Using Multiple Biological and Habitat Condition Indices for Bioassessment of California Streams

Prepared by: Andrew C. Rehn[1]

[1]*California Department of Fish and Wildlife*

SWAMP Technical Memorandum
SWAMP-TM-SB-2016-0003
July 2016

TABLE OF CONTENTS

KEY POINTS

- Multiple biological indicators (i.e, benthic macroinvertebrates, algae and physical habitat) provide a better assessment of condition than a single indicator used alone.
- Disagreements between indicators can support identification of stressors.
- Algal indices developed in southern coastal California can provisionally be used elsewhere in the state until statewide indices are available.

EXECUTIVE SUMMARY

A long-term goal of the SWAMP bioassessment program has been to use multiple indices of ecological condition in conjunction to produce more complete assessments of stream health than provided by any single index alone. Ideally, combined assessments should be based on different taxonomic assemblages, taking advantage of their different responses to various stressors deriving from upstream land use practices. In this study, the combined use of three ecological indices currently used to assess stream condition in California was explored. The indices used were the California Stream Condition Index (CSCI) based on benthic macroinvertebrates, the "H20" index based on diatoms and soft algae, and the California Rapid Assessment Method (CRAM) for riparian habitat condition. First, comparisons of index performance were used to assess whether cases of disagreement among indices indicate moderate levels of stress, to which some taxonomic assemblages (or physical habitat indicator in the case of CRAM) have responded but not others, or whether disagreement among indices was more likely to be "noise" due to poor performance in one or more index. H20 and CRAM did not perform as well as CSCI on a statewide scale for some performance measures, but often performed better than null CSCI, which was used as a benchmark for what would constitute poor index performance. Second, the frequency with which the three indices agreed and disagreed about site condition was assessed to identify whether cases of agreement and disagreement occur in systematic and predictable ways that relate to the particular stressor(s) affecting the site. Patterns of agreement and disagreement among the 3 indices were non-random: the indices frequently agreed that reference sites were not degraded, and that high-activity sites were degraded. Disagreements were most common at sites with moderate amounts of human activity. Where the indices disagreed, CSCI and CRAM were more sensitive to physical habitat stressors, whereas H20 was more sensitive to chemical stressors. The use of multiple indices in conjunction to infer the ecological condition of streams, each based on a different taxonomic assemblage or data type, greatly strengthens confidence in results from bioassessment surveys, reduces the likelihood of incorrect conclusions from sampling error or natural variability, and improves our ability to diagnose causes of degradation.

Introduction

Bioassessment has been used in California since 1994 to evaluate the ecological condition, or health, of streams and rivers throughout the state. For much of that time, benthic macroinvertebrates (BMIs) have been the taxonomic assemblage most frequently used to develop indices of ecological condition[1]. Those indices have typically been regional in scope, each covering only a portion of the state (e.g., Ode et al. 2005; Rehn 2009), although the recently developed California Stream Condition Index (CSCI, Mazor et al. 2016) has statewide applicability. While BMIs are powerful indicators of stream health because of their integrated response to multiple stressors over time and space, other assemblages, such as fish or algae, often respond differently to various stressors and restoration activities and over different time scales (Griffith et al. 2005; Resh 2008). A primary long-term goal of SWAMP's bioassessment program, and one that is in-line with recommendations from the U.S. Environmental Protection Agency (USEPA 2013), has been to develop multiple indices of stream condition so that results from different assemblages can be used in conjunction to produce a more complete and rigorous assessment of stream condition than provided by any single assemblage alone. At sites where multiple assemblages are all in agreement about condition, inference of human-caused alteration to the system, or lack thereof, is strengthened. By contrast, at sites where multiple assemblages are in disagreement, it may be possible to elucidate the effects of different stressors. Progress towards the goal of assessments based on more than one assemblage has recently been achieved in southern California through the development of ecological indices for diatoms and non-diatom (i.e., "soft") algae, and for the two in combination (Fetscher et al. 2014), and statewide through the development of a riverine wetland condition index using the California Rapid Assessment Method (CRAM; California Wetland Monitoring Workgroup 2013). The latter index (CRAM) is not based on a particular taxonomic assemblage, but instead integrates many aspects of riparian structure and disturbance into an overall condition score for the study area.

Several studies published over the last decade have compared the responses of different taxonomic assemblages to gradients of anthropogenic stress. In general, all of them found that different assemblages are sensitive to different stressors emanating from the same land use activities. For example, in a study of the effects of urbanization on streams in Victoria, Australia, diatoms and BMIs were both found to be sensitive to urban-derived impacts, but diatoms were better indicators of nutrient enrichment while BMIs were better indicators of catchment disturbance (Sonneman et al. 2001). Johnson et al. (2009) found similar results in lowland European streams, but in mountain streams found that BMIs were more sensitive to a nutrient gradient than macrophytes, diatoms, or fish, highlighting that response trajectories can differ among assemblages depending on stream type. Most of these studies have not compared the responses of composite, integrative indices of condition for each assemblage, such as indices of biotic integrity (IBIs), or

[1] Moyle and Randall (1998) developed an index for the Sierra based on native fish and frogs, and the EMAP Western Pilot (Stoddard et al. 2005) developed an index based on fish and riparian herpetofauna that was applicable across 12 western states. California monitoring programs have not included fish in statewide surveys since 2005.

observed-to-expected (O/E) ratios of taxonomic completeness. Rather, the responses of different assemblages are usually evaluated with multiple raw metrics, or by converting site-by-abundance data into ordination axes which are then used as the response variable. One of the few exceptions is Mazor et al. (2006), who found that indices based on BMIs were more sensitive than those based on diatoms to overall human disturbance in the Fraser River catchment, British Columbia.

Objectives

The purpose of this report is twofold. First, performance of three ecological indices currently used to assess stream condition in California is compared. Second, the frequency with which the three indices agree and disagree about site condition is assessed to identify whether cases of agreement and disagreement occur in systematic and predictable ways that relate to the particular stressor(s) affecting the site, thereby informing stressor identification and restoration options. The indices compared are the CSCI, the "H20" index based on soft algae and diatoms in conjunction (Fetscher et al. 2014), and CRAM. The CSCI and H20 are measures of in-stream biological condition, while CRAM combines in-stream and riparian measures to indicate the condition of habitat and overall stream setting. The full CRAM protocol measures four attributes (buffer and landscape context, hydrology, biotic structure and physical structure), scores each on a 25-100 scale, then averages individual attribute scores for a final CRAM score. The first two attributes include measures of human disturbance, whereas the second two attributes measure responses to disturbance. Because some performance measures evaluate index response to human disturbance, only the second two attributes were averaged here to produce a final "$CRAM_{bio\text{-}phys}$" score for comparisons to avoid circularity in evaluation of CRAM performance.

Part 1: Comparison of index performance

The first objective was to compare key performance measures of the 3 different indices. Comparisons of index performance were used to assess whether disagreement between indices indicates moderate levels of stress, to which some taxonomic assemblages (or riparian indicator in the case of CRAM) have responded but not others, or whether disagreement among indices was more likely to be "noise" due to poor performance in one or more index. Several performance measures of the CSCI were evaluated as part of its development, namely: accuracy, bias, precision, responsiveness and sensitivity (Mazor et al. 2016; see definitions in Appendix 1). Some, but not all, of these performance measures were evaluated for H20 (Fetscher et al. 2014) and for CRAM (Stein et al. 2009), but those evaluations were based on smaller data sets. In addition, H20 was developed for use in southern coastal California, but here was applied to statewide data sets, making evaluation of its performance in statewide assessments especially important. Therefore, performance measures used by Mazor et al. (2016) to evaluate the CSCI were applied to H20 and $CRAM_{bio\text{-}phys}$ to allow parallel comparison of all measures among all three indices.

Part 1 Data sets and analysis

Performance measures for the CSCI were taken directly from Mazor et al. (2016) and were based on calibration data described therein. For the H20 and $CRAM_{bio\text{-}phys}$ indices, performance measures were calculated by combining data collected by the Reference Condition Monitoring Program (RCMP), the Perennial Streams Assessment (PSA), and the Southern Monitoring Coalition (SMC) from 2008-2012 (Table 1). The RCMP targets high-quality (i.e., "reference") sites to define expected biological, chemical and physical conditions when human disturbance in the environment is absent or minimal. By contrast,

the PSA and SMC programs sample randomly selected (i.e., "probabilistic") sites that provide unbiased estimates of statewide and regional stream condition, some of which pass reference screening criteria. Some performance measures (e.g., accuracy and precision) required data from reference sites only, whereas others (e.g., responsiveness and sensitivity) required data from high-activity sites (see Appendices 2a and 2b for definitions of reference and high-activity sites, respectively).

Mazor et al. (2016) evaluated sensitivity of the CSCI as the percentage of high-activity sites that scored below the 10th percentile of reference sites, i.e., the threshold that defined the boundary between "possibly altered" and "likely altered" condition categories for that index[2]. For CSCI, which was calibrated so that the mean score at reference sites equals 1 and has an observed scoring range across sites of roughly 0.1 to 1.4, the threshold score was 0.79. The same approach was used to establish threshold scores for H20 and $CRAM_{bio\text{-}phys}$. For H20 (scored on a 0-100 scale), the threshold score was 60. For $CRAM_{bio\text{-}phys}$ (scored on a 25-100 scale), the threshold score was 63.

Development of the CSCI included statistical modeling that, based on an assessment site's unique environmental setting, allowed site-specific predictions of which BMI taxa, and what BMI metric values, were expected to occur there if the site were in reference condition (Mazor et al. 2016). Statistical modeling also helped reduce, or "factor out", responsiveness of CSCI to natural environmental gradients (e.g., stream size and elevation), which often co-vary with human disturbance gradients, thereby confounding index response to disturbance. Null models (Van Sickle et al. 2005), where every site was expected to have the same taxa, and where metric responsiveness to natural gradients was not factored out, were used to estimate the lowest possible precision for CSCI. Therefore, performance measures of CSCI null models (Mazor et al. 2016) were used here as a benchmark of poor performance in H20 and $CRAM_{bio\text{-}phys}$, since the latter indices were not developed using a modeling approach.

Table 1. Number of sites per condition category used for calculation of performance measures for three condition indicators. For CSCI, numbers of sites per category are as described in Mazor et al. (2016) for calibration data. For H20 and $CRAM_{bio\text{-}phys}$, sites in each category were obtained by combining data from RCMP, PSA and SMC as described in the text.

Index	Reference	Moderate-activity	High-activity	Sites w/ repeat visits
CSCI	473	626	491	220
H20	292	314	271	31
$CRAM_{bio\text{-}phys}$	285	344	333	14

[2] These categories were referred to as "fair" and "poor", respectively, in the latest statewide assessment of stream condition (Rehn 2015). Also see footnote 5 below.

Part 1 Results

$CRAM_{bio\text{-}phys}$ performed nearly as well as the predictive CSCI for some performance measures (Table 2). For example, $CRAM_{bio\text{-}phys}$ had a low proportion of variance explained by natural gradients at reference sites, among-site precision nearly as good as predictive CSCI, and proportion of variance explained by human disturbance gradients nearly identical to predictive CSCI. However, $CRAM_{bio\text{-}phys}$ showed bias among PSA regions (Fig. 1, center graph) as indicated by a statistically significant F-statistic from an ANOVA with PSA region as factor (Table 2, F = 4.2; p=0.003). *Post-hoc* analysis indicated that $CRAM_{bio\text{-}phys}$ scores at reference sites were significantly higher in the Chaparral and North Coast than in other regions. $CRAM_{bio\text{-}phys}$ was less sensitive than CSCI and H20 in all PSA regions except the Sierra Nevada, and was especially insensitive in the North Coast (Table 3).

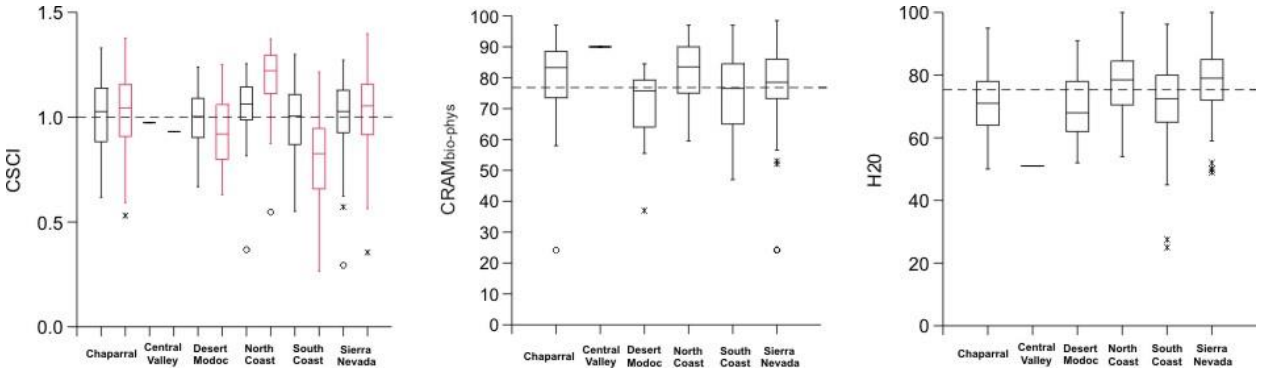

Figure 1. Distribution of index scores at reference sites by PSA region. CSCI box plots were modified from Mazor et al. (2016) based on calibration data; scores from CSCI null models are shown in red (left panel). $CRAM_{bio\text{-}phys}$ scores were significantly higher at reference sites in the Chaparral and North Coast. H20 scores were significantly higher at reference sites in the Sierra Nevada and North Coast. While bias among regions was not as pronounced as for null CSCI (red boxes in left panel), it was more than that observed for predictive CSCI, which did not show bias among PSA regions (block boxes in left panel). Dashed lines show mean score at reference sites for each index.

H20 had slightly better among- and within-site precision than predictive CSCI (Table 2), but had a higher proportion of variance at reference sites explained by natural gradients, and showed significant bias among PSA regions (Figure 1, right graph; F = 9.7; p<0.0001). *Post-hoc* analysis indicated that H20 scores at reference sites were significantly higher in the Sierra Nevada and North Coast than in other regions. H20 had better responsiveness than CSCI or $CRAM_{bio\text{-}phys}$ (Table 2) and was more sensitive than predictive CSCI in half of PSA regions and statewide, but was the least-sensitive index in the Sierra Nevada (Table 3). H20's better responsiveness and higher sensitivity than CSCI in some regions may have been a spurious (i.e., artificially good) result related to the fact that 17% of its variance among sites was explained by natural gradients that probably co-vary with human disturbance gradients, since statistical modeling was not used to factor out response to natural gradients during H20 development.

Table 2. Performance measures to evaluate indices (see definitions in Appendix 1). Only reference sites were used for accuracy and precision tests. Accuracy measures: F-statistic: The F-statistic for differences in reference site scores among 5 PSA regions (excluding the Central Valley which had only one reference site). Var: Variance in index scores explained by natural gradients at reference sites. Precision measures: Among sites: Standard deviation of scores at reference sites. Within sites: Pooled standard deviation of within-site residuals for reference sites with multiple samples. Note that for precision measures, $CRAM_{bio\text{-}phys}$ and H20 scores were adjusted to the same scale as CSCI by dividing by the mean reference score. Responsiveness measures: t-statistic: t-statistic for difference between mean scores at reference and high-activity sites. Var: Variance in index scores explained by human activity gradients at all sites.

		Bias & Accuracy		Precision		Responsiveness	
Index	Type	F-statistic	Var	Among sites	Within sites	t-statistic	Var
CSCI	Predictive	1.3 (p = 0.4)	-0.08	0.16	0.11	28.5	0.49
	Null	52.9 (p < 0.0001)	0.41	0.21	0.11	28.6	0.64
$CRAM_{bio\text{-}phys}$		4.2 (p = 0.003)	0.08	0.17	0.11	18.9	0.47
H20		9.7 (p < 0.0001)	0.17	0.15	0.09	30.5	0.57

Part 1 Conclusions

$CRAM_{bio\text{-}phys}$ and H20 did show some performance shortcomings relative to predictive CSCI, especially bias among PSA regions. For $CRAM_{bio\text{-}phys}$, bias may have been introduced by using just 2 of 4 CRAM attributes. Use of the full CRAM score reduced regional bias to non-significant levels, but as discussed above, introduced concerns of circularity in evaluation of responsiveness, since the buffer and hydrology attributes are partially redundant with measures of high activity. For H20, bias was likely introduced by statewide use of a southern California index. That said, there are several additional points to consider: 1) $CRAM_{bio\text{-}phys}$ and H20 performed as well as predictive CSCI for at least some performance measures, despite the fact that neither index was modeled to factor out response to natural gradients; 2) Performance comparisons may have been somewhat "apples-to-oranges", since performance measures were calculated from overlapping, but not identical data sets; 3) The mean $CRAM_{bio\text{-}phys}$ score at Chaparral reference sites was only 5 points higher than the mean reference score in "unbiased" regions, and in the North Coast was only 7 points higher than in unbiased regions; likewise, the mean H20 score at Sierra Nevada and North Coast reference sites was only 7 points higher than the mean reference score in unbiased regions. While these differences were statistically significant, rescaling of $CRAM_{bio\text{-}phys}$ and H20 scores to "correct" for bias among regions had little affect on, and was not incorporated into, results presented in Part 2 below. Therefore, given that $CRAM_{bio\text{-}phys}$ and H20 performed reasonably well on a statewide scale and that each is the state-of-the-art index for its respective type of indicator, disagreement between indices is more likely to indicate moderate levels of stress, to which some indices have responded but not others, rather than "noise" due to one or more index performing poorly. These relationships are evaluated further in Part 2.

Table 3. Sensitivity of the three indicators, i.e., the percentage of high-activity sites that have index scores below the 10th percentile of reference sites. CSCI values are for calibration data described in Mazor et al. (2016).

Region	CSCI	$CRAM_{bio\text{-}phys}$	H20
Statewide	76	66	89
North Coast	42	8	67
Chaparral	69	42	89
South Coast	86	73	95
Sierra Nevada	37	40	20
Central Valley	96	65	82
Desert-Modoc	100	50	75

Part 2: Frequency of Index Agreement and Disagreement

The second objective was to compare the frequency with which the three indices agreed and disagreed about site condition to identify whether cases of agreement and disagreement occurred in systematic and predictable ways according to which stressor(s) affect sites. Sites were first labeled as either "degraded" or "not degraded" for each index, where degradation was defined as a score below the 10th percentile of reference scores (see "*Data sets and analysis*" under Part 1 above for threshold definitions). Sites were also categorized as either agricultural, urban, forest or other according to upstream land use in the local and full upstream watershed, or as reference, moderate-activity or high-activity according to upstream land use and reach-scale criteria where available[3]. Finally, to associate biological and riparian condition with stressor conditions at a site, each site was categorized as "exceeding" or "not exceeding" thresholds for 11 different reach-scale stressors measured by bioassessment programs in California (Appendix 3). Attempts to evaluate index-stressor relationships based on more quantitative, continuous stressor variables were unsuccessful as described below.

Part 2 Data sets and analysis

Data from 628 probabilistic sites sampled by the PSA and SMC from 2008-2012 where all 3 indicators were present were used for agreement/disagreement assessments. Analyses were restricted to probabilistic sites because they provide an unbiased representation of the range of conditions across California watersheds. Reference sites sampled by the RCMP since 2008 were omitted from these analyses because 1) they were targeted based on minimal human disturbance in the upstream watershed and at the reach scale, and 2)

[3] Agricultural sites had ≥50% agricultural land use at either local or watershed scale; urban sites had ≥25% urban land use at either local or watershed scale; forest sites had ≥75% forest land cover at either local or watershed scale; "other" sites did not meet any of these criteria. Criteria for reference, moderate-activity and high-activity classifications are listed in Appendices 2a and 2b.

degradation for each index was defined as a score at the lower end of the reference distribution, meaning most reference sites were not degraded by definition. RCMP sites were therefore presumably biased towards agreement between all three indices that no degradation had occurred[4].

Part 2 Results

Full agreement between all 3 indices occurred at nearly 55% of sites (n = 344), with some form of disagreement between indices occurring at the remaining 45% of sites (n = 284; Table 4)[5]. Most sites where all 3 indices agreed that degradation had occurred (n = 159, or 25% of the total) were dominated by urban land use and had high levels of human activity in the upstream watershed. Conversely, most sites where all 3 indices agreed that no degradation had occurred (n = 185, or 29% of the total) were in watersheds with forested or "other" upstream land use, and nearly half were reference sites (i.e., not targeted reference sites from RCMP, but probabilistic sites that passed reference criteria). Sites with disagreement between indices were characterized by varied land use and human activity levels, but the majority had "other" land use in the upstream watershed and moderate levels of activity (Table 4). A chi-square goodness-of-fit test, assuming an equal number of sites in each of the three categories if agreement/disagreement patterns among indices were random, was highly significant (chi-square statistic = 41.57, df = 2, p <0.0001), indicating that agreement/disagreement patterns among indices were not random.

Geographically, the North Coast had the highest rate of agreement among indices, with all 3 indices agreeing that no degradation had occurred at the majority of sites; no North Coast sites were degraded for all 3 indices (Figure 2). The Sierra Nevada also had no sites degraded for all 3 indices, but all 3 indices had low sensitivity in that region (Table 3). It is also important to note that high-activity sites in the North Coast and Sierra Nevada are often less stressed than high-activity sites in other regions, usually failing just the road density criterion, whereas high-activity sites in other regions often fail multiple criteria. Conversely, the Central Valley had the highest rate of disagreement among indices, but also had the highest percentage of sites degraded for all 3 indices. The other regions (Desert-Modoc, Chaparral and South Coast) were somewhere between these extremes (Figure 2).

---

[4] This assumption was subsequently tested: all 3 indicators agreed that no degradation had occurred at 134 out of 165 RCMP sites (81%) sampled from 2008-2012. Twenty-eight RCMP sites were degraded for 1 indicator (either CSCI, $CRAM_{bio\text{-}phys}$, or H20), 2 sites were degraded for 2 indicators (CSCI and $CRAM_{bio\text{-}phys}$), and none was degraded for all 3.

[5] Mazor (2015) found only 40% agreement among multiple indices in a recent assessment of the SMC region, although that study kept soft algae ("S2") and diatom ("D18") indices separate (so was based on 4 indices), used the full CRAM index, and used the 2.5th percentile of reference sites as the degradation threshold for each index. In the SMC, 15% of stream length was degraded for all 4 indices, all in watersheds dominated by either agricultural or urban land use. Conversely, 25% of stream length was not degraded for all 4 indices, mostly in undeveloped watersheds. Degradation according to algae indices, but not CSCI or CRAM, was the largest source of disagreement.

Table 4. Summary of agreement and disagreement between CSCI, $CRAM_{bio\text{-}phys}$ and H20 at 628 statewide probability sites sampled 2008-2012, and land use characteristics of those sites.

Condition	*n*	Ag	Urban	Forest	Other	Reference	Moderate-activity	High-activity
All 3 degraded	159	24	108	0	27	0	20	139
All 3 not degraded	185	0	2	67	116	77	98	10
Sites w/ disagreement	284	16	78	40	150	34	141	109

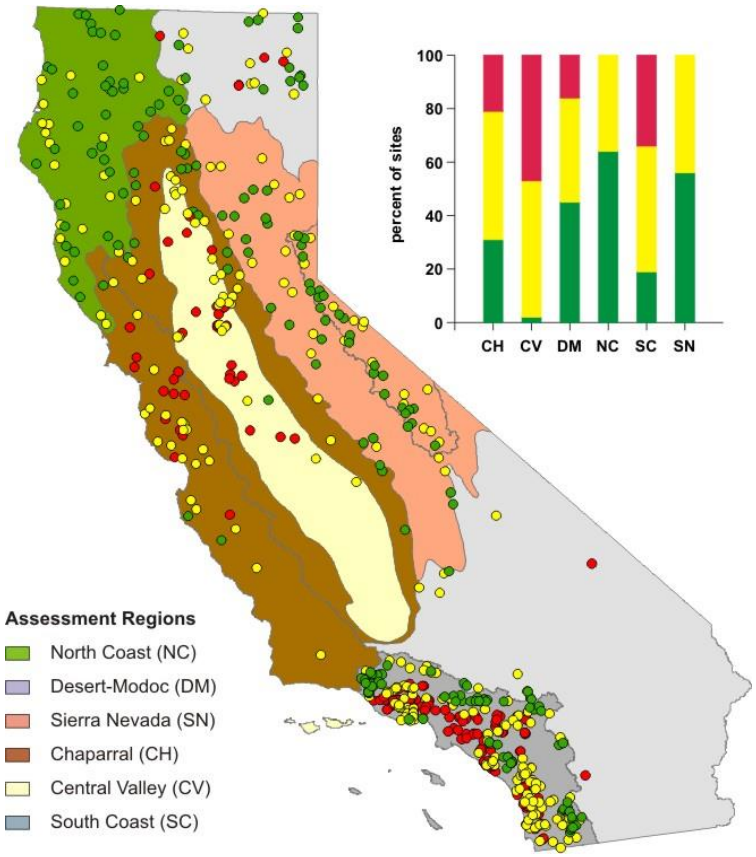

Figure 2. Geographic patterns of agreement/disagreement among CSCI, H20 and $CRAM_{bio\text{-}phys}$ at 628 probabilistic sites, and the percentage of sites per PSA region where agreement/disagreement occurred (bar chart inset). In the map, green dots are sites where all 3 indices agreed no degradation had occurred, red dots are sites where all 3 agreed degradation had occurred, and yellow dots are sites where indices disagreed. The same color-coding applies to percentage bars in bar-chart inset.

The subset of 284 sites where disagreement between indices occurred was divided into disagreement classes (Table 5). Patterns in the number of stressor exceedences were then evaluated for each disagreement class. For example, the first disagreement class listed in Table 5 comprises 24 sites where CSCI and $CRAM_{bio\text{-}phys}$ indicated degradation, but H20 scores indicated no degradation. Fourteen of those 24 sites had chemical exceedences, 22 had PHAB exceedences, 13 had both, 1 had *only* chemical exceedences, and 9 had *only* PHAB exceedences. Disagreement patterns among each of the 6 independent disagreement classes were moderately informative, but the real signal emerged when classes were combined (see the summaries in the last 3 rows of Table 5). For example, at the 69 sites where CSCI indicated degaradation but H20 did not

(and ignoring $CRAM_{bio\text{-}phys}$), only 5 of those sites had only chemical exceedences, whereas 30 of them had only PHAB exceedences. By contrast, at the 93 sites where H20 indicated degradation but CSCI did not (and ignoring $CRAM_{bio\text{-}phys}$), an opposite pattern emerged: 30 of those sites had only chemical exceedences, whereas 14 of them had only PHAB exceedences. When $CRAM_{bio\text{-}phys}$ indicated degradation (regardless of CSCI and H20 condition), sites were far more likely to have only PHAB exceedences than only chemical exceedences.

Table 5. Summary of the number of sites where thresholds for chemical and physical habitat stressor variables were exceeded per disagreement class. For disagreement classes, D = Degraded and N = Not Degraded

Disagreement class				Number of Exceedences				
CSCI	CRAM	H20	*n*	Chem	PHAB	Chem & PHAB	Only Chem	Only PHAB
D	D	N	24	14	22	13	1	9
D	N	D	90	71	65	53	18	12
D	N	N	45	12	29	8	4	21
N	D	D	27	19	20	13	6	7
N	D	N	32	8	22	7	1	15
N	N	D	66	46	29	22	24	7
Summary								
CSCI degraded, H20 not (ignoring $CRAM_{bio\text{-}phys}$)			69	26	51	21	**5**	**30**
H20 degraded, CSCI not (ignoring $CRAM_{bio\text{-}phys}$)			93	65	49	35	**30**	**14**
$CRAM_{bio\text{-}phys}$ degraded (ignoring CSCI and H20)			110	41	64	33	**8**	**31**

Overall, CSCI and $CRAM_{bio\text{-}phys}$ were more sensitive to PHAB exceedences, whereas H20 was more sensitive to chemical exceedences. As a follow-up to this result, Principal Components Analyses (PCA) of select chemical and PHAB variables[6] were conducted, and the three indices were each plotted against the first component from each PCA ordination (Figures 3 and 4). As suggested by patterns in the number of stressor exceedences per combined disagreement class, H20 showed the tightest response to the chemical principal component (Figure 3), while CSCI and $CRAM_{bio\text{-}phys}$ showed tighter responses to the PHAB principal component (Figure 4). It is perhaps unsurprising that $CRAM_{bio\text{-}phys}$ showed the tightest response to the PHAB principal component, given the partial redundancy between $CRAM_{bio\text{-}phys}$ and standard PHAB protocols (i.e., they measure similar things in different ways).

[6] Individual chemistry and PHAB analytes were sometimes missing for any given probability site. Since PCA cannot be performed if data are missing, a subset of variables was selected based on availability of all analytes for the greatest number of sites.

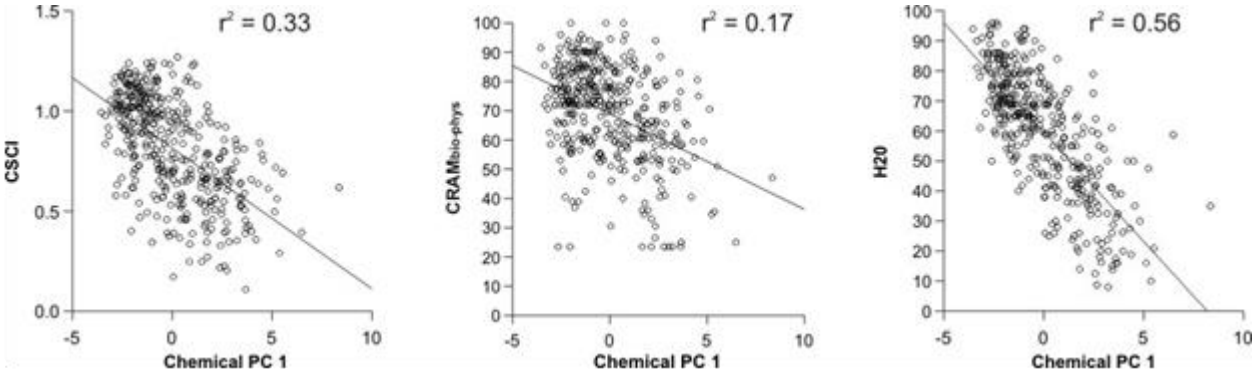

Figure 3. CSCI, $CRAM_{bio\text{-}phys}$ and H20 scores plotted against the first axis from PCA of 8 chemical variables: conductivity (µS/cm), acid neutralizing capacity (mg/L), turbidity (NTU), dissolved organic carbon (mg/L), total nitrogen (mg/L), total phosphorous (mg/L), sulphate (mg/L), chloride mg/L). Variables were log (x+1) transformed prior to analysis to improve normality. Variance explained by 1st PCA axis = 54.4%.

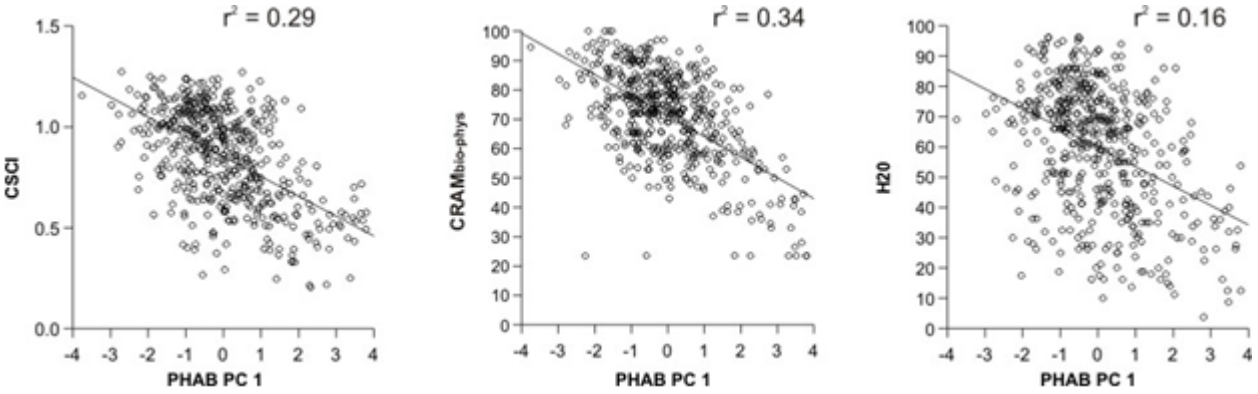

Figure 4. CSCI, $CRAM_{bio\text{-}phys}$ and H20 scores plotted against the first axis from PCA of 4 physical habitat variables: woody riparian cover index (XCMGW), in-stream habitat diversity index (XFC_NAT), riparian disturbance index (W1_HALL), and percent sand and fine substrate (PCT_SAFN). Variance explained by 1st PCA axis = 44.1%. Physical habitat variables are from Kaufmann et al. (1999).

Part 2 Conclusions

H20 showed a tighter response to chemical stressors than did CSCI and $CRAM_{bio\text{-}phys}$, not only in terms of the number of sites where degradation co-occurred with exceedence of chemical thresholds, but also in terms of its response to a multivariate axis of chemical variables. By contrast, degradation based on CSCI and $CRAM_{bio\text{-}phys}$ was more likely to co-occur with exceedence of PHAB thresholds, and those indicators showed a tighter response to a multivariate axis of PHAB variables than did H20. It is important to note that BMIs *did* respond to chemistry, just not as strongly as algae, and H20 *did* respond to physical habitat conditions, just not as strongly as BMIs. These results are consistent with those of several (but not all) studies published in the last ~10 years that showed different taxonomic assemblages often respond differently to the same stressors emanating from upstream land use practices (Sonneman et al. 2001; Mazor et al. 2006; Johnson et al. 2009). However, standard bioassessment protocols measure a relatively small suite of potential stressors at each site such that other important patterns were not taken into consideration due to lack of applicable data.

Several attempts were made to evaluate index-stressor relationships based on more quantitative, continuous variables. For example, index scores were converted to "difference from degradation threshold", so that scores much higher than the degradation threshold had large positive values, while scores much lower than the threshold had large negative values. The "distance from threshold" measures were then used as response variables in multivariate regression trees where the explanatory variables were raw stressor values. Results were not very interpretable and the trees did not explain much variance in the response variables. In addition, the same "difference from threshold" approach was applied to stressor variables using the exceedence thresholds in Appendix 3: the distribution of the "distance from threshold" variables was then plotted for each of the summary disagreement categories listed in Table 5. Again, the results were not easily interpretable, and when they were, showed more-or-less the same patterns described by categorical analyses. In the end, the clearest patterns were achieved by lumping disagreement categories (e.g., "CSCI degraded, H20 not") and by treating single exceedences the same as multiple exceedences for both chemical and PHAB variables, e.g., sites were in exceedence for chemistry whether one or many variables were exceeded, and regardless of how much the observed value(s) for different analytes exceeded the threshold value(s).

Final Conclusions and Closing Remarks

The use of multiple indices in conjunction to infer the ecological condition of streams, each based on a different taxonomic assemblage or data type, greatly strengthens confidence in results from bioassessment surveys. Patterns of agreement and disagreement among the 3 indices evaluated here were highly non-random: the indices frequently agreed that reference sites were not degraded, and that high-activity sites were degraded. This was not just a circular result derived from assessing the same pool of sites used to develop degradation thresholds; rather, thresholds were established based on a pool of reference sites that was largely independent of assessed probability sites. Thus, the 3 indices are independently well-calibrated to detect the extremes of human disturbance gradients, and their redundancy in such cases improves the precision of our assessments and reduces the likelihood of incorrect conclusions from sampling error or natural variability. Disagreement among indices was most frequent at sites with intermediate levels of disturbance and mixed upstream land use, which makes intuitive sense. Moreover, results presented here showed that the 3 indices respond differently to different types of disturbance, corroborating similar results from the published literature (see Introduction for examples), improving the overall sensitivity of our assessments to different stressors, and providing greater opportunities to diagnose causation. In sum, SWAMP's investment in multiple indices has yielded valuable returns, and their combined use should continue to be a central component of regional and statewide bioassessment programs.

While not explored in this study, analyses from the recent report from the SMC's stream bioassessment survey showed the value of multiple indicators in discerning physical habitat and water quality stress in concrete channels (Mazor 2015). The SMC found that CSCI and CRAM scores were invariably low in concrete channels, while diatom and soft algae indices showed a range of conditions, presumably because they reflected differences in water quality among concrete channels.

Averaging index scores into a single, weight-of evidence composite score is an approach sometimes advocated by bioassessment practitioners (e.g. Williams et al. 2009; Jessup and Pappani 2015). A composite approach was deliberately avoided here in preference for an approach of independent applicability, i.e., in recognition of the fact that indices based on different taxonomic assemblages (or data types in the case of CRAM) can provide unique responses to potential stressors affecting a site, are equally valid, and results from any one are independent of confirmation by the others (Yoder 1995). While

attractive from a simplicity of communication perspective, a composite index could potentially obscure unique responses of each assemblage to different stressors, thereby diminishing opportunities to diagnose causation (Carlisle et al. 2008). Also, reporting distinct scores for each index, rather than a single combined score, facilitates greater regulatory flexibility for how bioassessment results will be interpreted and used. An approach of independent applicability provides a higher level of protection to California's streams and rivers, and is echoed in the precautionary principle of the European Water Framework Directive, where overall degradation status is determined by the most sensitive indicator (Simboura et al. 2005).

Finally, this study points to the need for an algae index with statewide applicability, analogous to the CSCI. Despite its development for use in southern coastal California, the H20 index based on diatoms and soft algae in conjunction had generally good performance and applicability in statewide assessment, but a new indicator calibrated across the full range of natural environmental settings in California should only improve the utility of algae as an additional indicator, particularly given its potential importance in helping to discern physical habitat degradation from water quality stress.

REFERENCES

California Wetlands Monitoring Workgroup. 2013. California Rapid Assessment for Wetlands. Riverine Wetlands Field Book version 6.1. Available at *www.cramwetlands.org*.

Carlisle, D.M., C.P. Hawkins, M.R. Meador, M. Potapova and J. Falcone. 2008. Biological assessments of Appalachian streams based on predictive models for fish, macroinvertebrate, and diatom assemblages. Journal of the North American Benthological Society 27: 16-37.

Fetscher, A.E., R. Stancheva, J.P. Kociolek, R.G. Sheath, E.D. Stein, R.D. Mazor, P.R. Ode and L.B. Busse. 2014 Development and comparison of stream indices of biotic integrity using diatoms vs. non-diatom algae vs. a combination. Journal of Applied Phycology 26: 433-450.

Griffith, M.B., B.H. Hill, F.H. McCormick, P.R. Kaufmann, A.T. Herlihy and A.R. Selle. 2005. Comparative application of indices of biotic integrity based on periphyton, macroinvertebrates and fish to southern Rocky Mountain streams. Ecological Indicators 5(2): 117-136.

Jessup, B. and J. Pappani. 2015. Combination of biological and habitat indices for assessment of Idaho streams. Journal of the American Water Resources Association 51: 1408-1417.

Johnson, R.K. and D Hering. 2009. Response of taxonomic groups in streams to gradients in resource and habitat characteristics. Journal of Applied Ecology 46: 175-186.

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. U.S. Environmental Protection Agency. Washington, D.C.

Mazor, R.D., T.B. Reynoldson, D.M. Rosenberg, and V.H. Resh. 2006. Effects of biotic assemblage, classification, and assessment method on bioassessment performance. Canadian Journal of Fisheries and Aquatic Science 63: 394-411.

Mazor, R.D., 2015. Bioassessment of perennial streams in southern California: a report on the first five years of the Stormwater Monitoring Coalition's regional stream survey. Southern California Coastal Water Research Project Technical Report #844

Mazor, R.D., A.C. Rehn, P.R. Ode, M. Engeln, K.C. 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(1): 249-271.

Moyle, P.B., and P.J. Randall. 1998. Evaluating the biotic integrity of watersheds in the Sierra Nevada, California. Conservation Biology 12: 1318-1326.

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

Ode, P.R., T.M. Kincaid, T. Fleming and A.C. Rehn. 2011. Ecological Condition Assessments of California's Perennial Wadeable Streams: Highlights from the Surface Water Ambient MonitoringProgram's Perennial Streams Assessment (PSA) (2000-2007). Available at: http://www.waterboards.ca.gov/water_issues/programs/swamp/docs/reports/psa_smmry_rpt.pdf.

Ode, P.R., A.C. Rehn, R.D. Mazor, K.C. Schiff, E.D. Stein, J.T. May, L.R. Brown, D.B. Herbst, D.J. Gillett, K. Lunde and C.P. Hawkins. 2016. Evaluating the adequacy of a reference-site pool for ecological assessments in environmentally complex regions. Freshwater Science 35(1) 237-248.

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

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. 2015. The Perennial Streams Assessment (PSA): An Assessment of Biological Condition Using the New California Stream Condition Index (CSCI). SWAMP-MM-2015-0001. Available at: waterboards.ca.gov/water_issues/programs/swamp/bioassessment/docs/ psa_memo_121015.pdf

Resh, V. 2008. Which group is best? Attributes of different biological assemblages used in freshwater biomonitoring programs. Environmental Monitoring and Assessment 138: 131-138.

Simboura, N., P. Panayotidis and E. Papathanassiou. 2005. A synthesis of the biological quality elements for the implementation of the European Water Framework Directive in the Mediterranean ecoregion: the case of Saranikos Gulf. Ecological Indicators 5: 253-266.

Sonneman, J.A., C.J. Walsh, P.F. Breen and A.K. Sharpe. 2001. Effects of urbanization on streams of the Melbourne region, Victoria, Australia. II. Benthic diatom communities. Freshwater Biology 46: 553-565.

Stein, E.D., A.E. Fetscher, R. P.Clark, A. Wiskind, J.L. Grenier, M. Sutula, J.N. Collins and C. Grosso. 2009. Validation of a wetland rapid assessment method: use of EPA's level 1-2-3 framework for method testing and refinement. Wetlands 29: 648-665.

Stoddard, J.L., D.V. Peck, S.G. Paulsen, J. Van Sickle, C.P Hawkins, A.T. Herlihy, R.M. Hughes, P.R. Kaufmann, D.P. Larsen, G. Lomnicky, A.R. Olsen, S.A. Peterson, P.L. Ringold and T.R. Whittier. 2005. *An Ecological Assessment of Western Streams and Rivers*. EPA 620/R-05/005. U.S. Environmental Protection Agency, Office of Research and Development: Washington, D.C.

USEPA (U.S. Environmental Protection Agency). 2013. *Biological Assessment Program Review: Assessing Level of Technical Rigor to Support Water Quality Management*. Office of Science and Technology, Washington, D.C. EPA 820-R-13-001.

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

Williams, M., B. Longstaff, C. Buchanan, R. Llanso and W. Dennison. 2009. Development and evaluation of a spatially-explicit index of Cheasapeake Bay health. Marine Pollution Bulletin 59: 14-25.

Yoder, C.O. 1995. Policy Issues and Management Applications of Biological Criteria. pp 327-344. *In*: Biological Assessment and Criteria: Tools for Water Resource Planning and Decision Making, W.S. Davis and T.P Simon (eds). Lewis Publishers, Boca Raton, Florida.

SUGGESTED CITATION

Rehn, A.C. 2016. Using Multiple Biological and Habitat Condition Indices for Bioassessment of California Streams. SWAMP Technical Memorandum SWAMP-TM-SB-2016-0003

APPENDIX I. SUMMARY OF PERFORMANCE EVALUATIONS FROM MAZOR ET AL. (2016)

Aspect	Description	Indication of good performance
Accuracy and Bias	Scores are minimally influenced by natural gradients	-Approximately 90% of validation reference sites have scores above the 10th percentile of calibration reference sites. -Landscape-scale natural gradients explain little variability in scores at reference sites, as indicated by a low pseudo-R^2 for a 500-tree random forest model. -No significant difference in mean score of reference sites among major geographic regions.
Precision	Scores are similar when measured under similar settings	-Low standard deviation of scores among reference sites (one sample per site) -Low pooled standard deviation of scores among samples at reference sites with multiple sampling events.
Responsiveness	Scores change in response to human activity gradients	-Large t-statistic in comparison of mean scores at reference and high-activity sites. -Landscape-scale human activity gradients explain variability in scores, as indicated by a high pseudo-R^2 for a 500-tree random forest model.
Sensitivity	Scores indicate poor condition at high-activity sites	-High percentage of high-activity sites with scores below the 10th percentile of calibration reference sites.

APPENDIX 2A. STRESSOR AND HUMAN ACTIVITY GRADIENTS USED TO IDENTIFY REFERENCE SITES

See Ode et al. (2016) for additional information on development of reference criteria. Sites that did not exceed listed thresholds were used as reference sites. WS: Watershed. 5 km: Watershed clipped to a 5-km buffer upstream of the sample point. 1 km: Watershed clipped to a 1-km buffer upstream of the sample point. W1_HALL: proximity-weighted riparian disturbance index (Kaufmann et al. 1999). Data sources are as follows: A: National Landcover Data Set. B: Custom roads layer. C: National Hydrography Dataset Plus. D: National Inventory of Dams. E: Mineral Resource Data System. F: Predicted specific conductance (Olson and Hawkins 2012). G: Field-measured variable. Code 21 is a land use category that corresponds to managed vegetation, such as roadsides, lawns, cemeteries, and golf courses.

Variable	Scale	Threshold	Unit	Data source
% Agriculture	1 km, 5 km, WS	<3	%	A
% Urban	1 km, 5 km, WS	<3	%	A
% Ag + % Urban	1 km, 5 km, WS	<5	%	A
% Code 21	1 km and 5 km	<7	%	A
	WS	<10	%	A
Road density	1 km, 5 km, WS	<2	km/km^2	B
Road crossings	1 km	<5	crossings/ km^2	B, C
	5 km	<10	crossings/ km^2	B, C
	WS	<50	crossings/ km^2	B, C
Dam distance	WS	<10	km	D
% Canals and pipelines	WS	<10	%	C
Instream gravel mines	5 km	<0.1	mines/km	C, E
Producer mines	5 km	0	mines	E
Specific conductance	Site	99/1**	prediction interval	F
W1_HALL	Sample reach	<1.5	NA	G

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

APPENDIX 2B. CRITERIA USED TO DEFINE HIGH-ACTIVITY SITES

Criteria are from Mazor et al. (2016). Sites that were not defined as either reference or high-activity were classified as moderate-activity.

Variable	Scale	Threshold	Unit
% Developed land (i.e., % Ag + % Urban)	Any (1 km *or* 5 km *or* WS)	>50	%
Road density	Any (1 km *or* 5 km *or* WS)	>5	km/km^2
W1_HALL	Sample reach	<5	NA

APPENDIX 3. CRITERIA FOR IDENTIFYING STRESSOR EXCEEDENCES IN 4 AGGREGATE LEVEL III ECOREGIONS

See Stoddard et al. (2005) for aggregate ecoregion definitions. Criteria were developed using the biology-based approach suggested (but not actually used) by Ode et al. (2011). The 90th percentile of stressor values at sites in good biological condition (based on CSCI scores) defined the exceedence threshold for variables where higher values indicate more disturbance (i.e., chloride, conductivity, total nitrogen, % sand and fines, total phosphorous, total suspended solids, turbidity, riparian disturbance index, mean embeddedness). The 10th percentile of stressor values at sites in good biological condition defined the exceedence threshold for variables where lower values indicate more disturbance (i.e., woody riparian cover index, stream habitat diversity index). Aggregate ecoregions were used to define thresholds rather than PSA regions because the Central Valley has too few sites in good biological condition to establish robust thresholds, and because xeric and mountainous regions in the South Coast had very different distributions for the stressors evaluated. Physical habitat variables are from Kaufmann et al. (1999).

	Chloride mg/L (CL)	Conductivity µS/cm (COND)	Total Nitrogen mg/L (NTL)	Percent sand & fines (PCT_SAFN)	Total Phosphorous mg/L (PTL)	Total Suspended Solids mg/L (TSS)	Turbidity NTU (TURB)	Riparian disturbance index (W1_HALL)	Woody riparian cover index (XCMGW)	Mean percent embeddedness (XEMBED)	Stream habitat diversity index (XFC_NAT)
Sierra and North Coast	10.1	282	0.27	35	0.056	5.5	2.4	1.27	0.55	46	0.18
Southern California Mtns	25	930	0.586	54	0.19	10.1	3.2	0.73	0.37	59	0.27
Xeric California (= xeric SoCal, Central Valley and Chaparral)	122	1460	2.3	69	0.122	7.2	5.1	1.3	0.54	54	0.14
Xeric Southwest (= Desert-Modoc)	3.2	205	0.173	47	0.048	9.2	4.2	1.9	0.45	57	0.19