CENTER for PRODUCE SAFETY

CPS 2016 RFP
FINAL PROJECT REPORT

Project Title
Remotely-sensed and field-collected hydrological, landscape and weather data can predict the quality of surface water used for produce production

Project Period
January 1, 2017 – December 31, 2018 (extended to January 31, 2019)

Principal Investigator
Martin Wiedmann
Cornell University
Department of Food Science
341 Stocking Hall
Ithaca, NY 14853
T: 607-254-2838
E: mw16@cornell.edu

Co-Principal Investigator
Channah Rock
University of Arizona
Dept. of Soil, Water and Environmental Science
37860 W. Smith Enke Rd
Maricopa, AZ 85138
T: 520-374-6258
E: channah@cals.arizona.edu

Objectives
1. *Perform sampling on 4* streams in New York and 4* canals in Arizona throughout one growing season to examine changes in generic E. coli levels as well as Listeria monocytogenes, Salmonella, and STEC presence at a fine temporal scale (daily/weekly); data will be analyzed as part of Obj. 3.* (*Revised to 6 in NY and 2 in AZ)
2. *Perform sampling on 30* streams in NY and 30* canals in AZ throughout one growing season to allow assessment of spatial and temporal variation in generic E. coli levels, and L. monocytogenes, Salmonella, and STEC presence; data will be analyzed as part of Obj. 3.* (* Revised to 60 each)
3. *Identify and prioritize consistent and region-specific landscape, weather and hydrological factors that are associated with generic E. coli levels, as well as the presence of L. monocytogenes, Salmonella, and STEC in surface water, and to use these data to (i) develop and compare a series of geospatial models to predict surface water quality in NY and AZ, and (ii) quantify the association between generic E. coli levels and pathogen presence in surface water used for produce production.*

Funding for this project provided by the Center for Produce Safety through:
CDFA SCBGP grant# 16070

FINAL REPORT

Abstract

Water used for produce production has emerged as a focus of produce safety programs. However, interpreting water quality tests is difficult because the relationship between indicator organisms and pathogens varies between studies, and microbial water quality varies over time. Alternative approaches that account for this variation are thus needed to improve growers' ability to identify and address produce safety risks associated with preharvest surface water use. This project was designed to support development of these approaches by (i) assessing the association between *E. coli* levels, and foodborne pathogen presence in surface water used for produce production, (ii) identifying and ranking factors associated with pathogen detection in surface water, and (iii) developing models to predict when and where surface water sources are likely to be contaminated by pathogens. In this study, water quality was assessed in two produce-growing regions (Arizona [AZ] and New York [NY]), by enumerating *E. coli* levels, and testing for key pathogens (enterohemorrhagic *E. coli* [EHEC]*, Salmonella*, and *Listeria monocytogenes*) in surface water samples. In 2017, 2 AZ and 6 NY waterways were sampled longitudinally to (i) characterize the relationship between *E. coli* levels and likelihood of pathogen presence in surface water, and (ii) identify and rank environmental factors associated with microbial water quality. In 2018, 60 AZ and 68 NY waterways were sampled to (i) identify spatial factors associated with pathogen detection in surface water samples, and (ii) develop models that predict when and where pathogen contamination of surface water is likely to occur.

In this study, microbial water quality varied over space and time. For instance, PCR-based EHEC detection in AZ was substantially higher in 2017 (51%; 43/83) than in 2018 (22%; 39/178). Regression analysis showed that *Salmonella* isolation in AZ ($P = 0.009$) and EHEC detection in AZ ($P = 0.002$) and NY ($P < 0.001$) were significantly associated with the $\log_{10}$ MPN of *E. coli* per 100 mL, but *L. monocytogenes* and *Salmonella* isolation in NY were not associated with *E. coli* levels ($P > 0.050$). To determine if meeting *E. coli*-based agricultural water standards was associated with a reduced likelihood of pathogen presence in surface water, we simulated water sampling using bootstrapping (N=10,000 20-sample subsets). Our findings suggest that meeting these standards is (i) largely a function of when the samples that comprise a subset were collected, and (ii) a poor approximation of microbial water quality at the time of water use. As such, alternative approaches are needed to improve growers' ability to identify food safety risks associated with preharvest surface water use in real-time.

In 2017, two sampling methods were used to detect pathogens in AZ and NY surface water, 24-h Moore swabs (MS) and 10-L grab samples (GS). Regression analysis indicated that the odds of a detecting *Salmonella* ($P = 0.001$) and EHEC ($P < 0.001$) using a MS was significantly greater than the odds of detecting *Salmonella* and EHEC using a single paired GS. For example, the odds of isolating *Salmonella* from MS were 2.2 times greater than the odds of isolating *Salmonella* from a single paired GS. Conversely, the odds of a detecting *L. monocytogenes* using a MS was significantly lower than the odds of detecting *L. monocytogenes* using a single paired GS ($P = 0.041$). Overall, our findings indicate that appropriate water sampling methods may depend on the organism of concern (*Salmonella* versus *Listeria*).

This study also found that microbial water quality is associated with environmental heterogeneity (i.e., changes in and interactions between environmental factors). According to random forest (RF) analysis, weather and water quality factors (e.g., *E. coli* levels and rainfall, dissolved oxygen and solar radiation) interacted to affect the likelihood of detecting pathogens in surface water. However, untangling the nature of these interactions is difficult due to correlation between weather and water quality factors. The produce safety risks associated with preharvest surface water use therefore appear to be dependent on environmental conditions at the time of water use. Although our study found that interactions between environmental factors were

associated with microbial water quality, these findings also indicate that some factors may be useful as supplemental indicators of microbial water quality (e.g., as a covariate in models that predict the likelihood of pathogen contamination). Indeed, the best geospatial models that predict pathogen presence incorporated multiple data types (e.g., weather, water quality, spatial). For example, a model that predicts EHEC presence in AZ canals used the MPN of *E. coli*/100 mL and average solar radiation to identify samples with a higher or lower likelihood of testing positive for EHEC.

Overall this study provides data that growers can use to generally assess food safety risks associated with preharvest surface water use (e.g., a ranked list of risk factors that are associated with a higher risk of surface water being pathogen contaminated). This study also provides data that stakeholders can use to optimize water sampling protocols, such as information on the relative ability of 24-h MS compared to 10-L GS to detect different pathogens in surface water.

Background

Preharvest surface water use has been repeatedly identified as a potential source of foodborne pathogens on produce farms (1–4). For instance, irrigation with untreated surface water has been associated with the isolation of foodborne pathogens from preharvest environments (1, 5, 6), and identified as a potential cause of produce-associated outbreaks (4, 7–9). Low-level preharvest contamination can also result in contamination of post-harvest environments, where pathogens can persist and contaminate product (10). Recognizing the role of surface water as a source of and pathway for produce contamination, the FDA proposed microbial standards for agricultural water in the Produce Safety Rule (PSR). The proposed agricultural water standard (AWS) states that *E. coli* levels in water applied to preharvest produce cannot exceed a geometric mean of 126 CFU/100 mL and a statistical threshold of 410 CFU/100 mL. However, understanding and complying with the AWS has been cited as a key challenge facing growers (11–13). This may partially be because temporal variation in the microbial quality of surface water (14–16) and conflicting data on the relationship between *E. coli* levels and pathogen presence (17–21) complicate the interpretation of *E. coli*-based water quality tests. Past studies have suggested that variability in microbial water quality may be due to the heterogeneity inherent to farm and freshwater environments [e.g., changes in weather and physiochemical water quality over time (17, 18)]. Thus, targeted approaches that account for this heterogeneity may improve growers' ability to identify and address produce safety risks associated with preharvest water use; however, additional data is needed to develop these approaches. This project addressed these needs by (i) assessing the association between *E. coli* levels and the presence of key pathogens (*Listeria monocytogenes*, *Salmonella,* enterohemorrhagic *E. coli* (EHEC), and enteropathogenic *E. coli* [EPEC]) in AZ and NY waterways used for produce production, (ii) identifying and ranking factors associated with pathogen detection in these waterways, and (iii) developing models to predict when and where AZ and NY waterways are more likely to be contaminated by foodborne pathogens.

Research Methods

Obj. 1: Perform sampling on 6 NY and 2 AZ waterways

Six NY and 2 AZ waterways were sampled longitudinally between February and December 2017. Sampling was performed to assess variation in water quality over the (i) growing season, (ii) course of a week ("weekly sampling"), and (iii) course of a day ("daily sampling"). At each sampling, a Moore swab (MS) set and a grab sample (GS) set were collected. Each GS set

consisted of three 10-L GS (i.e., one GS for detection of each of the 3 target pathogen groups [*Listeria, Salmonella,* and pathogenic *E. coli*]) and one 1-L GS for determining the MPN of *E. coli*/100 mL. Each MS set consisted of three MS (i.e., one MS for detection of each of the target pathogen groups). On the first day of weekly sampling, a MS set was anchored in the waterway, and a GS set was collected. When the MS set was collected 24 h later, a second GS set was collected and a second MS set was placed in the waterway. This was repeated daily for up to 6 days. During daily sampling, a MS set was anchored in a waterway for 24 h, and during this time six GS sets were collected between 5 am and 8 pm. Weekly sampling was performed on 2 AZ canals and all 6 NY streams three times in 2017. Daily sampling was performed on 2 AZ canals and 5 NY streams three times in 2017. Daily sampling was performed on fewer waterways than weekly sampling due to the substantial time needed to perform a single daily sampling. While daily sampling was performed once on a third canal, this canal was removed from the study for the aforementioned reasons. In total, 83 GS sets and 33 MS sets were collected in AZ, and 181 GS sets and 88 MS sets were collected in NY.

Waterway Enrollment: Remotely-sensed data (e.g., flow accumulation raster) were obtained from publicly available databases to facilitate waterway enrollment. Hydrological, land use and other spatial data were downloaded from federal and state geodata portals (e.g., https://nassgeodata.gmu.edu/CropScape/). NY streams were enrolled by identifying watersheds (i) with an area of ≥15 km^2 and (ii) where produce had been grown in ≥4 of the last 8 years. Six publicly accessible sites that were along streams in these watersheds and were ≤400 m from a produce field were randomly selected and enrolled in this study. Publicly accessible sites were sites near stream-road intersections, on public or university lands (e.g., parks, Cornell farms), or with public-right-of-way (e.g., fishing access). AZ canals were enrolled by identifying canals in produce-growing regions and then selecting three sites along these canals that were ≤400 m from a produce field and were accessible to the sampling team.

Metadata Collection: Physical characteristics for each waterway (e.g., sediment substrate size) were recorded on the first day of sampling; any changes during the course of the study were noted. Data on water quality, flow and weather were collected each time a sample was collected. Specifically, dissolved oxygen (DO) levels, pH, conductivity, and water temperature were measured in-field using a Hach HQ40d meter; turbidity was measured in lab using a Hach turbidimeter. While a flow meter was used to measure flow rate in NY, the float method was used to estimate surface flow in AZ (22). Meteorological data were obtained from the AZMet or NEWA weather station closest to each sampling site. Solar radiation, temperature, rainfall, relative humidity, and wind speed data were obtained for the duration of the study.

Grab Sample Processing: The 10-L GS were filtered using modified Moore swabs [mMS; (23)]. After all 10 L of water was filtered, each mMS was transferred to a separate Whirl-Pak bag and processed as described below. A 100-mL aliquot of the 1-L GS was used for *E. coli* enumeration, which was performed using the IDEXX Colilert Quanti-Tray 2000 kit per manufacturer instructions.

Listeria Isolation: *Listeria* enrichment was performed as previously described (1, 5). Briefly, 225 mL of buffered *Listeria* enrichment broth was added to each Whirl-Pak bag containing a MS or mMS. Following incubation at 30°C for 4 h, *Listeria* selective enrichment supplement was added to each enrichment. After incubating at 30°C for a total of 24 h and 48 h, 50 µl of each enrichment was streaked onto *L. monocytogenes* plating medium (LMPM) and Modified Oxford agar (MOX); the plates were incubated for 48 h at 35°C and 30°C, respectively. Following incubation, ≤4 presumptive *Listeria* colonies were sub-streaked from MOX to LMPM and

incubated at 35°C for 48 h. The species of one presumptive *Listeria* colony (excluding *L. monocytogenes*) and one presumptive *L. monocytogenes* colony per sample was determined by PCR amplification and sequencing of the *sigB* gene (24, 25).

Salmonella Isolation: Two-hundred and twenty-five mL of buffered peptone water supplemented with novobiocin to a concentration of 20 mg/L (BPW+N) was added to each Whirl-Pak containing a MS or mMS. Following incubation at 35°C for 24 h, *Salmonella* negative samples and presumptive *Salmonella*-positive samples were identified using real-time BAX *Salmonella* assays. BAX negative samples were considered *Salmonella* negative, while BAX positive samples were culture-confirmed as *Salmonella* positive as previously described (5). Briefly, 1 mL of the BPW+N enrichment was added to 9 mL of tetrathionate (TT) broth supplemented with 200 uL of I_2-KI and 100 uL of Brilliant Green. In parallel, 0.1 mL of the BPW+N enrichment was added to 9.9 mL of Rappaport Vassiliadis (RV) broth. After incubating the TT and RV broths at 42°C in a shaking water bath for 24 h, 50 uL of each broth was streaked onto Salmonella CHROMagar and xylose lysine deoxycholate agar (XLD) plates. The CHROMagar and XLD plates were incubated for 24 h at 37 and 35°C, respectively. Following incubation, ≤12 presumptive *Salmonella* colonies per sample were confirmed as *Salmonella* by PCR amplification of the *invA* gene (26).

Pathogenic *E. coli* Detection: Two-hundred and twenty-five mL of tryptic soy broth supplemented with casamino acids and novobiocin (TSB+N) to a final concentration of 10 g/L and 8 mg/L, respectively, was added to each Whirl-Pak containing a MS or mMS. Following incubation at 41°C for 24 h, EHEC negative samples, presumptive EHEC positive, and presumptive EPEC positive samples were identified using real-time BAX EHEC assays. Samples that were BAX negative were considered EHEC negative, while BAX positive samples initially underwent culture-confirmation for EHEC. Briefly, 100 uL of the TSB+N enrichment was transferred to 9.9 mL of modified buffered peptone water with pyruvate. After incubating at 37°C for 5 h, 10 uL of a 10 mg/L acriflavin solution, 10 uL of a 10 mg/L cefsulodin solution, and 8 uL of a 10 mg/L vancomycin solution were added. After incubation at 41°C for 24 h, 50 uL of the secondary enrichment was streaked onto sorbitol MacConkey plates and STEC CHROMagar plates. Following incubation, ≤24 presumptive EHEC colonies were substreaked onto brain heart infusion (BHI) plates; the BHI plates were incubated at 37°C for 24 h. Samples positive for EHEC were then confirmed by PCR amplification of the partial *stx1, stx2* and *eaeA* genes (27, 28). After processing ~1/3 of all samples, none of the 108 BAX positive samples were culture-confirmed as EHEC positive. As this was consistent with previous publications that reported a high false-negative rate for culture-based EHEC detection methods (29–32), we decided that after July 2017 the culture confirmation of BAX positive samples would not be performed. We therefore used data on the frequency of samples that were PCR-screen positive for EHEC and EPEC instead of frequency of EHEC or EPEC isolation in the analyses reported here.

Obj. 2: Perform sampling on 60 NY streams and 60 AZ canals

60 AZ and 68 NY waterways were sampled between February and October 2018. Due to reduced water flow and loss of access (e.g., construction, development of unsafe conditions), 2 of the AZ canals could only be sampled two times each. Similarly, 8 of the 60 NY streams initially enrolled in the study had to be replaced after two samplings. All replacement sites were sampled three times each. A total of 178 GS sets were collected in AZ (i.e., 58 sites were sampled three times, 2 sites were sampled two times), and 196 GS sets were collected in NY (i.e., 60 sites were sampled three times, 8 sites were sampled two times). AZ canals were enrolled by identifying and randomly selecting 60 sites along canals that were ≤400 m from a produce field and were accessible to the AZ sampling team. NY streams were enrolled as described in the Obj. 1 methods section. GS and metadata collection (e.g., physiochemical water quality and weather

data), GS processing, pathogen detection, and *E. coli* enumeration were also performed as described in the Obj. 1 methods section. Due to the low prevalence of *Listeria* in AZ in 2017 (4%), *Listeria* enrichment and isolation were not performed in AZ in 2018.

Obj. 3: Analysis of data collected as part of Obj. 1 and Obj. 2

All analyses were performed in R. Correlation between environmental factors was quantified and visualized as previously described (33, 34). The prevalence of each of the target organisms (*Listeria* spp. [including *L. monocytogenes*], *L. monocytogenes*, *Salmonella*, EHEC, and EPEC) as well as the median and geometric mean *E. coli* level (MPN/100 mL) were calculated for each state and year.

NOTE: An in-depth analysis of the data collected as part of Obj.1 will be made available on the CPS website in a Draft Manuscript (as of August 2019) and a final Publication (when available), and includes supplementary figures that illustrate the results discussed below. The supplementary figures and tables are also available from the Principal Investigator upon request.

Comparison of Pathogen Detection by MS and Paired GS: As part of Obj.1, we collected two sample types (24-h MS and 10-L GS). For each sampling day, we determined if a contamination event had occurred. A waterway was considered contaminated by a given organism if the MS or one of its paired GS tested positive for the organism; each MS had between 2 and 7 paired GS. We used generalized linear mixed models (GLMMs) to determine if MS was significantly more or less likely to detect the target organisms than a paired GS. Since the outcome of the GLMMs was the presence or absence of the target organism we used a binomial distribution with a logit link. The explanatory variable was sample type (GS was the reference level). Site nested in state and year day were included as random effects; year day is the numerical day of the year (e.g., Jan. 1st is year day 1, Jan. 2nd is year day 2). Since the ability of a MS compared to a paired GS to detect foodborne pathogens in a given waterway at a given time should not differ between states, AZ and NY data were combined for these analyses.

Relationship Between *E. coli* Levels and Pathogen Detection: GLMMs (35) were developed to characterize the relationship between *E. coli* levels and likelihood of pathogen detection using the data collected as part of Obj. 1. Since the outcome of the GLMMs was the presence or absence of the target organism we used a binomial distribution with a logit link. The $\log_{10}$ MPN of *E. coli*/100 mL was included as a fixed effect. Site and year day were included as random effects. Bootstrapping was used to simulate water sampling to create a microbial water quality profile (MWQP) composed of 20 samples (N=10,000 MWQPs per waterway), and to quantify the ability of the proposed AWS [geometric mean <126 CFU/100 mL and STV <410 CFU/100 mL; (36)] to identify water sources with a high or low risk of pathogen presence at the time the water source is used for produce production. The last GS selected for inclusion in each subset (i.e., the 20th sample selected) represented microbial water quality at the time of water use for the given waterway (e.g., if the 20th GS selected was positive for *Salmonella* then the water source was considered positive for *Salmonella* at the time of water use). The sensitivity, specificity, and diagnostic odds ratio were then calculated for each target organism in each state. The AZ and NY data were analyzed separately since differences in environmental conditions and water type (managed canals versus free-flowing streams) may affect the relationship between pathogen detection and *E. coli* levels.

Random Forest Analysis: Using the data collected as part of Obj. 1, random forest (RF) analysis was performed to separately identify and rank weather and physiochemical water quality factors associated with detection of each target pathogen in each sample type (MS and GS) (see Tables S1 and S2 in the Draft Manuscript for a full list of factors). RF was also performed to rank factors

associated with *E. coli* levels. The AZ and NY data were analyzed separately to allow for identification of region-specific factors. Five, overlapping time frames (0-1, 0-2, 0-3, 0-4, or 0-5 d before sample collection) were used to calculate the values of the weather factors with the exception of rainfall. Separate RF were then developed for each outcome and time period in each state. Unbiased conditional RF was performed using the party package (37). Repeated, 10-fold cross-validation was performed to tune hyperparameters and calculate the Kappa score (38). The RFs with the highest Kappa score for each outcome are reported here. RF results are interpreted by quantifying conditional variable importance (VI). A higher VI, relative to all other factors included in the RF, indicates a stronger association between the outcome and factor. A VI ≤0 indicates no association with the outcome. Since VI is relative, normalized VI (NVI) was calculated to facilitate interpretation and visualization of RF results. Partial dependence plots (PDPs) were developed to graphically characterize (i) relationships between top-ranked factors and RF outcomes (one-way PDPs), and (ii) the impact of two-way interactions between factors on RF outcomes (39). Interactions were defined as occurring if the marginal effect of a factor on the outcome was not constant over all values of a second factor (40). The y-axis of each one-way PDP shows the marginal effect of the given factor on the RF outcome.

Geospatial Model Development: Geospatial predictive models were developed using the data collected as part of Obj 2. Separate conditional inference trees (CTrees) were developed to predict the likelihood of foodborne pathogen contamination of surface water in AZ and NY. Weather, physiochemical water quality, and spatial factors were included as potential predictors in the CTrees. Six time frames (0-1, 1-5, 5-10, 10-20, 20-30, and 30-60 d before sample collection) were used to calculate the values for all weather factors in AZ and NY. Five overlapping distance ranges (0-100 m, 0-250 m, 0-500 m, 0-1 km, and 0-2 km) were used to calculate values of spatial factors in AZ, and 8 overlapping distance ranges (0-100 m, 0-250 m, 0-500 m, 0-1 km, 0-2 km, 0-5 km, 0-10 km, 0-15 km, and 0-stream source [i.e., total watershed area]) were used to calculate values of spatial factors in NY. Different distance ranges were used in AZ and NY due to differences between canals and streams (e.g., streams have well-delineated watersheds but canals do not). CTrees were developed using the party package (37). Repeated, 3-fold cross-validation was performed to tune hyperparameters and perform internal validation (38). The ability of the CTree to correctly predict pathogen status levels for a de novo dataset (i.e., external validation) was assessed using the data collected as part of Obj. 1. If external validation returned an accuracy score <60% then a new CTree was developed using the combined Obj. 1 and Obj. 2 data.

Research Results

In total, 17,868 liters of water were collected and analyzed in this study, including 453 10-L GS used for *Listeria* isolation, 632 10-L GS used for *Salmonella* isolation, 638 10-L GS used for pathogenic *E. coli* detection, and 638 1-L GS used for enumeration of *E. coli* levels (**Table 1**). Additionally, 362 MS were collected and analyzed for pathogen presence (120 for *Listeria*, 121 for *Salmonella*, and 121 for pathogenic *E. coli*). Different numbers of samples were analyzed for different organisms due to loss of samples in the field (e.g., MS were lost during storms and due to human tampering). Additionally, due to the low prevalence of *Listeria* in AZ in 2017, *Listeria* detection was not performed in AZ in 2018 as part of Obj. 2. As a result, we have data on pathogenic *E. coli* for 203 sampling days (121 d in 2017; 82 d in 2018), on *Salmonella* for 202 sampling days (120 d in 2017; 82 d in 2018), and *Listeria* for 161 sampling days (120 d in 2017; 41 sampling d in 2018) (**Table 2**).

E. coli levels: *E. coli* levels ranged between <1.0 and >2,419.6 MPN/100 mL in AZ (Geometric Mean [GM] = 26.5; Median = 25.6), and between 2.0 and >2419.6 MPN/100-mL in NY (GM = 198.2; Median = 193.5). According to RF analysis, *the top-ranked factors associated with E. coli levels in AZ in 2017 were site (i.e., the waterway being sampled), dissolved oxygen (DO), and average (avg.) and minimum (min.) air temperature 0-5 d before sample collection (BSC)* (**Fig. 1**). PDPs indicate that, on average, *E. coli* levels in the sampled canals (i) decreased as DO increased from 7 to 10 mg/L, and (ii) increased as avg. and min. air temperature 0-5 d BSC increased from 13 to 33°C and from 3 to 24°C, respectively. *The top-ranked factors associated with E. coli levels in NY in 2017 were turbidity, flow rate, pH, and min. air temperature 0-5 d BSC.* On average, *E. coli* levels in the sampled streams increased as (i) turbidity increased from 0 to 50 NTUs, (ii) flow rate increased from 0.0 to 1.0 m/s, and (iii) min. air temperature 0-5 d BSC increased from 5 to 18°C. On average, *E. coli* levels in NY decreased as pH increased from 7.0 to 8.5.

Listeria: In 2017, *Listeria* spp. (including *L. monocytogenes*) was isolated from 4% (3/76) of GS collected in AZ and from 47% (85/181) of GS collected in NY, while *L. monocytogenes* was isolated from 4% (3/76) of GS collected in AZ and 15% (27/179) of GS collected in NY (**Table 1**). *Listeria* was isolated from 0 of the 34 MS collected in AZ and from 27% (23/86) of MS collected in NY; *L. monocytogenes* was isolated from 7% (6/86) of MS collected in NY. In 2017, *Listeria* spp. was isolated from samples collected on 72 of the 120 sampling days. Of these 72 contamination events, two were detected by MS only (i.e., all paired GS were *Listeria* spp. negative), while 49 were detected by one or more of the paired GS but not by the MS (**Table 2**). According to GLMM, the odds of isolating *Listeria* spp. from MS was 4.2 times lower than the odds of isolating *Listeria* spp. from a single paired GS (Odds Ratio [OR] = 0.24; 95% Confidence Interval [CI] = 0.12, 0.48; $P < 0.001$; Table 2). Similarly, the odds of isolating *L. monocytogenes* from MS was 2.6 times lower than the odds of isolating *L. monocytogenes* from a single paired GS (OR = 0.38, 95% CI = 0.15, 0.96; $P < 0.041$; Table 2).

While *Listeria* detection was not performed in AZ in 2018, 38% (75/196) of GS collected in NY in 2018 were *Listeria* spp. positive. Due to the low prevalence of *Listeria* in AZ in 2017, RF and CTrees were not developed for *Listeria* in AZ. According to RF analysis, *the top-ranked factors associated with Listeria spp. isolation in NY GS were site, year day, avg. wind speed 0-1 d BSC, and water temperature* (**Fig. 1**). On average the likelihood of *Listeria* spp. isolation was highest for GS collected from Stream E and lowest for GS collected from Streams A and C. The likelihood of *Listeria* spp. isolation was constant from May to July but increased from July to September. Additionally, the likelihood of *Listeria* spp. isolation from GS (i) increased as avg. wind speed 0-1 d BSC increased from 0 to 15 kmph, and (ii) decreased as water temperature decreased from 10 to 23°C. *The top-ranked factors associated with Listeria spp. isolation from NY MS were rainfall 0-1 d BSC, min. and avg. air temperature 0-5 d BSC, and flow rate*.

According to RF analysis, *flow rate was among the 4 top-ranked factors associated with L. monocytogenes isolation from NY GS and MS* (**Fig. 1**). The likelihood of *L. monocytogenes* isolation from NY GS decreased as flow increased from 0.0 to 1.0 m/s, while the likelihood of *L. monocytogenes* isolation from NY MS increased as flow increased from 0.0 to 1.0 m/s. *The other top-ranked factors associated with L. monocytogenes isolation from GS were year day, rainfall 3-4 d BSC, and avg. relative humidity 0-4 d BSC* (**Fig. 1**); year day is the numerical day of the year (e.g., Jan. 1st is year day 1, Jan. 2nd is year day 2). The likelihood of *L. monocytogenes* isolation from NY GS decreased from May to July and increased from July to September. Additionally, the likelihood of *L. monocytogenes* isolation from NY GS (i) increased as rainfall 3-4 d BSC increased from 0.0 to 2.0 cm, and (ii) decreased as avg. relative humidity 0-4 d BSC increased from 60% to 100%. *In addition to flow rate, the other top-ranked factors associated with L. monocytogenes isolation from MS were pH, and min. and max. air temperature 0-1 d BSC.*

Salmonella: In 2017, *Salmonella* was isolated from 34% (26/77) of GS and 64% (21/33) of MS collected in AZ, and from 44% (80/181) of GS and 57% (50/88) of MS collected in NY (**Table 1**). In 2017, *Salmonella* was isolated from samples collected on 99 of the 120 sampling days. Of these 99 contamination events, 21 were detected by MS only (i.e., all paired GS were *Salmonella* negative), while 28 were detected by one or more of the paired GS but not by the MS (**Table 2**). According to GLMM, the odds of isolating *Salmonella* from MS were 2.2 times greater than the odds of isolating *Salmonella* from a single paired GS (OR = 2.2; 95% CI = 1.36, 3.45; *P* = 0.001; Table 2). In 2018 *Salmonella* was isolated from 37% (65/178) of GS collected in AZ, and 40% (79/196) of GS collected in NY.

According to RF analysis, *the two top-ranked factors associated with Salmonella isolation from AZ GS and MS were max. and avg. air temperature* (**Fig. 1**). The likelihood of *Salmonella* isolation from AZ GS (i) increased as max. and avg. air temperature increased from 20 to 41°C, and from 13 to 30°C, respectively, and (ii) decreased as max. and avg. air temperature increased from 41 to 47°C and from 30 to 38°C, respectively. *The other top-ranked factors associated with Salmonella isolation from AZ GS were weekday and conductivity.* The likelihood of isolating *Salmonella* from AZ GS was highest for samples collected on Tuesdays and Wednesdays. Additionally, the likelihood of isolating *Salmonella* from AZ GS increased as conductivity increased from 750 to 1,300 uS/cm.

The top-ranked factors associated with Salmonella isolation from NY GS were avg. wind speed 0-1 d BSC, max and avg. air temperature 0-1 d BSC, and weekday (**Fig. 1**). The likelihood of *Salmonella* isolation from NY GS decreased as (i) avg. wind speed increased from 0 to 4 kmph, and (ii) avg. and max. air temperature increased from 10 to 19°C, and from 15 to 26°C, respectively. The likelihood of *Salmonella* isolation from NY GS increased as (i) avg. wind speed increased from 4 to 13 kmph, and (ii) avg. and max. air temperature increased from 19 to 26°C and from 26 to 33°C, respectively. The likelihood of isolating *Salmonella* from NY GS was highest for samples collected on Saturdays and lowest for samples collected on Wednesdays and Fridays. *The top-ranked factors associated with Salmonella isolation from NY MS were rainfall 3-4 and 4-5 d BSC, turbidity and year day.*

EPEC: In 2017, 83% (69/83) of GS and 97% (32/33) of MS collected in AZ, and 94% (171/181) of GS and 97% (85/88) of MS collected in NY were PCR-screen positive for EPEC (defined as samples positive for *eaeA;* **Table 1**). In 2017, EPEC was detected in samples collected on 118 of the 121 sampling days. Of these 118 contamination events, 5 were detected by MS only (i.e., all paired GS were EPEC negative), while 3 were detected by one or more of the paired GS but not by the MS (**Table 2**). According to GLMM, the odds of detecting EPEC in a MS was 8.8 times greater than the odds of detecting EPEC in a single paired GS (OR = 8.8; 95% CI = 1.3, 61.5; *P* = 0.028; Table 2). In 2018, 35% (62/178) and 97% (190/196) of GS collected in AZ and NY, respectively, were PCR-screen positive for EPEC.

While RF could not be performed to identify factors associated with detecting EPEC in MS due to the limited number of EPEC-negative MS in both states, RF was performed to identify factors associated with detecting EPEC in GS. According to RF analysis, *the top-ranked factors associated with detecting EPEC in AZ GS were max. and avg. air temperature 0-1 d BSC, water temperature, and turbidity* (**Fig. 1**). The likelihood of detecting EPEC in AZ increased as (i) max. and avg. air temperature 0-1 d BSC increased from 17 to 32°C and from 9 to 24°C, respectively, (ii) water temperature increased from 15 to 26°C, and (iii) turbidity increased from 0 to 8 NTUS. The likelihood of detecting EPEC in AZ decreased as (i) max. and avg. air temperature 0-1 d BSC increased from 32 to 43°C and from 24 to 36°C, respectively, (ii) water temperature increased from 26 to 30°C, and (iii) turbidity increased from 8 to 16 NTUS.

The top-ranked factors associated with detecting EPEC in NY GS were avg. wind speed 0-3 d BSC, pH, min. air temperature 0-3 d BSC, and year day (**Fig. 1**). The likelihood of detecting EPEC in NY increased as (i) avg. wind speed 0-3 d BSC increased from 2.5 to 10 kmph, (ii) min. air temperature 0-3 d BSC increased from 3 to 19°C, and (iii) pH increased from 7.0 to 7.8. The likelihood of detecting EPEC in NY decreased as pH increased from 7.8 to 8.7. Additionally, the likelihood of detecting EPEC in NY increased from May to July and decreased from July to September.

EHEC: In 2017, 48% (44/83) of GS and 91% (30/33) of MS collected in AZ, and 69% (125/181) of GS and 88% (77/88) of MS collected in NY were PCR-screen positive for EHEC (defined as samples positive for *eaeA* and *stx*) (**Table 1**). In 2017, EHEC was detected in samples collected on 113 of the 121 sampling days (**Table 2**). Of these 113 contamination events, 13 were detected by MS only (i.e., all paired GS were EHEC negative), while 9 were detected by one or more of the paired GS but not by the MS. According to GLMM, the odds of detecting EHEC in a MS was 6.7 times greater than the odds of detecting EHEC in a single paired GS according to GLMM (OR = 6.7; 95% CI= 3.2, 14.0; $P < 0.001$; Table 2). In 2018, 22% (39/178) and 68% (133/196) of GS collected in AZ and NY, respectively, were PCR-screen positive for EHEC.

While RF could not be performed to identify factors associated with detecting EHEC in AZ MS due to the limited number of EHEC negative MS in AZ, RF was performed to identify factors associated with detecting EHEC in AZ GS. According to RF analysis, *the top-ranked factors associated with detecting EHEC in AZ GS were avg. solar radiation 0-3 d BSC, and avg., max., and min. air temperature 0-3 d BSC* (**Fig. 1**). The likelihood of detecting EHEC in AZ increased as (i) avg. solar radiation 0-3 d BSC increased from 11 to 31 Ly, (ii) avg. air temperature 0-3 d BSC increased from 10 to 27°C, (iii) max. air temperature 0-3 d BSC increased from 20 to 42°C, and (iv) min. air temperature 0-3 d BSC increased from 1 to 18°C. The likelihood of detecting EHEC in AZ decreased as avg. air temperature and min. air temperature 0-3 d BSC increased from 27 to 36°C and from 18 to 28°C, respectively.

The top-ranked factors associated with detecting EHEC in NY GS were pH, min. air temperature 0-4 d BSC, MPN of E. coli/100 mL, and site (**Fig. 1**). The likelihood of detecting EHEC in the NY GS increased as (i) min. air temperature 0-4 d BSC increased from 5 to 15°C, and (ii) *E. coli* levels increased from 18 to 1,000 MPN/100 mL. The likelihood of detecting EHEC in the NY GS decreased as (i) pH increased from 7.4 to 8.8, and (ii) min. air temperature 0-4 d BSC increased from 15 to 18°C. The likelihood of detecting EHEC was greatest in samples collected from Streams A, F, and E and lowest in samples collected from Stream B. *The top-ranked factors associated with detecting EHEC in NY MS were rainfall 3-4 d BSC, conductivity, flow rate, and avg. air temperature 0-2 d BSC.*

Effect of Two-Way Interactions on Microbial Water Quality: Interaction effects (i.e., the interaction between two environmental [e.g., weather, water quality] factors and their effect on the likelihood of pathogen detection) are discussed across pathogens and not as part of the pathogen-specific sections above due to the number of interaction effects identified in this study that were consistent across pathogens, produce-growing regions and/or sample types. Similarly, due to the large number of interactions observed, only selected interactions are discussed here in-text (and in the supplemental materials in the Draft Manuscript). Specifically, we focused on the impact of biologically plausible interactions on the likelihood of detecting pathogens in GS as opposed to MS. We focused on GS because (i) ~two times as many GS (N=264) were collected as MS (N=121) in 2017, and (ii) industry stakeholders typically use GS to monitor surface water quality.

We found evidence of interactions between DO and multiple factors. For example, the likelihood of isolating *Salmonella* from AZ GS appeared to be higher when DO <8.5 mg/L and avg. air temperature was >20°C compared to when DO was >8.5 mg/L or avg. air temperature

was <20°C. We also found evidence of interactions between rainfall and physiochemical water quality. For instance, *E. coli* levels in NY were highest when rainfall 0-1 d BSC was >1 cm and turbidity was >10 NTU as compared to when rainfall 0-1 d BSC was <1 cm or turbidity was <10 NTU (see Figure S12, Draft Manuscript). Overall, these findings suggest that (i) temporal environmental heterogeneity (i.e., interactions between environmental factors and changes in environmental conditions over time) is associated with microbial water quality, and (ii) produce safety risks associated with preharvest surface water use are dependent on environmental conditions at the time of water use.

Furthermore, interactions between *E. coli* levels and multiple other factors also appear to affect the likelihood of pathogen detection. For instance, the likelihood of detecting EPEC in NY GS was higher when *E. coli* levels were >200 MPN/100 mL and turbidity was >7 NTUs compared to when turbidity <7 NTUs or <200 MPN/100 mL. Similarly, the likelihood of isolating *Salmonella* from AZ GS appears to be highest when *E. coli* levels were >200 MPN/100 mL and DO <8.5 mg/L as opposed to when *E. coli* levels were <200 MPN/100 mL or DO was >8.5 mg/L. Overall, this indicates that (i) the relationship between *E. coli* levels and pathogen presence in surface water is mediated by environmental conditions, and (ii) *E. coli* levels alone may not be a suitable indicator of the food safety risks associated with preharvest surface water use for produce production. Instead, these findings suggest that data on environmental conditions should be incorporated into *E. coli*-based monitoring tools to facilitate preharvest surface water treatment, testing, and use.

Relationship Between *E. coli* Levels and Pathogens in Grab Samples: The relationship between the $\log_{10}$ MPN of *E. coli*/100 mL and pathogen detection was characterized using GLMM. A GLMM could not be developed to characterize the relationship between *Listeria* isolation and *E. coli* levels in AZ due to the low frequency of *Listeria* in AZ. Regression analysis showed that *Salmonella* isolation in AZ, EPEC detection in AZ, and EHEC detection in AZ and NY were significantly associated with *E. coli* levels, while *L. monocytogenes*, *Listeria* spp., *Salmonella,* and EPEC detection in NY were not significantly associated with *E. coli* levels (**Table 3**). For instance, the odds of isolating *Salmonella* from AZ GS increased by a factor of 4 (95% CI = 1.5, 13.5; $P = 0.009$) for each increase in the $\log_{10}$ MPN of *E. coli*/100 mL. Similarly, the odds of detecting EHEC in AZ and NY GS increased by a factor of 50 (95% CI = 4.1, 621.9; $P = 0.002$) and a factor of 19 (95% CI = 5.4, 63.31; $P < 0.001$), respectively, for each $\log_{10}$ increase in the MPN of *E. coli*/100 mL. These findings suggest that the utility of *E. coli* levels as an indicator of the food safety risks associated with preharvest water is pathogen-specific (e.g., *Salmonella* versus *Listeria*) and region-specific (AZ vs NY). Specifically, *E. coli* may be an appropriate indicator of the presence of EHEC but not *L. monocytogenes*.

We also assessed the predictive accuracy of the AWS to identify water sources that were contaminated by the target pathogens at the time of water use. Briefly, bootstrapping was used to simulate water sampling to create a microbial water quality profile (MWQP) composed of 20 samples (N=10,000 MWQPs per waterway). The last GS selected for inclusion in each MWQP (i.e., the 20th sample selected) represented water quality at the time of water use (e.g., if this GS was positive for *Salmonella* then the water source was considered positive for *Salmonella* at the time of water use). While approx. 50% of MWQPs in AZ and 27% of MWQPs in NY met the AWS (**Table 4**), the percent of pathogen-positive MWQPs that met the AWS ranged between 20% (EHEC in NY) and 72% (*L. monocytogenes* in AZ). We found that the geometric mean and STV varied substantially among the simulated MWQP for a given waterway (**Fig. 2**). In general, the efficacy of the AWS for identifying water sources contaminated by foodborne pathogens also appears to be region- and pathogen-specific. For instance, while the odds of *E. coli* levels exceeding the AWS was 2.6 times greater for EHEC-positive streams compared to EHEC-negative streams (DOR=2.6), the odds of *E. coli* levels exceeding the AWS were approx. the

same for EHEC-positive and EHEC-negative canals (DOR=0.99). We found the opposite trend for *Salmonella* (DOR=1.8 in AZ; DOR=1.0 in NY). Overall, these findings indicate that meeting the AWS is largely a function of when the water samples that comprise the MWQP were collected, and that meeting the AWS may be a poor approximation of *E. coli* levels in surface water at the time of water use.

Geospatial Predictive Models: Conditional inference trees were developed to predict the microbial quality of surface water in AZ and NY. Specifically, classification trees were developed to separately predict the presence of each target pathogen in each state (**Fig. 3**). Multiple weather factors (e.g., average solar radiation, rainfall), physiochemical water quality factors (e.g., conductivity, turbidity) and spatial factors (e.g., percent of the upstream area used for livestock production) were included as covariates in two or more trees. However, based on repeated 3-fold cross-validation, the ability of the classification trees to correctly predict pathogen presence varied, depending on the outcome (e.g., EHEC vs *Listeria* spp.) and state (**Fig. 3**). For instance, the accuracy of the trees developed to predict *Salmonella* presence in AZ and NY was 67% and 41%, respectively. The poor predictive accuracy of some of the trees may be due to (i) the small number of samples available for training the models (N =261 in AZ and 377 in NY), and (ii) that the most recent land cover data available was for 2011. As such, the trees developed here will be refined using (i) large datasets provided by collaborators (e.g., a dataset with information on *E. coli* levels for 46 NY waterways between 2002 and 2018), and (ii) the 2016 land cover dataset, which should be published in 2019. Despite the low accuracy of some of the trees, the retention of multiple data types (e.g., weather, water quality, microbial, remotely-sensed) as covariates indicates the value of using multiple data types to inform decision-making.

Overall, the trees that predicted EHEC presence (Accuracy = 61% in AZ, 70% in NY) had the greatest predictive ability. Based on these trees, EHEC was more likely to be present in NY waterways when the avg. air temperature 5-10 d before sampling was >21°C, compared to when avg. air temperature 5-10 d before sampling was ≤21°C ($P < 0.001$). EHEC was more likely to be present in AZ waterways when the MPN of *E. coli*/100 mL was >25.6 compared to when the MPN of *E. coli*/100 mL was ≤25.6. This finding suggests that the use of *E. coli*-based water quality tests to identify and control EHEC risks associated with the preharvest surface water use in AZ may be reasonable. However, given the sample size of the study reported here, we are not recommending growers use this cut-off of 25.6 MPN/100 mL; instead, additional data is needed to refine the trees reported here and to develop specific guidance.

Outcomes and Accomplishments

Overall this project provides data on foodborne pathogen prevalence in AZ and NY surface water, including data on factors associated with an increased likelihood of pathogen contamination of water sources used for produce production. Specifically, 2,723 samples (2,361 GS and 362 MS) representing 128 waterways (60 AZ and 68 NY waterways) were collected as part of the study reported here. *This study is therefore one of the most comprehensive surveys of foodborne pathogen prevalence in NY agricultural water to date.* As such, this study also provides baseline data that will allow AZ and NY growers to better address food safety risks associated with preharvest surface water use. The data generated here will be used to develop three peer-reviewed publications; one manuscript draft is complete as of March 2019. Additionally, the samples and data collected as part of this study were able to support other CPS-funded projects. For example, while collecting samples for the study presented here, we also collected samples for the CPS-funded project, "*Developing cross-assembly phage as a viral indicator for irrigation waters.*" Similarly, the isolates collected as part of the study reported here will be used to complete Objective 3 (Perform WGS of Listeria isolates obtained from throughout

the produce chain) of the CPS-funded project, "*Listeria whole genome sequence data reference sets are needed to allow for improved persistence assessment and source tracking."*

Importantly, this study provides industry with a ranked list of risk factors that can be used to generally assess times and locations with an increased risk of pathogen detection and/or elevated *E. coli* levels (see **Figure 1** and **Table 5**). Overall, the top-ranked factors associated with pathogen detection appear to be pathogen and region-specific. For example, the 3 top-ranked factors associated with *E. coli* levels in AZ were the canal being sampled, dissolved oxygen level, and minimum air temperature 0–5 days before sample collection, while the 3 top-ranked factors associated with *E. coli* levels in NY were turbidity, flow rate and pH (Fig. 1; Table 5). However, certain factors were among the top-ranked factors for multiple combinations of pathogen and region. For example, temperature (i.e., *avg. air*, *max. air*, *min. air* and/or *water temperature*) was identified as a top-ranked factor for all pathogens in both states with the exception of *L. monocytogenes* in NY. Temperature was the top-ranked factor associated with *Salmonella*, EPEC and EHEC detection in AZ (Fig. 1; Table 5). In general, there was an optimal temperature for detection of each pathogen in each state; for instance, for the waterways sampled as part of the study reported here, the likelihood of EHEC detection was greatest in (i) AZ when avg. air temperature was approx. 27°C, and (ii) NY when min. air temperature was approx. 18°C (see Figures S6, S7-S10, Draft Manuscript). Overall, this study found that microbial water quality is associated with temporal environmental heterogeneity (i.e., interactions between environmental factors and changes in environmental conditions over time), and as such, the produce safety risks associated with the preharvest use of a specific surface water source are not constant over time and depend on the environmental conditions at the time the water is used.

The study reported here also provides data on the relationship between *E. coli* levels and pathogen detection in surface water. Specifically, the findings presented here indicate that (i) meeting the proposed AWS is largely a function of when water samples were collected, (ii) that meeting *E. coli*-based water quality standards may not reduce produce safety risks associated with preharvest water use, (iii) the relationship between *E. coli* levels and pathogen presence in surface water is mediated by environmental conditions, and (iv) the utility of *E. coli* levels as an indicator of the food safety risks associated with preharvest water use is pathogen- and region-specific. For example, our findings indicate that the *MPN of E. coli/100 mL is not a suitable indicator of L. monocytogenes presence in NY streams but may be a suitable indicator of EHEC presence in AZ canals*. As such, *basing risk management decisions solely on water quality tests that quantify E. coli levels may not be appropriate*. Instead, data on environmental conditions should be incorporated into *E. coli*-based monitoring tools to facilitate preharvest surface water treatment, testing, and use. Moreover, given the dynamic and complex nature of surface water systems these tools need to (i) account for temporal variation in environmental conditions, and (ii) provide insights on microbial water quality at the time of water use. This study provides a blueprint for how these tools can be developed by (i) identifying factors that may be useful as supplemental indicators of microbial water quality (e.g., as covariates in a model to predict when EHEC is likely to be present), and (ii) providing a conceptual framework on how multiple data types (e.g., weather, remotely-sensed land use) and novel technologies (e.g., remote sensing, geographic information systems, machine learning) can be integrated to manage preharvest produce safety risks. Specifically, the geospatial models developed here provide a framework that can be refined and adapted to other produce-growing regions as data becomes available.

This study also provides data on the ability of two water sampling methods (24-h MS and 10-L GS) to detect foodborne pathogens in surface water. Our findings indicate that appropriate water sampling methods depend on the organism of concern. Specifically, we found that *24-h MS were significantly better than 10-L GS at detecting Salmonella and pathogenic E. coli in surface water sources, while GS were significantly better at detecting Listeria*; stakeholders can use this information to optimize their water sampling protocols. The study's findings also suggest that

study aims (e.g., is the goal to optimize pathogen detection or replicate industry practices?), time constraints (i.e., MS require 2 site visits but GS require 1 visit), the outcome of interest (e.g., MS cannot be used to calculate concentrations since the volume of water that flows through the MS is unknown), and the potential loss of MS to storms or human tampering should also be considered when designing sampling protocols.

In the short-term, findings from this study will increase our understanding of risk factors that are associated with foodborne pathogen presence in surface water and will assist growers to develop produce safety risk management plans. Broadly speaking, the produce industry will benefit from these findings as the data generated here provide for more effective, scientifically justified pre-harvest risk management strategies, which has direct implications for contamination risks throughout the produce supply chain.

Summary of Findings and Recommendations

- In total, 2,723 samples (2,361 grab samples [GS] and 362 Moore swabs [MS]) were collected, representing 128 waterways (60 AZ and 68 NY waterways).
- In the study reported here, 24-h Moore swabs were better than 10-L grab samples at detecting pathogenic *E. coli* and *Salmonella*, while grab samples were better at detecting *Listeria*, indicating that appropriate water sampling methods depend on the organism of concern. This information will help growers optimize their agricultural water sampling protocols (e.g., AZ leafy green growers concerned about EHEC may decide to use MS as opposed to GS to test water sources during traceback sampling).
- While the utility of *E. coli* as an indicator of the food safety risks associated with preharvest water appears to be pathogen-specific and region-specific, monitoring tools that incorporate data on *E. coli* levels may be particularly useful for managing produce safety risks associated with EHEC contamination in AZ.
- Changes in environmental conditions over time and complex interactions between environmental factors are associated with microbial water quality. As such, environmental conditions should be taken into account when designing strategies for managing the produce safety risks associated with preharvest surface water use.
- The ranked lists of risk factors included here (Figure 1 and Table 5) provide a way to incorporate environmental data into decision-making, by allowing growers to generally assess times and locations with an increased risk of pathogen detection. For example, an average air temperature of approx. 27°C 0 to 3 days before sample collection (or water use) was associated with a higher likelihood of EHEC contamination of the AZ canals sampled as part of the study reported here.

APPENDICES

Publications and Presentations

Initial results were presented at the 2017 meeting of the International Association for Food Protection (IAFP) as (i) a poster entitled, "*Salmonella isolation is not associated with E. coli levels in agricultural water samples collected from New York streams"*, and (ii) an invited symposium talk entitled, "*Optimizing agricultural water sampling strategies to account for variability across time and space".* Interim results were also presented at the annual Center for Produce Safety (CPS) Symposia as a lightning talk in 2017, and a presentation in 2018; final results will be presented at the 2019 CPS Symposium. Preliminary study results were also presented as case studies in invited guest lectures at Ithaca College, the Rochester Institute of Technology, the State University of New York (SUNY) at Oswego, and SUNY Environmental Science and Forestry. An abstract entitled, "*The relationship between E. coli levels and pathogen detection in surface water samples is mediated by environmental conditions"*, has been submitted to present the study findings at the 2019 IAFP meeting. Additionally, a draft manuscript based on analysis of the data collected as part of Obj. 1 has been prepared – a copy of the Draft manuscript, which contains supplementary figures and tables referred to in this final report, will be made available on the CPS website by August 2019. We will complete two additional manuscripts based on analysis of the data collected as part of Obj. 2.

Budget Summary

A total of $358,214 was awarded for this project.
Funds utilized as of 2/13/19:

Salaries and Wages	158,016.00
Travel	13,018.53
Materials and Supplies	24,995.39
Services	59.50
Subcontracts	129,478.00
Other Direct Expenses	49.36
Indirect Costs	9,479.99
Total:	**335,096.80**

Figures and Tables

(see below)

Figures 1–3 and Tables 1–5

Figure 1: Results of random forests (RF) that identified factors associated with *E. coli* levels and the likelihood of detecting pathogens in AZ and NY GS. The y-axis shows the factors ranked from most to least important. The x-axis shows normalized variable importance (NVI). A higher NVI (relative to all other factors in the RF) equates to a stronger association between outcome and factor; a NVI ≤0 indicates no association. Five overlapping time frames (0-1, 0-2, 0-3, 0-4, or 0-5 d before sample collection [BSC]) were used to calculate the values of all weather factors with the exception of rainfall; instead, rainfall was calculated on a daily basis (i.e., 0-1, 1-2, 2-3, 3-4, 4-5 d BSC). Separate RF were then developed for each outcome and time frame in each state, and the results for the RF with the highest Kappa score for each outcome are reported here. Thus, the time frame for the RF reported here differs for each combination of outcome and state. Factors marked with * were ranked among the top four factors across all time frames considered. Factors marked with † received the same ranking across all time frames (see also Tables S3–S6, Draft Manuscript). RF analysis was not performed to identify factors associated with *Listeria* in AZ due to the low prevalence of *Listeria* in AZ.

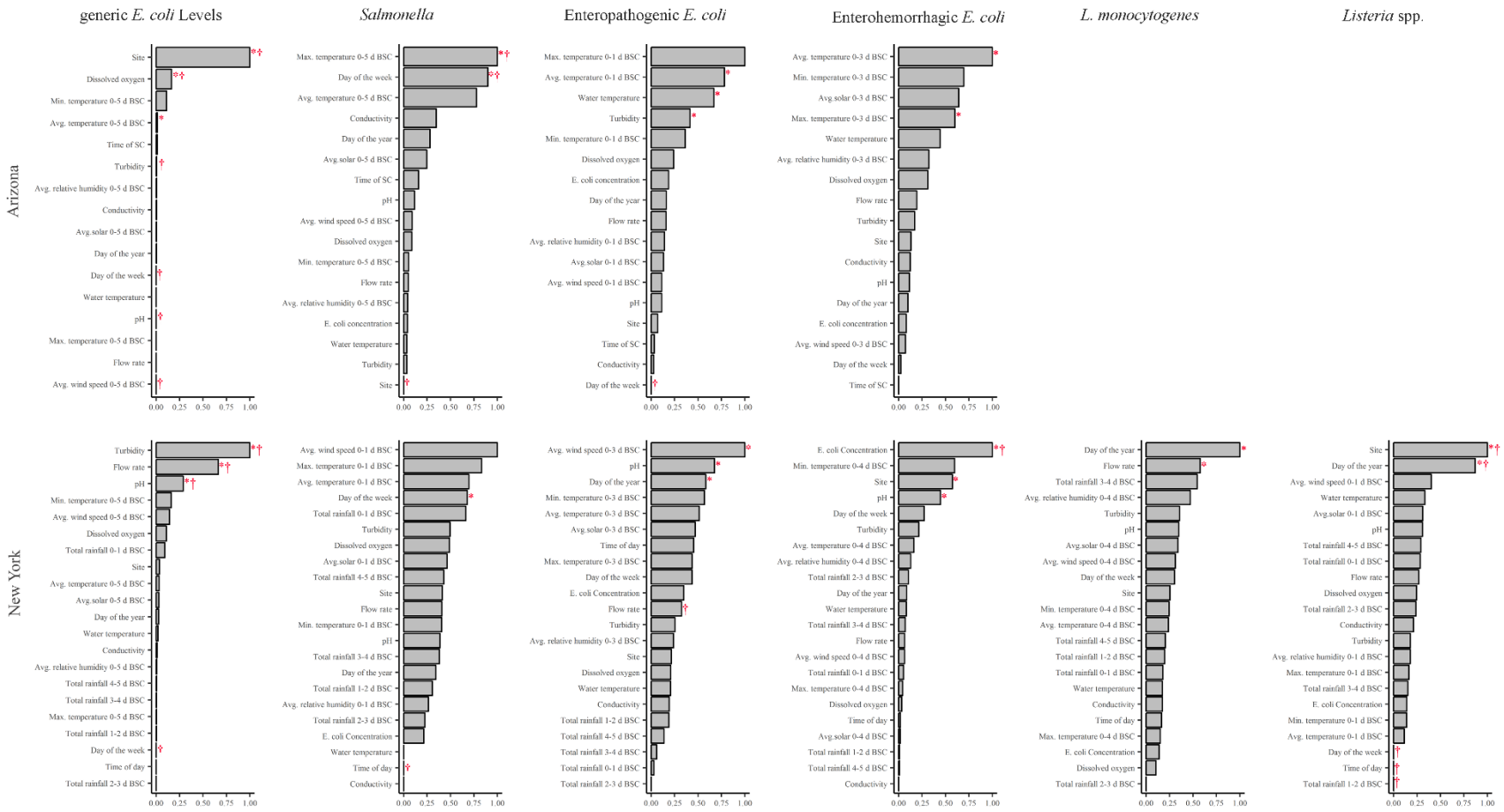
generic E. coli Levels
Salmonella
Enteropathogenic E. coli
Enterohemorrhagic E. coli
L. monocytogenes
Listeria spp.
Arizona
New York
Site
Dissolved oxygen
Min. temperature 0-5 d BSC
Avg. temperature 0-5 d BSC
Time of SC
Turbidity
Avg. relative humidity 0-5 d BSC
Conductivity
Avg.solar 0-5 d BSC
Day of the year
Day of the week
Water temperature
pH
Max. temperature 0-5 d BSC
Flow rate
Avg. wind speed 0-5 d BSC
Max. temperature 0-5 d BSC
Day of the week
Avg. temperature 0-5 d BSC
Conductivity
Day of the year
Avg.solar 0-5 d BSC
Time of SC
pH
Avg. wind speed 0-5 d BSC
Dissolved oxygen
Min. temperature 0-5 d BSC
Flow rate
Avg. relative humidity 0-5 d BSC
E. coli concentration
Water temperature
Turbidity
Site
Max. temperature 0-1 d BSC
Avg. temperature 0-1 d BSC
Water temperature
Turbidity
Min. temperature 0-1 d BSC
Dissolved oxygen
E. coli concentration
Day of the year
Flow rate
Avg. relative humidity 0-1 d BSC
Avg.solar 0-1 d BSC
Avg. wind speed 0-1 d BSC
pH
Site
Time of SC
Conductivity
Day of the week
Avg. temperature 0-3 d BSC
Min. temperature 0-3 d BSC
Avg.solar 0-3 d BSC
Max. temperature 0-3 d BSC
Water temperature
Avg. relative humidity 0-3 d BSC
Dissolved oxygen
Flow rate
Turbidity
Site
Conductivity
pH
Day of the year
E. coli concentration
Avg. wind speed 0-3 d BSC
Day of the week
Time of SC
Turbidity
Flow rate
pH
Min. temperature 0-5 d BSC
Avg. wind speed 0-5 d BSC
Dissolved oxygen
Total rainfall 0-1 d BSC
Site
Avg. temperature 0-5 d BSC
Avg.solar 0-5 d BSC
Day of the year
Water temperature
Conductivity
Avg. relative humidity 0-5 d BSC
Total rainfall 4-5 d BSC
Total rainfall 3-4 d BSC
Max. temperature 0-5 d BSC
Total rainfall 1-2 d BSC
Day of the week
Time of day
Total rainfall 2-3 d BSC
Avg. wind speed 0-1 d BSC
Max. temperature 0-1 d BSC
Avg. temperature 0-1 d BSC
Day of the week
Total rainfall 0-1 d BSC
Turbidity
Dissolved oxygen
Avg.solar 0-1 d BSC
Total rainfall 4-5 d BSC
Site
Flow rate
Min. temperature 0-1 d BSC
pH
Total rainfall 3-4 d BSC
Day of the year
Total rainfall 1-2 d BSC
Avg. relative humidity 0-1 d BSC
Total rainfall 2-3 d BSC
E. coli Concentration
Water temperature
Time of day
Conductivity
Avg. wind speed 0-3 d BSC
pH
Day of the year
Min. temperature 0-3 d BSC
Avg. temperature 0-3 d BSC
Avg.solar 0-3 d BSC
Time of day
Max. temperature 0-3 d BSC
Day of the week
E. coli Concentration
Flow rate
Turbidity
Avg. relative humidity 0-3 d BSC
Site
Dissolved oxygen
Water temperature
Conductivity
Total rainfall 1-2 d BSC
Total rainfall 4-5 d BSC
Total rainfall 3-4 d BSC
Total rainfall 0-1 d BSC
Total rainfall 2-3 d BSC
E. coli Concentration
Min. temperature 0-4 d BSC
Site
pH
Day of the week
Turbidity
Avg. temperature 0-4 d BSC
Avg. relative humidity 0-4 d BSC
Total rainfall 2-3 d BSC
Day of the year
Water temperature
Total rainfall 3-4 d BSC
Flow rate
Avg. wind speed 0-4 d BSC
Total rainfall 0-1 d BSC
Max. temperature 0-4 d BSC
Dissolved oxygen
Time of day
Avg.solar 0-4 d BSC
Total rainfall 1-2 d BSC
Total rainfall 4-5 d BSC
Conductivity
Day of the year
Flow rate
Total rainfall 3-4 d BSC
Avg. relative humidity 0-4 d BSC
Turbidity
pH
Avg.solar 0-4 d BSC
Avg. wind speed 0-4 d BSC
Day of the week
Site
Min. temperature 0-4 d BSC
Avg. temperature 0-4 d BSC
Total rainfall 4-5 d BSC
Total rainfall 1-2 d BSC
Total rainfall 0-1 d BSC
Water temperature
Conductivity
Max. temperature 0-4 d BSC
E. coli Concentration
Dissolved oxygen
Total rainfall 2-3 d BSC
Site
Day of the year
Avg. wind speed 0-1 d BSC
Water temperature
Avg.solar 0-1 d BSC
pH
Total rainfall 4-5 d BSC
Total rainfall 0-1 d BSC
Flow rate
Dissolved oxygen
Total rainfall 2-3 d BSC
Conductivity
Turbidity
Avg. relative humidity 0-1 d BSC
Max. temperature 0-1 d BSC
Total rainfall 3-4 d BSC
E. coli Concentration
Min. temperature 0-1 d BSC
Avg. temperature 0-1 d BSC
Day of the week
Time of day
Total rainfall 1-2 d BSC
0.00 0.25 0.50 0.75 1.00

Figure 2: Bootstrapping was used to simulate water sampling to create a microbial water quality profile (MWQP) composed of 20 samples (N=10,000 MWQPs per waterway). The graphs show the geometric mean and statistical threshold value for all MWQPs for each waterway. The blue line represents the Produce Safety Rule's proposed agricultural water standard cut-offs (i.e., geometric mean < 126 CFU/100 mL and statistical threshold value < 410 CFU/100 mL).

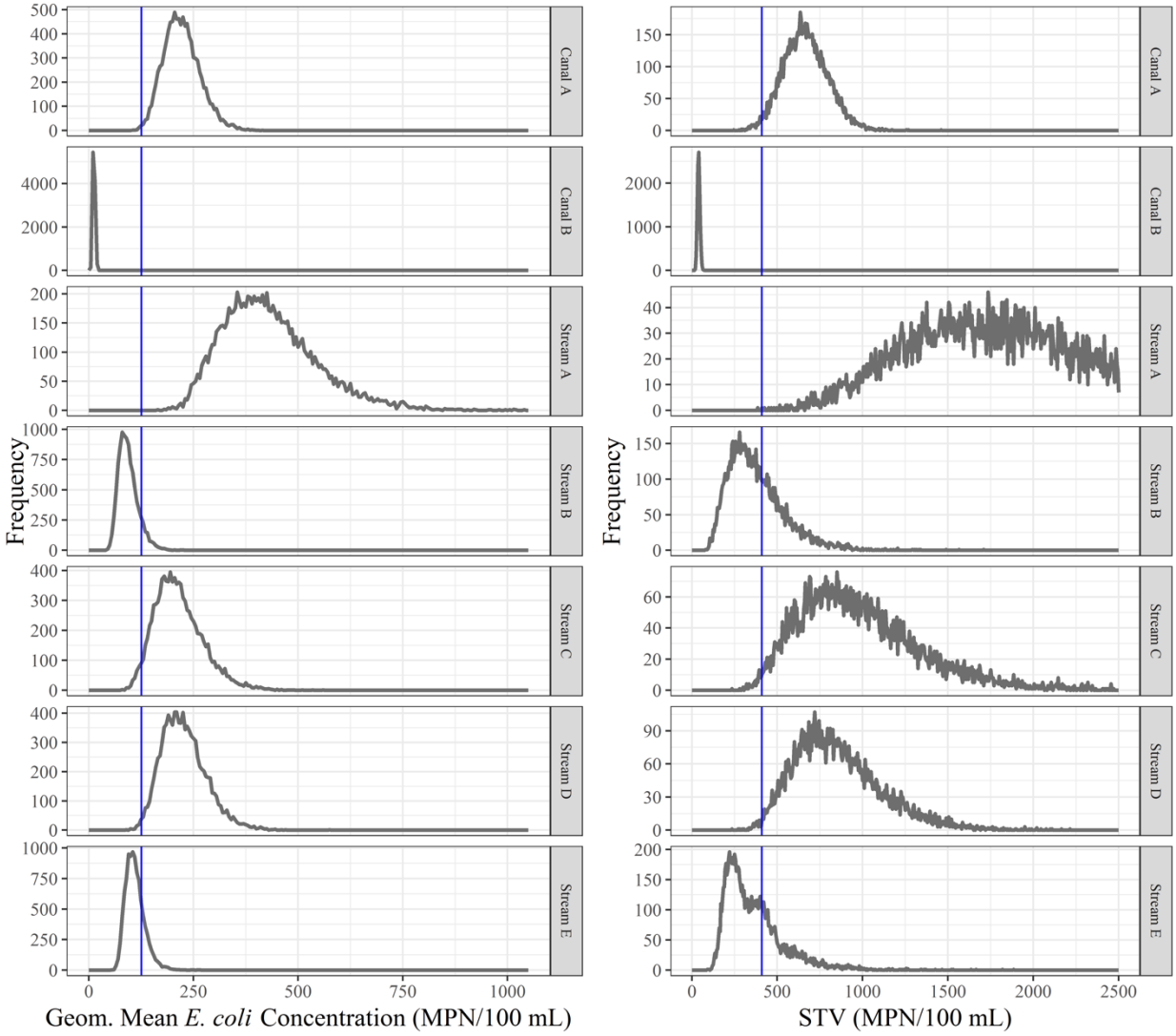

Figure 3: Conditional inference trees that use hierarchical rules to predict if GS are more likely to be positive or negative for *Salmonella*, EHEC or *Listeria* spp. At the top of each split is a rule that partitions samples into as homogenous subsets (i.e., mostly positive or mostly negative samples) as possible using weather, water quality, or spatial factors. The *P-value* for each split is shown next to the split descriptor. The expected numbers of positive (P) and negative (N) samples for each terminal node are shown as bar graphs; the number of samples in the training data that fall into each node is noted at the top of each graph. Trees were initially built using data collected as part of Obj. 2. The ability of each tree to correctly predict pathogen status for a de novo dataset (i.e. external validation) was then performed using the data collected as part of Obj. 1. Since external validation returned an accuracy score <60% for trees that predicted pathogen status in NY, new NY trees were developed using the combined Obj. 1 and 2 data; repeated, 3-fold cross-validation was then used to estimate the ability of each these trees to correctly predict pathogen status for a de novo dataset. The accuracy of the trees that predicted (i) *Salmonella* status in AZ and NY were 67% and 41%, respectively, (ii) EHEC status in AZ and NY were 67%, and 70%, respectively, and (iii) *Listeria* spp. status in NY was 47%.

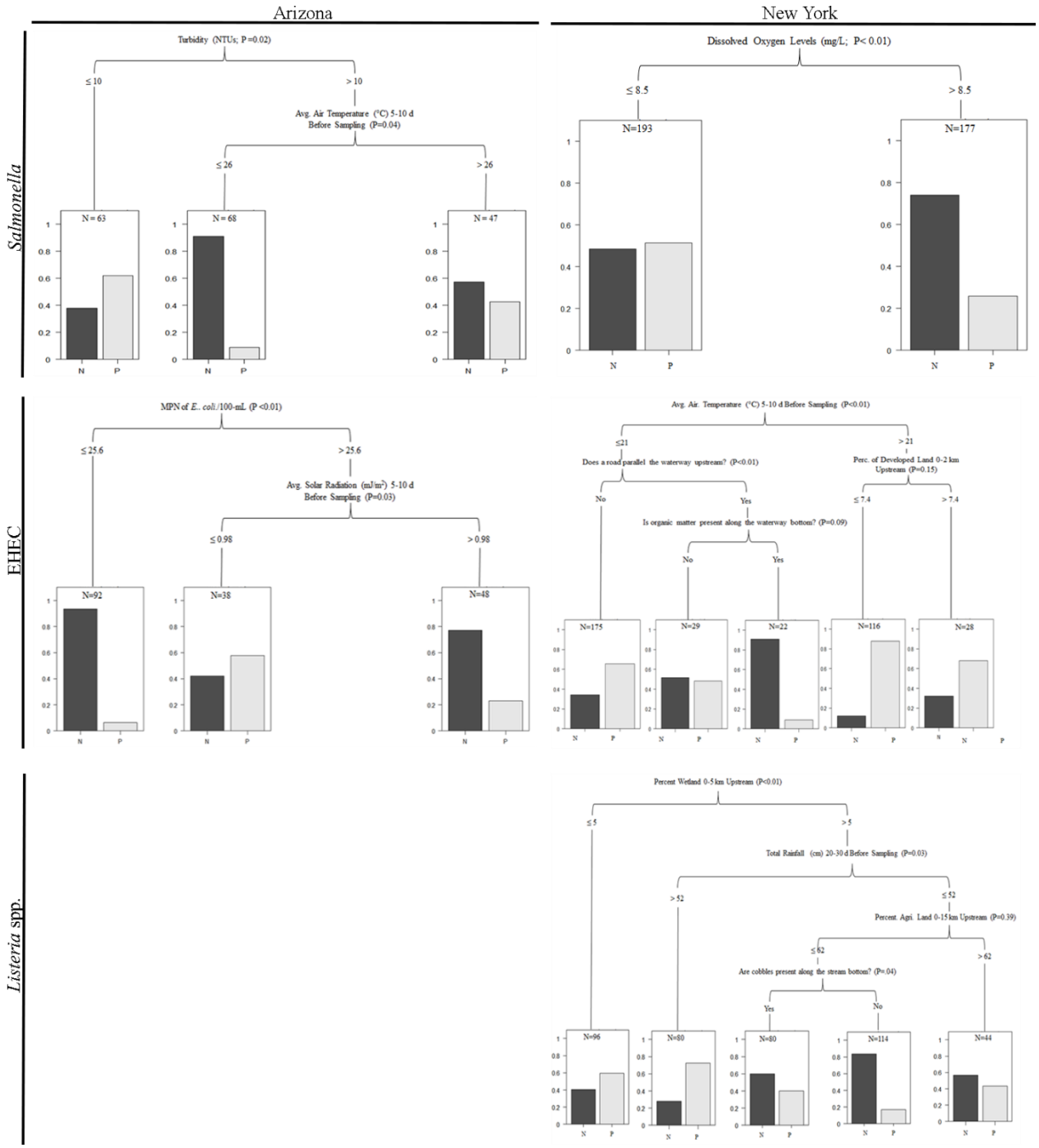
Arizona
New York
Salmonella
EHEC
Listeria spp.
Turbidity (NTUs; P=0.02)
≤ 10
> 10
Avg. Air Temperature (°C) 5-10 d Before Sampling (P=0.04)
≤ 26
> 26
N = 63
N = 68
N = 47
Dissolved Oxygen Levels (mg/L; P< 0.01)
≤ 8.5
> 8.5
N=193
N=177
MPN of E. coli/100-mL (P <0.01)
≤ 25.6
> 25.6
Avg. Solar Radiation (mJ/m²) 5-10 d Before Sampling (P=0.03)
≤ 0.98
> 0.98
N=92
N=38
N=48
Avg. Air. Temperature (°C) 5-10 d Before Sampling (P<0.01)
≤21
> 21
Does a road parallel the waterway upstream? (P<0.01)
Perc. of Developed Land 0-2 km Upstream (P=0.15)
No
Yes
≤ 7.4
> 7.4
Is organic matter present along the waterway bottom? (P=0.09)
N=175
N=29
N=22
N=116
N=28
Percent Wetland 0-5 km Upstream (P<0.01)
≤ 5
> 5
Total Rainfall (cm) 20-30 d Before Sampling (P=0.03)
> 52
≤ 52
Percent. Agri. Land 0-15 km Upstream (P=0.39)
≤ 62
> 62
Are cobbles present along the stream bottom? (P=.04)
N=96
N=80
N=80
N=114
N=44
N
P

Table 1: Frequency of target organism detection, and median *E. coli* levels in each year and state. Samples collected in 2017 were collected as part of Obj. 1. Samples collected in 2018 were collected as part of Obj. 2.

Year & Water Source	Prevalence (No. of Positive Samples/No. of Negative Samples)				Median MPN of *E. coli*/100 mL (Min.-Max.)
	Culture-Confirmed		PCR-Screen Positive [b]		
	Listeria spp.[a]	*Salmonella*	EPEC	EHEC	
Arizona					
2017	4% (3/76)	34% (26/77)	83% (69/83)	51% (43/83)	28.8 (< 1.0 - 770.1)
2018	-	37% (65/178)	35% (62/178)	22% (39/178)	25.3 (< 1.0 - >2,419.6)
Total	4% (3/76)	36% (91/255)	50% (131/261)	31% (82/261)	25.6 (< 1.0 - >2,419.6)
New York					
2017	47% (85/181)	44% (80/181)	94% (171/181)	69% (125/181)	160.4 (18.5 - >2,419.6)
2018	38% (75/196)	40% (79/196)	97% (190/196)	68% (133/196)	211.4 (2.0 - >2,419.6)
Total	42% (160/377)	42% (159/377)	96% (361/377)	68% (258/377)	193.5 (2.0 - >2,419.6)

[a] *Listeria* spp. includes *L. monocytogenes*.

[b] Samples that were PCR-screen positive for the *eaeA* gene were considered presumptive EPEC positive samples, while samples that were positive for the *eaeA* and *stx* genes were considered presumptive EHEC positive samples.

Table 2: Comparison of the ability of 24-h Moore swabs (MS) and paired 10-L grab samples (GS) to detect foodborne pathogens in surface water; data used in these analyses were collected as part of Obj. 1. A contamination event was defined as occurring if the MS or one of its paired GS tested positive for the target organism.

Microbe	No. of Contamination Events/Total No. of Sampling Days	No. of Events Detected By		Disagreement [b]	Regression Results [c]					
		Paired GS Only [a]	MS Only		Fixed Effects			Variance of Random Effects (Standard Deviation)		
					OR [d]	95% CI [e]	*P-value*	Year Day	State	Site
L. monocytogenes	37/120	31	3	92%	0.38	0.15, 0.96	0.041	0.1 (0.3)	0.5 (0.7)	$1.7*10^{-10}$ ($1.3*10^{-5}$)
Listeria spp. [f]	72/120	49	2	71%	0.24	0.12, 0.48	< 0.001	1.4 (1.2)	3.6 (1.9)	1.2 (1.1)
Salmonella	99/120	28	21	49%	2.17	1.36, 3.45	0.001	0.4 (0.6)	0.0 (0.0)	$1.0*10^{-2}$ (0.1)
EPEC	118/121	3	5	7%	8.81	1.26, 61.50	0.028	41.0 (6.4)	$1.7*10^{-10}$ ($1.3*10^{-5}$)	1.2 (1.1)
EHEC	113/121	9	13	19%	6.65	3.16, 14.01	< 0.001	1.7 (1.3)	0.0 (0.0)	0.7 (0.8)

[a] GS collected during the 24 h that the MS was in the waterway; each MS had between 2 and 7 paired GS.

[b] Disagreement = (No. of events detected by GS Only + No. of events detected by MS Only)/Total No. of Contamination Events.

[c] Results of generalized linear mixed models that compared pathogen detection by 24-h MS and individual paired 10-L GS; GS were the reference level.

[d] Odds ratio (i.e., the odds of detecting the target organism in a MS/Odds of detecting the target organism in a single paired GS).

[e] 95% Confidence Interval.

[f] Includes *L. monocytogenes.*

Table 3: Results of generalized linear mixed models that characterized the relationship between the $\log_{10}$ MPN of *E. coli* level/100 mL and pathogen detection in grab samples; data used in these analyses were collected as part of Obj. 1.

Target Organism	Fixed Effects			Variance of Random Effects (Standard Deviation)	
	Change in Odds [a]	95% CI [b]	*P-value*	Year Day	Site
L. monocytogenes					
New York	1.1	0.45, 2.70	0.828	0.5 (0.7)	0.0 (0.0)
Listeria spp. [c]					
New York	1.0	0.37, 2.54	0.955	2.4 (1.6)	1.5 (1.2)
Salmonella					
Arizona	4.4	1.46, 13.47	0.009	2.1 (1.4)	0.0 (0.0)
New York	1.5	0.70, 3.30	0.292	1.2 (1.1)	0.2 (0.5)
Enteropathogenic *E. coli*					
Arizona	3.5	1.48, 8.28	0.004	0.0 (0.0)	0.0 (0.0)
New York	13.1	0.31, 555.00	0.179	72.1 (8.5)	0.0 (0.0)
Enterohemorrhagic *E. coli*					
Arizona	50.2	4.05, 621.88	0.002	12.1 (3.5)	6.6 (2.6)
New York	18.6	5.43, 63.31	< 0.001	1.1 (1.0)	$5.0*10^{-2}$ (0.2)

[a] Change in the odds of detecting the target organism for a one $\log_{10}$ increase in the *E. coli* concentration.

[b] 95% Confidence Interval.

[c] Includes *L. monocytogenes.*

Table 4: Ability of the proposed agricultural water standard [AWS; (36)] to correctly predict the pathogen status of surface water samples; analyses were conducted using a simulated dataset generated by bootstrapping.

Target Organism	Proportion of Pathogen Positive Subsets Below AWS [a]	DOR [b]	Sensitivity	Specificity
L. monocytogenes				
Arizona	72%	0.39	0.28	0.49
New York	32%	0.75	0.68	0.26
Listeria spp. [c]				
Arizona	72%	0.39	0.28	0.49
New York	39%	0.29	0.61	0.16
Salmonella				
Arizona	40%	1.80	0.60	0.55
New York	27%	1.03	0.74	0.27
EPEC				
Arizona	45%	3.35	0.54	0.74
New York	26%	1.83	0.74	0.39
EHEC				
Arizona	51%	0.99	0.49	0.50
New York	20%	2.60	0.80	0.40

[a] A sample exceeded the proposed AWS if the geometric mean was ≥126 CFU/100 mL or the statistical threshold value was ≥410 CFU/100 mL. In total, 50% of AZ subsets and 27% of NY subsets met the AWS (regardless of pathogen status).

[b] Diagnostic odds ratio = (Odds of *E. coli* levels in waterways contaminated by the target pathogen exceeding the AWS thresholds)/(Odds of *E. coli* levels in waterways not contaminated by the target pathogen exceeding the AWS thresholds). DOR <1 indicates that not meeting the AWS is protective, a DOR ~1 indicates that exceeding the AWS does not relate to pathogen contamination status, and a DOR >1 indicates that exceeding the AWS is a risk factor that is predictive of pathogen presence.

[c] Includes *L. monocytogenes.*

Table 5: Results of random forests (RF) that identified factors associated with *E. coli* levels and the likelihood of detecting pathogens in AZ and NY GS. Factors marked with an X were among the four top-ranked factors associated with the given outcome. Five, overlapping time frames (0-1, 0-2, 0-3, 0-4, or 0-5 d before sample collection) were used to calculate the values of all weather factors with the exception of rainfall; instead, rainfall was calculated daily (i.e., 0-1, 1-2, 2-3, 3-4, 4-5 d before sample collection). Separate RF were then developed for each outcome and time frame in each state, and the results for the RF with the highest Kappa score for each outcome are reported here. Thus, the time frame for the RF reported here differs for each combination of outcome and state. As such, the results reported here are aggregated for each given factor type (e.g., avg. solar radiation, avg. wind speed).

Factor	*E. coli* Levels		*Salmonella*		EPEC		EHEC		*L. monocytogenes*	*Listeria* spp. [a]	Total Times in Top 4 [b]
	AZ	NY	AZ	NY	AZ	NY	AZ	NY	NY	NY	
Avg. Relative Humidity									X		1
Avg. Solar Radiation							X				1
Avg. Wind Speed				X		X				X	3
Conductivity			X								1
Day of the Week			X	X							2
Day of the Year						X			X	X	3
Dissolved Oxygen	X										1
Flow Rate		X							X		2
MPN of *E. coli*/100 mL								X			1
pH		X				X		X			3
Rainfall									X		1
Sample Site	X							X		X	3
Temperature											
Avg. Air	X		X	X	X		X				5
Max. Air			X	X	X		X				4
Min. Air	X	X				X	X	X			5
Water					X					X	2
Turbidity		X			X						2

[a] Includes *L. monocytogenes.*

[b] The total number of times the given factor was among the four top-ranked factors across all outcomes.

References cited

1. **Weller D**, **Wiedmann M**, **Strawn LK**. 2015. Irrigation is significantly associated with an increased prevalence of *Listeria monocytogenes* in produce production environments in New York State. J Food Prot **78**:1132–1141.
2. **Strawn LK**, **Gröhn YT**, **Warchocki S**, **Worobo RW**, **Bihn E**, **Wiedmann M**. 2013. Risk factors associated with *Salmonella* and *Listeria monocytogenes* contamination of produce fields. Appl Environ Microbiol **79**:7618–7627.
3. **Holvoet K**, **Sampers I**, **Seynnaeve M**, **Uyttendaele M**. 2014. Relationships among hygiene indicators and enteric pathogens in irrigation water, soil and lettuce and the impact of climatic conditions on contamination in the lettuce primary production. Int J Food Microbiol **171**:21–31.
4. **Mody RK**, **Greene S**, **Gaul L**, **Sever A**, **Pichette S**, **Zambrana I**, **Dang T**, **Gass A**, **Wood R**, **Herman K**, **Cantwell L**, **Falkenhorst G**, **Wannemuehler K**, **Hoekstra R**, **McCullum I**, **Cone A**, **Franklin L**, **Austin J**, **Delea K**, **Behravesh C**, **Sodha S**, **Yee JC**, **Emanuel B**, **Al-Khaldi SF**, **Jefferson V**, **Williams I**, **Griffin PM**, **Swerdlow D**. 2011. National outbreak of *Salmonella* serotype Saintpaul infections: importance of Texas restaurant investigations in implicating Jalapeño peppers. PLoS One **6**:e16579.
5. **Strawn LK**, **Fortes ED**, **Bihn E**, **Nightingale KK**, **Gröhn YT**, **Worobo RW**, **Wiedmann M**, **Bergholz PW**. 2013. Landscape and meteorological factors affecting prevalence of three food-borne pathogens in fruit and vegetable farms. Appl Environ Microbiol **79**:588–600.
6. **Guan TY**, **Blank G**, **Ismond A**, **Van Acker R**. 2001. Fate of foodborne bacterial pathogens in pesticide products. J Sci Food Agric **81**:503–512.
7. **Food and Drug Administration (U.S.)**. 2018. Environmental assessment of factors potentially contributing to the contamination of Romaine lettuce implicated in a multi-state outbreak of E. coli O157:H7.
8. **Centers for Disease Control and Prevention (U.S.)**. 2018. Oubtreak of *E. coli* infections linked to Romaine lettuce.
9. **Baloch MA**. 2014. Leafy greens: the case study and real-life lessons from a Shiga-toxin-producing *Escherichia coli* (STEC) O145 outbreak in romaine lettuce. Glob Saf Fresh Prod.
10. **Materon L a.**, **Martinez-Garcia M**, **McDonald V**. 2007. Identification of sources of microbial pathogens on cantaloupe rinds from pre-harvest to post-harvest operations. World J Microbiol Biotechnol **23**:1281–1287.
11. **Alexander LM**. 2015. Figuring Out The Food Safety Modernization Act. Grow Prod. Willoughby, OH.
12. **McEntire J**, **Gorny J**. 2017. Fixing FSMA's Ag Water Requirements. Food Saf Mag.
13. **Wall G**, **Clements D**, **Fisk C**, **Stoeckel D**, **Woods K**, **Bihn E**. 2019. Meeting Report: Key Outcomes from a Collaborative Summit on Agricultural Water Standards for Fresh Produce. Compr Rev Food Sci Food Saf : pre-print.
14. **Pandey PK**, **Soupir ML**, **Haddad M**, **Rothwell JJ**. 2012. Assessing the impacts of watershed indexes and precipitation on spatial in-stream *E. coli* concentrations. Ecol Indic **23**:641–652.
15. **Goyal SM**, **Gerba CP**, **Melnick JL**. 1977. Occurrence and distribution of bacterial indicators and pathogens in canal communities along the Texas coast. Appl Environ Microbiol **34**:139–49.
16. **Hipsey MR**, **Antenucci JP**, **Brookes JD**. 2008. A generic, process-based model of

microbial pollution in aquatic systems. Water Resour Res **44**:published online.

17. **Benjamin L**, **Atwill ER**, **Jay-Russell M**, **Cooley M**, **Carychao D**, **Gorski L**, **Mandrell RE**. 2013. Occurrence of generic *Escherichia coli*, *E. coli* O157 and *Salmonella* spp. in water and sediment from leafy green produce farms and streams on the central California coast. Int J Food Microbiol **165**:65–76.
18. **McEgan R**, **Mootian G**, **Goodridge LD**, **Schaffner DW**, **Danyluk MD**. 2013. Predicting *Salmonella* populations from biological, chemical, and physical indicators in Florida surface waters. Appl Environ Microbiol **79**:4094–4105.
19. **Wilkes G**, **Edge T**, **Gannon V**, **Jokinen C**, **Lyautey E**, **Medeiros D**, **Neumann N**, **Ruecker N**, **Topp E**, **Lapen DR**. 2009. Seasonal relationships among indicator bacteria, pathogenic bacteria, Cryptosporidium oocysts, *Giardia* cysts, and hydrological indices for surface waters within an agricultural landscape. Water Res **43**:2209–23.
20. **Harwood VJ**, **Levine AD**, **Scott TM**, **Chivukula V**, **Lukasik J**, **Farrah SR**, **Rose JB**. 2005. Validity of the indicator organism paradigm for pathogen reduction in reclaimed water and public health protection. Appl Environ Microbiol **71**:3163–70.
21. **Pachepsky Y**, **Shelton D**, **Dorner S**, **Whelan G**. 2015. Can *E. coli* or thermotolerant coliform concentrations predict pathogen presence or prevalence in irrigation waters? Crit Rev Microbiol 1–10.
22. **Gore J**. 2006. Discharge Measurements and Streamflow Analysis, p. 69–70. *In* Hauer, F, Lamberti, G (eds.), Methods in Stream Ecology, 2nd ed. Elsevier, Burlington, MA.
23. **Sbodio A**, **Maeda S**, **Lopez-Velasco G**, **Suslow T V.** 2013. Modified Moore swab optimization and validation in capturing *E. coli* O157:H7 and *Salmonella enterica* in large volume field samples of irrigation water. Food Res Int **51**:654–662.
24. **Bundrant BN**, **Hutchins T**, **den Bakker HC**, **Fortes E**, **Wiedmann M**. 2011. Listeriosis outbreak in dairy cattle caused by an unusual *Listeria monocytogenes* serotype 4b strain. J Vet Diagn Invest **23**:155–158.
25. **Den Bakker HC**, **Bundrant BN**, **Fortes ED**, **Orsi RH**, **Wiedmann M**. 2010. A population genetics-based and phylogenetic approach to understanding the evolution of virulence in the genus *Listeria*. Appl Environ Microbiol **76**:6085–6100.
26. **Kim JS**, **Lee GG**, **Park JS**, **Jung YH**, **Kwak HS**, **Kim SB**, **Nam YS**, **Kwon S-T**. 2007. A novel multiplex PCR assay for rapid and simultaneous detection of five pathogenic bacteria: *Escherichia coli* O157:H7, *Salmonella*, *Staphylococcus aureus*, *Listeria monocytogenes*, and *Vibrio parahaemolyticus*. J Food Prot **70**:1656–1662.
27. **Carlson BA**, **Nightingale K**, **Scanga JA**, **Tatum JD**, **Sofos JN**, **Smith GC**, **Belk KE**, **Summar O**. 2006. Identification and evaluation of cattle persistently shedding vs. cattle non- persistently shedding Escherichia coli O157:H7.
28. **Hu Y**, **Zhang Q.**, **Meitzler JC**. 1999. Rapid and sensitive detection of *Escherichia coli* O157:H7 in bovine faeces by a multiplex PCR. J Appl Microbiol **87**:867–876.
29. **Vallières E**, **Saint-Jean M**, **Rallu F**. 2013. Comparison of Three Different Methods for Detection of Shiga Toxin-Producing *Escherichia coli* in a Tertiary Pediatric Care Center. J Clin Microbiol **51**:481–486.
30. **Parsons BD**, **Zelyas N**, **Berenger BM**, **Chui L**. 2016. Detection, characterization, and typing of Shiga toxin-producing *Escherichia coli*. Front Microbiol.
31. **Noll LW**, **Shridhar PB**, **Dewsbury DM**, **Shi X**, **Cernicchiaro N**, **Renter DG**, **Nagaraja TG**. 2015. A Comparison of Culture- and PCR-Based Methods to Detect Six Major Non-O157 Serogroups of Shiga Toxin-Producing *Escherichia coli* in Cattle Feces. PLoS One **10**:e0135446.

32. **Newell DG**, **La Ragione RM**. 2018. Enterohaemorrhagic and other Shiga toxin-producing *Escherichia coli* (STEC): Where are we now regarding diagnostics and control strategies? Transbound Emerg Dis **65**:49–71.
33. **Weller D**, **Wiedmann M**, **Strawn L**. 2015. Spatial and temporal factors associated with an increased prevalence of *L. monocytogenes* in spinach fields in New York State. Appl Environ Microbiol **81**:6059–6069.
34. **Wei T**. 2013. corrplot: Visualization of a correlation matrix. R package version 0.73. 0.73.
35. **Bates D**, **Maechler M**, **Bolker B**, **Walker S**. 2014. _lme4: Linear mixed-effects models using Eigen and S4_. R package version 1.1-7. J Stat Softw.
36. **Food and Drug Administration**. 2015. Standards for the Growing, Harvesting, Packing, and Holding of Produce for Human Consumption, Food Safety Modernization Act.
37. **Strobl C**, **Boulesteix A-L**, **Zeileis A**, **Hothorn T**. 2007. Variable Importance Measures: Illustrations, Sources and a Solution. BMC Bioinformatics **8**.
38. **Kuhn M**. caret: Classification and Regression Training. 2018.
39. **Greenwell BM**. 2017. pdp: An R Package for Constructing Partial Dependence Plots. R J **9**:421–436.
40. **Boulesteix A-L**, **Janitza S**, **Hapfelmeier A**, **Van Steen K**, **Strobl C**. 2015. Letter to the Editor: On the term “interaction” and related phrases in the literature on Random Forests. Brief Bioinform **16**:338–345.