Stevenson's Responses to Charge Questions:

Is the report organization optimal for a document of this length? Would it be better if each of the report sections was a stand-alone piece?

The report is organized well into chapters of appropriate length and topics.

Is the cited literature sufficiently comprehensive? Are there any key references that have been omitted?

Recommendations for some literature have been added in comments on the manuscript. Often, only book chapters reviewing primary literature were cited, rather than the actual primary literature. More citations of primary literature documenting research and results should be cited, rather than book chapters.

Are any limitations of the datasets and monitoring approaches employed to support the analyses in this report adequately addressed?

Although the sample size is relatively large, the role of confounding factors and temporal variability in conditions is likely limiting ability to observe relationships among some variables. Relationships between algal biomass and nutrients, for example, are notoriously challenging, and often require repeated sampling of a site and 75 or more sites to resolve nonlinear relationships between algal biomass and nutrients in streams when natural variability is controlled. Given the great variability in climatic, geologic, hydrologic, topographic, and riparian conditions of streams in California, we would expect challenges in establishing biomass-nutrient relationships. Variability in substratum and current velocity among habitats in streams also affect algal biomass. The EPA/CA protocols calling for sampling random locations along 11 transects in a reach does not provide sufficient sample size to sample stratification for within-stream habitats, such as riffle, pools, and runs, to provide sufficiently precise characterizations of algal biomass in streams to develop good algal biomass-nutrient relationships without enormous sample sizes. Alternatively, the NAWQA approach of sampling richest target habitat or a habitat specific sampling protocol for characterizing algal biomass would improve assessments given the great problem with spatial and temporal variability in algal biomass in streams. Given relative costs of sampling versus field travel and the importance of algal biomass as an indicator of ecosystem services, perhaps more intensive sampling, stratified sampling, and separate composite samples for different habitats is warranted. Could data from the National Rivers and Streams Assessments be used to supplement CA's data?

Executive Summary

Does the executive summary adequately capture the major findings of the report? Are the summary statements adequately supported in the body of the report?

The findings of the report provide a good overview and support conclusions from chapters. Recommendations 1 and 3 are well supported by the results. Recommendation 2 provides a complementary approach for developing nutrient management targets, but not an alternative approach unless ecosystem services that are affected by algae, other than benthic macroinvertebrates, are not important.

Chapter 1

Does Chapter 1 provide sufficient background information to put the rest of the report into context with respect to information needs for the state of California?

The title of the report "IMPROVING TOOLS TO LINK NUTRIENTS TO ADVERSE EFFECTS ON STREAM ECOSYSTEM SERVICES IN CALIFORNIA" indicates tools will link nutrients to stream ecosystem services. Ecosystems services are not related well to the endpoints measured in the research. Ecosystem services are related to beneficial uses in the introduction, but only benthic algal biomass is those uses. And all that linkage takes place in the first paragraph of the introduction. Ecosystem services are not mentioned throughout the rest of the document. Maintaining high levels of biological condition is also beneficial use and ecosystems service, even though its relationship to valuation by most people may be debated. It is a goal of the Clean Water Act. Since many endpoints measured in the project are related to taxonomic metrics, they should be explicitly related to ecosystems services if it is to take a prominent position in the title of the report.

Cite primary literature that documents that relationships exist between measurable stream attributes (algal blooms in wadeable streams) and ecosystem services and human well-being exist rather than reviews that state these relationship exist (e.g. Biggs 2000, Lembi 2003). Suplee et al. 2009 is an important reference for linking benthic algal biomass to valuation of aesthetics, and thus establishing benthic algal biomass is an appropriate measure of aesthetics, which is an important ecosystem service. It should be used in the first paragraph of the introduction. Do you have evidence that wadeable stream algal blooms negatively impact human health and other ecosystem services or beneficial uses through toxin-forming harmful algal blooms, proliferation of pathogenic bacteria, and taste/odor problems in municipal drinking water supplies, or are problems in downstream lakes and coastal zones?

Aquatic life use, as measured with metrics using taxonomic information, an important endpoint in analyses, is not integrated into the introduction. The introduction is about benthic biomass and does not relate biodiversity elements of biological condition as measured with taxonomic metrics to ecosystem services, beneficial uses, and goals of the research and report.

Chapter 2

Are the methods used to estimate reference and ambient values for stream eutrophication indicators pooling data across multiple monitoring programs scientifically valid?

What was the variation in quality (e.g. % undisturbed watershed) in reference sites among ecoregions? Without that qualification, comparison across ecoregions is less informative because quality of reference sites in addition to natural factors remains in important variable. Is there a standardized scale for reference condition across ecoregions?

Chapter 3

Have the different methods for evaluating response thresholds of primary producer biomass and nutrient effects been described adequately so that someone previously unfamiliar with these methods can understand the approach and the method strengths and weaknesses and interpret the results?

The methods are described really well. I would recommend that results be described more thoroughly. Often there seems to be a disconnect between the results and highlights developed in the discussion. Results should be more thorough reported in the text and linked to referenced tables and figures. Currently substantial amounts of the results that are important were not reviewed well in the text.

Are the methods used to estimate response thresholds scientifically valid? Have statistical assumptions been adequately tested?

Yes, the methods used to estimate response thresholds scientifically valid, and assumptions for analyses have probably been met in most cases. However, I prefer transforming independent variables to get an even distribution to provide equal weight to values along the stressor gradient, rather than the common right skew with most data in the lowest levels of the stressor gradient. The piecewise linear regression results show the problem with the highly skewed results, and potential problems with loosing pattern recognition of assimilative capacity at really low nutrient concentrations and levels of productivity.

The use of sample weights in analysis of stressor-response relationship should be explained. I can see how weights help in stressor-response analyses to make them more representative of the state. But, aren't weights related to stratification variables which are natural classification variables? So when results with weighted and unweighted data agree, then would natural variables used in stratification probably have little effect on stressor-response relationships? When they don't agree, then stratification factors do affect stressor-response relationships and the current analysis that doesn't account for natural factors (e.g. piecewise regression) should be downweighted relative to results from BRT because BRT did account for natural factors in the partial dependence plots.

Although the statistical methods are scientifically valid, they are conceptually flawed for application in policy for two reasons. First, do the thresholds exist? Are the thresholds real or just blips along a stressor gradient? For example on pgs 38 and 40, there is little analysis of the strength of non-linear responses, and thus the importance of the thresholds for driving consensus among stakeholders. How much more variation in stressor-response relationships were explained by CART model fit that would have occurred by linear regression? Percent variance explained by CART is very similar to a coefficient of determination for a regression model (r^2), which could be added in a column to Table 3.5 for comparison with CART model fit.

Second reason: all thresholds are not equal for policy development. The rationale for using thresholds is they provide a point (threshold) along the stress or gradient above which adverse effects get greater, which focuses stakeholder consensus for a specific level of protection for a valued ecological attribute (Stevenson et al. 2008, Stevenson 2014). A figure like the BCG (Fig 3.1) with some assimilative capacity provides that tipping point that stakeholders do not want to pass.

Thresholds with assimilative capacity focus stakeholder consensus on a specific level of stressor that would protect ecosystem services, above which great losses would be expected. Linear responses do not provide that focus on a stressor level that provides stakeholder consensus.

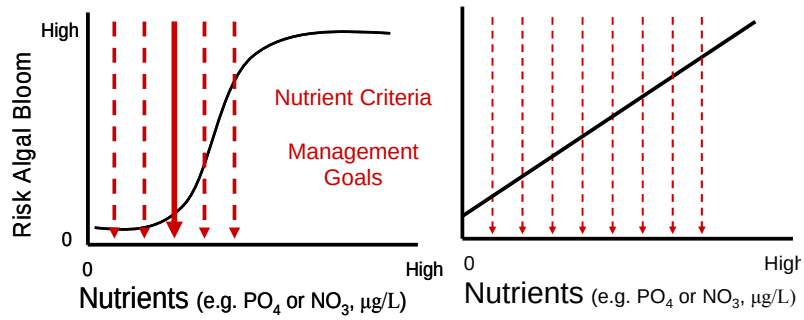

Figure showing differ challenge with stressor level selection when responses are linear.

Some responses of biological condition to productivity and nutrient gradients did illustrate assimilative capacity, such as in Figure 3.4 from the report.

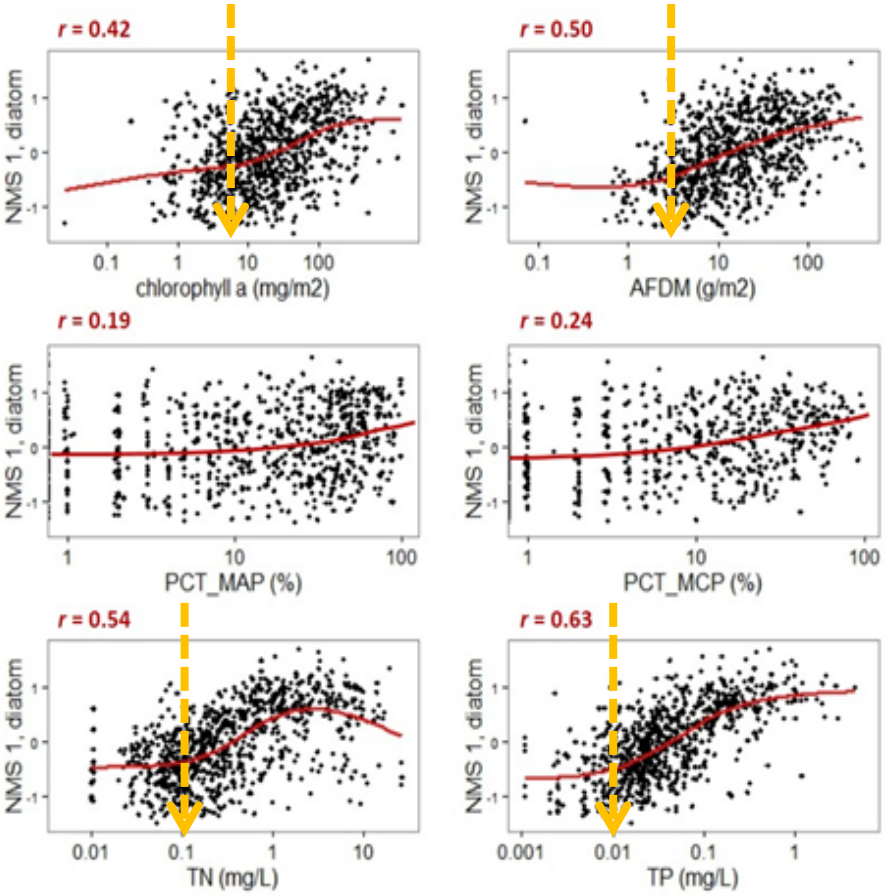

Figure 3.4. Scatterplots and splines for NMS axis 1 values from the diatom community against biomass and selected cover and nutrient gradients on log scale, using the statewide data set. The Pearson correlation coefficient for each relationship is provided to the upper left of each graph. Correlation analyses were performed on rank-transformed data, and sample weights were used in the analyses. All relationships were highly significant (p <0.0001).

Positively and negatively asymptotic responses have similar problems as linear responses with respect to policy because there is not a clear stressor level with a tipping point before the attribute is essentially lost or severely degraded. Figure 3.14 from the report provides an example of such a response illustrated with SiZer and piecewise regression. The delineated threshold is too high to protect the illustrated attribute. Yet these thresholds are used in the results to guide policy with no apparent modification. Some states have used the stressor level at the mid-point of the relatively linear decrease as a management target to protect at least half of the condition that would be lost.

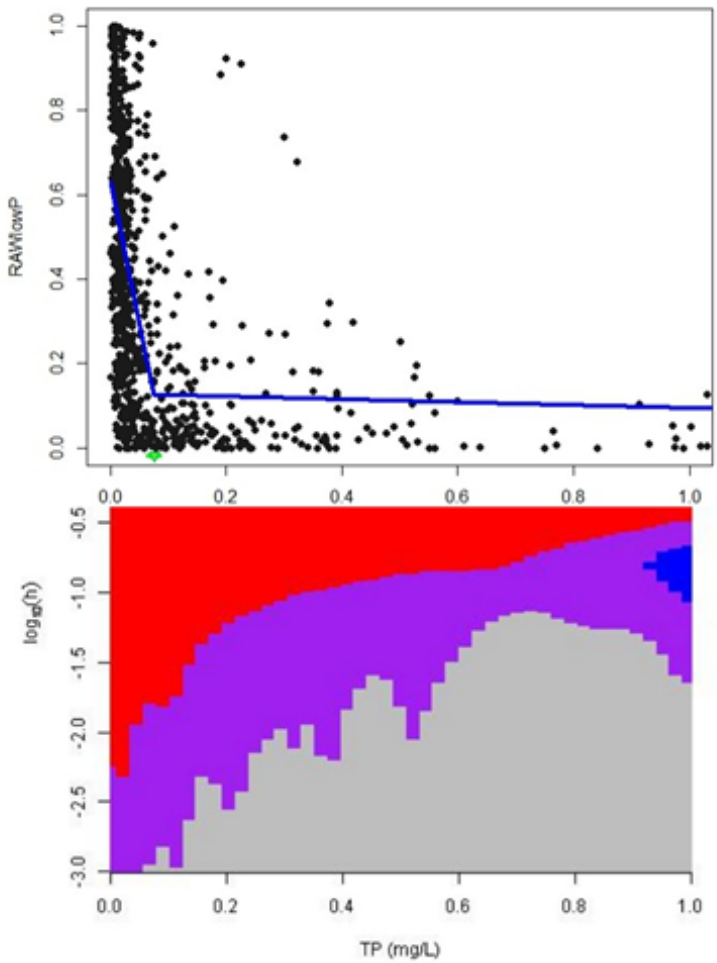

Figure 3.14. Piecewise regression plot of diatom ALU variable, RAWlowP on a TP gradient (top) and SiZer map from analysis of the same two variables (bottom). On the piecewise regression, the green line under data points along the x-axis is the 95% confidence interval for the breakpoint, and the green triangle is the breakpoint. On the SiZer map, the downward-extended ("downward peak") red portion of the graph indicates an area of well-supported, significantly negative slope that holds even at narrow bandwidths (see Methods for interpretation of color coding on a SiZer map, as well as definition of the y-axis).

Another problem with the threshold analysis is that statistic analyses are not well designed to isolate the lower bound of threshold response ranges. For example, the results highlight the upper bounds of the stressor range where the attributes respond most sensitively: "Furthermore, the scatterplots show that the most pronounced relationship between the diatom community (NMS axis 1), and TP occurs <0.1 mg/L, whereas for TN, the most pronounced relationship occurs <1 mg/L. These observations are corroborated by the results of the CART analyses of diatom and BMI NMS axis 1 scores (Table 3.5, Figure 3.5), in which median cut point values for TP and TN were consistently <0.1 and <1 mg/L, respectively. All median cut points for chlorophyll *a* were <31 mg/m^2, and for AFDM were <42 g/m^2."

In Figure 3.4 above, I've marked the upper bounds of the stressor range with assimilative capacity and the lower bounds of the stressor range with greatest response with the vertical dashed arrow. These are not the thresholds highlighted in the results or by the statistical analyses. CART often finds the upper change point, rather than the lower change point in curves with assimilative capacity. I would have highlighted the following. Below 0.01 mg TP/L and 0.1 mg TN/L for nutrients and below 2 mg chl/m2 and 2 g AFDM/m2 for algal biomass, little response was observed in valued attributes (response variables). So these thresholds are the good protection levels because there is assimilative capacity below 0.01 mg TP/L, 0.1 mg TN/L, 2 mg chl/m2 and 2 g AFDM/m2 and unwanted response above those levels. Thus, these are the kinds of responses that develop stakeholder consensus for management targets.

The CART and Piecewise Regression analyses are not well designed for threshold analyses of assimilative capacity.

Are the conclusions of this chapter adequately supported by the analyses and results?

The analyses of results do not distinguish whether thresholds mark the upper bounds or lower bounds of a response range and whether the response is below the threshold is okay or bad.

See above.

Does this chapter do a good job of synthesizing the results of multiple analyses contributing to a weight-of-evidence approach that could be used to support numeric nutrient endpoint development? Can you suggest any improvements?

Medians of thresholds in responses of ecosystem services should not be used to guide environmental policy. Policy should protect known elements of the ecosystem services of water resources. In the Everglades, some researchers suggested a similar approach and it was rejected because sensitive changes in the Everglades occurred at stressor levels below median response thresholds. Part of this rationale could be related to the idea that we are just using indicators of change in systems, not changes that we really care about. However, the focus of this report is ecosystem services and many of the responses to nutrients can be related to ecosystems services.

We should better understand the progressive degradation in water resources with increasing stressor levels and multi-stressor effects, and then set management targets to provide appropriate levels of protection given regional extents of landscape alteration and tiered uses that protect ecosystems services within a region and not necessarily at all sites.

Figure 3.25 does a good job of synthesizing results, but many of these thresholds do not account for the distinction of whether thresholds mark the lower or upper stressor bounds of the sensitive responses and whether they would protect ecosystem services (e.g. the mid-point of linear range of non-linear asymptotic responses). Update the figure with management targets based on statistically derived breakpoints in the data versus just the statistically derived breakpoints in the data. This additional correction is important for application in policy development.

Chapter 4

Is the evolution of the NNE benthic biomass spreadsheet tool adequately explained to allow the reader to understand its use and potential strengths and weaknesses?

I understood its use. I'm not sure what the "potential" strengths and weakness were, but the strengths and weaknesses before and after this report are evident and clear.

Are the methods used to evaluate the performance of the NNE benthic biomass spreadsheet tool scientifically valid? Have statistical assumptions been adequately tested?

The methods used to evaluate the performance of the NNE benthic biomass spreadsheet tool are scientifically valid. There was little discussion of statistical assumptions for tests, nor transformation of variables and subsampling data to meet assumptions, or qualifying interpretation of results based on lack of assumptions being met.

Were response variables transformed to reduce skewness in the data and overemphasis on high nutrient and biomass levels? Although I have heard it argued that CART and BRT like analyses are robust to assumptions of normality in variables, given the importance of variance analyses, I'm not sure how that could be the case. Some simple runs of analyses using transformed and untransformed variables would provide a valuable evaluation of this potential problem if variables were not transformed.

The Bayesian Cart models included factors that could be affected by human activities (e.g. NH_4 and conductivity), and even some measures of human activities (e.g. urban5K). Most classification of ecological data to develop relationships between ecosystem services (endpoints) and stressors only includes naturally varying factors that humans affect relatively little, like precipitation, temperature, geology, soils, ecoregion….. This confounding of the types of predictor variables is probably really only in assessment, whereas here the BBST tool is really to set site-specific nutrient management targets. So distinction between these variable types is probably not an issue.

Are the conclusions of this chapter adequately supported by the analyses and results?

Yes. The models show promise but they need more work and simplification if possible.

Does the analysis of residuals for model predictions presented in this chapter help to guide future improvements in these models?

The method for sampling habitats and expectation to come up with a model for the reach-scale average algal biomass may limit the potential success of the analysis, and linkage to ecosystem

services. Consider within stream habitat units as levels of assessment and try this analysis with the rapid periphyton surveys conducted with the 105 assessment points, if microhabitat data were collected with the assessment points.

References for potential use

Riseng CM, Wiley MJ, Stevenson RJ. 2004. Hydrologic disturbance and nutrient effects on benthic community structure in midwestern US streams: a covariance structure analysis. Journal of the North American Benthological Society 23:309-326.

Stevenson, R.J., B.J. Bennett, D.N. Jordan, R.D. French. 2012. Phosphorus regulates stream injury by filamentous green algae, thresholds, DO, and pH. Hydrobiologia 695:25-42.

Wootton JT, Parker MS, Power ME. 1996. Effects of disturbance on river food webs. Science 273:1558-1561.