Running Head: Assessing reference network performance

Title: An approach for evaluating the suitability of a reference site network for the ecological assessment of streams in environmentally complex regions

Authors: Peter Ode, Andrew Rehn, Raphael Mazor, Kenneth Schiff, Eric Stein, Jason May, Larry Brown, David Gillett, Kevin Lunde and David Herbst

Peter Ode[1], Andrew Rehn[1], Raphael Mazor [1,2], Kenneth Schiff[2], Eric Stein[2], Jason May[3], Larry Brown[3] , David Gillett[2] , Kevin Lunde[4] and David Herbst[5]

[1] Aquatic Bioassessment Laboratory, California Department of Fish and Wildlife, 2005 Nimbus Road, Rancho Cordova, CA 95670

[2]Southern California Coastal Water Research Project, 3535 Harbor Blvd., Suite 110, Costa Mesa, CA

[3]United States Geological Survey, 6000 J Street, Sacramento, CA 95819

[4]San Francisco Bay Regional Water Quality Control Board, 1515 Clay Street, Oakland, CA 94612

[5]Sierra Nevada Aquatic Research Laboratory, 1016 Mt. Morrison Road, Mammoth Lakes, CA 93546

Abstract:

The definition of reference conditions is now widely accepted as an essential element of stream bioassessments. Many of the advances in this field have focused on approaches for objectively selecting reference sites, but much less emphasis has been placed on evaluating the suitability of the reference network for its intended application(s). We present an approach for evaluating the suitability of a reference network for supporting biological integrity scoring tools in environmentally heterogeneous and pervasively altered regions. We screened 1,985 candidate stream reaches to create a 590 site reference network for perennial wadeable streams in California, USA. We first characterized all sites in terms of their natural environmental characteristics and potential sources of anthropogenic stress. We then used non-biological screening metrics and criteria to select reference sites following standard approaches. We assessed the resulting set of reference sites against two primary performance criteria. First, we evaluated natural environmental representativeness with univariate and multivariate comparisons of the range of environmental conditions in the reference network to the full range of these gradients found in the region. Second, we evaluated the degree to which we minimized the influence of anthropogenic stress by: a) measuring the reduction of sources of biological variance associated with human activity and b) comparing biological metric scores at a subset of reference sites that would have passed very strict screens to those of passing sites that had higher levels of human activity. Using this approach, we demonstrated strong coverage of environmental heterogeneity as well as low levels of anthropogenic stress in the reference network, indicating that we did not sacrifice biological integrity in order to achieve adequate environmental representation. This approach should be widely applicable and easily customizable to particular regional or programmatic constraints.

Key Words: reference condition, bioassessment, environmental heterogeneity, performance measures, benthic macroinvertebrates

Introduction

The worldwide use of biological indicators in water quality monitoring programs has evolved rapidly in the last 30 years (Rosenberg and Resh 1993, Gibson et al. 1996, Wright et al. 2000, Bonada et al. 2006, Collier 2011, Pardo et al. 2012). Many of the refinements to biological monitoring techniques over this period have centered on strengthening the theoretical and practical basis for predicting the biological expectation for sites with low levels of human-derived disturbance, the “reference state” or “reference condition” (Hughes et al. 1986, Reynoldson et al. 1997, Stoddard et al. 2006, reviewed by Bonada et al. 2006, Hawkins et al. 2010a and Dallas 2012). As a result, the need to anchor biological expectations to a reference state is now widely regarded as highly desirable: to the extent possible, the expected biological state of a monitoring site should be based on the biological state observed at sites having similar environmental settings, but low levels of human disturbance.

Although early efforts to use a reference condition approach often relied on subjective criteria and best professional judgments (e.g., Wright et al. 1984, Hughes et al. 1986, Barbour et al. 1995, 1996, Reynoldson et al. 1995, 1997, Rosenberg et al. 1999), most recent treatments of the subject recognize that objective criteria can greatly enhance the defensibility of reference condition determinations (Whittier et al. 2007, Yates and Bailey 2010). Examples of objective site selection are increasingly common (e.g., Stoddard et al. 2006, Collier et al. 2007, Sanchez-Montoya et al. 2009 Whittier et al. 2007, Yates and Bailey 2010). A robust approach to selecting reference sites in environmentally complex landscapes should account for a variety of potential stressor types as well as natural sources of disturbance and variation. However, multiple criteria can complicate the achievement of uniform reference definitions in such complex regions (Statzner et al. 2001, Herlihy et al. 2008, Mykrä et al. 2008, Ode et al. 2008, Ode and Schiff 2009).

Because truly pristine streams are rare or non-existent throughout the world, programs that measure biological integrity typically use a ”minimally-disturbed” or “least-disturbed” standard for selecting reference sites (*sensu* Stoddard et al. 2006). The main challenge is to choose selection criteria that retain sites with high biological integrity and thus maintain the philosophical integrity of the reference condition approach. This involves balancing two potentially conflicting demands: 1) reference criteria should select sites that uniformly represent the least disturbed conditions throughout the region of interest, and 2) reference sites should represent stream types from the full range of environmental settings in the region and in adequate numbers to cover all habitats of interest for assessment. Because meeting the second demand usually requires at least some loosening of reference screening criteria, reference site selection becomes an exercise in balancing the risk of allowing some disturbed sites in the reference network (decreased naturalness) versus unnecessarily rejecting minimally disturbed sites from under-represented stream types (decreased representativeness).

In a perfect world with a large number of undisturbed streams of all types, we could focus exclusively on avoiding contamination of the reference pool with biologically-impaired sites. However, overly restrictive criteria can result in under-representation of important natural gradients, particularly regions with diverse natural conditions (Mapstone 2006, Osenberg et al. 2006, Yuan et al. 2008, Dallas 2012). Thus, excessive rejection of candidate sites can reduce the performance (i.e., accuracy and precision) of scoring tools. This is especially critical in regulatory applications where errors in site specific accuracy can have significant financial and resource protection consequences. Evaluating the performance of reference criteria allows scientists and resource managers to make informed decisions about this balance.

This paper outlines the use of an approach we created to measure the robustness of a reference site network in California, an environmentally complex region of the USA overlain with large areas of pervasive development. This reference network was established as the foundation of a statewide biological integrity scoring tool that had high site-specific assessment accuracy (Mazor et al. in prep). This work built on previous efforts to identify reference conditions in similarly complex regions (e.g., Collier et al. 2007, Sánchez-Montoya et al. 2009, Falcone et al. 2010, Yates and Bailey 2010). We drew on these efforts to identify an initial suite of stressor screens and thresholds, expanded them to accommodate a broad array of anthropogenic activities known to be important in California (Gillett et al. *in prep*), then evaluated the degree to which we met our objectives.

Methods

A set of 1,985 candidate sites with bioassessment, habitat and water chemistry data and which represented a wide range of stream types was assembled to support the development of screening criteria. Site selection was restricted to wadeable, perennial streams, although some sites were included in the screening pool that were non-wadeable or non-perennial. Each site was characterized with a suite of landuse and landcover metrics that quantified both its natural characteristics and potential anthropogenic stressors near the site or in its upstream drainage basin. Sites were then screened with a subset of metrics using thresholds that represented low levels of anthropogenic stress ("least disturbed" *sensu* Stoddard et al. 2006). Finally, the pool of passing reference sites was evaluated to assess whether the objectives of balancing naturalness and representativeness were achieved to a degree sufficient to support defensible biological scoring tools and condition thresholds (i.e., biocriteria).

Setting

California's stream network is approximately 280,000 km long according to the NHD medium resolution (1:100k) stream hydrology (approximately 30% of which is perennial) and drains a large (424,000 km²) and remarkably diverse landscape. Spanning latitudes between 33° and 42° (N), California's geography is characterized by its extremes. California boasts both the highest and lowest elevations in the continental US and its ecoregions range from temperate rainforests in the Northwest to deserts in the Northeast and Southeast, with the majority of the state having a Mediterranean climate (Omernik 1987). California's geology is also complex, ranging from Coast Ranges comprised of recently uplifted and poorly consolidated marine sediments, broad internal valleys to granitic batholiths along the eastern border to recent volcanism in the northern mountains. This geographical diversity is associated with a high degree of biological diversity and endemism in the stream fauna (Erman 1996, Moyle et al. 1996, Moyle and Randall 1996). California's natural diversity is further complicated by an equally complex pattern of land use. The native landscapes of some regions of the state have been nearly completely converted to agricultural or urban land uses (e.g., the Central Valley, the San Francisco Bay Area and the South Coast) (Sleeter et al. 2011). Other regions are still largely natural but contain pockets of agricultural and urban land use and also support timber harvest, livestock grazing, mining and recreational uses. To facilitate data evaluation, the state was divided into six regions based on modified ecoregional (Omernik 1987) and hydrological boundaries (Figure 1).

<u>Aggregation of site data</u>

More than 20 federal, state, and regional monitoring programs were inventoried to assemble data sets used for screening reference sites. All unique sites sampled between 1999 and 2010 were aggregated into a single database (Figure 1). From the population of > 10,000 California sites with bioassessment data, sites were prioritized for inclusion if they had benthic macroinvertebrate data available and met at least one of two criteria: 1) they were reasonably likely to pass screening thresholds (e.g., ones identified as reference in previous biological integrity index development in California), 2) they were sampled under probabilistic survey designs. Randomly selected probability sites served several functions in this effort: they helped ensure coverage of the full range of stream types in the state, they were used to infer the full range of natural gradients in different regions of the state, and a large proportion of the probability sites were also good reference candidates. When multiple programs sampled identical candidate sites or sites in close proximity (within 300 m), data were treated as a single site to minimize redundancy.

Assembled data included benthic macroinvertebrate (BMI) taxa lists, water chemistry and physical habitat characteristics. Field protocols often varied among programs and not all programs collected all data types, but most analytes were available for most sites (Tables 1, 2). The majority of BMI data were collected using the reachwide protocol of the US EPA's Environmental Monitoring and Assessment Program (EMAP, Peck et al. 2006), but some of the older data were collected with targeted riffle proocols. Previous studies have documented that these protocols are generally compatible (Ode et al. 2005, Gerth and Herlihy 2006, Herbst and Silldorff 2006, Rehn et al. 2007). BMI taxa lists were standardized for analyses (metrics and ordinations and variance partitioning) with a database that converted all taxonomic data to conform to California's standard taxonomic effort levels (SAFIT 2011), generally genus-level identifications with chironomid midges identified to subfamily.

For calculation of local scale physical habitat metrics, preference was given to programs that used quantitative field protocols (e.g., Peck et al. 2006, Ode 2007) and allowed calculation of quantitative reach-scale habitat condition variables defined by Kaufmann et al. (1999).

<u>Integration of probability data sets</u>

A subset of the data set collected under probabilistic survey designs (919 sites) was used to evaluate whether our final pool of reference sites adequately represented the full range of natural stream settings occurring in California. Probability datasets provide objective statistical estimates of the true distribution of characteristics of a population (in this case, natural characteristics of California's perennial stream network) (Stevens and Olsen 2004). Data from 10 probabilistic surveys were combined for this effort. Although most surveys had similar design characteristics, they were different enough to require synchronization before they could be integrated. First a common sample frame was created so that the relative contribution of each site to the overall distribution could be calculated for each site in the combined data set. All probabilistic sites were registered to a uniform stream network (National Hydrography Database - NHD 1:100,000), which was attributed with strata defined by the design parameters of all integrated programs (e.g., land use, stream order, survey boundaries, etc.). Weights were calculated for each site by dividing total stream length in each stratum by the number of site evaluations in that stratum. All weight calculations were conducted using the *spsurvey* package (Kincaid and Olsen 2009) in R v 2.11.1 (The R Foundation for Statistical Computing 2010). These weights were used to estimate regional distributions for environmental variables using the Horvitz-Thompson estimator (Horvitz-Thomson 1952). Confidence intervals were based on local neighborhood variance estimators (Stevens and Olsen 2004).

<u>GIS data and metric calculation</u>

A large number of spatial data sources were assembled to characterize natural and anthropogenic gradients that may affect biological condition at each site, such as land cover and land use, road density, hydrologic alteration, mining, geology, elevation and climate (Table 1). Data sets were evaluated for statewide consistency and layers with poor or variable reliability were excluded. All spatial data sources were publicly available except for the roads layer, which was customized for this project by appending unimproved and logging roads obtained from the United States Forest Service and California Department of Forestry and Fire Protection to a base roads layer (TeleAtlas 2009).

Land cover, land use and other measures of human activity were quantified into metrics (Table 2) that were calculated at three spatial scales: within the entire upstream drainage area (watershed), within 5 km upstream and within 1 km upstream. Polygons defining these spatial analysis units were created using ArcGIS tools (ESRI 2009). Upstream watershed polygons were aligned to NHD polygons and the downstream portion of each watershed was adjusted with standard flow direction and flow accumulation techniques using 30 m digital elevation models (National Elevation Dataset). The local (5k and 1k) scales were created by intersecting a 5km or 1km radius circle with the primary watershed polygon. Site metrics associated with each sampling location also were calculated based on each site's latitude and longitude (e.g., mean annual temperature, elevation, NHD+ attributes, etc.).

Selection of screening metrics and thresholds

A primary set of screening metrics was selected based on land use frequently associated with impairment to the biological integrity in streams and rivers. The specific metrics and thresholds were initially identified from a combination of prior reference development (Ode et al. 2005; Rehn et al. 2005, Stoddard et al. 2006, Rehn 2008) or values obtained from literature (e.g., Collier et al. 2007, Angradi et al. 2009, Falcone et al. 2010). This initial list was augmented after examining the distribution of stressors in watersheds in California (Gillett et al. *in prep*). Stressor values representing least disturbed conditions were used to setting thresholds for metrics or particular spatial scales (e.g., 1k or 5k) that lacked published values.

A set of secondary thresholds was established to further refine reference site selection. In contrast to our primary screens, secondary thresholds were not chosen to minimize the influence of anthropogenic stressors but to eliminate sites with other sources of disturbance that were not eliminated by primary metrics. Secondary thresholds were applied in the same manner as primary screens but were intentionally set at higher values: 1) for land use at the watershed scale because distant disturbance generally has less impact on biological condition than near-site disturbance (Munn et al. 2009), and 2) for number of upstream road crossings because inaccuracies in GIS layers (specifically, the line work that forms stream networks and road layers) make this metric difficult to quantify accurately.

Exploration of metric thresholds

Regions often vary in the relative dominance of different types of stressors. Thus, the relative contribution of these to overall disturbance at candidate sites also varies regionally. To explore regional differences in reference site selection and the degree of inter-correlation of stressor metrics, thresholds for each primary metric were adjusted individually while all others were held constant and the number of passing sites (i.e., threshold sensitivity) was plotted for each region. This gave us a measure of among-regional differences in the number of reference sites that could be gained by relaxation of individual screening criteria. Examination of these partial-dependence curves was used to evaluate the number of reference sites that could be gained by relaxing thresholds for each screening metric in each region.

<u>Performance Measures</u>

Evaluation of reference network representativeness

Evaluations focused on two properties: 1) the number of reference sites identified, both statewide and within major regions of California (i.e., adequacy, Diamond et al. 2012), and 2) the degree to which those reference sites represented the range of natural variability in California streams (i.e., environmental representativeness).

The robustness of the reference site density for developing biological integrity indices was first assessed by counting the number of reference sites statewide and within major sub-regions. A target minimum number of sites was not set, but if low numbers of reference sites were available in a given region, these regions might need to be aggregated with similar regions or excluded from subsequent reference-based analyses.

Because geographic representation alone is not sufficient for evaluating representativeness, we also compared the distribution of reference sites against important natural gradients, both individually and with multivariate gradients identified by principal components analysis (PCA). All the natural gradients listed in Table 2 were used in the PCA analysis except the three atmospheric deposition variables (AtmCa, AtmMg, AtmSO4). Additionally, predicted conductivity (Olson and Hawkins 2012) was also used. Because geographic patterns obscure the distribution of these gradients at reference sites, locational variables (i.e., latitude, longitude, and elevation) were excluded from analysis, and residuals of gradients of interest were used in the PCA instead of raw variables.

Evaluation of sources of variance in the reference network

Because all thresholds allowed at least some degree of upstream disturbance (i.e., none were pristine), responsiveness of representative biological metrics to disturbance levels allowed by our screens was evaluated in three ways. First, the variance in BMI metrics explained by the residual levels of disturbance that remained in reference sites was compared to the variance explained within the overall data set to examine the extent to which reference thresholds minimized the impact of major stressors. If Pearson's R^2 was < 0.1 for correlations between individual stressors and BMI metrics at reference sites, the biological response to disturbance levels below reference thresholds was considered to be negligible and thresholds were considered to be adequately protective of biological integrity. Second, variance partitioning was used to evaluate the residual effects of stress on benthic macroinvertebrates at reference sites. Taxonomic identifications were converted to operational taxonomic units, subsampled to 400, and converted to presence-absence data. Then, variance partitioning analysis was then performed using the *varpart* function in the *vegan* package in R (Oksanen et al. 2012) to estimate the proportion of the variance attributable to natural variables, stressor variables, and their interaction. All the variables in Table 2 were included in this analysis. The amount of variance explained by stress in the full data set was compared to the amount explained in the subset of reference calibration sites.

Although the use of biological data in the process of selecting screening metrics and thresholds was deliberately avoided, biological metric values in reference sites affected the least amount of stressors were compared to those in passing sites that had more disturbance. Because the biological metric values indicative of healthy biological condition vary in different environmental settings, metric values were adjusted for major natural gradients by using residuals of random forest models of natural gradients as the response variable instead of the raw metric values. Equivalent metric scores in the more stressed and less stressed reference groups would be considered evidence that biological integrity was maintained.

Results

Reference status by region

Of the 1,985 sites evaluated for potential use as reference sites, 590 passed our screening thresholds (Table 4). The number of reference sites varied by region, with highest concentrations in mountainous regions (e.g., the Sierra Nevada, the North Coast and South Coast Mountains), which also contain the majority of the state's perennial stream length (NHD). Lower elevation, drier sub-regions generally had few reference sites (South Coast Xeric = 33, Interior Chaparral = 32), and only a single reference site was identified in the Central Valley.

Based on sampling weight estimates from the probability data, 29% (± 2% standard error) of California's stream-length was estimated to meet our reference criteria (Table 5). Reference quality streams were predominant in mountainous regions, comprising approximately 76% and 53% of the stream length in the Central Lahontan and South Coast Mountain regions, respectively. Only 2-3% of stream length in the Central Valley and the South Coast Xeric regions were estimated to be in reference, whereas 43% and 32 of the Sierra Nevada and Deserts / Modoc stream length met our reference criteria, respectively. Despite the large number of reference sites in the North Coast, only 26% of North Coast stream length is estimated to meet reference criteria (similar to levels seen in Chaparral regions), suggesting that the abundance of reference sites in the North Coast is due more to the overall large extent of streams than the lack of anthropogenic stressors in the region.

Threshold sensitivity

There were strong regional differences in the number and types of stressor metrics that contributed to the removal of individual candidate sites from the reference pool (Table 4). For example, whereas most non-reference sites in the Sierra Nevada and the South Coast Mountains failed only one or two metrics (typically road density and Code 21), a large majority (i.e., > 85%) of non-reference sites in the Central Valley and the South Coast Xeric regions failed five or more metrics. The other regions had intermediate failure rates. 44% of Chaparral sites were rejected on the basis of only one or two stressors (most typically road density), whereas 39% of Chaparral sites failed 5 or more criteria. The majority of non-reference North Coast sites (57%) failed 3 to 5 criteria and Desert – Modoc sites were generally less stressed than Chaparral sites, with most 51% of sites failing only one or two criteria.

Related patterns were reflected in threshold sensitivity plots (Figure 2), where the number of passing sites was plotted as a function of changing stressor thresholds using four example metrics. Adjusting thresholds for the two landuse metrics (% agricultural land and % urban landuse) had little influence on the number of sites that passed reference screens in most regions, indicating that other metrics were limiting or co-limiting in all regions. This pattern was common for most metrics. In contrast, the metrics Road Density and Code21 (an NLCD landcover class closely associated with roadside and urban vegetation) were distinctly sensitive to changing thresholds. Even modest relaxation of thresholds for these metrics resulted in increased numbers of sites passing our reference screens in most regions. For road density, this was true for all regions, but especially the North Coast and Chaparral. For Code 21, this was true for the North Coast, Chaparral and South Coastal Mountains. We took advantage of this sensitivity to increase the screening thresholds for road density and Code21 and thereby increased the number of sites in several regions, improving a critical shortage in the Interior Chaparral. Thus, slight relaxation of the statewide screening thresholds for these two metrics allowed us to significantly improve the representation of sites in several regions, whereas we would have had to adjust many other metric thresholds concurrently to achieve a comparable result.

Reference site representativeness

The large number of sites in our probability data set (919 sites) allowed us to produce well-resolved distribution curves for a suite of natural gradients in each region (Figure 3 illustrates several examples of biologically-important gradients). For nearly all of the natural gradients and regions we examined, the distribution of reference sites was a very good match to the overall distribution of gradients in most regions of the California, with a few exceptions. Very large (i.e., > 500 km^2) watersheds were under-represented, but most of these sites were from non-wadeable rivers, which were not part of the scope of this effort. Very high elevation streams (i.e., > 3,000 m) may also be under-represented. Most of the other minor gaps were associated with a class of streams that represented the tails of distributions for several related environmental variables (low elevation, low-gradient, low precipitation, large watersheds). Gaps were most conspicuous for nearly all gradients in regions with few reference sites (i.e., the Central Valley and Deserts / Modoc), but these examples represented minor exceptions to the overall high degree of concordance between the reference and overall distributions.

Multivariate analysis (PCA) also showed that the reference sites represented natural gradients well (Figure 6), as there were few identifiable gaps in ordination space. Gaps were generally restricted to the extremes of the gradients. For example, investigation of the first two axes (Figure 6) identified a cluster of sites in the upper-left part of the graph, corresponding to large river sites with the largest watersheds. Sparse coverage in the upper-right of the graph corresponds to sites receiving little rainfall, where perennial streams are predominantly a product of urban or agricultural runoff.

Biological response to stressors

Nearly all stressors investigated had negative relationships with selected bioassessment metrics when evaluated against the full screening data set of 1,985 sites (see examples in Figure 5). However, these relationships were always weaker (and frequently absent) when only reference sites were examined (Figure 4). Variance partitioning indicated that much of the variance in BMI taxa at reference sites (87%) was not associated with either natural or stressor gradients used in the analysis (Table 6). Although the 13% explained is appears low, it is similar to other numbers reported for regional factors from similar analyses (e.g., Sandin and Johnson 2004). Of the explained fraction, 76% was attributable to pure natural sources, 13% to pure stressors, and 11% to their interaction, for a total of 23% explained by stress. In contrast, although the amount of total variance attributable to natural and stress gradients was the same in the total dataset, the interaction term increased greatly (from 1% to 6%), suggesting that the influence of stress was reduced in the reference data set in particular environmental settings.

Reduction of the effects of residual stress was even more strongly evident when bioassessment metrics were analyzed. The amount of biological variance in our reference sites explained by various stressors (as contrasted to the variance in the whole dataset) is a demonstration of the amount of residual anthropogenic impairment in our reference pool (Figure 5). Although reference thresholds did not completely eliminate the influence of disturbance on biological metrics in our reference pools, this influence was greatly reduced across all the metrics we evaluated. Furthermore, thresholds successfully reduced the influence of stressors that were not specifically included in reference screens, such as percent sand and fines, presumably because these stressors are associated with other stressors included in screens (Figure 5). The low amount of biological variability in our reference network that was associated with anthropogenic sources indicates that we did not sacrifice a significant amount of biological integrity in order to achieve adequate natural gradient representation.

Biological metric scores evaluated at reference sites with different levels of stress were nearly indistinguishable from each other (all comparisons were not significant at Bonferroni-adjusted p-values of 0.01), implying that reference sites with lowest disturbance levels did not have higher biological quality than the remainder of reference sites.

Discussion

As the focus of water quality monitoring programs shifts toward greater emphasis on ecological condition (Rosenberg and Resh 1993, Davies and Jackson 2006, Collier 2011, Pardo et al. 2012), reference concepts can enhance multiple components of watershed management programs, including non-biological endpoints. To ensure optimal use of reference condition - based tools, programs need to evaluate whether selection criteria produce a set of reference sites that are suitable for the intended uses of the reference network (Bailey et al. 2004, 2012). Although programs developing and using reference sites networks traditionally tend to focus on minimizing degradation of reference site quality, representativeness may be just as important a performance criterion for many applications. In particular, we argue that explicit attention to environmental representativeness could help improve overall accuracy of condition assessments and reduce prediction bias (see Hawkins 2010a) in all reference applications.

<u>Performance summary</u>

Our reference thresholds yielded an unexpectedly large data set, with 590 unique reference sites distributed throughout California. With the exception of one major region of the state, the Central Valley, sites in the reference pool represent nearly the full range of all the natural gradients we evaluated. Thus, we have confidence that analyses and assessment tools developed from this reference data set are valid for the vast majority of perennial streams in California. Although our thresholds did not eliminate all anthropogenic disturbances from the pool of reference sites, we demonstrated that the influence of these disturbances on the reference pool fauna has been greatly minimized, suggesting that impacts on ecological integrity are likely to be small or negligible. Furthermore, although we anticipated that we might need to make regional adjustments in either the choice of stressors or specific thresholds used for screening reference sites, we were able to achieve adequate reference condition representation for most regions of the state with a common set of stressors and thresholds, maintaining inter-regional comparability (i.e., no need for region specific threshold adjustments). Furthermore, we were able to demonstrate that stress-associated variation in reference site biological metrics was greatly minimized. These performance evaluations give us confidence that the balance of environmental representativeness and biological integrity is sufficient to support robust regulatory applications for wadeable perennial streams in California.

<u>Managing inter-regional complexity</u>

Programs attempting to apply a consistent set of criteria for ecological benchmarks across a diverse geographical and anthropogenic landscape are faced with a common problem: Because regions can vary widely in extent of different stressors, a uniform approach is often unable to provide satisfactory results (Herlihy et al. 2008, Mykrä et al. 2008, Dallas 2012). Restrictive criteria may minimize natural stress within the reference network at the expense of spatial or environmental representativeness. In contrast, lowering the bar enough to accommodate highly altered regions can sever the connection to the theoretical anchor of naturalness.

Using the terminology of Stoddard et al. (2006), our reference network could be viewed as a version of the “least disturbed” model. We found that a combination of two strategies allowed us to achieve broad representation of most perennial, wadeable streams in California with a single set of statewide reference criteria: 1) the selective and systematic relaxation of reference screens, and 2) exclusion of pervasively altered regions (e.g., Central Valley) from the population of interest.

Because relaxing thresholds potentially degrades biological integrity, it is critical that impacts to biological integrity be quantified in least disturbed regions (as we did in this study). In highly altered regions, the choice is often between greatly relaxing the overall definition of reference and thus weakening the ability to predict biological potential in less developed regions (Cao and

Hawkins 2011) or excluding a region or category of streams from the main stream network. If this is necessary, condition benchmarks could still be developed using other approaches such as modeling of expected biological indicator scores based on empirical or theoretical relationships with stress (e.g., Chessman 1999, Chessman and Royal 2004, Carter and Fend 2005, Birk et al. 2012). Regardless of which alternate approach is used, benchmarks in excluded regions will need to be related to those used minimally or moderately disturbed regions in order to make sensible state-wide assessments and management decisions (see Herlihy et al. 2008, Bennett et al. 2011).

Applications of the reference condition approach

A well-established reference network has several potential applications for stream and watershed management. Reference concepts provide defensible regulatory frameworks for protecting and managing aquatic resources, and providing a "common currency" for the integration of multiple biological indicators (e.g., algal and fish assemblages). Beyond perennial streams, the approach outlined in this paper can be used to define reference sites for a wide range of habitat types, including non-perennial streams, lakes, depressional wetlands, and estuaries (e.g., Solek et al. 2010). Further, the process of defining reference criteria can be part of the process of identifying streams and watersheds deserving of special protections and application of anti-degradation policies, which are often under-applied in the United States and globally (Linke et al. 2011, Collier 2011).

Two general applications extend these uses to management of non-biological parameters: `1) objective regulatory thresholds for non-biological indicators and 2) context for interpreting targeted and probabilistic monitoring data. The process of establishing regulatory standards for management of water quality parameters with non-zero expected values (e.g., nutrients, chloride, conductivity, and fine sediment) is more subjective than for novel pollutants that do not occur naturally, like pesticides. The range of parameter values found at reference sites can help standardize the way regulatory benchmarks are set for these pollutants. Examples of this concept have appeared in peer-reviewed literature (Yates and Bailey 2010, see Hawkins et al.

2010a, 2010b for a variety of physical and chemical endpoints), but management applications are rare. Comparisons of reference to the full range of stressor values in a region (i.e., as obtained from probability surveys as we did for natural variable values in Figure 3) can establish a framework for evaluating the success of site-specific restoration projects. This context gives management programs the ability to distinguish between relatively small differences in pollutant concentration and environmentally meaningful differences.

<u>Limits of this analysis</u>

Two major types of data limitations have potentially large impacts on any approach to identify reference sites: 1) inadequate or inaccurate GIS layers; and 2) lack of information about reach scale stressors. Although improvements in availability and accuracy of spatial data over the last two decades have greatly enhanced our ability to apply consistent screening criteria across large areas, reliance on these screens can underestimate impairment (Yates and Bailey 2010). The most accurate and uniform spatial data tend to be associated with urban and agricultural stressors (e.g., landcover, roads, hydrologic alteration), so impacts in non-agricultural rural areas (e.g., recreation, livestock grazing, riparian disturbance, invasive species) are typically underestimated (Herbst et al. 2011). Other stressors, such as climate change and aerial deposition of nutrients or pollutants, are even more challenging to screen. Reach scale stressors (proximate stressors) have a large influence on aquatic assemblages (e.g., Waite et al. 2000, Munn et al. 2009), but are challenging to assess unless adequate quantitative data were collected along with biological samples, as this context is often essential for interpreting proximate sources of stress (e.g., Poff et al. 2009). We were fortunate to have access to good reach scale chemical and physical habitat data at many sites, but we undoubtedly missed locally important variables in some cases. We anticipate that this will improve over time as the availability and quality of stressor data sets improves (a pattern we have witnessed over the last 15 years).

Likewise, highly heterogeneous regions like California are likely to contain some rare environmental settings (e.g., Gasith and Resh 1999, Millan et al. 2011) that are difficult to identify and might slip through a screening process such as the one we employed, unless they are actively included in the screening pool. We attempted to include as much environmental diversity as possible, but there are probably some stream types with unique physical or chemical characteristics that were undersampled (e.g., mountain streams > 10,000 ft.). However, the framework we developed provides a means of explicitly testing the degree to which such stream types are represented by the overall network.

Conclusions

An increasing amount of attention has been paid in recent years to the importance of measuring the performance of various components of bioassessment (Cao and Hawkins 2011, Diamond et al. 1996, 2012), particularly as they relate to the assessment of among data set comparability. This attention to validation of performance is likely to help solidify the increasing adoption of biological endpoints in water quality programs worldwide. We believe that similar attention to measuring the performance of reference site networks relative to their intended uses will likewise be of significant benefit. We have provided a number of different examples of tests that can be applied to measure key performance criteria for effective reference networks, environmental coverage and maintenance of biological integrity. These tests should be applicable in other regions and for other reference network purposes, since they were successful in perennial wadeable streams of California, one of the most environmentally heterogeneous regions of the USA.

Acknowledgements <to be added later>

Literature Cited

Angradi, T.R., M.S. Pearson, T.M. Jicha, D.L. Taylor, D.W. Bolgrien, M.F. Moffett, K.A. Blocksom, and B.H. Hill. 2009. Using stressor gradients to determine reference expectations for great river fish assemblages. Ecological Indicators 9: 748-764.

Bailey, R.C., R.H. Norris, and T.B. Reynoldson. 2004. Bioassessment of freshwater ecosystems: Using the reference condition approach. Kluwer Academic Publishers. Boston, MA.

Bailey, R.C., G. Scrimgeour, D. Cote, D. Kehler, S. Linke and Y. Cao. 2012. Bioassessment of stream ecosystems enduring a decade of simulated degradation: lessons for the real world. Canadian Journal of Fisheries and Aquatic Sciences 69: 784-796.

Barbour, M.T., J.B. Stribling, and J.R. Karr. 1995. Multimetric approach for establishing biocriteria and measuring biological condition. Pages 63-77 in W.S. Davis and T.P. Simon (editors). Biological Assessment and Criteria: Tools for water resource planning and decision making. Lewis Pubishers. Boca Raton, FL.

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. JNABS 15: 185-211.

Bennett, C., R. Owen, S. Birk, A. Buffagni, S. Erba, N. Mengin, J. Murray-Bligh, G. Ofenböck, I. Pardo, W. van de Bund, F. Wagner, and J.G. Wasson. 2011. Bringing European river quality into line: an exercise to intercalibrate macro-invertebrate classification methods. Hydrobiologia 667: 31-48.

Birk, S., L. Van Kouwen and N. Willby. 2012. Harmonising the bioassessment of large rivers in the absence of near-natural reference conditions – a case study of the Danube River Freshwater Biology 57: 1716-1732.

Bonada, N., N. Prat, V.H. Resh, and B. Statzner. 2006. Developments in Aquatic Insect Biomonitoring: A comparative analysis of recent approaches. Annual Review of Entomology 51: 495 – 523.

Cao, Y., and C.P. Hawkins. 2011. The comparability of bioassessments: A review of conceptual and Methodological issues. Journal of the North American Benthological Society 30:680-701.

Carter, J.L and S.V. Fend. 2005. Setting limits: the development and use of factor-ceiling distributions for an urban assessment using benthic macroinvertebrates. Pages 179-191 in

L.R. Brown, R.H. Gray, R.M. Hughes, and M.R. Meador, editors 47. Bethesda, MD.X: Effects of urbanization on stream ecosystems. American Fisheries Society Symposium.

Chessman, B.C. 1999. Predicting the macroinvertebrate faunas of rivers by multiple regression of biological and environmental differences. Freshwater Biology 41: 747-757.

Chessman, B.C and M.J. Royal. 2004. Bioassessment without reference sites: use of environmental filters to predict natural assemblages of river macroinvertebrates. Journal of the North American Benthological Society 23: 599-615.

Collier, K.J. 2011. The rapid rise of streams and rivers in conservation assessment. Aquatic Conservation: Marine and Freshwater Ecosystems 21: 397-400.

Collier, K.J., A. Haigh, and J. Kelly. 2007. Coupling GIS and multivariate approaches to reference site selection for wadeable stream monitoring. Environmental Monitoring and Assessment 127: 29-45.

Dallas, H. 2012. Ecological status assessment in Mediterranean rivers: complexities and challenges in developing tools for assessing ecological status and defining reference conditions. Hydrobiologia DOI 10.1007/s10750-012-1305-8

Davies, S.P. and S.K. Jackson. 2006. The biological condition gradient: A descriptive model for interpreting change in aquatic ecosystems. Ecological Applications 16: 1251-1266.

Diamond, J., M. Barbour, J.B. Stribling. 1996. Characterizing and comparing bioassessment methods and their results: A perspective. Journal of the North American Benthological Society 15: 713-727.

Diamond, J., J.B. Stribling, L. Huff, and J. Gilliam. 2012. An approach for determining bioassessment performance and comparability. Environmental Monitoring and Assessment 184: 2247-2260.

Erman, N. 1996. Status of aquatic invertebrates. In: Sierra Nevada Ecosystem Project: Final report to Congress, Vol. II, chap. 35. Davis: University of California, Centers for Water and Wildland Resources.

Falcone, J.A., D.M. Carlisle, and L.C., Weber. 2010. Quantifying human disturbance in watersheds: Variable selection and performance of a GIS- based disturbance index for predicting the biological condition of perennial streams. Ecological Indicators 10: 264-273.

Gasith, A. and V.H. Resh.1999. Streams in mediterranean climate regions: abiotic influences and biotic responses to predictable seasonal events. Ann. Rev. Ecol. Syst. 30:51-81.

Gerth, W.J. and A.T. Herlihy. 2006. The effect of sampling different habitat types in regional macroinvertebrate bioassessment surveys. The Journal of the North American Benthological Society 25: 501-512.

Gibson, J.R., M.T. Barbour, J.B. Stribling, J. Gerritsen, and J.R. Karr. 1996. Biological criteria: Technical guidance for streams and rivers (revised edition). EPA 822-B-96-001. Office of Water. US Environmental Protection Agency. Washington, DC.

Gillett, D.J., P.R. Ode, K. Schiff, R.D. Mazor, K. Ritter, A. Rehn, E. Stein, J. May, and L. Brown. In Prep. Distribution of stressors in California's perennial wadeable streams.

Hawkins, C.P. J.R. Olson, and R.A. Hill. 2010a. The Reference Condition: Predicting benchmarks for ecological water-quality assessments. Journal of the North American Benthological Society 29: 312-343.

Hawkins, C.P., Y. Cao, and B. Roper. 2010b. Method of predicting reference condition biota affects the performance and interpretation of ecological indices. Freshwater Biology 55: 1066-1085.

Herbst, D.B. and E.L. Silldorff. 2006. Comparison of the performance of different bioassessment methods: similar evaluations of biotic integrity from separate programs and procedures. Journal of the North American Benthological Society 25: 513-530.

Herbst, D. B., M.T., Bogan, S.K.Roll and H.D. Safford. 2011. Effects of livestock exclusion on in-stream habitat and benthic invertebrate assemblages in montane streams. Freshwater Biology 57: 204-217.

Herlihy, A.T., S.G. Paulsen, J. Van Sickle, J.L. Stoddard, C.P. Hawkins, and L. Yuan. 2008. Striving for consistency in a national assessment: The challenges of applying a reference condition approach on a continental scale. Journal of the North American Benthological Society 27: 860-877.

Horvitz and Thompson. 1952. A Generalization of Sampling Without Replacement from a Finite Universe, Journal of the American Statistical Association 47, Iss. 260, 1952.

Hughes, R.M., D.P. Larsen, and J.M. Omernik. 1986. Regional reference sites: A method for assessing stream potentials. Environmental Management 10: 629-635.

Kaufmann, P.R., P. Levine, E.G. Robinson, C. Seeliger, and D.V. Peck. 1999. Quantifying physical habitat in wadeable streams. EPA/620/R-99/003. US Environmental Protection Agency. Research Ecology Branch. Corvallis, OR.

Kincaid, T. and T. Olsen. 2009. SPSurvey package for R. Available from http://www.epa.gov/nheerl/arm

Linke, S. , E. Turak, and J. Neale. 2011. Freshwater conservation planning: The case for systematic approaches. Freshwater Biology 56: 6 – 20.

Mapstone, B.D. 2006. Scalable decision criteria for environmental impact assessment: Effect size, Type I, and Type II errors. Pages 67-82 *in* R.J. Schmitt and C.W. Osenberg, editors. Detecting Ecological Impacts. Academic Press. New York, NY.

Mazor, P.R., A.C. Rehn, P.R. Ode, M.Engeln and K.Schiff. In Prep. Development of a bioassessment tool for streams in heterogeneous regions: Accommodating environmental complexity through site specificity in the California Stream Condition Index

Meinshausen, N. 2006. Quantile regression forests. Journal of Machine Learning Research 7: 983-999.

Millan, A., J. Velasco, C. Gutierrez-Canovas, P. Arribas, F. Picazo, D. Sanchez-Fernandez and P. Abellan. 2011. Mediterranean saline streams in southeastern Spain: What do we know?. Journal of Arid Environments 75: 1352-1359.

Moyle, P.B. 1996. Potential aquatic diversity management areas. In: Sierra Nevada Ecosystem Project: Final report to Congress, Vol. II, chap. 57. Davis: University of California, Centers for Water and Wildland Resources.

Moyle, P.B.; Randall, P.J. 1996. Biotic integrity of watersheds. In: Sierra Nevada Ecosystem Project: Final report to Congress, Vol. II, chap. 34. Davis: University of California, Centers for Water and Wildland Resources.

Munn, M.D., I.R. Waite, D.P. Larsen, and A.T. Herlihy. 2009. The relative influence of geographic location and reach-scale habitat on benthic invertebrate assemblage in six ecoregions. Environmental Monitoring and Assessment 154: 1-14.

Mykrä H., J. Aroviita, J. Kotanen, H. Hämäläinen, and T. Muotka. 2008. Predicting the stream macroinvertebrate fauna across regional scales: influence of geographical extent on model performance. Journal of the North American Benthological Society 27: 705-716.

Ode, P.R. 2007. Standard operating procedures for collecting benthic macroinvertebrate samples and associated physical and chemical data for ambient bioassessment in California. Surface Water Ambient Monitoring Program. Sacramento, CA.

Ode, P.R., and K. Schiff. 2009. Recommendations for the development and maintenance of a reference condition management program (RCMP) to support biological assessment of

California's wadeable streams. Report to the State Water Resources Control Board's Surface Water Ambient Monitoring Program (SWAMP). Technical Report 581. Southern California Coastal Water Research Project. Costa Mesa, CA.

Ode, P.R., A.C. Rehn, and J.T. May. 2005. A quantitative tool for assessing the integrity of Southern California coastal streams. Environmental Management 35: 493-504.

Ode, P.R., C.P. Hawkins, and R.D. Mazor. 2008. Comparability of biological assessments derived from predictive models and multimetric indices of increasing geographic scope. Journal of the North American Benthological Society 27: 967-985.

Oksanen et al. 2012. VegAn scripts for R

Olson, J.R. and C.P. Hawkins. 2012. Predicting natural base-flow stream water chemistry in the western United States. Water Resources Research 48: W02504, doi:10.1029/ 2011WR011088

Omernik, J. M. 1987. Ecoregions of the conterminous United States. Map (scale 1:7,500,000). Annals of the Association of American Geographers 77(1):118–125.

Osenberg, C.W. , R.J. Schmitt, S.J. Holbrook, K.E. Abu-Saba, and A.R. Flegal. 2006. Detection of environmental impacts: Natural variability, effect size, and power analysis. Pages 83 – 108 *in* R.J. Schmitt and C.W. Osenberg, editors. Detecting Ecological Impacts. Academic Press. New York, NY.

Pardo, I., C. Gómez-Rodríguez, J. Wasson, R. Owen, W. van de Bund, M. Kelly, C. Bennett, S. Birk, A. Buffagni, S. Erba, N. Mengin, J. Murray-Bligh, G. Ofenböeck. 2012. The European reference condition concept: A scientific and technical approach to identify minimally-impacted river ecosystems. Science of the Total Environment 420: 33-42.

Peck, D.V., A.T. Herlihy, B.H. Hill, R.M. Hughes, P.R. Kaufmann, D.J. Klemm, J.M. Lazorchak, F.H. McCormick, S.A. Peterson, S.A. Ringold, T. Magee, and M. Cappaert. 2006. Environmental Monitoring and Assessment Program—Surface Waters Western Pilot study: Field operations manual for wadeable streams. EPA/620/R-06/003. US Environmental Protection Agency. Office of research and Development Corvallis, OR.

Poff, N.L, B.D. Richter, A.H. Arthington, S.E. Bunn, R.J. Naiman, E. Kendy, M. Acreman, C. Apse, B.P. Bledsoe, M.C. Freeman, J. Henriksen, R.B. Jacobson, J.G. Kennen, D.M. Merritt, J.H. O'Keefe, J.D. Olden, K. Rogers, R.E. Tharme, and A. Warner. 20009. The ecological limits of hydrological alteration (ELOHA): A new framework for developing regional environmental flow standards. Freshwater Biology 55: 147-170.

The R Foundation for Statistical Computing. 2010. R. Version 2.11.1. Available from http://www.r-project.org/

Rehn, A.C. 2008. Benthic macroinvertebrates as indicators of biological condition below hydropower dams on west slope Sierra Nevada streams, California, USA. River Research and Applications 25: 208-228.

Rehn, A.C., P.R. Ode, and J.T. May. 2005. Development of a benthic index of biotic integrity (B-IBI) for wadeable streams in northern coastal California and its application to regional 305(b) assessment. Report to the State Water Resources Control Board. California Department of Fish and Game. Aquatic Bioassessment Laboratory. Rancho Cordova, CA.

Rehn, A.C., P.R. Ode, and C.P. Hawkins. 2007. Comparison of targeted-riffle and reach-wide benthic macroinvertebrate samples: implications for data sharing in stream-condition assessments. Journal of the North American Benthological Society 26: 332-348.

Resh, V.H., L.A. Bêche, J.E. Lawrence, R.D. Mazor, E.P. McElravy, A.H. Purcell, and S.M. Carlson. 2013. Long-term Patterns in Fish and Benthic Macroinvertebrates in Northern California Mediterranean-climate Streams. Hydrobiologia. DOI 10.1007/s10750-012-1373-9

Reynoldson, T.B. 1995. Biological guidelines for freshwater sediment based on BEnthic Assessement of SedimenT (BEAST) using a multivariate approach for predicting biological state. Australian Journal of Ecology 20: 198-219.

Reynoldson, T.B., R.H. Norris, V.H. Resh, K.E. Day, and D.M. Rosenberg. 1997. The reference condition: A comparison of multimetric and multivariate approaches to assess water quality impairment using benthic macroinvertebrates. Journal of the North American Benthological Society 16: 833-852.

Richards, A.B. and D.C. Rogers. 2006. List of freshwater macroinvertebrate taxa from California and adjacent states including Standard taxonomic effort levels. Southwest Association of Freshwater Invertebrate Taxonomists. Chico, CA. Available from www.safit.org

Rosenberg, D.M., T.B. Reynoldson, and V.H. Resh (1999). Establishing reference conditions for benthic invertebrate monitoring in the Fraser River Catchment. British Columbia, Canada. Fraser River Action Plan. Environment Canada. Vancouver, B.C. FRAP Rep. No. DOE-FRAP 1998-32.

Rosenberg, D.M., and V. Resh. 1993. Freshwater biomonitoring and benthic macro-invertebrates. Chapman and Hall, New York.

Sanchez-Montoya MM, Vidal-Abarca MR, Punti T, Poquet JM, Prat N, Rieradevall M, Alba-Tercedor J, Zamora-Munoz C, Toro M, Robles S, Alvarez M, Suarez ML. 2009. Defining criteria to select reference sites in Mediterranean streams. Hydrobiologia 619:39-54

Sandin, L. and R.K. Johnson. 2004. Local, landscape and regional factors structuring benthic macroinvertebrate assemblages in Swedish streams. Landscape Ecology 19: 501-514.

Sleeter, B.M., T.S. Wilson, C. E. Soulard, and J. Liu. 2011. Estimation of the late twentieth century land-cover change in California. Environmental Monitoring and Assessment. 173: 251-266.

Solek, C.W., E. Stein, J.N. Collins, L. Grenier, J.R. Clark, K. O'Connor, C. Clark, and C. Roberts. 2010. Developing a statewide network of reference wetlands for California: Conceptual approach and process for prioritizing and selecting reference sites. Southern California Coastal Water Research Project. Costa Mesa, CA.

Stevens, D.L. and A.R. Olsen. 2004. Spatially balanced sampling of natural resources. Journal of the American Statistical Association 99 (465): 262-278.

Stoddard, J. L., P. Larsen, C. P. Hawkins, R. K. Johnson, and R. H. Norris. 2006. Setting expectations for the ecological condition of running waters: the concept of reference condition. Ecological Applications 16:1267-1276.

TeleAtlas. 2009.

Waite, I.R., A.T. Herlihy, D.P. Larsen, and D.L. Klemm. 2000. Comparing strengths of geographic and nongeographic classifications of stream benthic macroinvertebrates in the Mid-Atlantic Highlands, USA. Journal of the North American Benthological Society 19: 429-441.

Whittier, T.R., J.L. Stoddard, D.P. Larsen, and A.T. Herlihy. 2007. Selecting reference sites for stream biological assessments: Best professional judgment or objective criteria. Journal of the North American Benthological Society 26: 349-360.

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

Wright, J.F., D.W. Sutcliffe, and M.T. Furse, 2000. Assessing the biological quality of fresh waters: RIVPACS and other techniques. The Freshwater Biological Association. Ambleside, UK.

Yates, A.G. and R.C. Bailey. 2010. Selecting objectively defined reference streams for bioassessment programs. Environmental Monitoring and Assessment 170: 129-140.

Yuan, L.L., C.P. Hawkins, and J. Van Sickle. 2008. Effects of regionalization decisions on an O/E index for the national assessment. Journal of the North American Benthological Society 27: 892-905.

Table 1. Sources of spatial data used in this analysis.

Type of spatial data	Source or Model	Reference	Code
Climate	PRISM	http://www.prism.oregonstate.edu	a
Geology and mineral content	Generalized geology and mineralogy data	Olson and Hawkins (2012)	c
Atmospheric deposition	National Atmospheric Deposition Program National Trends Network	http://nadp.sws.uiuc.edu/ntn/	d
Predicted surface water conductivity	Quantile regression forest model (Meinshausen 2006)	Olson and Hawkins (2012)	e
Groundwater	MRI-Darcy Model (Baker et al. 2003)	Olson and Hawkins (2012)	h
Waterbody location and attribute data	NHD Plus	http://www.horizon-systems.com/nhdplus/	i
Dam location, storage	National Inventory of Dams	http://geo.usace.army.mil/	j
Land cover, imperviousness	National Land Cover Dataset (2001)	http://www.epa.gov/mrlc/nlcd-2006.html	k
Elevation	National Elevation Dataset	http://ned.usgs.gov/	m
Mine location and attribute data	Mineral Resource Data System	http://tin.er.usgs.gov/mrds/	n
Discharge location and attribute data	California Integrated Water Quality System	http://www.swrcb.ca.gov/ciwqs/	o
Road location and attribute data	CSU Chico Geographic Information Center	CSU Chico Geographic Information Center	q
Railroad location and attribute data	CSU Chico Geographic Information Center	CSU Chico Geographic Information Center	r
Invasive invertebrate records	CA Aquatic Bioasssessment Lab	http://www.dfg.ca.gov/abl/	u
	University of Montana	http://www.esg.montana.edu/aim/mollusca/nzms/index.html	
	Santa Monica Baykeeper	Abramson et al. (2009)	
	USGS Non-indigenous Aquatic Species Database	http://nas.er.usgs.gov	

Table 2. Natural and stressor metrics used in these analyses. Unless noted in column “n”, metrics were calculated for 1985 sites. “Sources” codes refer to sources listed in Table 1.

					Scales			
Metric	**Description**	**n**	**Source(s)**	**Unit**	**Point**	**WS**	**5k**	**1k**
Natural gradient								
Location								
logWSA	Area of the unit of analysis		l, m	m^2		X		
ELEV	Elevation of site		m	m	X			
MAX_ELEV	Maximum elevation in catchment		m	M		X		
ELEV RANGE	Elevation range of catchment		m	m		X		
New_Lat	Latitude				X			
New_Long	Longitude		m	m	X			
Climate					X			
PPT_00_09	10-y (2000-2009) average annual precipitation		a	mm	X			
TEMP_00_09	10-y (2000-2009) average monthly temperature		a	°C	X			
AtmCa	Catchment mean of mean 1994-2006 annual ppt-weighted mean Ca concentration		d	mg/L		X		
AtmMg	Catchment mean of mean 1994-2006 annual ppt-weighted mean Mg concentration		d	mg/L		X		
AtmSO4	Catchment mean of mean 1994-2006 annual ppt-weighted mean SO4 concentration		d	mg/L		X		
LST32AVE	Average of mean 1961 to 1990 first and last day of freeze		D	Days		X		

MINP_WS	Catchment mean of mean 1971-2000 min monthly ppt	d	mm/month	X
MEANP_WS	Catchment mean of mean 1971-2000 annual ppt	d	mm/month	X
SumAve_P	Catchment mean of mean June-Sep 1971-2000 monthly ppt	d	mm/month	X
TMAX_WS	Catchment mean of mean 1971-2000 max temperature	d	°C	X
XWD_WS	Catchment mean of mean 1961-1990 annual number of wet days	d	# days	X
MAXWD_WS	Catchment mean of 1961-1990 annual max number of wet days	d	# days	X
Geology				
CaO_Avg	Calcite mineral content	c	%	X
MgO_Avg	Magnesium oxide mineral content	c	%	X
N_Avg	Nitrogenous mineral content	c	%	X
P_Avg	Phosphorus mineral content	c	%	X
PCT_SEDIM	Sedimentary geology in catchment	C	%	X
S_Avg	Sulphur mineral content	c	%	X
UCS_Mean	Catchment mean unconfined Compressive Strength	f	MPa	X
LPREM_mean	Catchment mean log geometric mean hydraulic conductivity	h	10^{-6} m/s	X
BDH_AVE	Catchment mean bulk density	f	g/cm^3	X
KFCT_AVE	Catchment mean soil erodability (K) factor	f	None	X
PRMH_AVE	Catchment mean soil permeability	f	In/hour	X

Stressor								
Hydrology								
PerManMade	Percent canals or pipes at the 100k scale		i	%		X		
InvDamDist	Inverse distance to nearest upstream dam in catchment		j	km	X			
Land use								
Ag	% Agricultural (row crop and pasture, NLCD 2001 codes 81 and 82)		k	%		X	X	X
Urban	% Urban (NLCD 2001 codes 21 - 24		K	%		X	X	X
CODE_21	% Urban/Recreational Grass (NLCD code 21)		k	%		X	X	X
Mining								
GravelMinesDensL	Linear density of gravel mines within 250 m of stream channel		n	mines/km		X	X	X
MinesDens	Density of mines (producers only)		n	mines/km^2			X	
Transportation								
PAVED_INT	Number of paved road crossings		q, r	Count		X	X	X
RoadDens	Road density (includes rail)		q, r	km/km^2		X	X	X
Habitat								
P_SAFN	Percent sands and fines	1191	Field measurements	%	X			
W1_HALL	Weighted human influence	964	Field measurements	None	X			
Water chemistry								
CondQR50	Median predicted conductivity	1155	e	uS/cm	X			

Table 3. Thresholds used to select reference sites

Variable	Scale	Threshold	Unit
% Agriculture	1k, 5k, WS	3	%
% Urban	1k, 5k, WS	3	%
% Ag + % Urban	1k, 5k	5	%
% Code 21	1k, 5k	7	%
	WS	10	%
Road density	1k, 5k, WS	2	km/km^2
Road crossings	1k	5	crossings/ km^2
	5k	10	crossings/ km^2
	WS	50	crossings/ km^2
Dam distance	WS	10	km
% canals and pipelines	WS	10	%
Instream gravel mines	5k	0.1	mines/km
Producer mines	5k	0	mines
Specific conductance	site	99/1*	prediction interval
W1_HALL	site	1.5	NA

* The 99th and 1st percentiles of predictions were used to generate site-specific thresholds for specific conductance. Because the model was observed to under-predict at higher levels of specific conductance (data not shown), a threshold of 2000 µS/cm was used as an upper bound if the prediction interval included 1000 µS/cm.

Table 4. Number (n) and percent (%) of reference, and non-reference sites, by region and sub-region as shown in Figure 1.

Region	Total stream network length (km)	Non-reference n	Non-reference %	Reference n	Reference %	% of non-reference sites failing: 1 to 2 thresholds	3 to 5	5 or more
North Coast	9,278	168	69	76	31	26	57	18
Chaparral	8,126	334	78	93	22	44	17	39
--Coastal Chaparral	5,495	275	82	61	18	47	16	37
--Interior Chaparral	2,631	59	65	32	35	34	22	44
South Coast	2,945	555	82	119	18	22	10	68
--South Coast Mountains	1,123	121	58	86	42	62	23	15
--South Coast Xeric	1,821	434	93	33	7	11	6	83
Central Valley	2,407	69	99	1	1	1	7	91
Sierra Nevada	11,313	218	44	276	56	56	26	18
--Western Sierra Nevada	8577	118	47	131	53	58	29	14
--Central Lahontan	2,736	100	41	145	59	54	23	23
Deserts / Modoc	2,531	51	67	25	33	51	29	20
Total	36,599	1395	70	590	30	33	20	47

Table 5. Extent of streams estimated to be reference by region (based on probability data only).

Region	n prob	n prob and ref	% ref (length)	SE
North Coast	162	40	26	3
Chaparral	147	26	19	4
--Coastal Chaparral	97	11	14	5
--Interior Chaparral	50	15	28	6
South Coast	387	54	23	4
--South Coast Mountains	94	42	53	7
--South Coast Xeric	293	12	3	1
Central Valley	60	1	2	2
Sierra Nevada	106	42	43	5
--Western Sierra Nevada	63	18	34	6
--Central Lahontan	43	24	76	5
Deserts / Modoc	57	14	32	10
Total	919	177	29	2

Table 6. Variance partitioning results (DF =number of variables tested minus 1)

Component	DF	Ref R^2 (n = 473)	All sites R^2 (n = 1985)
Pure natural	30	0.095	0.100
Interaction	0	0.014	0.065
Pure stress	17	0.016	0.015
Residual		0.874	0.819

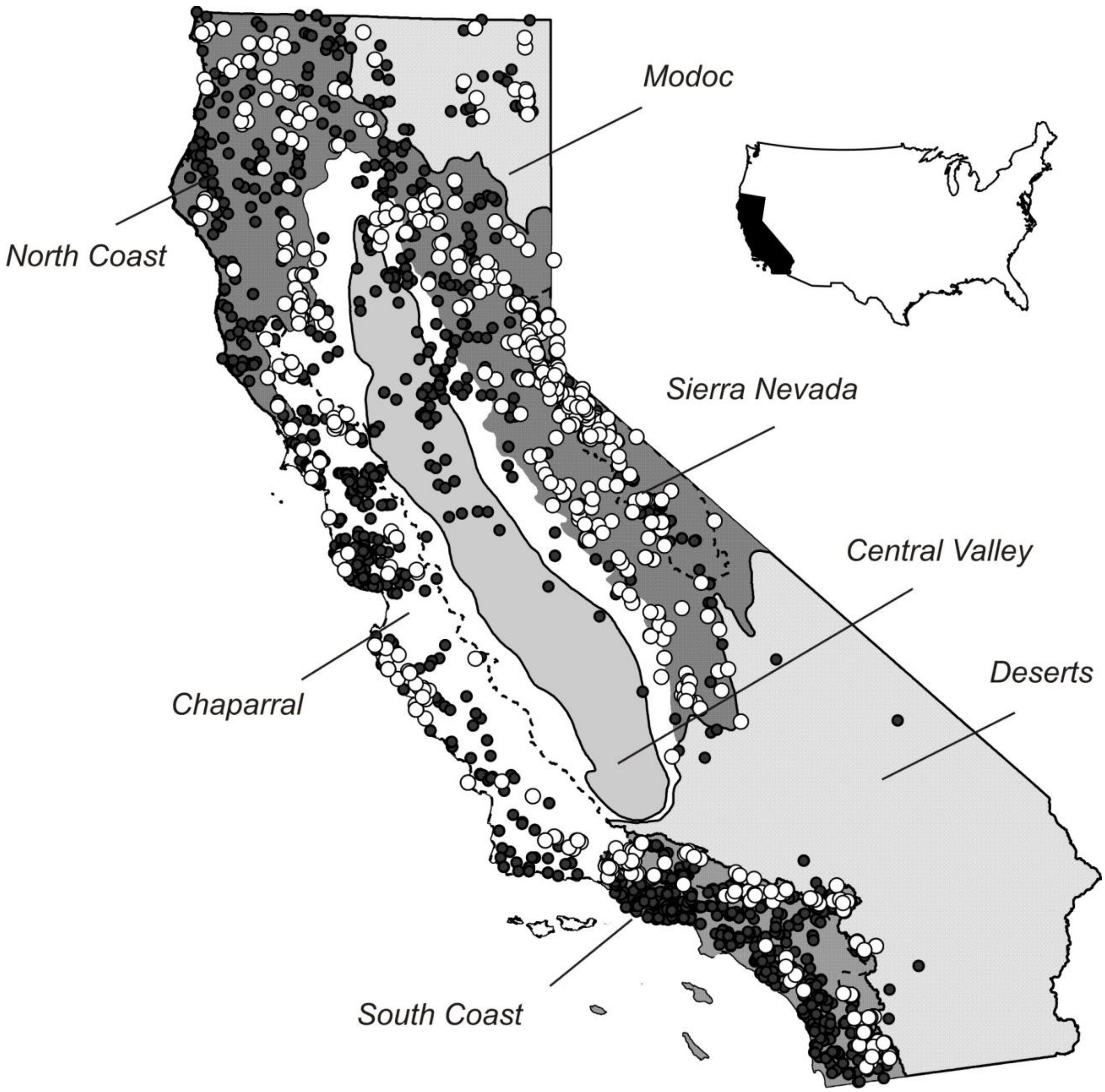

Figure 1. Distribution of 1985 candidate sites screened for inclusion in California's reference pool. White circles represent passing sites and black circles represent sites that failed one or more screening criteria. Thick solid lines indicate boundaries of major ecological regions referred to in the text. Lighter dashed lines indicate sub-regional boundaries referred to in the text (not labeled).

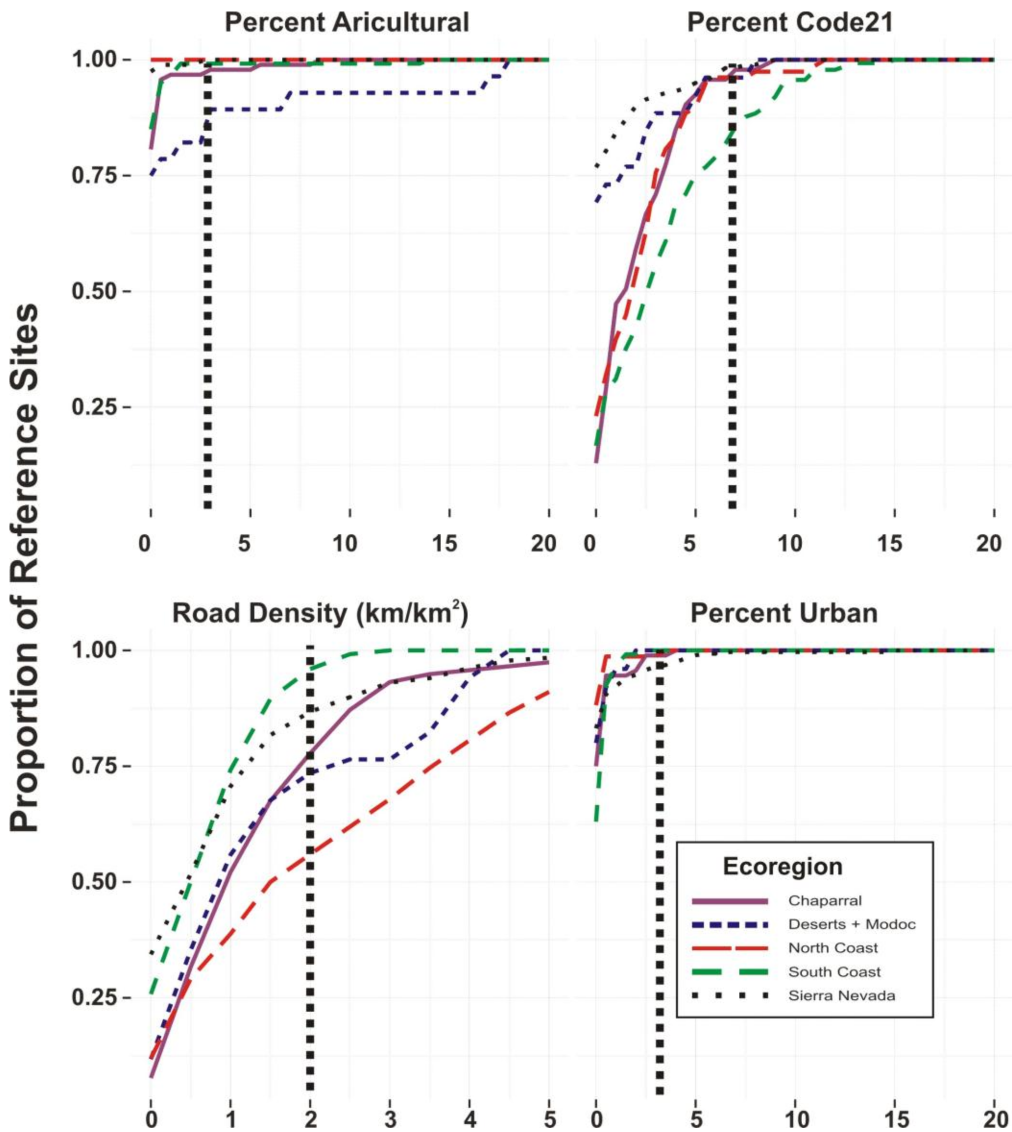

Figure 2. Example threshold sensitivity (partial dependence) curves showing the relationship between numbers of reference sites and thresholds for selected stressors (% Urban, Road Density, % Agricultural, and % Code 21). All other stressors were held constant using the thresholds listed in Table 3. Vertical dotted lines indicate position of impairment thresholds for each metric.

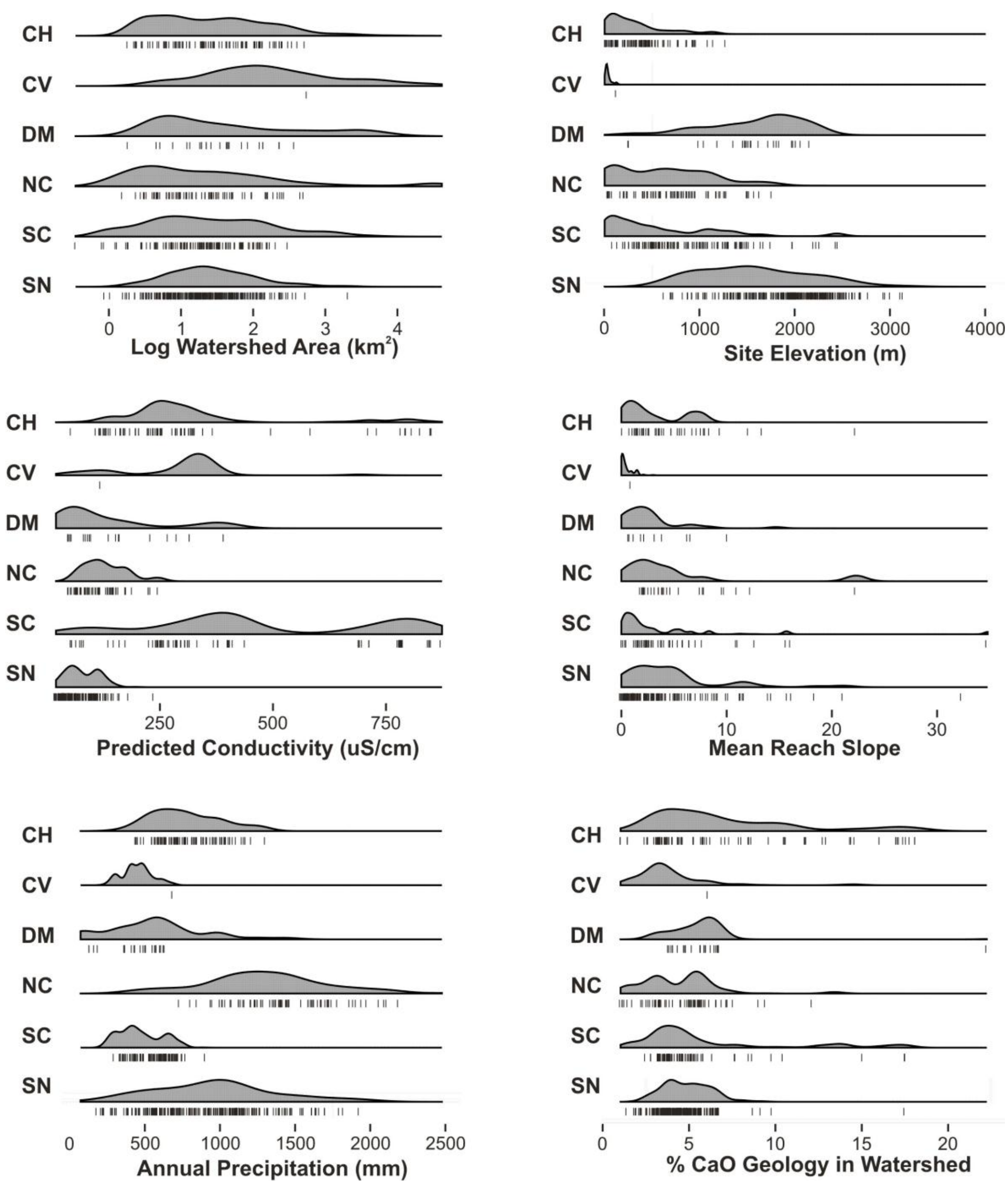

Figure 3. Comparison of reference site representation along several natural gradients. Full distributions (kernel density estimates) of natural gradients estimated from probabilistic sampling surveys within major regions of California. Values of individual reference sites are shown as small vertical lines. Regions (see Figure 1) are abbreviated as follows: SN = Sierra Nevada, SC = South Coast, NC = North Coast, DM = Deserts / Modoc, CV = Central Valley, CH = Chaparral.

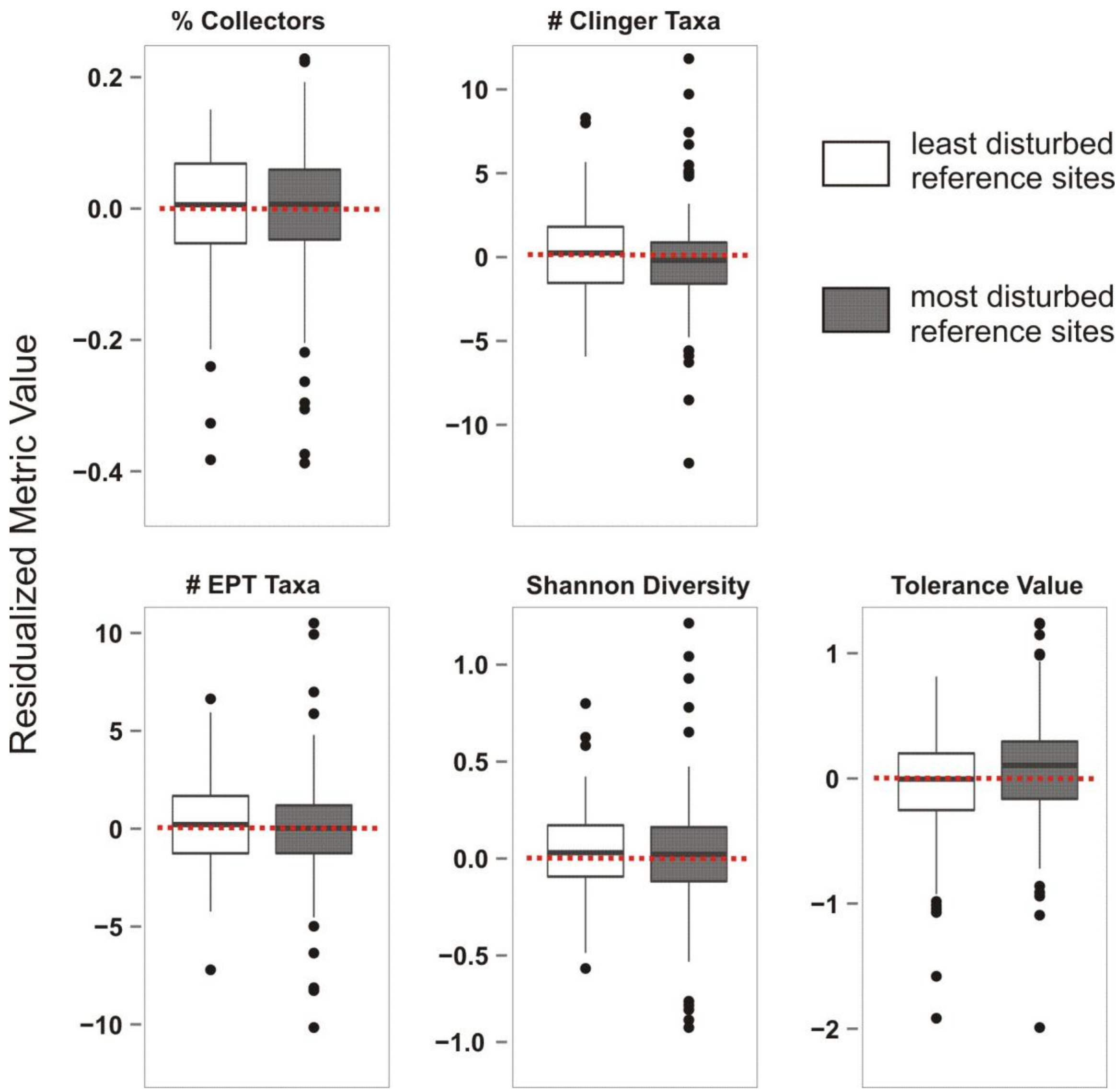

Figure 4. Boxplots comparing biological metric scores at a subset of reference sites that would have passed very strict screens (open boxes) to those of passing sites that had higher levels of human activity (dark boxes).

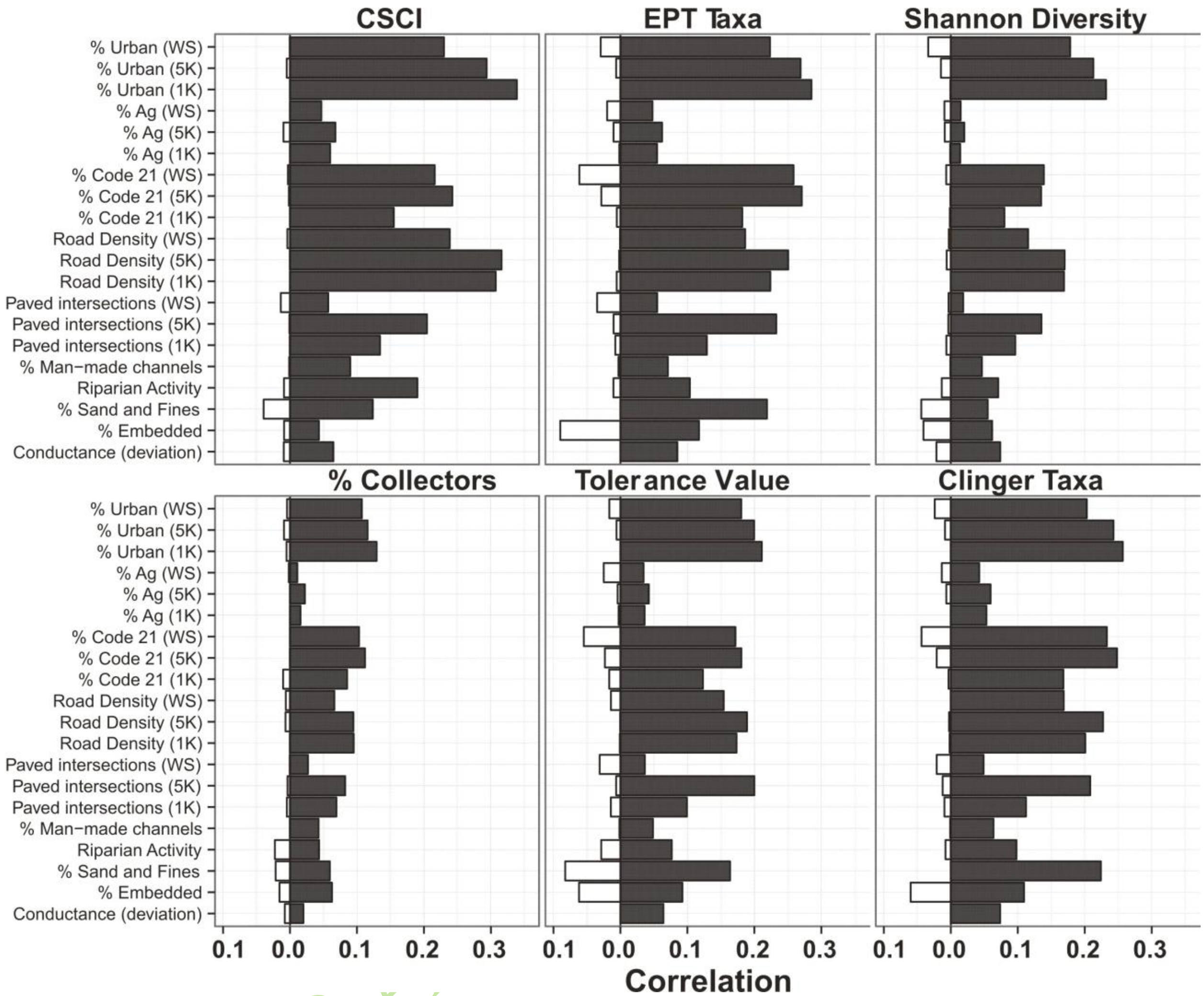

Figure 5. Butterfly plots illustrating the strength of correlations between several bioassessment indicators and common anthropogenic stressors. Open bars on the left of each plot indicate correlations measured at reference sites, and the dark bars on the right of each plot indicate correlations with all sites. (note that CSCI is included here for reviewers benefit, but will be removed in journal version)

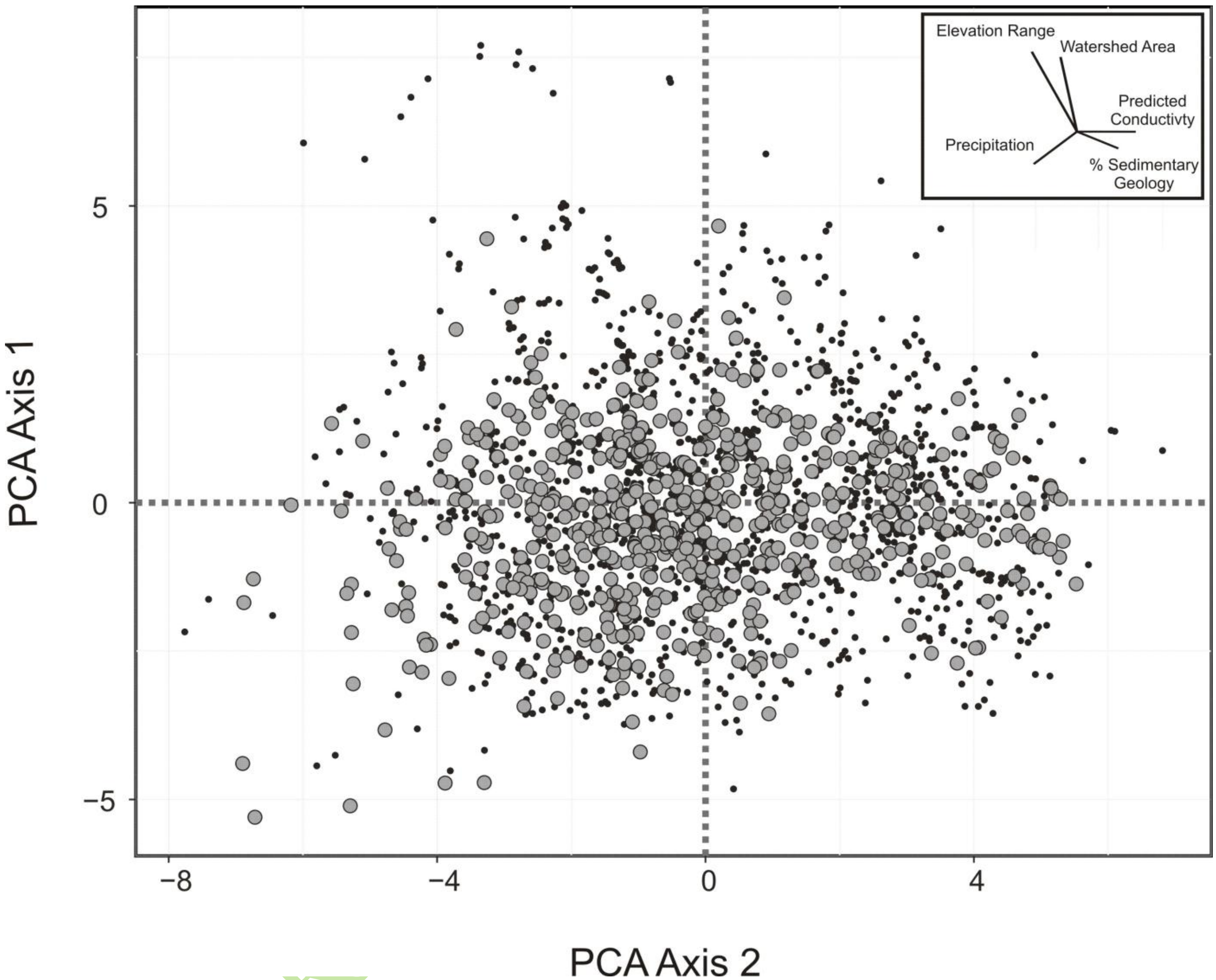

Figure 6. Ordination of benthic invertebrate assemblage data at 1,985 sites at the two primary principle component axes based on primary natural gradients. Grey circles indicate reference sites and black dots indicate non-reference sites. The inset depicts vectors of selected natural variables as estimated from correlation with the PCA axes.