FINAL REPORT

Development of Preservation Strategies to Preserve Water Resources – Coastal Zone Demonstration

Submitted to:

Maryland Department of Natural Resources
Coastal Zone Management Division

By:

Dr. Glenn E. Moglen
Principal Investigator
Department of Civil and Environmental Engineering
National Center for Smart Growth Research and Education
University of Maryland
College Park, MD 20742

December 14, 2005
(last revised July 26, 2006)

Disclaimer:

Although this project is funded in part by the Environmental Agency, it does not necessarily reflect the opinion or position of the EPA.

Acknowledgements:

This project was funded in part by the U.S. EPA Chesapeake Bay Implementation Grant. We gratefully acknowledge this support.

GISHydro2000+

Using GISHydro2000 to make pollutant/nutrient loading and Stream Biodiversity Estimates

Glenn E. Moglen and Michael J. Paul

This document presents the annual pollutant loading background for both the EPA-PLOAD (EPA, 2001) and the USGS (Driver and Tasker, 1990) methods. Also provided is a worked-out example using GISHydro2000 to make these estimates.

Background

EPA-PLOAD Method

Imperviousness is determined from a coefficient table that assigns a value based on land use. This table is largely based on imperviousness values from the NRCS TR-55 model (SCS, 1986) but modified to cover all land uses defined by the Maryland Department of Planning in their generalized land use mapping.

MDP Imperviousness by Land Use

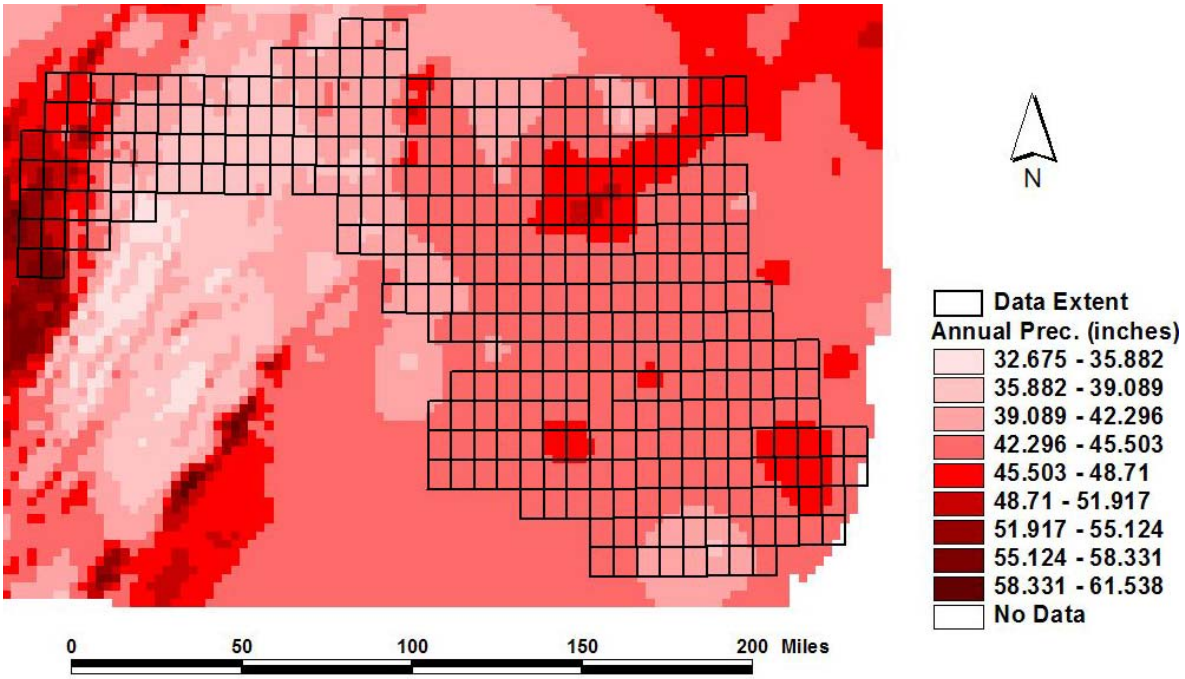
Land Use Code	Description	Imperviousness (%)
11	Low Density Residential	25
12	Medium Density Residential	38
13	High Density Residential	65
14	Commercial	85
15	Industrial	72
16	Institutional	50
17	Extractive	11
18	Open Urban Land	11
21	Cropland	0
22	Pasture	0
23	Orchards	0
24	Feeding Operations	0
25	Row Crops	0
41	Deciduous Forest	0
42	Evergreen Forest	0
43	Mixed Forest	0
44	Brush	0
50	Water	0
60	Wetlands	0
70	Barren Land	50
71	Beaches	0
72	Bare Exposed Rock	100
73	Bare Ground	50
80	Transportation	75
191	Large Lot Agricultural	15
192	Large Lot Forest	15
241	Feeding Operations	10
242	Agricultural Buildings	10

Based on imperviousness calculated from above, the spatial distribution of runoff coefficients can be determined on a pixel by pixel basis based on the following relationship:

$$R_v = 0.05 + (0.009 \cdot I) \qquad (1)$$

where R_v is the runoff coefficient for a land surface and I is the percent imperviousness of the pixel being examined.

Mean annual precipitation is determined as a spatial distributed quantity based on a GIS grid obtained from the Spatial Climate Analysis Service (http://www.ocs.orst.edu/prism/) at Oregon State University. The gridded data covers the



entire GISHydro2000 spatial extent at 4 km resolution. The data represent mean annual precipitation between 1971 and 2000. A representation of these data are shown in the figure below.

Pollutant loadings are then calculated as follows:

$$L = \frac{2.72}{12} \cdot P_j \sum_u \overline{P_u} \cdot R_{vu} \cdot C_u \cdot A_u \qquad (2)$$

where L is the pollutant loading in lbs/year, $\overline{P}_u$ is the mean annual precipitation in the area of land use u in inches, P_j is the fraction of storms producing runoff (P_j = 0.9 is used by default), the subscript, u, is used to denote land use type, R_{vu} is the runoff coefficient for land use u (defined previously in equation 1), C_u is the event mean concentration in milligrams/liter for land use u, and A_u is the area of land use u in acres within the watershed. Tables showing event mean concentrations by different land use source are provided in the appendix.

USGS – Driver and Tasker (1990) method

The equation for developing the average annual load from Driver and Tasker is:

$$l = 10^x \cdot BCF \qquad (3)$$

where l is the mean storm load in lbs, x is an exponent dependent on several watershed characteristics (discussed below), BCF is a bias correction factor. The exponent, x, is further defined by the equation:

$$x = \beta_0 + \beta_1 \cdot A^{0.5} + \beta_2 \cdot I + \beta_3 \cdot \overline{P} + \beta_4 \cdot \overline{T_J} + \beta_5 \cdot X \quad (4)$$

where A is the watershed area in mi^2, I is the percent imperviousness of the watershed, $\overline{P}$ is the mean annual precipitation in inches, $\overline{T_J}$ is the mean January temperature in ^{o}F, and X is an indicator variable equal to 1 if the sum of commercial and industrial land use in the watershed exceeds 75 percent, 0 otherwise. Values of β_0, … β_5 are tabulated below.

Quantity	β_0	β_1	β_2	β_3	β_4	β_5	**BCF**
Nitrogen	-0.2433	1.6383	0.0061	--	--	-0.4442	1.345
Phosphorus	-1.3884	2.0825	--	0.0234	-0.0213	--	1.314
TSS	1.543	1.5906	--	0.0264	-0.0297	--	1.521

Annual loading of a given pollutant or nutrient is calculated by multiplying the mean storm load by the annual average number of storms, n, exceeding 0.05 inches:

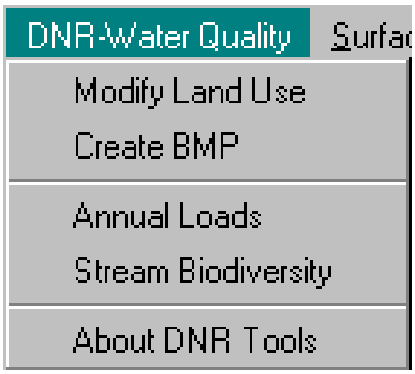
$$L = \frac{n \cdot l}{2000} \quad (5)$$

where L is the pollutant loading in tons/year.

This method depends on an estimate of the number of storms where a storm is defined as an "a rainfall event in which the total rainfall is at least 0.05 inch. Storms are separated by at least six consecutive hours of zero rainfall." The number of storms appropriate for the state of Maryland was not feasible to determine rigorously. From the data provided by Driver and Tasker (1990) it appears that Washington, DC was taken to have 42 storms during the April through September (6 months) period while Baltimore had 39 storms during this same period. Simply doubling these figures gives 84 and 78 storms/year for Washington, DC and Baltimore, MD, respectively from their data.

For the Rockville, MD COOP gage (ID 187705) from January 1, 1950 through January 31, 2002 (18,325 days of record available, approximately 50.2 years of record) the average number of ***days*** with rain exceeding 0.5 inches per year was approximately 85.9. Note that days and storms may not equate. For instance, a single storm might accrue more than 0.05 inches on each of two consecutive days. Similarly, a single day might potentially contain 2 or even 3 individual storms if there are short intense cloudbursts that are separated by 6 hours or more of zero rainfall.

Interpreting all this information, it would seem reasonable to assume that an average annual number of storms in the Baltimore/Washington region is approximately 82 storms/year. This number will be used for n in equation 4, throughout Maryland, but clearly this is subject to error, especially in the wetter Blue Ridge and Appalachian Plateau regions.



Using the new tools within GISHydro2000

Once the "Basin Statistics" have been determined all menu choices under this overall menu are available for use. The menu is divided into three groups as shown at right. The top group, "Modify Land Use" and "Create BMP" menu choices allow the user to specify the land use and BMP characteristics of specific

geographic areas that differ from the default characteristics built into GISHydro2000. The default land use characteristics are governed by the land use layer selected from the dialog shown in Figure 2. There are no default BMP characteristics. The middle group on this menu does the actual calculation of annual pollutant loading and stream biodiversity. The bottom group gives some background, software build date, and contact information for Dr. Moglen. We will cover the use of each of the menu choices here.

Modify Land Use Menu Choice/Tool

There are several reasons why one might wish to use this menu choice/tool:

1. When working with ultimate zoning data, the base information contained within the GISHydro2000 database may not be current in the location of a particular watershed analysis. This tool can be used to update the base information to reflect recent zoning changes.
2. The most likely land use data to be used in GISHydro2000 to reflect "current" conditions are the data supplied by the Maryland Department of Planning (MDP). These data indicate generalize land cover across approximately 25 land cover categories. The hydrologic characteristics of some of these categories (e.g. "Institutional" are not particularly well-defined and may vary considerably from one location to another. This tool can be used to create a new land use category that reflects land cover/land use conditions that are well-understood by the engineer making the change through paper maps or field reconnaissance.
3. The MDP data uses a broad "low density residential" land use category which includes housing densities from half-acre lots up to 2-acre lots. The imperviousness and/or curve numbers associated with this range of housing densities can vary considerably depending on whether the actual density is close to the upper or lower bound of this range. This tool can be used to create a new land use category that more precisely captures the actual housing density through the specification of curve numbers or degree of imperviousness specified directly by the engineer for this new land use category.
4. The user may wish to specify very specific nutrient EMC values that differ from the default ones.

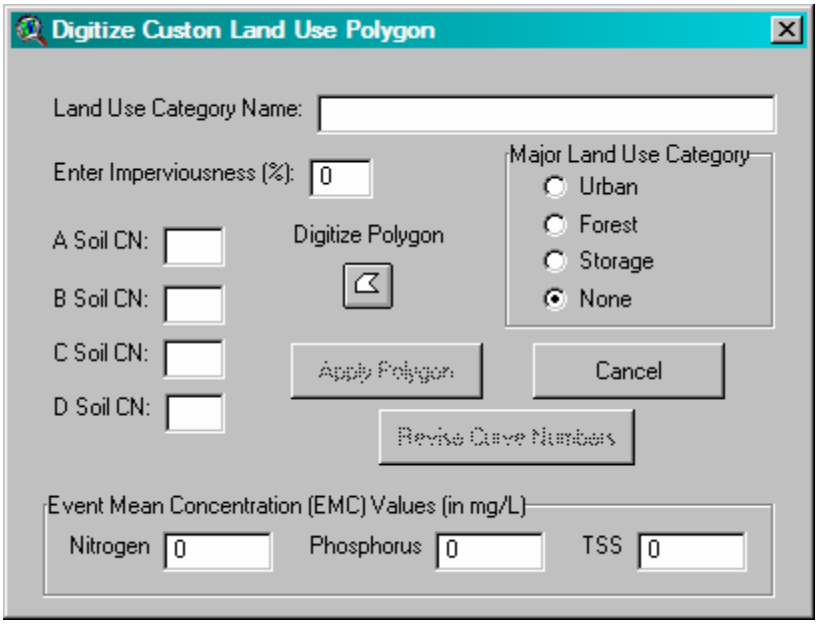
To use the Modify Land Use menu choice/tool, please perform the following steps:

Step 1: Select the Quadrangles/Delineate the Study Watershed (as usual): The analysis performed by the engineer proceeds as before with the engineer using the "Q" button to define the quadrangles that are indicated for a particular analysis. GISHydro2000 will create the "Area of Interest" view with focused on the data for the selected quadrangles. The land use modification tool can now be used, although I suggest the user go one step further and also delineate the watershed before proceeding to use this tool since only the land use within the watershed need be updated.

Step 2: Invoke the Land Use Modification Dialog: Select the "Modify Land Use" menu choice from the "DNR-Water Quality" menu or press the "*LU*" (

) button, located to the right of the "Q" button used earlier to initiate the analysis. This will bring up the dialog box shown at the top of the next page.

Note: Steps 3 through 7 below can be performed in any order provided the directions in these steps are followed appropriately.



Step 3: Entering the Land Use Category Name: Enter in this box the text describing the land use category. You may want to include a special parenthetical comment indicating that this is a special, user defined category. For example, "Residential, 1-acre houses (user defined)." This field is for informational purposes only and is not a required input.

Step 4: Indicating the Major Land Use Category: There exist three special classes of land use that need to be indicated for correct calculation of the "Basin Statistics" and/or the USGS regression equations. These categories are, "urban", "forest", and "storage". User simply needs to click on the category that applies to the new land use category being specified. If none of these categories apply, leave the selection set as the category, "none". Please note that the "forest" and "storage" categories assume and impose an imperviousness of 0%.

Step 5: Indicating the Curve Numbers and/or Imperviousness: The default imperviousness is 0% as the dialog box opens. There are no default curve number values. So long as the major land use category is "urban" or "none" the imperviousness box is editable. Any numerical entry in imperviousness box will result in the calculation of the associated A, B, C, and D curve numbers according to the formulas:

$$x \cdot 98 + (1 - x) \cdot 39 = CN_A \qquad \text{(A Soil)}$$
$$x \cdot 98 + (1 - x) \cdot 61 = CN_B \qquad \text{(B Soil)}$$
$$x \cdot 98 + (1 - x) \cdot 75 = CN_c \qquad \text{(C Soil)}$$
$$x \cdot 98 + (1 - x) \cdot 80 = CN_D \qquad \text{(D Soil)}$$

where x is the imperviousness expressed as a fraction of 1. All curve numbers are rounded to the nearest integer value. Please note that any manual entry in the imperviousness box after the curve number boxes have been filled out will undo entries manually entered in the curve number boxes. If you wish to manually *both* specify curve numbers and imperviousness, you should first specify the imperviousness and then the curve numbers.

Step 6: Indicate Event Mean Concentration (EMC) values: The default values for nitrogen, phosphorus, and total suspended solids, as shown in the dialog, are zero. Enter new values in units of mg/L.

Step 7: Digitizing the Land Use Polygon: Press the "Digitize Polygon" button () and digitize on the computer screen the outline of the polygon of land use you are specifying. Two things to note: 1) To end the digitizing of the polygon, double-click rapidly at the last location of the polygon your are updating; 2) You can digitize multiple polygons for a given category simultaneously. If you have multiple polygons you wish to digitize that you wish to have the same land use, you simply digitize as many polygons you wish before pressing the "Apply Polygon" button.

Step 8: Applying the Polygon: Only after both a polygon has been digitized and curve number/imperviousness information has been entered will the "Apply Polygon" button become active (black). At the time this button is pressed, the text information indicated in the dialog box along with all digitized polygons (see Step 7 above) are written to disk. If the "Apply Polygon" button is not pressed and the dialog box is exited (through the use of the "Cancel" button or the "X" box at the upper-right corner of the dialog) then any information contained in the dialog box at the time of exiting is lost. The Land Use Modification Dialog may be opened once and multiple polygons of land use entered and applied, or the dialog may be opened multiple times each time specifying one or more polygons of land use.

Step 8: Revising the Curve Numbers: After one or more polygons of modified land use are entered and applied, the "Revise Curve Numbers" button becomes active "black". Until this button has been pressed, the land use and curve number themes have not been revised to reflect any of the changes entered in this dialog. This button needs to be pressed only once, at the conclusion of the entry of all modified land use polygons, but may actually be pressed anytime after the first land use change polygon has been completely entered. Note that once this button has been pressed, the legend colors for the display of the "Land Use" and "Curve Number" themes are changed. Since it is impossible to anticipate what kinds of land use will be entered by the engineer, no effort has been made to control the color legends for these themes. For the land use theme, the engineer must manually modify the legends for these themes with the appropriate colors associated with all previously existing and new categories of land use. This is chronologically the last button you will press when using this dialog. Once you are finished with this dialog you can proceed with your hydrologic analysis as done previously.

Step 9: Using the "Cancel" Button: Pressing this button (or the "X" button at the upper-right corner of the dialog) cause the dialog box to close with any information contained in the dialog at the time of exiting being permanently lost. For instance, you may wish to use this button if you are unhappy with the polygon you have digitized. You could then re-open the dialog box by selecting the "Modify Land Use" menu choice or pressing the "*LU*" button with no memory of any information entered previously (the

defined polygon or other text information) being retained since the last time the "Apply Polygon" button was pressed.

Documenting Modified Land Use: The "Digitize Custom Land Use Polygon" dialog stores information in two places during and after use of this dialog is completed. Non-GIS information is stored in the landuse lookup table. The digitized polygons are stored in a shapefile (3 physical files make up 1 shapefile). Both of these entities are written to the c:\temp\#####\ directory.

The Landuse Lookup Table: This table is visible within the GIS as one of the table called, "Landuse Lookup Table." The file that contains the information in displayed in this table is located on the machines hard-drive at, "c:\temp\#####\templutab.dbf". The default version of this table corresponding to the selection of Maryland Department of Planning land use data is shown below:

Landuse Lookup Table

Lucode	Classifica	Hyd_a	Hyd_b	Hyd_c	Hyd_d	Imp	Lucat	Cond	Nitrogen	Phosphorus	Tss
10	Urban	61	75	83	87	0.38	u	Good			
11	Low Density Residential	54	70	80	85	0.25	u	Good	12	2	221
12	Medium Density Residential	61	75	83	87	0.38	u	Good	17	2	305
13	High Density Residential	77	85	90	92	0.65	u	Good	27	3	477
14	Commercial	89	92	94	95	0.85	u	Good	31	4	542
15	Industrial	81	88	91	93	0.72	u	Good	33	4	578
16	Institutional	69	80	86	89	0.50	n	Good	24	3	419
17	Extractive	77	86	91	94	0.11	n	Good	0	0	0
18	Open Urban Land	39	61	74	80	0.11	n	Good	11	2	200
20	Agriculture	67	78	85	89	0.00	n	Good			
21	Cropland	67	78	85	89	0.00	n	Good	11	2	192
22	Pasture	39	61	74	80	0.00	n	Good	11	2	192
23	Orchards	32	58	72	79	0.00	n	Good	0	0	0
24	Feeding Operations	59	74	82	86	0.00	n	Good	0	0	0
25	Row Crops	67	78	85	89	0.00	n	Good	11	2	192
40	Forest	30	55	70	77	0.00	f	Good			
41	Deciduous Forest	30	55	70	77	0.00	f	Good	11	2	190
42	Evergreen Forest	30	55	70	77	0.00	f	Good	11	2	190
43	Mixed Forest	30	55	70	77	0.00	f	Good	11	2	190
44	Brush	30	48	65	73	0.00	f	Good	11	2	190
50	Water	100	100	100	100	0.00	s	Good	0	0	0
60	Wetlands	100	100	100	100	0.00	s	Good	11	2	190
70	Barren Land	77	86	91	94	0.50	n	Good	15	2	268
71	Beaches	77	86	91	94	0.00	n	Good	0	0	0
72	Bare Exposed Rock	77	86	91	94	1.00	n	Good	0	0	0
73	Bare Ground	77	86	91	94	0.50	n	Good	0	0	0
80	Transportation	83	89	92	94	0.75	n	Good	33	4	578
191	Large Lot Agricultural	67	78	85	89	0.15	n	Good	11	2	192
192	Large Lot Forest	30	55	70	77	0.15	f	Good	11	2	190
241	Feeding Operations	59	74	82	86	0.10	n	Good	0	0	0
242	Agricultural Buildings	59	74	82	86	0.10	n	Good	0	0	0

The "Hyd_x" fields (columns) indicate the curve numbers that apply to this land use category for soil type "x." The "Imp" field shows the default imperviousness associated with each land use category as a decimal fraction. The "Lucat" field indicates the major land use class (see Step 4) that applies to each land use category ("u"=urban, "f"=forest, "s"=storage, and "n"=none. The values and category descriptions appearing in the leftmost two fields will vary depending on the land use coverage selected by the engineer at the time the analysis is initiated. Additional records (rows) starting with values of

Lucode = 501 will be added to this table if the land use modification dialog is used to indicate new land use polygons. This table should be included as a standard part of all hydrologic analysis reports.

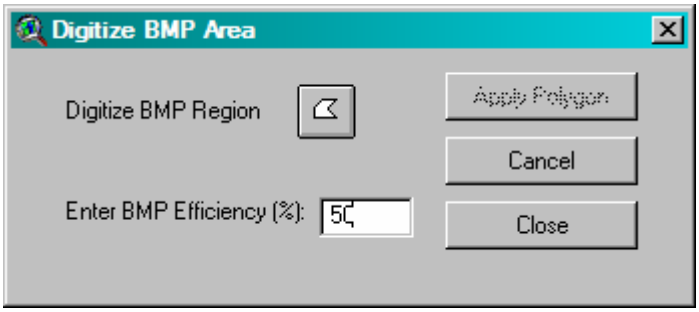
The "lumod" shapefile: This file is not loaded into the GIS. It exists only on disk as "c:\temp\#####\lumod.xxx" (where "#####" is the specific path off the temp directory that corresponds to the given GISHydro2000 session, and "xxx" are the 3 file extensions: "shp", "shx", and "dbf" that make up a shapefile.) If land use is changed as part of a given analysis, this shapefile should be included electronically as a standard part of the reporting of that analysis.

Create BMP Tool

This tool exists to allow the user to specify arbitrary geographic polygons that are subject to some form of best management practice (BMP) that serves to reduce nutrient/pollutant loading rates. This tool is similar to, but somewhat simpler to use, the modify land use tool.

To use the Create BMP menu choice/tool, please perform the following steps:

Step 1: Digitize the BMP region: Press the "Digitize BMP Region" button () and



digitize on the computer screen the outline of the polygon of BMP applicability you are specifying. As with the land use digitizing, there are two things to note: 1) To end the digitizing of the polygon, double-click rapidly at the last location of the polygon your are updating; 2) You can digitize multiple polygons for a given BMP efficiency simultaneously. Simply digitize all polygons you wish to apply a given efficiency to (e.g. 50% efficiency is shown in dialog above) before pressing the "Apply Polygon" button.

Step 2: Enter BMP Efficiency: BMP efficiencies are entered in units of percent. A 0% efficiency means there is effectively no BMP present, a 100% efficiency will eliminate all pollutant loadings from the specified polygon area.

Step 3: Applying the Polygon: After both a polygon has been digitized and a BMP efficiency has been entered will the "Apply Polygon" button become active (black). At the time this button is pressed, the efficiency entered in the dialog box along with all digitized polygons (see Step 2 above) are written to disk. If the "Apply Polygon" button is not pressed and the dialog box is exited (through the use of the "Cancel" or "Close" buttons or the "X" box at the upper-right corner of the dialog) then any information contained in the dialog box at the time of exiting is lost. The Create BMP tool may be opened once and multiple polygons of BMP areas entered and applied, or the dialog may

be opened multiple times each time specifying one or more polygons of BMP applicability.

Step 4: Using the "Cancel" or "Close" Buttons: There is no functional difference between these buttons. Pressing these buttons (or the "X" button at the upper-right corner of the dialog) causes the dialog box to close with any information contained in the dialog at the time of exiting being permanently lost.

An Example

Figure 1 presents the opening screen of GISHydro2000. The red rectangular outlines correspond to individual 7.5 minute USGS quadrangles. The quads shown comprise the entire set of quads necessary to account for all watersheds both in the state of Maryland and those watersheds which drain into the state of Maryland. The only exceptions to this are along the main stem of both the Potomac and Susquehanna rivers which would necessarily include much of Virginia (in the case of the Potomac) and much of Pennsylvania and even New York (in the case of the Susquehanna).

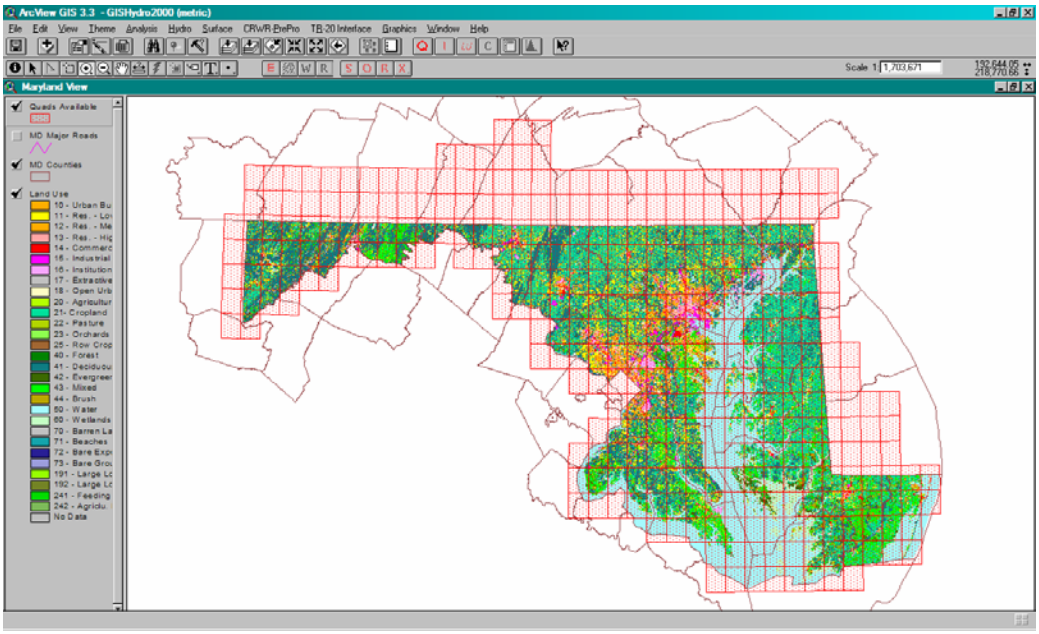


Figure 1. The opening "Maryland View" screen of GISHydro2000.

Pressing the "Q" button initiates an analysis in which the first step is the specification of the planned spatial extent of the analysis. The user indicates this spatial extent by selecting one or more quads for analysis. This step is shown in Figure 2 which illustrates the "Select Quadrangle(s)" dialog box. Shown selected is a single (Kensington) quadrangle which defines the areal extent of the planned analysis. There are three categories of spatial data that must be selected by the user at this point: DEM data (topography), land use, and soils. GISHydro2000 contains three different sources of DEM data (with the shown NED – National Elevation Dataset (USGS, 2005a) data being the best quality), nine different sources of land use data which include recent satellite imagery from the national land cover dataset (USEPA, 2005) and several statewide generalized land use products from 1985 to 2000 (shown), and soils data are available in

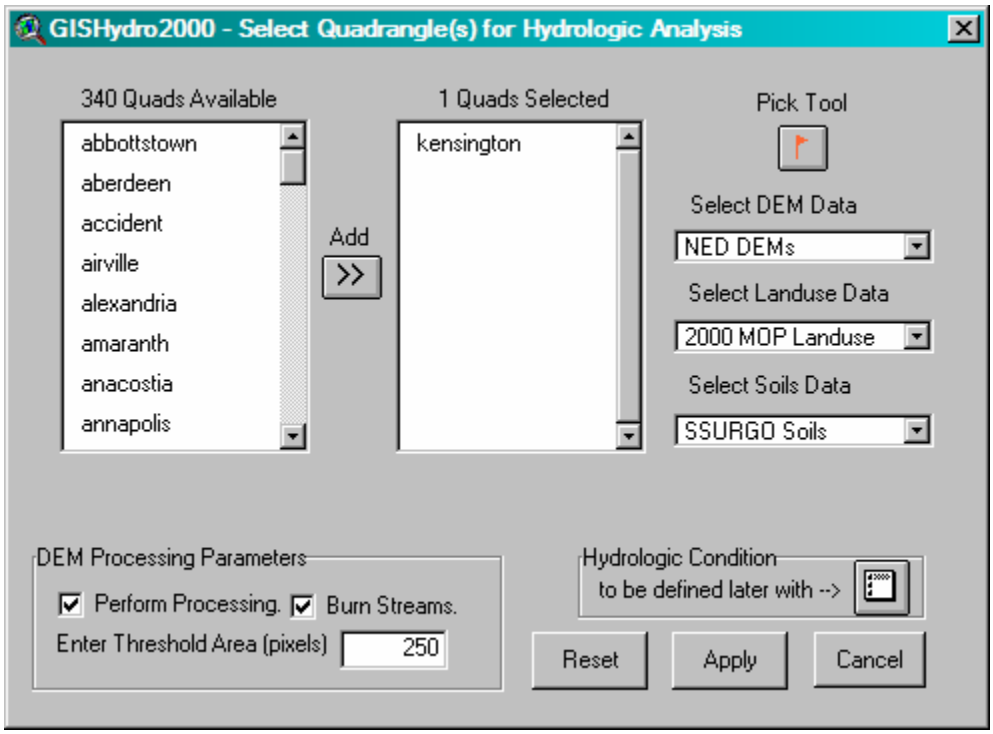


Figure 2. The "Select Quadrangle(s)" dialog box.

three different sources: SSURGO (NRCS, 2005a), STATSGO (NRCS, 2005b) and a homegrown layer developed from scanned county soil maps. Finally, the check boxes in the lower right concern whether the DEM will be interpreted for flow directions and flow accumulation (yes, if the "Perform Processing" box is selected) and whether streams should be imposed where they have been digitized in the National Hydrography Dataset – NHD (USGS, 2005b) (yes, if the "Burn Streams" box is selected). The value of 250 entered for the threshold area is simply a count in 30 meter x 30 meter pixels used to define the minimum area required to form a stream. This is a visualizing device that will be used to infer stream locations in the next part of the analysis. Once the "Apply" button is selected the data indicated in the dialog box are extracted from the GISHydro2000 database, and any indicated processing is performed.

Watershed Delineation, Characteristics, and Flood Frequency

At the completion of all data extraction and processing, GISHydro2000 creates the "Area of Interest" view as shown in Figure 3. Figure 3 shows the view window after the next step of watershed delineation has taken place. Watershed delineation is a task that is easily accomplished by the GIS, with the user only needing to indicate, with a mouse click, the location of the outlet of the watershed. In this example, a watershed has been delineated at the intersection of a road and the stream network. This is a typical application of GISHydro2000 since hydrologic engineers are concerned with such locations where bridges or culverts must be constructed at road/stream intersections.

The delineated watershed is now used by the GIS to overlap the existing data layers and determine a suite of watershed characteristics. These characteristics are determined by the user simply choosing the "Basin Statistics" menu choice which produces the dialog box shown in Figure 4. These statistics include estimates of all the

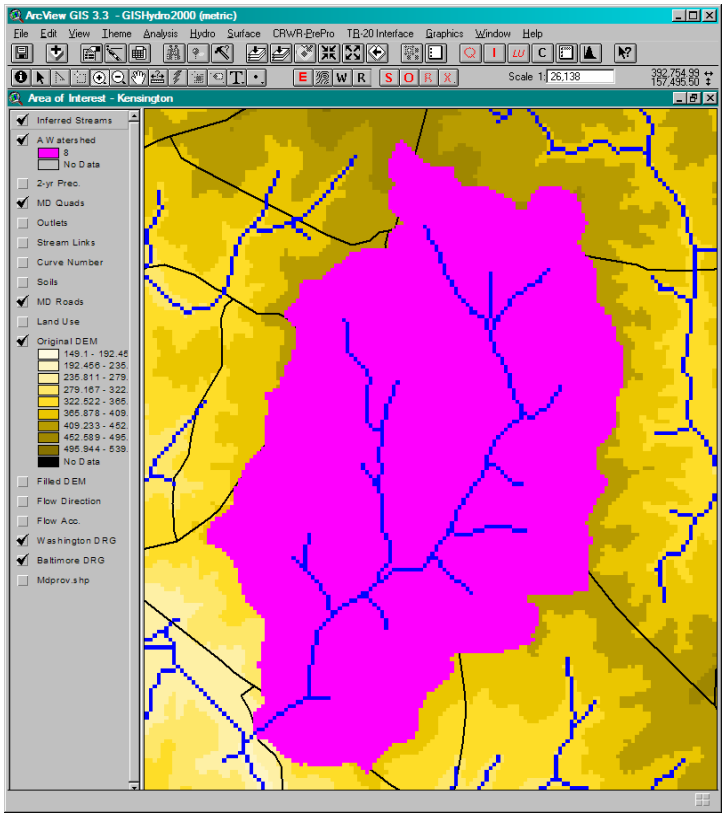


Figure 3. The "Area of Interest" view focused on a delineated example watershed.

watershed characteristics necessary to estimate flood frequency using the current USGS rural regression equations (Dillow, 1996) and some more recent regression equations based on USGS gage data but that also include measures of urbanization where urbanization is sufficient to influence flood frequency relations (Ragan et al., 2004).

The two-year flood from regression equations by Dillow (1996) and Ragan et al. (2004) are provided in equations 6 and 7, respectively,

$$Q_2 = 451A^{0.635}(F+10)^{-0.266} \quad (6)$$

where Q_2 is the 2-year peak discharge in ft^3/s, A is the drainage area in mi^2 and F is the percent forest cover, and,

$$Q_2 = 37.01A^{0.635}(IA+1)^{0.588} \quad (7)$$

where IA is the percent impervious area. Figure 4 shows that the drainage area of the watershed shown in Figure 3 is 3.8 mi^2 with 7.3 percent forest cover and 37.1 percent impervious area. We insert these values into equations 1 and 2, respectively,

$$Q_2 = 451\cdot(3.8)^{0.635}(7.3+10)^{-0.266} = 492\ ft^3/s \quad (8)$$

$$Q_2 = 37.01\cdot(3.8)^{0.635}(37.1+1)^{0.588} = 734\ ft^3/s \quad (9)$$

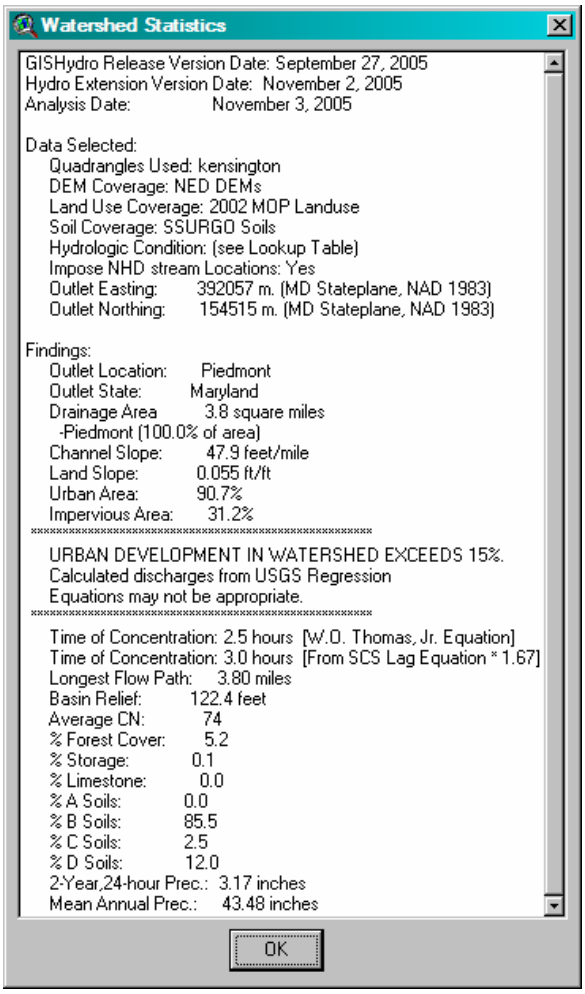


Figure 4. Watershed statistics for example watershed.

The calculations shown in equations 3 and 4 do not need to be performed by hand. Instead, the user may simply choose two menu choices in GISHydro2000, "Calculate Dillow Discharges" and "Calculate Thomas Discharges", respectively. The top part of the dialog for the Thomas discharges response is shown in Figure 5. A similar dialog is produced for the Dillow discharges but is not shown. As shown in Figure 5, GISHydro2000 determines the 1.25- through 500-year peak discharges (along with confidence intervals (not shown) around each of these discharges based on the standard errors of the regression equations). It is worthwhile to pause momentarily here to note that with just a few mouse clicks to indicate analysis extent, desired data sets, and watershed outlet location, GISHydro2000 has allowed the rapid determination of watershed characteristics and flood frequency. Computation speed varies somewhat with analysis extent, but for illustrative purposes the results shown here can be produced within less than 5 minutes on a 650MHz CPU. These same calculations, performed by manual techniques, could easily consume many hours or even days depending on watershed size.

Not shown here are additional capabilities within GISHydro2000. For example, this tool can readily determine the analogous flood frequency outcomes for the Ragan et al., (2004) regression equations. Further, GISHydro2000 has an ancillary supporting layer of the location of all USGS stream gages in and around Maryland. If the watershed outlet indicated by the user should correspond to a stream location within +/- 50 percent of the drainage area of a gaging location, GISHydro2000 will prompt the user as to whether he/she wishes to produce a weighted flood frequency estimate based on a combination of the regression equations for

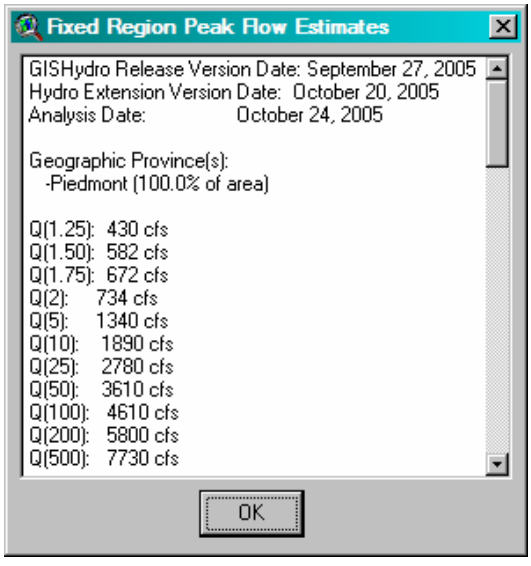


Figure 5. Thomas discharges for example watershed.

the observed flood frequency at this gage and the flood frequency that would result from the regression equations alone.

Using the DNR – Water Quality Menu

Imagine that the northern third of the example watershed is to undergo urbanization such that it is 30% impervious with the following EMC values: nitrogen (2.0 mg/L), phosphorus (0.5 mg/L), and TSS (50 mg/L). The modify land use tool is shown in Figure x above, creating these entries. Also shown, the "Landuse Lookup Table" was also revised by inserting an additional record that reflects: 1) that the new digitized area is considered urban, 2) the new curve numbers, and 3) the new EMC values.

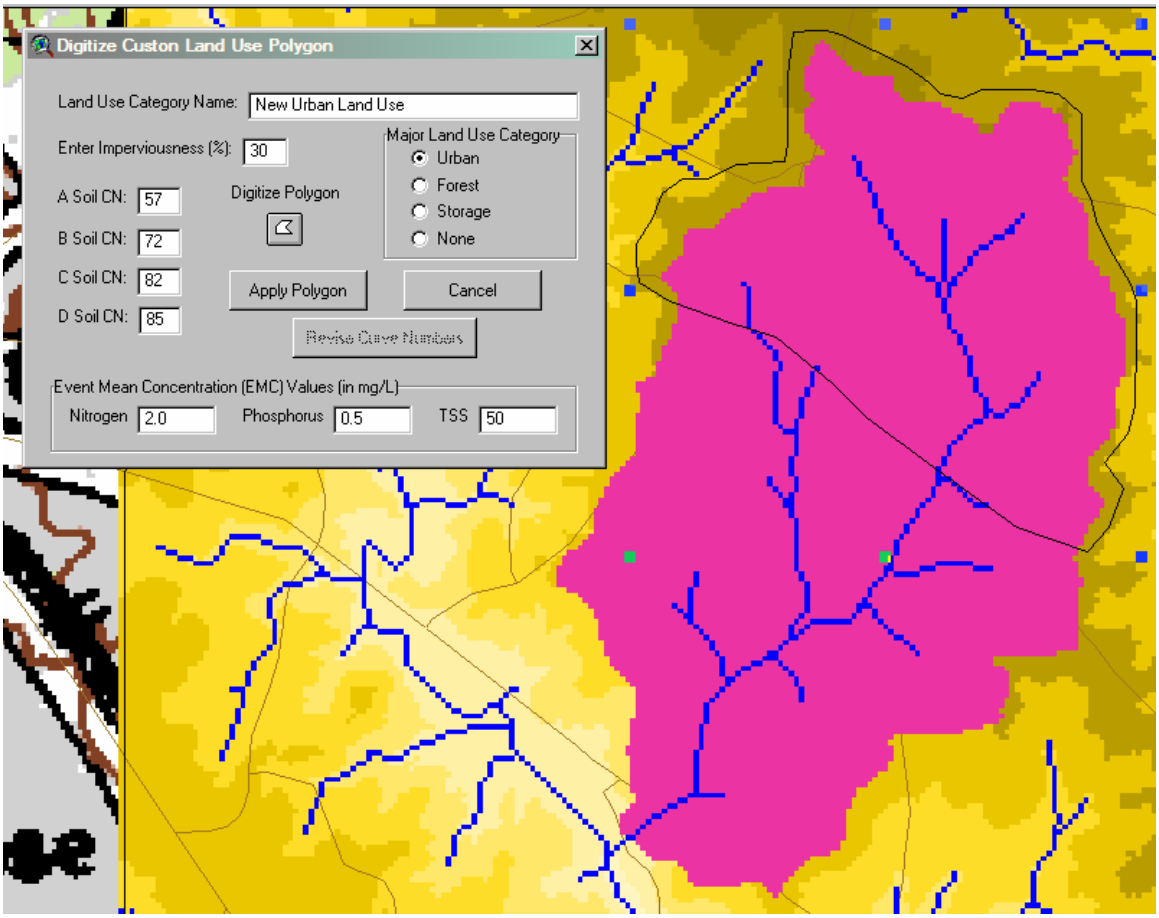


Figure 6. Digitizing a new land use area and specifying its characteristics (imperviousness, curve numbers, and EMC values).

Landuse Lookup Table

Lucode	Classifica	Hyd_a	Hyd_b	Hyd_c	Hyd_d	Imp	Lucat	Cond	Nitrogen	Phosphorus	Tss
501	New Urban Land Use	57	72	82	85	0.30	u		20	5	500
11	Low Density Residential	54	70	80	85	0.25	u	Good	12	2	221
12	Medium Density Residential	61	75	83	87	0.38	u	Good	17	2	305
13	High Density Residential	77	85	90	92	0.65	u	Good	27	3	477
14	Commercial	89	92	94	95	0.85	u	Good	31	4	542
15	Industrial	81	88	91	93	0.72	u	Good	33	4	578
16	Institutional	69	80	86	89	0.50	n	Good	24	3	419

Figure 7. Landuse Lookup Table with record reflecting new digitized land use. (Note that EMC values are in mg/L x 10.)

EMC values shown in the Landuse Lookup Table are stored as 10 times greater than their actual value in mg/L. This is a brought on by a limitation of GIS functionality. However, it does not affect the calculations. As shown the "20" for nitrogen for the "New Urban Land Use" will be applied as 2.0 mg/L.

Similar to land use, now imagine that two BMP areas exist, one in the southern third of the watershed that results in a 25% reduction in loadings, and one in the northeastern quarter of the watershed that produces a 50% reduction in loadings. These BMP areas and the BMP digitizing tool are shown in Figure x. The "Digitize BMP Area" dialog is shown just before the northeastern BMP polygon is applied.

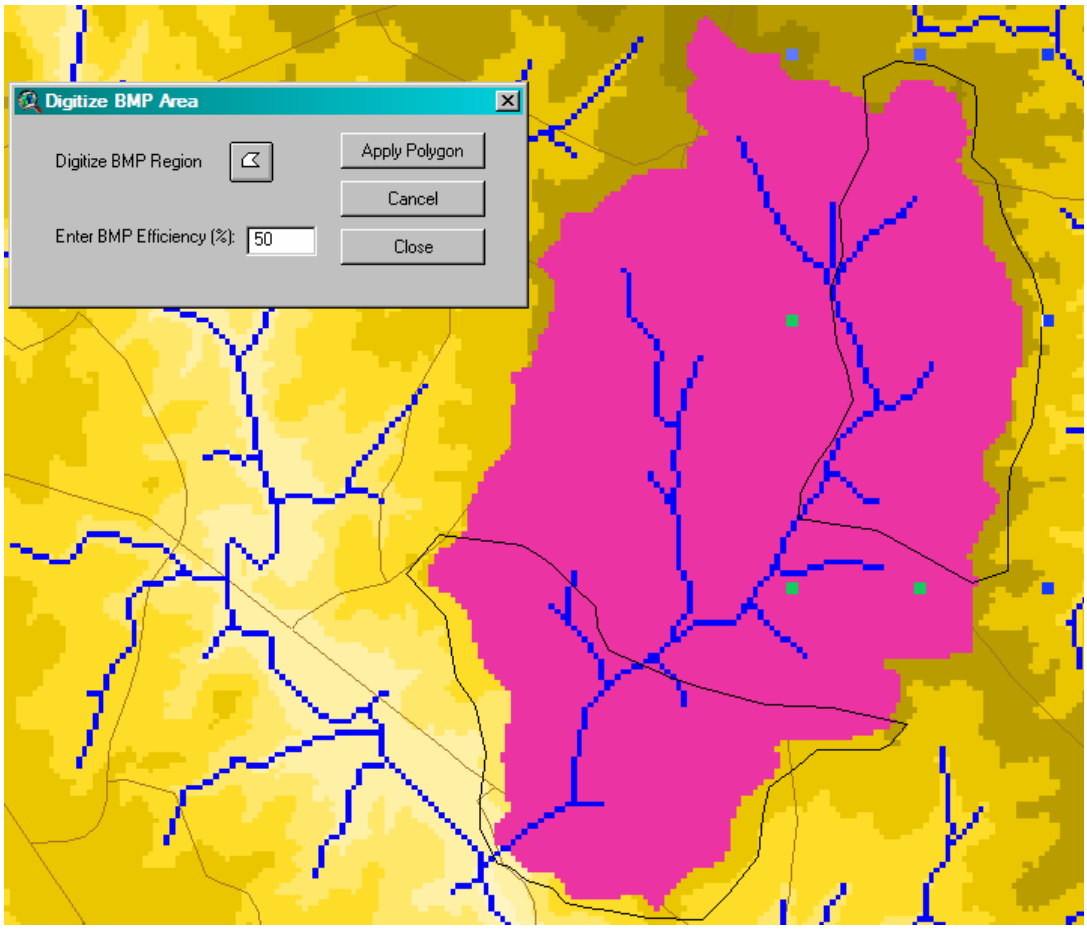


Figure 8. Digitizing BMP areas. (The southern third region has a BMP effectiveness of 25%. The northeastern corner has an effectiveness of 50% as shown.)

Once any/all land use and BMPs have been entered, it is now time to calculate annual loads. Choosing the "DNR-Water Quality: Annual Loads" menu choice produces the report dialog box shown in Figure x.

The dialog box shows several things. First and foremost, there is a large disparity in the estimates of loading produced by the PLOAD vs. the USGS methods. The dialog shows that, for the watershed being examined, the annual loading estimates from the PLOAD method is greater than an order of magnitude smaller than the equivalent

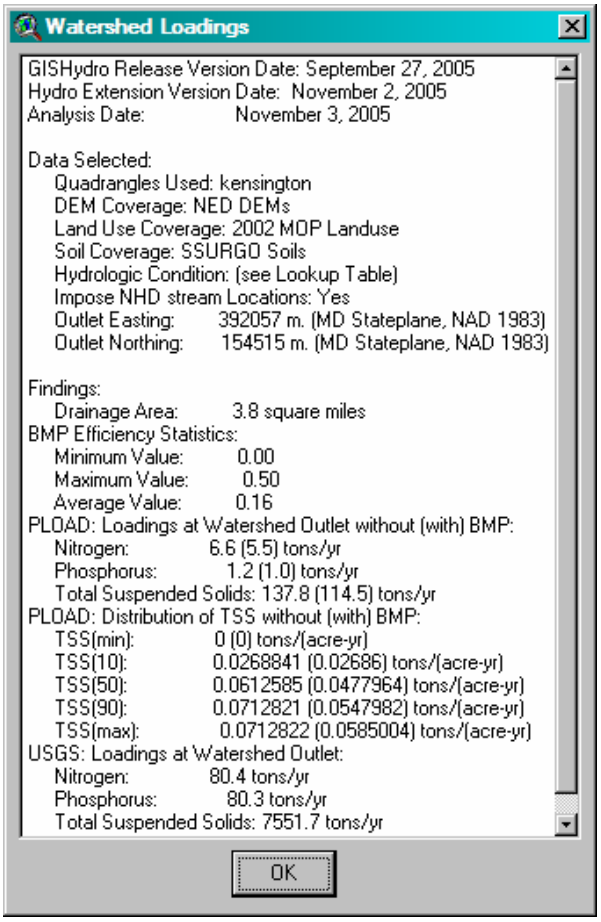


Figure 9. Watershed loadings dialog for example watershed.

estimates from the USGS method. The PI has carefully checked the calculations in GISHydro2000 underlying the execution of both methods and feels certain that both methods are being carried out correctly. It is the PI's feeling that the difference between the two methods simply illustrates the uncertainty underlying such estimates.

That said, it is the PI's feeling that the best way to interpret these results is to use them in *relative comparison* with other watersheds, rather than to accept the absolute magnitudes of the values calculated. In other words, it is meaningful to compare differences in results between two different watersheds for the same method. Hypothetically, if "Watershed A" produces loadings (from either or both methods) that are twice those estimated by "Watershed B", it is a fair assumption that this ratio in loadings is accurate. Similarly, comparing the ratio of loadings for a single watershed in two different land use conditions is appropriate. Based on the large difference in estimates from both methods, the PI is ***not*** comfortable using these estimates to make absolute assessments of loadings that a given watershed produces. It is best to consider these estimates for planning (not design) purposes only.

The BMP areas indicated earlier are applied to the PLOAD estimates only. Two BMP areas were indicated using the "Create BMP" tool shown in Figure xx. One BMP area covered the southern third of the watershed with 25% efficiency. The other area covered the northeastern quarter of the watershed with 50% efficiency. The "Watershed Loadings" dialog shows that averaged over the entire area of the watershed this amounts to a 16% reduction. The values in parentheses following each PLOAD entry indicate the determined loading with the BMP in place. The larger value outside the parentheses indicates the loading estimate in the absence of any BMPs.

It should be noted that the PLOAD method can be executed two ways: based on one-time application of watershed averaged values or on the sum of the individual application of this method to each individual pixel within the watershed. Depending on the way this method is executed, the results will differ. The method is carried out in GISHydro2000 in the latter fashion, applying equation 2 at each individual pixel and then summing these incremental loadings across the entire watershed. Using this approach the result for annual loadings of nitrogen were determined to be 6.57tons/year (this number is rounded to 6.6 tons/year in the watersheds loading dialog). In contrast, if we take

watershed averaged values as reported in the basin statistics dialog, the PLOAD method would estimate:

$$\overline{R_u} = 0.05 + (0.009 \cdot I) = 0.05 + \left[(0.009) \cdot (31.19)\right] = 0.3307$$

$$L = \frac{2.72}{12} \cdot P_j \cdot \overline{P} \cdot \overline{R_u} \cdot \overline{C} \cdot A = \frac{2.72}{12} \cdot (0.9) \cdot (43.48) \cdot (0.3307) \cdot (1.445) \cdot (2412) \qquad \text{(x)}$$

$$L = 10{,}220 \text{ lbs/year} = 5.11 \text{ tons/year}$$

The difference in this case is that the watershed averaged value is about 22 percent smaller than of the pixel-based sum across the same watershed. It is almost certainly the case that the pixel-based execution of this method represents a more precise implementation of the method than the data from which the method was derived. Nevertheless, the method, as implemented in GISHydro2000 is based on the pixel-based approach for consistency with the PLOAD documentation.

The TSS results are presented two different ways. They are presented as outright loadings similar to the nitrogen and phosphorus entries in the dialog box. Here the units are in tons/year and represent the loading that applies at the delineated watershed outlet. The TSS results are also presented as a loading "rate" distribution in units of tons/(acre-yr) and apply at all locations along the "blue lines" indicated in the National Hydrography Dataset (NHD) that are within the delineated watershed. These loading "rates" represent the average loading values per unit area to provide a sense of loading intensity along the drainage network. It is the PI's feeling that this is a more meaningful way to look at nutrient/pollutant loadings because it turns the measure into a reach-based rather than point measure. It also provides the user some sense of the variation in loadings along a reach rather than providing just a single value.

For those that are more comfortable working with GIS, rather than pure tabular data, the "Annual Loads" menu choice produces several themes that may be of interest. There are three themes the represent the at-site EMC values (in mg/L) for nitrogen, phosphorus, and TSS with BMPs (if any) applied. There are also three more themes that represent the cumulative loadings (in tons/year) for each of these three quantities. Also provided for background information is a theme called, "Average TSS" which presents the loading rate of TSS discussed above in units of tons/(acre-year) and a "Nhdpoints.shp" which is a point file that approximates the location of any NHD blue lines within the watershed. The values contained in this point file can be readily queried or visualized so as to examine spatial patterns in loading characteristics. Finally, there is a theme called simply, "Imp" which conveys the spatial distribution in imperviousness (in percent) used in the PLOAD and USGS calculations.

Some final comments are useful here. The current "Watershed Loadings" dialog as presented is not expected to represent its final form. Several different kinds of information are conveyed in this dialog to motivate/illustrate various possibilities. The PI awaits feedback from DNR staff as to what ways of analyzing and quantifying the information are most useful. Modifications to this dialog are certainly possible.

Stream Biodiversity

Generating an expected macroinvertebrate richness within GISHydro2000

Model Background

The predictive bioassessment approach used for this application was based on the River InVertebrate Prediction And Classification System (RIVPACS) approach (Wright 2000). RIVPACS, developed as one bioassessment model for Britain, and AUSRIVAS (AUStralian RIVer Assessment System) are methods of bioassessment that predict an expected invertebrate community in a stream based on physical features of the stream reach and surrounding landscape (Wright et al 1984, Furse et al 1984, Moss et al 1987, Marchant et al 1995, Wright 1995, Davies 2000, Simpson and Norris 2000, Wright 2000). These assessment models compare the observed community of insects at a test site to that expected in the absence of human disturbance (Observed:Expected; O/E) and assess biological condition based on a significant departure from 1.0 (where Observed = Expected). The observed community is that found using standard sampling methods, whereas the expected community is built using a model based on reference (minimally disturbed) sites from across the sampling region. The approach is based on the concept that any site would most likely have those taxa commonly found from physically similar reference sites. So, in essence, one constructs a site-specific reference condition for each test site that is the most probable community of invertebrates expected at that test site in the absence of human disturbance. The expected taxa list is conceptually a weighted average of taxa frequencies found in reference sites, where the weights are the probability a site is in a particular group of reference sites; average taxa frequencies from reference sites that are physically very similar to a test site are weighted most. The approach has been applied successfully in the UK and Australia and in several US states (Wright et al. 1993, Hawkins et al. 2000, Paul et al. 2002). For this application, the focus is strictly on estimating the expected taxa richness (E) for any site.

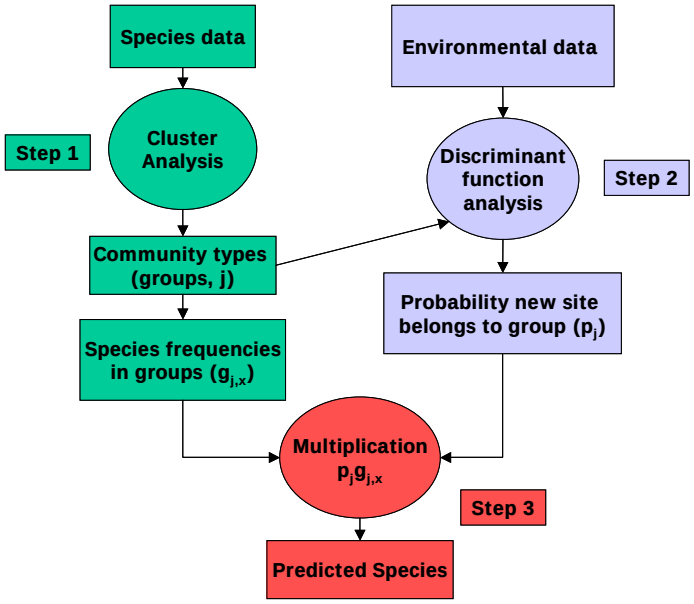


Figure 1 - Schematic showing the three main steps involved in building RIVPACS-type bioassessment models.

RIVPACS-type analysis proceeds in three main steps (Figure 1): 1) a cluster analysis of reference sites based on taxonomic composition to classify reference community groups, 2) a discriminant analysis to develop linear models using physical variables to estimate the probability with which a test site belongs to each of the reference community groups created in step 1, and 3) the prediction of the taxonomic

composition of test sites based on group membership probabilities (step 2) and the frequency of taxa occurrence in each reference group.

RIVPACS predictive models are built using predictor variables considered relatively invariant to human disturbance (Wright et al. 1984, Hawkins et al. 2000, Wright 2000). Using established biogeographic factors that are minimally affected by human activity, it is possible to predict the expected community for altered streams. If alterable variables were used (e.g., nutrient concentrations, conductivity, forest cover), it would be difficult to discriminate the natural gradient from that caused by human activity, and confident prediction of an expected community in the absence of human disturbance for a test site would be impossible. Commonly used variables for building RIVPACS models are shown in Table 1.

TABLE 1. Predictor variables commonly used for building multivariate predictive models.

Predictor Variables Used		Reference
RIVPACS in United Kingdom		Wright 2000
Mean depth	Slope	
Mean width	Discharge category	
Mean substratum	Mean air temperature	
Alkalinity	Annual air temperature range	
Altitude	Latitude	
Distance from source	Longitude	
AUSRIVAS in Australia		Simpson and Norris 2000
Longitude	Macrophyte taxa	
Latitude	Flow pattern	
Alkalinity	Macrophyte cover	
Altitude	Shading	
Distance from source	Bedrock	
Catchment area	Stream width	
Conductivity	Riffle depth	
Stream slope	Percent pebble	
Riparian width	Edge/bank vegetation	
Percent cobble	Vegetation category	
Percent boulder	Annual air temperature range	
Stream order	Percent gravel	
Discharge	Percent silt	
Percent sand	Percent clay	
Models from California		Hawkins et al.2000
Conductivity	Stream length	
Longitude	Mean width	
Catchment area	Sampling date	
Altitude	Slope	
Mean depth	Azimuth	
Latitude		

The RIVPACS is a natural fit for generating a biodiversity estimate (E) within GISHydro. GISHydro generates a number of the typical non-human influenced, landscape predictors generally used in RIVPACS modeling for any stream location in the state. Since RIVPACS models use these types of predictors to generate an estimate of the number of taxa and a probability of capturing any taxon, this approach was ideal for generating expected richness within GISHydro. The goal was to be able to produce information on biodiversity as an additional element of information generated in standard GISHydro output. This information has potential utility for project and landscape planning, among other applications.

One important note. The biodiversity module for GISHydro is not intended to supplant existing tools for assessing stream condition or impairment. The State of Maryland has an accepted and tested standard sampling program and an accepted stream condition tool – the Benthic Index of Biological Integrity, used by the Maryland Biological Stream Survey (MBSS) (http://www.dnr.state.md.us/streams/mbss/index.html). A relative measure of stream condition (O/E score) can be generated by comparing the observed community from a stream to that predicted with this model output. Since MBSS data were used to build these models, any estimate of the observed community must use the same MBSS stream sampling protocol.

Methods

Data on macroinvertebrate communities were assembled from the Maryland Biological Stream Survey database from 1995-1997 and a subset of sites from 2000-2002. Raw taxa count data for each of 1679 sites were extracted and used. In order to have consistent taxonomic resolution across all the sites, a set of operational taxonomic unit designations were developed. There were a few ambiguous genus and corresponding family level resolution records within the same database. This is not uncommon, especially with potentially damaged or juvenile individuals that can be difficult to place into genera. Such decisions were relatively rare within the very clean MBSS database. For the most part, operational taxonomic units were kept at the genus level, and some family level records were dropped. In addition, a second data matrix was developed with raw taxa counts to family level for each site. Both genus and family level models were constructed.

RIVPACS models are built using reference sites – sites minimally impacted by human disturbance. MBSS had developed reference criteria for building their assessment models. In order to be consistent with their approach, the same criteria were used to designate reference sites in this process. These criteria can be found in the MBSS IBI development report (Roth et al. 2000). MBSS chemistry, habitat, and land cover data were all used for this selection process. The reference selection resulted in the identification of 176 reference sites, 158 of which could be used for modeling. Of this, 29 were set aside for model validation and not used to build the model.

Once the biological data were assembled and reference sites identified, coordinates for each of the 1679 sites used in the models were entered in GISHydro to generate a set of variables for each watershed to use as predictors in the predictive

modeling (Table 2). The predictors were, subsequently analyzed and transformed as necessary to meet assumptions of normality and equal variance (Table 2).

Table 2 Predictor variables generated by GISHydro for use in predictive modeling.

Predictor Variables Generated	Definition	Transformation
AREA	Watershed Area	$Log_{10}(x+1)$
WSLOPE	Watershed Slope	$Log_{10}(x+1)$
CSLOPE	Channel Slope	$Log_{10}(x+1)$
RELIEF	Basin Relief	$Log_{10}(x+1)$
LIME	Percent Limestone	Presence/Absence
PERIM	Watershed Perimeter Length	$Log_{10}(x+1)$
LENGTH	Main Channel Length	$Log_{10}(x+1)$
SA, SB, SC, SD	Percent Watershed as Hydrologic Soil Types A-D	$\text{Arcsine}(\sqrt{x})$
ELEV	Sampling Point Elevation	$Log_{10}(x+1)$
SIN	Channel Sinuosity	
P2	2-Year Precipitation	
P100	100-Year Precipitation	
HIGHELEV	Percent Watershed Above 2000 feet	$\text{Arcsine}(\sqrt{x})$
HYPSO	Hypsometric Area Ratio	
EASTING	Coordinate	
NORTHING	Coordinate	
LAT	Coordinate	
LONG	Coordinate	

The first step in the predictive modeling process was cluster analysis. The biological data for reference sites only were assembled into genus level and family level site by count matrices. Counts were converted into presence-absence data. Rare taxa (< 5% of reference sites) were removed from each matrix for the cluster step only. In general, rare taxa (occurring at less than 5% of reference sites) are often excluded because they contribute too much unique information for only a few sites and lead to under-clustering (over-splitting) (Hawkins et al. 2000). Again, these taxa are not eliminated from the whole process, only from the cluster analysis. They are used later in the construction of expected communities for each site.

The goal of cluster analysis is not only to produce as many groups as possible to simulate the continuous and dynamic community structure that exists across any region, but also to minimize the number of unique small groups that would be too hard to predict accurately without over-fitting the discriminant function models. Organisms exist along continuous environmental gradients with optima under certain conditions. Of course, there are a multitude of different environmental gradients and many different taxa, so modeling the distribution of all of those taxa and all of those continuous gradients would not be a trivial exercise. The cluster analysis step is used to dissect the continuous distributions of taxa into as many small groups of co-occurring taxa as possible, much like one learns to approximate curves by breaking them into small pieces using integral calculus. The ultimate result is a series of unique site clusters with similar taxonomic composition.

Cluster analysis actually refers to a suite of different methods that group sites together based on their similarity with regards to many elements. Different cluster

analysis approaches have been used in building bioassessment models. An agglomerative clustering approach within the PC-ORD software (Flexible-beta linkage method, beta = -0.5) was run using a matrix based on Bray-Curtis distance measures.

After the cluster analysis was finished, decisions about where to prune the cluster dendrogram had to be made in order to assign sites to groups. Obviously, the final cluster (one group) would not work. Likewise, using every individual site would not work. There is a point between these two extremes that represents the optimum number of clusters. The goal was to have as many clusters as possible to resolve the continuous distribution well, while at the same time avoiding very small clusters (<5 sites). The first cut used the 50% information line and was adjusted up and down as needed to optimize the number of clusters while avoiding over-splitting. A variety of grouping schemes were modeled for both the family and genus level models, but the 7 group genus and 9 group family models worked best (Figure 2a and b).

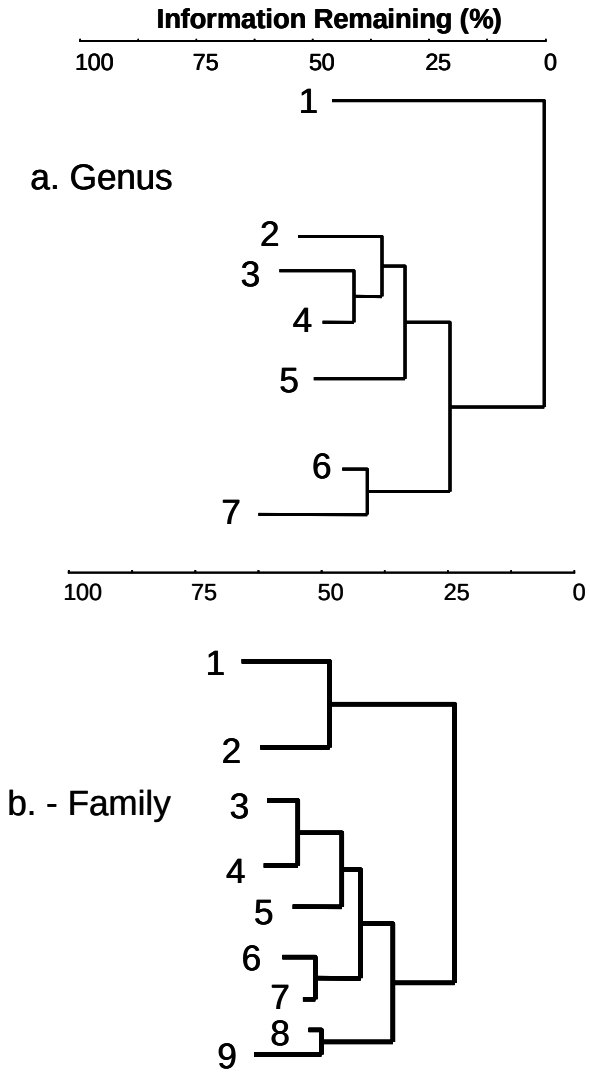


Figure 2 Dendrograms of reference site cluster analysis for genus (a) and family (b) level taxonomic resolution.

Once the cluster analysis was complete, a series of discriminant function models were run. The goal of discriminant function analysis in predictive modeling is to generate a probability with which a site belongs to each of the reference cluster groups generated by the cluster analysis. This probability is generated using environmental predictor variables available for each site. Discriminant function analysis itself is a technique used when one has an existing grouping structure and wants to develop a model to predict the group membership of a new observation (Legendre and Legendre 1998). In some applications, one only wants to know into which one group to assign a site. But in the RIVPACS approach, the object is to generate the probability with which a new site belongs to each of the cluster groups. When a non-reference site has physical characteristics that resemble a mixture of a few different reference groups (e.g. along an ecotone), one would expect to find a mixture of the most common taxa found in each of those different groups. The degree of mixture is generated using probabilities derived from discriminant function analysis.

Discriminant analysis requires that the predictor variables meet the assumptions of parametric statistics, although some departure from normality and equal variance is allowed. In addition, collinear variables cannot be used. Collinear variables (a rule of thumb is correlation coefficients (Pearson's r) greater than 0.7) will lead to redundant variables and can disrupt the discriminant analysis, so no collinear combinations were used.

As described, the actual goal of the discriminant function analysis is to generate the probability with which each site belongs to each reference group. The cluster analysis was used to break the continuous distribution of communities into discrete pieces and the discriminant function analysis uses the physical characteristics of those groups, in a sense, to place a site back along that continuous gradient. The membership probabilities are generated using the Mehalanobis distance. The Mehalanobis distance is a multivariate distance measure. It is the distance from any one site to the centroid (a multivariate average) of each of the different groups in multivariate space and is calculated as:

$$D^2 = \overline{d_j} V^{-1} \overline{d_j}'$$

Where D^2 is the squared Mehalanobis distance, $\overline{d_j}$ is a vector of the distances of each predictor between a site and the mean predictor value for group j ($\overline{d_j}'$ is its transpose) and V^{-1} is the inverted covariance matrix of predictors.

The probability a site belongs in each group is derived from those distances – the closer a site is to one centroid, the higher the probability it belongs to that group. These probabilities were calculated using the formula:

$$p_j = q_j / \sum_{j=1}^{k} q_j ,$$

where p_j is the probability a site belongs to group j (of k different groups). The value q_j is a weighted distance measure and is defined as:

$$q_j = n_j \times e^{\left(\frac{-d_j^2}{2}\right)},$$

where n_j is the number of sites in group j and d_j^2 is the squared Mehalanobis distance between the site score and each group mean discriminant function score (Moss et al. 1987). These probabilities are the important outcome of the discriminant function analysis. They are combined with taxa frequencies in each group to predict the final taxonomic composition of a site.

The next calculation is to generate a set of per taxon capture probabalities (P_c). As mentioned all along, the predicted taxa list for a site is not only based on the taxa composition of the one reference group to which a site is most similar. If that were the case, one could simply find the group to which the site had the highest probability of

belonging and compare the observed community to the average community composition of that one group. If all test sites looked exactly like only one reference group, this would be fine. But sites are often physically similar to several groups, since the groupings frequently reflect very subtle differences among reference sites (e.g. low gradient vs. high gradient reaches within one basin). Therefore, this approach predicts a mixture of taxa based on 1) which reference groups a site is most similar to and 2) which taxa are most frequently found in those groups. The P_c, therefore, is a weighted average expected taxon frequency for a site. It weights the per taxon frequencies in each reference group by the probability a site belongs to each of those groups. For example, common taxa from groups to which a site is most similar would have the highest probability of being captured.

In order to do this, the frequency of each taxon in each reference group needs to be calculated. This is done by calculating the frequency with which each taxon is found in each group; $g_{j,x}$ = proportion of reference sites in group j containing taxon x. This value is calculated for each taxon in the master taxa list (over all sites). In the end, each taxon has a frequency with which it occurs in each reference group. Many taxa from the master list are not found in every group; therefore, they will have a frequency of zero where they are absent; others are ubiquitous and have a value near 1.0 for every reference group.

Now that the probability of membership of any site in each reference group (p_j) from the discriminant function analysis and the frequency of every taxon *x* in each reference group ($g_{j,x}$) have been calculated, the probability of capturing (P_c) each taxon *x* at any site can be estimated using the equation:

$$P_{c,x} = \sum_{j=1}^{k} p_j \times g_{j,x}$$, for *k* reference groups.

Note that each probability of capturing a taxon is a continuous probability and not a discrete number. It is derived from the probability of group membership and the distribution of taxa frequencies. The expected number of taxa (E), then, is the sum of the capture probabilities of all the taxa at a site:

$$\mathrm{E} = \sum_{x=1}^{i} P_{c_x} .$$

This total can be the sum of all taxa, but it is common to only sum taxa with a capture probability greater than 0.01 (most taxa) or 0.5 (common taxa). Common taxa models worked best in this exercise.

Generating E was the goal of this exercise –an expected taxa richness for any site. In standard RIVPACS assessment models, this E would then be compared to the observed taxa richness (O) to generate and O/E score – or the percent of expected taxa found at a site. O/E scores were calculated for reference sites used to build the model, since current model diagnostics are based on the distribution of O/E scores for reference sites and not on E alone.

There are a number of potential discriminant models that can be developed using any set of predictors, and discriminant model selection is, obviously, critical. One option is to use stepwise discriminant analysis, but this can lead to locally solved and/or over-fit models. A newer option is to explore the subset of all possible predictor combinations.

An all-subsets routine was developed in the R programming language and was used to identify the best performing models for this project (Vansickle et al., in review). The all-subsets program routine explores all possible predictor combinations and evaluates the 5 best models of each predictor order (1 predictor, 2 predictor, etc.) based on their discrimination of the reference groups using Wilks' lambda, a measure of model discrimination. The program also calculates an O/E score using observed data, and calculates a number of model diagnostics: the standard deviation of O/E among sites (SDOE, a measure of precision), the standard deviation of replicate sampling (SDRS, a measure of the best possible model, Van Sickle et al., in review), a null model O/E score (NULLOE, which calculates E as the average taxon frequency among all reference sites ignoring classification and discriminant models, Van Sickle et al. 2005), and evaluates the extent of model over-fitting by comparing re-substituted and cross-validated model classification efficiencies. All of these criteria were considered in selecting the best overall model.

Model Results

The best model produced, that balanced precision with sensitivity, was a model using family level taxonomy, a 9-group cluster, and a capture probability (P_c) of 0.5. This is considered the principal model. In order to generate a genus prediction as well, the best genus level model, using a 7-group cluster and P_c>0.5, is also included in the output, as are E for all taxa (P_c>0.01) for each taxonomic resolution (genus and family).

The models produce an overall estimate of the expected (E) number of families with P_c>0.5 and P_c>0.01 and the expected number of genera with P_c>0.5 and P_c>0.01. They will also produce a list of which families and genera are expected along with their individual capture probabilities.

Table 3 Model results for best family and genus models.

Family		Mean	SD	Best	Null	% Explained
9 group	Pc>0.5	1	0.25	0.18	0.32	50
Predictors: Area, Elevation, Hypsometry, Longitude, Percent soil type D						
Genus		Mean	SD	Best	Null	% Explained
7 group	Pc>0.5	1	0.31	0.27	0.46	79
Predictors: Area, Elevation, Hypsometry, Longitude, 2-year precipitation, and Relief						

Family Model

The family model used the following predictors: Area, Elevation, Hypsometry, Longitude, and Percent of Soils in Hydrologic Group D (Table 3). Distribution of O/E scores in reference sites using the family model are shown in Figure 3. Since these are all reference sites, the first diagnostic is whether or not the average score is equal to 1. Significant departure from 1 would indicate some error or bias in the model. Clearly, the model means were comparable to 1 (Figure 3 and Table 3). The second diagnostic is

how precise the estimates are. The standard deviation of reference O/E scores (SDOE) is an estimate of the model precision and is an indicator of the range around 1 that can be considered comparable to reference. Small SDOEs (approximately 0.15) are the goal of models. These models were relatively less precise than this goal (family model SDOE = 0.25, Table 3). To put this number in context, however, the models produce two other values – a null model SD and a best possible model SD. The null model capture probability is simple the frequency of each taxon across all reference sites. It removes the weighting by reference group – since there is no grouping. It is a measure of how much additional benefit is gained by classification and discriminant modeling. The null model SD was 0.32, so the model is performing better than null. In addition, the best possible model SD was 0.18 – this is an estimate of the best possible precision that can be achieved with the data. So, the model is accounting for approximately 50% of the explainable variation.

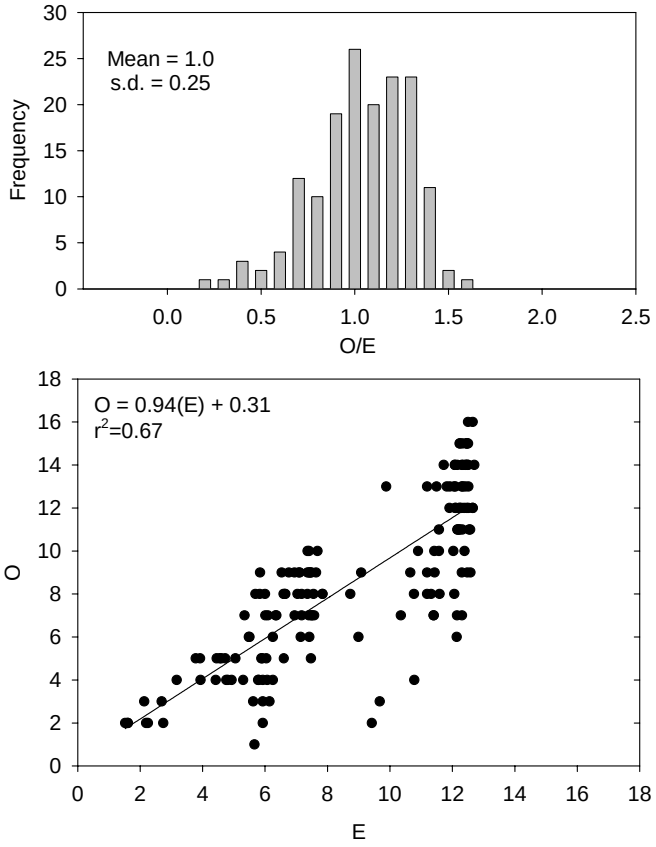


Figure 3 Plot of the frequency of O/E scores and E vs O values using the Family, 9-group, P_c>0.5 model

The comparison of the modeled E to O is also an indicator of model quality. In Figure 3, this is plotted as a linear regression. Clearly there is good agreement between the two, the slope is similar to 1, meaning that E is a good predictor of O and the model explains two-thirds of the variance in the observed data.

With these results in mind, the 9 group model using P_c>0.5 was selected as the best Family model. It is not performing as well as generally hoped for these types of predictive models. This is likely a function of the small sample sizes (100 individuals) which do not provide enough information to consistently characterize site richness. However, the model does provide an accurate prediction of the expected taxa in reference sites with known precision (0.25). The model output in GISHydro presents the expected family taxa richness for taxa with a capture probability > 0.5 and a list of those families with their capture probabilities.

Genus Model

The genus model used the following predictors: Area, Elevation, Hypsometry, Longitude, 2-year Precipitation, and Relief (Table 3 and Figure 4). Similar to the family model, the reference distribution was centered on 1, as expected. However, the standard

deviation of the reference distribution for these sites was larger (0.31) than the family model (Table 3). The genus null model was larger than that for the family model (Null model SD = 0.46) and the best possible model was less precise than that for the family model (best model SD = 0.27). As a result, the genus model did predict a larger percent of the predictable variance (79%) than the family model, even though the overall model precision was lower.

Looking at the model prediction visually, the genus model E also provided a reliable prediction of O (Figure 4) with a comparable amount of variance explained (0.70).

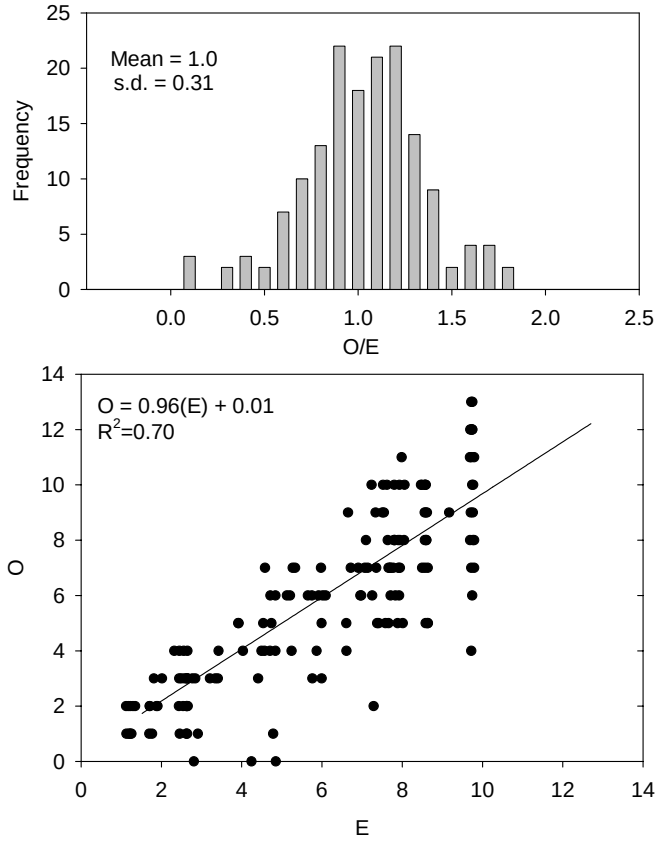


Figure 4 Plot of the frequency of O/E scores and E vs O values using the Genus, 7-group, P_c>0.5 model

Model Output

The biological models in GISHydro produce an expected family and genus taxa richness based on the above models. They will also generate expected richness estimates based on the P_c>0.01 criteria. These models were less precise and that output should be used more conservatively. The models also produce a list of the predicted taxa along with their capture probabilities (Table 4). These results will hopefully be useful for modeling purposes. It is tempting to extrapolate these results to a variety of applications, but a few important caveats need to be kept in mind:

- Not all taxa are predicted – only those expected based on reference sites. New taxa encountered at a site are not considered
- These models do not accurately predict rare taxa
- These are not assessment models. They are designed as an indicator of relative biological richness. Users interested in approved bioassessment tools should consult the Maryland Biological Stream Survey for appropriate methods (http://www.dnr.state.md.us/streams/mbss/).
- Again, these models are imprecise relative to other predictive models that have been developed

Table 4 Sample output for the family model from biological module of GISHydro using the predictive models.

Overall Expected Family Richness:

E(0.5):	7.8254
E(1.0):	13.8378

Family:	Probability:	Family:	Probability:
Aeshnidae	0.03	Hydropsychidae	0.76
Asellidae	0.13	Hydroptilidae	0.07
Baetidae	0.62	Lepidostomatidae	0.09
Brachycentridae	0.01	Leptoceridae	0.07
Branchiobdellidae	0.02	Leptophlebiidae	0.11
Calopterygidae	0.03	Leuctridae	0.24
Cambaridae	0.01	Limnephilidae	0.35
Capniidae	0.13	Lumbriculidae	0.07
Ceratopogonidae	0.16	Naididae	0.3
Chironomidae	1	Nemouridae	0.83
Chloroperlidae	0.08	Odontoceridae	0.04
Collembola	0.03	Oligoneuriidae	0.19
Corydalidae	0.23	Palaemonidae	0.03
Crangonyctidae	0.03	Peltoperlidae	0.05
Dixidae	0.05	Perlidae	0.35
Elmidae	0.84	Perlodidae	0.21
Empididae	0.58	Philopotamidae	0.49
Enchytraeidae	0.07	Physidae	0.01
Ephemerellidae	0.73	Pisidiidae	0.05
Ephemeridae	0.09	Planariidae	0.09
Gammaridae	0.16	Pleuroceridae	0.03
Glossosomatidae	0.34	Polycentropodidae	0.3
Gomphidae	0.01	Psephenidae	0.05
Gordiidae	0.04	Psychomyiidae	0.11
Heptageniidae	0.84	Pteronarcidae	0.09

Land Use Effect Model

In addition to the predictive models, a very simple land use effects model was built. This model is intended to model the potential impact of land use transformation on O/E scores from Maryland streams based on the predictive models developed using GISHydro predictors. This can be used as one guide in considering the potential impacts, in a very general sense, of land use change on biological resources. These are not per taxon models and cannot be used to look at impacts on rare, threatened, or endangered taxa – arguably the most important to consider.

The first approach considered was modeling the probability or likelihood of O/E scores for a given level of taxa loss (e.g., 0.5) at different land cover amounts. However, these models were fairly difficult to produce and interpret. In addition, the O/E cutoffs were arbitrary and it seemed as if users may be more interested in continuous response models. For this reason, simple correlation and regression models were built.

The GISHydro output for this modeling exercise only included gross level information for a few land cover variables: agricultural, forested and impervious land cover. It also produced storage area estimates, but these were not considered. So, these 3 land use variable were used to predict O/E scores.

Table 5 Table of correlation and regression model results of land cover and O/E scores.

Correlation Coefficients

Variable	Family O/E	Genus O/E
Forest	0.19	0.11
Agriculture	0.09	0.11
Impervious Area	-0.26	-0.18

Multiple Regression Models

Variable	Slope	s.e	t	p-level	R2
Family Pc>0.5 model					
Intercept	0.945	0.020	47.48	0.00	0.07
Agriculture	-0.006	0.022	-0.27	0.78	
Imperviousness	-0.408	0.041	-10.04	0.00	
Genus Pc>0.5 model					
Intercept	0.862	0.026	33.11	0.00	0.03
Agriculture	0.060	0.029	2.07	0.04	
Imperviousness	-0.319	0.053	-5.99	0.00	

Linear Regression Models

Variable	Slope	s.e	t	p-level	R2
Family Pc>0.5 model					
Intercept	0.973	0.014	69.48	0.00	0.09
Imperviousness	-0.487	0.042	-11.55	0.00	
Genus Pc>0.5 model					
Intercept	0.952	0.018	51.68	0.00	0.05
Imperviousness	-0.476	0.055	-8.59	0.00	

All land cover variables were first transformed using the standard arcsine-square root transformation used for percentile data (Sokal and Rohlf 1995). These were then explored for their relation to the family and genus models using correlation analysis (Table 5). All correlations were significant; however, they were weak and explained little of the variability – not surprising given the large sample size and low number of land cover variables. Biological condition increased with forested land cover and decreased with imperviousness. Strangely, condition also increased with agriculture, although very weakly. This last

relationship is due, likely, in part to the fact that urbanization is having a larger impact than agriculture. Therefore, watersheds with low agricultural cover have, understandably, potentially larger urban land cover and, therefore, lower condition scores leading to the perceived effect. This inherent co-linearity among land cover variables is well known and precludes the use of all 3 variables in any one model. Since impervious area incorporates forest loss, the forested land cover variable was removed. The remaining variables, imperviousness and agriculture, were placed into a multiple regression model, with imperviousness first and then agriculture.

The multiple regression models showed a strong effect of imperviousness and a relatively weak or insignificant effect of agriculture (Table 5). For these reasons, simple linear regression models were built using imperviousness data alone and excluding values of imperviousness equal to 0, since it was difficult to discriminate where this was due to no impervious cover versus missing data. These models indicate a fairly strong decline in O/E scores with imperviousness (Table 5).

The final equations recommended at this time are to model O/E scores using imperviousness alone. These equations are:

$$\text{Family O/E Score} = 0.97 - 0.49\,(\arcsin\sqrt{\text{Imperviousness}})$$

$$\text{Genus O/E Score} = 0.95 - 0.48\,(\arcsin\sqrt{\text{Imperviousness}})$$

These equations yield the general linear relationship between imperviousness and O/E scores seen in Figure 5.

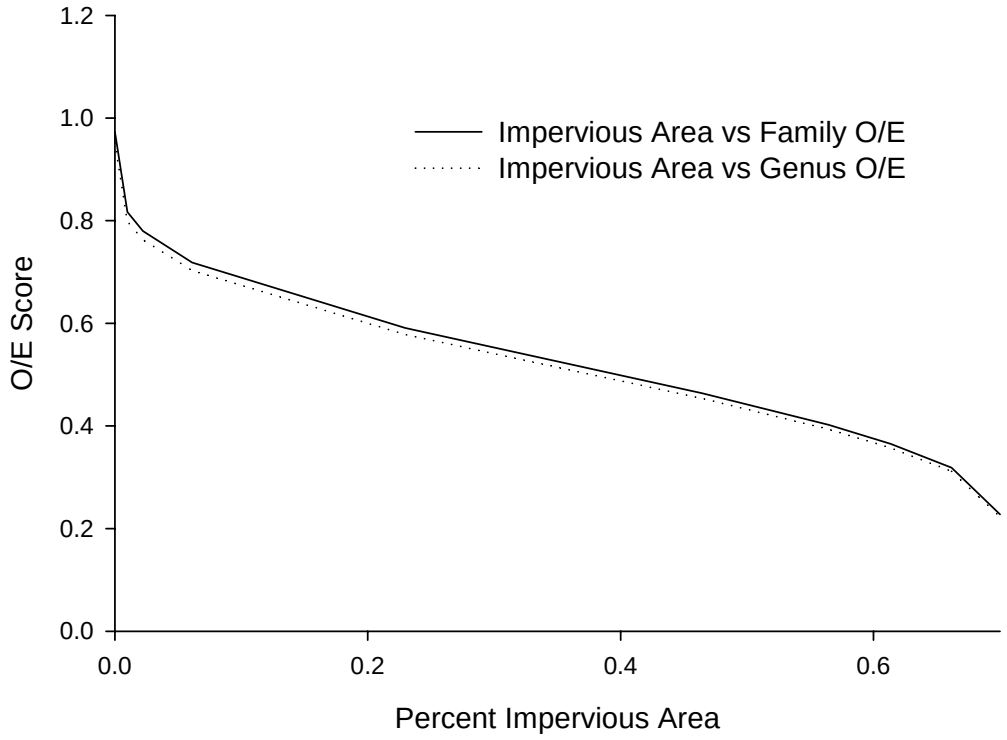


Figure 5 Model of family and genus O/E decline with impervious area

References

CH2M-Hill (2000). Technical Memorandum 1, Urban Stormwater Pollution Assessment, prepared for North Carolina Department of Environment and Natural Resources, Division of Water Quality.

Davies, P.E. 2000. Development of a national river bioassessment system (AUSRIVAS) in Australia. in Pages 113-124. Wright, J.F., D.W. Sutcliffe, and M.T. Furse (editors). Assessing the biological quality of freshwaters: RIVPACS and other techniques. Freshwater Biological Association, Ambleside, Cumbria, United Kingdom.

Dillow, J. J. A. (1996). "Technique for estimating magnitude and frequency of peak flows in Maryland." U.S. Geological Survey, Water-Resources Investigations Report 95-4154, 55p.

Driver, N.E. and G.D. Tasker (1990). "Techniques for estimation of storm-runoff loads, volumes, and selected constituent concentrations in urban watersheds in the United States." U.S. Geological Survey, Water Supply Paper 2363, 44 p.

Furse, M.T., D. Moss, J.F. Wright, P.D. Armitage, and R.J.M. Gunn. 1984. The influence of seasonal and taxonomic factors on the ordination and classification of running-water sites in Great Britain and on the prediction of their macroinvertebrate communities. Freshw. Biol. 14:257-280.

Hawkins, C.P., R.H. Norris, J.N. Hogue, and J.W. Feminella. 2000. Development and evaluation of predictive models for measuring the biological integrity of streams. Ecological Applications 10:1456-1477.

Legendre, P., and L. Legendre. 1998. Numerical ecology: Second english edition. Elsevier, New York.

Marchant, R., L. Barmuta, and B.C. Chessman. 1995. A preliminary study of the ordination and classification of macroinvertebrate communities from running waters in Victoria, Australia. Austral. J. Mar. Freshw. Res. 45:945-962.

Moss, D., M.T. Furse, J.F. Wright, and P.D. Armitage. 1987. The prediction of the macro-invertebrate fauna of unpolluted running-water sites in Great Britain using environmental data. Freshw, Biol. 17:41-52.

Natural Resources Conservation Service (2005a). "Soil Survey Geographic (SSURGO) Database" <http://www.ncgc.nrcs.usda.gov/products/datasets/ssurgo/>(January 21, 2005).

Natural Resources Conservation Service (2005b). "State Soil Geographic (STATSGO) Database" <http://www.ncgc.nrcs.usda.gov/products/datasets/statsgo/>(January 21, 2005).

Paul, M.J., J. Gerritsen, and E. Leppo. 2002. Multivariate predictive invertebrate bioassessment models for Illinois streams. Draft Technical Report. Prepared by Tetra Tech, Inc. for Illinois Environmental Protection Agency, Springfield, IL.

Ragan, R. M., Berich, R. H, Merkel, W. H., Moglen, G. E., Thomas, W. O., Jr., and Woodward, D. E. (2004). "Application of Hydrologic Methods in Maryland: A Report Prepared by the Maryland Hydrology Panel." <http://www.gishydro.umd.edu/HydroPanel/index.html> (January 21, 2005).

Roth, N.E., Southerland, M.T., Chaillou, J.C., Kazyak, P.F. and Stranko, S.A. (2000). Refinement and Validation of a Fish Index of Biotic Integrity for Maryland

Streams. Prepared by Versar, Inc., Columbia, MD, with Maryland Department of Natural Resources, Monitoring and Non-Tidal Assessment Division.

Simpson, J.C., and R.H. Norris 2000. Biological assessment of river quality: development of AUSRIVAS models and outputs. in Pages 125-142. Wright, J.F., D.W. Sutcliffe, and M.T. Furse (editors). Assessing the biological quality of freshwaters: RIVPACS and other techniques. Freshwater Biological Association, Ambleside, Cumbria, United Kingdom.

Soil Conservation Service (1984). "Computer Program for Project Formulation, *Technical Release* 20." Washington, DC.

Soil Conservation Service (1986). "Urban Hydrology for Small Watersheds, *Technical Release* 55." Washington, DC.

Sokal, R.R. and F.J. Rohlf. 1995. Biometry. 3rd Edition. W.H. Freeman, New York.

United States Environmental Protection Agency (2001). "PLOAD version 3.0: An ArcView GIS Tool to Calculate Nonpoint Sources of Pollution in Watershed and Stormwater Projects: User's Manual". January, 2001, 44p.

United States Environmental Protection Agency (2005). "EPA MRLC National Land Cover Data (NLCD)", < http://www.epa.gov/mrlc/nlcd.html>(January 21, 2005).

United States Geological Survey (2005a). "National Elevation Dataset." <http://ned.usgs.gov/>(January 20, 2005).

United States Geological Survey (2005b). "National Hydrography Dataset Home Page." <http://nhd.usgs.gov/>(January 20, 2005).

Van Sickle, J., C.P. Hawkins, D.P. Larsen, and A. T. Herlihy. 2005. A null model for the expected macroinvertebrate assemblage in streams. Journal of the North American Benthological Society 24:178-191.

Van Sickle, J., D.D. Huff, and C.P. Hawkins. In review. Selecting discriminant function models for predicting the expected richness of aquatic macroinvertebrates.

Wright, J.F. 1995. Development and use of a system for predicting the macroinvertebrate fauna in flowing waters. Australian J. Ecol. 20:181-197.

Wright, J.F. 2000. An introduction to RIVPACS. in Pages 1-24. Wright, J.F., D.W. Sutcliffe, and M.T. Furse (editors). Assessing the biological quality of freshwaters: RIVPACS and other techniques. Freshwater Biological Association, Ambleside, Cumbria, United Kingdom.

Wright, J.F., M.T. Furse, and P.D. Armitage. 1993. RIVPACS: a technique for evaluating the biological water quality of rivers in the UK. Euro. Wat. Poll. Control 3:15-25.

Wright, J.F., D. Moss, P.D. Armitage, and M.T. Furse. 1984. A preliminary classification of running-water sites in Great Britain based on macroinvertebrate species and the prediction of community type using environmental data.. Freshw. Biol. 14:221-256.

Appendices

EMC Values – Maryland Department of Planning Generalized Land Use and "Ragan" Land Use. (based on CH2M-Hill, 2000)

ID	Classification	Nitrogen (mg/L)	Phosphorus (mg/L)	TSS (mg/L)
11	Low Density Residential	1.2	0.2	22.1
12	Medium Density Residential	1.7	0.2	30.5
13	High Density Residential	2.7	0.3	47.7
14	Commercial	3.1	0.4	54.2
15	Industrial	3.3	0.4	57.8
16	Institutional	2.4	0.3	41.9
17	Extractive	0	0	0
18	Open Urban Land	1.1	0.2	20.0
21	Cropland	1.1	0.2	19.2
22	Pasture	1.1	0.2	19.2
23	Orchards	1.1	0.2	19.0
24	Feeding Operations	0	0	0
25	Row Crops	1.1	0.2	19.2
41	Deciduous Forest	1.1	0.2	19.0
42	Evergreen Forest	1.1	0.2	19.0
43	Mixed Forest	1.1	0.2	19.0
44	Brush	1.1	0.2	19.0
50	Water	0	0	0
60	Wetlands	1.1	0.2	19.0
70	Barren Land	1.5	0.2	26.8
71	Beaches	0	0	0
72	Bare Exposed Rock	1.5	0.2	26.8
73	Bare Ground	0	0	0
80	Transportation	3.3	0.4	57.8
191	Large Lot Agricultural	1.1	0.2	19.2
192	Large Lot Forest	1.1	0.2	19.0
241	Feeding Operations	0	0	0
242	Agricultural Buildings	0	0	0

* Note: if the user of this program wishes to change these default values, he/she should edit the tab-delimited file located at: drive:/umdgism/mdinterface/mdpemclookup.txt.

EMC Values – USGS-GIRAS Land Use (1970's Land Use). (based on CH2M-Hill, 2000)

ID	Classification	Nitrogen (mg/L)	Phosphorus (mg/L)	TSS (mg/L)
11	Residential	1.7	0.2	30.5
12	Commercial	3.1	0.4	54.2
13	Industrial	3.3	0.4	57.8
14	Transportation	3.3	0.4	57.8
15	Ind./Comm. Comp.	3.3	0.4	57.8
16	Mixed Urban	1.7	0.2	30.5
17	Other Urban	1.1	0.2	20.0
21	Cropland/Pasture	1.1	0.2	19.2
22	Orchards	1.1	0.2	19.2
23	Feeding Operations	0	0	0
24	Other Ag. Land	1.1	0.2	19.2
41	Deciduous Forest	1.1	0.2	19.0
42	Evergreen Forest	1.1	0.2	19.0
43	Mixed Forest	1.1	0.2	19.0
51	Streams/Canals	0	0	0
52	Lakes	0	0	0
53	Reservoirs	0	0	0
54	Bays/Estuaries	0	0	0
61	Forested Wetlands	1.1	0.2	19
62	Non-Forested Wetlands	1.1	0.2	19.0
70	Barren Land	1.5	0.2	26.8
71	Dry Salt Flats	0	0	0
72	Beaches	0	0	0
73	Sandy Areas	0	0	0
74	Bared Exposed Rock	1.5	0.2	26.8
75	Strip Mines/Quaries	0	0	0
76	Transitional	1.5	0.2	26.8
77	Mixed Barren	1.5	0.2	26.8

* Note: if the user of this program wishes to change these default values, he/she should edit the tab-delimited file located at: drive:/umdgism/mdinterface/usgsemclookup.txt.

EMC Values – EPA-MRLC Land Cover. (based on CH2M-Hill, 2000)

ID	Classification	Nitrogen (mg/L)	Phosphorus (mg/L)	TSS (mg/L)
11	Water	0	0	0
21	Low intensity developed	1.2	0.2	22.1
22	High intensity residential	2.7	0.3	47.7
23	High intensity commercial/industrial	3.3	0.4	57.8
24	Urban - Unknown	1.1	0.2	20.0
31	Bare rock/sand	1.5	0.2	26.8
32	Quarries/strip mines/gravel pits	0	0	0
33	Transitional barren	1.5	0.2	26.8
41	Deciduous forest	1.1	0.2	19.0
42	Evergreen forest	1.1	0.2	19.0
43	Mixed Forest	1.1	0.2	19.0
81	Hay/pasture	1.1	0.2	19.2
82	Row crops	1.1	0.2	19.2
85	Other grass (lawns	1.1	0.2	20.0
90	Wetland - Unknown	1.1	0.2	19.0
91	Woody wetland	1.1	0.2	19.0
92	Emergent herbaceous wetland	1.1	0.2	19.0
95	Wetland - Unknown	1.1	0.2	19.0

* Note: if the user of this program wishes to change these default values, he/she should edit the tab-delimited file located at: drive:/umdgism/mdinterface/mrlcemclookup.txt.

EMC Values – Zoned Land Use. (based on CH2M-Hill, 2000)

ID	Classification	Nitrogen (mg/L)	Phosphorus (mg/L)	TSS (mg/L)
10	Urban	1.2	0.2	22.1
11	Low Density Residential	1.2	0.2	22.1
12	Medium Density Residential	1.7	0.2	30.5
13	High Density Residential	2.7	0.3	47.7
14	Commercial	3.1	0.4	54.2
15	Industrial	3.3	0.4	57.8
16	Institutional	2.4	0.3	41.9
17	Extractive	0	0	0
18	Open Urban Land	1.1	0.2	20.0
20	Agriculture	1.1	0.2	19.2
21	Cropland	1.1	0.2	19.2
22	Pasture	1.1	0.2	19.2
23	Orchards	1.1	0.2	19.0
24	Feeding Operations	0	0	0
25	Row Crops	1.1	0.2	19.2
40	Forest	1.1	0.2	19.0
41	Deciduous Forest	1.1	0.2	19.0
42	Evergreen Forest	1.1	0.2	19.0
43	Mixed Forest	1.1	0.2	19.0
44	Brush	1.1	0.2	19.0
50	Water	0	0	0
60	Wetlands	1.1	0.2	19.0
70	Barren Land	1.5	0.2	26.8
71	Beaches	0	0	0
72	Bare Exposed Rock	1.5	0.2	26.8
73	Bare Ground	0	0	0
80	Transportation	3.3	0.4	57.8
111	Res.: 2.00 ac <=x	1.1	0.2	20.0
112	Res.: 1.00<=x<2.00 ac	1.1	0.2	21.1
113	Res.: 0.50<=x<1.00 ac	1.2	0.2	22.1
114	Res.: 0.33<=x<0.50 ac	1.5	0.2	26.3
115	Res.: 0.25<=x<0.33 ac	1.7	0.2	30.5
116	Res.: x <0.25 ac	2.7	0.3	47.7
191	Large Lot Agricultural	1.1	0.2	19.2
192	Large Lot Forest	1.1	0.2	19.0
241	Feeding Operations	0	0	0
242	Agricultural Buildings	0	0	0

* Note: if the user of this program wishes to change these default values, he/she should edit the tab-delimited file located at: drive:/umdgism/mdinterface/zoningemclookup.txt.