2002

Geospatial and statistical foundations for streamflow synthesis in West Virginia

Annie J. Morris
West Virginia University

Geospatial and Statistical Foundations for Streamflow Synthesis in West Virginia

Annie J. Morris

Thesis submitted to the
Eberly College of Arts and Sciences at
West Virginia University in partial fulfillment
of the requirements for the degree of

Masters of Science
In
Geology

Joseph J. Donovan, chair
Michael Strager
Patricia Miller

Department of Geology and Geography

Morgantown, West Virginia
2002

Key Words: Streamflow Prediction, Principal Component Analysis

Abstract

Geospatial and Statistical Foundations for Streamflow Synthesis in West Virginia

Annie J. Morris

Streamflow values must be synthesized for locations where flow measurement stations, in applications such as the West Virginia SWAP program, are lacking or where only intermittent measurements are available (West Virginia Department of Health and Human Resources, 1999). This research describes an effort to improve upon the current synthetic streamflow model by incorporating geomorphic, geologic, and hydrogeologic measurements. Principal components analysis (PCA) was used to derive a set of master variables that characterize stream flow in West Virginia based on historical data from 29 watersheds. The relationships between variables affecting stream flow were also analyzed using cluster and correlation analysis to derive an optimum set of variables for predicting stream flow in the state. Based on this analysis, there are two categories of watersheds in West Virginia. The first is strongly correlated to climatic variables: precipitation, temperature, elevation, and groundwater recharge. The second is strongly correlated to two geomorphic variables; watershed slope, and percentage of forested area. The spatial distribution of the watershed groupings shows that watersheds dominated by the climatic component are located along the Allegheny Front while watersheds dominated by the geomorphic component are located in the Allegheny plateau and Valley and Ridge physiographic provinces.

Acknowledegments

I would like to thank my professors and colleagues that have helped me complete this research. My advisor Dr. Joe Donovan was an invaluable source of information, and ideas. Thank you for your support, patience, and sense of humor. Mike Strager generously gave me his time to act as a sounding board, technical advice about GIS, and access too much of the data used in this study. I would also like to thank Dr. Patricia Miller for her time and insight.

Thanks must also be extended to Dr. Steven Kite. Without your understanding, support, and insight, this would not have been possible.

Finally I would like to thank my family. I will never be able to adequately express my gratitude for your unconditional support of every endeavor I have ever attempted. I could not have succeeded without you.

Table of Contents

List of Tables

List of Figures

Introduction

Synthesis of stream flow rates is practiced in a number of applications. Many regulatory agencies require stream flow measurements or estimates to regulate pollutant discharges and to prepare for potential pollutant spills into streams. Infrequently, however, are locations of interest at U.S. Geological Survey gauging stations. Therefore, to generate modeled or "synthetic" estimates of flow at such data-poor locations, it is common practice to estimate flows from surrogate data. One such approach is to perform multivariate regression analysis on historical climatic and stream flow data to estimate flow. The current regression-based model for estimating stream flow used by the West Virginia Bureau of Public Health (WVBPH) in the Source Water Area Protection (SWAP) program is a model of the formula:

$$\text{Flow} = 1232 + 0.00304\ A - 23.6\ T_{max} + 0.338\ S_s \quad (1)$$

where A = catchment area (acres), T_{max} = maximum temperature (°F), and S_s = stream slope (Spatial Analytics, LLC, 2000).

A multivariate analysis was performed on 13 watersheds in West Virginia using combinations of these variables along with T_{min} (minimum temperature), P (mean annual precipitation), and S_w (watershed slope) (Spatial Analytics, LLC, 2000). The goal of the analysis was to determine the best predictors for stream discharge in West Virginia. Statewide, the variable set in equation (1) (A, T_{min}, and S_s) yielded the highest correlation of all possible combinations. A regression analysis was then run using these three variables within each stream basin in order to derive local values for regression parameters.

The cited analysis neglected several factors that may be relevant in West Virginia. Variables for groundwater recharge, geology soil, land use/vegetation characteristics, or elevation were not included in this study. This is despite the fact that elevation correlates strongly with both temperature and precipitation at many locations of the state. On average, temperature decreases 6.5°C for each kilometer increase in elevation, and the upward motion of air moving across areas of high elevation causes condensation of moisture in the atmosphere (Trentwartha , 1980). Thus higher elevations tend to be cooler and wetter.

Understanding the water budget is an essential first step in estimating streamflow. When precipitation falls to the ground, it is either used by vegetation, infiltrates below the surface of the earth, or runs off to streams and rivers (Figure 1). Water above the surface can be intercepted by trees or evapo-transpired. It can also fall directly on a stream or river, or it can fall on the ground and run off to rivers and stream. When water infiltrates through the vadose zone, it also has the opportunity to evapo-transpire and be used by vegetation. The remaining water infiltrates to the phreatic zone to become part of the groundwater. There can also be evapo-transpiration from the phreatic zone. Eventually, groundwater discharges to streams and rivers becoming base flow. By calculating a mass balance of the water budget, the following water budget can be derived (Dunne and Leopold, 1978):

$$P = I + AET + OF + \Delta SM + \Delta GWR + GWR \quad (2)$$

Where P = precipitation, OF = overland flow, I = interception, AET = actual evapotranspiration, ΔSM = change in soil moisture, ΔGWS = change in groundwater

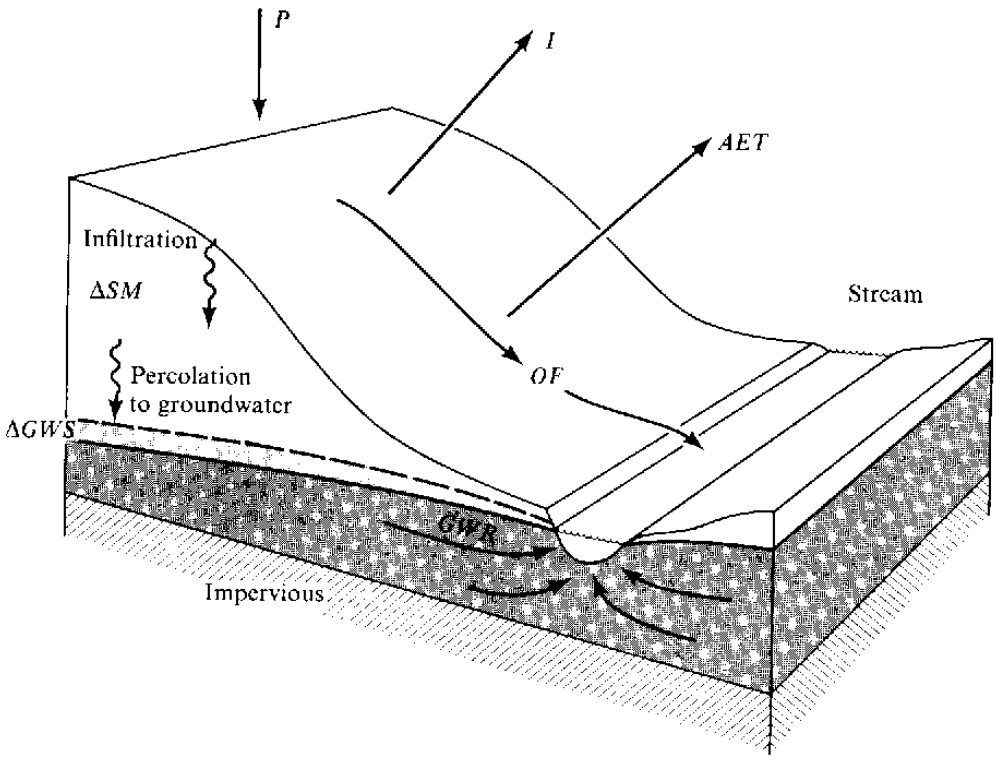


Figure 1 – Water budget, after Dunne and Leopold, 1978: Where P = precipitation, OF = overland flow, I = interception, AET = actual evapotranspiration, ΔSM = change in soil moisture, ΔGWS = change in groundwater storage, and GWR = groundwater runoff

storage, and GWR = groundwater runoff. Changes in soil moisture and groundwater storage can be attributed to evapotranspiration in the vadose and phreatic zones.

While direct precipitation and runoff contribute to stream flow, groundwater discharge also plays an important role. The role of groundwater is especially important during low-flow seasons or droughts. The base flow of a stream is groundwater discharge, by definition (Keller, 1988).

Geology influences the proportions of runoff vs. infiltration, and the rate of runoff to streams. Local permeability of rocks or sediment beneath streams may influence the amount of groundwater loss or gain (Dunne and Leopold, 1978). The geomorphic character of drainage basins may also affect surface flow. Elongate drainage basins with straight channels tend to transfer runoff to streams far faster than equant drainage basins, leaving little time for infiltration or evapotranspiration (Keller, 1988). Geology may also influence runoff amounts by the types of soils that evolve from local parent material. More clay-rich soils (as derived from shale) may induce high runoff and less groundwater recharge (Dunne and Leopold, 1978). Soils with high permeability allow more rapid infiltration.

Land use, soil type, and vegetation cover all play a large role in how much water is delivered to a stream during a precipitation event, yet these variables were not included in the WVBPH analysis. A site-specific analysis in the midwestern United States showed that 49% of a precipitation event falling on barren land will run off and 51% will either infiltrate or evapotranspire (Dunne and Leopold, 1978). In contrast, less than 1% of a precipitation event will run off in a forested region of the midwestern United States (Dunne and Leopold, 1978). Despite the high percentage of forestation in West Virginia,

there are local unforested areas due to urbanization, agriculture, and mining. These areas may have a much higher runoff rate than forested areas. This potential for large differences in runoff rates warrants further analysis of the impact of land use and vegetation type on stream flow.

These variables (precipitation, land use, vegetation, soils, underlying geology and groundwater recharge rate) also affect stream flow and should not be ignored in the synthesis of streamflow. The challenge, however, is to determine the optimum selection and quantification of these variables of available, as well as the necessary data distillation.

Relevant Prior Investigations

In a study of relationships between flow and several geomorphic variables for small watersheds in Kentucky, Haan and Read (1970) generated a predictive flow equation for estimation of mean annual runoff. This analysis used multiple regression to derive a prediction equation employing the following variables; stream discharge, mean annual rainfall, drainage basin area, average land slope, axial basin length, perimeter, basin diameter (the largest circle that fits within the basin), basin shape factor, stream frequency and a relief ratio (Haan, 1977). Using these data, a PCA was performed on all variables except drainage basin area in order to reduce them to components for use as independent variables (Haan, 1977). It was assumed that mean annual rainfall was an independent variable and mean annual runoff was the only dependent variable. Using the PCA results, a multiple regression analysis was performed to create a flow prediction equation (Haan, 1977).

A similar study was performed by Benson (1962b) on factors that influence floods in humid regions of diverse terrain. Data from the humid New England Region, where the historical climate record is long and spatially uniform, was used for this study. The author split these variables into two types, topographic and meteorologic. Topographic variables included drainage area, channel slope, profile curvature, shape factors, storage area, altitude index, stream density, soils, cover, land use, and urbanization (Benson, 1962b). Meteorologic variables were rainfall, snowfall, and temperature (Benson, 1962b). Based on a correlation analysis the author found that drainage area, channel slope, storage, rainfall intensity, temperature, and an orographic factor were most influential on floods in New England.

Benson (1964d) compared annual peak discharges to a number of hydrologic factors in the western Gulf of Mexico basin. Similar to Benson (1962b), drainage area, channel slope, altitude, length of basin, a shape factor, channel geometry, mean annual rainfall, and mean annual snowfall were used in multiple regression analysis. Benson added variables for stream order, soil, geology, orientation, forested area, basin rise, rainfall intensity, thunderstorm days, water content of snow, winter temperature, spring temperature, wind, evaporation, and a monthly runoff ratio. Benson's analysis showed that discharge was best correlated with drainage basin area, channel slope, storage, rainfall intensity, channel length, ratio of runoff to precipitation during the month of peak annual discharge, and number of thunderstorm days.

A procedure for estimating annual streamflow was developed for the unglaciated Allegheny Plateau Division by Brakenseik (1961). The variables mean annual

streamflow, mean annual precipitation, watershed area, perimeter, length of principal watercourse, and maximum relief were used to create the equation

$$Q_t = Q(1 + CV_q K) \quad (3)$$

where Q_t is the predicted annual stream flow, Q is the average annual streamflow, CV_q is a variability parameter for annual streamflow, and K is the standard normal deviate for a probability level of (1/T). The streamflow variability parameter is the coefficient of variation associated with an appropriate record of annual precipitation.

This study created a predictive equation using the variables: watershed area, perimeter, relief ratio, and annual average precipitation. It was found that either watershed area or watershed perimeter, along with precipitation, best predicted average stream flow.

$$Q = 0.01334\ P^{-1.915} A^{0.026} \quad (4)$$

were Q is the average streamflow, P is the average precipitation, and A is the watershed area. Their data set yielded an R^2 of 0.86.

Diaz and others (1968) applied principal components analysis (PCA) and factor analysis to annual precipitation and runoff data for fourteen small agricultural watersheds in Ohio and seven watersheds in Texas. PCA was used to establish an orthogonal set of variables from the original data. The ratio of runoff to precipitation was used as the dependent variable, while drainage basin area, soil type, internal drainage, degree of erosion, land capability class, cultural practices, and drainage basin slope were chosen as independent variables. Factor 1 accounted for 45.5 % of the variance and determined that the greatest variation of the runoff/ precipitation ratio had the largest covariance with watershed area. Factor 2 only accounted for 12.9% of the variance and was associated with watershed slope. For Texas watersheds, the greatest variance (Factor 1) was due to

the soil type. Factor 2 was associated with cultural practices while Factor 3 was associated with land capability. These three factors accounted for 97.5% of the total variation in the data set. The use of PCA in this study showed the effect of several geomorphic factors on stream flow.

Haan and Allen (1972) performed a study comparing multiple regression and PCA regression for predicting stream flow. Eight geomorphic variables were collected for 13 small agricultural watersheds in Kentucky. These variables were area, land slope, axial length, perimeter, diameter of largest circle that can be drawn completely within the watershed, a dimensionless shape factor, streams per square mile, relief ratio, and mean annual precipitation. Results showed that area, slope, shape, and stream frequency were the most important variables for predicting streamflow. This study suggests that PCA should be used to extract the important variables for describing a system (Haan and Allen, 1970).

Geographic Information Systems

Geographic Information Systems (GIS) are a computer-based system that enters, stores, manages, analyzes, and displays spatial, and associated nonspatial data (Davis, 1996). These systems are composed of software, data, hardware and organization and people. The combination of these components allows for visualization of spatial data, and the creation of spatial models. This study used the ESRI software ArcView GIS 3.2 for data collection and spatial analysis.

Spatial analysis, among the most important applications of GIS, requires logical connection between attribute data and map features. Two types of operations can be performed within spatial analysis: (1) spatial queries and (2) the generation of new data

sets from the original data. To accomplish this, ArcView GIS 3.2 utilizes spatial data of one of two types: vector or raster. Vector data are points, lines, or polygons, all composed of a location and a direction. Raster data is composed of regularly spaced cells arranged in rows and columns, called "grids". Each cell has a uniform size and is assigned a value according to the data that is being analyzed. To perform calculations on raster data, the Spatial Analyst extension must be used, a tool kit for understanding and analyzing spatial data. It allows the user to perform analysis on and between multiple data layers. The Spatial Analyst functions used in this study were *map query*, *find distance*, *summarize zones*, *map algebra*, *cell statistics*, *derive slope*, *fill sinks*, *flow direction*, and *watershed*.

A customized ArcView GIS interface called *Watershed Characterization and Modeling System 2.8* (WCMS) was created by the Natural Resource Analysis Center (2001) for the West Virginia Division of Environmental Protection (WVDEP). WCMS combines a wide variety of spatial data layer and water quality modeling components, along with watershed delineation, determination of average flow conditions, and tracking overland flow capabilities. This interface was created to facilitate access to West Virginia watershed data.

One of the tools available in WCMS is the *watershed delineation* tool. This tool defines the upstream watershed from any point on a stream by using several Spatial Analyst functions. First, raw DEMs are converted to ARC GRID format. Then, all sinks in the grid are filled. This is done in order to remove imperfections in the data and enable *flow direction* to run properly (Perez, 2000). Then a *flow direction* grid is created using the *flow direction* function to show the direction of flow from each cell in the elevation grid to its steepest down slope neighbor. Finally, the *Watershed* function is performed

using a user defined point. The *Watershed* function returns the upstream drainage area based on the *flow direction* grid from the user-defined point (Jenson and Domingue, 1988).

Purpose

The purpose of this study is to improve the understanding of the interaction between variables used for generating synthetic stream flow in West Virginia by incorporating appropriate variables for geomorphic, climatologic, and geologic characteristics. Data from 29 gauging stations, with at least thirty years of continuous record, will be used to develop this dataset. Potential applications include improved delineation of Zones of Critical Concern (ZCC) in source-water analysis by West Virginia or other states. ZCCs are areas within 1000 ft one each bank of a principal stream an d a 5 hour travel time upstream of a water intake (West Virginia Department of Health and Homan Resources, 1999).

Objectives

The first objective of this study is to choose optimum variables for predicting streamflow in West Virginia. This will be based on previous research and availability of data for West Virginia. The second objective is to use statistical analysis, including correlation analysis, PCA, and cluster analysis, to define the optimum quantification of these variables for predicting streamflow in West Virginia. Finally, spatial analysis will be used for the visualization and interpretation of the results of the statistical analysis.

Methodology

Dependent Variables

Stream flow

Monthly average stream flow records were collected from 29 USGS gauging stations throughout the state (Figure 2). Gauging stations were chosen based on three criteria. First, to avoid redundancy, the drainage basin for one station could not overlap that of another station used in this study. The second criterion was that the stations were not below a dam. The third criterion was that record length for flow matched or exceeded the record length for precipitation and temperature (1960 to 1990). Only twenty stations were acceptable under all three criteria, although many gauges in the state met the only the first two criteria. To maximize the size of the data set, the assumption was made that long term climatic averages could be used (Table 1). Stations were chosen with periods of records as close to that for climatic data as possible.

Independent Variables

Spatial Variables

Watershed Delineation

Watersheds were measured and employed as drainage areas upstream from each gauging station. This area was found using a 30-meter Digital Elevation Model (DEM) and the *Watershed Characterization and Modeling System* (WCMS) version 2.8 (Natural Resource Analysis Center, 2001). The DEM used in this study is part of the National Elevation Dataset of 1999 created by the U.S. Geological Survey. It uses 30 m cells and has been corrected to fill areas where there were inaccuracies between adjacent 7 1/2-minute quads. WCMS was used to delineate the upstream drainage basins for each of the

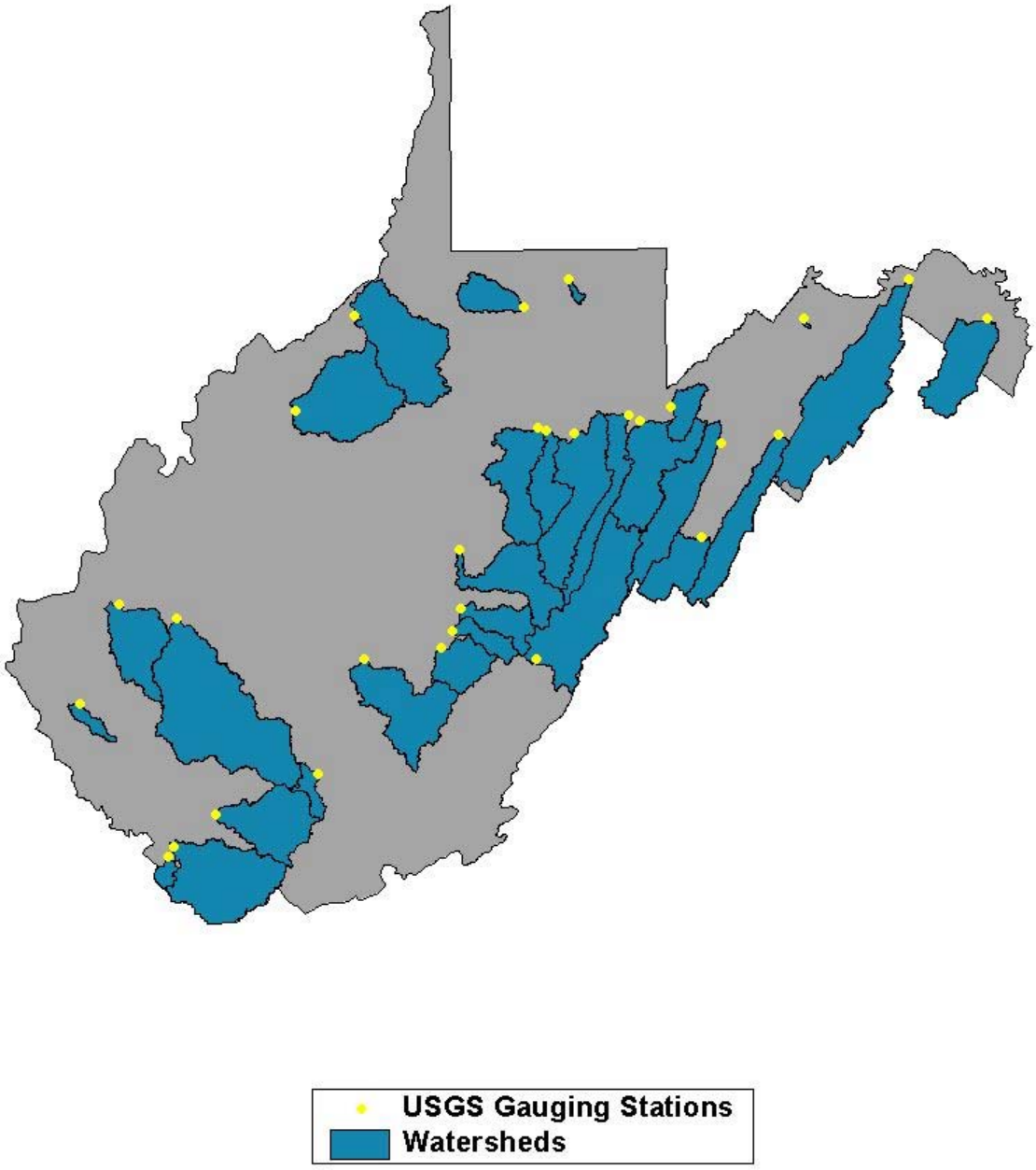


Figure 2 – Location of associated studied watersheds and USGS gauging stations

Table 1 - USGS Gauging stations used in this study

ID	STATION NAME	RECORD
1604500	PATTERSON CREEK NEAR HEADSVILLE, WV	1939-1995
1605500	SOUTH BRANCH POTOMAC RIVER AT FRANKLIN, WV	1941-1995
1606000	N F SOUTH BR POTOMAC R AT CABINS, WV	1941-1980
1608000	SO FK SOUTH BRANCH POTOMAC R NR MOOREFIELD, WV	1929-1995
1611500	CACAPON RIVER NEAR GREAT CACAPON, WV	1923-1995
1616500	OPEQUON CREEK NEAR MARTINSBURG, WV	1948-1995
3051000	TYGART VALLEY RIVER AT BELINGTON, WV	1908-1995
3052000	MIDDLE FORK RIVER AT AUDRA, WV	1943-1995
3053500	BUCKHANNON RIVER AT HALL, WV	1916-1995
3061500	BUFFALO CREEK AT BARRACKVILLE, WV	1908-1995
3062400	COBUN CREEK AT MORGANTOWN, WV	1966-1994
3065000	DRY FORK AT HENDRICKS, WV	1941-1993
3066000	BLACKWATER R AT DAVIS, WV	1922-1995
3069000	SHAVERS FORK AT PARSONS, WV	1911-1993
3114500	MIDDLE ISLAND CREEK AT LITTLE, WV	1916-1995
3155500	HUGHES RIVER AT CISCO, WV	1929-1994
3182500	GREENBRIER RIVER AT BUCKEYE, WV	1930-1995
3185000	PINEY CREEK AT RALEIGH, WV	1952-1982
3186500	WILLIAMS RIVER AT DYER, WV	1930-1995
3187500	CRANBERRY RIVER NEAR RICHWOOD, WV	1945-1995
3189000	CHERRY RIVER AT FENWICK, WV	1930-1982
3190400	MEADOW RIVER NEAR MT. LOOKOUT, WV	1967-1995
3200500	COAL RIVER AT TORNADO, WV	1909-1995
3202400	GUYANDOTTE RIVER NEAR BAILEYSVILLE, WV	1969-1995
3204500	MUD RIVER NEAR MILTON, WV	1939-1980
3206600	EAST FORK TWELVEPOLE CREEK NEAR DUNLOW, WV	1965-1995
3213000	TUG FORK AT LITWAR, WV	1931-1984
3213500	PANTHER CREEK NEAR PANTHER, WV	1947-1986
3194700	ELK RIVER BELOW WEBSTER SPRINGS, WV	1909-1916

gauging stations used in this study (Figure 2). In ArcView, the resulting temporary shapefiles were converted to permanent shape files for vector calculations and to grids with 30m cells for raster calculations. Drainage basin area was calculated using WCMS (Table 2).

Elevation (E_{max}, E_{min}, E_{mean}, E_{SD}, and E_{range})

Elevation of each watershed was determined using ArcView from the DEM. The analysis was masked to the extent of each watershed. The map calculator was used to multiply a grid of each watershed by the DEM, creating an output DEM the same size as the input watershed with a 30 m cell size. Statistics were then calculated on the value field of this grid's attribute table, yielding the maximum elevation, minimum elevation, mean elevation, change in elevation, and the standard deviation of the elevation (Table 2).

Stream Length (L)

Stream length was calculated as the length of all contributing streams in each watershed. Previous studies calculated stream length as the length of the mainstem however, in this case, distinguishing the mainstem from tributaries in this study was impractical due to the extremely dendritic nature of several of the small watersheds, with numerous long tributaries.

For length calculation, the *select by theme* tool was applied to all stream centerlines within each watershed. The stream centerlines consists of the centerline of each stream represented by numerous stream segments mapped a 1:100,000 National Hydrography Dataset (NHD) compiled by the USGS. Portions of the shapefile were updated to 1:24,000 scale as an intermediate product before the 1:24,000 scale NHD was available. All the stream centerline segments that were completely contained within each

Table 2 - Geomorphic Variables

Watershed	A	L	Ss	Ws	Emax	Emin	Emean	Erange	Esd
Blackwater River	86	170059.16	0.3	6.08	1347	923	1133	424	121
Buckhannon River	277	718368.84	1.4	11.5	1199	412	805	787	226
Buffalo Creek	115	260493.11	1.5	13.39	578	270	394	248	71
Cacapon River	677	1360389.57	2.7	10.78	1026	139	584	887	256
Cherry River	150	289923.32	5.5	13.35	1373	636	1005	737	213
Coal River	862	1628308.75	1.4	20.4	1069	167	616	902	259
Cobun Creek	11	24432.26	1.1	9.86	676	270	473	406	118
Cranberry River	80	152024.54	3.2	14	1400	650	1025	750	217
Dryfork	345	656574.15	1.9	14.24	1466	575	989	951	274
East Fork Twelvepole	39	72039.67	1.6	17.72	559	220	384	339	95
Elk River	266	486888.09	5	17.57	1498	304	896	1185	342
Greenbrier River	540	794243.6	0.6	14.44	1484	632	1059	852	246
Guyandotte River	306	7518883.21	4.9	18.12	1084	347	716	737	213
Hughes River	452	1133355.81	1.9	13.4	429	182	305	247	71
Meadow River	365	743383.31	0.2	11.99	1328	373	851	955	276
Middle Fork River	149	373967.79	4.4	12.23	1171	520	846	651	188
Middle Island Creek	458	986810.07	1.1	15.01	506	189	346	317	91
Mud River	256	516758.817	0.8	14.35	512	172	340	307	97
N Fork S Br Potomac	314	560115.48	4.9	20.13	1481	381	900	1163	336
Opequon Creek	272	483326.06	2.3	4.1	506	106	304	400	115
Panther Creek	31	71144.77	5.1	19.7	731	312	521	419	119
Patterson Creek	219	6393.73	2.4	10.84	519	195	355	324	93
Piney Creek	52	108489.73	1	11.14	1041	635	838	406	118
S fork S Br Potomac	283	517354.54	4.7	18.27	1329	261	793	1068	308
Shavers Fork	214	385162.23	5	14.47	1490	0	933	1490	288
So Branch Potomac	182	236136.82	1	13.11	1391	520	953	871	250
Tug Fork River	504	1060948.1	1	20.35	1046	263	662	783	222
Tygart River	408	965539.13	2.2	13.63	1469	501	985	968	280
Williams River	128	142337.37	1.1	14.52	1437	664	1049	773	222

watershed were selected. Stream centerlines were used because the length of the right and left bank of a river can vary. The vector stream centerline shapefile was clipped using the *geoprocessing wizard* to include only stream centerlines within each watershed boundary. Field statistics were then performed on the length field in the attribute table. The results are shown in Table 2 as total stream length.

Stream Slope (S_S)

The stream slope for each watershed, except North Fork of the South Branch of the Potomac, South Fork of the South Branch of the Potomac, South Branch of the Potomac, and Opequon Creek, was found by analysis of a stream slope grid found in WCMS. The stream slope grid was developed from a raster version of the stream centerline shape file, with each cell containing an interpolated elevation value from the DEM. Then stream reaches were defined as the segment of a stream that begins either at the headwater or the confluence of two streams and ends at the next confluence of two streams. The slope was then calculated for each stream reach as the change in elevation over the stream length. Each stream reach has the same slope value (Natural Resources Analysis Center, 2000). In order to calculate the average stream slope of each watershed in this study, a grid of each watershed was multiplied by the stream slope grid. Field statistics were performed on the resulting attribute tables. The median stream slope for each watershed was taken from these statistics (Table 2).

Because the headwaters of North Fork of the South Branch of the Potomac, South Fork of the South Branch of the Potomac, South Branch of the Potomac, and Opequon Creek are in Virginia, the above method for finding stream slope could not be used, as the stream slope grid is only available in West Virginia. For streams with headwaters in Virginia, stream slope was calculated as the change in elevation divided by the length of

the mainstem of the stream (Table 2). Mainstem length was found by selecting the stream centerline segments along the mainstem, then summing their lengths. The change in elevation was calculated from the DEM. The high and low elevations along the mainstem of the stream or river were found by using the *identification* tool with the DEM active. The slope was then calculated by dividing the change in elevation by the stream length.

Steepness Index (S_w)

GIS was used to calculate the average slope within each watershed. The *calculate slope* function in Spatial Analyst was used to creates a 30m grid of the average slope in the study area from a 30m DEM (Figure 3). ArcView calculates slope by identifying the maximum rate of change from each cell to its neighbors. The output grid values express slope in degrees. The map calculator was used to multiply a grid of each watershed by the grid of the slope with an analysis mask of each watershed. The result was a grid of the slope within each watershed. The *summarize zones* tool was then used to determine the average slope within each watershed (Table 2).

Geologic Variables

Land Use / Land Cover (F)

The land use/ land cover (LULC) variable was calculated from the Gap LULC grid. This grid was created by the Natural Resources Analysis Center (NRAC) at West Virginia University from 1:50,000 scale base maps (Figure 4). The LULC grid was multiplied by a grid of each watershed. The resulting attribute table for each watershed was used to calculate the percent forested area. This was done by summing the number of cells representing the 15 forest classes, then dividing this sum by the total number of cells in each watershed (Table 3).

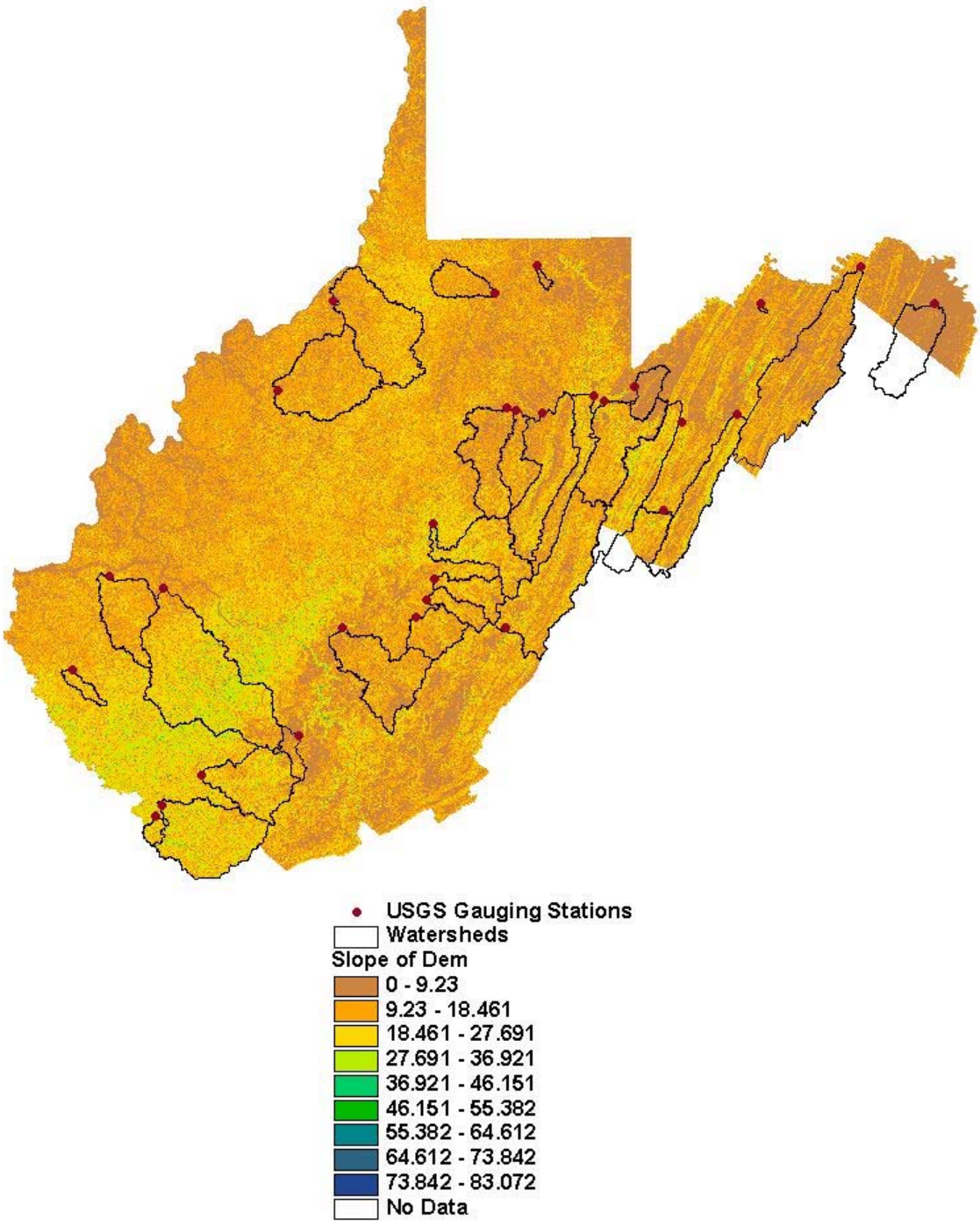


Figure 3 - %Slope of land surface derived from Digital Elevation Model

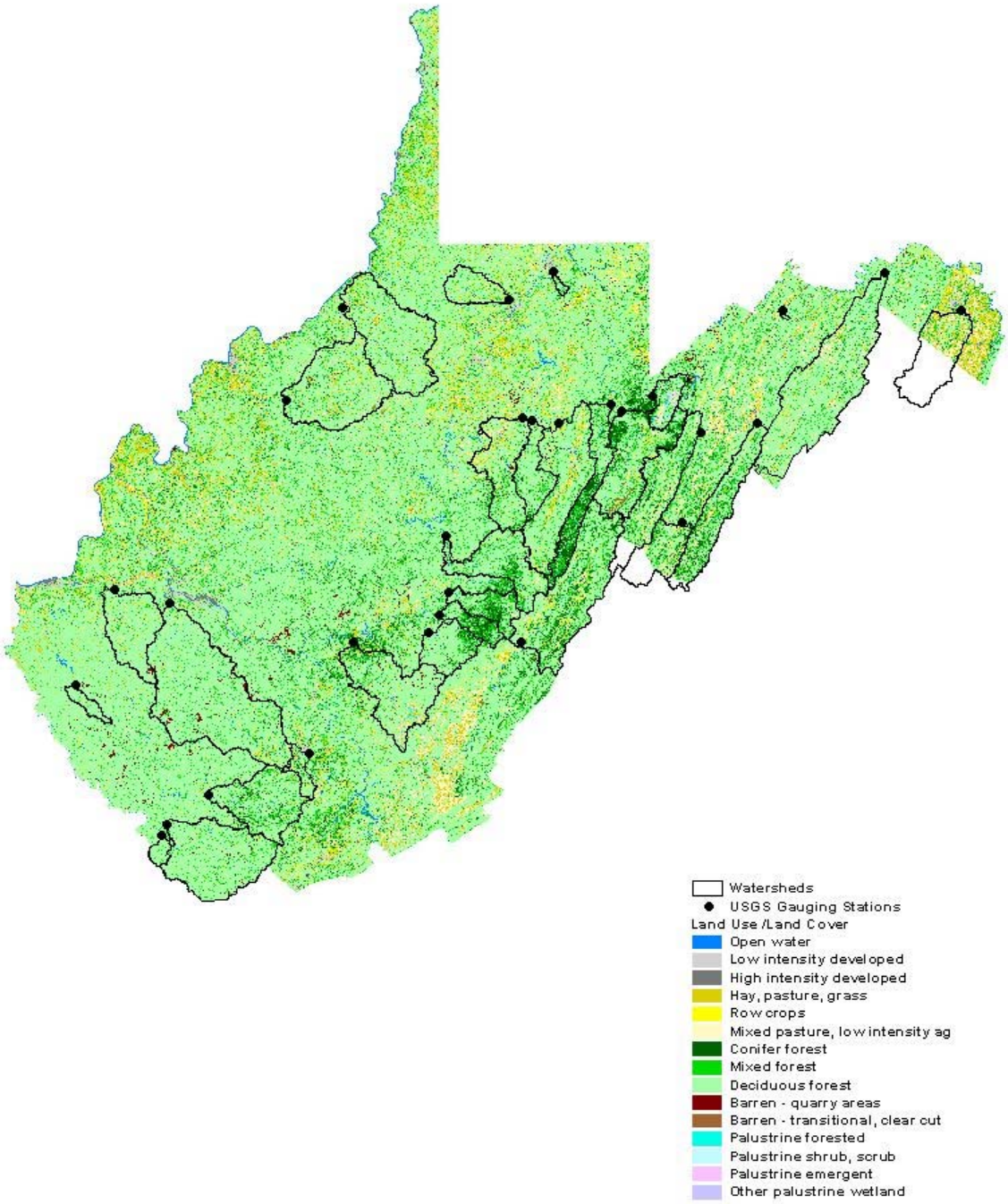


Figure 4 – Land use/ land cover data for West Virginia

Table 3 - Geologic Variables

	%F	R	S	S	G	L/W
Blackwater River	71	22.5	c	3	3.34	0.4705
Buckhannon River	81	21.4	c	3	1.13	3.667
Buffalo Creek	85	21.4	b	2	1.13	2.1845
Cacapon River	85	8.7	c	3	1.26	4.2003
Cherry River	97	27.8	c	3	3.34	0.9403
Coal River	93	11.9	d	4	3.34	2.2263
Cobun Creek	82	21.4	c	3	1.13	4.2212
Cranberry River	97	31.6	c	3	3.34	3.6107
Dry Fork	89	21.4	c	3	0.4	2.7259
East Fork Twelvepole	97	12.4	c	3	3.34	3.5389
Elk River	95	23.9	c	3	1.13	5.815
Greenbrier River	88	21.1	c	3	3.34	3.6836
Guyandotte River	94	14.5	c	3	1.13	1.624
Hughes River	83	7.1	b	2	1.13	1.3853
Meadow River	84	20.6	c	3	3.34	2.7506
Middle Fork River	94	24.5	c	3	3.34	4.8624
Middle Island Creek	87	8	b	2	1.13	1.9023
Mud River	91	11.9	d	4	1.13	1.7573
N Fork S Br Potomac	88	11	c	3	6.45	6.3809
Opequon Creek	36	9.8	b	2	0.46	2.7001
Panther Creek	99	11.1	b	2	3.34	1.6779
Patterson Creek	97	7.3	c	3	6.45	2.7997
Piney Creek	80	11.9	c	3	3.34	0.4994
S Fk S Br Potomac	89	9	c	3	3.34	14.9865
Shavers Fork	96	24.8	c	3	6.45	7.7616
S Branch Potomac	73	11.6	b	2	1.26	2.2989
Tug Fork	96	11.3	b	2	3.34	1.365
Tygart River	74	15.4	c	3	0.4	5.7814
Williams River	97	26.4	c	3	3.34	2.7285

Groundwater Recharge (R)

Kozar and Mathes(2001) published mean ground-water recharge rates estimated from stream flow data at 41 West Virginia gauging stations. Recharge rates were calculated by using the recession-curve displacement method (Rorabaugh, 1964). The study did not incorporate evaporation or transpiration into the recharge model. Twenty-three of the gauging stations used by Kozar and Mathes (2001) were in the watersheds used in this study. The values calculated by Kozar and Mathes (2001) were used as the recharge value for those watersheds (Table 3).

Kozar and Mathes derived 41 spatially-averaged recharge estimates from six large river basins that "share similar geologic, topographic, and climatalogical settings" (Kozar and Mathes, 2001). Kozar and Mathes did not use any stations on the Buckhannon River, Buffalo Creek, Cobun Creek, Mud Creek, or Dry Fork in their study. As an estimate of recharge for these stream basins, the average values for the large river basins (Potomac, Little Kanawha, Ohio tributaries, Monongahela , Kanawha (western portion), Kanawha (eastern portion), Tug, Twelvepole and Guyandotte basins were employed (Table 3). It is recognized that use of average data for some watersheds and calculated data for others introduces some inconsistency into the data set.

Soils (S)

Soils data for the study area were obtained from the State Soil Geographic (STATSGO) database for Virginia and West Virginia. This data set consists of digital soil maps originally generated by the National Cooperative Soil Survey (STATSGO Metadata) These maps are at a 1:250,000 scale and are in the form of shape files. The

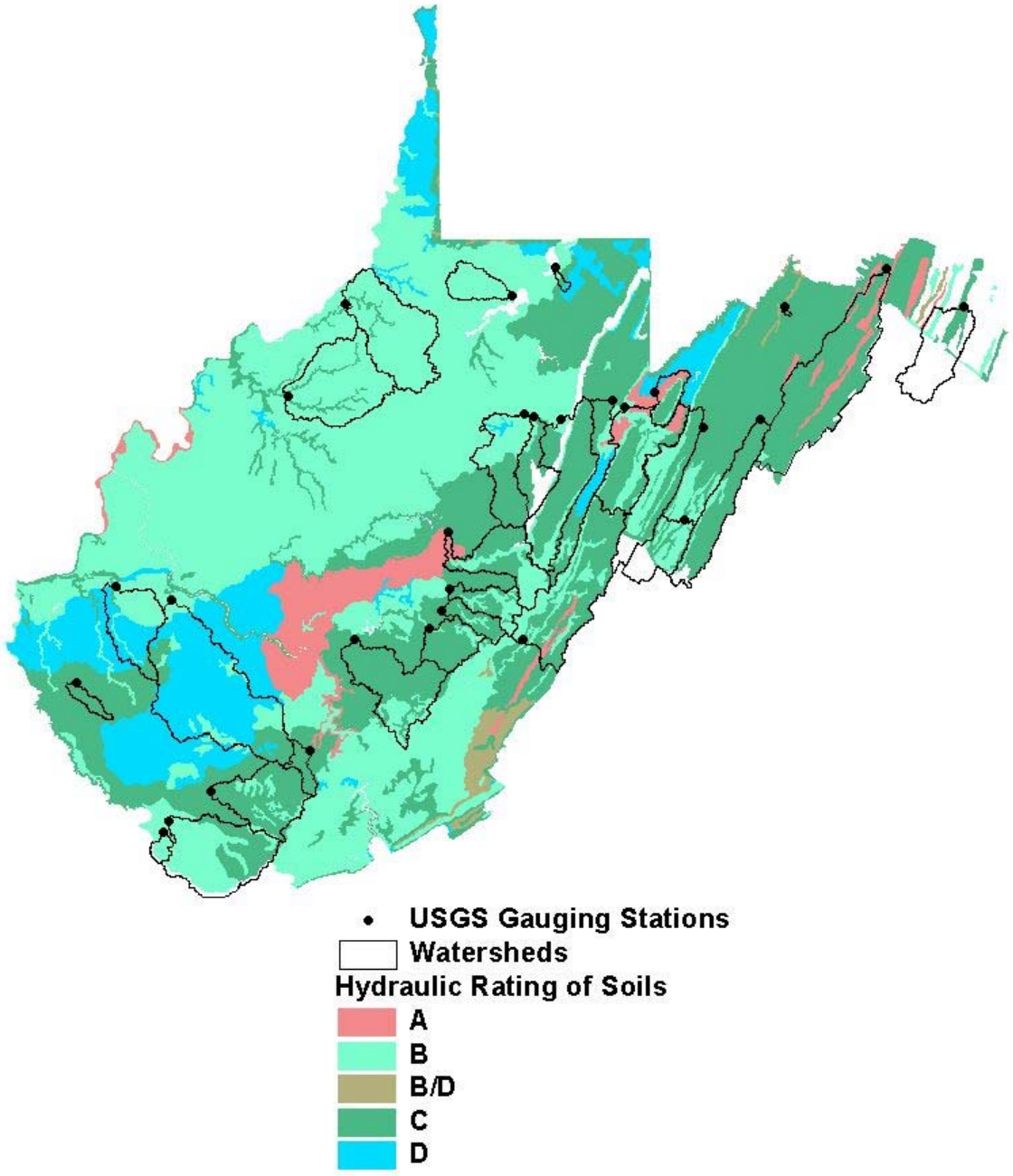


Figure 5 – Hydraulic rating of soils: A = high infiltration capability, B = moderate infiltration capability, C = low infiltration capability, D = very low infiltration capability

soils maps for both states were merged using the merge theme command in the ArcView *geoprocessing wizard*.

A polygon of each watershed was clipped from the soils shapefile using the *geoprocessing wizard*, then converted to a grid. Once each grid was complete, the legend was set to display the attribute hydraulic rating (Figure 5). The hydraulic rating of a soil is a classification system from A to D based on the soil's infiltration capacity. According to the National Cooperative Soil Survey (1976),

> "Hydrologic soil groups . . . Refers to soils grouped according to their runoff-producing characteristics. The chief consideration is the inherent capacity of soil bare of vegetation to permit infiltration. The slope and the kind of plant cover are not considered, but are separate factors in predicting runoff. Soils area assigned to four groups. In group A are soils having a high infiltration rate when thoroughly wet and having low runoff potential. They are mainly deep, well drained, and sandy or gravelly. In group D, at the other extreme are soils having a very slow infiltration rate and thus a high runoff potential. They have a claypan or clay layer at or near the surface, have a permanent high water table, or are shallow over nearly impervious bedrock or other material. A soil is assigned to two hydrologic groups if part of the acreage is artificially drained and part is undrained."

Statistics were performed on the hydraulic rating field of the attribute table to determine the majority rating for each watershed.

As nominal data can not be used in PCA, a data conversion to ordinal scale was performed. The hydraulic rating for each watershed was assigned a value 1 for Hydraulic rating A (minimum runoff yield) through 4 for Hydraulic rating D (maximum runoff yield) was assigned a 1 because this conversion gives ordinal values that can be used in PCA. Non-ratio data cannot be normally distributed unless the number of cases is very large.

Geology (G)

To derive quantitative variables for geology in West Virginia and Virginia, a digital version of the 1:250,000 State Geologic Map (West Virginia Geological and

Economic Survey, 1968) found in WCMS was used. In 1998 the WV Department of Environmental Protection digitized the map. A new attribute field was created for this study integrating group and formation names (Figure 6).

A shale-to-sandstone ratio was calculated for each of the categories shown in Table 4 using the bedding thickness of each lithology from county reports. The location used for each formation was one of the watersheds in this study (Figure 6). The average value of sandstone to shale was calculated using the *summarize-by-zones* tool in ArcView.

Shape Factor (L/W)

The length to width ratio (L/W) was calculated by using the *measure distance* tool in ArcView. First the median length was calculated, then the median width. The width was measured upstream of where the watershed tapers to its outlet. Length was measured from the outlet to the headwaters. Then median length was divided by median width as a measure of basin shape factor (Table 3).

Climatic Variables

Temperature (T_{max}, T_{min})

Temperature data for the study area was obtained from ZedX (2001), a compilation of climatological data from 1960 - 1990 from climatological station records compiled by the National Climatic Data Center (NCDC). An algorithm was created to transform climate data for 1960-1990, including average maximum temperature (T_{max}), average minimum temperature (T_{min}), and mean annual precipitation (P_{mean}), to find

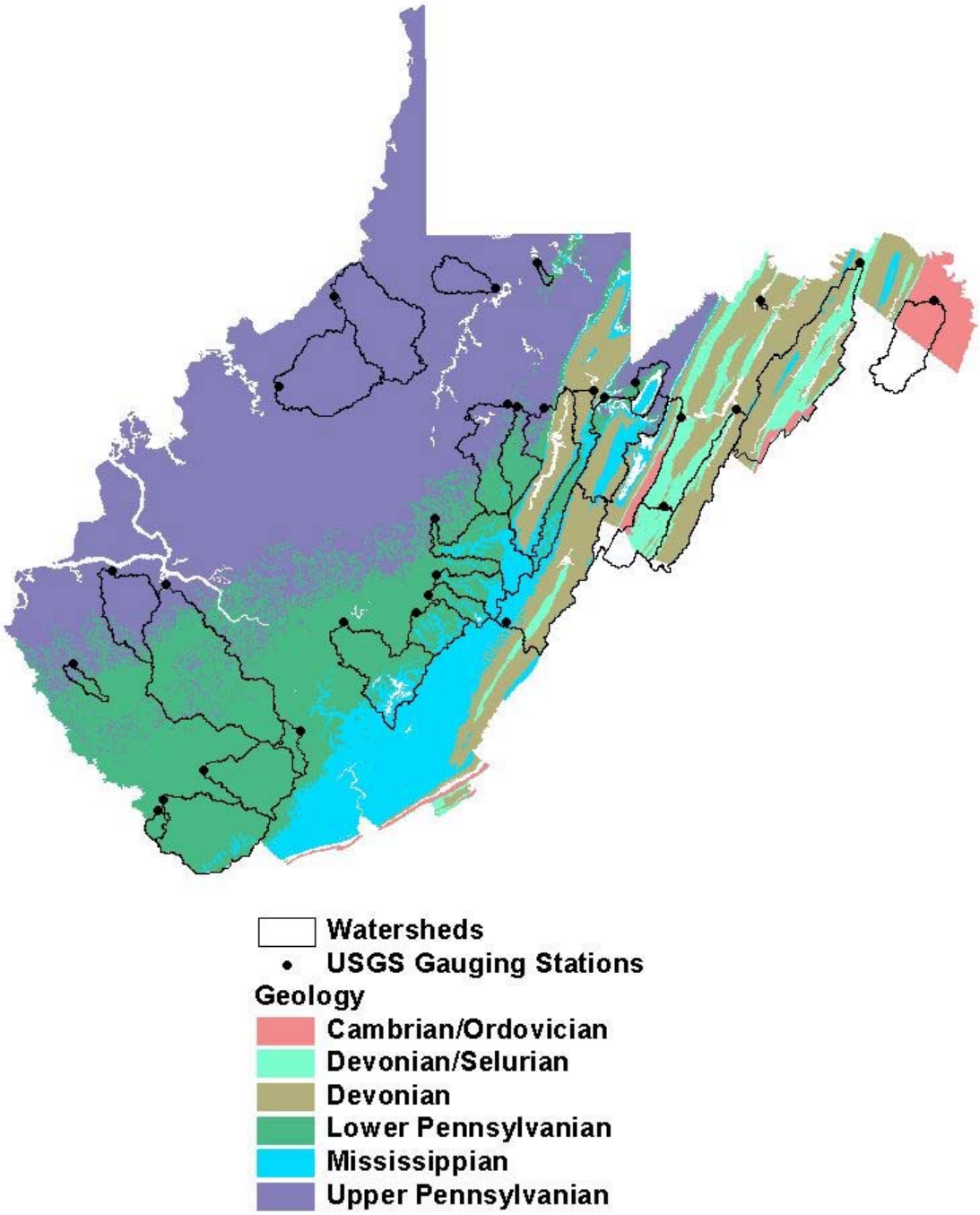


Figure 6 – Modified geologic map for West Virginia

Table 4 - Conversion of geology to shale/sandstone ratio

Geologic Age Range	Map Unit	Shale/Sandstone
Upper Penn	Dunkard - Conemaugh	1.13
Lower Penn	Pottsville – Allegheny	3.34
Mississippian	Bluestone - Pocono	0.4
Devonian	Hampshire - Marcellus	6.45
Dev/Sel	Oriskany - Tuscarora	1.26
Cam/Ord	Stonehenge - Weverton	0.46

monthly and annual average temperaturesfor stations across West Virginia and Virginia. These data were then gridded to approximately one square kilometer cell size.

Figures 7 and 8 show the annual T_{min} and $T_{max,}$ respectively, for the study area from 1960-1990. Analysis masks of each watershed were used in the calculation of T_{max} and T_{min}, and the output grid was set to have a cell size equal to the cell size of the original temperature grids. A grid of each watershed multiplied both the T_{max} grid and the T_{min} grids. Then field statistics were performed on the value field in the attribute table of each output grid. The statistics returned the average T_{max} and T_{min} for each watershed (Table 5).

Precipitation (P_{mean}, P_{SD})

The precipitation data used in this study was also taken from ZedX (2001) as 30 year averages (1960 to 1990) interpolated across the state (Figure 9). Similar to the temperature data, the average monthly precipitation from climatalogical stations were averaged to create the annual average precipitation (P_{mean}) (ZedX, 2001)

The precipitation grid was multiplied by a grid of each watershed. The result was a precipitation grid for each watershed. Statistics were performed on the value field of the attribute tables to calculate the P_{mean} and the standard deviation of the annual average precipitation (P_{SD}).

Correlation Analysis

The covariance between variables is a measure of their joint variation about a common bivariate mean (Davis, 1986). In order to compare the two variables, the correlation coefficient, a measure of the strength of this covariance, is calculated by the cross-variation of variables (Davis, 1986). A correlation coefficient of +1 indicates a perfectly-linear positive relationship, whereas a correlation of –1 indicates a perfectly-

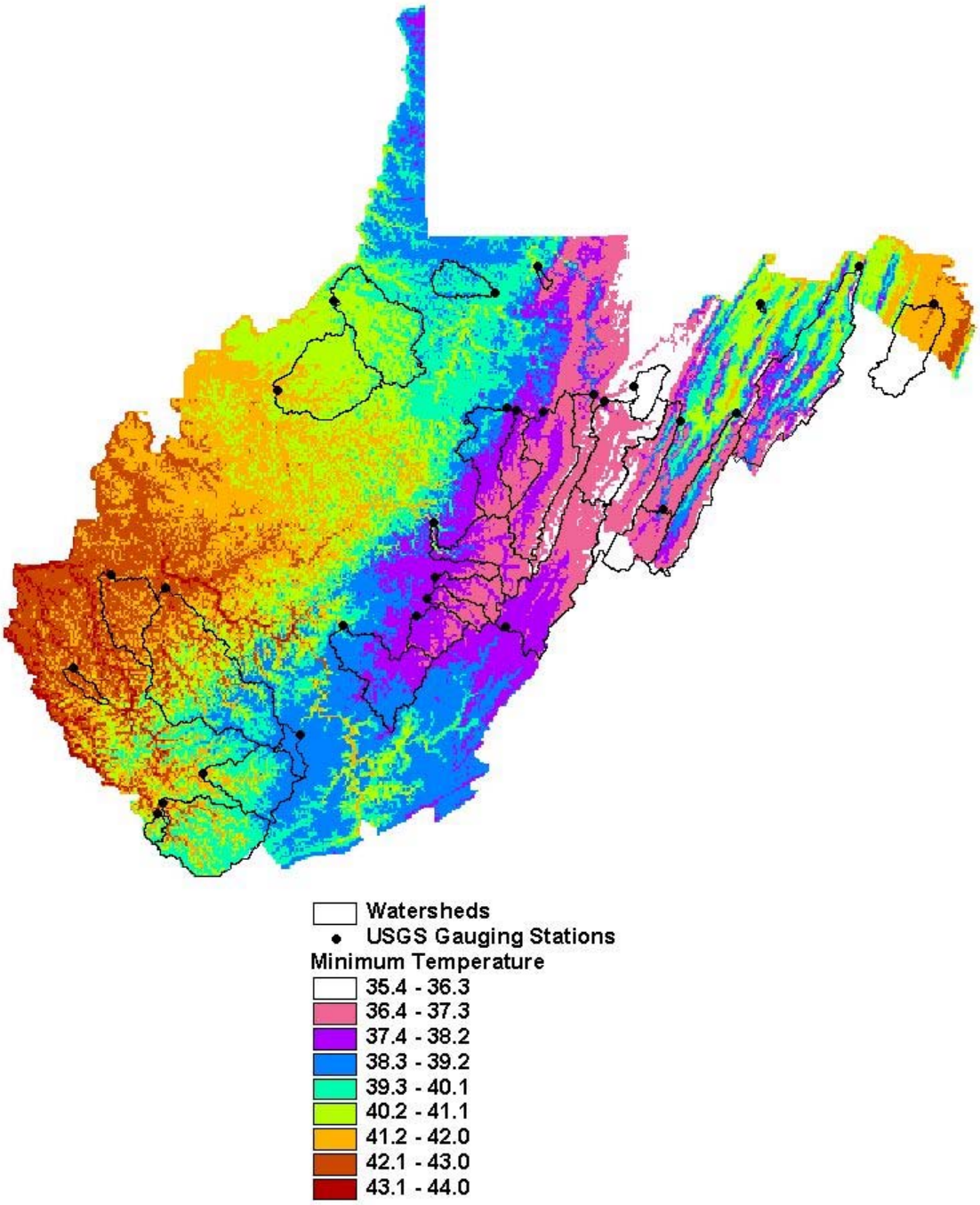


Figure 7 – Average annual minimum temperature

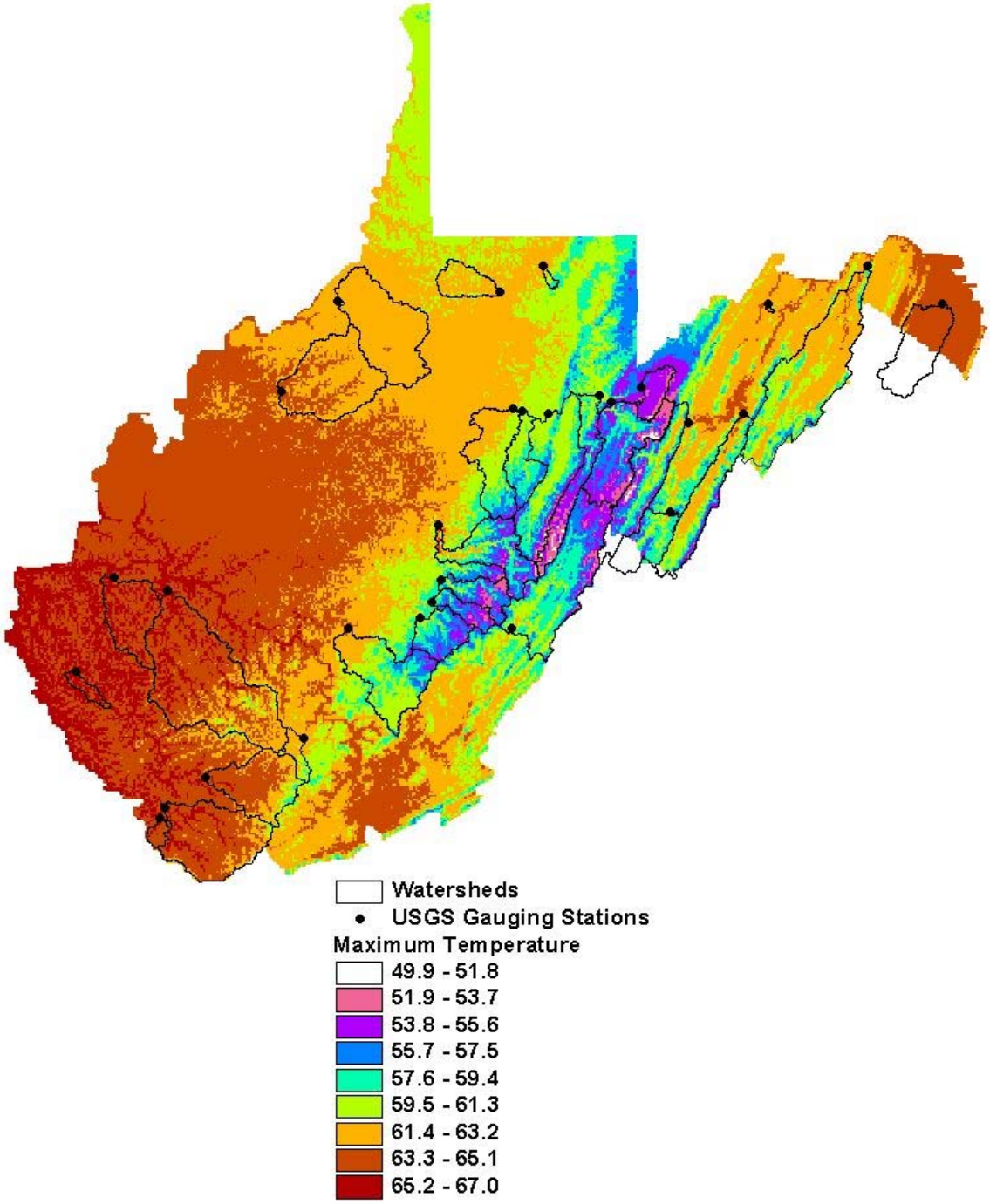


Figure 8 – Average annual maximum temperature

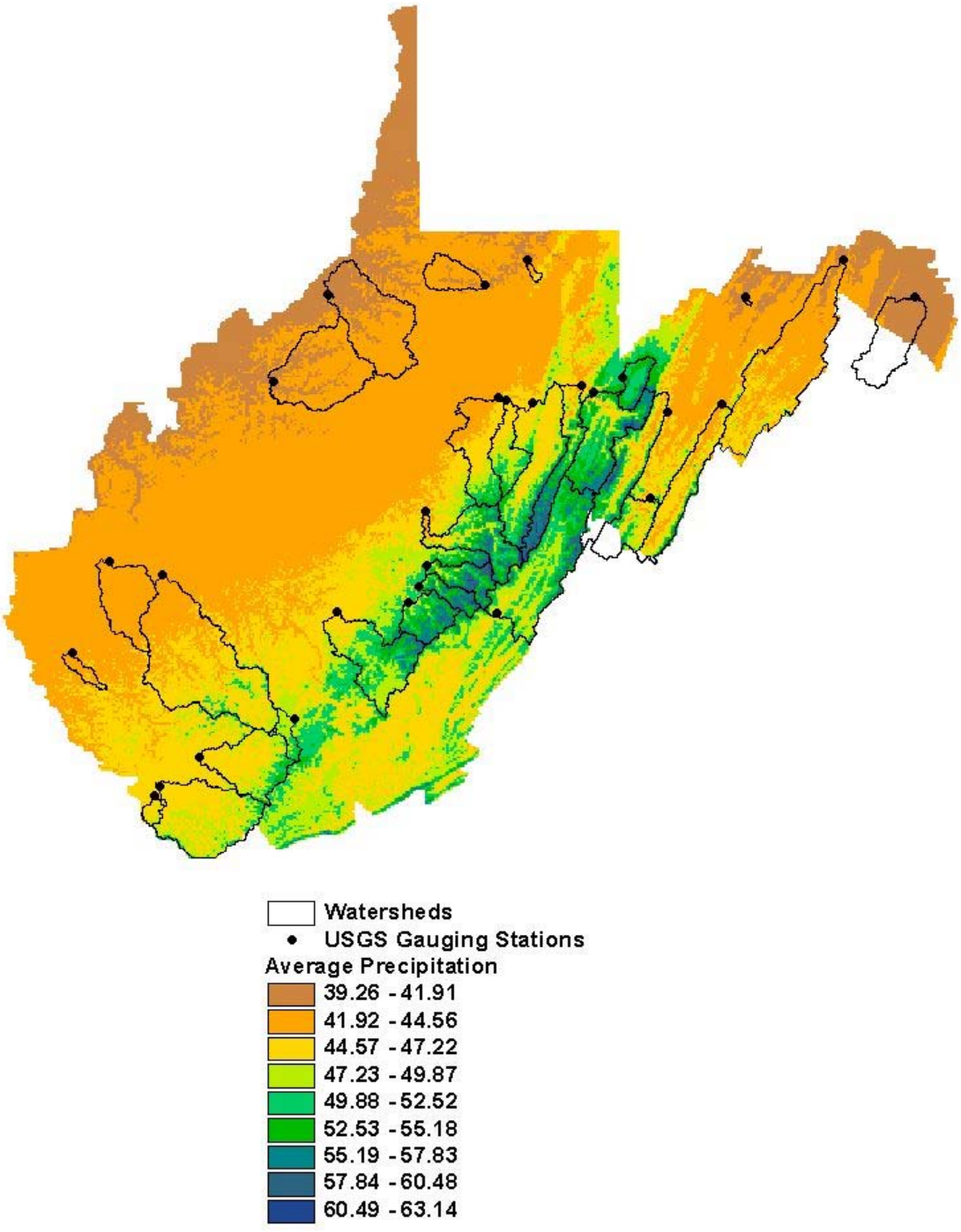


Figure 9 – Average annual precipitation

Table 5. - Temporal Variables

	Tmax	Tmin	Pmean	Psd
Blackwater River	53.9	36	53.97	1.95
Buckhannon River	58.3	38.1	47.99	3.54
Buffalo Creek	61.6	39.4	42.35	0.34
Cacapon River	59.8	38.7	44.3	1.94
Cherry River	57.5	37.6	54.12	2.95
Coal River	62.6	40.8	46.46	2.56
Cobun Creek	60.5	38.6	43.76	1.46
Cranberry River	56.3	37.4	55.27	2.89
Dry Fork	55.8	36.6	52.45	3.62
East Fork Twelvepole	64.9	41.8	44.18	0.47
Elk River	57.5	38.4	51.85	4.18
Greenbrier River	56.3	37.4	55.27	2.89
Guyandotte River	61.9	40.3	48.63	2.37
Hughes River	63	40.7	42.24	0.41
Meadow River	58.8	38.6	50.99	3.34
Middle Fork River	58.2	37.5	48.53	2.54
Middle Island Creek	62.3	40	42.11	0.6
Mud River	64.7	41.8	43.44	0.63
N Fork S Br Potomac	57.7	36.8	49.63	4.09
Opequon Creek	57.7	36.8	49.63	1.5
Panther Creek	64.6	41	46.88	0.75
Patterson Creek	62.4	40.1	42042	0.42
Piney Creek	61.1	39.7	49.84	1.85
S Fk S Br Potomac	63.6	37.7	46.22	2.69
Shavers Fork	55.7	36.8	54.19	4.77
S Branch Potomac	59.2	37.1	48.92	2.45
Tug Fork	63.1	40.9	46.1	1.93
Tygart River	56.4	37.1	49.98	4.07
Williams River	56.5	37.3	54.88	3.36

linear inverse relationship (Davis, 1986). Variables may be strongly correlated in a nonlinear way but have a low correlation coefficient.

Principal Component Analysis

PCA is a multivariate technique used to remove or filter intercorrelation between variables. Its objective is to transform the original variables into an identical number of uncorrelated or orthogonal components called eigenvectors (Haan, 1977). The number ofeigenvectors is equal to the number of original variables . These eigenvectors have eigenvalues, which are measures of their relative strength. An eigenvector with a high eigenvalue accounts for a large portion of the variance in the data set. The eigenvalues also are normalized to the sum of the number of original variables.

Each variable used in the analysis has a loading on each eigenvector. These loadings can be used to attach physical significance to the components. If a particular component is highly correlated with 1, 2, or 3 variables, then the component is a reflection of these variables (Haan, 1977). The sign of the loading indicates the relationship between each variable and the vector. For example, as elevation increases, temperature decreases. These two variables are strongly, but inversely, correlated. Both variables might display a high loading on the same eigenvector but with opposite signs. Observations are also given scores on eigenvectors. Analogous to the loadings, observations with high scores on an eigenvector are strongly influenced by that eigenvector.

PCA was chosen for this study because of the expected strong intercorrelation between variables that influence stream flow. The goal of the PCA is to create a short list of eigenvectors (2-5) that are independent (orthogonal) of one another but contain most

of the covariance information in the original dataset. These data may then be used in a regression analysis, as has been done many times in investigations of factors affecting streamflow. For example, McCuen and Snyder (1986) compared monthly runoff and precipitation to streamflow. PCA was performed first to remove the correlation between runoff and precipitation, then regression was performed to relate the resulting eigenvectors to streamflow.

This study did not include rotation of the eigenvectors, which involves moving the component axes (i.e., changing the eigenvector loadings) such that each variable has a factor loading near either zero or one (Davis, 1986). This causes the maximization of the variance on the factors but as a result they are no longer orthogonal (Davis, 1986). While PCA may produce eigenvectors with less total variance than the rotation techniques, it is an objective and unique analysis that require no operator intervention except for selection of appropriate input variables

Cluster Analysis

Cluster analysis is a technique to classify objects into groups on categories so that the relationships between the objects will be revealed (Davis 1986). First, the correlation coefficient is calculated for use as a similarity measure (Davis, 1986). Observations with the highest similarity area clustered first. When two observations are connected, that means that they have the highest correlations with each other (Davis, 1986). Once two observations are clustered, their correlation is averaged with all other observations in the analysis (Davis, 1986). A dendrogram is usually produced for visualizing the results of the cluster analysis. This is a tree diagram that shows the linkages and similarities of the observations.

Results

Correlation Analysis

The variables chosen in this study include five measures of elevation, two measures of slope, two measures of temperature, and two measures of precipitation. Because they are multiple measures of similar phenomena, there is undesired correlation within the dataset. It was therefore necessary to screen these variables and determine the optimum measure for each category to predict streamflow. To examine correlation, a correlation matrix (Table 6) was used.

In the unaltered dataset, annual discharge (Q) has a high positive correlation with A, E_{range}, E_{SD}, and P_{SD}. There is a slight positive correlation between annual discharge, S, E_{min}, E_{max}, L, and P_{mean}. Q has low correlation with T_{min}, F, R, and Em_{in}.

Understandably, the two temperature variables are highly correlated with each other, and correlate similarly with every other variable except Q. This covariation suggests that only one measure of temperature need be retained. T_{max} and T_{min} show moderate to strong negative correlation with P_{mean}, P_{SD}, R, E_{max}, E_{min}, E_{mean}, E_{range}, and E_{SD}. These variables also display moderate to strong positive correlation with S_w. T_{Max} and T_{Min} exhibit low correlation with A, L, S_S, and S.

P_{mean} shows a strong positive correlation with, P_{SD}, R, E_{max}, E_{mean}, E_{mange}, and, E_{SD}, and has moderate correlation with S, and S_S. P_{mean} has low correlation with F, A, and S_w. P_{SD} show similar behavior to P_{mean}, correlating strongly with R, E_{max}, E_{mean}, E_{mange}, and, E_{SD} and moderately with P_{SD} A, S_w, S_s, and S. Since both measures of precipitation are strongly correlated, only one need be retained. F has a strong correlation with only S_W and moderate correlation with S_S, S, and E_{range}. E_{mean} is strongly correlated with all of the

Table 6- Correlation Analysis

	Q	Tmax	Tmin	Pmean	Psd	F	R	A	L	Ss	Sw	Emax	Emin	Emean	Erange	Esd	S
Q	1																
Tmax	-0.252	1															
Tmin	-0.097	0.918	1														
Pmean	0.19	-0.802	-0.722	1													
Psd	0.446	-0.646	-0.662	0.731	1												
F	0.078	0.263	0.378	-0.011	0.082	1											
R	0.047	-0.676	-0.546	0.7	0.493	0.232	1										
A	0.854	0.007	0.084	-0.084	0.174	-0.084	-0.31	1									
L	0.247	0.151	0.223	-0.057	0.045	0.084	-0.16	0.33	1								
Ss	-0.11	0.006	-0.116	0.176	0.362	0.327	0.108	-0.19	0.204	1							
Sw	0.249	0.441	0.44	-0.093	0.182	0.687	-0.16	0.24	0.258	0.385	1						
Emax	0.364	-0.7	-0.689	0.83	0.913	0.165	0.567	0.12	0.015	0.252	0.172	1					
Emin	-0.089	-0.575	-0.519	0.649	0.274	-0.007	0.538	-0.27	-0.123	-0.169	-0.216	0.564	1				
Emean	0.234	-0.752	-0.716	0.882	0.8	0.133	0.644	-0.02	-0.041	0.144	0.034	0.959	0.75	1			
Erange	0.502	-0.467	-0.495	0.571	0.925	0.2	0.309	0.33	0.101	0.422	0.357	0.834	0.02	0.657	1		
Esd	0.53	-0.44	-0.48	0.553	0.914	0.177	0.274	0.38	0.13	0.382	0.387	0.857	0.12	0.68	0.961	1	
S	0.217	-0.011	0.006	0.184	0.382	0.306	0.175	0.09	0.069	0.132	0.177	0.323	0.07	0.266	0.34	0.374	1

Legend

Tmax = Maximum Temperature
Tmin = Minimum Temperature
Pmean = Mean Annual Precipitation
Psd = Standard Deviation of Annual Precipiatation
F = % Forested Area
R = Recharge
A = Area
S = Soils
L = Stream Length
Ss = Stream Slope
Sw = Watershed Slope
Emax = Maximum Elevation
Emin = Minimum Elevation
Emean = Mean Elevation
Erange = Elevation Range
Esd = Standard Deviation of Elevation

elevation variables but especially E_{max} and E_{min}. E_{Min} and E_{max} may be discarded due to their strong correlation to E_{Mean}.

Discussion

The correlation analysis suggests that several variables contain redundant information. T_{Max} and T_{Min} are very similar measure of temperature and should be reduced to a single variable. The same is true for P_{Mean} and P_{SD}, as well as E_{range} and E_{SD}. Of the three variables E_{Max}, E_{Min}, and E_{Mean} should be reduced to E_{mean} because both E_{Max}, and E_{Min} correlate strongly with $E_{mean.}$ The correlation between F and S_W is most likely related to the fact that steep slopes are less likely to be developed for forestation. Along steep slopes, there is generally also a large elevation range, explaining some of the correlation between E_{Range} and F.

Principal Component Analysis

RUN 1

Variables Included

Although redundancy was recognized in the dataset, all variables were used in PCA run 1 to test the hypothesis of the redundancy. Table 7 shows the results of the PCA. Since there are 18 variables, an average or random variable would explain 1/18 or 5.5% of the variation in the data set. The first eigenvector (principal component) in run 1, or PC1, accounted for 42% of the total variation and has the largest eigenvalue (7.16). The second principal component accounted for 20% of the variation with an eigenvalue of 3.39, and the third accounts for 12% of the variation with an eigenvalue of 2.04. The fourth principal component (eigenvalue 1.2) accounted for only 7% of the variation in

Table 7 - Run 1: All Variables

	PC1	PC2	PC3	PC4	PC5	PC6
Eigenvalue	**7.1577**	**3.3923**	**2.0458**	**1.2108**	**0.9268**	**0.7891**
Proportion	**0.421**	**0.2**	**0.12**	**0.071**	**0.055**	**0.046**
Cumulative	**0.421**	**0.621**	**0.741**	**0.812**	**0.867**	**0.913**
	PC7	PC8	PC9	PC10	PC11	PC12
Eigenvalue	**0.6035**	**0.2932**	**0.2031**	**0.1323**	**0.0917**	**0.0845**
Proportion	**0.035**	**0.017**	**0.012**	**0.008**	**0.005**	**0.005**
Cumulative	**0.949**	**0.966**	**0.978**	**0.986**	**0.991**	**0.996**
	PC13	PC14	PC15	PC16	PC17	
Eigenvalue	**0.0318**	**0.0219**	**0.0146**	**0.0005**	**0.0002**	
Proportion	**0.002**	**0.001**	**0.001**	**0**	**0**	
Cumulative	**0.998**	**0.999**	**1**	**1**	**1**	

Variable	PC1	PC2	PC3	PC4
Q	*0.147*	**0.292**	**0.415**	**0.254**
Tmax	**-0.301**	**0.247**	*-0.167*	0.041
Tmin	-0.29	**0.255**	-0.14	**0.252**
Pmean	**0.329**	*-0.136*	-0.051	0.053
Psd	**0.348**	*0.117*	0.019	*-0.157*
F	0.033	**0.275**	**-0.468**	**0.339**
R	**0.246**	-0.19	**-0.215**	*0.196*
A	0.041	**0.321**	**0.511**	*0.168*
L	-0.005	**0.251**	0.086	0.004
Ss	0.099	*0.192*	-0.36	**-0.553**
Sw	0.02	**0.439**	**-0.267**	*0.104*
Emax	**0.364**	0.044	-0.037	0.045
Emin	**0.202**	**-0.287**	*-0.114*	**0.411**
Emean	**0.351**	-0.075	-0.078	*0.159*
Erange	**0.306**	**0.247**	0.035	**-0.227**
Esd	**0.308**	**0.254**	0.055	*-0.152*
S	*0.121*	*0.182*	*-0.137*	**0.271**

the data set. All other eigenvalues were less than one and were considered statistical noise.

Loadings

On PC1, T_{max}, T_{min}, P_{mean}, P_{SD}, R, E_{max}, E_{min}, E_{mean}, E_{mange}, and E_{SD} all had high magnitude loadings. All these variables are associated with climate, even though the association is indirect for elevation variables. PC1 is designated the "climate component".

The remaining variables Q and S also load positively on the climate component. Thus as elevation and precipitation increases annually, discharge increases, and run off potential is high.

The second principal component, PC2, was less straightforward than the first. The variables Q, T_{max}, T_{min}, A, F, L, S_w, E_{range}, and E_{SD} all plot high in a positive direction while E_{min} shows a strong negative loading. This suggests that watersheds which score highly on component two have high annual discharge, high temperature, are highly forested, have large areas, long lengths, steep watershed slopes, large ranges in elevation, and low minimum elevations. The strong negative loading of E_{min} is unusual because the other measures of elevation load less strongly on this component, and in a positive direction.

The variables P_{SD}, S_S, and S have moderate positive loadings on PC2, while P_{mean} and R have moderate negative loadings..

PC3 is less complex than PC2. On this vector, Q and A show strong positive loadings on PC2 and F, S_S, and S_W have strong negative loadings. T_{max}, T_{min}, E_{mean} and S have moderate loadings on PC2. This indicates that watersheds strongly influenced by this component have large discharges, large areas, a high percentage of forested area, a

low stream slope, and a low watershed slope. They also have low temperatures, low minimum elevation, low recharge, and low runoff potential. The high loading of forested area on this vector could be due to the correlation between forested area and watershed slope.

On PC4, Q, T_{min}, F, E_{min} and S had strong positive loadings, whereas S_S and E_{range} have strong negative loadings. Watersheds that are strongly influenced by PC4 have high discharge, high minimum temperatures, high minimum elevation, high potential for run off, and high percentage of forested land. E_{mean}, S_W, R and A had moderate positive loadings, whereas P_{SD} and E_{SD} had moderate negative loadings on PC4. PC4 varied from the first three components in that E_{min} did not load in the same direction as the other four elevation variables.

Interpretation of Components

The variable loadings on PC1 suggest that as temperature decreases, precipitation and recharge increase in areas influenced by PC1 (i.e., with high positive scores on PC1). Also as elevation increases, precipitation increases but temperature decreases. The factor loadings also show that as the range in elevation increases, the standard deviation of the precipitation increases. These statistics are consistent with the contrast in environment between lowland and mountainous settings.

Both minimum temperature and maximum temperature had strong positive loadings on PC2, meaning that annual low temperatures, and annual high temperatures are high in watersheds strongly influenced by PC2. PC2 is also associated with areas that have large watershed slopes. In order to have a large watershed slope, the maximum elevation must be high, the minimum elevation must be low, and the stream length must be short. The negative correlation between E_{min} and the other variables loading on PC2

could be due to the correlation between variables, such as elevation and slope, rather than true geomorphic relationships. Both components two and three are combinations of geomorphic and climate variables, in contrast to Component 1 (elevation and climatic variables).

This PCA analysis confirmed that there is redundancy in this data set. The first component showed a climatic trend where variables like precipitation and temperature tended to have high scores. Many of the climate variables also scored high on other components

RUN 2

Variables Included

The second PCA used the results of PCA run 1 and the correlation analysis to discard several variables. Among temperature variables, T_{min} was removed and T_{max} was retained because it had higher correlation coefficients with P_{mean}, R, E_{max}, E_{min}, and E_{mean}. For precipitation variables, P_{mean} was retained and P_{SD} deleted due to its higher correlation with E_{max}, E_{range}, and E_{SD}. For elevation variables, E_{SD} was discarded and E_{range} was retained because the former had lower correlation with R P_{mean} and T_{max}. E_{mean}, was retained due to its higher correlation with E_{max} and E_{min}. E_{max} and E_{min} were also discarded. The variables included in Run 2 were Q, T_{max}, P_{mean}, R, F, A, L, S_S, S_W, E_{mean}, E_{range}, and S (Table 8).

Eigenvalues

The first component had an eigenvalue of 3.86, the second 2.74, the third 1.92 and the fourth 1.01. All other components had an eigenvalue of less than one and were not

Table 8 - PCA Run 2

	PC1	PC2	PC3	PC4	PC5	PC6
Eigenvalue	**3.8671**	**2.7486**	**1.9285**	**1.0096**	**0.7784**	**0.7117**
Proportion	**0.322**	**0.229**	**0.161**	**0.084**	**0.065**	**0.059**
Cumulative	**0.322**	**0.551**	**0.712**	**0.796**	**0.861**	**0.92**
	PC7	PC8	PC9	PC10	PC11	PC12
Eigenvalue	**0.42**	**0.1782**	**0.1697**	**0.0788**	**0.0612**	**0.0482**
Proportion	**0.035**	**0.015**	**0.014**	**0.007**	**0.005**	**0.004**
Cumulative	**0.955**	**0.97**	**0.984**	**0.991**	**0.996**	**1**

Variable	PC1	PC2	PC3	PC4
Q	*0.19*	**0.348**	**-0.462**	*-0.126*
Tmax	**-0.414**	**0.254**	**0.208**	*-0.107*
Tmin	**0.464**	*-0.129*	0.001	*0.095*
R	**0.384**	*-0.196*	*0.172*	*-0.164*
F	0.084	**0.331**	**0.442**	**-0.333**
A	-0.002	**0.405**	**-0.505**	-0.056
L	-0.012	**0.303**	-0.084	**0.559**
Ss	*0.135*	*0.197*	**0.411**	**0.514**
Sw	0.019	**0.489**	**0.281**	-0.031
Emean	**0.47**	-0.048	0.019	-0.009
Erange	**0.39**	**0.254**	-0.04	*0.131*
S	*0.175*	**0.22**	0.082	**-0.477**

considered. In combination, the first four components account for 79.6% of the variation in this data set.

Loadings

The variables that had high loadings on the PC1 (Table 8) were P_{mean}, R, E_{mean}, E_{range}, and T_{max}. S_S, Q, and S showed moderate loadings on PC1. Watersheds that scored strongly on this component are areas that have high precipitation, high groundwater recharge, high elevations, and cool temperatures. This is clearly a climatic component

Similar to Run 1, variables with high factor loadings on PC2 include climatic and geomorphic variables. Q, T_{max}, F, A, L, S_W, and E_{range} all have a high positive factor loading on this component. Watersheds influenced by this component have high discharges, high temperatures, large areas, long stream lengths, steep topography, and extreme relief. They are also largely forested. The two climatic variables that load on PC2 are T_{max} and E_{range}

PC3 has a strong geomorphic character, with F, S_S, S_W, Q, and A load strongly. None of the climatic variables load significantly on this variable. Watersheds scoring on PC2 would tend to be small, very steep, and highly forested. The relationships between the high factor loadings for this component make sense. This is clearly a “geomorphic” component.

The variables with high loadings on PC4 are L, S_S, F, and S. Similar to PC3, none of the climate variables load on this component. While explaining little (8.4%) variation in the dataset PC4 could, however, influence a small number of watersheds. These are steep and dendritic but with little forestation.

Interpretation of Components

Similar to Run 1, PC1 in Run 2 is a climatic component. PC2, PC3, and PC4 are also a combination of geomorphic and climatic variables, though more clearly expressed in Run 2 than Run 1. E_{range}, an indirect climate variable, could be showing up on PC2 because it is inherently correlated with S_W.

RUN 3

Variables Included

To further clarify the components, more variables were removed for Run 3. The correlation between area and all of the other variables affecting stream flow is very strong. Haan and Reed (1970) found area to be the most important factor for predicting stream flow. To remove such intercorrelation, A was removed and Q and other variables were normalized with respect to area. Instead of using L/A, a basin shape factor, L/W, was employed.

Eigenvalues

Run 3 yielded only three important components. The first 3 components had eigenvalues of 3.91, 2.37, and 1.17 respectively for a total of 75% of variance. The other seven components had Eigenvalues of less than one therefore were not considered.

Loadings

PC1 had high loadings for all climate variables as in all of the other runs (Table 9). However there were positive loadings for P_{mean}, R, E_{mean}, E_{range} and negative loadings for T_{max}. PC1 Run 3 differs slightly from Runs 1 and 2 in that T_{max} has a positive loading. Watersheds that score strongly on PC1 area at high elevation, have low temperatures, and receive large amounts of precipitation.

Table 9 - PCA Run 3

	PC1	PC2	PC3	PC4	PC5	PC6
Eigenvalue	**4.1598**	**2.3907**	**1.4497**	**0.8967**	**0.6654**	**0.6022**
Proportion	**0.378**	**0.217**	**0.132**	**0.082**	**0.06**	**0.055**
Cumulative	**0.378**	**0.596**	**0.727**	**0.809**	**0.869**	**0.924**
	PC7	PC8	PC9	PC10	PC11	
Eigenvalue	**0.4067**	**0.1627**	**0.1385**	**0.0776**	**0.05**	
Proportion	**0.037**	**0.015**	**0.013**	**0.007**	**0.005**	
Cumulative	**0.961**	**0.976**	**0.988**	**0.995**	**1**	

Variable	PC1	PC2	PC3
G	**0.272**	-0.1	**-0.516**
Tmax	**-0.382**	**-0.359**	*-0.103*
Pmean	**0.448**	*0.144*	-0.077
R	**0.384**	*0.111*	**-0.247**
F	*0.11*	**-0.522**	**-0.262**
Ss	*0.15*	**-0.348**	**0.269**
Sw	0.028	**-0.581**	0.003
Emean	**0.447**	0.061	-0.051
Erange	**0.368**	*-0.192*	**0.375**
S	*0.157*	**-0.232**	-0.083
L/W	*0.193*	-0.058	**0.604**

PC2 is, as for Runs 1 and 2, a mixture of climatic and geomorphic variables. Watersheds scoring on PC2 are above average temperature, forested and steep. Since E_{mean} had a low loading, PC2 watersheds may occur at either high or low elevation but have extreme relief, little precipitation or recharge.

Interpretation of Components

The component scores from Run 3 of the individual watersheds on PC1 and PC2 are plotted against normalized discharge in Figures 10&11. In PC1 there are two groups. The first group has a low normalized discharge and the second has high normalized discharge. There is a somewhat linear trend (with many outliers) between 27 and 40 inches per year. PC2 does not, however, have a linear trend, but does have two groups, one with high discharge and the other with low discharge.

Run 3 still showed that the strongest component is the climate component and the second two components are still mixtures of climatic and geomorphic variables. There should be a linear relationship between component one and discharge and between component two and discharge. In order to clarify the signal more variables were removed with the hope of removing redundancy.

RUN 4

Variables Included

In Run 4, additional variable elimination was undertaken. S was removed from Run 4 due to ambiguous loadings in Runs 1-3. Of the slope variables S_w and $S_{S,}$ S_S was discarded because the data collection methods for this variable were less uniform than for S_w, and its loadings were more ambiguous. Of the very similar variables E_{range} and $W_{s,}$

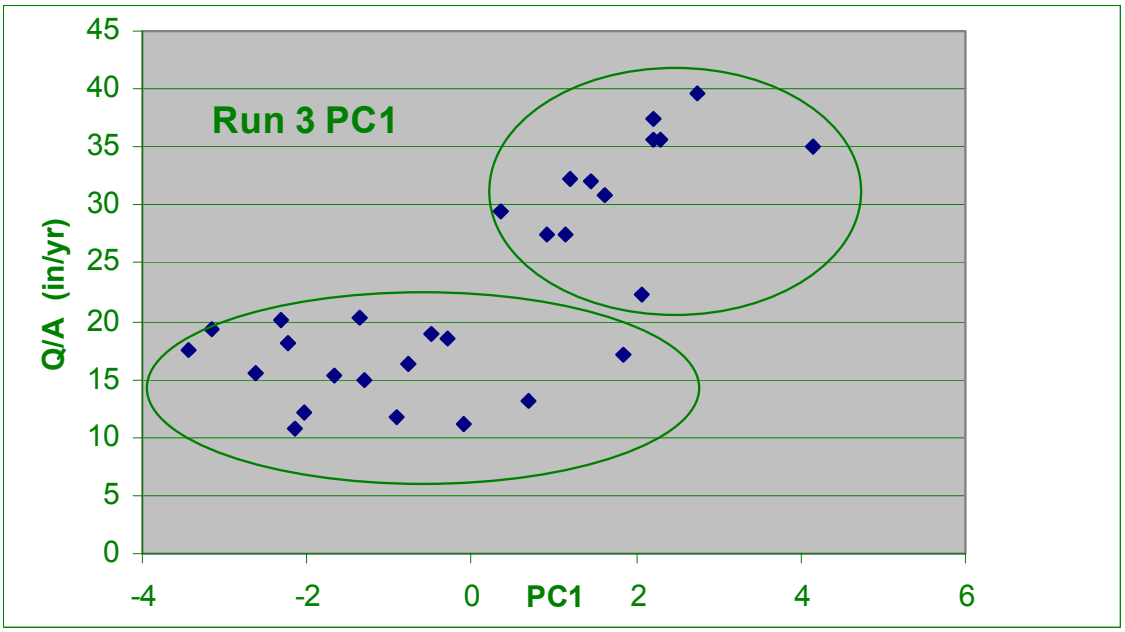


Figure 10 – Run 3 PC1 v. Discharge

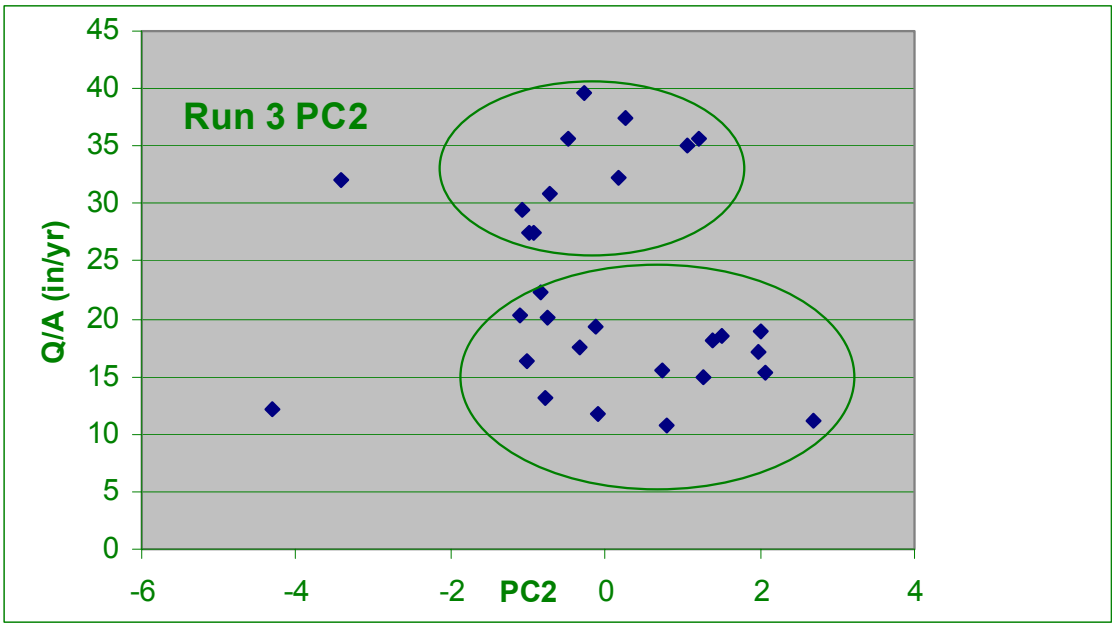


Figure 11 – Run 3 PC2 v. Discharge

E_{range} was discarded because it loaded ambiguously. Finally, G was removed because it also had ambiguous loadings. This leaves the variables T_{max}, P_{mean}, R, E_{mean}, S_w, F, and L/W

Eigenvalues

PC1 and PC2 had eigenvalues of 3.36 and 1.81 respectively. PC3 had an eigenvalue of 0.93 but was retained because it was the only component L/W loaded on.

Loadings

PC1 in Run 4 is again the climate component with only P_{mean}, R, E_{mean}, and T_{max} loading strongly (Table 10). This component reflects climate but not geomorphic variables. PC2 is a geomorphic component, with only S_w and F loading strongly. PC3 had a high loading only for L/W. Despite the low eigenvalue, no other variables load on this component suggesting that the component has minor significance.

Interpretation of Components

Watershed scores (Figure 12) for PC1 vs. normalized discharge show a clear linear trend. There are still two groupings of watersheds with respect to Q/A, as in Run 3, but separation between the two groups is more distinct. Figure 13 shows watershed scores on PC2, plotted against Q/A, also showing these two distinct groups. In the high discharge group (Group 1) there is a clear linear relationship between PC2 and Q/A. The low discharge group (Group 2) shows no clear a linear relationship between PC2 and Q/A.

Cluster Analysis

Run 3 and Run 4 showed two groupings of watersheds for PC1 and PC2. One group had high relative discharge (Group 1) and the other had low relative (Group 2) discharge. The relationship between discharge and both PC1 and PC2 is linear for the

Table 10 - PCA Run 4

	PC1	PC2	PC3	PC4	PC5	PC6
Eigenvalue	**3.3652**	**1.8188**	**0.9297**	**0.5297**	**0.1696**	**0.104**
Proportion	**0.481**	**0.26**	**0.133**	**0.076**	**0.024**	**0.015**
Cumulative	**0.481**	**0.741**	**0.873**	**0.949**	**0.973**	**0.988**

	PC7
Eigenvalue	**0.0829**
Proportion	**0.012**
Cumulative	**1**

Variable	PC1	PC2	PC3
Tmax	**-0.506**	0.186	-0.058
Pmean	**0.506**	0.069	-0.124
R	**0.447**	0.136	-0.186
F	-0.024	**0.684**	-0.101
Sw	-0.14	**0.656**	0.039
Emean	**0.484**	0.177	-0.116
L/W	0.183	0.113	**0.96**

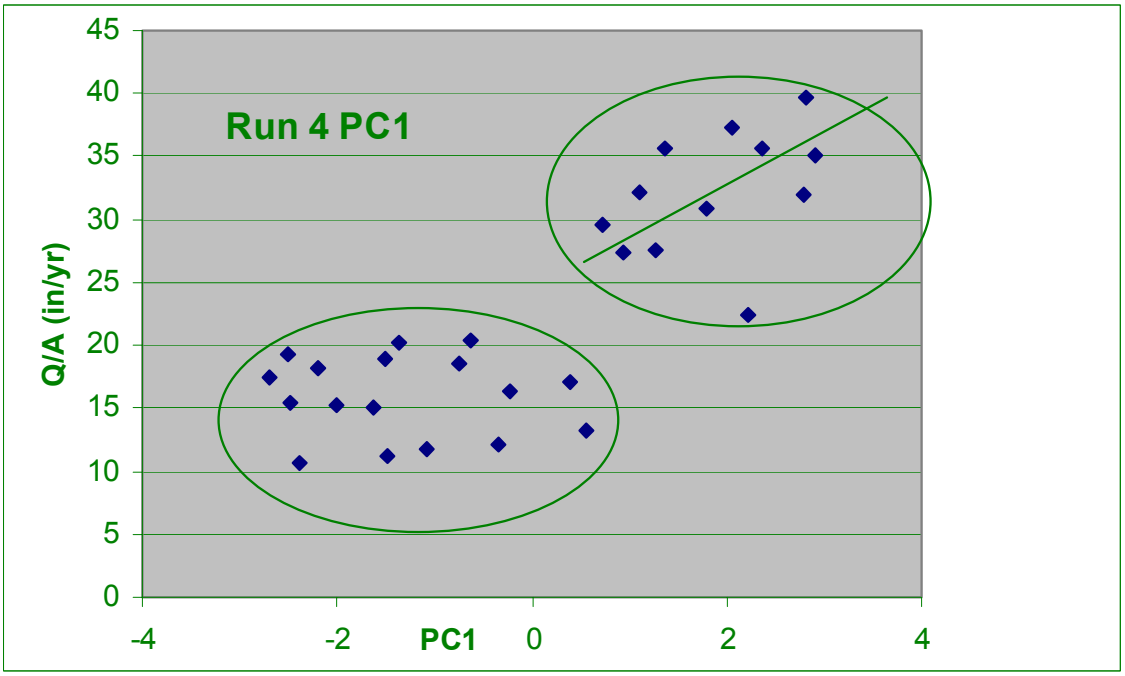
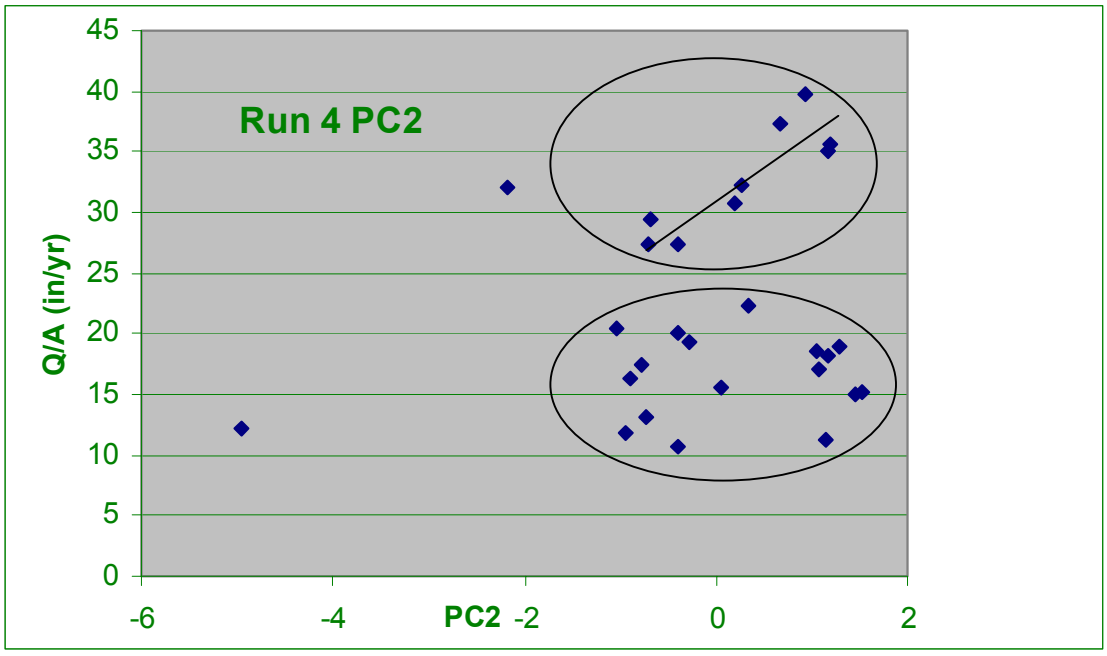


Figure 12 – Run 4 PC1 v. Discharge

Figure 13 – Run 4 PC2 v. Discharge



group 1. The relationship between the Group 2 and Q/A is not linear. In order to clarify this a cluster analysis was performed (Figure 14).

Run 4 PC1

The cluster analysis was performed using single linkage and the Euclidean distance method on PC1. The first level of clusters, 82% similarity, are the same as the groupings in the plot of PC1 versus discharge. Table 11 shows that elevation as well as discharge separates the groups. The group high Q/A watersheds also had higher E_{mean}. The Greenbrier River is the last watershed to join indicating it to be an outlier.

Further interpretation of the cluster analysis shows the high discharge group splits into two clusters at approximately 89.79% similarity (Figure 15). Cluster 1 has a slightly lower P_{mean}, E_{mean}, S_w, R, F, and L/W, than Cluster 2. The low discharge group also splits into two clusters, clusters 3 and 4. Cluster 3 has higher P_{mean}, R and E_{mean} than cluster four, but cluster four has the higher values for T_{max}, Sw, and F.

Run 4 PC2

A cluster analysis was also performed on PC2 from Run 4 (Fig 16). The observations in each cluster for PC2 are the same as the observations in each cluster for PC1 with the exception of Opequon Creek. This watershed is an outlier on PC2 but not PC1. The only other difference between the results of this cluster analysis and PC1 lies in the order of the linkages and the levels of similarity. The watersheds join the clusters in a different order in PC2 and have slightly higher percent similarities.

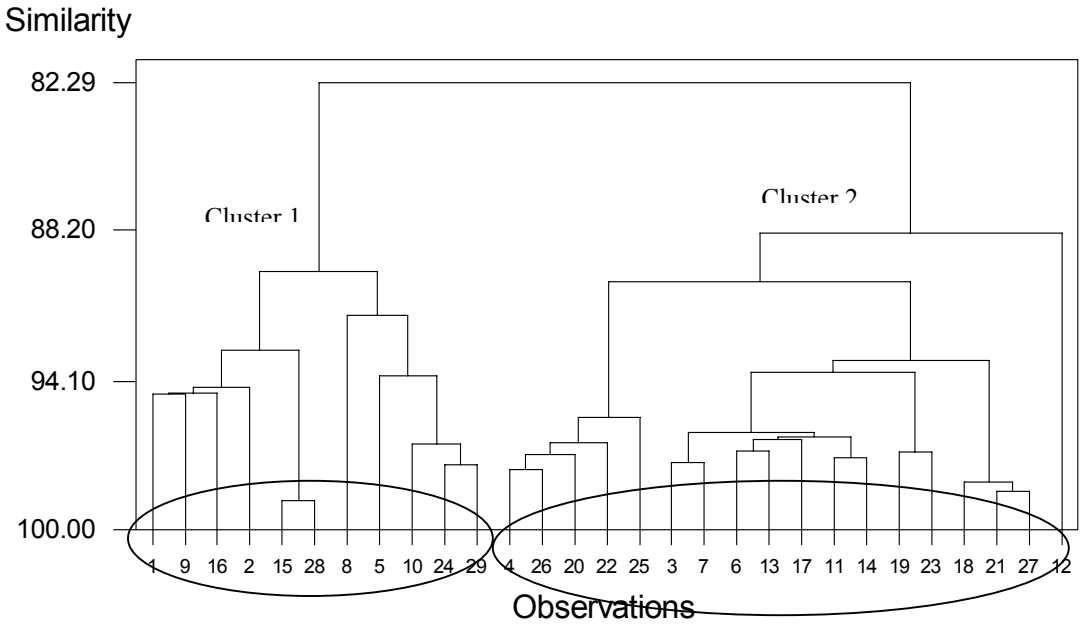


Figure 14 – Dendrogram of Cluster analysis, PC1 Level one clusters
Figure 15 – Dendrogram of Cluster Analysis, PC2 Level 2 clusters

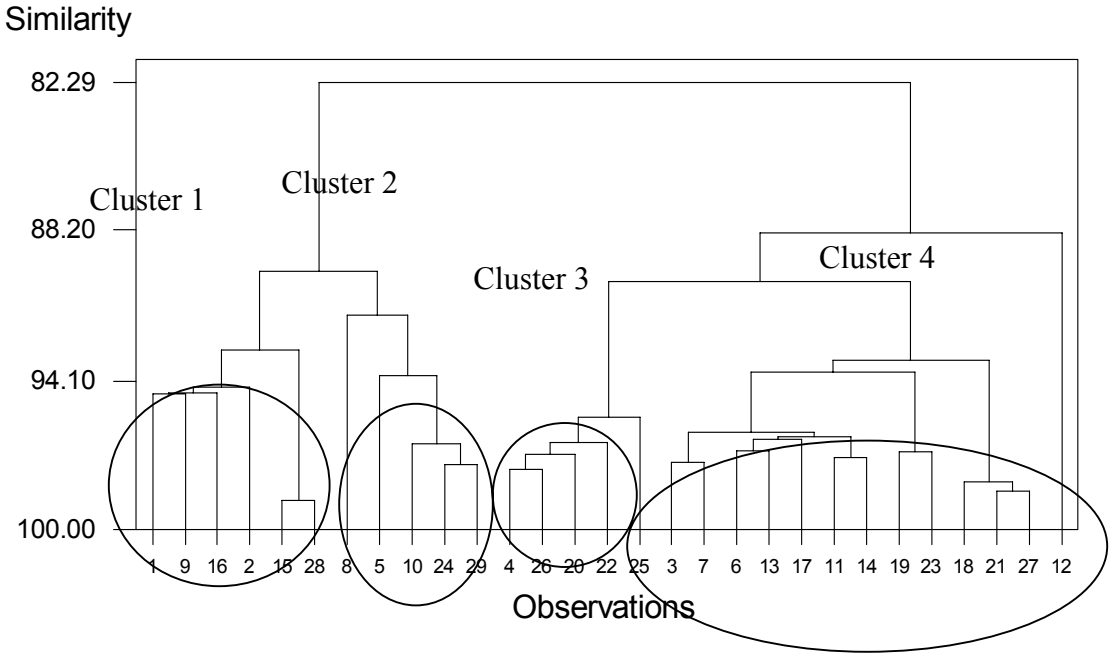


Table 11 Summary of Cluster Analysis

		ID	Tmax	Pmean	R	Sw	Emean	F	G	L/W
	Blackwater	1	53.9	53.97	22.5	6.08	1133	71	3.34	0.470477
	Buckhannon	2	58.3	47.99	21.4	11.5	805	81	1.13	3.667036
Cluster 1	Dryfork	9	55.8	52.45	21.4	14.24	989	89	0.4	2.725922
	MiddleFork	16	58.2	48.53	24.5	12.23	846	94	3.34	4.862353
	Meadow	15	58.8	50.99	20.6	11.99	851	84	3.34	2.750569
	Tygart	28	56.4	49.98	15.4	13.63	985	74	0.4	5.781449
	Average		56.9	50.6517	20.97	11.612	934.8	82.167	1.992	3.376301
	Cranberry	8	56.3	55.27	31.6	14	1025	97	3.34	3.610742
	Cherry	5	57.5	54.12	27.8	13.35	1005	97	3.34	0.940284
Cluster 2	Elk	10	57.5	51.85	23.9	17.57	896	95	3.34	3.538898
	Williams	29	56.5	54.88	26.4	14.52	1049	97	3.34	2.728533
	ShaversFork	24	55.7	54.19	24.8	14.47	933	96	3.34	14.98655
	Average		56.7	54.062	26.9	14.782	981.6	96.4	3.34	5.161
	Cacapon	4	59.8	44.3	8.7	10.78	584	85	1.26	4.200261
	PattersonCr	22	62.4	42	7.3	10.84	355	97	0.53	2.799697
Cluster 3	Sobrpot	25	59.2	48.92	11.6	13.11	953	73	0.53	7.761568
	SfkSBrPot	26	64.6	47.22	10	19.27	794	90	1.26	2.298945
	Average		61.5	47.3004	12.9	13.5	671.5	86.25	0.895	4.444294
	Buffallo	3	61.6	42.35	21.4	13.39	394	85	1.13	2.18449
	Cobun	7	60.5	43.76	21.4	9.86	473	82	1.13	4.221172
	MiddleIslandCr	17	62.3	42.11	8	15.01	346	87	1.13	1.902307
	Coal	6	62.6	46.46	11.9	20.4	616	93	3.34	2.226349
	EastFork	11	64.9	44.18	12.4	17.72	384	97	1.13	5.814969
	Guyandotte	13	61.9	48.63	14.5	18.12	716	94	1.13	1.623974
Cluster 4	NfSBrPot	19	57.7	49.63	11	20.13	900	88	0.53	6.380927
	Hughes	14	63	42.24	7.1	13.4	305	83	1.13	1.385285
	PineyCr	23	61.1	49.84	11.9	11.14	838	80	3.34	0.499358
	Mud	18	64.7	43.44	11.9	14.35	340	91	1.13	1.757326
	Panther	21	64.6	46.88	11.1	19.7	521	99	3.34	1.67792
	TugFork	27	63.1	46.1	11.3	20.35	662	96	3.34	1.365009
	Average		63.37	45.7	10.66	16.131	541.3	89.583	1.817	1.33698
	Opequon	20	57.7	49.63	9.8	4.1	304	36	0.046	2.700058
	Greenbrier	12	56.3	55.27	21.1	14.44	1059	88	3.34	3.683631

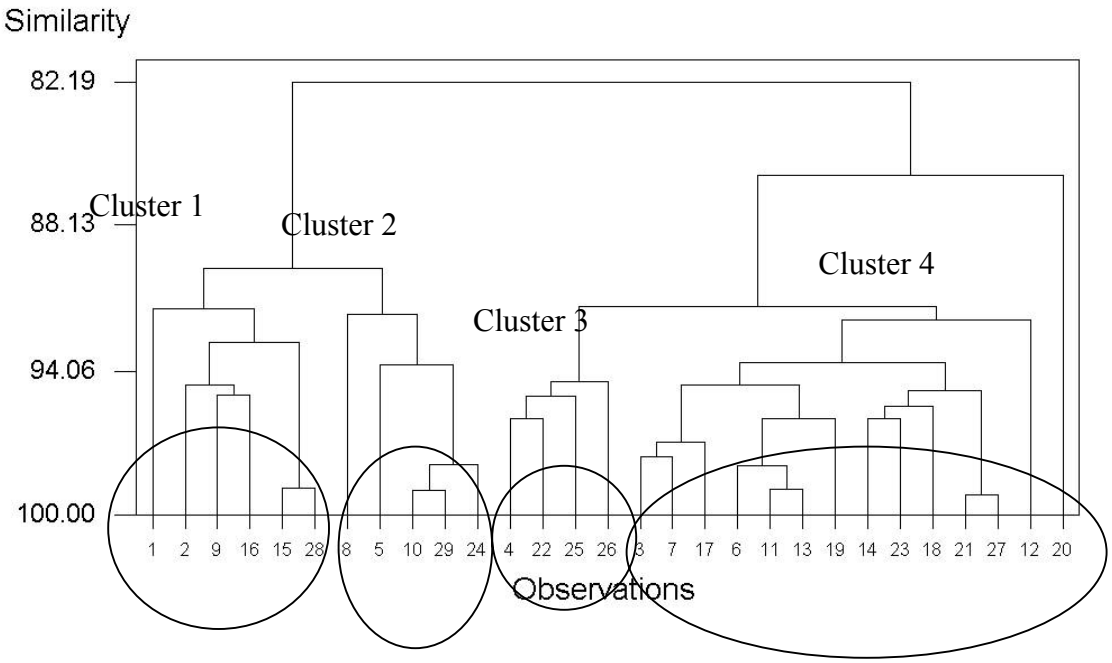


Figure 16 - Dendrogram of Cluster Analysis; PC2 Level 2 clusters

Interpretations

The variables used in this study do not relate to one another in a uniform manner across the study area. Watersheds with large relative discharges have different relationships between climatic variables and geomorphic variables than watersheds with small relative discharges. The similarity between the cluster analysis for PC1 and PC2 indicates that the climate component is so strong, that it influences the geomorphic component.

Spatial Analysis

Watershed Loadings

In order to help visualize the relationships shown above, spatial figures were created. Figure 17 shows watershed loadings from Run 4 on PC1. There is a spatial relationship on PC1, however, it is not very clear. The majority of watersheds with high scores on component one are along the mountainous region along the eastern edge of the state. The watersheds with low scores on the first component are mostly in the Eastern Panhandle and the Allegheny Plateau. Figure 18 shows watershed loadings on PC2. There is no clear spatial pattern to PC2.

Cluster Analysis

Similar figures were generated for the results of the cluster analysis for PC1 and PC2. Figure 19 shows PC1 grouped into four clusters. There is a clear spatial relationship shown in this figure. Cluster 3 watersheds are located in the Valley and Ridge physiographic province, eastern panhandle, cluster 4 watersheds are located in the

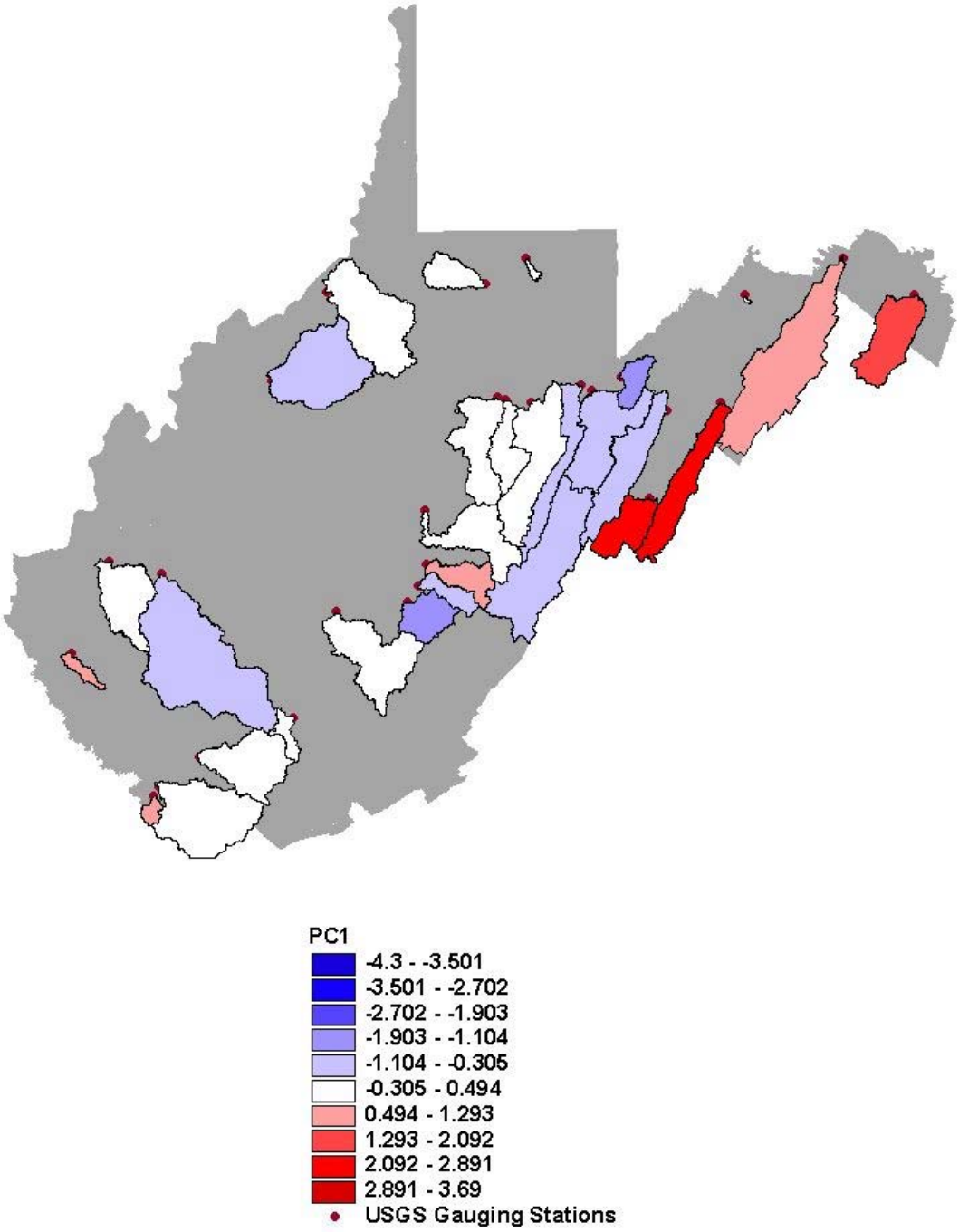


Figure 17 – Watershed scores on run 4 PC1

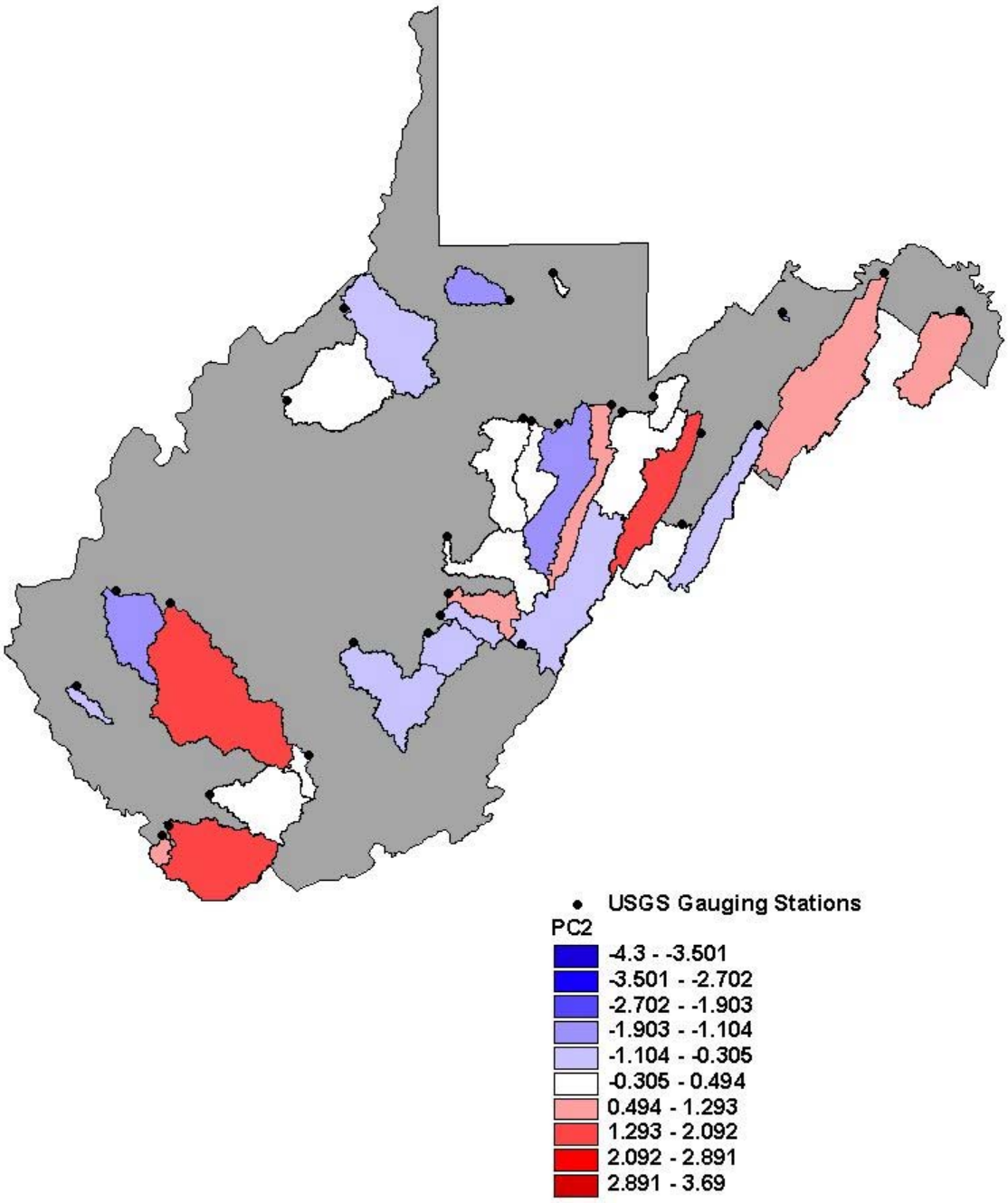


Figure 18 – Watershed scores on run 4 PC2

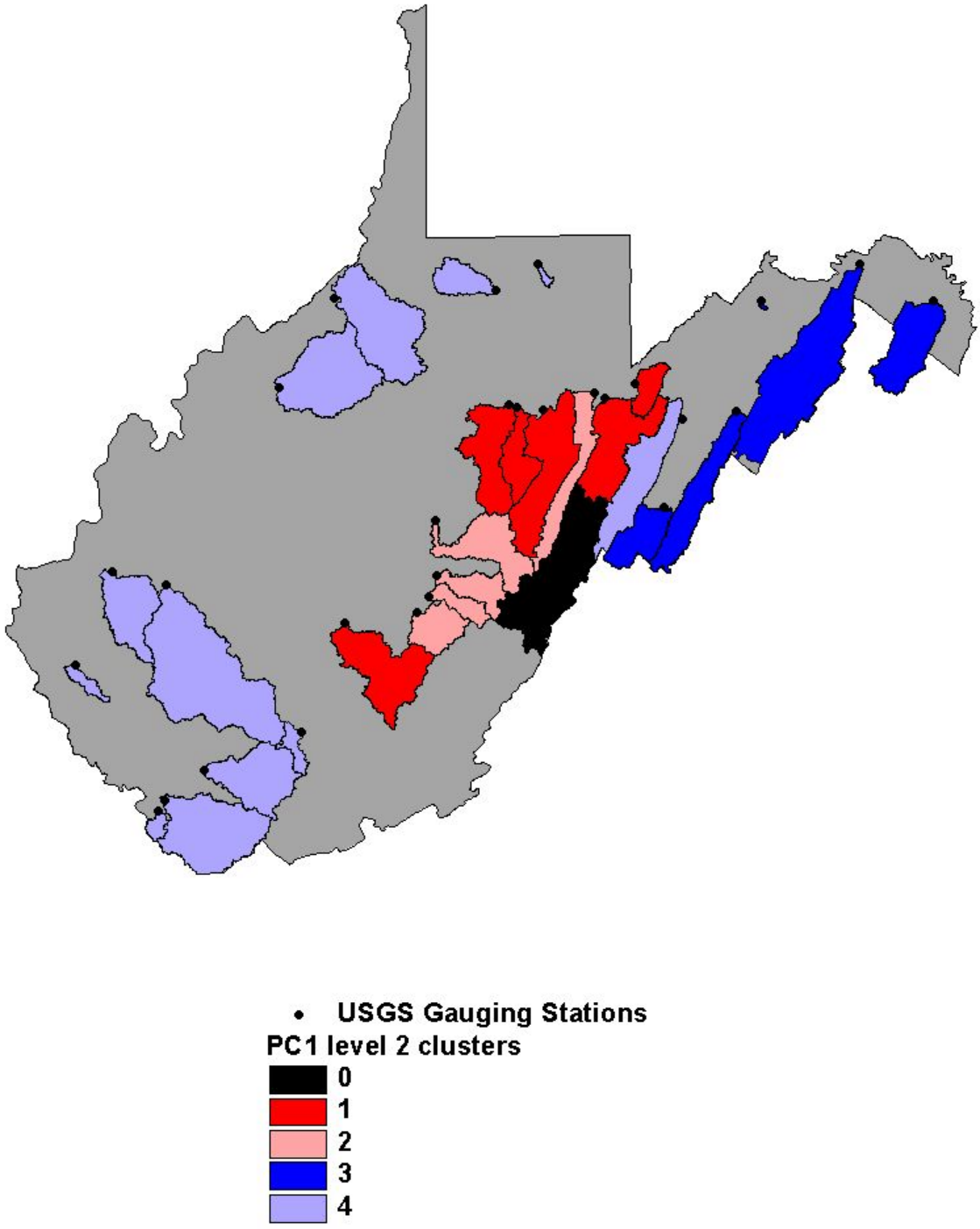


Figure 19 – Cluster Analysis: PC1 Level 2 Clusters

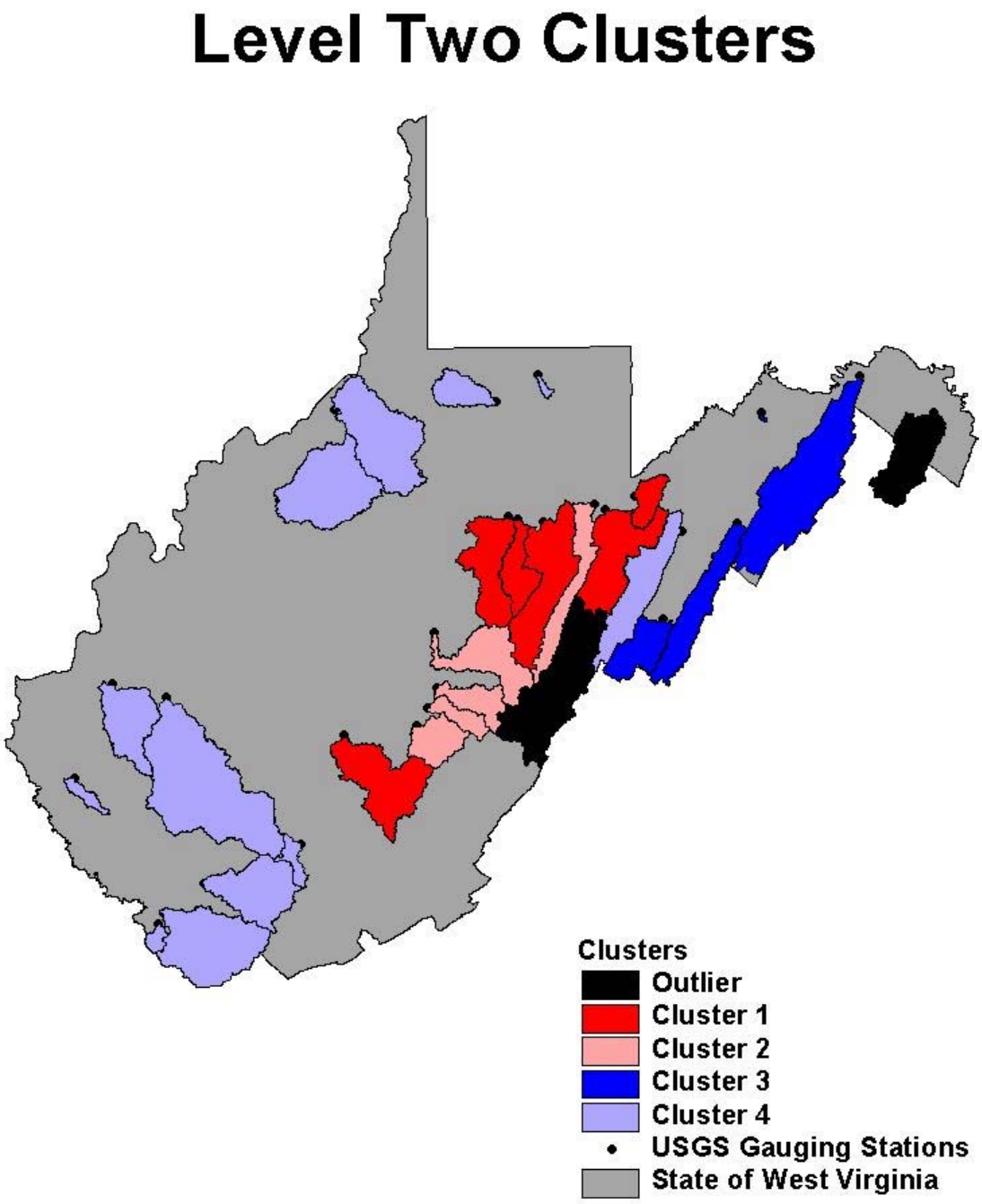


Figure 20 – Cluster Analysis: Level 2 Clusters, PC2

Allegheny Plateau, Cluster 1 and 2 watersheds are located along the Allegheny Front. One exception to this is the North Fork of the South Branch of the Potomac River, which is in Cluster 4. This watershed lies on the boundary between the Allegheny Front and the Valley and Ridge. It is also in the rain shadow of the Allegheny Front. These factors are likely causing this watershed to be an intermediate watershed and not cluster with the other Valley and Ridge watersheds. Clusters 1 and 2 are all located along the Allegheny Front. Cluster 2 is at high elevation and cluster 1 is at lower elevation around Cluster 2. Grouping PC2 into four clusters (Figure 20) yields very similar results. The one difference is Opequon Creek. This watershed is part of cluster 3 in PC1 but is an outlier in PC2.

Interpretations

Watersheds that are located along the Allegheny Front have a linear relationship between PC1 and PC2 and Q/A. All other watersheds in this study do not. These watersheds also score high on PC1. This area is split into two groups, a group at high elevation and a group with slightly lower elevation but still in the Allegheny front. Watersheds in the valley and ridge and Allegheny Plateau do not have a linear relationship between PC1 or PC2 and Q/A. These watersheds have different relationships between geomorphic variables and Q/A than the other watersheds. Clusters 3 and 4 show that the relationships between these variables are not the same in the Allegheny Plateau and the ridge and valley regions of the state. The similarity between the cluster analysis for PC1 and PC2 suggests that in this state the climate variables are dominant. Possibly to the extent of driving the geomorphology.

Conclusions

The results of the statistical analysis were as follows.

- the optimum combination of variables for predicting stream flow in West Virginia were T_{max}, P_{mean}, R, E_{mean}, S_w, F, and L/W;
- variables for geology and soils were no load variables;
- drainage basin are skews the importance of other variables when included in the dataset;
- there are three major components of streamflow in West Virginia: climate, geomorphology, and basin aspect ratio;
- the relationship between the PC1 and PC2, and Q/A differs in the Allegheny Plateau, Allegheny Front, and Valley and Ridge physiographic provinces; and
- the climate component (PC1) is so strong that it may drive the geomorphology (PC2) in the state.

The results of this study indicate that the optimum combination of variables for predicting stream flow in West Virginia is T_{max}, P_{mean}, R, E_{mean}, S_w, F, and L/W. The geology and soils data used in this study were found to add noise to the dataset rather than explain variance. These variables are "no load" variables because they did not load distinctively on any one component. This could be due to poor or inconsistent data. These conclusions are based on discharge that is normalized to area. Drainage basin area is such a strong factor and is highly correlated with so many of the other characteristics of watersheds that influence stream flow that it skews the importance of the other variables.

Using the above variables, there are three major components of stream flow as expressed in PCA run 4: climate (PC1), geomorphology (PC2), and basin aspect ratio (PC3). PC3 has very little importance to explaining the variance in the data set. PC1 however, is very strong and explains 48% of the variation in the data set on its own. The

importance of PC1 is reinforced by the results of the spatial analysis. The watershed scores on PC1 had a slight spatial relationship while the watershed scores on PC2 did not have a spatial relationship. When looking at the results of the cluster analysis for PC1 and PC2, there are three groups of watersheds in West Virginia. Clusters 1 and 2 have high Q/A and are located along the Allegheny front similar to the spatial pattern in the spatial analysis of the watershed scores.

As previously discussed there are four groupings of watersheds based on the cluster analysis of PC2. These groupings fall within the three physiographic provinces in the state: Allegheny Plateau, Allegheny Front, and Valley and Ridge. The cluster analysis in this study showed that the variables affecting stream flow in these three regions do not interact in a uniform manner. Thus, a single equation for stream flow in the state may not be practical. The watersheds in the Allegheny Front and Valley and Ridge regions have clear spatial relationships. The Allegheny Front is not as distinct. This region has two sub-regions and intermediate watersheds along its eastern side. Because of the variability in this area further study should be done to clarify these relationships.

These conclusions are based on several limitations. First, the current available statewide data are mapped at very small scale. Larger scale data would be more accurate and might remove noise from variables, particularly, geology and soils. Another limitation was that realistic recharge estimates were not available for every watershed. Several watersheds used regional-average numbers instead of locally measured estimates. Finally, evapotranspiration was not included in this study. Evapotranspiration can occur above or below the ground making it difficult to measure. The inclusion of R and LULC are both variables that influence evapotranspiration, but are thought to be very imprecise

approximations, as confirmed by the PCA. The effects of underground and surface mining were also not looked at in this study. Surface mining is widespread in the southern part of the state and could impact streamflow. Similarly, underground mining is widespread throughout the entire state and can impact streamflow when flooded mines discharge to streams or when streams loose water to the mines.

Future studies should include larger scale studies in each of the three groups of watersheds found in the spatial analysis. Because the areas are smaller, better data may be available. Individual studies on these areas may also better define the differences in the relationships between the variables. Other future studies should include studies using precipitation and temperature data at a seasonal time scale. Seasonal variations in streamflow are very strong in West Virginia and need to be included in streamflow prediction models.

Discussion

Regulatory agencies in West Virginia use stream flow prediction models for various things including protecting drinking water sources (West Virginia Department of Health and Human Resources, 1999). Currently a single prediction equation is used for the entire state. This study found that one stream flow prediction equation is not adequate for the state of West Virginia. There are three regions of similar streamflow characteristics within the state. These regions are the Allegheny Plateau, the Valley and Ridge, and the Allegheny Front areas. Separate streamflow equations should be developed based on large-scale studies for each of these regions. If only one prediction equation can be used, the variables were T_{max}, P_{mean}, R, E_{mean}, S_w, F, and L/W should be used as the independent variables while Q/A should be used as the dependant variable. Q/A should be used rather than Q, because area is so closely correlated to streamflow, that it is impossible to have independent variables when it is included.

Several variables looked at in this study, such as S and G, were discarded because they had ambiguous loadings in the PCA. Most likely this is due to the quality of the data. The current available data for S and G is at very small scale and uses a nominal classification. Larger scale data and better quantification of the data may show that these variables play an important roll in estimating streamflow. An example of this is the two outliers Opequon Creek and the Greenbrier River. These are the only two watersheds included in the study that are limestone-dominated watersheds. It is possible that better geology data could show the reason these watersheds are outliers is due to the karst geology. Another limitation in this study is the impact of mining was not included in this study although it can play a large roll in streamflow. Streams can gain or loose water

from underground mines. Surface mines also impact streamflow by changing vegetation and runoff patterns.

Of the three streamflow regions, the Allegheny Front is the most complex. There are two sub-regions within the region; high elevation, and low elevation. Watersheds along the eastern fringe of the Allegheny Front are intermediate watersheds and do not group with either the Front or the Valley and Ridge. The variability in area could be better understood from a larger scale study performed on the Allegheny Front region.

References Cited

Benson, M. A., 1962. Factors Influencing the Occurrence of Floods in a Humid Region of Diverse Terrain, *United States Geological Survey Water Supply Paper 1580-b*: B1 – B64.

Benson, M. A., 1964. Factors Affecting the Occurrences of Floods in the Southwest, *United States Geological Survey Water Supply Paper 1580-d*: D1 – D72.

Brackensiek, D. L., 1961. "Estimating dependable annual streamflow in the unglaciated allegheny plateau", Agricultural Research Services, Vol. 41: 1-34.

Davis, Bruce, 1996. *GIS A Visual Approach*, Albany, New York: On Word Press.

Davis, John, C., 1973. *Statistics and Data Analysis in Geology*. New York: John Wiley & Sons.

Diaz, Guillernmo, Sewell, J. I., and Shelton, C. H., 1968. "An application of principal component analysis and factor analysis in the study of water yield*". Water Resources Research* 4 (2): 299-305.

Dunne and Leopold, 1978. *Water in Environmental Planning*. San Francisco: W. H. Freeman and Company

Garren, D.C., 1992. "Improved techniques in regression-based stream flow volume forecasting". *Journal of Water Resources Planning and Management* 118 (6): 654-669.

Haan, Charles T., 1977. *Statistical Methods in Hydrology*. Ames, Iowa: Iowa State University Press.

Haan, C. T., and Allen, David, M., 1972. "Comparison of multiple regression and principal component regression for predicting water yields in Kentucky*". Water Resources Research*: 8 (6) 1593-1596.

Hann, Charles T., and Read, H. R., 1970. "Prediction of monthly, seasonal and annual Runoff Volumes for small agricultural watersheds in Kentucky", Bulletin 711, Kentucky Agricultural Experiment Station, Lexington Kentucky: University of Kentucky.

Jenson, S., K., and J., O., Domingue, 1988. "Extracting Topographic Structure from Digital Elevation Data for Geographic Information Systems Analysis". *Photogrammetric Engineering and Remote Sensing* 54 (11): 1593-1600.

Keller, Edward, A., 1988. *Environmental Geology*; Seventh Edition, Upper Saddle, New Jersey: Prentice Hall.

Kozar, Mark, D., and Mathes, Melvin, V., 2001. Aquifer-characteristics data for West Virginia, U.S. Geological Survey Water-Resources Investigations Report 01-4036.

McCuen, Richard, H., and Snyder, Willard, M., 1986. *Hydrologic Modeling: Statistical Methods and Applications*. Englewood Cliffs, New Jersey: Prentice-Hall.

National Cooperative Soil Survey, 1976. *Soil Survey of Kanawha County, West Virginia*.

Natural Resources Analysis Center, 2001. *Watershed Characterization and Modeling System Version 2.8 Users Manual*, Morgantown, West Virginia: West Virginia University Natural Resources Analysis Center.

Perez, Albert, 2000. Source water protection project: a comparison of watershed delineation methods in ARC/INFO and ArcView GIS, in hydrologic and hydraulic modeling support: with Geographic Information Systems. In *Hydrologic and Hydraulic Modeling Support with Geographic Information System*s. Redlands, California: ESRI Press, p. 53-64

Rorabaugh, M.I., 1964. Estimating changes in bank storage and ground-water contribution to streamflow. *International Association of Scientific Hydrology Publication 63*, p. 352-363.

Spatial Analytics, LLC, 2000. Surface Water Protection System : Final Technical Report, Spatial Analytics, 40 p.

Trentwartha, Glen, Thomas, 1980. *An Introduction to Climate*, New York: McGraw-Hill, Inc, 416 p.

West Virginia Department of Health and Human Resources, 1999. State of West Virginia Source Water Assessment and Protection Program. West Virginia Department of Health and Human Resources, Bureau for Public Health, Office of Environmental Engineering Division, 60p.