BASIN-WIDE ANNUAL BASEFLOW ANALYSIS FOR THE FRACTURED BEDROCK UNIT IN THE POTOMAC RIVER BASIN

Nebiyu Tiruneh, Ph.D.

Interstate Commission on the Potomac River Basin

June 2007

ICPRB Report No. 07-6

Table of Contents

Tables

Figures

Executive Summary

Estimation of annual baseflow statistics for a stream, or a network of streams for a watershed, is an alternative method used in determining water availability over a specified period of time for water resources management or watershed management purposes. Annual stream baseflow provides an estimate of annual recharge to the aquifers which discharge to the stream. Studying the availability of both surface water and groundwater resources of a basin provides the tools necessary to develop a sustainable water resources management plan and very helpful input for decision support systems.

The Potomac River Basin covers an area of approximately 14,600 square miles with interstate sub-basins in Maryland, Virginia, West Virginia, Pennsylvania, and the District of Columbia. The basin has very distinct geologic, geomorphic, hydrological and landuse/landcover characteristics that vary across the region. Both surface water and groundwater resources of the basin support about 5.3 million people according to the 2000 census. Increasing demand for water associated with population growth and weather anomalies that result in drought may strain the water resources of the basin.

The objective of the study is to determine water availability in the fractured bedrock units of the Potomac River Basin by analyzing annual baseflow statistics and developing statistical models. Baseflow magnitudes were estimated for continuous and partial record stream flow gages operated by the U.S. Geological Survey and the results were then used in the development of statistical equations that help determine magnitudes of baseflow in ungaged watersheds.

The statistical equations relate baseflow magnitudes in watersheds to hydrological parameters and basin characteristics. The main basin characteristics investigated in this study include drainage area, geology, soil type, landuse, hydrogeomorphology (HGMR), precipitation, and potential evapotranspiration. Some derived parameters such as slope, dryness index, HGMR index, geology index, and landuse index were also explored. The selection of the final basin characteristics and hydrologic parameters was based on a series of stepwise regression procedures, statistical results, and the persistence of the parameters in the specified recurrence interval.

Based on the results, drainage area, dryness index, and the hydrogeomorphic properties of the basin were found to be significant in quantifying average annual baseflow of a watershed. Therefore given the watershed average values of dryness index, drainage area, and percentage coverage of certain hydrogeomorphic groupings it is possible to determine annual baseflow magnitudes of specified recurrence interval.

In general, the study demonstrated the suitability of statistical regional baseflow models as an alternative to more comprehensive and data intensive numerical groundwater flow models. Further improvements in the model could be achieved by the availability of a more extensive record of continuous streamflow data where the flow is not significantly affected by regulation. In addition investigating other robust and innovative techniques

such as Artificial Neural Networks might result in a significant departure from the traditional statistical approach and lead to a novel approach to modeling with more efficient algorithms. As explained in the final chapters of this report this kind of approach requires training a system with records of sufficient length before forecasting.

Introduction

The Potomac River Basin covers an area of approximately 14,600 square miles with interstate sub-basins in Maryland, Virginia, West Virginia, Pennsylvania, and the District of Columbia. The basin has very distinct geologic, geomorphic, hydrological and landuse/landcover characteristics that vary across the region. Both surface water and groundwater resources of the basin support about 5.3 million people according to the 2000 census. Increasing demand for water associated with population growth and weather anomalies that result in drought may strain the water resources of the basin.

Studying the availability of both surface water and groundwater resources of the basin helps answer key questions which are crucial in the development of a sustainable water resources management plan. Perhaps the more rigorous scientific approach requires developing an integrated surface-ground water model that simulates the regional flow and withdrawals over a period of time. Developing such a model is very data intensive and requires data collected over a longer period of time. In the absence of such a model other scientifically correct substitute methods provide some of the answers. One such approach involves quantifying the baseflow contribution from groundwater and using statistical methods to estimate the value of annual baseflow for specified recurrence intervals. The resulting values for annual baseflow provide estimates of annual recharge to basin aquifers.

The Potomac River Basin encompasses five hydrogeomorphic regions, which are groupings based primarily on geology and physiography. These regions are, Appalachian Plateau, Valley and Ridge, Blue Ridge, Piedmont Plateau, and Coastal Plain. Apart from the Coastal Plain, the basin is underlain predominantly by fractured bedrock. The fractured bedrock unit is generally understood to be a major contributor to the baseflow of the network of streams that form tributaries of the Potomac River.

The current study is an extension of an earlier study that assessed seasonal and annual water budgets for the Monocacy and Catoctin sub-basins. This study is part of the Potomac River Basin Groundwater Assessment Project, a collaboration between ICPRB and the U.S. Geological Survey (USGS). The federal funding was made available through the USGS for the duration of the study.

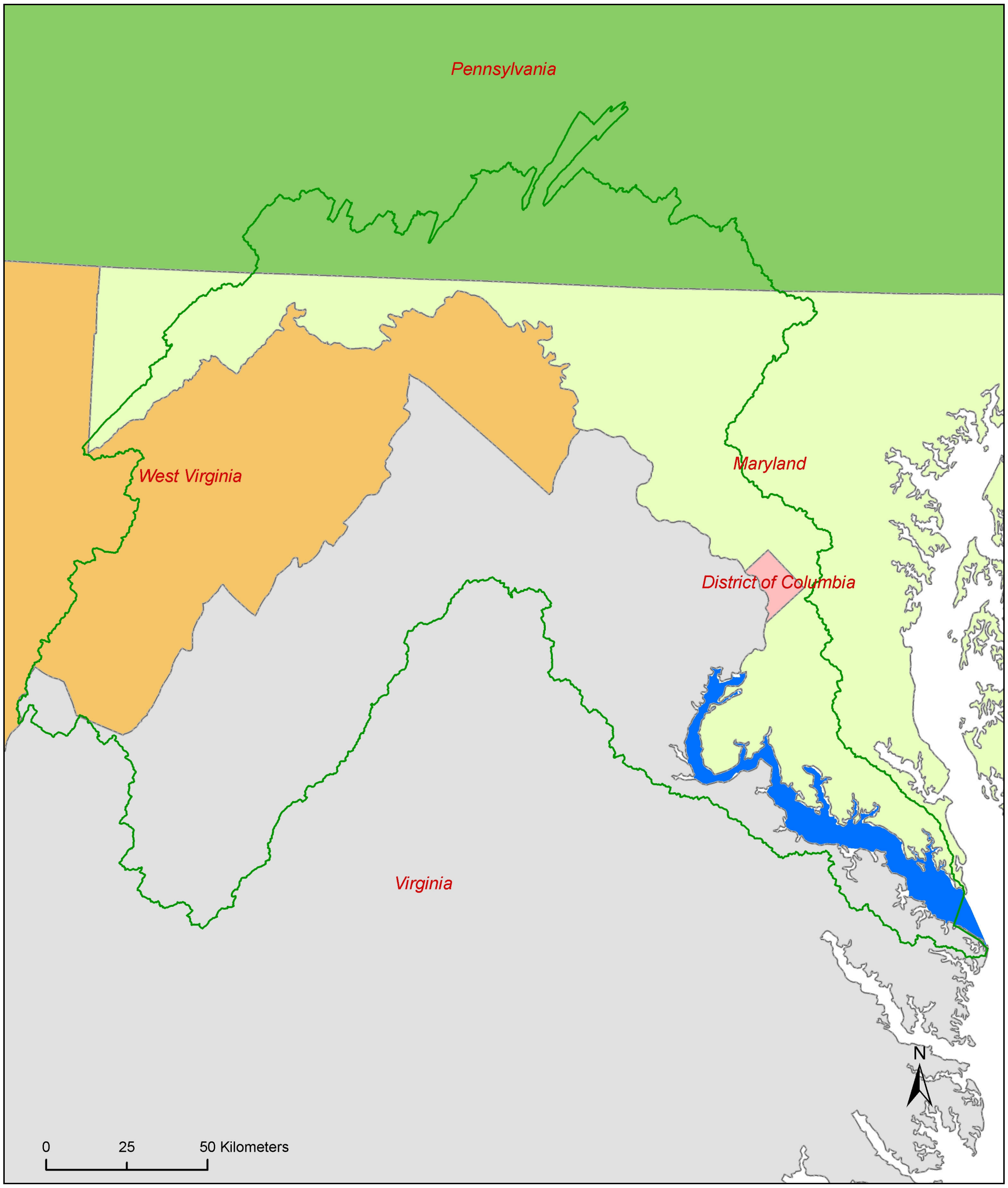


Figure 1. Potomac River Basin and States

Objective of the Study

The objective of the study is to determine water availability in the fractured bedrock units of the Potomac River Basin by analyzing annual baseflow statistics and developing statistical models. The previous water budget analysis completed for the Monocacy/Catoctin watersheds (Schultz et al., 2005) formed the basis for this study. The basic techniques adopted in this study are similar while the area covers the entire fractured bedrock aquifer of the Potomac River Basin.

Description of Main Basin Characteristics

The primary basin characteristics considered in this study include drainage area, geology, soil type, landuse, hydrogeomorphology (HGMR), precipitation, and potential evapotranspiration. Some derived parameters such as slope, dryness index, HGMR index, geology index, and landuse index were also explored.

Topography

The topography of the Potomac River Basin could broadly fall under three categories. The first category includes the Appalachian Mountains, the Valley and Ridge and the Blue Ridge areas. These areas have hills, mountains, valleys, and gorges that extend in a north east –south west direction. The second category is the Piedmont region and its surroundings which form a plateau while still maintaining the north east-south west bearing. The third group is the Coastal Plain, which has more or less flat lands that extend to the Atlantic coast. The elevation of the basin ranges from close to 1500 meters in the mountainous areas to zero in the Coastal Plain. Most of the gorges in the mountainous areas have very steep slopes. There are numerous ravines and gorges that were formed over the past millennia and now form a very intricate drainage pattern for the perennial and intermittent streams.

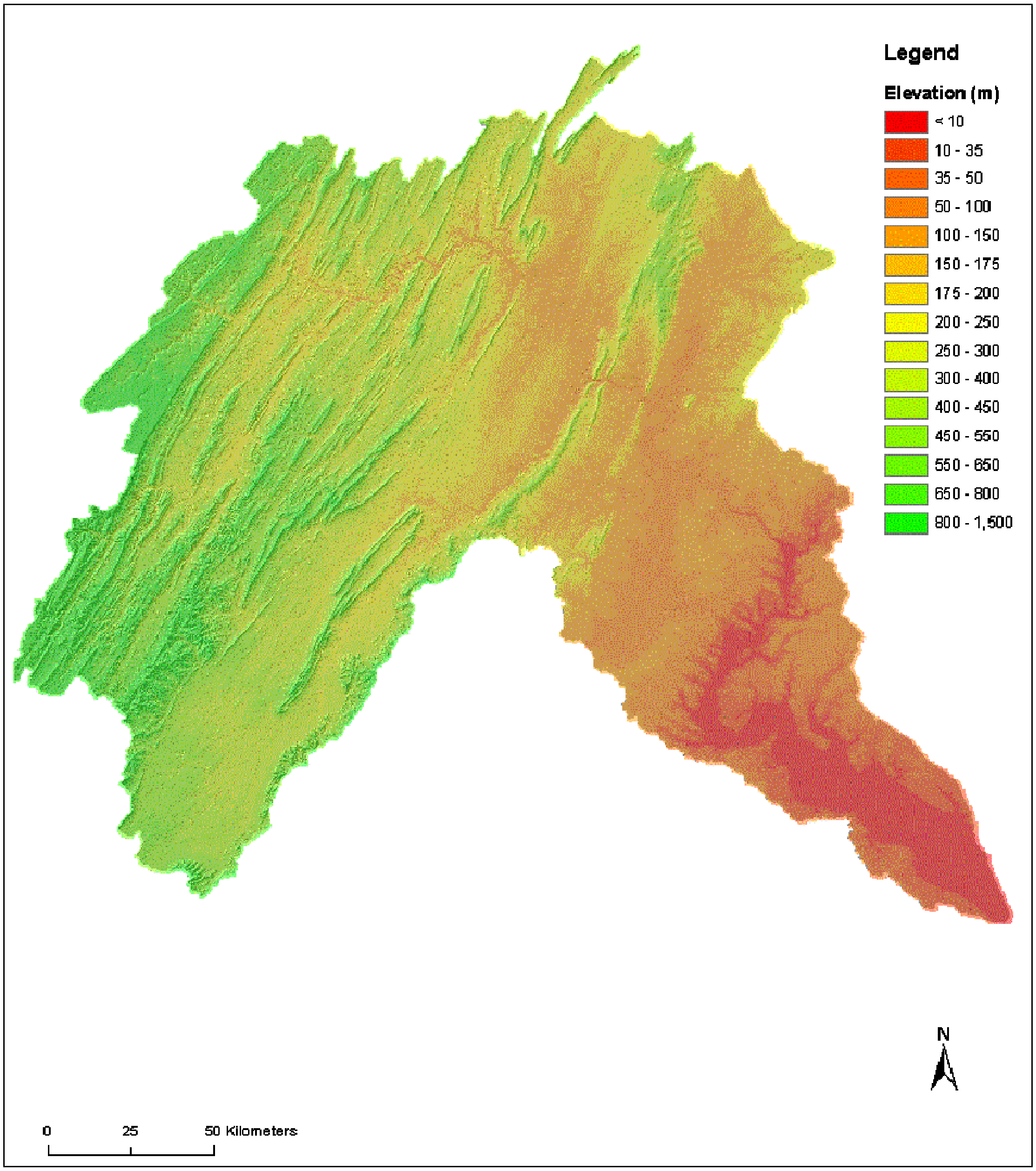


Figure 2. Potomac River Basin hillshade of digital elevation

Geology

The Potomac River Basin exhibits a varied geology consisting of rocks that make up the Appalachian Plateau, Blue Ridge, Coastal Plain, Valley and Ridge, and Piedmont regions. The rock types in the basin include carbonates, siliciclastic carbonates, siliciclastics, crystallines, and unconsolidated sediments. The geology data used in this study was derived from geospatially referenced Geographic Information Systems (GIS) data produced by the USGS for the USGS project Water-Quality Assessment of the

Potomac River Basin (Derosier, et al, 1998). The surficial geology data for this layer was compiled and digitized from state geologic maps for the entire area within the Appalachian Valleys-Piedmont Regional Aquifer System (APRASA) study area (Swain et al, 1991; Mesko, 1992).

The southeastern part of the basin, known as the Coastal Plain Province, is a lowland area that borders the Atlantic Ocean, and is underlain predominantly by semi-consolidated to unconsolidated sediments that consist of silt, clay, sand, and gravel. In parts of the Coastal Plain there are some consolidated beds of limestone and sandstone. The Coastal Plain sediments range in age from Jurassic to Holocene and generally dip toward the southeast. The Coastal Plain geology does not exhibit any direct connection with the fractured bedrock system of the Potomac River Basin. Therefore the Coastal Plain portion of the basin is excluded from this study as it does not satisfy the underlying assumption of connected fractures.

The topography of the Piedmont Province ranges from lowlands to peaks and ridges of moderate altitude and relief. The region has sheared, fractured, and folded metamorphic and igneous rocks that range in age from Precambrian to Paleozoic. There are also sedimentary basins that formed along rifts in the Earth's crust and contain shale, sandstone, and conglomerate of early Mesozoic age, interbedded locally with basaltic lava flows and minor coal beds. The sedimentary rocks and basalt flows are intruded in places by diabase dikes and sills.

The Blue Ridge Province is a mountain belt found in the northwestern margin of the Piedmont. Predominant rock types are igneous and high-rank metamorphic rocks but also include low-rank metamorphic rocks of late Precambrian age and small areas of sedimentary rocks of Early Cambrian age along its western margin.

The Valley and Ridge Province is characterized by layered sedimentary rocks that have been complexly folded and locally thrust faulted. As the result of repeated cycles of uplift and erosion, resistant layers of well-cemented sandstone and conglomerate form elongated mountain ridges and less resistant, easily eroded layers of limestone, dolomite, and shale form valleys. The rocks of the province range in age from Cambrian to Pennsylvanian. Parts of this province from central Pennsylvania into New Jersey have been glaciated, and glacial deposits fill or partially fill some of the valleys.

The Appalachian Plateaus Province is underlain by rocks that are continuous with those of the Valley and Ridge Province, but in the Appalachian Plateaus the layered rocks are nearly flat-lying or gently tilted and warped, rather than being intensively folded and faulted.

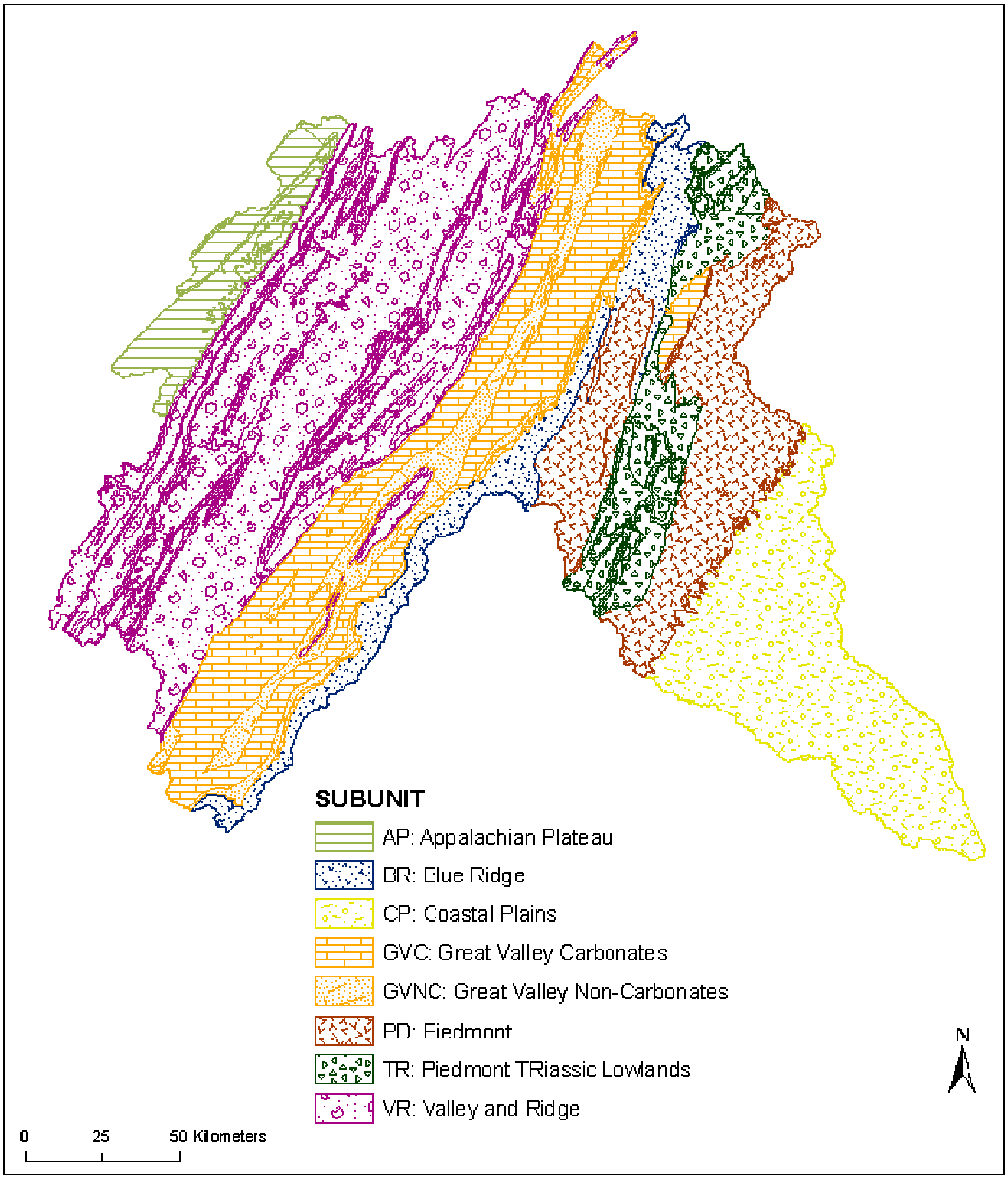


Figure 3. Potomac River Basin geology subunits

Hydrogeomorphology

The HydroGeoMorphic Regions (HGMR) present a generalized classification based on lithology and physiography of the region. The HGMR data was originally prepared for the Chesapeake Bay Program and used in conjunction with existing data to assess the significance of ground water discharge as a source of nitrate load to non tidal streams

(Bachman et al, 1998). A considerable part of the digital geology used to create this data set was compiled by the USGS Appalachian-Piedmont Regional Aquifer Systems Analysis (APRASA) study (Swain & others, 1991; Mesko, 1992).

The HGMR classification is based on the assumption that a given HGMR has similar amounts of groundwater discharge and similar responses to application of nutrients (Bachman et al., 1998). The HGMR data set was created by combining digital data sets of physiography and rock types from numerous sources. The physiographic provinces found in the data set include the Appalachian Plateau Province, Appalachian Mountain Section, Great Valley Section, Blue Ridge Province, Reading Prong Section, Mesozoic Lowlands Section, Piedmont Lowland Section, Piedmont Upland Section, and Coastal Plain Province. The major rock types consist of carbonates, siliciclastics, crystallines, and unconsolidated sediments.

Carbonates are characterized by the existence of solution channels created by dissolution of fractures (Fetter, 2000). As a result of the enlarged fractures, carbonates generally have high permeability values. The Appalachian Plateau, Valley and Ridge, and Piedmont physiographic regions have carbonates that cover most of the area. Siliciclastics predominantly have sandstone, silt-stone, and shales. Siliciclastics are found in the Appalachian Plateau (Trapp and Horn, 1997) where the groundwater flow occurs primarily as fracture flow. The Valley and Ridge region also has extensive siliciclastics. The extent of fracturing varies across the provinces and it affects the permeability of the formation accordingly.

During the testing of data integrity and completeness it was revealed that there were some inconsistencies in the classifications of HGMRs as a result of map shifts (Harlow and Nelms, 1998). The physiographic provinces were determined based on geologic provinces in each state which already had some edge matching problems at state boundaries due to mapping inconsistencies. Since the inconsistencies occurred primarily in the northern most parts in the original Chesapeake Bay map, the anomalies were assumed to be minimal in the smaller section clipped to create the Potomac Basin HGMR.

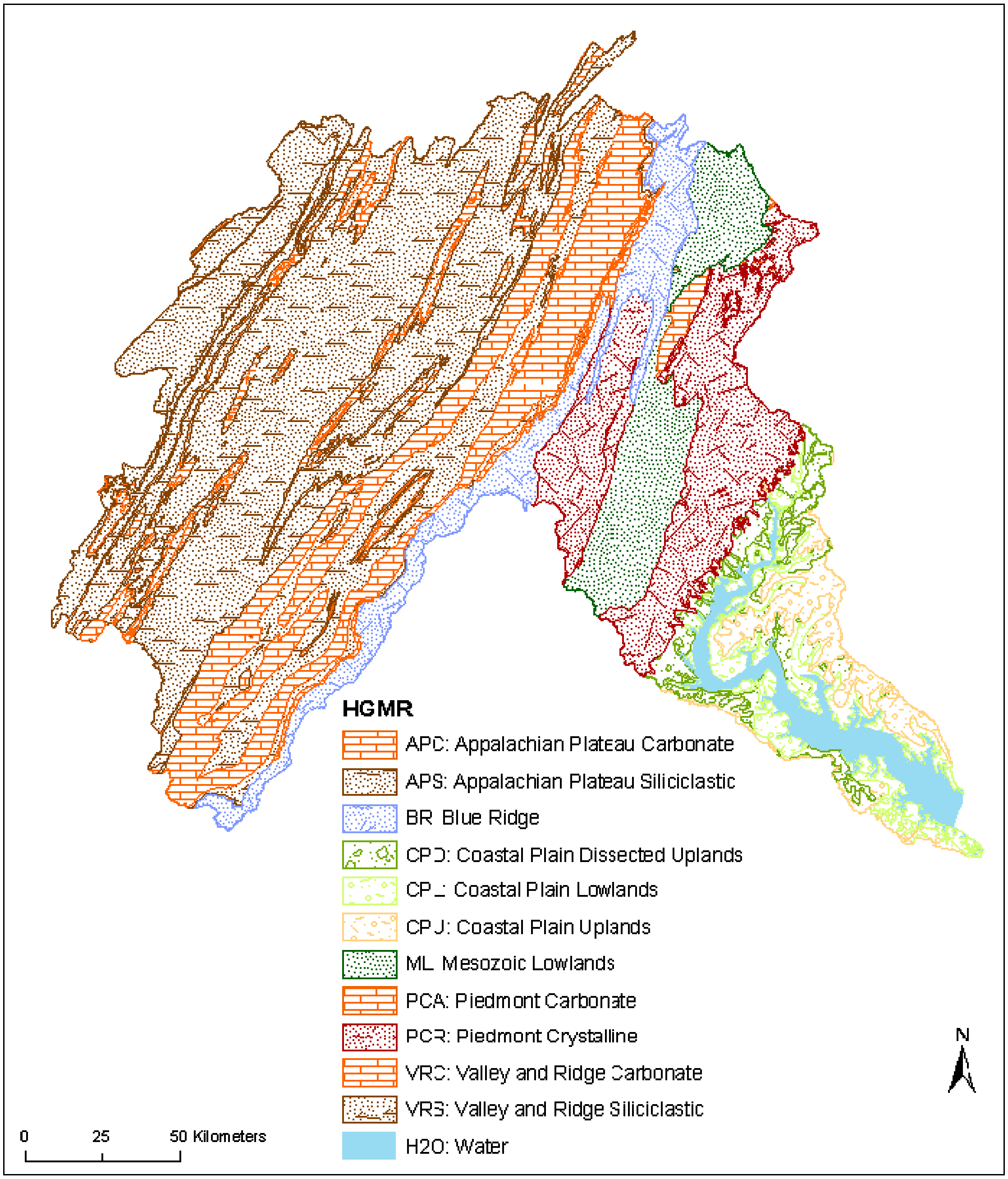

HGMR
APC: Appalachian Plateau Carbonate
APS: Appalachian Plateau Siliciclastic
BR Blue Ridge
CPD: Coastal Plain Dissected Uplands
CPL: Coastal Plain Lowlands
CPU: Coastal Plain Uplands
ML Mesozoic Lowlands
PCA: Piedmont Carbonate
PCR: Piedmont Crystalline
VRC: Valley and Ridge Carbonate
VRS: Valley and Ridge Siliciclastic
H2O: Water
N
0
25
50 Kilometers

Figure 4. Potomac River Basin HGMR provinces

Soil Type

The soils data is extracted from the State Soil Geographic (STATSGO) database. STATSGO is a relational database that has a unique identifier for each state soil classification known as Map Unit Identification symbol (MUID). A number of soil properties are mapped through the MUID for individual polygons in every state. Using the MUIDs, specific soil properties of interest are extracted from the tabulated definition of soil data elements and codes. From hydraulics point of view the soil property of interest with possible physical influence on baseflow generation is the permeability or infiltration capacity of the soil. This property primarily governs the flow of water through the soil pores. From the available categories of soil properties in STATSGO, the parameter that provides values for infiltration properties is tabulated as HYDGRP (the hydrologic group of the soil). Subsequent analysis was carried out using the GIS data to process the relational database and create a layer of HYDGRP for the Potomac River Basin. The HYDGRP classes available in the Potomac River Basin are:

- Class – A: High infiltration rates. Soils are deep, well drained to excessively drained sands and gravels.
- Class - A/D: Drained/undrained hydrology class of soils that can be drained and are classified.
- Class – B: Moderate infiltration rates. Deep and moderately deep, moderately well and well drained soils with moderately coarse textures.
- Class - B/D: Drained/undrained hydrology class of soils that can be drained and are classified.
- Class – C: Slow infiltration rates. Soils with layers impeding downward movement of water, or soils with moderately fine or fine textures.
- Class - C/D: Drained/undrained hydrology class of soils that can be drained and classified.
- Class – D: Very slow infiltration rates. Soils are clayey, have a high water table, or are shallow to an impervious layer.

The soil classes, D, A/D, B/D, and C/D, are poorly drained soils with very fine texture and very slow infiltration rates.

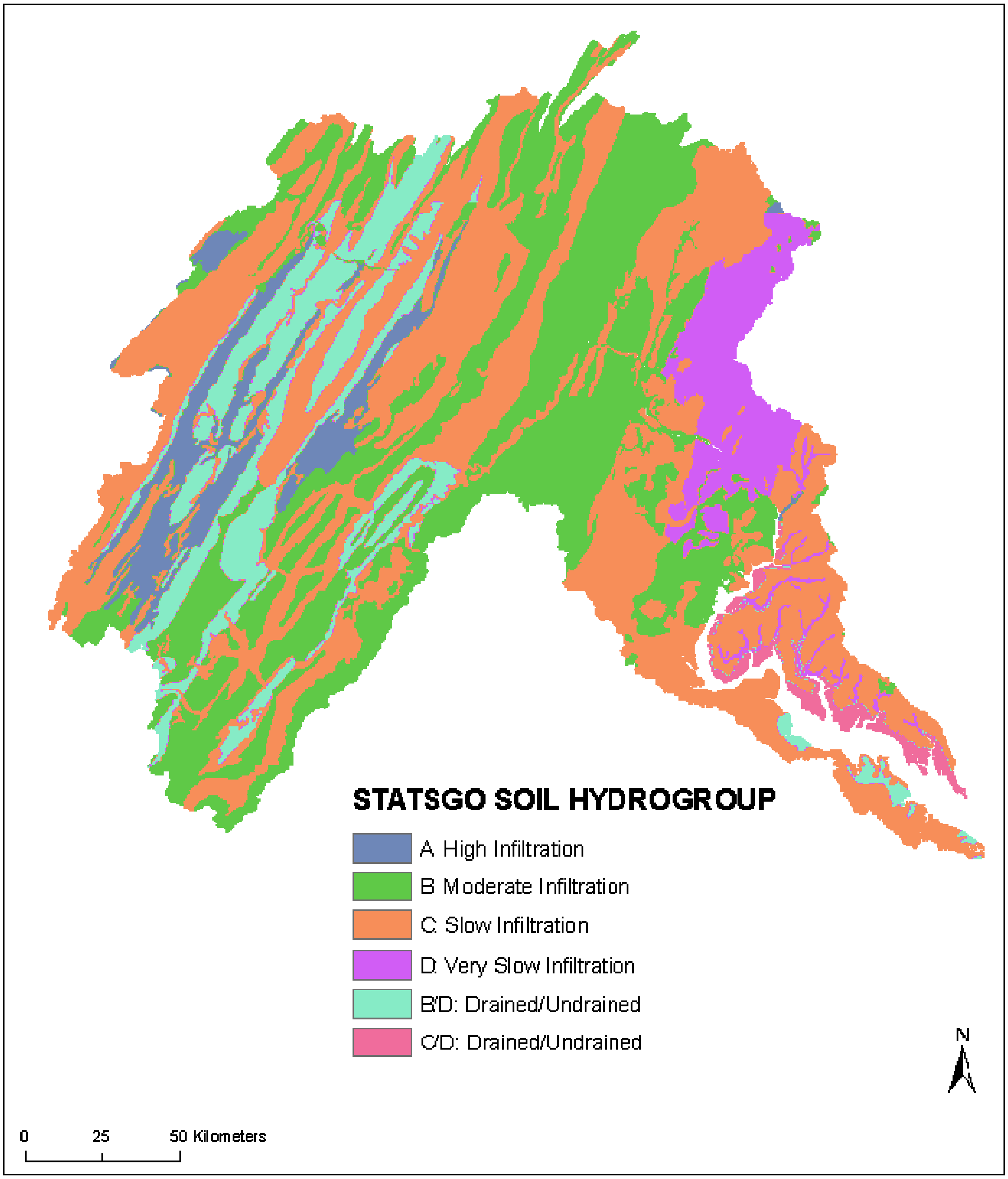


Figure 5. Potomac River Basin soil HYDGRP classes

Landuse/Landcover

The landuse/land cover data used in this study was originally created by the Land Cover Characterization Program of the USGS Earth Resources Observation Systems (EROS) Data Center for the Water Resources Division of the USGS in response to the request of the Chesapeake Bay Program. The data set is a 1997 land cover data set comparable to the 1992 National Land Cover Dataset. These two land cover data sets were used in the SPARROW (SPAtially Referenced Regressions On Watershed attributes) model. Satellite imagery acquired in 1997 was used to derive a vegetation cover-type data. As described in the documentation of the data, the intended use of the 1997 data set was to compare 1997 land cover with 1992 vintage land cover data. The classification is a Multi-Resolution Land Characterization (MRLC). There are about 20 different landuse/land cover classes in the original classification. In order to create a manageable size of explanatory variables, a reclassification was performed by lumping closely related types. Reclassification of the landuse classes resulted in 5 lumped land cover classes as listed in Table 1. The reclassification is based on the land cover type irrespective of the secondary attributes, such as the density of the land cover type. For example, residential areas, industrial areas, commercial areas, and transportation were lumped as one, since they all represent developed areas with relatively high runoff rates.

Table 1. Reclassified landuse/landcover types

Land Cover MRLC 97	Reclassified Class Name
Water	Open Water
Low Intensity Residential High Intensity Residential Commercial/Industrial/Transportation	Residential and Industrial
Bare Rock/Sand Quarries/Mines/Gravel Pits Transitional	Impervious
Deciduous Forest Evergreen Forest Mixed Forest	Forest
Pasture/Hay Row Crops Small Grains Bare Soil Other Grasses	Pasture and Grassland
Forested Wetlands Emergent Wetlands	Wetland

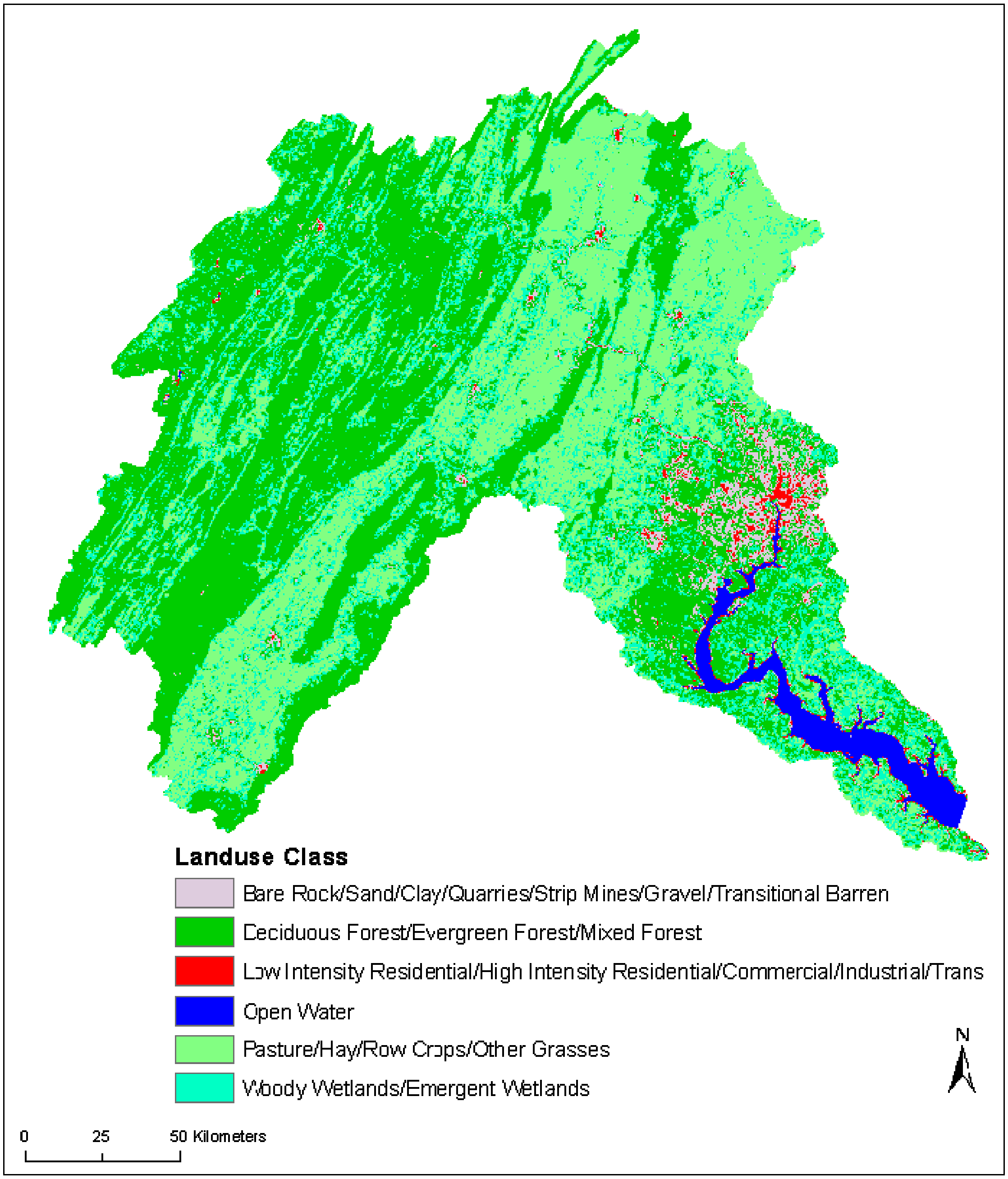

Landuse Class
Bare Rock/Sand/Clay/Quarries/Strip Mines/Gravel/Transitional Barren
Deciduous Forest/Evergreen Forest/Mixed Forest
Low Intensity Residential/High Intensity Residential/Commercial/Industrial/Trans
Open Water
Pasture/Hay/Row Crops/Other Grasses
Woody Wetlands/Emergent Wetlands
N
0
25
50 Kilometers

Figure 6. Potomac River Basin landuse classes

Hydrology

Hydrologic data used in this study include precipitation, potential evapotranspiration dryness index, and streamflow. The precipitation and potential evapotranspiration data were obtained from the Chesapeake Bay Program point data measurements. Streamflow data was obtained from the US Geological Survey continuous record gages and partial record flow measurements. The point data of precipitation and evapotranspiration were geo-referenced, projected, and then converted to raster coverages of average interpolated precipitation and potential evapotranspiration. The initial raster cell size was about 10 km by 10 km; this was re-sampled to a 500 m by 500 m cell size. The creation of a smaller cell size helped to avoid the creation of cells with no data values which were observed at the original scale.

The relationship between climatic demand for water (potential evapotranspiration) and climatic supply (precipitation) is expressed as the index of dryness (Budyko, 1974; Milly, 1994). Dryness index is defined as the ratio of potential evapotranpiration to precipitation. Customarily, regions with an index of dryness less than one can be considered humid, while regions with a dryness index greater than one can be classified as arid. The dryness index for the study area was derived from the precipitation and evapotranspiration GIS coverages.

The hydrology of the Potomac River Basin exhibits noticeable patterns of precipitation, evapotranspiration, and streamflow that loosely follow the physiographic features and landuse/landcover classes. The influence of orographic and landcover variabilities on the hydrology of the basin is manifested by a relatively higher precipitation in the higher altitude areas or mountain ranges with dense vegetation cover. The Appalachian Mountains and the ridges of the Blue Ridge, and Valley and Ridge physiographic regions have higher precipitation values. The Mesozoic Lowlands and parts of the Piedmont Crystalline also have higher precipitation that follows the topography, although it is not as pronounced as the western part of the basin. Evapotranspiration in the western part of the basin follows the landuse/landcover classes rather than the orographic trends. In the low lying areas of the eastern part of the basin, both landuse and lower altitude seem to be determining factors. The landuse/landcover classes with higher evapotranspiration are usually agricultural croplands, pasture lands, residential, and industrial landuse classes.

Annual precipitation in the Potomac River Basin ranges from close to 60 inches in the Appalachian Plateau to above 30 inches in the Valley and Ridge and about 40 inches in the Piedmont and Coastal Plain and the Blue Ridge. The potential evapotranspiration ranges from above 20 inches in the Appalachian Plateau to about 35 inches in the Piedmont and Coastal Plain where the highest evapotranspiration rate is observed. The average evapotranspiration exhibits a generally increasing trend from west of the basin to east of the basin.

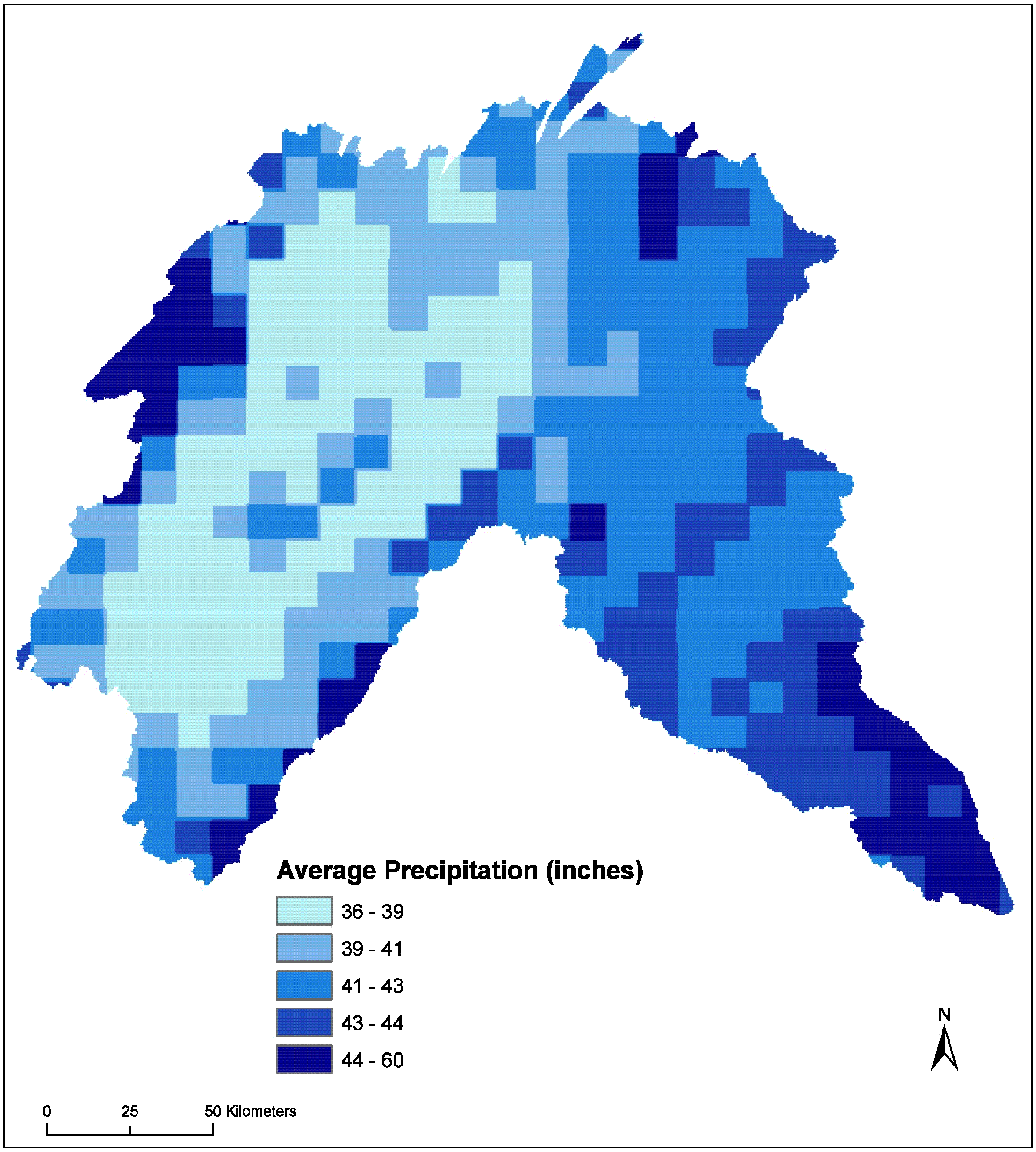


Figure 7. Potomac River Basin average precipitation (1984-2003)

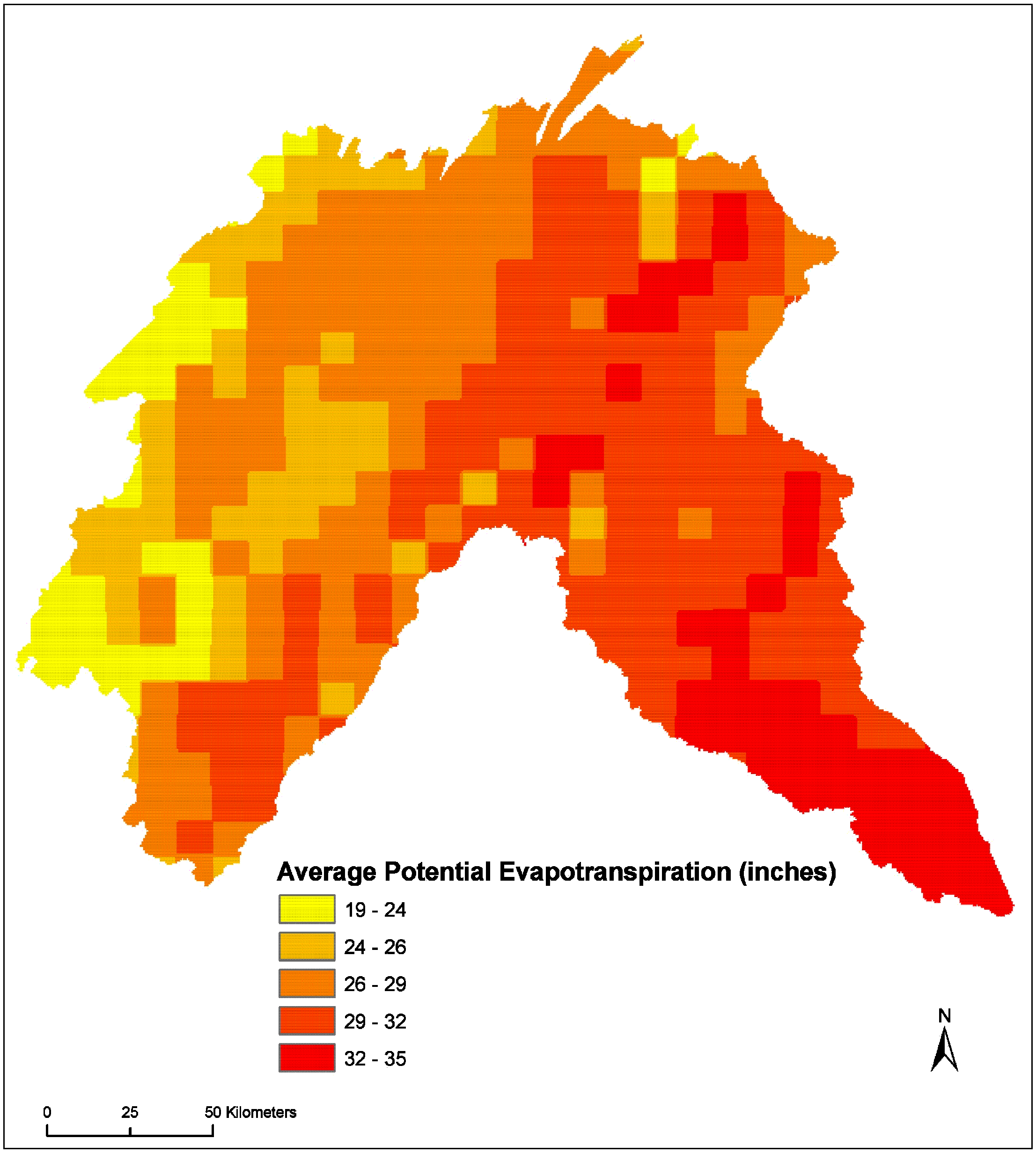


Figure 8. Potomac River Basin average potential evapotranspiration (1984-2003)

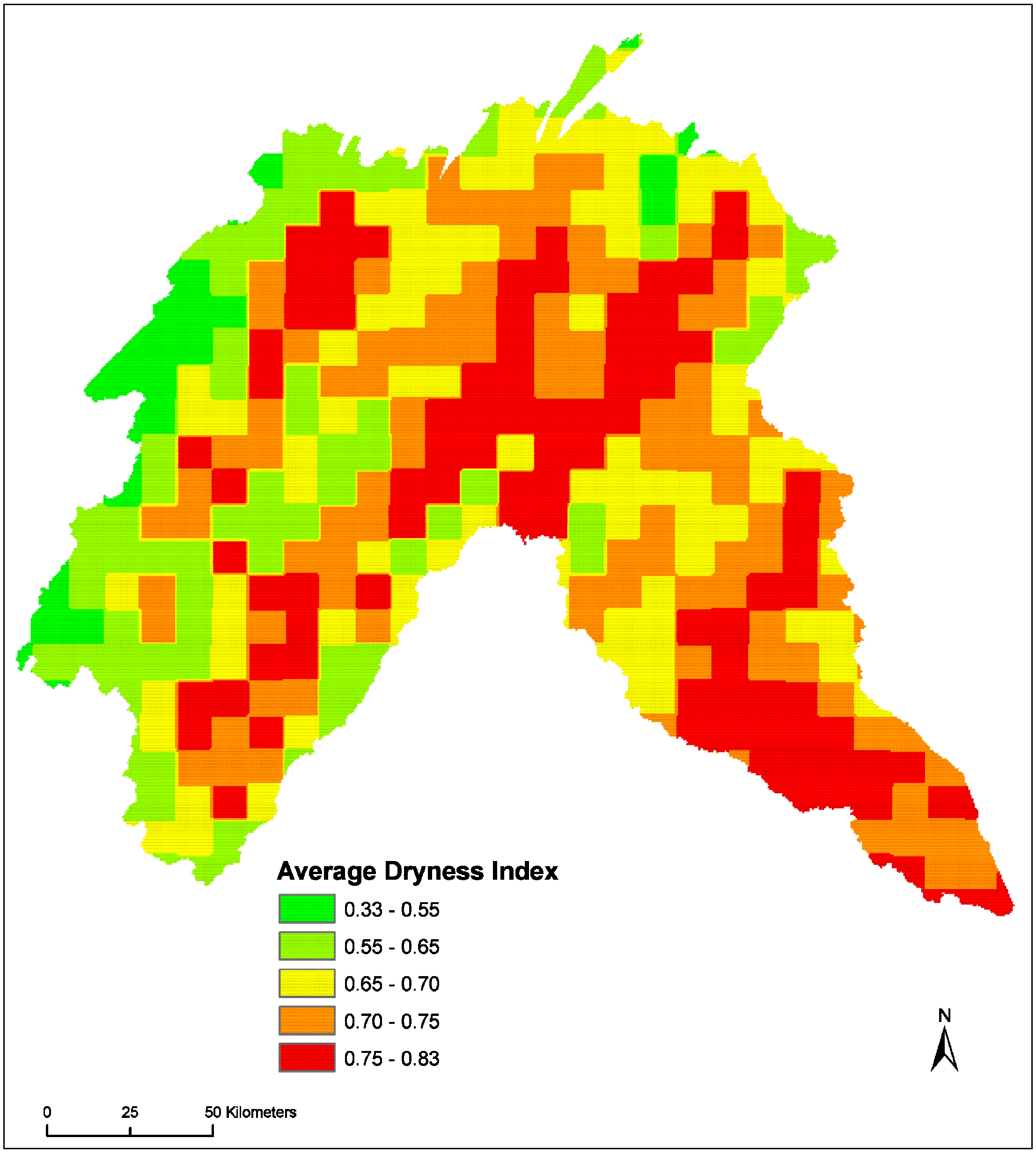


Figure 9. Potomac River Basin average dryness index (1984-2003)

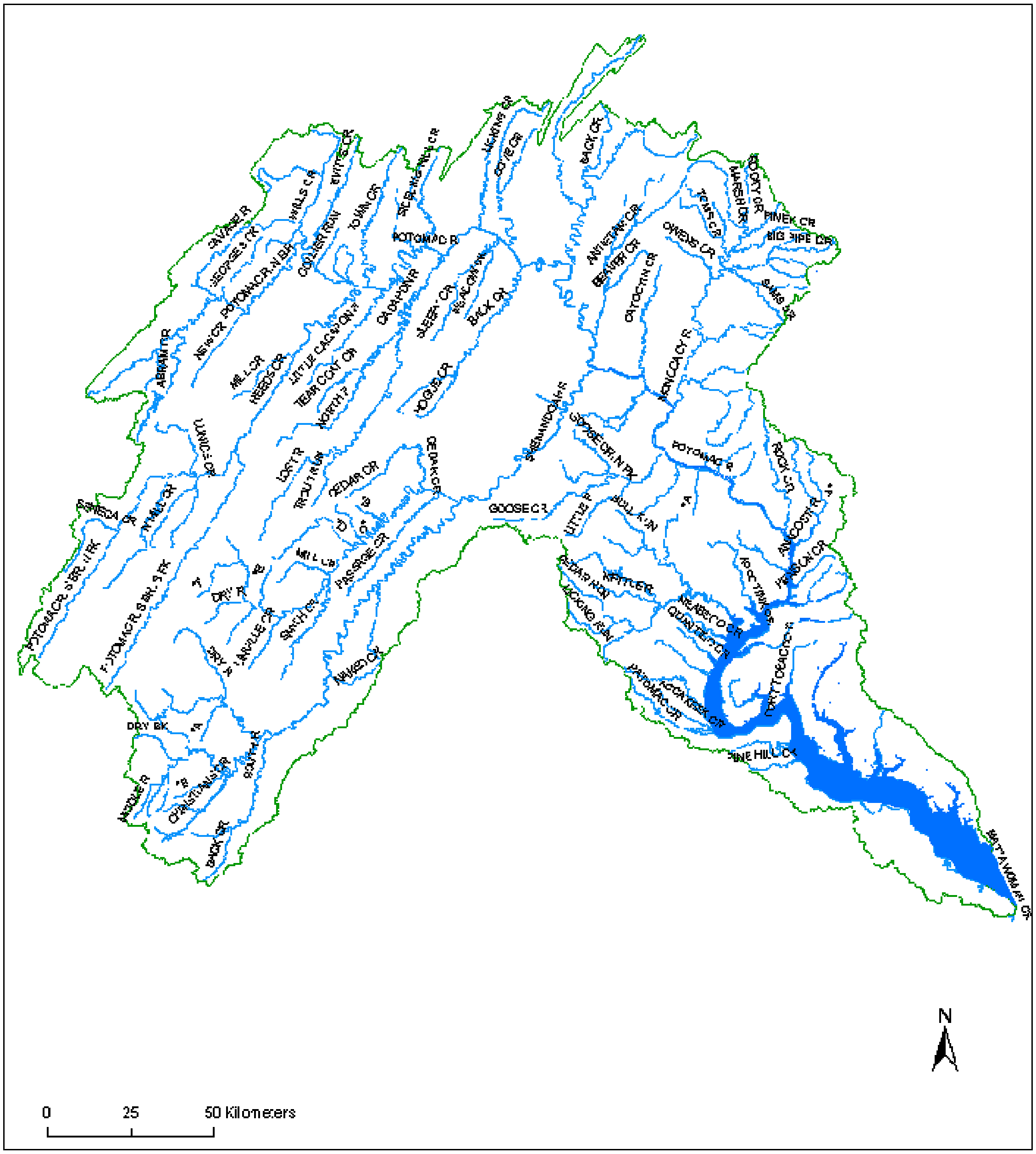


Figure 10. Potomac River Basin main tributaries and creeks

Annual Water Budget Analysis

Water budget analysis is a common hydrologic analysis method employed in the estimation of storage or water available in a watershed or aquifer. The method makes generalized assumptions in determining the balance of water available in a watershed or basin based on averaged inflow and outflow quantities of water into and from the watershed. The hydrologic components of the watershed are identified and accounted for in the detailed analysis. The more general approach involves using annual average values over a longer period of time, thus resulting in long-term average estimates. Over a shorter period the analysis is also performed to determine dry and wet season water availability. The second approach requires seasonal analysis to highlight the seasonal variability of hydrologic parameters.

Methodology

The primary goal of this study is investigating the annual recharge to groundwater in the fractured bedrock portions of the Potomac River Basin by developing regional statistical models of annual baseflow. Baseflow represents the long-term contribution of groundwater from storage to streams and maintains the streamflow in between rainy seasons. Baseflow is computed graphically or analytically from streamflow measurements. The streamflow gage stations are selected on the basis of specified length of record which usually is a minimum of 10 years and information on regulation of the flow. Stations with less than 10 years of record are also included in the regression equations development by extending their record using data from adjacent continuous record stations. Low-flow partial record stations are also selected and appropriate statistical techniques used to create baseflow data of specified magnitude and recurrence interval. For the final set of continuous record and extended record stations, the baseflow is separated from the streamflow data and a suitable statistical distribution is fitted to create a set of baseflow statistics of specified magnitude and recurrence interval.

Sub-watersheds are delineated based on the location of the selected gage stations and the digital elevation of the area. Basin characteristics data is generated for each watershed defined by a gage station. The basin characteristics are either percentage of areas covered or average values for the watershed. The models are regression equations that relate baseflow statistics to basin characteristics data such as geology, soil type, physiography, and land cover, and hydrologic parameters such as precipitation and evapotranspiration. The selection of equations is based on a series of statistical goodness of fit tests. A series of matrix plots are also investigated to determine visible relationships between magnitudes of baseflow and basin characteristics. Equations for specified magnitudes and recurrence interval are selected after satisfying the statistical criteria.

Statistical Analysis of Annual Baseflow

The statistical analysis of annual baseflow requires developing regression models which are calibrated by fitting basin characteristics and hydrologic data. The final regression models are then used to estimate baseflow statistics for ungaged catchments with appropriate explanatory variables data. The set of possible explanatory variables usually include more variables than are statistically necessary, therefore, preliminary analyses have to be conducted to determine the most significant set of parameters. The most common tests performed to determine the significance of explanatory variables include examining the p-value, F-statistic, t-statistic and standard error terms. In addition to testing for significance of statistical values, one has to create and analyze correlation plots of the explanatory variables.

During the statistical analysis of the data null and alternate hypotheses have to be established. The parametric tests that follow will be based on the hypothesis test for each case. The null hypothesis, commonly represented as H_0, is usually assumed to be true. For example, in linear regression the null hypothesis could be the slope of the equation is zero. The alternate hypothesis H_1 is the opposite of H_0, and it is accepted if data based evidence suggests that H_0 is not true. The statistical analysis also requires defining a level of significance known as α -value. This value signifies the probability of rejecting a null hypothesis H_0 while it is true (Type I error). The most common value of the level of significance α is 5% or 0.05. If the p-value is greater than α , H_0 is not rejected.

The p-value of a variable helps determine the probability that the assumed relationship occurs by chance, and that it represents the population from which the sample was drawn. It also represents the probability of error that is involved in accepting the observed result as valid, therefore the higher the p-value, the less reliable the result. In many instances the p-values do not necessarily stand out the as the true stand-alone tests that determine the importance of an explanatory variable or a regressor. Experience has shown that an important regressor can have a large p-value for a number of reasons. Some of the reasons include a small sample size, if the regressor is measured over a narrow range, whether there are large measurement errors, or if another closely related regressor is included in the equation. Similarly, a truly less significant regressor can also have a very small p-value if the sample size is large. Therefore, p-values should be used within a computed confidence interval for a parameter estimate. The p-values computed over a given confidence interval provide more reliable information than simply the value alone. In other words, the p-value is simply the probability of obtaining the computed test statistic and measures the believability of the null hypothesis. The null hypothesis is rejected while the p-value is less than the level of significance α .

The F value and probability of the F statistic test the overall significance of the regression model. More specifically, F values test a null hypothesis that assumes all of the regression coefficients are equal to zero, while the converse is the real hypotheses stating that at least one of the coefficients is significantly different from zero. This tests the full model against a model with no variables and with the estimate of the dependent variable being the mean of the values of the dependent variable. Statistically the F value is the ratio of the mean regression sum of squares divided by the mean error sum of squares and its values range from zero to an arbitrarily large number. A brief description of the nested F test is presented next (Helsel and Hirsch, 2002). Consider a multiple linear regression model of the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon \qquad \textbf{(1)}$$

where

y is the response variable

β_0 , β_1 , β_2 , β_k are slope coefficients for the first, second and kth response variables and

ε is the error term

Let's consider a simpler model "s" which has fewer explanatory variables.

$$y_s = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon_s \qquad \textbf{(2)}$$

The model has k+1 parameters including the intercept , n-(k+1) degrees of freedom (df$_s$), and sum of squared errors SSE$_S$.

Let's also consider a more complex model "c" of the form

$$y_c = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \beta_{k+1} x_{k+1} + \ldots + \beta_m x_m + \varepsilon_c \qquad \textbf{(3)}$$

with m+1 parameters, residual degrees of freedom (df$_c$) of n-(m+1), and sum of squared errors SSE$_C$.

The nested F test helps in determining whether the simpler model with fewer parameters is sufficient in explaining the variation of y as compared to the more complex model that has more number of variables. The test is called nested because all the variables in the simpler model are also present in the complex model. The test requires formulating a null hypothesis and an alternative. The null hypothesis assumes that the coefficients of the extra variables in the complex model are all zero, whereas the alternative hypothesis suggests that at least one of the extra variable coefficients is different from zero.
The null hypothesis is;

$$\mathrm{H}_0 : \beta_{k+1} = \beta_{k+2} = \ldots = \beta_m = 0$$

and the alternative hypothesis is,

H$_1$: at least one of these m-k coefficients is not equal to zero.

The cost of adding the extra variables in the complex model has to be compensated by the added power of explaining the variability in y. This cost is measured by the loss in the degrees of freedom (= m-k), the number of additional variables in the more complex equation. The test statistic is calculated as

$$\mathrm{F} = \frac{(\mathrm{SSE_s} - \mathrm{SSE_c}) / (\mathrm{df_s} - \mathrm{df_c})}{(\mathrm{SSE_c} / \mathrm{df_c})}$$ where (df$_s$ – df$_c$) = m-k

The F statistic should not exceed the tabulated value of the F distribution with (df_s – df_c) and df_c degrees of freedom for the selected α (say α =0.05) , otherwise H_0 is rejected. If the null hypothesis is rejected then it is an indicator of better performance of the more complex model in terms of explaining the variability in y.

The t-test uses the t statistic to evaluate the differences in means between two sample groups. The t statistic is computed by dividing the estimated value of the parameter by its standard error. This statistic is a measure of the likelihood that the actual value of the parameter is not zero. The larger the absolute value of t, the less likely that the actual value of the parameter could be zero. The p-level reported with a t-test represents the probability of error involved in accepting the hypothesis about the existence of a difference.

The hypothesis test of greatest interest in regression is the test for a significant slope (β_1).
Typically, the null hypothesis is;
H_0: $\beta_1 = 0$

versus the alternative hypothesis

H_0: $\beta_1 \neq 0$

The null hypothesis also states that the value of y does not vary as a linear function of x. Thus for the case of a single explanatory variable this also tests for whether the regression model has statistical significance. A third interpretation is as a test for whether the linear correlation coefficient significantly differs from zero. The latter two interpretations are not applicable for multiple explanatory variables. The test statistic computed is the t-ratio (the fitted coefficient divided by its standard error). The null hypothesis H_0 is rejected if $|t| > t_{crit}$, where t_{crit} is the point on the Student's t distribution with n-2 degrees of freedom, and with probability of exceedance of $\alpha /2$.

If there is more than one explanatory variable involved, then the one-way t-test will no longer be valid to test the differences between the means and medians of the data. Assuming the data within each of the groups are normally distributed and possess identical variances, the analysis of variance (ANOVA) test could be used. Analysis of variance is a parametric test, which helps to determine whether each data group's mean is identical. In the case of only two groups (one dependent and one explanatory), the ANOVA becomes identical to a t-test. Thus ANOVA is like a t-test between three or more groups of data, and is restricted by the same types of assumptions as was the t-test.

Correlation plots are matrix plots that show the correlation between different dependent and independent variables. The correlation coefficient indicates the existence of a linear relationship between two variables. The absolute value of the correlation coefficient ranges between zero and one. If the correlation value is zero then the parameters under consideration have no relationship, while a value of one signifies perfect relationship.

Correlation coefficients can have positive or negative signs, with negative values showing inverse relationships.

Data Pre-processing

The pre-processing of data involved screening all available stations and classifying them based on record length and then delineating watersheds. In order to identify the record length of the stations, the USGS computer program SCREEN was used. Based on the results, all stations with record length of 10 years or more and drainage area that did not exceed 500 square miles were selected to be used as continuous record stations. Also, gage stations where the regulation is thought to influence the flow regime were not selected for the final analysis. The attempt to use continuous record stations with concurrent periods of record was abandoned after carefully examining the advantages and disadvantages. The advantages lie in the fact that it describes the hydrologic conditions that persisted in the time period of interest, thus creating a favorable condition for analyzing the seasonal variations in the hydrologic cycle of the area. However, in the absence of a sufficient number of stations with longer periods of continuous record, the sampling of gage stations with only concurrent periods of record leaves out many stations that do not satisfy the requirement. Moreover, the use of a specific period from a station that has a longer record (for example, using the 1951 to 1980 period only from a station that has a record extending from 1910 to present) leaves out data that would have enhanced the statistical significance of the data from the station.

For the purpose of this study where the goal is developing regional regression equations that utilize long-term average statistics of specified recurrence interval, using data from all stations with the most number of years of record provides more data and thus better statistics. Also, the areal coverage of the study area with delineated sub-watersheds increases significantly resulting in more stations and more continuous records in each station. Since this approach maximizes the number of stations with longer periods of record, it provides a better representation of the long-term average climatological conditions and thus better estimates of baseflow. This also minimizes the bias that tends to influence the estimates if the records were taken during extremes such as consecutive dry periods. The task of extending records at stations with shorter periods of record also benefits from a larger pool of possible index stations of longer periods of record. From the available list of stations obtained from the USGS for this study, only 82 satisfied all the requirements. Further investigation of the estimated baseflow estimates through preliminary plotting and regression revealed some values that appeared to be outliers due to the significantly higher magnitudes. The outlier stations were then removed from the list, and the final list of 73 stations is presented in Table 2. The list includes continuous record stations, extended record stations and partial record stations.

Table 2. Final list of gage stations used in the analysis

Station ID	Station Name	HUC8	Station Type
01595000	North Branch Potomac River at Steyer, MD	02070002	Continuous
01595200	Stony River near Mount Storm,WV	02070002	Continuous
01595300	Abram Creek at Oakmont, WV	02070002	Continuous
01596500	Savage River near Barton, MD	02070002	Continuous
01596600	Big Run near Swanton, MD	02070002	Partial
01597000	Crabtree Creek near Swanton, MD	02070002	Continuous
01597100	Middle Fork Crabtree Creek near Swanton, MD	02070002	Partial
01598980	Mill Run at Morrison, MD	02070002	Partial
01599500	New Creek near Keyser, WV	02070002	Continuous
01601000	Wills Creek below Hyndman, PA	02070002	Continuous
01601325	Jennings Run at Corriganville, MD	02070002	Partial
01603500	Evitts Creek near Centerville, PA	02070002	Continuous
01604500	Patterson Creek near Headsville, WV	02070002	Continuous
01605425	Mill Run at Oldtown, MD	02070002	Partial
01605500	South Branch Potomac River at Franklin, WV	02070001	Continuous
01606000	North Fork South Branch Potomac River at Cabins, WV	02070001	Continuous
01608000	South Fork South Branch Potomac RIver near Moorefield, WV	02070001	Continuous
01609000	Town Creek near Oldtown, MD	02070003	Continuous
01609500	Sawpit Run near Oldtown, MD	02070003	Continuous
01609800	Little Cacapon River near Levels, WV	02070003	Continuous
01612500	Little Tonoloway Creek near Hancock, MD	02070004	Continuous
01613500	Licking Creek near Sylvan, PA	02070004	Continuous
01614000	Back Creek near Jones Springs, WV	02070004	Continuous
01614500	Conococheague Creek at Fairview, MD	02070004	Continuous
01615000	Opequon Creek near Berryville, VA	02070004	Continuous
01616000	Abrams Creeks near Winchester, VA	02070004	Continuous
01617000	Tuscarora Creek Above Martinsburg, WV	02070004	Continuous
01617600	Downey Branch near Downville, MD	02070004	Partial
01617800	Marsh Run at Grimes, MD	02070004	Continuous
01619000	Antietam Creek near Waynesboro, PA	02070004	Continuous
01619325	Beaver Creek at Benevola, MD	02070004	Partial
01619350	Little Beaver Creek at Benevola, MD	02070004	Partial
01619480	Little Antietam Creek at Keedysville, MD	02070004	Partial
01622000	North River near Burketown, VA	02070005	Continuous
01624800	Christians Creek near Fishersville, VA	02070005	Continuous
01625000	Middle River near Grottoes, VA	02070005	Continuous
01626000	South River near Waynesboro, VA	02070005	Continuous
01626500	South River at Waynesboro, VA	02070005	Continuous
01627500	South River at Harriston, VA	02070005	Continuous
01628060	White Oak Run near Grottoes, VA	02070005	Continuous
01632000	North Fork Shenandoah River at Cootes Store, VA	02070006	Continuous
01632082	Linville Creek at Broadway, VA	02070006	Continuous
01632900	Smith Creek near New Market, VA	02070006	Continuous
01634500	Cedar Creek near Winchester, VA	02070006	Continuous
01635500	Passage Creek near Buckton, VA	02070006	Continuous
01636730	Israel Creek near Weaverton, MD	02070008	Partial
01636850	Little Catoctin Creek near Brunswick, MD	02070008	Partial

Station ID	Station Name	HUC8	Station Type
01637500	Catoctin Creek near Middletown, MD	02070008	Continuous
01638480	Catoctin Creek at Taylorstown, VA	02070008	Continuous
01638600	Tuscarora Creek at Tuscarora, MD	02070008	Partial
01639000	Monocacy River at Bridgeport, MD	02070009	Continuous
01639140	Piney Creek near Taneytown, MD	02070009	Continuous
01639500	Big Pipe Creek at Bruceville, MD	02070009	Continuous
01640500	Owens Creek at Lantz, MD	02070009	Continuous
01641000	Hunting Creek at Jimtown, MD	02070009	Continuous
01641500	Fishing Creek near Lewistown, MD	02070009	Continuous
01641900	Tuscarora Creek near Frederick, MD	02070009	Partial
01642050	Israel Creek near Walkersville, MD	02070009	Partial
01642500	Linganore Creek near Frederick, MD	02070009	Continuous
01643125	Ballenger Creek near Lime Kiln, MD	02070009	Partial
01643500	Bennett Creek at Park Mills, MD	02070009	Continuous
01644000	Goose Creek near Leesburg, VA	02070008	Continuous
01645000	Seneca Creek at Dawsonville, MD	02070008	Continuous
01646000	Difficult Run near Great Falls, VA	02070008	Continuous
01650500	Northwest Branch Anacostia River near Colesville, MD	02070010	Continuous
01652500	Fourmile Run at Alexandria, VA	02070010	Continuous
01653000	Cameron Run at Alexandria, VA	02070010	Continuous
01654000	Accotink Creek near Annandale, VA	02070010	Continuous
01656100	Cedar Run near Aden, VA	02070010	Continuous
01656650	Broad Run near Bristow, VA	02070010	Continuous
01657000	Bull Run near Manassas, VA	02070010	Continuous
01658500	South Fork Quantico Creek near Independent Hill, VA	02070011	Continuous
01660400	Aquia Creek near Garrisonville, VA	02070011	Continuous

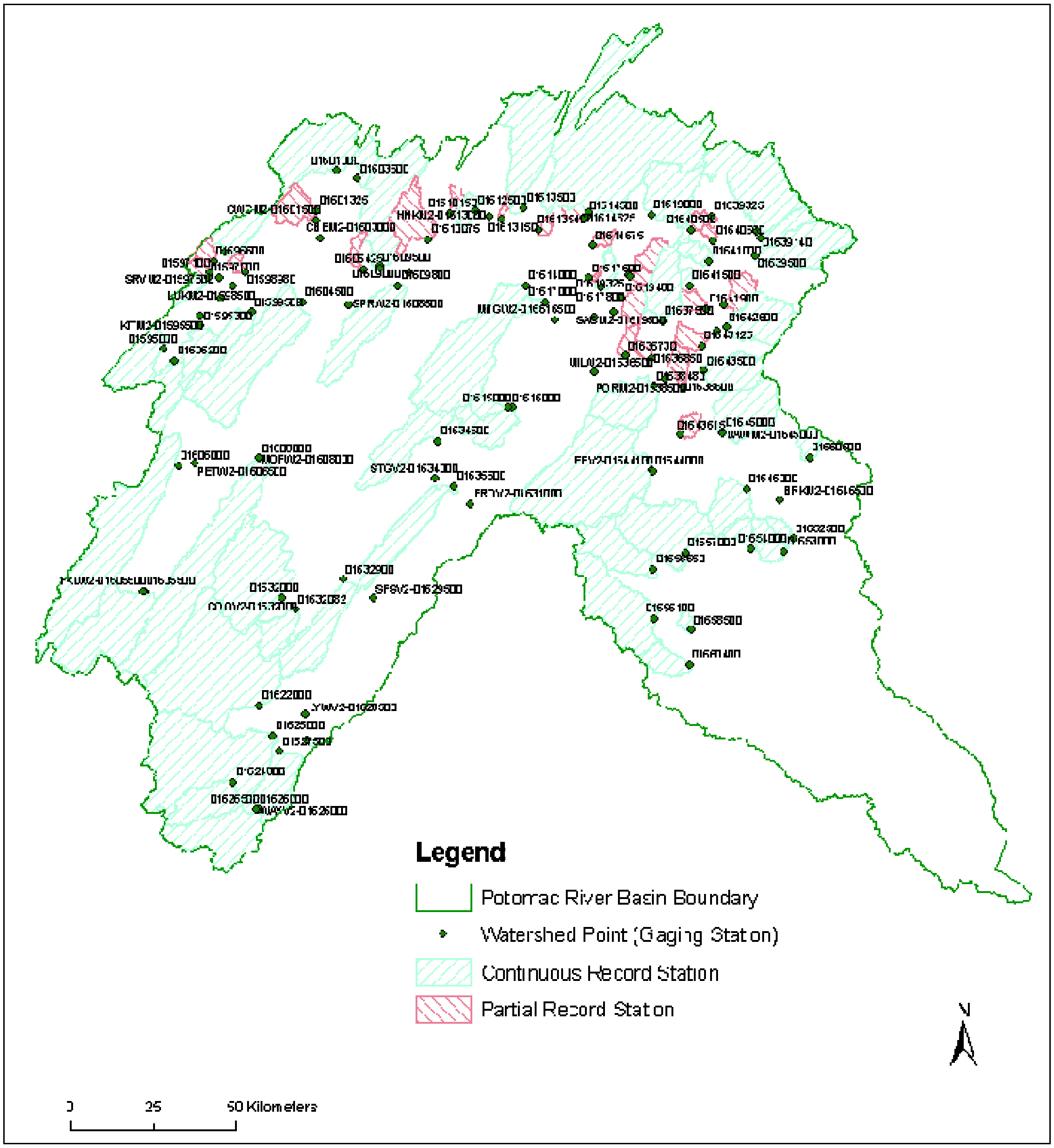


Figure 11. Delineated watersheds and corresponding gage stations

Baseflow Analysis

Currently there are a number of methods to estimate portions of streamflow that constitute baseflow or interflow. Horton (1933) developed a method based on identifying segments of the hydrograph where the streamflow is essentially equal to baseflow. `Once these segments are identified, baseflow during excess runoff is computed by connecting the segments and estimating the departure of the streamflow hydrograph from the depletion curve. Researchers have also used characteristic curves of groundwater discharge combined with other hydrologic and meteorological data (Olmsted and Hely, 1962). There are two commonly used methodologies used in the separation of streamflow hydrograph into baseflow and surface runoff. These methodologies are base-flow-recession methods (Olmsted and Hely, 1962; Riggs, 1963; Rorabaugh, 1963) and curve-fitting methods (Pettyjohn and Henning, 1979; Linsley et al., 1982). The recession methods rely on some form of recession constants that describe the baseflow as a function of time at a given point. The curve-fitting methods usually try to fit a straight line below a streamflow hydrograph and quantify the baseflow quantity as the area below the lines. More recently, new techniques of hydrograph separation have been in use. Two of the new techniques are isotope or tracer based techniques and digital filtering.

Estimation of baseflow for this study was done using the computer program PART (Rutledge, 1998), a recession based method that uses a streamflow partitioning algorithm. The algorithm has antecedent recession requirements during which the groundwater discharge or baseflow is assumed to be equal to the streamflow. During execution of the program the streamflow record is scanned to identify periods that meet the requirement for antecedent recession of streamflow. During the periods of streamflow record that fit the requirement for antecedent recession, the algorithm assumes the groundwater discharge to be equal to the streamflow. For periods of the record where the antecedent streamflow recession (which is defined by a daily decline of more than 0.1 log cycle) does not meet the antecedent recession requirements, a linear interpolation algorithm is used to linearly interpolate the groundwater discharge. The underlying assumptions in the development of PART include a groundwater flow system with diffuse areal recharge to the water table and groundwater discharge to a stream. The method is appropriate if all or most ground water in the basin discharges to the stream and if a streamflow-gage station at the downstream end of the basin measures all or most outflow. Regulation and diversion of streamflow should be negligible.

Flow Frequency Analysis at Gaged and Ungaged Stations

The analysis of baseflow frequency at gage stations is primarily aimed at estimating the annual D-day baseflow statistic. The annual D-day, T-year baseflow statistic Q at a given station is estimated by fitting the Log-Pearson Type III equation, which is the method widely used by the USGS and other federal and state agencies to calculate low-flow statistics. The general procedure in the analysis includes the following;

- Use stations with 10 or more years of record (usually based on climatic year April 1- March 31).
- Determine annual minimum D-day flows and examine suitability for frequency analysis (e.g. Freedom from gross trends).
- Fit logarithms of D-day flows to Pearson Type III distribution.
- Adjust the frequency curve for zero events using conditional probability adjustment.

- Inspect the fitted Log-Pearson curve for adequate fit and adjust graphically, if necessary, using Weibull (m/N+1) plotting positions where m is the rank of the data value and N is the total number of data values used in the analysis.

Selected D-day, T-year low-flow characteristics (for nonzero annual D-day events) are computed from the Log-Pearson Type III equation of the form

$$X_T = \overline{X} + KS \qquad \textbf{(4)}$$

where

X_T = a T year event for Pearson Type II distribution
$\overline{X}$ = mean of the logarithms of the annual D-day events
S = standard deviation of the logarithms of the annual D-day events
K = Pearson Type III frequency factor for a skewness of *G* and recurrence interval of *T* and exceedance probability *p=1/T*

Regional estimates of flow statistics are usually performed using regional statistical regression tools. These methods use basin characteristics to calibrate the regional regression models used to estimate flow statistics for ungaged catchments. In the absence of reliable regression equations, drainage area ratio, regional statistics, or baseflow correlation methods could be used to determine the baseflow statistics of ungaged catchments (Stedinger et al., 1993). Some of the common techniques are described as follows.

Drainage Area Ratio Method: A simple approach, which estimates the flow quantile, y_p , for an ungaged site as

$$y_p = \left(\frac{A_y}{A_x}\right) x_p \qquad \textbf{(5)}$$

where
x_p is the corresponding flow quantile for a nearby gage station
A_x is drainage area of the gage station
A_y is the drainage area of the ungaged site

Regional Statistics Methods: Requires using a gage station record to construct a monthly streamflow record at an ungaged site using

$$y(i,j) = M(y_i) + \frac{S(y_i)\left[x(i,j) - M(x_i)\right]}{S(x_i)} \qquad \textbf{(6)}$$

where
$y(i,j)$ is monthly streamflow at the ungaged site in month *i* and year *j*
$x(i,j)$ is monthly streamflow at the nearby gaged site in month *i* and year *j*

$M(y_i)$ is the mean of the observed flows at the gaged site

$S(y_i)$ is the standard deviation of the observed flows at the gaged site

Baseflow Correlation Methods: If instantaneous or daily average values of baseflow measurements are available in an ungaged site, it is possible to create a correlation with concurrent streamflows at nearby gaged sites for which long flow record is available. It is possible to develop estimators of low flow moments at the ungaged site using bivariate and multivariate regression and estimators of their standard errors.

$$y = a + bx + \varepsilon \qquad \textbf{(7)}$$

with Var(ε) = $S_\varepsilon{}^2$ and estimators of the mean and variance of annual minimum D-day average flows y are

$$M(y) = a + b\,M(x)$$
$$S^2(y) = b^2\,S^2(x) + S_\varepsilon^2$$

where

$M(x)$ is the mean of the annual minimum D-day averages at the gaged *x* site

$S^2(x)$ is estimator of the variance of the annual minimum D-day averages at the gaged *x* site.

The study documented in this report is based on annual baseflow quantities and the subsequent analyses were based on low-flow techniques. Annual 365-day baseflow quantiles of select recurrence intervals were computed using a Windows® based hydrological frequency analysis program. This program, HYFRAN, is specifically designed to solve hydrologic frequency problems by fitting a variety of statistical distributions for both high and low flow analyses. HYFRAN performs statistical analyses of extreme events using a choice of fitting methods which includes Method of Moments (normal/weighted), Method of Moments (WRC, SAM, BOB), Method of Maximum Likelihood, and estimation of quantiles X_T of return period T with confidence intervals. HYFRAN was developed at by Chair's statistical hydrology team at L'Institut National de la Recherche Scientifique, in Quebec Canada.

Low-flow Partial Record Stations Record Extension

In the analysis of flow records, some of the gage stations in the study area may not have sufficient length of continuous record to meet the criteria. In such instances it is customary to extend the available records to a period that satisfies the requirements of length of record. There are a variety of techniques that could be used to extend records and the techniques primarily rely on creating a statistical correlation between the short duration record stations and longer duration record stations in the vicinity. Some of the techniques are;

- Ordinary Least Squares (OLS) regression

- OLS regression plus noise
- Maintenance of Variance Extension (MOVE) i.e., MOVE 1, MOVE 2, MOVE 3, MOVE 4

Maintenance of Variance Extension (MOVE 1) (Hirsch, 1982) is a statistical method that gives estimates of D-day, T-year low flows at partial record stations. The MOVE 1 technique assumes that a linear relation exists between the concurrent flows at the short and long-term stations. Because streamflow data are highly skewed, a log transformation is commonly done in order to linearize the data. Once linearity is confirmed, the means and standard deviations of the logs of the concurrent streamflow data are calculated. The MOVE 1 equation is then written as follows:

$$\hat{Y}_{D,T} = \bar{Y} + \frac{S_Y}{S_X}\left(X_{D,T} - \bar{X}\right) \qquad \textbf{(8)}$$

where:

$\hat{Y}_{D,T}$ = D-day, T-year low flow at the partial record site for MOVE 1

$X_{D,T}$ = D-day, T-year low flow at the index station

$\bar{Y}$ and $\bar{X}$ are mean of baseflows and concurrent daily flows

S_Y and S_X are standard deviations of baseflows and concurrent daily flows

Partial record stations are sites where only baseflow measurements are made. The general procedure for estimation of low-flow characteristics at partial record stations involves:

- Developing a relationship (usually in log units) between baseflow measurements at partial record stations and concurrent daily flows at nearby continuous record stations (index stations).
- Define the relationship using graphical or least squares regression analysis.
- Use the relation and low-flow statistics at the nearby index stations to estimate the desired characteristics at the partial record stations.

Table 3. Methods for estimations of flow characteristics at partial record stations

Number of Discharge Measurements Available	Method
0	Regional regression of $Q_{D,T}$ on basin characteristics
1 to several	Control-point graphical method or Discharge ratio method
3 to several	Graphical method
10 to several	Stedinger-Thomas regression method

The analysis of low-flow partial record stations requires using one or more continuous record gage stations and transferring the statistical parameters of the continuous gage stations to the low-flow partial record station. The main assumption involved in the analysis is that the low-flow partial records are actually the baseflow records for that period. Prior to using record extension the stations have to be screened to make sure that this assumption is justified. In order to test the records it is important to screen the records by comparing them with the concurrent continuous record gage station data. The data need to be screened to make sure that there is satisfactory correlation between the index station and the partial record station. The screening methods are a mix of graphical and statistical tools that help examine the available data. These methods include;

1. Plotting the partial records and comparing them with the hydrograph of the continuous records.
2. Correlation tests for both linear and log-transformed records of the partial record and continuous record concurrent data.
3. Testing the correlation for low mean square error values.
4. Data which fail to correlate well with the continuous record might not be baseflow records as assumed and might have to be excluded from further analysis

From the published annual low-flow partial records for the study area 15 stations were found to satisfy the above criteria and the availability of an index station in the proximity. Correlation plots and data statistic were computed for both the partial record and continuous flow record stations and MOVE 1 method was used to transfer records from index stations to partial record stations.

Table 4. Correlation table for partial record and index stations

Partial Record Station ID	Concurrent Index Station ID	R^2
01596600	01597000	0.8566
01597100	01597000	0.9193
01598980	01596500	0.9227
01601325	01596500	0.9064
01605425	01603500	0.8785
01617600	01617800	0.9306
01619325	01619000	0.9552
01619350	01619000	0.8985
01619480	01617800	0.7871
01636730	01637500	0.8962
01636850	01637500	0.9064
01638600	01641500	0.7789
01641900	01639000	0.8287
01642050	01639000	0.7899
01643125	01641500	0.8062

Selection of Explanatory Variables

Prior to selection of the significant variables, preliminary analyses were conducted by applying different techniques. The primary goal of the analyses was determining the existence of statistically meaningful relationships between response variables and explanatory variables and identifying explanatory variables that could be used in the development of the final regression equations. The analyses performed included;

- Inspection of matrix plots of all variables
- Conducting stepwise regression
- Variable transformation and simple regression fits

In addition to the existing variables, some derived variables were also created. For example the carbonates and silisiclastics in the basin are subdivided on the basis of the physiographic region they are found. In order to create smaller number of parameters for the analysis all the carbonates were lumped into one. Similarly the silisiclastics were also categorized into one group irrespective of their physiographic region.

Table 5. List of possible regression parameters

INDEPENDENT PARAMETERS					**DEPENDENT PARAMETERS**
GEOLOGY (NAWQA) (%) Area	**HGMR (HYDROGEOMORPHIC REGIONS) (%) Area**	**LANDUSE 1997 (%) Area**	**SOIL (STATSGO HYDGRP) (%) Area**	**Hydrologic and Basin Physical Parameters**	
Silisiclastic (**GEOL1**) Carbonate Silisiclastic Unidivided (**GEOL2**) Carbonate (**GEOL3**) Crystalline (**GEOL4**) Unconsolidated Sediments (**GEOL5**)	Valley and Ridge Carbonates :**HGMRVRC** Valley and Ridge Silisiclastics :**HGMRVRS** Piedmont Crystalline :**HGMRPCR** Mesozoic Lowlands :**HGMRML** Piedmont Carbonates :**HGMRPCA** Blue Ridge :**HGMRBR** Appalachian Silisiclastics :**HGMRAPS** Appalachian Carbonates :**HGMRAPC** combined HGMRs Carbonates :**HGMRC**	Low Intensity Residential/High Intensity Residential/Commercial/Industrial/Trans (**LAND1**) Bare Rock/Sand/Clay/Quarries/Strip Mines/Gravel/Transitional Barren (**LAND2**) Deciduous Forest/Evergreen Forest/Mixed Forest (**LAND3**) Woody Wetlands/Emergent Wetlands (**LAND4**) Pasture/Hay/Row Crops/Other Grasses (**LAND5**)	A: High Infiltration (**SOIL1**) B: Moderate Infiltration (**SOIL2**) C: Slow Infiltration (**SOIL3**) D: Very Slow Infiltration (**SOIL4**) B/D: Drained/Undrained (**SOIL5**)	Drainage Area **(DAREA)** Mean Slope (**SLOPE.MEAN**) Average Precipitation (1983-2004) (in) (**PRCP.AVE**) Average Potential Evapotranspiration (1983-2004) (in) (**PET.AVE**) Average Dryness Index (1983-2004) (**DRYNESS.AVE**)	Annual Baseflow (cfs) for 2, 5, 10, and 20 years recurrence **Q365,2, Q,365,5, Q,365,10, Q,365,20**

Matrix plots, exploratory linear regression fits, and a series of stepwise regression analyses resulted in a pool of variables which provided discernible relationships with baseflow estimates of 2,5,10 and 20 years recurrence interval. The data transformation involved logarithmic transformation and scaling of basin characteristics in each watershed. The percent value of basin characteristic in each watershed was converted to percent coverage *0.01 +1 prior to log-transformation. This transformation helps eliminate the problems encountered in using zero percentage values. Matrix plots of parameters are presented in Appendix A.

Data transformation facilitates the creation of recognizable correlation if the data do not correlate well in linear space. Log transforming the data frequently results in a better correlation with the dependent variable. Transformation also helps reduce the effects of heteroscedasticity, which is a common problem in hydrologic data. In heteroscedastic data the variance of the residuals is a function of the explanatory variables, therefore, the absolute magnitude of the variance along the regression line increases as the magnitude of the dependent variable increases or vice versa. The automated stepwise regression procedure calculates Mallow's C_p statistics for the current model as well as those for all reduced and augmented models, then adds or drops the term that reduces C_p the most.

Table 6. Total percent coverage of parameters in the study watersheds

PARAMETER TYPE	PARAMETER NAME	% Area
GEOLOGY NAWQA	Silisiclastic (GEOL1)	4.8
	Carbonate Silisiclastic Unidivided (GEOL2)	41.9
	Carbonate (GEOL3)	35.6
	Crystalline (GEOL4)	6.2
	Unconsolidated Sediments (GEOL5)	11.5
HGMR Hydrogeomorphic Regions	Valley and Ridge Carbonate (HGMRVRC)	20.1
	Valley and Ridge Siliciclastic (HGMRVRS)	43.6
	Piedmont Crystalline (HGMRPCR)	14.3
	Mesozoic Lowlands (HGMRML)	8.2
	Piedmont Carbonate (HGMRPCA)	0.5
	Blue Ridge (HGMRBR)	7.4
	Appalachian Plateau Siliciclastic (HGMRAPS)	5.7
	Appalachian Plateau Carbonate (HGMRAPC)	0.1
STATSGO Soil Type HYDGRP Classes	A: High Infiltration (SOIL1)	4.8
	B: Moderate Infiltration (SOIL2)	41.9
	C: Slow Infiltration (SOIL3)	35.6
	D: Very Slow Infiltration (SOIL4)	6.2
	B/D: Drained/Undrained (SOIL5)	11.5

Correlated Regressors

Statistical analysis that involves estimation of parameters using regressors is best performed when the regressors are orthogonal (not correlated with each other). With orthogonal regressors, the parameter estimate for a given regressor does not depend on which other regressors are included in the model, although other statistics such as standard errors and p-values may change. If the regressors are correlated, it becomes difficult to make a clear distinction between the effects of one regressor and another, and the parameter estimates may be highly dependent on which regressors are used in the model. Two correlated regressors may be non-significant when tested separately but highly significant when considered together. If two regressors have a correlation of 1.0, it is impossible to separate their effects. It may be possible to recode correlated regressors to make interpretation easier. For example, if X and Y are highly correlated, they could be replaced in a linear regression by X+Y and X-Y without changing the fit of the model or statistics for other regressors. In this study correlation between geology and HGMR parameters was anticipated due to the fact that HGMRs are derived from geology and physiography. As a result, correlation tests were performed and the results are presented in Table 7. Based on the results there is a medium to high correlation between geology and HGMR, therefore redundancy needs to be minimized by removing parameters which are highly correlated but not as statistically significant during the stepwise regression.

Table 7. Correlation between geology and HGMR parameters

	HGMRVRC	HGMRVRS	HGMRPCR	HGMRML	HGMRPCA	HGMRBR	HGMRAPS
GEOL1	-0.24	0.51	-0.54	0.12	-0.10	-0.42	0.55
GEOL2	0.01	0.43	-0.24	-0.15	-0.10	-0.21	0.09
GEOL3	0.92	0.04	-0.34	-0.16	0.07	-0.05	-0.26
GEOL4	-0.40	-0.60	0.76	0.07	0.08	0.50	-0.33
GEOL5	-0.09	0.20	0.03	-0.08	-0.06	-0.12	-0.09

After a series of tests, the HGMRs were found to consistently be better performing. The following parameters were identified to be significant in most of the recurrence intervals tested, and were included in the development of the final equation:

- Drainage Area
- Dryness Index
- Combined Silisiclastics
- Piedmont Crystallines
- Mesozoic Lowlands
- Blue Ridge

Development of a Least Squares Statistical Model

The techniques of multiple regression enable estimation of values for dependent parameters when the explanatory variables are many in number. The general category of multiple regression techniques primarily utilizes the Ordinary Least Squares (OLS) estimator techniques to fit values and solve the equations. A general linear regression model for response variable *y* and explanatory variables *x* has the form:

$$y_i = \beta_0 + \sum_{j=1}^{n} \beta_j x_{ij} + \varepsilon_i \qquad \textbf{(9)}$$

where the error term, ε_i, is assumed to be independent and identically distributed, the values of ε_i have mean zero and finite variance σ^2, and the ε_i are normally distributed. When the elements of the response variable have unequal variances and/or are correlated, σ^2 is no longer a scalar variance-covariance matrix, and hence there is no guarantee that the OLS estimator is the most efficient within the class of linear unbiased (or the class of unbiased) estimators. The method of generalized least squares (GLS) is introduced to improve upon estimation efficiency when σ^2 is not a scalar variance-covariance matrix.

Generalized least squares models are regression (or ANOVA) models in which the errors have a nonstandard covariance structure. Like simple least squares regression, the GLS method uses maximum likelihood or restricted maximum likelihood to fit a continuous, univariate response as a linear function of a single predictor variable. In GLS, however, the errors are allowed to be correlated and/or to have unequal variances. Without loosing the benefit of the generalized assumptions with regard to the variance structure of the generalized nonlinear regression model the general form of the equation that relates baseflow to basin characteristics parameters and hydrologic parameters is of the form:

$$\log(Q_T) = \beta_0 + \beta_1 \log(DA) + \beta_2 \log(DI) + \sum_{i=3}^{n} \beta_i \log(HRC)_i \qquad \textbf{(9)}$$

where Q_T is the annual baseflow estimate of T years recurrence interval, *DA* is the drainage area of the watershed, *DI* is the long-term averaged dryness index of the watershed, HRC is the coefficient derived from percentage of each HGMR covering the area transformed to HGMR*0.01+1.

Results

The primary result of the study is watershed baseflow estimates for the watersheds that have continuous record, extended record, or partial record data. The estimates are presented in Table 8. The equations were developed for 2, 5, 10, and 20 years recurrence interval.

The final equation was found to be:

$$log(Q_T) = \beta_0 + \beta_1 * log(DAREA) + \beta_2 * log(DRYNESS.AVE) + \beta_3 * log(HGMRS * 0.01 + 1) + \beta_4 * log(HGMRPCR * 0.01 + 1) + \beta_5 * log(HGMRML * 0.01 + 1) + \beta_6 * log(HGMRBR * 0.01 + 1)$$

where DAREA is drainage area (sq. mi.), DRYNESS.AVE average dryness of the watershed, HGMRS is percentage of combined siliciclastic HGMR, HGMRPCR is percentage of Piedmont Crystallines, HGMRML is percentage of Mesozoic Lowlands, HGMRBR is percentage of Blue Ridge.

Performing linear regression in S-PLUS provides a detailed summary of the fit, including information on quantiles of the residuals, coefficients, standard errors of the coefficients, t-statistics testing whether the coefficient is significantly different from zero, and p-values for such test. It also calculates the residual standard error and F-statistic, testing whether at least one of the coefficients, excluding the intercept, are significantly different from zero, with the corresponding p-value. The results are presented in Table 9.

Table 8. List of stations and baseflow estimates

Station ID	DAREA (sq.mi.)	$Q_{365,2}$ (cfs)	$Q_{365,5}$ (cfs)	$Q_{365,10}$ (cfs)	$Q_{365,20}$ (cfs)
01595000	70.77	108.24	94.79	89.14	85.08
01595200	47.31	57.33	43.08	36.59	31.74
01595300	41.56	45.46	37.92	34.34	31.56
01596500	48.06	43.58	37.50	34.95	33.11
01596600	12.99	15.72	11.74	9.96	8.65
01597000	16.54	19.20	16.15	14.65	13.47
01597100	10.68	32.37	20.28	15.58	12.42
01598980	7.10	2.94	2.52	2.30	2.14
01599500	47.06	30.81	24.23	20.09	16.63
01601000	145.33	103.39	89.08	82.14	76.68
01601325	38.13	19.97	15.17	12.99	11.37
01603500	27.59	21.58	17.02	14.89	13.26
01604500	218.04	93.17	66.25	55.50	47.95
01605425	10.49	2.24	1.80	1.59	1.43
01605500	171.00	108.82	89.34	80.03	72.81
01606000	306.20	219.51	188.81	172.47	159.03
01608000	273.02	125.56	98.79	87.29	78.85
01609000	148.82	102.48	76.91	63.80	53.59

Station	DAREA	$Q_{365,2}$	$Q_{365,5}$	$Q_{365,10}$	$Q_{365,20}$
ID	(sq.mi.)	(cfs)	(cfs)	(cfs)	(cfs)
01609500	5.02	1.66	1.33	1.15	1.00
01609800	107.68	42.04	28.98	21.40	16.64
01612500	16.96	7.95	6.53	5.83	5.26
01613500	157.73	95.06	72.63	61.05	51.94
01614000	234.84	100.65	76.04	64.65	56.07
01614500	496.39	383.53	296.26	256.93	227.52
01615000	57.11	20.71	14.99	12.7	11.1
01616000	18.32	14.32	11.75	10.39	9.29
01617000	11.87	15.10	9.03	6.21	4.30
01617600	2.25	5.29	4.32	3.86	3.50
01617800	18.57	12.09	7.912	6.05	4.726
01619000	94.21	97.73	71.35	58.22	48.18
01619325	22.84	11.05	10.03	9.49	9.06
01619350	8.82	2.48	2.15	1.98	1.85
01619480	24.96	34.28	28.33	25.44	23.19
01622000	375.33	235.78	187.06	164.21	146.73
01624800	73.15	51.73	36.90	29.78	24.44
01625000	370.90	214.24	156.60	130.50	111.16
01626000	125.61	103.94	76.83	64.92	56.18
01626500	132.12	105.08	79.79	68.64	60.40
01627500	205.04	167.54	130.46	113.50	100.72
01628060	2.10	1.35	0.85	0.61	0.44
01632000	210.20	87.79	65.00	54.76	47.17
01632082	44.59	23.84	16.21	13.12	10.97
01632900	94.90	51.74	36.04	29.26	24.39
01634500	101.88	54.88	41.21	35.36	31.12
01635500	86.51	39.78	28.95	24.26	20.84
01636730	13.12	3.97	3.41	3.13	2.91
01636850	8.56	1.58	1.30	1.17	1.06
01637500	67.31	49.91	36.14	30.36	26.21
01638480	87.13	54.32	37.18	29.91	24.74
01638600	20.25	22.45	19.91	18.61	17.55
01639000	170.04	73.93	57.23	50.14	44.96
01639140	31.00	16.86	11.34	9.06	7.46
01639500	98.41	69.41	51.29	43.37	37.56
01640500	6.12	6.88	5.18	4.39	3.80
01641000	18.70	18.75	13.95	11.73	10.06
01641500	7.27	10.22	7.62	6.40	5.47
01641900	15.74	5.31	4.81	4.55	4.33
01642050	28.37	6.88	6.11	5.71	5.39
01642500	81.53	56.14	42.09	35.85	31.25
01643125	19.69	17.18	15.66	14.86	14.21
01643500	63.13	44.82	32.69	27.54	23.83
01644000	329.98	180.39	127.62	105.12	88.94
01645000	100.39	68.03	50.39	42.74	37.16
01646000	57.81	35.03	26.87	23.21	20.49
01650500	21.13	12.67	9.49	8.05	6.99
01652500	13.97	5.55	4.70	4.34	4.07

Station ID	DAREA (sq.mi.)	$Q_{365,2}$ (cfs)	$Q_{365,5}$ (cfs)	$Q_{365,10}$ (cfs)	$Q_{365,20}$ (cfs)
01653000	33.85	14.01	11.26	10.00	9.05
01654000	23.88	10.91	7.64	6.15	5.06
01656100	154.02	54.80	36.02	27.66	21.70
01656650	89.58	43.24	26.07	18.97	14.17
01657000	147.29	46.71	33.67	28.01	23.89
01658500	7.66	3.04	2.14	1.74	1.44
01660400	35.31	18.44	13.12	10.72	8.95

Table 9. Baseflow equations' coefficients and statistical fitness test values

Annual baseflow recurrence interval	Coefficient value and Test Statistic	Equation Coefficients						
		β_0	β_1	β_2	β_3	β_4	β_5	β_6
2-years	value	-0.2555	0.9814	-2.4386	-2.0826	-1.6339	-1.9533	-1.1494
	t-value	-0.8766	20.4080	-5.2907	-5.0359	-3.5509	-5.3268	-2.7059
	p-value	0.3839	0.0000	0.0000	0.0000	0.0007	0.0000	0.0087
5-years	value	-0.5142	0.9810	-2.3869	-2.0036	-1.6832	-1.9920	-1.0967
	t-value	-1.8475	21.3640	-5.4234	-5.0741	-3.8311	-5.6893	-2.7039
	p-value	0.0692	<.0001	<.0001	<.0001	0.0003	<.0001	0.0087
10-years	value	-0.6835	0.9833	-2.3569	-1.9461	-1.7022	-1.9941	-1.0559
	t-value	-2.4298	21.1861	-5.2981	-4.8758	-3.8331	-5.6345	-2.5756
	p-value	0.0178	<.0001	<.0001	<.0001	0.0003	<.0001	0.0123
20-years	value	-0.8428	0.9869	-2.3309	-1.8908	-1.7151	-1.9869	-1.0155
	t-value	-2.8877	20.4934	-5.0501	-4.5659	-3.7222	-5.4111	-2.3874
	p-value	0.0052	<.0001	<.0001	<.0001	0.0004	<.0001	0.0198

Flow prediction was conducted for a hypothetical watershed with 100% of a given HGMR at 95% confidence interval. As shown in Table 10 the percentage of Blue Ridge appears to yield consistently higher values of baseflow compared to the others.

Table 10. Predicted Flows with 95% confidence interval for a basin with 30 sq. mi. drainage area and dryness index of 0.15

Predicted Annual Baseflows (in/yr)	Hydrogeomorphic Regions			
	100% Siliciclastics	100% Piedmont Crystallines	100% Mesozoic Lowland	100% Blue Ridge
2-years	5.95	6.19	6.81	7.88
5- years	4.51	4.52	4.96	5.92
10- years	3.79	3.73	4.08	4.95
20- years	3.23	3.14	3.41	4.21

Performance Metrics

The performance metrics reported in this section serve as measures that test how well the model fits the data and how well it might perform during prediction. The tests have been used in a number of articles, such as the study by Reilly and Kroll (2003), where it was used to measure the performance of their statistical model. The model was used for the estimation of 7-day, 10-year low-streamflow statistics using baseflow correlation.

Average relative absolute difference (ARAD):

$$ARAD = \frac{\sum_{i=1}^{N}\left(\frac{\left|\hat{Q}-Q\right|}{Q}\right)}{N}$$

Relative bias (R-BIAS):

$$R-BIAS = \frac{\sum_{i=1}^{N}\left(\frac{\hat{Q}-Q}{Q}\right)}{N}$$

Relative mean square error (R-MSE):

$$R-MSE=\frac{\sum_{i=1}^{N}\left(\frac{\hat{Q}-Q}{Q}\right)^{2}}{N}$$

where $\hat{Q}$ is the estimate of the Q at the ungaged site, Q is the ‘‘true value’’ of the Q at the ungaged site, and M is the number of baseflow segments with an associated $\hat{Q}$. Q is the at-site estimator obtained using the entire historic record at the site, fitting a log-Pearson type III distribution by method of moments (Stedinger et al., 1993), and estimating the tenth percentile of the distribution. The ARAD performance metric is primarily used to discuss the results of this experiment. The ARAD measures the average percent deviation of the Q estimator, and thus is easily interpretable. For example, an ARAD of 0.05 indicates a 5% error on average, and an ARAD of 1.00 indicates a 100% error on average. R-BIAS and R-MSE are also included to further assess method performance.

Table 11. Performance metrics of the model

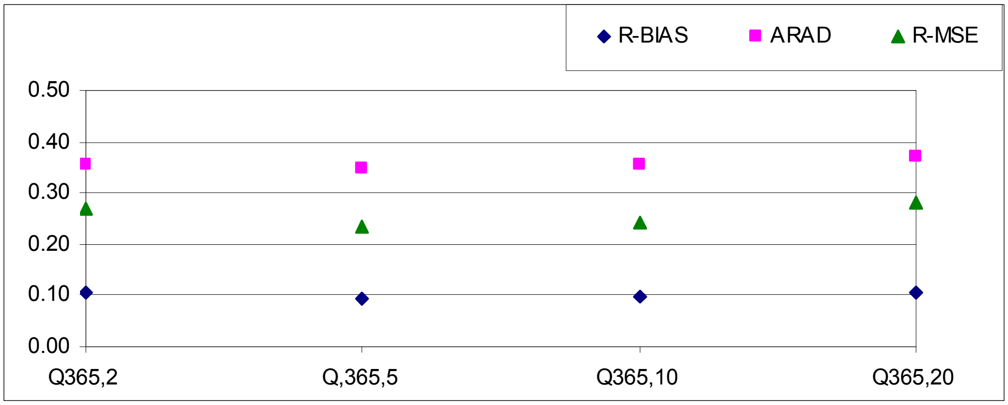


Table 12. S-PLUS estimated model error terms

Error terms	Annual Baseflow Recurrence Interval			
	2 yrs	5 yrs	10 yrs	20 yrs
estimated residuals	-0.41807	-0.74402	-0.93359	-1.07179
error sum of squares SSE	2.383473	2.265976	2.364755	2.580108
mean square error MSE	0.03357	0.031915	0.033306	0.03634
standard error of regression OR standard deviation of residuals	0.183221	0.178648	0.1825	0.190629

Application of Artificial Neural Networks

Artificial neural networks are computational models that are loosely based on the neuron cell structure of the biological nervous system. Given a training set of data, a neural network can learn the data with a learning algorithm such as the most common algorithm, backpropagation. Through the learning algorithm, the neural network forms a mapping between inputs and desired outputs from the training set by altering weighted connections within the network. Feed-forward neural networks provide a computationally robust method to generalize and model linear regression functions.

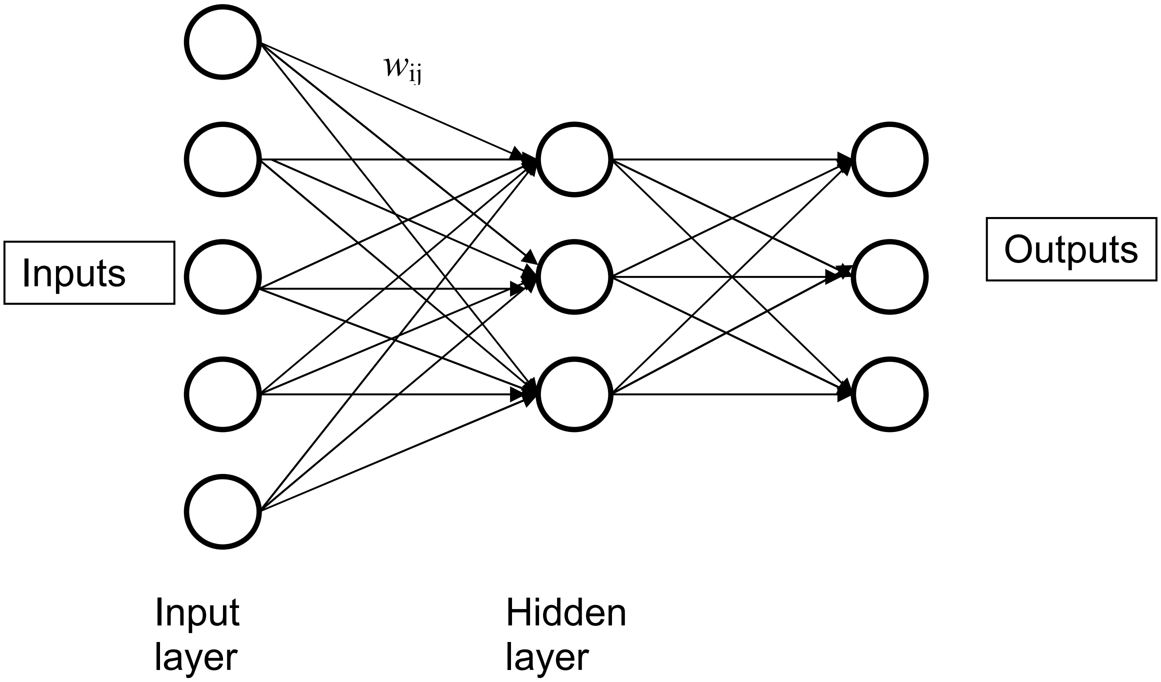


Figure 12. Representation of simple feed forward neural network system

For a simple one hidden layer feed forward neural network as shown in Figure 12, the input units distribute the inputs into the hidden layer. These units sum their inputs, add a constant (the bias) and take a fixed function ϕ_h of the result. The output units are of the same form, but with output function ϕ_o. Thus the formulation becomes (Venables and Ripley, 2002):

$$y_k = \phi_0\left(\alpha_k + \sum_h w_{hk}\phi_h\left(\alpha_h + \sum_h w_{ih}x_{ih}\right)\right)$$

and the activation function ϕ_h of the hidden layer is customarily a logistic function of the form:

$$l(z) = \frac{\exp(z)}{1+\exp(z)}$$

Current application of artificial neural networks in this study was limited to one hidden layer feed-forward network. In order to achieve perform the computation a neural networks algorithm known as *nnet* was programmed into S-PLUS. The results showed that the model performed better during training phase and was not very good during forecasting. The performance of the model during forecasting did not encourage further testing. The inadequate performance for such a robust modeling approach is due to limitations of data used for training. The available data was used by applying a split sampling where a portion of the data is used for training and the remaining of the data is used for forecasting. This resulted in insufficient data for training and forecasting. However, with a larger data size this method could be investigated as an alternative and novel approach. The plots in Figure 13 depict the performance of the model during training and forecasting for the four recurrence intervals.

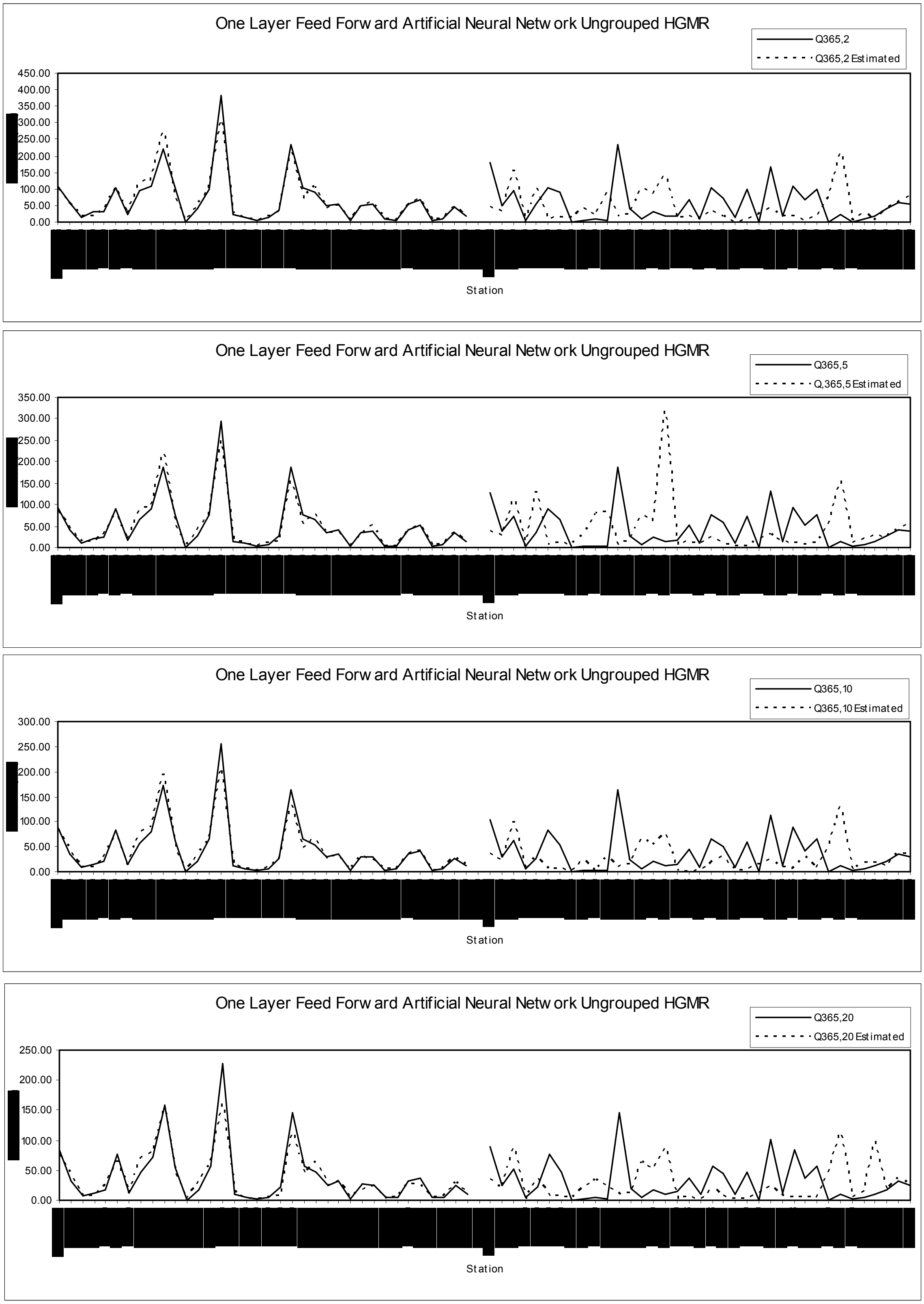


Figure 13. Training and prediction results of a simple feed forward neural network system

Summary and Conclusions

The results of this study were primarily based on the inherent model selection criteria of the modeling environment S-PLUS which provides a robust stepwise regression procedure. As described in the documentations of S-PLUS one of the major statistics of concern in stepwise processes is Cp, which is optimized by a series of iterations. This is consistent with the common practice used in the selection of regression models in hydrological analyses.

From hydrological point of view, Carbonates, especially the ones with interconnected pores, are important geologic formations. However, in this study the carbonates failed to make it to the final list. The requirement was each formation has to be consistently significant in the 2, 5, 10, and 20 years recurrence interval. Some of the Carbonate groups and the combined Carbonates made it in one or two recurrence intervals but not consistently in all. As a result they were dropped from the final equation.

The results are constrained by the sparse coverage of gage stations and the impacts of regulation which further reduced the number of stations that could be used. Given the size of the area to be modeled and the hydrogeomorphological diversity, it would have been desirable to have additional flow data, especially in the Appalachian, Piedmont, and Valley and Ridge Carbonate regions. This could be one plausible explanation for the failure of Carbonates to appear in the final equations.

The Blue Ridge, Mesozoic Lowlands, Piedmont Crystallines, and Siliciclastics have been identified as the significant HGMR units that contribute to generation of baseflow in the study area. During the initial phase of the study the idea of separating the basin into homogeneous regions with similar hydrogeomorphic, soil, or geologic units was entertained. However that turned out to be very difficult due to the existence of two or more units in any given region.

Other hydrologically important parameters such as landuse and soil infiltration property were excluded in the preliminary stepwise regression model development phase due to their statistical insignificance. This reflects the fact that no weighting was added based on known hydraulic properties. The unavailability of a method that gives more weight to hydrologically important landuse or soil types on the basis of their hydraulic conductance is also considered to be a contributing factor for their poor performance in the stepwise process.

This study could be particularly useful if it is combined with more extensive water usage data, as it can help in the identification of hotspots and facilitate future planning efforts. There is room for improvement of the model by increasing the number of continuous gages and their distribution in some of the hydrogeomorphic regions. Statistical analysis of baseflow provides an alternative to a computationally and data intensive 3-dimensional groundwater flow modeling in the assessment of water availability.

Outcomes and Opportunities

The primary outcome of this study is the set of equations that could be used to forecast baseflow quantiles of 2, 5, 10, and 20 years recurrence interval in ungaged catchments. The process of generating the data required to create a list of possible descriptor parameters resulted in a geo-database that includes the following:

- Digital Elevation and derived parameters
- Delineated watersheds
- STATSGO soil hydraulic properties
- Geology and Hydrogeomorphology
- Precipitation, Potential Evapotranspiration, and Dryness Index (1984-2003)
- Continuous record gage stations
- Baseflow estimates

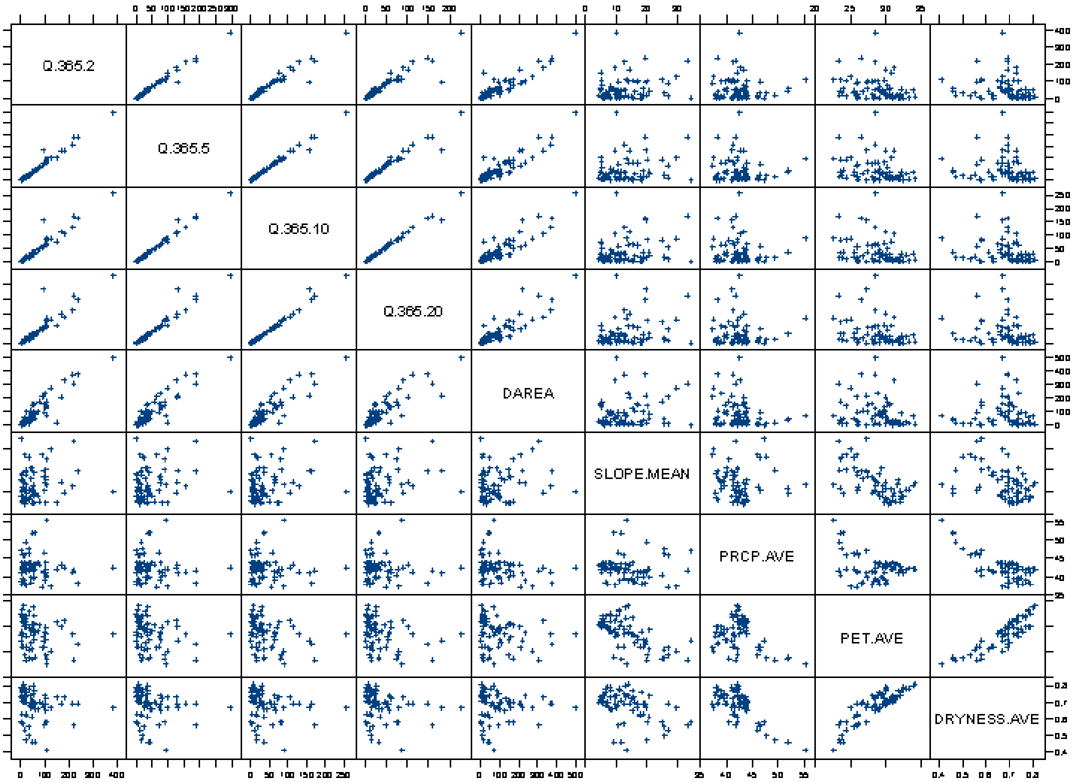
Also some precipitation data were downloaded from the PRISM website. PRISM (Parameter-elevation Regressions on Independent Slopes Model) is an analytical tool that uses point data, a digital elevation model, and other spatial data sets to generate gridded estimates of monthly, yearly, and event-based climatic parameters, such as precipitation, temperature, and dew point. (PRISM is a project undertaken by the Oregon State University Spatial Climate Analysis Service.)

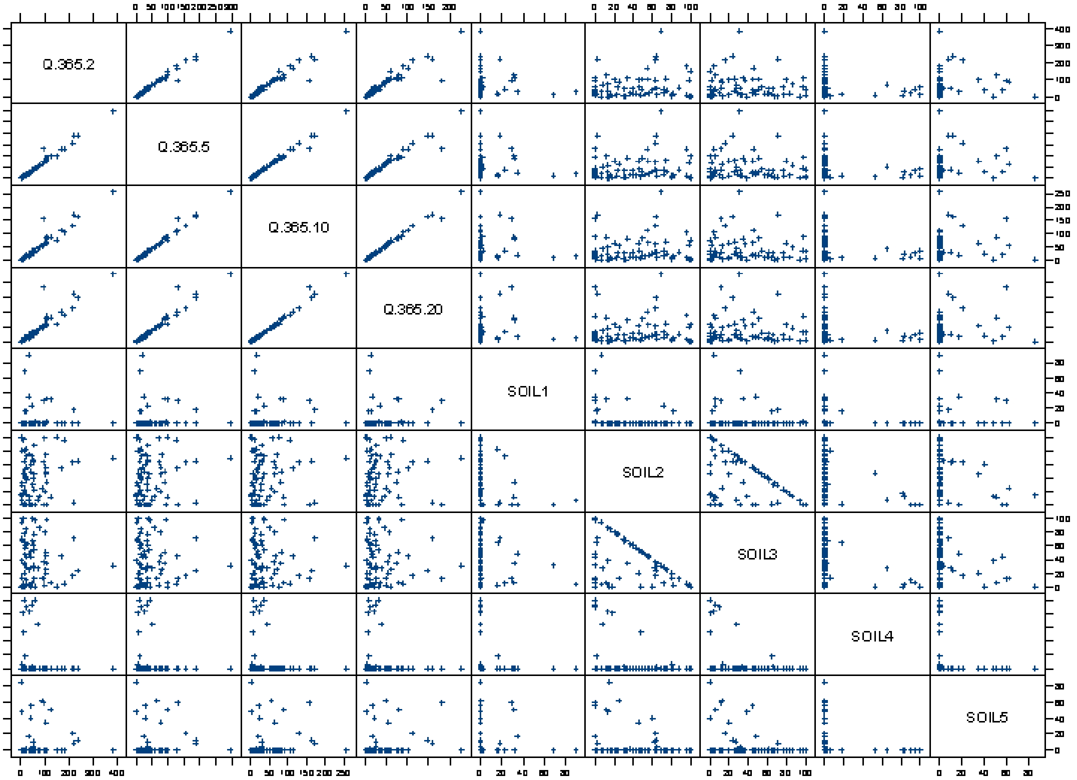
Data Resolution and type

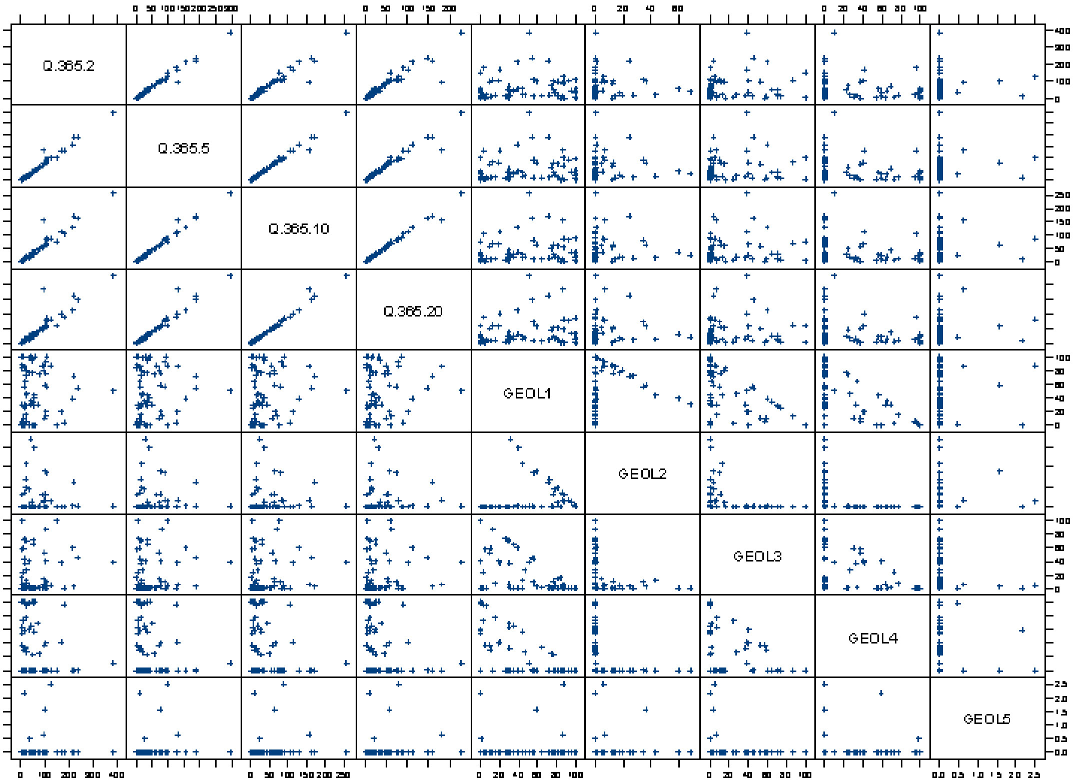
For future analysis it is possible to conduct a more intensive and robust study utilizing the spatially distributed hydrological parameters of PRISM and generate baseflow and other indices that are based on the geospatial data available. One suggested method that makes use of the PRISM data and the available geological, soil, landuse/landcover, hydrogeomorphic, and physiographic data is:

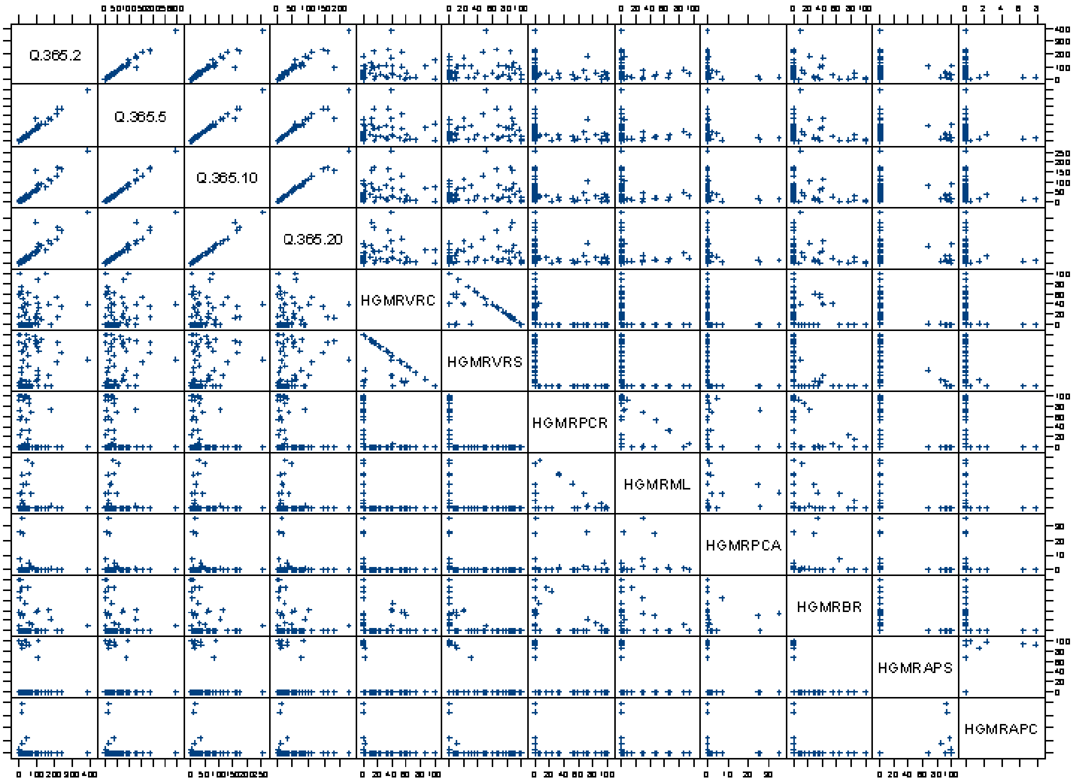
- Create a raster coverage of precipitation
- Calculate potential evapotranspiration using Thornthwaite’s equation
- Using basin characteristics data and hydrologic data perform GIS analysis and calculate raster–based estimates of recharge for the basin
- Using the basin-wide estimates of recharge, basin hydrogeology data, and consumptive use data, calibrate and develop a basin-wide numerical groundwater flow model
- Using the basin-wide data calibrate distributed or lumped physically based surface water models that could be used to conduct surface water-groundwater interaction studies

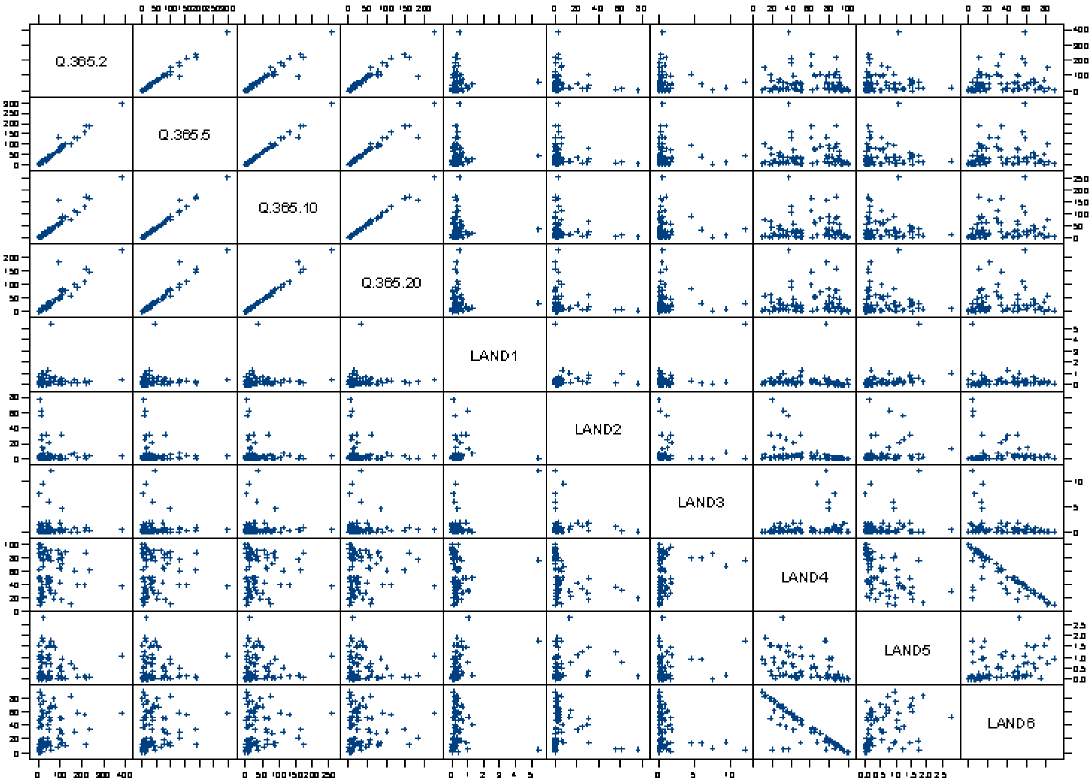
Appendix A: Matrix plots of parameters


Q.365.2
Q.365.5
Q.365.10
Q.365.20
DAREA
SLOPE.MEAN
PRCP.AVE
PET.AVE
DRYNESS.AVE
0 50 100 150 200 250 300
0 50 100 150 200
0 10 20 30
20 25 30 35
400 300 200 100 0
250 200 150 100 50 0
500 400 300 200 100 0
55 50 45 40 35
0.8 0.7 0.6 0.5 0.4
300 250 200 150 100 50 0
200 150 100 50 0
30 20 10 0
35 30 25 20
0 100 200 300 400
0 50 100 150 200 250
0 100 200 300 400 500
35 40 45 50 55
0.4 0.5 0.6 0.7 0.8


Q.365.2
Q.365.5
Q.365.10
Q.365.20
SOIL1
SOIL2
SOIL3
SOIL4
SOIL5
0 50 100 150 200 250 300
0 50 100 150 200
0 20 40 60 80 100
0 20 40 60 80 100
400 300 200 100 0
250 200 150 100 50 0
80 60 40 20 0
100 80 60 40 20 0
80 60 40 20 0
300 250 200 150 100 50 0
200 150 100 50 0
100 80 60 40 20 0
100 80 60 40 20 0
0 100 200 300 400
0 50 100 150 200 250
0 20 40 60 80
0 20 40 60 80 100
0 20 40 60 80


Q.365.2
Q.365.5
Q.365.10
Q.365.20
GEOL1
GEOL2
GEOL3
GEOL4
GEOL5


Q.365.2
Q.365.5
Q.365.10
Q.365.20
HGMRVRC
HGMRVRS
HGMRPCR
HGMRML
HGMRPCA
HGMRBR
HGMRAPS
HGMRAPC


Q.365.2
Q.365.5
Q.365.10
Q.365.20
LAND1
LAND2
LAND3
LAND4
LAND5
LAND6
0 50 100 150 200 250 300
0 50 100 150 200
0 20 40 60 80
0 20 40 60 80 100
0 20 40 60 80
400
300
200
100
0
250
200
150
100
50
0
5
4
3
2
1
0
10
5
0
2.5
2.0
1.5
1.0
0.5
0.0
0 100 200 300 400
0 50 100 150 200 250
0 1 2 3 4 5
0 5 10
0.0 0.5 1.0 1.5 2.0 2.5

References

Bachman, J., Lindsey, B., Brakebill, J., Powars, D. 1998. Ground-Water Discharge and Base-Flow Nitrate Loads of Nontidal Streams, and Their Relation to a Hydrogeomorphic Classification of the Chesapeake Bay Watershed, Middle Atlantic Coast. U.S. Geological Survey Water-Resources Investigations Report 98-4059.

Fetter, C.W. 2000. Applied Hydrogeology. Prentice-Hall.

Budyko, M.I. (1974). Climate and Life, Academic Press, New York, 508 pp.

Derosier, Brakebill, Denis, & Kelley. 1998. Water Quality Assessment of the Potomac River Basin: Water Quality and Selected Spatial Data, 1992 - 1996. (Map; 1:500,000). U.S. Geological Survey.

Harlow, G.E., Nelms, D.L. 1998. Hydrogeomorphic Regions Geospatial Data (Map; 1:250,000). U.S. Geological Survey

Helsel, D.R., and Hirsch, R.M. 2002. Statistical Methods in Water Resources. Techniques of Water-Resources Investigations of the United States Geological Survey. Book 4, Hydrologic Analysis and Interpretation.

Hirsch, R.M. 1982. A comparison of four streamflow record extension techniques: Water Resources Research, v. 18, no. 4., p. 1081-1088.

Hutson, S.S., Barber, N.L., Kenny, J.F., Linsey, K.S., Lumia, D.S., and Maupin, M.A., 2004, Estimated use of water in the United States in 2000: Reston, VA., U.S. Geological Survey Circular 1268, 46 p.

Mesko, T.O. 1992. Appalachian-Piedmont Regional Aquifer System Analysis (APRASA) (Provisional data): U.S. Geological Survey.

Milly PCD. 1994. Climate, soil water storage, and the average annual water balance. Water Resour Res 30:2143–2156.

Reilly, F.C. and Kroll, C.N. 2003. Estimation of 7-day, 10-year low-streamflow statistics using baseflow correlation. Water Resources Research, VOL. 39, NO. 9

Rutledge, A.T. 1998. Computer programs for describing the recession of ground-water discharge and for estimating mean ground-water recharge and discharge from streamflow data – update: U.S. Geological Survey Water-Resources Investigations Report 98-4148, 43 p.

Schultz, C, Tipton, D, and Palmer J. 2005. Annual and Seasonal Water Budget for the Monocacy/Catoctin Drainage Area. Interstate Commission on the Potomac River Basin. Report No. 04-04.

Stedinger, J.R., Vogel, R.M., and Foufoula-Georgiou E. in Maidment, D. (ed.). 1993. Handbook of Hydrology

Stedinger, J.R., and Thomas, W.O., Jr. 1985. Low-flow frequency estimation using base-flow measurements: U.S. Geological Survey Open-File Report 85-95, p.21.

Swain, L.E., Hollyday, C.D., and Zapecza, O., 1991, Plan of study for the regional aquifer-system analysis of the Appalachian Valley and Ridge, Piedmont, and Blue Ridge physiographic provinces of the eastern and southeastern United States, with a description of study-area geology and hydrogeology, U.S. Geological Survey, Water Resources Report, 91-4066, 44 p.

Trapp, H. Jr.,, and Horn, M.A. 1997. Ground-Water Atlas of the United States, Segment 11, Delaware, Maryland, New Jersey, North Carolina, Virginia, West Virginia. U.S. Geological Survey Hydrologic Investigations Atlas 730-L.

Venables, W.N., and Ripley, B.D. 2002. Statistics and Computing – Modern Applied Statistics with S (4th ed.). Springer-Verlag, New York Inc.