

SUPPLEMENTARY MATERIAL

Environmental prediction of enterococci analytical states across sampled coastal locations and laboratory batches in the United States
Blinded review version

Three-state model

Table S1. Three-state multinomial performance in the marine cohort

Validation scheme	Model log-loss	Null log-loss	Relative reduction, % (95% CI)	Macro ROC-AUC	Balanced accuracy
Analytical batches held out	0.9433	0.9966	5.11 (1.67 to 8.53)	0.665	0.452
Stable locations held out	0.9233	0.9959	7.21 (4.20 to 10.28)	0.68	0.459

Note: The three states were non-detect, numerical estimate below the lower limit of quantification, and unflagged numerical result. The last category denotes a numerical result with no analytical qualifier; it is an operational reporting category, not a claim of complete analytical validity.

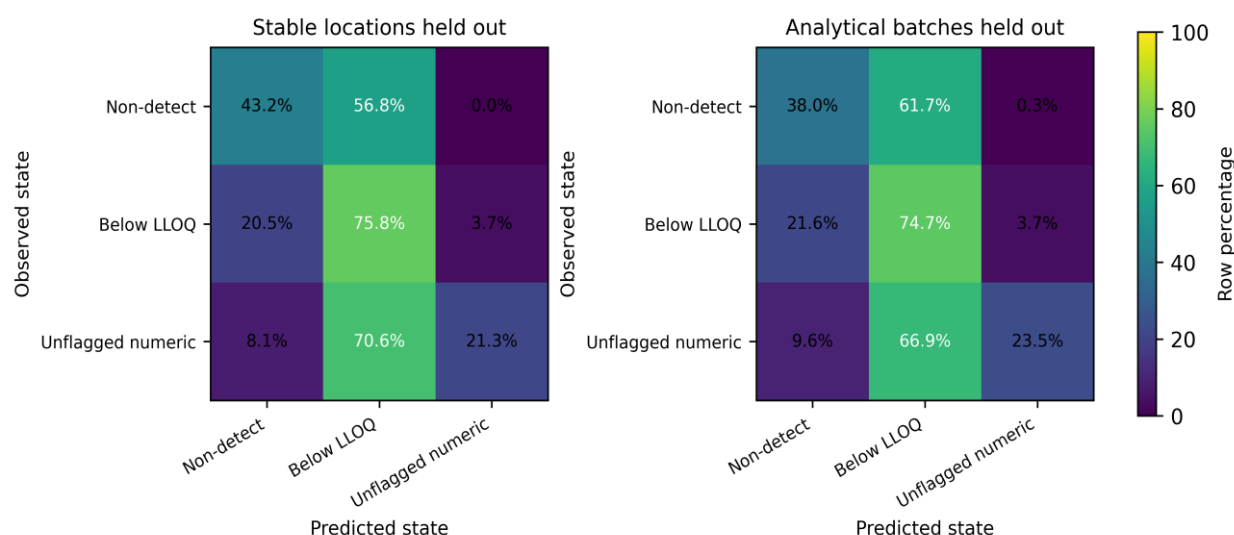


Figure S1. Row-normalized confusion matrices for the three-state model. Percentages sum to 100% within each observed state.

Table S2. Event counts and row percentages for the three-state confusion matrices

Validation scheme	Actual state	Predicted state	Events	Row percentage
Analytical batches held out	Non-detect	Non-detect	115	38.0
Analytical batches held out	Non-detect	Below LLOQ	187	61.7
Analytical batches held out	Non-detect	Unflagged numeric	1	0.3
Analytical batches held out	Below LLOQ	Non-detect	99	21.6
Analytical batches held out	Below LLOQ	Below LLOQ	342	74.7
Analytical batches held out	Below LLOQ	Unflagged numeric	17	3.7
Analytical batches held out	Unflagged numeric	Non-detect	13	9.6
Analytical batches held out	Unflagged numeric	Below LLOQ	91	66.9

Validation scheme	Actual state	Predicted state	Events	Row percentage
Analytical batches held out	Unflagged numeric	Unflagged numeric	32	23.5
Stable locations held out	Non-detect	Non-detect	131	43.2
Stable locations held out	Non-detect	Below LLOQ	172	56.8
Stable locations held out	Non-detect	Unflagged numeric	0	0.0
Stable locations held out	Below LLOQ	Non-detect	94	20.5
Stable locations held out	Below LLOQ	Below LLOQ	347	75.8
Stable locations held out	Below LLOQ	Unflagged numeric	17	3.7
Stable locations held out	Unflagged numeric	Non-detect	11	8.1
Stable locations held out	Unflagged numeric	Below LLOQ	96	70.6
Stable locations held out	Unflagged numeric	Unflagged numeric	29	21.3

Batch-conditioned environmental associations

Table S3. All environmental associations conditioned on enterococci analytical batch

Environment	Predictor	Events	Informative batches	OR per SD	CI lower	CI upper	FDR-adjusted p
Marine	Total nitrogen, log(1+x)	868	32	1.2079	0.8844	1.6497	0.4029
Marine	Total phosphorus, log(1+x)	868	32	1.2436	0.9196	1.6817	0.3138
Marine	Dissolved inorganic nitrogen, log(1+x)	868	32	0.9356	0.7017	1.2473	0.709
Marine	Chlorophyll a, log(1+x)	868	32	1.1703	0.8808	1.555	0.4173
Marine	Surface dissolved oxygen	868	32	1.272	0.9767	1.6565	0.1782
Marine	Surface pH	868	32	0.6306	0.4999	0.7954	0.0012
Marine	Surface temperature	868	32	0.7365	0.5654	0.9593	0.0933
Marine	Station depth, log(1+x)	868	32	0.7307	0.5827	0.9163	0.0395
Marine	Secchi depth, log(1+x)	868	32	0.8902	0.6582	1.204	0.6004
Marine	Day-of-year sine	868	32	0.6589	0.1597	2.7181	0.6767
Marine	Day-of-year cosine	868	32	0.3027	0.0891	1.0282	0.1663
Marine	Surface salinity	868	32	0.9672	0.7748	1.2072	0.7678
Great Lakes	Total nitrogen, log(1+x)	404	30	0.2498	0.0637	0.9787	0.2002
Great Lakes	Total phosphorus, log(1+x)	404	30	1.0655	0.5957	1.9058	0.8307
Great Lakes	Dissolved inorganic nitrogen, log(1+x)	404	30	3.098	0.9249	10.377	0.2002
Great Lakes	Chlorophyll a, log(1+x)	404	30	1.3678	0.7651	2.4454	0.5813

Environment	Predictor	Events	Informative batches	OR per SD	CI lower	CI upper	FDR-adjusted p
Great Lakes	Surface dissolved oxygen	404	30	0.6803	0.372	1.2442	0.5066
Great Lakes	Surface pH	404	30	0.9117	0.6101	1.3624	0.7668
Great Lakes	Surface temperature	404	30	0.8761	0.444	1.7288	0.7668
Great Lakes	Station depth, log(1+x)	404	30	0.5864	0.4058	0.8472	0.0536
Great Lakes	Secchi depth, log(1+x)	404	30	0.8293	0.519	1.3251	0.6506
Great Lakes	Day-of-year sine	404	30	0.4179	0.0147	11.8998	0.7668
Great Lakes	Day-of-year cosine	404	30	0.2304	0.049	1.0831	0.2002
Great Lakes	Surface conductivity, log(1+x)	404	30	0.8182	0.5086	1.3165	0.6506

Note: Associations were evaluated separately by environment. They are observational and do not identify causal effects or fecal sources.

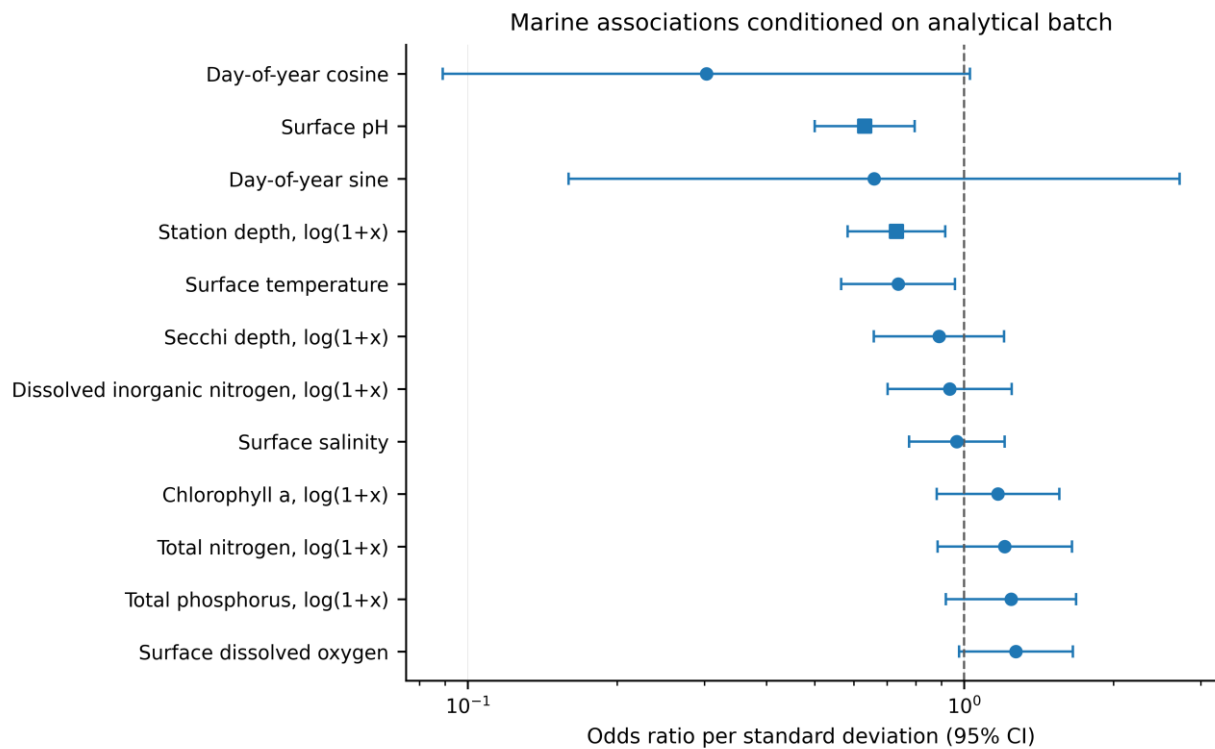


Figure S2. Marine environmental associations conditioned on analytical batch. Squares identify associations surviving false-discovery-rate control.

Temporal and domain transportability

Table S4. Temporal and leave-domain-out validation

Environment	Validation	Held-out domain	Training events	Test events	Model log-loss	Null log-loss	Relative improvement	ROC-AUC
Marine	Temporal	2021	711	186	0.5279	0.5847	9.71%	0.6591
Marine	Leave-domain-out	Gulf	632	265	0.7061	0.726	2.75%	0.6366
Marine	Leave-domain-out	Northeast	514	383	0.5703	0.6246	8.70%	0.7181
Marine	Leave-	Southeast	783	114	0.6545	0.574	-14.02%	0.4725

Environment	Validation	Held-out domain	Training events	Test events	Model log-loss	Null log-loss	Relative improvement	ROC-AUC
	domain-out							
Marine	Leave-domain-out	West	762	135	0.6227	0.6269	0.67%	0.6704
Great Lakes	Temporal	2021	314	69	0.6847	0.6664	-2.75%	0.542
Great Lakes	Temporal	2022	383	37	0.5829	0.5988	2.65%	0.619
Great Lakes	Leave-domain-out	Lake Erie	323	97	0.8674	0.7151	-21.29%	0.465
Great Lakes	Leave-domain-out	Lake Huron	370	50	0.6474	0.6919	6.42%	0.6526
Great Lakes	Leave-domain-out	Lake Michigan	284	136	0.6848	0.6513	-5.15%	0.4253
Great Lakes	Leave-domain-out	Lake Ontario	370	50	0.6567	0.6285	-4.49%	0.496
Great Lakes	Leave-domain-out	Lake Superior	333	87	0.8855	0.7352	-20.45%	0.4678

Advanced-method comparison

Table S5. Penalized logistic regression versus LightGBM

Environment	Validation	Logistic log-loss	LightGBM log-loss	LightGBM change vs logistic, % (95% CI)
Marine	Locations held out	0.5911	0.5869	0.72 (-1.65 to 3.06)
Marine	Batches held out	0.5956	0.6031	-1.26 (-4.02 to 1.33)
Great Lakes	Locations held out	0.6585	0.6474	1.69 (-2.64 to 6.03)
Great Lakes	Batches held out	0.6795	0.6783	0.18 (-3.60 to 4.01)

Note: Positive change indicates lower log-loss for LightGBM. Values in parentheses are 95% UNIQUE_ID-cluster bootstrap confidence intervals. The advanced method required at least 2% improvement in both grouped validation schemes to be retained.

Analytical-batch heterogeneity

Table S6. Heterogeneity of numerical-result proportions among analytical batches

Environment	Events	Batches	Chi-square	df	p
Great Lakes	420	31	64.51	30	0.000253
Marine	897	32	37.677	31	0.190197

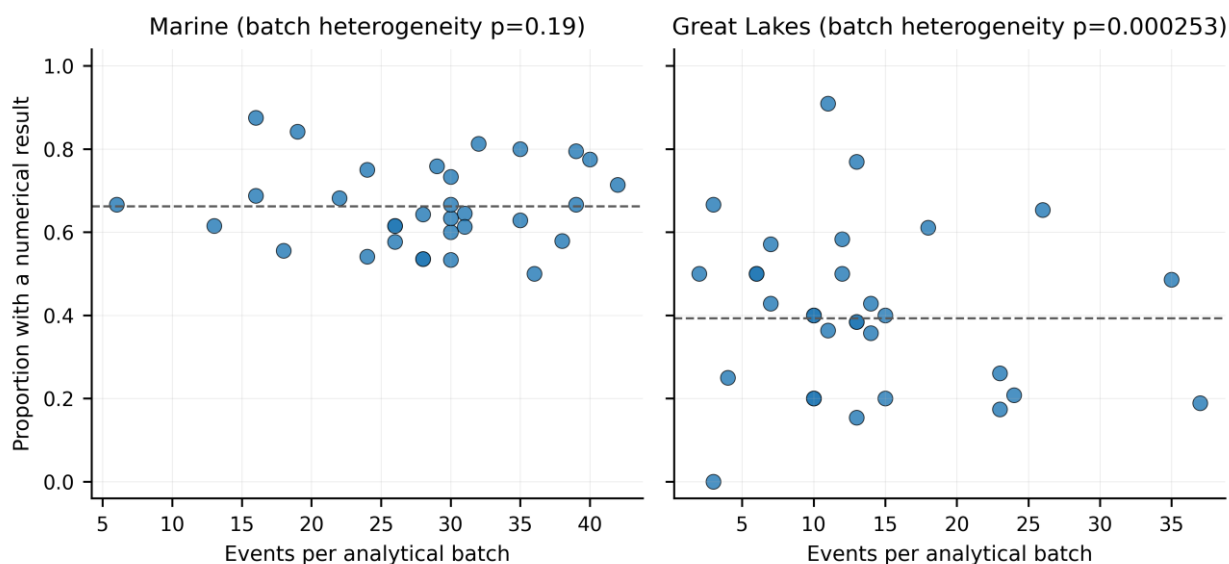


Figure S3. Batch-level proportion of enterococci records with a numerical result. Dashed lines show environment-wide proportions.

Source-file and variable crosswalk

Table S7. Original source files used in data assembly

Analytical component	Original source file	Analytical use
Enterococci measurements	ncca20_enterococci.csv	Outcome reconstruction and analytical-batch lineage
Water chemistry	ncca20_waterchemistry.csv	Nutrient and chlorophyll a predictors
Hydrographic profiles	ncca20_hydrographicprofile.csv	Surface physicochemical predictors and station depth
Water transparency	ncca20_secchi.csv	Secchi-depth predictor
Auxiliary site metadata	ncca20_siteinfo_estuarine.csv; ncca20_siteinfo_greatlakes.csv	Audited but not required for primary eligibility

Table S8. Original source-variable and internal-label crosswalk

Main-text term	Original source label	Analytical role
Sampling-event identifier	UID	Event-level linkage key
Stable site identifier	UNIQUE_ID	Grouping repeated visits
Analytical batch identifier	BATCH_ID	Grouped validation and conditional analysis
Enterococci calibrator cell	ENT_1X_CCE_100ML	Outcome reconstruction
equivalents per 100 mL		
Official analytical flag	NARS_FLAG	Classification of analytical states
Total nitrogen	NTL	Water-chemistry predictor
Total phosphorus	PTL	Water-chemistry predictor
Dissolved inorganic nitrogen	DIN	Water-chemistry predictor
Chlorophyll a	WCHL	Water-chemistry predictor
Non-detect	NONDETECT_ND	Reproducible outcome class
Below-quantification estimate	NUMERIC_BELOW_LLOQ_ESTIMATE	Reproducible outcome class
Unflagged numerical result	NUMERIC_UNFLAGGED_QUANTIFIABLE	Reproducible outcome class