Supporting Information:
Performance of conditional random forest and regression models at predicting human fecal contamination of produce irrigation ponds in the southeastern United States

Jessica Hofstetter[1,2,3], David Holcomb[1], Amy Kahler[1], Camila Rodrigues[3], Andre Luiz Biscaia Ribeiro da Silva[3], Mia Mattioli[1]

1. Waterborne Disease Prevention Branch, Centers for Disease Control and Prevention, Atlanta, GA 30333, USA
2. Chenega Enterprise Systems & Solutions, LLC, Chesapeake, VA 23320, USA
3. Department of Horticulture, Auburn University, Auburn, AL 36849, USA

Table of Contents

S1 Building Septic Systems and Proximity to Training Dataset Ponds

The Georgia Department of Public Health *Manual for On-Site Sewage Management Systems* specifies that the residential trench absorption field absorption area for a two-bedroom house should be 500 feet^2, and for a commercial building should be 2190 feet^2.[1] Therefore, a distance of a building within 500 feet or a commercial building within 2000 feet was chosen as the criteria for if a building was close to the pond or not. Pond A1 had a commercial septic system rated at the capacity of <2000 gallons per day approximately 670 feet away and a residential building approximately 420 feet away (the commercial building had a septic record, the closer building did not). Pond A2 was approximately 500 feet from the nearest building with a field in between (septic system was recorded as single-family residence with 2 bedrooms), Pond A3 was approximately 100 feet from the nearest building (no septic record for this building). Pond A4 was approximately 1200 feet from a commercial processing facility (septic record for <2000 gallons/day) and approximately 500 feet from a building with no septic record. Ponds B1-B3 did not have any buildings located within 500 feet or commercial buildings within 2000 feet. Pond B4 had several buildings within 400 feet but no publicly available septic record.

S2 Explanatory Variable Descriptive Statistics

Table S1. Descriptive statistics for continuous explanatory variables in the training dataset (2020 – 2021) and test dataset (2015 – 2016).

Explanatory variable	Training dataset		Test dataset	
	Mean (SD)*	**Range**	**Mean (SD)**	**Range**
Rain 0-2, inches	0.2 (0.5)	0.0 - 2.3	0.4 (0.6)	0.0 - 1.7
Rain 2-7, inches	0.7 (1.0)	0.0 - 5.8	0.9 (1.1)	0.0 - 4.1
Solar 0-2, $\log_{10}(MJ/m^2)$	2.8 (0.2)	2.1 - 3.1	2.7 (0.3)	1.7 - 3.1
Solar 2-7, $\log_{10}(MJ/m^2)$	3.2 (0.2)	2.8 - 3.4	3.2 (0.2)	2.8 - 3.5
Wind 0-2, mph	6.1 (2.4)	3.2 - 12.0	7.9 (2.6)	4.3 - 12.2
Wind 2-7, mph	16.5 (6.4)	9.2 - 30.7	17.8 (4.8)	9.9 - 27.5
Temperature, °C	22.0 (5.4)	11.0 - 34.6	22.4 (6.5)	10.6 - 31.6
pH	8.4 (0.8)	6.7 - 11.2	7.7 (0.6)	6.8 - 8.8
Dissolved oxygen, mg/L	8.2 (2.8)	0.9 - 15.6	10.1 (2.3)	6.5 - 16.3
Conductivity, $\log_{10}(\mu S/cm)$	2.3 (0.1)	1.8 - 2.8	2.2 (0.2)	1.8 - 2.5
Turbidity, $\log_{10}(NTU + 1)$	1.1 (0.5)	0.0 - 2.5	0.9 (0.5)	0.0 - 2.0

*SD, standard deviation; Range, minimum value to maximum value

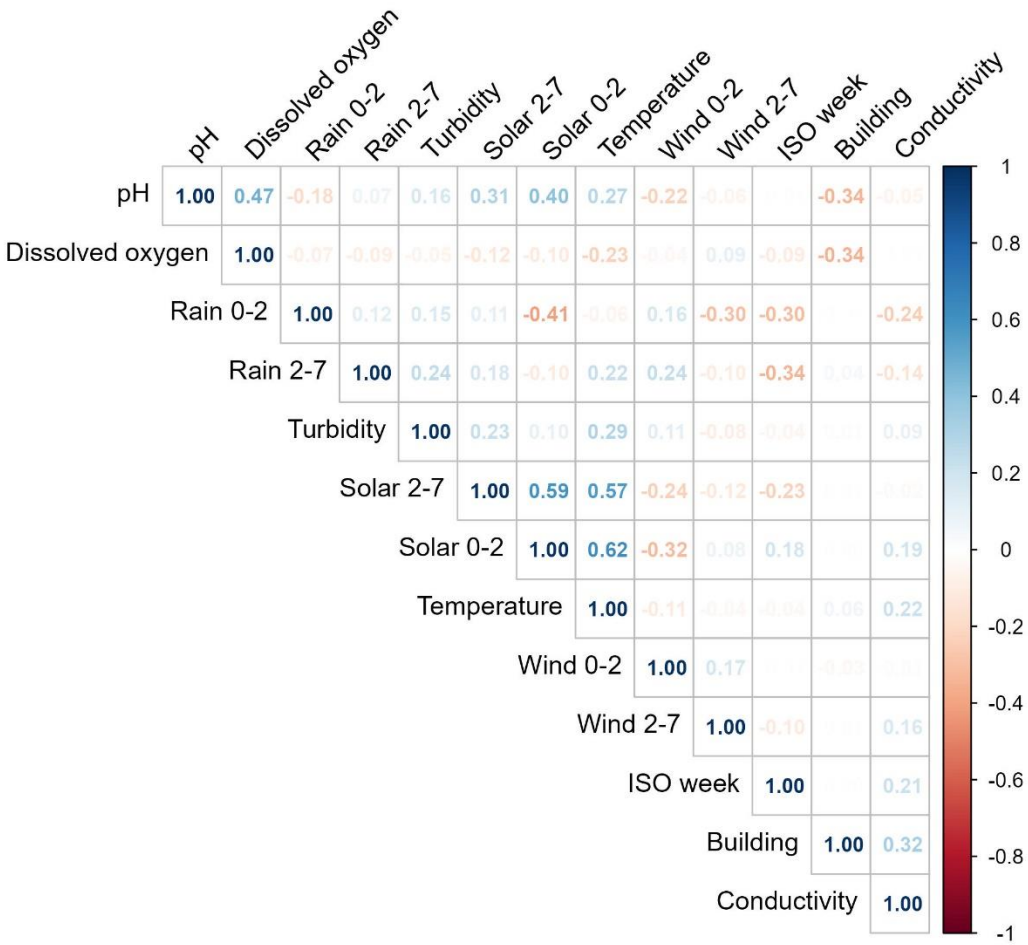

Figure S1. Pairwise Pearson correlations for explanatory variables. Blue and red indicate positive and negative correlation, respectively, with darker shades indicating greater correlation magnitude.

S3 Univariable Logistic Regression Analysis

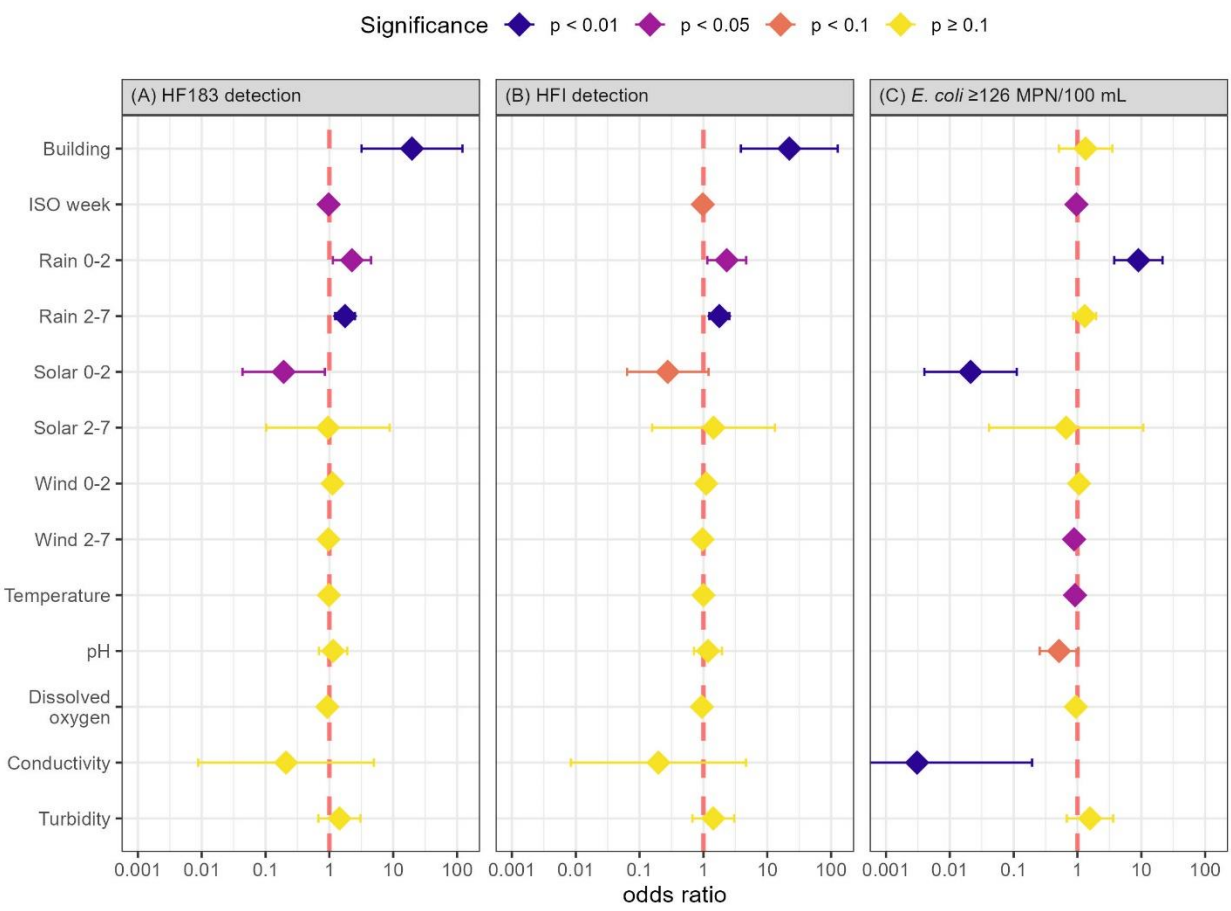

Figure S2. Odds ratio (95% confidence interval) estimates for explanatory variables in univariable mixed effects logistic regression models for the three fecal indicators, HF183 (A), human fecal indicator (HFI; HF183 and crAssphage) (B), and *E. coli* ≥126 MPN/100 mL (C).

S4 Elevated *E. coli* as a Predictor of Human Fecal Contamination

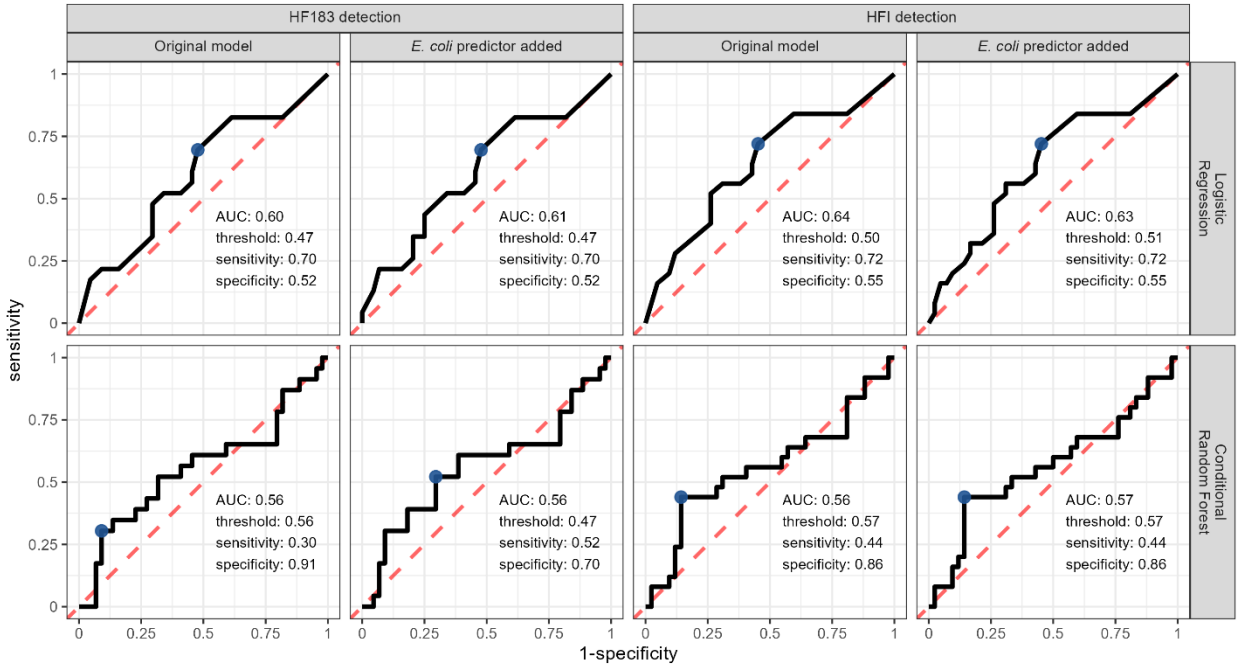

Figure S3. Receiver operating characteristic (ROC) curves (black lines) for logistic regression (top row) and conditional random forest (CRF, bottom row) model predictions of HF183 and human fecal indicator (HFI; HF183 and/or FRNA GII coliphage) in the test dataset (2015-2016) comparing the predictions from the models as originally specified to the same models retrained with *E. coli* ≥126 MPN/100 mL included as an additional explanatory variable.

S5 Hyperparameter Tuning and Class Imbalance Correction

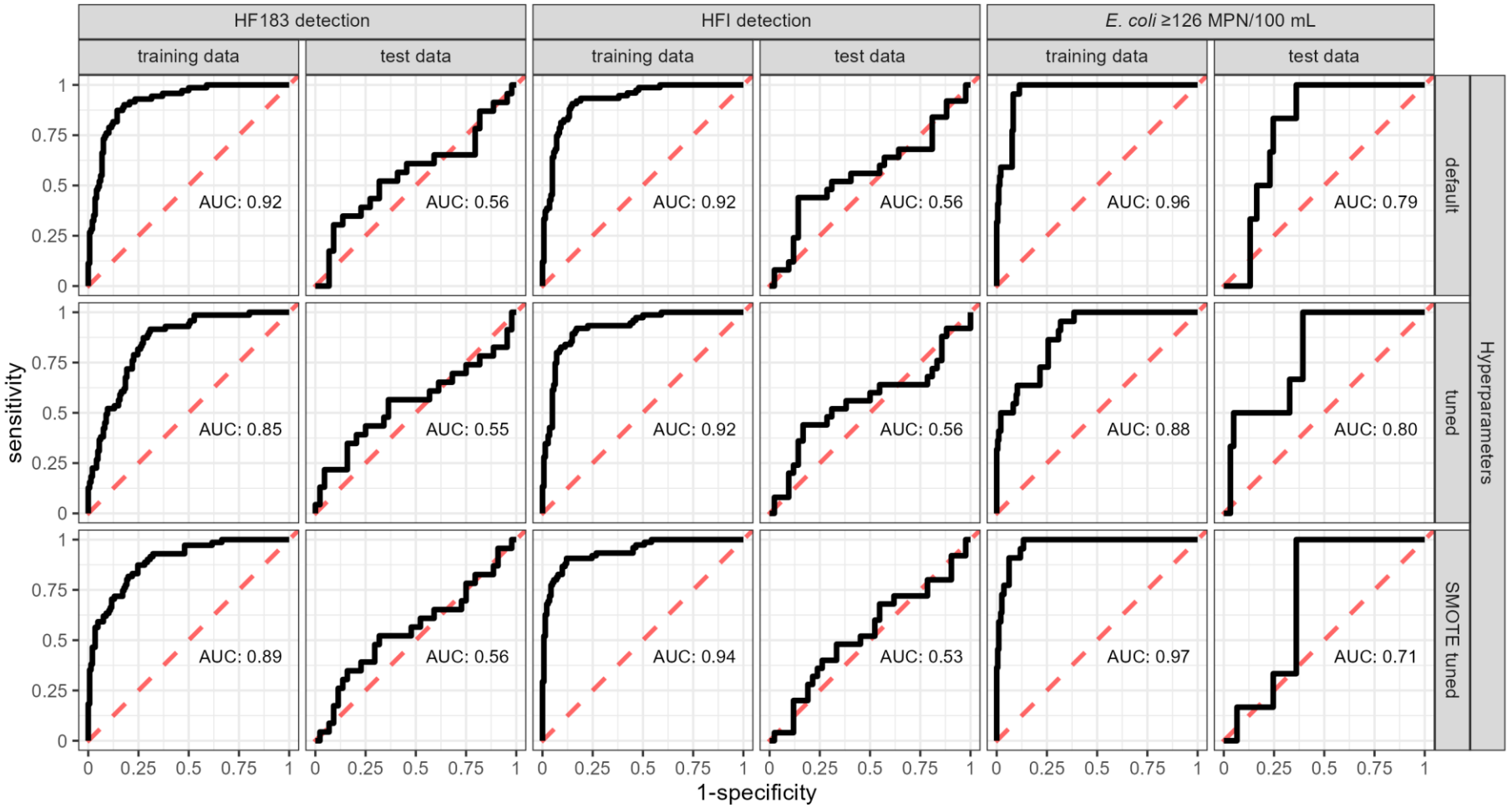

Figure S4. Receiver operating characteristic (ROC) curves (black lines) for predictions of both the training and test datasets by conditional random forest models using default hyperparameters (top row), hyperparameter values tuned by repeated three-fold cross-validation (middle row), and hyperparameters tuned on training data corrected for outcome class imbalance using synthetic minority oversampling technique (SMOTE).

S6 Supplemental References

(1) 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 (accessed 2024-07-09).