

MULTISCALE ADVANCED RASTER MAP ANALYSIS SYSTEM
Network-Based Analysis of Biological Integrity in Freshwater Streams

G. P. Patil
Center for Statistical Ecology and Environmental Statistics
Department of Statistics
The Pennsylvania State University
University Park, PA 16802
<http://www.stat.psu.edu/~gpp>

Research and Outreach Team

Ganapati P. Patil, Center for Statistical Ecology and Environmental Statistics,
The Pennsylvania State University

Wayne L. Myers, School of Forest Resources and Office of Remote Sensing of Earth
Resources, The Pennsylvania State University

Denice H. Wardrop, Cooperative Wetlands Center, The Pennsylvania State University

C. Taillie, Center for Statistical Ecology and Environmental Statistics,
The Pennsylvania State University

Joe Bishop, School of Forest Resources and Office of Remote Sensing of Earth Resources,
The Pennsylvania State University

Kristen Hychka, Cooperative Wetlands Center, The Pennsylvania State University

MULTISCALE ADVANCED RASTER MAP ANALYSIS SYSTEM
Network-Based Analysis of Biological Integrity in Freshwater Streams

PROJECT SUMMARY	3
PROJECT DESCRIPTION.....	4
PROBLEM STATEMENT.....	4
SCOPE OF WORK.....	6
INFORMATION RESOURCES.....	7
PHASE 1. IDENTIFICATION OF CRITICAL SUB-NETWORKS OF THE STREAM-BASED NETWORK.....	9
PHASE 2. COMPILATION, ANALYSIS, AND APPLICATION OF RELEVANT FILTERS	10
PHASE 3. DEVELOPMENT OF PROCESS MODELS OF STREAM SYSTEMS AS NODAL NETWORKS	11
SPECIFIC METHODOLOGICAL TASKS FOR PHASES 1 AND 2	13
<i>Network Scan Statistic (Task 1)</i>	14
<i>Continuous Response Distributions (Task 2)</i>	15
<i>Filtering for Explanatory Variables (Task 3)</i>	16
NETWORK MODELING METHODS FOR PHASE 3	17
PROJECT OUTCOMES AND DELIVERABLES.....	17
PROJECT WORK GROUPS	17
INTEGRATION OF RESEARCH, EDUCATION, AND DISSEMINATION	18
PERTINENT BACKGROUND AND SUPPORT	18
REFERENCES	19

Project Summary

Human impacts on hydrologic systems of the Mid-Atlantic Region are well documented (e.g., Paul et al., 1998). Degradation of surface waters and associated threats to human and aquatic life have spawned responses to remediate both point and nonpoint pollution [e.g., Clean Water Act of 1972 (CWA), Chesapeake Bay Agreement], but the original intent to restore and maintain the chemical, physical, and biological integrity of the Nation's waters has not been completely attained. Thus, improvement of the ecological condition of the waters of the U.S. continues to be a broadly pursued goal of both the regulatory and the scientific communities. We propose that the current state of the science identifies several issues: 1) need for a hierarchical process to identify critical stream segments with progressively greater specificity; 2) need for improvement in the articulation of landscape properties relevant to aquatic resource condition for those critical areas that remain unexplainable with existing information.

This project proposes a new analysis of stream networks that allows three integrative tasks to be accomplished:

- Identification of critical portions (i.e., subnetworks) of low biological integrity on the stream systems of the Commonwealth of Pennsylvania. This will be accomplished by a network version of the scan statistic originally developed for geographical surveillance of disease (Kulldorff and Nagarwalla, 1995; Kulldorff, 1997), and adapted to regional environmental and ecological setting (Myers et al., 2002; Patil et al., 2003).
- Reassessment of critical subnetworks taking into account potential explanatory factors. Three categories of factors will be evaluated: descriptive, structural, and network-related.
- Network modeling of those critical areas which are not satisfactorily accounted for in the preceding step by general landscape and drainage variables, to provide the etiology necessary for full understanding of the relationship between aquatic resource condition and landscape factors.

The treatment of the stream system as a formal network is critical to the accomplishment of all three, and the project goals could not be achieved with alternative representations of the aquatic resource, using spatial units such as watersheds. Identification of critical subnetworks will be achieved by reformulating a statistical hotspot-detection methodology (scan statistic) so that scanning is across a dendritic network instead of across spatial polygons. A key feature of the scan statistic is its explicit evaluation of statistical significance, thus controlling for false alarms in its localization of impairment. New characterizations for articulating linkage between streams and landscapes will be obtained by direct parameterizations of the formal networks and indirect association with the physical network by having nodes as objects of attribution. Incorporation of explanatory variables in evaluating criticality will be achieved by using these variables as 'filters' in the scan statistic.

Completion of the project will serve both to extend substantially the understanding of the biological condition of Pennsylvania's freshwater aquatic systems and to provide a scientific template and a computational statistical toolkit for addressing freshwater aquatic systems of other states in like manner.

Project Description

Problem Statement

Along the mid-Atlantic of the U.S., most major cities (e.g., Philadelphia, Baltimore, Richmond) are located along the fall line between the Piedmont and Coastal Plain, where historic navigation was halted by steep stream gradients and where hydropower was plentiful. Many other cities and towns sprang up on peninsulas (e.g., Norfolk, VA), confluences (e.g., Washington DC), and floodplains of major and minor rivers (e.g., Harrisburg, PA). The impacts to freshwater aquatic resources by these human habitations have been devastating and well-documented (e.g., Paul et al., 1998). Threats to human and aquatic life continue to be major issues of concern. The degradation of surface waters has spawned responses to remediate both point and nonpoint pollution [e.g., Clean Water Act of 1972 (CWA), Chesapeake Bay Agreement]. Through implementation of the CWA and derivative state laws, “fishable and swimmable” conditions have been restored in some areas. However, the original intent of the CWA “to restore and maintain the chemical, physical, and biological integrity of the Nation’s waters,” has not been completely attained. Thus, the improvement of the ecological condition of the waters of the U.S. continues to be a broadly pursued goal of both the regulatory and scientific communities.

Within this general pursuit, the range of individual scientific questions is almost limitless. However, management efforts, and their attendant scientific issues, are generally and collectively directed at answering four questions:

1. *Management:* Where is the resource (i.e., resource inventory)? *Science:* What are the relevant natural factors that determine both ecosystem structure and an individual ecosystem’s response to stress?
2. *Management:* What is the condition of the resource? Is there an impairment of condition and, if so, how do we characterize it? *Science:* What is the role of disturbance in structuring ecosystems? Can we evaluate ecological functioning and/or susceptibility? Can ecosystem function and disturbance be evaluated remotely?
3. *Management:* Where does the impairment occur? *Science:* Can susceptibility be related to a combination of environmental conditions that can be set in a hierarchical framework?
4. *Management:* How can we improve the ecological functioning of the impaired system? *Science:* Which conditions can be altered to change ecosystem susceptibility and functioning, and where in the hierarchical framework do these occur?

Most states are in the process of addressing the first two questions, through requisite monitoring and assessment activities mandated by the CWA, and a wealth of scientific effort has supported these activities. The latter two questions remain largely open to the scientific community, and the quality of our response is dependent upon our ability to build a conceptual model of the role of natural factors in structuring an ecosystem, evaluate the potential impacts of anthropogenic activities on that ecosystem, and elucidate our ability to restore portions of the impacted ecosystem.

Based on indicators of biotic integrity, many streams are in poor condition (USEPA, 2000). Both the abiotic (Poff and Ward, 1990; Johnson et al., 1997) and biotic (Harding et al., 1998) conditions of an aquatic community are affected by historic and current characteristics of its basin. Within the mid-Atlantic, some of the main stressors to local streams are excess sediment,

riparian degradation, mine drainage, acid deposition, excess nutrients, and exotic species (USEPA, 2000). Region-wide stressors include loss of riparian habitat, farming in riparian areas and on steep slopes, road crossings, and forest fragmentation (Jones et al., 1997). Most streams are degraded by more than one stressor (Karr, 1981; Minshall et al., 1983).

In light of multiple stressors, classification tools are often utilized as a basis for understanding the role of natural factors in structuring an ecosystem and/or its response to stressors. Many of these assume that the ecological health of a watershed reflects attributes of the transmission, storage, and release of water. The influences of climate, soils, and topography, for example, drive the channel's hydrologic and fluvial geomorphologic processes (Leopold et al., 1964; Lotspeich, 1980). These processes in turn affect water quality, flow regime, physical habitat, food and energy sources, and biotic interactions (Karr and Chu, 1999), which collectively affect biotic communities. Classification tools can be geographically dependent or independent, hierarchical or non-hierarchical, incorporate a fixed or sliding scale, and based on structure or function (Detenbeck et al., 2000). Examples of hierarchical, geographically dependent classifications include ecoregions (Omernik, 1987; Bryce and Clarke, 1996), aquatic regions (Maxwell et al., 1995), and functional groupings (Hawkins et al., 2000). Maxwell et al. (1995) grouped watersheds by indicators of watershed function, including geoclimate, zoogeographic pattern, watershed morphology, and disturbance history. To reveal function in a geographically dependent scheme, the Nature Conservancy developed a nested framework for biotic and abiotic aquatic classification based on two landscape-level regions and two levels of smaller-scale habitat characterization (Lammert et al., 1997).

Geographically independent, hierarchical classifications rely on stream hydrogeomorphic (HGM) structure (Rosgen, 1996) and wetland HGM function (Brinson, 1993). Imhof et al. (1996) developed a scheme based on physical processes that drive fish abundance. Other examples classify watersheds based on water chemistry and/or temperature (e.g., Richards, 1990; Seelbach et al., 1997), and hydrology and geomorphology and/or sediment size (e.g., Whiting and Bradley, 1993; Seelbach et al., 1997).

Ecosystems are also classified by measures of ecological resistance, or the ability of the system to withstand perturbation (Forman and Godron, 1986). Resistance, plus resilience, which is the ability of a system to return to its original state after a disturbance (Gunderson, 2000), are the primary determinants of ecosystem stability. Aquatic systems are dynamic, but if resistance is overwhelmed by disturbance, then its average long-term state changes (Reeves et al., 1995). Poff and Ward (1990) suggested that because stream communities reflect the history of disturbance they are resistant. Others have suggested that resistance is enhanced by material retention (Minshall et al., 1983) or storage capacity (Detenbeck et al., 2000).

Classifications may also be based on measures of disturbance, which require the articulation of relevant landscape metrics that can be used to monitor aquatic integrity regionally. Studies have investigated the relationship between in-stream measures of fish and invertebrate communities to landscape metrics, specifically land use measured across the contributing area or in a fixed 50- to 200-m stream buffer zone in the contributing area to a sample point. These studies have shown conflicting results, with one showing the highest correlation with local or buffer characteristics (Lammert and Allan, 1999), others with the highest correlation to catchment-wide

characterization (Roth et al., 1996; Allan et al., 1997), and some showing correlation to both buffer and catchment-wide variables (Pan et al., 1999; Richards et al., 1996). The conflicting results of these studies may be the result of differing spatial scales of the projects study areas (Lammert and Allan, 1999) and the processes influencing the aquatic community (Wiley et al., 1997). Strong interactions between basin and local conditions exist, particularly in terms of channel morphology (Richards et al., 1997) and physical habitat (Imnof et al., 1996).

This review illustrates that the ability to build a conceptual model of the role of natural factors in structuring an ecosystem, evaluate the potential impacts of anthropogenic activities on that ecosystem, and restore portions of the impacted ecosystem has not been effectively demonstrated. The current state of the science suggests the following gaps, or issues:

Issue 1: Current studies need to be focused in their efforts, implying the need for a hierarchical identification of critical areas, i.e., areas of low biological condition for which explanatory variables are not known. The power of existing studies in contributing to this understanding must be testable, and areas which require new approaches need to be identified. In addition, the identification of critical areas must be at a scale compatible with restoration efforts. For example, the identification of a cluster of watersheds as an area of concern may assist the formation of regional policy, but does not assist a local watershed group in determining relevant and appropriate sections of stream for riparian restoration. Restoration efforts are generally performed at the scale of individual stream reaches. Therefore, the ability to identify critical sections of a stream network is necessary.

Issue 2: Improvement in the articulation of both the structure of stream drainages and landscape properties relevant to aquatic resource condition is necessary for two reasons: 1) to further increase the explanatory power of those landscape properties which are recognized as important predictors of aquatic resource condition; and 2) to construct and test relationships for those areas of low biological condition for which explanatory variables are not known. The representation of stream drainages as stream-based networks could potentially improve our evaluation of the potential impacts of anthropogenic activities on aquatic ecosystems, by opening up a whole new series of network analysis tools for use in understanding these critical relationships. Use of network-based analysis has been effectively used in a number of ecological studies, and its expanded use has been called for (Dale, 2002). This new representation of the aquatic resource of interest (streams) can serve both as a spatial basis for the identification of critical sub-networks, as well as the basis of a network modeling effort. This network-based analysis will require development of relevant expressions of landscape properties, e.g., description of land use at network junctures may be more useful than that catchment-wide.

Scope of Work

The work proposed herein is depicted in project format in Figure 1. This project proposes a new representation of the stream network, termed a stream-based network, which allows three integrative exercises to be performed;

- Identification of critical sub-networks of low biological integrity on the stream systems of the Commonwealth of Pennsylvania,
- Elucidation of effective explanatory variables (termed herein as filters), and

- Network modeling of those critical areas which are not resolved with the application of the specified filters, to provide the etiology necessary for full understanding of the relationship between aquatic resource condition and landscape factors.

The use of network modeling techniques in only critical sub-networks that are not adequately explained by the filters maximizes the effectiveness of this process; network modeling of the entire network would be too process-intensive to be feasible. The representation of the stream network as a network, per se, is critical to the performance of all three, and the project could not be achieved with alternative representations of the aquatic resource, such as watersheds. A detailed description of the proposed information resource and each phase follows.

Information Resources

Pennsylvania's plan for achieving a comprehensive statewide assessment of its surface waters includes implementation of a program to evaluate all free-flowing streams. The resulting dataset is particularly qualified as a basis for this study. In 1996, Pennsylvania's Department of Environmental Protection (DEP) developed a strategy for these assessments that involves preliminary screening of each of the state's 104 identified watersheds (referred to as State Water Plan Watersheds), followed by a field-level biological assessment. Full-scale fieldwork for the unassessed waters project began in 1997. This is a cooperative effort, with assessments being conducted by staff biologists from the Department's six Field Offices and the Bureau of Water Supply and Wastewater Management, the Susquehanna River Basin Commission, the Interstate Commission on the Potomac River Basin, the Pennsylvania Fish & Boat Commission, and the Erie County Department of Health. After five complete assessment seasons ending September of 2001, a total of 53 State Water Plan (SWP) Watersheds have been completed and assessments in 28 SWP Watersheds are either nearly completed or still in progress. In addition to those SWP Watersheds still in progress, 14 have been scheduled for the 2002 season. The entire coverage is expected to be completed during the 2003 field season.

The Surface Water Assessment Program uses a biological screening protocol to determine impairment of aquatic life uses. Biological screening is conducted on wadeable waters using a modification of EPA's Rapid Bioassessment Protocol (RBP II), which includes field identification of benthic macroinvertebrates to the family level and an RBP habitat assessment. The field biologist begins the assessment process by conducting a reconnaissance of local watershed units to verify current land-use patterns, confirm known point source discharge locations, categorize stream habitat types, flow conditions, accessibility, and other conditions that would determine where sampling stations are placed. In-stream assessments begin after the reconnaissance is completed. Both the number and the location of assessment points is at the discretion of the field biologist. Beginning at the top of the watershed, i.e., headwater, first order streams, and moving downstream, the field biologist identifies locations at which a transitioning of water quality is expected to occur, and above which water quality is homogeneous. For example, a typical situation for Pennsylvania streams is forested, first order and second order streams, gradually converging to form a third order stream which is located on a valley floor occupied by agriculture. In this instance, absence of point sources and consistency of land use may result in the first sampling location being chosen to characterize all forested, low order streams, and the second sampling location characterizing the transition to agricultural land use

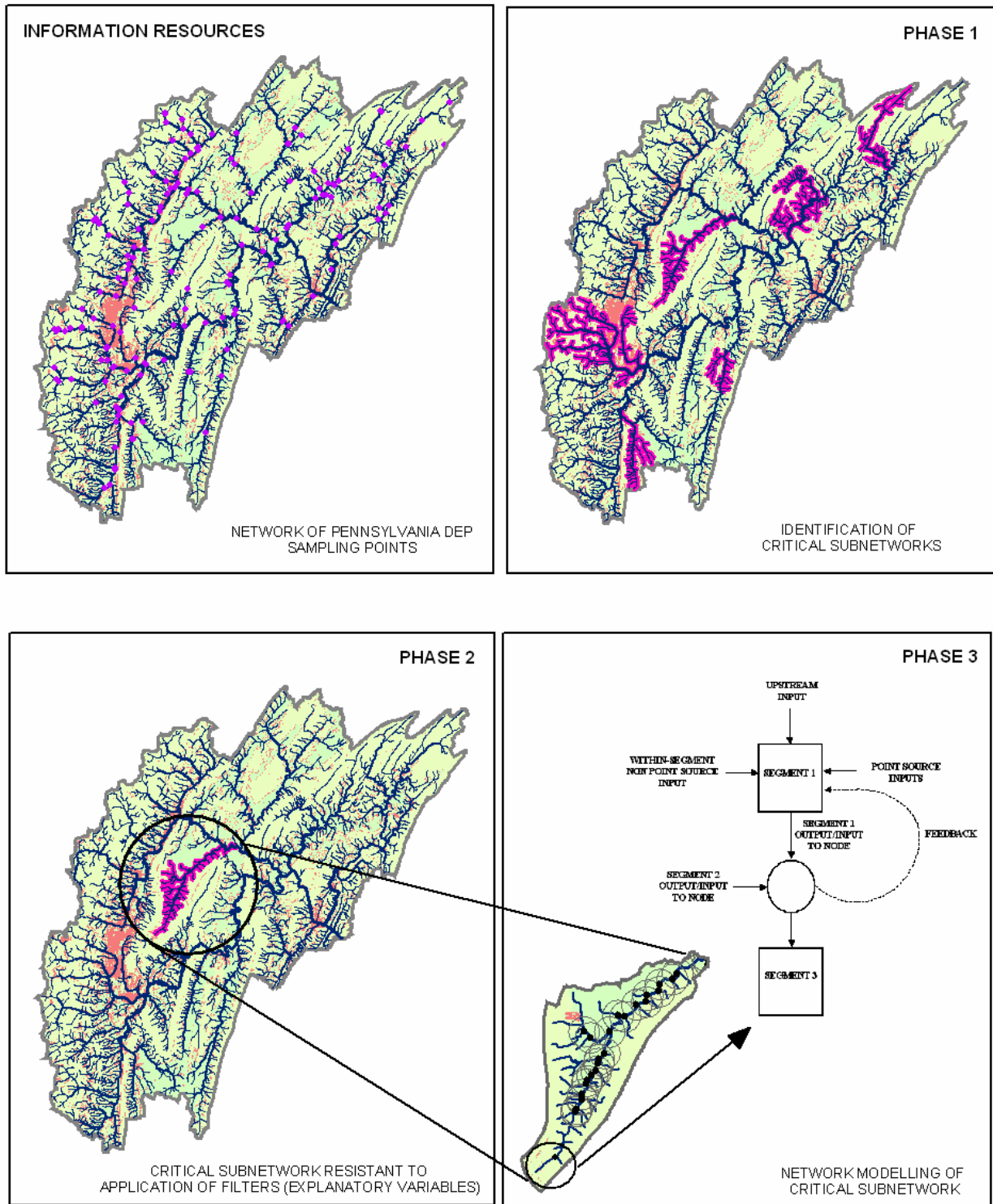


Figure 1. Upper Juniata river network with superimposed network of Pennsylvania Department of Environmental Protection (DEP) sampling stations. In Phase 1, scan statistics methods will be used to identify hotspots of biological impairment. Phase 2 incorporates potential explanatory factors and identifies residual hotspots that are subjected to detailed network modeling in Phase 3.

along the third order stream. As the biologist moves downstream, and land use transitions and point sources occur with higher frequency, the distance between sampling locations decreases, and the stream is characterized by many smaller segments. Thus, this data set incorporates professional field judgement to assure that all stream/land use combinations are characterized, and point sources are clearly identified and their impact assessed. This has enormous advantages over a randomly stratified design when the ultimate goal of the study is not the general assessment of condition of the resource, but the prediction of the relationships between natural and anthropogenic factors and the biological conditions within a watershed.

The biological water quality data, stream assessment maps, and geographic information are electronically compiled and stored in DEP databases. Geographic information includes biological sample and stream reach locations, ecoregions, municipalities, counties, and watershed boundaries. Assessment data include items such as habitat and biological results, source and causes of pollution, and investigator comments and observations. Preliminary inquiries into the availability of all portions of the database have indicated that database access would be complete.

Phase 1. Identification of Critical Sub-networks of the Stream-based Network

Critical sub-network identification will be based upon biological assessment data available for the Commonwealth of Pennsylvania. Hot spots, or critical sub-networks, are initially identified as sub-networks on the entire stream/sample point network. It is impossible to know a priori what the average spatial scale of a critical sub-network would be, although current land use patterns provide clues. Land use patterns are generally consistent at the reach scale for streams in a number of physiographic provinces, implying that a critical sub-network may occur in the spatial range generally referred to as a small watershed. These areas are typically sized as tens to hundreds of km², and encompass several stream or river reaches, adjacent riparian corridors, associated wetlands and waterbodies, and the contributing drainage basin. There are a number of reasons why areas equivalent in size to a small watershed are desirable as the focal units for critical sub-network identification:

- Units of this size are central to the entire dimensional range of ecological information currently collected on freshwater aquatic resources, spanning several orders of magnitude. The majority of environmental data are collected at smaller scales, such as points, plots, reaches, and sites. These data can be efficiently aggregated up to the scale of a small watershed without significant loss of resolution.
- Assessment and restoration activities for aquatic ecosystems can be effectively targeted at and reported for units of this scale.
- A nested or hierarchical approach to data collection and assessment provides opportunities to address larger emergent properties and regional issues at landscape and ecoregion scales, but still allows investigators to trace the origins of those data.

The product of Phase 1 is a first generation surface mapping of critical sub-networks of low biological condition, which is carried to Phase 2 for the revision by application of explanatory variables, or filters. Identification of critical sub-networks will utilize an adaptation of the scan statistic methodology (see below for details).

Phase 2. Compilation, Analysis, and Application of Relevant Filters

A number of parameters will be compiled for use as explanatory variables (filters) to be applied to the determination of critical sub-networks. In this step, each explanatory variable is applied to the scan statistic procedure. The overall explanatory power of each variable will be assessed utilizing the disappearance rate of critical sub-networks. For example, if the explanatory variable of drainage pattern type is selected, the relevant drainage pattern category of each data point is utilized to adjust the expected response rate in the scan statistic protocol. If the variable is useful as an explanatory one, the majority of critical sub-networks will “disappear” in the surficial map. Any critical sub-network that remains persistent, i.e., unexplained, after application of all the available parameters will be assessed via network modeling in Phase 3.

We will utilize three categories of parameters: descriptive, structural, and network related. Descriptive parameters are watershed-wide attributes previously documented in the literature, including watershed-wide land use characterization, riparian buffer characterization, and wetland condition (where available). Additional parameters will be constructed from current literature and utilized in a raw format or summarized through statistical techniques.

Structural parameters are primarily hierarchical, geographically dependent classifications such as drainage pattern type, watershed morphometry, physiographic province, and ecoregions. A subset of existing watershed classification systems, such as the classification system utilized to represent a range of social choices in the Atlantic Slope Project, will also be assessed. These parameters are all attributable to the stream network in the study region. In addition to these currently available classifications (Wardrop et al., 2002), we propose to examine the relationship between existing drainage pattern types described in the literature, and a classification of types based upon parameters of the stream-based network. The “drainage pattern” or structural configuration of the channel network is strongly indicative of underlying geomorphic control on the hydrologic system. The penetration, convergence, angularity and sinuosity are all reflective of geologic substrate and its fracturing, gradient, soil porosity and erodability, and precipitative energy absorption by vegetative cover. These influence flushing, transport, chemistry, turbidity, and temperature regimes of the aquatic environment. Drainage pattern has traditionally been classified subjectively by comparison with idealized diagrams. For hydrologic modeling and quantitative analysis purposes, it is important to capture this integrative expression in an objective manner. For example, dendritic, parallel, and trellis drainage patterns could be quantitatively described (and thus determined) by network parameters such as nodal distribution, nodal density, and percent of nodes in the upper third of the watershed. GIS capabilities for distance, buffering, intersection, routing and flow analysis provide systematic methods for extracting composite properties of channel networks as a basis for discriminating representative types. Spacing, perimeter, composition and overlap relations among points and circles representative of confluence positions (nodes) in channel networks can be used to parameterize the patterns for statistical discrimination. Sub-networks can then be quantitatively compared for characterization of homogeneity and heterogeneity in conformation.

Additional stream-based network parameters will be identified as part of the proposed research. As a general rule, relevant parameters utilized in the construction of the network model (Phase 3) would be summarized for use as explanatory variables (filters). Potential parameters of this type

are nodal density, differences in within and between reach nodal overlap, summary statistics of nodal land use, internodal segment land use and other characteristics, and gravity measures between nodes. Vector summaries of nodal land use (constellations) and nodal land use history trajectories will also be investigated.

Phase 3. Development of Process Models of Stream Systems as Nodal Networks

Having established the capability through statistical scanning of indicators to identify specific portions of stream systems as being biologically degraded, we propose to develop a network-based process modeling framework by which to determine the nature of potential for recovery within those hydrologic subsystems. The proposed approach uses series of nodes at stream confluences and internodal segments as functional frames for parameterization and prediction.

This approach arises from a number of studies relating landscape properties in a bounded area to biological information at a point. Bounded areas have been contributing watersheds, riparian buffers, riparian buffers within a circle, etc. We intend to combine this concept with systems analysis to represent the stream structure as a circuit comprised of nodes, internodal segments, and associated catchments. In this manner, a typology of stream drainages emerges that matches the information currently available. For example, stream 'layers' for Geographic Information Systems (GIS) allow analysis of stream reach lengths, locations of stream convergences, gradients, land use associated with riparian zones, land use intensity at stream convergences, etc. In addition, biological assessment information is associated with a designated stream reach, reflecting the cumulative effect of all upstream reaches and convergences. We strive particularly to increase the explanatory power of the relationship between landscape metrics and ecological condition of streams (Forman and Godron, 1986).

Nodes of confluence function primarily as relays of water movement rather than as destinations, but these locations of mixing also provide special environments for biota. A nodal network representation of the aquatic resource provides amenability to formal network analysis. Nodes are a direct reflection of the complexity of the drainage network, and will increase in density along with drainage complexity as often occurs in a downstream progression. The nodal approach thus allows for differential weighting of landscape parameters in more complex components of the hydrologic system.

AVGWLF is a recently developed assessment tool for predicting nutrient and sediment loadings in Pennsylvania watersheds that utilizes GIS for automation of the Generalized Watershed Loading Function (Evans, 2002; Evans, Sheeder and Corradini, 2001; Evans et al., in review). AVGWLF features intelligent and automated integration of model components with regard to interfaces and transfer of information, and accounts for both nonpoint and point sources. It will serve the purpose of predictive parameterization for the chemical and sediment environment for biota. Landscape metrics and demographic information will further contribute predictive parameters for biological influence modeling.

The network modeling structure will entail functional, statistical, and adaptive linkages between frames. Frame-based inheritance is potentially advantageous for effecting downstream

propagation of upstream influences. Cascading increments will enable internal determination of sensitivity, thus indicating potential effectiveness of prospective remediation.

Scan Statistic Methodology

Three central problems arise in geographical surveillance for a spatially distributed response variable. These are (i) identification of areas having exceptionally high (or low) response, (ii) determination of whether the elevated response can be attributed to chance variation (false alarm) or is statistically significant, and (iii) evaluation of explanatory factors that may account for the elevated response. Although a wide variety of methods have been proposed for modeling and analyzing spatial data (Cressie, 1991), the spatial scan statistic (Kulldorff and Nagarwalla, 1995; Kulldorff, 1997) has quickly become a popular method for detection and evaluation of disease clusters, and is now widely used by many health departments and academic scientists. With suitable modifications, the scan statistic approach can be used for hotspot analysis in fields other than the health sciences (Myers et al., 2002; Patil et al., 2003). We propose to develop methodology and corresponding software for application of the scan statistic to analysis of biological impairment along river networks. We first review the spatial scan statistic as it has been used for disease surveillance and then we indicate the modifications that are needed to carry out Phases 1 and 2 of this prospectus.

Spatial Scan Statistic Background

The spatial scan statistic deals with the following situation. A region R of Euclidian space is tessellated or subdivided into cells that will be labeled by the symbol a . Data is available in the form of a count Y_a (non-negative integer) on each cell a . In addition, a “size” value A_a is associated with each cell a . The cell sizes A_a are regarded as known and fixed, while the cell counts Y_a are random variables. In the disease setting, the response Y_a is the number of diseased individuals within the cell and the size A_a is the total number of individuals in the cell. Generally, however, the size variable is adjusted for factors such as age, gender, environmental exposures, etc., that might affect incidence of the disease. The disease rate or relative risk within the cell is the ratio

$$\text{Rate} = Y_a / A_a .$$

The spatial scan statistic seeks to identify “hotspots” or clusters of cells that have an elevated rate compared with the rest of the region, and to evaluate the statistical significance (p -value) of each identified hotspot. These goals are accomplished by setting up a formal hypothesis-testing model for a hotspot. The null hypothesis asserts that there is no hotspot, i.e., that all cells have (statistically) the same rate. The alternative states that there is a cluster Z such that the rate for cells in Z is higher than for cells outside Z . An essential point is that the cluster Z is an unknown parameter that has to be estimated. Likelihood methods are employed for both the estimation and significance testing. Candidate clusters Z are referred to as **zones**. Ideally, maximization of the likelihood should search across all possible zones, but their number is generally too large for practical implementation. Various devices (e.g., expanding circles) are employed to reduce the list of candidate zones to manageable proportions.

Explication of the likelihood function requires a distributional model (response distribution) for the response Y_a in cell a . This distribution is allowed to vary from cell to cell but in a manner that is regulated by the size variable A_a . Thus, A_a enters into the parametric structure of the response distribution and in such a manner that the mean response is proportional to A_a while the relative variability, $\text{var}(Y_a / A_a)$, decreases with increasing A_a . In disease surveillance, response distributions are generally taken as either binomial or Poisson, leading to comparatively simple likelihood functions. The scan statistic needed for the present prospectus requires continuous response distributions and comparatively complex likelihood functions.

Significance testing for the spatial scan statistic employs the likelihood ratio test. In “standard” settings a chi-squared distribution with appropriate degrees of freedom can be used as reference or null distribution. Unfortunately, the spatial scan statistic is not a standard setting---in part because the zonal parameter Z is discrete. Accordingly, Monte Carlo simulation (Dwass, 1957) is used to determine the needed null distributions.

Specific Methodological Tasks for Phases 1 and 2

We will be searching for hotspots of biological impairment along the river systems of Pennsylvania. Impairment will be measured by a complemented form of the index of biological integrity determined at each sampling station of the Pennsylvania DEP sampling network (Figure 1). Although we will use scan statistic methods, the basic carrier of information is not the cell of a tessellated region but is the sampling station within a network of sampling stations. Therefore, in our approach to the scan statistic, the geometric structure that carries the numerical information is an **abstract graph** consisting of objects a (sampling stations) and edges joining certain pairs of distinct objects. These edges are determined by direct flow of water from one station to the next. The sampling stations (about 12,000 currently; 15,000 eventually) become the **vertices** of the graph while the stream network determines the edges. In actuality, each sampling station corresponds to and represents a subjectively homogeneous stream sub-network; the detailed structure of each sub-network will be analyzed in Phase 3 of the prospectus while Phases 1 and 2 treat each sub-network as an aggregate entity.

Each object a in the graph carries three pieces of information: (i) a **size variable** that is treated as known and non-random, (ii) a **response variable** whose recorded value is regarded as a realization of some probability distribution, and (iii) the probability distribution itself, which is called the **response distribution**. Parameters of the response distribution may vary from object to object, but the mean response should be proportional to the value of the size variable for that object. On the other hand, relative response variability (e.g., coefficient of variation) should decrease as the size variable increases. The **response rate** is the ratio

$$\text{Response Rate} = \frac{\text{Response}}{\text{Size}},$$

and a **hotspot** is a collection of objects (i.e., a zone) whose response rates are unusually large.

Methodological Task 1: To develop a new version of the scan statistic designed for detection of hotspots on a network. Candidate zones will be all connected components of upper level sets of the response rate. The method is adaptive since the list of candidate zones is determined by the actual data rather than by some *a priori* prescription like expanding circles or ellipses. Adaptivity entails data-dependence and this dependence must be accounted for in the null distributions used for significance testing.

Methodological Task 2: The scan statistic methodology will be extended to include continuous response distributions. We will focus on three parametric families of distributions: gamma distribution, lognormal distribution, and scaled beta distribution. The first two families apply to responses that can range from zero to infinity, while the third is for bounded responses. The overall approach is to model the mean and relative variance of each response distribution in terms of the size variable. The mean and variance are functions of the parameters of the response distribution, so that the likelihood function can be written down and parameters estimated by maximum likelihood. In Phase 1 of the proposed work, size variables will all be taken as unity.

Methodological Task 3: Hotspots are always relative to the size variable and one assesses the impact of an explanatory variable by using the explanatory variable to adjust the size variable (which is proportional to expected object response). The hotspot analysis is then redone with the adjusted sizes, and before-adjustment and after-adjustment hotspot maps are compared. Methods for making these adjustments need to be developed for application in Phase 2.

Network Scan Statistic (Task 1)

The key element here is determination of the list of candidate zones Z . A zone is, first of all, a collection of vertices from the sampling network. Secondly, those vertices should be connected in the network sense because a geographically scattered collection of vertices would not be a reasonable candidate for a “hotspot.” Even with this connectedness limitation, the number of candidate zones is too large for a maximum likelihood search in all but the smallest of networks. We propose to reduce the list of zones to searchable size in the following way. The response rate at vertex (sampling station) a is given by $G_a = Y_a / A_a$. These rates determine a function $a \rightarrow G_a$ defined over the vertices in the graph. This function has only finitely many values (called levels) and each level g determines an **upper level set** defined by

$$U_g = \{a : G_a \geq g\}.$$

Upper level sets do not have to be connected but each upper level set can be decomposed into the disjoint union of connected components. The list of candidate zones Z for the network scan statistic consists of all connected components of all upper level sets. Since they are portions of upper level sets, these zones are certainly plausible as potential hotspots. Their number is small enough for practical maximum likelihood search and they comprise a tree under set inclusion, thus facilitating computer representation.

Finding the connected components for an upper level set is essentially the issue of determining the transitive closure of the adjacency relation defined by the edges of the graph. Several generic algorithms are available in the computer science literature (Cormen et al, 2001, Section 22.3 for depth first search; Knuth, 1973, p. 353 or Press et al, 1992, Section 8.6 for transitive closure).

Continuous Response Distributions (Task 2)

In our general scan statistic setup, there is a collection of objects which are indexed by the symbol a . These objects are the vertices in an abstract graph that, in concrete situations might be the cells of a tessellation or the nodes a stream network. Attached to each object a is a response value Y_a and a size value A_a . Our strategy for handling continuous responses is to model the mean and variance of each response distribution in term of the size variable using as guiding principle that the mean response for object a should be proportional to A_a and the relative variability should decrease with A_a . This is best illustrated in the context of the gamma family of distributions.

Gamma Distribution. We parameterize the gamma distribution by (k, β) where k is the index parameter and β is the scale parameter. Thus, if Y is a gamma-distributed variate

$$E[Y] = k\beta \quad \text{and} \quad \text{Var}[Y] = k\beta^2.$$

Both k and β can vary from object to object but additivity with respect to the index parameter suggests that we take k proportional to the size variable:

$$k_a = A_a / c$$

where c is an unknown parameter but whose value is the same for all objects a . This gives the following mean and squared coefficient of variation for the response on object a :

$$E[Y_a] = \beta_a A_a / c \quad \text{and} \quad \text{CV}^2[Y_a] = c / A_a.$$

The hotspot testing model is analogous to that used in disease surveillance:

H_0 : β_a is the same for all objects a , i.e., there is no hotspot.

H_1 : There is a non-empty zone Z and parameter values $0 < \beta_0, \beta_1$ such that

$$\beta_a = \begin{cases} \beta_1 & \text{for all objects } a \text{ in } Z \\ \beta_0 & \text{for all objects } a \text{ outside of } Z \end{cases} \quad \text{and} \quad \beta_1 > \beta_0.$$

Notice that the null model implies that the expected rate, $E[Y_a / A_a] = \beta_a / c$, is the same for all a . The full model has four unknown parameters Z, c, β_0, β_1 that need to be estimated. The profile likelihood function for Z is obtained by fixing an arbitrary candidate zone Z and maximizing the likelihood with respect to the other three parameters. The latter optimization problem reduces to the solution of three likelihood equations. Two of these equations can be solved in closed form while the third requires Newton-Raphson iteration. The maximum likelihood estimate for the hotspot zone Z is obtained by maximizing the profile likelihood by direct search over the entire list of candidate zones.

Lognormal and Other Continuous Distributions. A similar approach is applicable to other two-parameter families of distributions on the positive real line. Specifically, for the lognormal distribution, we take

$$E[Y_a] = \beta_a A_a / c \quad \text{and} \quad \text{CV}^2[Y_a] = [c / A_a]^d,$$

where d is either user-specified (e.g., $d = 1$) or is an unknown parameter to be estimated. The conventional lognormal parameters (μ_a, σ_a^2) can then be expressed in terms of the model parameters (β_a, c, d) , allowing the likelihood function to be written down explicitly.

Simulating the Null Distribution to Obtain p -Values. Conditional simulation is used to obtain the null distribution in the cases of the binomial and Poisson response distributions (Kulldorff, 1997). One conditions on the sufficient statistic (under H_0) to eliminate the unknown parameters from the null model. The resulting parameter-free distributions are hypergeometric and multinomial, respectively, and are easily simulated. This parameter-removal by conditioning is not available for most continuous distributions. Accordingly, simulation will be done by replacing unknown parameters with their maximum likelihood estimates under H_0 .

Filtering for Explanatory Variables (Task 3)

The scan statistic searches for regions of high response relative to a geo-referenced set of prior expected responses. Thus, a hotspot map depicts regions of extreme departure from expectation in the multiplicative sense, i.e., multiplicative residuals. The size values A_a , which are proportional to model expectations, are the link between the response variable and potential explanatory variables. In disease surveillance, the A_a are routinely adjusted for factors like age, gender and population size before beginning the analysis (Bithell et al., 1995; Kulldorff et al., 1997; Rogerson, 2001; Waller, 2002; Walsh and Fenster, 1997; Walsh and DeChello, 2001). Such standard, agreed-upon, factors are unavailable in the environmental sciences and the initial Phase 1 analysis will identify absolute impairment highs by setting all A_a equal to unity. Locations of these highs may provide clues in identifying potential explanatory factors. Next, in Phase 2, the size values are adjusted for these factors and the scan statistic is rerun with the adjusted sizes. Comparative configuration of new and old hotspots reveals the impact of these factors on the response under study.

There are several methods available for adjusting the A_a . Suppose, first, that there is only one explanatory variable X . A nonparametric approach partitions the X -values into intervals and calculates the mean response for each interval. These calculations should utilize all available pertinent data. For example, data from the entire Pennsylvanian Ridge and Valley physiographic province might be used in calculating adjustment factors for the Juniata river basin. The adjusted size value for object a becomes

$$A'_a = \frac{m_a}{m} A_a,$$

where A_a is the old size value, m_a is the mean response for the interval containing object a , and m is an overall mean response. Regression of Y upon X can also be the basis for adjustment provided an appropriate functional relation could be identified. The same approaches work, in principle, for multiple factors. However, the curse of dimensionality comes into play and data sparseness prevents calculation of dependable local means. Our approach, in such cases, will be to cluster the data points in factor space. A mean response is then calculated for each cluster.

Network Modeling Methods for Phase 3

The network modeling itself will be an experimental undertaking, since existing models can generally be considered as either area-based routing or flow-based models (AWRA, 2001). The proposed representation does not conform to either of these conventional types. The primary purpose of our network model is to account for relative change in biological integrity as a function of chemical, physical, hydrologic, and habitat variables with special emphasis on anthropogenic factors. We will first develop a modeling testbed in a database that will be configured to support informal object-oriented representations as frames operated upon by modular prototype software written in Visual C and/or Visual Basic. Informational, functional, and logical components will all be incorporated symbolically in the database for transparency in studying performance of provisional models and easy alteration in successively adapting system representation to findings. When provisional models have been suitably validated, the prototype representation will be transformed into a suitable software environment that is formally object-oriented and configured for both operability and run-time performance.

Project Outcomes and Deliverables

Year 1

- 1) Obtain and recompile biological and habitat data for stream systems of Pennsylvania as collected by the Dept. of Environmental Protection for the Surface Water Assessment Program.
- 2) Develop first-tier (broad-scale) representation of Pennsylvania stream systems as a formal network emphasizing nodes as objects for attribution.
- 3) Reformulate scan statistic methodology for a network domain instead of a spatial domain.

Year 2

- 1) First-level network scan for regional subsystems exhibiting severe biological degradation.
- 2) Study network representations of streams at different scales with a view to formulating systematic and objective approaches for classifying drainages and segregating hydrologic environments to enable scanning for hotspots with hotspots.
- 3) Compile landscape information for nodal characterization from GIS and remote imagery.
- 4) Likelihood modeling for continuous responses and efficient optimization algorithms.

Year 3

- 1) Conduct nested scans for successively more detailed determinations of criticality using habitat and landscape filter variables.
- 2) Select initial subsystems for hydrologic process modeling.
- 3) Assemble software and databases to support hydrologic process modeling.

Year 4

- 1) Conduct and test hydrologic process modeling of critical network components.
- 2) Disseminate results of scanning protocols, and package scanning technology for distribution.
- 3) Journal publications, monograph preparation, and workshops.

Project Work Groups

Phase 1: GP, DW, CT, JB, GSS-1

Phase 2: GP, DW, CT, KH, GSS-1

Phase 3: GP, WM, DW, JB, KH, GSS-2

Methods & Tools: GP, CT, GSS-1, GSS-2

GP: G. P. Patil; DW: D. Wardrop, CT: C. Taillie; JB: J. Bishop; KH: K. Hychka;
 WM: W. Myers; GSS: Graduate Statistics Student

Integration of Research, Education, and Dissemination

In graduate education, we will integrate project methods and approaches into the related graduate courses offered at the University. We believe that students in both disciplinary and interdisciplinary courses will find the project methodology of great value to their professional preparation, and expect significant student experimentation with the innovative issues, approaches, and software. Graduate students supported on this project will present their research goals and progress during summer workshops. The project is fortunate to have two competent graduate research assistants with extensive expertise and involvement in both GIS and water resource systems. Statistics graduate research students for the project will be selected on the basis of a well-publicized mini-competition.

To enhance synergistics, three weekly group activities will be organized: (a) Monday methodological cross-disciplinary miniseminar; (b) Wednesday methodological disciplinary miniseminar; and (c) Friday young professional miniseminar. These will help strengthen both vertical and horizontal integration. For effective technology transfer, the following are planned: one monograph, one case book, two thematic journal issues, summer research workshops, journal publications, and distributed information management. These products and outcomes will help evaluate the success of the educational activities of the project. Four additions to the project will be requested for interactive undergraduate dimension of the project.

Pertinent Background and Support

A keystone motivation and foundation for the current prospectus comes from a recently submitted final report on the NSF research grant DEB-9524722 for the NSF Water and Watersheds Program on “Multiscale Statistical Approach to Critical-Area Analysis and Modeling of Watersheds and Landscapes,” for October 1995 to March 1999. Details are on the website <http://www.stat.psu.edu/~gpp> together with thirty-five related publications, describing the methods, tools, applications, and outcomes (website: <http://www.stat.psu.edu/~gpp/newpage11.htm>). The grant project has been reported as a success story by the responsible NSF program officer in her report to the NSF Director.

The American Water Resources Association has this year announced its William Boggess Best Paper Award to this NSF project based paper, “Predictability of Surface Water Pollution Loading in Pennsylvania Using Watershed-Based Landscape Measurements.” The project assistant, a co-author with the PI and the Co-PI, did his MS in Environmental Statistics and Ph.D in Statistical Ecology with the PI. The award paper (website: <http://www.stat.psu.edu/~gpp/PDFfiles/TR99-0303.pdf>) uses parts of the multiscale advanced raster map analysis system in progress (website: <http://www.stat.psu.edu/~gpp/newpage11.htm>).

The NSF project prompted an Indo-US Workshop on Regional Policy Research Using Remote Imagery and Multiscale Analysis under the joint auspices of the Geography Panel of the University Grants Commission, the Center for Policy Research (CPR), and

the Forest Survey of India with the PI and the CPR President as Joint Conveners. The NSF project work was also instrumental for the appointment of PI as Visiting Research Professor of Regional Policy during 1998-1999 at the Center for Policy Research, New Delhi, a major think tank of India.

References

- Allan JD, Erickson DL, and Fay J (1997). The influence of catchment land use on stream integrity across multiple spatial scales. *Freshwater Biology*, **37**, 149-161.
- AWRA (2001). Water quality monitoring and modeling. *Proceedings. American Water Resources Association*. 284 pp.
- Bithell JF, Dutton SJ, Neary NM, and Vincent TJ (1995). Controlling for socioeconomic confounding using regression methods. *Community Health*, **49**, S15-S19.
- Brinson MM (1993). A hydrogeomorphic classification for wetlands, pp. 1-18. U.S. Army Engineer Waterways Experiment Station, Vicksburg, MS.
- Bryce SA, and Clarke SE (1996). Landscape-level ecological regions: linking state-level ecoregion frameworks with stream habitat classifications. *Environmental Management*, **20**, 297-311.
- Cormen TH, Leieron CE, Rivest RL, Stein C (2001). *Introduction to Algorithms*, Second Edition. MIT Press, Cambridge, Massachusetts.
- Cressie N, (1991). *Statistics for Spatial Data*. Wiley, New York.
- Dale VH (2002). A Systems Approach to Environmental Security. Plenary Address, June 10, 2002, Healthy Ecosystems, Healthy People: Linkages between biodiversity, ecosystem health, and human health. Presented by the International Society for Ecosystem Health, June 6-11, 2002, Washington, DC.
- Detenbeck NE, Batterman SL, Brady VJ, Brazner JC, Snarski VM, Taylor DL, and Thompson JA (2000). A test of watershed classification systems for ecological risk assessment. *Environmental Toxicology and Chemistry*, **19**, 1174-1181.
- Dwass M (1957). Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics*, **28**, 181-187.
- EPA, US (2000). Mid-Atlantic Highlands Stream Assessment. U.S. Environmental Protection Agency, Environmental Monitoring and Assessment Program National Health and Environmental Effects and Region III, Washington, D.C.
- Evans BM (2002). Development of an Automated GIS-Based Modeling Approach to Support Regional Watershed Assessments. Ph.D. Dissertation in Soil Science, The Pennsylvania State University, 231 p.
- Evans BM, Sheeder SA, and Corradini KJ (2001). AVGWLF, Version 4.04: User's Guide. Environmental Resources Research Institute, Penn State University, 73 p.
- Evans BM, Lehning DW, Corradini KJ, Petersen GW, Nizeyimana E, Hamlett JM, Robillard PD,

- and R. L. Day RL (in review). A comprehensive GIS-based modeling approach for predicting nutrient loads in watersheds. *Journal of Spatial Hydrology*.
- Forman RTT, and Godron M (1986). *Landscape Ecology*. John Wiley & Sons, New York.
- Gunderson LH (2000). Resilience in theory and practice. *Annual Review of Ecology and Systematics*, **31**, 425-439.
- Harding JS, Benfiels EF, Bolstad PV, Helfman GS, and Jones EBD (1998). Stream biodiversity: the ghost of land use past. *Proceedings of the National Academy of Science*, **95**, 14843-14847.
- Hawkins CP, Norris RH, Gerritsen J, Hughes RM, Jackson SK, Johnson RK, and Stevenson RJ (2000). Evaluation of the Use of Landscape Classifications for the Prediction of Freshwater Biota: Synthesis and Recommendations. *Journal of North American Benthological Society*, **19**, 541-556.
- Imhof JG, Fitzgibbon J, and Annable WK (1996). A hierarchical evaluation system for characterizing watershed ecosystems for fish habitat. *Canadian Journal of Fisheries and Aquatic Sciences*, **53**, 312-326.
- Johnson LB, Richards C, Host GE, and Arthur JW (1997). Landscape influences on water chemistry in Midwestern stream ecosystems. *Freshwater Biology*, **37**, 193-208.
- Jones KB, Ritters KH, Wickham JD, Tankersley RD, O'Neill RV, Chaloud DJ, Smith ER, and Neale AC (1997). An ecological assessment of the United States Mid-Atlantic region: a landscape atlas. U.S. Environmental Protection Agency, Las Vegas, NV.
- Karr JR (1981). Assessment of biotic integrity using fish communities. *Fisheries*, **6**, 21-26.
- Karr JR, and Chu EW (1999). *Restoring Life in Running Waters: Better Biological Monitoring*. Island Press, Washington, DC.
- Knuth DE (1973). *The Art of Computer Programming: Volume 1, Fundamental Algorithms*, Second Edition. Addison-Wesley, Reading, Massachusetts.
- Kulldorff M (1997). A spatial scan statistic. *Communications in Statistics: Theory and Methods*, **26**, 1481-1496.
- Kulldorff M, Feuer E, Miller B, Freedman L (1997). Breast cancer in northeast United States: a geographic analysis. *American Journal of Epidemiology*, **146**, 161-170.
- Kulldorff M, Nagarwalla N (1995). Spatial disease clusters: Detection and inference. *Statistics in Medicine*, **14**, 799-810.
- Lammert M, and Allan JD (1999). Assessing biotic integrity of streams: effects of scale in measuring the influence of land use/cover and habitat structure on fish and macroinvertebrates. *Environmental Management*, **23**, 257-270.

- Lammert M, Higgins J, Grossman D, and Bryer M (1997). A classification framework for freshwater communities: proceedings of the Nature Conservancy's Aquatic Community Classification Workshop, pp. 1-16. The Nature Conservancy, New Haven, Missouri.
- Leopold LB, Wolman MG, and Miller JP (1964). *Fluvial Processes in Geomorphology*. WH Freeman, San Francisco, California.
- Lotspeich FB (1980). Watersheds as the basic ecosystem: this conceptual framework provides a basis for a natural classification system. *Water Resources Bulletin*, **16**, 581-586.
- Maxwell JR, Edwards CJ, Jensen ME, Paustian SJ, Parrott H, and Hill DM (1995). A hierarchical framework of aquatic ecological units in North America (nearctic zone), pp. 1-76. U.S. Department of Agriculture, Forest Service, Washington, D.C.
- Minshall GW, Peterson RC, Cummins KW, Bott TL, Sedell JR, Cushing CE, and Vannote RL (1983). Interbiome comparison of stream ecosystem dynamics. *Ecological Monographs*, **53**, 1-25.
- Myers WL, Kurihara K, Patil GP, Vraney R (2002). Finding upper level sets in cellular surface data using echelons and SaTScan. Invited paper, 2002 Joint Statistical Meetings, New York.
- Omernik JM (1987). Ecoregions of the conterminous United States. *Annals of the Association of American Geographers*, **77**, 118-125.
- Pan Y, Stevenson RJ, Hill BH, Kaufmann PR, and Herlihy AT (1999). Spatial patterns and ecological determinants of benthic algal assemblages in Mid-Atlantic streams, USA. *Journal of Phycology*, **35**, 460-468.
- Patil GP, Bishop J, Myers WL, Taillie C, Vraney R, and Wardrop DH (2003). Detection and delineation of critical areas using echelons and spatial scan statistics with synoptic cellular data. *Environmental and Ecological Statistics* (to appear).
- Paul JF et al. (1998). Condition of the Mid-Atlantic Estuaries. USEPA, Office of Research and Development, Washington, DC.
- Poff NL, and Ward JV (1990). Physical habitat template of lotic systems: recovery in the context of historical pattern of spatiotemporal heterogeneity. *Environmental Management*, **14**, 629-645.
- Press WH, Teukolsky SA, Vetterling WT, and Flannery BP (1992). *Numerical Recipes in C*, Second Edition. Cambridge University Press, Cambridge.
- Reeves GH, Benda LE, Burnett KM, Bisson PA, and Sedell JR (1995). A disturbance-based ecosystem approach to maintaining and restoring freshwater habitats of evolutionarily significant units of anadromous salmonids in the Pacific Northwest. *American Fisheries Society Symposium*, **17**, 334-349.

- Richards C, Haro RJ, Johnson LB, and Host GE (1997). Catchment and reach-scale properties as indicators of macroinvertebrate species traits. *Freshwater Biology*, **37**, 219-230.
- Richards C, Johnson LB, and Host GE (1996). Landscape-scale influences on stream habitats and biota. *Canadian Journal of Fisheries and Aquatic Sciences*, **53**, 295-311.
- Richards RP (1990). Measures of flow variability and a new flow-based classification of Great Lakes tributaries. *Journal of Great Lakes Research*, **16**, 53-70.
- Rogerson PA (2001). Monitoring point patterns for the development of space-time clusters. *Journal of the Royal Statistical Society, Series A*, **164**, 87-96.
- Rosgen D (1996). *Applied River Morphology*. Printed Media Companies, Pagosa Springs, Colorado.
- Roth NE, Allan JD, and Erickson DL (1996). Landscape influences on stream biotic integrity assessed at multiple spatial scales. *Landscape Ecology*, **11**, 141-156.
- Seelbach PW, Wiley MJ, Kotanchik JC, and Baker ME (1997). A landscape-based ecological classification system for river valley segments in lower Michigan (MI-VSEC Version 1). State of Michigan, Department of Natural Resources.
- Waller L (2002). Methods for detecting disease clustering in time or space. In *Statistical Methods and Principles in Public Health Surveillance*. R. Brookmeyer and D. Stroup, eds. Oxford University (to appear).
- Walsh SJ, and DeChello LM (2001). Geographical variation in mortality from systemic lupus erythematosus in the United States. *Lupus*, **10**, 637-646.
- Walsh SJ, and Fenster JR (1997). Geographical clustering of mortality from systemic sclerosis in the Southeastern United States, 1981-90. *The Journal of Rheumatology*, **24(12)**, 2348-2352.
- Wardrop DH, Bishop JA, Easterling M, Hychka K, Myers WL, Patil GP, and Taillie C (2002). Use of landscape and land use parameters for classification and characterization of watersheds in the Mid-Atlantic across five physiographic provinces. Healthy Ecosystems, Healthy People Conference, International Society for Ecosystem Health, Washington, DC (Invited paper).
- Whiting PJ, and Bradley JB (1993). A process-based classification system for headwater streams. *Earth Surface Processes and Landforms*, **18**, 603-612.
- Wiley MJ, Kohler SL, and Seelbach PW (1997). Reconciling landscape and local views of aquatic communities: lessons from Michigan trout streams. *Freshwater Biology*, **37**, 133-148.