Appendix A
Chemistry Indicators and Evaluation Methods
June 19, 2006

CHEMICAL INDICATORS

The chemical indicators evaluated in this study were based on chemical-specific sediment quality guidelines (SQGs) obtained from several sources. SQGs are numeric values intended to help in the interpretation of sediment chemistry data. SQGs are not intended to be a final assessment of environmental condition at a site, but rather to assist in the determination of the potential for biological effects. Numerical SQGs have been developed using both mechanistic and empirical relationships between chemistry and biological effect. Both types of approaches were evaluated in the early phases of the SQO project, but the mechanistic approaches (i.e., equilibrium partitioning) were not included in the final statistical evaluations based on the results of preliminary analyses and the recommendation of the SSC.

Three types of empirical chemical indicators were compared and evaluated: **established indicators** that were based on existing published SQGs that were developed for application on a national level, **regional indicators** that represent established indicator approaches calibrated to California data, and **new indicators** developed specifically for this project. All of the chemical indicators were based on chemical mixtures in order to represent the joint effects of multiple chemicals present in a sample. The individual chemical SQGs were integrated using a method specific to each approach to describe mixture effects. The chemicals included in each candidate indicator are shown in Table 1.

Established Indicators

Effects Range Median (NOAA ERM)

The Effects Range Median (ERM) approach (Long *et al.*, 1995) is one of the most commonly used SQGs. This method is used to identify adverse effects to sediment dwelling marine organisms. The ERM values were created from a national database of paired biological effects and sediment contaminant data. Multiple biological effects indicators were included in the database (this approach is not endpoint specific) and evaluated for the degree of concordance between chemical and different types biological responses. Only the data for which a biological effect was observed in association with elevated chemical concentrations were used for ERM derivation.

The ERMs were calculated by sorting the data in ascending order of concentration to calculate percentiles. The ERM corresponds to the 50^{th} percentile (median value) for each chemical and represents the concentration above which adverse effects are frequently observed. Individual ERMs were combined as a mean quotient to represent chemical mixture effects. The quotients were calculated by normalizing each chemical to its respective ERM and subsequently averaging them for each sample.

Mean Sediment Quality Guideline Quotient 1 (SQGQ1)
The mean sediment quality guideline quotient 1 (SQGQ1) is a subset of chemical-specific SQGs from various empirical and mechanical approaches (Fairey *et al*. 2001). The chemical suite includes five metals and four organics (Table 1). This suite of chemicals was selected because it was found to contain the chemicals with the strongest relationships to adverse biological effects in the data used for evaluation.

SQGQ1 quotients are calculated by normalizing each chemical value by its corresponding SQG. Then the normalized values for the suite of chemicals are averaged.

Consensus Midpoint Effect Concentration (Consensus)
The Consensus guidelines represent the integration of different types of SQGs. This approach collected and collated existing SQGs for chemicals of interest and evaluated them to determine their applicability (Swartz, 1999). Consensus SQG values have been developed for the threshold effect concentration (TEC), contaminant concentrations below which harmful effects on organisms are expected to occur infrequently; probable effect concentration (PEC), which represent contaminant concentrations above which harmful effects are frequently observed; and the midpoint effect concentration (MEC), an intermediate level of effect between the TEC and PEC. The MEC values were used to derive the Consensus chemical indicator evaluated in the SQO project.

Consensus MEC values were calculated by determining the geometric mean of three or more SQGs. Consensus values were previously derived for PAHs and PCBs in marine and freshwater systems, as well as for metals and several pesticides in freshwater systems (Swartz, 1999; McDonald et al., 2000). This project also used consensus MEC values calculated by SCCWRP for other chemicals: DDTs, dieldrin, arsenic, cadmium, chromium, copper, lead, mercury, nickel, silver, and zinc (Vidal and Bay, 2005).

The Consensus chemical indicator evaluated in this project was the mean quotient of the individual consensus MECs. Individual chemical values were normalized by dividing them by their corresponding consensus MEC value, then the normalized values where averaged for each sample.

Logistic Regression Modeling ($NatP_{max)}$
The Logistic Regression Modeling (LRM) approach is based on statistical analysis of matching chemistry and biological effects for a single endpoint (e.g., amphipod toxicity) (Field *et al.*, 1999). Chemistry and toxicity data from national databases were used for this approach. The LRM method does not yield specific SQG values for each chemical, but rather describes the relationship between contaminant concentrations and the probability of toxicity. This relationship can be used to calculate SQGs based on the level of protection desired.

In the LRM approach, data for individual sediment samples were sorted by ascending concentrations for each particular contaminant. The data were screened to reduce the influence of samples that did not contribute to the toxic effects associated with the

specific contaminant of interest. A logistic regression model was then applied to the screened data that described the relationship between the concentration of a selected contaminant and the probability of observed toxicity. The logistic model can be simplified and described by the following equation:

$$p = e^{B0+B1(x)} / (1 + e^{B0+B1(x)})$$

Where: p= probability of observing a toxic effect;
B0= intercept parameter;
B1= slope parameter; and,
x= concentration or log concentration of the chemical.

Individual chemical regression models were combined into a single mixture effects model based on the maximum probability of effects or P_{max} (Field et al., 2002). The maximum probability obtained from the individual chemical models is selected to represent the chemical mixture present in a sample.

Regional Indicators

Regional chemical indicators were developed based on two established SQG approaches: NatPmax and NOAA ERM. Three versions of each indicator were developed: a statewide version that was calibrated to data from throughout California, and two region specific versions. The region-specific versions were calibrated separately for northern and southern California data using Point Conception as the separation point.

CA ERM, SoCA ERM, NorCA ERM

SQGs analogous to ERMs were calculated using California data. The data were screened to identify toxic samples ($\geq$ 20% mortality) with chemical concentrations > 2x median concentration of non-toxic samples. After screening, the data were sorted in ascending order and the median concentration for each chemical was calculated (for chemicals with $\geq$ 10 samples). CA ERM values were calculated for 27 chemicals for the statewide and southern California indicators and for 25 chemicals for the northern California indicator (Table 1).

CA Pmax, SoCA Pmax, NorCA Pmax

Development of California LRM models and the P_{max} approach followed the methods described in Field et al. (2002). California-specific models were selected from a library of models that included national models as well as models derived using the California data sets. The selected models were developed and evaluated based on toxicity (using control-adjusted amphipod survival < 80% as the definition of toxic samples). The selected models were chosen based on the goodness of fit with the observed probability of toxicity. Models with high false positive rates were not used for analysis.

New Indicators

Mean Weighted Toxicity or Benthic Category Score (TCS, BCS)

The mean weighted category score approach is a novel approach based on the association between chemicals and the magnitude of biological response (i.e., category prediction

based on toxicity or benthic community disturbance). Three thresholds defining the biological response categories and a weighting factor reflecting the strength of association were calculated for each chemical. The thresholds and weights were determined for each chemical by optimizing the weighted kappa statistic for each chemical with respect to agreement with biological response. First, we calculated prediction thresholds for individual constituents to coincide with effects levels based on the BRI (Benthic Response Index) or amphipod mortality. Calibrations for biological response were developed separately for the BRI and amphipod mortality. Based on these chemical thresholds, an effects level prediction was determined for each chemical in the sample, given as a categorical response (e.g. 1 = reference, 2 = low, 3 = moderate, 4 =high). In addition, we calculated the overall weighted kappa value for that constituent for use as a weighting factor. Each constituent's predicted effect level was multiplied by its respective weighting factor to produce a "kscore". These "kscores" were then summed across all constituents in the sample and divided by the sum of all kappa values, giving a mean weighted score for either toxicity (TCS) or benthic disturbance (BCS).

Mean Weighted Score (mnwks):

$$mnkws = \frac{\sum \kappa \times cat}{\sum \kappa}$$

where cat = predicted toxicity or benthic impact category, and κ is the associated weighted kappa value for that constituent.

North and South versions of the TCS and BCS were calculated. A statewide TCS was also developed, which was the average of the north and south versions.

INDICATOR COMPARISON AND EVALUATION METHODS

The comparison and evaluation of the chemical indicators were based on statistical analyses conducted on independent validation data sets, composed of matched chemistry and toxicity (or chemistry and benthic disturbance) data from California embayments. The validation data were grouped into two regions: North, consisting of samples located north of Pt. Conception (N=147 for toxicity and 25 for benthos); and South, consisting of samples located south of Pt. Conception (N=249 for toxicity and 146 for benthos). Results for a statewide data set are reported for some analyses; these results were calculated by combining the north and south validation data sets. Indicator development and calibration was conducted on a separate set of data (development data set) that contained approximately double the number of samples present in each validation data set.

Comparison and evaluation of the chemical indicators were based on two characteristics: strength of association between chemistry and biological response and classification accuracy. The strength of association was measured as the nonparametric Spearman's correlation coefficient between the chemical indicator value and either sediment toxicity

(amphipod mortality) or benthic community response (benthic community category based on a combination of four indices).

Classification accuracy represents the ability of the chemical indicator to correctly predict the measured toxicity response (i.e., nontoxic, low, moderate, or high) or benthic response category for the sample. Two measures of classification accuracy were calculated: agreement and weighted kappa. Agreement is the percentage of samples where the chemical indicator correctly predicted the biological response category. The weighted kappa statistic is a measure of the magnitude of agreement between two sets of category predictions. The weighted kappa statistic differs from the percent agreement in that it corrects for the agreement expected purely by chance and gives partial credit that is related to the magnitude of disagreement (i.e., an incorrect prediction that is close to the measured category is given more partial credit than a prediction that differs greatly).

The thresholds used for the analysis of classification accuracy were selected using a statistical optimization procedure based on maximizing overall agreement between the chemical indicator and the biological impact category. Choosing the "best" set of thresholds for each indicator assured that the thresholds were comparable among indicators.

Threshold selection and calculation of correlation, agreement, and weighted kappa used a resampling (bootstrap) approach based on even numbers of samples within each biological impact category, where possible. Randomly selecting equal numbers of test samples from each biological impact category reduces testing bias and provides a more reliable measure of classification accuracy, particularly for the weighted kappa statistic which is sensitive to unequal sample proportions among categories. The values reported in this document are the median of 50 resamples. The results for nonbootstrapped analyses are also reported to assess whether the ranking of candidates changed when applied to a more representative distribution of the data. These results are based on a single analysis of the entire validation data set.

Both correlation and classification accuracy were used to evaluate the candidate chemical indicators and select the recommended approaches. The 90^{th} percentile confidence limits of the bootstrapped results were used to identify the best performing indicators with respect to correlation and classification accuracy. The approach having the best overall performance for both correlation and classification accuracy was selected as the recommended indicator. The correlation results were given greater weight when the rankings were variable among the performance measures.

REFERENCES

Fairey, R., E. R. Long, C. A. Roberts, B. S. Anderson, B. M. Phillips, J. W. Hunt, H. R. Puckett and C. J. Wilson. 2001. An evaluation of methods for calculating mean sediment quality guideline quotients as indicators of contamination and acute toxicity to amphipods by chemical mixtures. Environ. Tox. Chem. 20: 2276-2286.

Field L. J., MacDonald D. D., Norton S. B., Severn C. G., and C. G. Ingersoll. 1999. Evaluating sediment chemistry and toxicity data using logistic regression modeling. Environ. Toxicol. Chem. 18: 1311-1322.

Field L. J., MacDonald D. D., Norton S. B., Severn C. G., Ingersoll C. G., Severn C. G., Smorong D., and R. Lindskoog. 2002. Predicting amphipod toxicity form chemistry using logistic regression models. Environ. Toxicol. Chem. 21: 1993-2005.

Long E. R, Mac Donald D. D., Smith S. L. and F. D. Calder. 1995. Incidence of adverse biological effects within ranges of chemical concentrations in marine and estuarine sediments. Environ. Manag. 19: 81-97.

MacDonald D. D., Di Pinto L. M., Field J., Ingersoll C. G., Long E. R., and R. C. Swartz. 2000. Development and evaluation of consensus-based sediment effect concentrations for polychlorinated biphenyls (PCB). Environ. Toxicol. Chem. 19: 1403-1413.

Swartz R. C. 1999. Consensus sediment quality guidelines for PAH mixtures. Environ. Toxicol. Chem. 18: 780-787.

Vidal, D. E., and S. M. Bay. 2005. Comparative sediment quality guideline performance for predicting sediment toxicity in southern California, USA. Environ. Tox. Chem. 24: 3173-3182.

Table 1. List of analytes included in each chemical indicator evaluated.

	Indicator											
Chemicals	**NOAA ERM**	**CA ERM**	**NorCA ERM**	**SoCA ERM**	**SQGQ1**	**Consensus**	**TCS**	**BCS**	**National Pmax**	**CA Pmax**	**NorCA Pmax**	**SoCA Pmax**
Arsenic	•	•		•		•						
Cadmium	•	•	•	•	•	•	•	•	•	•	•	•
Chromium	•	•	•	•		•						
Copper	•	•	•	•	•	•	•	•	•	•	•	•
Lead	•	•	•	•	•	•	•	•	•	•		•
Mercury	•	•	•	•		•	•	•	•	•	•	
Nickel	•	•		•		•						•
Silver	•	•	•	•	•	•						
Zinc	•	•	•	•	•	•	•	•	•	•	•	•
Chlordanes		•	•	•	•							
Alpha Chlordane							•	•		•		•
Gamma Chlordane							•	•				•
DDTs	•	•	•	•		•					•	
o,p'-DDE							•	•				
o,p'-DDD							•	•				•
o,p'-DDT							•	•			•	
p,p'-DDD							•	•			•	•
p,p'-DDE							•	•			•	
p,p'-DDT							•	•	•	•		•
Dieldrin		•	•	•	•	•	•		•	•		•
Nonachlor										•		•
PAHs					•	•						
Low molecular weight PAHs							•	•		•	•	•
High molecular weight PAHs							•	•		•	•	•

	Indicator											
Chemicals	**NOAA ERM**	**CA ERM**	**NorCA ERM**	**SoCA ERM**	**SQGQ1**	**Consensus**	**TCS**	**BCS**	**National Pmax**	**CA Pmax**	**NorCA Pmax**	**SoCA Pmax**
1-Methylnaphthalene									•			
1-Methylphenanthrene									•			
2,6-Dimethylnaphthalene									•			
2-Methylnaphthalene	•	•	•	•					•			
Acenaphthalene	•	•	•	•					•			
Acenaphthylene	•	•	•	•					•			
Anthracene	•	•	•	•					•			
Benz(a)anthracene	•	•	•	•					•			
Benzo(a)pyrene	•	•	•	•					•			
Benzo(b)fluoranthene									•		•	
Benzo(e)pyrene											•	
Benzo(g,h,I)perylene									•			
Benzo(k)fluoranthene									•			
Biphenyl									•			
Chrysene	•	•	•	•					•		•	
Dibenz(a,h)anthracene	•	•	•	•					•			
Fluoranthene	•	•	•	•					•			
Fluorene	•	•	•	•					•			
Indeno(1,2,3-c,d)pyrene									•			
Naphthalene	•	•	•	•								
Perylene									•			
Phenanthrene	•	•	•	•					•			
Pyrene	•	•	•	•					•			
PCBs	•	•	•	•	•	•	•	•	•	•	•	•
Tributyltin		•	•	•								