Supporting Information for "Know before you go: Data-driven beach water quality forecasting"

Ryan T. Searcy and Alexandria B. Boehm

Department of Civil and Environmental Engineering, Stanford University
Stanford, CA 94305, USA

Table of Contents

Study Sites - Description and FIB Monitoring

Cowell Beach (CB, 36.962 N, 122.023 W) is a semi-enclosed beach in Santa Cruz, California bordered by rock cliffs and a municipal wharf. It receives freshwater inflow nearly year-round from buried stormwater pipes and the nearby San Lorenzo River (which drain a watershed with semi-urban and agricultural land uses). Huntington State Beach (HSB, 33.633 N, 117.966 W) is an open-facing beach and receives seasonal freshwater inflow from the Santa Ana River (which drains a heavily-urbanized watershed). Mean annual precipitation is 790 mm for CB and 290 mm for HSB. Historically, CB has elevated to high levels of FIB in routinely-collected water samples, whereas HSB is less chronically impaired yet occasionally experiences spikes in FIB concentrations. There have been several recent infrastructure changes at CB during the data coverage period in this study that may have had impacts on water quality, including the implementation of routine cleaning of sewer mains and installation of bird netting; no infrastructure changes that could affect water quality at HSB are known. Study sites are mapped in Figure S1.

In the summer months (April-October), water samples are collected and analyzed for FIB on average 1 time per week at CB by the Santa Cruz County Environmental Health, and 2 times per week at HSB by the Orange County Environmental Health Division. While total coliforms are also monitored, we focused only on *Escherichia coli* (EC) and Enterococcus (ENT) data in this study. EC was enumerated from water samples at CB using membrane filtration (Standard Method 9222D) and at HSB using both membrane filtration and multiple tube fermentation (Standard Method 9221E); while the two methods yield EC concentrations in different units, here we treat them equivalently and henceforth use colony forming units (CFU)/100ml to maintain consistency. ENT was enumerated at both sites using membrane filtration (Standard Method 9230C at CB, EPA Method 1600 at HSB).

Tuning Case Study

The binary classifiers used in this study operate by first outputting a probability of the response being in the positive class and then assessing if that probability is greater than or equal to the decision threshold probability (DTP). In other words, probability predictions greater than the DTP were assigned to the positive class (i.e. prediction of FIB standard exceedance), while probability predictions below this value were assigned to the negative class (i.e. prediction of attainment of FIB standard). The default DTP used to develop models in this study was 0.5

Here we assess the potential of *DTP tuning* to improve sensitivity while preserving acceptable specificity in our models. As model AUC (area under the ROC curve) increases, the higher the model sensitivity can be tuned without sacrificing specificity. We randomly selected a subset of 80 models (~20% of the models developed in our study) equally distributed between the four model types (BLR, SVM, RF, GBM). Using the model's training data, we adjusted the DTP such that training sensitivity was maximized while training specificity was at minimum 80%. This tuning criteria is based on that determined acceptable by beach managers and FIB modelers in California (see Searcy et al. 2018 for more details). 'Tuned' model performance on the test data was then compared to the test performance prior to tuning.

We found that upon tuning, median sensitivity of the models in the subset on the test datasets rose from 0.00 [IQR 0.00 - 0.34] to 0.25 [0.10 - 0.45], while median specificity dropped from 0.93 [0.77 - 0.99] to 0.78 [0.72 - 0.84]. The median DTP used to optimize models was 0.363 [0.111 - 0.542]. Random forest models had the largest gain in sensitivity (median from 0.12 to 0.21) with the least tradeoff in specificity (median from 0.85 to 0.83), while support vector machine was the model type with the largest overall improvement in performance: median sensitivity of SVM models rose from 0.00 to 0.32 when tuned, while median specificity dropped from 1.00 to 0.78, nearly the optimal level set by the criteria. See Table S9 for all data.

Tables

Table S1: FIB and Environmental Data Sources and Stations

Data Type	Source	Station(s)	Temporal Interval	% Missing*
FIB	Santa Cruz County Environmental Health	CB: West of Wharf (36.962 N, 122.023 W)	Approx. weekly	-
	Orange County Environmental Health Division	HSB: 3N (33.633 N, 117.966 W)	Approx twice weekly	-
Tide	NOAA CO-OPS	CB: Monterey HSB: Long Beach	6 minutes	0%
Waves	CDIP	CB: Monterey Bay West HSB: San Pedro	30 minutes	4% < 1 %
Water Quality	CenCOOS	CB: Santa Cruz Municipal Wharf HSB: Newport Pier	5 minutes 4 minutes	56% 13%
Meteorological	Wind: CenCOOS (CB) or NCDC (HB)	CB: Santa Cruz Municipal Wharf HSB: John Wayne Airport	1 hour	23% 0%
	Other Parameters: CIMIS	CB: Santa Cruz HSB: Irvine		0% 3%
Currents	SCOOS	CB: N/A HSB: 33.620 N, 117.961 W	1 hour	- 54%
Streamflow	USGS	CB: San Lorenzo River HSB: Santa Ana River	1 day	0%

* Refers to the missing data on days in which FIB measurements were made between the years 2007 and 2021.

Table S2.1: Environmental variable descriptions - Meteorological variables

Variable	Description
atemp[i]	Air temperature (mean daily on day *i*)
atemp_min[i]	Air temperature (mean daily on day i)
dtemp[i]	Dew point temperature (mean daily on day i)
relhum[i]	Relative humidity (mean daily on day i)
rad[i]	Solar irradiance (mean daily on day i)
wspd[i]	Wind speed (mean daily on day i)
awind[i]	Alongshore wind speed (mean daily on day i)
owind[i]	Offshore wind speed (mean daily on day i)
wspd[i]_q75	Wind speed (mean daily on day *i*) greater than 75th percentile value (Binary)
awind_bin[i]	Alongshore wind speed (mean daily on day i) positive (upshore) or negative (downshore) (Binary)
owind_bin[i]	Offshore wind speed (mean daily on day i) positive (offshore) or negative (onshore) (Binary)
lograin[j]	Precipitation (Total on day *j* , log10-transformed)
lograin[k]T	Precipitation (Total over previous *k* days, log10-transformed)
lograin[i]_[m]T	Precipitation (Total between days *k* and *m*, log10-transformed)

[i] = 1, 2, 3, 4, 5
[j] = 1, 2, 3, 4, 5, 6, 7
[k] = 2, 3, 4, 5, 6, 7, 14, 30
[m] = i + 1, 2, 4, 6, 13, 29

Table S2.2: Environmental variable descriptions - Tidal variables

Variable	Description
tide	Instantaneous tide level (Predicted)
tide_max[n]	Maximimum tide level (daily) (Predicted)
tide_min[n]	Minimum tide level (daily) (Predicted)
tide_range[n]	Tidal range (daily) (Predicted)
tide_spring	Tide cycle in spring or neap (Binary)

[n] = 0, 1, 2, 3, 4, 5

Table S2.3: Environmental variable descriptions - Wave variables

Variable	Description
WVHT[i]	Significant wave height (mean daily)
DPD[i]	Dominant wave direction (mean daily)
APD[i]	Average wave period (mean daily)
WVHT[i]_q75	Significant wave height (mean daily) greater than 75th percentile (Binary)
DPD[i]_q75	Dominant wave direction (mean daily) greater than 75th percentile (Binary)

[i] = 1, 2, 3, 4, 5

Table S2.4: Environmental variable descriptions - Water quality variables

Variable	Description
wtemp[i]	Water temperature (mean daily)
logchl[i]	Chlorophyll concentration (mean daily, log10-transformed)
logturb[i]	Turbidity (mean daily, log10-transformed)
cond[i]	Conductivity (mean daily)
DO[i]	Dissolved oxygen (mean daily)
sal[i]	Salinity (mean daily)
pH[i]	pH (mean daily)
chl[i]_q75	Chlorophyll concentration (mean daily) greater than 75th percentile (Binary)
turb[i]_q75	Turbidity (mean daily) greater than 75th percentile (Binary)

[i] = 1, 2, 3, 4, 5

Table S2.5: Environmental variable descriptions - Streamflow variables

Variable	Description
logflow[i]	Stream discharge (mean daily, log10-transformed)
flow[i]_q50	Stream discharge (mean daily) on previous day greater than 50th percentile (Binary)
flow[i]_q75	Stream discharge (mean daily) on previous day greater than 75th percentile (Binary)
flow[i]_q90	Stream discharge (mean daily) on previous day greater than 90th percentile (Binary)

[i] = 1, 2, 3, 4, 5

Table S2.6: Environmental variable descriptions - Current variables

Variable	Description
u[i]	Current speed in east-west direction (mean daily)
v[i]	Current speed in north-south direction (mean daily)
along[i]	Current speed in alongshore direction (mean daily)
cross[i]	Current speed in crossshore direction (mean daily)
current_mag[i]	Magnitude of current speed (mean daily)
current_q75[_i]	Magnitude of current speed (mean daily) greater than 75th percentile (Binary)
cross_bin[i]	Alongshore current speed (mean daily) positive (upshore) or negative (downshore) (Binary)
cross_mag[i]	Magnitude of current speed in crossshore direction (mean daily)
along_bin[i]	Crossshore current speed (mean daily) positive (offshore) or negative (onshore) (Binary)
along_mag[i]	Magnitude of current speed in alongshore direction (mean daily)

[i] = 1, 2, 3, 4, 5

Table S2.7: Environmental variable descriptions - Date variables

Variable	Description
year	Sampling year
month	Sampling month
doy	Day of year
dow	Day of week
weekend	Sampling performed on weekend (Friday-Sunday) (Binary)
weekend1	Sampling performed on day before weekend (Thursday-Saturday) (Binary)

Table S3: Parameter value sets used in grid search for each binary classifier. Classifiers were implemented using the scikit-learn package in python. Values marked with a * indicate the default parameter used for variable selection. The default max_features parameter for RF was 0.75., while the default kernel parameter for SVM was 'linear'.

Classifier	Parameter	Description	Values Searched
BLR	C	Regularization strength	0.0001, 0.001, 0.01, 0.1, 1*, 10
	l1_ratio	Elastic Net mixing parameter (Ratio of L1 to L2 penalty)	0.0, 0.1, 0.25, 0.5*, .75, 0.9, 1.0
SVM	C	Regularization strength	0.001, 0.01*, .1, 1, 10
	kernel	Kernel type	'rbf','poly','sigmoid'
	gamma	Kernel coefficient	'auto'*, 'scale'
RF	max_depth	Maximum depth of tree	3*, 5
	max_features	Proportion of features used for node splitting	0.25, 0.5
	min_samples_leaf	Minimum samples at a leaf node	1, 3*, 5
GBM	n_estimators	Number of estimators	100*, 200, 300
	max_depth	Maximum depth of estimator	3*, 5, 7
	learning_rate	Learning rate	0.01, 0.1*, 0.3
	subsample	Fraction of samples used to fit each estimator	0.5, 0.75*, 1

Table S4: FIB data from 2007-2021 (15 seasons). EXC - concentrations measured above CA regulatory standard; BLOQ - concentrations measured below LOQ

Beach	FIB	N Total	N EXC	(%)	N BLOQ	(%)
CB	EC	735	142	19%	56	8%
	ENT	709	91	13%	305	43%
HSB	EC	1411	81	6%	938	66%
	ENT	1426	117	8%	863	61%

Table S5: Summary of FIB distributions by model dataset partition.

Partition		CB - EC				CB - ENT				HSB - EC				HSB - ENT			
		Train		*Test*		*Train*		*Test*		*Train*		*Test*		*Train*		*Test*	
Train Years	*Test Years*	N	Exc	N	Exc	N	Exc	N	Exc	N	Exc	N	Exc	N	Exc	N	Exc
2007 - 2012	2013 - 2014	290	85	94	20	263	43	94	15	763	53	112	9	764	73	115	13
2008 - 2013	2014 - 2015	292	82	85	17	289	46	85	12	688	42	114	10	693	54	118	15
2009 - 2014	2015 - 2016	278	79	68	3	276	52	68	1	611	41	111	5	619	50	116	9
2010 - 2015	2016 - 2017	257	58	64	8	256	39	64	1	506	36	101	1	513	46	104	4
2011 - 2016	2017 - 2018	246	47	69	13	245	38	69	6	419	31	110	2	427	44	109	4
2012 - 2017	2018 - 2019	235	42	88	9	235	25	88	13	380	29	111	4	387	42	116	10
2013 - 2018	2019 - 2020	231	36	96	9	231	22	96	12	319	16	108	3	326	26	114	9
2014 - 2019	2020 - 2021	237	34	93	8	237	26	92	13	322	15	100	4	334	29	100	5
Average		258	58	82	11	254	36	82	9	501	33	108	5	508	46	112	9

Table S6: Top 10 most-common variables across all models and by beach and FIB type. Percentages indicate how frequently a variable was selected for inclusion in models.

Rank	All Models		CB - EC		CB - ENT		HSB - EC		HSB - ENT	
	Variable	%	Variable	%	Variable	%	Variable	%	Variable	%
1	tide	43%	tide_max4	43%	doy	44%	tide	68%	tide	50%
2	rad4	29%	APD4	42%	tide_max1	34%	tide_range1	50%	awind3	44%
3	APD3	26%	logflow4	33%	tide	33%	rad4	38%	wspd4	42%
4	atemp_min5	25%	APD3	33%	logflow4	30%	rad3	34%	tide_range1	35%
5	lograin6	24%	APD2	33%	DPD4	29%	lograin5_11T	32%	APD3	32%
6	rad3	23%	lograin6	32%	rad2	25%	lograin6	31%	tide_max1	31%
7	rad5	22%	rad4	29%	rad3	24%	atemp_min5	28%	rad5	31%
8	tide_range1	21%	WVHT4	28%	APD3	24%	APD3	21%	atemp_min5	27%
9	APD4	19%	atemp_min5	27%	APD2	24%	rad2	18%	WVHT3	25%
10	APD2	19%	APD5	23%	rad4	23%	DPD_q75_5	18%	lograin6	24%

Table S7: Predictive performance by model type. PER - persistence model. IQR - Interquartile range.

Model	Sensitivity			Specificity			AUC		
	Median	IQR		Median	IQR		Median	IQR	
BLR	0.40	0.22	0.56	0.69	0.62	0.76	0.57	0.52	0.67
SVM	0.00	0.00	0.00	1.00	1.00	1.00	0.56	0.48	0.65
RF	0.13	0.00	0.40	0.85	0.79	0.93	0.60	0.54	0.68
GBM	0.00	0.00	0.15	0.95	0.92	0.97	0.59	0.51	0.65
PER	0.00	0.00	0.15	0.95	0.89	0.98	0.50	0.48	0.51

Table S8: Model performance by beach and FIB type. IQR - Interquartile range.

Models		Sensitivity			Specificity			AUC		
Beach	FIB	Median	IQR		Median	IQR		Median	IQR	
CB	EC	0.12	0.00	0.34	0.91	0.71	0.98	0.58	0.48	0.65
	ENT	0.00	0.00	0.14	0.94	0.81	1.00	0.54	0.46	0.58
HSB	EC	0.00	0.00	0.25	0.93	0.82	0.99	0.62	0.54	0.71
	ENT	0.10	0.00	0.41	0.92	0.77	0.99	0.62	0.56	0.70
Persistence		**Sensitivity**			**Specificity**			**AUC**		
Beach	FIB	Median	IQR		Median	IQR		Median	IQR	
CB	EC	0.11	0.00	0.26	0.90	0.85	0.95	0.51	0.47	0.58
	ENT	0.15	0.05	0.17	0.90	0.86	0.96	0.51	0.50	0.54
HSB	EC	0.00	0.00	0.00	0.98	0.95	0.99	0.49	0.48	0.50
	ENT	0.00	0.00	0.00	0.97	0.94	1.00	0.50	0.48	0.50

Table S9: Nowcast and forecast performance by lead time. Metrics are aggregated from all beaches, FIB type, data partition, and model type. The performance of the persistence method ('PER') at a lead time of 0 is also provided for comparison. IQR - interquartile range

Lead Time	Sensitivity			Specificity			AUC		
	Median	IQR		Median	IQR		Median	IQR	
0	0.08	*0.00*	*0.28*	0.93	*0.84*	*0.98*	0.58	*0.49*	*0.67*
1	0.07	*0.00*	*0.33*	0.94	*0.79*	*0.99*	0.60	*0.52*	*0.67*
2	0.00	*0.00*	*0.25*	0.92	*0.78*	*0.98*	0.58	*0.50*	*0.65*
3	0.00	*0.00*	*0.33*	0.91	*0.74*	*1.00*	0.58	*0.49*	*0.68*
PER	0.04	*0.0*	*0.18*	0.92	*0.88*	*0.96*	0.49	*0.48*	*0.53*

Table S10: Proportion of models with improved performance over the persistence method. Calculated across all beach, FIB type, data partition, and lead time combinations. A star indicates that the metric for a given model type was significantly greater on average than the persistence method (Wilcoxon Signed Rank test $p < 0.05$)

Model	Sensitivity	Specificity	AUC
BLR	**0.76***	0.01	0.76*
SVM	0.03	**0.77***	0.67*
RF	0.54*	0.18	**0.82***
GBM	0.25	0.47	0.71*
Overall	0.4	0.36	0.74

Figures

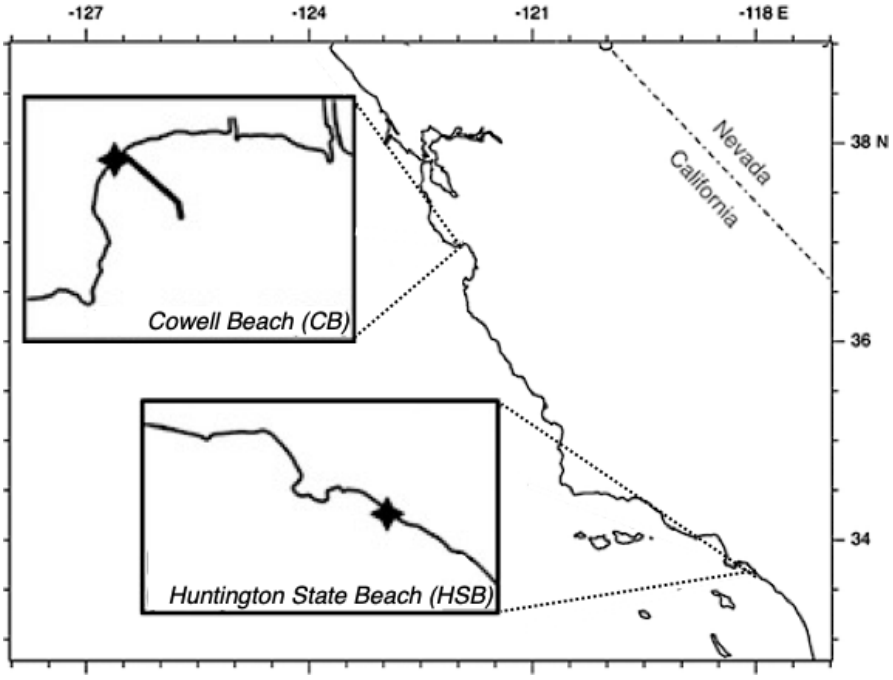

Figure S1 - Map of study sites on California coast.

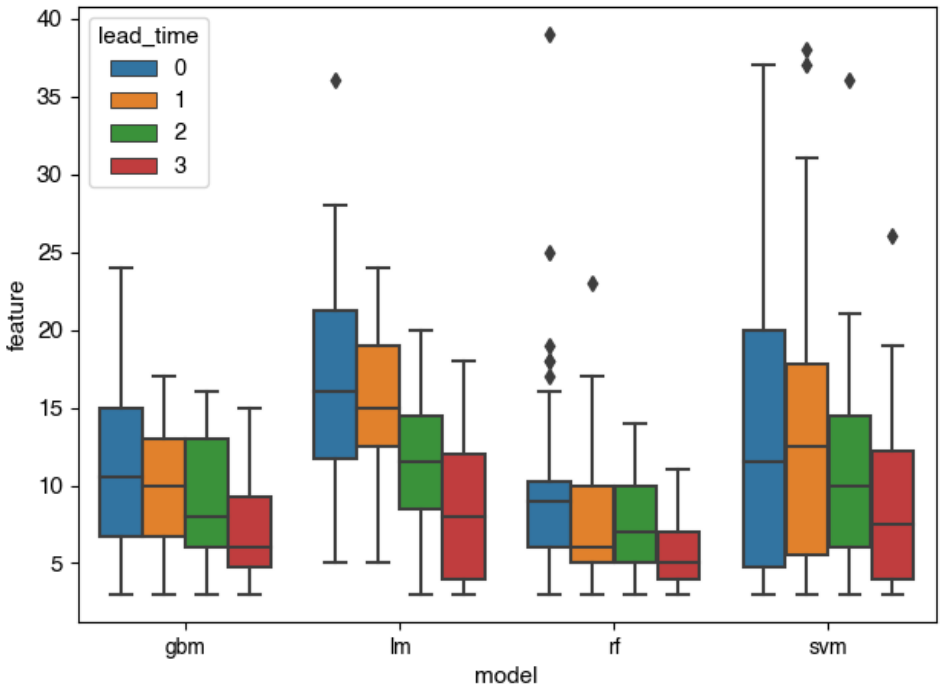

Figure S2: Distribution of the number of environmental variables included in models. The middle line in the boxplots represents the median; the upper and lower edges of the boxes represent the 75th and 25th quantiles, respectively. The whiskers extend to 1.5 times the interquartile range (75th quartile−25th quartile).

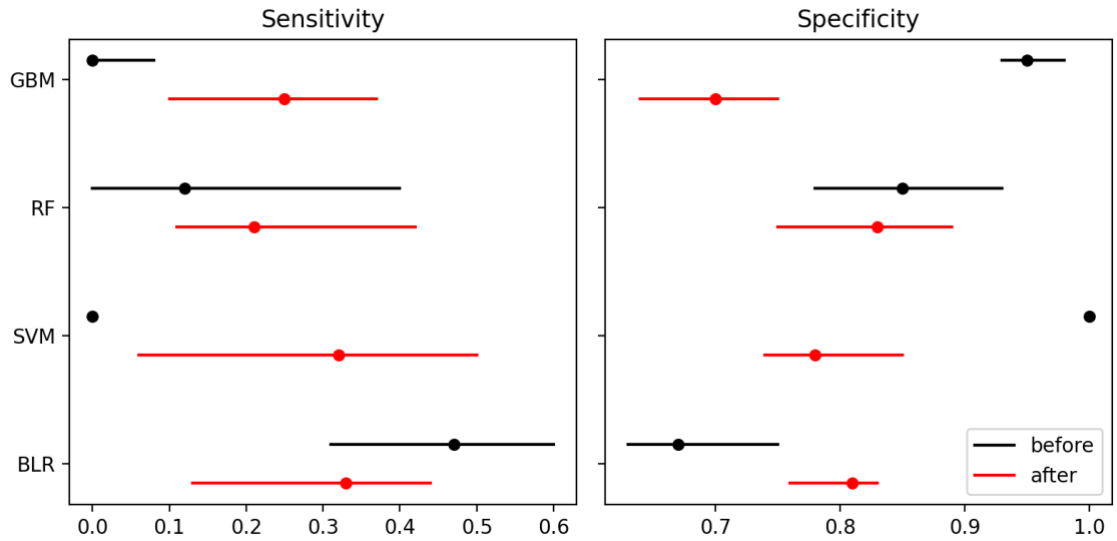

Figure S3: Change in model sensitivity and specificity before (black) and after (red) tuning. Dots represent the median of the 20 models tested per model type, while the extents of the lines span the IQR.