Methods

## Website descriptions

Four primary datasets were evaluated to answer key management questions for review of the Orange County Monitoring Program. For each dataset, an R Shiny website (Chang et al. [2020](#ref-Chang20); RDCT (R Development Core Team) [2020](#ref-RDCT20)) was created to provide online access to analyses that addressed key management questions.

* Mass emissions, concentration data: <https://sccwrp.shinyapps.io/ocms4review/mass_emissions.Rmd>
* Mass emissions, load data: <https://sccwrp.shinyapps.io/ocms4review/mass_emissions_loads.Rmd>
* Dry weather or illicit discharge/illicit contaminant data: <https://sccwrp.shinyapps.io/ocms4review/dry_weather.Rmd>
* Tissue contaminant monitoring: <https://sccwrp.shinyapps.io/ocms4review/tissue.Rmd>

The structure of each website was similar, with some variation depending on specific questions relevant to each dataset. In general, each website included the following tabs to access and view results for information about each dataset:

1. Inventory: Map-based and tabular summaries of monitoring effort and basic characteristics of the data
2. Trends and power analyses: Changes over time by select constituents and locations, including power analyses to help identify optimal sampling effort
3. Station differences: A comparison of time series between stations to identify similarities among trends
4. Overall trends: Map-based and tabular summary of trend tests for all stations shown together

The website for the dry weather monitoring dataset also included the following tabs:

1. Hotspots: Assessment of hotspot sites based on threshold exceedances over time
2. Analyses by waterbody: A simple analysis of threshold exceedances for select sites shown for complete time series

Each main tab on the website (for the numbered items above) included additional sub-tabs or drop-down menus for selecting and viewing different content relevant to each dataset. For datasets where parameters lists were large (i.e., including all parameters in the website was impractical), select constituents were often included that were the most often observed or otherwise important for regulatory or other management needs. Decisions on specific constituents to include were vetted by the advisory committee.

## Trend Analyses

All websites, excluding that for the dry weather monitoring program, included an assessment of trends over time to understand 1) long-term changes at specific locations for parameters of interest, 2) whether any changes were above or below thresholds of management or regulatory concern, and 3) if management actions to reduce pollutant concentrations or loads have been successful. Trends were evaluated visually using plots of concentration or loading of observed data over time, including an evaluation of monthly and annual trends using boxplots (McGill, Tukey, and Larsen [1978](#ref-McGill78)). Formal hypothesis tests for trends also included linear regression analyses and non-parametric Kendall tests of annual averages. For both linear regression and Kendall tests, trends were evaluated based on deviations of the annual average of a parameter from the grand mean at an individual monitoring station. The regression analysis reported an estimate of slope (change in concentration or load per year) and overall significance of the regressions. Kendall tests provided an alternative indication of trend by evaluating magnitude, direction, and significance of a change over time using a non-parametric approach. For the latter, the kendallTrendTest function from the EnvStats package for R was used (Millard [2013](#ref-Millard13)).

## Power analysis

A critical question addressed during the evaluation of the Orange County Monitoring Program was how well the current sampling design was able to detect trends of interest. In particular, questions were evaluated regarding the ability to detect a specified magnitude of change (e.g., 30% decrease over ten years) and the likelihood of observing an exceedance of a concentration (e.g., is the true mean above a regulatory threshold) for a given parameter over a period of time. For both questions, power analyses were conducted for specific parameters and locations where sufficient data were available. In essence, power describes the probability of observing a true event in a population, based on a sample of the population and if the true event actually occurred. This is analogous to observing an actual change in water quality conditions for a given sample design with the knowledge that sampling is discontinuous over time and at varying time intervals depending on location.

The first power analysis estimated the ability to detect a specific trend for a desired sampling frequency. For a chosen parameter and location, the observed time series was first detrended by taking the residuals of a regression of concentration or load vs time. From the residuals, the variance of the dataset around the mean was estimated and used to simulate new time series from which power was evaluated. For example, if a 50% change (increase or decrease) was considered the true change, a simulated time series was created by first estimating the linear change over the length of time that the true time series was observed (e.g., ten years) and then imposing uncertainty in the linear estimate by adding variance from the residuals to the linear trend. The observed level of sample effort was considered 100% of the current effort if the number of observations in the simulated time series was the same as the observed. Evaluating power at different levels of effort required subsampling of each simulated time series for the selected level of effort. For a large number of simulated time series (n = 1000), power was estimated as the percentage of simulations where the change was significant based on linear regression. This was repeated for varying sample effort from 10% to 200% of the current for a given time series.

A second power analysis was conducted to quantify the likelihood of observing an exceedance of a concentration or load of a parameter at a given sample density. Similar methods as the first analysis were used such that power was estimated by repeated sampling of a simulated time series using an estimate of variance for each parameter. Rather than estimating significance of a trend at a simulated magnitude of change, power was defined as the percentage of simulations where a simple t-test identified a significant difference of the simulated values above a threshold. The evaluated thresholds for each parameter were chosen across the range of observations of a parameter from the mean value to the 95th percentile. The simulated values were created as before, except that zero change in the linear difference from the beginning of each time series was assumed (i.e., time series had 0% magnitude change across the period of record). As for the first power analysis, power was evaluated for varying levels of sample effort from 10% to 200% of the observed. Importantly, the interpretation of the power estimates were slightly different than the first analysis. Rather than showing the percentage of time for which a trend would be detected, the power values shows the ability to determine that the true mean of a time series is equal to the threshold as an indication of confidence in exceeding a value of interest for a given sample design.

## Optimal sample effort

An optimal level of sample effort was derived from the power analyses to describe the balance between over- and under-sampling. From a programmatic perspective, the optimal level of effort minimizes sampling cost by identifying the level of effort where any additional samples do not substantially increase the ability to detect a trend, whereas reductions in sample effort cause a disproportionate increase in the magnitude of the trend to be detected for a desired level of power. Graphically, the optimal level of effort is the inflection point on a power curve where the y-axis shows the magnitude of the trend to detect and the x-axis is the level of sampling effort. This inflection point was determined quantitatively for each water quality parameter as the point in a monotonic power curve where the slope of y vs x (i.e., trend to detect vs sample effort) exceeded that of x vs y (i.e., sample effort vs trend to detect). Given that sample effort and variance of each time series differed considerably among the observed time series, optimal effort was identified for power curves only where sufficient data were available. A power estimate of 80% was considered a sufficient target for optimal effort.

Optimal effort was estimated for each parameter at each monitoring station and based on aggregates of optimal effort across parameters and stations. For aggregate estimates, optimal effort was identified for one parameter across all stations as the median optimal sample effort across stations. Similarly, optimal effort for multiple parameters and multiple stations was simply the median across all estimates. In cases, where the aggregate optimal effort varied considerably from the median, boxplot summaries were used to characterize the spread. Separate boxplots were also retained for parameters across stations so that variation across parameters could be determined, i.e., optimal effort for one parameter may be larger than another and an a priori preference for one parameter could determine future sampling design. A final management question relative to optimal effort addressed the need to focus attention at locations where observed values were close to important regulatory thresholds. For the plots of optimal effort, points showing the inflection point were sized based on proximity of the average value for a parameter at a station to an appropriate threshold. Larger points indicated a station had a parameter close to a threshold, whereas smaller points indicated the average value was either much less or much greater than the threshold. Proximity to the threshold provided added context that additional attention could be focused at sites where management intervention may be needed.

## Station similarity

The interactive websites also included a tab for identifying similarities among stations within each monitoring dataset. These analyses were provided to support decisions where potentially redundant sites with similar characteristics may be dropped or one site may be preferred over another for continued sampling if no additional information is gained by sampling multiple sites. Conversely, future sampling designs could be focused on locations where parameters or groups of parameters had maximum differences, thereby focusing efforts at locations with the highest variation among all sites. For individual parameters, dissimilarity measures across all sites were estimated by calculating the Euclidean distances of the standardized (zero mean, unit variance) values between each time series (Oksanen et al. [2018](#ref-Oksanen18)). This resulted in pairwise estimates of dissimilarity between all sites as a single relative number to quantitatively evaluate which sites had time series with similar characteristics. Further, a principal components analysis (PCA) was used on the average values of each time series across all parameters and sites to group similar sites using biplots (Venables and Ripley [2002](#ref-Venables02)). Each biplot showed site groupings relative to the dominant principal components and vectors for each parameter used in the PCA. Pairs of sites could also be selected on the website to view relative overlap for assessing similarity and which vectors (i.e., parameters) explained the groupings.

## Hotspots

The dry weather monitoring dataset (illicit discharge, illicit contaminants) included an analysis of “hotspots” that was not provided for the other datasets given the priority management questions of the review committee. This analysis was motivated by the monitoring question of which sites were more likely to have exceedances for important thresholds relative to those that were stable and of low concentration over time. Hotspots were identified for individual parameters across all stations as the proportion of observations that were above a threshold of interest, defined as a threshold of regulatory concern (e.g., for TMDL compliance) or as the median value across all stations if a threshold was not applicable. These estimates produced a map where site points were sized and colored relative to the number of exceedances to identify groups of sites where exceedances were more common. A similar analysis was conducted by grouping all parameters across all sites, where the percentage of exceedances were estimated as all instances of an exceedance across parameters divided by all observations at a site. This provided a similar map where hotspots could be evaluated relative to multiple parameters. For both analyses, sites could be filtered by relative sample effort (i.e., total number of observations) and by date ranges to subset sites with similar sampling characteristics. Options to filter sites by receiving waterbodies where TMDLs are currently in place was also provided.

# References

Chang, Winston, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson. 2020. *Shiny: Web Application Framework for R*. <https://CRAN.R-project.org/package=shiny>.

McGill, R., J. W. Tukey, and W. A. Larsen. 1978. “Variations of Boxplots.” *The American Statistician* 32 (1): 12–16.

Millard, Steven P. 2013. *EnvStats: An R Package for Environmental Statistics*. New York: Springer. <http://www.springer.com>.

Oksanen, Jari, F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R. Minchin, et al. 2018. *Vegan: Community Ecology Package*. <https://CRAN.R-project.org/package=vegan>.

RDCT (R Development Core Team). 2020. “R: A language and environment for statistical computing, v4.0.1. R Foundation for Statistical Computing, Vienna, Austria.”

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with S*. Fourth. New York, New York: Springer-Verlag. <http://www.stats.ox.ac.uk/pub/MASS4>.