Supplementary information for

Frequent detection of a human fecal indicator in the urban ocean: environmental drivers and covariation with enterococci

Wiley C. Jennings[a], Eunice C. Chern[b,c], Diane O'Donohue[d], Michael G. Kellogg[d], Alexandria B. Boehm[a]*

[a] *Department of Civil and Environmental Engineering, Environmental Engineering and Science, Stanford University, 94305-4020, USA. *Fax: (650) 725-3164; Tel: (650) 724-9128; E-mail: aboehm@stanford.edu*
[b] *San Francisco Public Utilities Commission, Water Quality Laboratory, 1000 El Camino Real, Millbrae, CA 94030*
[c] *EPA Region 10 Laboratory, 7411 Beach Dr. E, Port Orchard, WA 98366*
[d] *San Francisco Public Utilities Commission, Oceanside Biology Laboratory, 3500 Great Highway, San Francisco, CA 94132*

S1. Analysis of CSD impact

To assess whether CSDs had a measurable impact on indicator levels, only samples from monitoring/CSD stations were utilized. For this assessment, follow up samples were not used because they were collected on a conditional basis (conditional on cENT ≥104 MPN/100 mL), which could introduce bias; CSD only stations were not used because background samples (when CSDs were not occurring) did not exist for comparison; and monitoring only stations were not used because, by definition, it was not possible for a CSD to occur at these stations. The effect of precipitation was accounted for by including the precipitation variable as a covariate in each generalized linear mixed model (GLMM).

The statistically significant coefficients (one coefficient in each of three indicator models) for CSD occurrence estimated by the GLMMs indicate the log odds ratio of high indicator concentrations (defined in the main text) accompanying CSD occurrence (as compared to non-occurrence). A GLMM with a time-changing predictor, in this case the

"CSD occurrence" predictor (sometimes CSD=1 and sometimes CSD=0), incorporates both "within-station effects" (the longitudinal effect of a CSD on indicator levels at a given station) and "between station effects" (the cross-sectional effect of a CSD occurring at one station but not another on the same date) into its estimation of the regression coefficient for CSD occurrence. Thus, the significant coefficients for CSD in each of these GLMMs (one for each indicator) represent some combination of longitudinal and cross-sectional effects.

S2. Censored data analysis

In datasets with multiple reporting limits for a variable and many values below reporting limits, substituting arbitrary values, such as ½ LDC for measurements below an LDC, can produce spurious statistical results. Thus, substitution was avoided for hypothesis tests and correlation. Instead, data were recorded in interval-censored format, such that every value was represented by an interval (1). Measurements below an LDC were assigned an interval [0, LDC). Measurements ≥ LDC at a value *m* were assigned an interval [m, m]. Duplicate qPCR measurements were averaged to yield a single interval estimate for each sample by square-root transforming the interval bounds, averaging them (left with left bound, right with right bound), and then squaring the averaged bounds, yielding an interval estimate for each sample in original units. To test rank-order correlation between indicators, intervals were treated as left-censored, and Kendall's tau-a correlation coefficient for left-censored data (hereafter, KTC) was used. This measure correctly identifies ties for data with multiple reporting limits and accurately estimates p-values for data with a large number of ties (1). KTC is expected to be about 0.15 units smaller than Spearman's coefficient (which is also rank-order) for the same degree of correlation (2).

However, it was not possible to use interval-censored format for binary regression models because data needed to be dichotomized at a single number (i.e., the median). Thus, to calculate medians and dichotomize indicator concentrations for binary logistic regression, intervals of each sample were collapsed by square-root transforming interval bounds, averaging them, and then squaring this single value.

S3. Multivariate regression models

Multivariate binary logistic regression models incorporated environmental variables expected to be important in controlling indicator concentrations. Each monitoring station was modeled separately. Models were fit to data from weekly monitoring samples of cENT, qENT, and HF183, respectively, from all 14 monitoring stations, yielding a total of 42 models. LASSO regularization was used for variable selection and coefficient estimation (3), implemented in the R package 'glmnet' (4). The LASSO selects models by penalizing large coefficient estimates, avoiding model variance inflation which can be a problem for models with a relatively large number of predictors or co-linear predictors, such as the seasonal variables in this study (3).

Binary logistic models were selected to model these data because of the relatively high proportion of indicator concentrations below a limit of detection, especially for cENT. While other regression methods have been shown to perform as well or better than binary logistic regression for predicting environmental FIB levels (5,6), the high proportion of data <LDC and the presence of multiple LDCs in this dataset make dichotomizing outcome variables appropriate. Indicator concentrations were dichotomized at their respective median values (Table 2) to facilitate comparison between models of the three indicators

and maximize the number of binary events in each model (models suffered from non-convergence when dichotomized at presence-absence). Binary logistic models assume that the logit-transformed outcome variable is linearly related to predictor variables. This assumption was checked visually with data aggregated across all stations, and resulted in modeling CSD occurrence instead of CSD volume and square-root transforming the precipitation variable.

Table S1. Summary of 42 regression models – excluding precipitation as a predictor variable. Excluding precipitation and using CSD volume instead of CSD occurrence yields an identical summary table to the one below.

	Outcome variable			
Predictor	**cENT**	**qENT**	**HF183**	**Sum**
Solar insolation, mean daily	4	2	2	8
Water temperature, mean daily	0	3	5	8
Tidal range, daily	0	0	3	3
Tide level	0	0	1	1
Tidal gradient	0	0	1	1
Significant wave height	0	1	0	1
CSD occurrence	0	0	0	0
Time since solar noon	0	0	0	0
Wind speed, mean daily	0	0	0	0

Table S2. Regression coefficients for 42 models including precipitation and CSD occurrence (same models as in the summary table presented in the main text). Descriptions of variables, including units, are given in Table 3. Models are binary logistic. Therefore, a coefficient value of 0.23 signifies that for every unit increase in the predictor variable, the log odds of the indicator being elevated (>median) increase by 0.23. Exponentiating these coefficients yields odds ratios.

Indicator	Station	Intercept	CSD occurrence	Precipitation, 72-hr	Significant wave height	Solar insolation, mean daily	Tidal gradient	Tidal range, daily	Tide level	Time since solar noon	Water temperature, mean daily	Wind speed, mean daily
cENT	O1	-1.08	0	0	0	0	0	0	0	0	0	0
	O2	-1.08	0	0	0	0	0	0	0	0	0	0
	O3	-1.20	0	0.23	0	-0.31	0	0	0	0	0	0
	O4	-1.56	0	0	0	0	0	0	0	0	0	0
	O5	-1.15	0	0	0	0	0	0	0	0	0	0
	O6	-0.30	0	0	0	0	0	0	0	0	0	0
	O7	-0.63	0	0	0	0	0	0	0	0	0	0
	B1	-0.70	0	0.12	0	0	0	0	0	0	0	0
	B2	-0.58	0	0	0	0	0	0	0	0	0	0
	B3	-0.03	0	0	0	0	0	0	0	0	0	0
	B4	-0.94	0	0	0	0	0	0	0	0	0	0
	B5	-0.47	0	0	0	-0.23	0	0	0	0	0	0
	B6	-0.46	0	0	0	0	0	0	0	0	0	0
	B7	-0.03	0	0	0	-0.25	0	0	0	0	0	0
qENT	O1	-0.03	0	0	0	0	0	0	0	0	0	0
	O2	0.00	0	0	0	0	0	0	0	0	0	0
	O3	-0.03	0	0	0.02	-0.31	0	0	0	0	-0.04	0
	O4	0.01	0	0.13	0	-0.16	-0.06	0	0	0.08	-0.46	0
	O5	-0.03	0	0	0	0	0	0	0	0	0	0
	O6	-0.03	0	0	0	0	0	0	0	0	0	0
	O7	0.00	0	0	0	0	0	0	0	0	0	0
	B1	-0.03	0	0	0	0	0	0	0	0	-0.17	0
	B2	-0.03	0	0	0	0	0	0	0	0	0	0
	B3	-0.03	0	0	0	0	0	0	0	0	0	0
	B4	-0.03	0	0.10	0	0	0	0	0	0	0	0
	B5	0.00	0	0.11	0	0	0	0	0	0	0	0
	B6	-0.03	0	0	0	0	0	0	0	0	0	0
	B7	-0.03	0	0	0	0	0	0	0	0	0	0

HF183	O1	-0.03	0	0	0	0	0	0	0	0	0	0
	O2	-0.03	0	0.11	0	0	0	0	0	0	0	0
	O3	-0.03	0	0	0	0	0	0	0	0	0	0
	O4	-0.08	0	0.11	0	0	0	0	0	0	-0.22	0
	O5	-0.03	0	0.17	0	0	0	0.29	0	0	-0.41	0
	O6	-0.08	0	0.24	0	0	0	0	0	0	-0.03	0
	O7	-0.02	0	0.18	0	0	0	0	0	0	-0.43	0
	B1	-0.03	0	0.05	0	0	0	0	0	0	0	0
	B2	-0.03	0	0	0	0	0	0	0	0	0	0
	B3	-0.02	0	0.26	0	0	0	0.17	0.07	0.02	-0.13	0
	B4	-0.03	0	0.17	0	0	0	0	0	0	0	0
	B5	-0.03	0	0	0	0	0	0	0	0	0	0
	B6	-0.03	0	0.10	0	0	0	0	0	0	0	0
	B7	-0.36	0	0	0	-0.31	0	0	0	0	0	0

References for supplementary information

1. Helsel DR. Statistics for censored environmental data using Minitab and R. 2nd ed. NJ, USA: Wiley; 2012.

2. Helsel DR, Hirsch RM. Statistical methods in water resources. In: US Geological Survey, Techniques of Water-Resources Investigations Book 4, Chapter A3. U.S. Geological Survey; 2002.

3. Tibshirani R. Regression shrinkage and selection via the LASSO. J R Stat Soc Ser B. 1994;58:267–288.

4. Friedman JH, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33(1):1–22.

5. Brooks W, Corsi S, Fienen M, Carvin R. Predicting recreational water quality advisories: A comparison of statistical methods. Environ Model Softw. 2016 Feb;76:81–94.

6. Thoe W, Gold M, Griesbach A, Grimmer M, Taggart ML, Boehm AB. Sunny with a Chance of Gastroenteritis: Predicting Swimmer Risk at California Beaches. Environ Sci Technol. 2015 Jan 6;49(1):423–31.