APPENDIX I

Answers for Regional Board

1. *Please describe in detail the statistical methods used to identify statistically significant differences between two watersheds for a particular constituent in a particular year, when monitoring data has been collected for two storm events. Please provide an example using existing data. Compare the statistical methods used and analysis results obtained when using monitoring data from two storm events versus three storm events.*

Analyses performed to date on the data collected for the San Diego County CoPermittee's Monitoring and Reporting Program have focused on comparing all watersheds in the county, and comparisons between just two watersheds have not been performed. In a particular year, Analysis of Variance (ANOVA) has been used with a Tukey multiple range test to compare data from mass loading stations (MLS) on a constituent by constituent basis. The ANOVA uses storm events as replication to determine differences between MLS means. The statistical methodology does not change with a reduction in the number of sampling events.

To compare results of this analysis performed with three storm events to the same analysis performed with two storm events, two constituents were selected that showed some differences between MLS during the 2003-04 monitoring year, but were not dominated by one particular site. For the comparison, two storms were selected from the three at each MLS by picking the first storm event sampled and the first storm event sampled after February 1, reflecting the present requirement that at least one storm event be collected after this date. Results of the ANOVAs with three storm events and two storm events are shown below. In each representation, the means for each station are shown in descending order and the colored lines below the means show those MLS that are not significantly different with the Tukey multiple range test.

Total Dissolved Solids (TDS)

	Prob>F	SDC	SR	SLR	PC	TC	EC	SDR	SM	AH	TJ	CC
3 events	<0.001	2050	1963	1927	1770	1640	1092	1050	638	604	539	162
Power=0.99												

		SDC	TC	SR	SLR	PC	SDR	EC	AH	SM	TJ	CC
2 events	0.003	2400	1925	1830	1685	1635	1330	983	901	638	571	178
Power=0.98												

Total Nickel

		TJ	AH	CC	TC	SDR	EC	SM	SR	PC	SDC	SLR
3 events	0.044	0.048	0.013	0.011	0.008	0.003	0.003	0.003	0.002	0.002	0.002	0.002
Power=0.82												

		TJ	AH	CC	TC	SDR	EC	SM	SR	PC	SDC	SLR
2 events	0.057	0.065	0.017	0.015	0.010	0.004	0.004	0.003	0.003	0.002	0.002	0.002
Power=0.72												

The analyses on TDS were both statistically significant and the Tukey multiple range test showed similar differences between MLS. SDC is significantly higher in TDS than SM, TJ, and CC whether three or two storm events are used. There are some changes in order of MLS due to changes in mean values, changes range from 6% (TJ) to 39% (AH); but MLS at the extremes with three events remain at the extremes with two events. The total nickel ANOVA with three events was statistically significant ($p < 0.05$) with three events and showed differences between

TJ and PC, SDC, and SLR. When the number of storm events was reduced to two per MLS, the test was not statistically significant ($p > 0.05$) and differences between TJ and the three MLS stations with the lowest nickel concentrations could not be determined.

Forty-seven ANOVAs were run on the 2003-04 data. Four of the ANOVAs had results with multiple groups of stations (similar to that above for TDS); three of these resulted in less spatial differentiation with two storm events than with three. Seven ANOVAs showed similar patterns of two groups of stations (like those for total nickel); with only two storm events in the ANOVA three tests had no change and four resulted in no differences between stations. Eleven of the ANOVA/Tukey tests were dominated by high concentrations at the Tijuana River station; nine of the eleven test results did not change with only two storm events, the other two grouped Tijuana River with two or more other stations. The remaining 25 ANOVAs were non-significant results.

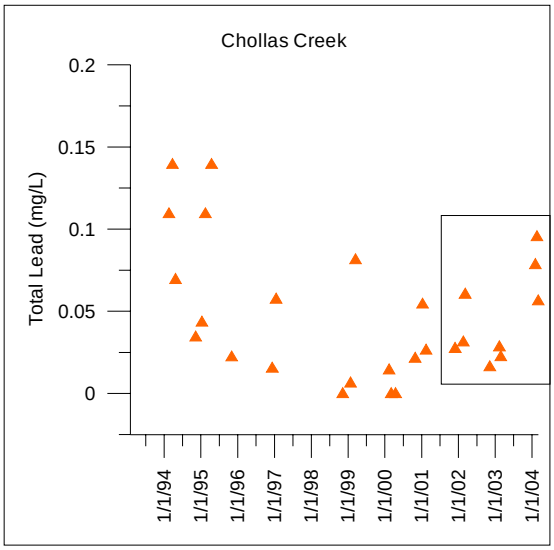
To answer the specific question asked above, identification of differences between two watersheds for a particular constituent in a particular year; a t-test was performed for these same two constituents to compare two randomly selected MLS based on three and two storm events. San Luis Rey (SLR) and Chollas (CC) were selected for comparison. The results of the t-tests and the corresponding power are shown. The t-test included a test for equality of variance; when variances were determined to be unequal, Satterthwaite's approximation of the degrees of freedom was applied (SAS 1989).

Measure (mg/L)	Events	SLR Mean	CC Mean	SLR S.D.	CC S.D.	Prob > T	Power (alpha=0.05)
TDS	3	1927	162	446	52.3	0.002	1.00
	2	1685	178	219	62.2	0.011	1.00
Total Nickel	3	0.0017	0.0107	0.0006	0.0076	0.175	0.53
	2	0.0020	0.0150	0.0001	0.0014	0.049	1.00

The results do not change for TDS whether three or two events are used in the analysis. The means are so different and the standard deviations for each are relatively small in comparison, yielding a significant t-test and strong power. For total nickel, the t-test is not significant when three events are considered and is significant with only two events. In this particular case, both data points that were excluded from the two event analysis were non-detectable results, the four remaining data points were considered detectable at or above the detection limit. Because non-detectable values are tested at one-half of the detection limit, in this case additional variability was introduced by the lower values.

2. *Please describe in detail the statistical methods used to identify statistically significant differences between two years for a particular constituent in a particular watershed, when monitoring data has been collected for two storm events each year. Please provide an example using existing data. Compare the statistical methods used and analysis results obtained when using monitoring data from two storm events annually versus three storm events annually.*

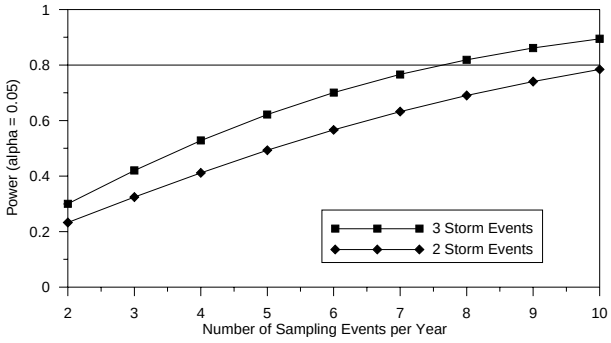
The present monitoring program has not been designed to answer this question, whether two or three annual storm events are considered. If one were to perform this analysis, the results would be more dependent on the magnitude of differences between years and the variability of the data within each of the years than the number of storm events sampled.



This is illustrated here with two examples taken from three consecutive years of data for total lead. The years used for this example are the last three years in the graph at right, delineated by the box. These years were chosen as they provide examples with both small and large variability within a year in addition to having means that are potentially distinguishable from each other. A t-test was performed on each set of two years of data, (i.e., 2001-02 vs. 2002-03 and 2002-03 vs. 2003-04) first using all three sampling events, then selecting only the first event of the season and the first event after February 1 in each year. The results of the t-tests and corresponding power for each test are summarized below. As could be expected from viewing the graph at right, 2001-02 was not significantly different from 2002-03 whereas 2002-03 was determined to be different from 2003-04 whether three storm events or two storm events were considered.

Events	2001-02 Mean	2002-03 Mean	2001-02 S.D.	2002-03 S.D.	Prob > T	Power (alpha=0.05)
3	0.040	0.023	0.018	0.006	0.189	0.42
2	0.030	0.023	0.003	0.009	0.384	0.23
Events	**2002-03 Mean**	**2003-04 Mean**	**2002-03 S.D.**	**2003-04 S.D.**	**Prob > T**	**Power**
3	0.023	0.077	0.006	0.020	0.010	0.99
2	0.023	0.088	0.009	0.012	0.025	0.99

To have a t-test with 0.80 power for the comparison of 2001-02 vs. 2002-03, eight storm events would be needed assuming the means and standard deviations for the three event test and ten storm events would be needed using the data in the two sample comparison. The power curves for each of these scenarios are shown at right.



This example demonstrates that the magnitude of the variation between storm events is more influential in determining differences between years than the number of storm events monitored when choosing between three events and two events.

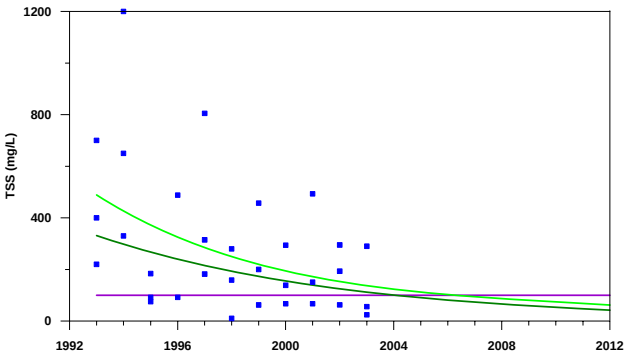
To further demonstrate the potential impacts of observing changes between years when reducing the number of events sampled, ten station/constituents were selected at random to test the same way as done for total lead above. The results are shown in the table below. No t-tests changed from non-significant to significant (and vice-versa) with a reduction of storm events.

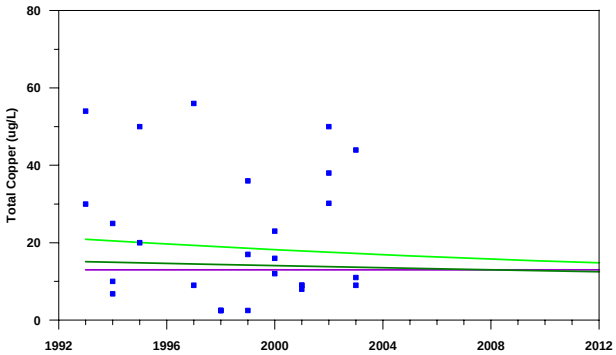
Location	Measure	Events	2001-02 Mean	2002-03 Mean	2001-02 S.D.	2002-03 S.D.	Prob > T	Power (alpha=0.05)
AH	Diazinon	3	0.196	0.326	0.076	0.135	0.218	0.66
AH	Diazinon	2	0.155	0.329	0.038	0.191	0.332	0.56
CC	Nitrite as N	3	0.170	0.094	0.056	0.025	0.097	0.92
CC	Nitrite as N	2	0.165	0.106	0.078	0.020	0.408	0.43
EC	Dissolved Cadmium	3	0.001	0.000	0.000	0.000	0.199	0.96
EC	Dissolved Cadmium	2	0.001	0.000	0.000	0.000	0.500	0.41
EC	Total Arsenic	3	0.001	0.003	0.001	0.001	0.022	0.99
EC	Total Arsenic	2	0.001	0.003	0.000	0.000	0.027	1.00
SLR	Disssolved Zinc	3	0.010	0.033	0.000	0.033	0.341	0.54
SLR	Disssolved Zinc	2	0.010	0.038	0.000	0.045	0.537	0.35
SLR	pH	3	7.533	7.130	0.058	0.656	0.399	0.44
SLR	pH	2	7.500	6.860	0.000	0.651	0.397	0.63
SLR	Total Coliforms	3	2.651	2.728	0.642	0.435	0.871	0.08
SLR	Total Coliforms	2	2.738	2.477	0.882	0.000	0.748	0.15
SR	Total Zinc	3	0.022	0.032	0.020	0.009	0.463	0.30
SR	Total Zinc	2	0.028	0.036	0.025	0.009	0.710	0.15
TJ	Oil and Grease	3	2.333	4.573	1.528	3.707	0.388	0.39
TJ	Oil and Grease	2	3.000	2.580	1.414	1.909	0.826	0.10
TJ	Total Arsenic	3	0.007	0.010	0.001	0.007	0.506	0.30
TJ	Total Arsenic	2	0.008	0.006	0.001	0.002	0.572	0.26
Location	**Measure**	**Events**	**2002-03 Mean**	**2003-04 Mean**	**2002-03 S.D.**	**2003-04 S.D.**	**Prob > T**	**Power (alpha=0.05)**
AH	Diazinon	3	0.326	0.094	0.135	0.062	0.054	0.99
AH	Diazinon	2	0.329	0.116	0.191	0.069	0.276	0.67
CC	Nitrite as N	3	0.094	0.055	0.025	0.028	0.146	0.82
CC	Nitrite as N	2	0.106	0.070	0.020	0.014	0.172	0.91
EC	Dissolved Cadmium	3	0.000	0.001	0.000	0.000	0.199	0.97
EC	Dissolved Cadmium	2	0.000	0.001	0.000	0.000	0.500	0.41
EC	Total Arsenic	3	0.003	0.003	0.001	0.002	0.565	0.22
EC	Total Arsenic	2	0.003	0.004	0.000	0.001	0.360	0.52
SLR	Disssolved Zinc	3	0.033	0.010	0.033	0.000	0.341	0.54
SLR	Disssolved Zinc	2	0.038	0.010	0.045	0.000	0.537	0.35
SLR	pH	3	7.130	7.860	0.656	0.386	0.172	0.76
SLR	pH	2	6.860	7.690	0.651	0.354	0.254	0.72
SLR	Total Coliforms	3	2.728	2.974	0.435	0.556	0.579	0.21
SLR	Total Coliforms	2	2.477	3.281	0.000	0.236	0.130	1.00
SR	Total Zinc	3	0.032	0.025	0.009	0.013	0.494	0.28
SR	Total Zinc	2	0.036	0.023	0.009	0.018	0.480	0.34
TJ	Oil and Grease	3	4.573	5.973	3.707	3.384	0.654	0.17
TJ	Oil and Grease	2	2.580	7.770	1.909	1.881	0.112	0.99
TJ	Total Arsenic	3	0.010	0.025	0.007	0.026	0.400	0.38
TJ	Total Arsenic	2	0.006	0.033	0.002	0.031	0.350	0.53

3. *Please describe in detail the statistical methods used to identify statistically significant differences between multiple years for a particular constituent in a particular watershed, where monitoring data has been collected for two storm events in some of the years and three storm events in other years. Please provide an example using existing data. Compare the statistical methods used and analysis results obtained when using monitoring data from multiple years where data from two and three storm events was collected versus data from multiple years with data from three storm events only.*

This question basically describes the manner in which temporal trends are investigated at a MLS. Trend analysis (regression) has been used to determine whether a significant trend exists in the data. ANOVA has not been used to assess whether differences exist between multiple years. While an ANOVA with differing numbers of storm events as an unbalanced design could be performed to assess differences between multiple years, the trend analysis gives the same information plus information about the pattern of the change. The analysis uses $\log_{10}$ transformed data regressed with year to determine whether the slope of the regression line is significantly different from a horizontal line drawn through the data. Log transformation is commonly used to normalize data when concentrations are of different orders of magnitude. Measured constituents can then be classified as to whether they exhibit a significant decreasing (more recent values are lower than earlier values), increasing (more recent values higher), or no trend. In particular, this approach has been used to estimate how many years of sampling would need to be performed to be confident that a mean concentration for a constituent is under the pertinent water quality objective. Two examples of this analysis are shown below; the first shows a currently significant decreasing trend for total suspended solids (TSS), the second shows a currently non-significant but visually decreasing trend for total copper.

The approach in this analysis is to determine the equation of the regression line of the existing data and from the equation estimate the predicted mean values in future years based on the line. This has been used further to look at the differences of sampling two or three events every year or in alternate years. This analysis and results are presented later in this paper.





4. *Please describe in detail the various statistical methods that can be used to analyze data across regions, when one region is monitoring three wet weather events annually and the other is monitoring two wet weather events annually. How would data from a region monitoring two events annually be used in combination with data from regions monitoring three events annually in order to identify overall trends in constituent concentrations? How would data from a region monitoring two events annually be compared o data from regions monitoring three events annually in order to identify statistically significant differences? Please provide examples of how these and other pertinent types of statistical analysis would be conducted.*

Any analysis performed to compare sets of data with unequal replication is termed an "unbalanced design". Although most sampling plans originate with plans for the same replication in all sets of data, in reality no dataset is perfectly balanced. This imbalance can be the result of many factors such as: sample breakage, improper sample handling, questionable laboratory results, as well as redesign of the sampling plan. Statistical procedures in established statistical software (SAS, SPSS, Systat, and others) have been developed to be robust and able to handle unbalanced designs in the correct manner. For example in SAS (used by Weston for this program), several procedures can perform an ANOVA. PROC ANOVA is designed for balanced designs and the description of the program emphasizes this fact and refers the user to PROC GLM (General Linear Models) for unbalanced designs. T-tests are a special case of ANOVA where there are only two categories to compare; these tests are also robust and designed to handle unequal numbers of samples.

Therefore, an analysis to compare regions with different sampling frequencies is an unbalanced statistical analysis. To illustrate examples of a comparison between regions, a subset of data from another region in California (Region A)[1] and a subset of data from the San Diego County database (Region B) are used here in a t-test. Three constituents were selected (before examining the data for each) representing biological, organic, and inorganic indicators. The t-test was run twice for each constituent, first comparing the data with balanced replication (3 storm events each) and then comparing the three storm events in Region A to two storm events in Region B.

Constituent	Events	Region A Mean (4 MLS)	Region B Mean (6 MLS)	Region A S.D.	Region B S.D.	Prob > T	Power (alpha=0.05)
Fecal coliform (log)	3	3.981	3.204	0.929	0.781	0.020	0.70
	2	3.981	3.217	0.929	0.762	0.038	0.60
Nitrate-N	3	2.008	1.006	2.234	1.040	0.168	0.31
	2	2.008	1.122	2.234	1.166	0.240	0.23
Total Copper	3	27.98	14.00	25.20	14.82	0.102	0.41
	2	27.98	15.58	25.20	17.22	0.209	0.26

As can be seen in the table, the overall result of the statistical test does not change when the number of storm events changes but the power of the test decreases by about 0.10. None of the power estimates are strong; power is dependent on number of replicates, the means being compared, and the standard deviation, but the biggest factor in the power is the size of the standard deviation compared to the means. This is also true of a statistical comparison of means. In the examples above, the constituent with standard deviations smaller than the means

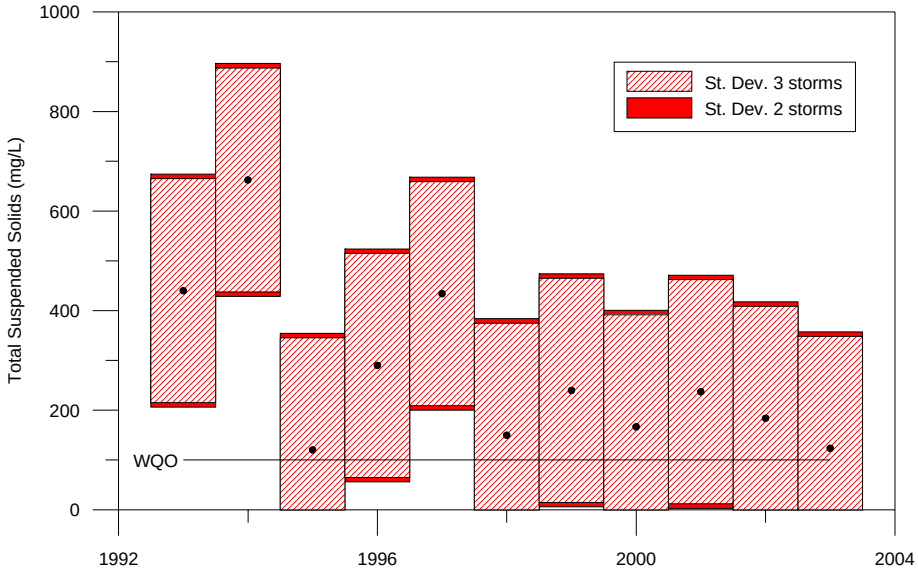
[1] ***Disclaimer***
This information is used solely for an analysis of replication and power. It should not and cannot be used to imply similarities or differences between regions. To ensure this, we have randomly selected a subset of watersheds within each region and even though the data used are actual data from the public record, the region used shall remain anonymous.

shows significant differences between regions regardless of the number of storm events. On the other hand, the constituents with standard deviations larger than the means are not significantly different between regions even though, in both cases, Region A has a mean concentration about twice that of Region B.

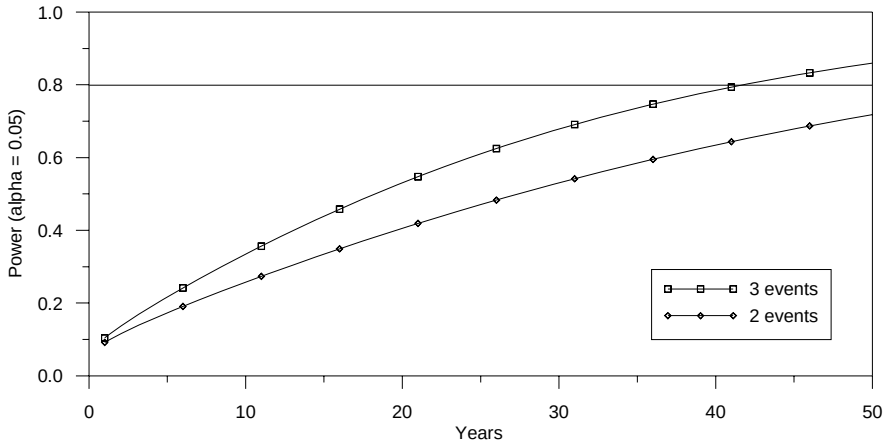
5. *Please provide all of the power analysis curves which were used to come to the conclusions reached in your October 1, 2004 letter. Identify what level of confidence was used and what percent change in constituent concentration was addressed during the analysis. Explain the rationale for using these values.*

The graphs and data shown in the October 1, 2004 letter were based on using the variability in the data from an ANOVA on all of the years sampled to that date. Subsequent to that letter, the question of the impact of sampling two storm events rather than three events has been re-evaluated and a more refined and appropriate statistical analysis has been applied to answer the question. This approach is discussed after the answer to this specific question.

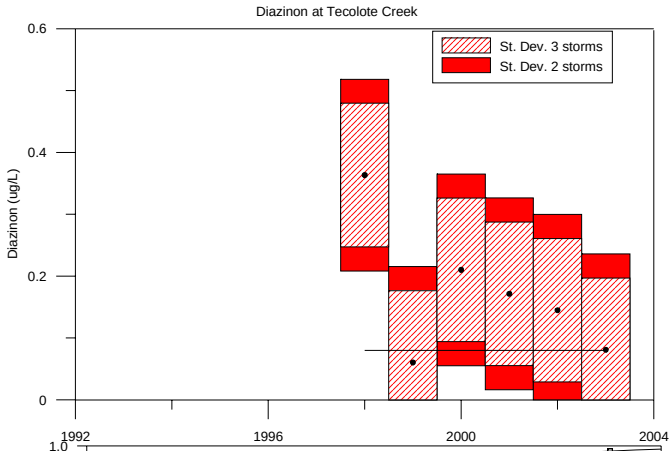
One example of the approach used in the October 1, 2004 letter was for total suspended solids. The bar chart shows the means and common standard deviation (based on ANOVA) for three and two storm events. In the original analysis, the number of sampling years necessary to have 80% power was based on determining that the mean (in any year) was at 50% of the water quality objective. This concentration was chosen as one that could be confidently considered to be under the water quality objective. The common standard deviation, as shown in the bar chart, was used and alpha was set at 0.05.

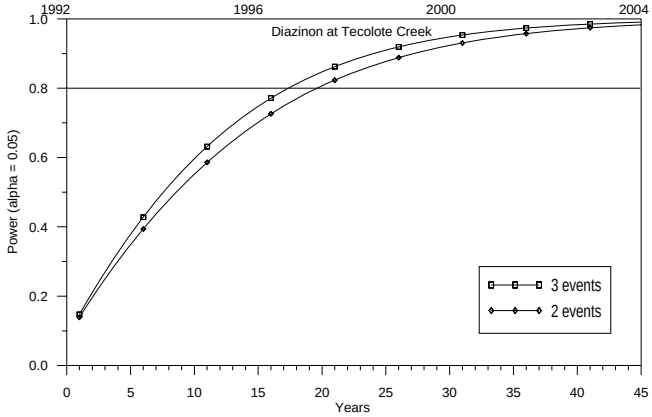


This analysis produced the power curves shown at right. From these curves, it can be seen that many years of sampling would be necessary to attain 80% power with either three storm events or two storm events.

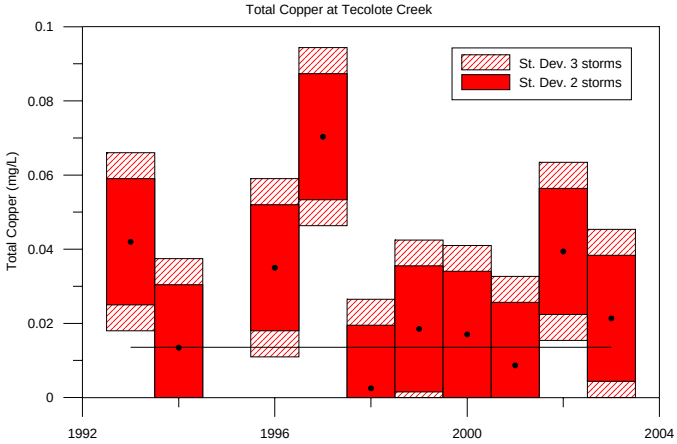


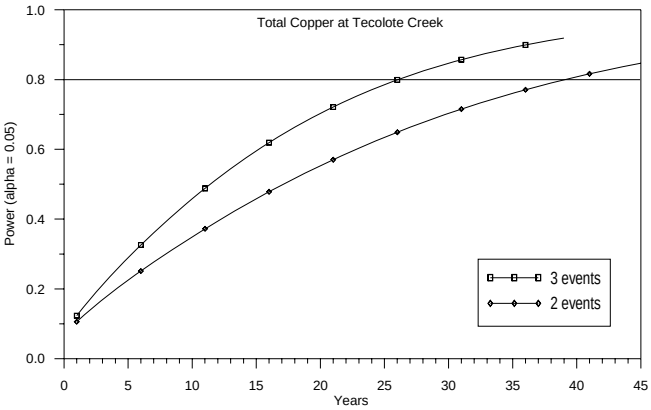
The following constituents were looked at in the same manner. Diazinon at Tecolote Creek was evaluated to determine how many years of sampling. The comparison was made between the WQO of 0.08 mg/L and a concentration of one-half the WQO (0.04 mg/L). Observed mean concentrations with the common standard deviations for three and two storm events per year are shown at the right along with the power curves for the evaluation. The power curves for this constituent/location are virtually identical.



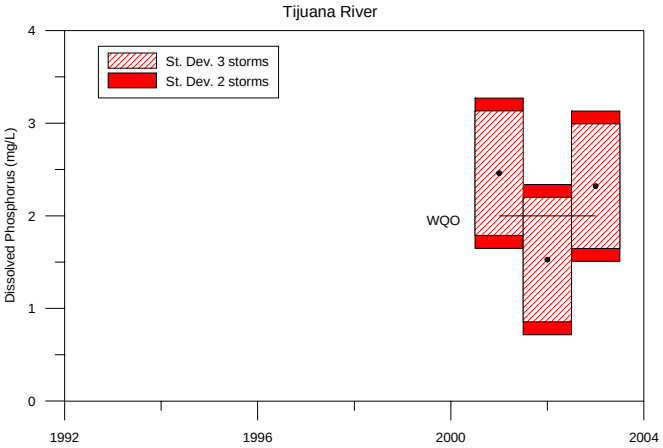


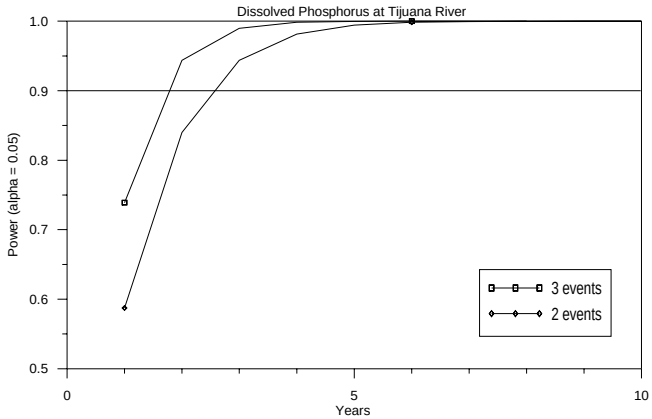
Total copper at Tecolote Creek is a little different than the other measures looked at for this evalution in that the standard deviation calculated for two storms was smaller than that for all three storms in the dataset for each year. Means and standard deviations are shown at right. The power curves represent the ability to differentiate the WQO of 0.0135 mg/L (based on a default hardness of 100 mg/L $CaCo_3$) from a concentration of one-half the WQO.



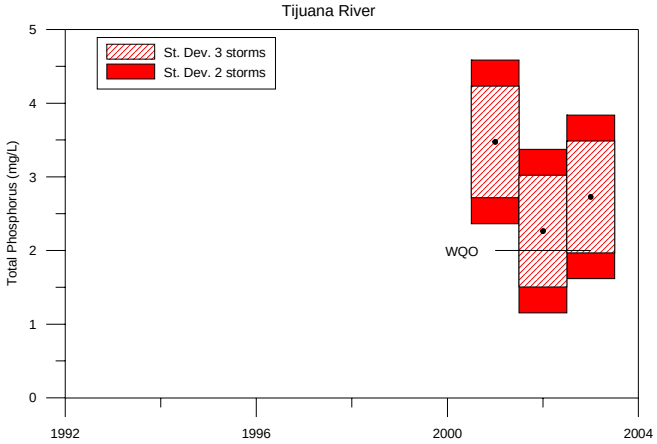


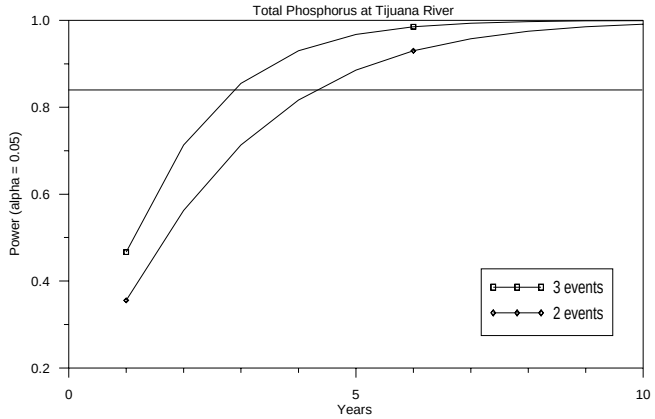
Dissolved and total phosphorus at Tijuana River were used to demonstrate the same evaluation with a shorter timeline of existing data. These data are also less variable than those previously shown. For dissolved phosphorus, mean concentrations are currently bouncing around the WQO of 2 mg/L. With the variability of the existing data, the power curves show that only two years of data are needed to achieve 80% power, regardless of the number of events sampled per year.





Likewise for total phosphorus, variability is low and 80% power to confirm that a mean concentration of 1 mg/L is different from the WQO of 2 mg/L would take about 8-9 sampling events, or 3 years with three events and 4 years with 2 events per year.



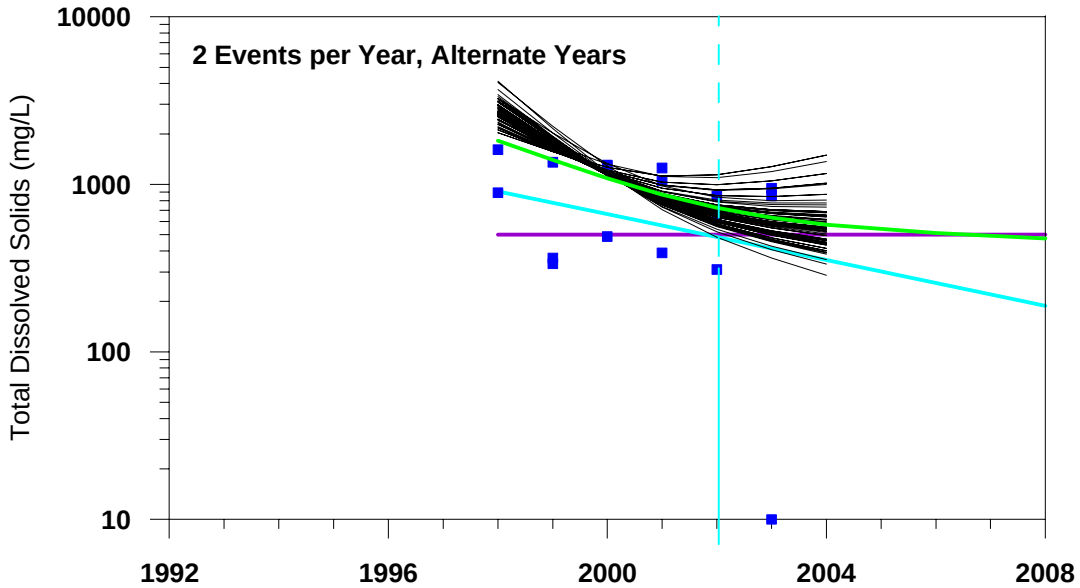


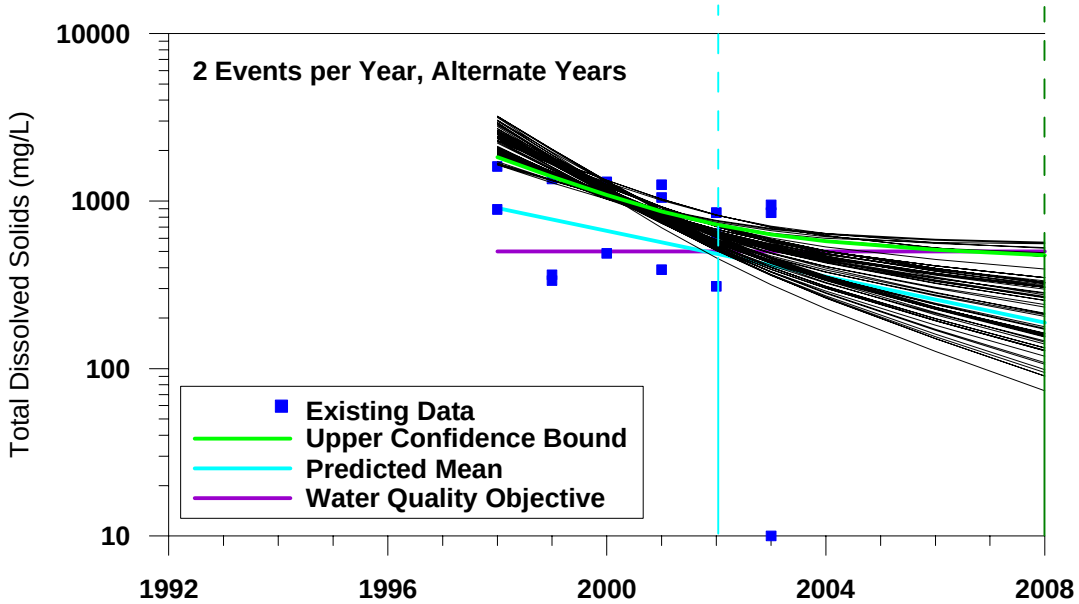
Revised Approach to Determining Impacts of Reducing Wet Weather Sampling Effort

A major consideration of any redesign of a monitoring program is an evaluation of the impacts the changes may have on the analysis and interpretation of the results. The recommendation to change the quantity or frequency of sampling therefore necessitated a thorough look at the impacts on (1) following established trends in the data, (2) detecting the development of changes or trends in the data, and (3) determining the cessation of a trend.

Enough data exist for three of the MLS to determine whether trends exist at the present time. Temporal plots of data for 31 constituents at Agua Hedionda, Tecolote, and Chollas Creeks were evaluated to determine the existence, at the present time, of a decreasing or increasing trend. Of the 93 plots evaluated, 14% exhibited decreasing trends, 22% showed increasing trends, and the remaining 63% had no apparent trend.

The analysis uses $\log_{10}$ transformed data regressed with year to determine the equation of the regression line drawn through the data. Log transformation is commonly used to normalize data when concentrations are of different orders of magnitude. The regression equation is used to compute the predicted mean value in future years and the standard deviation from the regression analysis, data are generated for each future year that have a mean equal to the predicted mean and are randomly distributed within the bounds of the standard deviation. Using these simulated data with the existing data, the regression is rerun and the point in time when the upper 95% confidence bound crosses below the water quality objective is determined. Because this is just one random simulation that may be anomalous, this process is repeated with 100 sets of randomly simulated data based on the original equation. The entire set of 100 regressions is then evaluated to determine when the upper confidence bound is below the water quality objective 95 out of 100 times. This is the number of years of sampling that must occur to be confident that the concentration meets the objective. The process starts with the addition of one more year of sampling and evaluates whether the confidence bound meets this criteria, if not, sequential years are added until the upper bound is below the water quality objective for 95 of the 100 datasets. Examples of the first year and a final year evaluation are shown below.





This whole process is performed for sampling a) three storm events every year into the future, b) three sampling events in alternate years, c) two storm events every year, and d) two storm events in alternate years. Examples of this analysis for constituents with decreasing trends, increasing trends, and no apparent trends are discussed in this document.

A. Decreasing Trends

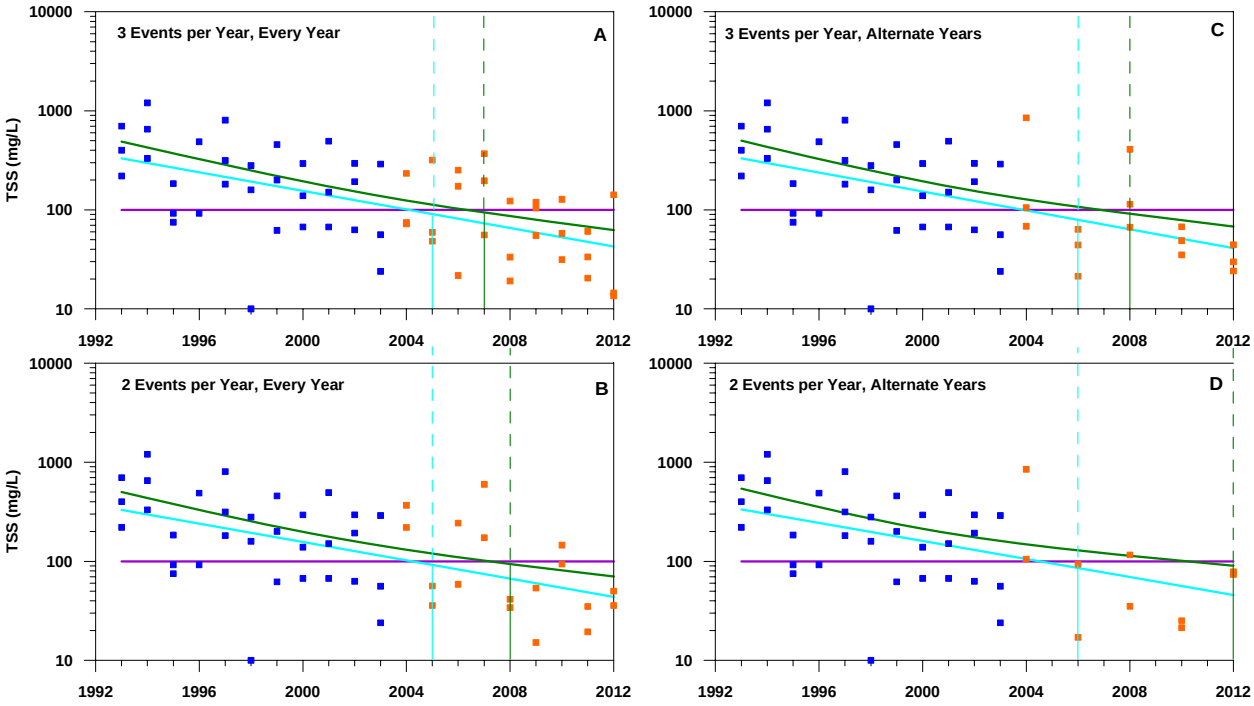
Four constituents were selected because they have established water quality objectives (WQO) that are below the current concentrations and provide cases with differing numbers of existing data points as well as a variety of slopes with decreasing concentrations through time.

Example constituents include:

1. TSS data from Chollas Creek with 11 years of existing data with a statistically significant trend at the present time.

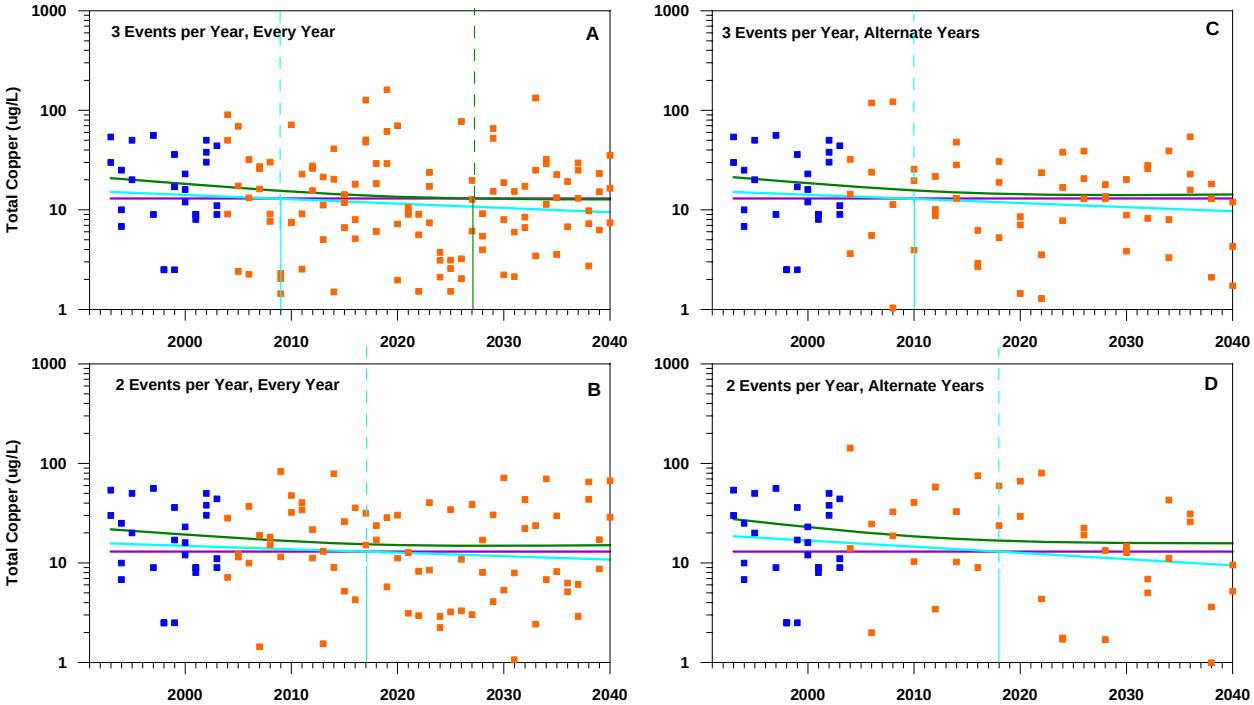
2. Total copper from Tecolote Creek with 10 years of existing data and a non-significant trend.
3. Diazinon from Tecolote Creek with 6 years of existing data and a non-significant trend.
4. TDS from Agua Hedionda Creek with 6 years of existing data and a non-significant trend.

The plots shown below for each example constituent and frequency of sampling represent one of the randomly generated datasets for which the upper confidence bound crosses the water quality objective at the point where 95 of the 100 regression lines would be below this line. Vertical lines on the plots indicate the years in which the mean and upper bound cross below the water quality objective (horizontal line).



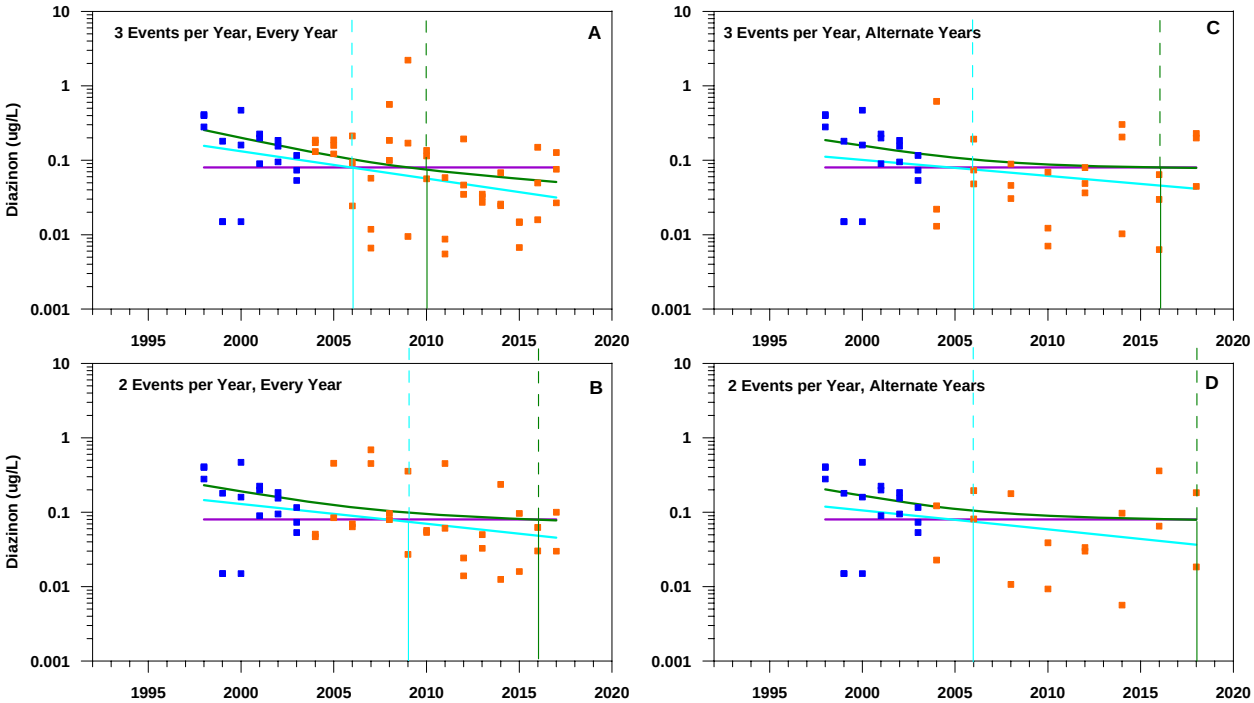
TSS at Chollas Creek

The water quality objective for TSS is 100 mg/L (shown by horizontal purple line). As can be seen above in Plot A, if sampling during wet weather were to continue at three events per year with sampling every year, the mean value for TSS would be below the water quality objective in 2005 and the upper confidence bound would drop below in 2007. Plot B shows the change if sampling was to occur only in alternate years: the mean would go below the WQO in 2005 also, but the upper confidence bound would not cross until 2008, one year later than in plot a. If sampling were reduced to two storm events per year (Plot C), the mean and upper confidence bound would cross the WQO in 2006 and 2008, respectively, a year later than in plot a. Sampling two events per year in alternate years (Plot D), would result in the upper confidence bound not crossing below the WQO until 2012.



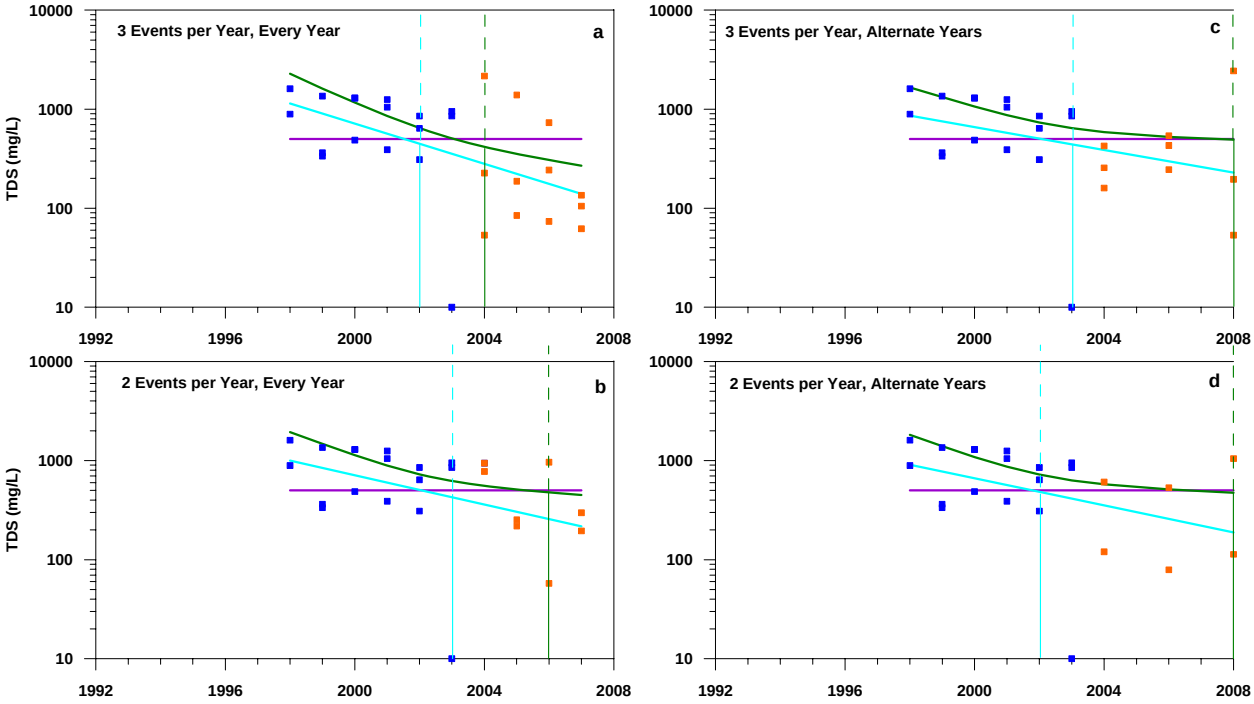
Total Copper at Tecolote Creek

The WQO for total copper varies dependent on the hardness of the water, for purposes of this analysis the objective is shown at 13 ug/L (100 mg/L $CaCO_3$). Concentrations of total copper vary widely in Tecolote Creek, with existing data ranging between 2 and 50 ug/L even within one storm season. This variability along with a slowly decreasing trend extends the time period until a decrease can be seen below the WQO until about 2027 with the current sampling frequency.



Diazinon at Tecolote Creek

Diazinon in Tecolote Creek presently shows a significant decreasing trend that, if continued, will result in mean concentrations below the WQO of 0.08 ug/L in the next few years. 95% confidence limits will vary with the four sampling scenarios, following four to ten years after the mean value goes below the WQO. The difference in the number of years is dependent on the overall number of observations for this measure (existing data plus simulated data).

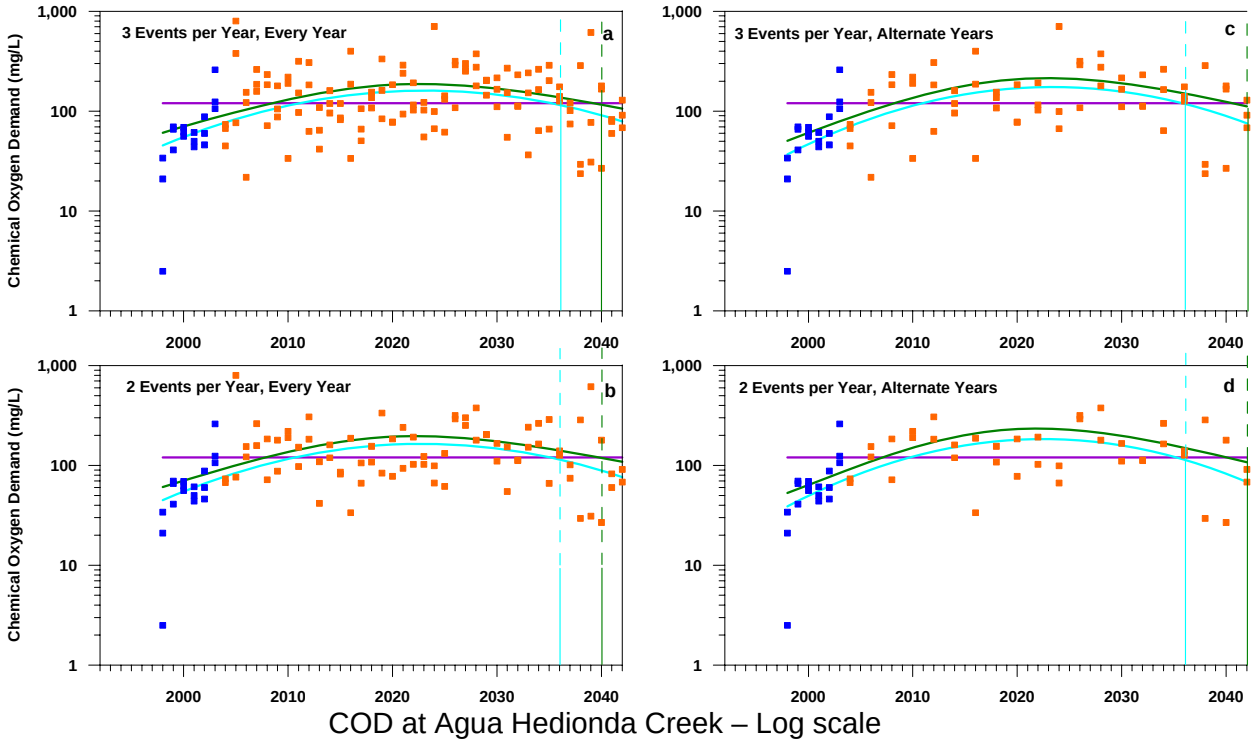


TDS at Agua Hedionda Creek

Similarly, TDS at Agua Hedionda Creek shows a decreasing trend that will soon pass through the WQO of 500 mg/L. The longer time span until the 95% confidence bound also falls below the WQO for the alternate year scenarios is based on the number of samples, both existing and simulated).

B. Increasing Trends

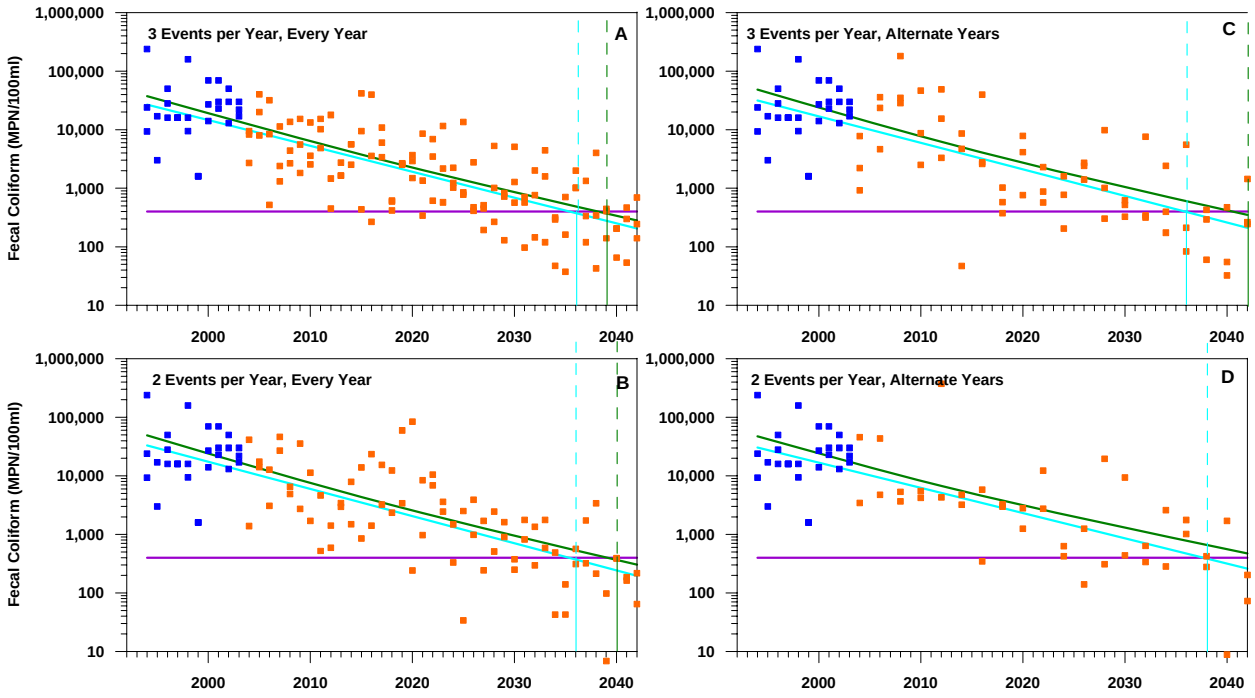
Constituents with apparent increasing trends are assumed to level off at some point in time. Based on this premise, the regression equation is used to compute the predicted mean value in future years and the standard deviation from the regression analysis. Using these simulated data with the existing data, the regression is rerun and the point in time when the upper 95% confidence bound is determined to stay below the WQO. This approach is illustrated with chemical oxygen demand (COD) data from Agua Hedionda Creek. The simulated data were allowed to continue to increase until they reached the WQO of 120 and then the mean was forced to remain at 120 for the duration of the simulation. The regression line based on the whole data set curves above the WQO and then goes below the line again in the future.



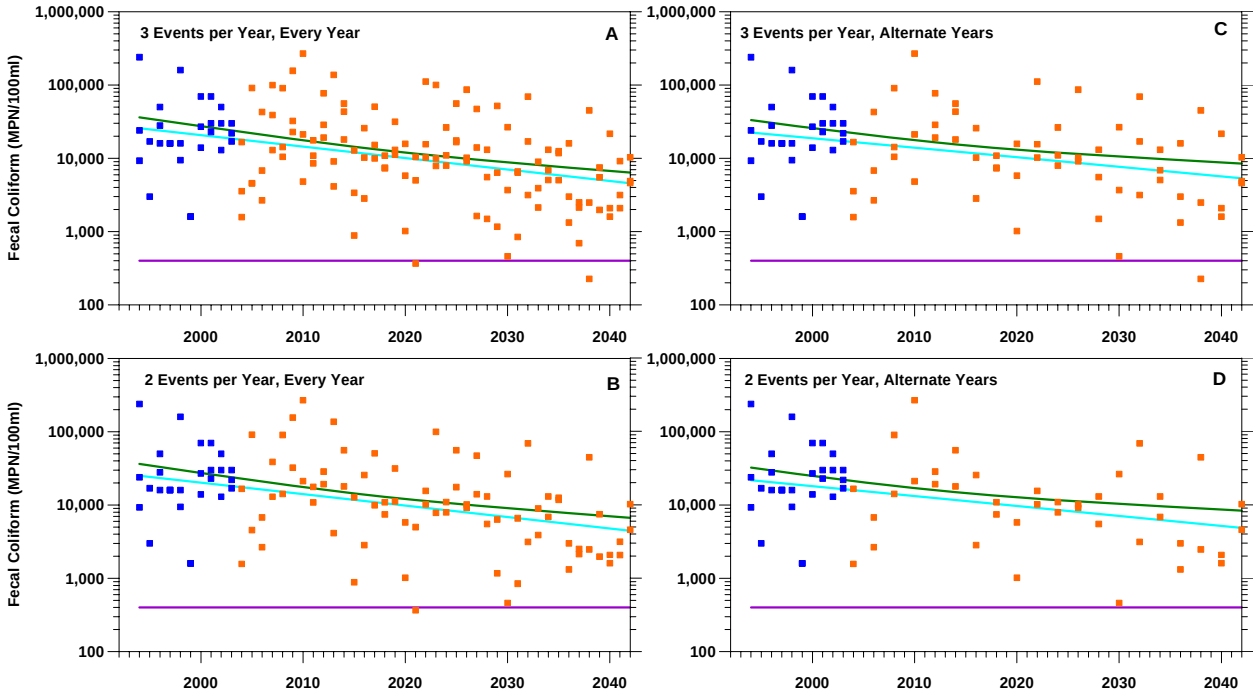
COD at Agua Hedionda Creek – Log scale

C. No Apparent Trend

Constituents without an apparent trend at the present time present a problem in determining the impact of reducing the number or frequency of sampling. The approach taken for this set of data is to create two scenarios where the data will change due to some intervention (e.g., BMPs) to lower concentrations. For this example, fecal coliforms in Chollas Creek were selected as an example dataset. The first scenario assumes that concentrations decrease by 50% in the simulated data; the second assumes concentrations decrease by 5% per year. The data are then treated in the same manner as for the constituents with decreasing trends and the year in which the upper confidence bound falls below the WQO is estimated for each of the potential sampling alternatives. With the first alternative, the upper confidence limit drops below the WQO within the time frame of the simulation; in the second alternative scenario the upper confidence never reaches the WQO.



Fecal Coliforms at Chollas Creek – 50% decrease every five years



Fecal Coliforms at Chollas Creek – 5% decrease every year