Landscape and Meteorological Factors Affecting Prevalence of Three Foodborne Pathogens in Fruit and Vegetable Farms

Strawn, Laura K1, Esther D. Fortes†, Elizabeth A. Bihn2, Kendra K. Nightingale3, Yrjo T. Gröhn4, Randy W. Worobo2, Martin Wiedmann1, and Peter W. Bergholz1

1Department of Food Science, Cornell University, 410 Stocking Hall, Ithaca, NY 14853, USA
2Department of Food Science, Cornell University, New York State Agricultural Experiment Station, 630 W. North St., Geneva, NY 14456, USA
3Department of Animal and Food Sciences, Texas Tech University, Lubbock, TX
4Department of Population Medicine and Diagnostic Sciences, Cornell University, Ithaca, New York 14853
†Present address: Hinton State Laboratory Institute, Massachusetts Department of Public Health, 305 South Street, Jamaica Plain, MA 02130, USA

*Corresponding author:
Peter W. Bergholz
Department of Food Science
401 Stocking Hall
Cornell University
Ithaca, NY 14853
Phone: 607-255-1266
Fax: 607-254-4868
E-mail: pwb49@cornell.edu

Running title: Predicting Foodborne Pathogens in Produce Fields

Key Words: Listeria monocytogenes, Salmonella, STEC, Classification Tree, Prevalence, Diversity, PFGE, Meteorological Events, Landscape Factors, Produce, Food Safety
Abstract

Produce related outbreaks have been traced back to the preharvest environment. A longitudinal study was conducted on five farms in New York State to characterize prevalence, persistence, and diversity of foodborne pathogens in fresh produce fields and to determine landscape and meteorological factors that predict their presence. Produce fields were sampled four times per year for two years. A total of 588 samples were analyzed for *L. monocytogenes*, *Salmonella* and Shiga toxin-producing *E. coli* (STEC). The prevalence measures of *L. monocytogenes*, *Salmonella* and STEC were 15.0, 4.6, and 2.7%, respectively. *L. monocytogenes* and *Salmonella* were detected more frequently in water samples, while STEC was detected with equal frequency across all sample types (soil, water, feces and drag swabs). *L. monocytogenes sigB* gene allelic types 57, 58 and 61, and *Salmonella Cerro*, were repeatedly isolated from water samples. Soil available water storage (AWS), temperature, and proximity to three land cover classes: water, roads and urban development, and pasture/hay grass, influenced the likelihood of detecting *L. monocytogenes*. Drainage class, AWS, and precipitation were identified as important factors in *Salmonella* detection. This information was used in a geographic information systems framework to hypothesize locations of environmental reservoirs where the prevalence of foodborne pathogens may be elevated. The map indicated that not all croplands are equally likely to contain environmental reservoirs of *L. monocytogenes*. These findings advance recommendations to minimize the risk of preharvest contamination by enhancing models of the environmental constraints on the survival and persistence of foodborne pathogens in fields.
Introduction

Despite the implementation of produce safety practices, foodborne outbreaks associated with fresh produce continue to result in significant illnesses, hospitalizations and deaths. Approximately 13% of reported foodborne outbreaks were linked to produce from 1990 to 2005 (1). While this increase in produce-associated outbreaks may, in part, be due to improved surveillance of produce commodities, fresh produce will remain a vehicle for foodborne disease for two reasons: i) increased consumption due to promotion of a healthy lifestyle associated with eating fresh produce and ii) fresh produce commodities are often consumed raw. As a result, contamination with foodborne pathogens at any point in the supply chain from farm to fork has a heightened chance of causing disease (2). Three major bacterial foodborne pathogens: *Listeria monocytogenes*, Shiga toxin-producing *E. coli* (STEC) and *Salmonella*, have been associated with disease outbreaks linked to produce. Together, these pathogens account for an estimated 76% (653/861) of deaths attributed to known bacterial foodborne pathogens in the United States (3).

STEC and *Salmonella* have been responsible for the majority of produce-associated outbreaks (1, 4, 5). Many of these outbreaks were traced back to environmental reservoirs located on the implicated farms. In 2006, one of the first major produce-associated outbreaks linked to preharvest contamination was in spinach. The outbreak-associated strain of *E. coli* O157:H7 was isolated from feral swine, cattle, surface water, sediment and soil (6). An investigation of an outbreak of *Salmonella* Saintpaul in jalapeño peppers recovered the outbreak strain from peppers in the field and irrigation water from one of the implicated farms (7). *L. monocytogenes* has the potential to cause produce-associated outbreaks (8, 9). In 2011, a cantaloupe-borne *L. monocytogenes* outbreak caused 146 illnesses, 30 deaths and 1 miscarriage.
in 28 US states (10). The source of the outbreak was suspected to be a piece of contaminated equipment in the farm packinghouse (11). Although the cause of this outbreak was not due to preharvest contamination of the melons, the results of the outbreak investigation demonstrate the potential risk of *L. monocytogenes* contamination in produce, and the difficulties associated with managing this pathogen in the food safety system, because incidental contamination originating from food or environmental sources can persist through food processing/handling facilities (12).

Studies have demonstrated the ability of foodborne pathogens to survive for extended periods of time in the soil and water, with the potential to infect new hosts and/or contaminate food products (13-15). Laboratory and field studies have identified a number of likely sources of preharvest contamination, such as irrigation water, application of untreated manure, runoff water from livestock operations, and wildlife intrusion into fields (16, 17). Management of farms at the farm or production block scale might greatly influence the local movements of the pathogens and the chance for produce contamination. However, contamination of produce in the preharvest environment remains a complex challenge because the conditions that promote persistence of pathogens in the preharvest environment and subsequent produce contamination are not well-understood. Each farm landscape is a unique combination of numerous environmental characteristics that we hypothesize to set the baseline conditions for persistence of pathogens in or near produce fields.

The focus of the presented research was to better understand *L. monocytogenes*, STEC and *Salmonella* in the produce preharvest environment and, more specifically, to identify specific, remotely-sensed, topographical properties (e.g., proximity to forests), soil properties (e.g., available water storage), and meteorological events (e.g., precipitation) that may influence pathogen prevalence. To that end, we assessed the prevalence, persistence, and diversity of
foodborne pathogens among farms, seasons and sample types. Classification tree (CT) models were used to identify remotely-sensed landscape (i.e., topographic and edaphic) and meteorological characteristics that delineate the presence and absence of foodborne pathogens in the preharvest environment (18, 19). By modeling foodborne pathogen contamination as an ecological process, we seek to supply the food safety researchers and professionals with recommendations to minimize the risk of preharvest contamination.

Materials and Methods

Field Sampling Design. A longitudinal field study was performed on five produce farms in New York State. Farms were selected based on the willingness of growers to participate and to sample farms geographically distributed across New York State (NYS). Farms were not selected based on management practices. Farms were sampled nine times from June 2009 to August 2011. Farms were located in three regions of NYS: Central New York \( n = 1 \), the Finger Lakes \( n = 3 \), and Western New York \( n = 1 \). Distance between farms ranged from 33 to 205 km. Sample size calculations were performed using the lower end of the reported range of prevalence estimates for \( L.\ monocytogenes \) (20), \( Salmonella \) (21), and STEC (6) in order to reach 50 isolates for each targeted pathogen. However, due to time and budget constraints, 588 samples were collected that yielded 107, 27 and 16 representative \( L.\ monocytogenes \), \( Salmonella \), and STEC isolates, respectively. Farms were sampled every astronomical season (summer, fall, winter, and spring). Samples were not collected during snow cover in winter.

Within each farm, four fields were selected to standardize sample sizes among farms since the overall farm sizes varied considerably. Fields were selected that had produce commodities generally consumed raw and to capture topographical field diversity such as low
and high elevation in the field. During each sampling excursion, a single soil sample was collected, consisting of five subsamples of topsoil from five locations in each field. Soil samples were pooled because pathogens were expected to have high spatial variability and small population sizes within the fields (22, 23). One area drag swab and where available, up to five water and fecal samples were collected for each field. In total, seventy-seven (68 surface and 9 engineered), 9 (all engineered), 45 (44 surface and 1 engineered), 18 (9 surface and 9 engineered), and 25 (all surface) water samples were collected from each of the five farms. Fecal samples represented only 10% of the 588 total samples and the majority of fields did not contain feces. Global positioning system (GPS) coordinates were recorded for each sample collected within the field and revisited upon each subsequent visit, in order to access possible persistence of the targeted foodborne pathogens in the preharvest environment. General farm characteristics were documented (Table 1).

**Sample Collection.** Latex gloves and disposable plastic boot covers (Nasco, Fort Atkinson, WI) were worn for sample collection. Gloves and boot covers were changed between each field and gloves were disinfected with 70% ethanol prior to sample collection. A total of 588 samples were collected. Approximately six inch (15.2 cm) deep soil samples and fecal deposits were gathered into sterile Whirl-Pak bags (Nasco, Fort Atkinson, WI) using sterile scoops (Fisher Scientific, Hampton, NH). Pre-moistened drag swabs, as described by Uesugi et al. (14), were dragged around the field perimeter and diagonally back and forth for \( \geq 10 \) min, covering an average field area of 0.75 ha. Drag swabs were deposited into a Whirl-Pak sample bag containing 45 mL of phosphate buffered tryptic soy broth (pTSB, Becton Dickinson, Franklin Lakes, NJ). Water samples were collected directly into sterile 250 mL jars by use of a 3.66 m sampling pole (Nasco, Fort Atkinson, WI). These water samples were taken a minimum of 2 m from the
shoreline and 0.3 m below the surface. All samples were transported on ice, stored at 4±2°C and processed within 24 h of collection.

**Sample Preparation.** All samples were used for three separate enrichment schemes to allow for the isolation and detection of (i) *L. monocytogenes*, (ii) *E. coli* O157:H7 and (iii) a combined enrichment for non-O157 STEC and *Salmonella*. The five soil samples collected in each field were weighed into 5 g portions, and the portions were combined to form a 25 g pooled sample. Three 25 g pooled soil samples were prepared and deposited in sterile filter Whirl-Pak bags. For fecal samples, 2 to 10 g of each fecal sample collected was deposited into three sterile filter Whirl-Pak bags. Drag swab samples were mixed with 90 mL pTSB in the Whirl-Pak bag by hand massaging for 2 min. The drag swab was then squeezed and 10 mL of the liquid contents from the sample bag was aseptically transferred to each of three sterile filter Whirl-Pak bags. Water samples were processed according to Environmental Protection Agency (EPA) standard methods (24, 25). Each water sample collected (250 mL) was passed through a 0.45 μm filter unit (Nalgene, Rochester, NY). This filter was then aseptically removed, cut into three equal-sized pieces, transferred to separate sterile filter Whirl-Pak bags.

*L. monocytogenes enrichment and isolation.* *L. monocytogenes* detection and isolation from environmental samples collected was performed as in previous studies (26-28). Briefly, samples were diluted 1:10 with buffered *Listeria* enrichment broth (BLEB; Becton Dickinson, Franklin Lakes, NJ). These enrichments were incubated at 30±2°C for 4 h. At 4 h, *Listeria Selective* Enrichment Supplement (Oxoid, Cambridge, UK) was added. After 24 h and 48 h incubations at 30±2°C, 50 μl of each enrichment was streaked onto modified Oxford agar (MOX, Becton Dickinson, Franklin Lakes, NJ) and *L. monocytogenes* plating medium (LMPM, Biosynth International, Itasca, IL). MOX and LMPM plates were incubated for 48 h at 30 and 35±2°C,
respectively. Up to ten *L. monocytogenes* presumptive colonies were sub-streaked to brain heart infusion agar (BHI; Becton Dickinson, Franklin Lakes, NJ). BHI agar plates were incubated for 37±2°C for 24 h. Presumptive *L. monocytogenes* colonies were confirmed by polymerase chain reaction (PCR) and partial *sigB* gene sequencing (29-31).

**E. coli O157:H7 enrichment and isolation.** Samples in Whirl-pak bags were diluted 1:10 with pTSB and incubated for 2 h at room temperature (23±2°C) to aid in the recovery of injured cells (32). Enrichments were transferred to 42±2°C and incubated for 24 h. Enrichments were subjected to immunomagnetic separation (IMS) to concentrate *E. coli* O157:H7 cells as previously described (33). Washed IMS beads (50 μl) were plated onto two selective and differential media: modified sorbitol-MacConkey agar (mSMAC; Becton Dickinson, Franklin Lakes, NJ) supplemented with 20 mg/L of novobiocin and 2.5 mg/L of potassium tellurite (Sigma-Aldrich, St. Louis, MO) and CHROMagar O157 agar (CHROMagar, Paris, France). CHROMagar O157 and mSMAC plates were incubated at 37±2°C for 24 and 48 h, respectively. Up to ten presumptive *E. coli* O157:H7 colonies were sub-streaked onto BHI and incubated at 37±2°C for 24 h. Presumptive *E. coli* O157:H7 colonies were confirmed using a multiplex PCR assay that simultaneously screens for *hlyE*, *fliC*, *eaeA*, *rfbE*, *stx*-I and *stx*-II as previously described (34, 35).

**Non-O157 STEC and Salmonella enrichment and isolation.** The non-selective enrichment step (pTSB) for both pathogens was the same. Environmental samples were diluted 1:10 with pTSB and incubated for 2 h at 23±2°C, followed by a 24 h incubation at 35±2°C. To isolate non-O157 STEC, a 1 mL aliquot of the non-selective enrichment was transferred to 9 mL of *E. coli* broth (EC broth; Oxoid) and incubated at 37°C with shaking for 24 h. A 50 μl aliquot of EC broth was plated onto washed sheep’s blood agar (Hemostat, Dixon,
with 10 mM CaCl₂ and 0.5 ug/mL mitomycin C (WBMA; Fisher Scientific, Hampton, NH) and incubated at 35±2°C for 24 h. Up to 20 colonies that demonstrated enterohemolysis were sub-streaked to SMAC plates and incubated at 37±2°C for 24 h. Up to ten colonies that rapidly fermented sorbitol were sub-streaked to BHI and incubated at 37±2°C for 24 h. Presumptive non-O157 STEC colonies were confirmed by the multiplex PCR described above (35) and considered positive if one or both stx gene(s) were detected. 

*Salmonella* detection and isolation was performed using a modified version of the procedures outlined in the Food and Drug Administration’s Bacteriological Analytical Manual (36). A 1.0 and 0.1 mL aliquot of non-selective pTSB enrichment was transferred to 9 and 9.9 mL of tetrathionate (TT; Oxoid) and Rappaport Vassiliadis (RV; Oxoid; Fisher; Acros Organic, Belgium), respectively. These selective enrichment cultures were incubated in a shaking water bath at 42±2°C for 24 h. A 50 μl aliquot selective enrichment was plated onto xylose lysine deoxycholate agar (XLD; Neogen, Lansing, MI) and CHROMagar *Salmonella* (CHROMagar) agar, and incubated at 35 and 37±2°C for 24 and 48 h, respectively. Up to 20 presumptive *Salmonella* colonies were sub-streaked to BHI and incubated at 37±2°C for 24 h. Presumptive *Salmonella* colonies were confirmed using a previously described PCR assay that detects invA, a gene specific to *Salmonella enterica* (37).

**Controls and storage.** Positive and negative controls were processed in parallel with each pathogen detection and isolation scheme. The following strains were used as positive controls: FSL R3-001 for *Listeria monocytogenes* (actA deletion mutant; (38)), ATCC 43895 strain tagged with GFP (FSL F6-825) for *E. coli* O157:H7 (39), FSL F6-704 for non-O157 STEC (*E. coli* O26:H11), and ATCC 700408 strain tagged with GFP (FSL F6-826) for *Salmonella* (56).
Negative controls were sterile enrichment media. All isolates were preserved at -80°C in 15% glycerol.

**Characterization of Isolates.** All *L. monocytogenes*, *STEC* (*E. coli* O157:H7 and non-O157 STEC), and *Salmonella* isolates were streaked from frozen culture onto BHI and incubated at 37°C for 18 h and a well-isolated colony was selected. Nucleotide sequences of *sigB* from *L. monocytogenes* isolates were obtained by Sanger sequencing preformed by the Cornell University Life Sciences Core Laboratories Center and compared with those in the GenBank database using BLASTN to assign allelic types, i.e., a unique combination of polymorphisms (40, 41). *L. monocytogenes* isolates that shared the same allelic type from the same location at least three times were considered possible persistent subtypes and were further subtyped by pulsed field gel electrophoresis (PFGE). PFGE typing was performed using the standard CDC PulseNet protocol with the restriction enzymes AscI and Apal (42). *Salmonella* ser. Braenderup digested with XbaI was used as the reference standard, which allowed for normalization and comparison of gel images (43). Pattern images were captured with a Bio-Rad Gel Doc and the Multi Analyst software (Bio-Rad Laboratories, Hercules, CA). PFGE banding patterns were analyzed using BioNumerics (Applied Maths, Saint-Matins-Latem, Belgium). Comparisons were performed using similarity analyses by using the unweighted pair group-matching algorithm (UPGMA) and the Dice correlation coefficient with a maximum space tolerance of 1.5%.

To confirm the identity of O157:H7 isolates and to determine the serotype of non-O157:H7 STEC isolates, comprehensive O serotyping and H typing were performed on one representative STEC isolate per positive sample at the *E. coli* Reference Center at Pennsylvania State University (State College, PA), as previously described (44, 45).
Salmonella cultivation methods use four combinations of selective enrichments and plating media. To account for the possibility of different strains of Salmonella being isolated from the sample, one representative isolate per Salmonella positive sample from each isolation scheme (e.g., TT to XLD or RV to XLD) was selected for molecular subtyping (46). Serotyping and PFGE were performed on isolates selected. Serotyping, using the White-Kauffman-Le Minor (formerly known as the Kauffman-White scheme) scheme, was performed by the Wadsworth Center, New York State Department of Health (Albany, NY) (47). PFGE typing was performed according to the standard CDC PulseNet protocol for Salmonella using the restriction enzyme XbaI (48).

Descriptive Data Analysis. Associations of pathogen positive cultures with farm, season, or sample type were determined using a chi-square test. A Fisher’s exact test was used if the expected frequency in any cell was less than 5. Confidence intervals were calculated assuming a binomial distribution. Individual p values were calculated and were considered statistically significant if less than 0.05. Bonferroni’s correction was used to account for multiple testing of the three statistical hypotheses (farm, season and sample type) (49). The diversity of subtypes within farm, season and sample type was quantified using Simpson’s Index of Diversity (D) (50). All statistics for descriptive analyses were performed in SAS 9.1 (SAS Institute Inc., Cary, NC).

Topographical and Soil (spatial) data. Spatially dependent predictor data were obtained for each sample site (Table S1). Global positioning system (GPS) coordinates of samples were imported into the Geographical Resources Analysis Support System (GRASS) Geographic Information Systems (GIS) environment (51). Site coordinates were re-projected from latitude-longitude into the Universal Transverse Mercator (UTM) coordinate system, North American Datum of 1983. Map layers for land cover (NLCD; National Land Cover Database 2006) and...
digital elevation model (DEM; Shuttle Radar Topography Mission, 1 arc-second dataset) were acquired from the US Geological Survey (USGS) EarthExplorer geographical databank (http://earthexplorer.usgs.gov/). Map layers for soil characteristics were acquired from the US Department of Agriculture Soil Survey Geographic database (SSURGO) (http://soils.usda.gov/survey/geography/ssurgo/). Road and hydrologic line graphs were obtained from the Cornell University Geospatial Information Repository (CUGIR; http://cugir.mannlib.cornell.edu/). Proximity data were derived from the NLCD land cover basemap by calculating euclidean nearest-neighbor distance to the desired land cover type. Proximity to urban areas was calculated from a map combining road lines with all four classes of developed land cover. Proximity to water was calculated from a map combining water body areas and flow lines. Percent slope was derived from the DEM. In total, 15 different landscape factors were obtained for CT model development, such as soil type; slope; drainage class; available water storage; organic matter; and proximity to urban development, pastures, forests, and water (Table S1).

**Meteorological (temporal) data.** For each sample collection date, meteorological variables were obtained from the major airport nearest to each farm, using the airport weather stations in the National Oceanic and Atmospheric Administration (NOAA) National Climate Data Center (NCDC) Local Climatology Database (http://gis.ncdc.noaa.gov/map/lcd/). A major airport was within 60 miles of each farm used in the study. While small scale differences in weather may be observed between airport and farm, the study aimed to capture the association between remotely-sensed meteorological data and pathogen prevalence. In total, 70 different meteorological factors were obtained for CT model development, including temperature (maximum, minimum and daily average) and precipitation amounts (Table S1). Direct measures of temperature and
precipitation were acquired for the day of sampling and three days antecedent. The average temperature and precipitation was calculated for each time period ranging from 1 to 10 days prior to sample collection. Frost cycles were counted by summing the number of times the surface air temperature fluctuated above and below 0°C for each time period ranging from 1 to 10 days prior to sampling. Averaging and frost calculations were performed using a Perl script (Code available from PWB).

Spatial and Temporal Data Analysis. Methods used in our analysis of spatial and temporal factors were adapted from Ivanek et al. (19). Large numbers of landscape and meteorological variables were included in our classification analysis as possible predictors of pathogen presence (Table S1). Since there was high potential for covariation among landscape and meteorological predictors of pathogen presence, detrending and principal components analysis (PCA) techniques were applied to account for the linear covariation among predictors. PCA was performed using the ade4 package in R 2.13.1 (52).

It was desirable to account for season, temperature and precipitation as independent factors predicting pathogen presence, but season and the meteorological variables did not behave independently at monthly time scales. Temperature and precipitation were de-trended for the seasonal effect by performing linear regressions and retaining the residuals from these regressions to represent variation of temperature and precipitation within seasons. Soil properties and elevation were also dependent on the general farm properties, so to examine the effects of soil property and elevation variation within farms, these characteristics were de-trended against farm using linear regressions.

De-trended residuals were standardized and used as input for two PCAs to synthesize variation among meteorological and landscape data, respectively, into eigenvectors representing
the characteristic behavior of these variables. PCA on meteorological variables yielded an
eigenvector that represented 56.1% of the total variation and corresponded well to all
temperature variables except average temperature 3 days prior to sampling. The same PCA
yielded a second eigenvector describing 18.4% of the total variation that corresponded well to all
precipitation variables except precipitation on the day before and the day of sampling. A second
PCA showed that landscape data demonstrated less covariation among landscape data. This PCA
yielded a single useful eigenvector, representing 51% of the total variation. Available water
storage and soil organic matter properties were loaded on this eigenvector, but topographic data
were retained as independent predictors of pathogen presence. These three eigenvectors were
used as predictor variables in CT models, as they synthesized characteristics of multiple, co-
varying temperature, precipitation, and soil variables, respectively. This allowed us to minimize
the number of predictor variables in the CT models to those that behaved independently.

**Classification Tree (CT) Model Development.** Tree-based modeling was used to determine
rules that classified sampled sites by pathogen presence or absence. Splits were formed by
maximizing homogeneity of presence vs. absence results in each node according to the Gini
index (18). The CTs were built using the rpart package in R 2.13.1 (53). To assess predictive
power of resulting trees, a cross-validation procedure was performed 25 times for each tree. The
detection methods for foodborne pathogens in the environment are not 100% sensitive or
specific; therefore the response variable was weighted to maximize the predictive power of the
resulting tree. To limit the potential effect of different CT outcomes based on weighting the
response variable, we performed a sensitivity analysis in which different weights were applied to
negative samples to reflect probabilities of false negatives. The weight of positive samples was
always set to 1. The weights for negative results were varied until the weight that minimized
cross-validation error was discovered. This weighting scheme was used to produce CTs with 25-fold cross-validation. CTs were pruned to the number of splits that minimized cross-validation error within the selected weighting scheme. This combination of procedures resulted in CTs of reproducible size, predictive power, and split rules and the subsequent analysis of *L. monocytogenes* results by random forest using the CT model supported the CT outcome entirely (data not shown; (54)).

**Geospatial search for *L. monocytogenes* reservoirs.** Classification trees, and related techniques, result in rules that can be used to predict the most likely areas to observe a species (18). Using *L. monocytogenes* as an example, the rules from the *L. monocytogenes* CT were applied in a GIS framework to explore the potential for croplands to harbor persistent *L. monocytogenes* in a central New York State landscape (see Results for rule definitions). All calculations on maps were performed using GRASS GIS 6.4.1. Raster maps of a) water features and flow lines and b) pasture areas were extended to reflect proximity-based split rules from the CT using the spatial buffering function r.buffer. These rasters were then converted to vector maps, and used in vector map queries using v.overlay to determine cropland areas corresponding to three categories of hypothetical *L. monocytogenes* reservoir based on CT results. Reservoir polygon areas and minimum reservoir distance from pasture class land areas were calculated using statistical functions in GRASS GIS.

**Accession Numbers.** Isolate information and subtyping data from this study are archived and available through the Food Microbe Tracker database ([http://www.foodmicrobetracker.com](http://www.foodmicrobetracker.com)) using a guest user login. To facilitate batch retrieval of isolate records, the accession numbers have been tabulated (Table S4).
Results

*L. monocytogenes* prevalence. *L. monocytogenes* prevalence was estimated to be 15.0% (88/588) across all samples collected. Farm, season, and sample type were found to be significantly associated with the frequency of *L. monocytogenes* positive samples (Table 2). Over the nine collection periods, winter had a consistently higher prevalence of *L. monocytogenes* than all other seasons, the only exception being the summer of 2010. Farm 1 showed a significantly higher prevalence of *L. monocytogenes* compared to Farm 2 (Table 2). The prevalence of *L. monocytogenes* was highest amongst water samples (48/174). All *L. monocytogenes* positive water samples were from surface water (e.g., creek or pond); none of the 28 samples from engineered water sources (e.g., municipal or well) were positive for *L. monocytogenes* (Table 2).

*Salmonella* prevalence. The prevalence of *Salmonella* across all samples was 4.6% (27/588). Farm and sample type were significantly associated with the frequency of *Salmonella* positive samples (Table 2). While there was no significant seasonal association, *Salmonella* prevalence was greatest in the 2010 and 2011 summers (7.8 and 8.3%, respectively). Farm 1 showed a significantly higher prevalence of *Salmonella* compared to Farms 2 and 5 (Table 2) possibly due to the co-management of the produce operations on farm 1 with livestock operations located nearby. The prevalence of *Salmonella* was significantly higher in water samples (16/174) compared to soil and drag swab samples (4/178 and 3/175, respectively), but similar to fecal samples (4/61) (Table 2). All of the 16 *Salmonella* positive water samples originated from surface water (Table 2).

*STEC* prevalence. The prevalence of STEC was 2.7% (16/588) across all samples. Four samples tested positive for *E. coli* O157:H7, including a (i) drag swab sample from a pepper field, (ii) drag swab sample from a sweet corn field, (iii) water sample from a drainage ditch, and
(iv) water sample from a creek. None of the factors (e.g., farm, season and sample type) were shown to have a significant association with the frequency of STEC positive samples (Table 2). Similar to findings for *L. monocytogenes* and *Salmonella*, all four STEC positive water samples were from surface water (Table 2).

*L. monocytogenes* diversity. A total of 107 *L. monocytogenes* isolates were obtained from the collection of 88 environmental samples in which this pathogen was detected. Alignment of *sigB* nucleotide sequences for the 107 isolates showed 12 different *sigB* allelic types. Allelic types belonged to *L. monocytogenes* lineage I, II and IIIa (6, 5, and 1 allelic types, respectively). There was a high diversity of *L. monocytogenes* allelic types amongst farms, seasons, and sample types (D=0.80, 0.78, and 0.85, respectively).

There were four cases of repeat isolation which was defined as the same *sigB* allelic type being isolated three or more times from the same sample site over time. These isolates were subtyped further using PFGE with restriction enzymes *AscI* and *ApaI* (Fig. 1). Analysis of *L. monocytogenes* PFGE showed multiple PFGE patterns for three of the four cases of repeat isolation; however, one case showed identical PFGE patterns for 3 of the 4 isolates obtained from the same water sample site.

*Salmonella* diversity. Serotyping and PFGE was conducted on one representative isolate from each isolation scheme in 26 out of 27 positive samples for a total of 57 *Salmonella* isolates. No isolate was available for typing from one *Salmonella* sample, because preservation failed. One of the 26 available samples yielded a different PFGE type under the four isolation schemes. The two PFGE types were confirmed as *Salmonella* serovars Newport and Thompson. All other isolates from the isolation schemes had identical PFGE types within a sample. The remaining 25 *Salmonella* positive samples contained *Salmonella* serotypes Cerro (10 samples), Newport (5
samples), Thompson (4 samples), Give (2 samples), IV 40:z4,z32:- (2 samples), Typhimurium (1 sample), and I 6,8:i:- (1 sample) (Fig. 2). The 7 *Salmonella* serotypes corresponded to eleven different PFGE types (Fig. 2). Overall, there was a high level of diversity amongst *Salmonella* serotypes and PFGE types in the produce preharvest environment (serotype D=0.84 and PFGE type D=0.80). *Salmonella* Newport was isolated from two fecal samples and one soil sample from the same field on farm 4 (Fig. 2).

Repeat isolation of a *Salmonella* serotype was also observed. *Salmonella* Cerro was isolated from a water source three times during the nine collection periods (Fig. 2). The water sample was collected from a creek that was across the road from a field on Farm 1. *Salmonella* Cerro is highly clonal and this particular PFGE pattern matches 89% of *Salmonella* Cerro PFGE patterns (55, 56).

**STEC diversity.** Serotyping was conducted to further characterize the 16 STEC isolates. Six distinct O and H serotype results were observed. Serotypes O157:H7 (4/16 isolates) and O8:H19 (4/16 isolates) represented half of the 16 STEC isolates. Additional, serotypes identified were O26:H11 (1/16 isolates), O:-H- (2/16 isolates), OX25:H11 (3/16 isolates), and O91:H49 (2/16 isolates).

**Classification of High and Low Prevalence Samples.** Classification tree (CT) models were fit using sample presence/absence data in order to further explore the environmental and topographic variables that were associated with the detection of *L. monocytogenes* and *Salmonella* at smaller scales of variation than farm, season or sample type (Fig. 3 and 4). CTs start with a root node containing all samples and recursively split sample sites by minimizing the mixture between positive and negative environmental samples for the selected foodborne pathogen. CTs often determine multiple possible rules useful for splitting samples (Fig. 3 and 4).
Primary splits exhibited the best improvement score for dividing positive and negative samples into separate nodes; the rule with the second best improvement score was considered a competitor against the primary rule, except in cases where this rule was informationally redundant. In these cases, the next best competitor was selected for display (Text S2 and S3).

Surrogate rules represent the predictor that best correlates to the primary rule for the split, and they are used by the algorithm to fill in missing data for the primary rule. These surrogates mimic the primary rule and produce a split with a similar division of positive and negative samples in daughter nodes. No CT was developed for STEC because the trees only produced a root node.

The *L. monocytogenes* CT that gave negative samples one quarter the weight of positive samples resulted in the lowest relative cross-validation error at 0.65 (Fig. 3). The CT determined that the prevalence of *L. monocytogenes* in samples collected within 37.5 m of mapped waterways was 39% (29/74). All 74 *L. monocytogenes* samples within 37.5 m of mapped waterways were surface water samples. All *L. monocytogenes* positive terrestrial samples (n=40) were located farther than 37.5 m from mapped waterways. In samples locations ≥ 37.5 m from water, the eigenvector describing temperature variables split samples such that temperatures that were lower than approximately 2°C below average had 21% prevalence, but samples from warmer temperatures had only 7% prevalence. Using the remotely-sensed average temperature over 5 days prior to sampling, the split rule (Eigenvector_1_Temporal 1.483) corresponded to < 14°C in summer, < 10°C in spring, or < 5°C in winter. The eigenvector for soil properties (Eigenvector_1_Spatial in Fig. 3) included available water storage and soil organic matter. Use of predicted available water storage from the SSURGO database as a representative value revealed soils with available water storage in 0-25 cm depth > 4 cm yielded samples with 31%
prevalence versus 10% prevalence in less moist soils. Proximity to pastures was also identified as an important factor in the prediction of *L. monocytogenes* prevalence. Moist terrestrial soil locations sampled at cooler temperatures within 62.5 m of a pasture had 50% (n=25/50) versus 7.5% (3/40) *L. monocytogenes* prevalence in similar sample locations further than 62.5 m from pasture-class land areas. Land use classes were highly interspersed in the areas surrounding sampled farms, so the sample locations meeting the criteria for 50% *L. monocytogenes* prevalence occurred on four of five sampled farms, indicating the proximity to pastures rule was not the product of bias due to the fact that farm 1 had higher prevalence and shorter distances to pasture than the four others. A terminal node was also identified for proximity to urban development or roads (i.e., impervious surface coverage). Locations within 9.5 m of an impervious surface had a predicted *L. monocytogenes* prevalence of 20%, compared to 5% for locations farther than 9.5 m (Fig. 3).

The *Salmonella* CT that gave negative samples one twentieth the weight of positive samples yielded the lowest relative cross-validation error, 0.67, across all our attempted weighting schemes (Fig 4). Drainage class was identified as the most important factor delineating locations of high or low *Salmonella* prevalence. A location where drainage is classified as very poor, somewhat poor, poor, and somewhat drained was determined to have a higher *Salmonella* prevalence (9%) than a location where drainage is classified as moderately well drained and well drained (1.2%). After the tree determined that poorly drained soils contained more *Salmonella*, the algorithm then produced a rule indicating that soils near the upper limit of available water storage were more likely to be negative. *Salmonella* was less likely to occur in soils with available water storage (at 0-25 cm) of 10 cm, which was the maximum value in the soil database. The second temporal eigenvector, which described temporal variation...
in precipitation, formed another split. In areas with poorly drained soils, *Salmonella* was more prevalent (12%) when measurable precipitation occurred within 3 days prior to sampling.

**Geospatial prediction of terrestrial *L. monocytogenes* reservoirs.** The CT contained topographic and soil property rules that might be useful to map the locations of environmental reservoirs of *L. monocytogenes*. In order to explore the usefulness of the CT rules, a map was generated to represent a CT-based hypothesis about the locations and spatial extents of *L. monocytogenes* environmental reservoirs in a 9024 ha mixed land cover area in central New York State (Fig 5). We hypothesize that these locations can harbor *L. monocytogenes* within croplands and are more likely to be positive when sampled than other parts of produce fields.

Three classes of environmental reservoirs were extracted to produce this map: i) areas within 38 m of mapped surface water, ii) areas outside class (i) with soil available water storage (AWS) > 4.0 cm in the top 25 cm and within 62 m of mapped pasture land cover, and iii) areas with AWS > 4.0 cm in the top 25 cm but outside classes (i) and (ii). All classes were clustered spatially to the west of the main stream in the map, indicating that not all portions of the landscape provide equally good reservoir habitats for *L. monocytogenes*, because soil properties differed east of the stream. The algorithm identified 221 cropland reservoirs within 38 m of water (Fig 5, light blue; 45% positive samples) and had areas ranging from 0.01 to 4.79 ha (median=0.07 ha). One hundred ninety-two reservoirs in higher available water storage soil areas (Fig 5, cyan; 7.5% positive samples) ranged from 0.1 to 16.2 ha in area (median=0.6 ha). Two hundred thirty-five reservoirs within 62 m of pasture class land cover and in soils with higher available water storage (Fig 5, dark blue; 50% positive samples) ranged from 0.1 to 5.2 ha in area (median=0.5 ha).

**Discussion**
One of the largest practical challenges to ensuring produce safety is to optimize the expenditure of financial and labor resources such that the pathogen will be detected where it is a public health risk. While many aspects of pathogen biology in the environment have been the topic of research (13, 57), field data are lacking on the nature of environmental pathogen reservoirs in produce fields. As a result, farmers and food safety professionals have little data on which to base sampling schemes that are intended to detect foodborne pathogens in the preharvest environment. This study describes the environmental distribution of three foodborne pathogens at the farm and field scales. We propose that using these analyses provides a means to improve surveillance for foodborne pathogens in produce fields by describing environmental variables that constrain the prevalence of pathogens. These data may also be used to identify areas of high and low predicted pathogen prevalence within farms, enabling more informed decisions about the management of crops associated with foodborne disease outbreaks. While this study does not directly measure the risk of produce contamination, these data can be used to support the development of risk models.

The present research was conducted by collecting diverse sample types on privately operated farms. While this has the advantage of enabling the collection of data in farms that are faced with the management challenges and practical considerations of businesses, it had the disadvantage that sample collection was dictated in large part by convenience to the land owner, so the sampling design quickly became unbalanced and smaller than was originally planned. However, it is important to note that field ecology studies frequently feature unbalanced and under-sampled designs (18). As an alternative to violating the assumptions of logistic regression analysis, classification trees are an ideal method to analyze such unbalanced and under-sampled designs (19). The method makes a single, simple, assumption about the distribution of
presence/absence observations: that these data can be subdivided according to environmental
data to maximize the homogeneity of the observed results.

Prevalence of key foodborne pathogens is higher in water and fecal samples. Water has been
identified as both a reservoir and transmission pathway for foodborne pathogen contamination of
produce (17, 58). Moreover, all three pathogens examined in this study are known to be common
contaminants of agricultural watersheds (59-62). The range of L. monocytogenes and Salmonella
prevalence in watersheds has been estimated from 6.4 to 62% and 6 to 80%, respectively, based
on the region of study (62-66). A 7% prevalence of Salmonella was observed in water samples
obtained from a produce growing region in California (67). Similarly, in the participating NYS
produce farms we obtained the highest prevalence of L. monocytogenes and Salmonella from
water samples demonstrating water sources are a potential pathogen reservoir.

However, we recognize there may be a potential bias toward a higher prevalence of
pathogens in water samples compared to soil samples in this study. The abundance of foodborne
pathogens in the environment is expected to be low so there may be a greater chance of detecting
a pathogen in water samples because the pathogen may be distributed more uniformly in water
compared to soil samples. The five subsamples of soil collected were pooled in one composite
sample per field, which may contribute to a lower prevalence estimate. Positive L.
monocytogenes and Salmonella samples collected from on-site surface water sources were
mostly from small waterways, no more than 3 m wide, as many of the sampled produce farms
are surrounded by residential land. Previous research has shown that lower order waterways
often receive more direct agricultural drainage and runoff from livestock production
environments compared to higher order systems (60, 61).
Foraging wildlife may also contribute to contamination of fields as demonstrated by isolation of *Salmonella* Newport from soil and wildlife feces (deer) from a single field. Wild and domestic animals are widely known to harbor foodborne pathogens (6, 67, 68). A review of *Salmonella* in wild and domesticated animals (68) determined the prevalence of *Salmonella* in reptiles (6-100%), poultry (0-60%), cattle (2-42%), swine (3.5-28%), rodents (1-15%), birds (3-13%), and domestic cats and dogs (1-5%), which reflects the potential for animals to be sources or vectors of preharvest contamination. An optimal solution for the co-management of wildlife habitat, environmental quality, and food safety is the target of current produce safety research and our research facilitates these efforts by describing the environmental prevalence of pathogens in preharvest environments relative to potential wildlife habitats on the same landscape (69).

Soil properties and topographic features were identified as constraints on pathogen prevalence in produce fields. The *L. monocytogenes* and *Salmonella* CTs both demonstrate that not all croplands have an equal risk of foodborne pathogen contamination. Soil characteristics and topographic variables corresponding to proximity of sampled areas to other landscape types, including impervious surface (e.g., buildings, roadways), water, or pasture were identified as factors (i.e., primary rules in the CT) for predicting locations containing pathogens. The primary split for both pathogen CTs were associated with water features, specifically proximity to water and soil drainage class for *L. monocytogenes* and *Salmonella*, respectively. This finding further demonstrates the ability of on-site water sources as potential reservoirs and transmission pathways for foodborne pathogen contamination in the preharvest environment (4, 17, 70).

Proximity to pastures was a factor influencing the likelihood of detecting an *L. monocytogenes* positive sample. It is important to note that proximity to pastures was obtained
through remotely sensed data, and pasture class land-cover can indicate active pasturages, livestock pens, and hay grass fields. It has been shown that livestock shedding and subsequent run-off of foodborne pathogens may be one of the major sources of preharvest contamination (17, 71, 72). Ruminants can shed significant numbers of *L. monocytogenes* while being asymptomatic, and may release the pathogen into the environment (26, 56). A strong association was shown between the prevalence of *L. monocytogenes* and proximity to cattle and dairy farms in watersheds impacted in agricultural landscapes (27, 60).

Proximity to impervious surfaces was also identified as a factor for the classification of high or low *L. monocytogenes* prevalence locations. Some wildlife carriers of foodborne disease, e.g. rodents and ground nesting birds, use roadside ditches as nesting habitats and may enter croplands from ditches while foraging for food (73). This behavior, particularly in short-dispersal distance species, may cause pathogen prevalence to be amplified in the edges between residential land and cropland. Impervious surfaces are constructed to remove excess water, usually into bordering drainage ditches, and these may channel fecal contamination. This contamination could subsequently be spread out of the ditch by local wildlife, heavy precipitation, or human activities.

Soil characteristics, specifically available water storage, soil drainage class, or soil organic matter, were important factors in CTs generated from both *L. monocytogenes* and *Salmonella* prevalence data. Pathogen survival has been shown to increase in moist soils (74, 75). *E. coli* and *Salmonella* held in dry soil for 14 d demonstrated the ability for growth after the soil was moistened with sterile distilled deionized water (76). In addition, it was shown that *E. coli* and *Salmonella* persisted longer in moist soils compared to drier soils. Similarly, *E. coli* numbers were shown to decrease faster in dry soils compared to moist soils (77).
Weather plays a role in pathogen prevalence in the preharvest environment. Meteorological factors were identified in both the *L. monocytogenes* and *Salmonella* CTs, indicating that recent temperature and precipitation can influence foodborne pathogens in the environment, which is consistent with previous findings (19, 62, 67, 78). *L. monocytogenes* was more frequently detected in cooler temperatures that were above freezing prior to sample collection in our study. Two previous studies (66, 79) observed a higher prevalence of *L. monocytogenes* in spring and winter-spring. Ivanek et al. (2009) determined that fewer freeze/thaw cycles increased the detection of *Listeria* spp. in vegetation samples (19). *L. monocytogenes* can grow and survive in a wide range of temperatures, but freezing can have an inhibitory effect (80, 81). However, temperature was not found to be a factor in the CT for *Salmonella*. In contrast, other studies (62, 78) have identified temperature or season to be associated with frequencies of *Salmonella* in environmental settings. It may be concluded the ability to predict an increased likelihood of detecting a positive *Salmonella* sample, as for other species, may be dependent on local ecology and agriculture practices of the specific location of the study.

Precipitation was identified as a factor influencing the detection of *Salmonella* positive samples in our study. It has been suggested that heavy rain and storm events may aid in the transport of pathogens (82, 83) and potentially lead to higher loads of bacteria in the water along with sediment (84, 85). Foodborne pathogens can survive in sediments for substantial periods (86). High water flow rates have also been observed to influence pathogen incidence levels in the water, and may transport pathogens up to 32 km (87). Such long transport distances in waterways may have important implications for the diversity and source tracking of foodborne pathogens impacting preharvest environments (64, 71). It may be necessary to analyze flux of pathogens from potential sources, like livestock pasture, through hydrologic connectivity.
networks in a dynamic framework that accounts for precipitation in order to accurately estimate risk of *Salmonella* contamination in croplands where it is expected to result from contaminated water sources. The Soil and Water Assessment Tool (SWAT), while unsophisticated in its treatment of bacteria as simple particles, is one tool available to model this process.

**Benefits to Predicting the Presence of Foodborne Pathogens in Produce Environments.**

Transmission of foodborne pathogens to produce in the preharvest environment is a complex process, involving multiple vehicles that transport pathogens from sources (e.g., pasture areas) to sinks (e.g., moist soils in fields). The development of practical tools to predict the presence of pathogens in produce fields or the risk of produce contamination is further complicated by the fact that every produce farm contains a unique combination of spatial and temporal variables. These variables undoubtedly influence the ecology of foodborne pathogens in the environment and may influence the potential for product contamination. Indeed, the presented analysis indicates that, while the average prevalence of foodborne pathogens in randomly-collected environmental samples is low, local prevalence can be significantly higher under specific combinations of environmental and local land-use conditions. An essential component to developing a mechanistic approach to understanding foodborne pathogen transmission to produce is to characterize environmental reservoirs as favorable or unfavorable pathogen persistence sites. Since CT analysis generates concise rules to delineate pathogen positive and negative sites, application of the resulting rules to remotely-sensed data about farm landscapes can enable the development of specific predictions about expected pathogen presence for any individual produce field. While the CTs presented here require further validation to determine their ultimate usefulness in produce farms, these models advance GIS-enabled modeling to predict the risk of produce contamination. Additionally, fully developed models of how
pathogens disperse and persist in the preharvest environment may also permit the development of land management strategies to minimize produce contamination by possibly allowing growers to select crops for these sites that are less susceptible to contamination.

Acknowledgements.

This research was supported by USDA NIFSI-2008-51110-0688 (R.W. Worobo) and NIFSI-2008-51110-04333 (K.N. Nightingale and M. Wiedmann). We are grateful for the technical assistance of Steven Warchoki and Emily Wright.

References.


11. FDA. 2012. Final Update on Multistate Outbreak of Listeriosis Linked to Whole Cantaloupes.


35. Hu, Y., Q. Zhang, and J. C. Meitzler. 1999. Rapid and sensitive detection of 


87:867-876.


Bacteriological Analytical Manual, Chapter 5, Salmonella. Food and Drug 

Administration. [Online.]


S. T. Kwon. 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.


mutagenesis probe the contribution of ActA amino-acid variability to phosphorylation 

and virulence-associated phenotypes among *Listeria monocytogenes* strains. FEMS 


Development of green fluorescent protein-expressing bacterial strains and evaluation for 


40. Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. 


FIGURES LEGENDS

FIG. 1. AscI and ApaI PFGE patterns of the four repeat isolation cases of *L. monocytogenes*. The four cases of repeat isolation are as follows: FSL-S10-009, 084, 366 and 598; FSL-S10-020, 086, and 601; FSL-S10-298, 604, and 1591; and FSL-S10-306, 335, 609, 1363, and 1490. In one case (top three PFGE patterns), three of four isolates (FSL-S10-009, 084, and 366) have identical PFGE patterns. Band sizes (kb) are displayed at the top of the PFGE pattern images. PFGE pattern order displayed is result of sample site (manually ordered within BioNumerics).

FIG. 2. XbaI PFGE patterns of the 27 *Salmonella* isolates representing the 26 *Salmonella* positive samples available for typing. One isolate per isolation scheme was typed; only one representative PFGE pattern is shown. One sample yielded two *Salmonella* PFGE patterns from the four isolation schemes, which represented *Salmonella* Newport and Thompson (FSL-S10-1570 and 1574). One case of repeat isolation, *Salmonella* Cerro from a surface water sample site, was identified (FSL S10-550, 1253, and 1411; depicted with boxes and bold text). Band sizes (kb) are displayed at the top of the PFGE pattern images. PFGE pattern order displayed is result of BioNumerics similarity analyses using the unweighted pair group-matching algorithm (UPGMA) and the Dice correlation coefficient with a maximum space tolerance of 1.5%.

FIG. 3. CT dividing *L. monocytogenes* environmental samples based on remotely-sensed topographical, edaphic, and meteorological data. On the top of each node there is a rule used for partitioning samples into homogenous subsets. Primary rules are those used to make the depicted split. Competitor rules represent the rule with the second best improvement score. Surrogate rules mimic the primary rule and produce a split with a similar division of cultured positive and negative samples in daughter nodes. Percent cultured positive samples are displayed in each node. Rules partition toward the left-hand daughter node. Left-hand daughter nodes are enriched for negative samples, and right-hand daughter nodes are enriched for positive samples.
Abbreviations used: L = number of samples partitioned in the left daughter node, R = number of samples partitioned in the right daughter node, N = number of cultured negative samples, and P = number of cultured positive samples. Text S3 provides a full summary of the CT. Briefly, Eigenvector_1_Temporal was an axis negatively correlated with temperature data and splits based on this rule indicated that cooler temperatures are more likely to predict a pathogen positive sample. Eigenvector_1_Spatial was an axis that was positively correlated with available water storage and organic matter values for soils. Locations with higher values on this eigenvector were likely to be pathogen positive.

FIG. 4. CT dividing Salmonella environmental samples based on remotely-sensed topographical and meteorological data. On the top of each node there is a rule used for partitioning samples into homogenous subsets. Primary rules are those used to make the depicted split. Competitor rules represent the rule with the second best improvement score. Surrogate rules mimic the primary rule and produce a split with a similar division of cultured positive and negative samples in daughter nodes. Percent cultured positive samples are displayed in each node. Rules partition to the left-hand daughter node. Left-hand daughter nodes are enriched for negative samples, and right-hand daughter nodes are enriched for positive samples. Abbreviations used: L = number of samples partitioned in the left daughter node, R = number of samples partitioned in the right daughter node, N = number of cultured negative samples, and P = number of cultured positive samples. Text S4 provides a full summary of the CT. Briefly, Eigenvector_1_Spatial was positively correlated with soil moisture, and the primary split based on this variable indicated that soils below the maximum soil moisture in our data base were likely to yield a positive sample. Eigenvector_2_Temporal was positively correlated with precipitation data and indicated that when measurable precipitation occurred within 3 days prior to sampling a positive sample.
was more likely. Farm was determined to be an important variable, though this split appeared lower in the tree than soil characteristics and precipitation, indicated that the latter variables have larger scale effects on the detection of *Salmonella*.

FIG. 5. Map predicting environmental locations and spatial extents of areas with increased *L. monocytogenes* prevalence based on CT results (Fig. 3). Percentages in legend indicate the prevalence of *L. monocytogenes* from the CT analysis. Cropland polygons were isolated and were assigned values as reservoirs based on three levels of the CT: i) areas within 38 m of mapped surface water (light blue, 45%), ii) areas outside class (i) with soil available water storage (AWS) > 4.0 cm in the top 25 cm and within 62 m of mapped pasture land cover (dark blue, 50%), and iii) areas with AWS > 4.0 cm in the top 25 cm but outside classes (i) and (ii) with 7.5% prevalence (cyan). The color surface represents proximity to nearest pasture class land cover. Brown cropland polygons are not expected to contain significant *L. monocytogenes* reservoirs.
TABLE 1 General Farm Characteristics and Key Management

<table>
<thead>
<tr>
<th>Farm</th>
<th>Size (acres)</th>
<th>Organic</th>
<th>Irrigate$^b$</th>
<th>Manure$^c$</th>
<th>Compost or composted manure$^d$</th>
<th>Staff (no. of employees)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Year-round</td>
</tr>
<tr>
<td>1</td>
<td>&gt; 1000</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes (31-40)</td>
</tr>
<tr>
<td>2</td>
<td>&lt; 250</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>&gt; 1000</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes (6-10)</td>
</tr>
<tr>
<td>4</td>
<td>&lt; 250</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>&gt; 1000</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes (11-15)</td>
</tr>
</tbody>
</table>

$^a$ All farms answered yes to wildlife control measures (e.g., hunting or fences), worker training (e.g., sessions or videos on good hygiene and sanitation practices) and having Good Agricultural Practices (GAP) plans (i.e., third party audits of food safety practices).

$^b$ Farm 2 used a combination of drip and overhead irrigation depending on crop, farm 3 used overhead irrigation, and Farm 4 uses drip irrigation.

$^c$ Manure slurry has been applied to a field within the past year. Produce was not planted before 120 days.

$^d$ Compost and composted manure is a treated product.
### TABLE 2 Effect of Factors (Farm, Season and Sample Type) on frequency of positive *L. monocytogenes*, *Salmonella*, and STEC samples found in produce preharvest environments.

<table>
<thead>
<tr>
<th>Factor (No. of samples)</th>
<th>Frequency (Percent)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><em>L. monocytogenes</em></td>
</tr>
<tr>
<td>Farm</td>
<td></td>
</tr>
<tr>
<td>1 (166)</td>
<td>39 (23)(^A)</td>
</tr>
<tr>
<td>2 (103)</td>
<td>5 (5)(^B)</td>
</tr>
<tr>
<td>3 (113)</td>
<td>13 (12)(^AB)</td>
</tr>
<tr>
<td>4 (100)</td>
<td>14 (14)(^AB)</td>
</tr>
<tr>
<td>5 (106)</td>
<td>17 (16)(^AB)</td>
</tr>
<tr>
<td>Season</td>
<td></td>
</tr>
<tr>
<td>Fall (136)</td>
<td>9 (7)(^B)</td>
</tr>
<tr>
<td>Winter (125)</td>
<td>30 (24)(^A)</td>
</tr>
<tr>
<td>Spring (134)</td>
<td>23 (17)(^AB)</td>
</tr>
<tr>
<td>Summer (193)</td>
<td>26 (19)(^A)</td>
</tr>
<tr>
<td>Sample Type</td>
<td></td>
</tr>
<tr>
<td>Soil (178)</td>
<td>16 (9)(^B)</td>
</tr>
<tr>
<td>Drag Swab (175)</td>
<td>15 (9)(^B)</td>
</tr>
<tr>
<td>Fecal (61)</td>
<td>9 (15)(^AB)</td>
</tr>
<tr>
<td>Water (174)</td>
<td>48 (28)(^A)</td>
</tr>
<tr>
<td>Engineered (28)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Surface (146)</td>
<td>48 (33)</td>
</tr>
</tbody>
</table>
Superscript letters represent different statistical populations of values that are significantly different with $P$-value < 0.016. No letters represent values that are not significantly different.
**Primary Rule: Drainage Class**
- Competing Rule: Urban Proximity > 14 m
- Surrogate Rule: Water Table Depth > 35.5 cm
  - N=320 & P=4
  - 1.2%

**Moderately and well drained**

**Very poorly, poorly, somewhat poorly and somewhat well drained**

**9%**

**Primary Rule: Eigenvector_1_Spatial < -3.642**
- Competing Rule: Forest Proximity > 321 m
- Surrogate Rule: Urban Proximity >263.5 m
  - N=89 & P=2
  - 2.2%

**9%**

**Primary Rule: Eigenvector_2_Temporal < -1.689**
- Competing Rule: Urban Proximity > 18.5 m
- Surrogate Rule: Water Proximity >
  - L=34 & R=139
  - N=34 & P=0
  - 0%

**12%**

**Primary Rule: Farm**
- Competing Rule: Urban Proximity > 156 m
- Surrogate Rule: Urban Proximity > 213 m
  - 4 and 5
  - N=14 & P=0
  - 0%
  - 1, 2 and 3
  - N=104 & P=21
  - 16.8%