HHS Public Access

Author manuscript

ACS ES T Water. Author manuscript; available in PMC 2024 December 27.

Published in final edited form as:

ACS ES T Water. 2024 November 27; 4(12): 5844–5855. doi:10.1021/acsestwater.4c00839.

Performance of Conditional Random Forest and Regression Models at Predicting Human Fecal Contamination of Produce Irrigation Ponds in the Southeastern United States

Jessica Hofstetter[‖],
Waterborne Disease Prevention Branch, Centers for Disease Control and Prevention, Atlanta, Georgia 30333, United States; Chenega Enterprise Systems & Solutions, LLC, Chesapeake, Virginia 23320, United States; Department of Horticulture, Auburn University, Auburn, Alabama 36849, United States

David A. Holcomb[‖],
Waterborne Disease Prevention Branch, Centers for Disease Control and Prevention, Atlanta, Georgia 30333, United States

Amy M. Kahler,
Waterborne Disease Prevention Branch, Centers for Disease Control and Prevention, Atlanta, Georgia 30333, United States

Camila Rodrigues,
Department of Horticulture, Auburn University, Auburn, Alabama 36849, United States

Andre Luiz Biscaia Ribeiro da Silva,
Department of Horticulture, Auburn University, Auburn, Alabama 36849, United States

Mia C. Mattioli
Waterborne Disease Prevention Branch, Centers for Disease Control and Prevention, Atlanta, Georgia 30333, United States

Abstract

Corresponding Author: Mia C. Mattioli – Waterborne Disease Prevention Branch, Centers for Disease Control and Prevention, Atlanta, Georgia 30333, United States; mmattioli@cdc.gov.
[‖]Jessica Hofstetter and David A. Holcomb contributed equally and should be designated co-first authors.

Author Contributions

CRediT: **Jessica Hofstetter** data curation, formal analysis, investigation, methodology, project administration, software, supervision, writing - original draft, writing - review & editing; **David A. Holcomb** formal analysis, methodology, software, validation, visualization, writing - original draft, writing - review & editing; **Amy M. Kahler** conceptualization, data curation, investigation, methodology, project administration, writing - review & editing; **Camila Rodrigues** project administration, supervision, writing - review & editing; **Andre Luiz Biscaia Ribeiro da Silva** conceptualization, funding acquisition, methodology, project administration, supervision; **Mia C. Mattioli** conceptualization, funding acquisition, investigation, methodology, project administration, resources, supervision, writing - review & editing.

Supporting Information

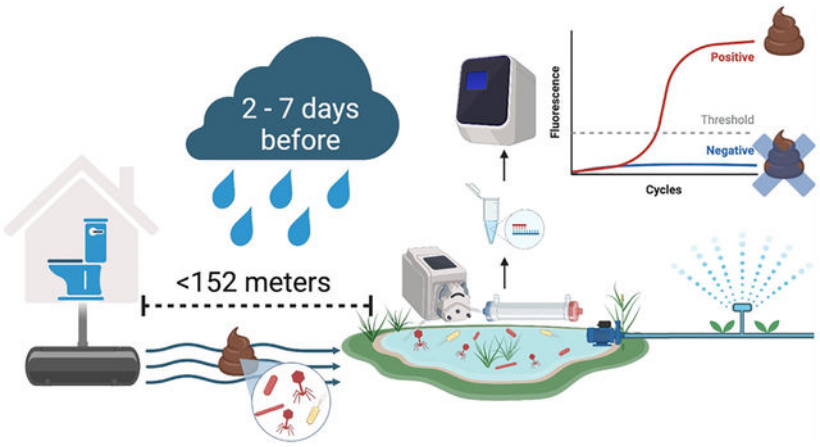
The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsestwater.4c00839.
Additional building proximity and septic system details; descriptive statistics; univariable odds ratios; predictions of human fecal contamination using *E. coli* as explanatory variable; and hyperparameter tuning and imbalance correction sensitivity analyses (PDF)

Complete contact information is available at: https://pubs.acs.org/10.1021/acsestwater.4c00839

The authors declare no competing financial interest.

Irrigating fresh produce with contaminated water contributes to the burden of foodborne illness. Identifying fecal contamination of irrigation waters and characterizing fecal sources and associated environmental factors can help inform fresh produce safety and health hazard management. Using two previously collected data sets, we developed and evaluated the performance of logistic regression and conditional random forest models for predicting general and human-specific fecal contamination of ponds in southwest Georgia used for fresh produce irrigation. Generic *Escherichia coli* served as a general fecal indicator, and human-associated *Bacteroides* (HF183), crAssphage, and F+ coliphage genogroup II were used as indicators of human fecal contamination. Increased rainfall in the previous 7 days and the presence of a building within 152 m (a proxy for proximity to septic systems) were associated with increased odds of human fecal contamination in the training data set. However, the models did not accurately predict the presence of human-associated fecal indicators in a second data set collected from nearby irrigation ponds in different years. Predictive statistical models should be used with caution to assess produce irrigation water quality as models may not reliably predict fecal contamination at other locations and times, even within the same growing region.

Graphical Abstract



Keywords

microbial source tracking; quantitative polymerase chain reaction (qPCR); dead-end ultrafiltration (DEUF); predictive modeling; conditional random forest; agricultural water; fresh produce safety; foodborne illness

INTRODUCTION

The United States Interagency Food Safety Analytics Collaboration (IFSAC) estimates that among the 1,322 foodborne outbreaks between 1998 and 2021, produce was the vehicle for 43% of foodborne illnesses from *Salmonella*, 52% of *Listeria monocytogenes* illnesses, and 67% of *Escherichia coli* O157 illnesses.[1,2] Preharvest application of poor microbial-quality water is one way that fruits and vegetables can become contaminated with foodborne pathogens.[3] Surface water is more likely than groundwater to be exposed to fecal contamination from humans and animals and may pose a greater risk to human health

when used for irrigation.[3] As such, an important component of fresh produce safety hazard management is the ability to identify times when irrigation water may be contaminated and the sources and factors contributing to contamination. One of the most widely used methods for evaluating microbial water quality is measuring generic *E. coli* as a general indicator of fecal contamination.[4] However, the utility of this fecal indicator in untreated irrigation water for fresh produce production is debated. The US Environmental Protection Agency (EPA) has recommended threshold values for generic *E. coli* levels, such as a geometric mean concentration ≥126 *E. coli* per 100 mL, to identify impaired microbial water quality in surface water used for recreation.[5] *E. coli* concentrations exceeding these thresholds are associated with higher rates of illness among swimmers. It has also been suggested that these thresholds be applied to irrigation waters,[6] but generic *E. coli* levels are not consistently associated with pathogen presence in irrigation water,[7] and several pathogens that cause significant human foodborne illness, such as *Salmonella* and pathogenic *E. coli*, have been detected in irrigation water sources even when generic *E. coli* was not detected or levels were below the EPA recreational water quality thresholds.[8,9]

Generic *E. coli* can arise from many animals and other aquatic sources, which limits its use for characterizing of fecal sources.[10] Many foodborne illnesses associated with produce, such as norovirus GI, GII, and GIV; hepatitis A types I, II, and III; hepatitis E types 1–4 and 7; and the parasite *Cyclospora cayetanensis*, are solely associated with human contamination.[11–14] These pathogens have been found in water impacted by human fecal contamination and subsequently in produce grown using these water sources.[15,16] This highlights the importance of characterizing human-specific fecal contamination in irrigation water for remediation and the mitigation of health risks.

Testing produce irrigation water for microbial source tracking (MST) markers is a strategy for determining fecal contamination sources.[17] Previous studies have highlighted the importance of considering multiple MST markers to account for differences in marker decay rates and abundance in the host feces, particularly when low levels of contamination are suspected.[18–20] Molecular assays that target gene sequences from *Bacteroides* in human feces (e.g., HF183) have been developed and widely implemented as MST markers to infer the presence of human fecal contamination in environmental samples.[21–25] CrAssphage, a recently identified virus of *Bacteroides* that is abundant in human feces, has also been used as a sensitive and human-specific MST marker.[26,27] As a virus, crAssphage is more biologically similar to human-specific enteric viral pathogens than bacterial fecal indicators and has been associated with enteric viruses in environmental waters.[28] The associations between *E. coli* levels and the presence of HF183 or crAssphage in surface waters reported previously have been inconsistent and limited, and the environmental factors associated with generic *E. coli* levels differ from those associated with HF183 occurrence.[28–32] Male-specific (F+) coliphages, which infect coliform bacteria like *E. coli*, have also been widely used as fecal indicator viruses and can be detected in environmental samples by conventional culture methods.[33] Although F+ coliphages in general are not host-specific, F+ RNA (FRNA) coliphage genogroup II (GII) has been associated primarily with human feces and used as a human MST marker.[34–36]

Previous water quality modeling studies have found associations between microbial contamination measured by *E. coli* and HF183 and various environmental factors, including pH, conductivity, turbidity, season, temperature, precipitation, and land use.[8,37–42] However, the specific factors associated with fecal contamination varied between studies, water source type, and location. Regression modeling has previously been used to determine significant factors for predicting human fecal contamination in ambient recreational water,[40] while recent advances in machine learning modeling approaches have been applied to predicting pathogens in certain agricultural settings.[43] The environmental factors identified as driving contamination often vary based on modeling approach. For example, regression and machine learning models previously identified different explanatory variables as important predictors of *Salmonella* and enterohemorrhagic *E. coli* markers in irrigation water.[44]

Tools to identify fecal contamination and characterize fecal sources and associated environmental factors in irrigation water could help growers manage hazards for fresh produce safety. The discrepancies in the apparent drivers of contamination identified in the literature suggest that conducting agricultural setting-specific water quality assessments that consider multiple microbial targets, environmental factors, and modeling approaches may be necessary to adequately characterize microbial hazards in irrigation water. Previous models of human fecal contamination in US irrigation waters have focused on predicting the presence of bacterial MST markers (e.g., HF183) in streams;[45] to our knowledge, comparable models for viral markers like crAssphage and for nonflowing water sources like irrigation ponds have not previously been reported. In this study, we developed models to evaluate environmental factors associated with the detection of four fecal markers in irrigation ponds in the southeast United States: generic *E. coli*, HF183, crAssphage, and FRNA GII coliphage. We then tested the predictive performance of the models on a separate data set that had been collected previously from the same growing region.

MATERIALS AND METHODS

Study Area.

The data used in this study were collected from irrigation ponds on farms in southwest Georgia. Sites were located in a region with subtropical environmental conditions characterized by coarse-textured and well-drained soils used for agriculture, pasture, and mixed forests.[46] The ponds used for produce irrigation were located in the Little River watershed in the headwaters of the Suwannee River basin.[47] The irrigation ponds in the test data set were located within 0.5 to 10 miles (0.8–16 km) of the ponds sampled for the training data set. All ponds in the training data set and two of the three ponds in the test data set were reported to be surface water-fed, while one pond in the test data set was groundwater-fed.

Training Data Set.

The training data set was collected as part of a study monitoring the occurrence of *C. cayetanensis* in produce irrigation water.[48] Large-volume pond water samples (50 L) were collected by dead-end ultrafiltration (DEUF) from eight ponds serving two growers (A and B) one or two times per month from September 2020 through December 2021.

Generic *E. coli* was enumerated from 100 mL grab samples collected alongside the DEUF samples within six hours of collection using the IDEXX Colilert-18 Quanti-Tray 2000 method (IDEXX Laboratories, Westbrooke, ME). DEUF samples were shipped on ice to the US Centers for Disease Control and Prevention (CDC) to be backflushed and further concentrated by centrifugation (4000*g* for 15 min) within 48 h of collection, as previously described.[48,49]

DNA was extracted from 200 μL of DEUF concentrates using the Qiagen AllPrep PowerViral DNA/RNA Kit (Qiagen, Hilden, Germany). Isolated DNA was immediately subjected to molecular analysis. Detailed sample processing and molecular analysis methods have been described previously.[48] Briefly, human-associated genetic markers were amplified by quantitative polymerase chain reaction (qPCR) in triplicate following EPA Method 1696 to detect HF183 and the CPQ_056 assay to detect crAssphage.[26,48,50] HF183 and crAssphage were considered to have been detected in a sample when two of the three qPCR replicates for a given assay with demonstrated amplification above a threshold of 0.03 ΔRn and a quantification cycle (C_q) value below 40.[48]

Test Data Set.

The test data set consisted of samples collected from three additional ponds between May 2015 and May 2016 as part of a study to evaluate large-volume sample collection for characterizing foodborne pathogens and indicators in irrigation water.[51] During each sampling visit, 1 L grab samples and 50 L DEUF samples were collected at two pond-edge locations per pond. Generic *E. coli* concentrations were measured separately in each grab sample by the same Colilert-18 method and averaged for further analysis. The two DEUF samples collected per pond were separately backflushed using the same procedure as for the training set ultrafilters. Prior to secondary concentration, the backflush was analyzed for male-specific (F+) coliphage viruses using the EPA Single Agar Layer (SAL) method,[52] followed by F+ RNA coliphage (FRNA) genotyping as described elsewhere.[51,53] Half the remaining backflush volume was further concentrated by poly(ethylene glycol) (PEG) precipitation and centrifugation at 10,000*g* for 30 min,[54] and the other half was concentrated by centrifugation alone (4000*g* for 30 min), yielding four ultrafilter concentrates per pond per sampling visit. Nucleic acid was extracted from 750 μl of each concentrate by the Universal Nucleic Acid Extraction (UNEX) method.[55] Each concentrate was analyzed for HF183 by qPCR as described previously.[22,51] HF183 was determined to have been detected in a sample if two qPCR replicates from either DEUF sample showed amplification before the C_q value of 40.

Though produced using somewhat different workflows, the training and test data sets both ultimately consisted of single observations of the generic *E. coli* concentration, the presence of human-associated bacteria (HF183), and the presence of a human-associated virus in each pond for each sampling event. For both data sets, generic *E. coli* was measured in grab samples by culture, and HF183 was detected by qPCR in large-volume water samples processed by DEUF with secondary concentration by centrifugation. The human-associated virus assessed for the training data set, crAssphage, was detected by qPCR in centrifuge-concentrated DEUF sample backflush. For the test data set, FRNA GII coliphage was

assessed as the human-associated virus in DEUF sample backflush (prior to any secondary concentration) using culture methods coupled with qPCR-based genotyping.

Environmental Explanatory Variables.

Water quality parameters, including dissolved oxygen (mg/L), turbidity (NTU), pH, conductivity (μS/cm), and temperature (°C), were measured using a ProDSS Multiparameter Digital Water Quality Meter (YSI, Yellow Springs, OH) at the time of sample collection for both the training and test data sets. Negative turbidity measurements were censored at 0 NTU for analysis. Each parameter was measured four times over approximately 30 min during the training set collection, and the measurements were averaged, as described previously.[48]

For the training data set, daily rainfall accumulation (inches) was collected using Rain101A Rainfall Data Loggers (MadgeTech, Inc., Warner, NH) or WatchDog 1120 Data Logging Rain Gauges (Spectrum Technologies, Inc. Aurora, IL) placed at each pond. Rainfall data were intermittently unavailable at individual ponds due to equipment failures. However, rainfall data were successfully collected from at least one of the four Grower A pond gauges during the entire study period. Since all Grower A ponds were located within a three-mile radius, data from all working gauges were averaged each day to create a complete data set for the study period. Rain gauge malfunctions occurred at two of the four Grower B ponds, but as these ponds were located less than 1 mile apart, their rain data were merged or averaged if data from both were available. Test data set rainfall measurements were retrieved from a University of Georgia-managed publicly accessible weather logging system stationed within 10 miles (16 km) of all study ponds.[56]

Daily average wind speed (miles per hour [mph]) and daily solar radiation (MJ/m^2) were obtained for both data sets from the US Department of Agriculture (USDA) Soil Climate Analysis Network monitoring station located within 15 miles (24 km) of all of the ponds.[57] Rainfall, wind, and solar radiation data were each aggregated into two variables: accumulation within the previous 2 days (Rain 0–2, Wind 0–2, and Solar 0–2) and in the previous 2-to-7 days (Rain 2–7, Wind 2–7, and Solar 2–7). These categories were constructed to represent more recent events and events occurring further in the past, respectively, a distinction shown to be meaningful in previous predictive models of fecal contamination in surface water.[58,59] The International Organization for Standardization (ISO) week of sample collection was also included as a continuous explanatory variable to account for recurring temporal patterns.

Proximity of the ponds to septic systems was considered a potential source of human fecal contamination in this region. Because public records of septic installations were incomplete, we used the proximity to a building as a proxy for the proximity to potential septic pollution sources. Due to the rural setting, it was likely that any buildings were served by septic systems. The Georgia Department of Public Health requires an absorption field area of 500 $feet^2$ for a two-bedroom house with a residential trench septic system and a 2190 ft^2 absorption field area for commercial buildings.[60] Therefore, a pond was classified as “close” to a building if it was located within 2000 ft (610 m) of any commercial building or within 500 ft (152 m) of all other building types, regardless of septic record. We ascertained

building proximity from satellite imagery in Google Maps (maps.google.com). A detailed examination of building proximity and septic records near the ponds is provided in the Supporting Information (SI).

Descriptive Statistical Analysis.

Conductivity and the two cumulative solar radiation variables (Solar 0–2 and Solar 2–7) were $\log_{10}$-transformed prior to statistical analyses. Turbidity was also $\log_{10}$-transformed after adding one to each measurement to address zero values. Samples with *E. coli* most probable number (MPN) concentrations ≥126 MPN/100 mL were classified as having elevated generic *E. coli* levels based on the EPA recreational water guidance and previous irrigation water models that predicted pathogenic *E. coli* gene occurrence using a 126 MPN/100 mL generic *E. coli* threshold.[43] Associations between the frequency of elevated generic *E. coli* ≥126 MPN/100 mL and frequency of HF183 detection and between HF183 and crAssphage detection frequencies were assessed using a Cochran–Mantel–Haenszel (CMH) test stratified by pond. Pearson correlation analysis was conducted to assess pairwise correlations between all environmental explanatory variables. All analyses were conducted using R version 4.4.0.[61] Analysis code and study data are available at https://cdcgov.github.io/WDPB_EMEL/manuscripts/irrigation_models/.

Model Development.

Logistic regression and conditional random forest (CRF) models were developed to predict the detection of HF183, detection of any human fecal indicator (HFI), and elevated generic *E. coli* ≥126 MPN/100 mL. HFI detection was defined as detecting either or both HF183 and a human-associated virus. For the training data set, the HFI variable used crAssphage detection as the second indicator of human fecal contamination. Detection of FRNA GII coliphage was used as the human-associated viral indicator to define the HFI variable in the test data set.[34–36]

The same set of explanatory variables was considered for both the logistic regression and CRF models; the final variable sets were selected separately for each modeling approach and outcome (HF183, HFI, and generic *E. coli* ≥ 126 MPN/100 mL). The full explanatory variable set evaluated included: ISO week of sample collection; water sample temperature, dissolved oxygen, pH, $\log_{10}$ conductivity, and $\log_{10}$ turbidity; cumulative rain, wind, and $\log_{10}$ solar radiation in the previous 0–2 days (Rain 0–2, Wind 0–2, and Solar 0–2) and previous 2–7 days (Rain 2–7, Wind 2–7, and Solar 2–7); and a binary variable indicating building proximity. To limit colinearity, explanatory variables with pairwise correlation absolute value ≥0.5 were not included in the same model.

Model training and tuning, including variable selection, were conducted using the training data set. The final trained models were then applied to the test data set to evaluate out-of-sample predictive performance. Models for the detection of HF183 and HFI were also retrained including a binary variable indicating *E. coli* ≥ 126 MPN/100 mL as an additional explanatory variable to evaluate whether elevated generic *E. coli* levels were predictive of human fecal contamination.

Training Regression Models.—Mixed-effects logistic regression models were implemented with the *lme4* package in R and included the pond of sample collection as a random effect to account for repeated measures.[62] Variable selection for the logistic regression models proceeded in two stages. First, univariable associations were evaluated in separate models for each binary outcome variable (detection of HF183, HFI detection, and generic *E. coli* ≥ 126 MPN/100 mL) and each explanatory variable. Explanatory variables with p-value <0.1 in the univariable models were considered for inclusion in multivariable models. Second, backward stepwise selection was performed to select the explanatory variables to retain in the multivariable mixed-effects logistic regression models. After specifying the full model with all variables retained from the univariable models, the explanatory variable with the highest p-value was removed and the full and reduced models were compared using a chi-squared test with one degree of freedom.[62] A nonsignificant chi-squared test at the 10% significance level indicated that the full model did not meaningfully reduce the deviance and was used as the decision criterion in favor of the simpler model. The procedure was repeated until a significant chi-square test was obtained. Due to convergence issues for models predicting generic *E. coli* ≥ 126 MPN/100 mL, a forward stepwise selection procedure was used instead, beginning with an intercept-only model (including the pond random effect) and adding variables until a nonsignificant chi-squared test was obtained.

Training Conditional Random Forest Models.—CRF models build on the advantages of random forest analysis, including the ability to explore complex and nonlinear interactions between numerous explanatory variables without needing to prespecify the model structure, by incorporating conditional inference approaches to mitigate the overfitting and bias toward correlated variables exhibited by conventional random forest.[63–65] CRF models from the *party* package were developed using the *mlr* package framework in R.[63,66] Models were trained using 10,001 conditional inference trees and the default hyperparameter values suggested for unbiased variable selection.[64] As a sensitivity analysis, we also constructed CRF models with hyperparameter values for the number of explanatory variables randomly considered for splitting each node (“mtry”) and the minimum number of observations to construct a terminal node (“minbucket”) tuned by maximizing the mean area under the receiver operating characteristic curve (AUC) using repeated 3-fold cross-validation (five iterations).[44,65] Synthetic minority oversampling technique (SMOTE) was implemented during hyperparameter tuning for an additional set of models to address class imbalance of the three binary outcome variables as an additional sensitivity analysis.[67,68] While resampling-based imbalance corrections have been reported to improve the predictive accuracy of previous CRF models of foodborne pathogen presence in water,[69] the practice has been criticized for producing poorly calibrated probabilistic predictions with inconsistent impacts on classification performance.[70–72] We assessed variable importance as the independent impact of each variable on the AUC using a conditional permutation approach to address potential bias from correlated explanatory variables and outcome variable class imbalance.[73,74]

Predictive Performance.

The trained logistic regression and CRF models were applied to the test data set to generate predicted probabilities for the detection of HF183, detection of HFI, and generic *E. coli* ≥ 126 MPN/100 mL in different ponds from the same growing region. Predictive performance was assessed by receiver operating characteristic (ROC) curve analysis using the *pROC* package in R.[75] For consistency with the CRF variable importance procedure, which utilized AUC as a less-biased alternative to the traditional accuracy metric for determining variable importance,[74] we estimated the area under the ROC curve as a dimensionless metric of the overall ability of each model to discriminate between the presence and absence of the outcome.[76] An AUC of 1 denotes perfect concordance between predicted and observed outcome values, indicating ideal model performance, and an AUC of 0.5 corresponds to model classification performance equivalent to random chance.[77,78] We also calculated predictive sensitivity (the proportion of test samples positive for the outcome correctly predicted to be positive by the model) and specificity (the proportion of test samples negative for the outcome correctly predicted to be negative) at model-specific classification thresholds identified by maximizing Youden's *J* statistic.[79] The classification threshold is the minimum predicted probability of the outcome required to classify a sample as positive; increasing the threshold generally increases specificity (i.e., reduces the false positive rate) at the expense of decreasing the sensitivity (the true positive rate). The threshold that maximizes *J* balances sensitivity and specificity by minimizing the overall proportion of misclassified samples, weighting false positives and false negatives equally.

RESULTS

Training Data Set.

Of the 217 training data set samples, HF183 was detected in 71 (33%) water samples, crAssphage was detected in 14 (7%) samples, and these two human-associated markers were codetected in 10 (5%) samples (Table 1). HF183 was detected in ≥25% of the samples from ponds A1, A2, A3, A4, and B4, all of which were considered near buildings (<610 m from a commercial or <152 m from any other building). Likewise, all ponds in which crAssphage was detected were near buildings. CrAssphage detections were significantly associated with HF183 detections (CMH $\chi^2_{df=1} = 5.11$, $p = 0.02$). Generic *E. coli* exceeded 126 MPN/100 mL at least once during the sampling period in every pond (4–22% of samples per pond, Table 1). Elevated *E. coli* ≥126 MPN/100 mL were not associated with HF183 detection (CMH $\chi^2_{df=1} = 1.82$, $p = 0.18$). Descriptive statistics of explanatory variables are summarized for each pond in Table S1, and pairwise Pearson correlation coefficients are presented in Figure S1.

Test Data Set.

All ponds in the test data set were located within 500 ft (152 m) of a building or 2000 ft (610 m) of a commercial building. HF183 was detected in about a third of the samples from each pond (Table 1). Human-associated FRNA GII coliphage was detected less frequently than HF183 but at a similar frequency to crAssphage in the training data set. HF183 and FRNA GII coliphage were codetected in 5 (8%) samples. Similarly, generic *E. coli* levels

were ≥126 MPN/100 mL at comparable frequencies in both the training (10%) and test (9%) data sets.

Models.

Human Fecal Indicators.—Results of univariable logistic regression models used to inform explanatory variable selection are presented in Figure S2. Following backward stepwise variable selection, building presence, cumulative rainfall in the previous 0–2 days, and cumulative rainfall in the previous 2–7 days were retained in the final multivariable logistic regression models for both HF183 detection and HFI detection. The presence of a building was associated with elevated odds of HF183 and HFI (HF183 odds ratio [OR]: 24.8, 95% confidence interval [CI]: 3.6–172.5; HFI OR: 28.6, 95%CI: 4.4–187.1; Figure 1). Rainfall was also positively associated with HFI presence. The odds of detection approximately doubled for each additional inch (2.5 cm) of rain 0–2 days before sample collection for both HF183 (OR: 2.0, 95%CI: 0.99–4.2) and HFI (OR: 2.1, 95%CI: 1.0–4.5). An additional inch of rain 2–7 days before sample collection was associated with a 70% increase in the odds of detecting both HF183 (OR: 1.7, 95%CI: 1.2–2.5) and HFI (OR: 1.7, 95%CI: 1.2–2.5). Although the magnitude of the estimated associations was lower for rain 2–7 days prior, the relationships were more precise than the larger associations estimated for rain in the previous 0–2 days, which also included the null. Similarly, the top two ranked explanatory variables by variable importance in the CRF models for the HF183 and HFI were the presence of a building and rainfall in the previous 2–7 days (Figure 2). All other variables had negligible importance values. Generic *E. coli* ≥ 126 MPN/100 mL was not significant when included as an additional explanatory variable in logistic regression models (HF183 OR: 1.1, 95%CI: 0.26–5.0; HFI OR: 0.77, 95%CI: 0.18–3.4) and was of negligible variable importance in CRF models.

Generic E. coli.—Rainfall and solar radiation in the previous 0–2 days were the only variables retained in the multivariable logistic regression model for elevated generic *E. coli*. Rainfall in the previous 0–2 days was associated with increased odds of generic *E. coli* ≥126 MPN/100 mL (OR: 6.7, 95%CI: 2.6–17.6; Figure 1). Conversely, a $\log_{10}$-increase in solar radiation 0–2 days prior was associated with lower odds of generic *E. coli* ≥ 126 MPN/100 mL (OR: 0.12, 95%CI: 0.01–1.1), although the association was not significant. CRF analysis also ranked rainfall and solar radiation in the previous 0–2 days as the most important variables for predicting generic *E. coli* ≥126 MPN/100 mL (Figure 2).

Model Prediction Performance.—Model predictions for the test data set outcomes (2015–2016) were analyzed with ROC curves (Figure 3), using the AUC to evaluate overall predictive performance. Logistic regression models and CRF models demonstrated comparable discriminatory ability. Logistic regression model AUCs were slightly higher than the CRF AUC for the human-associated outcomes but lower for elevated levels of *E. coli*. Performance was lowest for predicting HF183 detection (AUC: 0.56–0.60). Despite the substitution of human-associated FRNA GII coliphage for crAssphage in the HFI variable definition, logistic regression predictions of HFI detection were more accurate (AUC: 0.64) but did not achieve the AUC > 0.7 target conventionally viewed as acceptable predictive performance.[78] Models built for predicting generic *E. coli* ≥ 126 MPN/100 mL had higher

predictive performance (AUC: 0.77–0.79) than models for either human-associated outcome, attributable to the high sensitivity (100%) attained at moderate specificities (54–64%). However, sensitivity declined rapidly with any further increase in specificity, reflected in the low classification thresholds identified by Youden's *J* statistic at probabilities of 0.08–0.11. Such low thresholds indicate that the application of more stringent criteria to discriminate between detects and non-detects sharply reduced identification of true positives (of which there were only 6 in the test data set) without providing a corresponding reduction in the false positive rate. Including *E. coli* ≥ 126 MPN/100 mL as an additional explanatory variable in the logistic regression and CRF models did not improve predictions of either human-associated outcome (Figure S3). Similarly, CRF hyperparameter tuning and imbalance correction did not improve AUC for test data set predictions (Figure S4).

DISCUSSION

Our results can be used to identify factors associated with human fecal contamination in southeastern US produce irrigation water. However, predictive statistical models should be used with caution in irrigation water quality assessments, as predictions for locations and times beyond those on which the models were trained may be unreliable. In the current study, no model produced accurate out-of-sample predictions of the presence of human fecal indicators in additional ponds from the same growing area sampled in different years. Although the negligible influence of nearly all explanatory variables and the dominance of a single, static site characteristic (building proximity) suggest limited opportunity to improve predictions through increased data collection, expanding the training data set with observations at additional locations and times could potentially provide greater generalizability to inform out-of-sample predictions.

The most influential factor in detecting molecular human fecal indicators in irrigation water was being located near a building, which, in this rural area, indicates a high likelihood of proximity to a septic system. The soil in this area is rated "very limited" for septic tank absorption fields, meaning septic systems are expected to perform poorly and may introduce human fecal contamination to adjacent environments.[80] Recent rainfall was also associated with an increased risk of detecting human fecal indicators and generic *E. coli* ≥126 MPN/100 mL. Increased rainfall in the previous 48 h was the strongest predictor of elevated *E. coli* and was associated with larger, but more variable, increases in odds of HF183 and HFI detection than less recent rainfall. The impact of increased rainfall in the previous 2–7 days was smaller in magnitude but more consistently associated with increased odds of HF183 and HFI detection. Although the ponds sampled for the training data set were all reported to be fed by surface water, this suggests that contamination of the subsurface water through septic pollution could be a contributor to human fecal contamination in this growing region.

While human MST markers have been reported in produce irrigation water,[45,81,82] predictors of these markers in irrigation ponds have not previously been characterized. Studies in beach waters have consistently found that precipitation is an important predictor of HF183.[40,83,84] Rainfall was also significantly associated with HF183 in private well water in Pennsylvania[85] and in rural waterways where onsite wastewater treatment was suspected

as the source of contamination.[86] While other studies have found significant associations between solar radiation and HF183,[40] we did not observe human marker associations with solar radiation in this study. Rainfall variables and solar radiation have also been determined as critical factors in modeling unsafe ambient recreational water conditions due to elevated *E. coli*.[87,88] A systematic review of predictive models of *E. coli* in beach water found that rainfall was the most frequently included variable in final models.[89] The second most commonly included variable was turbidity, which corresponds to more suspended particles in the water column that can provide protection against solar inactivation for particle-associated microorganisms.[90] However, turbidity was not associated with any of the fecal indicator outcomes in this study. The frequent presence of algae during sample collection may partially account for the lack of association with turbidity. Algae can interfere with probe-based turbidity measurements and have the potential to both inhibit and stimulate bacterial growth.[91] Future studies may consider using more robust laboratory-based turbidity measurements and quantifying algae in surface water samples to address inconsistencies potentially introduced by heavy algal loads.

Elevated generic *E. coli* and the presence of HF183 were not correlated in this study. This finding is consistent with previous research, showing that the drivers of human fecal contamination vary from those for generic fecal indicator bacteria.[28,32,40] Multiple lines of evidence suggest the likely presence of nonhuman fecal contamination, including previous research that identified wildlife- and livestock-shed foodborne pathogens *Campylobacter jejuni*, *Salmonella enterica*, and pathogenic *E. coli* in surface waters used for irrigation in this growing region.[8,92–95]

This study allowed us not only to assess predictors of general and human-specific fecal contamination but also to compare different predictive modeling approaches. Previous studies have suggested that machine learning-based predictive models could be used to determine when pathogens are most likely to be present in irrigation water.[43,96] In particular, conditional random forest models were previously found to more accurately capture relationships with environmental factors to predict *Salmonella* and pathogenic *E. coli* presence in produce irrigation water in northeastern and southwestern US growing regions.[43,44] A comparison of predictive modeling approaches also identified random forest models as the most accurate approach for predicting fecal indicator bacteria in ambient recreational water.[97] However, our study observed out-of-sample predictive performance by CRF models for human fecal indicator presence that was only marginally better than chance and slightly inferior to the predictive performance of logistic regression models. A recent systematic review of clinical prediction models for a range of binary outcomes likewise found no consistent advantage of random forest and other machine learning approaches over logistic regression.[98] Alternative performance metrics to AUC could have yielded different relative performance rankings of the two approaches, but the absolute performance was sufficiently poor that any reasonable metric should have captured the predictive inadequacy of both approaches. Previous comparisons of random forest- and regression-based approaches found that the different methods identified different explanatory variables as important for predicting pathogen presence in irrigation water.[44] By contrast, in this study, both regression and CRF approaches identified the same influential explanatory variables for each fecal indicator outcome.

A strength of our study was the inclusion of multiple markers of human fecal contamination to address the limitations of the individual markers. HF183 has been shown to cross-react with poultry and dog feces in many settings,[24,99–101] while crAssphage, though less extensively validated, has previously demonstrated superior host specificity.[99] Domestic dogs were observed during sample collection at residences near the irrigation ponds, which could have served as a potential source of the HF183 assay cross-reaction. Therefore, we used a conservative detection criterion of two or more positive qPCR replicates. CrAssphage may be a less-sensitive human fecal indicator than HF183, though it is often correlated with HF183, as was observed in the present study.[18,28] The human MST markers were codetected too infrequently to develop predictive models of HF183 and crAssphage codetection, but all samples in which human markers were codetected occurred in irrigation ponds near buildings, further suggesting the influence of buildings (with presumed septic systems) on human fecal contamination of irrigation waters.

Because crAssphage was not measured in the test data set, we substituted FRNA GII coliphage as the human-associated fecal indicator virus. Though less human-specific than crAssphage,[35,36] coliphage was detected with similar frequency in the test ponds (all close to buildings) as the frequency of crAssphage detection in the training ponds with nearby buildings. Furthermore, the models developed to predict HF183 and/or crAssphage produced more accurate predictions for HF183 and/or FRNA GII coliphage than the HF183-trained model predictions of HF183 alone, supporting FRNA GII coliphage as a reasonable substitute for crAssphage as a human fecal indicator virus in this setting. Future studies should consider the addition of a viral concentration step, such as PEG precipitation or cellulose ester membrane filtration, to increase the recovery of human-associated viral markers and improve the sensitivity of human fecal contamination detection.[18,27]

CONCLUSIONS

This research demonstrated significantly more human fecal marker intrusion into irrigation ponds in an agricultural region of southwest Georgia when a building was present and with greater rainfall in the previous week. This should be considered when a preharvest water assessment is completed for the introduction of hazards onto produce. Human fecal contamination from nearby buildings should be assessed prior to using an irrigation pond for produce production. Predictive models have previously been suggested for preharvest assessment; however, this study demonstrated that while our modeling approaches were able to determine risk factors, they could not reliably predict water contamination over multiple years. Our findings highlight the continued role for water quality testing, including MST approaches, in protecting the safety of fresh produce.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We thank Daniel Weller for guidance on analytical methods. Funding for this project was provided by the Center for Produce Safety through a CDFA 2019 Specialty Crop Block Grant Program & CPS Campaign for Research and

through CDFA SCBGP grant #SBC14060. The table of contents graphic was created with BioRender. The use of trade names and names of commercial sources is for identification only and does not imply endorsement by the Centers for Disease Control and Prevention or the U.S. Department of Health and Human Services. The findings and conclusions are those of the authors and do not necessarily represent those of the Centers for Disease Control and Prevention.

REFERENCES

(1). Batz MB; Richardson LC; Bazaco MC; Parker CC; Chirtel SJ; Cole D; Golden NJ; Griffin PM; Gu W; Schmitt SK; Wolpert BJ; Kufel JSZ; Hoekstra RM Recency-Weighted Statistical Modeling Approach to Attribute Illnesses Caused by 4 Pathogens to Food Sources Using Outbreak Data, United States. Emerg. Infect. Dis 2021, 27 (1), 214–222. [PubMed: 33350919]

(2). Interagency Food Safety Analytics Collaboration. Foodborne Illness Source Attribution Estimates for 2021 for Salmonella, Escherichia coli O157, and Listeria monocytogenes Using Multi-Year Outbreak Surveillance Data, United States; U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, Food and Drug Administration, U.S. Department of Agriculture's Food Safety and Inspection Service: Atlanta, GA and Washington, D.C., 2023. https://www.cdc.gov/ifsac/media/pdfs/P19-2021-report-TriAgency-508.pdf.

(3). Steele M; Odumeru J Irrigation Water as Source of Foodborne Pathogens on Fruit and Vegetables. J. Food Prot 2004, 67 (12), 2839–2849. [PubMed: 15633699]

(4). Holcomb DA; Stewart JR Microbial Indicators of Fecal Pollution: Recent Progress and Challenges in Assessing Water Quality. Curr. Environ. Health Rep 2020, 7 (3), 311–324. [PubMed: 32542574]

(5). U.S. Environmental Protection Agency. Factsheet on Water Quality Parameters: E. coli (Escherichia coli); EPA 841F21007F, 2021. https://www.epa.gov/system/files/documents/2021-07/parameter-factsheet_e.-coli.pdf.

(6). California Leafy Greens Marketing Agreement. Commodity Specific Food Safety Guidelines For the Production and Harvest of Lettuce and Leafy Greens; Western Growers Association: Irvine, CA, 2021. https://lgma-assets.sfo2.digitaloceanspaces.com/downloads/August-2021-CA-LGMA-Metrics_FINAL-v20211208_A11Y.pdf.

(7). Shelton DR; Karns JS; Coppock C; Patel J; Sharma M; Pachepsky YA Relationship between *eae* and *stx* Virulence Genes and *Escherichia coli* in an Agricultural Watershed: Implications for Irrigation Water Standards and Leafy Green Commodities. J. Food Prot 2011, 74 (1), 18–23. [PubMed: 21219758]

(8). Harris CS; Tertuliano M; Rajeev S; Vellidis G; Levy K Impact of Storm Runoff on *Salmonella* and *Escherichia coli* Prevalence in Irrigation Ponds of Fresh Produce Farms in Southern Georgia. J. Appl. Microbiol 2018, 124 (3), 910–921. [PubMed: 29316043]

(9). Antaki EM; Vellidis G; Harris C; Aminabadi P; Levy K; Jay-Russell MT Low Concentration of *Salmonella enterica* and Generic *Escherichia coli* in Farm Ponds and Irrigation Distribution Systems Used for Mixed Produce Production in Southern Georgia. Foodborne Pathog. Dis 2016, 13 (10), 551–558. [PubMed: 27400147]

(10). Jokinen CC; Hillman E; Tymensen L Sources of Generic *Escherichia coli* and Factors Impacting Guideline Exceedances for Food Safety in an Irrigation Reservoir Outlet and Two Canals. Water Res 2019, 156, 148–158. [PubMed: 30913418]

(11). Villabruna N; Koopmans MPG; De Graaf M Animals as Reservoir for Human Norovirus. Viruses 2019, 11 (5), 478. [PubMed: 31130647]

(12). Aggarwal R; Jameel S Hepatitis E. Hepatology 2010, 54 (6), 2218–2226.

(13). Di Cola G; Fantilli AC; Pisano MB; Ré VE Foodborne Transmission of Hepatitis A and Hepatitis E Viruses: A Literature Review. Int. J. Food Microbiol 2021, 338, No. 108986.

(14). Eberhard ML; Ortega YR; Hanes DE; Nace EK; Quy Do R; Robl MG; Won KY; Gavidia C; Sass NL; Mansfield K; Gozalo A; Griffiths J; Gilman R; Sterling CR; Arrowood MJ Attempts to Establish Experimental *Cyclospora cayetanensis* Infection in Laboratory Animals. J. Parasitol 2000, 86 (3), 577–582. [PubMed: 10864257]

(15). Kokkinos P; Kozyra I; Lazic S; Söderberg K; Vasickova P; Bouwknegt M; Rutjes S; Willems K; Moloney R; De Roda Husman AM; Kaupke A; Legaki E; D'Agostino M; Cook N; Von

Bonsdorff C-H; Rzezutka A; Petrovic T; Maunula L; Pavlik I; Vantarakis A Virological Quality of Irrigation Water in Leafy Green Vegetables and Berry Fruits Production Chains. Food Environ. Virol 2017, 9 (1), 72–78. [PubMed: 27709435]

(16). Giangaspero A; Marangi M; Koehler AV; Papini R; Normanno G; Lacasella V; Lonigro A; Gasser RB Molecular Detection of *Cyclospora* in Water, Soil, Vegetables and Humans in Southern Italy Signals a Need for Improved Monitoring by Health Authorities. Int. J. Food Microbiol 2015, 211, 95–100. [PubMed: 26188495]

(17). Harwood VJ; Staley C; Badgley BD; Borges K; Korajkic A Microbial Source Tracking Markers for Detection of Fecal Contamination in Environmental Waters: Relationships between Pathogens and Human Health Outcomes. FEMS Microbiol. Rev 2014, 38 (1), 1–40. [PubMed: 23815638]

(18). Sala-Comorera L; Reynolds LJ; Martin NA; Pascual-Benito M; Stephens JH; Nolan TM; Gitto A; O'Hare GMP; O'Sullivan JJ; García-Aljaro C; Meijer WG crAssphage as a Human Molecular Marker to Evaluate Temporal and Spatial Variability in Faecal Contamination of Urban Marine Bathing Waters. Sci. Total Environ 2021, 789, No. 147828.

(19). Ahmed W; Gyawali P; Feng S; McLellan SL Host Specificity and Sensitivity of Established and Novel Sewage-Associated Marker Genes in Human and Nonhuman Fecal Samples. Appl. Environ. Microbiol 2019, 85 (14), No. e00641–19.

(20). Boehm AB; Graham KE; Jennings WC Can We Swim Yet? Systematic Review, Meta-Analysis, and Risk Assessment of Aging Sewage in Surface Waters. Environ. Sci. Technol 2018, 52 (17), 9634–9645. [PubMed: 30080397]

(21). Bernhard AE; Field KG A PCR Assay To Discriminate Human and Ruminant Feces on the Basis of Host Differences in *Bacteroides-Prevotella* Genes Encoding 16S rRNA. Appl. Environ. Microbiol 2000, 66 (10), 4571–4574. [PubMed: 11010920]

(22). Haugland RA; Varma M; Sivaganesan M; Kelty C; Peed L; Shanks OC Evaluation of Genetic Markers from the 16S rRNA Gene V2 Region for Use in Quantitative Detection of Selected Bacteroidales Species and Human Fecal Waste by qPCR. Syst. Appl. Microbiol 2010, 33 (6), 348–357. [PubMed: 20655680]

(23). Green HC; Haugland RA; Varma M; Millen HT; Borchardt MA; Field KG; Walters WA; Knight R; Sivaganesan M; Kelty CA; Shanks OC Improved HF183 Quantitative Real-Time PCR Assay for Characterization of Human Fecal Pollution in Ambient Surface Water Samples. Appl. Environ. Microbiol 2014, 80 (10), 3086–3094. [PubMed: 24610857]

(24). Layton BA; Cao Y; Ebentier DL; Hanley K; Ballesté E; Brandão J; Byappanahalli M; Converse R; Farnleitner AH; Gentry-Shields J; Gidley ML; Gourmelon M; Lee CS; Lee J; Lozach S; Madi T; Meijer WG; Noble R; Peed L; Reischer GH; Rodrigues R; Rose JB; Schriewer A; Sinigalliano C; Srinivasan S; Stewart J; Van De Werfhorst LC; Wang D; Whitman R; Wuertz S; Jay J; Holden PA; Boehm AB; Shanks O; Griffith JF Performance of Human Fecal Anaerobe-Associated PCR-Based Assays in a Multi-Laboratory Method Evaluation Study. Water Res 2013, 47 (18), 6897–6908. [PubMed: 23992621]

(25). Li X; Sivaganesan M; Kelty CA; Zimmer-Faust A; Clinton P; Reichman JR; Johnson Y; Matthews W; Bailey S; Shanks OC Large-Scale Implementation of Standardized Quantitative Real-Time PCR Fecal Source Identification Procedures in the Tillamook Bay Watershed. PLoS One 2019, 14 (6), No. e0216827.

(26). Stachler E; Kelty C; Sivaganesan M; Li X; Bibby K; Shanks OC Quantitative crAssphage PCR Assays for Human Fecal Pollution Measurement. Environ. Sci. Technol 2017, 51 (16), 9146–9154. [PubMed: 28700235]

(27). Sabar MA; Honda R; Haramoto E CrAssphage as an Indicator of Human-Fecal Contamination in Water Environment and Virus Reduction in Wastewater Treatment. Water Res 2022, 221, No. 118827.

(28). Jennings WC; Gálvez-Arango E; Prieto AL; Boehm AB CrAssphage for Fecal Source Tracking in Chile: Covariation with Norovirus, HF183, and Bacterial Indicators. Water Res. X 2020, 9, No. 100071.

(29). Nguyen KH; Smith S; Roundtree A; Feistel DJ; Kirby AE; Levy K; Mattioli MC Fecal Indicators and Antibiotic Resistance Genes Exhibit Diurnal Trends in the Chattahoochee River: Implications for Water Quality Monitoring. Front. Microbiol 2022, 13, No. 1029176.

(30). Staley ZR; Vogel L; Robinson C; Edge TA Differential Occurrence of *Escherichia coli* and Human Bacteroidales at Two Great Lakes Beaches. J. Great Lakes Res 2015, 41 (2), 530–535.

(31). Stachler E; Akyon B; De Carvalho NA; Ference C; Bibby K Correlation of crAssphage qPCR Markers with Culturable and Molecular Indicators of Human Fecal Pollution in an Impacted Urban Watershed. Environ. Sci. Technol 2018, 52 (13), 7505–7512. [PubMed: 29874457]

(32). Shahin SA; Keevy H; Dada AC; Gyawali P; Sherchan SP Incidence of Human Associated HF183 *Bacteroides* Marker and *E. coli* Levels in New Orleans Canals. Sci. Total Environ 2022, 806 (1), No. 150356.

(33). Nappier SP; Hong T; Ichida A; Goldstone A; Eftim SE Occurrence of Coliphage in Raw Wastewater and in Ambient Water: A Meta-Analysis. Water Res 2019, 153, 263–273. [PubMed: 30735956]

(34). Havelaar AH; Pot-Hogeboom WM; Furuse K; Pot R; Hormann MP F-specific RNA Bacteriophages and Sensitive Host Strains in Faeces and Wastewater of Human and Animal Origin. J. Appl. Bacteriol 1990, 69 (1), 30–37. [PubMed: 2204615]

(35). Schaper M; Jofre J; Uys M; Grabow WOK Distribution of Genotypes of F-Specific RNA Bacteriophages in Human and Non-Human Sources of Faecal Pollution in South Africa and Spain. J. Appl. Microbiol 2002, 92 (4), 657–667. [PubMed: 11966906]

(36). Stewart-Pullaro J; Daugomah JW; Chestnut DE; Graves DA; Sobsey MD; Scott GIF $^{+}$ RNA Coliphage Typing for Microbial Source Tracking in Surface Waters. J. Appl. Microbiol 2006, 101 (5), 1015–1026. [PubMed: 17040225]

(37). Sowah RA; Molina M; Georgacopoulos O; Snyder B; Cyterski M Sources and Drivers of ARGs in Urban Streams in Atlanta, Georgia. Microorganisms 2022, 10 (9), 1804. [PubMed: 36144405]

(38). Bihn EA; Mangione KJ; Lyons B; Wszelaki AL; Churey JJ; Stoeckel DM; Worobo RW Development of an Irrigation Water Quality Database to Identify Water Resources and Assess Microbiological Risks During the Production of Fresh Fruits and Vegetables. Front. Water 2021, 3, No. 741653.

(39). McKee BA; Molina M; Cyterski M; Couch A Microbial Source Tracking (MST) in Chattahoochee River National Recreation Area: Seasonal and Precipitation Trends in MST Marker Concentrations, and Associations with *E. coli* Levels, Pathogenic Marker Presence, and Land Use. Water Res 2020, 171, No. 115435.

(40). Jennings WC; Chern EC; O'Donohue D; Kellogg MG; Boehm AB Frequent Detection of a Human Fecal Indicator in the Urban Ocean: Environmental Drivers and Covariation with Enterococci. Environ. Sci.: Processes Impacts 2018, 20 (3), 480–492.

(41). Staley C; Reckhow KH; Lukasik J; Harwood VJ Assessment of Sources of Human Pathogens and Fecal Contamination in a Florida Freshwater Lake. Water Res 2012, 46 (17), 5799–5812. [PubMed: 22939220]

(42). Weller DL; Love TMT; Wiedmann M Interpretability Versus Accuracy: A Comparison of Machine Learning Models Built Using Different Algorithms, Performance Measures, and Features to Predict *E. coli* Levels in Agricultural Water. Front. Artif. Intell 2021, 4, No. 628441.

(43). Weller DL; Love TMT; Belias A; Wiedmann M Predictive Models May Complement or Provide an Alternative to Existing Strategies for Assessing the Enteric Pathogen Contamination Status of Northeastern Streams Used to Provide Water for Produce Production. Front Sustainable Food Syst 2020, 4, No. 561517.

(44). Belias A; Brassill N; Roof S; Rock C; Wiedmann M; Weller D Cross-Validation Indicates Predictive Models May Provide an Alternative to Indicator Organism Monitoring for Evaluating Pathogen Presence in Southwestern US Agricultural Water. Front. Water 2021, 3, No. 693631.

(45). Green H; Wilder M; Wiedmann M; Weller D Integrative Survey of 68 Non-Overlapping Upstate New York Watersheds Reveals Stream Features Associated With Aquatic Fecal Contamination. Front. Microbiol 2021, 12, No. 684533.

(46). Griffith GE; Omernik JM; Comstock JA; Lawrence S; Martin G; Goddard A; Hulcher VJ; Foster T Ecoregions of Alabama and Georgia (Color Poster with Map, Descriptive Text, Summary Tables, and Photographs); U.S. Geological Survey: Reston, VA, 2001. https://www.epa.gov/eco-research/ecoregion-download-files-state-region-4.

(47). Sullivan DG; Batten HL; Bosch D; Sheridan J; Strickland T Little River Experimental Watershed, Tifton, Georgia, United States: A Geographic Database. Water Resour. Res 2007, 43 (9), No. 2006WR005836.

(48). Kahler AM; Hofstetter J; Arrowood M; Peterson A; Jacobson D; Barratt J; da Silva ALBR; Rodrigues C; Mattioli MC Sources and Prevalence of *Cyclospora cayetanensis* in Southeastern U.S. Growing Environments. J. Food Prot 2024, 87, No. 100309.

(49). Kahler AM; Hill VR Detection of *Cryptosporidium* Recovered from Large-Volume Water Samples Using Dead-End Ultrafiltration. In Cryptosporidium; Mead JR; Arrowood MJ, Eds.; Methods in Molecular Biology; Humana: New York, NY, 2020; Vol. 2052.

(50). Environmental Protection Agency US. Method 1696: Characterization of Human Fecal Pollution in Water by TaqMan Quantitative Polymerase Chain Reaction (qPCR) Assay; EPA 821-R-19–002; U.S. EPA Office of Research and Development: Cincinnati, OH, 2019. https://www.epa.gov/sites/default/files/2019-03/documents/method_1696_draft_2019.pdf.

(51). Hill V; Vellidis G; Levy K Improved Sampling and Analytical Methods for Testing Agricultural Water for Pathogens, Surrogates and Source Tracking Indicators; Center for Produce Safety, 2017. https://www.centerforproducesafety.org/assets/research-database/Hill-2014-Final-Report.pdf.

(52). U.S. Environmental Protection Agency. Method 1602: Male-Specific (F+) and Somatic Coliphage in Water by Single Agar Layer (SAL) Procedure; EPA 821-R-01–029; U.S. EPA Office of Water: Washington, DC, 2001. https://www.epa.gov/sites/default/files/2015-12/documents/method_1602_2001.pdf.

(53). Friedman SD; Cooper EM; Calci KR; Genthner FJ Design and Assessment of a Real Time Reverse Transcription-PCR Method to Genotype Single-Stranded RNA Male-Specific Coliphages (Family Leviviridae). J. Virol. Methods 2011, 173 (2), 196–202. [PubMed: 21320531]

(54). Polaczyk AL; Narayanan J; Cromeans TL; Hahn D; Roberts JM; Amburgey JE; Hill VR Ultrafiltration-Based Techniques for Rapid and Simultaneous Concentration of Multiple Microbe Classes from 100-L Tap Water Samples. J. Microbiol. Methods 2008, 73 (2), 92–99. [PubMed: 18395278]

(55). Hill VR; Narayanan J; Gallen RR; Ferdinand KL; Cromeans T; Vinjé J Development of a Nucleic Acid Extraction Procedure for Simultaneous Recovery of DNA and RNA from Diverse Microbes in Water. Pathogens 2015, 4 (2), 335–354. [PubMed: 26016775]

(56). University of Georgia. UGA Weather—Automated Environmental Monitoring Network Page; UGA Weather Network, 2023. http://www.georgiaweather.net/?variable=HI&site=TYTY.

(57). National Water and Climate Center. Little River-Site Information and Reports; Soil Climate Analysis Network; Site 2027; Natural Resources Conservation Service, United States Department of Agriculture, 2023. https://wcc.sc.egov.usda.gov/nwcc/site?sitenum=2027.

(58). Wiesner-Friedman C; Beattie RE; Stewart JR; Hristova KR; Serre ML Microbial Find, Inform, and Test Model for Identifying Spatially Distributed Contamination Sources: Framework Foundation and Demonstration of Ruminant *Bacteroides* Abundance in River Sediments. Environ. Sci. Technol 2021, 55 (15), 10451–10461. [PubMed: 34291905]

(59). Holcomb DA; Messier KP; Serre ML; Rowny JG; Stewart JR Geostatistical Prediction of Microbial Water Quality throughout a Stream Network Using Meteorology, Land Cover, and Spatiotemporal Autocorrelation. Environ. Sci. Technol 2018, 52 (14), 7775–7784. [PubMed: 29886747]

(60). Georgia Department of Public Health. Manual for On-Site Sewage Management Systems, 2019. https://dph.georgia.gov/document/document/manual-site-sewage-management-systems-rules/download.

(61). R Core Team. R: A Language and Environment for Statistical Computing; R Foundation for Statistical Computing: Vienna, Austria, 2024. https://www.R-project.org/.

(62). Bates D; Mächler M; Bolker B; Walker S Fitting Linear Mixed-Effects Models Using lme4. J. Stat. Softw 2015, 67 (1), 1–48.

(63). Hothorn T; Hornik K; Zeileis A Unbiased Recursive Partitioning: A Conditional Inference Framework. J. Comput. Graph. Stat 2006, 15 (3), 651–674.

(64). Strobl C; Boulesteix A-L; Zeileis A; Hothorn T Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. BMC Bioinf 2007, 8 (1), 25.

(65). Strobl C; Malley J; Tutz G An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests. Psychol. Methods 2009, 14 (4), 323–348. [PubMed: 19968396]

(66). Bischl B; Lang M; Kotthoff L; Schiffner J; Richter J; Studerus E; Casalicchio G; Jones ZM mlr: Machine Learning in R. J. Mach. Learn. Res 2016, 17 (170), 1–5.

(67). Chawla NV; Bowyer KW; Hall LO; Kegelmeyer WP SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res 2002, 16, 321–357.

(68). Bischl B; Kühn T; Szepannek G On Class Imbalance Correction for Classification Algorithms in Credit Scoring. In Operations Research Proceedings 2014; Lübbecke M; Koster A; Letmathe P; Madlener R; Peis B; Walther G, Eds.; Operations Research Proceedings; Springer International Publishing: Cham, 2016; pp 37–43.

(69). Weller DL; Love TMT; Wiedmann M Comparison of Resampling Algorithms to Address Class Imbalance When Developing Machine Learning Models to Predict Foodborne Pathogen Presence in Agricultural Water. Front. Environ. Sci 2021, 9, No. 701288.

(70). Kim M; Hwang K-B An Empirical Evaluation of Sampling Methods for the Classification of Imbalanced Data. PLoS One 2022, 17 (7), No. e0271260.

(71). Van Den Goorbergh R; Van Smeden M; Timmerman D; Van Calster B The Harm of Class Imbalance Corrections for Risk Prediction Models: Illustration and Simulation Using Logistic Regression. J. Am. Med. Inform. Assoc 2022, 29 (9), 1525–1534. [PubMed: 35686364]

(72). Piccininni M; Wechsung M; Van Calster B; Rohmann JL; Konigorski S; Van Smeden M Understanding Random Resampling Techniques for Class Imbalance Correction and Their Consequences on Calibration and Discrimination of Clinical Risk Prediction Models. J. Biomed. Inf 2024, 155, No. 104666.

(73). Strobl C; Boulesteix A-L; Kneib T; Augustin T; Zeileis A Conditional Variable Importance for Random Forests. BMC Bioinf 2008, 9 (1), 307.

(74). Janitza S; Strobl C; Boulesteix A-L An AUC-Based Permutation Variable Importance Measure for Random Forests. BMC Bioinf 2013, 14 (1), 119.

(75). Robin X; Turck N; Hainard A; Tiberti N; Lisacek F; Sanchez J-C; Müller M pROC: An Open-Source Package for R and S+ to Analyze and Compare ROC Curves. BMC Bioinf 2011, 12 (1), 77.

(76). Steyerberg EW; Vickers AJ; Cook NR; Gerds T; Gonen M; Obuchowski N; Pencina MJ; Kattan MW Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures. Epidemiology 2010, 21 (1), 128–138. [PubMed: 20010215]

(77). Fawcett T An Introduction to ROC Analysis. Pattern Recognit. Lett 2006, 27 (8), 861–874.

(78). Mandrekar JN Receiver Operating Characteristic Curve in Diagnostic Test Assessment. J. Thorac. Oncol 2010, 5 (9), 1315–1316. [PubMed: 20736804]

(79). Youden WJ Index for Rating Diagnostic Tests. Cancer 1950, 3 (1), 32–35. [PubMed: 15405679]

(80). Natural Resources Conservation Service. Web Soil Survey; U.S. Department of Agriculture, 2023. http://websoilsurvey.sc.egov.usda.gov/.

(81). Silverman AI; Akrong MO; Amoah P; Drechsel P; Nelson KL Quantification of Human Norovirus GII, Human Adenovirus, and Fecal Indicator Organisms in Wastewater Used for Irrigation in Accra, Ghana. J. Water Health 2013, 11 (3), 473–488. [PubMed: 23981876]

(82). Ravaliya K; Gentry-Shields J; Garcia S; Heredia N; Fabiszewski De Aceituno A; Bartz FE; Leon JS; Jaykus L-A Use of Bacteroidales Microbial Source Tracking To Monitor Fecal Contamination in Fresh Produce Production. Appl. Environ. Microbiol 2014, 80 (2), 612–617. [PubMed: 24212583]

(83). Ahmed W; Payyappat S; Cassidy M; Harrison N; Besley C Sewage-Associated Marker Genes Illustrate the Impact of Wet Weather Overflows and Dry Weather Leakage in Urban Estuarine Waters of Sydney, Australia. Sci. Total Environ 2020, 705, No. 135390.

(84). Schiff K; Griffith J; Steele J; Zimmer-Faust A Dry and Wet Weather Survey for Human Fecal Sources in the San Diego River Watershed. Water 2023, 15 (12), 2239.

(85). Murphy HM; McGinnis S; Blunt R; Stokdyk J; Wu J; Cagle A; Denno DM; Spencer S; Firnstahl A; Borchardt MA Septic Systems and Rainfall Influence Human Fecal Marker and Indicator Organism Occurrence in Private Wells in Southeastern Pennsylvania. Environ. Sci. Technol 2020, 54 (6), 3159–3168. [PubMed: 32073835]

(86). Stea EC; Truelstrup Hansen L; Jamieson RC; Yost CK Fecal Contamination in the Surface Waters of a Rural- and an Urban-Source Watershed. J. Environ. Qual 2015, 44 (5), 1556–1567. [PubMed: 26436273]

(87). Thoe W; Gold M; Griesbach A; Grimmer M; Taggart ML; Boehm AB Predicting Water Quality at Santa Monica Beach: Evaluation of Five Different Models for Public Notification of Unsafe Swimming Conditions. Water Res 2014, 67, 105–117. [PubMed: 25262555]

(88). Whitman RL; Nevers MB; Korinek GC; Byappanahalli MN Solar and Temporal Effects on *Escherichia coli* Concentration at a Lake Michigan Swimming Beach. Appl. Environ. Microbiol 2004, 70 (7), 4276–4285. [PubMed: 15240311]

(89). Heasley C; Sanchez JJ; Tustin J; Young I Systematic Review of Predictive Models of Microbial Water Quality at Freshwater Recreational Beaches. PLoS One 2021, 16 (8), No. e0256785.

(90). Walters E; Graml M; Behle C; Müller E; Horn H Influence of Particle Association and Suspended Solids on UV Inactivation of Fecal Indicator Bacteria in an Urban River. Water, Air, Soil Pollut 2014, 225 (1), 1822.

(91). Cole JJ Interactions Between Bacteria and Algae in Aquatic Ecosystems. Annu. Rev. Ecol. Syst 1982, 13 (1), 291–314.

(92). Luo Z; Gu G; Ginn A; Giurcanu MC; Adams P; Vellidis G; Van Bruggen AHC; Danyluk MD; Wright AC Distribution and Characterization of *Salmonella enterica* Isolates from Irrigation Ponds in the Southeastern United States. Appl. Environ. Microbiol 2015, 81 (13), 4376–4387. [PubMed: 25911476]

(93). Gu G; Luo Z; Cevallos-Cevallos JM; Adams P; Vellidis G; Wright A; Van Bruggen AHC Factors Affecting the Occurrence of *Escherichia coli* O157 Contamination in Irrigation Ponds on Produce Farms in the Suwannee River Watershed. Can. J. Microbiol 2013, 59 (3), 175–182. [PubMed: 23540335]

(94). Gu G; Luo Z; Cevallos-Cevallos JM; Adams P; Vellidis G; Wright A; Van Bruggen AHC Occurrence and Population Density of *Campylobacter jejuni* in Irrigation Ponds on Produce Farms in the Suwannee River Watershed. Can. J. Microbiol 2013, 59 (5), 339–346. [PubMed: 23647347]

(95). Li B; Vellidis G; Liu H; Jay-Russell M; Zhao S; Hu Z; Wright A; Elkins CA Diversity and Antimicrobial Resistance of *Salmonella enterica* Isolates from Surface Water in Southeastern United States. Appl. Environ. Microbiol 2014, 80 (20), 6355–6365. [PubMed: 25107969]

(96). Polat H; Topalcengiz Z; Danyluk MD Prediction of *Salmonella* Presence and Absence in Agricultural Surface Waters by Artificial Intelligence Approaches. J. Food Saf 2020, 40 (1), No. e12733.

(97). Brooks W; Corsi S; Fienen M; Carvin R Predicting Recreational Water Quality Advisories: A Comparison of Statistical Methods. Environ. Model. Softw 2016, 76, 81–94.

(98). Christodoulou E; Ma J; Collins GS; Steyerberg EW; Verbakel JY; Van Calster B A Systematic Review Shows No Performance Benefit of Machine Learning over Logistic Regression for Clinical Prediction Models. J. Clin. Epidemiol 2019, 110, 12–22. [PubMed: 30763612]

(99). Ahmed W; Payyappat S; Cassidy M; Harrison N; Besley C Microbial Source Tracking of Untreated Human Wastewater and Animal Scats in Urbanized Estuarine Waters. Sci. Total Environ 2023, 877, No. 162764.

(100). Holcomb DA; Knee J; Capone D; Sumner T; Adriano Z; Nalá R; Cumming O; Brown J; Stewart JR Impacts of an Urban Sanitation Intervention on Fecal Indicators and the Prevalence of Human Fecal Contamination in Mozambique. Environ. Sci. Technol 2021, 55 (17), 11667–11679. [PubMed: 34382777]

(101). Nshimyimana JP; Cruz MC; Thompson RJ; Wuertz S Bacteroidales Markers for Microbial Source Tracking in Southeast Asia. Water Res 2017, 118, 239–248. [PubMed: 28433694]

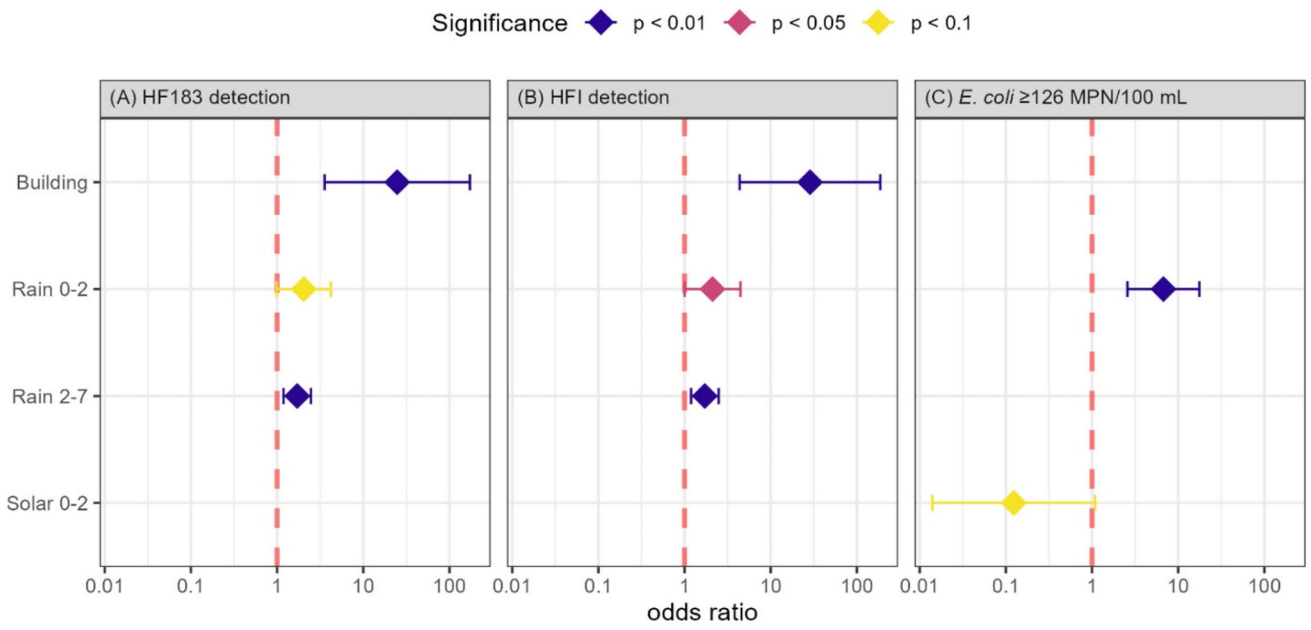


Figure 1.

Odds ratio (95% confidence interval) estimates for exposure variables in the final mixed-effects logistic regression models for the three fecal indicators, HF183 (A), human fecal indicator (HFI; HF183 and crAssphage) (B), and *E. coli* ≥ 126 MPN/100 mL (C).

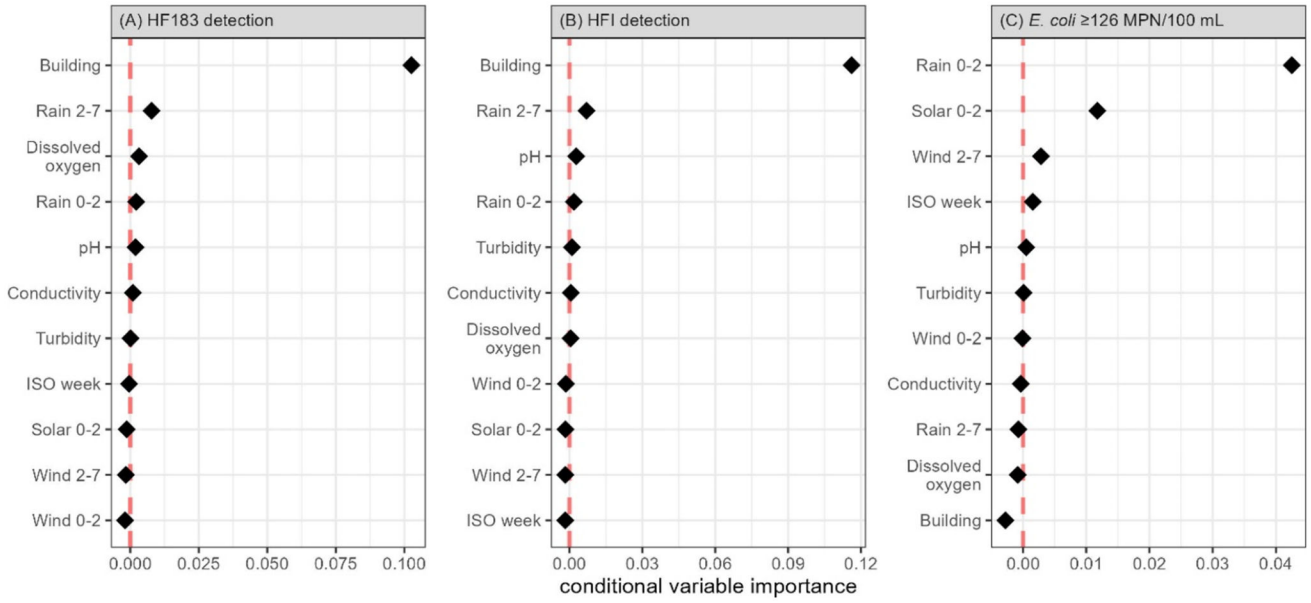


Figure 2.

Conditional variable importance for each conditional random forest model (CRF): HF183 (A), human fecal indicator (HFI; HF183 and/or crAssphage) (B), and *E. coli* ≥ 126 MPN/100 mL (C). The *y*-axis shows the explanatory variables ranked from most important to least important. The *x*-axis shows the variable importance on the basis of reduction in the area under the curve (AUC) by conditional permutation; higher relative variable importance indicates stronger association between the variable and the outcome. Variable importance ≤0 indicates negligible association.

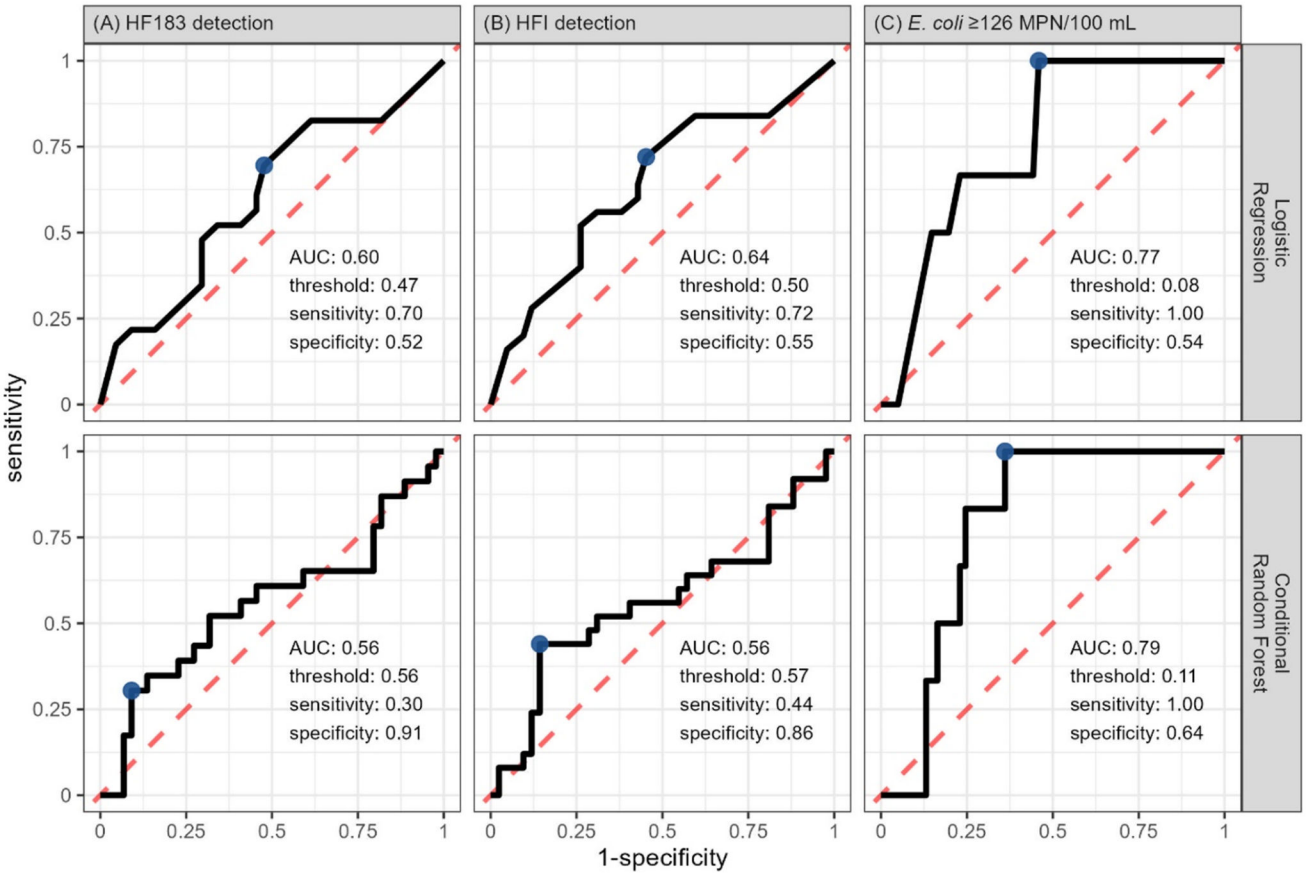


Figure 3.

Receiver operating characteristic (ROC) curves (black lines) for logistic regression (top row) and conditional random forest (CRF, bottom row) model predictions of HF183 (A), human fecal indicator (HFI; HF183 and/or FRNA GII coliphage) (B), and *E. coli* ≥ 126 MPN/100 mL (C) in the test data set (2015–2016). The area under the curve (AUC) summarizes overall predictive performance, and the classification threshold is the predicted probability that minimizes misclassification, corresponding to the blue point on the ROC curve. The red-dashed line represents the performance of an unskilled classifier (no discriminatory ability) with an AUC of 0.5.

Table 1.

Fecal Indicator Occurrence and Building Presence by Pond in the Training (2020–2021) and Test (2015–2016) Data Sets

data set	pond	building (Y/N)	no. HF183 detection (%)	no. human-associated phage[a] detection (%)	no. HF183 and phage codetection (%)	no. generic *E. coli* ≥126 MPN/100 mL (%)
training	A1	Y	12 (44)	3 (11)	2 (7)	2 (7)
	A2	Y	7 (26)	1 (4)	1 (4)	1 (4)
	A3	Y	14 (52)	6 (22)	5 (19)	4 (15)
	A4	Y	26 (96)	0 (0)	0 (0)	2 (7)
	B1	N	1 (4)	0 (0)	0 (0)	1 (4)
	B2[b]	N	2 (7)	0 (0)	0 (0)	2 (7)
	B3	N	2 (7)	0 (0)	0 (0)	4 (15)
	B4	Y	7 (26)	4 (15)	2 (7)	6 (22)
test	LV[c]	Y	8 (35)	2 (9)	1 (4)	0 (0)
	NP	Y	8 (36)	2 (9)	1 (5)	4 (18)
	SC	Y	7 (32)	3 (14)	3 (14)	2 (9)

[a] crAssphage was assessed in the training data set and FRNA GII coliphage was assessed in the test data set.

[b] Sample size was 27 for each training data set pond except B2, from which 28 samples were collected.

[c] Sample size was 22 for test data set ponds NP and SC and 23 for pond LV.