

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

Alberto Franco Cerna-Cueva^{1*} , César Samuel López-López² 

¹ Professional School of Environmental Engineering, Faculty of Renewable Natural Resources, Universidad Nacional Agraria de la Selva, Carretera Central km 1.21, Tingo María, Huanuco 10131, Peru

² Faculty of Renewable Natural Resources, Universidad Nacional Agraria de la Selva, Carretera Central km 1.21, Tingo María 10131, Huánuco, Peru

* Corresponding author's e-mail: alberto.cerna@unas.edu.pe

ABSTRACT

Quantitative polymerase chain reaction measurements of enterococci often include non-detects and numerical estimates below the lower limit of quantification; collapsing these analytically distinct records into a single quantitative outcome can obscure censoring. This study evaluated whether concurrent environmental conditions predict enterococci analytical states and whether predictive performance transfers to unseen sampling locations and laboratory batches. Record-level field and laboratory data from the United States National Coastal Condition Assessment 2020 included 897 eligible marine events and a secondary contrast of 420 Great Lakes events. Penalized logistic regression was evaluated using repeated five-fold cross-validation grouped first by stable location and then by complete analytical batch, with all preprocessing estimated within training folds. A secondary multinomial model retained non-detect, below-quantification estimate, and unflagged numerical states. In marine events, the environmental model reduced log-loss relative to a prevalence-only comparator by 7.58% under location-held-out validation and 6.94% under batch-held-out validation; corresponding areas under the receiver operating characteristic curve were 0.699 and 0.692, and 95% UNIQUE_ID-cluster bootstrap intervals for absolute improvement excluded zero in both schemes. The three-state model retained relative log-loss reductions of 7.21% and 5.11%, respectively. A prespecified boosting model did not improve batch-held-out performance and was not retained. Performance varied among regions and did not transfer robustly to Great Lakes batches. Concurrent environmental conditions therefore provided moderate predictive information about enterococci analytical states in sampled marine events, but performance remained insufficient for operational use and depended on analytical censoring, laboratory batch, and geographic domain.

Keywords: enterococci, quantitative polymerase chain reaction, detection and quantification limits, grouped cross-validation, laboratory batch, coastal monitoring.

INTRODUCTION

Enterococci are widely used as fecal-indicator bacteria in coastal and recreational-water monitoring, and quantitative polymerase chain reaction (qPCR) methods can provide same-day measurements of enterococcal DNA targets. The analytical construct nevertheless depends on the method. Culture-based procedures enumerate culturable organisms as colony-forming units, whereas qPCR

reports molecular target abundance in calibrator cell equivalents; the two measurements are related but not directly interchangeable, and their relationship can vary among beaches, temporal scales, and assay contexts (Converse et al., 2012; Wymer et al., 2021). Molecular results must therefore be interpreted as assay-specific observations rather than direct measures of viable cells, pathogen abundance, or individual infection risk (U.S. Environmental Protection Agency [U.S. EPA], 2012).

Environmental qPCR records also contain qualitatively different analytical states. Detection and quantification limits answer different questions, and a numerical estimate below the lower limit of quantification does not have the same evidential precision as a fully quantifiable result (Klymus et al., 2020; Borchardt et al., 2021). For enterococci, multi-laboratory evaluations have documented between-laboratory variation and sensitivity to environmental matrices, inhibition, and quality-control implementation, while site-specific assessments likewise recommend verifying performance under local conditions (Haugland et al., 2016; Saleem et al., 2022). Treating non-detects, below-quantification estimates, and quantifiable results as a homogeneous continuous outcome can consequently conceal both censoring and analytical heterogeneity.

Environmental prediction of fecal-indicator bacteria is already a mature research area. A systematic review identified 53 predictive-model studies at freshwater recreational sites and highlighted recurrent limitations involving validation, missing-data handling, and incomplete reporting (Heasley et al., 2021). Site-specific models have also predicted qPCR enterococci and other microbial indicators at Great Lakes beaches, and advanced ensemble approaches have been evaluated for beach-water-quality forecasting (Cyterski et al., 2022; Wang et al., 2021). Accordingly, novelty cannot rest on applying a new algorithm or on showing that environmental variables correlate with enterococci. A more specific question is whether concurrent environmental conditions predict the analytical reporting state itself after censoring is retained explicitly.

A second requirement is credible evaluation of transportability. Randomly splitting clustered observations can yield optimistic performance when repeated measurements share locations, periods, or other higher-level structures; validation groups should instead reflect the intended prediction task (Roberts et al., 2017). The same principle applies to laboratory lineage. Because qPCR performance can vary among analytical settings and matrices, leaving entire batches out of model training provides a direct test of whether an environmental signal extends beyond batch-specific regularities (Haugland et al., 2016). Location-grouped and batch-held-out validation therefore address different, complementary forms of generalization.

This study used record-level field and laboratory data from the National Coastal Condition

Assessment 2020, a component of the standardized National Aquatic Resource Surveys monitoring framework (Nahlik et al., 2025; U.S. EPA, 2025b). The primary objective was to determine whether concurrent environmental conditions improve probabilistic prediction of a numerical enterococci quantitative polymerase chain reaction result versus a non-detect in sampled marine events when stable locations and complete analytical batches are held out. A secondary multinomial analysis retained three states – non-detect, numerical estimate below the lower limit of quantification, and unflagged numerical result – and additional analyses evaluated regional, temporal, and Great Lakes transportability. Penalized logistic regression was prespecified as the interpretable baseline, whereas an advanced boosting model was retained only if it produced a material and transportable improvement. All conclusions were restricted to analytical-state predictability in the sampled events and did not infer viable-organism abundance, fecal source, causality, health risk, or population-wide coastal prevalence.

MATERIALS AND METHODS

Study design, data source and inferential scope

This observational secondary analysis used record-level files from the United States Environmental Protection Agency National Coastal Condition Assessment 2020, part of the standardized National Aquatic Resource Surveys monitoring framework (Nahlik et al., 2025; U.S. EPA, 2025b). The assessment-cycle label encompasses visits conducted in 2020, 2021 and 2022. Four official tables were used: enterococci, water chemistry, hydrographic profiles and Secchi depth, together with their data dictionaries. The exact unmodified source tables and companion metadata are supplied in Supplementary File S2 with official URLs, acquisition records and SHA-256 hashes. Supplementary File S1 provides the ordered reconstruction code and verification records. Laboratory and quality-assurance procedures were interpreted using the official laboratory manual (U.S. EPA, 2025a). Records were linked at the sampling-event level using the official event identifier. A stable site identifier tracked repeat visits, and a laboratory batch identifier tracked analytical lineage. Marine events constituted the primary

analysis; Great Lakes events were analyzed separately as a transportability contrast and were never pooled with marine records. A feasibility screen preceded the final analysis plan; consequently, the work is described as rigorously grouped internal validation rather than fully independent external validation. The delivered files did not contain survey weights or the complete design variables required for population estimates, so the inferential scope was restricted to the sampled and analytically eligible events.

Cohort construction and exclusions

All 1330 enterococci records were audited before linkage. The analytical state was interpretable for 1317 records. Thirteen records classified as having no laboratory result, accompanied by laboratory quality-control failure comments, were excluded from both the binary and three-state outcomes; they were not coded as non-detects or zeros. The remaining events were retained even when one or more environmental predictors were missing. Environmental records were left-joined by the official event identifier, thereby preserving the eligible outcome cohort, and missing predictor values were handled only within the training portion of each validation fold. Dates were parsed explicitly in month/day/year format. Events classified as marine in the official water-body-type field formed the primary cohort, whereas the remaining eligible Great Lakes events formed the separate secondary cohort. The auxiliary site-metadata file was not required for eligibility because none of the primary predictors depended on it and requiring it would have introduced avoidable attrition.

Enterococci analytical-state outcomes

The outcome was reconstructed from the reported enterococci calibrator cell equivalents per 100 mL and the official National Aquatic Resource Surveys analytical flag. Records carrying the non-detect qualifier were classified as non-detects. Numerical records carrying the below-quantification qualifier were classified as numerical estimates below the lower limit of quantification because the data dictionary states that reported calibrator cell equivalents include non-reportable estimates below that limit. Numerical records with an empty analytical flag were classified operationally as unflagged numerical results.

An empty flag was not treated as independent confirmation that every analytical quality attribute had been satisfied. Original source-variable names and internal class labels are reported only in Supplementary Appendix S1 for reproducibility. The primary binary outcome combined the two numerical categories and contrasted any numerical result with a non-detect. The secondary multinomial outcome retained all three categories separately. Continuous calibrator cell equivalents were not used as the primary target because most numerical records were analytically qualified as below the quantification limit. These outcomes represent molecular analytical state or reportability (Borchardt et al., 2021; Klymus et al., 2020). They were not interpreted as viable cells, colony-forming units, pathogens, fecal source, or infection risk (U.S. EPA, 2015).

Environmental predictor derivation

Predictors were fixed before final model comparison and derived at the sampling-event level. Water-chemistry predictors comprised total nitrogen (mg/L), total phosphorus (mg/L), dissolved inorganic nitrogen (mg/L), and chlorophyll a ($\mu\text{g/L}$). Results were converted to numeric values and collapsed by the median when more than one row occurred for an event. For total nitrogen, total phosphorus, and chlorophyll a, a separate binary quality indicator recorded whether any non-empty analytical qualifier was present, preventing qualified values from being treated as having the same status as unflagged values. Laboratory pH, laboratory conductivity, ammonia nitrogen, and soluble reactive phosphorus were imported during data assembly but were not included because they were outside the frozen predictor set. For hydrographic profiles, the surface stratum was defined as the minimum recorded depth within each event; when several readings occurred at that depth, their median was used. Surface dissolved oxygen (mg/L), pH, and temperature ($^{\circ}\text{C}$) were retained in both environments, surface salinity (practical salinity units) was used for marine events, and surface conductivity ($\mu\text{S/cm}$) was used for Great Lakes events. Station depth (m) was obtained from the same surface-profile rows. Secchi depth (m) was collapsed by the event median. Collection dates supplied calendar year and day of year; seasonal position was represented by sine and cosine terms using a 365.25-day period. Total nitrogen, total phosphorus, dissolved

inorganic nitrogen, chlorophyll a, station depth, and Secchi depth were transformed as $\log(1 + x)$ after enforcing their nonnegative measurement domain; surface conductivity received the same transformation in Great Lakes. Marine models additionally included National Coastal Condition Assessment region, whereas Great Lakes models included lake identity; year was categorical in both. Administrative and laboratory-lineage identifiers were retained for linkage or grouping only and were blocked as predictors.

Primary model and fold-wise preprocessing

The primary model was L2-penalized logistic regression with $C = 0.1$, the liblinear solver, a maximum of 300 iterations and random seed 20260827. The regularization setting was fixed rather than selected after examining outcome performance. Numerical variables were median-imputed with an accompanying missingness indicator and then standardized. Categorical variables were imputed with the most frequent training category and one-hot encoded; previously unseen test categories were ignored rather than used to refit the encoder. Every imputation value, scale parameter and category map was estimated exclusively from the training partition and then applied unchanged to the corresponding test partition. The null comparator assigned every test event the numerical-result prevalence observed in its training partition. The marine and Great Lakes models used the same pipeline but retained their environment-specific salinity or conductivity and geographic categories.

Grouped internal validation

Performance was estimated with stratified, grouped five-fold cross-validation repeated ten times. For repetition r , fold assignment used random seed $20260827 + r$. Two complementary grouping schemes were analyzed. In location-grouped validation, all events sharing a stable site identifier were kept in the same fold, preventing repeated visits to a stable location from appearing in both training and test data. In batch-held-out validation, all events sharing an enterococci analytical batch were kept together, so each test fold represented batches that were completely absent from model fitting. The environmental model and the prevalence comparator used identical test sets within each fold. Each eligible event therefore

received one out-of-fold prediction per repetition under each grouping scheme. These schemes directly evaluate generalization to unseen locations and unseen analytical batches and avoid the optimistic leakage that may arise from random row-level splitting of structured ecological data (Roberts et al., 2017). They remain internal validation procedures and do not substitute for prospective or locally external validation.

Performance estimands, calibration and uncertainty

The primary performance measure was log-loss because it evaluates the full predicted probability and penalizes confident errors. Secondary measures were the area under the receiver operating characteristic curve, average precision, Brier score, calibration intercept and calibration slope. ROC-AUC, average precision and Brier score were calculated within each test fold and summarized as unweighted means across completed folds. The primary estimand was the absolute reduction in out-of-fold log-loss relative to the training-prevalence comparator; relative reduction was the absolute difference divided by null log-loss. For each model and validation scheme, binary log-loss was calculated for every out-of-fold prediction, averaged across repeats within UID, and then averaged across UIDs. This is equivalent to weighting fold-level log-loss by fold test size. The prevalence-only null was aggregated identically. Predictions were scored before averaging; log-loss calculated after averaging repeated probabilities within UID was retained only as a sensitivity analysis. For calibration, predicted probabilities were first averaged across repetitions for each stable site; a binomial generalized linear model of the observed outcome on the logit of the averaged probability yielded the intercept and slope, for which ideal values are zero and one. Uncertainty in model, null, absolute-reduction and relative-reduction estimates was quantified with 5000 percentile-bootstrap resamples of UNIQUE_ID clusters, retaining every UID belonging to each sampled cluster. A marine result was considered supported only when relative log-loss improvement was at least 2%, the 95% interval for absolute improvement excluded zero under both location- and batch-grouped validation, calibration was not materially deficient, and the signal was not attributable to a single domain.

Secondary three-state model

A secondary multinomial L2-penalized logistic regression retained non-detects, numerical estimates below the lower limit of quantification, and unflagged numerical results as separate operational classes. It used $C = 0.1$, the lbfgs solver, a maximum of 600 iterations, and the same fold-wise preprocessing, predictors, ten repeated five-fold grouped validation schemes, and random seeds as the binary analysis. The null comparator used the three class proportions estimated in the training partition. Multinomial log-loss was primary; macro-averaged one-versus-rest area under the curve, balanced accuracy, and the event-level confusion matrix were secondary. Incremental log-loss intervals were obtained from 5,000 stable-site bootstrap resamples. This analysis tested whether binary grouping had created the apparent signal and whether the minority unflagged numerical class could be discriminated separately.

Batch-conditioned associations and multiplicity

Environmental associations were examined as a secondary explanatory analysis, separate from predictive performance. Within each environment, complete cases for the 12 continuous environmental covariates were standardized to one standard deviation and entered simultaneously into conditional logistic regression with enterococci analytical batch as the stratum. Only batches containing both binary outcome states contributed to estimation. Odds ratios therefore compare events within informative analytical batches and reduce confounding by batch-level differences, but they remain observational associations. False-discovery-rate adjustment used the Benjamini-Hochberg procedure across the 12 covariates separately within each environment. Batch heterogeneity in the binary outcome was also screened with a batch-by-state chi-square test. Predictor-level significance was not inferred from penalized-model coefficients, and no variable-wise search was opened after observing predictive results.

Temporal, geographic and ecosystem transportability

Temporal transfer was evaluated by fitting the fixed logistic pipeline to all eligible events from

earlier years and testing it in each later year with at least 100 training events, 20 test events and both outcome classes. Geographic transportability was examined by leaving one NCCA marine region, or one Great Lake in the freshwater contrast, completely outside training. Great Lakes analyses repeated the binary location- and batch-grouped validation as a secondary ecosystem contrast but were interpreted independently because the environmental setting, predictor distributions and analytical-state frequencies differed from marine waters. Failure in a Great Lakes or held-out regional domain restricted transportability claims rather than being averaged away through pooled modeling.

Advanced-model comparison

A constrained Light Gradient Boosting Machine (LightGBM) classifier was prespecified as the single advanced comparator. Settings were 150 trees, learning rate 0.03, seven leaves, minimum 30 observations per child, L2 regularization 2, L1 regularization 0.5, one processing thread and random seed 20260827. It was evaluated with five repetitions of the same five-fold location- and batch-grouped schemes, compared directly with the ten-repetition logistic-regression OOF results using the canonical UID-weighted aggregation, and assessed with 5,000 UNIQUE_ID-cluster bootstrap resamples. Relative improvement versus the prevalence-only null and incremental improvement versus penalized logistic regression were calculated and labeled separately. The advanced model was eligible for retention only if it reduced logistic-regression log-loss by at least 2% in both grouping schemes without materially worse calibration; percentages versus the null were descriptive and were not used for model promotion. No additional algorithms were explored to rescue an unfavorable result.

Reproducibility and analytical restrictions

The analysis was implemented in Python 3.13.5 using pandas 2.2.3, NumPy 2.3.5, scikit-learn 1.8.0, SciPy 1.17.0, statsmodels 0.14.6 and LightGBM 4.6.0. Fixed seeds, out-of-fold predictions, fold metrics, model settings, cohort classifications and file-level hashes were retained with the executable scripts. Official files remained immutable; all state classifications and predictor tables were derived copies with recorded lineage.

An anonymized Supplementary File S1 contains the analysis code, derivation rules, derived non-identifying data, locked predictions, reconciled R1 outputs, complete execution logs and package hashes. Supplementary File S2 contains the four exact official NCCA source tables used in the study, their companion metadata and a source manifest. The S1 clean-verification route checks the official hashes, reconstructs the 1317-row, 68-column analytical table, verifies value equivalence at relative and absolute tolerances of 1×10^{-12} , links reconstructed outcomes to all locked OOF predictions, and recomputes the R1 estimates without refitting the models. Because the design is observational, internally validated and unweighted, the analysis does not estimate causal effects, viable-enterococcus abundance, infection risk, operational beach-management thresholds or population-wide prevalence across coastal waters of the United States.

RESULTS

Cohort and analytical-state distribution

The marine cohort included 897 events: 303 non-detects, 458 numerical estimates below the lower limit of quantification, and 136 unflagged numerical results. Numerical results were available for 594 events (66.2%), but 458 of these (77.1%) were analytically qualified as below the quantification limit. The Great Lakes contrast included 420 events, of which only 18 (4.3%) were unflagged numerical results (Table 1).

Predictive performance in the marine cohort

The penalized environmental model outperformed the prevalence-only comparator under both grouped validation schemes. With stable locations held out, log-loss was 0.5911 versus 0.6396 for the comparator, an absolute reduction

of 0.0485 (95% bootstrap CI 0.0239–0.0721) and a relative reduction of 7.58% (95% CI 3.74–11.19%). The ROC-AUC was 0.699, the Brier score was 0.2021, and the calibration slope was 0.880. With complete analytical batches held out, log-loss was 0.5956 versus 0.6400; the absolute reduction was 0.0444 (95% CI 0.0197–0.0686) and the relative reduction was 6.94% (95% CI 3.09–10.68%). The corresponding ROC-AUC was 0.692, the Brier score was 0.2037, and the calibration slope was 0.823. These values indicate moderate discrimination and probabilistic improvement, but not calibration sufficient for operational use (Figure 1).

Three-state multinomial analysis

Explicitly retaining non-detect, below-quantification estimate, and unflagged numerical states produced the same qualitative result. Relative log-loss was reduced by 7.21% (95% CI 4.20–10.28%) with locations held out and by 5.11% (95% CI 1.67–8.53%) with analytical batches held out. Macro-averaged ROC-AUC values were 0.680 and 0.665, respectively. Row-normalized recall for the unflagged numerical class was 21.3% in location-grouped validation and 23.5% in batch-grouped validation, confirming that unflagged numerical results remained difficult to distinguish from below-quantification estimates.

Detailed multinomial metrics and row-normalized confusion matrices are provided in Supplementary Table S1 and Supplementary Figure S1.

Batch-conditioned environmental associations

The complete-case conditional analysis included 868 events across 32 informative marine batches. After false-discovery-rate control, higher surface pH was associated with lower odds of a numerical result (OR per SD 0.631, 95% CI 0.500–0.795, adjusted $p=0.0012$), as was greater log-transformed station depth (OR 0.731, 95% CI

Table 1. Distribution of enterococci analytical states in the eligible cohorts

Setting	Events	Non-detect	Non-detect, %	Below LLOQ	Below LLOQ, %	Unflagged numeric	Unflagged numeric, %
Marine	897	303	33.8	458	51.1	136	15.2
Great Lakes	420	255	60.7	147	35.0	18	4.3

Note: LLOQ, lower limit of quantification. “Unflagged numeric” denotes a numerical result with no analytical qualifier; it is an operational reporting category, not a claim of complete analytical validity.

Table 2. Binary predictive performance in the marine cohort

Validation scheme	Model log-loss	Null log-loss	Relative reduction, % (95% CI)	ROC-AUC	Brier score	Calibration slope
Stable locations held out	0.5911	0.6396	7.58 (3.74 to 11.19)	0.699	0.2021	0.88
Analytical batches held out	0.5956	0.6400	6.94 (3.09 to 10.68)	0.692	0.2037	0.823

Note: Lower log-loss and Brier scores indicate better probabilistic prediction. ROC-AUC = area under the receiver operating characteristic curve. Log-loss confidence intervals were obtained from 5,000 percentile-bootstrap resamples of UNIQUE_ID clusters, retaining all UIDs within each sampled cluster.

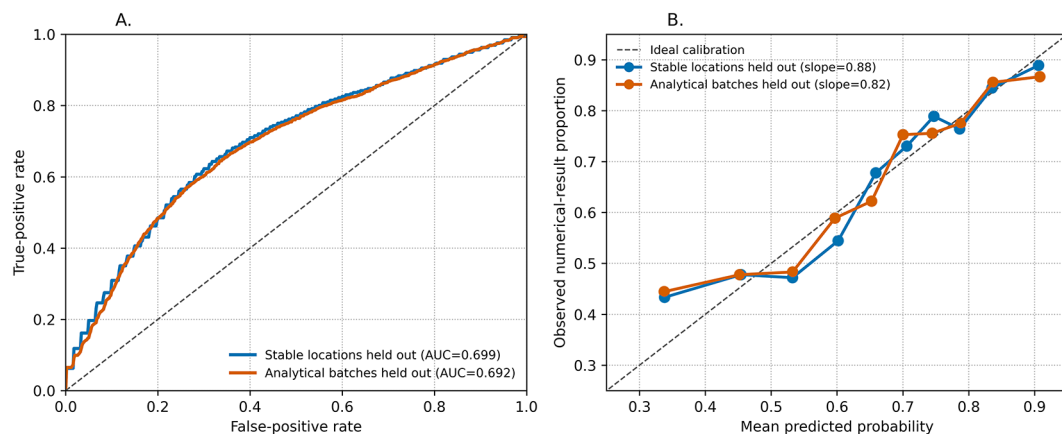


Figure 1. Discrimination and calibration of the marine binary model. Panel A shows mean receiver operating characteristic curves across the held-out folds of the ten repeated grouped-validation runs; legend AUC values are the means of the corresponding fold-level AUCs reported in Table 2. Panel B shows quantile-binned calibration using event-level probabilities averaged across repetitions. Point size in Panel B reflects the number of observations per probability bin

0.583–0.916, adjusted $p=0.0395$). These estimates are conditional observational associations and do not identify causal effects, fecal sources, viable-organism abundance, or health risk (Table 3).

Transportability and advanced-method comparison

Geographic transportability was heterogeneous. Relative log-loss reduction in leave-one-region-out analysis was 8.70% for Northeast, 2.75% for Gulf, 0.67% for West, and -14.02% for Southeast. In the Great Lakes contrast, location-grouped improvement was 1.72% (95% CI -3.29–6.51%) and batch-grouped improvement was -1.16% (95% CI -6.05–3.68%). The direct incremental LightGBM comparison with penalized logistic regression was 0.72% (95% CI -1.65–3.06%) under marine location-grouped validation and -1.26% (95% CI -4.02–1.33%) under marine batch-grouped validation. It therefore failed the prespecified 2% retention rule and was not retained (Figure 2).

DISCUSSION

Concurrent environmental conditions contained moderate and reproducible information about the analytical reporting state of enterococci in sampled marine waters. The penalized logistic model reduced log-loss by 7.58% when stable locations were held out and by 6.94% when complete analytical batches were held out, with areas under the receiver operating characteristic curve close to 0.70. These gains were measurable but not large enough to justify operational substitution for laboratory analysis. Previous microbial-water-quality models have generally focused on concentrations, threshold exceedance, or site-specific beach forecasts (Heasley et al., 2021; Cyterski et al., 2022). Rapid qPCR monitoring has also been evaluated at individual beaches (Saleem et al., 2022). The present analysis addresses a narrower measurement question: whether the environmental context helps distinguish the analytical state returned by the molecular assay after censoring is retained explicitly.

Table 3. Environmental associations with a numerical enterococci result after conditioning on analytical batch

Predictor	Events	Informative batches	Odds ratio per SD	95% confidence interval	FDR-adjusted p
Surface pH	868	32	0.631	0.500-0.795	0.0012
Station depth, log(1+x) transformed	868	32	0.731	0.583-0.916	0.0395

Note: Odds ratios are reported per one standard deviation. Only associations surviving false-discovery-rate control within the marine environment are shown; all tested associations are reported in Supplementary Table S3.

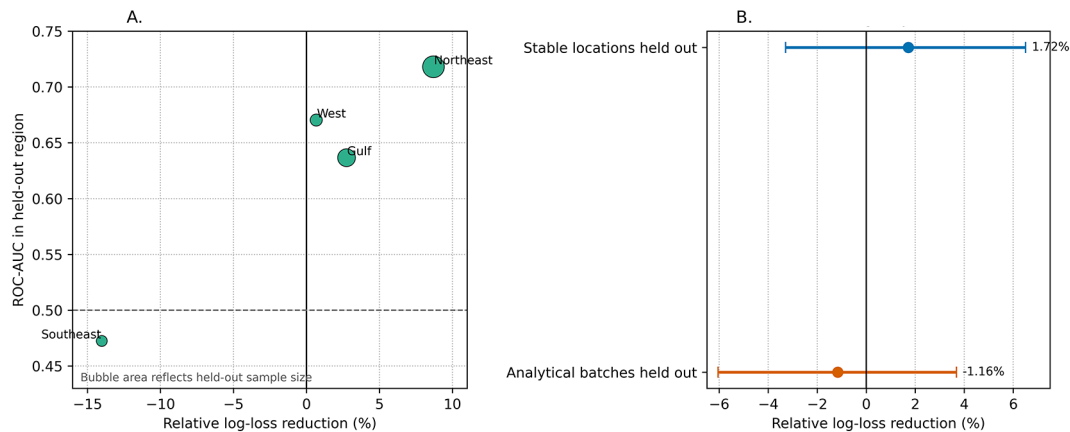


Figure 2. Geographic and ecosystem transportability. Panel A positions each held-out marine region by relative log-loss reduction and ROC-AUC; bubble area reflects held-out sample size. Panel B shows Great Lakes relative log-loss reduction with 95% UNIQUE_ID-cluster bootstrap intervals for location- and batch-grouped validation

Censoring was central rather than incidental to that question. Of the 594 numerical marine results, 458 (77.1%) were estimates below the lower limit of quantification. The three-state model preserved this distinction and retained relative log-loss reductions of 7.21% under location-held-out validation and 5.11% under batch-held-out validation. Nevertheless, it correctly identified the unflagged numerical state in only 21.3% and 23.5% of cases, respectively. Environmental information therefore distinguished non-detection from a numerical assay response more reliably than it separated below-quantification estimates from unflagged numerical results. This finding supports recommendations to report detection and quantification limits as distinct analytical properties rather than treating all numerical values as equally precise observations (Klymus et al., 2020; Borchardt et al., 2021). Calibrator cell equivalents and culture-based counts measure different constructs (Converse et al., 2012; Wymer et al., 2021). Accordingly, calibrator cell equivalents should not be interpreted as direct counts of viable organisms, colony-forming units, pathogens, or individual health risk (U.S. EPA, 2012).

The batch-held-out result is particularly important because the analytical batch identifier was used only to define validation groups and was never supplied as a predictor. The model therefore had to predict events processed in analytical batches absent from training. Retaining a 6.94% improvement under this design indicates that the environmental signal was not solely memorization of batch identity. At the same time, the batch-level state distributions and the broader multi-laboratory literature show that analytical lineage remains a substantive source of heterogeneity in enterococci quantitative polymerase chain reaction measurements (Haugland et al., 2016). Within informative batches, higher surface pH and greater station depth were associated with lower odds of a numerical result, but these conditional associations do not identify causal ecological or analytical mechanisms. Dilution, hydrodynamics, DNA persistence, matrix inhibition, source differences, and unmeasured environmental conditions are plausible explanations, but they were not distinguished by this observational design.

Geographic transportability was the main limitation to generalization. The leave-one-region-out analysis retained an 8.70% improvement

in Northeast waters, but the corresponding improvement was 2.75% in Gulf waters, 0.67% in West waters, and -14.02% in Southeast waters. The Great Lakes contrast likewise failed to show robust improvement when locations or batches were held out. Such domain dependence is consistent with the site-specific character of much of the microbial-water-quality prediction literature and with the general principle that random validation can obscure spatial, temporal, or hierarchical dependence (Roberts et al., 2017; Heasley et al., 2021). Differences in indicator prevalence, hydrodynamics, salinity, seasonality, fecal-source mixtures, and laboratory or matrix behavior could all contribute to the observed domain shift, but these explanations remain hypotheses. The marine and Great Lakes results should consequently be interpreted as separate evaluations rather than pooled evidence for a universal model.

Algorithmic complexity did not resolve these limitations. The direct LightGBM comparison with penalized logistic regression showed an improvement of only 0.72% (95% CI -1.65–3.06%) when locations were held out and an improvement of -1.26% (95% CI -4.02–1.33%) when complete analytical batches were held out. It therefore failed the prespecified retention rule. Ensemble and other machine-learning methods can improve beach-water-quality forecasts in some settings (Wang et al., 2021), but their value depends on the prediction target, the available sample size, the validation structure, and the intended domain. Here, the simpler penalized model offered comparable or better transportability and clearer interpretation. The scientific contribution is therefore not the use of a more complex algorithm, but the explicit modeling of analytical states and the use of validation schemes aligned with location and laboratory lineage.

These results have methodological implications for environmental monitoring studies that use molecular indicators. Non-detects, below-quantification estimates, and unflagged numerical results should be preserved as distinct states when their analytical meaning differs; batch identifiers should be retained for validation and quality assessment; and model performance should be reported for new locations and new analytical batches rather than only for randomly partitioned records. The present model may help characterize how environmental context relates to assay reporting states and may support future quality-assurance research, but it is not a replacement for

microbiological measurement or a basis for beach closure, public-health advice, or site classification. Prospective evaluation with new sampling campaigns and prespecified local domains would be required before any operational use.

Strengths and limitations

This study has several strengths. It used record-level observations from a standardized multi-regional monitoring program, reconstructed analytical states from the numerical result and quality flags, retained all eligible outcomes rather than applying complete-case exclusion, and estimated imputation and scaling parameters only within training folds. Stable locations and complete laboratory batches were evaluated separately, the three-state analysis checked whether the binary result depended on outcome collapse, and an advanced method was rejected when it failed the prespecified transportability criterion. Negative regional and ecosystem results were retained rather than used to redefine the primary population after analysis.

The observational and concurrent design does not establish that any environmental variable caused an enterococci analytical state. The target was the laboratory reporting state, not viable-cell abundance, fecal-source attribution, pathogen occurrence, or infection risk. Although the analysis accounted for measured water chemistry, hydrographic conditions, season, region, year, location, and analytical batch, the delivered files did not provide every potentially relevant meteorological, hydrodynamic, source-tracking, matrix-inhibition, or laboratory-process variable in a harmonized event-level form. Residual confounding and measurement heterogeneity are therefore expected. The two batch-conditioned associations should be viewed as descriptive adjusted associations requiring independent confirmation.

This study evaluates internal transportability within a single national assessment program and cycle and therefore does not constitute independent external validation. Repeated visits were grouped by stable location and entire analytical batches were held out, but all observations came from the same assessment cycle and analytical program. The negative Southeast result and the absence of a robust confirmatory signal in the Great Lakes analysis restrict geographic generalization; independent validation across monitoring programs, assessment cycles, and regions is required. Class

imbalance further limited the three-state model, especially in Great Lakes events, where only 18 results were unflagged numerical results. Calibration was acceptable for comparative research but not sufficient for operational deployment, and no decision threshold was developed or validated.

Finally, the supplied data did not contain the survey weights, inclusion probabilities, strata, and variance variables required for population-level coastal estimates. Results therefore describe associations and predictive performance among the sampled and analytically eligible events, not the prevalence of analytical states across all United States coastal waters. The auxiliary site-metadata file was not required for inclusion because it was unnecessary for the selected predictors and would have introduced avoidable attrition. Prospective evaluation in independently collected marine data, with preserved analytical-batch lineage and detailed inhibition and quantification controls, would be required before operational use. Domain-specific recalibration should be evaluated without obscuring failure in new regions.

CONCLUSIONS

Concurrent environmental conditions provided moderate predictive information about enterococci analytical states in 897 sampled marine events. Improvement over a prevalence-only comparator persisted when stable locations and complete analytical batches were held out and when non-detect, below-quantification, and unflagged numerical states were retained separately. The model was less successful at identifying the unflagged numerical state, performance varied sharply among marine regions, and the signal did not transport robustly to Great Lakes events. Penalized logistic regression was retained because the prespecified boosting model did not improve batch-held-out performance. Overall, the evidence supports moderate event-level analytical-state predictability under grouped internal validation, with limited geographic and ecosystem transportability.

Acknowledgements

The U.S. Environmental Protection Agency National Aquatic Resource Surveys program is acknowledged for making the National Coastal Condition Assessment 2020 record-level data and documentation publicly available.

REFERENCES

1. Borchardt, M. A., Boehm, A. B., Salit, M., Spencer, S. K., Wigginton, K. R., Noble, R. T. (2021). The environmental microbiology minimum information (EMMI) guidelines: qPCR and dPCR quality and reporting for environmental microbiology. *Environmental Science & Technology*, 55(15), 10210–10223. <https://doi.org/10.1021/acs.est.1c01767>
2. Converse, R. R., Griffith, J. F., Noble, R. T., Haugland, R. A., Schiff, K. C., Weisberg, S. B. (2012). Correlation between quantitative PCR and culture-based methods for measuring *Enterococcus* spp. over various temporal scales at three California marine beaches. *Applied and Environmental Microbiology*, 78(4), 1237–1242. <https://doi.org/10.1128/AEM.07136-11>
3. Cyterski, M., Shanks, O. C., Wanjugi, P., Mc-Minn, B., Korajkic, A., Oshima, K., Haugland, R. (2022). Bacterial and viral fecal indicator predictive modeling at three Great Lakes recreational beach sites. *Water Research*, 223, 118970. <https://doi.org/10.1016/j.watres.2022.118970>
4. Haugland, R. A., Siefring, S., Varma, M., Oshima, K. H., Sivaganesan, M., Cao, Y., Raith, M., Griffith, J., Weisberg, S. B., Noble, R. T., Blackwood, A. D., Kinzelman, J., Anan'eva, T., Bushon, R. N., Stelzer, E. A., Harwood, V. J., Gordon, K. V., Sinigalliano, C. (2016). Multi-laboratory survey of qPCR enterococci analysis method performance in U.S. coastal and inland surface waters. *Journal of Microbiological Methods*, 123, 114–125. <https://doi.org/10.1016/j.mimet.2016.01.017>
5. Heasley, C., Sanchez, J. J., Tustin, J., Young, I. (2021). Systematic review of predictive models of microbial water quality at freshwater recreational beaches. *PLOS ONE*, 16(8), e0256785. <https://doi.org/10.1371/journal.pone.0256785>
6. Klymus, K. E., Merkes, C. M., Allison, M. J., Goldberg, C. S., Helbing, C. C., Hunter, M. E., Jackson, C. A., Lance, R. F., Mangan, A. M., Monroe, E. M., Piaggio, A. J., Stokdyk, J. P., Wilson, C. C., Richter, C. A. (2020). Reporting the limits of detection and quantification for environmental DNA assays. *Environmental DNA*, 2(3), 271–282. <https://doi.org/10.1002/edn3.29>
7. Nahlik, A. M., Paulsen, S. G., Dumelle, M., Holdsworth, S., Lehmann, S., Tulve, N. S., Paul, S. J., Frey, H. C. (2025). National Aquatic Resource Surveys (NARS): The foundation for long-term aquatic monitoring data across the United States. *Environmental Monitoring and Assessment*, 197(12), 1291. <https://doi.org/10.1007/s10661-025-14629-8>
8. Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Aroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., Dormann, C. F. (2017). Cross-validation strategies for data

- with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8), 913–929. <https://doi.org/10.1111/ecog.02881>
9. Saleem, F., Edge, T. A., Schellhorn, H. E. (2022). Validation of qPCR method for enterococci quantification at Toronto beaches: Application for rapid recreational water monitoring. *Journal of Great Lakes Research*, 48(3), 707–716. <https://doi.org/10.1016/j.jglr.2022.02.008>
 10. U.S. Environmental Protection Agency. (2012). Recreational water quality criteria (EPA 820-F-12-058). Office of Water, U.S. Environmental Protection Agency. <https://www.epa.gov/sites/default/files/2015-10/documents/rwqc2012.pdf>
 11. U.S. Environmental Protection Agency. (2015). Method 1609.1: Enterococci in water by TaqMan quantitative polymerase chain reaction (qPCR) with internal amplification control (IAC) assay (EPA 820-R-15-099). *Office of Water, U.S. Environmental Protection Agency*. <https://www.epa.gov/cwa-methods/other-clean-water-act-test-methods-microbiological>
 12. U.S. Environmental Protection Agency. (2025a). *National Coastal Condition Assessment 2020: Laboratory operations manual (Version 1.3)* (EPA 841-F-19-004). Office of Water, U.S. Environmental Protection Agency. <https://www.epa.gov/national-aquatic-resource-surveys/national-coastal-condition-assessment-2020-manuals>
 13. U.S. Environmental Protection Agency. (2025b). *National Coastal Condition Assessment 2020: Quality assurance project plan (Version 1.3)* (EPA 841-F-19-003). Office of Water, U.S. Environmental Protection Agency. <https://www.epa.gov/national-aquatic-resource-surveys/national-coastal-condition-assessment-2020-manuals>
 14. Wang, L., Zhu, Z., Sassoubre, L., Yu, G., Liao, C., Hu, Q., Wang, Y. (2021). Improving the robustness of beach water quality modeling using an ensemble machine learning approach. *Science of the Total Environment*, 765, 142760. <https://doi.org/10.1016/j.scitotenv.2020.142760>
 15. Wymer, L. J., Wade, T. J., Sams, E. A., Oshima, K. H., Dufour, A. P. (2021). Comparative stability of assay results of enterococci measured by culture and qPCR over time in bathing beach waters. *Journal of Microbiological Methods*, 188, 106274. <https://doi.org/10.1016/j.mimet.2021.106274>