Power Analysis and Optimization of the RMP Status and Trends Program

FINAL REPORT

Aroon R. Melwani[1], Ben K. Greenfield[1], Andy Jahn[2],
John J. Oram[1], Meg Sedlak[1], and Jay Davis[1]

1. San Francisco Estuary Institute
   and the Regional Monitoring Program for Trace Substances
2. Statistical Consultant

SAN FRANCISCO ESTUARY INSTITUTE
7770 Pardee Lane, Second floor, Oakland, CA 94621
p: 510-746-7334 (SFEI), f: 510-746-7300, www.sfei.org

Executive Summary

The Regional Monitoring Program for Water Quality in the San Francisco Estuary (RMP or Program) evaluates the concentration of pollutants in a wide range of matrices. Trace metals and organic compounds are monitored in water, sediments, bivalves, and sport fish as part of the Status and Trends element of the RMP (SFEI 2005, 2006b). Periodically, the cost-effectiveness and statistical power of the major elements of the RMP are evaluated using power analysis in combination with an evaluation of the information needs and priorities of water quality managers.

This evaluation of the Program was motivated by new understanding of bay processes (e.g., presence of phytoplankton blooms), changes in the regulatory focus from the water column to biota (e.g., bird eggs, sport fish and small fish), and significant management actions that may impact the Bay (e.g., large-scale wetland restorations). This report describes methods and selected findings for power analyses conducted in 2006, and documents the rationale for Technical Review Committee (TRC) and Steering Committee (SC) decisions regarding the design of RMP Status and Trends monitoring.

The objective of the power analysis was to evaluate whether current sampling size and frequency are appropriate for meeting the needs of RMP stakeholders, including the San Francisco Bay Regional Water Quality Control Board. Specifically, the statistical power analysis was conducted for two scenarios. The first scenario compared RMP data to thresholds. This was analogous to the previous RMP power analysis (Lowe *et al.* 2004), with the difference that updated thresholds and data were used. The second scenario evaluated the ability of the RMP random sampling design to detect long-term trends. The analysis focused on pollutants that are of currently high management priority

for the Bay, which include polychlorinated biphenyls (PCBs), mercury (Hg), copper (Cu), and nickel (Ni).

On the basis of power analysis results and other considerations, the TRC made recommendations for future monitoring of contaminants in the Bay. The power for Status and Trends matrices (water, sediments, bivalves, and sport fish) with the current sampling design and assumed rates of decline will be very high (> 95% in most cases). Furthermore, sport fish and bivalves appear to be the best indicators for trends. The TRC recommended that the sampling design for matrices other than sport fish be modified, due to adequate characterization of contaminant variability in these matrices for the majority of the Bay. Sport fish monitoring would achieve adequate power for trend detection, but due to high concentrations of PCBs (shiner surfperch and white croaker) and mercury (white croaker), the ability to distinguish concentrations below thresholds could not be achieved with the current design. Nevertheless, the TRC deemed the value of the current sport fish monitoring design to be very high, and suggested that modifications not be made. Finally, bird eggs such as of cormorants have previously only been monitored under special studies. A proposal for monitoring of bird eggs (cormorants and terns) every three years was supported by the TRC and SC.

Introduction

The Regional Monitoring Program for Water Quality in the San Francisco Estuary (RMP) evaluates the concentration of pollutants in a wide range of matrices. Currently, trace metals and organic compounds are monitored in water, sediments, bivalves, and sport fish as part of the Status and Trends element of the RMP (SFEI 2005, 2006b). Additionally, contaminant concentrations in cormorant eggs (Davis et al. 2007b), mercury in tern eggs, and mercury concentrations in small forage fish have been monitored as part of special studies. Periodically, the cost-effectiveness and statistical power of RMP elements are evaluated using power analysis in combination with an evaluation of the information needs and priorities of water quality managers. The previous redesign of the Program and power analysis, completed in 2002, resulted in substantial changes to the monitoring design (Lowe *et al.* 2004).

In 2006, the RMP Technical Review Committee (TRC) and Steering Committee (SC) undertook a review of the Program. This review was motivated by a number of reasons including: changes in our understanding of Bay processes (e.g., increases in phytoplankton blooms); changes in regulatory focus from the water column to the source of impairment (e.g., bird eggs, sport fish and small fish); and changes in the management of the Bay or adjacent wetlands (e.g., restoring salt ponds to wetlands). Where possible, a power analysis was used to evaluate the cost-effectiveness and statistical power of the major elements of the Status and Trends Program. For certain elements of the Program, it was not possible to conduct a statistical analysis. This report describes methods and

selected findings from the power analyses and summarizes the rationale for changes made to the Program in cases where it was not feasible to conduct statistical analyses.

The results of power analysis are dependent on the questions asked and the specific assumptions incorporated into the analysis. The objective of this analysis was to evaluate whether current sampling size and frequency are appropriate for meeting the needs of RMP stakeholders, including the San Francisco Bay Regional Water Quality Control Board (Regional Board). In particular, there is interest in three questions: 1. What power does the sample size of RMP stations provide to distinguish concentrations from relevant regulatory thresholds? 2. What is the power of the current sampling design to determine long-term trends? 3. Are there regions in San Francisco Bay where sampling intensity can be reduced in order to reallocate funds to higher priority items?

The focus of this power analysis was on pollutants that are of currently high management priority for the Bay. These include polychlorinated biphenyls (PCBs), mercury (Hg), copper (Cu), and nickel (Ni) (Table 1). Although there are many new compounds detected in Bay waters and sediments, datasets on these compounds are very limited to date (Oros *et al.* 2003).

Methods

Based on the stated study objectives, we evaluated power using the RMP data set for two scenarios (Table 1). The design of each power analysis was tailored to the study questions being addressed (Table 2), which included explicit estimates of variability, effect size, and null and alternative hypotheses. Further details on the analyses can be

found in the methods description below. The first scenario compared the RMP data to thresholds. This is analogous to the previous power analysis (Lowe *et al.* 2004), with the difference that updated thresholds and data were used. The analysis only focused on thresholds of management significance for the Bay (Table 3). These were the US Environmental Protection Agency's California Toxics Rule water quality criteria, Total Maximum Daily Load (TMDL) thresholds, and site-specific objectives for the Bay. Site-specific aquatic life water quality objectives for copper and nickel were adopted by the State of California in 2003. Notably, of the pollutants recently measured in water, only PCBs have shown a high incidence of exceeding thresholds (n = 54 of 60, 90%; Table 4). The second scenario evaluated the power of the RMP random sampling design to detect long-term trends. The specific question that was addressed can be summarized as, "*given an expected rate of decline over a specified time frame, what is the power of the sampling design to detect a significant negative trend?"*

The first scenario compared water and sport fish data to thresholds using a power analysis for a one-tailed non-central t-distribution in Systat 11 software (Systat Software Inc., San Jose, CA). The rationale for a one-tailed test is that the TRC and RWQCB expressed interest in detecting contaminant concentrations that are significantly lower than the applicable threshold. The hypothetical distributions of the test statistic (t) illustrated in Figure 1 demonstrate the difference between a one-tailed and a two-tailed test. In Plot A, a two-tailed distribution is illustrated. The region under the curve representing a type-1 error rate (α) is split between the two tails, and significant test results can be obtained for either being higher or lower than the threshold. In Plot B, a one-tailed distribution is illustrated. The overall type-1 error rate is the same as for a two-

tailed distribution, except it is tested in one direction only. Therefore, the power analysis addressed whether the mean concentration of a contaminant in water or sport fish was below its threshold of concern (H_0: $\mu \geq$ Threshold vs. H_A: $\mu <$ Threshold).

Some contaminants are currently well above threshold values. To simulate power to detect future pollutant levels that are below thresholds in these scenarios, concentrations were adjusted to 20% below its threshold, and the one-tailed comparison was made with this simulated data. This was necessary for mercury in white croaker, PCBs in white croaker and shiner surfperch, and mercury and PCBs in water. For water comparisons that did not require adjustments, mean contaminant concentrations were calculated based on the log-average for RMP sampling years since the redesign (2002 and 2003). For sport fish, the log-average for all available years (1997, 2000, and 2003) was used. In both water and sport fish, variability was represented by the standard deviation of these yearly averages ($n = 2$ and $n = 3$, respectively). Using the estimates of mean concentration and variability, we calculated the number of samples required to be significantly below a particular threshold with 80% and 95% power. Table 3 lists the contaminants and thresholds that were applied.

To address the second objective of evaluating the ability to detect long-term trends, a Monte Carlo simulation program was developed in the mathematical program Matlab (The MathWorks, Natick, MA). The program simulates contaminant concentrations for expected trends, with variability estimated based on current RMP data. In essence, the program creates a large number of simulated data points, based on an assumed model of contaminant concentrations and variability. This simulated data set

was then statistically evaluated using a range of sampling designs to determine which designs would have a high probability of detecting significant trends.

The model for data simulation was:

$$y_i = Y_o - R(t) + \varepsilon_1 + \varepsilon_2 \qquad \textit{(Equation 1)}$$

Where, y_i = an individual simulated contaminant concentration sample, Y_o = the initial average concentration, R = annual rate of decline, t = time (in years), and ε_1 and ε_2 are normally distributed error terms that represent the intra- and inter-annual variation, respectively. This is essentially a linear model with error terms (Figure 2). ε_1 and ε_2 were estimated using current RMP data as described below. We realize that the majority of the intra-annual variation likely represents spatial variability, while temporal variability is encompassed by the estimates of inter-annual variation. Power analyses were performed either Bay-wide, or based on individual segments, depending on the matrix tested. RMP data were log-transformed prior to evaluation, and the model was run using log-scale parameters. As a result of log transformation, the model depicts contaminant declines as an exponential decay function, a common assumption for contaminant fate data (*e.g.,* Stow *et al.* 1999).

After the redesign of the sampling program (in 2002), RMP station selection for water and sediments has followed the Generalized Random Tessellation Stratified (GRTS) design used by EPA's Environmental Monitoring and Assessment Program (Lowe *et al.* 2004). This redesign included switching from targeted sampling of 22

stations for both water and sediments to 31 stations for water and 47 stations for sediment. A few stations still remain targeted (n = 5 [water] and n= 7 [sediments]), with the remainder being sampled randomly (GRTS design). Estimates of the contaminant mean, variance, standard deviation, and standard error using the randomly collected data were calculated in version 2.9 of the psurvey.analysis statistical library, using the R system. R is free software downloadable through the Comprehensive R Archive Network (CRAN) web site at http://cran.r-project.org. The psurvey.analysis library for the analysis of probability surveys may be obtained from the Monitoring Design and Analysis section of the U.S. Environmental Protection Agency Aquatic Resources Monitoring web site (http://www.epa.gov/nheerl/arm/analysispages/software.htm).

Monte Carlo simulations generated 1,000 simulated data sets (*e.g.,* Figure 3), based on the statistical parameters in Equation 1. All parameters were estimated using RMP Status and Trends data for a given contaminant and matrix. For sediment and water, the intra-annual (within-year) variation ε_1 was estimated using psurvey.analysis based on the standard deviation of log-normalized data for each RMP sampling year since the redesign (generally 2002 and 2003). The inter-annual (between-year) variation ε_2 was estimated by simply calculating the log average for each RMP sampling year (1994 – 2003), and then determining the standard deviation of these averages. Scatter plots and regression analyses were performed to confirm lack of trend over this period.

An RMP Exposure and Effects Pilot Study for Double-crested Cormorant eggs was performed in 2002 and 2004. The pilot study monitored egg composites from three locations in the Bay (i.e., Wheeler Island, Richmond Bridge and Don Edwards in the South Bay). Of these, Richmond Bridge had previously been sampled by SFEI as a

special study in 1999 – 2001. Richmond Bridge was therefore selected for power analysis evaluation as this station represented the longest running dataset on cormorant eggs (Davis et al. 2007b). Variance estimates (ε_1 and ε_2) were based on two samples per year for the five years of data, which were derived using the same method as for sediment and water, except that data pertained to a single station.

The variability estimates for sport fish and bivalves were derived using a modification of the methodology for sediment and water. Since RMP tissue data are collected at discrete stations distributed around San Francisco Bay, the variability in contaminant concentration due to the effects of time and station were accounted for. Transplanted bivalves are deployed once a year (dry season) at nine fixed mooring stations in the Bay for a period of 90 – 100 days, and subsequently analyzed for tissue contaminant concentrations. Description of the data treatment for bivalves can be found in Appendices I and II. Sport fish are collected every three years in a non-random fashion, at five targeted stations distributed around the Bay. For sport fish, the effects of lipid, length, and station were tested first using a one-way analysis-of-variance (ANOVA) in Systat 11 (Systat Software Inc., San Jose, CA). If any significant effects were found, the residuals of the ANOVA were saved, and the average of these residuals for each available RMP sampling year (1994, 1997, 2000, and 2003) was calculated. In general, lipids never constituted a significant effect, while length and station commonly did. The residuals of the ANOVA represent the variation in contaminant concentration once the influence of these effects has been removed. The between-year variation ε_2 was then determined by calculating the standard deviation of these yearly averages. If no significant effects of length or station were found, ε_2 was simply represented by the

standard deviation of the annual average concentrations of the log-transformed raw data. The within-year variation ε_1 was estimated by performing a one-way ANOVA with a year effect, on either the raw data or the residuals of the first ANOVA, depending on whether a station or length effect was found initially. The within-year variability was then represented by the standard deviation of the residuals of this ANOVA.

The power to detect trends over time was evaluated for various pollutants. In water, DDTs (sum of o,p' and p,p' isomers of DDT, DDD, and DDE), PCBs (sum of 40 congeners) and total mercury were used. Rates of change that are consistent with the current understanding of long-term trends in these pollutants were examined. Historical data on PCBs and DDTs suggest that these pollutants are declining at a rate of approximately 5% per year (Davis et al. 2006, Connor et al. 2007, Davis et al. 2007a). To be conservative, these pollutants were examined at a decline rate of 3.5% per year for 20 years. Mercury is generally considered to be declining at a much slower rate (Greenfield et al. 2005, Conaway et al. 2007). Therefore, this pollutant was examined at a decline rate of 1% per year for 30 years. The same PCB and mercury decline rates for water were used to evaluate sediments and sport fish. Sport fish analyses were only conducted on shiner surfperch and white croaker, as these species represented the largest sport fish datasets, although other sport fish are routinely monitored. For bivalves, PCBs (3.5 % per year), DDTs (3.5% per year) and PBDE 047 were examined. Notably, PBDEs have only been sampled at seven sites, visited annually since 2002. Evaluation of this dataset showed that the average concentration of PBDE 047 has declined at an average rate of 9% per year. However, with such a limited dataset, there was little basis for projecting the average decline into the future. To be conservative, we examined PBDE 047 with a decay

rate of 3.5% per year (same as PCBs and DDTs). For cormorant eggs, the same pollutants evaluated for water were used (PCBs, DDTs, and mercury), but with modified decline rates. Scenarios for PCBs were employed at a decline rate of 6% per year for 20 years, based on the expected decline rate for transplanted bivalves in the Bay (Davis *et al.* 2006). DDT is expected to decline more rapidly, and therefore an annual decline rate of 4% and 8% per year for 20 years were evaluated. Mercury was evaluated with slower annual decline rates of 1% and 3% per year for 30 years. Power analyses for PBDEs in cormorant eggs were also performed, but this dataset only represented two years. Therefore, the PBDE results are not presented here, but have been summarized by Davis *et al.* (2007).

Linear regression analysis was performed on each simulated data set to determine slope and statistical significance (p-value). The proportion of results that exhibited statistically significant declining slopes ($p < 0.05$) was then calculated to determine statistical power. Therefore, the program evaluates the statistical power to correctly discern a declining trend given the underlying model. Power was evaluated across a range of sampling designs, including varying sampling frequency (*e.g.,* every 1 to 5 years) and sample size per year (*e.g.,* Figure 4).

Results and Discussion

Water

The current design for monitoring water in the Bay is nine random samples in the South Bay, five random samples in Lower South Bay, and four random samples in the remaining three segments, collected annually. A total of 31 (26 random and 5 fixed) sites are currently monitored (SFEI 2006a). This design would be sufficient to detect trends of PCBs and DDT in water over the next 20 years. In all Bay segments, the power to detect a 3.5% annual decline would be > 95%, using the current sampling design (red cells in Table 5). Analysis of mercury indicated similarly high power estimates over 30 years in segments other than San Pablo Bay. Four samples collected annually in San Pablo Bay would be sufficient to detect a 1% annual decline in mercury with 74% power. This lower power can largely be attributed to higher within-year variation (s.d. = 0.41) compared to other segments (s.d. = 0.12 – 0.32). Furthermore, Table 6 suggests that the coefficient of variation (CV) of some contaminants differ greatly between segments (*e.g.,* PCBs and DDTs in water). Differential mixing processes, loadings, and hot spots may explain why the two largest segments (Central Bay and San Pablo Bay) often had the highest CVs. Hot spots in San Francisco Bay (*e.g.,* Hunter's Point, Oakland Harbor, Richmond Harbor) have often coincided with the margins of the largest Bay segments, and thus could contribute to the higher variability in contaminant concentrations. Future modeling work in the Bay will examine the contributions of these characteristics of the Bay to contaminant variability estimates.

Water contaminant concentrations have generally been used for comparison to water quality objectives, and not for detecting trends. Recently, the target matrix for PCB and mercury TMDL development has begun shifting from a focus on water to biota (*e.g.,* SFRWQCB 2006). Given this, the power analysis results suggest that it may be possible to reduce either the frequency of monitoring or the number of stations (*e.g.,* Appendix III.1). Nevertheless, one of the concerns with switching to less frequent sampling is the possibility of missing short-term fluctuations. Mechanisms underlying natural variability in water constituents, such as rate of plankton growth and rainfall-induced loading during individual sampling events may cause short-term fluctuations in contaminant concentrations. If surprisingly different concentrations are found in the future based on an infrequent monitoring strategy, there is concern that trends and sources of variability would take much longer to recognize. In light of the generally high power estimates across contaminants and segments, the TRC recommended continuing annual sampling but collecting fewer samples each year (Table 7).

RMP water data were also evaluated for the power to determine that concentrations are below water quality objectives (Table 3). Results indicated that we currently have adequate power to distinguish copper, nickel, and lead in water from thresholds with 95% power (Table 8). This can largely be attributed to current concentrations that are already well below guidelines. Future PCB and mercury concentrations were simulated by adjustment to 20% below their respective thresholds as described in the Methods. As a result, the analysis suggested that only in the South Bay would the current design detect PCBs below the CTR threshold with 80% power. For mercury, many more samples than are currently being collected would be required to

detect concentrations below the TMDL threshold with 80% power. Mercury concentrations have been well above guidelines in numerous matrices since measurements began (SFEI 2006a). The current design would not distinguish PCBs or mercury below their respective thresholds in any of the segments with 95% power. The current water sampling design was largely driven by copper in the South Bay, in order to have sufficient data to discern whether the segment was below previous thresholds. With the revised copper guideline, it is not necessary to have so many stations located in this segment for this purpose. In addition, for pollutants such as PCBs, much of the water column exceeds the CTR threshold and will likely exceed it for an extended period of time. The threshold analysis suggests a reduction in sampling stations would still allow us to confirm that concentrations of copper, nickel, and lead are below management thresholds with relatively high power. For PCBs and mercury, concentrations are not expected to be below thresholds for quite some time.

Sediment

The current design for monitoring sediment in the Bay is eight samples in each segment, collected annually. The existing number of sites (40 random and 7 fixed sites) was chosen to provide good coverage across the Bay. The power analysis results suggested that this design obtains the necessary power to detect long-term trends in mercury and PCBs. The power of the current design to detect a 3.5% annual decline in PCBs over the next 20 years would be > 99% in all Bay segments (red cells in Table 9).

For mercury, power would be similarly high. The power analysis indicated that a reduction in sample size or monitoring frequency would be appropriate.

Reallocation of samples to the three largest sample areas (South, Central, and San Pablo Bays) was considered given the low number of samples per unit area. The TRC noted that by reducing the number of samples below six, we would begin to lose the power to adequately characterize trends in Suisun Bay. Suisun Bay is one of the two most dynamic segments (Lower South Bay is the other), being heavily influenced by freshwater inflows and high storm flow regimes (SFEI 2006a). The high within-year variability and CV values (Table 6) would therefore be expected given these regimes. Therefore, modifying the number of stations to be proportional to the size of the segment is not warranted. Currently, all samples are collected in the dry season (July – August). To better characterize each segment with the current design, the TRC recommended a small portion of samples be reallocated to the winter months every other year as this is when highest sediment toxicity is frequently observed. Given the high power estimates across contaminants and segments with the current design, the TRC recommended that future sediment monitoring could alternate between summer and winter sampling every other year (Table 7).

Bivalves

The current design for monitoring bivalves is annual sampling of eight targeted stations distributed throughout the Bay. Bivalve collections (California mussel) have been made since 1981, but were generally not spaced evenly over time, with some years

sampled in both the spring and fall, while other years were only sampled in the fall. Power analyses to evaluate long-term trend detection using the bivalve dataset are detailed in Appendices I and II. These appendices describe bivalve analyses using two techniques to demonstrate their robustness to multiple power analysis approaches. Both analyses included a rigorous treatment of spatial and temporal effects on bivalve contaminant concentrations. Similar evaluations were performed for other matrices, but given special attention here due to the fixed monitoring design and the significant spatial and temporal variations reported previously (Gunther *et al.* 1999). An overall declining trend was evident in PCBs (Appendix I, Figures 5 and 6). However, the trend was significantly different for a few stations in the Bay. Notably, Fort Baker/Horseshoe Bay, Pinole Point, and Red Rock/Richmond Bridge had significantly lower bivalve PCB concentrations than the other stations. Estimates of variability in both analyses accounted for the spatial and temporal effects on bivalve contaminant concentrations.

Power estimates for all three pollutants were very high. The power to detect a 3.5% annual decline over 20 years would be > 99% in all cases. The results of both analyses suggested that a significant reduction in sampling frequency or number of sampling locations would still achieve > 90% power (Appendix I - Table 2; Appendix II - Table 5). Therefore, the RMP could substantially reduce bivalve monitoring frequency or number of sites without affecting the program's ability to detect trends. For example, a reduction of sampling frequency to every 2 – 5 years would not impair the ability to determine long-term trends (Appendix II - Table 5). Neither would a reduction in sample size from 10 to 7. Therefore, due to the relatively high cost to mobilize the bivalve

sampling, the TRC recommended modifying the current sampling regime (Table 7) to biennial sampling of the same stations monitored in previous years.

Sport fish

A total of five fixed sites are currently monitored for shiner surfperch and white croaker every three years (SFEI 2006a). The power of the current design to detect the assumed decline rates would be > 90% in both species (red cells in Table 10). Power for sample designs with less frequent monitoring and reduced samples indicated that PCBs would be more sensitive to such changes than mercury.

Sport fish were also evaluated for the ability to detect concentrations that are below thresholds (Table 3). Results for mercury in shiner surfperch indicated that a sample size could be reduced considerably while still retaining relatively high power (Table 11). To achieve 95% power to be below thresholds, only four samples would be needed to detect mercury concentrations. However, for mercury in croaker and PCBs in croaker and shiner surfperch concentrations have been well above management thresholds (SFEI 2006a). Therefore, the average concentrations were adjusted to 20% below the respective threshold to facilitate the one-tailed comparison. Consequently, for mercury in white croaker the power analysis estimated that the number of samples needed to achieve 95% power to be more than 40. For PCBs in both croaker and shiner, the estimated sample size for 95% power was more than 50. In general, the number of samples required to achieve either 80% or 95% power in these scenarios are much higher than could be reasonably collected with the available resources.

Reduction in the number of samples for each species was considered by the TRC. However, the RMP sport fish dataset is considered to be extremely important for management decisions and human health evaluations. A reduction in number of samples was not considered to represent a significant savings in cost or time. Therefore, the TRC recommended that no changes be made to the current sport fish monitoring design (Table 7).

Cormorant Eggs

Monitoring of bird eggs to date has only been performed in special studies. A recommended design for continued cormorant egg monitoring has recently been proposed based on our power analysis results (Davis et al. 2007b). For DDTs (8% annual decline), PCBs (6% annual decline), and mercury (3% annual decline) there would be > 80% power to detect these trends if at least three samples were collected every 1 – 3 years (Table 12). However, if mercury declines more slowly (1% per year), there would be a much smaller chance of detecting that trend. Davis *et al.* (2007b) recommended that three composites of seven eggs per site be collected every three years to detect the expected trends in all three pollutants. This judgment was largely based on a power requirement of 80%, cost considerations, and priority relative to other monitoring elements. Notably, the current dataset for cormorant eggs is relatively small, and exhibits higher variability than other matrices. Given budgetary constraints, 95% power could not be expected for most of the contaminant scenarios. Therefore, the ability to detect trends in cormorant eggs will not be as powerful as predicted for other matrices.

Non-statistical Evaluations of Other Elements of the Program

In a number of instances, it was not possible to conduct statistical analyses to evaluate program elements. The TRC and SC evaluated each element in light of the following information: regulatory context; management guidelines; highlights of scientific findings; and recommended staff options. For each of the elements not addressed through the power analyses, a summary of the discussion and key points are presented below. A summary of changes made to future RMP sampling is included on Table 7. Further discussion of the changes can be found in the minutes from the redesign meetings that occurred in 2006.

Episodic Toxicity (renamed Causes of Toxicity)

With the changing use of pesticides (a shift from organophosphates to pyrethroids), the episodic toxicity program has moved from a focus on water column toxicity to sediment toxicity (see for example the discussion of this issue in the 2003 Pulse of the Estuary). This element addresses the narrative objective in the Basin Plan that states "all waters shall be free of toxic substances in concentrations that are lethal to or that produce other detrimental responses in aquatic organisms". At present, the causes of episodic toxicity are not well understood. Considerable research has been conducted by Dr. Donald Weston and his group at University of California-Berkeley to demonstrate that the use of pyrethroids in urban areas may be responsible for toxicity observed in

urban streams. Preliminary work to date in the San Francisco Bay watershed has identified toxicity in select urban creeks (e.g., Lower San Mateo Creek). However, the causes of this toxicity have not been identified. The TRC and SC recommended that work towards this element be changed from annual to biennial.

Sediment Toxicity and Benthos

As discussed in the sediment chemistry section above, the TRC and SC approved the continuation of monitoring sediment chemistry annually. Monitoring sediment toxicity is conducted in concert with the sediment chemistry element, though at a more limited number of sites (27 vs. 47). Sediment toxicity addresses the narrative objective in the Basin Plan that states "all waters shall be free of toxic substances in concentrations that are lethal to or that produce other detrimental responses in aquatic organisms". To better understand the causes of sediment toxicity that are observed in the Bay, particularly in the winter, sediments will be monitored in alternating wet and dry seasons. It was originally proposed that sediment toxicity be reduced to 14 sites; however, this does not allow for sufficient sample coverage for each of the RMP segments of the Bay. The TRC and SC approved the analysis of sediment toxicity at 27 sites annually.

The State Regional Water Quality Control Board will soon adopt sediment quality objectives that are based on a triad approach – sediment chemistry, sediment toxicity, and benthos. To facilitate these evaluations in the San Francisco Bay, the RMP will conduct benthic assessments at the 27 sites where sediment toxicity is being conducted.

Small Fish

To date, monitoring of contaminants in small fish has been conducted as a pilot study under the review of the Exposure and Effects workgroup. Small fish are good spatial indicators as they have limited home range and are good temporal indicators as the sampled fish are less than a year old. The small fish pilot study commenced in 2005 with an evaluation of mercury concentrations in benthic and pelagic species at eight locations within the Bay, and this sampling has since been conducted annually. The preliminary results indicate significant mercury variation among species and location. In 2007, small fish will also be evaluated for organic compounds (e.g., PCBs and PBDEs). This element addresses the small fish monitoring objective listed in the mercury TMDL (0.03 ug/g), assists in the evaluation of risks to piscivorous wildlife, and aids in the development of food web models.

The small fish monitoring program is an important tool for evaluating the bioavailability of contaminants for uptake into the food web. The Contaminant Fate workgroup has placed a high priority on this monitoring element and has proposed that it be expanded for 2008. The scope of the program has not yet been determined and will be reviewed by both the Contaminant Fate and Exposure and Effects workgroups.

Tributary Loading

The RMP has conducted a number of special studies to evaluate contaminant loads introduced into the Bay from the Delta (San Joaquin and Sacramento Rivers) to the

north and the Guadalupe River to the south. This loading information has been helpful in the evaluation of trends, critical for model development, and important for TMDL development. It was recommended that the Delta and Guadalupe River loading studies be continued on a triennial basis.

Small tributaries represent one of the largest contributions of contaminant loadings to the Bay and as such are very important to the development of TMDLs and management actions to prevent the release of contaminants. In 2006, the RMP embarked on a special study to begin to characterize the loads from small watersheds surrounding the Bay. The first study has focused on a small industrial watershed located in Hayward. The Sources, Pathways and Loading workgroup, which oversees the tributary studies, recommended that characterization of small tributary loads be undertaken on an annual basis and this element characterize loads from a variety of watersheds. The TRC and SC endorsed this concept.

USGS Hydrography Studies

The USGS conducts monthly monitoring of phytoplankton, suspended sediment concentrations, temperature, dissolved oxygen, and light at 36 stations located along the spine of the Bay. The RMP contributes to a portion of the costs associated with this monitoring program (approximately 20%). This element addresses a narrative objective in the Basin Plan prohibiting biostimulatory substances in harmful amounts. The Basin Plan has also adopted standards for dissolved oxygen, temperature, and salinity.

Data from this program element has been used to track trends and recently identified a significant change in phytoplankton blooms (increased spring blooms and appearance of a fall bloom). Understanding phytoplankton dynamics is also important for understanding metal cycling in the Bay as metals are frequently taken up and released by phytoplankton. It is also useful for modeling Bay hydrodynamics and understanding the impacts of invasive species (e.g., the dearth of phytoplankton populations after the Asian clam invasion in the late 1980s). The TRC/SC endorsed continuation of this program element at the existing effort.

Summary

On the basis of the power analysis results and other considerations, the TRC made recommendations for future monitoring of contaminants in the Bay (Table 7). Water, sediments, bivalves, and sport fish are currently monitored under the RMP Status and Trends Program. The power for these matrices with the current sampling design and assumed rates of decline will be very high (> 95% in most cases). Sport fish and bivalves, in particular, appear to be the best indicators for trends (Table 10 and Appendix II, Table 5). This can be attributed to the rigorous treatment of spatial and temporal effects in our analyses. Specifically, the removal of these effects resulted in lower variability estimates in sport fish and bivalves (the unexplained variability), relative to water and sediments. For example, the between-year standard deviation for PCBs in shiner surfperch and white croaker were 0.02 and 0.04, respectively. However, the standard deviations for PCBs in water were much higher, ranging from 0.20 – 0.26, depending on the segment. Therefore,

in some cases, the power results for sport fish and bivalves were higher than for water and sediments. Future analyses could attempt to evaluate power based on estimates of unexplained variability in water and sediments, as was performed for bivalves and sport fish.

General discussion by the TRC also included an assessment of the contribution that each matrix has made to thresholds evaluations, management questions, and regulatory interest (*e.g.,* Appendix III.1). The TRC recommended that the sampling design for Status and Trends matrices other than sport fish be modified, due to adequate characterization of contaminant variability in these matrices for the majority of the Bay. Sport fish monitoring would achieve adequate power for trend detection, but due to high concentrations of mercury in white croaker, the ability to distinguish concentrations below thresholds could not be achieved with the current design. Nevertheless, the Committee deemed the value of the current sport fish monitoring design to be very high, particularly since the TMDL target matrix has begun shifting from water to sport fish (SFRWQCB 2006), and suggested that modifications not be made. Finally, cormorant eggs have previously only been monitored under special studies. A proposal for inclusion of cormorant egg monitoring every three years that included an evaluation of power was supported by the TRC, due to their value for assessment of long-term trends and regional patterns in mercury, PCBs, and other analytes such as dioxins and bioaccumulative emerging pollutants.

References

Connor, M. S., J. A. Davis, J. Leatherbarrow, B. K. Greenfield, A. Gunther, D. Hardin, T. Mumley, C. Werme, and J. J. Oram. 2007, in press. The slow recovery of the San Francisco Estuary from the legacy of organochlorine pesticides. Environmental Management.

Davis, J. A. 2004. The long term fate of polychlorinated biphenyls in San Francisco Bay (USA). Environmental Toxicology and Chemistry 23:2396-2409.

Davis, J. A., F. Hetzel, and J. Oram. 2006. PCBs in San Francisco Bay: Impairment assessment/Conceptual model report. San Francisco Estuary Institute and Clean Estuary Partnership, Oakland, CA.

Davis, J. A., J. A. Hunt, J. R. M. Ross, A. R. Melwani, M. Sedlak, T. Adelsbach, D. Crane, and L. Phillips. 2007. Monitoring pollutant concentrations in eggs of Double-crested cormorants from San Francisco Bay in 2002 and 2004: A Regional Monitoring Program Pilot Study. SFEI Contribution #434, San Francisco Estuary Institute, Oakland, CA.

Draper, N. R., and H. Smith. 1998. Applied Regression Analysis, 3rd edition. Wiley-Interscience, New York.

Greenfield, B. K., and J. A. Davis. 2005. A PAH fate model for San Francisco Bay. Chemosphere 60:515-530.

Greenfield, B. K., J. A. Davis, R. Fairey, C. Roberts, D. Crane, and G. Ichikawa. 2005. Seasonal, interannual, and long-term variation in sport fish contamination, San Francisco Bay. Sci. Tot. Environ. 336:25-43.

Gunther, A. J., J. A. Davis, D. Hardin, J. Gold, D. Bell, J. Crick, G. Scelfo, J. Sericano, and M. Stephenson. 1999. Long term bioaccumulation monitoring with transplanted bivalves in San Francisco Bay. Mar. Poll. Bull. 38:170-181.

Lowe, S., B. Thompson, R. Hoenicke, J. Leatherbarrow, K. Taberski, R. Smith, and D. Stevens, Jr. 2004. Re-design Process of the San Francisco Estuary Regional Monitoring Program for Trace Substances (RMP) Status & Trends Monitoring Component for Water and Sediment. SFEI Contribution 109, SFEI, Oakland, CA. 86 pp. http://www.sfei.org/rmp/rmp_docs_author.html.

Oros, D. R., W. M. Jarman, T. Lowe, N. David, S. Lowe, and J. A. Davis. 2003. Surveillance for previously unmonitored organic contaminants in the San Francisco Estuary. Marine Pollution Bulletin 46:1102-1110.

SFEI. 2005. 2003 Annual Monitoring Results. The San Francisco Estuary Regional Monitoring Program for Trace Substances (RMP). San Francisco Estuary Institute (SFEI), Oakland, CA. http://www.sfei.org/rmp/2003/2003_Annual_Results.htm.

SFEI. 2006a. 2005 Annual Monitoring Results. The San Francisco Estuary Regional Monitoring Program for Trace Substances (RMP). San Francisco Estuary Institute (SFEI), Oakland, CA.

SFEI. 2006b. The Pulse of the Estuary: Monitoring and Managing Water Quality in the San Francisco Estuary. SFEI Contribution #517, San Francisco Estuary Institute (SFEI), Oakland, CA. 88 pp. http://www.sfei.org/rmp/pulse/2006/index.html.

SFRWQCB. 2006. Mercury in San Francisco Bay. Proposed Basin Plan Amendment and Staff Report for Revised Total Maximum Daily Load (TMDL) and Proposed

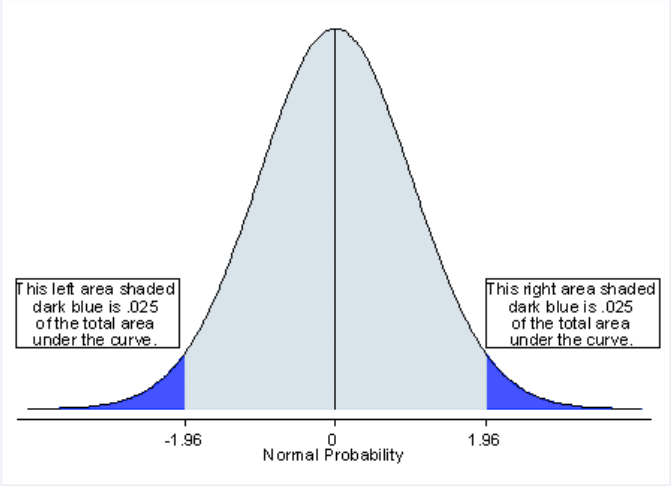
Mercury Water Quality Objectives. San Francisco Regional Water Quality Control Board, Oakland, CA.

Stow, C. A., L. J. Jackson, and S. R. Carpenter. 1999. A mixed-order model to assess contaminant declines. Environmental Monitoring and Assessment 55:435-444.

Wilkinson, L., G. Blank, and C. Gruber. 1996. Desktop data analysis with SYSTAT. Prentice Hall, Upper Saddle River, New Jersey.

Figure 1. Normal distribution of probability values associated with the test statistic (t) for one and two-tailed tests ($\alpha = 0.05$). In the two-tailed test (A), the "tails" of the distribution indicate significant lower (left side) or higher (right side) values. In the one-tailed test (B), the tail of the distribution only indicates a significant lower value. Therefore, the "rejection region" of the null hypothesis is larger, because a significant higher value is of no interest. Source: http://www.ats.ucla.edu/stat/sas/faq/pvalue.htm

(A)



(B)

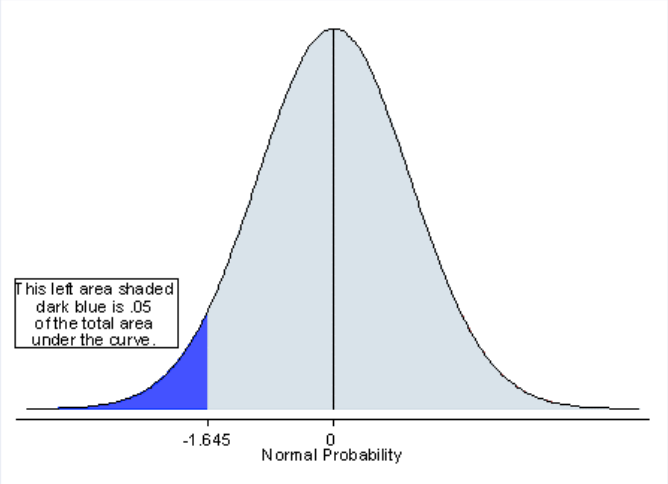


Figure 2. Schematic depiction of model to generate simulated data. Dashed line represents the simulated trend. Solid line with large dots represents simulated inter-annual variation (i.e., the trend plus ε_1). The small dots represent individual simulated data points (y_i), incorporating both inter-annual variation and within-year variation ε_2.

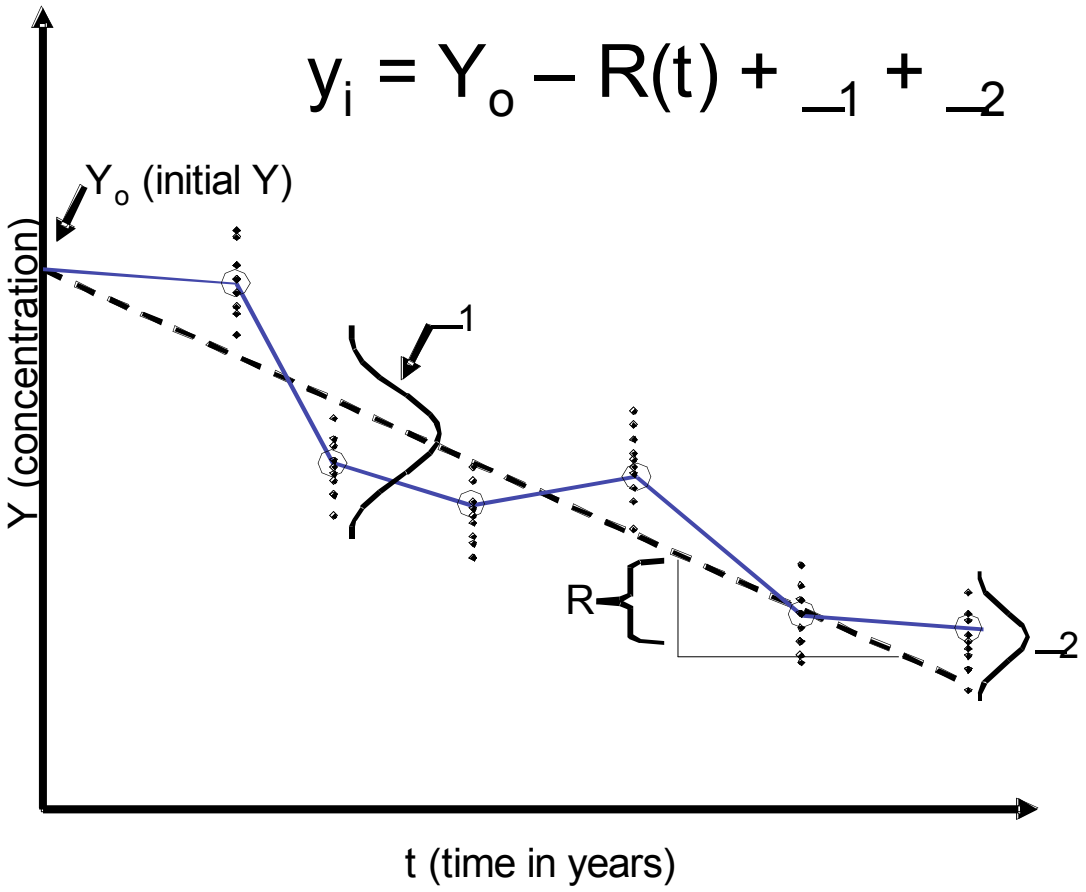
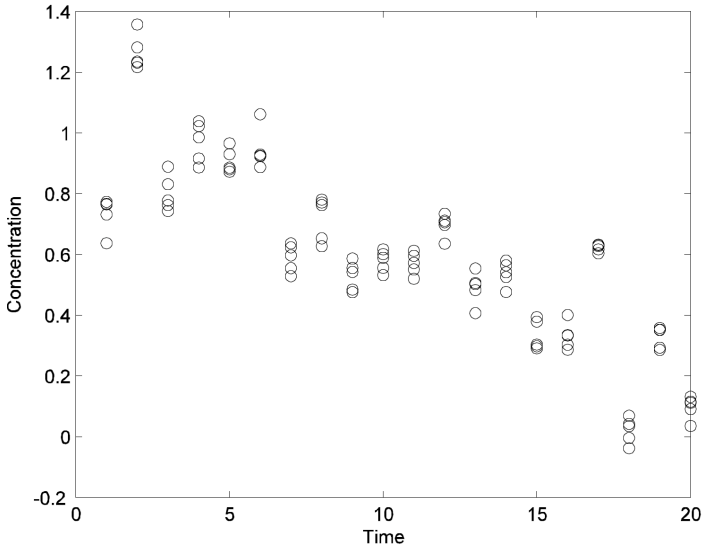


Figure 3. Example outputs from Monte Carlo simulations. Both simulations include 5 samples collected annually and are based on log-transformed data. **a**. Sample simulation result for relatively low variation, as compared to rate of decline. Parameter inputs were $Y_o = 1$, $R = -0.02$, $\varepsilon_1 = 0.15$, and $\varepsilon_2 = 0.05$. **b**. Sample simulation result for relatively high variation, as compared to rate of decline. Parameter inputs were $Y_o = 1$, $R = -0.01$, $\varepsilon_1 = 0.5$, and $\varepsilon_2 = 0.3$.

a.



b.

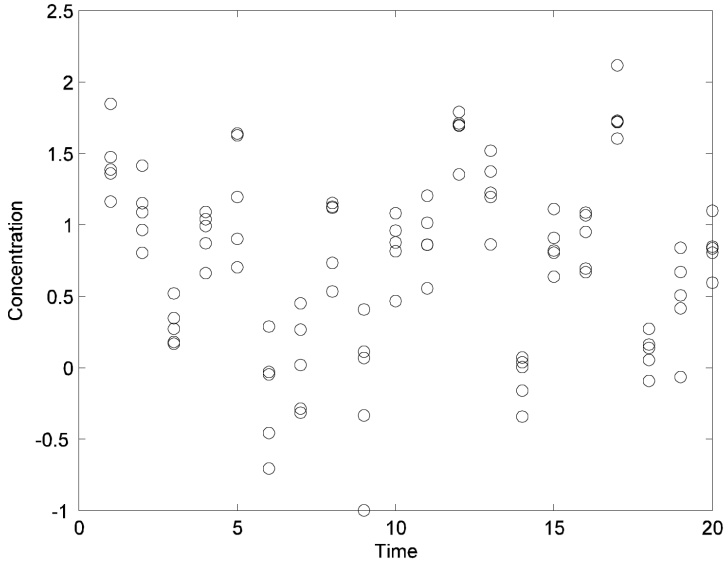


Figure 4. Example of Monte Carlo simulation results. Results are for evaluation of white croaker PCB concentrations, assuming a 1.12 % annual rate of decline (R = - 0.0112, ε_1 = 0.081, and ε_2 = 0.039). Sample size and interval combinations in red are expected to have adequate power. For example, on a 3 year sampling interval (current regime), 5 samples should be collected at each sampling event to have power > 80%.

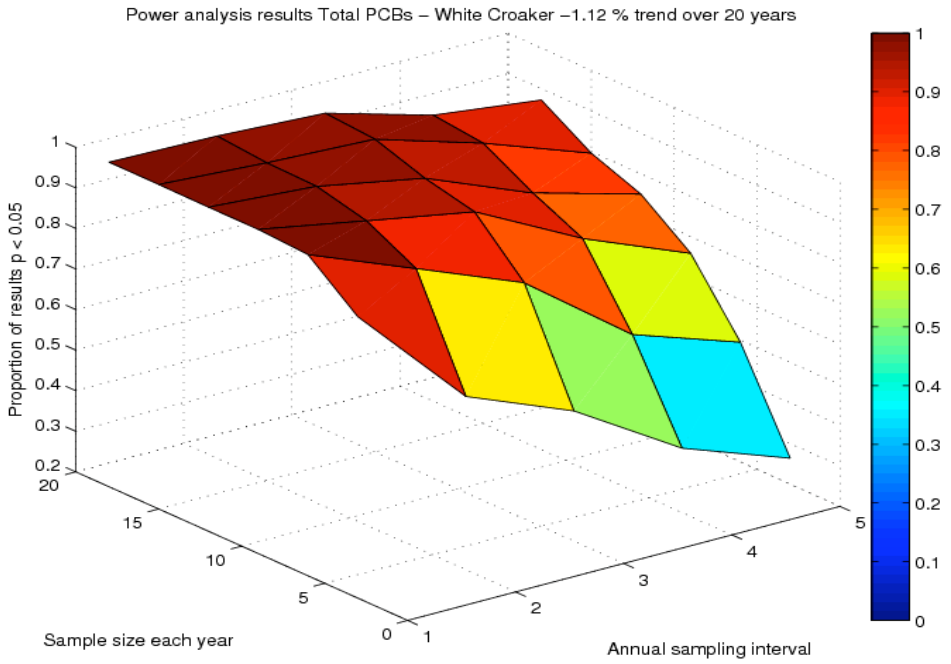


Table 1. Priority analyses and compounds. Trend Analysis: Evaluated power to detect trend given the decline rate and time frame. Threshold Analysis: Evaluated power to detect whether concentrations in a given year were below the threshold of concern.

	Trend analysis		Threshold analysis	Priority analyses					
Compound	**Time frame**	**Rate of decline**	**Relevant thresholds for current comparison**	**Water (total)**	**Water (dissolved)**	**Sediment**	**Bivalves**	**Bird eggs**	**Sport fish**
DDT	20 yr	3.5%/year	None	Trends			Trends	Trends (2)	
PCBs	20 yr	3.5%/year	PCB TMDL (fish tissue), CTR (water)	Trends/ Threshold		Trends	Trends	Trends (2)	Threshold/ Trends
Mercury	30 yr	1%/year	Hg TMDL (sediment, sport fish), CTR (water)	Trends/ Threshold (1)		Trends	Trends	Trends (2)	Threshold/ Trends
PBDE 047	20 yr	3.5%/year	None				Trends		
Copper	No trend work	No trend work	Revised thresholds (dissolved - 6 and 6.9)		Threshold				
Nickel	No trend work	No trend work	Revised thresholds (dissolved - 11.9)		Threshold				
Lead	No trend work	No trend work	CTR (occasional exceedances)		Threshold				

1. Note that the Hg TMDL requires evaluation of Hg in TSS, which was therefore used for the water evaluation
2. Modified rates of decline were used for bird eggs: For DDT, 8% per yr and 4% per yr; for PCBs, 6% per yr; and for mercury, 3% per yr and 1% per yr.

Table 2. Summary of power analysis for each matrix ($\alpha = 0.05$). Within-year standard deviation was estimated using the standard deviation of log-averages for data collected since the redesign (2002-2003). Between-year standard deviation was estimated using standard deviation of log-averages of all years (1994 – 2003). Methods section provides further details on these analyses.

Matrix	Power Analysis Scenario	Sampling Design	Contaminants	Estimated Variability	Effect Size Tested	Null Hypothesis	Alternative Hypothesis
Water	Scenario 1 – Power for concentration to be below threshold of regulatory significance	Random (annual)	PCBs (total) and Mercury (in TSS)	Within-year standard deviation	Concentration set to 80% of regulatory threshold compared to that threshold	≥ Threshold	< Threshold
		Random (annual)	Copper, Nickel, and Lead (dissolved)	Within-year standard deviation	Current mean concentration compared to regulatory threshold	≥ Threshold	< Threshold
Sport fish	Scenario 1 – Power for concentration to be below threshold of regulatory significance	Fixed (triennial)	Mercury and PCBs (in white croaker)	Within-year standard deviation	Concentration set to 80% of regulatory threshold compared to that threshold	≥ Threshold	< Threshold
		Fixed (triennial)	Mercury (in shiner surfperch)	Within-year standard deviation	Current mean concentration compared to regulatory threshold	≥ Threshold	< Threshold
		Fixed (triennial)	PCBs (in shiner surfperch)	Within-year standard deviation	Concentration set to 80% of regulatory threshold compared to that threshold	≥ Threshold	< Threshold
Water	Scenario 2 – Power for simulated decline in concentration over time	Random (annual)	PCBs and DDTs (total)	Between- and within-year standard deviation	Exponential annual decline of 3.5% over 20 years	Lack of trend	Significant declining trend
		Random (annual)	Mercury (total)	Between- and within-year standard deviation	Exponential annual decline of 1% over 30 years	Lack of trend	Significant declining trend
Sediment	Scenario 2 – Power for simulated decline in concentration over time	Random (annual)	PCBs	Between- and within-year standard deviation	Exponential annual decline of 3.5% over 20 years	Lack of trend	Significant declining trend
		Random (annual)	Mercury	Between- and within-year standard deviation	Exponential annual decline of 1% over 30 years	Lack of trend	Significant declining trend
Sport fish	Scenario 2 – Power for simulated decline in concentration over time	Fixed (triennial)	PCBs	Between- and within-year standard deviation	Exponential annual decline of 3.5% over 20 years	Lack of trend	Significant declining trend
		Fixed (triennial)	Mercury	Between- and within-year standard deviation	Exponential annual decline of 1% over 30 years	Lack of trend	Significant declining trend
Bivalves	Scenario 2 – Power for simulated decline in concentration over time	Fixed (annual)	PCBs, DDTs, and PBDE 047	ANCOVA (site and year variance)	Exponential annual decline of 3.5% over 20 years	Lack of trend	Significant declining trend
Double-crested Cormorant Eggs	Scenario 2 – Power for simulated decline in concentration over time	Fixed (special study)	PCBs	Between- and within-year standard deviation	Exponential annual decline of 6% over 20 years	Lack of trend	Significant declining trend
		Fixed (special study)	DDTs	Between- and within-year standard deviation	Exponential annual declines of 4% and 8% over 20 years	Lack of trend	Significant declining trend
		Fixed (special study)	Mercury	Between- and within-year standard deviation	Exponential annual declines of 1% and 3% over 30 years	Lack of trend	Significant declining trend

Table 3. The thresholds of management significance for RMP stakeholders and the Regional Board. These were used both in threshold analyses, and in determining the power to detect a declining trend sufficient to reach thresholds.

Compound	Matrix	Threshold	Source and basis	Segments[1]
Cu	Dissolved in Water	6.9 µg/L (ppb)	Tom Hall, Richard Looker, Peter Schafer, *pers. comm.* Revised Cu guidelines.	LSB
Cu	Dissolved in Water	6.0 µg/L (ppb)	Revised Cu guidelines.	SB, CB, SPB, SU
Hg	Total in TSS	0.2 ng/g (ppb) dry sediment	2006 TMDL Draft Basin Plan Amendment. Appendix A. Pg. 7	All
Hg	Sport Fish	0.2 µg/g (ppm) wet	2006 TMDL Draft Basin Plan Amendment. Appendix A. Pg. 4	All
Ni	Dissolved in Water	11.9 µg/L (ppb)	Tom Hall, Richard Looker, Peter Schafer, *pers. comm.* Revised Ni guidelines.	LSB
Ni	Dissolved in Water	8.2 µg/L (ppb)	Revised Ni guidelines.	CB, SPB, SU, SB
PCB	Total in Water	170 pg/L (ppq)	CTR to protect human health	All
PCB	Sport Fish	10 ng/g (ppb) wet	PCB TMDL. Fred Hetzel *pers. comm.*	All

Table 4. Compounds observed to exceed thresholds in recent RMP monitoring. These were the focus of the threshold component of the power analysis. Results are based on evaluation of Annual Monitoring Results (SFEI 2005).

				2002/2003		
Matrix	Constituent	Threshold (µg/L)	Threshold type	Number exceedances	N	Location of Exceedances
Water	Total Copper	3.7	"non-regulatory saltwater effects threshold"	15	60	SU, SPB
Water	Total Lead	3.2	"non-regulatory freshwater effects threshold"	3	60	SPB, LSB
Water	Total Hg	0.051	"lower South Bay site specific objective"	1	11	LSB
Water	Total Hg	0.025	"regulatory objective"	2	49	SPB
Water	Total Ni	7.1	"non-regulatory effects threshold"	4	49	SU, SPB
Water	Total Sum PCBs	0.17	"human health criterion"	54	60	All Segments

[1] Abbreviations for Bay segments in this document follow RMP conventions: SU – Suisun Bay, SPB – San Pablo Bay, CB – Central Bay, SB – South Bay, LSB – Lower South Bay.

Table 5. Power analysis results for detecting long-term trends in PCBs, DDT, and mercury in water. Results are based on estimated inter- and intra-annual variability for each segment, and assumed rates of decline. Red text represents the current monitoring design for each segment, and the blue areas highlight results that are > 95% power.

			Lower South Bay					South Bay					Central Bay					San Pablo Bay					Suisun Bay				
			Sampling Interval (years)					Sampling Interval (years)					Sampling Interval (years)					Sampling Interval (years)					Sampling Interval (years)				
			1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Scenario:		3	99%	89%	86%	73%	66%	100%	94%	87%	78%	69%	98%	84%	73%	59%	52%	99%	84%	76%	60%	53%	99%	86%	82%	68%	64%
PCBs		4	99%	93%	90%	79%	76%	100%	96%	93%	83%	76%	98%	89%	81%	68%	65%	99%	91%	83%	71%	63%	99%	92%	87%	77%	72%
Water (total)	Samples/year	5	100%	95%	91%	83%	78%	100%	97%	93%	87%	83%	99%	91%	83%	74%	70%	99%	95%	88%	77%	68%	100%	93%	89%	78%	74%
20 Year		6	100%	95%	93%	84%	80%	100%	97%	94%	86%	81%	99%	91%	87%	80%	70%	100%	93%	90%	82%	73%	100%	95%	90%	84%	77%
3.5% Annual Dedine		7	100%	96%	90%	85%	81%	100%	98%	96%	90%	86%	99%	92%	88%	80%	77%	100%	96%	94%	82%	77%	100%	94%	93%	85%	78%
		8	99%	97%	93%	86%	84%	100%	98%	95%	90%	85%	100%	93%	90%	83%	76%	100%	97%	94%	85%	79%	100%	96%	93%	86%	83%
Scenario:		3	96%	82%	74%	64%	56%	96%	84%	78%	66%	62%	100%	100%	98%	92%	87%	97%	81%	73%	59%	51%	99%	88%	81%	71%	69%
DDTs		4	98%	85%	78%	71%	65%	96%	86%	82%	73%	68%	100%	100%	99%	95%	90%	98%	88%	80%	69%	62%	99%	92%	87%	78%	74%
Water (total)	Samples/year	5	97%	88%	81%	73%	68%	99%	89%	84%	75%	74%	100%	100%	100%	96%	93%	99%	89%	83%	74%	65%	99%	93%	89%	80%	76%
20 Year		6	97%	89%	83%	75%	74%	98%	91%	86%	77%	73%	100%	100%	99%	98%	96%	99%	93%	87%	77%	72%	100%	93%	91%	82%	80%
3.5% Annual Dedine		7	99%	91%	86%	77%	74%	99%	91%	88%	80%	78%	100%	100%	100%	99%	96%	99%	94%	90%	80%	72%	99%	94%	90%	81%	81%
		8	98%	90%	87%	79%	73%	99%	93%	90%	82%	77%	100%	100%	100%	99%	97%	100%	94%	89%	80%	77%	100%	96%	92%	83%	83%
Scenario:		3	99%	88%	72%	67%	50%	96%	80%	67%	65%	57%	79%	52%	40%	33%	25%	60%	36%	29%	25%	17%	100%	96%	85%	85%	68%
Mercury		4	100%	92%	82%	77%	64%	97%	87%	75%	73%	60%	89%	64%	50%	43%	33%	74%	47%	33%	31%	21%	100%	98%	91%	90%	76%
Water (total)	Samples/year	5	100%	96%	87%	83%	71%	98%	90%	80%	76%	68%	92%	68%	54%	51%	39%	80%	53%	43%	36%	26%	100%	98%	94%	90%	85%
30 Year		6	100%	98%	91%	90%	75%	98%	90%	80%	82%	70%	95%	76%	58%	54%	45%	86%	60%	48%	46%	29%	100%	99%	96%	95%	83%
1% Annual Dedine		7	100%	99%	94%	91%	78%	99%	92%	85%	81%	75%	96%	79%	63%	59%	50%	89%	67%	54%	49%	36%	100%	99%	97%	95%	87%
		8	100%	100%	95%	94%	83%	99%	93%	87%	85%	74%	97%	84%	66%	67%	54%	93%	70%	57%	52%	38%	100%	100%	98%	97%	91%

Table 6. Comparison of mean, intra-annual variance, and coefficient-of-variation (CV) for PCBs, mercury, and DDT measured in each Bay segment for water and sediments (2002 – 2005). Red represents the highest CV for each contaminant-matrix combination.

Contaminant	Matrix	Segment	Intra-annual Variance	Mean	CV	Segment Area[2]
PCB (pg/L)	Water	LSB	184.4	943.1	19.55	5
		SB	151.6	452.7	33.48	144
		CB	316.5	642.1	**49.29**	382
		SPB	229.3	478.5	**47.91**	181
		SU	72.86	200.0	36.43	72
Hg (µg/L)	Water	LSB	0.005	0.009	**60.95**	5
		SB	0.002	0.006	37.86	144
		CB	0.003	0.006	50.00	382
		SPB	0.014	0.022	**65.36**	181
		SU	0.004	0.014	31.06	72
DDT (pg/L)	Water	LSB	104.5	285.6	36.58	5
		SB	23.75	101.3	23.44	144
		CB	73.30	191.2	**38.35**	382
		SPB	165.0	335.5	**49.19**	181
		SU	53.37	329.4	16.21	72
PCB (µg/kg)	Sediment	LSB	1.770	5.792	30.61	8
		SB	1.880	5.356	35.12	185
		CB	2.530	6.321	40.08	396
		SPB	4.130	4.686	**88.18**	227
		SU	1.270	1.768	**71.75**	80
Hg (µg/kg)	Sediment	LSB	0.04	0.257	15.52	8
		SB	0.05	0.218	21.91	185
		CB	0.05	0.245	22.23	396
		SPB	0.08	0.265	**30.67**	227
		SU	0.09	0.141	**61.10**	80

[2] Segment areas for sediment are larger than water because sediment sampling accounts for topography whereas water sampling does not.

Table 7. Summary of recommendations for redesign of specific matrices monitored in San Francisco Bay.

Element	Current Design	Recommended Option	Comments
	No. of Sites	No. of Sites	
Water Chemistry	31	22	Recommend to reduce the number of sites.
Sediment Chemistry	47	47 in summer, 27 in winter	Recommend to continue annual sampling, but alternating between wet and dry seasons.
Sediment Toxicity	27	27	Strong interest in determining causes of toxicity on an annual basis. Toxicity signal is stronger in winter.
Benthos	0	27	Benthos samples will be collected at the same sites as sediment toxicity on an annual basis.
Bivalves	11	11	Recommend to reduce frequency to biennial sampling.
Sport fish	5	5	Stay with the status quo. Five sites sampled triennially.
Small fish	8 as a pilot study	TBD	Recommended annual small fish sampling. To be expanded in 2008.
Double-crested Cormorant Eggs	3	3	Recommend to add matrix to Status and Trends. Monitor three stations triennially.
Tern Eggs	Pilot study in 2002 and 2003	TBD	Tern egg monitoring has largely been conducted at one colony in the South Bay. This triennial element will be developed in concert with the USGS.

Element	Current Design	Recommended Option	Comments
	No. of Sites	**No. of Sites**	
Large Tributary Loading and Guadalupe loading	Pilot study	1	Tributary loads from the Delta and the Guadalupe river will be monitored on a triennial basis
Small Tributary Loading	Pilot study	1	Rotate through Bay Area watersheds to quantify loads from small watersheds on an annual basis.
Causes of Toxicity (formerly Episodic Toxicity)	Variable	TBD	Recommended that this element be conducted every two years.
USGS Hydrography Monitoring	36	36	Recommended that this element be continued at its current level on an annual basis.

Table 8. Evaluation of power to determine that pollutant levels in water are below management thresholds. Future PCB and mercury concentrations were simulated by adjustment of the mean to 20% below their respective thresholds as described in the Methods.

	Total PCBs (Total)	Mercury (Total in TSS)	Copper (Dissolved)	Nickel (Dissolved)	Lead (Dissolved)	Current Design
Segment	**Number of Samples Required to Achieve 80% Power**					
Lower South Bay	5	12	2	2	2	4
South Bay	8	17	3	2	2	8
Central Bay	19	23	3	2	2	4
San Pablo Bay	23	14	2	2	Insufficient data	4
Suisun Bay	11	9	2	2	3	4
Segment	**Number of Samples Required to Achieve 95% Power**					
Lower South Bay	7	20	2	3	2	4
South Bay	12	28	4	3	3	8
Central Bay	32	40	3	3	2	4
San Pablo Bay	40	23	2	2	Insufficient data	4
Suisun Bay	18	14	2	2	3	4

Table 9. Power analysis results for detecting long-term trends in PCBs and DDT in sediment. Results are based on estimated inter- and intra-annual variability for each segment, and assumed rates of decline. Red text represents the current monitoring design for each segment, and the blue areas highlight results that are > 95% power.

		Lower South Bay					South Bay					Central Bay					San Pablo Bay					Suisun Bay				
		Sampling Interval (years)					Sampling Interval (years)					Sampling Interval (years)					Sampling Interval (years)					Sampling Interval (years)				
		1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Scenario:	2	100%	97%	92%	78%	67%	100%	99%	97%	85%	74%	100%	94%	87%	73%	60%	96%	75%	64%	43%	36%	76%	51%	37%	26%	23%
PCBs	4	100%	100%	98%	95%	89%	100%	100%	100%	99%	94%	100%	99%	97%	91%	86%	100%	95%	88%	73%	61%	93%	71%	64%	50%	43%
Sediment	6	100%	100%	100%	97%	95%	100%	100%	100%	100%	99%	100%	100%	100%	96%	92%	100%	97%	95%	86%	78%	98%	84%	74%	63%	55%
20 Year	8	100%	100%	100%	98%	96%	100%	100%	100%	100%	99%	100%	100%	100%	97%	95%	100%	99%	97%	90%	84%	99%	88%	82%	68%	62%
3.5% Annual Dedine	10	100%	100%	99%	99%	97%	100%	100%	100%	100%	100%	100%	100%	100%	99%	96%	100%	99%	98%	93%	89%	100%	92%	88%	72%	67%
Samples/year	12	100%	100%	100%	99%	97%	100%	100%	100%	100%	99%	100%	100%	100%	98%	97%	100%	100%	99%	95%	91%	99%	92%	87%	78%	71%
Scenario:	2	99%	87%	74%	72%	56%	100%	97%	88%	81%	62%	100%	99%	95%	93%	79%	100%	100%	95%	93%	81%	51%	29%	20%	17%	16%
Mercury	4	100%	94%	87%	83%	76%	100%	100%	100%	97%	89%	100%	100%	99%	99%	95%	100%	100%	99%	98%	93%	76%	48%	35%	34%	22%
Sediment	6	100%	96%	91%	90%	81%	100%	100%	100%	100%	96%	100%	100%	100%	100%	98%	100%	100%	99%	99%	97%	91%	60%	46%	44%	34%
30 Year	8	100%	98%	92%	92%	84%	100%	100%	100%	100%	99%	100%	100%	100%	100%	99%	100%	100%	100%	99%	97%	96%	74%	60%	55%	43%
1% Annual Dedine	10	100%	97%	94%	93%	85%	100%	100%	100%	100%	100%	100%	100%	100%	100%	99%	100%	100%	100%	100%	97%	98%	83%	67%	65%	45%
Samples/year	12	100%	98%	94%	93%	88%	100%	100%	100%	100%	100%	100%	100%	100%	100%	99%	100%	100%	100%	100%	98%	98%	89%	73%	68%	54%

Table 10. Power analysis results for detecting long-term trends in PCBs and DDT in sport fish. Results are based on estimated inter- and intra-annual variability across all sites, and assumed rates of decline. Red text represents the current monitoring design for each segment, and the blue areas highlight results that are > 95% power.

			Shiner Surfperch					White Croaker				
			Sampling Interval (years)					Sampling Interval (years)				
			1	2	3	4	5	1	2	3	4	5
Scenario:	Samples/year	**3**	100%	100%	100%	100%	100%	100%	100%	100%	100%	98%
PCBs		**6**	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Sportfish		**9**	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
20 Year		**12**	100%	100%	**100%**	100%	100%	100%	100%	**100%**	100%	100%
3.5% Annual Decline		**15**	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		18	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Scenario:	Samples/year	**3**	100%	100%	100%	100%	97%	100%	100%	100%	100%	98%
Mercury		**6**	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Sportfish		**9**	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
30 Year		**12**	100%	100%	**100%**	100%	100%	100%	100%	**100%**	100%	100%
1% Annual Decline		**15**	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		18	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

Table 11. Evaluation of power to detect pollutant levels in sport fish that are below thresholds. Future mercury concentrations in white croaker and PCB concentrations in croaker and shiner surfperch were simulated by adjustment of the mean to 20% below its threshold as described in the Methods.

	Mercury	Total PCBs	Current Design
Species	**Number of Samples Required to Achieve 80% Power**		
Shiner Surfperch	3	> 50	12
White Croaker	24	> 50	12
Species	**Number of Samples Required to Achieve 95% Power**		
Shiner Surfperch	4	> 50	12
White Croaker	41	> 50	12

Table 12. Power analysis results for detecting long-term trends in PCBs, DDT, and mercury in Double-crested Cormorants. Results are based on estimated inter- and intra-annual variability across two years of data collected at Richmond Bridge, and assumed rates of decline. Blue and green areas highlight results that are > 95% and > 80% power, respectively.

			Double-crested Cormorants Sampling Interval (years)				
			1	2	3	4	5
Scenario:	Samples/year	1	100%	96%	82%	51%	31%
DDT		2	100%	100%	99%	93%	88%
Cormorants		3	100%	100%	99%	98%	95%
20 Year		4	100%	100%	100%	98%	96%
8% Annual Decline		5	100%	100%	100%	98%	98%
Scenario:	Samples/year	1	79%	44%	31%	18%	14%
DDT		2	95%	73%	67%	56%	52%
Cormorants		3	98%	83%	77%	67%	62%
20 Year		4	96%	88%	83%	74%	69%
4% Annual Decline		5	98%	89%	85%	76%	72%
Scenario:	Samples/year	1	86%	51%	36%	20%	13%
PCB		2	97%	81%	72%	58%	53%
Cormorants		3	98%	88%	82%	70%	67%
20 Year		4	99%	91%	87%	77%	72%
6% Annual Decline		5	99%	93%	89%	80%	76%
Scenario:	Samples/year	1	99%	82%	62%	52%	32%
Mercury		2	100%	97%	88%	85%	70%
Cormorants		3	100%	99%	94%	93%	83%
30 Year		4	100%	99%	97%	96%	88%
3% Annual Decline		5	100%	100%	98%	97%	90%
Scenario:	Samples/year	1	29%	17%	11%	10%	8%
Mercury		2	54%	36%	29%	27%	22%
Cormorants		3	65%	47%	38%	37%	31%
30 Year		4	72%	55%	44%	43%	39%
1% Annual Decline		5	76%	58%	50%	50%	43%

List of Appendices

Appendix I. Ben Greenfield's analysis of PCBs in bivalves

Appendix II. Andy Jahn's analysis of PCB, DDT, and PBDEs in bivalves

Appendix III. Example of the format for presentation of power analysis results to the TRC. Table III.1 illustrates water chemistry re-design options, including RMP objectives addressed through water sampling, power results for both trend and threshold scenarios, and cost estimates for each design. Appendix III.2 lists the RMP objectives and management questions.

Appendix IV. Peer review comments from Dr. Don L. Stevens, Jr. (Stevens Environmental Statistics, LLC).

Evaluation of PCB Trends in Bivalves to Determine the Power of Future Bivalve Monitoring

Ben Greenfield, SFEI (ben@sfei.org)

Background and Description of Data

The Regional Monitoring Program for Water Quality in San Francisco Bay (RMP), in combination with other monitoring programs, has collected considerable monitoring data on contaminant concentrations and other chemical parameters in water, sediments, and animal tissues in San Francisco Bay. Of the various types of matrices collected, bivalves have the longest time-series available. In some cases, one to two data points have been collected every year since 1981. However, these data may exhibit serial autocorrelation, violating required assumptions for standard linear modeling approaches, such as linear regression. Furthermore, collection dates have not been evenly spaced, with some years sampled in both the spring and the fall, and other years only sampled in the fall. This variability in sampling design, although frequently observed in long-term monitoring programs, makes it more difficult to apply standard time-series analysis methods.

In many cases, data exhibit apparent linear trends. Figure 1 depicts a time-series of polychlorinated biphenyl (PCB) concentrations in transplanted bivalves placed at Pinole Point in San Francisco Bay. PCB concentration in bivalves are presented on a log-scale. Linear regression analysis of these data indicated a declining trend over the entire sampling period. The regression model was based on log transformed concentration data, to minimize variance heteroskedasticity. Significant serial autocorrelation is not present in the data set (Figure 2), indicating that ARIMA models are not necessary. This is fortunate, given the fact that the limited sample size (N = 25) and inconsistent annual sampling frequency would have made application of time-series models very difficult. Similar results of a log-linear declining trend with no serial autocorrelation were observed for bivalves transplanted to another site (Treasure Island). However, there is strong correlation among the different sampling sites in seasonal PCB concentration results (e.g., Figure 3). For a number of sites, data collection only began in 1994, making detection of long-term trends more difficult (e.g., Figure 4). The use of least-square methods allows assessment of the simultaneous probability of observing declining trends at multiple sites.

Figure 1. Total PCB concentrations in *Mytilus sp.* bivalve mollusks collected from Pinole Point in San Francisco Bay between 1981 and 2000. All samples are bivalves collected from a relatively uncontaminated location and transplanted to Pinole Point for 3 to 6 months prior to collection (Gunther et al. 1999).

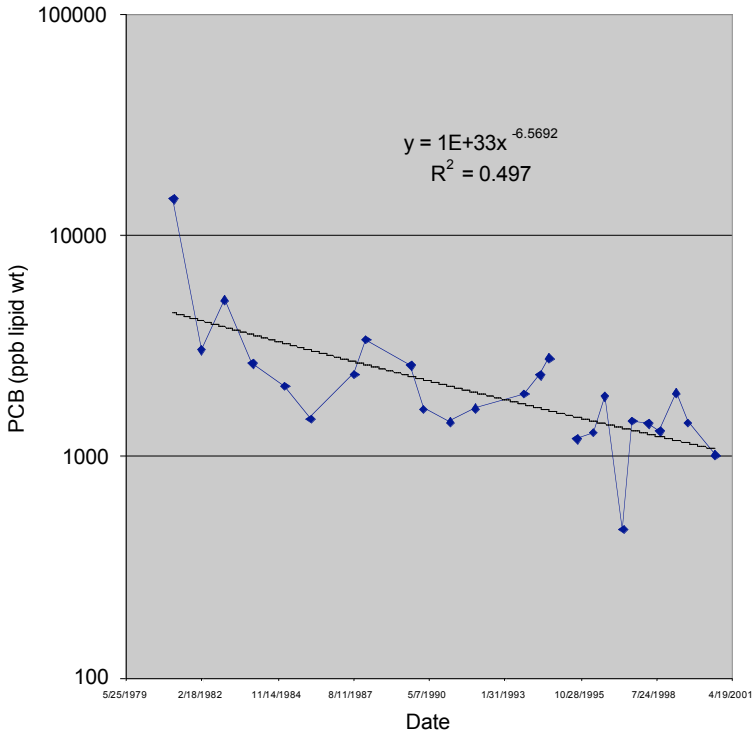


Figure 2. Plot of autocorrelation function for lags one through five using the residuals of the regression model in presented in Figure 1.

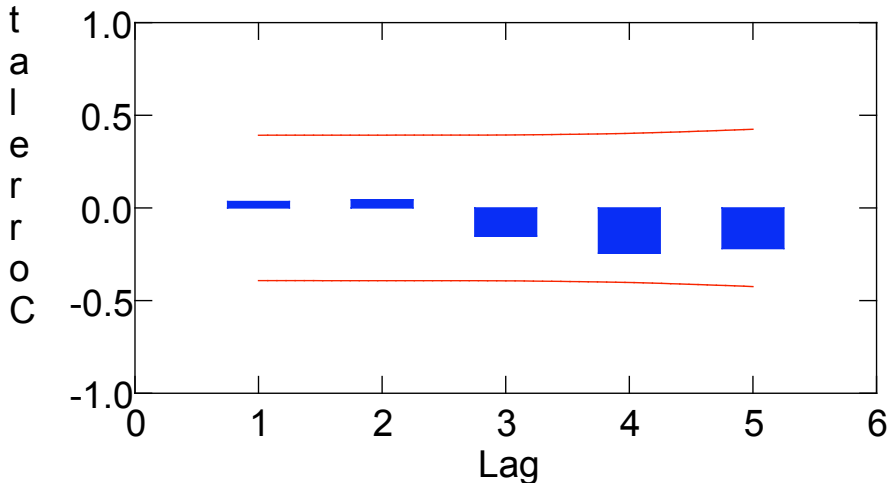


Figure 3. Total PCB concentrations in *Mytilus sp.* bivalve mollusks simultaneously collected from Yerba Buena Island, vs. Pinole Point.

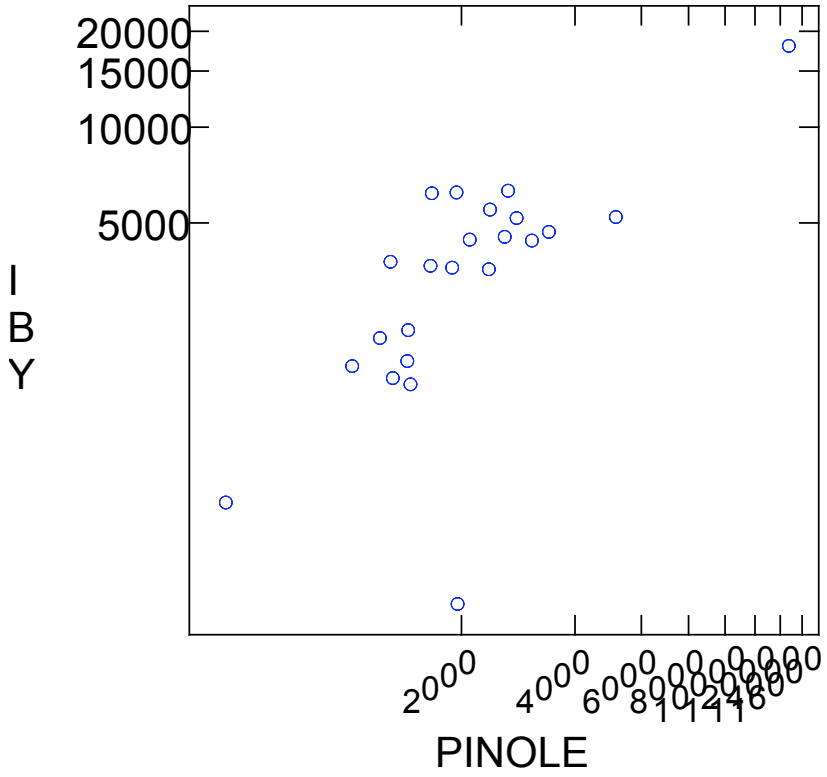
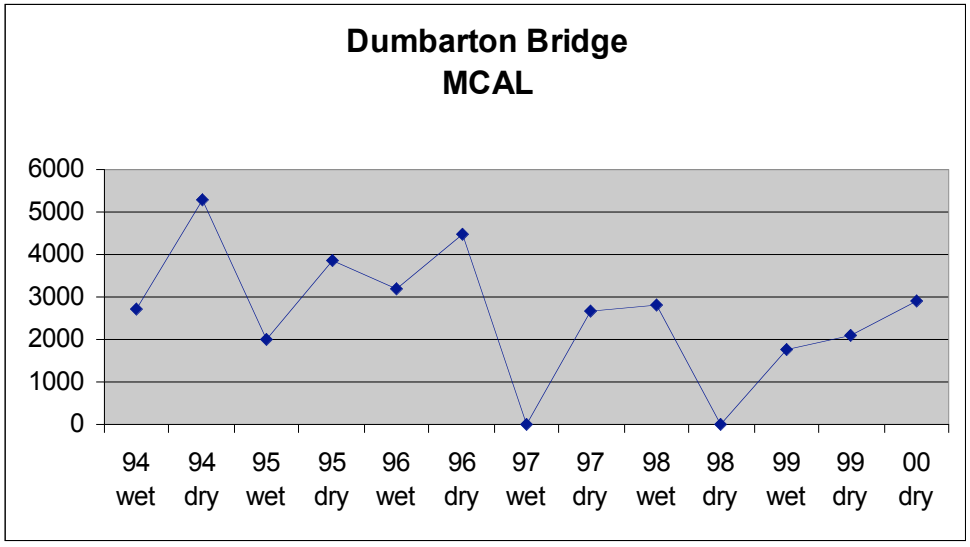


Figure 4. Total PCB concentrations in *Mytilus sp.* bivalve mollusks collected from the Dumbarton Bridge in San Francisco Bay between 1994 and 2000. All samples are prepared as described in Figure 1.



General approach

Linear regression was used to determine PCB trends in bivalves of San Francisco Bay. The goal was to determine the most likely value and range of possible values for the rate parameter that determine concentration declines of PCBs in the Bay. Analyses were performed on log-transformed PCB concentrations. Therefore, the results for the slope term represent a parameter for the exponential decay function. Both the slope and intercept terms may vary among sites due to different proximity from sources (and hence different concentrations or rates of change in concentration). To test this possibility, we tested for the effect of sampling station on slope and intercept using general linear modeling and analysis of covariance.

Finally, after fitting the best model, we used the residuals to estimate variability intrinsic to the bivalve data. This was used in power analysis to estimate the impact of changing sampling frequency on ability to detect long-term future trends with RMP bivalve monitoring.

Statistical methods

Prior to modeling, data were prepared by converting collection dates to a continuous variable representing time rounded to the nearest year and month (e.g., dates from March 15 to April 15, 1993 were converted to 1993.25). Date values were centered by subtracting the mean of all dates (mean = 1995.0811; N = 148). Bivalve concentrations were $\log_{10}$ transformed to stabilize variances, and because declines of legacy contaminants often follow exponential decay functions.

Linear regression was performed using standard parametric methods (Draper and Smith 1998) in SYSTAT 11.0 (Wilkinson et al. 1996), following the model:

$$Y = \alpha + \beta X + \varepsilon$$

where α represents the y-intercept parameter, β represents the slope parameter, ε represents a normally distributed variance term, X represents the centered date data, and Y represents the bivalve tissue concentration (lipid weight and log transformed).

General linear models using least squared regression indicated a significant effect of station on the bivalve concentrations, but not a significant effect of station on rate of decline. Based on this observation, dummy variables were created for changes in intercept (i.e., α) due to station, but not changes in slope (i.e., β) due to station. Analysis of covariance simulations were then performed following the model:

$$Y = \alpha_0 + \alpha_i D_i + \beta X + \varepsilon$$
$$i = [1, 2, 3, 4, 5, 6]$$

where α_0 represents an overall y-intercept parameter for the model, α_i represents six parameters for differences in y-intercept from α_0 as a function of sampling station, D_i represents a categorical (i.e., dummy) variable for six of the seven stations.

Results and Discussion

Linear regression

The parameter and confidence intervals were well estimated with low variability. The α (intercept) estimate was 3.349 with confidence intervals[3] (2.5% and 97.5%) of 3.308 and 3.389, respectively. The average β (slope) estimate was -0.0266 with confidence intervals of -0.0335 and -0.0197. The linear regression model was highly significant ($p < 0.0001$) but only explained a modest portion of the variability in the data set ($R^2 = 0.29$). In summary, the least squared regression indicated a statistically significant decline in PCBs in bivalves across the entire data set.

Least squared general linear models were generated to determine whether overall station effects were present for slope or intercept terms. Results indicated that in addition to the α and β, a highly significant station effect was observed for the overall model (p for station effect < 0.0001; final model $R^2 = 0.52$). A significant station*slope interaction was not present (f-ratio = 1.27; p = 0.28).[4] The interpretation of this is that overall, there was a significant difference in intercepts but not slopes among the stations examined. Another way of stating this is that the bivalves in different stations were sometimes different in PCB concentration overall, but that the rate of log-linear decay of PCBs did not vary among stations. Graphical analysis supported this interpretation, with the best linear fits for each station being generally similar in slope but having variable intercepts (Figure 5).

Linear regression with dummy variables (Analysis of covariance)

Based on the general linear model results, analysis of covariance was performed to determine which stations had significantly different y-intercepts due to station effects. As described in methods, six dummy variables were incorporated into the model ($D_{1,2,...6}$), representing differences in the intercept term ($\alpha_{1,2,...6}$). A backwards stepwise elimination model (p to remove of 0.05) retained three parameters for changes in intercept (α_2, α_4, and α_5). The final least squared regression model R^2 was 0.49, indicating that the model accounted for about half of the variability in the data set. Based on these statistical findings, we can conclude that Fort Baker/Horseshoe Bay, Pinole Point, and Red Rock/Richmond Bridge had significantly lower bivalve PCB concentrations over time than the other stations (Table 1).

[3] Calculated by adding or subtracting 2*SE to/from the parameter estimate

[4] This result was also obtained running a forward addition stepwise regression model, with dummy variables for effect of each individual station on slope or intercept. The final model included intercept effects but not slope effects.

Table 1. Least squared regression results for the final model. Only significant parameters were included.

Parameter	Mean	SD	Effect
alpha	3.459	0.023	Intercept (Baseline)
alpha2	-0.300	0.053	Fort Baker/Horseshoe Bay
alpha4	-0.265	0.047	Pinole Point
alpha5	-0.211	0.057	Richmond Bridge/Red Rock
beta	-0.026	0.003	Slope
sigma	0.044	NA	Variance

To evaluate the parameter estimates in a parsimonious model, the analyses was re-run including only parameters that were significantly different from zero. This model ($Y = \alpha_0 + \alpha_2 D_2 + \alpha_4 D_4 + \alpha_5 D_5 + \beta X + \varepsilon$) would be expected to have more accurate estimates for all parameters, because non-significant or marginally significant parameters would not exert undue influence on the important parameters.

These results may be used to estimate rate of decay of PCBs in bivalve tissues, and also to predict future rates of decay. The curve based on best model fit indicates a half-life of approximately 12 years (Figure 6). Such analyses can be used to estimate probable changes in PCBs over time, to develop management plans for the future (Davis 2004, Greenfield and Davis 2005).

Figure 5. Graphical depiction of trend data with best fit linear curves applied to the separate stations.

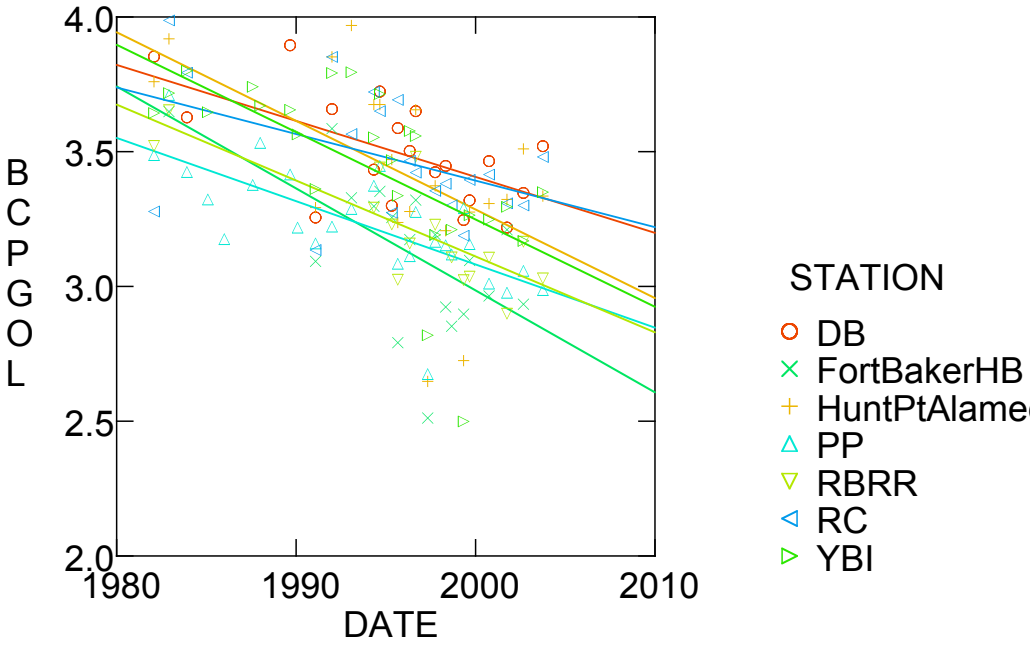
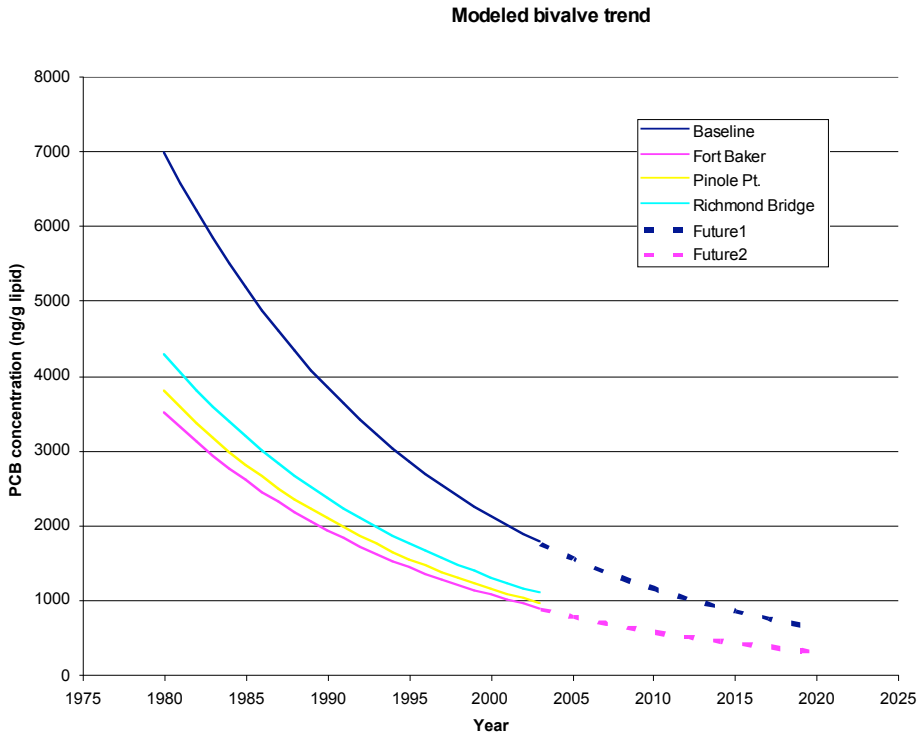


Figure 6. Bivalve trends predicted by the final selected model. Baseline represents trends for most stations (Yerba Buena Island, Dumbarton Bridge, Hunter's Point, and Redwood Creek). Other trends are as indicated. Dotted lines past 2004 represent future forecasted concentrations.



Power Analysis

To run the power analysis, estimates of both the within year and among year variability in the data set were calculated. Based on the above modeling efforts, I determined that variability could be inflated due to differences in intercept among station, as well as the trend in the data. Therefore, to estimate variability in the data, I obtained the residuals from the following regression equation:[5]

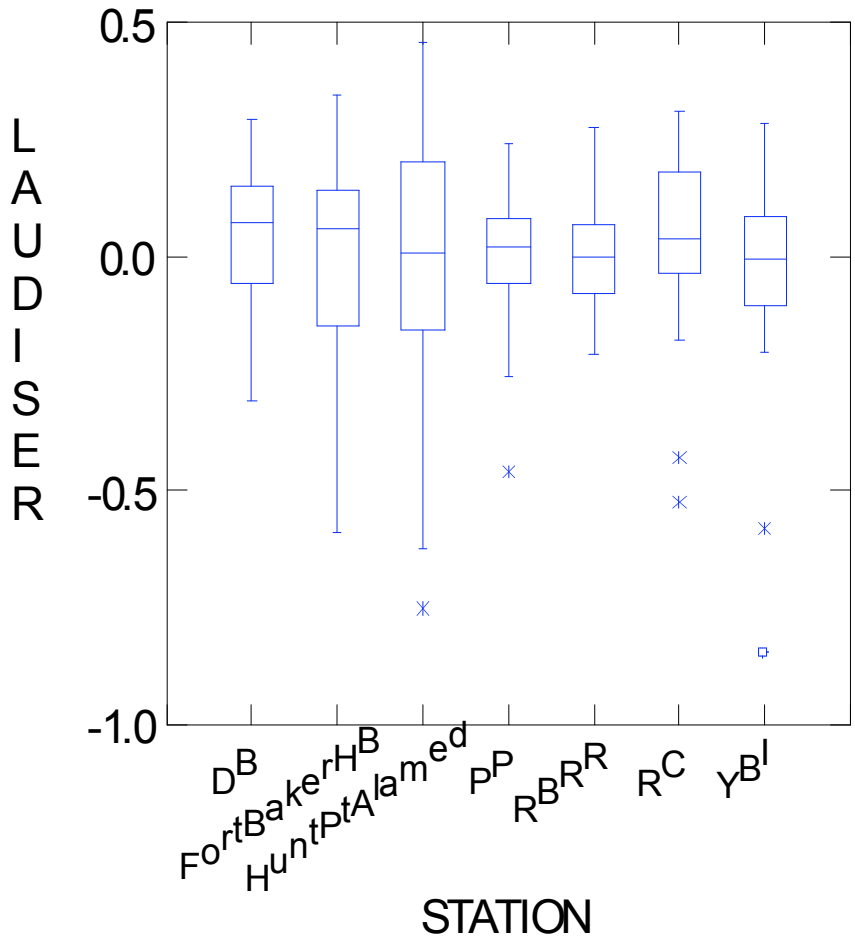
$$\log(\text{PCB concentration}) = \alpha_0 + \alpha_2 D_2 + \alpha_4 D_4 + \alpha_5 D_5 + \beta X + \varepsilon$$

Recall that α_0 is an intercept term, α_2, α_4, and α_5 are changes in intercept for Fort Baker/Horseshoe Bay, Pinole Point, and Red Rock/Richmond Bridge, respectively, and β is a slope term for change in time. X represents time, which has been centered by subtracting 1995.0811.

[5] Note that this is the same as the final selected model from the least squared regressions in the previous section.

These residuals should indicate the intrinsic variability in the data set, irrespective of date or sampling location. Plotting these residuals, we see that they are generally similar in average value across stations (Figure 7). There are some differences in variability across stations (Figure 10). Looking at the residuals across time, this variability appears to be due to some particularly low PCB years for some of the stations (Figure 8). For example, 1997 and 1999 had low residuals for Hunter's Point/Alameda and Yerba Buena Island. Perhaps the high flood event in 1997 reduced PCB bioavailability in those years.

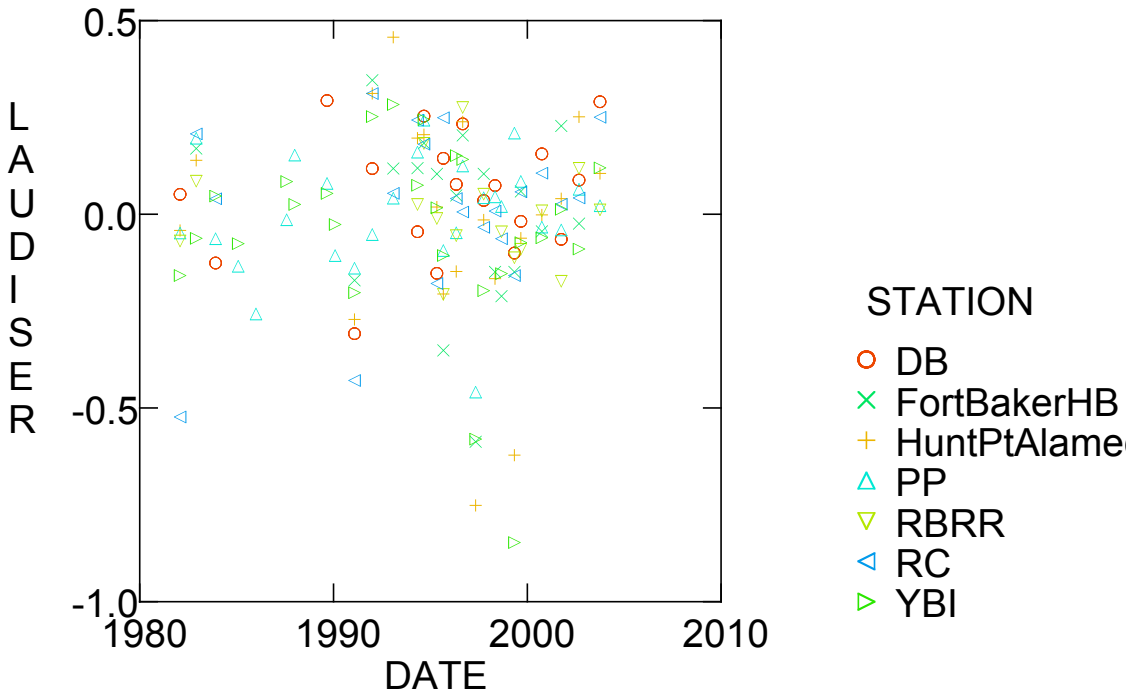
Figure 7. Residuals vs. station for estimating variance for power analysis.



These residuals were used to estimate interannual (among-year) and intraannual (within-year) variability for the bivalves. These variability estimates are required for the Matlab power analysis regression program. To estimate among-year variability, I took the average residual from each year and determined the standard deviation of these averages. To estimate within-year variability, I took the standard deviation of all the residuals collected within each given year, and then averaged all of these standard deviations. This resulted in an estimated within-year variability (standard deviation) of 0.116 and an estimated among-year variability (standard deviation) of 0.175. Note that the overall standard deviation of the residuals was 0.207.

These two components of variability were then combined into the MatLab trend detection program and 1000 simulations were run. A scenario of 50% decay in 20 years was run, with the power being equivalent to the ability of a linear regression model to detect a significant decline at $p < 0.05$. As summarized in Figure 9 and Table 2, using these assumptions, the bivalve monitoring would have plenty of power to detect long-term declines in PCBs in the Bay. In particular, in all scenarios, the power was well above 80% (often closer to 100%) (Table 2, Figure 9). If we were to treat the stations as independent replicates[6], a reduction in number of stations would not have any substantive effect on the ability to detect a log-linear decay over the 20 year time period. A reduction of the sampling frequency to every 2 or even 3, 4, or 5 years would not impair the ability to determine log-linear trends on a long-term basis. This result might change if the variance of other compounds (DDTs or PBDEs) turns out to be greater than that observed for PCBs.

Figure 8. Residuals from the best fitted model, vs. date, with symbols representing stations.

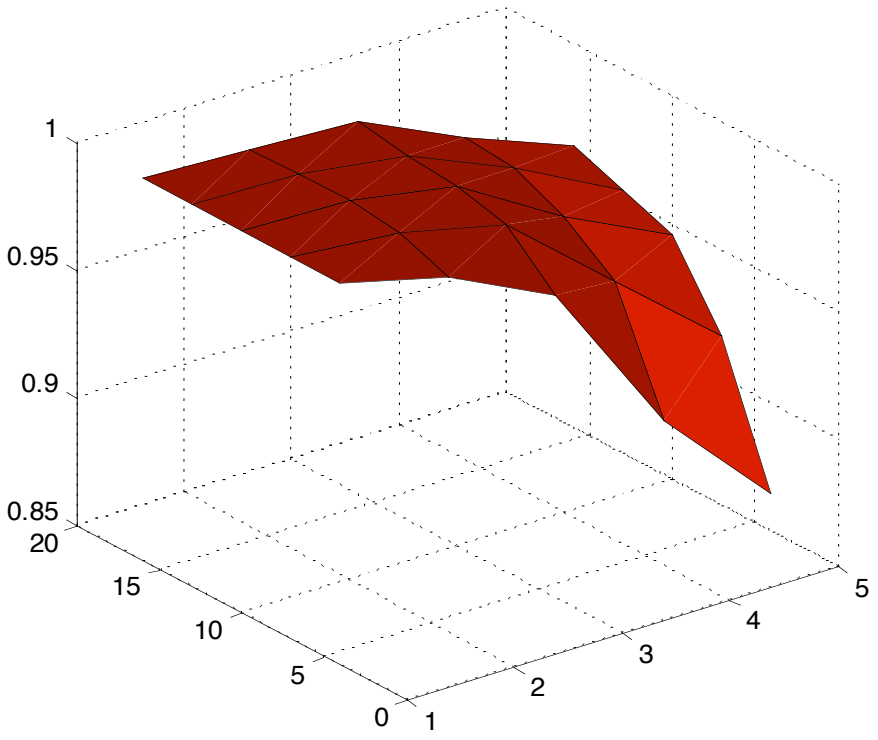


[6] Note that the Stations are NOT independent, as they strongly track each other (e.g., Figure 3). This statistical issue has not been addressed in this exercise.

Table 2. Power of the RMP bivalve sampling design to detect a significant ($p < 0.05$) decline when a 50% decline occurs over 20 years. Results are based on variance estimates from the historic RMP and State Mussel Watch PCB data, and other assumptions described in the text.

	Sampling interval (sample every X years)				
Samples/event	1	2	3	4	5
4	100%	99%	97%	91%	85%
7	100%	100%	98%	96%	92%
10	100%	100%	100%	96%	95%
13	100%	100%	99%	97%	94%
16	100%	100%	100%	98%	95%

Figure 9. Matlab output for power analysis of bivalves. Model variability estimates are based on PCB data from 7 RMP/SMW monitoring stations. Model is run for a range of sampling frequencies (sampling every 1 to 5 years) and sample sizes (3 to 16 samples per sampling period).



References:

Conaway, C. H., J. R. M. Ross, R. Looker, R. P. Mason, and A. R. Flegal. 2007. Decadal mercury trends in San Francisco Bay sediments. Environmental Research 105:53-66.

Connor, M. S., J. A. Davis, J. Leatherbarrow, B. K. Greenfield, A. Gunther, D. Hardin, T. Mumley, J. J. Oram, and C. Werme. 2007. The slow recovery of the San Francisco Estuary from the legacy of organochlorine pesticides. Environmental Research 105:87-100.

Davis, J. A. 2004. The long term fate of polychlorinated biphenyls in San Francisco Bay (USA). Environmental Toxicology and Chemistry 23:2396-2409.

Davis, J. A., F. Hetzel, and J. Oram. 2006. PCBs in San Francisco Bay: Impairment assessment/Conceptual model report. San Francisco Estuary Institute and Clean Estuary Partnership, Oakland, CA.

Davis, J. A., F. Hetzel, J. J. Oram, and L. J. McKee. 2007a. Polychlorinated biphenyls (PCBs) in San Francisco Bay. Environmental Research 105:67-86.

Davis, J. A., J. A. Hunt, J. R. M. Ross, A. R. Melwani, M. Sedlak, T. Adelsbach, D. Crane, and L. Phillips. 2007b. Monitoring pollutant concentrations in eggs of Double-crested cormorants from San Francisco Bay in 2002 and 2004: A Regional Monitoring Program Pilot Study. SFEI Contribution #434, San Francisco Estuary Institute, Oakland, CA.

Draper, N. R., and H. Smith. 1998. Applied Regression Analysis, 3rd edition. Wiley-Interscience, New York.

Greenfield, B. K., and J. A. Davis. 2005. A PAH fate model for San Francisco Bay. Chemosphere 60:515-530.

Greenfield, B. K., J. A. Davis, R. Fairey, C. Roberts, D. Crane, and G. Ichikawa. 2005. Seasonal, interannual, and long-term variation in sport fish contamination, San Francisco Bay. Sci. Tot. Environ. 336:25-43.

Gunther, A. J., J. A. Davis, D. Hardin, J. Gold, D. Bell, J. Crick, G. Scelfo, J. Sericano, and M. Stephenson. 1999. Long term bioaccumulation monitoring with transplanted bivalves in San Francisco Bay. Mar. Poll. Bull. 38:170-181.

Lowe, S., B. Thompson, R. Hoenicke, J. Leatherbarrow, K. Taberski, R. Smith, and D. Stevens, Jr. 2004. Re-design Process of the San Francisco Estuary Regional Monitoring Program for Trace Substances (RMP) Status & Trends Monitoring Component for Water and Sediment. SFEI Contribution 109, SFEI, Oakland, CA. 86 pp. http://www.sfei.org/rmp/rmp_docs_author.html.

Oros, D. R., W. M. Jarman, T. Lowe, N. David, S. Lowe, and J. A. Davis. 2003. Surveillance for previously unmonitored organic contaminants in the San Francisco Estuary. Marine Pollution Bulletin 46:1102-1110.

SFEI. 2005. 2003 Annual Monitoring Results. The San Francisco Estuary Regional Monitoring Program for Trace Substances (RMP). San Francisco Estuary Institute (SFEI), Oakland, CA. http://www.sfei.org/rmp/2003/2003_Annual_Results.htm.

SFEI. 2006a. 2005 Annual Monitoring Results. The San Francisco Estuary Regional Monitoring Program for Trace Substances (RMP). San Francisco Estuary Institute (SFEI), Oakland, CA.

SFEI. 2006b. The Pulse of the Estuary: Monitoring and Managing Water Quality in the San Francisco Estuary. SFEI Contribution #517, San Francisco Estuary Institute (SFEI), Oakland, CA. 88 pp. http://www.sfei.org/rmp/pulse/2006/index.html.

SFRWQCB. 2006. Mercury in San Francisco Bay. Proposed Basin Plan Amendment and Staff Report for Revised Total Maximum Daily Load (TMDL) and Proposed Mercury Water Quality Objectives. San Francisco Regional Water Quality Control Board, Oakland, CA.

Stow, C. A., L. J. Jackson, and S. R. Carpenter. 1999. A mixed-order model to assess contaminant declines. Environmental Monitoring and Assessment 55:435-444.

Wilkinson, L., G. Blank, and C. Gruber. 1996. Desktop data analysis with SYSTAT. Prentice Hall, Upper Saddle River, New Jersey.

Statistical Power Analysis of RMP Bivalve Tissue Samples

Andy Jahn, Statistical Consultant

Introduction

Statistical power is the complement of the type-two error rate, the probability of accepting the null hypothesis of no effect when, in fact, the effect exists. Thus in designing a program to have 95% power to detect an effect of a given magnitude, we intend to effect a type-two error rate ≤ 5%, or no greater than the customary type-1 error rate used in ecological work. Power is positively affected by the type-1 error rate (here held constant), size of the effect sought, and the size of the sample, and negatively affected by error variance. Ignoring the type-1 error rate, the factors under control of the investigator are the effect size, the sample size (usually equated with cost), and a sampling design that minimizes the proportion of the variance that ends up in the error term[7]. These factors will be explained, with examples, in the next section.

My task was to determine statistical power of the bivalve data set under future scenarios:

- Scenario 1. Power to detect a 50% decline in contaminant concentration over a 20-year monitoring period (3.4% per year average linear decline).
- Scenario(s) 2: Explore ways to reduce sampling costs while maintaining 95% power to detect a decline rate of 3.4% per year.
  A secondary task was to estimate the recent rates of decline in PCBs, DDTs, and PBDE047.

Methods

I used files of combined State Mussel Watch (SMW) and Regional Monitoring Program (RMP) data for California mussel supplied by Ben Greenfield and Jay Davis (PCBs) and Jennifer Hunt (DDTs and PBDE047). Only dry-weather data (June through October) were accepted. Lipid-normalized contaminant data, on a dry weight basis were log-transformed prior to analysis.

In the case where an investigator has analyzed a data set, accepted the null hypothesis, and wishes to know the power to detect an effect if one were present, the assumptions of the power analysis are the same as those attending the original hypothesis test, and results of a power analysis can be accepted with the same confidence. However, it is often the case that the investigator wishes to predict the power of a future data set. This is the situation here. To do this, we must make an assumption about the nature of the future

[7] As noted by Stevens (Appendix 4), this definition implies that the type of test and the form of the null and alternative hypotheses are already given. Explicitly, the test assumed here is an ANOVA in which the variance is partitioned into a fixed SITE effect and a linear trend in time called YEAR. The null hypothesis is that the slope due to YEAR is zero, and the alternative is that the slope is negative, i.e., that the tissue analyte concentration is declining.

data, i.e., that its variance structure will closely resemble that of the existing data set, except to the extent imposed by the stipulated effect size that we project on the data. In the present case, we maximize the power to detect a trend by first partitioning the variance into spatial and temporal components, as shown below (Table 1). The relative effects of site differences on future contaminant concentrations is impossible to know, and so acceptance of the analysis requires faith that the future will look like the recent past in this regard, i.e., that the ratio of variance due to SITE (SS_{SITE}) to that in the residual error (SS_{Error}) will remain the same[8] - in this example, 244:106. (I also assumed that, unlike the recent past, there will be no missing data as we go forward, i.e., that each analyte will generate a value at each site in each year.)

As shown in Table 1, the design of the sampling is basically that of a one-way ANOVA on sampling site (SITE) with time (YEAR) as a covariate. In effect, we are factoring out site differences to test for trends through time. To determine power, I used the methods of Cohen (1977), in which the effect size in the power analysis is calculated as **L**, the product of the degrees of freedom in the error term times a quantity called $\mathbf{f^2}$, the ratio of the proportion of variance due to YEAR ($\mathbf{PV}_{YEAR}$) to $\mathbf{PV}_{Error}$. **L** is thus a sort of signal-to-noise ratio, scaled to sample size.

Table 1. Partitioning of variance of log-transformed PBDE047 in California mussel from seven sites sampled in 2002, 2003, and 2005 (some sites dropped in some years due to missing data, or because the site had data for only one year; n=20).

Source	Sum-of-Squares	**PV**	df
SITE	.244	.480	6
YEAR	.158	.311	1
Residual Error	.106	.209	12
Total	.508	1	

From Table 1, we obtain $\mathbf{f^2}$ = .311/.209 = 1.49, and $\mathbf{L}=\mathbf{f^2} \bullet df_{Error} = 1.49 \bullet 12 = 17.9$. From Cohen's Table 9.3.2 for an ANCOVA with a type-1 error rate of 5% with a single covariate, we obtain power >99%. This is in accord with the F test for Table 1, which gives a highly significant result ($p \approx 0.001$) for the effect of YEAR.

[8] The discerning reader will note, as has Stevens (Appendix 4), that part, or even all of the trend in a short time series can be due simply to random error. In using this empirical estimate of the residual error proportion, my method probably over-estimates power, especially for PBDE047 (see the discussion of the slope in Figure 1, next section). For DDTs and PCBs, based on 13 and 19 years, respectively, the probable error is relatively small, and may be compensated for by my use of the non-directional F test for slope. That is, for the one-sided tests anticipated in the future (see Footnote 1), the rejection zone for the F tests will be double the size used here.

From Table 1 we can also calculate a partial correlation coefficient for the effect of YEAR on log PBDE concentration by dividing the sum of squares for YEAR by the sum of SS_{YEAR} and SS_{Error}, and taking the square root, i.e., $\mathbf{r} = (.158/(.158+.106))^{1/2} = -0.77$ (the sign will be confirmed in the next section). The power analysis of Scenario 1 proceeds in the reverse direction, by turning the slope of the projected decay rate into a correlation coefficient, and then adjusting the SS_{YEAR} (leaving SS_{SITE} and SS_{Error} unchanged, as discussed above) in Table 1 and re-calculating **L**:

A slope of 3.4% per year equates in log space to a regression coefficient (β) = -0.015 (log of 0.966). We then use the relation $\mathbf{r} = \beta \bullet (\sigma_x/\sigma_y)$, where σ_x and σ_y are the standard deviations of YEAR (x) and the part of the PBDE concentration not due to SITE effects (estimated as the std. dev. of the residuals from a one-way ANOVA of log PBDE vs. SITE). From the PBDE data used to generate Table 1, we get $\sigma_y = 0.118$. The standard deviation of 20 sequential digits is 5.916, leading to $r = -0.015 \bullet 5.916/0.118 = -0.752$. Squaring this result ($R^2 = .566$) and doing some algebra yields Table 2, from which we obtain $\mathbf{f^2} = .283/.217 = 1.304$ and $\mathbf{L} = 132 \bullet \mathbf{f^2} = 172$. Power under Scenario 1 is expected to be > 99%. This is not a surprise, because the modeled **r** (-.75) is only slightly less than the observed **r** (-.77), which generated a highly significant ANCOVA result with only three years of data (12 df in the error term vs. 132 under Scenario 1). Actual sums of squares in a 20-year data set will be much larger than the numbers in Table 2, but their ratios (**PV**s) under our assumptions are expected to be about as indicated. Scenario(s) #2 are calculated simply by figuring the degrees of freedom for various numbers of samplings and stations over a 20-year time period, as given in the results.

Table 2. Scenario 1 for PBDE047, in which a which a 50% decline in tissue concentration is modeled by adjusting SS_{YEAR} so that the ratio of SS_{YEAR} to the sum of $SS_{YEAR} + SS_{Error}$ = the modeled partial R^2 for YEAR.

Source	Sum-of-Squares	**PV**	df
SITE	.244	.500	6
YEAR	.138	.283	1
Residual Error	.106	.217	132
Total	.488	1	

Results

The results of the power projections are summarized in the last section, both for clarity and for the sake of readers with limited time for details.

PCBs

The RMP sum of PCBs, adjusted for lipid content and expressed as μg PCB per kg tissue, produced a data set spanning the years 1983 through 2003 with coverage of seven sites during the dry season. Geometric mean PCB concentration decreased from about 4500 ppb in 1983 to about 1900 ppb in 2003, for an average rate of -4.2% per year, somewhat greater than the rate of decline modeled in Scenario 1. The PCB file has month information for all the samples, so it was possible to express the date as a decimal fraction of the year (DYEAR). The variance partitioning is shown in Table 3. Here, **L** is 134 and Power >99%, in keeping with a highly significant ANCOVA.

The std. dev. of the residuals of a one-factor ANOVA on SITE is 0.219, giving a modeled variance structure, as calculated in the Methods section, of SS_{YEAR}= 0.626, PV_{YEAR} = .164, PV_{Error}= .362, and $\mathbf{f^2}$ = 0.453. For a 20-year program with no missing data going forward, we would have df_{Error} = 132 and L=60 for Power > 99%. Ninety-five percent power for an F test on an ANCOVA with a single covariate would be obtained when **L** = 13.34, which with $\mathbf{f^2}$ = 0.453 would require 29 degrees of freedom. We could approximate this by sampling every other year for 20 years at only 4 stations, giving df_{Error} =34 and Power = 97%. We might not want to reduce the number of stations so drastically, but it is obvious that there is a good deal of safety from the occasional missing data point, even with analyzing for PCBs every other year.

Table 3. Partitioning of variance of log-transformed sum of PCB compounds in California mussel from seven sites sampled from 1983 to 2003 (some sites dropped in some years due to missing data, n=82).

Source	Sum-of-Squares	**PV**	df
SITE	1.811	.318	6
DYEAR	2.504	.439	1
Residual Error	1.381	.243	74
Total	.508	1	

DDTs

After filtering for dry-weather data, the data set for DDTs reduced to RMP samplings from 1993 to 2005. Overall geometric mean DDT concentration ranged from 776 ppb in 1993 to 234 in 2005 for an average annual decline of -8.8%, far more rapid than the decline of -3.4% per year modeled in Scenario 1. The variance partitioning is shown in Table 4. Here, **L** is 102 and Power >99%, in keeping once again with a highly significant ANCOVA.

Table 4. Partitioning of variance of log-transformed sum of DDTs in California mussel from seven sites sampled from 1993 to 2005 (some sites dropped in some years due to missing data, n=79).

Source	Sum-of-Squares	**PV**	df
SITE	.592	.126	6
YEAR	2.428	.515	1
Residual Error	1.691	.359	71
Total	4.711	1	

The std. dev. of the residuals of a one-factor ANOVA on SITE is 0.230, giving a modeled variance structure, as calculated in the Methods section, of SS_{YEAR}= 0.296, PV_{YEAR} = .230, PV_{Error}= .656, and $\mathbf{f^2}$ = 0.175. For a 20-year program with no missing data going forward, we would have df $_{Error}$ = 132 and L=23 for Power > 99%. Sampling every other year would give 91% power. Sampling every year at four sites would give 95% power, and sampling 2 out of every 3 years (14 out of 20) at all seven sites would give 97% power.

PBDEs

PBDEs have only been sampled at seven sites since 2002 (Table 1). The average tissue concentration of PBDE047 in 2005 was 69% of that in 2002 (Figure 1), for an average annual decline of -8.8%. The back-transformed fitted slope from a log-linear regression is -15% with a 95% confidence band from -4% to -24%. (Of course, with such a small data set, there is no assurance that the trend is log-linear, and there is little basis for projecting it forward in time.)

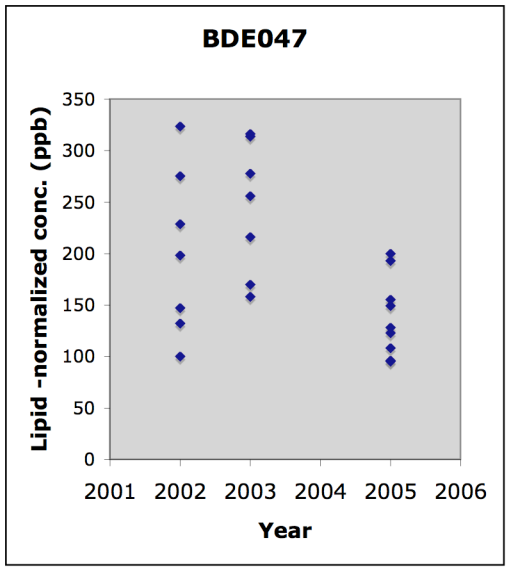


Figure 1. Scatter plot of PBDE concentration by year, all data.

As described in the Methods section, power for this small data set is >99%, in keeping with a highly significant ANCOVA. From Table 2, $\mathbf{f^2} = 1.304$ and the power to detect a 50% decline over 20 years is also >99%. Even under the reduced slope of Scenario 1, a three-year sampling program at all seven stations would have >99% power to reject the null hypothesis; the same program with four missing data points scattered among years and sites would have 95% power, and a 4-year program with four sites would give 96% power.

Summary

Power for the three tested classes of organic compounds in California mussel tissue is very good (Table 5), even under scenarios of reduced sampling frequency and reduced number of sites. In all three analytes, the observed rate of reduction in tissue concentration was greater than 3.4% per year, which is equivalent to a 50% decline over 20 years.

Table 5. Summary of the power analyses assuming a 50% decline in tissue concentration over a 20-year period.

Analyte	# Years*	# Sites	Power (%)
PCBs	20	7	>99
"	10	7	>99
"	10	4	97
DDTs	20	7	>99
"	10	7	91
"	14	7	97
"	20	4	95
PBDE047	20	7	>99
"	10	3	>99
"	4	4	96

* The scenarios are based on sampling every year (20) or every other year (10 of 20) except for the third scenario for DDTs, in which a two years on, one off schedule (14 of 20) is envisioned.

Reference

Cohen, J. 1977. Statistical Power Analysis for the Behavioral Sciences. (Rev.) Academic Press, 474 pp.

Appendix III.1 An example of the format for presentation of power analyses to the TRC. Description of RMP management objectives can be found in the Appendix III.2.

Design	Number of sites	Frequency	Season	Objectives Addressed 1 2 3 4 5	95% Trend Power in Each Segment	LSB SB CB SPB SUB	95% Threshold Power in Each Segment	LSB SB CB SPB SUB	Cost/yr
Status Quo	31	Annual	Summer		PCBs Hg DDTs		Hg on SSC PCBs Ni Pb		$450,000
4 sites per segment	25	Annual	Summer		PCBs Hg Se DDTs		Hg on SSC PCBs Cu Ni Pb		$400,000
3 sites per segment	20	Annual	Summer		PCBs Hg Se DDTs		Hg on SSC PCBs Cu Ni Pb		$340,000
Biennial	31	Biennial	Summer	?	PCBs Hg Se DDTs		Hg on SSC PCBs Cu Ni Pb	2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	$225,000
Triennial	31	Triennial	Summer	?	PCBs Hg Se DDTs		Hg on SSC PCBs Cu Ni Pb	3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3	$150,000

LEGEND KEY

High value for this objective
Medium value for this objective
Some limited value for this objective

Power greater than 95%

2 Assessment made every second year
3 Assessment made every third year

Appendix III.2 RMP OBJECTIVES AND MANAGEMENT QUESTIONS

Every five years, an outside group of scientific experts reviews the RMP to assure its fulfilling its objectives and providing useful and timely information regarding the Estuary. As part of the 2003 Program Review, the Review Panel stated "… that the Program must continue to evolve to ensure its long-term relevance." In response to this comment, the RMP reviewed lessons learned from data collected over the last ten years and developed a new set of management objectives based on this data and water quality management questions. These new management objectives were reviewed by the Technical Review and Steering Committees and approved in 2005.

1. Describe the distribution and trends of pollutant concentrations in the Estuary

1.1 Which pollutants should be monitored in the Estuary, in what media, and at what frequency?
1.2 Are pollutants of concern increasing, decreasing, or remaining the same in different media?
1.3 How are contaminant patterns and trends in the Estuary over time affected by remediation and source control or pollution prevention in the watersheds?
1.4 Do pollutant concentration distributions indicate particular areas of origin or regions of potential ecological concern?
1.5 What effects on beneficial uses or attainment of Water Quality Standards will occur due to large-scale habitat restoration in the Estuary in decades to come?

2. Project future contaminant status and trends using current understanding of ecosystem processes and human activities

2.1 Can reasonably accurate recovery forecasts be developed for major segments and the Estuary as a whole under various management scenarios?
2.2 Can potential impairment and degradation be better anticipated in the face of projected changes in land and water use and management, as well as product use and disposal?
2.3 Which pollutant categories are predicted to accumulate in the Estuary faster then they can be assimilated?
2.4 Do pollutant trends reflect historical changes in use patterns, transport and transformation processes, or control actions?
2.5 How will the importance of each pathway change through time under various management and development scenarios?
2.6 What is the projected future loading of pollutants of concern under various management and development scenarios?
2.7 What are the likely consequences of various management actions or risk reduction measures?
2.8 Do pollutants show existing distributions that fit our current understanding or models of their origin, loads, and transport?

2.9 What changes in loadings or ecosystem characteristics (e.g., extent of restored tidal marsh, Estuary circulation and flushing, food web shifts) would reduce or increase pollutant exposures and effects?
2.10 How are distributions and long-term trends in pollutants affected by current and predicted estuarine processes (e.g. sediment erosion, deposition, river inflows)?

3. Describe sources, pathways, and loading of pollutants entering the Estuary

3.1 Where are/were the largest pollutant sources, in what context are/were these pollutants applied or used, and what are/were their ultimate points of release into the aquatic environment?
3.2 What are the circumstances and processes that cause the release of pollutants from both internal and external source areas?
3.3 Once released, how do pollutants travel from source areas to the Estuary, what are the temporal and spatial patterns of storage, and are they transformed along the way or after deposition?
3.4 What is the annual mass of each pollutant of concern entering the Bay from each pathway?
3.5 Can data with high temporal resolution from a few watersheds be projected to other watersheds and the Basin as a whole?
3.6 For each pollutant of concern, what forms are released from each pathway and what are the magnitude and temporal variation of concentrations and loadings?
3.7 How do loads change over time in relation to management activities?
3.8 What is the relative importance of pollutant loadings from different sources and pathways, including internal inputs, in terms of beneficial use impairment?

4. Measure pollution exposure and effects on selected parts of the Estuary ecosystem (including humans)

4.1 How are emerging problems reflected in exposure and effects measurements?
4.2 Which (co-)factors (e.g., food web structure) influence exposure and effects of specific pollutants on biota?
4.3 What ecological risks are caused by pollutants of concern?
4.4 What human exposure to pollutants of concern results from consumption of fish and game?
4.5 To what extent does exposure to multiple pollutants lead to effects?
4.6 Which forms of pollutants cause impairment?
4.7 To what extent do factors other than specific pollutants (invasive species, flow diversions, land use changes, toxic algal blooms) contribute to beneficial use impairment?

5. Compare monitoring information to relevant benchmarks, such as TMDL targets, tissue screening levels, water quality objectives, and sediment quality objectives

5.1 What percentage of the Estuary is supporting beneficial uses?

5.2 Which segments should be considered impaired and why, and how do segments compare in terms of recovery targets?
5.3 How can specific source limitations, controls, and mitigation be best linked to appropriate beneficial use endpoints and recovery targets?

6. Effectively communicate information from a range of sources to present a more complete picture of the sources, distribution, fate, and effects of pollutants and beneficial use attainment or impairment in the Estuary ecosystem.

This objective applies to all of the questions listed under objectives 1 – 5.

Appendix IV

In the appendix that follows, peer review comments from Dr. Don Stevens, Jr. on the external draft of this report are provided. Dr. Stevens' comments suggested changes to the report to clarify analyses and improve upon the statistical techniques employed. Firstly, Dr. Stevens identified areas of the report that required further detail. For example, Dr. Stevens suggested that the description of calculating variance for the comparison of concentrations to regulatory thresholds could be improved upon, and explicit statements of null and alternative hypotheses should be provided. Response to these comments resulted in changes to the methods and results sections of the report. Secondly, Dr. Stevens commented on statistical approaches that could have been performed differently to those presented in this report. However, upon further discussion between co-authors and Dr. Stevens, SFEI decided to not pursue such analyses for this report, as Dr. Stevens indicated that the modified approaches would likely not change our overall interpretations. Future attempts by the RMP to perform power analyses and optimization of the Program should consider Don Stevens' recommendations of more sophisticated statistical techniques. These can be summarized as:

1. The use of power curves to evaluate designs in comparison of concentrations to regulatory thresholds.
2. Encompass GRTS design in the evaluation of spatial trends and inter-annual variation.
3. Improvements on regression techniques: specifically, in the presence of trends, the estimate of variance of the slope term must be modified.
4. In sport fish trend analyses, multifactor ANOVA may be more appropriate that an iterative forward stepwise regression, as the residual mean square error may be inflated due to variance explained by other factors.

Comments on
Power Analysis and Optimization of the
RMP Status and Trends Program

Don L. Stevens, Jr.
Stevens Environmental Statistics, LLC
December 28, 2007

Introduction

The Power Analysis and Optimization of the RMP Status and Trends Program Draft Report (Melwani, et al. (2007)) (hereinafter referred to as "PAO") consists of a main body and three appendices. The main body of the report discusses analyses to distinguish ambient conditions from relevant regulatory thresholds and to detect trend in ambient conditions. Details of power analyses to evaluate long-term trend detection using the bivalve dataset are in Appendices I and II. Appendix III provides an example format for presentation of the results and a summary of RMP management objectives.

The stated objectives of the report were to address the questions:

1. What power does the sample size of RMP stations provide to distinguish concentrations from relevant regulatory thresholds?
2. What is the power of the current sampling design to determine long-term trends?
3. Are there regions in San Francisco Bay where sampling intensity can be reduced in order to reallocate funds to higher priority items?

Power can be difficult to evaluate for a complex sampling design and its accompanying analysis. Both the design for data collection and the data analysis must be tailored to the complexities of population to be described. Overall, the PAO does a reasonable job of acknowledging of complexities of the RMP Status & Trends program and accommodating them in the power analysis. There are a number of instances where the analysis could be strengthened by using more appropriate techniques. Nevertheless, I think overall conclusions of the report would not be substantially impacted if the repeated using more sophisticated analytical methods.

One of the shortcomings of the PAO is a lack of clarity and precision in the discussion of how the objectives are addressed. To some extent, this may be due to the organization of the report. I found it difficult to determine exactly what analyses were carried out of which data sets. Also, it is not evident that the authors have a thorough understanding of statistical power and the appropriate methodology to assess power. This may be due to a lack of clarity in the discussion of methods and, in particular, exactly what null and alternative hypotheses were being evaluated. In order to give context to my comments, I begin with a background discussion of statistical power.

Andy Jahn began Appendix II with a very nice, concise explanation of statistical power and the factors that affect power. His opening paragraph is repeated here for reference:

> Statistical power is the complement of the type-two error rate, the probability of accepting the null hypothesis of no effect when, in fact, the effect exists. Thus in designing a program to have 95% power to detect an effect of a given magnitude, we intend to effect a type-two error rate ≤ 5%, or no greater than the customary type-1 error rate used in ecological work. Power is positively affected by the type-1 error rate (here held constant), size of the effect sought, and the size of the sample, and negatively affected by error variance. Ignoring the type-1 error rate, the factors under control of the investigator are the effect size, the sample size (usually equated with cost), and a sampling design that minimizes the proportion of the variance that ends up in the error term.

The only things that are missing from his discussion is mention of two implicit factors: the statement of null and alternative hypotheses (the null is not always one of "no effect") and the statistical test that is applied (some tests are more powerful than others). Both factors need to be explicit before a sensible power analyses can be performed.

Within the context of testing statistical hypotheses, null hypotheses (H_0), alternative hypotheses (H_A), and allied probability distributions are defined. Power is simply the probability of making the right decision, where the decision is based on outcome of an appropriate statistical test. More carefully, Power = Prob(Rejecting H_0 given that the true value is a point in H_A). In most real situations, the null and alternate hypotheses are not simple hypotheses consisting of a single value for a parameter, but are composite hypotheses. For example, in the regulatory context that is the backdrop for the RMP, a meaningful null hypothesis might have the form

> H_0: μ, the mean concentration of a contaminant in the target waterbody, exceeds a defined standard C

or

> H_0: $\mu \geq C$

and the alternative hypothesis the form

> H_A: the mean concentration is less than the standard

or

> H_A: $\mu < C$.

This is consistent with the formulation used in Lowe, et al., (2004):

> The tests addressed whether, within a survey period, the mean concentration of a chemical constituent in a sub-region was above the guideline for causing potential environmental harm. The null hypothesis of

> the statistical test addressing this question is that the actual mean concentration (μ) is above the guideline value.

It is also consistent with the approach used in evaluating Objective 1 (or Scenario 1), I think.

When both null is a composite hypothesis, the size of the test is usually computed at the boundary of the hypothesis, e.g., for the null $H_0 : \mu \geq C$, the size is evaluated at $\mu = C$. In effect, the composite null is replaced with a simple null hypothesis. When the alternative is a composite, the test doesn't have a single value of power. Rather, the power depends on the particular point in the alternative that happens to be true. The definition of power given above makes this point, as does Andy Jahn's discussion of effect size. The effect size is just the difference between the value of the null hypothesis and the selected point in H_A where power is evaluated. For this reason, power analyses frequently are expressed in the form of *power curves* where the effect size is varied over the values in the alternative hypothesis while the other factors are held constant. Power curves can be vary useful for informing decisions on sampling design and resource allocation, because alternative scenarios can be compared easily.

Summarizing, a proper power analysis needs explicit statements of sampling design, null and alternative hypotheses, analytic procedure used to test the null hypothesis, specification of the Type I error size, and the effect size (or the point(s) in H_A where power is to be evaluated.). The PAO would benefit greatly if these elements were clearly identified for all of the analyses carried out.

Comments on Scenario 1 Power Analysis

The analysis would benefit greatly from explicit statements of null and alternative hypotheses, and exactly where power is being evaluated. From the discussion on one-tailed versus two-tailed tests, I think the PAO uses $H_0 : \mu \geq C$ versus $H_A : \mu < C$. The power is apparently evaluated at the observed mean value, which was also done in Lowe, et al., 2004. That seemed odd to me when I initially reviewed Chapter 3 of Lowe, et al, (2004), and still seems like a very limiting approach to a power analysis. Furthermore, in some cases, it leads to contortions like

> "Some contaminants are currently well above threshold values. To simulate power to detect future pollutant levels that are below thresholds in these scenarios, concentrations were adjusted to 20% below its threshold, and the one-tailed comparison was made with this simulated data. (PAO, p.7)

Why not do an analysis that gives power curves as a function of departure from threshold? This avoids the necessity of special handling of cases where the observed mean is close to the standard, provides much more information for design consideration, and really doesn't take much more effort. Similarly, if sample size were one of the design parameters up for consideration, a power curve could be generated for a fixed effect size

and variable sample size. More generally, one could generate a power surface as a function of sample and effect size. The power surface could be presented as a two-way table. One could also determine the combinations of effect size and sample size that would give a specified power.

The PAO addressed Objective 1 assuming that H_0 was assessed using a t-test and referenced SYSTAT software for power. I'm not familiar with SYSTAT, but I presume that the power was calculated using a non-central t distribution. It would be reassuring if that were explicitly stated.

For a t-test applied to a particular population, the only population parameter that influences power is the variance. Everything else is under the direct control of the investigator. Thus, a good estimate of variance is essential to a realistic power analysis, so variance estimation should be the focus of the power analysis. This is the weakest point of the PAO: variance estimation is either not well-described or done in a very off-hand fashion.

A brief description of the design of the data collection should be provided. I'm familiar with the RMP design for water and sediment, but not for sport fish or bivalves. There is no mention made of how the sport fish data was collected (there is a citation that may have design information, but at least a summary should be included here). Design discussion is critical, because the statistical test used should be consistent with the manner in which data were collected.

The GRTS design used to collect sediment and water is not a simple random sample (SRS), and analysis of data collected with a GRTS design using a t-test is not strictly appropriate. However, a GRTS design will yield a lower standard error of the mean than an SRS design in the presence of spatial correlation of the response, so a t-test is likely conservative. It follows that the power analysis based on t-test should be conservative, that is, power using a test that is consistent with a GRTS design will be higher than the power using a t-test.

Comments on Scenario 1 Power Analysis

Most of the power analyses for trend were carried out using a simulation model. (A different approach was taken in Appendix II). The choice to evaluate power using a simulation model is very reasonable. However, when the choice is made to evaluate power via simulation, it is critical that the model chosen reflect all of the complexities of the actual population and design and that the parameters of the model are appropriately estimated. Ecological variation frequently has a very meaningful structure. If the design for gathering the data to be analyzed for trend detection accommodates the variation present in the population and measurement processes, some of that variation may have virtually no effect on the ability to detect trend. The ability to detect trends in a regional population sampled in annual (or longer) time steps can be influenced by four major components of variance: (1) population variance, (2) interannual variance, (3) site–year

interaction, and (4) response variance. Larsen, et al., (1995) provides an extensive discussion of these components in an ecological context.

From the PAO,

The model used in the simulation study was

$$y_i = Y_o - R(t) + \varepsilon_1 + \varepsilon_2 \qquad \textit{(Equation 1)}$$

Where, y_i = an individual simulated contaminant concentration sample, Y_o = the initial average concentration, R = annual rate of decline, t = time (in years), and ε_1 and ε_2 are normally distributed error terms that represent the intra- and inter-annual variation, respectively.

Larsen's interannual variance is equivalent to the ε_2 component of PAO Eq 1. The ε_1 component then encompasses every other source of variation.

The model used in the PAO may not be complex enough to form the basis for a realistic power analysis. A dominant feature of most ecological populations is that responses exhibit spatial pattern. A GRTS design exploits that spatial pattern by ensuring spatial balance of the sample locations, so that the resulting estimates are more precise that those resulting from simple random sampling. Since spatial correlation was not mentioned, I assume that the simulated error terms were generated independently, which probably does not reflect the true nature of intra-annual variation. Furthermore, the responses may also be temporally correlated, so that the inter-annual error term is correlated. The presence of correlation will tend to decrease the effective sample size, and hence decrease power. This may be balanced by the increased precision that a GRTS design provides. Furthermore, the several components of variation impact the ability to detect trend in different ways. If the components are present but not included in the model, a misleading estimate of precision can result.

According to the PAO (P12, L1), trend detection for simulated values was performed by linear regression. If statistical significance was carried out by the usual t-test from a standard regression package, then the significance level is almost certainly wrong. The slope estimate produced by ordinary least squares is appropriate, but in the presence of interannual variation, the standard error of the estimate must be modified. See, for example, the discussion in Larsen et al., (1995), Larsen et al., (2004), Urquhart, et al., (1998), or VanLeeuwen, et al.(1996) . For example, Larsen, et al. (2004) give an expression for the variance of the slope estimate that explicitly identifies the role of the role of the several components of variance.

Sport fish

I'm not sure that I follow the discussion of the analysis of the sport fish data. It sounds as if determination of the various components of variance were determined by applying a sequence of one-way ANOVAs, with each successive ANOVA applied to the residuals of the previous. If a factor did not show significance, then it was dropped in subsequent

analyses. The correct approach would be to do a multi-factor ANOVA. The problem in doing multiple single factor ANOVAs is that factor significance is understated so that important factors may be dropped. For example, below are three ANOVAs of some artificial data. The first is a one-way, & shows no significant treatment effect. The second is also a one-way, and shows a highly significant effect of Factor B. The third includes both factors, and shows significant effects for both factors. The lack of significance in the first ANOVA occurs because the residual mean square is inflated by the effects of Factor B. The one-way ANOVA for Factor B indicates a significant effect of B, but the residual mean square is still inflated by the effects of Factor A in comparison the two-way ANOVA. The residual mean square is going to have the most influence on the power to detect trend, and smaller is definitely better.

One-way Analysis of Variance Table (Factor A)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
A	4	38.012	9.503	1.7079	0.1798
Residuals	25	139.105	5.564		

One-way Analysis of Variance Table (Factor B)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
B	2	76.217	38.109	10.198	0.0005023 ***
Residuals	27	100.900	3.737		

Two-way Analysis of Variance Table (Factors A and B)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
A	4	38.012	9.503	3.4756	0.0232104 *
B	2	76.217	38.109	13.9375	0.0001084 ***
Residuals	23	62.887	2.734		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Another approach would be to estimate components of variance and trends simultaneously, say by using the approach given in VanLeeuwen, et al.(1996). They consider a mixed model with a trend, and show how to separate random effects of components such as site and year from fixed effects such as trend.

Appendices I & II

The methodology seems basically correct, but I'm concerned that relevant components of variance were not properly incorporated. Of particular concern is the inter-annual variation. For short time series, inter-annual variation can very easily be confused with trend. If inter-annual variation is not included in the model, a spurious estimate of trend and precision can result. For example, the ANOVA given in Table 1, Appendix II, estimates trend from a data set with six sites and three years of data. The ANOVA accounts for site variation but not inter-annual variation. The effect attributed to trend may very well be due to inter-annual variation, so the resulting power estimates could be

misleading. Also, if there is inter-annual variation, the estimate of the standard of the slope estimate must be modified, as commented on above.

Unfortunately, distinguishing random temporal variation from change due to real trend with only a short time series available is beset with difficulty. The split of the temporal variation into a stochastic component and a deterministic component is essentially arbitrary. Even though an estimate of the temporal slope may appear to be significantly non-zero, an alternative model that attributes the temporal variation to random interannual variation may provide an equally reasonable explanation. The difficulty is greatly reduced with longer time series: extraction of a trend component is much more feasible with 10 to 20 years of record. An analysis of trend detection power based on less than ten years of data should be interpreted with caution.

References

Larsen, D. P., N. S. Urquhart, and D. L. Kugler. 1995. Regional-scale trend monitoring of indicators of trophic conditions of lakes. *Water Resources Bulletin* 31:117–140.

Larsen, D. P., P. R. Kaufmann, T. M. Kincaid, and N. S. Urquhart. 2004. Detecting persistent change in the habitat of salmon-bearing streams in the Pacific Northwest. *Can. J. Fish. Aquat. Sci.* 61: 283–291.

Lowe, S., B. Thompson, R. Hoenicke, J. Leatherbarrow, K. Taberski, R. Smith, and D. Stevens, Jr. 2004. Re-design Process of the San Francisco Estuary Regional Monitoring Program for Trace Substances (RMP) Status & Trends Monitoring Component for Water and Sediment. SFEI Contribution 109, SFEI, Oakland, CA. 86 pp. http://www.sfei.org/rmp/rmp_docs_author.html.

Melwani, A.R., B. K. Greenfield, A. Jahn, J. J. Oram, M. Sedlak, and J. Davis (2007) Power Analysis and Optimization of the RMP Status and Trends Program. Draft report, SFEI.

Urquhart, N.S., Paulsen, S.G., and Larsen, D.P. 1998. Monitoring for policy-relevant regional trends over time. *Ecol. Appl*. 8: 246–257.

VanLeeuwen, D. M., L.W. Murray, and N. S. Urquhart. 1996. A mixed model with both fixed and random trend components across time. *Journal of Agricultural, Biological and Environmental Statistics* 1:435–453.