2016

Evaluating the adequacy of a reference-site pool for ecological assessments in environmentally complex regions

Charles P Hawkins

Evaluating the adequacy of a reference-site pool for ecological assessments in environmentally complex regions

Peter R. Ode[1,7], Andrew C. Rehn[1,8], Raphael D. Mazor[1,2,9], Kenneth C. Schiff[2,10], Eric D. Stein[2,11], Jason T. May[3,12], Larry R. Brown[3,13], David B. Herbst[4,14], David Gillett[2,15], Kevin Lunde[5,16], and Charles P. Hawkins[6,17]

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

[2]Southern California Coastal Water Research Project, 3535 Harbor Boulevard, Suite 110, Costa Mesa, California 92626 USA

[3]US Geological Survey, 6000 J Street, Sacramento, California 95819 USA

[4]Sierra Nevada Aquatic Research Laboratory, 1016 Mount Morrison Road, Mammoth Lakes, California 93546 USA

[5]San Francisco Bay Regional Water Quality Control Board, 1515 Clay Street, Oakland, California 94612 USA

[6]Western Center for Monitoring and Assessment of Freshwater Ecosystems, Department of Watershed Sciences, and the Ecology Center, Utah State University, Logan, Utah 84322-5210 USA

Abstract: Many advances in the field of bioassessment have focused on approaches for objectively selecting the pool of reference sites used to establish expectations for healthy waterbodies, but little emphasis has been placed on ways to evaluate the suitability of the reference-site pool for its intended applications (e.g., compliance assessment vs ambient monitoring). These evaluations are critical because an inadequately evaluated reference pool may bias assessments in some settings. We present an approach for evaluating the adequacy of a reference-site pool for supporting biotic-index development in environmentally heterogeneous and pervasively altered regions. We followed common approaches for selecting sites with low levels of anthropogenic stress to screen 1985 candidate stream reaches to create a pool of 590 reference sites for assessing the biological integrity of streams in California, USA. We assessed the resulting pool of reference sites against 2 performance criteria. First, we evaluated how well the reference-site pool represented the range of natural gradients present in the entire population of streams 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 comparing biological metric scores at reference sites with the most vs fewest potential sources of stress. Using this approach, we established a reference-site pool with low levels of human-associated stress and broad coverage of environmental heterogeneity. This approach should be widely applicable and customizable to particular regional or programmatic needs.

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

Many of the refinements to biological monitoring techniques over the past 30 y have centered on strengthening the theoretical and practical basis for predicting the biological expectation for a given location 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. 2010b, Dallas 2013). The need to anchor biological expectations to a reference condition is now often regarded as essential. However, discussion regarding how to evaluate whether the properties of a pool of reference sites are adequate for its intended uses has been rare (Herlihy et al. 2008).

Authors of many recent treatments of the reference-site selection process recognize that objective criteria can greatly enhance the consistency of reference-condition determinations (Whittier et al. 2007, Herlihy et al. 2008, Yates and Bailey 2010, Dobbie and Negus 2013, Lunde et al. 2013), and examples of objective site-selection processes are increasingly common (e.g., Hawkins et al. 2000,

E-mail addresses: [7]peter.ode@wildlife.ca.gov; [8]andy.rehn@wildlife.ca.gov; [9]raphaelm@sccwrp.org; [10]kens@sccwrp.org; [11]erics@sccwrp.org; [12]jasonmay@usgs.gov; [13]lrbrown@usgs.gov; [14]david.herbst@lifesci.ucsb.edu; [15]davidg@sccwrp.org; [16]kevin.lunde@waterboards.ca.gov; [17]chuck.hawkins@usu.edu

DOI: 10.1086/684003. Received 4 November 2014; Accepted 3 February 2015; Published online 24 September 2015.
Freshwater Science. 2016. 35(1):237–248.

Stoddard et al. 2006, Collier et al. 2007, Whittier et al. 2007, Sánchez-Montoya et al. 2009, Yates and Bailey 2010). Several different approaches reflecting philosophical differences of practitioners and the varied monitoring questions each program addresses exist for identifying reference sites (e.g., Herlihy et al. 2008, Sánchez-Montoya et al. 2009, Yates and Bailey 2010). Programs in which biological integrity is assessed often call for a 'minimally disturbed' or 'least disturbed' standard (sensu Stoddard et al. 2006) for selecting reference sites because truly pristine streams are rare or nonexistent throughout the world. The main challenge is to choose site-selection criteria that retain sites with the highest biological integrity possible, thereby maintaining the philosophical ideal of the reference condition. However, geographic variability in the importance of different stressors that affect biological condition can complicate the establishment of uniform reference definitions (Statzner et al. 2001, Herlihy et al. 2008, Mykrä et al. 2008, Ode et al. 2008).

Robust reference-site selection involves balancing 2 potentially conflicting goals: 1) reference criteria should select sites that uniformly represent the least disturbed conditions throughout the region(s) of interest, minimizing the effects of anthropogenic stress on the indicator of interest, and 2) reference sites should represent the full range of environmental settings in the region in sufficient numbers to adequately characterize natural variability in the indicator(s) of interest. Restrictive criteria may minimize anthropogenic stress within the reference pool 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 2013, Feio et al. 2014). On the other hand, relaxing criteria to allow enough sites in the reference-site pool weakens the ability to detect deviation from the natural biological state. The consideration of environmental representativeness is especially critical in regulatory applications where errors in estimating site-specific reference expectations may have significant financial and resource-protection consequences. Evaluating the influence of the selected reference criteria on characteristics of the reference-site pool allows scientists and resource managers to make informed decisions about this balance.

We describe an approach used to evaluate the adequacy of a reference-site pool for assessing biological condition of streams in California, an environmentally complex region of the USA overlain with large areas of pervasive development. Our work is built 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). We followed common approaches to identify reference sites, then conducted an extensive characterization of the pool of reference sites, with an emphasis on assessing how well the natural diversity at streams in the region was represented by the reference-site pool and whether the biological integrity of the pool was reduced when maximizing representativeness.

METHODS

We assembled a set of 1985 candidate reference sites representing a wide range of stream types to support development of screening criteria. We characterized each site with a suite of landuse and land-cover metrics that quantified both its natural characteristics and potential anthropogenic stressors at the site or in its upstream drainage area. We then screened sites with a subset of landuse metrics (e.g., road density and % urban land use in the upstream watershed) based on thresholds that represented low levels of anthropogenic activity (least disturbed sensu Stoddard et al. 2006). We evaluated the pool of reference sites that passed screening criteria 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 ~280,000 km long, and 30% of the length is perennial according to the National Hydrography Dataset (NHD) medium-resolution (1 : 100k) stream hydrology data set (http://nhd.usgs.gov/data.html). These streams drain a large (424,000 km^2) and remarkably diverse landscape. California spans latitudes between 33 and 42°N, and its geography is characterized by extreme natural gradients. It encompasses both the highest and lowest elevations in the conterminous US, temperate rainforests in the northwest, deserts in the northeast and southeast, and a Mediterranean climate in most remaining regions. California's geology is also complex, with recently uplifted and poorly consolidated marine sediments in the Coast Ranges, alluvium in its broad internal valleys, granitic batholiths along the eastern border and recent volcanic lithology in the northern mountains. The state's environmental heterogeneity is associated with a high degree of biological diversity and endemism in the stream fauna (Erman 1996, Moyle and Randall 1996, Moyle et al. 1996).

California's natural diversity is accompanied by an equally complex pattern of land use. The natural landscapes of some regions of the state have been nearly completely converted to agricultural or urban land uses (e.g., the Central Valley and the South Coast) (Sleeter et al. 2011). Other regions are still largely natural but contain pockets of agricultural and urban land use and support timber harvest, livestock grazing, mining, and intensive recreational uses. Our analyses generally treated environmental variation as continuous, but to facilitate some assessments, we divided the state into 6 regions, 4 of which

were further divided into subregions (north coast, Central Valley, chaparral [coastal and interior], Sierra Nevada [western and central Lahontan], south coast [xeric and mountains], deserts + Modoc) based on modified ecoregional (Omernik 1987) and hydrological boundaries (Fig. 1).

Aggregation of site data

We inventoried >20 federal, state, and regional monitoring programs to assemble the data sets used for screening reference sites. Candidate data sets were mostly restricted to wadeable, perennial streams, but some nonwadeable rivers were included, as were some nonperennial streams, because of unavoidable imprecision in the assignment of flow status to stream reaches. All 1985 unique sites (Fig. 1) were sampled between 1999 and 2010, and resulting data were compiled into a single database. We considered sites sampled within 300 m of one another to be replicates and used only the most recent sample.

We used physical habitat data to characterize gradients related to natural (e.g., slope) and anthropogenic (e.g., riparian disturbance) factors (Tables S1, S2). All physical habitat data were collected with standard protocols from the US Environmental Protection Agency's [EPA's] Environmental Monitoring and Assessment Program protocol (EMAP; Peck et al. 2006) or California's modification of EMAP protocols (Ode 2007). For calculation of reach-scale physical-habitat metrics, we followed Kaufmann et al. (1999).

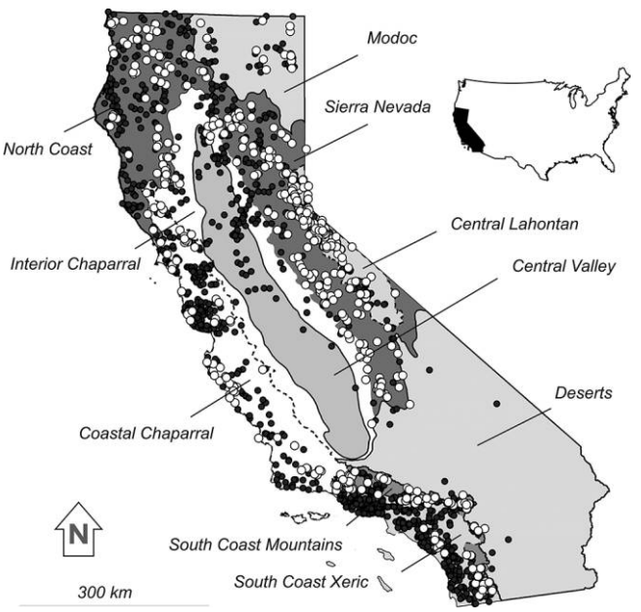

Figure 1. Distribution of 1985 candidate sites screened for inclusion in California's reference pool. White circles represent passing sites and black circles represent sites that failed ≥1 screening criteria. Thick solid lines indicate boundaries of major ecological regions referred to in the text. Lighter dashed lines indicate subregional boundaries.

Most benthic macroinvertebrate (BMI) data used to evaluate the extent of degradation within the reference-site data set were collected following the EMAP reach-wide protocol, but some older data were collected following the EMAP targeted-riffle protocol (Peck et al. 2006). Previous studies have shown these protocols produce similar bioassessments in the western USA (Ode et al. 2005, Gerth and Herlihy 2006, Herbst and Silldorff 2006, Rehn et al. 2007, Mazor et al. 2010). Prior to all analyses, BMI data were converted to standard taxonomic effort levels (generally genus-level identifications except chironomid midges were identified to subfamily; see Richards and Rogers 2006) and subsampled when necessary to 500-count.

Combination of probability data sets

We used data from a subset of sites (919 of 1985 sites) that were sampled under probabilistic survey designs to evaluate whether our final pool of reference sites adequately represented the full range of natural stream settings in California. Probability data sets 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, we created a common sample frame such that the relative contribution of each site to the overall distribution of stream length (the site's weight) could be calculated in the combined data set. All probabilistic sites were registered to a uniform stream network (NHD Plus version 1; http://www.horizon-systems.com/nhdplus/) and attributed with strata defined by the design parameters of all integrated programs (e.g., land use, stream order, survey boundaries, etc.). Second, we calculated site weights for each site by dividing total stream length in each stratum (e.g., all 2^{nd}-order streams draining agricultural areas in the north coast region) by the number of sampled sites in that stratum using the *spsurvey* package (Kincaid and Olsen 2009) in R (version 2.11.1; R Project for Statistical Computing, Vienna, Austria).

Geographical information system (GIS) data and metric calculation

We assembled a large number of spatial data sources to characterize natural and anthropogenic gradients that may affect biological condition at each site, e.g., land cover and land use, road density, hydrologic alteration, mining, geology, elevation, and climate (Table S1). We evaluated data sets for statewide consistency and excluded layers with poor or variable reliability. All spatial data sources were publicly available except for the roads layer, which was customized for this project by appending unimproved

and logging road coverages obtained from the US Forest Service and California Department of Forestry and Fire Protection to a base roads layer (ESRI 2009).

We converted land cover, land use, and other measures of human activity into metrics (Table S2) expressed within the entire upstream drainage area of the site (watershed), within 5 km upstream (5k) and within 1 km upstream (1k). We created polygons defining these spatial units with ArcGIS tools (ESRI 2009). Local polygons were created by intersecting a 5-km- or 1-km-radius circle centered at the stream site with the primary watershed polygon. We calculated metrics associated with sampling location (e.g., mean annual temperature, elevation, NHD+ attributes, etc.), based on each site's latitude and longitude. Data for all screening variables were available for all sites, except for W1_HALL (Kaufmann et al. 1999), a field-based quantitative measure of anthropogenic stressors at the reach scale that was available for ~½ of the sites (Table S2). We also calculated predicted electrical conductivity (Olson and Hawkins 2012), a site-specific estimate of natural background conductivity based on modeled relationships between observed conductivity and a suite of natural geographic, geological, climatic, and atmospheric variables (Table S2).

Selection of stressor screening variables and thresholds

To maximize the naturalness of the reference-site pool, we eliminated sites that exceeded specific thresholds for human activity (Table 1). Failure of any one screen was sufficient to eliminate a candidate site from the reference pool. Strict initial screening criteria (human influence variables set at 0) resulted in a set of only ~100 reference sites that occurred almost exclusively in mountainous regions (Sierra Nevada and North Coast Mountains) and that poorly represented most streams in California. We then relaxed thresholds after consulting reports from prior reference development projects (Ode et al. 2005, Rehn et al. 2005, Stoddard et al. 2006, Rehn 2009) and the literature (e.g., Collier et al. 2007, Angradi et al. 2009, Falcone et al. 2010) for previously established criteria, except for specific conductance. For specific conductance, we rejected sites whose observed conductance values fell outside the 1 and 99% prediction interval. If the prediction was >1000 μS/cm, we used a fixed rejection threshold of >2000 μS/cm. The new goal for the screening criteria was maximum representativeness in all regions of the state with the least relaxation of human-influence criteria possible.

Sensitivity of site counts to different screening thresholds The relative dominance of different stressors and their contribution to overall disturbance at candidate sites vary regionally. To explore the effect of threshold adjustments on site counts in different regions, we adjusted thresholds for each primary metric individually while all others were held constant and plotted the number of passing sites (i.e., threshold sensitivity) for each region. These partial-dependence curves were used to evaluate the number of reference sites potentially gained by relaxing thresholds for each screening metric in each region (see

Table 1. Distribution of reference and nonreference sites (number [*n*] and %) failing different numbers of thresholds (each screen and spatial-scale combination is counted independently). Number of streams and extent of stream length estimated to be reference by region (% ref ± 1 SE) based on probability data only.

Region	Total stream length (km)	Reference				Nonreference		% of nonreference sites failing (thresholds)		
		n	% of sites	% of stream length	SE	*n*	% of sites	1–2	3–5	≥5
North Coast	9278	76	31	26	3	168	69	26	57	18
Chaparral	8126	93	22	19	4	334	78	44	17	39
Coastal Chaparral	5495	61	18	14	5	275	82	47	16	37
Interior Chaparral	2631	32	35	28	6	59	65	34	22	44
South Coast	2945	119	18	23	4	555	82	22	10	68
South Coast Mountains	1123	86	42	53	7	121	58	62	23	15
South Coast Xeric	1821	33	7	3	1	434	93	11	6	83
Central Valley	2407	1	1	2	2	69	99	1	7	91
Sierra Nevada	11,313	276	56	43	5	218	44	56	26	18
Western Sierra Nevada	8577	131	53	34	6	118	47	58	29	14
Central Lahontan	2736	145	59	76	5	100	41	54	23	23
Deserts + Modoc	2531	25	33	32	10	51	67	51	29	20
Total	36,599	590	30	29	2	1395	70	33	20	47

Hill et al. 2013 for a similar example). We distinguished stressor variables with thresholds whose adjustment had a large influence on the number of accepted reference sites (and therefore, might improve overall environmental representativeness of the reference pool) from variables whose threshold adjustment had little influence on final site numbers. In the latter case, thresholds could be kept strict to minimize the risk of compromising biological integrity.

Performance measures

Evaluation of reference-site-pool representativeness We evaluated 2 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). We compared the dispersion of reference sites to the distribution of natural gradients in each region (as estimated from our probability distributions) and in multivariate environmental space described by a principal components analysis (PCA)-based ordination. We used 11 natural gradients that are associated with benthic invertebrate composition in California streams (Mazor et al. 2016; listed in Table S2) in the PCA analysis.

Evaluation of anthropogenic stress in the reference-site pool All thresholds allowed at least some degree of upstream human activity, so we determined if these stress levels were biologically important by assessing the responsiveness of a set of common BMI metrics to different stress-related variables. We used *t*-tests to determine if means of BMI metrics at a subset of sites passing more stringent screens were different from those at sites passing only 'standard' screens (Table 2). The more stringent screens were: <1% nonnatural land use (agricultural, urban, or Code 21 [a development-associated vegetation class in the NLCD data set that corresponds to lawns and recreational grasses in urban areas and roadside vegetation in rural and exurban areas]) at all spatial scales; road density <1 km/km^2 for all spatial scales; W1_HALL < 0.5; and all other criteria as listed in Table 2.

BMI metric values indicative of healthy biological condition vary naturally in different environmental settings. Natural variation could reduce our ability to discern biologically meaningful differences between stringent and less stringent reference groups. To correct for this potential confounding influence and to apply a more conservative test of the null hypothesis (no difference between groups), residuals from random-forest models of metric response to natural gradients were used in *t*-tests instead of raw metrics. Lack of significant differences in residuals between the high and low threshold groups was taken as evidence that biological integrity was not sacrificed by the less strict thresholds.

Table 2. Thresholds used to select reference sites. Scale refers to spatial area of analysis (WS = upstream watershed, 1k = watershed area within 1 k of site, 5k = watershed area within 5 k of site). NA = not applicable, W1_HALL = Index of human disturbance.

Variable	Scale	Threshold	Unit
% agriculture	1k, 5k, WS	3	%
% urban	1k, 5k, WS	3	%
% agriculture + % urban	1k, 5k, WS	5	%
% Code 21	1k, 5k	7	%
	WS	10	%
Road density	1k, 5k, WS	2	km/km^2
Road crossings	1k	5	crossings
	5k	10	crossings
	WS	50	crossings
Dam distance	WS	10	km
% canals and pipelines	WS	10	%
In-stream gravel mines	5k	0.1	mines/km
Producer mines	5k	0	mines
Specific conductance	site	99/1[a]	Prediction interval
W1_HALL	site	1.5	_

[a] 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 underpredict 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.

RESULTS

Reference status by region

Of the 1985 sites evaluated for potential use as reference sites, 590 passed all screening thresholds (Table 1). The number of reference sites varied by region, with the highest concentrations in mountainous regions (e.g., the Sierra Nevada, North Coast, and South Coast Mountains). Lower elevation, drier subregions had fewer reference sites (South Coast Xeric = 33, Interior Chaparral = 32), and only a single reference site was identified in the Central Valley (near the boundary of the Interior Chaparral).

Based on probability survey data, 29 ± 2% (SE) of California's stream length was estimated to meet our reference criteria (Table 1). Streams that met reference criteria were most extensive in mountainous regions, contributing ~76 and 53% of the stream length in the Central

Lahontan and South Coast Mountain subregions, respectively. Only 2 to 3% of stream length in the Central Valley and the South Coast Xeric regions was estimated to meet reference criteria, 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 (n = 76) and Western Sierra (n = 131) regions, relatively limited extents of stream length met reference criteria in these regions (26 and 34% of stream length, respectively). These levels are similar to levels seen in Chaparral regions, suggesting that the abundance of reference sites in some regions is caused more by the large extent of perennial streams than lack of anthropogenic stressors in the region.

Sensitivity of site counts to threshold levels

Large regional differences were present in the number and types of stressor metrics that contributed to the removal of candidate sites from the reference pool (Table 1). For example, most nonreference sites in the Sierra Nevada, South Coast Mountains, Chaparral, and Desert + Modoc regions failed only 1 or 2 thresholds (typically road density and Code 21), but a large majority (i.e., >80%) of nonreference sites in the Central Valley and the South Coast Xeric regions failed ≥5 thresholds. For other regions, the percentage of streams failing 1 or 2 thresholds ranged from 26 to 47%.

Similar regional patterns were reflected in threshold sensitivity plots (Fig. 2A–D; all example metrics were calculated at the watershed scale). For example, adjusting thresholds for the landuse metrics, % agricultural and % urban (Fig. 2A, D), had little influence on the proportion of sites that passed reference screens in most regions, indicating that other screening thresholds were limiting. In contrast, even a modest increase of the threshold for Code 21 greatly increased the number of passing sites in most regions, especially in the North Coast, Chaparral, and South Coast (Fig. 2B). Threshold adjustments for road density had similarly large impacts in the North Coast, Chaparral, and Desert + Modoc regions (Fig. 2C). This sensitivity allowed us to selectively relax screening thresholds for road density and Code 21, thereby increasing the number of passing sites and improving representation in several regions, particularly in the Chaparral, a region with relatively few sites prior to the adjustment. We would have had to adjust many other metric thresholds concurrently to achieve a comparable result had we not identified this pattern of differential threshold sensitivity.

Reference-site representativeness

The large number of sites in our data set that came from probabilistic surveys (n = 919) allowed us to produce well resolved distribution curves for a suite of natural environmental factors in each region (Fig. 3A–F). For

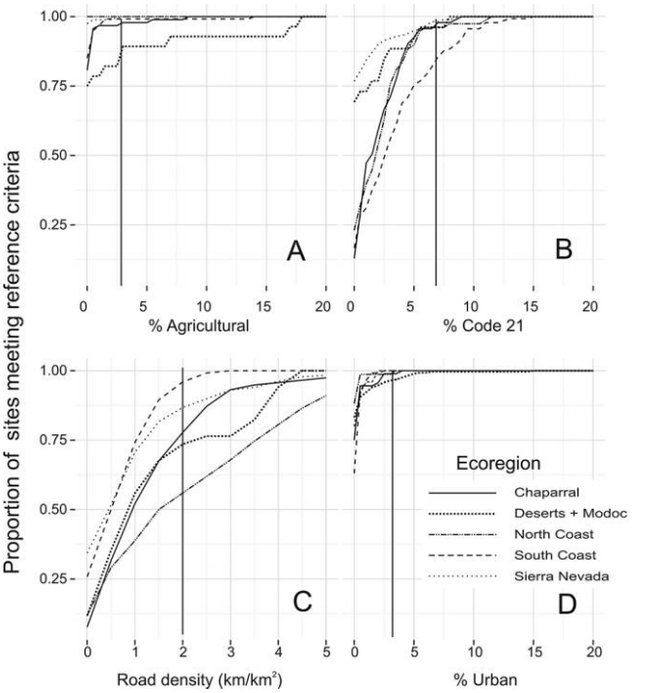

Figure 2. Example threshold sensitivity (partial dependence) curves showing the relationship between proportion of potential reference sites and thresholds for % agricultural (A), % Code 21 (B), road density (C), and % urban (D). All other stressors were held constant using the thresholds listed in Table 2. Vertical lines indicate reference thresholds for each metric.

nearly all natural factors and regions examined, the dispersion of reference-site values along environmental gradients matched the overall distribution of values for these gradients well. However, small but potentially important gaps were evident. For example, streams with very large watersheds (i.e., >500 km^2; Fig. 3A) and very high-elevation streams (i.e., >2500 m; Fig. 3B) and were represented by only a few reference sites. 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; Fig. 3B, D, E).

PCA of environmental variables provided additional evidence that the reference-site pool represented natural environmental gradients well (Fig. 4). Gaps generally were restricted to the extremes of gradients. For example, in a 2-dimensional plot of PCA Axes 1 and 2, a cluster of sites that lacked reference-site coverage was evident at the right side of PCA Axis 1 (Fig. 4), which corresponds to portions of the Central Valley and a group of low-gradient, low-elevation, low-precipitation, and large watershed streams in southern coastal California. Other axis combinations indicated similarly good coverage of natural environmental gradients and identified similar gaps.

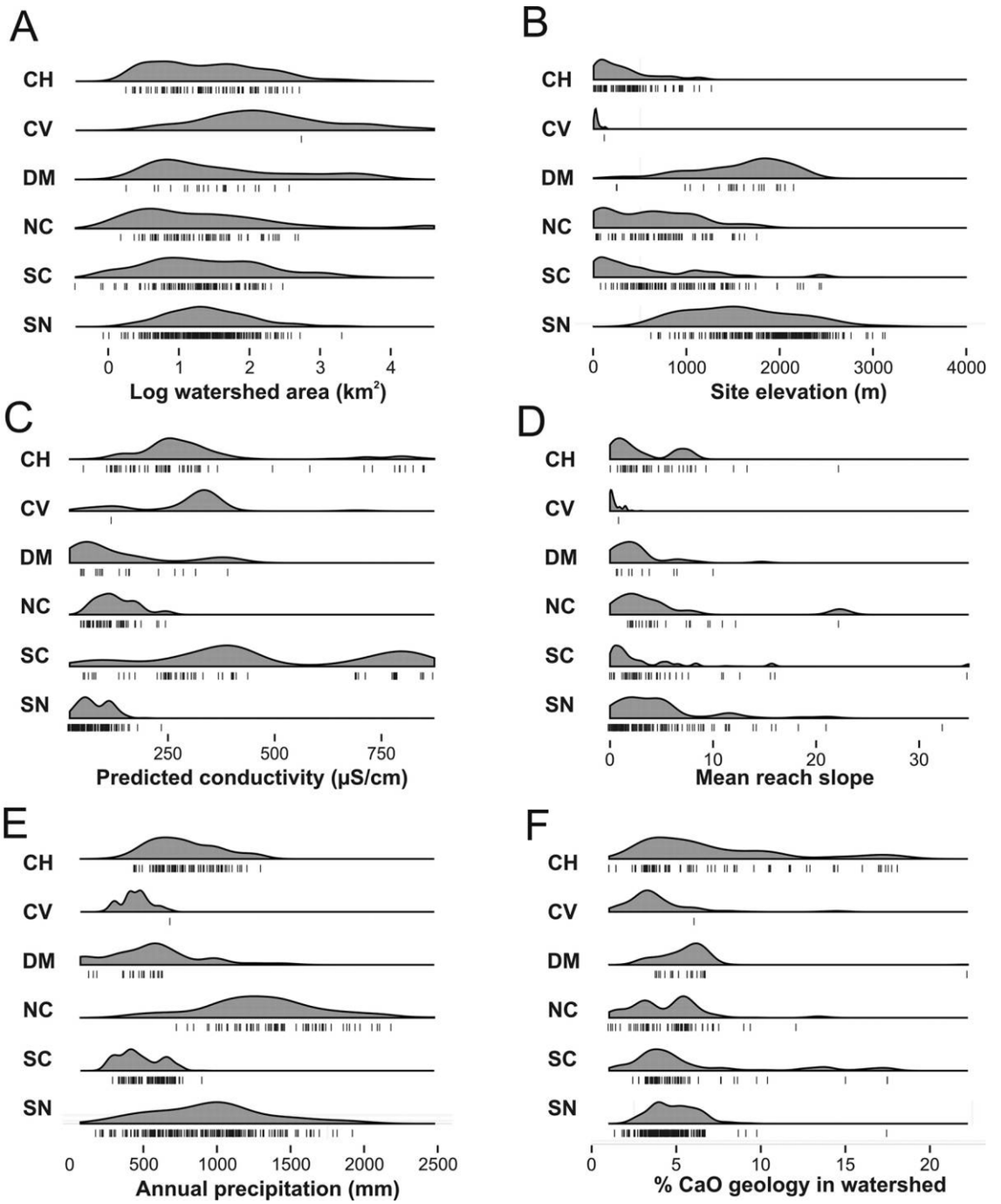

Figure 3. Comparison of reference-site representation along biologically influential natural gradients of watershed area (A), site elevation (B), conductivity (C), % reach slope (D), annual precipitation (E), and CaO geology in the watershed (F). Full distributions of natural gradients estimated from 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 Fig. 1) are abbreviated as: SN = Sierra Nevada, SC = South Coast, NC = North Coast, DM = Deserts + Modoc, CV = Central Valley, CH = Chaparral.

Biological response to stressors

BMI metric scores at reference sites that passed the most stringent screening criteria (n = 294) were indistinguishable from scores at those reference sites that passed more relaxed standard screens (Fig. 5) and were clearly different from scores at nonreference sites. All t-tests for differences in mean BMI metric scores between the 2 sets of reference sites were not significant (Fig. 5), indicating that we did not sacrifice biological integrity to achieve adequate representation of natural gradients.

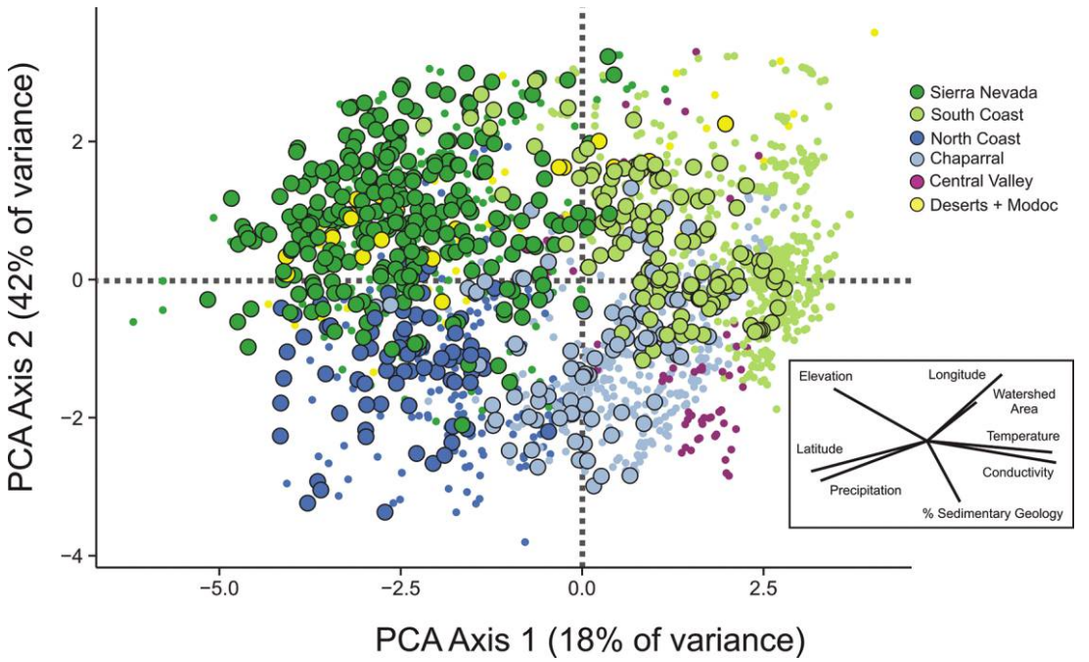

Figure 4. Principal components analysis (PCA) ordination of 1985 sites based on natural environmental characteristics (geology and climate variables listed in Table S2), showing the 2 primary principal component axes. Larger outlined circles indicate reference sites and smaller dots indicate nonreference sites. Colors represent regions shown in Fig. 1. The inset depicts vectors of selected natural variables as estimated from correlation with the PCA axes.

DISCUSSION

Rigorous consideration of reference concepts can enhance multiple components of watershed-management programs, including development and application of biological (e.g., indices of biotic integrity) and nonbiological (e.g., streambed substrate composition) endpoints. To ensure optimal use of reference-condition-based tools, program personnel need to evaluate whether selection criteria

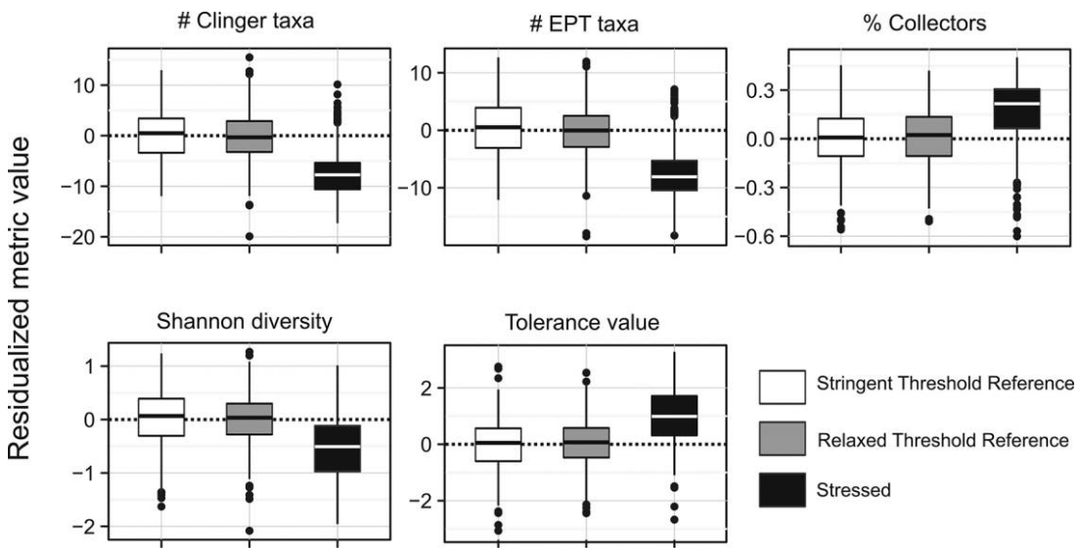

Figure 5. Boxplots comparing benthic macroinvertebrate (BMI) metric scores at a subset of reference sites that passed very strict screens (n = 292) and reference sites that passed less strict screens (n = 298). Significant differences ($p < 0.05$) were not observed for any comparison of the reference groups. Boxplots of nonreference, stressed sites (n = 613) are included for visual comparison. Lines in boxes are medians, box ends are quartiles, whiskers represent 1.5× the interquartile distance, and dots are outliers. EPT = Ephemeroptera, Plecoptera, Trichoptera.

produce a pool of reference sites suited to intended uses (Bailey et al. 2004, 2012, Herlihy et al. 2008). Our selection process yielded 590 unique reference sites that, except for the Central Valley, represented nearly the full range of all natural gradients evaluated. Thus, we have confidence that analyses and assessment tools developed from this pool of reference sites will be representative for most perennial streams in California. Our thresholds did not eliminate all anthropogenic disturbances from the reference-site pool, but the influence of these disturbances on the reference-site fauna was minimal, and the balance we achieved between environmental representativeness and biological integrity is sufficient to support robust regulatory applications for wadeable perennial streams in California. Furthermore, although we anticipated that we would 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, thereby maintaining interregional comparability. Thus, no need exists for region-specific threshold adjustments, and the complications they create for management interpretation can be avoided (see Herlihy et al. 2008, Yuan et al. 2008).

In the terminology of Stoddard et al. (2006), our reference-site pool was initially identified based on a least disturbed definition, but the sites probably are minimally disturbed given the limited response to watershed alteration in BMI metrics. The selective and systematic relaxation of reference screens allowed us to achieve broad representation of most perennial, wadeable streams in California with a single set of statewide reference criteria.

Applications of the reference-condition approach

A robust reference-site pool is needed to achieve several stream and watershed management objectives. Reference-condition concepts provide defensible regulatory frameworks for protecting and managing aquatic resources and a consistent basis for combining multiple biological indicators (e.g., algal, fish, and BMI assemblages) in integrated assessments. The process of defining reference criteria also can be used to help identify candidate streams and watersheds deserving of special protections and application of antidegradation policies, which are often under-applied in the USA and elsewhere (Collier 2011, Linke et al. 2011).

The reference-condition approach also is potentially useful in: 1) establishing objective regulatory thresholds for nonbiological indicators and 2) providing context for interpreting targeted and probabilistic nonbiological monitoring data. Establishment of regulatory standards for water-quality constituents that vary naturally in space and time (e.g., nutrients, Cl^-, conductivity, and fine sediment) can be arbitrary and contentious, especially compared to the process for establishing objectives for manufactured pollutants, like pesticides. The range of concentrations occurring at reference sites could be used to guide criteria development for physical and chemical pollutants with non-0 expectations (Hawkins et al. 2010a, b, Yates and Bailey 2010, Vander Laan et al. 2013). The physiochemical conditions found at reference sites can be used to predict the condition of test sites in a natural state (e.g., Vander Laan et al. 2013). Furthermore, the range of values expected in the natural reference state can give management program personnel the perspective needed to distinguish relatively small differences in pollutant concentration from environmentally meaningful differences. Ultimately, the broad success of these nonbiological applications will depend on rigorous evaluations of the reference data set, just as they do for biological applications of the reference concept.

Limits of this analysis

At least 3 types of data limitations can influence the adequacy of a reference-site pool: 1) inadequate or inaccurate GIS layers, 2) limited or imprecise information about reach-scale stressors, and 3) inadequate or uneven sampling effort. Improvements in the availability and accuracy of spatial data over the last 2 decades have greatly enhanced our ability to apply consistent screening criteria across large areas, but reliance on these screens can underestimate the amount of biological impairment that actually exists at a site (Herlihy et al. 2008, Yates and Bailey 2010). The most accurate and uniformly available spatial data tend to be associated with urban stressors (e.g., land cover, roads, hydrologic alteration), but estimates of recreation, livestock grazing, timber harvest, mining, and their probable effects on biota typically are under- and more variably estimated (Herbst et al. 2011). Other potential stressors, such as climate change and aerial deposition of nutrients or pollutants, are even more challenging to quantify and use to screen reference sites. Reach-scale (proximate) alterations can have a large influence on aquatic assemblages (e.g., Waite et al. 2000, Munn et al. 2009), but are difficult 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 disturbance at sites where reach-scale data were lacking. Unintentional inclusion of stressed sites probably affected biota in our reference-site pool, but we anticipate these effects can be reduced over time as availability and quality of stressor data sets improve.

Highly heterogeneous regions like California are likely to contain unique environmental settings (Erman 1996, Moyle and Randall 1996) that are infrequently sampled and might not be included in reference-site screening unless intentionally targeted. For example, we added additional reference sites with naturally high conductivity

when we identified a lack of sites at the high end of this gradient. We attempted to include as much environmental diversity as possible, but some stream types with unique physical or chemical characteristics probably were under-sampled (e.g., mountain streams >2500 m asl). However, our framework provides a means of explicitly testing the degree to which such stream types are represented by the overall pool. Applicability of existing assessment tools to sites in these gaps may require further investigation, and additional targeted sampling (e.g., in high-elevation headwater streams) is likely to yield needed data. In contrast, some data gaps occur in pervasively disturbed regions (e.g., the Central Valley) that are unlikely to yield additional sites.

We used objective reference criteria based largely on GIS-measured variables, but the approach we used for evaluating performance of the reference-site pool could be applied to other strategies for selecting reference sites, such as one that emphasizes field-measured criteria (e.g., Herlihy et al. 2008), or best professional judgment (e.g., Lunde et al. 2013). The approach outlined in our paper is general and can be used to evaluate the suitability of a reference-site pool for a wide range of habitat types, including nonperennial streams, lakes, depressional wetlands, and estuaries (e.g., Solek et al. 2010). For applications where different reference criteria are applied to different regions or stream types (e.g., Herlihy et al. 2008, Yuan et al. 2008, Yates and Bailey 2010), these analyses provide essential context for performing multiregion comparisons.

Conclusions

Increased attention has been paid in recent years to the importance of quantifying the performance of various components of bioassessment (Diamond et al. 1996, 2012, Cao and Hawkins 2011), particularly as they relate to comparability among data sets. This attention to performance validation is likely to facilitate the adoption of biological endpoints in water-quality programs worldwide. Similar attention to measuring the performance of reference-site pools relative to their intended uses also will be of significant benefit. In particular, explicit attention to environmental representativeness should help improve overall accuracy of condition assessments and reduce prediction bias (see Hawkins et al. 2010b) in all reference-condition applications.

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 EPA Region IX and the California State Water Resources Control Board's Surface Water Ambient Monitoring Program. 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 his valuable editorial contributions and Joseph Furnish for constructive advice throughout the process. We thank Tony Olsen, Tom Kincaid, and Kerry Ritter for help combining the multiple probability surveys. We also thank Christoph Matthaei and 2 anonymous referees for helpful comments that improved this manuscript.

This work would not have been possible without 10 y of effort by field crews, taxonomists, and program staff from multiple state and federal monitoring programs, including the US Forest Service, US EPA, California Department of Fish and Wildlife, State and Regional Water Resources Control Boards, and the Stormwater Monitoring Coalition.

LITERATURE CITED

Abramson, M. 2009. Tracking the invasion of the New Zealand mudsnail, *Potamopyrgus antipodarum*, in the Santa Monica Mountains. Urban Coast 1:21–27.

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, Massachusetts.

Bailey, R. C., G. Scrimgeour, D. Coté, 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.

Baker, M. E., M. J. Wiley, and P. W. Seelbach. 2001. GIS-based hydrologic modelling of riparian areas: implications for stream water quality. Journal of the American Water Resources Association 37:1615–1628.

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.

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. 2013. Ecological status assessment in Mediterranean rivers: complexities and challenges in developing tools for assessing ecological status and defining reference conditions. Hydrobiologia 719:483–507.

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.

Dobbie, M. J., and P. Negus. 2013. Addressing statistical and operational challenges in designing large-scale stream condition surveys. Environmental Monitoring and Assessment 185:7231–7243.

Erman, N. 1996. Status of aquatic invertebrates. Pages 987–1008 *in* Sierra Nevada Ecosystem Project: final report to Congress, Vol. II. Centers for Water and Wildland Resources, University of California Davis, Davis, California.

ESRI (Environmental Systems Research Institute). 2009. Business analyst. Environmental Systems Research Institute, Redlands, California. (Available from: http://www.esri.com/software/businessanalyst)

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. Chauvin, F. Delmas, G. Dörflinger, S. Erba, N. For, M. Ferréol, M. Germ, L. Mancini, P. Manolaki, S. Marcheggiani, M. R. Minciardi, A. Munné, E. Papastergiadou, N. Prat, C. Puccinelli, J. Rosebery, S. Sabater, S. Ciadamidaro, E. Tornés, I. Tziortzis, G. Urbanic, and C. Vieira. 2014. Least disturbed condition for European Mediterranean rivers. Science for the Total Environment 476/477:745–756.

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

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

Hawkins, C. P., R. H. Norris, J. Gerritsen, R. M. Hughes, S. K. Jackson, R. H. Johnson, and R. J. Stevenson. 2000. Evaluation of landscape classifications for biological assessment of freshwater ecosystems: synthesis and recommendations. Journal of the North American Benthological Society 19:541–556.

Hawkins, C. P., J. R. Olson, and R. A. Hill. 2010b. The reference condition: predicting benchmarks for ecological water-quality assessments. Journal of the North American Benthological Society 29:312–343.

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.

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.

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.

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. Research Ecology Branch, US Environmental Protection Agency, Corvallis, Oregon.

Kincaid, T., and T. Olsen. 2009. SPSurvey package for R. National Health and Environmental Effects Research Laboratory, Western Ecology Division, US Environmental Protection Agency, Corvallis, Oregon. (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.

Lunde, K. B., M. R. Cover, R. D. Mazor, C. A. Sommers, and V. A. Resh. 2013. Identifying reference conditions and quantifying biological variability within benthic macroinvertebrate communities in perennial and non-perennial Northern California streams. Environmental Management 51:1262–1273.

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.

Mazor, R. D., K. Schiff, K. Ritter, A. C. Rehn, and P. R. Ode. 2010. Bioassessment tools in novel habitats: an evaluation of indices and sampling methods in low-gradient streams in California. Environmental Monitoring and Assessment 167:91–104.

Mazor, R. D., A. C. Rehn, P. R. Ode, M. Engeln, K. Schiff, E. D. Stein, D. J. Gillett, D. B. Herbst, and C. P. Hawkins. 2016. Bioassessment in complex environments: designing an index for consistent meaning in different settings. Freshwater Science 35:249–271.

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

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

Moyle, P. B., R. M. Yoshiyama, and R. A. Knapp. 1996. Status of fish and fisheries. Pages 953–974 *in* Sierra Nevada Ecosystem Project: final report to Congress, Vol. II. Centers for Water and Wildland Resources, University of California Davis, Davis, California.

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 assemblages 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, California. (Available from http://www.swrcb.ca.gov/water_issues/programs/swamp/docs/swamp_sop_bio.pdf)

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.

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.

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. Pages 83–108 *in* R. J. Schmitt and C. W. Osenberg (editors). Detecting ecological impacts. Academic Press, New York.

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. Office of Research and Development, US Environmental Protection Agency, Corvallis, Oregon.

Rehn, A. C. 2009. 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 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.

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. Aquatic Bioassessment Laboratory, California Department of Fish and Game, Rancho Cordova, California. (Available from: http://www.waterboards.ca.gov/water_issues?programs/swamp/docs/reports/final_north_calif_ibi.pdf)

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, California. (Available from: http://www.safit.org/ste.html)

Sánchez-Montoya, M. M., M. R. Vidal-Abarca, T. Puntí, J. M. Poquet, N. Prat, M. Rieradevall, J. Alba-Tercedor, C. Zamora-Muñoz, M. Toro, S. Robles, M. Álvarez, and M. L. Suárez. 2009. Defining criteria to select reference sites in Mediterranean streams. Hydrobiologia 619:39–54.

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, California. (Available from: http://www.sccwrp.org)

Statzner, B., B. Bis, S. Dolédec, and P. Usseglio-Polatera. 2001. Perspectives for biomonitoring at large spatial scales: a unified measure for the functional classification of invertebrate communities in European running waters. Basic and Applied Ecology 2:73–85.

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.

Vander Laan, J. J., C. P. Hawkins, J. R. Olson, and R. A. Hill. 2013. Linking land use, in-stream stressors, and biological condition to infer causes of regional ecological impairment in streams. Freshwater Science 32:801–820.

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.

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 27:192–905.