Evaluating the adequacy of a reference site pool for the ecological assessment of streams in environmentally complex regions

Peter Ode[1], Andrew Rehn[1], Raphael Mazor [1,2], Kenneth Schiff[2], Eric Stein[2], Jason May[3], Larry Brown[3], David Herbst[4], David Gillett[2], Kevin Lunde[5] and Charles P. Hawkins[6]

Abstract

The characterization 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 less emphasis has been placed on evaluating the suitability of the reference pool for its intended applications. We present an approach for evaluating the adequacy of a reference pool for supporting biotic index development in environmentally heterogeneous and pervasively altered regions. We screened 1,985 candidate stream reaches to create a pool of 590 reference sites for assessing the biological integrity of streams in California, USA, following standard approaches for selecting sites with low levels of anthropogenic stress. We assessed the resulting pool of reference sites against two primary types of performance criteria. First, we evaluated how well the reference pool represented the range of natural gradients present in the full stream population as estimated by sites sampled through probabilistic surveys. Second, we evaluated the degree to which we were successful in rejecting sites influenced by anthropogenic stress by: a) measuring the biological variance associated with remaining human activity at reference sites, and: b) comparing biological metric scores at a subset of near-pristine reference sites that passed very strict screens with scores at sites that passed less stringent (standard) screening thresholds. Using this approach, we validated a reference pool with minimal human-associated stress that also provided nearly full coverage of environmental heterogeneity. This approach should be widely applicable and customizable to particular regional or programmatic needs.

Introduction

The worldwide use of biological indicators in water quality monitoring programs has evolved rapidly in the last 30 years (Rosenberg and Resh

[1]*CA Department of Fish and Wildlife, Aquatic Bioassessment Laboratory, Rancho Cordova, CA*

[2]*Southern California Coastal Water Research Project, Costa Mesa, CA*

[3]*US Geological Survey, Sacramento, CA*

[4]*Sierra Nevada Aquatic Research Laboratory, Mammoth Lakes, CA*

[5]*San Francisco Bay Regional Water Quality Control Board, Oakland, CA*

[6]*Utah State University, Department of Watershed Sciences, Western Center for Monitoring and Assessment of Freshwater Ecosystems, Ecology Center, Logan, UT*

1993, Gibson *et al.* 1996, Wright *et al.* 2000, Bonada *et al.* 2006, European Commission 2000, 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 that should occur in the absence of human-derived disturbance, i.e., 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). The need to anchor biological expectations to a reference condition is now widely regarded as highly desirable. Furthermore, reference condition for an individual site should be based on the biological states observed at reference sites having similar natural environmental settings. However, there remains little discussion of how to evaluate if a pool of reference sites is adequate for its intended uses.

Many recent treatments of the reference site selection process recognize that objective criteria can greatly enhance the defensibility of reference condition determinations (Whittier *et al.* 2007, Herlihy *et al.* 2008, Yates and Bailey 2010, Lunde *et al.* In press) and examples of objective site-selection processes are increasingly common (e.g., Hawkins *et al.* 2000, Stoddard *et al.* 2006, Collier *et al.* 2007, Sanchez-Montoya *et al.* 2009 Whittier *et al.* 2007, Yates and Bailey 2010). There are several different approaches to identifying reference sites (e.g., Herlihy *et al.* 2008, Sánchez-Montoya *et al.* 2009, Yates and Bailey 2010), reflecting philosophical differences of practitioners and the varied monitoring questions each program addresses. Programs that measure biological integrity often use a "minimally-disturbed" or "least-disturbed" standard (*sensu* Stoddard *et al.* 2006) for selecting reference sites because truly pristine streams are rare or non-existent throughout the world. The main challenge is to choose site selection criteria that retain sites with the highest biological integrity possible, thus maintaining the philosophical integrity of the reference condition concept. However, geographic variation in the importance of different stressors that affect biological condition can complicate the achievement of uniform reference definitions (Statzner *et al.* 2001, Herlihy *et al.* 2008, Mykrä *et al.* 2008, Ode *et al.* 2008). Thus, robust reference site selection involves balancing two potentially conflicting goals: 1) reference criteria should select sites that uniformly represent the least disturbed conditions throughout the region of interest, minimizing the effects of remaining stress on the indicator of interest, and 2) reference sites should represent the full range of environmental settings in the region and with sufficient numbers to adequately characterize variability in the indicator of interest. Restrictive criteria may minimize anthropogenic stress within the reference network at the expense of spatial or environmental representativeness, particularly in regions with diverse environmental settings or pervasive alteration (Mapstone 2006, Osenberg *et al.* 2006, Yuan *et al.* 2008, Dallas 2012, Feio *et al.* 2013). In contrast, lowering the bar enough to accommodate highly altered regions can weaken the ability to measure deviation from the natural biological state.

Ideally, a large number of undisturbed streams of all types would allow us to focus exclusively on avoiding contamination of the reference pool with biologically-impaired sites. However, overly stringent criteria may result in under-representation of biologically important natural gradients, Thus, high rejection rates of candidate sites can reduce the performance (i.e., accuracy and precision) of biological condition indices based on the reference pool to the extent that biologically important sources of natural variation are not accounted for. The consideration of environmental representativeness is especially critical in regulatory applications where errors in estimating site-specific reference conditions may 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 describes an approach we used to evaluate the adequacy of a reference site pool for assessing of biological condition of streams in California, an environmentally complex region of the USA overlain with large areas of pervasive development. This work builds on previous efforts to identify reference conditions in similarly complex regions (e.g., Collier *et al.* 2007, Herlihy *et al.* 2008, Sánchez-Montoya *et al.* 2009, Falcone *et al.* 2010, Yates and Bailey 2010). We drew on these efforts to identify an initial suite of land use stressors and screening thresholds, expanded them to accommodate a broad array of anthropogenic activities known to be important in California, then evaluated the degree to which we met our objectives of environmental representativeness and maintenance of biological integrity in the final pool of reference

sites. This approach can be applied to any method of generating a pool of reference sites, or integrating pools from different programs.

Methods

A set of 1,985 candidate reference sites representing a wide range of stream types was assembled to support development of screening criteria. Each site was characterized with a suite of land use and land cover metrics that quantified both its natural characteristics and potential anthropogenic stressors at the site or in its upstream drainage basin. Sites were then screened with a subset of land use metrics (e.g., road density and percent urban in the upstream watershed) using thresholds that represented low levels of anthropogenic activity ("least disturbed" *sensu* Stoddard *et al.* 2006). Finally, the pool of reference sites that passed screening criteria was evaluated to assess whether the objectives of balancing naturalness and representativeness were achieved to a degree sufficient to support the development and defensible application of biological scoring tools and condition thresholds (i.e., biocriteria).

Setting

California's stream network is approximately 280,000 km long and 30% perennial according to the NHD medium resolution (1:100 k) stream hydrology dataset (US Geological Survey 2009) and drains a large (424,000 km²) and remarkably diverse landscape. Spanning latitudes between 33° and 42° (N), California's geography is characterized by extreme natural gradients. California boasts both the highest and lowest elevations in the conterminous 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 recently uplifted and poorly consolidated marine sediments in the Coast Ranges to alluvium in its broad internal valleys, to granitic batholiths along the eastern border to recent volcanic lithology in the northern mountains. The State's environmental 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 accompanied 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 intensive recreational uses. Our analyses generally treated environmental variation as multiple continuous variables, but to facilitate some assessments, the state was divided into six discrete regions based on modified ecoregional (Omernik 1987) and hydrological boundaries (Figure 1).

Aggregation of Site Data

More than 20 federal, state, and regional monitoring programs were inventoried to assemble data sets used for screening reference sites. Candidate data sets were mostly restricted to wadeable, perennial streams, but some non-wadeable rivers were included, as were some non-perennial streams because of unavoidable imprecision in the assignment of flow status to stream reaches. We retained these sites in this analysis because they helped illustrate patterns of

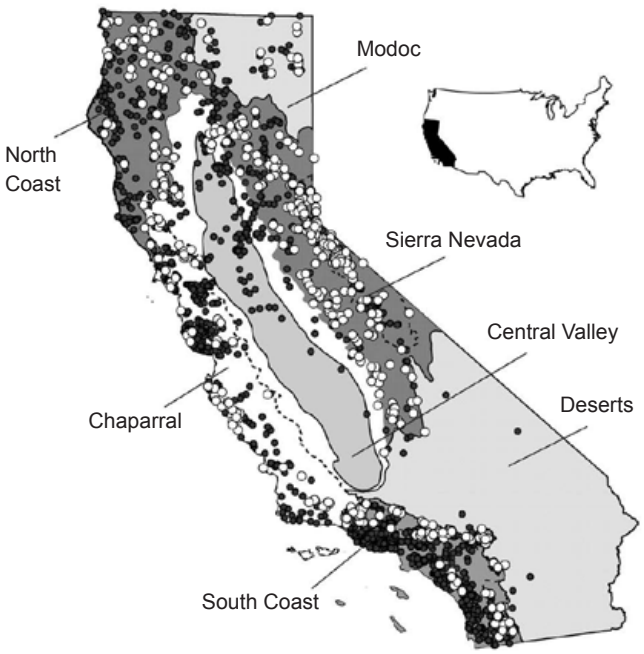

Figure 1. Distribution of 1,985 candidate sites screened for inclusion in California's reference pool. White circles represent passing sites and black circles represent sites that failed one or more screening criteria. Thick solid lines indicate boundaries of major ecological regions referred to in the text. Lighter dashed lines indicate sub-regional boundaries referred to in the text (not labeled).

environmental representativeness. All 1,985 unique sites were sampled between 1999 and 2010 (Figure 1) and resulting data were compiled into a single database. Sites sampled within 300 m of one another were considered redundant.

Assembled field data included benthic macroinvertebrate (BMI) counts 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 and 2). Most BMI data were collected with the reachwide protocol of the USEPA Environmental Monitoring and Assessment Program (EMAP; Peck *et al.* 2006), but some older data were collected with targeted riffle protocols. Previous studies have shown these protocols to produce similar bioassessments in the western USA (Ode *et al.* 2005, Gerth and Herlihy 2006, Herbst and Silldorff 2006, Rehn *et al.* 2007). Prior to all analyses, BMI data were converted to standard taxonomic effort levels (generally genus-level identifications except chironomid midges identified to subfamily; see SAFIT 2011), and subsampled when necessary to 500-count. For calculation of reach-scale physical habitat metrics, we prioritized data from sampling programs that used quantitative field protocols (e.g., Peck *et al.* 2006, Ode 2007) and allowed calculation of variables defined by Kaufmann *et al.* (1999).

Integration of Probability Data Sets

A subset of data 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 in California. Probability datasets provide objective statistical estimates of the true distribution of population parameters (in this case, natural characteristics of California's perennial stream network; Stevens and Olsen 2004). First, a common sample frame was created so that the relative contribution of each site to the overall distribution of stream length (the site's weight) could be calculated in the combined

Table 1. Sources of spatial data used for screening of reference sites and evaluation of reference site characteristics. Codes refer to application in Table 2.

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	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	Mineral Resource Data System	http://tin.er.usgs.gov/mrds/	n
Discharge location and attribute	California Integrated Water Quality System	http://www.swrcb.ca.gov/ciwqs/	o
Road location and attribute	CSU Chico Geographic Information Center	CSU Chico Geographic Information Center	q
Railroad location and attribute	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 variables used for screening reference sites and evaluating reference site characteristics. Unless noted in column "n" (= sample size), metrics were calculated for 1,985 sites. "Source(s)" codes refer to sources listed in Table 1. Scale refers to spatial area of analysis (WS= upstream watershed, 1k = watershed area within 1k of site, 5k = watershed area within 5k of site). Variables preceded by an asterisk were used in the calculation of predicted conductivity (CondQR50).

Metric	Description	n	Source(s)	Unit	Scale			
					Point	WS	5 k	1 k
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		X			
Climate								
*PPT_00_09	10-y (2000-2009) average annual ppt		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		

Table 2. Continued

Metric	Description	n	Source(s)	Unit	Scale			
					Point	WS	5 k	1 k
*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		
CondQR50	Median predicted conductivity	1155	E	µS/cm		X		
Stressor								
Hydrology								
PerManMade	Percent canals or pipes at the 100 k 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								
GravelMines	Linear density of gravel mines within 250 m		n	mines/km		X	X	X
MinesDens	Number of mines (producers only)		n	mines			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								
Embeddedness	Average % cobble embeddedness	576	Field measured	%	X			
P_SAFN	Percent sands and fines	1191	Field measured	%	X			
W1_HALL	Weighted human influence	964	Field measured	None	X			

data set. All probabilistic sites were registered to a uniform stream network (NHD Plus v1, Horizon Data Systems 2006), attributed with strata defined by the design parameters of all integrated programs (e.g., land use, stream order, survey boundaries, etc.). Second, site weights were calculated for each site by dividing total stream length in each stratum (e.g., all second order streams draining agricultural areas in the north coast region) by the number of sampled sites 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). Finally, site weights were used to estimate regional distributions for environmental variables using the Horvitz-Thompson estimator (Horvitz-Thomson 1952). Confidence intervals for estimates of the proportion of California's stream length meeting reference criteria were based on local neighborhood variance estimators (Stevens and Olsen 2004).

GIS Data and Metric Calculation

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 (ESRI 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 of the site (watershed), within 5 km upstream (5 k) and within 1 km upstream (1 k). Polygons defining these spatial units were created with 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, Gesch *et al.* 2002). Local polygons were created by intersecting a 5-km or 1-km radius circle centered at the stream site with the primary watershed polygon. Metrics associated with sampling location, but not upstream polygons (e.g., mean annual temperature, elevation, NHD+ attributes, etc.), were calculated based on each site's latitude and longitude. Data for all screening variables was available for all sites, except for W1_HALL (Kaufmann *et al.* 1999), a measure of anthropogenic activity at the reach scale that was available at approximately half of the sites (Table 2).

Selection of Stressor Screening Variables and Thresholds

To restrict the reference pool to sites with low amounts of anthropogenic disturbance, we eliminated sites that exceeded screening thresholds for a set of human activity variables (Table 3). Failure of any one screen was sufficient to eliminate any candidate site from the reference pool. Specific metrics and thresholds were initially identified from a combination of prior reference development projects (Ode *et al.* 2005; Rehn *et al.* 2005, Stoddard *et al.* 2006, Rehn 2008) or values obtained from the literature (e.g., Collier *et al.* 2007, Angradi *et al.* 2009, Falcone *et al.* 2010). Stressor values representing low levels of human activity were used to set thresholds for metrics of different spatial scales (e.g., 1 k or 5 k) that lacked published values. Screening thresholds were intentionally set at higher (i.e., less stringent) values 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 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.

Table 3. Thresholds used to select reference sites. Scale refers to spatial area of analysis (WS= upstream watershed, 1 k = watershed area within 1 km of site, 5 k = watershed area within 5 km of site). * indicates that the 99th and 1st percentiles of predictions were used to generate site-specific thresholds for specific conductance; because the predicted conductivity model (Olson and Hawkins 2012) 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.

Variable	Scale	Threshold	Unit
% Agriculture	1 k, 5 k, WS	3	%
% Urban	1 k, 5 k, WS	3	%
% Ag + % Urban	1 k, 5 k	5	%
% Code 21	1 k, 5 k	7	%
	WS	10	%
Road Density	1 k, 5 k, WS	2	km/km^2
Road Crossings	1 k	5	crossings/km^2
	5 k	10	crossings/km^2
	WS	50	crossings/km^2
Dam Distance	WS	10	km
% Canals and Pipelines	WS	10	%
Instream Gravel Mines	5 k	0.1	mines/km
Producer Mines	5 k	0	mines
Specific Conductance	site	99/1*	prediction interval
W1_HALL	site	1.5	NA

Exploration of Screening Threshold Sensitivity

Regions often vary in the relative dominance of different types of stressors. Thus, the relative contribution of different stressors to overall disturbance at candidate sites also varies regionally. To explore regional differences in reference site selection, 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. Examination of resulting partial-dependence curves was used to evaluate the number of reference sites

potentially gained by relaxing thresholds for each screening metric in each region (see Hill *et al.* 2013 for a similar example). We used this information to differentiate stressor thresholds whose adjustment had a large influence on accepted reference sites (and might therefore improve overall environmental representativeness of the reference pool) from thresholds whose adjustment had little influence on the numbers of final reference sites.

Performance Measures

Evaluation of Reference Pool Representativeness

We evaluated two aspects of representativeness: 1) the number of reference sites identified 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 physical and chemical gradients associated with California streams (i.e., environmental representativeness). A target minimum number of sites per region was not set, but regions with few reference sites may need to be pooled with other similar regions or excluded from subsequent reference-based analyses in the future. Geographic representation alone is not sufficient for evaluating representativeness, so we also evaluated the distribution of reference sites across individual natural gradients and in multivariate environmental space identified by principal components analysis (PCA). All natural gradients listed in Table 2 were used in the PCA analysis except the three atmospheric deposition variables (AtmCa, AtmMg, and AtmSO4). Also included in the PCA was predicted conductivity (Olson and Hawkins 2012), an estimate of site-specific conductivity based on modeled relationships between observed conductivity and a suite of natural geographic, geological, climatic and atmospheric variables (Table 2).

Because many natural gradients (e.g., temperature and precipitation) are correlated with locational variables (i.e., latitude, longitude, and elevation), geographic patterns may obscure interpretation of environmental gradient representation. To compensate for this potential effect, we built multiple linear regression models of each non-locational variable (all variables in Table 2, except stressors, elevation, latitude and longitude) against the three locational variables. Residuals from these models were used in the PCA instead of raw variables.

Evaluation of Anthropogenic Stress in the Reference Network

All thresholds allowed at least some degree of upstream human activity (i.e., reference sites were not pristine); responsiveness of BMI metrics to sources of stress allowed by our screens was evaluated in two ways. First, variance in BMI metrics associated with remaining stressor sources at reference sites (measured as Pearson's r) was compared to variance in the overall data set to examine the extent to which reference screening thresholds minimized the impact of major stressors on biology. If Pearson's r was <0.1 for associations 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, BMI metric values were used to verify biological condition in the final reference population (note that use of biological data in selection of land use metrics and thresholds was deliberately avoided in the reference site screening process). BMI metrics at a subset of sites passing more stringent screens were compared to metrics from remaining sites passing only "standard" screens (Table 3) using t-tests. The more stringent screens were: <1% agricultural, urban and Code 21 (a development-associated vegetation class in the NLCD dataset) at all spatial scales; road density <1 km/km^2 for all spatial scales; W1_HALL <0.5; all other criteria as listed in Table 3.

Because the BMI metric values indicative of healthy biological condition vary in different environmental settings, metric values were adjusted for major natural gradients by using residuals from random forest models of metric response to natural gradients as the response variable instead of raw metrics. Eighty percent of the reference sites were used to calibrate random forest models. First, recursive feature elimination (RFE) was used to select the simplest random forest model whose mean squared error was within 2% of the mean squared error of the optimal model. RFE was implemented with the caret package in R (Kuhn *et al.* 2012). Once the predictors were selected through RFE, 250-tree forests were created for each metric using the randomForest package (Liaw and Wiener 2002). These models were then used to predict metric values and calculate residuals for all sites, using out-of-bag predictions for the reference calibration set. Equivalent residuals in the most stressed and least stressed reference groups

would be considered evidence that the measurement of biological integrity for sites of variable quality was equitably maintained.

Equivalent metric scores in most-stressed and least-stressed reference groups was considered evidence that biological integrity among reference sites of variable quality was equitably maintained.

Results

Reference Status by Region

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

Based on probability survey 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.

Table 5. Number of streams and extent of stream length estimated to be reference by region (% ref ±1 standard error) based on probability data only), indicating the number of probabilistic sites used for estimates (n prob) and the number of probabilistic sites meeting reference criteria (n prob and ref).

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

Only 2 to 3 % of stream length in the Central Valley and the South Coast Xeric regions were estimated to be in reference condition, 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 was estimated to meet reference criteria (similar to

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

Region	Total Stream Network (km)	Non-Reference		Reference		% of Non-Reference Sites Failing		
		n	%	n	%	1 - 2 Thresholds	3 - 5 Thresholds	≥5 Thresholds
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

levels seen in Chaparral regions), suggesting that the abundance of reference sites in the North Coast is due more to the large extent of perennial streams than 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, most non-reference sites in the Sierra Nevada and the South Coast Mountains failed only one or two metrics (typically road density and NLCD Code 21), but 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. Other regions had intermediate failure rates. For example, 44% of Chaparral sites were rejected on the basis of only one or two stressors (usually 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 Deserts + Modoc sites were generally less stressed than Chaparral sites, with 51% of sites failing only one or two criteria.

Related patterns were reflected in threshold sensitivity plots (Figure 2). For example, adjusting thresholds for the land use metrics % agricultural and % urban had little influence on the proportion of sites that passed reference screens in most regions, indicating that other screening thresholds were limiting. This pattern was typical for most metrics except road density and Code 21, where even modest relaxation of thresholds resulted in more passing sites in most regions, especially for road density in the North Coast and Chaparral, and Code 21 in the North Coast, Chaparral and South Coastal Mountains. This sensitivity allowed us to selectively increase screening thresholds for road density and Code 21, thereby increasing the number of passing sites in several regions, particularly in the Interior Chaparral, a region with relatively few sites prior to the adjustment. Thus, slight relaxation of 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 (n = 919) allowed us to produce well-resolved

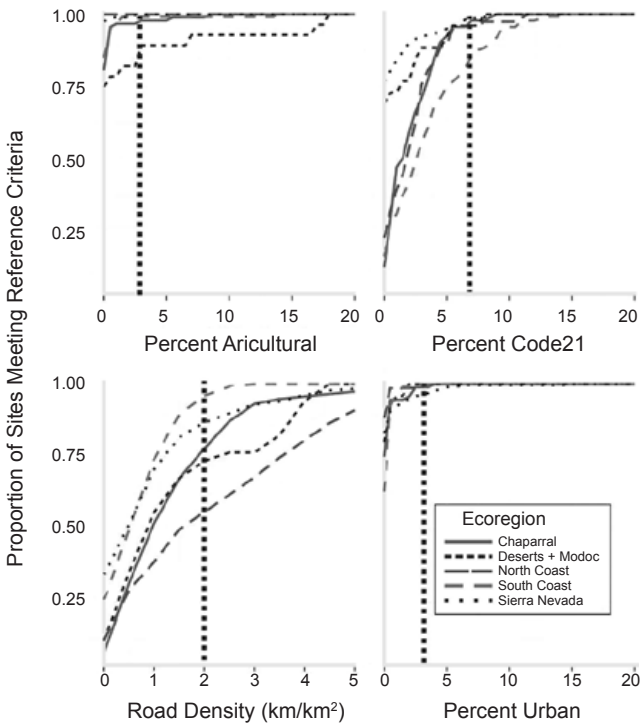

Figure 2. Example threshold sensitivity (partial dependence) curves showing the relationship between proportion of potential reference sites and thresholds for selected stressors (Percent Agricultural, Percent Code 21, Road Density, and Percent Urban). All other stressors were held constant using the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

distribution curves for a suite of natural gradients in each region (Figure 3). For nearly all natural gradients and regions examined, the distribution of reference sites was well-matched to the overall distribution of gradients in most regions of the California. Possible exceptions were very high elevation streams (i.e., >3,000 m) and very large watersheds (i.e., >500 km^2). Most of the other 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).

Principal components analysis (PCA) of environmental variables also showed that the reference pool represented natural gradients well (Figure 4), as there were few identifiable gaps in ordination space. Gaps were generally restricted to the extremes of gradients. For example, in a two dimensional plot of PCA Axis 1 (22% of variation) and PCA Axis

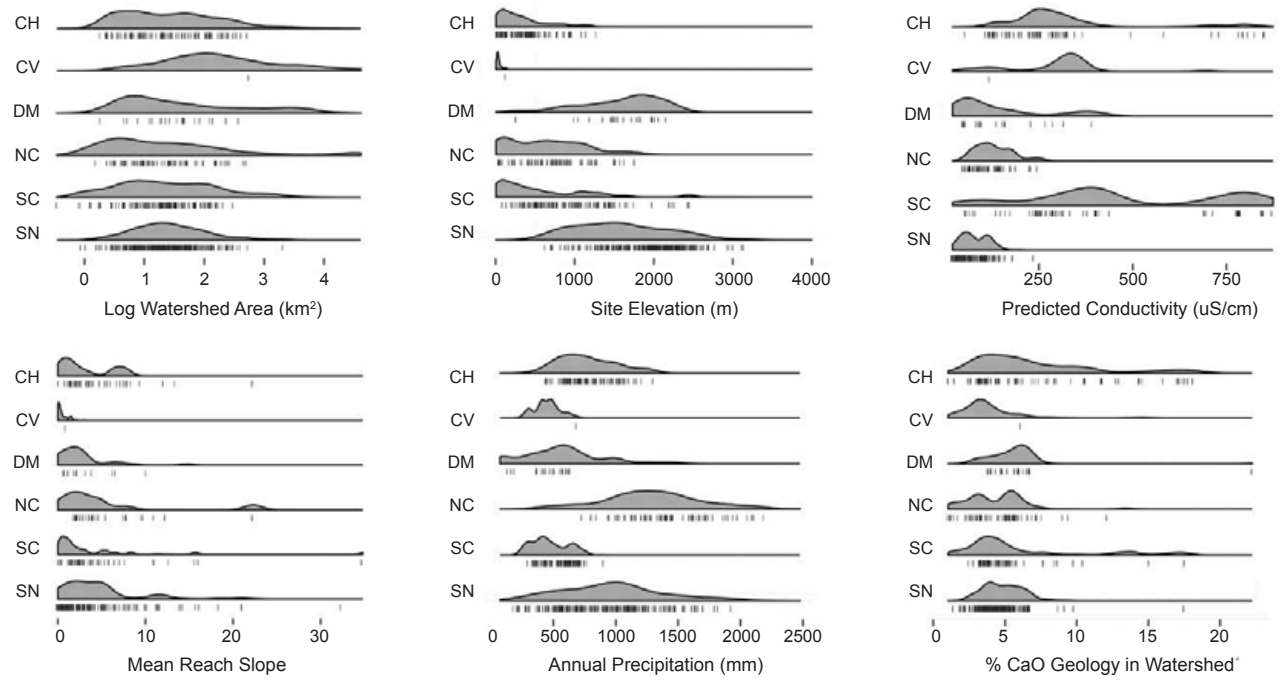

Figure 3. Comparison of reference site representation along several biologically-influential natural gradients. Full distributions of natural gradients estimated from probabilistic sampling surveys within major regions of California are shown as kernel density estimates. Values of individual reference 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.

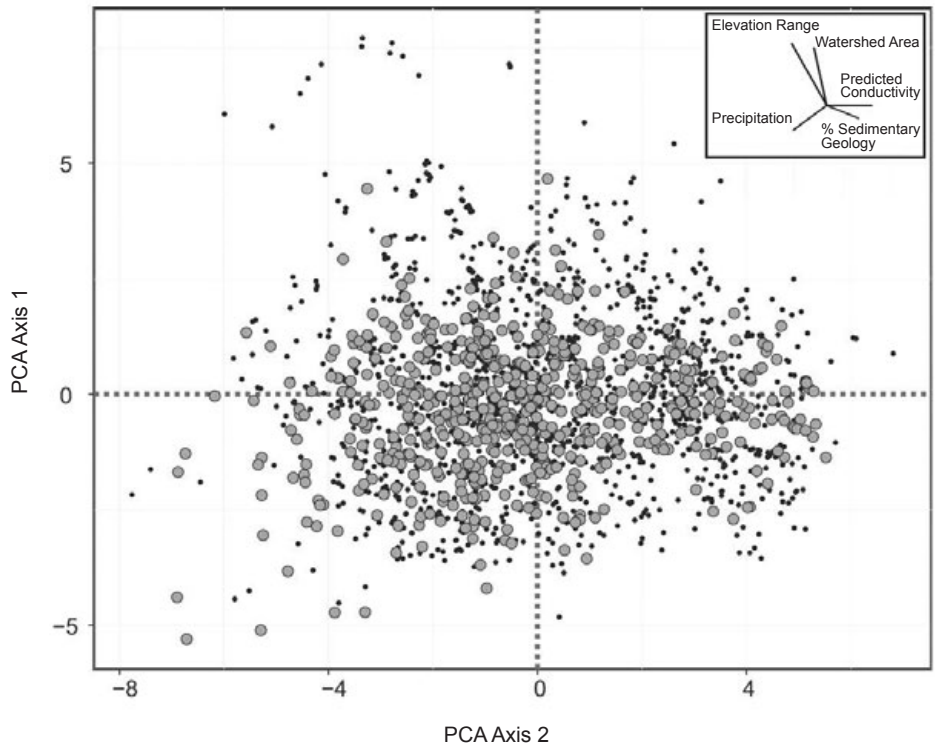

Figure 4. Ordination of 1,985 sites based on natural environmental characteristics (i.e., geology and climate variables listed in Table 2), showing the two primary principle component axes. Grey circles indicate reference sites and black dots indicate non-reference sites. The inset depicts vectors of selected natural variables as estimated from correlation with the PCA axes.

2 (11%), a cluster of sites that lacks reference site coverage is evident in the upper-left quadrant (Figure 4), and corresponds to large river (nonwadeable) sites with the largest watersheds. Sparse coverage in the upper-right quadrant of the ordination corresponds to sites receiving little rainfall, where perennial streams are predominantly a product of urban or agricultural runoff. Other axis combinations indicated similarly strong coverage of natural gradients.

Biological Response to Stressors

Biological variance was only minimally associated with anthropogenic stress at reference sites as compared to the full dataset (Figure 5), indicating low levels of residual anthropogenic impairment in our reference pool. Although reference thresholds did not completely eliminate the influence of disturbance on BMI metrics in our reference sites, this influence was greatly reduced across all the metrics evaluated. Furthermore, our screening thresholds successfully reduced the influence of stressors that were not specifically included in reference screens, such as percent sand and fine sediment, which is presumably associated with included stressors (Figure 5). The low variance of BMI metrics associated with anthropogenic stressors in the reference pool indicates that we did not sacrifice biological integrity in order to achieve adequate natural gradient representation. BMI metric scores at reference sites that passed the most stringent screening criteria (n = 294) were nearly indistinguishable from remaining reference sites that passed ”standard” screens (Table 3). All comparisons (t-tests) were non-significant at Bonferroni-adjusted p-values of 0.01 (Figure 6), implying that reference sites with lowest disturbance

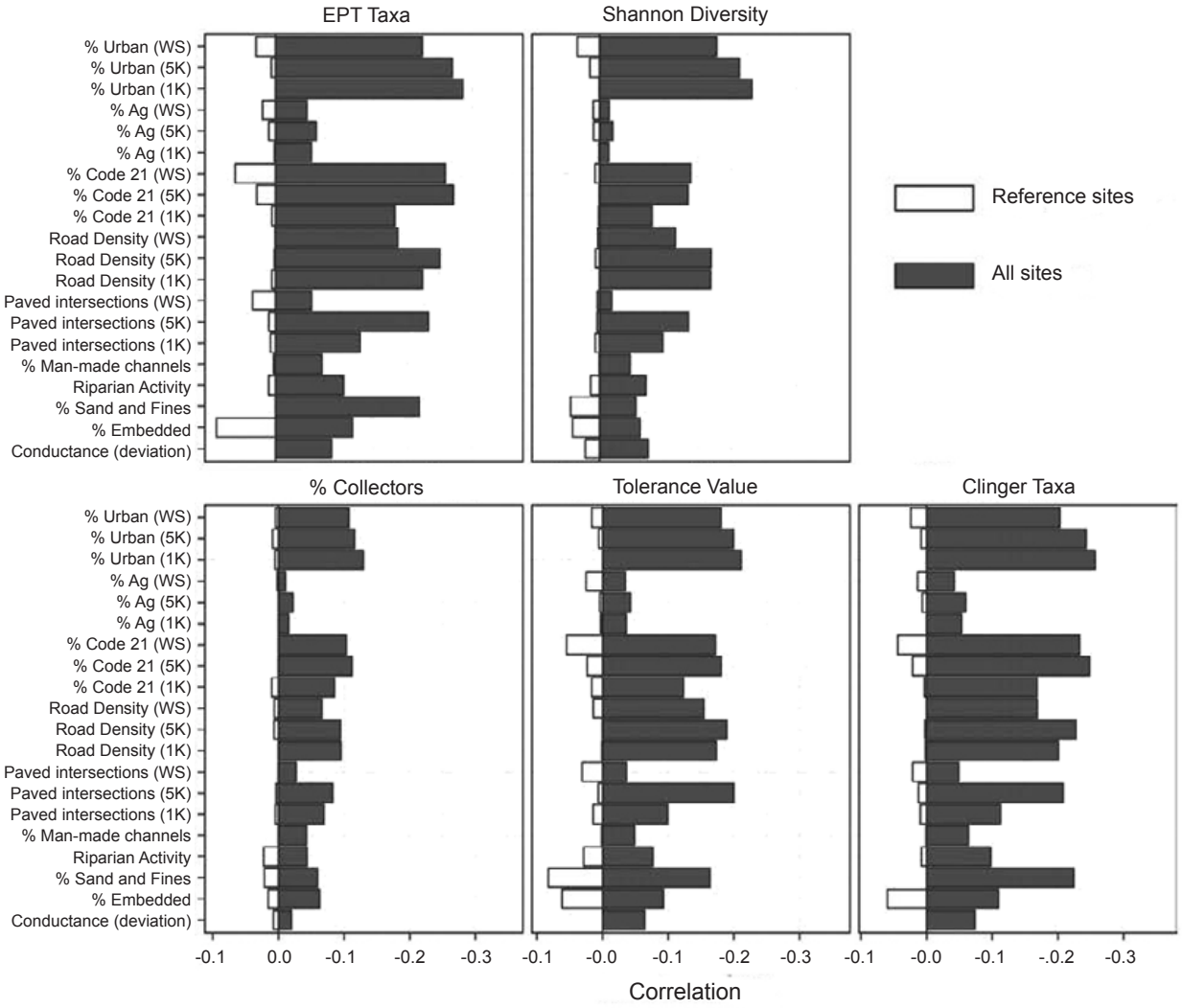

Figure 5. Butterfly plots illustrating the strength of correlations (Pearson's r) between several bioassessment indicators and common anthropogenic stressors. Open bars on the left of each plot indicate correlations measured at reference sites, and the solid bars on the right of each plot indicate correlations with all sites included.

levels did not have different biological quality from more disturbed reference sites.

Discussion

As the focus of water quality monitoring programs shifts toward increased emphasis on ecological condition (Rosenberg and Resh 1993, Davies and Jackson 2006, Collier 2011, Pardo *et al.* 2012), rigorous consideration of reference concepts can enhance multiple components of watershed management programs, including biological and 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 suited to the intended uses of the reference network (Bailey *et al.* 2004, 2012). Although programs developing and using reference site pools traditionally focus on minimizing degradation of reference site quality, representativeness may be just as important for many applications. In particular, we argue that explicit attention to environmental representativeness may help improve overall accuracy of condition assessments and reduce prediction bias (see Hawkins 2010a) in all reference condition applications.

Performance Summary

Our reference site selection process yielded a large data set, with 590 unique reference sites distributed throughout California. With the exception of the Central Valley, sites in the reference pool represented nearly the full range of all natural gradients evaluated. Thus, we have confidence that analyses and assessment tools developed from this reference pool are valid for the vast majority of perennial streams in California. Although our thresholds did not eliminate all anthropogenic disturbances from the reference pool, we demonstrated that the influence of these disturbances on the reference pool fauna was greatly minimized. Thus, impacts on ecological integrity are likely to be negligible and the balance of environmental representativeness and biological integrity is sufficient to support robust regulatory applications for wadeable perennial streams in California. Furthermore, although we anticipated would need

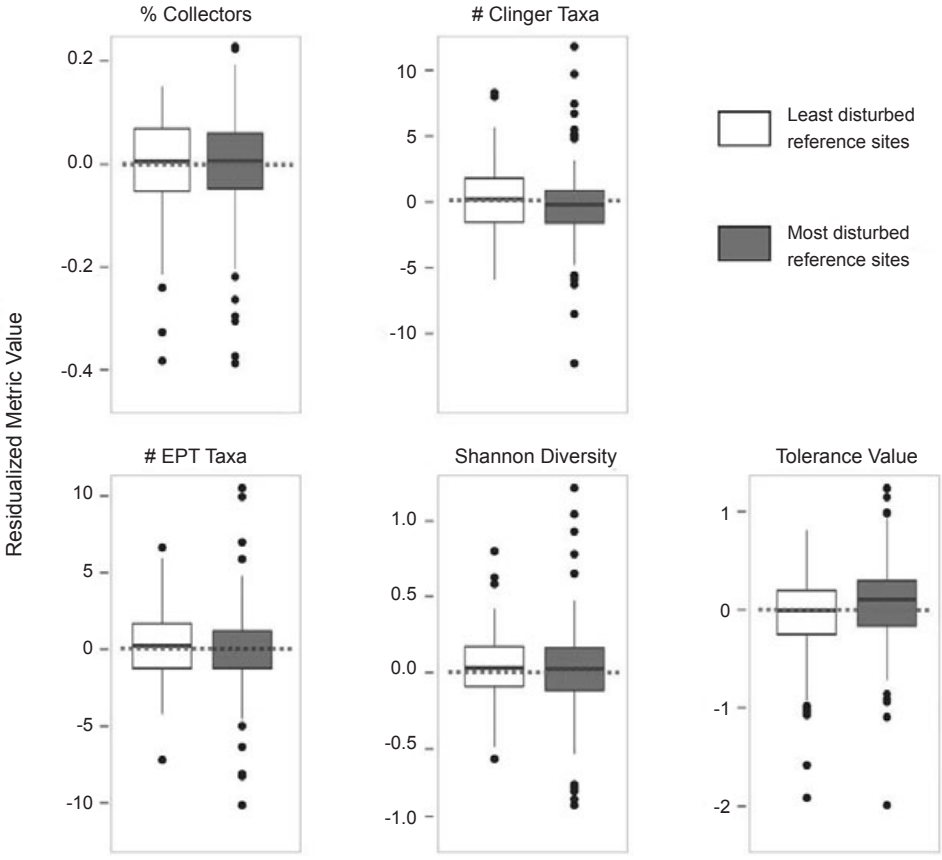

Figure 6. Boxplots comparing BMI metric scores at a subset of reference sites that passed very strict screens (open boxes, n = 294) and those that passed less strict screens (shaded boxes, n = 296). Significant differences were not observed for any comparison.

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., there is no need for region--specific threshold adjustments).

Managing Inter-Regional Complexity

Using the terminology of Stoddard *et al.* (2006), our reference pool could be viewed as a version of the "minimally disturbed" model, given the minimal amount of BMI metric response to stressors we observed. 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 to which the reference pool applies.

Because relaxing thresholds could potentially degrade the mean biological integrity of the reference pool, it is critical that relationships between biological integrity metrics and stressors be quantified in least- or minimally-disturbed regions. 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 selective exclusion of a subset of streams 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, Hill *et al.* 2013). Regardless of which alternate approach is used, benchmarks in excluded regions will need to be related to those used in minimally disturbed regions 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 pool has several potential applications for stream and watershed management. Reference concepts provide defensible regulatory frameworks for protecting and managing aquatic resources, and for providing a "common currency" for the integration of multiple biological indicators (e.g., algal, fish and BMI assemblages). The approach outlined in this paper is general and can be used to evaluate suitability of a reference pool for a wide range of habitat types, including non-perennial streams, lakes, depressional wetlands, and estuaries. Further, the process of defining reference criteria can also be used to identify streams and watersheds deserving of special protections and application of anti-degradation policies, which are often under-applied in the United States and elsewhere (Linke *et al.* 2011, Collier 2011).

Two general applications extend these uses to management of non-biological water quality constituents: 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 different water quality constituents with non-zero expected values (e.g., nutrients, chloride, conductivity, and fine sediment) is often more arbitrary than for novel pollutants that do not occur naturally, like pesticides. The range of concentrations 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; Hawkins *et al.* 2010a, 2010b) for a variety of physical and chemical endpoints, but management applications are rare. Comparisons of stressor values found at reference sites to the full range of 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 perspective needed to distinguish between relatively small differences in pollutant concentration and environmentally meaningful differences.

Limits of This Analysis

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) limited or imprecise 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., land cover, roads, hydrologic alteration), so impacts in 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) can 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 are collected along with biological samples. We included reach-scale anthropogenic disturbance data (W1_HALL) in our screens when available (~50% of sites), but we undoubtedly missed other locally important variables. Unintentional inclusion of stressed sites is likely to contribute to the total biological variability in our reference pool, but we anticipate this variability can be reduced over time as the availability and quality of stressor data sets improve.

Highly heterogeneous regions like California are likely to contain 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 intentionally 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 under-sampled (e.g., mountain streams > 3,000 m). However, our framework provides a means of explicitly testing the degree to which such stream types are represented by the overall network. Applicability of existing assessment tools to sites in these gaps may require further investigation.

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.

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. pp. 179-191 *in*: L.R. Brown, R.H. Gray, R.M. Hughes, and M.R. Meador (eds.), Effects of Urbanization on Stream Ecosystems. American Fisheries Society Symposium. Bethesda, MD.

Chessman, B.C. 1999. Predicting the macroinvertebrate faunas of rivers by multiple regression of biological and environmental differences. Fre*shwater 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 and 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. Chapter 35 *in*: Sierra Nevada Ecosystem Project: Final Report to Congress, Vol. II. University of California, Centers for Water and Wildland Resources. Davis, CA.

ESRI. 2009. Business Analyst. http://www.esri.com/software/businessanalyst.

European Commission. 2000. Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for community action in the field of water policy. *Official Journal of the European Communities* L327:1-73.

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.

Feio, M.J., F.C. Aguiar, S.F.P. Almeida, J.Ferreira, M.T. Ferreira, C.Elias, S.R.S. Serra, A. Buffagni, J.Cambra, C. Chauvan, F. Delmas, G. Dorflinger, S. Erba, N. For, M. Ferreol, M. Germ, L. Mancini, P. Manolaki, S. Marcheggianin, M.R. Miniciardi, A. Munne, E. Papastergiadou, N. Prat, C. Puccinelli, J. Rosebery, S. Sabater, S. Ciadamidaro, E. Tornes, I. Tziortzis, G. Urbanic and C. Vieira. 2013. Least disturbed condition for European mediterranean rivers. *Science for the Total Environment* doi. org/10.1016/j.scitoenv.2013.05.056.

Gasith, A. and V.H. Resh. 1999. Streams in Mediterranean climate regions: abiotic influences and biotic responses to predictable seasonal events. *Annual Review of Ecology and Systematics* 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.

Gesch, D., M. Oimoen, S. Greenlee, C. Nelson, M. Steuck and D. Tyler. 2002. The National Elevation Dataset. *Photogrammetric Engineering and Remote Sensing* 68:5-11.

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. US Environmental Protection Agency, Office of Water.. Washington, DC.

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.

Hill, R.A., C.P. Hawkins and D.M. Carlisle. 2013. Predicting thermal reference conditions for USA streams and rivers. *Freshwater Science* 32:39-55.

Horizon Data Systems. 2006. NHD Plus Version 1. http://www.horizon-systems.com/nhdplus/.

Horvitz, D.G. and D.J. Thompson. 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47:663-685.

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. http://www.epa.gov/nheerl/arm.

Kuhn, M., J. Wing, S. Weston, A. Williams, C. Keefer and A. Engelhardt. 2012. caret: Classification and Regression Training. R package version 5.15-045. The R Foundation for Statistical Computing. Vienna, Austria.

Liaw, A. and M. Wiener. 2002. Classification and Regression by random Forest. *R News* 2:18-22.

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

Lunde, K.B., M.R. Cover, R.D. Mazor, C.A. Sommers and V.A. Resh. In Press. Identifying reference conditions and quantifying biological variability within benthic macroinvertebrate communities in perennial and non-perennial northern California streams. *Environmental Management* DOI 10.1007/s00267-013-0057-1

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

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. pp. 1493-1502 *in*: Sierra Nevada Ecosystem Project: Final report to Congress, Vol. II. University of California, Centers for Water and Wildland Resources. Davis, CA.

Moyle, P.B. and P.J. Randall. 1996. Biotic integrity of watersheds. pp. 975-986 *in*: Sierra Nevada Ecosystem Project: Final report to Congress, Vol. II. University of California, Centers for Water and Wildland Resources. Davis, CA.

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, J., F. Guillaume Blanchet, Roeland Kindt, Pierre Legendre, Peter R. Minchin, R.B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry, H. Stevens and Helene Wagner. 2013. vegan: Community Ecology Package. R package version 2.0-7. http://CRAN.R-project.org/package=vegan

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.

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

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. pp. 83-108 *in*: R.J. Schmitt and C.W. Osenberg (eds.), 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 and 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. 2009. 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. 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., 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.

Rosenberg, D.M. and V. Resh. 1993. Freshwater Biomonitoring and Benthic Macro-Invertebrates. Chapman and Hall. New York, NY.

Sanchez-Montoya, M.M., M.R. Vidal-Abarca, T. Punti, J.M. Poquet, N. Prat, M. Rieradevall, J. Alba-Tercedor, C. Zamora-Munoz, M. Toro, S. Robles, M.

Alvarez and M.L. Suarez. 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.

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

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.

US Geological Survey. 2009. National hydrography dataset (high resolution), digital data. http://nhd.usgs.gov/data.html.

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.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.

Acknowledgements

The analyses reported here were developed in support of California's biological assessment program and serve as the foundation of the State's regulatory biological criteria. Financial support for this effort was provided by grants from the US Environmental Protection Agency (USEPA) Region IX and the California State Water Resources Control Board. The development process was supported by stakeholder and regulatory development advisory groups, whose contributions strongly influenced the objectives and approach we used. We are especially grateful to the considerable contributions of members of our scientific advisory panel who provided constructive guidance throughout the project: Dave Buchwalter, Rick Hafele, Chris Konrad, LeRoy Poff, John Van Sickle and Lester Yuan. We further thank John Van Sickle for valuable editorial contributions, Kerry Ritter for statistical support, and Joseph Furnish for constructive advice throughout the process. This work would not have been possible without 10 years of efforts of field crews, taxonomists and program staff from multiple state and federal monitoring programs, including the US Forest Service, USEPA, California Department of Fish and Wildlife, State and Regional Water Resources Control Boards and the Stormwater Monitoring Coalition.