10.6
Topography	2.0	7.1	0.0	30.2
		Low flows		
Climatic	16.1	5.1	75.4	15.0
LULC	3.7	0	0	5.0
Soil	56.9	86.3	0.0	34.2
Topography	23.3	8.5	24.6	45.8

† The percent proportions are aggregate results of the most selected variables after implementing each algorithm 26 times by deleting a single watershed with replacement on each run for high flows ($Q_{0.01} - Q_{10}$), medium flows ($Q_{20} - Q_{70}$), and low flows ($Q_{80} - Q_{99.9}$). Based on the accuracy of the HITON–MB compared to other methods (refer to chapter 3), one can suppose that the high and medium flows for the Applachian Plateaus are driven by topographic and climatic variables while the low flows are driven by topographic and soil descriptors.

Table C.2: Piedmont: Percent proportion of most selected variable classes

Variable class[‡]	GS	interIAMBnPC	LCD2	HITON–MB
		High flows		
Climatic	16.7	14.7	25	34.6
LULC	18.3	21.2	15.7	5.5
Soil	26.4	12	28.4	19
Topography	38.7	52.1	30.8	40.9
		Medium flows		
Climatic	10.6	36.9	4	19.5
LULC	0	0	6.4	11.7
Soil	59.5	22.5	51.9	20.8
Topography	29.8	40.5	37.8	48
		Low flows		
Climatic	17	21	34.8	15.4
LULC	0	0	0	23
Soil	40.1	30.1	2.1	37.6
Topography	42.9	48.9	63.1	24

[‡] The percent proportions are aggregate results of the most selected variables after implementing each algorithm 26 times by deleting a single watershed with replacement on each run for high flows ($Q_{0.01} - Q_{10}$), medium flows ($Q_{20} - Q_{70}$), and low flows ($Q_{80} - Q_{99.9}$). Based on the accuracy of the HITON–MB compared to other methods (refer to chapter 3), one can suppose that the high flows for the Piedmont are driven by topographic and climatic conditions, the medium flows by topographic and soil descriptors, while the low flows by soil descriptors and topography.

Table C.3: Ridge and Valley: Percent proportion of most selected variable classes

Variable class⊛	GS	interIAMBnPC	LCD2	HITON–MB
		High flows		
Climatic	46.5	29.1	46.9	14.0
LULC	6.0	14.2	19.5	4.0
Soil	25.5	17.2	5.7	40.6
Topography	22.0	39.5	27.9	41.4
		Medium flows		
Climatic	53.0	11.3	100.0	51.3
LULC	0.0	10.4	0.0	8.0
Soil	21.1	32.5	0.0	40.7
Topography	25.9	45.9	0.0	0.0
		Low flows		
Climatic	0.0	0.0	38.5	28.7
LULC	8.3	12.6	10.9	3.0
Soil	19.8	39.4	24.1	45.8
Topography	71.9	48.0	26.4	22.5

⊛ The percent proportions are aggregate results of the most selected variables after implementing each algorithm 26 times by deleting a single watershed with replacement on each run for high flows ($Q_{0.01} - Q_{10}$), medium flows ($Q_{20} - Q_{70}$), and low flows ($Q_{80} - Q_{99.9}$). Based on the accuracy of the HITON–MB compared to other methods (refer to chapter 3), one can suppose that the high flows for the Ridge and Valley are driven by topographic and soil descriptors, the medium flows by climatic and soil descriptors, while the low flows by soil and climatic descriptors.

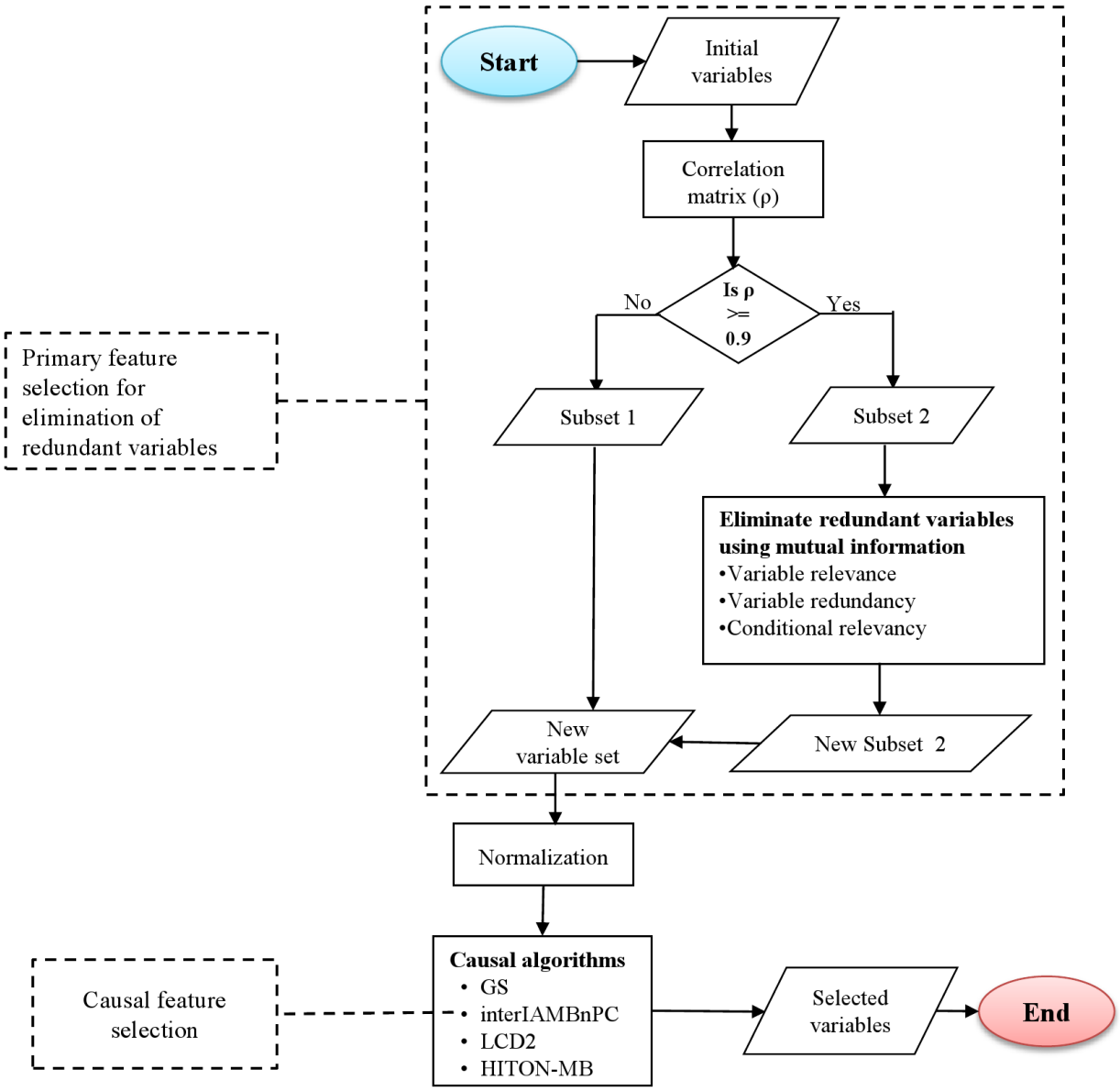

Figure C.1: The process chart depicts the variable selection process implemented in this study. The first step is the primary selection which seeks to minimize variable redudancy by selecting one of each pairs of highly correlated variables ($r \geq 0.9$). This is achieved by selecting the variable that provides the highest mutual information about the response variable from the pair. The second step normalizes all variables such that variable structure rather than the magnitude influences the selected variables. The normalization transforms variables from a scale of $[-\infty - +\infty]$ to [0 – 1]. The third and last step implements each algorithm onto the normalized variables.

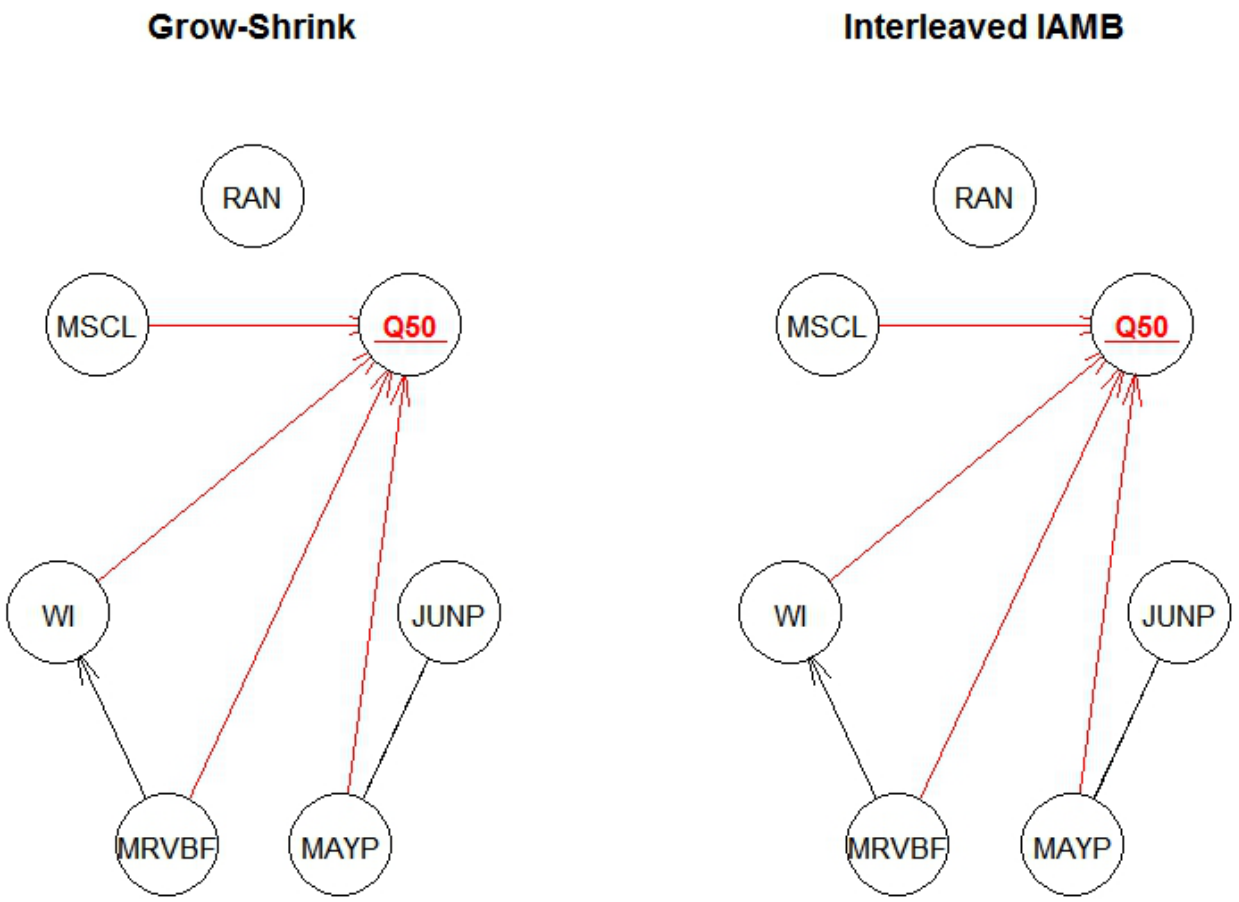

Figure C.2: Two Markov Blankets (MB) constructed by two causal algorithms. The causal selection methods select variables by reconstructing an MB around a response variable (e.g medium flow; Q50). The direction of the arrows indicate whether the variables are direct causes (pointing towards the response variable) or direct effects (pointing away from the response variable). The MB consists of direct causes (parents), direct effects (children), and direct causes of direct effects (parents of children). For this example, a random variable (RAN) is included in the data however, none of the algorithms selects it.

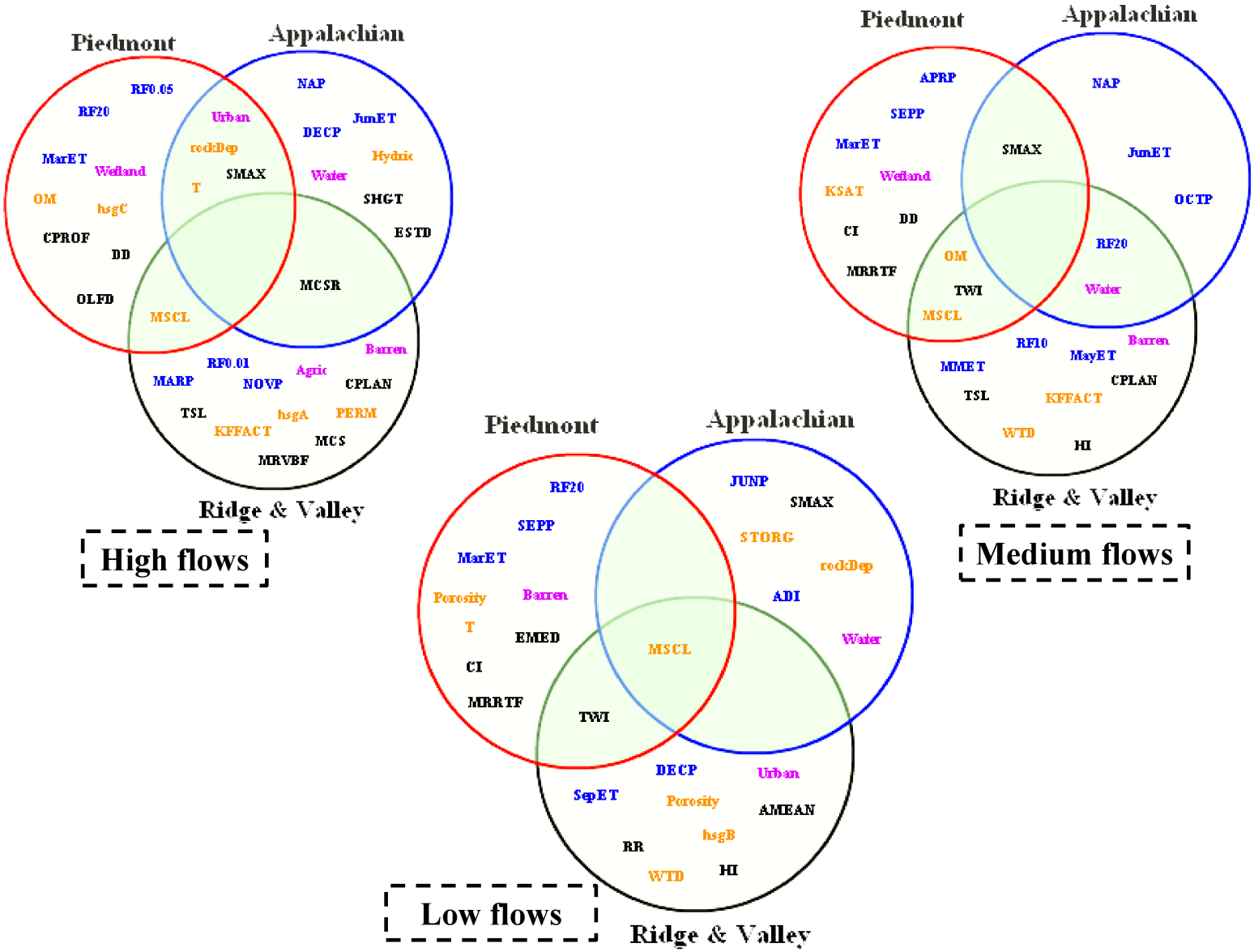

Figure C.3: The figure depicts the inter–connectivity of dominant variables across physiographic provinces for each flow regime. For example, for high and medium flows there is no single variable that s dominant across the three provinces. However, for low flows, the soil macroscopic capillary lengnth (MSCL) is dominant across the three provinces. The MSCL is a measure of the effect of capillarity on water movement and is associated with characteristic times for flows to reach steady state flow.

APPENDIX D

EFFECT OF SAMPLE SIZE ON STABILITY OF SELECTED VARIABLES

Overview and approach

Algorithm reliability also referred to as the similarity index (SI) in this study referred to the ability of an algorithm to select the same set of variables on subsequent runs when the initial sample data was slightly changed. Algorithm reliability provides confidence on the robustness and stability of both the algorithm and the selected variables. For example if an algorithm selects the same top three variables when a different watershed is removed on multiple runs, that algorithm is more reliable and provides a sense of confidence in the selected variables. Reliability of algorithms was estimated by computing the similarity index (SI) for variable subsets generated by the same algorithm after changing the sample data. The sample dataset was changed by excluding a single watershed with replacement from the original sample and implementing the variable selection procedure. The RI varies from zero meaning different variables are selected on subsequent runs to 1.0 meaning the exact variables are selected on subsquent runs. If the interest is the top five variables, an RI=0.6 implies that on average, exactly three of the five variables are selected on subsquent runs.

The effect of initial sample size on the final variables selected by the algorithms was assessed using the reliability index (RI). The process included: (1) aggregation of the top 5 variables for each flow percentile across all algorithms using data from the 26 runs. These variable sets for each flow were assumed to be the baseline variables; (2) a single watersheds was deleted from the original dataset with replacement and variable selection implemented on the remaining watersheds. Top 5 variables for each flow across all algorithms were summarized and a reliability index for each flow was determined using the baseline variables in step 1; (3) step 2 was repeated 10 times; (4) steps 2 and 3 were repeated by deleting 2 to 15 watersheds at a time such that the initial number of watersheds varied from 10 to 28; (5) average values (average of 10 runs for each sample size) of reliability index (RI) for each flow corresponding to different initial number of watersheds were computed; (6) the RI values were further aggregated based on the flow class such as average of $Q_{0.01} - Q_{10}$ for high flows, average of $Q_{20} - Q_{70}$ for medium flows, and average of $Q_{80} - Q_{99.9}$ for low flows with corresponding initial number of watersheds.

Therefore, the sample size of the watersheds has an effect on the final selected variables. Overall, of the three physiographic provinces, one needs more watersheds in the Ridge and Valley, then the Appalachian Plateaus, and least in the Piedmont to have the s ame level of reliability of the selected variables.

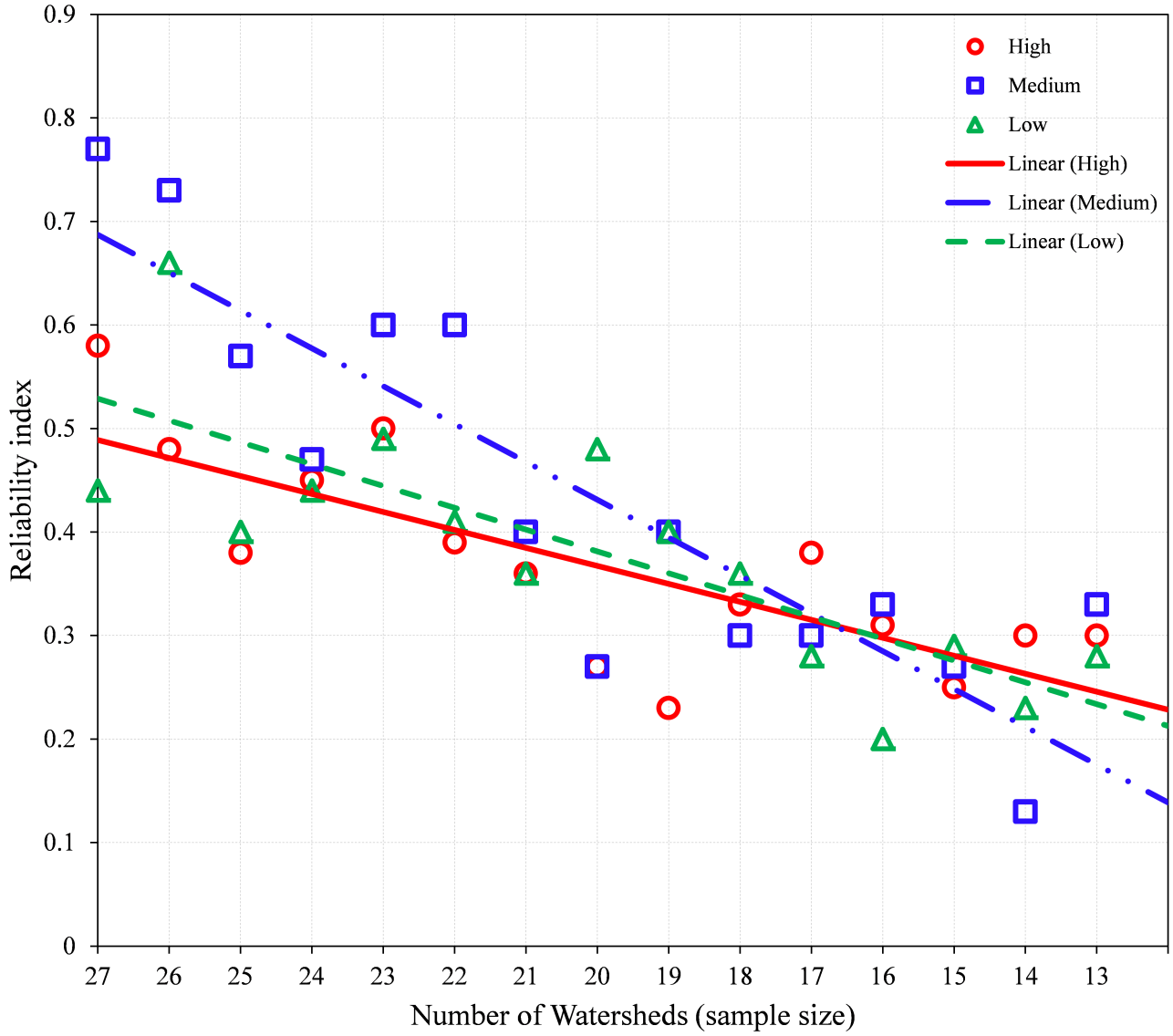

Figure D.1: Generally, the reliability index (RI) for Appalachian Plateau watersheds declined for high, medium, and low flows as the initial number of watersheds decreased. The rate of decline was more pronounced in the medium flows and less prevalent in the high flows. Thus the probability of algorithms selecting the same variable subset with less initial number of watersheds was high for medium flows and relatively low for low and high flows. For the appalachian Plateau, use of less than 25 watersheds decreased the reliability to less than 0.5 for high flows while use of less than 20 initial watersheds had the same effect on medium and low flows. Therefore for the Appalachian Plateaus, high flows are more sensitive to sample size compared to medium and low flows.

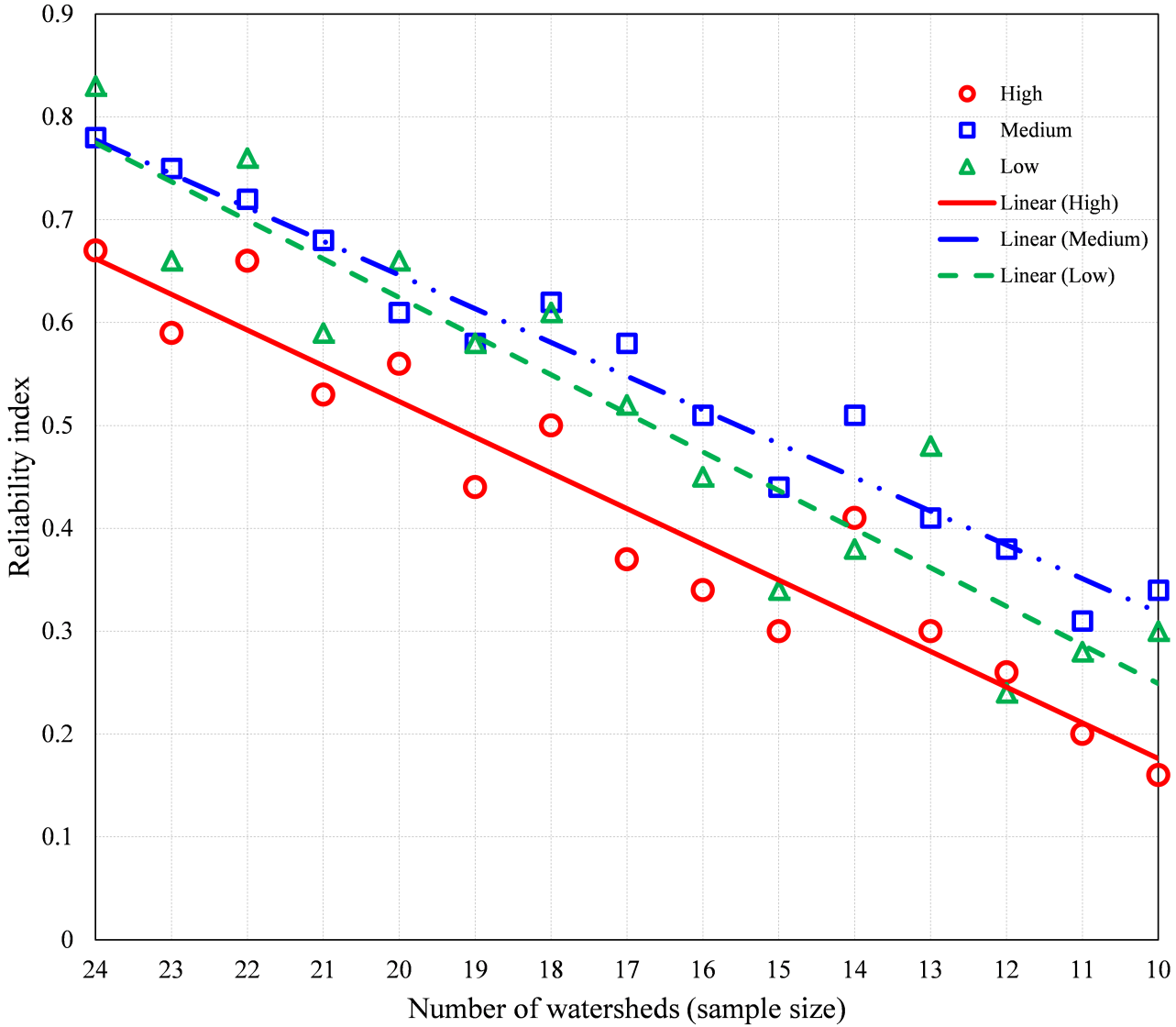

Figure D.2: Generally, the reliability index (RI) for Piedmont watersheds declined for high, medium, and low flows as the initial number of watersheds decreased. For the Piedmont, use of less than 28 watersheds decreased the reliability to less than 0.5 for high flows while use of less than 16 initial watersheds had the same effect on medium and low flows. Therefore for the Piedmont province, high flows are more sensitive to sample size compared to medium and low flows.

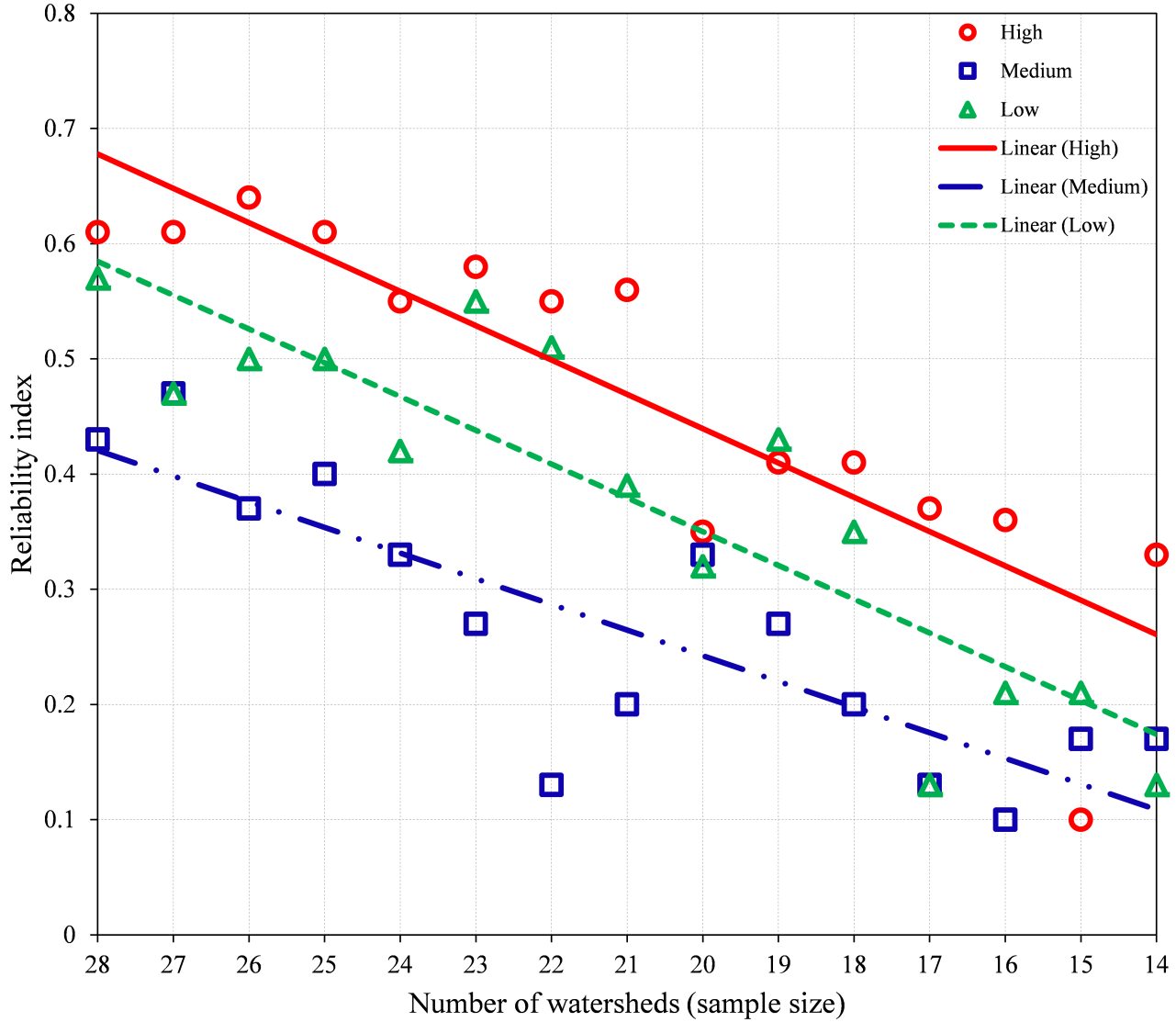

Figure D.3: Generally, the reliability index (RI) for Ridge and Valley watersheds declined for high, medium, and low flows as the initial number of watersheds decreased. The rate of decline was more pronounced in the medium flows and less prevalent in the high flows. For the Ridge and Valley, use of less than 28 watersheds decreased the reliability to less than 0.5 for medium flows while use of less than 25 and 23 initial watersheds had the same effect on low and high flows, respectively. Therefore for the Ridge and Valley province, medium flows are more sensitive to sample size compared to high and low flows.