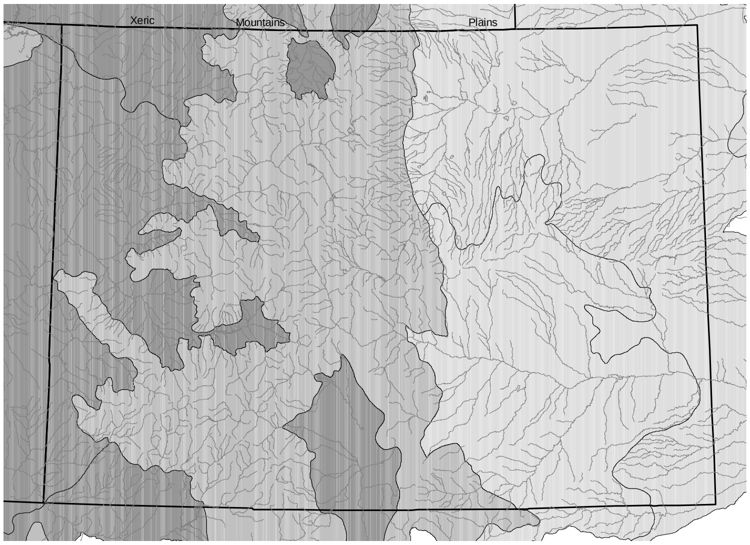
Development of Biological Assessment Tools for Colorado

Prepared for:

Colorado Department of Public Health and Environment

Water Quality Control Division – Monitoring Unit

Denver, CO



Prepared by:

Michael J. Paul[1], Jeroen Gerritsen[1], Chuck Hawkins[2], Erik Leppo[1]

[1]Tetra Tech, Inc.
400 Red Brook Boulevard, Suite 200
Owings Mills, MD 21117

[2]*Western Center for Monitoring and Assessment of Freshwater Ecosystems*
5210 Old Main Hill
Utah State University
Logan, UT 84322-5210

October 2005
Draft Report

ABSTRACT

Colorado is a unique state with an abundance of high quality natural freshwater resources. The continued protection of those resources will depend on the development of reliable assessment tools. Principal among these are methods for assessing biological integrity. In this study, multimetric and multivariate predictive indexes were developed for bioassessment of streams in Colorado. Macroinvertebrate data were assembled from an existing Department of Public Health and Environment database. Both modeling approaches rely on reference sites for constructing the models. In order to increase the available number of reference sites, the DPHE database was amended with biological samples from USGS NAWQA, USEPA EMAP, and Utah State STAR programs which were consistent in methodology with the DPHE program. Consistent reference criteria were applied across the dataset to generate a set of reference and stressed sites. Next, a set of operational rules for standardizing taxonomic resolution within and across sites was developed and applied to the data, one that uses chironomid taxa at the sub-family level of resolution to be consistent with US Forest Service standard taxonomy. Two indexes were developed. The multimetric index (MMI) development process began by classification of reference sites into bioregions. Three regions – the Mountains, Plains, and Xeric regions, were identified. Individual MMIs were developed for each region. Taxonomic data from each site was combined to calculate a series of biological metrics representing benthic macroinvertebrate taxa composition, richness, pollution tolerance, trophic or functional feeding behavior, and habit. Index construction consisted of screening potential metrics for sensitivity and variability, establishing a set of candidate metrics, assembling non-redundant candidate metrics into a number of potential multimetric indexes, and evaluating the variability and sensitivity of these to identify one model for each region. The final models consisted of 5 to 6 metrics representing each ecological category. All the models showed excellent discrimination between reference and stressed sites and low variability, with the exception of the Plains model, which was more variable. Model construction in the Xeric and Plains regions was hampered by low sample size of stressed and especially reference sites. Care is recommended in applying these two models and recalibration encouraged as soon as more data become available. The multivariate predictive model proceeded by classifying reference sites into groups based on taxonomic similarity, calculating the frequency of taxa within these groups, developing a discriminant model to predict the probability of a new site belonging to each of the groups using important determinants of taxon distributions, estimating taxon capture probabilities at a site as the frequencies of occurrence among classes weighted by the probabilities of a site belonging to a class, estimating the expected number of taxa (E) as the sum of capture probabilities, and comparing the number of observed (O) to expected taxa (E). It ranges from 0 to 1 and values less than 1 measure the loss of expected taxa from a site. The predictive model developed for Colorado is comparable to most models in use in the US or elsewhere. It accounted for substantial natural variability in taxonomic composition and was precise, accurate, and responsive, and showed little spatial bias across ecoregions or river basins. The model also uses three easily derived map-based predictor variables, simplifying its implementation. Despite the paucity of reference sites in lower elevation regions, the model appeared to be surprisingly robust in those regions. Future refinements of the model with data collected from additional reference sites should only improve confidence in assessments based on this approach.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

1.0 Introduction

Over the past century, land use activities such as mining, agriculture, silviculture, industrialization, and urbanization have contributed nonpoint sources of water pollution, and often degraded the quality of surface waters of the United States. In Colorado, investigating these nonpoint sources of water pollution has become a priority. It is the responsibility of Colorado's Department of Public Health and Environment (CDPHE) to maintain and protect the quality of the State's waters. In keeping with the Clean Water Act of 1972 (CWA, PL-92-500, and revisions of 1977, 1987) and technical guidance from the U. S. Environmental Protection Agency (USEPA), CDPHE has developed water quality standards for the protection of human and ecosystem health.

Through the 303(d) and Total Maximum Daily Load (TMDL) framework outlined in the CWA, waters considered impaired and threatened must be identified and improved to meet their designated uses. *Impairment*, as defined by natural resource management or regulatory agencies is typically based on attainment or non-attainment of numerical water quality standards associated with a waterbody's designated use. If those standards are not met (or attained), then the waterbody is considered to be impaired.

In support of its water quality standards, Colorado's ambient monitoring program has established an assessment "toolbox" that includes physical, chemical, and biological techniques. A core team of water resource professionals (biologists, hydrologists, naturalists, chemists and others) provide the technical resources to conduct the monitoring. Biological assessments are necessary for evaluating the health of Colorado's surface waters and for characterizing their biological condition. Resident aquatic biota in a watershed function as continous natural monitors of environmental quality, responding to the effects of both episodic as well as cumulative pollution and habitat alteration. Thus, the assessment of ecosystem health cannot be accomplished without a direct survey of those biota and comparison to regional ecological potential.

Biological integrity *is commonly defined as "the capability of supporting and maintaining a balanced, integrated, adaptive community or organisms having a species composition, diversity and functional organization comparable to that of the natural habitat of the regions"* (Karr and Dudley 1981, Gibson et al. 1996)

The Clean Water Act has as one of its primary goals the maintenance and restoration of biological integrity, which incorporates biological, physical, and chemical quality. This concept of biological integrity refers to the natural assemblage of indigenous organisms that inhabit a particular area that has not been affected by human activities (Frey 1977, Karr et al. 1986). The measurable definition of biological integrity is the *reference condition* (Barbour et al. 1995, 1999), which is characterized using data from minimally-disturbed sites within a region.

1.1 Colorado's Biomonitoring Program

Colorado DPHE has established a biological monitoring program for streams throughout the State. To date, the information collected in the biomonitoring program has not been incorporated into Colorado's 305(b) report or into its 303(d) list of impaired streams. The objective of this project was to develop biological indexes for interpreting the condition of Colorado streams, and to establish a database and assessment system for operational assessment, data storage, and future re-evaluation of indexes and methods by the Water Quality Control Division.

This report documents the development of two biological indexes, one a multimetric index, and one a multivariate predictive index, for use in the assessment of Colorado streams and to support the 305(b) report and 303(d) list of impaired waterbodies. Biological, chemical, and physical habitat data collected throughout Colorado between 1992 and 2003 were used to test for possible bioregion classifications and to develop the indexes. The specific questions investigated in this study were:

- Are the existing data sufficient to develop a biological index and biocriteria for Colorado?
- What is the most appropriate site classification for assessing ecosystem health across the diverse landscape and physiographic regions of Colorado?
- Which metrics are most appropriate for use in a Colorado multimetric macroinvertebrate stream condition index?
- What prediction model was best for Colorado streams?
- What biological index thresholds indicate the degree of comparability of Colorado streams to reference condition?
- What improvements can be made to better define the reference condition for ecosystem health of Colorado streams?

1.2 Tools for Biological Assessment

Careful sampling and analysis of aquatic systems and their resident biota can characterize biological condition relative to reference condition. Several key attributes are measured to determine the quality of the aquatic resources. Biological surveys establish the attributes or measures used to summarize several community characteristics, such as taxa richness, number of individuals in particular taxa groups or ecological categories, sensitive or insensitive taxa, observed pathologies, and the presence or absence of essential habitat elements.

1.2.1 Multimetric Indexes

Biological measurements, called metrics, represent elements of the structure and function of the bottom-dwelling (benthic) macroinvertebrate assemblage. Metrics change in some predictable way with increased human influence that alters environmental conditions (Barbour et al. 1996) and include specific measures of diversity, composition, functional feeding group representation, and information on tolerance to pollution. Multimetric indexes, such as an Index of Biotic Integrity (IBI), incorporate multiple biological community characteristics and measure the overall response of the community to environmental alteration and stress on the community (Karr et al. 1986, Barbour et al. 1995). Such a measure of the structure and function of the biota (using a regionally-calibrated multimetric index) is an appropriate indicator of ecological quality, reflecting biological responses to changes in physical habitat quality, the integrity of soil and water chemistry, geophysical process, and land use changes (to the degree that they affect the sampled habitat and water quality).

Multimetric, invertebrate indexes of biotic integrity (IBI), also variously called , ICI (Invertebrate Condition Index; Ohio EPA 1989), B-IBI (Benthic IBI; Kerans and Karr 1994), and SCI (Stream Condition Index; Barbour et al. 1996; Burton and Gerritsen 2003), have been developed for many regions of North America and are generally accepted for biological assessment of aquatic resource quality (e.g., Gibson et al. 1996, Plafkin et al. 1989; Barbour et al. 1999, Southerland and Stribling 1995, Karr 1991). The framework for bioassessment consists of characterizing reference conditions upon which comparisons can be made and identifying appropriate biological attributes with which to measure the condition. Reference conditions are typically the “best available” conditions where biological communities are the closest to natural for the particular region or area. These reference conditions are taken to be representative of healthy ecosystems.

1.2.2 Multivariate Predictive Models

RIVPACS assessments measure biological condition or quality by estimating the taxonomic completeness of a standard sample. Taxonomic completeness is a fundamental aspect of biological integrity and is defined here as the proportion of the taxa that should occur in a sample that were actually sampled. Values of the ratio, O/E, theoretically can range from 0 to 1, with values of 1 implying reference conditions and values less than 1 implying biological impairment. The accuracy and precision of RIVPACS assessments depend on the quality of the model used to predict the taxa expected to occur in a sample collected from an individual site. These models describe how probabilities of capture of all taxa vary across naturally occurring environmental gradients, information from which the taxa expected at individual sites can be derived. In contrast to multimetric indexes, the performance of these models does not depend on calibration against presumed stressed sites. Models are calibrated only with reference site data. If models accurately predict the assemblage that should occur at a site under reference conditions, any deviation from these predictions is a direct measure of biological impairment. Development and evaluation of RIVPACS models require the following steps:

1. Selection of a set of reference sites that adequately represent the naturally occurring environmental gradients in the region of interest (whole state, subregions, etc.).
2. Classification of reference sites based on their taxonomic similarity to one another.
3. Estimation of frequencies of occurrence of each taxon in each reference site class.
4. Development of a discriminant function model to predict the probability of a new site belonging to each reference site class from surrogate variables representing important determinants of taxon distributions.
5. Estimation of taxon probabilities of capture as the frequencies of occurrence among classes weighted by the probabilities of a site belonging to a class.
6. Estimation of the expected number of taxa at a site as the sum of the predicted probabilities of capture.

7. Assessment of the performance of the model by (1) comparing the observed number of predicted taxa (O) found at reference sites with the expected number of taxa (E) and (2) calculation of the precision in O/E estimates.

2.0 Data Sources and Organization

A robust dataset is the basis for developing any assessment tool. Colorado DPHE provided data in the Ecological Data Application System (EDAS; an Access database) for use in this study. This dataset included benthic macroinvertebrate, physical habitat (visual assessments and pebble count information), and water chemistry data. DPHE also provided hard copies and/or electronic files of additional data to be added to the database. To supplement the water chemistry data collected in conjunction with the benthic macroinvertebrate samples, additional DPHE water chemistry data was accessed from National STORET (a national environmental database warehouse maintained by USEPA) and added to EDAS. Even with a large dataset, there were areas of the State with poor site coverage (spatially or temporally). To increase the number of sites and samples, additional datasets were obtained from other agencies and added to EDAS. These additional datasets were of high quality and employed similar methods. The datasets used were Western EMAP and Southern Rockies Regional EMAP (US EPA Environmental Monitoring and Assessment Program), USU-STAR (Utah State University Western Center for Monitoring and Assessment of Freshwater Ecosystems Science to Achieve Results program), and NAWQA (US Geological Survey National Water Quality Assessment program).

Although different programs do not use the exact same methods, the protocols used by each program were relatively comparable (Table 1). All the programs sampled in riffles, used kick or D-frame nets with similar mesh, sampled a similar size area using comparable kick methods, and identified organisms to the lowest possible taxonomic level (see discussion on taxonomic resolution below). The principal differences among the programs were related to replication. Some combined replicates while other programs composited samples. A second difference was the subsample size. Most programs enumerated the whole sample, while DPHE identified a 300 count subsample.

Table 1 Methods comparison of programs whose data were combined for this analysis. (NEED STAR METHODS)

Methods Comparison	CO DPHE	REMAP	EMAP	NAWQA	USU STAR
Habitat Selection	Riffle/Run	Reachwide Riffle/Pool Separate	Richwide and Riffle	Riffle/Run/Pool	Riffle
Samping Net Type	Dipnet	Modified Kicknet	Modified D-frame	D-frame	
Sampling Net Size	8” x 18”	18”	12”	12”	
Sampling Mesh Size	500-600 µm	595 µm	425 µm	500 µm	
Sampling Method	1 m^2 riffle kick for 30-60 s	0.5 m^2 riffle kick for 20 s	1 ft^2 kick for 20-30 s	0.25 m^2 kick	
Replication	3 reps	9 reps Pool and riffle kept separate	Composite of 8 riffle samples	Field splitting until 750ml volume	
Subsampling and Enumeration	300 Organisms	Total Sample	Total Sample	Total Sample	
Taxonomic Level	LTU	LTU	LTU	LTU	LTU
LTU= Lowest Taxonomic Unit Identifiable					

The relational database structure greatly improved the ease of data manipulation for analysis. One principal adjustment for data comparability was to adjust the number of organisms subsampled across programs to 300 to be consistent with the DPHE method (Table 1). To accomplish this, non-DPHE samples were randomly subsampled to a 300 organism count before analysis.

By using data from other entities it was possible to increase the number of sites by almost 300 (Table 2). This provided an overall dataset with wider geographic coverage, especially in the Plains and Xeric bioregions (Figure 1). The additional datasets also increased the pool of candidate reference sites (next section).

Table 2. Number of stations with biological data by dataset and bioregion.

DataSet	Plains	Mountains	Xeric	*Total*
CO_DPHE	85	253	88	*426*
WEMAP	5	27	14	*46*
REMAP_SR		158	12	*170*
USU_STAR	4	28		*32*
NAWQA	25	10	8	*43*
Total	*119*	*476*	*122*	*717*

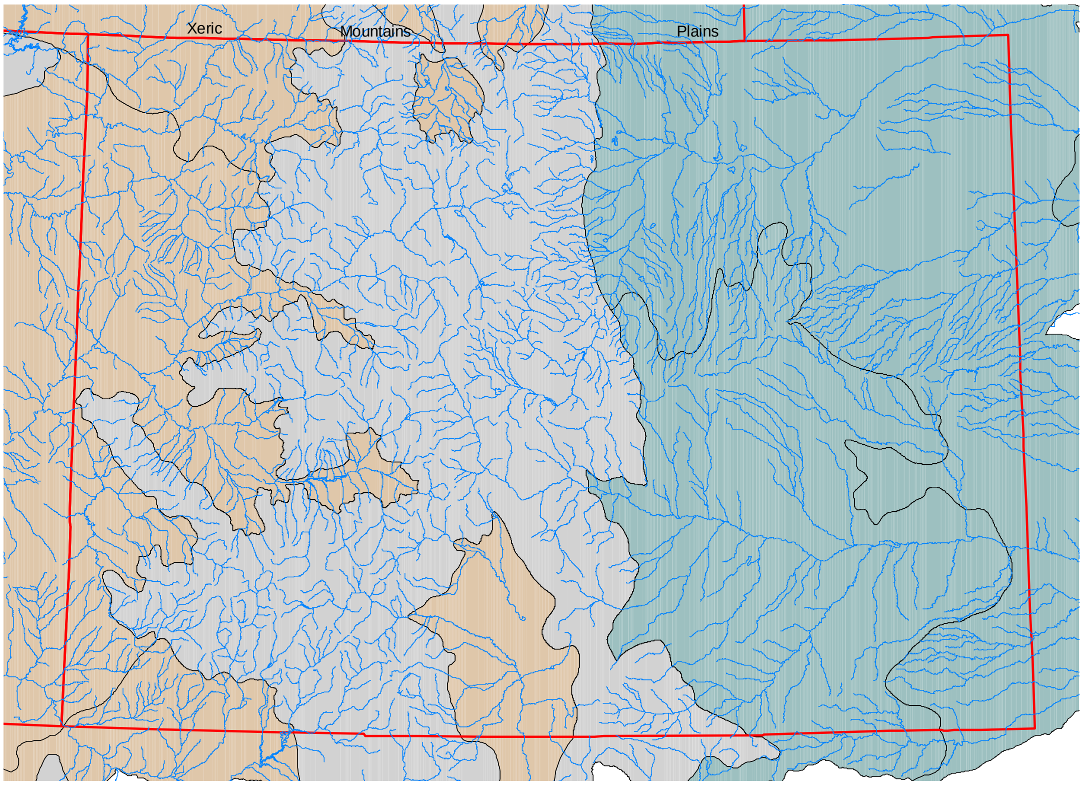


Figure 1. Colorado's bioregions (Xeric, Mountains, and Plains).

3.0 General Data Preparation

3.1 Reference Condition

Most biological assessment models evaluate the biological condition of a waterbody relative to some expected or reference condition. The biological communities of relatively undisturbed "reference" streams are representative of healthy ecological communities expected to occur under the natural range of relatively undisturbed habitat, climate, geomorphology, and other physico-chemical characteristics of a region. A simple metaphor would be the use of 98.6 degrees as a "reference" for human body temperature. That target represents an "average" for relatively healthy individuals. It was likely derived by defining a population of relatively "healthy" individuals using a set of criteria to define an expected healthy condition and then averaging the temperatures of all of those meeting the criteria.

It is critical to identify reference sites as both the multimetric and multivariate predictive models are built using them. Reference sites are ideally identified using a set of physical and chemical criteria that define a relatively undisturbed condition. There are two caveats: 1) The models need a sufficient number of reference sites to be robust, to precisely characterize the natural variability (noise) among reference conditions so that true departure from that condition can be detected (signal). As a result, initial criteria are often set to establish a sufficient number of reference sites as well. 2) Best professional judgement (BPJ) is often used to winnow certain sites from the pool of potential reference. In many cases, insufficient samples (one-time water chemical grabs) or temporally limiting data (decade old land cover data) are used to screen for reference sites. In some cases, the judgment of individuals more familiar with site conditions is used to remove sites from consideration as reference.

In addition to the use of reference sites, the approach used to build multimetric models applied here required the identification of stressed sites. In this approach, indexes are constructed based on stream biological community characteristics that best discriminate between reference and stressed streams. As a result, it was necessary to develop both reference and stressed site criteria to identify sites for building these models.

In order to identify these 2 classes of sites, a framework was developed based on the available data. The DPHE database lacked a sufficient pool of potential reference sites in some regions of the state. Reference sites were selected from the non-DPHE datasets using reference criteria developed by each data owner, but the reference criteria for each of these non-DPHE data sets was nearly identical and formed the basis of the selection process for the DPHE data.

There are 2 approaches to identifying reference sites, both use weight of evidence to identify reference sites. During the screening process, many different parameters are examined. Criteria are then developed such that a site has to meet all the criteria to be considered a reference site. These criteria are often very stringent and only a small subset of sites meets all the criteria. The criteria for each parameter are based on best professional judgment and knowledge about stream conditions and biological response. Once candidate reference sites are selected, sites are screened by those with site knowledge to remove sites that may have impacts not evident in the data.

Another method for identifying reference sites is based on excluding sites from a pool of candidate reference sites (inclusion by exclusion). Typically, this method is employed when many different data sets are used or in situations where the parameters collected have changed over time. Both of these situations occur in the DPHE analysis as a result of the variety of methods employed by different entities spatially and by DPHE through time. Ideally, the selection of reference sites would be based upon a large dataset with the same parameters collected for each site (e.g., water chemistry, physical habitat, landuse/landcover data). In this case, however, not all parameters were collected at all sites, so criteria were developed not to identify reference sites (as in the first method) but to exclude sites of poor quality. With this method a site was excluded from the pool of candidate reference sites if it failed any of the criteria for which data existed. Sites were not expected to have data for every parameter with a developed criterion, but they must have had enough collected data to pass a minimum number of criteria.. The criteria developed for this method were not necessarily less stringent than in the first approach. But as in the first method, criteria were not tied to any particular statute or regulation (though these were used as guidelines). Lastly, the pool of candidate reference sites was reviewed by DPHE personnel with knowledge of the sites to exclude sites with known problems not evident in the collected data.

The approaches and criteria employed in this study were consistent with EMAP and STAR and were not meant to identify pristine sites but to exclude those sites that were clearly stressed and should not be considered reference. For example, a percent urban land use of < 20 in the mountains was not necessarily indicative of reference conditions but values over this threshold were indicative of stress and excluded a site from consideration as reference. Stressed sites were those that failed a certain number of criteria. For the DPHE data, sites had to pass a minimum of 4 criteria and have no failures to be considered a candidate reference site, and any site with at least 4 failures was considered stressed.

There were 424 stations in the Colorado DPHE database with least one biological sample (benthic macroinvertebrates). Habitat data were collected at 175 sites and substrate particle size data (pebble count) were available from 24 sites. Field chemistry data (DO, pH, conductivity, and temperature) were collected at some of the sites. To increase the amount of data, water chemistry data were extracted from National STORET for as 141 of the DPHE sites with biological samples. To supplement sites with little data, land use data were generated. As a rough estimate of surrounding land use, a 1 kilometer buffer around the station was created and land use from MRLC 1992 was used to assign percent urban, forest, and agricultural land cover. This land use classification was revised with more recent data and with land cover in the actual catchments above each station during the model development.

All data were assembled in the Colorado EDAS database and final reference criteria were drawn from Colorado’s water quality standards, EMAP (Herlihy), nearby states (MT and WY), and best professional judgment (Appendix A).

Rather than a single set of criteria for the entire state, three sets of criteria were established, for three distinct regions (Plains, Mountains, and Xeric). These regions were defined by the level 3 ecoregions. The Plains in the east were defined as the Western High Plains and Southwestern Tablelands. The Mountains consisted of the Southern Rockies ecoregion. The Xeric region was defined as the Wyoming Basin, Colorado Plateau, and Arizona/New Mexico Plateau ecoregions. The reference criteria developed for the plains had to be different from the criteria developed for the mountains in order to generate a sufficient number of sites for model development.

Most stations were located in the Mountains bioregion (Table 3). This was also reflected in the number of candidate reference sites selected from the screening process (Table 4).

Table 3. Number of stations by bioregion and project.

Project	Plains	Mountains	Xeric	*Total*
CO_DPHE	85	253	88	*426*
WEMAP	0	84	6	*90*
REMAP_SR	0	80	6	*86*
NAWQA	25	10	8	*43*
STAR_CO	4	28	0	*32*
Total	*114*	*455*	*108*	*677*

In the non-DPHE projects, each monitoring agency had selected reference sites based on their own screening data. These non-DPHE data were, again, intended to supplement the CO DPHE reference sites to increase the sample size (Table 5).

Applying reference site criteria resulted in identifying 24 candidate reference sites in the DPHE data set. Of these 24 sites 13 were rejected by DPHE personnel as not representative of reference conditions. Reasons for rejecting a site as reference included being downstream of dams, hatcheries, or other dischargers. All reference sites from other datasets were retained (Table 5).

Table 4. Distribution of candidate reference sites in CO DPHE data set.

Project	Bioregion	Ref	Stressed
CO_DPHE	Mountains	15	61
CO_DPHE	Plains	2	4
CO_DPHE	Xeric	7	27
	Total	*24*	*92*

In the end, all USGS sites were removed from MMI development due to uncertainty about the nature of the abundance information included with those data. IN addition, all site duplicates were removed. All effort was made to assure reference sites were consistent between the two models. The reference sites used to build the MMI are listed in Appendix A.

Table 5. Final (Candidate) reference sites.

Project	Plains	Mountains	Xeric	*Total*
CO DPHE	0 (2)	3 (15)	0 (7)	*3 (24)*
WEMAP	2	8	0	*6*
REMAP	0	36	3	*38*
USU-STAR	4	28	0	*32*
Total	*6*	*75*	*4*	*85*

3.2 Taxonomic resolution

Assessment tools that rely on considering the number of taxa in a particular sample (e.g., richness metrics or O/E scores) require consistent taxonomic assignments of individual organisms to taxonomic groups. Ideally, every taxonomist would assign each individual invertebrate to the same taxon. However, the quality of samples and the expertise of taxonomists vary. As a result, specimens may not be identified to the same taxonomic resolution across all samples, and single samples may contain specimens identified to different hierarchical taxonomic levels. For example, one sample may have organisms identified to Diptera, Chironomidae, and *Chironomus*. In this example, it is impossible to tell whether these organisms represent a single taxon or three. Assuming that higher level identifications (order Diptera; family Chironomidae) are unique taxa when they are not, would result in an inflated richness estimate. Such ambiguities in taxonomy require correction. Commonly the taxonomy is corrected using consistent operational taxonomic rules.

Operational taxonomic units are decisions made based on the distribution of taxonomic identifications across hierarchical levels. This requires identifying at which level most individuals were assigned, and correcting non-conforming identifications. For example, if most samples have individuals identified to *Limnephilus* and only a few are identified to the family Limnephilidae, then the Limnephilidae individuals would be removed from the dataset. This loss of individuals is the cost of keeping the unique information provided to each site by having the specimens identified to genus. If all individuals were lumped to the family level, then every site with that family would “appear” more similar – even sites with very different limnephilid caddisfly genera. However, if most of the samples in a dataset were identified to the family level, then it would make sense to lift the genus samples to family since very little unique information would be lost.

For both assessment approaches, taxonomic resolution of the data was explored. Taxonomic ambiguities were corrected and used a practically identical set of operational taxonomic units for each analysis. The operational decisions made for the multimetric analysis were derived and matched to those made for the multivariate predictive model. Notable decisions include the exclusion of most Hemiptera taxa due to difficulty in obtaining representative samples for most the of the taxa in this order. Also, all chironomid taxa were identified to sub-family, to be consistent with taxonomic resolution of US Forest Service samples, a dataset that was anticipated to be accessed and incorporated into the CO DPHE database.

4.0 Multimetric Index (MMI) Development

4.1 Site Classification

Multimetric indexes are based on reference biological conditions and comparisons to those conditions. The reference condition is expected to vary due to natural differences among reference sites. If the differences are consistently associated with variable natural characteristics, then identification of multiple reference categories, or strata, would allow definition of multiple expectations of natural reference conditions. This would increase the chances of identifying truly degraded sites and decrease the chances of erroneously assessing a site as biologically impaired when it is actually of a different natural type.

Aquatic macroinvertebrate species are well-known to be specialized for certain water velocities, substrate types, water temperature, etc. (e.g., Merritt and Cummins 1996). Therefore, communities inhabiting fast riffles are very different from those inhabiting slow waters. Accordingly, we expected to find that Colorado Mountain streams would have different species compositions than Plains streams.

We examined alternative site classifications with non-metric multidimensional scaling (NMS) of the reference sites. Alternatives included the 3 aggregated bioregions (Mountains, Plains, Xeric) defined by EPA for the Western EMAP; catchment area; elevation; mean temperature; and mean precipitation. Due to the small number of reference sites in the Plains and Xeric areas, it was not feasible to break these down to Level 3 ecoregions.

As expected, ordination showed that Plains and some Xeric sites were separated in ordination space from the Mountain sites (Figure 2). Several of the Xeric sites appeared to group with the mountain sites; most but not all of these were on the Colorado Plateau. The first axis of the ordination was strongly associated with stream elevation, mean annual temperature, and mean annual precipitation, regardless of ecoregion or bioregion (elevation shown; Figure 3). Elevation, temperature, and precipitation are all strongly correlated with each other on the Rocky Mountain elevation gradient. Based on the ordination, elevation seems to be the single most important factor driving species composition of Colorado streams, similar to results found in developing the RIVPACS model approach (See Section 5)

Metrics are measurable attributes of the biological assemblages, and are calculated as aggregates of species or individuals in higher taxonomic, habit, or feeding groups (e.g., Plecoptera taxa, sprawlers, predators, sensitive taxa). Metrics may not indicate the same classification as the species composition. Accordingly, we examined raw metric values for association with elevation and catchment area. Calculation of metrics is described in Section 4.3, but we describe their relation with the classification variables here. Figure 4 shows the relationship of two candidate metrics, total taxa and percent tolerant individuals, to mean elevation. All taxa richness metrics were generally higher in mountain streams, but there was no smooth gradient corresponding to elevation: mountain streams as a class had higher richness than plains streams and low elevation xeric streams. Xeric reference streams at high elevations (> 2800 m mean catchment elevation) appear to be more like mountain streams in both species composition and metric values, and may have been misclassified. Currently, there are too few xeric reference sites to answer this question definitively.

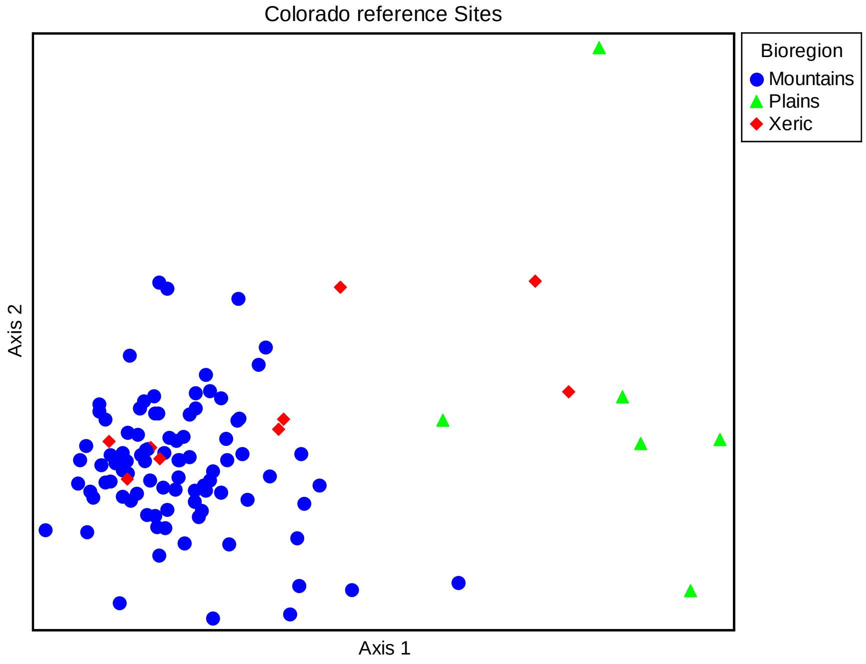


Figure 2. NMS ordination of Colorado reference sites, showing streams classified as Mountains, Plains, and Xeric

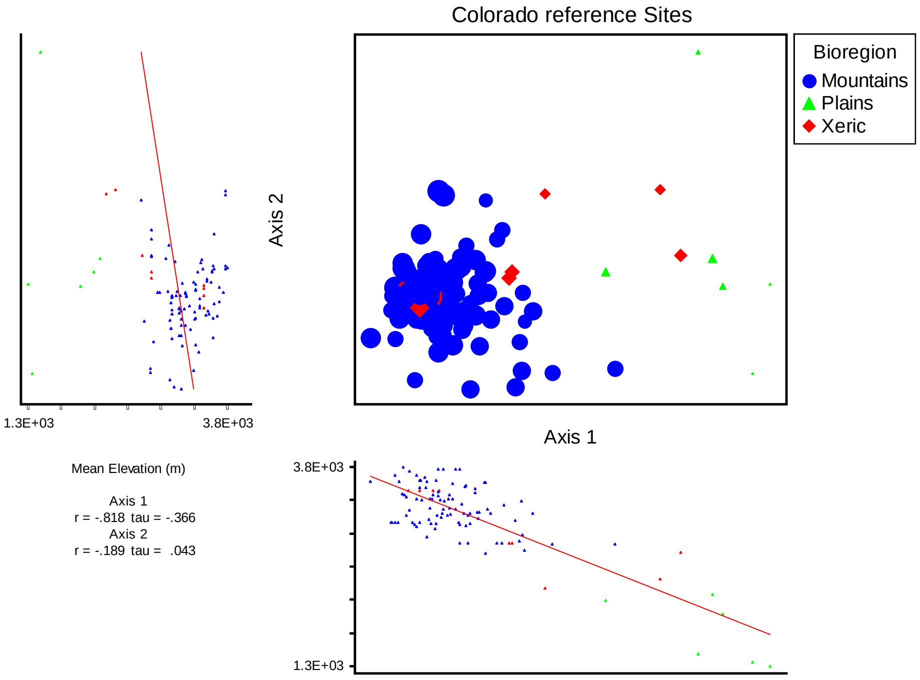


Figure 3. The NMS ordination of Fig. 2 showing correlation with mean catchment elevation. Left and lower scatter plots (with regression lines) show the relationship of elevation with the respective ordination axis: the lower graph shows ordination axis 1 on the x-axis and elevation on the y-axis. Symbol size in the ordination plot denotes the elevation: larger symbols represent sites at higher elevations.

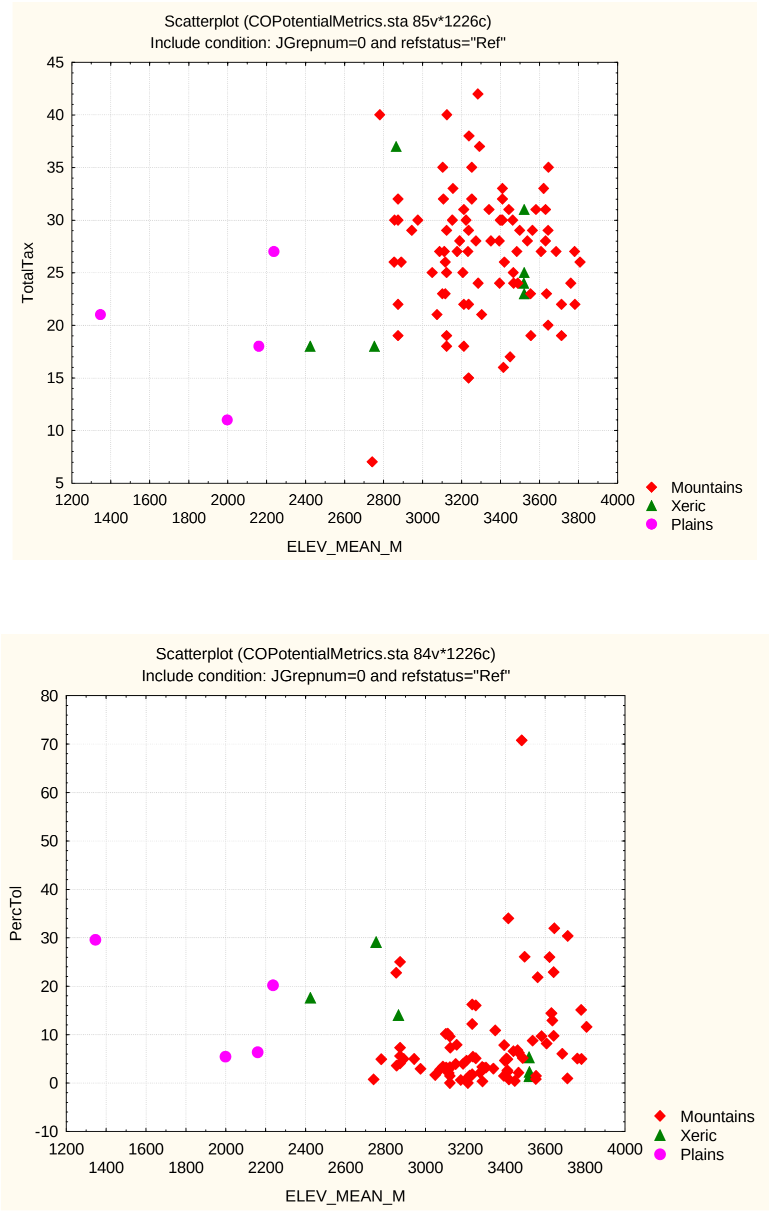


Figure 4. Metric values and mean elevation. a) total taxa. b) percent tolerant individuals.

Metric values at different elevations (Figure 4) suggested that a single division between high and low elevation streams would be sufficient, and that a regression calibration to elevation was unnecessary. From these graphs, the distinction between high and low elevation streams was approximately 2800 m, or 9200 feet (as mean catchment elevation, not measured site elevation – measured site elevation was not available at sufficient sites).

We tested metric responses (Section 4.2) using the elevation cutoff, as well as the 3 biological regions. When using elevation only, metrics from the low elevation xeric-plains class failed to discriminate between reference and stressed sites, that is, the responses were not consistent. When the xeric and plains sites were separated, metric responses were more consistent and interpretable.

In accordance with the above results, the three classes of Mountains, Plains and Xeric were retained for index development. We recommend that the xeric classification be reexamined when more reference site data are available. In particular, high elevation xeric sites may be more similar to mountain sites than to other xeric sites. We also recommend that actual site elevation be measured or estimated, as well as catchment elevation characteristics.

4.2 Metric Calculations and Responses to Stress

A biological metric is a numerical expression of a biological community attribute that responds to human disturbance in a predictable fashion. Metrics were considered for inclusion in this multimetric index on the basis of discrimination efficiency, low variability, ecological meaningfulness, contribution of representative and unique information, and sufficient range of values. They were organized into five categories: richness, composition, pollution tolerance, functional feeding group, and habit (mode of locomotion).

4.2.1 Methods

A suite of commonly applied, empirically proven, and theoretically responsive metrics was calculated for possible inclusion in a multimetric index (Table 6; see also Appendix B for descriptions of metrics). Tolerance metrics were based on amended Hilsenhoff tolerance values. Hilsenhoff tolerance values are on a 0 (most sensitive) to 10 (most tolerant) scale. The Hilsenhoff scale was derived primarily to address taxa tolerance to organic pollutants (Hilsenhoff 1987).

All richness metrics (e.g., insect taxa and non-insect taxa) were calculated such that only unique taxa were counted, through the use of the operational taxonomic unit (OTU) concept. Habit metrics were calculated using insect taxa only. All metrics were calculated in EDAS. Once calculated, the metrics were imported into the statistical package Statistica 7 for further analysis.

Discrimination efficiency

Discrimination efficiency (DE) is the capacity of the biological metric or index to detect stressed conditions. It is measured as the percentage of degraded sites that have values lower than the 25th percentile of reference values (Stribling et al. 2000). For metrics that increase with increasing stress, DE is the percentage of degraded sites that have values higher than the 75th percentile of reference values. DE can be visualized on box plots of reference and degraded metric or index values with the inter-quartile range plotted as the box (Figure 5). When there is no overlap of boxes representing reference and degraded sites, the DE is greater than 75%. A metric with a high DE thus has a greater ability to detect stress than metrics with low DEs. Metrics with DEs <25% do not discriminate and were not considered for inclusion in the index.

Table 6. Metric variability and discrimination efficiency. Discrimination efficiency (DE) is the percentage of degraded sites with metric values worse than the 25th or 75th percentile of reference. Metrics that decrease or increase with increasing stress are noted with D and I, respectively. Blanks cells were mathematically insoluble. (D)ecrease or (I)ncrease with stress:

Type				Mountains		Plains		Xeric	
	Metric	Stress Response	Response	DE	CV	DE	CV	DE	CV
Composition	Shannon Diversity	D	Y	0.45	0.18	1.0	0.26	0.72	0.12
	Percent Amphipoda (scuds)	I	N	0.00		0.05		0.68	1.90
	Percent Bivalvia (clams)	?	N	0.89	2.00	0.05	6.60	0.00	2.45
	Percent Chironomidae (midges)	I	N	0.72	0.39	0.20	0.86	22.45	0.78
	Percent Coleoptera (beetles)	D	Y	0.67	0.56	0.27	1.40	0.00	0.77
	Percent of Chironomidae that are *Cricotopus* or *Chironomus*	I	Y	0.28	1.52	0.71	2.45	0.00	1.83
	Percent Crustacea or Mollusca	I	N	1.00	2.00	0.15	2.81	2.38	1.40
	Percent Diptera (true flies)	I	Y	0.56	0.37	0.16	0.67	65.31	0.75
	Percent Ephemeroptera (mayflies)	D	Y	0.44	0.41	0.25	0.57	0.00	0.59
	Percent EPT (Ephemeroptera, Plecoptera, or Trichoptera)	D	Y	0.28	0.39	0.16	0.39	0.34	0.73
	Percent Gastropoda (snails)	I	N	0.00		0.13	8.47	1.70	1.25
	Percent non- insects	I	Y	0.67	1.23	0.22	1.23	34.35	1.39
	Percent Odonata (dragonflies)	D	Y	0.00		0.00		0.00	1.29
	Percent Oligochaeta (worms)	I	Y	0.61	1.57	0.44	4.65	31.97	
	Percent of Chironomidae that are Orthocladiinae	?	N	0.00		0.00	2.08	24.24	2.45
	Percent Plecoptera (stoneflies)	D	Y	0.72	0.89	0.35	0.73	0.00	1.68
	Percent Tanytarsini	I	N	0.94	0.88	0.16	2.08	0.68	1.34
	Percent of total Chironomidae that are Tanytarsini		N	0.83	0.59	0.25	1.18	3.03	1.17
	Percent Trichoptera (caddisflies)	D	N	0.44	0.36	0.24	1.06	0.34	1.43
Functional Feeding Group	Percent collectors	I	N	0.00	1.54	0.00	1.77	0.00	0.89
	Percent filterers	I	Y	0.06	1.40	0.00	2.14	0.00	1.36
	Percent predators	D	Y	0.72	0.54	0.33	0.55	8.50	0.85
	Percent scrapers	D	N	0.11	1.32	0.22	1.22	0.00	0.55
	Percent shredders	D	N	0.17	1.30	0.16	1.09	0.00	0.93
	Number of collector taxa	?	N	0.00	1.15	0.00	1.09	0.00	0.63
	Number of filterer taxa	?	N	0.06	0.86	0.00	1.20	0.00	0.73
	Number of predator taxa	D	Y	0.72	0.28	0.31	0.34	3.00	0.68
	Number of scraper taxa	D	N	0.06	0.58	0.07	0.55	0.00	0.70
	Number of shredder taxa	D	N	0.22	0.41	0.27	0.55	0.00	0.90

Table 6. Continued.

Type				Mountains		Plains		Xeric	
	Metric	**Stress Response**	**Response**	**DE**	**CV**	**DE**	**CV**	**DE**	**CV**
Habit	Percent burrowers	I	N	0.50	0.80	0.15	0.99	36.73	0.61
	Percent climbers	D	N	0.39	1.44	0.87	2.23	0.34	1.30
	Percent clingers	D	N	0.22	0.28	0.20	0.40	0.68	1.32
	Percent sprawlers	D	Y	0.72	0.90	0.36	1.23	8.84	0.96
	Percent swimmers	?	N	0.94	0.73	0.05	1.48	0.00	1.25
	Number of burrower taxa	?	N	0.61	0.67	0.22	0.65	3.00	0.37
	Number of climber taxa	D	N	0.00	1.15	0.00	1.19	1.00	0.89
	Number of clinger taxa	D		0.56	0.18	0.33	0.29	2.00	0.69
	Number of sprawler taxa	D	N	0.50	0.43	0.11	0.50	4.00	0.59
	Number of swimmer taxa	?	N	0.28	0.86	0.55	0.72	0.00	0.35
Richness	Chironomidae taxa	?	N	0.45	0.28	0.00	0.22	0.39	0.16
	Coleoptera taxa	?	N	0.25	0.81	0.33	0.37	0.22	0.41
	Crustacea or Mollusca taxa	?	N	0.20	2.43	0.33	1.26	0.61	2.00
	Diptera taxa	?	Y	0.31	0.30	0.00	0.32	0.56	0.09
	Ephemeroptera taxa	D	Y	0.44	0.32	1.00	0.59	0.67	0.43
	EPT taxa Ephemeroptera, Plecoptera, or Trichoptera)	D	Y	0.35	0.29	1.00	0.68	0.61	0.46
	Oligochaeta taxa	?	N	0.00	0.78	0.00	0.77	0.44	0.67
	Orthocladiinae taxa	?	N	0.07	0.16	0.00	0.49	0.00	0.00
	Plecoptera taxa	D	N	0.29	0.40	0.00	2.45	0.67	0.69
	Tanytarsini taxa	?	N	0.00	8.77	0.00		0.00	
	Total number of taxa	D	Y	0.27	0.39	0.00	0.00	0.17	0.00
	Trichoptera taxa	D	N	0.42	0.21	1.00	0.38	0.56	0.18
Tolerance	Beck's Biotic Index	D	Y	0.44	0.43	0.31	0.52	1.00	1.05
	Hilsenhoff Biotic Index	I	Y	0.44	0.49	0.36	0.30	4.00	0.95
	North Carolina Biotic Index	D	N	0.56	0.46	0.42	0.33	5.82	0.23
	Percent individuals in the most abundant taxon	I	Y	0.36	0.46	0.67	0.44	0.33	0.32
	Percent of Ephemeroptera that are Baetidae	I	N	0.17	0.14	0.42	0.46	37.77	0.42
	Pecent intolerant individuals	D	Y	0.67	0.37	0.33	0.53	0.00	0.90
	Pecent tolerant individuals	I	Y	0.28	0.85	0.24	0.47	0.00	1.60
	Percent of Trichoptera that are Hydropsychidae	I	Y	0.78	1.02	0.25	1.20	34.69	0.94
	Percent of EPT that are Hydropsychidae	I	N	0.44	1.04	0.55	1.72	0.00	0.85
	Number of intolerant taxa	D	N	0.61	1.35	0.49	2.28	0.00	1.29
	Number of tolerant taxa	I	Y	0.22	0.66	0.29	0.33	0.00	0.92

Metric variability

Metric variability was estimated for the reference site population. The coefficient of variability (CV) standardizes variability as a function of mean values (CV = standard deviation / mean). When comparing metrics, those with lower variability in the reference conditions are preferable to those with higher variability. Lower CVs indicate lower variability in relation to means. There was no threshold CV above which metrics would not be included in the index, but metrics with low CVs were preferred over those with high CVs.

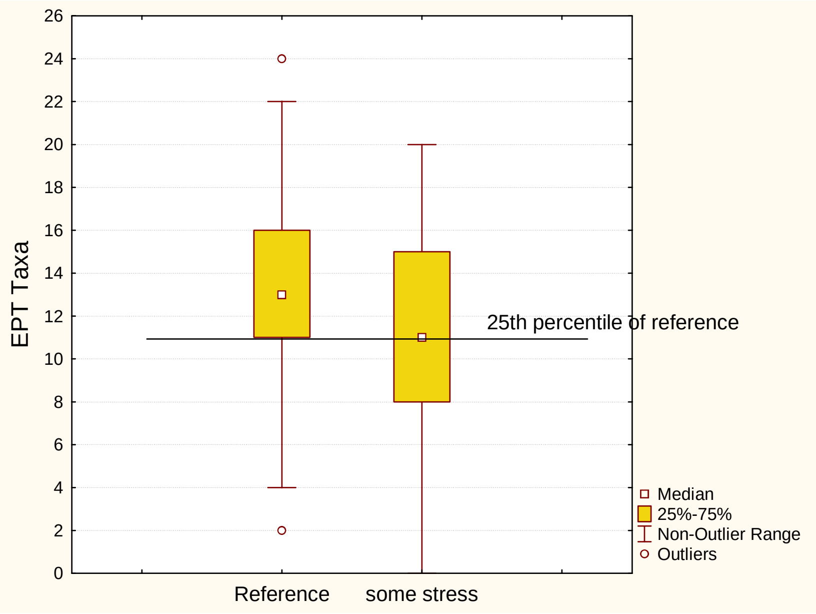


Figure 5. Illustration of metric discrimination efficiency (DE) between reference and stressed site samples, EPT taxa in Colorado Mountain region. In this example, DE is 50% because half of the stressed sites fall below the 25th percentile of the reference

Other metric considerations

Ecologically meaningful metrics are those for which the assemblage response mechanisms are understandable and are represented by the calculated value. Ecological meaningfulness is a professional judgment based on theoretical or observed response mechanisms. Those metrics that respond according to expectations established in other studies are defensible.

Metrics contribute information representative of integrity if they are from diverse metric categories. As many metric categories as practical should be represented in an index so that signals of various stressors can be integrated into the index. (Karr and Chu 1999) While several metrics should be included to represent biological integrity, those that are included should not be redundant with each other. Redundancy was evaluated using a Pearson Product-Moment correlation analysis.

For metrics to discriminate on a gradient of stress, they must have a sufficient range of values. Metrics with limited ranges (e.g., richness of taxa poor groups or percentages of rare taxa) may have good discrimination efficiency. However, small metric value changes will result in large and perhaps meaningless metric scoring changes.

4.2.2 Metric Results

Sixty three (63) metrics were calculated in the five metric categories (Table 6). Within the dataset, 27 metrics were minimally responsive in at least one of the 3 bioregions of Colorado.

Metrics were excluded from consideration in possible index alternatives if they did not discriminate or discriminated weakly between reference and degraded sites, were conceptually redundant with other, more discriminating metrics, or were not representative of the benthic community.

4.3 Index Composition

A multimetric index is a combination of metric scores that indicates a degree of biological stress in the stream community (Barbour et al. 1999). Individual metrics are candidates for inclusion in the index if they:

- discriminate well between reference and stressed sites;
- are ecologically meaningful (mechanisms of responses can be explained);
- represent diverse types of community information (multiple metric categories); and
- are not redundant with other metrics in the index.

Several index alternatives were calculated using an iterative process of adding and removing metrics, calculating the index using the average of individual metric scores, and evaluating index responsiveness and variability (Appendix C). The first few index alternatives included all the metrics, and then subsets of those that had the highest DEs within each metric category. Subsequent index alternatives were formulated by adding, removing, or replacing one metric at a time from the initial index alternatives that performed well. This was repeated for all the bioregions identified (mountains, plains, and xeric). The index alternatives recommended for each region were those that met the criteria listed above and that could not be improved (increased DE, lower variability) by substituting, adding, or removing metrics.

Metrics were scored on a common scale prior to combination in an index. The scale ranged from 0 to 100 and the optimal score was determined by the distribution of data. For metrics that decrease with increasing stress, the 95th percentile of all data was considered optimal and scored as 100 points using the equation

$$\text{Score} = 100 \times \frac{\text{Metric Value}}{95^{th}\ \text{Percentile}}.$$

All other metric values were scored as a percentage of the 95th percentile value (Figure 6) except those that exceeded 100, which were assigned a score of 100. The 95th percentile value was selected as optimal instead of the maximum so that outlying values would not skew the scoring scale. Metrics that increased with stress (reverse metrics) were scored based on the 95th and 5th percentile using the equation

$$Score = 100 \times \frac{(95^{\text{th}}\ \text{Percentile} - \text{Metric Value})}{(95^{\text{th}}\ \text{Percentile} - 5^{\text{th}}\ \text{Percentile})}.$$

It is important to note that the percentiles were derived from the data without any site replicates. Any site replicates within a 2 year period were removed for estimating percentiles since that would allow sites with more than one replicate to unduly influence metric scoring. Metric values for those replicates were, however, scored using the percentiles generated.

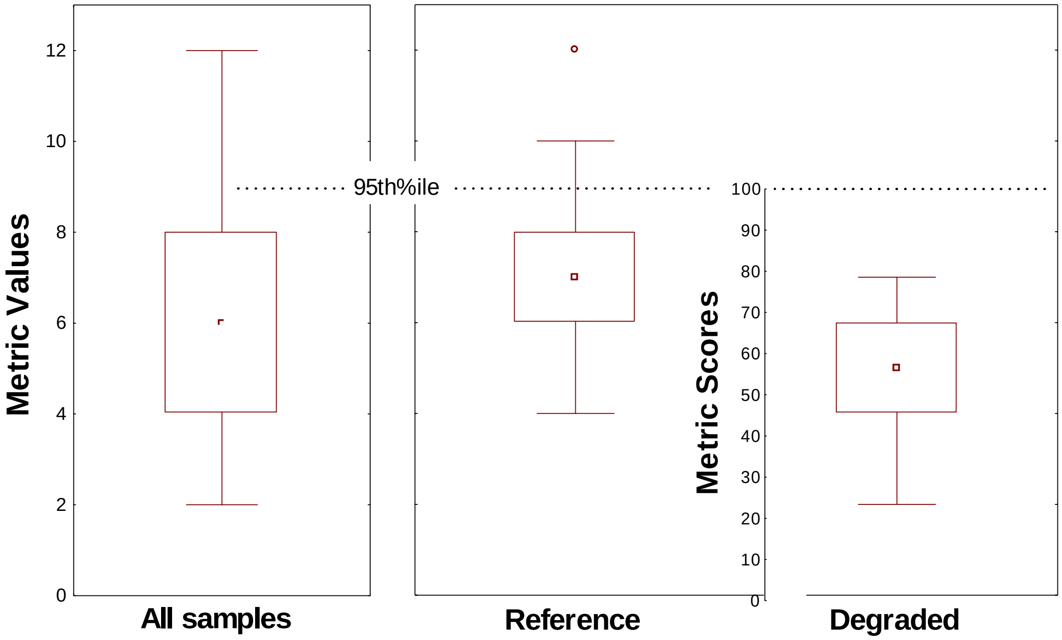


Figure 6. Metric scoring schematic for metrics that decrease with increasing stress. For metrics that increased with increasing stress, the 5th percentile of the data was considered optimal and assigned a value of 100 points. Metric values were scaled down towards 0. The lower end of the scoring scale is defined as the maximum metric value encountered in the dataset.

Each alternative index was calculated by averaging the candidate metric scores selected for that alternative. Each alternative index was evaluated based on discrimination efficiency (DE, calculated as for individual metrics), separation of reference and stressed index means, the inter-quartile range of reference index scores, and the relative variability of the reference site scores defined by the coefficient of variation (CV) within reference sites. Again, for this process, site replicates were removed to avoid any site unduly influencing the evaluation of any index.

4.3.1 Index Composition Results

Fifteen (15) index alternatives were calculated and tested, 5 each for every bioregion (Appendix C). These were reduced to 3 candidate indexes based on performance. We recommend the following index for each bioregion.

Mountains Index

The index alternative that is recommended for adoption in the Mountains contains five metrics, as follows:

- Percent Oligochaete (Composition)
- Total Taxa (Richness)
- Percent Climbers (Habit)
- Percent Trichoptera which are Hydropsychidae (Tolerance)

The index using the one richness metric (Total Taxa) was selected over alternatives which used dipteran taxa or chironomid taxa. Chironomid taxa was redundant with Diptera taxa, so the two could not be used together. Chironomid taxa also had a much lower DE, but did not contribute as much to overall model discrimination. Diptera taxa had a higher DE than total taxa, but did not contribute as well to overall model DE, so it was left out.

One composition metric was included in the index, percent Oligochaete. This was the only composition metric with a substantial DE. The habit metric, percent climbers, was selected over other habit metrics because it had the highest DE by a substantial margin. The remaining metric was a tolerance metric. The percent Trichoptera which are Hydropsychidae had the highest DE. It had the highest DE of the tolerance metrics and contributed most to the overall model.

No metrics used in the index were correlated at $r > 0.80$ or $r < -0.80$ (Table 7). Forty-seven of the 55 stressed sites had index scores lower than the 25th percentile of the reference site scores (Figure 7) resulting in an index DE of 85%. The mean separation of index scores between reference and stressed sites was 17 points. The inter-quartile range of reference index values was 7 points.

The index includes 4 of the 5 main metric categories (composition, richness, tolerance, habit, and trophic behavior). Ideally, all 5 categories are represented, but the lack of adequate discriminatory metrics in the trophic category precluded its representation.

Table 7. Correlations (Pearson Product-Moment) among candidate metrics of the Mountain Index.

	Percent Oligochaete	**Percent Climber**	**Percent Sprawler**	**Swimmer Taxa**	**Chironomid Taxa**	**Diptera Taxa**	**Total Taxa**	**Hilsenhoff Biotic Index**	**Percent Dominant**	**Percent Trichoptera which are Hydropsychidae**	**Percent EPT which are Hydropsychidae**
Percent Oligochaete	1.00										
Percent Climber	-0.14	1.00									
Percent Sprawler	0.01	-0.18	1.00								
Swimmer Taxa	-0.08	-0.11	-0.14	1.00							
Chironomid Taxa	-0.05	-0.11	0.08	0.23	1.00						
Diptera Taxa	-0.11	-0.02	-0.09	0.18	0.65	1.00					
Total Taxa	-0.20	0.10	-0.36	0.33	0.37	0.60	1.00				
Hilsenhoff Biotic Index	0.40	-0.24	0.32	-0.03	0.19	0.07	-0.30	1.00			
Percent Dominant	0.12	0.09	0.26	-0.20	-0.16	-0.28	-0.54	0.28	1.00		
Percent Trichoptera which are Hydropsychidae	-0.05	-0.04	-0.08	-0.05	-0.02	0.03	0.04	0.17	-0.05	1.00	
Percent EPT which are Hydropsychidae	-0.01	-0.14	-0.08	-0.05	-0.06	-0.01	0.07	0.13	-0.05	0.70	1

Mountains Index Interpretation

The metrics in the Mountains index are fairly straightforward in interpretation. Although the mechanisms by which aquatic macroinvertebrates responded to environmental stressors may not be fully understood because of a lack of adequate environmental data and mechanistic information, the fact that the metrics were responsive to a general gradient of stress (reference – degraded) suggests that they were responding to a common suite of stressors. The metrics in this and the other indexes were therefore selected largely based on their demonstrated responses in this data set.

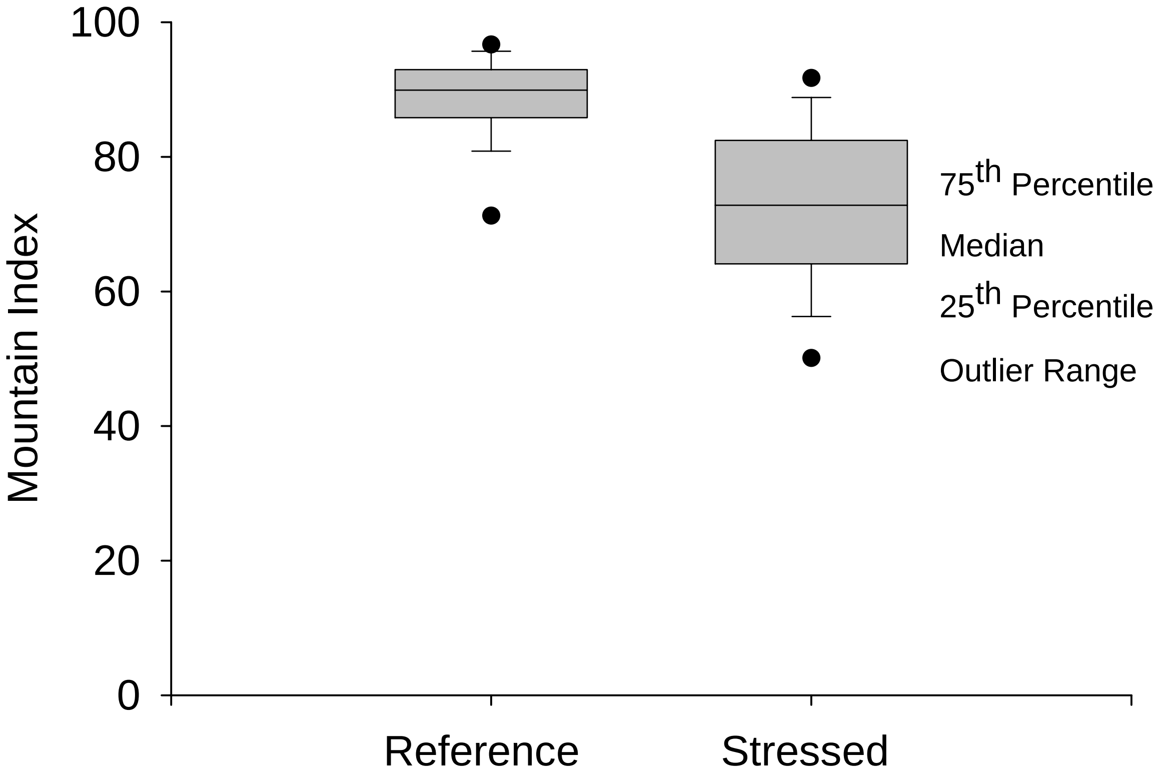


Figure 7. Multimetric index values in reference and stressed Mountain sites.

Oligochaete taxa, are well established tolerant taxa, able to exploit stressed conditions using their low oxygen tolerance and flexible feeding behavior. These taxa increased in stressed conditions, providing a good signal. High taxa richness usually correlates with increased ecological health of the stream and suggests that niche space, habitat, and food sources are adequate to support the survival and propagation of many species. As a result, taxa richness usually declines with stress, and this was true as evidenced by the selection of the Total Taxa metric. Climber taxa are those that move about the substrate and include many sensitive taxa. Lastly, tolerant taxa are expected to increase with stress as a percent of the community composition as sensitive taxa die or migrate away. The hydropsychid caddisflies are generally considered to contain more tolerant taxa than other caddisfly families. This metric responded positively to stress.

Plains Index

The index alternative that is recommended for adoption in the Plains contains five metrics:

- Percent Chironomidae (Composition)
- EPT Taxa (Richness)
- Hilsenhoff Biotic Index (Tolerance)
- Percent Burrowers (Habit)
- Percent Predators (Trophic)

The index used only one richness metric, EPT taxa, because this was redundant with Ephemeroptera taxa but had a larger range of values, providing more potential information and resolution. Other richness metrics lacked range, were too variable, or showed very little discrimination potential.

One composition metric were included in the index. Percent Chironomidae (Diptera) had among the highest DE among the candidate composition metrics, had the largest range, and contributed best to overall model DE. Other potential composition metrics were either redundant with percent chironomid or, again, did not improve the overall model. Only one tolerance metric was included, the Hilsenhoff Biotic Index (HBI). Percent tolerant and tolerant taxa richness were candidates, but both lowered overall model performance relative to the HBI. Percent burrower taxa was the best habit metric in terms of lowering model DE, and percent predators was the only functional metric to be considered. Others exhibited little or not discrimination.

No metrics used in the Plains index were correlated at $r > 0.8$ or $r < -0.8$ (Table 8). All of the stressed sites (3) had index scores lower than the 25th percentile of the reference site scores (Figure 8) resulting in an index DE of 100%. The mean separation of index scores between reference and stressed sites was 33 points. The inter-quartile range of reference index values was 16 points.

Clearly, the lack of a substantial number of stressed or reference sites may limit the applicability of the Plains model. Whenever a model is built with low replication, the potential for the model to be unique only to the data used to construct the model is high because the true natural variability among reference and stressed sites is not adequately characterized. This “overfitting” of models can be reduced by validating the model with independent data. This could be done using data collected subsequent to those used to construct this model. In any case, we recommend using this preliminary model with caution, and recalibrating the model as soon as more data from Plains streams become available.

Table 8. Correlations (Pearson Product-Moment) among candidate metrics of the Plains Index.

	Percent Chironomid	**Percent Diptera**	**Percent Ephemeroptera**	**Percent EPT**	**Percent Non-insects**	**Percent Oligochaete**	**Percent Predator**	**Percent Burrower**	**Clinger Taxa**	**Ephemeroptera Taxa**	**EPT Taxa**	**Hilsenhoff Biotic Index**	**Percent Tolerant**	**Tolerant Taxa**
Percent Chironomid	1.00													
Percent Diptera	0.88	1.00												
Percent Ephemeroptera	-0.46	-0.54	1.00											
Percent EPT	-0.48	-0.56	0.94	1.00										
Percent Non-insects	-0.52	-0.58	-0.28	-0.32	1.00									
Percent Oligochaete	-0.40	-0.45	-0.29	-0.32	0.88	1.00								
Percent Predator	0.04	0.07	-0.07	-0.05	-0.10	-0.21	1.00							
Percent Burrower	0.24	0.25	-0.19	-0.22	-0.08	-0.10	0.05	1.00						
Clinger Taxa	-0.19	-0.25	0.45	0.59	-0.35	-0.37	0.14	-0.14	1.00					
Ephemeroptera Taxa	-0.28	-0.34	0.74	0.72	-0.34	-0.33	0.02	-0.21	0.61	1.00				
EPT Taxa	-0.28	-0.34	0.66	0.75	-0.38	-0.37	0.07	-0.22	0.80	0.89	1.00			
Hilsenhoff Biotic Index	-0.26	-0.24	-0.40	-0.46	0.75	0.69	-0.07	0.12	-0.44	-0.49	-0.54	1.00		
Percent Tolerant	-0.38	-0.47	-0.13	-0.19	0.75	0.67	-0.03	0.15	-0.24	-0.29	-0.33	0.78	1.00	
Tolerant Taxa	0.08	0.07	0.00	-0.02	-0.08	-0.18	0.40	0.20	0.21	0.12	0.09	0.10	0.20	1.00

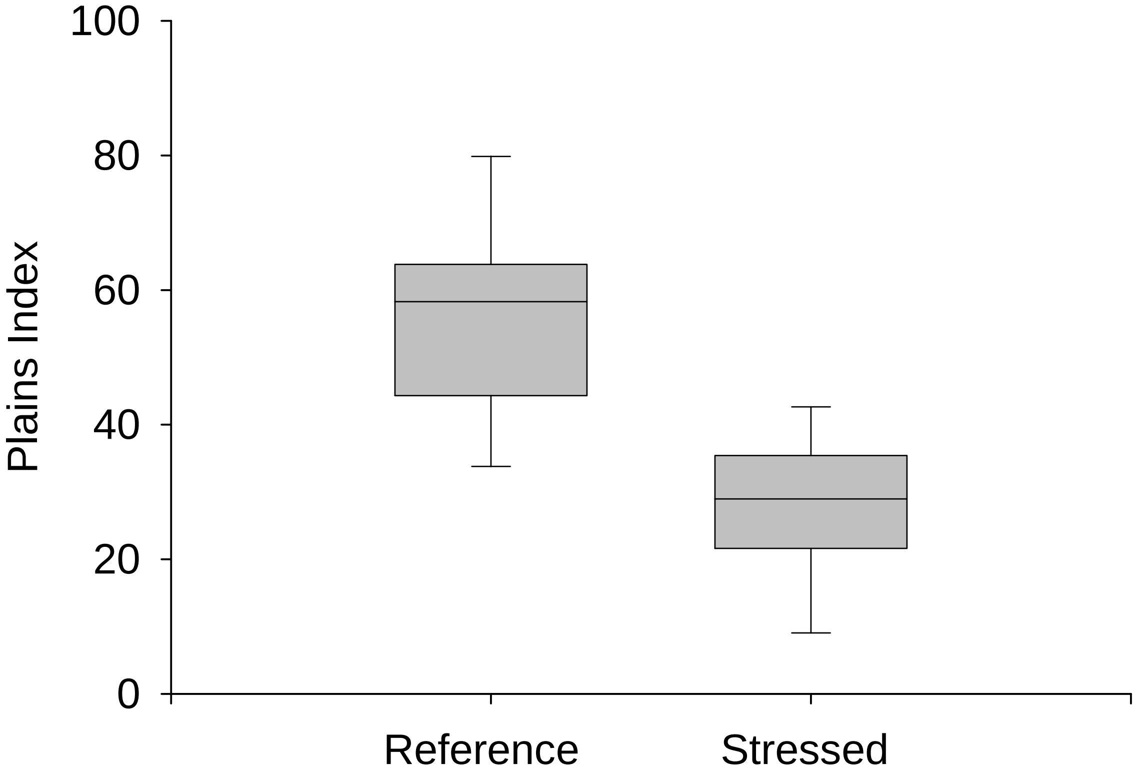


Figure 8. Multimetric index values in reference and stressed Plains sites.

The Plains index includes all 5 main metric categories.

Plains Index Interpretation

The metrics in the Plains index are also fairly straightforward in interpretation and behavior of a few of the metrics was explained earlier (see Mountains interpretation). The interpretation of metrics unique to this index are discussed here. Diptera, especially many chironomidae, are generally tolerant organisms and their percent contribution to community structure, would be expected to increase with stress. This was observed in the plains streams and explains the contribution of these metrics. EPT richness is the sum of sensitive mayfly, stonefly, and caddisfly taxa in the stream. Not surprisingly, it showed a strong response to disturbance in Plains streams. A unique tolerance metric to the Plains index was the HBI. This metric is an average tolerance score derived from tolerance values ascribed to each taxon based on their response to stress. The tolerance values of each taxon in a sample are weighted by their individual abundances. This weighted average is the overall HBI score for a sample. These scores increase with stress. Percent burrowers is a habit metric that reflects the physical position of organisms in the habitat. Burrower taxa burrow into and reside in sediment. In general, this metric increases in response to stress, as it reflects taxa (such as oligochaetes) which are generally tolerant to low oxygen and high fine sediment environments. Lastly, percent predatory taxa was included in the Plains index. These are generally larger, long-lived organisms that are sensitive to disturbance and often show strong response to changes in the environment, especially those affecting their preferred prey.

Xeric Index

The last index is for the Xeric region and the alternative recommended for adoption here contained seven metrics:

- Percent Coleoptera (Composition)
- Diptera Taxa (Richness)
- Percent dominant taxon (Tolerance)
- Percent Climbers (Habit)
- Predator Taxa (Trophic)

Percent Coleoptera had the second highest DE among the candidate composition metrics, but it greater range than the other and contributed better to the overall model. Percent EPT and percent Diptera had low DEs. Diptera taxa richness was included. Total taxa had a higher DE, but performed poorly in the overall model. Other metrics had lower DE values or had restricted ranges. One tolerance metric was included, percent of the dominant taxon. The other candidate tolerance metrics had very low DE values and contributed nothing to the model.

Percent climbers was the only habit metric included. It had a high DE and contributed to model DE. The other habit metrics exhibited lower discrimination between stressed and reference sites. Lastly, predator taxa was included as a trophic metric. It had the highest DE of the habit metrics and was selected over percent predators because of its greater contribution to overall model performance.

No metrics used in the Xeric index were correlated at $r > 0.8$ or $r < -0.8$ (Table 9). Sixteen of the 18 stressed sites (3) had index scores lower than the 25th percentile of the reference site scores (Figure 9) resulting in an index DE of 89%. The mean separation of index scores between reference and stressed sites was 20 points. The inter-quartile range of reference index values was 3 points.

As mentioned earlier for the Plains model, the lack of a substantial number of reference sites may limit the applicability of the Xeric model as well. The Xeric region did include a much larger number of stressed sites (18) than the Plains region (3). So, the characterization of natural variability of stressed sites is better. However, both regions had a similar number of reference

Table 9. Correlations (Pearson Product-Moment) among candidate metrics of the Xeric Index

	Percent Coleoptera	**Percent Diptera**	**Percent Ephemeroptera**	**Percent Plecoptera**	**Percent Trichoptera**	**Percent Predators**	**Predator Taxa**	**Percent Climbers**	**Clinger Taxa**	**Sprawler Taxa**	**Chironomid Taxa**	**Diptera Taxa**	**Ephemeroptera Taxa**	**EPT Taxa**
Percent Coleoptera	1.00													
Percent Diptera	-0.39	1.00												
Percent Ephemeroptera	-0.16	-0.51	1.00											
Percent Plecoptera	0.03	-0.36	0.25	1.00										
Percent Trichoptera	0.33	-0.45	-0.16	0.02	1.00									
Percent Predators	-0.09	0.15	-0.13	0.31	-0.11	1.00								
Predator Taxa	0.29	-0.26	0.01	0.20	0.24	0.16	1.00							
Percent Climbers	-0.15	-0.25	0.44	0.09	0.14	-0.13	0.02	1.00						
Clinger Taxa	0.41	-0.45	0.21	0.24	0.47	-0.10	0.57	0.13	1.00					
Sprawler Taxa	-0.13	0.38	-0.23	-0.13	-0.21	0.07	0.29	-0.10	-0.11	1.00				
Chironomid Taxa	0.00	0.37	-0.25	-0.13	-0.15	-0.06	0.28	-0.08	0.08	0.57	1.00			
Diptera Taxa	0.13	0.11	-0.16	0.03	0.13	-0.03	0.49	0.08	0.29	0.50	0.69	1.00		
Ephemeroptera Taxa	0.25	-0.55	0.47	0.21	0.36	-0.11	0.36	0.16	0.76	-0.14	-0.10	0.06	1.00	
EPT Taxa	0.37	-0.58	0.34	0.37	0.51	-0.09	0.50	0.21	0.88	-0.16	-0.09	0.16	0.86	1.00

Table 9. Continued.

	Plecoptera Taxa	Total Taxa	Hilsenhoff Biotic Index	North Carolina Biotic Index	Percent Dominant	Percent Trichoptera which are Hydropsychidae
Plecoptera Taxa	1.00					
Total Taxa	0.56	1.00				
Hilsenhoff Biotic Index	-0.64	-0.56	1.00			
North Carolina Biotic Index	-0.48	-0.24	0.56	1.00		
Percent Dominant	-0.35	-0.63	0.45	0.25	1.00	
Percent Trichoptera which are Hydropsychidae	0.03	0.10	-0.14	-0.19	-0.15	1.00

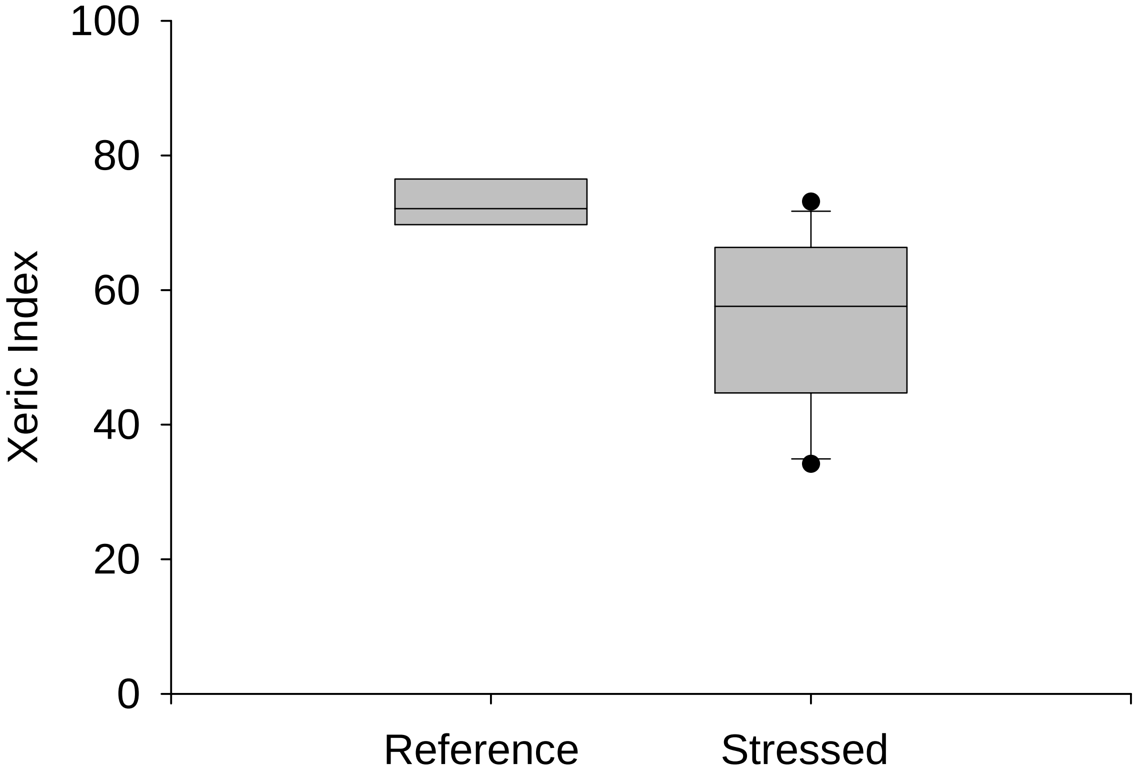


Figure 9. Multimetric index values in reference and stressed Xeric sites.

sites (6 in Plains, 5 in Xeric), so, arguably, the reference condition is not as well characterized. More reference sites would improve confidence in the model.

The Xeric index also included all 5 of the main metric categories.

Xeric Index Interpretation

Again, a subset of the metrics in the Xeric index were explained earlier (see Mountains and Plains index interpretation). Unique metrics to the Xeric region include the percent Coleoptera and percent dominant taxa. Beetles are generally sensitive insects and the percent contribution of these individuals decreases with stress. Both of these metrics responded predictably to stress in the Xeric region. The percent of the most dominant taxon is the percent that the most abundant taxon contributes to the overall abundance. In stressed conditions, communities commonly become dominated by only a few fairly tolerant taxa. The “evenness” of the community composition decreases dramatically, therefore, and the percent of a dominant taxon increases. This metric increased from reference to stressed Xeric sites.

4.3.2 Conclusions and recommendations

Three multimetric indexes (MMI) were developed as tools for identifying biological degradation in Colorado, one for each of the three bioregions – Mountains, Plains, and Xeric. The Mountain MMI used the following metrics:

- Percent Oligochaete (Composition)
- Total Taxa (Richness)
- Percent Climbers (Habit)
- Percent Trichoptera which are Hydropsychidae (Tolerance)

For averaged data, 47 of 55 (85%) of stressed sites had Mountain MMI scores lower than the 25th percentile of reference scores. This index also had a separation of average reference and stressed scores of 17 points, a reference site interquartile range of 7, and an overall coefficient of variation (standard deviation of reference scores/average reference score) of 9%.

The Plains MMI used the following metrics:

- Percent Chironomidae (Composition)
- EPT Taxa (Richness)
- Hilsenhoff Biotic Index (Tolerance)
- Percent Burrowers (Habit)
- Percent Predators (Trophic)

For averaged data, all 3 of the (100%) degraded sites had Plains MMI scores lower than the 25th percentile of reference scores. This index also had a separation of average reference and stressed scores of 33 points, a reference site interquartile range of 16 points, and an overall coefficient of variation (standard deviation of reference scores/average reference score) of 30%.

The Xeric MMI used the following metrics:

- Percent Coleoptera (Composition)
- Diptera Taxa (Richness)
- Percent dominant taxon (Tolerance)
- Percent Climbers (Habit)
- Predator Taxa (Trophic)

For averaged data, 16 of the 18 (89%) degraded sites had Xeric MMI scores lower than the 25th percentile of reference scores. This index also had a separation of average reference and stressed scores of 20 points, a reference site interquartile range of 3 points, and an overall coefficient of variation (standard deviation of reference scores/average reference score) of 8%.

We recommend applying the three MMI across Colorado. However, the Plains MMI was constructed with very few reference (6) and stressed (3) sites. As a result, little of the true natural variation in metric values is characterized by the model, which means the potential applicability of this model is limited until it can be verified with an independent set of data or recalibrated with larger sample sizes. Similarly, the Xeric model was built with very few reference sites (5), although there were significantly more stressed Xeric sites (18). Again, until additional data are made available to validate or recalibrate this model, care should be taken in applying this model. The mountains model was much more robust. It was built with a fairly large number of reference (72) and stressed (48) sites. This model is more broadly applicable and can be done with more confidence. As with any multimetric model, however, regular recalibration as more data become available is encouraged.

Using the inclusion by exclusion screening process and having to integrate data from a variety of programs all using different methods and having different accessory data is far from ideal. It restricted the number of reference sites and made the process far more involved. Ideally, the state will continue to build a dataset of comprehensive land cover information for all the sample sites along with a full suite of habitat and chemistry data. The collection of comparable data of these types for all sample sites will allow the state to build a more confident set of reference and known degraded sites for use in recalibrating the MMI models. This process may take several years, but through that process, a recalibration using state data alone and stricter reference and stressed site criteria will be possible.

In the near short term, comprehensive land cover data for Plains sites could be collected to generate a greater number of stressed sites for the Plains and improve those models during recalibration. The lack of a large number of Plains stressed sites seems odd, given the nature of land use and transformation in the Plains. Unfortunately, a lack of data was most responsible for this problem, rather than the condition of streams. Land cover data for each of the sites would allow the development of land cover criteria. By incorporating this information into the reference/stressed site selection process, explicit land cover data would better identify sites with significant land use disturbance and likely increase the number of stressed sites for recalibrating the models.

5.0 Predictive Model Development

Procedures for developing and evaluating RIVPACS models have been well documented and we only describe details germane to the Colorado model here (Clarke et al. 1996, 2002, 2003, Hawkins et al. 2000, Hawkins and Carlisle 2001, Ostermiller and Hawkins 2004, Van Sickle et al 2005)

5.1 Classification and Site Grouping

One hundred and seventy three OTUs were found in the samples collected from the 97 unique reference sites used to build the model. Eighty OTUs were observed in 5 or more samples and were used to create the biotic classification of sites on which the predictive model was based. A classification dendrogram was produced by first calculating all pair-wise Bray-Curtis similarities between samples and then clustering sites with the flexible-beta UPGMA algorithm (McCune and Grace 2002). Ten groups of sites were identified from this dendrogram (Figure 10), of which groups 9 and 10 were distinctly different from the others. Sites in these two groups occurred in the plains or lower-elevation valleys of Colorado (Figure 11). The other groups occurred in upland areas but otherwise showed little geographic clustering with the exception of group 8, sites in which were located along the western base of the Southern Rocky Mountains.

5.2 Discriminant Models

5.2.1 Predictors

We used the all subsets software developed by John Van Sickle of the USEPA (Corvallis, OR) to select the discriminate model that most effectively minimized bias and maximized precision of model predictions. This software evaluates up to 32,767 models based on all possible combinations of 15 or fewer predictor variables. Software output includes the 5 best performing models for each of 1, 2, ….. and 15-order (predictors) models. Performance measures include the mean, standard deviation, and root mean square error of O/E values derived from reference quality samples. These measures are compared with estimates of the error expected if no natural environmental gradients were accounted for (null model) and a theoretically perfect model in which the only error was the random variation expected among replicate samples (see Van Sickle et al. 2005). Ideally, models are evaluated with an independent set of validation samples collected from a range of reference-quality waterbodies. However, the small number of reference sites prohibited such an external validation. All performance measures reported here are based on internal validation in which the original data were run back through the models.

To be most easily derived and used, and to avoid use of variables that could potentially be influenced by human activities, we evaluated only map-derived predictor variables that were likely surrogates for local factors that actually influence the distribution of biota in space and time. Candidate variables included geographic coordinates (latitude and longitude), elevation, watershed area (log transformed), several climatic variables (mean annual precipitation, annual wet days, long-term mean, min, and max air temperature), several geology variables coded to represent different natural capacities to produce nutrients and sediments, major river basin from which the sample was taken, and day of the year the sample was collected.

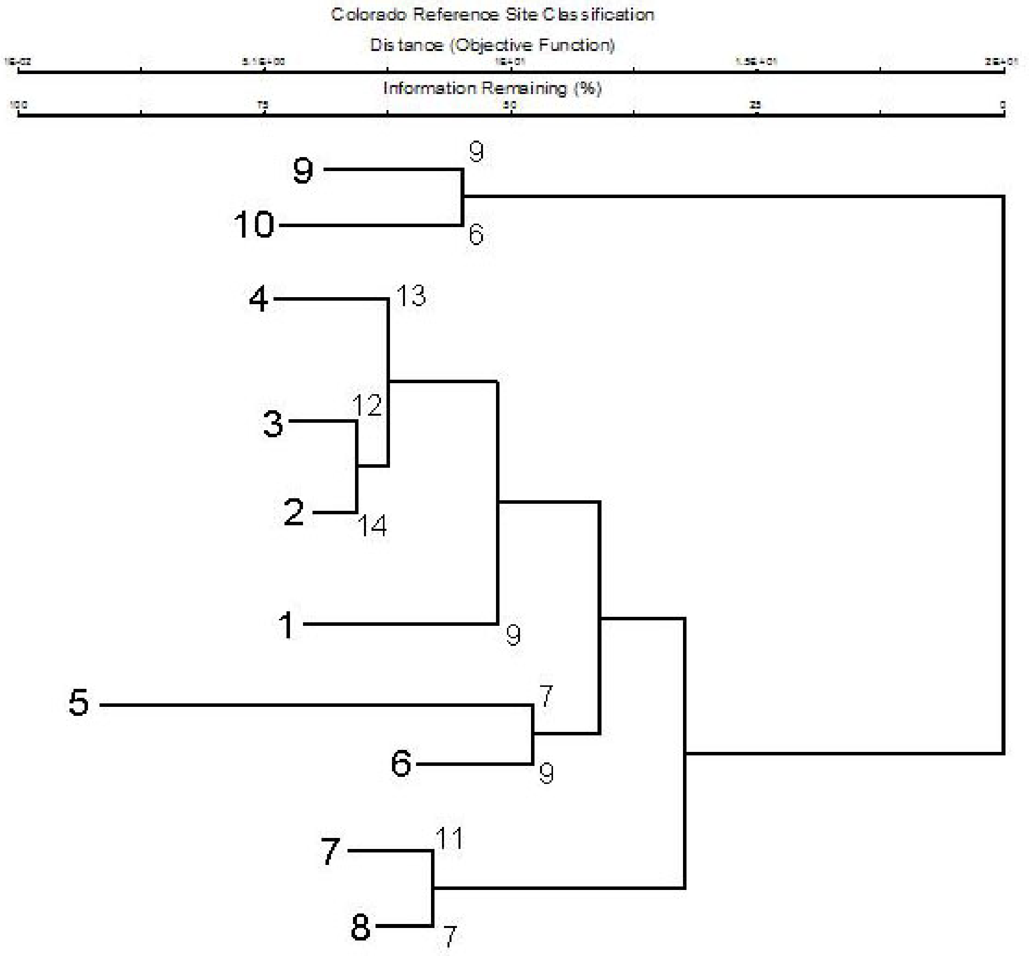

Colorado Reference Site Classification
Distance (Objective Function)
1E-02
5.1E+00
1E+01
1.5E+01
2E+01
Information Remaining (%)
100
75
50
25
0
9
10
9
6
4
13
3
12
2
14
1
9
5
7
6
9
7
11
8
7

Figure 10. Dendrogram showing the 10 site clusters. Numbers at the end of terminal nodes indicate group assignment. Numbers at the corners of branches represent the number of sites in each group. The position of the group node indicates how similar the sites within a group were to one another. The further a node is to the left side of the graph, the more similar sites within a group were to one another.

5.2.2 Final Model Performance

We chose to calculate O and E based on a probability of capture threshold of > 0.5. Use of lower thresholds increase the number of taxa on which assessments are based, but they usually result in increased error (lower precision) associated with the prediction of rare taxa (Hawkins et al. 2000, Ostermiller and Hawkins 2004).

The final model used only 3 predictor variables: longitude (decimal degrees), mean annual air temperature (^{o}C x 10), and log watershed area (km^{2}). Of these variables, mean annual air temperature varied the most among classes (Partial F-values: mean temperature = 9.81, longitude = 4.33, log watershed area = 4.13, Figure 12). The strong relationship between biotic classes and temperature implies that thermal variation across Colorado is the single most important factor affecting the distribution of stream taxa.

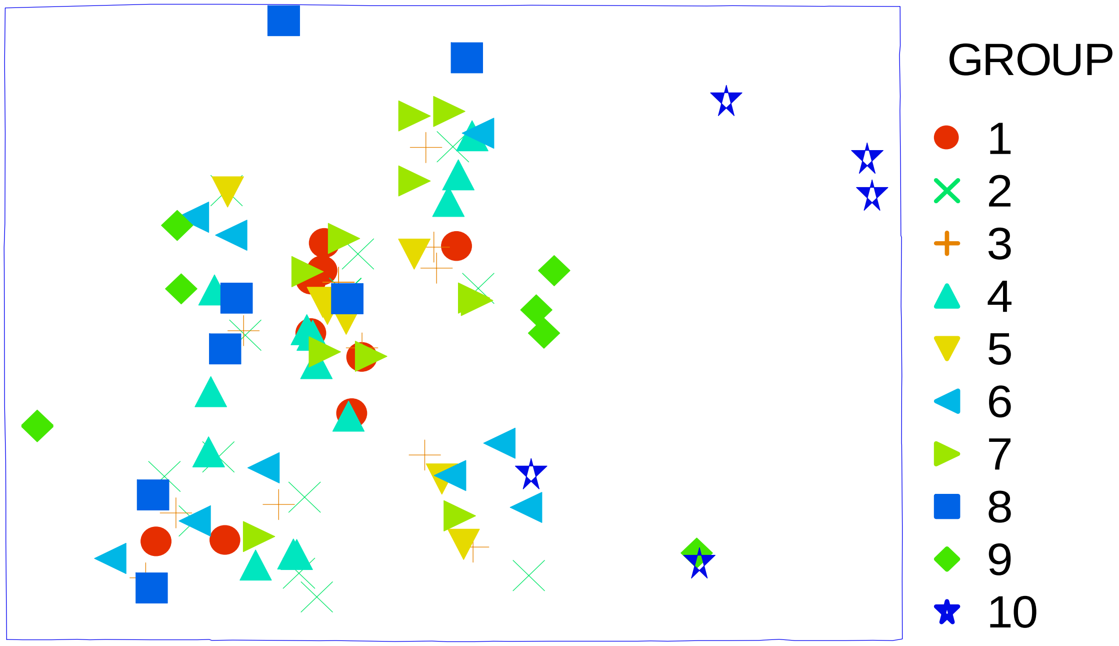


Figure 11. Location of the 97 reference sites within Colorado used in predictive model development and the biotic groups (classes) to which they were assigned. Note the lack of significant spatial clustering of sites in most classes.

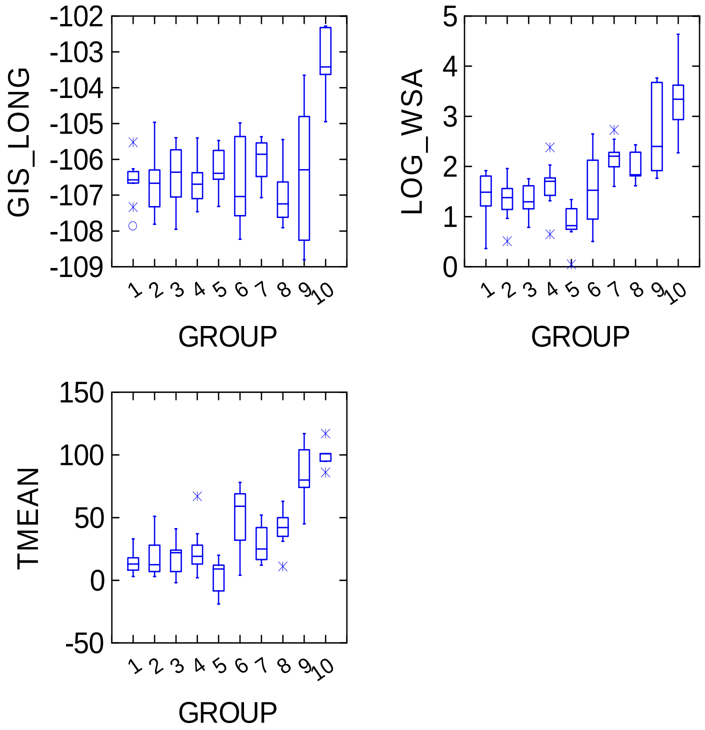


Figure 12. Box-whisker plots showing the variation in longitude, log watershed area, and mean annual temperature among and within the biotically defined classes (groups).

The mean O/E value of the calibration sites was 1.00 and the standard deviation was 0.17. This estimate of error was far better than that associated with the null model (0.29) and the model accounted for about 2/3 of the explainable variability in taxonomic composition among samples (Figure 13).

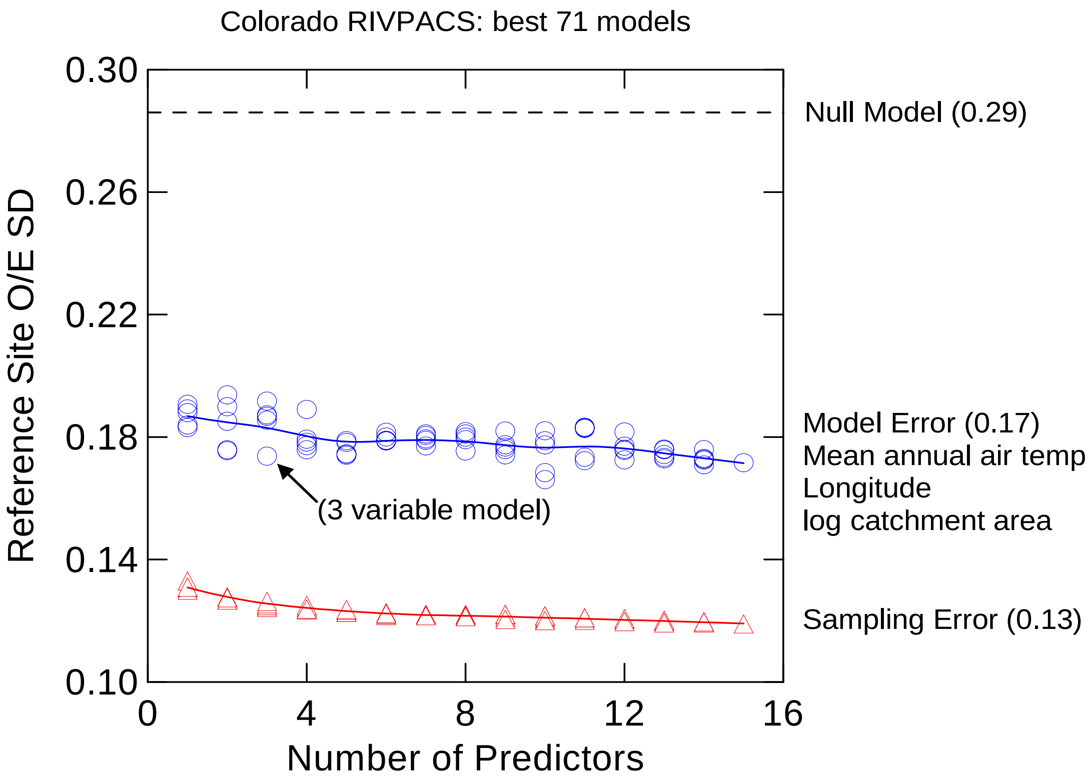


Figure 13. Relationship between model error (SD of reference quality samples) and the number of predictor variables used in 71 models. The maximum possible error is given by the null model and the lowest possible error by the estimate of random sampling error. O and E calculated with a probability of capture threshold of > 0.5.

To be most useful, predictive model assessments need to be accurate as well as precise. In general, the model was accurate in that the slope of the relationship between O and E was not significantly different from 1 (Figure 14) and there was no tendency for the model to over- or under-predict for any of the 10 groups (Figure 15). The model accounted for differences in richness observed both among reference site groups as well as within groups (Figure 14).

The model also showed little evidence that it produced biased predictions for streams that occurred in different regions of Colorado as defined either by major river basin or ecoregion (Tables 10 and 11). The apparent under-prediction of richness for the Arizona/New Mexico Plateau was based on one sample, and such outlier values occur in many data sets.

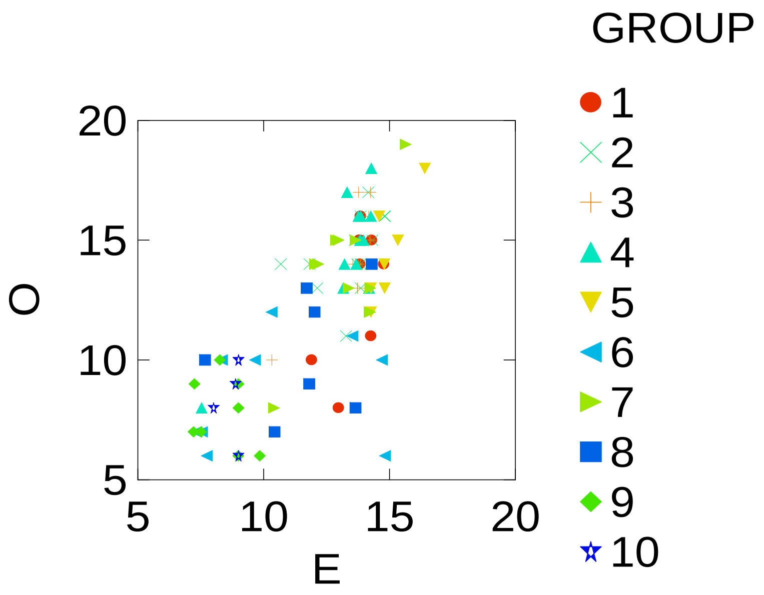


Figure 14. Relationship between observed richness (O) and expected richness (E) at reference sites. The model accounted for 57% of the variation in O and the slope of the relationship was not different from 1. O and E were calculated with a probability of capture threshold of > 0.5.

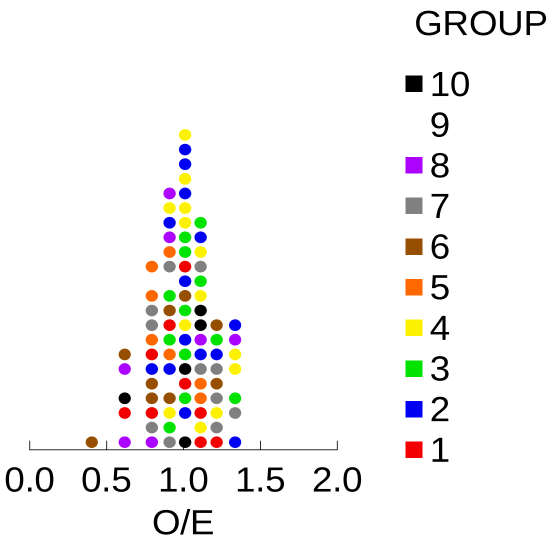


Figure 15. Frequency distribution of reference site O/E values. Sites are color coded based on their class membership. Note that there is no tendency to either over- or under-estimate O/E values based on the biotic class to which sites were assigned. Also note different color scheme than used in Figures 2 and 5

Table 10. Mean O/E values observed for samples taken from reference and non-reference sites in each of four major river basins and across all samples. Sample types are reference (R) and non-reference or test (T). ARB = Arkansas River Basin, CORB = Colorado River Basin, MSRB = Missouri River Basin, and RGRB = Rio Grande River Basin.

	River Basin				
Sample Type	**All**	**ARB**	**CORB**	**MSRB**	**RGRB**
R	1.00	0.98	0.99	1.06	1.04
T	0.73	0.62	0.75	0.77	0.69

Table 11. Mean O/E values observed for samples taken from reference and non-reference sites in each of six ecoregions. WB = Wyoming Basin, CP = Colorado Plateau, SR = Southern Rockies, ANMP = Arizona/New Mexico Plateau, WHP = Western High Plains, and SWTL = Southwest Table Lands.

	Ecoregion					
Sample Type	**WB**	**CP**	**SR**	**ANMP**	**WHP**	**SWTL**
R	-	1.04	1.00	1.30	0.96	0.97
T	0.59	0.76	0.74	0.56	0.76	0.62

5.2.3 O/E Sensitivity

Because RIVPACS models predict how taxa should be naturally distributed across sites, if the models are accurate, the only factor that should affect the sensitivity of assessments is the sensitivity or tolerance of the taxa in the region to the stressors that exist. Because the OTUs we used in the models generally represent relatively coarsely resolved taxa (e.g., many genera, some families, few species), these assessments will be conservative with respect to what we would see with models based on species-level data (Hawkins et al. 2000).

In spite of the fact that OTUs generally represented groupings of more than one species (and thus the response of sensitive species to stress could be masked by less sensitive species lumped in the OTU), application of the model to 741 samples taken from non-reference sites showed that on average these sites have lost a substantial number of taxa (Tables 10 and 11). The mean O/E value for all non-reference samples was less than ¾ than expected (0.73), and mean values varied between 0.56 and 0.77 depending on ecoregion and river basin.). Of the 741 non-reference samples assessed, 39% had O/E values > 0.8 (least impaired), 45% had values > 0.5 and < 0.8 (moderately impaired), and 16% had values < 0.5 (severely impaired).

Examination of the spatial distribution of O/E values showed that low, medium, and high values of O/E tended to occur in all regions of the state but that some localized clumping of degradation was apparent (Figure 16).

In general, the performance of the Colorado RIVPACS model is comparable to or better than most models in use in the USA and elsewhere. The fact that the model makes good predictions from just three easily derived map-based predictor variables means it will be easy to implement by most CDPHE staff. In spite of the paucity of reference sites in lower elevation regions of

Colorado, the model appeared to be surprisingly robust in those regions. Future refinements of the model with data collected from additional reference sites should only improve our confidence in assessments based on this approach.

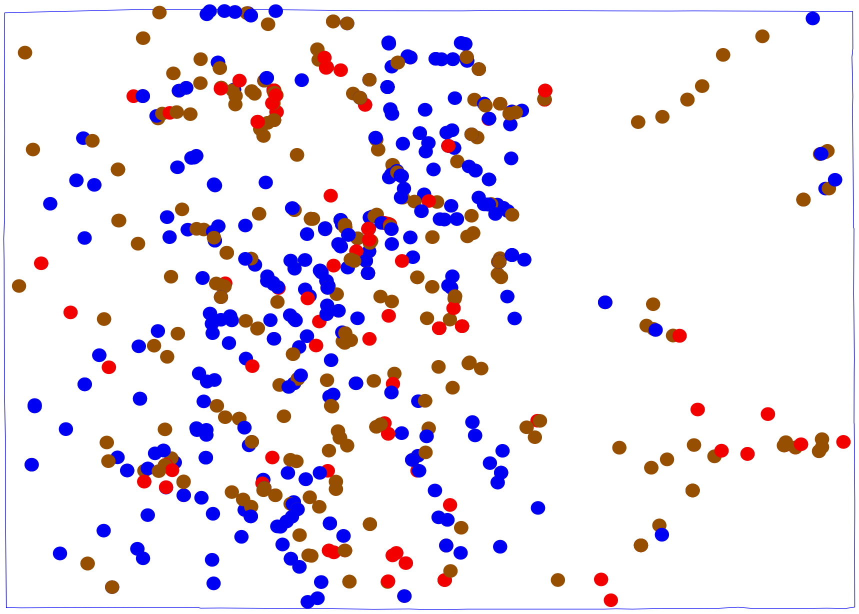


Figure 16. Distribution of least degraded (O/E > 0.8), moderately degraded (O/E < 0.8 and > 0.5), and severely degraded (O/E < 0.5) samples within Colorado.

6.0 Literature Cited

Barbour, M. T., J. B. Stribling, and J. R. Karr. 1995. The multimetric approach for establishing biocriteria and measuring biological condition. Pp. 63-76 in W. S. Davis and T. P. Simon, editors. Biological Assessment and Criteria: Tools for Water Resource Planning and Decision Making. Lewis Publishers, Ann Arbor, Michigan.

Barbour, M. T., J. Gerritsen, G. E. Griffith, R. Frydenborg, E. McCarron, J. S. White, and M. L. Bastian. 1996. A framework for biological criteria for Florida streams using benthic macroinvertebrates. Journal of the North American Benthological Society 15(2):185-211.

Barbour, M. T., J. Gerritsen, B. D. Snyder, and J. B. Stribling. 1999. Rapid Bioassessment Protocols for Use in Streams and Wadeable Rivers: Periphyton, Benthic Macroinvertebrates and Fish. Second Edition. EPA/841-B-99-002. U.S. Environmental Protection Agency, Office of Water, Washington, DC.

Burton, J. and J. Gerritsen. 2003. A stream condition index for Virginia non-coastal streams. Prepared for U.S. Environmental Protection Agency, Office of Water, Region 3 Environmental Services Division, and Virginia Department of Environmental Quality. Tetra Tech, Inc. Owings Mills, MD.

Clarke, R.T., M.T. Furse, J.F. Wright, and D. Moss. 1996. Derivation of a biological quality index for river sites: comparison of the observed with the expected fauna. Journal of Applied Statistics 23:311-332.

Clarke, R.T., M.T. Furse, R.J.M. Gunn, J.M. Winder, and J.F. Wright. 2002. Sampling variation in macroinvertebrate data and implications for river quality indexes. Freshwater Biology 47:1735-1751.

Clarke, R.T., J.F. Wright, and M.T. Furse. 2003. RIVPACS models for predicting the expected macroinvertebrate fauna and assessing the ecological quality of rivers. Ecological Modeling 160:219-233.

Frey, D. G. 1977. Biological integrity of water – an historical approach. Pages 127-140 in R. K. Ballantine and L. J. Guarraia (editors). The Integrity of Water. Proceedings of a Symposium, March 10-12, 1975, U. S. Environmental Protection Agency, Washington, DC.

Gibson, G. A., M. T. Barbour, J. B. Stribling, J. Gerritsen, and J. R. Karr. 1996. Biological criteria: Technical guidance for streams and rivers. EPA/822-B-94-001. U. S. Environmental Protection Agency, Office of Science and Technology, Washington, DC.

Hawkins, C.P. and D.M. Carlisle. 2001. Use of predictive models for assessing the biological integrity of wetlands and other aquatic habitats. Bioassessment and management of North American Wetlands Pages 59-83 in R.B. Rader, D.P. Batzer. John Wiley & Son, New York.

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.

Hawkins, C.P. and R.H. Norris. 2000. Effects of taxonomic resolution and use of subsets of the fauna on the performance of RIVPACS-type models. Pages 217-228 in J.F. Wight, D.W. Sutcliffe, and M.T. Furse, editors. Assessing the biological quality of fresh waters: RIVPACS and other techniques. Freshwater Biological Association, Ambleside, Cumbria, UK.

Hilsenhoff, W. L. 1987. An improved biotic index of organic stream pollution. Great Lakes Entomologist 20:31-39.

Karr, J. R. 1991. Biological integrity: A long-neglected aspect of water resource management. Ecological Applications 1:66-84.

Karr, J. R., and E.W. Chu. 1999. Restoring life in running waters: Better biological monitoring. Island Press, Washington , DC.

Karr, J. R. and D. R. Dudley. 1981. Ecological Perspectives on water quality goals. Environmental Management 5:55-68.

Karr, J. R., K. D. Fausch, P. L. Angermeier, P. R. Yant, and I. J. Schlosser. 1986. Assessment of biological integrity in running waters: A method and its rationale. Illinois Natural History Survey, Champaign, Illinois. Special Publication 5.

Kerans, B. L. and J. R. Karr. 1994. Development and testing of a benthic index of biotic integrity (B-IBI) for rivers of the Tennessee Valley. Ecological Applications 4(4): 768-785.

McCune, B., and J.B. Grace. 2002. Analysis of ecological communities. MjM Software Design, Gleneden Beach, OR.

Merritt, R.W. and K.W. Cummins. 1996. An introduction to the aquatic insects of North America. 3rd Edition, Kendall/Hunt Publishing Company, New York. 862p.

Ohio Environmental Protection Agency (Ohio EPA). 1989. Addendum to biological criteria for the protection of aquatic life, Volume II: Users manual for biological field assessment of Ohio surface water. Ohio EPA, Division of Water Quality Planning and Assessment, Ecological Assessment Section, Columbus, OH.

Ostermiller, J.D. and C.P. Hawkins. 2004. Effects of sampling error on bioassessments of stream ecosystems: application to RIVPACS-type models. Journal of the North American Benthological Society 23:363–382.

Plafkin, J.L, M.T. Barbour, K.D. Porter, S.K. Gross, and R.M. Hughes. 1989. Rapid Bioassessment Protocols for Use in Streams and Rivers: Benthic Macroinvertebrates and Fish.

EPA/440/4-89-001. U.S. Environmental Protection Agency, Office of Water, Washington, DC. http://www.epa.gov/owow/monitoring/rbp.

Southerland, M. T. and J. B. Stribling. 1995. Status of biological criteria development and implementation. Pages 81-96 in W. S. Davis and T. P. Simon, editors, Biological Assessment and Criteria: Tools for Water Resource Planning and Decision Making. Lewis Publishers, Boca Raton, FL.

Stribling, J. B., B. K. Jessup, and J. Gerritsen. 2000. Development of biological and physical habitat criteria for Wyoming streams and their use in the TMDL process. Tetra Tech. Prepared for the U.S. EPA Region 8, Denver, CO.

Van Sickle, J., C.P. Hawkins, D.P. Larsen, and A.H. Herlihy. 2005. A null model for the macroinvertebrate assemblage expected in unimpaired streams. Journal of the North American Benthological Society 24(1):178-191.