AMERICAN SOCIETY FOR MICROBIOLOGY | Applied and Environmental Microbiology

Check for updates

Food Microbiology | Full-Length Text

Methodological differences between studies confound one-size-fits-all approaches to managing surface waterways for food and water safety

Daniel L. Weller,[1,2] Claire M. Murphy,[2] Tanzy M. T. Love,[1] Michelle D. Danyluk,[3] Laura K. Strawn[2]

AUTHOR AFFILIATIONS See affiliation list on p. 23.

ABSTRACT Even though differences in methodology (e.g., sample volume and detection method) have been shown to affect observed microbial water quality, multiple sampling and laboratory protocols continue to be used for water quality monitoring. Research is needed to determine how these differences impact the comparability of findings to generate best management practices and the ability to perform meta-analyses. This study addresses this knowledge gap by compiling and analyzing a data set representing 2,429,990 unique data points on at least one microbial water quality target (e.g., *Salmonella* presence and *Escherichia coli* concentration). Variance partitioning analysis was used to quantify the variance in likelihood of detecting each pathogenic target that was uniquely and jointly attributable to non-methodological versus methodological factors. The strength of the association between microbial water quality and select methodological and non-methodological factors was quantified using conditional forest and regression analysis. Fecal indicator bacteria concentrations were more strongly associated with non-methodological factors than methodological factors based on conditional forest analysis. Variance partitioning analysis could not disentangle non-methodological and methodological signals for pathogenic *Escherichia coli*, *Salmonella*, and *Listeria*. This suggests our current perceptions of foodborne pathogen ecology in water systems are confounded by methodological differences between studies. For example, 31% of total variance in likelihood of *Salmonella* detection was explained by methodological and/or non-methodological factors, 18% was jointly attributable to both methodological and non-methodological factors. Only 13% of total variance was uniquely attributable to non-methodological factors for *Salmonella*, highlighting the need for standardization of methods for microbiological water quality testing for comparison across studies.

IMPORTANCE The microbial ecology of water is already complex, without the added complications of methodological differences between studies. This study highlights the difficulty in comparing water quality data from projects that used different sampling or laboratory methods. These findings have direct implications for end users as there is no clear way to generalize findings in order to characterize broad-scale ecological phenomenon and develop science-based guidance. To best support development of risk assessments and guidance for monitoring and managing waters, data collection and methods need to be standardized across studies. A minimum set of data attributes that all studies should collect and report in a standardized way is needed. Given the diversity of methods used within applied and environmental microbiology, similar studies are needed for other microbiology subfields to ensure that guidance and policy are based on a robust interpretation of the literature.

Editor Edward G. Dudley, The Pennsylvania State University, University Park, Pennsylvania, USA

Address correspondence to Laura K. Strawn, Lstrawn@vt.edu.

The authors declare no conflict of interest.

See the funding table on p. 23.

Received 15 October 2023
Accepted 14 November 2023
Published 12 January 2024

[This article was published on 12 January 2024 with an inaccuracy in Fig. 1. The figure was updated in the current version, posted on 18 January 2024.]

KEYWORDS produce safety, methods comparison, water quality, *Salmonella*, *Listeria*, shiga toxin *Escherichia coli*

Food and waterborne disease outbreaks have been linked to the use of contaminated water for food production or recreation (1–6). There is interest in developing models to predict when, where, and how much water is likely to become contaminated with enteric bacteria and foodborne pathogens. These models may support risk assessments and the development of guidance for monitoring and managing hazards in waters. Model development relies on having sufficiently large quantities of data for model training and testing. A multitude of studies have surveyed waters for foodborne pathogens, pathogen surrogates, and fecal indicator bacteria (7–18). The sampling and laboratory methods used by published studies are diverse and often vary by laboratory and available resources. For example, some studies collected water samples using Moore swabs, which capture microbial water quality flowing through a waterway during a given time frame (7, 16, 19), while other studies collected water using grab samples, which provides a snapshot of water quality at the moment of sample collection (16, 20). For grab samples, volumes vary substantially between studies with ranges between 3.33 mL (20) and 10 L (16, 21). Furthermore, both filtration methods (11, 16, 20, 22–24) and foodborne pathogen detection [culture-based (12, 25, 26) versus molecular-based methods (13, 21)] have varied across studies.

Previous studies have reported that methodological differences affect observed microbial water quality (15, 16, 27–32). The odds of pathogenic *Escherichia coli* and *Salmonella* detection were lower for 10-L grab samples filtered through modified Moore swabs, compared to 24-hour Moore swabs collected from the same waterways at the same time (16). A Mid-Atlantic study reported *Salmonella* detection was 26 and 44 times more likely when 10-L samples were collected compared to 1.0- and 0.1-L samples, respectively (32). Findings from these and other studies demonstrate that results may not be comparable when different sampling and laboratory methods are used. However, these past studies were limited geographically (e.g., focused on one or two regions and sampling a small number of waterways), temporally (e.g., conducted over a single growing season), and/or in sample size (e.g., small number of samples were collected). To address these limitations, the present study assessed how methodological differences impact the comparability of study findings using data from multiple studies to ensure sufficient sample size, geographical diversity, and temporal coverage.

Prior studies have also investigated and compared microbial water quality between water type and region (12, 16, 33–36). A 2020 review of agricultural water in the Southeastern United States noted that the geographical location of a water source played an important role in the prevalence and survival of foodborne pathogens (35). This conclusion is supported by other studies that compared microbial water quality between growing regions nationally (16, 37) and locally (12, 17, 38), and between water types (32, 39–41). Indeed, multiple studies have shown that this variability in water quality limits the efficacy of one-size-fits-all approaches to monitoring and mitigating microbial hazards in aquatic environments. Understanding how microbial water quality differs between water types (e.g., canal, pond, and stream) and regions is needed for the development of water type and regional guidance for managing and mitigating these hazards. Additionally, identification of strong regional and methodological signals by some studies (15, 16, 32) further highlights the challenge of comparing findings between studies conducted in different regions and on different water types when those studies also used different methods. Data are needed to determine how and if (i) such findings can be compared and (ii) methodological signals can be separated from signals of interest (e.g., region and water type).

To address these knowledge gaps, we compiled a large data set representing a diversity of regions, water types, and sampling and laboratory methods. We used these data to (i) quantify the impact of methodological differences on observed microbial water quality; (ii) determine if methodological signals can be disentangled from other

signals of interest (e.g., region and water type); and (iii) evaluate how using specific methodological differences (e.g., collecting 1 L versus 10 L of water) impacts observed microbial variability in water. To determine if strategies for managing microbial hazards in surface water should be region, water type, and/or waterway specific, we examined how strongly (i) water type and waterway-specific factors and (ii) regional classification scheme (i.e., how samples were assigned to a given regions, such as by ecoregion or climate region) were associated with microbial water quality after accounting for methodological differences.

RESULTS

Data quality and compatibility

Of the 3,211,254 data points collected from peer-reviewed papers, publicly available databases, citizen science groups, and government organizations, 2,429,990 (77%) were retained in the final data set (Fig. 1), and are also provided in an online database (https://github.com/wellerd2/Weller-et-al-2024-AEM-Datasets/tree/main). The retained datapoints represent 100,410 unique sampling sites and included data points from all US states, as well as multiple US territories, Canadian provinces, and Mexican states (Table 1; Fig. 2; Fig. S1). Additionally, data encompasses numerous major agricultural regions (e.g., the Central Valley, the Columbia Basin, and the Mid-Atlantic). Table 2 was developed to support future field studies on microbial water quality, highlighting recommended practices for data collection, recording, reporting, and future analyses. During data compilation, studies with insufficient methodological data were dropped, such as samples where volume for enumeration was not available. The vast majority of the 781,264 datapoints excluded were because methods for generic *E. coli*, total coliforms, fecal coliforms, or *Enterococcus* did not provide any sufficient information on sampling and/or enumeration methods (Fig. 1). One hundred and eight datapoints of foodborne pathogen data were excluded because methodological data were not available. Since complete GPS coordinates were needed to extract key non-methodological data (e.g., water type) assign unique site and waterway IDs, and apply the regional classification schemes, 13,010 datapoints were excluded where complete GPS coordinates were not reported or had missing information. Other datapoints that were excluded include 1,791 duplicates and 1,722 datapoints where waterway, water type, and water source could not be determined. Errors in sampling date resulted in 1,265 datapoints being excluded; most of these were excluded because the reported sample collection was listed as a year in the future or unrealistically far in the past. The vast majority of excluded datapoints were collected by local, state, or government agencies and were downloaded from government portals. Almost all data provided directly by research or citizen science groups or obtained from peer-reviewed papers were retained (Fig. 1).

Data compilation was complicated by challenges associated with translating data sheets into English. Similarly, the failure of multiple citizen science and government databases to provide method information or list methods online complicated data compilation. These challenges were substantially reduced by the willingness of the researchers to answer questions and share comprehensive data dictionaries as well as reports/certifications found online. Inconsistency in reporting methods information was also discovered. Many citizen science and government data sets used a standard or government method. While peer-reviewed papers often described their methods, these descriptions were often brief and seldom referred to the established methods by name. We also found that many enumeration protocols had multiple names (e.g., Standard Method 9223, Colilert, and Colisure refer to the same protocol). The use of internal laboratory nomenclature or abbreviations required extensive effort to link back to the published protocol. For instance, data available through the US National Water Quality Portal used >15 different United States Geological Survey (USGS)-specific terms to refer to the same protocol. Additional common errors associated with data included latitude as longitude and vice versa, not including the negative sign in the longitude, and reporting unrealistic values for attributes (e.g., pH >14).

TABLE 1 Summary of the data sets compiled as part of the study presented here[h]

Data source[a]	Organization type[b]	Years[c]	State[d]	No. of samples tested for[e]							Data sets[f]	Citation[g]
				Sites	Total samples	Fecal indicator bacteria				Pathogens		
						Coliforms		*E. coli*	*Enterococcus*			
						Fecal	Total					
ACAP, St. John	N/C	1995, 2019	NB	96	1,473	1,473	0	0	0	0	ACAP	(42)
Adhikari Lab, LA St U	U	2017–2018	LA	1	29	0	0	29	0	0	ADH	(43)
Bhullar Lab, KS St U	U	2017–2020	IA, KS, MO	511	511	0	0	511	0	0	BHU, KSM	(44, 45)
Black Warrior Riverkeeper	N/C	2019–2020	AL	9	83	0	1	83	0	0	BWR	(46)
CA Env Data Exchange Network	G	2009, 2019	CA	25	456	230	46	44	220	0	CEDEN	(47)
Canizalez-Roman Lab, U Autonoma de Sinaloa	U	2015	SI	173	405	118	393	390	0	405	ANF	(20)
Cary Inst for Ecosystem Studies	U	2001, 2017	MD	12	1,702	0	0	1,702	0	0	BES	(48)
Characklis Lab, U of NC at Chapel Hill	U	2004, 2008	NC	11	256	209	0	250	152	85	UNC, UNCS	(49–51)
Chesapeake Bay Foundation	N/C	2015–2016	MD, PA, and VA	64	591	50	0	193	398	0	CBF	(52)
Chesapeake Monitoring Cooperative	N/C	2011–2020	DC, DE, MD, and VA	415	8,383	91	0	7,889	1,463	0	CMC	(53)
City of Austin, TX	G	1997, 2021	TX	311	6,059	0	0	6,059	0	0	AUST	(54)
City of Chicago, IL	G	2006–2016	IL	25	17,598	0	0	2,634	14,964	0	BLD	(55)
Colwell Lab, U of MD	U	2019	DC	4	8	0	0	8	8	8	METG	(56)
Community Sci Inst	N/C	2002–2021	NY	273	6,511	0	4,541	6,454	0	0	CSI	(41)
Danyluk Lab, U of FL	U	2010–2016	FL	39	910	0	910	910	540	910	NUF, RUF, ZUF	(11–13)
Dept of the Env, Prince George's Co, MD	G	2008–2020	MD	2	402	34	0	368	0	0	BEAR	-
Food Safety Lab, Cornell U	U	2001, 2018	AZ, CA, and NY	587	1,745	0	800	800	0	1,730	FSL, PAWQ	(9, 15, 16, 33, 34, 36, 57–60)
Four Rivers Watershed Watch	N/C	2001–2018	KY and TN	279	2,769	543	0	2,226	0	0	FRSS	(61)
GA Env Monitoring and Assessment System	G	1999–2021	GA and SC	737	21,200	19,599	0	6,472	788	0	GOMAS	(62)
Green Lab, SUNY College of Env Sci and Forestry	U	2015, 2020	NY and UT	99	693	0	659	654	29	0	ESF, SLC	(63)
Hansen Lab, Dalhousie U	U	2008–2009	NS	12	333	0	333	333	0	333	LTH	(23)
Harwood Lab, U of South FL	U	2009–2012	FL	12	84	84	0	0	84	0	ZRS	(64)
Hornor Lab, Anne Arundel Co Community College	U	2001, 2019	MD	8	152	0	0	0	152	0	OPS	-

(*Continued on next page*)

TABLE 1 Summary of the data sets compiled as part of the study presented here[h] (*Continued*)

Data source[a]	Organization type[b]	Years[c]	State[d]	No. of samples tested for[e]							Data sets[f]	Citation[g]
				Sites	Total samples	Fecal indicator bacteria						
						Coliforms		*E. coli*	*Enterococcus*	Pathogens		
						Fecal	Total					
IA Dept of Natural Resources	G	1991–2020	IA	60	14,613	3,791	0	13,919	3,096	0	IDNR	(65)
Levy Lab, Emory U	U	2013	GA	16	107	0	0	105	0	107	CSH	(22)
LA Dept of Env Quality	G	1978, 2020	LA	9	1,205	1,205	0	0	0	0	LADEQ	(66)
Lower Colorado River Authority	G	1982–2019	TX	291	16,645	8,815	0	9,669	305	0	TXS	(67)
McLellan Lab, U of WI	U	2011–2013	MI, OH, and WI	30	539	319	0	536	537	0	GRTLK, MILK	(68, 69)
Milwaukee Riverkeeper	N/C	2014, 2019	WI	195	5,131	4,562	0	3,266	0	0	MRK	(70)
Mountain True	N/C	2018–2020	NC and TN	75	1,100	0	0	1,100	0	0	BRW, FBR	(71)
Nashwaak Watershed Watch	N/C	1996, 2019	NB	27	334	0	0	334	0	0	NASH	(72)
National Water Quality Portal	G	1919, 2022	[g]	79,121	2,141,249	961,883	155,571	1,257,650	99,214	1,979	[i]	(73)
OK Dept of Ag	G	2001–2014	MO and OK	275	4,599	3,459	0	3,832	3,867	0	OKW	(74)
OK Water Survey, U Of OK	U	2018	OK	24	306	0	0	296	273	0	OWS	(75)
Onondaga Env Inst	N/C	2008, 2017	NY	180	2,658	2,658	96	281	288	0	OEI	(17)
Pearl Riverkeeper	N/C	2018–2020	LA and MS	31	428	428	419	426	0	0	PRWS	(76)
Pickering Lab, Tufts U	U	2017	MA	2	4	0	0	0	0	4	PICK	(77)
Public Health Lab, Humboldt Co, CA/U of South FL	G, U	2020	CA	13	201	0	201	201	201	0	STRAWB	(78)
Richardson Lab, Cornell U	U	2017–2018	NJ and NY	36	179	0	0	163	174	65	RICH	(21, 79)
Rideout Lab, VA Tech	U	2013–2015	VA	4	434	0	191	191	0	431	GUF	(80)
Rock Lab, U of AZ	U	2012	AZ and CA	204	446	0	253	446	0	0	CPS	(81)
Saint Croix International Waterway Commission	N/C	1998, 2019	NB	96	245	0	0	245	0	0	SCIWC	(82)
Shariat Lab, U of GA	U	2018–2019	PA	28	112	0	111	111	0	112	SRB	(25)
Shrestha Lab, U of IL at Chicago	U	2016	IL	7	195	0	0	170	195	0	SHR	(83)
Smith Mountain Lake MP, Ferrum College	U	1995–2020	VA	26	4,068	0	4,066	2,681	0	0	FERR	(84, 85)
SC Adopt-a-Stream, Clemson U	U, N/C	2016–2020	SC	201	1,435	0	0	1,435	0	0	SCAAS	(86)
Spa Creek Conservancy	N/C	2016–2020	MD	16	672	0	0	0	672	0	SCC	(87)
Strawn Lab, VA Tech	U	2015, 2021	VA	63	1,120	0	1,000	1,000	0	520	DUCK, LW, TUF	(12, 14, 38, 88)

(*Continued on next page*)

TABLE 1 Summary of the data sets compiled as part of the study presented here[h] (*Continued*)

Data source[a]	Organization type[b]	Years[c]	State[d]	No. of samples tested for[e]							Data sets[f]	Citation[g]
				Sites	Total samples	Fecal indicator bacteria				Pathogens		
						Coliforms		*E. coli*	*Enterococcus*			
						Fecal	Total					
Surface Water Ambient MP, CA Water Boards	G	2018–2021	CA	26	1,992	0	0	1,992	0	0	SWA, SWAMP	(89)
Thornton Creek Alliance	N/C	2017–2020	WA	24	2,005	0	0	2,005	0	0	THC	(90)
US Dept of Ag	G	2011–2016	CA and IA	22	3,164	22	3,164	58	22	3,164	CEG, CLY, NASA	(7, 19, 91)
US Env Prot Agency	G	2012–2017	[i]	5,396	8,200	0	1,107	2,189	7,053	126	BRAD, CLCR, NARS, SFBR	(31, 92–96)
US Geological Survey	G	2001–2019	[j]	315	8,387	226	1,273	7,191	764	809	[k]	(24, 97–123)
Waccamaw Watershed Academy, Coastal Carolina U	U	2008–2020	NC and SC	28	7,364	1,322	6,240	6,242	0	0	CCU	(124)
Walters Lab, Stanford U	U	2008–2009	CA	14	241	0	0	241	227	241	SPW	(26)
Wang Lab, U of BC	U	2015–2016	BC	7	446	411	0	397	0	446	UBC	(125)
WA Dept of Ag	G	2014–2021	WA	823	41,184	38,985	0	2,201	0	0	WAWA	(126)
Watershed Watch, U of RI	U	1991, 2020	CT, MA, NY, and RI	560	19,508	10,335	0	2,866	15,892	0	URI	(127, 128)
WV Dept of Env Prot	G	1970–2020	WV	8,458	67,038	66,898	6,928	699	0	0	WVD	(129)
Western Center for Food Safety, U of CA at Davis	U	2012–2013	GA	2	83	0	0	83	0	0	EMA	(130)

[a]ACAP, Atlantic Coastal Action Program; Ag, Agriculture; Co, County; Dept, Department; Env, Environment/Environmental; Inst, Institute; MP, Monitoring Program; Prot, Protection; Sci, Science; St, State; U, University.
[b]Organization type: G, governmental organization; N/C, non-profit or citizen science organization; U, university.
[c]The first and last years that data from the given organization were available. Some organizations collected data every year during this period (indicated by the use of "–"), while other organizations did not continuously collect data during this period (indicated by the use of ",").
[d]US Post Office abbreviations are used for Canadian provinces, Mexican states, and US states and territories.
[e]The number of samples with data on fecal indicator bacteria concentrations and the presence/absence of any of the foodborne pathogens (e.g., *Listeria monocytogenes* and *Salmonella*) or indicator organisms (e.g., *Listeria* spp.) for foodborne pathogens.
[f]Some groups or databases provided data from multiple data sets. These codes correspond to each data set provided by each group or database except for the National Water Quality Portal (NWQP); these codes are used to link information in Supplemental Table 2. Because of the number of data sets available in and downloaded from the NWQP (*N*=534), all NWQP data sets could not be listed here.
[g]Data and attributes were downloaded from publicly available portals and peer-reviewed papers, or obtained through personal communication with project leads or points of contacts in the relevant organization, lab, or agency. Groups with no web presence or corresponding report/publication that were contacted by email are indicated by "–."
[h]Data were available for all 50 US states as well as American Samoa, British Columbia, the District of Columbia, Guam, Puerto Rico, Saskatchewan, and the US Virgin Islands.
[i]Data were available for the 48 continental United States but not for AK or HI.
[j]Data were available for DE, GA, IA, IL, IN, KS, KY, MI, MO, NC, NJ, OH, OR, PA, RI, SC, SD, TN, TX, VA, and WI.
[k]The datasets available from the US Geological Survey were assigned the following codes: AMRF; BCB; CHAT; CNP; DEGR; DSF; ERIE; FUR; INDPC; LAMI; LCM; LSC; MATH; MIR; MMF; RACI; SALAD; SCIT; SHV; ULRB; UPDU; USGS; WHCH; WHSKR.

TABLE 2 Recommended practices for the collection, recording, and reporting of data and attributes for research collecting water samples aimed at understanding microbial water quality

Category	Best practices for collection, recording, and reporting	Resources and references
Tidy data[a]		(131, 132)
Data structure	• Each column represents a unique variable with a unique column header that avoids spaces, capital letters, or special characters (e.g., water_type or watertype).	
	• Each row represents a unique observation with a unique row ID.	
	• Each cell is a single measurement without units within the cell (include units in data dictionary and, if needed, create a new column for units).	
	• Avoid visual formatting (e.g., colored cells and borders).	
Value formatting	• For categorical variables, use drop-down menus when possible so you are only selecting from pre-standardized categories. If drop-down menus are not possible, ensure consistency in value formatting and the case of the text (e.g., avoid using "Pond", "pond", and "P" to all refer to pond samples). Ensure consistency in spelling and using of white space/special characters (e.g., avoid "pond").	
	• For numeric, set upper and lower bounds so you can catch entry or measurement errors.	
Missing values	• Encode missing values as NA or as a blank. This will ensure it is read as missing data by analysis programs. Do not encode missing value as a number (e.g., 0 or 999), character (e.g., –), or word (e.g., missing).	
Data dictionary[b]	• Include the column header exactly as seen in the data set.	(133–135)
	• Define each variable, including all possible entries/factor levels and their meaning, range of possible numeric values, or accepted values for the variable (e.g., ≥0 and ≤100), and units (when applicable). If a numeric variable has upper and lower limits of detection, report these and how they should be dealt with for analysis.	
	• If any imputation or data transformations were or should be employed prior to use, describe these.	
	• If the variable can be used alone or in conjunction with other columns to calculate new columns, explain this as well.	
Recommended minimum collected and recorded attributes		
Methodology		(136–140)
Sample type	• Examples include grab sample and Moore swab.	
Sample volume	• Water volume tested for the given target, not the volume of the sample; these may be the same but often the sample is divided into aliquots used to test for different targets. Ensure all data are entered in the same units [e.g., most probable number (MPN)/100 mL, CFU/mL)	
Sample site	• Ideally, this would be at least three columns: the sampling site latitude, the sampling site longitude, and a descriptive column (e.g., near uptake pump or 1 m from into pond from pump). Additional columns that are recommended are sample depth and distance from shore.	
Unique sampler ID	• The person who collected the sample; if multiple individuals have the same initials, do not use initials for this column. Use an unambiguous method for this ID.	
Detection or quantification method	• If only one method is used throughout the study, clearly state the method used for target detection or enumeration in the data dictionary. The description should provide sufficient detail [e.g., was it a culture-based or molecular method? what was the detection limit(s)? if and how the sample was filtered (e.g., membrane, none, or modified Moore swab), if and how the target was enumerated (e.g., most probable number-based approach or membrane filter-based approach), volume used for reporting results (e.g., count per 100 mL or count per 10 mL)].	
	• If there is a standard name for the method [e.g., Environmental Protection Agency (EPA) Method 1603], include that. Do not use organization or lab-specific names in this column, as this information will not mean anything to folks outside the organization or lab.	
	• Be clear what the target is (e.g., is it the microbe, a specific gene, or multiple genes?).	
	• If the method has previously been validated/published in the peer-reviewed research literature, include this reference. Even if it was published, include key performance information for the method (e.g., sensitivity and specificity).	
	• If positive and negative controls were used, include what those controls were.	

(*Continued on next page*)

TABLE 2 Recommended practices for the collection, recording, and reporting of data and attributes for research collecting water samples aimed at understanding microbial water quality (*Continued*)

Category	Best practices for collection, recording, and reporting	Resources and references
Spatial		
GPS coordinates	• Use a consistent coordinate reference system (i.e., DATUM) and note this in the data dictionary. • Use a standard format; formats that include spaces, multiple symbols, or multiple "." can result in coordinates reading incorrectly. Confirm the appropriate presence/absence of "–" if that is part of the format you use. DO NOT drop the "–" just because all of your sites are in the same hemisphere. • Include longitude and latitude as separate columns	(133, 140–143)
Hydrologic units (HUC)	• Six or nine-digit code	
Type of waterway	• Examples include pond, stream, or canal.	
Location	• This could include separate columns for county, state, county, and/or city.	
Unique site ID	• Make it unique to each site and unambiguous. Use a standard way of naming that is relevant to your study design.	
Waterway name	• Common name(s) used to refer to the waterway. If multiple names, separate by ";" or ",".	
Temporal		
Date	• MM/DD/YYYY	(138, 140, 144)
Time of day	• Use military time (24 hours) to reduce the risk of incorrectly specifying a.m./p.m.	
Physiochemical		
Turbidity Water temperature pH Total suspended solids Dissolved oxygen Dissolved organic matter Conductivity Salinity	For categorical variables, use a drop-down menu to avoid spelling or entry errors. For numeric values, use a consistent number of significant digits reflects the accuracy of measuring the device and sets minimum and maximum thresholds to catch entry/measurement errors. Use consistent units (e.g., metric or imperial) and include units in the data dictionary.	(138, 145–147)
Meteorological		
Air temperature Precipitation volume Relative humidity UV intensity Wind speed	For categorical variables, use a drop-down menu to avoid spelling or entry errors. For numeric values, using a consistent number of significant digits reflects the accuracy of measuring the device and sets minimum and maximum thresholds to catch entry/measurement errors. Use consistent units (e.g., metric or imperial) and include units in the data dictionary.	(138, 147, 148)

[a]Tidy data provide a standard way to organize data values within a data set that has been cleaned in a way that is ready for analysis.
[b]A data dictionary is a centralized repository of attributes that provides a comprehensive description of the data used. Its main purpose is to provide additional context and information about each data point so that analysts can understand the data better.

Inconsistency in reporting methods and waterway information between and within organizations meant that waterway and water source characteristics had to be verified and corrected. For example, a single data set used >100 different names to refer to a single waterway. Waterway names are not unique to individual waterways, and multiple distinct streams, ponds, or rivers, even in a small geographical area, were identified with the same name. Waterway name categories that were commonly shared between and within geographical locations included color-based (e.g., silver lake), animal-based (e.g., deer creek), person-based (e.g., miller creek), and place-based (e.g., schoolhouse creek) names. For example, it is important to ensure that silver lake in State A will not be coded as the same waterway as silver lake in State B (color-based name). Here, any issues with sampling site ID resulted in its replacement with a geo-ID to avoid confusion.

Non-methodological and methodological signals could not be disentangled for foodborne pathogens and indicator organisms

Salmonella

Salmonella data (*N* = 9,348) were obtained from 35 studies representing 15 US states and the District of Colombia, two Canadian provinces, and one Mexican state (Table 1; Table S2). While samples were collected between 1973 and 2019, 75% (*N* = 7,043) were collected since 2012. Grab samples and Moore swabs represented 67% (*N* = 6,245) and 33% (*N* = 3,103) of datapoints, respectively. Of the grab samples, 2,282 (37%) were filtered using membrane filters; 748 (12%) were filtered through modified Moore swabs; 202 (3%) underwent tangential flow filtration; 1,944 (31%) were not filtered; and 1,069 (17%) samples did not describe a filtration method. Culture-based methods were used to detect *Salmonella* in 7,225 samples (77%), while molecular methods were used for 2,123 samples (23%). Seventy percent of studies (*N* = 5,075) that used a culture-based detection method used *invA* to confirm isolates as *Salmonella*.

Variance partitioning analysis demonstrated that 18% of variance in *Salmonella* detection was jointly attributable to methodological and non-methodological factors, while <1% and 13% of variance was uniquely attributable to methodological and non-methodological factors, respectively (Fig. 3A; Table S3). When variance attributable to waterway/sampling site, year/season, region, and methods was considered, the only source of unique variance greater than 0.01% was waterway/sampling site (4%). When the analysis was done to include waterway/sampling site with year/season, water type, and methodological differences, 10% of variance was jointly attributable to all four sources, while only 6% and 1% were uniquely attributable to waterway/sampling site and methodological differences, respectively; no variance was uniquely attributable to year/season and water type (Table S3).

After accounting for waterway and site-specific signals, the top-ranked factors associated with the likelihood of *Salmonella* detection by conditional forest analysis were sample filtration method, season, and sample volume (Fig. 4A; Table S4). Based on generalized linear models and post hoc testing, the odds of *Salmonella* detection differed significantly between all filtration methods (Table S5). When grab samples were filtered through membrane filters as opposed to modified Moore swabs, there was a fivefold [odds ratio (OR) = 4.70, standard errer (SE) = 1.26; $P < 0.001$] higher odds of *Salmonella* detection (Table S5). Increasing sample volume also significantly increased the odds of detecting *Salmonella*. As volume increased from 5 mL (OR = 0.13, 95% CI = 0.07–0.23) to 10 L (OR = 0.42, 95% CI = 0.27–0.59), *Salmonella* detection increased dramatically (Fig. 5A). Similar differences were observed when different sample types and detection methods were used (Table S5). After accounting for methodological, waterway, and site-specific signals, the top-ranked factors associated with the likelihood of *Salmonella* detection were ecoregion, state, and water type (Table S6).

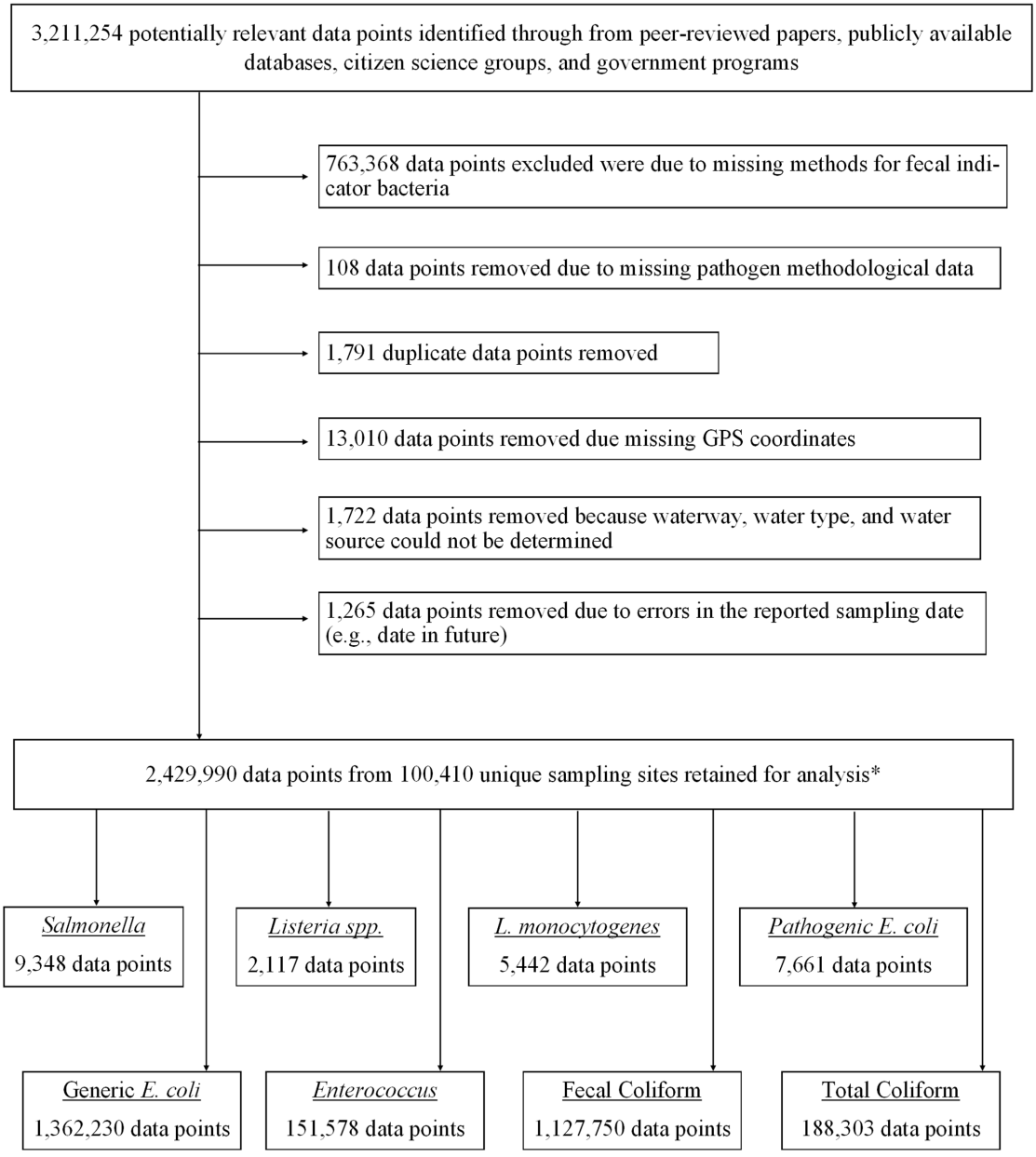

FIG 1 Schematic representation showing data exclusion due to data quality and compatibility issues. *Numerous samples were tested for more than one microbial target. GPS, Global Positioning System.

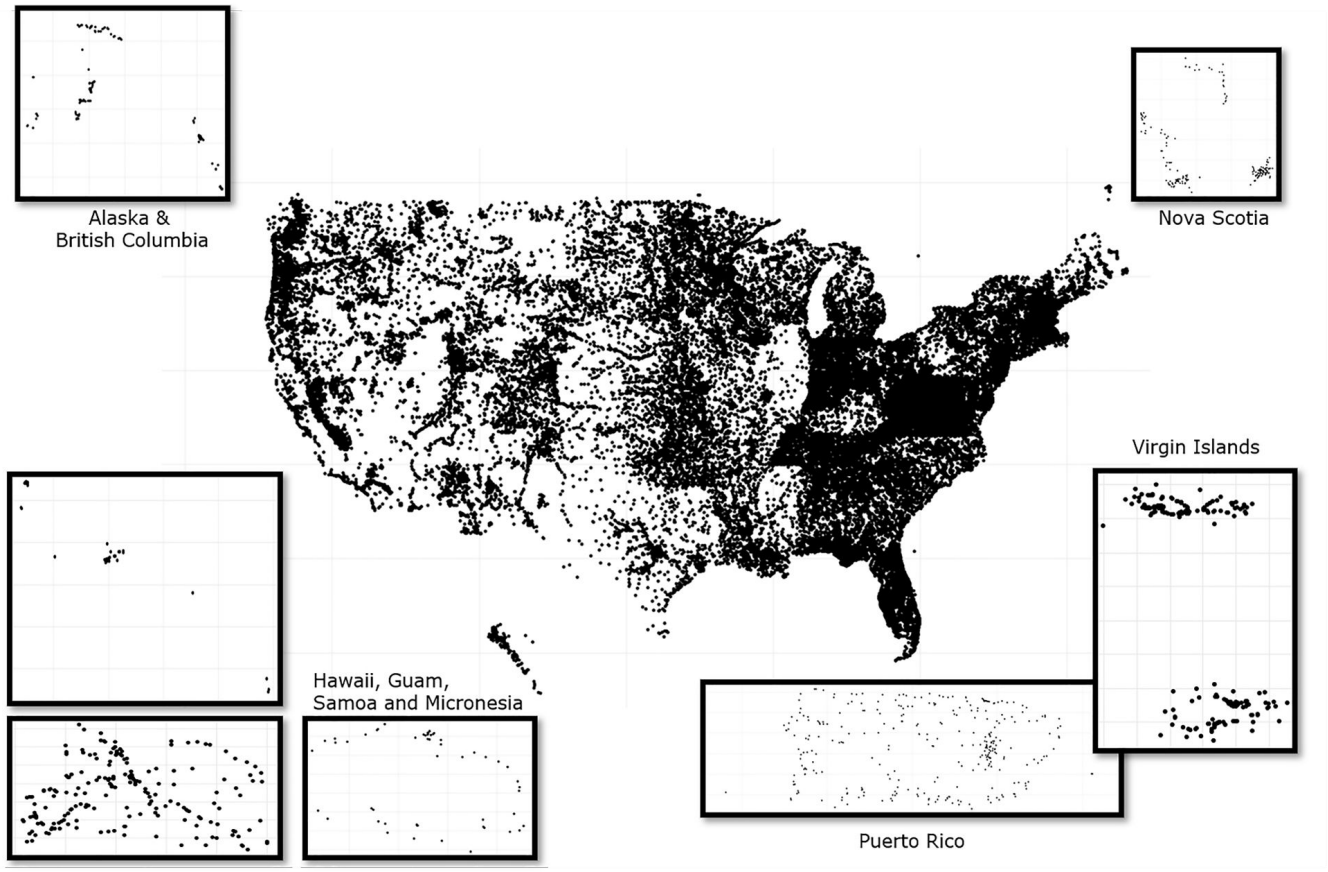

FIG 2 Location of the 100,410 unique sampling site locations represented by the data set compiled here. GPS coordinates were modified slightly to ensure confidentiality. (The maps were created in R using the ggplot2 and sf packages.)

Listeria spp. and L. monocytogenes

Listeria spp. data (*N* = 2,117) were obtained from five studies representing four US states and one Canadian province (Table 1; Table S2). *L. monocytogenes* data (*N* = 5,442) were obtained from eight studies representing four US states and two Canadian provinces (Table 1; Table S2). All *Listeria* spp. and *L. monocytogenes* data were collected between 2001 and 2018. Ninety-six percent of *Listeria* spp. data (*N* = 2,031) and 44% of *L. monocytogenes* data (*N* = 2,398) were grab samples (Table S2). Seventy-seven percent (*N* = 1,573) and 23% (*N* = 458) of grab samples tested for *Listeria* spp. were filtered through modified Moore swabs and membrane filters, respectively, while 19% (*N* = 457) and 79% (*N* = 1,893) of grab samples tested for *L. monocytogenes* were filtered through modified Moore swabs and membrane filters, respectively (Table S2). Two percent (*N* = 48) of grab samples tested for *L. monocytogenes* were not filtered. Culture-based methods were used by all studies to detect *Listeria* spp. and by seven studies (*N* = 5,406) to detect *L. monocytogenes*. Six studies (*N* = 4,060) confirmed isolates as *L. monocytogenes* through PCR and sequencing of *sigB*; one study (*N* = 36) used PCR and sequencing of *hly*; and one study (*N* = 446) used biochemical assays (Table S2).

Only 1% of variance in the likelihood of detecting *Listeria* spp. was jointly attributable to methodological and non-methodological factors, while 2% and 27% of variance were uniquely attributable to methodological and non-methodological factors, respectively (Table S3). Conversely, 8% of variance in the likelihood of detecting *L. monocytogenes* was jointly attributable to methodological and non-methodological factors, while 0% and 18% of variance were uniquely attributable to methodological and non-methodological factors, respectively (Fig. 3B; Table S3).

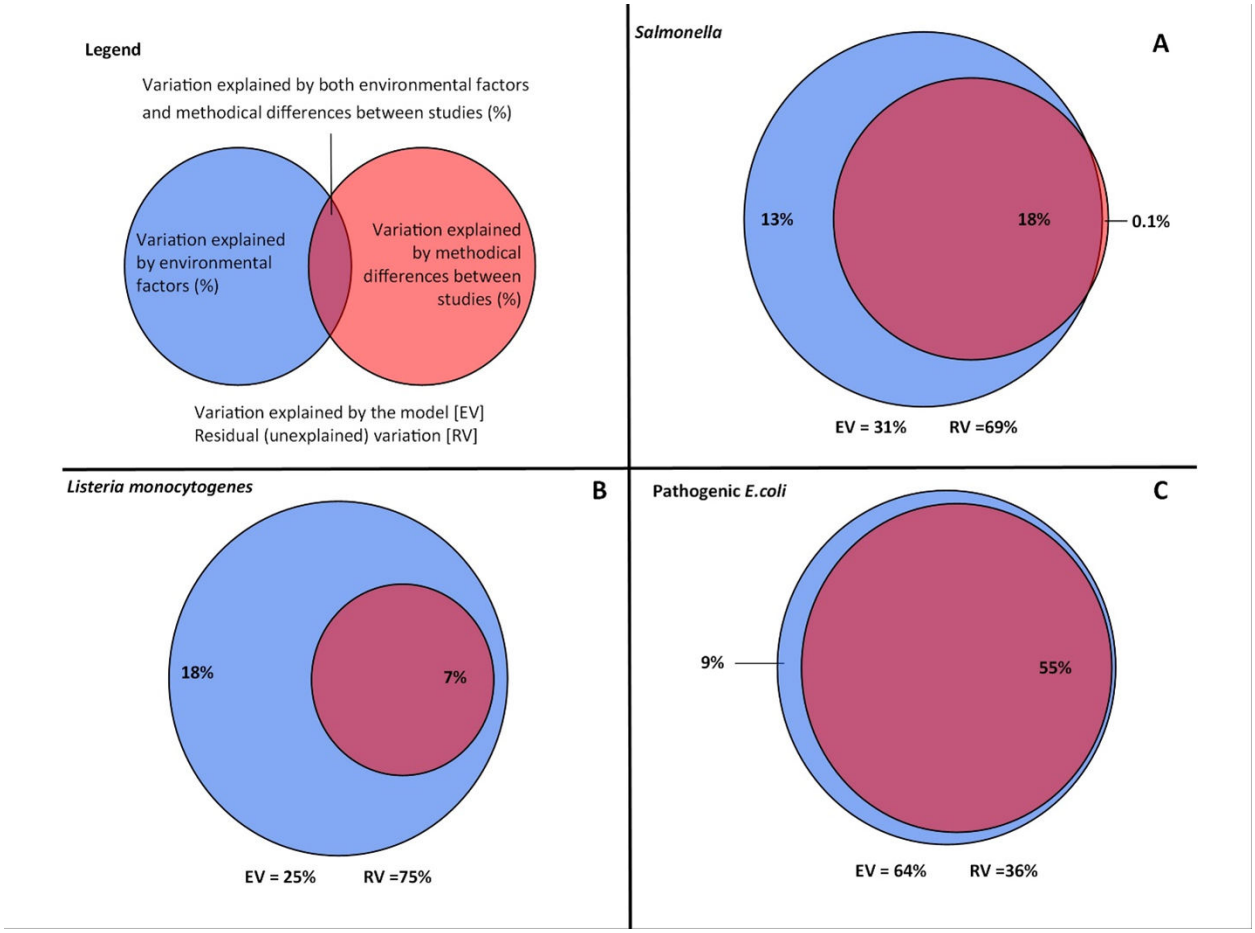

FIG 3 Variance in the likelihood of detecting (A) *Salmonella,* (B) *Listeria monocytogenes*, (C) and pathogenic *E. coli* that is jointly versus uniquely attributable to non-methodological (e.g., sampling site, season, water type, waterway, and year) and methodological (e.g., culture versus molecular-based detection, sample type, and volume) matrices.

After accounting for waterway and site-specific signals, the top-ranked factors associated with likelihood of detecting both *Listeria* spp. and *L. monocytogenes* were season, state, and sample type (Fig. 4B; Table S4). Based on generalized linear models and post hoc testing, the odds of isolating *L. monocytogenes* differed significantly between samples that were filtered through any type of filter compared to samples that were not filtered (Table S5). Sample volume was negatively associated with odds of *L. monocytogenes* isolation (OR = 0.92, SE = 0.11, P = 0.047) (Fig. 5B; Table S5). After accounting for methodological, waterway, and site-specific signals, the top-ranked factors associated with the likelihood of *Listeria* spp. detection were water type, state, and census region. The top-ranked factors associated with the likelihood of *L. monocytogenes* detection were water type, state, and EPA region (Table S6).

Pathogenic E. coli

Pathogenic *E. coli* data (N = 7,661) were obtained from 19 studies representing 20 US states, 2 Canadian provinces, and 1 Mexican state (Table 1; Table S2). All samples were collected between 2001 and 2019. Approximately half of samples tested for pathogenic *E. coli* were sampled using Moore swabs (N = 3,230) and grab samples (N = 4,431). Of the grab samples tested for pathogenic *E. coli*, 64% (N = 2,834) and 14% (N = 642) were filtered through membrane filters and modified Moore swabs, respectively, while 22% (N = 955) were not filtered. Culture-based methods were used to detect pathogenic *E. coli* in 58% (N = 4,407) of the samples, while molecular methods were used for 42% (N = 3,254).

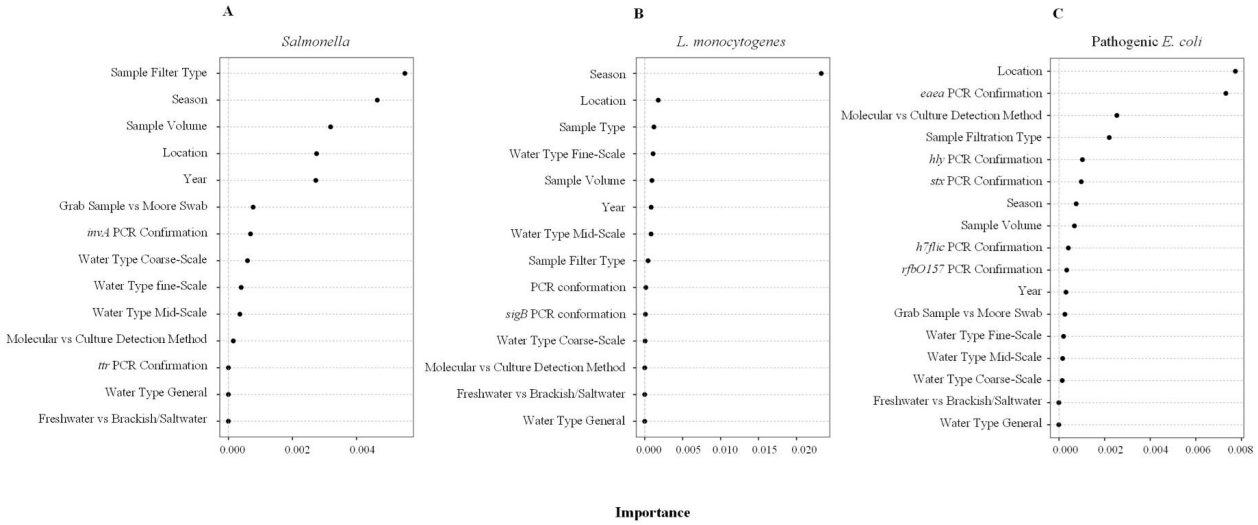

FIG 4 Results of conditional forest analysis that identified methodological and spatiotemporal factors associated with detection of (A) *Salmonella*, (B) *L. monocytogenes* and (C) pathogenic *E. coli* in water. The outcome of these forests was the residuals of a regression analysis that modeled likelihood of target pathogen detection as a function of two nested random effects (site and waterway). The *y*-axis shows the features ranked from highest to lowest variable importance. Variable importance is a unitless relative measure; thus, the importance of one variable should only be compared to another variable in the same plot, not between.

Based on variance partitioning analysis, 55% of variance in likelihood of detecting pathogenic *E. coli* was jointly attributable to methodological and non-methodological factors, with 0% and 9% of variance uniquely attributable to methodological and non-methodological factors, respectively (Fig. 3C; Table S3). When the variance in likelihood of detecting pathogenic *E. coli* attributable to waterway/sampling site, year/season, region, and methodological differences was considered, no variance was jointly attributable to all four sources, but 58% was jointly attributable to methods and at least one other source (Table S3). For example, 48% of variance in the likelihood of detecting pathogenic *E. coli* was jointly attributable to method, region, and waterway/sampling site. After accounting for waterway and site-specific signals, the top-ranked factors associated with the likelihood of pathogenic *E. coli* detection were state, the use of

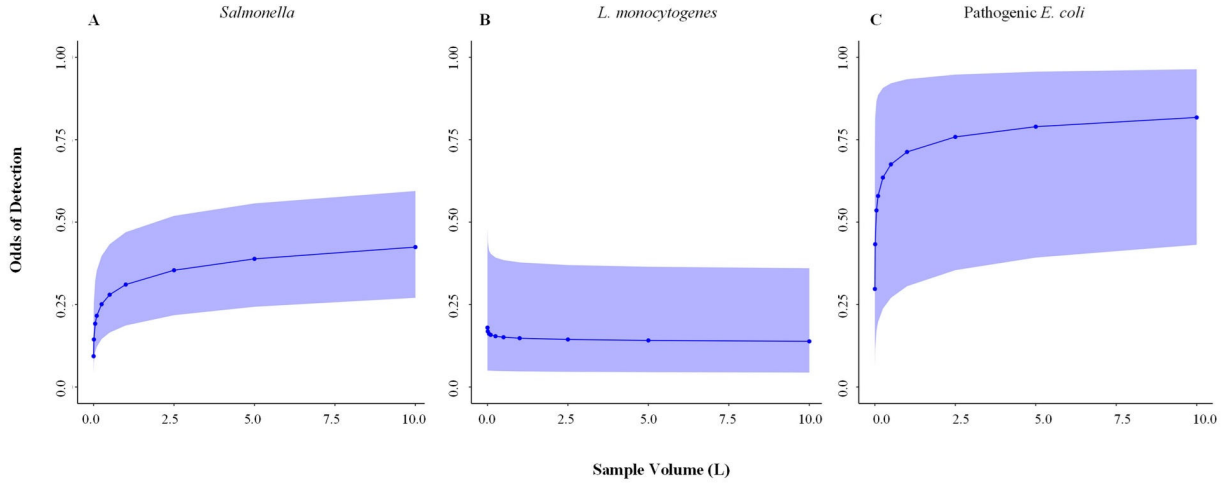

FIG 5 Impact of sample volume on probability of detection of (A) *Salmonella*, (B) *L. monocytogenes*, and (C) pathogenic *E. coli* according to generalized linear mixed models implemented with fixed effects of sample volume and season and random effects of site nested in waterway nested in state. No grab samples were tested for pathogen data in volumes greater than 10 L, and Moore swab volume was set to 10 L.

eaeA for detection/confirmation, and use of culture-based versus molecular detection methods (Fig. 4C; Table S4). Based on generalized linear mixed models and post hoc testing, the odds of pathogenic *E. coli* detection differed significantly ($P < 0.05$) by the genes used for detection/confirmation, if culture or molecular-based methods were used, sample filtration method, and sample volume (Table S5). After accounting for methodological, waterway, and site-specific signals, the top-ranked factors associated with likelihood of pathogenic *E. coli* detection were state, water type, and agricultural region (Table S6).

From the pathogenic *E. coli* samples, 1,978 samples were tested for *eaeA* allowing for identification of enteropathogenic *E. coli* (EPEC). The majority of samples tested for *eaeA* (95.5%; *N* = 1,890) were grab samples (Table S2). Based on variance partitioning analysis, 47% of variance in the likelihood of detecting EPEC was jointly attributable to methodological and non-methodological factors, while <1% and 15% of variance was uniquely attributable to methodological and non-methodological factors, respectively (Table S3). Season, state, and water type (general) were the top-ranked factors associated with EPEC detection (Table S4). In the forest that included regional factors, the highest-ranked factors associated with EPEC were water type, aquatic habitat type, and United States Department of Agriculture (USDA) region (Table S6).

Of the 7,661 samples tested for pathogenic *E. coli*, 6,589 were tested for *stx1* or *stx2* for identification of samples that were presumptively positive for Shiga toxin-producing *E. coli* (STEC; Table 1; Table S2). Approximately half of samples tested for STEC were sampled using Moore swabs (*N* = 3,230) and half were sampled using grab samples (*N* = 3,359). According to variable partitioning analysis, 16% of variance in the likelihood of detecting STEC was jointly attributable to methodological and non-methodological factors, while <1% and 21% of variance was uniquely attributable to methodological and non-methodological factors, respectively (Table S3). After accounting for waterway and site-specific signals, the top-ranked factors associated with likelihood of STEC detection were state, use of culture- versus molecular-based detection methods, and season (Table S4). After accounting for methodological, waterway, and site-specific signals, the top-ranked factors associated with the likelihood of STEC detection were water type, state, and EPA region (Table S6).

Of the 7,661 samples tested for pathogenic *E. coli*, 1,370 grab samples were tested for multiple genes allowing for presumptive identification of *E. coli* O157 positive samples (Table 1; Table S2). After accounting for waterway and site-specific signals, the top-ranked factors associated with likelihood of *E. coli* O157 detection were year, season, and water type (Table S4). After accounting for methodological, waterway, and site-specific signals, the top-ranked factors associated with the likelihood of *E. coli* O157 detection were aquatic habitat, USDA region, and agricultural region.

Non-methodological factors, as opposed to methodological factors, were more strongly associated with fecal indicator bacteria concentration

Generic *E. coli* (*N* = 1,362,230), *Enterococcus* (*N* = 151,578), fecal coliform (*N* = 1,127,750), and total coliform (*N* = 188,303) data were obtained from 57 US states and territories, Canadian provinces, and Mexican states (Table 1). While data on all four fecal indicator bacteria were collected from a diversity of water types, the proportion of data represented by each water type varied by indicator. For instance, oceans, tidal rivers, and estuaries represented 1% (*N* = 18,157) of *E. coli* data, but 8% (*N* = 12,108), 13% (*N* = 24,288), and 40% (*N* = 456,783) of *Enterococcus*, total coliform, and fecal coliform data, respectively. Conversely, ponds and lakes represented 5% (*N* = 53,818), 13% (*N* = 25,006), 21% (*N* = 38,309), and 24% (328,887) of fecal coliform, total coliform, *Enterococcus*, and *E. coli* data, respectively. Canals represented <5% of data for all four fecal indicators.

While *E. coli* concentrations were enumerated using MPN-based methods for 66% of samples (*N* = 918,919), 30% (*N* = 422,429), 3% (*N* = 48,585), and <1% (*N* = 1,464) were enumerated using membrane filtration, direct plating, or quantitative PCR-based approaches, respectively. IDEXX Quanti-Tray (65%; *N* = 904,814), EPA Method 1103 (10%;

N = 133,633), and EPA Method 1603 (8%; *N* = 106,500) were most frequently used for *E. coli* enumeration. Total coliform enumeration was most frequently performed using MPN-based methods (69%; *N* = 128,877), membrane filtration (23%; *N* = 43,120), and direct plating (7%; *N* = 13,562). Fecal coliform enumeration was most frequently performed using membrane filtration (58%; *N* = 650,123) and MPN-based methods (43%; *N* = 484,834). Standard Method 9222D was most frequently used for fecal coliform enumeration (54%; *N* = 610,324) followed by AOAC 978 (22%; *N* = 249,272) and APHA 3.2B (9%; *N* = 106,737). Approximately 40% of samples were tested for *Enterococcus* using membrane filtration (*N* = 66,455) and MPN-based methods (*N* = 64,393), while 14% (*N* = 20,730) used a molecular approach. IDEXX Enterolert (50%; *N* = 60,313), EPA Method 1600 (24%; *N* = 36,097), and Standard Method 9230C (20%; *N* = 29,581) were most frequently used for *Enterococcus* enumeration.

After accounting for waterway and site-specific signals, season was the top-ranked factor associated with *E. coli* concentrations in canals, rivers, streams, and other water types (e.g., estuaries, runoff, and wastewater) and the second ranked factor associated with *E. coli* concentrations in ponds, reservoirs, and lakes. Season was also the top-ranked factor associated with *Enterococcus* and fecal coliform concentrations in all water types and the second highest-ranked factor associated with total coliform concentrations (Table S7 and S8). After season, state was the second highest-ranked factor associated with *E. coli* concentrations in canals, rivers, streams, and other water types, *Enterococcus* concentrations, and fecal coliform concentrations in ponds, reservoirs, lakes, rivers, and streams (Table S7 and S8). State was the highest-ranked factor associated with *E. coli* concentrations in ponds, reservoirs, and lakes. Freshwater status and the mid-scale and/or coarse-scale methods factors were among the lowest ranked factors for all fecal indicator bacteria, regardless of water type. After accounting for methodological, waterway, and site-specific signals, the top-ranked factors associated with *Enterococcus* concentrations were ecoregion, water type, and hydrologic region. The top-ranked factors associated with total coliform concentration were state, terrestrial habitat type, and ecoregion (Table S8). Ecoregion was strongly associated with *E. coli* concentrations, and habitat type was strongly associated with fecal coliform concentrations in all water types considered.

DISCUSSION

Our analyses demonstrated water environments are intrinsically complex, and the use of different laboratory and sampling methods limited our ability to untangle this complexity, generating non-comparable results. Data collection and management need to be standardized across studies, and a minimum set of attributes that all studies should collect needs to be established (Table 2). While data from 2,429,990 samples were analyzed, these represent only 77% of the available data because 781,264 samples were discarded due to data quality issues. Even within a single organization, some data could be retained, while some were discarded due to inconsistencies in data collection, cleaning, and management within an organization. Furthermore, physiochemical water attributes and meteorological data, which are often considered during risk assessments for water quality, were not included in the present study due to inconsistencies in data collection and data quality issues. This highlights the need for (i) standardized data collection, cleaning, and management protocols within organizations, and (ii) reporting and archiving water quality data in a consistent way even when studies are conducted by unaffiliated organizations. Similar calls for uniform data standards have been made in other fields, and meeting such standards has become a requirement of certain funding agencies [e.g., US Health and Human Services Office of Minority Health Resource Center (149), US Office of Management and Budget (150), and US Centers for Disease Control and Prevention (151)]. Funding source-generated mandates show that method standardization and data reporting requirements are possible and provide a blueprint for implementing similar standards for water quality data. Such standards could be established by a consortium of key stakeholders, including funding agencies and/or

organizations from academia, industry, and government that generate large volumes of water data. Table 2 was developed using lessons learned from this study to help jumpstart the standardization conversation among microbial water quality researchers and end users. The table provides a brief overview of best practices and key considerations for collecting, recording, and reporting of microbial water quality data; Table 2 also cites resources that can be referred to for additional or more in-depth guidance.

Methodological differences between studies indicates findings are not comparable, limiting our ability to identify the broader ecological phenomenon driving foodborne pathogen contamination of waters and complicating the development of effective strategies for managing public health risks associated with contamination. Based on variance partitioning analysis, the impact of methodological differences on observed microbial water quality could not be disentangled from the impact of non-methodological (e.g., region, waterway, and water type) differences when the outcome was the detection of foodborne pathogens, indicator organisms, and fecal indicators. Conditional forest analysis showed that methodological differences were strongly associated with and predictive of observed water quality regardless of site-specific signals from the data. This is consistent with previous studies where the likelihood of detecting foodborne pathogens was strongly associated with the methods used to collect or test samples (15, 16, 27–29, 32). Past studies found that sample type, sample filtration method, and sample volume were strongly associated with the detection of *Listeria* spp., *L. monocytogenes*, *Salmonella*, and/or pathogenic *E. coli* (15, 16, 27–29, 32). Compared to *L. monocytogenes*, *Salmonella* and pathogenic *E. coli* are much more likely to be detected in 10-L grab samples filtered through modified Moore swabs compared to Moore swabs (15, 16, 27, 28). Our analysis also found significant differences in the likelihood of foodborne pathogen detection by filtration method but did not observe a significant difference in *L. monocytogenes* detection between membrane-filtered and modified Moore swab filtered samples, unlike a previous study (15). This difference may be due to substantially fewer samples (N = 29) in the previous study than the present study (N = 5,442) and most studies that tested for *L. monocytogenes* used the same or similar laboratory methods.

The impact of methodology on foodborne pathogen detection may be confounded by the heterogeneity of methods and the fact that some methods were only used by a single or small number of studies all conducted in a single region and/or on a single water type. For example, all grab samples that were not filtered and were tested for *Salmonella* came from a single study conducted in Sinaloa, Mexico. Therefore, we do not know if some of the larger odds ratios reported in Table S5 are due to actual methodological differences or confounding between methodology, region, study, laboratory, or water type. The size and heterogeneity of the data set reported here help reduce the impact of this confounding for commonly used methods (e.g., membrane filtration, modified Moore swabs, and Moore swabs).

With the exception of *Listeria* spp. and *L. monocytogenes*, larger sample volumes were associated with an increased likelihood of foodborne pathogen detection. Past studies demonstrating a similar trend have hypothesized that the inverse relationship between *Listeria* recovery and sample volume could be due to more competitive microflora in larger volume samples (15, 16). Regardless of why *Listeria* detection was inversely related to volume, *Salmonella* and pathogenic *E. coli* detection were positively associated with sample volume. This is consistent with a past study that found a 26-fold and 44-fold increase in odds of *Salmonella* recovery from 10-L grab samples compared to 1.0 and 0.1 L, respectively (32). A plateau effect was seen in the relationship between sample volume and odds of *Salmonella* and pathogenic *E. coli* detection; contrary to this previous study, increasing volume above 1 L does not substantially increase the odds of detection. Since collecting and processing larger volumes of water are more labor and capital intensive than collecting and processing smaller volumes, knowing the threshold for diminishing returns for sample volume is critical.

Similar to sample type, filtration method, sample volume, and pathogen detection method were all strongly associated with *Salmonella* and pathogenic *E. coli*

detection. While detection method was not significantly associated with *L. monocytogenes* detection, this may be an artifact of the fact that only one study representing 36 samples from a single water type and region used a molecular approach to detect *L. monocytogenes* (16). A previous study found *Listeria* detection was much more frequent when a TaqMan assay was used compared to a culture-based method (23). In the present study, odds of *Salmonella* detection were much lower, and odds of pathogenic *E. coli* detection were much higher when molecular-based as opposed to culture-based detection was used. A strong association between pathogenic *E. coli*, STEC, and EPEC detection and detection method is unsurprising, given the difficulties associated with culture-based detection of these foodborne pathogens (152–157). The finding that the odds of *Salmonella* detection were negatively associated with the use of molecular detection methods may be due to the coarse classification as either a culture-based or a molecular method. This binary classification ignored many of the substantial differences between protocols within these broad categories. For example, some studies used a culture-based methods that included PCR confirmation following isolation or a PCR-screen prior to isolation, while others used culture-based methods and did not include any form of molecular confirmation. The target gene was significantly associated with both likelihood of *Salmonella* and pathogenic *E. coli* detection in the present and previous studies (158–160). Similarly, the media used, incubation temperature, and timing are all known to affect foodborne pathogen recovery by culture-based methods (154, 156, 161, 162), and these were not considered here. Breaking down detection methods into culture and molecular-based approaches may have made it difficult to see the distinctions between various culture-based methods. This could have made it seem like there were fewer differences between culture-based and molecular approaches than there were. Despite this limitation, the findings of this and previous studies (16, 23, 158–160) highlight the impact of methodological differences between studies on observed water quality and suggest results may not be directly comparable between studies that used culture- and molecular-based methods. It is important to note that the difference in likelihood of detection by method could be due to higher false positive rates for culture-based or false negative rates for molecular-based methods.

Methodological factors were less impactful in the fecal indicator conditional forest analyses than non-methodological factors. This is unsurprising as many enumeration methods are considered equivalent to EPA Method 1603 (136, 137). Thus, existing fecal indicator data may be better suited for meta-analysis and use in large-scale modeling efforts than foodborne pathogen data. Much of the fecal indicator data (99.9% of *E. coli*, 86.3% of *Enterococcus*, and 100% of fecal and total coliform data) reported here were generated using culture-based as opposed to molecular methods; thus, differences due to the use of molecular methods could not be fully explored. Previously, higher *E. coli* concentrations using PCR-based as opposed to culture-based enumeration methods have been reported (163). While combining fecal indicator data enumerated using molecular and culture-based approaches will require further consideration, the level of methods standardization recommended for foodborne pathogen detection may not be needed for routine fecal indicator monitoring. There is still a need for standardization in data reporting and management since the majority of the 781,264 samples discarded for data issues were for fecal indicators.

Opportunities for site or waterway-specific management exist since substantial variance in microbial water quality was still uniquely attributable to non-methodological factors even though methodological signals could not be disentangled from non-methodological signals. While little to no variance in microbial water quality was uniquely attributable to methodological factors, a substantial amount of variance was uniquely attributable to non-methodological factors. More variance was uniquely attributable to sampling site and waterway compared to other non-methodological factors (e.g., region, year, and season), highlighting the dependency of microbial water quality on local environmental context. This is consistent with past studies that found evidence of strong site and waterway effects, and/or concluded that microbial water quality was

dependent on the local environmental context (16, 164, 165). Such dependency on local environmental factors complicates the establishment of one-size-fits-all water quality standards or universal best practices for the use of water.

While microbial water quality is strongly affected by local environmental factors, we implemented conditional forest analysis to see if and which regional, temporal, and water type factors were most strongly associated with each microbial water quality target. Water type-related factors were the top-ranked feature for five of seven foodborne pathogen forests, although the exact factor varied by microbial target. Conceptually, this is logical because different water types represent distinct environments where different processes drive water quality. For example, non-tidal rivers and streams are unidirectional and strongly influenced by upstream environmental conditions, but the impact of upstream conditions is reduced for other water types, such as the Great Lakes or seeps fed by groundwater. Past studies that sampled multiple water types often reported drastically different foodborne pathogen prevalence for the sampled water types (7, 12, 40, 166).

Every forest that considered regional factors in the present study included a regional feature among the top 10% of factors. Environmentally derived factors based on ecoregion (N = 9), habitat type (terrestrial = 6, aquatic = 2), climate (N = 1), or hydrologic region (N = 1) were the top-ranked regional features for 14 of the 17 forests. This is consistent with past studies that have repeatedly associated microbial water quality with environmental parameters, including weather and hydrologic characteristics (8–11, 167–172). Here, region and water type were strongly associated with several of the microbial water quality parameters considered in the present study.

Conclusion

This analysis shows that our current understanding of foodborne pathogen dynamics in water systems is limited by methodological confounders and brings into question the comparability of foodborne pathogen data generated by studies using different sampling and/or laboratory methods. Foodborne pathogen ecology in water is complex, without the added complications of methodological differences. Without comparability, it is difficult to identify the broader ecological phenomenon driving foodborne pathogen contamination of waters and complicating the development of effective strategies for managing public health risks associated with microbial contamination. This study highlights the need for standardizing sampling and laboratory methods used for microbial water quality testing. Future work could include comparing methods for equivalency. Similar standards are needed for data collection, management, and reporting. Since there is a diversity of methods used for a variety of sample types and fields within applied microbiology, similar analyses are needed to ensure the comparability of findings for other subfields of food and environmental microbiology. If methodological and data standards are not implemented, comparability will continue to be an issue, and there will always be a caveat to future findings.

MATERIALS AND METHODS

Data sets

Water quality data from peer-reviewed papers, publicly available databases, citizen science groups, and government programs were compiled (Table 1; database: https://github.com/wellerd2/Weller-et-al-2024-AEM-Datasets/tree/main). For each sampling site, water type, freshwater/saltwater status, and waterway name were determined; if these data were not available, Google Earth and Google Maps were used to obtain it. Four different, nested classifications for water type were used. The finest scale variable (fine-scale water type) included 27 categories; mid-scale water type included 20 categories; coarse-scale water type included 10 categories; and general-scale water type included four categories. For example, urban ponds (fine-scale water type) collapse

into ponds (mid-scale), which collapse into lakes/ponds/reservoirs (coarse-scale), which collapse into surface water (general). Footnote *h* in Table S9 describes all categories of the nested water type classification scheme used here.

Sampling and laboratory methods for each study were previously published, available online, or shared by the data set owners through personal correspondence. Key methods data were extracted and added to sample attributes. If methods data were not available for a sample, the sample was excluded from the study. Sample type, detection method, sample volume, and filtration method were recorded for samples with foodborne pathogen data. Since several of the analytical methods used here could not handle missing data, Moore swab samples were assigned a volume of 10 L, which is the largest grab sample volume collected in this study.

Studies often tested samples for a microbial target using a culture-based method and confirmed presumptive positives using a molecular method or vice versa. However, most of the data sets that included a confirmation step only provided data on if both the culture and molecular results were positive (e.g., included a single column for if it was culture-positive and molecular confirmed, not separate columns for culture and molecular results). As a result, only a single positive-negative designation was reported for most data sets. Thus, a sample was classified as being tested using a culture-based method, if there was any culture-based step (i.e., culture-based detection with no confirmation, molecular detection with culture confirmation, and culture-based detection with molecular confirmation). However, if results were available for both the culture and molecular tests separately, we included these as separate data rows with a common sample site ID and date; indeed, if any sample was tested by multiple methods (e.g., different filtration methods and sample volumes), then these results were treated as separate data rows linked by site ID and date. If a sample was screened by molecular detection with culture confirmation, it was considered molecular. If molecular methods were used, target genes were noted. Since there are multiple types of pathogenic *Escherichia coli*, samples were categorized as positive-negative for any pathogenic *E. coli*, and for EPEC (based on detection of the *eaeA* gene), STEC (based on detection of the *stx* genes), and *E. coli* O157 (using culture-based methods and PCR confirmation).

Fecal indicator methods were classified by the protocol used for enumeration and separately by the media used. Since not all studies used an established protocol, three different approaches were used to categorize methods (Table S1). A fine-scale method factor was created and reflects the established protocol used (e.g., Standard Method 9222B) or if a specific protocol could not be identified, study ID was used in place of the fine-scale method. A mid-level method factor was used to capture methods that were almost the same but with slight variations (e.g., Standard Method 9222B and Standard Method 9222C were Standard Method 9222). A coarse-level factor was used to capture if direct plating, membrane filtration, most probable number estimation, or a molecular approach was used for enumeration.

Assigning samples to regions

Using GPS coordinates and county, we classified samples separately into regions using different regional schemes. Different approaches for grouping samples into regions were used to determine if there was a regional scheme that accounted for the greatest variance for each microbial target. Fourteen regional schemes were considered, including schemes based on the biome, ecoregion, aquatic and terrestrial habitat type, climate region, hydrologic region (based on USGS HUC2 unit codes; 173) regional classifications used by US federal agencies (i.e., Census Bureau and US Environmental Protection Agency), and on agricultural practices and/or output (USDA-Farm Resource Regions, Farm Production Regions, National Agricultural Statistics Service Regions, and Human Geography Agricultural Regions). For US-specific schemes, sites outside the US were assigned to the region of the closest US site.

Statistical analysis

All analyses were performed in R version 4.0.3 (R Foundation for Statistical Computing, Vienna, Austria). A summary of all analyses implemented here, as well as their reason for being implemented and key model specifications, is outlined in Table 3 . Table S9 lists methodological and non-methodological factors used here.

Conditional forest analysis

Conditional forest analysis was used to characterize hierarchical associations between methodological, spatial, and temporal factors to better understand if and when methodological differences affected observed microbial water quality, and if interactions between methods and other factors might affect observed microbial water quality. Due to the large number of samples with *E. coli* or fecal coliform data, we lacked the computational resources to implement a single forest for either fecal indicator. Instead, separate *E. coli* and fecal coliform forests were implemented for samples collected from canals, lakes and ponds, streams, rivers, and all other water types (e.g., groundwater and ocean). As a result, five separate forests were run using the *E. coli* and fecal coliform data. Prior to training each forest, a general linear model (for continuous outcomes) or generalized linear model (for binary outcomes) with a random effect of site ID nested in waterway was fit using the *lme4* package (174); the dependent variable in the forest was the residuals from this model.

Separately, conditional forest analysis was used to determine if water type and/or regional scheme was more strongly associated with each microbial target after accounting for other confounding factors (e.g., methodological factors). As described above, separate water type-specific models were implemented for the *E. coli* and fecal coliform data. Prior to training each forest, a general linear model or generalized linear model was fit with random effects for each methodological factor available for the microbial target using the *lme4* package (174). If the model failed to converge or had singular fit, the model was re-parameterized (e.g., random effect shifted to fixed effect, or a factor was dropped); the dependent variable in the forest was the residuals from this model. For both sets of forests, unbiased conditional forest analysis was implemented using the *moreparty* package (175). Conditional variable importance was calculated to identify factors in each forest that were most strongly associated with each microbial target; conditional variable importance was used because it is unbiased by correlation between covariates.

Variance attributable to methodological versus non-methodological signals

Variance partitioning analysis was implemented using the *vegan* package (176) to quantify the variance in likelihood of foodborne pathogen detection uniquely and jointly attributable to methodological and non-methodological factors. For foodborne pathogen targets, four sets of variance partitioning analyses were performed using the following sets of matrices (i) state and region, site (waterway, site, water type, and freshwater status), temporal (season and year), and methodological factors; (ii) waterway (waterway and sampling site), water type (water type and freshwater status), temporal and methodological factors; (iii) methodological versus all other non-methodological factors; and (iv) sampling site, methodological factors, and all other non-methodological factors. Due to the computational intensity of these analyses and the large number of samples, variance partitioning analysis was not performed for fecal indicator bacteria.

Mixed models were implemented to identify "comparable" methods

To quantify how using a given sample processing or laboratory method influenced the likelihood of foodborne pathogen or indicator organism detection, generalized linear models were implemented using the *lme4* package (174). The models were implemented with the binomial family, a logit link, random effects of site nested in waterway nested in state, a random effect of season, and a fixed effect for the methodological

TABLE 3 Summary of statistical analysis performed, the goal of each analysis, R packages and model specifications needed for each analysis, and the data subsets that analysis was performed on

Analysis	Goal	R packages used	Model specifications	Data subset analysis were performed on
Conditional forest analysis using methodological, spatial, and temporal factors	To characterize hierarchical associations between methodological, spatial, and temporal factors	*lme4*, *moreparty*	• Prior to training each forest, a general linear model or generalized linear model was fit with random effect of site ID nested in waterway, and residuals from this model were the dependent variable in the forest.	• *Salmonella*. • *Listeria* spp. • *L. monocytogenes*. • Pathogenic *E. coli*. • EPEC. • STEC. • *E. coli* O157. • Generic *E. coli* by water type. • *Enterococcus*. • Total coliforms. • Fecal coliforms by water type.
Conditional forest analysis using water type and regional factors	To determine if water type and/or regional scheme was more strongly associated with each microbial target after accounting for other confounding factors (e.g., methodological).	*lme4*, *moreparty*	• Prior to training each forest, a general linear model or generalized linear model was fit with random effects for each methodological variable, and residuals from this model were the dependent variable in the forest.	
Variance partitioning analysis	To quantify the variance in likelihood of target detection that was uniquely versus jointly attributable to methodological versus non-methodological factors	*vegan*	Four sets of analyses were performed using the following matrices, each consisting of one or more factors: • A methodological matrix and a non-methodological matrix. • A methodological matrix and three non-methodological matrices (state and region, waterway/site and water type, and temporal). • A methodological matrix and three non-methodological matrices (waterway/site, water type, and temporal). • A methodological matrix and two non-methodological matrices (sampling site and all other non-methodological factors).	• *Salmonella*. • *Listeria* spp. • *L.monocytogenes*. • Pathogenic *E. coli*. • EPEC. • STEC.
Generalized linear models	To quantify how using a given method influenced the likelihood of detecting a microbial target	*lme4*	• Mixed model, binomial family, and logit link. • Random effects of site nested in waterway nested in state, and of season. • Fixed effect for the methodological factor of interest. • Benjamin-Hochberg multiple comparison correction applied.	• *Salmonella* • *Listeria* spp. • *L. monocytogenes*. • Pathogenic *E. coli*. • EPEC. • STEC.
Tukey's HSD[a]	To determine if specific methodological choices generated comparable data (e.g., if there was a significant difference in the likelihood of detection if	*multcomp*	• Tukey's HSD was applied post hoc to the model object returned from the generalized linear models run above when the methodological factor of interest was categorical.	

(*Continued on next page*)

TABLE 3 Summary of statistical analysis performed, the goal of each analysis, R packages and model specifications needed for each analysis, and the data subsets that analysis was performed on (*Continued*)

Analysis	Goal	R packages used	Model specifications	Data subset analysis were performed on
	membrane filtration, modified Moore swabs, or no filter was used).			

[a]HSD, honestly significant difference.

factor of interest. To determine if specific methodological choices generated comparable data (e.g., if there was a significant difference in the likelihood of detection if membrane filtration, modified Moore swabs, or no filter was used), Tukey's honestly significant difference was performed using the *multcomp* package (177). To account for multiple comparisons, the Benjamin-Hochberg multiple comparison correction was used. These analyses were only performed using the foodborne pathogen data (as opposed to the fecal indicator data) because (i) foodborne pathogen contamination is the primary outcome of interest and (ii) methodological factors were consistently among the top-ranked factors in the foodborne pathogen forests but not in fecal indicator forests.

ACKNOWLEDGMENTS

This work was supported by the Specialty Crops Research Initiative project 2019–51181-30016 from the United States Department of Agriculture (USDA) National Institute of Food and Agriculture, and Virginia Agricultural Experiment Station. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the view of the USDA.

We thank Dr. Jennifer McEntire for her time in reviewing the manuscript and Drs. Faith Critzer, Channah Rock, and Donald Schaffner for their time in reviewing Table 2.

AUTHOR AFFILIATIONS

[1]Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, New York, USA
[2]Department of Food Science and Technology, Virginia Tech, Blacksburg, Virginia, USA
[3]Department of Food Science and Human Nutrition, Citrus Research and Education Center, University of Florida, Lake Alfred, Florida, USA

AUTHOR ORCIDs

Daniel L. Weller http://orcid.org/0000-0001-7259-6331
Claire M. Murphy http://orcid.org/0009-0008-4082-6276
Tanzy M. T. Love http://orcid.org/0000-0002-7154-0229
Michelle D. Danyluk http://orcid.org/0000-0001-5780-7911
Laura K. Strawn http://orcid.org/0000-0002-9523-0081

FUNDING

Funder	Grant(s)	Author(s)
U.S. Department of Agriculture (USDA)	2019-51181-30016	Laura K. Strawn
VT | Virginia Agricultural Experiment Station, Virginia Polytechnic Institute and State University (VAES)		Laura K. Strawn

AUTHOR CONTRIBUTIONS

Daniel L. Weller, Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review and editing | Claire M. Murphy, Data curation, Investigation, Validation, Visualization, Writing – review and editing | Tanzy M. T. Love, Conceptualization, Resources, Software, Writing – review and editing | Michelle D. Danyluk, Conceptualization, Funding acquisition, Visualization, Writing – review and editing | Laura K. Strawn, Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – review and editing

DATA AVAILABILITY

Table 1 outlines the source of all data used in this study including all original study authors/data owners. Datasets and locations are available at: https://github.com/wellerd2/Weller-et-al-2024-AEM-Datasets/tree/main, or by contacting the study authors (L.S., C.M., D.W.) for confidential data. Confidential data including GPS coordinates, site names, and waterway names were dropped at request of the original study authors due to commercial agricultural farm privacy (also prior published studies have kept this information confidential), but is available upon request of the original study authors (listed in Table 1).

ADDITIONAL FILES

The following material is available online.

Supplemental Material

Figure S1, Tables S1 to S9 (AEM01835-23-s0001.docx). Supplemental materials providing additional details or outputs of data analyses.

REFERENCES

1. Craun MF, Craun GF, Calderon RL, Beach MJ. 2006. Waterborne outbreaks reported in the United States. J Water Health 4:19–30. https://doi.org/10.2166/wh.2006.016
2. Food and Drug Administration. 2019. Environmental assessment of factors potentially contributing to the contamination of Romaine lettuce implicated in a multi-state outbreak of *E. coli*. Available from: https://www.fda.gov/food/outbreaks-foodborne-illness/environmental-assessment-factors-potentially-contributing-contamination-romaine-lettuce-implicated
3. Graciaa DS, Cope JR, Roberts VA, Cikesh BL, Kahler AM, Vigar M, Hilborn ED, Wade TJ, Backer LC, Montgomery SP, Secor WE, Hill VR, Beach MJ, Fullerton KE, Yoder JS, Hlavsa MC. 2018. Outbreaks associated with untreated recreational water—United States. MMWR Morb Mortal Wkly Rep 67:701–706. https://doi.org/10.15585/mmwr.mm6725a1
4. Greene SK, Daly ER, Talbot EA, Demma LJ, Holzbauer S, Patel NJ, Hill TA, Walderhaug MO, Hoekstra RM, Lynch MF, Painter JA. 2008. Recurrent multistate outbreak of *Salmonella* Newport associated with tomatoes from contaminated fields, 2005. Epidemiol Infect 136:157–165. https://doi.org/10.1017/S095026880700859X
5. Lee SH, Levy DA, Craun GF, Beach MJ, Calderon RL. 2002. Surveillance for waterborne-disease outbreaks--United States, 1999-2000. Morb Mortal Wkly Rep Surveillance summaries 51:1–47.
6. US Environmental Protection Agency. 2009. Review of published studies to characterize relative risks from different sources of fecal contamination in recreational water
7. Cooley MB, Quiñones B, Oryang D, Mandrell RE, Gorski L. 2014. Prevalence of shiga toxin producing *Escherichia coli*, *Salmonella enterica*, and *Listeria monocytogenes* at public access watershed sites in a California central coast agricultural region. Front Cell Infect Microbiol 4:30. https://doi.org/10.3389/fcimb.2014.00030
8. Draper AD, Doores S, Gourama H, LaBorde LF. 2016. Microbial survey of Pennsylvania surface water used for irrigating produce crops. J Food Prot 79:902–912. https://doi.org/10.4315/0362-028X.JFP-15-479
9. Green H, Wilder M, Wiedmann M, Weller DL. 2021. Integrative survey of 68 non-overlapping upstate New York watersheds reveals stream features associated with aquatic fecal contamination. Front Microbiol 12:684533. https://doi.org/10.3389/fmicb.2021.684533
10. Gu G, Strawn LK, Ottesen AR, Ramachandran P, Reed EA, Zheng J, Boyer RR, Rideout SL. 2020. Correlation of *Salmonella enterica* and *Listeria monocytogenes* in irrigation water to environmental factors, fecal indicators, and bacterial communities. Front Microbiol 11:557289. https://doi.org/10.3389/fmicb.2020.557289
11. 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. https://doi.org/10.1128/AEM.00777-13
12. Murphy CM, Strawn LK, Chapin TK, McEgan R, Gopidi S, Friedrich L, Goodridge LD, Weller DL, Schneider KR, Danyluk MD. 2022. Factors associated with *E. coli* levels in and *Salmonella* contamination of agricultural water differed between North and South Florida waterways. Front Water 3. https://doi.org/10.3389/frwa.2021.750673
13. Topalcengiz Z, Strawn LK, Danyluk MD. 2017. Microbial quality of agricultural water in central Florida. PLoS One 12:e0174889. https://doi.org/10.1371/journal.pone.0174889
14. Truitt LN, Vazquez KM, Pfuntner RC, Rideout SL, Havelaar AH, Strawn LK. 2018. Microbial quality of agricultural water used in produce preharvest production on the Eastern shore of Virginia. J Food Prot 81:1661–1672. https://doi.org/10.4315/0362-028X.JFP-18-185
15. Weller D, Belias A, Green H, Roof S, Wiedmann M. 2020. Landscape, water quality, and weather factors associated with an increased likelihood of foodborne pathogen contamination of New York streams used to source water for produce production. Front Sustain Food Syst 3:124. https://doi.org/10.3389/fsufs.2019.00124
16. Weller D, Brassill N, Rock C, Ivanek R, Mudrak E, Roof S, Ganda E, Wiedmann M. 2020. Complex interactions between weather, and microbial and physicochemical water quality impact the likelihood of detecting foodborne pathogens in agricultural water. Front Microbiol 11:134. https://doi.org/10.3389/fmicb.2020.00134
17. Weller DL, Murphy CM, Johnson S, Green H, Michalenko EM, Love TMT, Strawn LK. 2022. Land use, weather, and water quality factors associated with fecal contamination of northeastern streams that span an urban-rural gradient. Front Water 3. https://doi.org/10.3389/frwa.2021.741676
18. 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 43:2209–2223. https://doi.org/10.1016/j.watres.2009.01.033
19. Cooley MB, Carychao D, Gorski L. 2018. Optimized co-extraction and quantification of DNA from enteric pathogens in surface water samples near produce fields in California. Front Microbiol 9:448. https://doi.org/10.3389/fmicb.2018.00448
20. Canizalez-Roman A, Velazquez-Roman J, Valdez-Flores MA, Flores-Villaseñor H, Vidal JE, Muro-Amador S, Guadrón-Llanos AM, Gonzalez-Nuñez E, Medina-Serrano J, Tapia-Pastrana G, León-Sicairos N. 2019. Detection of antimicrobial-resistance diarrheagenic *Escherichia coli* strains in surface water used to irrigate food products in the Northwest of Mexico. Int J Food Microbiol 304:1–10. https://doi.org/10.1016/j.ijfoodmicro.2019.05.017

21. Brooks YM, Spirito CM, Bae JS, Hong A, Mosier EM, Sausele DJ, Fernandez-Baca CP, Epstein JL, Shapley DJ, Goodman LB, Anderson RR, Glaser AL, Richardson RE. 2020. Fecal indicator bacteria, fecal source tracking markers, and pathogens detected in two Hudson river tributaries. Water Res 171:115342. https://doi.org/10.1016/j.watres.2019.115342
22. Harris CS, Tertuliano M, Rajeev S, Vellidis G, Levy K. 2018. Impact of storm runoff on *Salmonella* and *Escherichia coli* prevalence in irrigation ponds of fresh produce farms in southern Georgia. J Appl Microbiol 124:910–921. https://doi.org/10.1111/jam.13689
23. Stea EC, Purdue LM, Jamieson RC, Yost CK, Truelstrup Hansen L. 2015. Comparison of the prevalences and diversities of *Listeria* species and *Listeria monocytogenes* in an urban and a rural agricultural watershed. Appl Environ Microbiol 81:3812–3822. https://doi.org/10.1128/AEM.00416-15
24. Haack SK, Duris JW, Fogarty LR, Kolpin DW, Focazio MJ, Furlong ET, Meyer MT. 2009. Comparing wastewater chemicals, indicator bacteria concentrations, and bacterial pathogen genes as fecal pollution indicators. J Environ Qual 38:248–258. https://doi.org/10.2134/jeq2008.0173
25. Deaven AM, Ferreira CM, Reed EA, Chen See JR, Lee NA, Almaraz E, Rios PC, Marogi JG, Lamendella R, Zheng J, Bell RL, Shariat NW, Dozois CM. 2021. *Salmonella* genomics and population analyses reveal high inter- and intraserovar diversity in freshwater. Appl Environ Microbiol 87:e02594-20. https://doi.org/10.1128/AEM.02594-20
26. Walters SP, Thebo AL, Boehm AB. 2011. Impact of urbanization and agriculture on the occurrence of bacterial pathogens and STX genes in coastal waterbodies of central California. Water Res 45:1752–1762. https://doi.org/10.1016/j.watres.2010.11.032
27. 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. https://doi.org/10.1016/j.ijfoodmicro.2013.04.003
28. Colburn KG, Kaysner CA, Abeyta C Jr, Wekell MM. 1990. *Listeria* species in a California coast estuarine environment. Appl Environ Microbiol 56:2007–2011. https://doi.org/10.1128/aem.56.7.2007-2011.1990
29. Jimenez M, Chaidez C. 2012. Improving *Salmonella* determination in Sinaloa rivers with ultrafiltration and most probable number methods. Environ Monit Assess 184:4271–4277. https://doi.org/10.1007/s10661-011-2262-9
30. McEgan R, Rodrigues CAP, Sbodio A, Suslow TV, Goodridge LD, Danyluk MD. 2013. Detection of *Salmonella* spp. from large volumes of water by modified Moore swabs and tangential flow filtration. Lett Appl Microbiol 56:88–94. https://doi.org/10.1111/lam.12016
31. Meinersmann RJ, Berrang ME, Bradshaw JK, Molina M, Cosby DE, Genzlinger LL, Snyder BJ. 2020. Recovery of thermophilic *Campylobacter* by three sampling methods from river sites in northeast Georgia, USA, and their antimicrobial resistance genes. Lett Appl Microbiol 71:102–107. https://doi.org/10.1111/lam.13224
32. Sharma M, Handy ET, East CL, Kim S, Jiang C, Callahan MT, Allard SM, Micallef S, Craighead S, Anderson-Coughlin B, Gartley S, Vanore A, Kniel KE, Haymaker J, Duncan R, Foust D, White C, Taabodi M, Hashem F, Parveen S, May E, Bui A, Craddock H, Kulkarni P, Murray RT, Sapkota AR. 2020. Prevalence of *Salmonella* and *Listeria monocytogenes* in non-traditional irrigation waters in the Mid-Atlantic United States is affected by water type, season, and recovery method. PLoS One 15:e0229365. https://doi.org/10.1371/journal.pone.0229365
33. Chapin TK, Nightingale KK, Worobo RW, Wiedmann M, Strawn LK. 2014. Geographical and meteorological factors associated with isolation of *Listeria* species in New York State produce production and natural environments. J Food Prot 77:1919–1928. https://doi.org/10.4315/0362-028X.JFP-14-132
34. Ivanek R, Gröhn YT, Wells MT, Lembo AJ Jr, Sauders BD, Wiedmann M. 2009. Modeling of spatially referenced environmental and meteorological factors influencing the probability of *Listeria* species isolation from natural environments. Appl Environ Microbiol 75:5893–5909. https://doi.org/10.1128/AEM.02757-08
35. Rodrigues C, da Silva ALBR, Dunn LL. 2020. Factors impacting the prevalence of foodborne pathogens in agricultural water sources in the southeastern United States. Water 12:51. https://doi.org/10.3390/w12010051
36. Strawn LK, Danyluk MD, Worobo RW, Wiedmann M. 2014. Distributions of *Salmonella* subtypes differ between two US produce-growing regions. Appl Environ Microbiol 80:3982–3991. https://doi.org/10.1128/AEM.00348-14
37. Weller DL, Love TMT, Weller DE, Murphy CM, Rahm BG, Wiedmann M, Dudley EG. 2022. Structural equation models suggest that on-farm noncrop vegetation removal is not associated with improved food safety outcomes but is linked to impaired water quality. Appl Environ Microbiol 88:e0160022. https://doi.org/10.1128/aem.01600-22
38. Murphy CM, Weller DL, Strawn LK. 2023. *Salmonella* prevalence is strongly associated with spatial factors while *Listeria monocytogenes* prevalence is strongly associated with temporal factors on Virginia produce farms. Appl Environ Microbiol 89:e0152922. https://doi.org/10.1128/aem.01529-22
39. Acheamfour CL, Parveen S, Hashem F, Sharma M, Gerdes ME, May EB, Rogers K, Haymaker J, Duncan R, Foust D, et al. 2021. Levels of *Salmonella enterica* and *Listeria monocytogenes* in alternative irrigation water vary based on water source on the Eastern shore of Maryland. Microbiol Spectr 9:e0066921. https://doi.org/10.1128/Spectrum.00669-21
40. Belias A, Strawn LK, Wiedmann M, Weller D. 2021. Small produce farm environments can harbor diverse *Listeria monocytogenes* and *Listeria* spp. J Food Prot 84:113–121. https://doi.org/10.4315/JFP-20-179
41. Weller DL, Weller DE, Strawn LK, Love TMT. 2022. Scale of analysis drives the observed ratio of spatial to non-spatial variance in microbial water quality: insights from two decades of citizen science data. bioRxiv. https://doi.org/10.1101/2022.02.01.478743
42. ACAP Saint John. 2022. Water quality monitoring. Available from: https://www.acapsj.org/reports/?category=Water+Quality. Retrieved 10 Oct 2023.
43. Adhikari A, Chhetri VS, Camas A. 2020. Evaluation of microbiological quality of agricultural water and effect of water source and holding temperature on the stability of indicator organisms' levels by seven US environmental protection agency–approved methods. J Food Prot 83:249–255. https://doi.org/10.4315/0362-028X.JFP-19-381
44. Maher J, Nwadike L, Gragg S, Bhullar M. 2020. Survey of agricultural water microbial quality in Kansas and Missouri. IAFP
45. Bhullar MS, Shaw A, Hannan J, Andrews S. 2019. Extending the holding time for agricultural water testing EPA method 1603 for produce growers. Water 11:2020. https://doi.org/10.3390/w11102020
46. Riverkeeper BW. 2023. Ambient water monitoring. Available from: https://blackwarriorriver.org/ambient-water-monitoring/about.php. Retrieved 10 Oct 2023.
47. California state water resources control board. 2023. California environmental data exchange network. Available from: https://www.sfei.org/projects/california-environmental-data-exchange-network-ceden. Retrieved 10 Oct 2023.
48. Groffman P, Rosi E, Martel L, Kaushal S. 2018. Stream chemistry for core sites in Gwynns falls: concentration of Cl, NO3, PO4, total N and P, SO4, dissolved oxygen, *E. coli*, plus temperature, pH, clarity, turbidity, isotopes, and pharmaceuticals Ver 600. Environ Data Initiat
49. Krometis L-A, Characklis GW, Drummey PN, Sobsey MD. 2010. Comparison of the presence and partitioning behavior of indicator organisms and *Salmonella* Spp. in an urban watershed. J Water Health 8:44–59. https://doi.org/10.2166/wh.2009.032
50. Krometis L-AH. 2009. Microbial partitioning in urban stormwaters. The University of North Carolina at Chapel Hill.
51. Dilts MJ. 2004. Impact of Microbial-Particle Interaction on Microbial Fate and Transport in Stormwater.
52. Chesapeake Bay Foundation. 2023. Runoff pollution. Available from: https://www.cbf.org/issues/polluted-runoff/index.html. Retrieved 10 Oct 2023.
53. Chesapeake monitoring cooperative. 2023. Chesapeake data explorer. Available from: https://cmc.vims.edu/#/home. Retrieved 10 Oct 2023.
54. City of Austin. 2023. Watershed protection environmental resource management. Water quality sampling data. Available from: https://data.austintexas.gov/widgets/5tye-7ray?mobile_redirect=true. Retrieved 10 Oct 2023.

55. City of Chicago park district. 2006. Chicago data portal. Available from: https://data.cityofchicago.org/Parks-Recreation/Beach-Lab-Data/2ivx-z93u/data. Retrieved 10 Oct 2023.
56. Brumfield KD, Cotruvo JA, Shanks OC, Sivaganesan M, Hey J, Hasan NA, Huq A, Colwell RR, Leddy MB. 2021. Metagenomic sequencing and quantitative real-time PCR for fecal pollution assessment in an urban watershed. Front Water 3:626849. https://doi.org/10.3389/frwa.2021.626849
57. Belias A, Brassill N, Roof S, Rock C, Wiedmann M, Weller D. 2021. Cross-validation indicates predictive models may provide an alternative to indicator organism monitoring for evaluating pathogen presence in southwestern US agricultural water. Front Water 3. https://doi.org/10.3389/frwa.2021.693631
58. Sauders BD, Overdevest J, Fortes E, Windham K, Schukken Y, Lembo A, Wiedmann M. 2012. Diversity of *Listeria* species in urban and natural environments. Appl Environ Microbiol 78:4420–4433. https://doi.org/10.1128/AEM.00282-12
59. Strawn LK, Fortes ED, Bihn EA, 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. https://doi.org/10.1128/AEM.02491-12
60. Strawn LK, Gröhn YT, Warchocki S, Worobo RW, Bihn EA, Wiedmann M. 2013. Risk factors associated with *Salmonella* and *Listeria monocytogenes* contamination of produce fields. Appl Environ Microbiol 79:7618–7627. https://doi.org/10.1128/AEM.02831-13
61. Four Rivers Watershed Watch. 2023. Available from: https://www.frww.org/. Retrieved 10 Oct 2023.
62. Georgia Enviro. 2023. Monitoring and assessment system. Available from: https://gomaspublic.gaepd.org/. Retrieved 10 Oct 2023.
63. Green H, Weller D, Johnson S, Michalenko E. 2019. Microbial source-tracking reveals origins of fecal contamination in a recovering watershed. Water (Basel) 11:2162. https://doi.org/10.3390/w11102162
64. Staley ZR, Chase E, Mitraki C, Crisman TL, Harwood VJ. 2013. Microbial water quality in freshwater lakes with different land use. J Appl Microbiol 115:1240–1250. https://doi.org/10.1111/jam.12312
65. Iowa Department of Natural Resources. Ambient Stream Monitoring.
66. Louisiana Department of Environmental Quality. 2023. Water data. Available from: https://www.iowadnr.gov/environmental-protection/water-quality/water-monitoring/streams. Retrieved 10 Oct 2023.
67. Water Quality Data. 2023. Low color river auth. Available from: https://www.sariverauthority.org/whats-new/blog/using-data-management-analysis-improved-river-health. Retrieved 10 Oct 2023.
68. Dila DK, Corsi SR, Lenaker PL, Baldwin AK, Bootsma MJ, McLellan SL. 2018. Patterns of host-associated fecal indicators driven by hydrology, precipitation, and land use attributes in great lakes watersheds. Environ Sci Technol 52:11500–11509. https://doi.org/10.1021/acs.est.8b01945
69. Templar HA, Dila DK, Bootsma MJ, Corsi SR, McLellan SL. 2016. Quantification of human-associated fecal indicators reveal sewage from urban watersheds as a source of pollution to Lake Michigan. Water Res 100:556–567. https://doi.org/10.1016/j.watres.2016.05.056
70. Milwaukee Riverkeeper. 2022. Baseline water quality. Available from: https://milwaukeeriverkeeper.org/baseline-water-quality/. Retrieved 10 Oct 2023.
71. MountainTrue. 2022. Monitoring and waste programs. Available from: https://mountaintrue.org/waters/monitoring-and-waste-programs/. Retrieved 10 Oct 2023.
72. Nashwaak Watershed Water Quality Data. 2023. Available from: https://www.nashwaakwatershed.ca/watershed-monitoring/. Retrieved 10 Oct 2023.
73. National Water Quality Monitoring Council. 2022. National water quality portal. Available from: https://www.waterqualitydata.us/. Retrieved 10 Oct 2023.
74. McConaghy J. 2020. Oklahoma weather effects on *E. coli* in surface water and produce safety. IAFP
75. Bacteria data Oklahoma water Surv Univ Oklahoma. 2019. Available from: https://www.ou.edu/okh2o/monitoring/data-portal/bacteria-data. Retrieved 10 Oct 2023.
76. Pearl Riverkeeper Water Testing Results. 2023. Pearl Riverkeeper. Available from: https://www.pearlriverkeeper.com/water-testing-results.html. Retrieved 10 Oct 2023.
77. Fuhrmeister ER, Voth-Gaeddert LE, Metilda A, Tai A, Batorsky RE, Veeraraghavan B, Ward HD, Kang G, Pickering AJ. 2021. Surveillance of potential pathogens and antibiotic resistance in wastewater and surface water from Boston, USA and Vellore, India using long-read metagenomic sequencing. medRxiv. https://doi.org/10.1101/2021.04.22.21255864
78. Corrigan JA, Butkus SR, Ferris ME, Roberts JC. 2021. Microbial source tracking approach to investigate fecal waste at the strawberry creek watershed and clam beach. Int J Environ Res Public Health 18:6901. https://doi.org/10.3390/ijerph18136901
79. Fernández-Baca CP, Spirito CM, Bae JS, Szegletes ZM, Barott N, Sausele DJ, Brooks YM, Weller DL, Richardson RE. 2021. Rapid qPCR-based water quality monitoring in New York State recreational waters. Front Water 3:711477. https://doi.org/10.3389/frwa.2021.711477
80. Gu G, Strawn LK, Zheng J, Reed EA, Rideout SL. 2019. Diversity and dynamics of *Salmonella enterica* in water sources, poultry litters, and field soils amended with poultry litter in a major agricultural area of Virginia. Front Microbiol 10:2868. https://doi.org/10.3389/fmicb.2019.02868
81. Rock C, Gerba C, Bright K. 2012. Assessment of *E. coli* as an indicator of microbial quality or irrigation water use for produce
82. SCIWC. 2022. St Croix river watershed monitoring data. Available from: https://waterdata.usgs.gov/monitoring-location/01021060/. Retrieved 10 Oct 2023.
83. Shrestha A, Kelty CA, Sivaganesan M, Shanks OC, Dorevitch S. 2020. Fecal pollution source characterization at non-point source impacted beaches under dry and wet weather conditions. Water Res 182:116014. https://doi.org/10.1016/j.watres.2020.116014
84. Smith Mountain Lake Water Quality Monitoring Program. 2023. Available from: https://smlassociation.org/water-quality-monitoring/. Retrieved 10 Oct 2023.
85. Smith Mountain Lake Association. 2023. *Water quality monitoring*. Available from: https://smlassociation.org/water-quality-monitoring/. Retrieved 10 Oct 2023.
86. Clemson University. 2023. The South Carolina adopt-a-stream program. Available from: https://scdhec.gov/environment/your-water-coast/adopt-stream-program. Retrieved 10 Oct 2023.
87. Water Testing. 2023. Spa creek conserv. Available from: https://spacreek.net/about-us/spa-creek-water-testing/. Retrieved 10 Oct 2023.
88. Murphy CM, Weller DL, Ovissipour R, Boyer R, Strawn LK. 2023. Spatial versus non-spatial variance in fecal indicator bacteria differs within and between ponds. J Food Prot 86:100045. https://doi.org/10.1016/j.jfp.2023.100045
89. Surface Water Ambient Monitoring Program (SWAMP). 2023. Cent val reg water qual control board. Available from: https://www.waterboards.ca.gov/centralvalley/water_issues/swamp/. Retrieved 10 Oct 2023.
90. Community Scientist Water Quality Efforts on Thornton Creek. 2023. Thornt creek alliance. Available from: https://thorntoncreekalliance.info/project_tag/water-quality/
91. Givens CE, Kolpin DW, Borchardt MA, Duris JW, Moorman TB, Spencer SK. 2016. Detection of hepatitis E virus and other livestock-related pathogens in Iowa streams. Sci Total Environ 566–567:1042–1051. https://doi.org/10.1016/j.scitotenv.2016.05.123
92. Bradshaw J, Snyder B, Spidle D, Sidle R, Sullivan K, Molina M. 2017. Fecal indicator concentrations, and water quality physico-chemical parameters.
93. Bradshaw JK, Snyder B, Spidle D, Sidle RC, Sullivan K, Molina M. 2021. Sediment and fecal indicator bacteria loading in a mixed land use watershed: contributions from suspended sediment and bedload transport. J Environ Qual 50:598–611. https://doi.org/10.1002/jeq2.20166
94. Bradshaw JK, Snyder BJ, Oladeinde A, Spidle D, Berrang ME, Meinersmann RJ, Oakley B, Sidle RC, Sullivan K, Molina M. 2016. Characterizing relationships among fecal indicator bacteria, microbial source tracking markers, and associated waterborne pathogen occurrence in stream water and sediments in a mixed land use watershed. Water Res 101:498–509. https://doi.org/10.1016/j.watres.2016.05.014

95. Sowah RA, Bradshaw K, Snyder B, Spidle D, Molina M. 2020. Evaluation of the soil and water assessment tool (SWAT) for simulating *E. coli* concentrations at the watershed-scale. Sci Total Environ 746:140669. https://doi.org/10.1016/j.scitotenv.2020.140669
96. US environmental protection agency. 2022. Data from the national aquatic resource surveys
97. Fogarty LR, Duris JW, Crowley SL, Hardigan N. 2007. Antibiotic-resistant fecal bacteria, antibiotics, and mercury in surface waters of Oakland County, Michigan, 2005-2006. US Department of the Interior, US Geological Survey.
98. Duris J, Reif A, Olson L, Johnson H. 2009. Pathogenic bacteria and microbial-source tracking markers in Brandywine Creek Basin. In Pennsylvania and Delaware. Vol. 10. US Geological Survey.
99. Rowe J, Smalling K, Pearl C, Givens C, Sanders L, Iwanowicz L, Adams M. 2019. Nutrients, estrogenicity, and fecal indicators in surface water collected from wetlands in the klamath marsh national wildlife refuge. US Geological Survey.
100. McKee AM, Bradley PM, Shelley D, McCarthy S, Molina M. 2021. Feral swine as sources of fecal contamination in recreational waters. Sci Rep 11:1–13. https://doi.org/10.1038/s41598-021-83798-6
101. McKee BA, Molina M, Cyterski M, Couch A. 2020. Microbial source tracking (MST) in chattahoochee river national recreation area: seasonal and precipitation trends in MST marker concentrations, and associations with *E. coli* levels, pathogenic marker presence, and land use. Water Res 171:115435. https://doi.org/10.1016/j.watres.2019.115435
102. Duris JW, Haack SK, Fogarty LR. 2009. Gene and antigen markers of shiga - toxin producing *E. coli* from michigan and indiana river water: occurrence and relation to recreational water quality criteria. J Environ Qual 38:1878–1886. https://doi.org/10.2134/jeq2008.0225
103. Lenaker PL, Corsi SR, McLellan SL, Borchardt MA, Olds HT, Dila DK, Spencer SK, Baldwin AK. 2018. Human-associated indicator bacteria and human-specific viruses in surface water: a spatial assessment with implications on fate and transport. Environ Sci Technol 52:12162–12171. https://doi.org/10.1021/acs.est.8b03481
104. Lenaker P, Corsi S, Templar H, Borchardt M, McLellan S, Spencer S, Dila D, Baldwin A. 2017. Human-associated indicator bacteria and human specific virus loads, sample volumes, and drainage areas for six menomonee river watershed sampling locations. US Geological Survey.
105. Hubbard L, Givens C. 2020. Microbial and chemical contaminant occurrence and concentration in groundwater and surface water proximal to large-scale poultry facilities and poultry litter. US Geological Survey.
106. Byappanahalli MN, Nevers MB, Shively D, Nakatsu CH, Kinzelman JL, Phanikumar MS. 2021. Influence of filter pore size on composition and relative abundance of bacterial communities and select host-specific MST markers in coastal waters of Southern Lake Michigan. Front Microbiol 12:665664. https://doi.org/10.3389/fmicb.2021.665664
107. Byappanahalli M, Nevers M, Shively D. 2021. Influence of filter pore size on microbial communities and microbial source tracking (MST) markers on water in Racine Wisconsin; Chicago, Illinois; East Chicago, Indiana, 2015-2017. US Geological Survey.
108. Wilson J, Bussell A, Scheider M. 2019. Nutrients, *Escherichia coli*, and microbial source tracking markers in surface-water and known-source samples collected in salado creek and its watershed near Salado Texas, 2018
109. Nevers MB, Byappanahalli MN, Shively D, Buszka PM, Jackson PR, Phanikumar MS. 2018. Identifying and eliminating sources of recreational water quality degradation along an urban coast. J Env Quality 47:1042–1050. https://doi.org/10.2134/jeq2017.11.0461
110. Smith K. Water quality data from the providence water supply board for tributary streams to the scituate reservoir, water year 2015. US Geological Survey.
111. Rice KC, Monti MM, Ettinger MR. 2005. Water-quality data from ground- and surface-water sites near concentrated animal feeding operations (CAFOs) and non-CAFOs in the Shenandoah Valley and Eastern Shore of Virginia, January-February, 2004. US Geological Survey.
112. Crain AS, Cherry MA, Williamson TN, Bunch AR. 2017. Multiple-source 1045 tracking: investigating sources of pathogens, nutrients, and sediment in the upper little 1046 river Basin, Kentucky, water years 2013–14. US Geological Survey.
113. Murphy JC, Farmer J, Layton A A. 2016. Water-quality data and *Escherichia coli* 1048 predictions for selected karst catchments of the upper duck river watershed in central 1049 Tennessee, 2007–10. US Geological Survey.
114. Farmer J, Layton A, Murphy J. 2016. Water-quality datasets and *E. coli* predictions for selected streams in the upper duck river watershed. US Geological Survey.
115. Fargen C. 2019. Phytoplankton, microbial source tracking, and metagenomics data for evaluation of restoration efforts at Urban Beaches on Southern and Western Lake Michigan, 2016-2018. US Geological Survey.
116. Fargen C, Cole T. 2019. Phytoplankton, microbial source tracking, and metagenomics data for evaluation of restoration efforts at Urban Beaches on Southern and Western Lake Michigan, 2016-2018. US Geological Survey.
117. Baldwin AK, Graczyk DJ, Robertson DM, Saad DA, Magruder C. 2012. Use of real-time monitoring to predict concentrations of select constituents in the menomonee river drainage Basin, Southeast Wisconsin, 2008-9. US Geological Survey.
118. Francy DS, Stelzer EA, Duris JW, Brady AMG, Harrison JH, Johnson HE, Ware MW. 2013. Predictive models for *Escherichia coli* concentrations at inland lake beaches and relationship of model variables to pathogen detection. Appl Environ Microbiol 79:1676–1688. https://doi.org/10.1128/AEM.02995-12
119. Hittle E. 2017. Data collected in support of the longshore water-current velocity and the potential for transport of contaminants pilot study in lake erie. US Geological Survey.
120. Flickinger A, Christensen E. 2018. *Escherichia coli* concentrations and continuous hydrologic and physical parameters at USGS streamgage sites on the little blue river and its tributaries in independence, MO. US Geological Survey.
121. Nevers MB, Byappanahalli MN, Shively D. 2018. Identify sources of high *E. coli* concentrations, beaches of southern lake michigan. US Geological Survey.
122. Duris JW, Beeler S. 2008. Fecal-indicator bacteria and *Escherichia coli* pathogen data collected near a novel sub-irrigation water-treatment system in lenawee county, p 13. US geological survey.
123. Holtschlag DJ, Shively D, Whitman RL, Haack SK, Fogarty LR. 2008. Environmental factors and flow paths related to *Escherichia coli* concentrations at two beaches on Lake St. Clair, Michigan, 2002–2005. US Geological Survey.
124. Coastal Carolina University. Waccamaw watershed academy. Available from: https://www.coastal.edu/wwa/. Retrieved 10 Oct 2023. Accessed October 10, 2023
125. Falardeau J, Johnson RP, Pagotto F, Wang S. 2017. Occurrence, characterization, and potential predictors of verotoxigenic *Escherichia coli*, *Listeria monocytogenes*, and *Salmonella* in surface water used for produce irrigation in the lower mainland of British Columbia, Canada. PLoS One 12:e0185437. https://doi.org/10.1371/journal.pone.0185437
126. Washington State Department of Agriculture and the Whatcom Conservation District. Surface water monitoring for fecal Coliform data. Available from: https://www.whatcomcd.org/water-quality. Retrieved 10 Oct 2023. Accessed October 10, 2023
127. Green LTT, Herron EM. 2020. University of rhode island watershed watch program. Available from: https://web.uri.edu/watershedwatch/. Retrieved 10 Oct 2023.
128. University of Rhode Island Watershed Watch Analytical Laboratory. 2020. Quality assurance project plan. Kingston, Rhode Island. Available from: https://web.uri.edu/watershedwatch/resources/quality-assurance/. Retrieved 10 Oct 2023.
129. Water quality data report. 2017. Available from: https://dep.wv.gov/WWE/getinvolved/sos/Pages/Reports.aspx. Retrieved 10 Oct 2023.
130. Antaki EM, Vellidis G, Harris C, Aminabadi P, Levy K, Jay-Russell MT. 2016. Low concentration of *Salmonella enterica* and generic *Escherichia coli* in farm ponds and irrigation distribution systems used for mixed produce production in Southern Georgia. Foodborne Pathog Dis 13:551–558. https://doi.org/10.1089/fpd.2016.2117
131. Wickham H. 2016. Ggplot2. In Data analysis. Springer, Cham.
132. Wickham H. 2014. Tidy data. J Stat Soft 59:23. https://doi.org/10.18637/jss.v059.i10
133. Rashid SM, McCusker JP, Pinheiro P, Bax MP, Santos H, Stingone JA, Das AK, McGuinness DL. 2020. The semantic data dictionary - an approach for describing and Annotating data. Data Intell 2:443–486. https://doi.org/10.1162/dint_a_00058
134. Uhrowczik PP. 1973. Data dictionary/directories. IBM Syst. J 12:332–350. https://doi.org/10.1147/sj.124.0332

135. Wertz CJ. 1989. The data dictionary: concepts and uses. QED Information Sciences, Inc.
136. Center for Produce Safety. 2017. Report on agricultural water testing methods. Available from: https://www.centerforproducesafety.org/amass/documents/document/417/Water%20Colloquium%20Report%20-%20Final%20Release%208.2.17.pdf. Retrieved 10 Oct 2023.
137. Food and Drug Administration. 2021. Equivalent testing methodology for agricultural water. Available from: https://www.fda.gov/food/laboratory-methods-food/equivalent-testing-methodology-agricultural-water. Retrieved 10 Oct 2023.
138. Harris LJ, Bender J, Bihn EA, Blessington T, Danyluk MD, Delaquis P, Goodridge L, Ibekwe AM, Ilic S, Kniel K, Lejeune JT, Schaffner DW, Stoeckel D, Suslow TV. 2012. A framework for developing research protocols for evaluation of microbial hazards and controls during production that pertain to the quality of agricultural water contacting fresh produce that may be consumed raw. J Food Prot 75:2251–2273. https://doi.org/10.4315/0362-028X.JFP-12-252
139. US Environmental Protection Agency. 2012. Recreational water quality criteria. United States environmental protection agency. Available from: https://www.epa.gov/wqc/recreational-water-quality-criteria-and-methods#rec1. Retrieved 5 Aug 2023.
140. US Environmental Protection Agency. 2011. Sampling and consideration of variability (temporal and spatial) for monitoring of recreational waters. Available from: https://www.epa.gov/sites/default/files/2015-11/documents/sampling-consideration-recreational-waters.pdf. Retrieved 5 Aug 2023.
141. Federal Geographic Data Committee. 1998. Geospatial positioning accuracy standards part 3: National standard for spatial data accuracy. Available from: https://www.fgdc.gov/standards/projects/accuracy/part3/chapter3. Retrieved 5 Aug 2023.
142. US Environmental Protection Agency. 2023. Geospatial policies and standards. Available from: https://www.epa.gov/geospatial/geospatial-policies-and-standards. Retrieved 5 Aug 2023.
143. Jones KA, Niknami LS, Buto SG, Decker D. 2022. Federal standards and procedures for the national watershed boundary dataset (wbd): chapter 3 of section a, federal standards, book 11, collection and delineation of spatial data. US Geological Survey.
144. Snelder T, Fraser C, Larned S, Whitehead A. 2021. Guidance for the analysis of temporal trends in environmental data. NIWA Client Report. Christchurch, New Zealand NIWA
145. US Environmental Protection Agency. 2017. Water quality standards handbook. chapter 3: water quality criteria. Available from: https://www.epa.gov/sites/default/files/2014-10/documents/handbook-chapter3.pdf. Retrieved 5 Aug 2023.
146. US Geological Survey2008. US geological survey techniques of water-resources investigations, chapter A6. field measurements. Survey USG, Reston, VA.
147. US Environmental Protection Agency. 2014. Best practices for continuous monitoring of temperature and flow in Wadeable streams, . In US. environmental protection agency, office of research and development. National Center for Environmental Assessment, Washington, DC.
148. World Meteorological Organization. 2022. Standards and recommended practices. Available from: https://public.wmo.int/en/resources/standards-technical-regulations. Retrieved 5 Aug 2023.
149. US Department of Health and Human Services Office of Minority Health Resource Center. 2021. Explanation of data standards for race, ethnicity, sex, primary language, and disability. Available from: https://minorityhealth.hhs.gov/omh/browse.aspx?lvl=3&lvlid=54. Retrieved 05 May 2023.
150. US Office of Management and Budget. 2013. Uniform administrative requirements, cost principles, and audit requirements for federal awards. Available from: https://www.ecfr.gov/current/title-2/subtitle-A/chapter-II/part-200. Retrieved 05 May 2023.
151. Centers for Disease Control and Prevention. 2014. Standards to facilitate data sharing and use of surveillance data for public health action. Available from: https://www.cdc.gov/nchhstp/programintegration/sc-standards.htm. Retrieved 05 May 2023.
152. Cooley MB, Jay-Russell M, Atwill ER, Carychao D, Nguyen K, Quiñones B, Patel R, Walker S, Swimley M, Pierre-Jerome E, Gordus AG, Mandrell RE. 2013. Development of a robust method for isolation of shiga toxin-positive *Escherichia coli* (STEC) from fecal, plant, soil and water samples from a leafy greens production region in California. PLoS One 8:e65716. https://doi.org/10.1371/journal.pone.0065716
153. Hussein HS, Bollinger LM. 2008. Influence of selective media on successful detection of shiga toxin–producing *Escherichia coli* in food, fecal, and environmental samples. Foodborne Pathog Dis 5:227–244. https://doi.org/10.1089/fpd.2008.0081
154. Lejeune JT, Besser TE, Rice DH, Hancock DD. 2001. Methods for the isolation of water - borne *Escherichia coli* O157. Lett Appl Microbiol 32:316–320. https://doi.org/10.1046/j.1472-765x.2001.00905.x
155. uyttendaele m, van boxstael s, debevere j, , , , , , . 1999. pcr assay for detection of the *e. coli* o157: h7 eae-gene and effect of the sample preparation method on pcr detection of heat-killed *e. coli* o157: h7 in ground beef. Int J Food Microbiol 52:85–95. https://doi.org/10.1016/s0168-1605(99)00132-4
156. Ngwa GA, Schop R, Weir S, León-Velarde CG, Odumeru JA. 2013. Detection and enumeration of *E. coli* O157: H7 in water samples by culture and molecular methods. J Microbiol Methods 92:164–172. https://doi.org/10.1016/j.mimet.2012.11.018
157. Kim J-Y, Kim S-H, Kwon N-H, Bae W-K, Lim J-Y, Koo H-C, Kim J-M, Noh K-M, Jung W-K, Park K-T, Park Y-H. 2005. Isolation and identification of *Escherichia coli* O157:H7 using different detection methods and molecular determination by multiplex PCR and RAPD. J Vet Sci 6:7–19.
158. Kim HJ, Park SH, Lee TH, Nahm BH, Chung YH, Seo KH, Kim HY. 2006. Identification of *Salmonella* enterica serovar Typhimurium using specific PCR primers obtained by comparative genomics in *Salmonella* serovars. J Food Prot 69:1653–1661. https://doi.org/10.4315/0362-028x-69.7.1653
159. Kreitlow A, Becker A, Schotte U, Malorny B, Plötz M, Abdulmawjood A. 2021. Evaluation of different target genes for the detection of *Salmonella* sp. by loop - mediated isothermal amplification. Lett Appl Microbiol 72:420–426. https://doi.org/10.1111/lam.13409
160. Perelle S, Dilasser F, Grout J, Fach P. 2004. Detection by 5′-nuclease PCR of shiga-toxin producing *Escherichia coli* O26, O55, O91, O103, O111, O113, O145 and O157: H7, associated with the world's most frequent clinical cases. Mol Cell Probes 18:185–192. https://doi.org/10.1016/j.mcp.2003.12.004
161. Moriñigo MA, Borrego JJ, Romero P. 1986. Comparative study of different methods for detection and enumeration of *Salmonella* spp. in natural waters. J Appl Bacteriol 61:169–176. https://doi.org/10.1111/j.1365-2672.1986.tb04272.x
162. Worakhunpiset S, Tharnpoophasiam P. 2009. Influence of enrichment broths on multiplex PCR detection of total coliform bacteria, *Escherichia coli* and *Clostridium perfringens*, in spiked water samples. Southeast Asian J Trop Med Public Health 40:795–800.
163. Truchado P, Lopez-Galvez F, Gil MI, Pedrero-Salcedo F, Alarcón JJ, Allende A. 2016. Suitability of different *Escherichia coli* enumeration techniques to assess the microbial quality of different irrigation water sources. Food Microbiol 58:29–35. https://doi.org/10.1016/j.fm.2016.03.006
164. Duris JW, Reif AG, Krouse DA, Isaacs NM. 2013. Factors related to occurrence and distribution of selected bacterial and protozoan pathogens in Pennsylvania streams. Water Res 47:300–314. https://doi.org/10.1016/j.watres.2012.10.006
165. Pachepsky Y, Kierzewski R, Stocker M, Sellner K, Mulbry W, Lee H, Kim M. 2018. Temporal stability of *Escherichia coli* concentrations in waters of two irrigation ponds in Maryland. Appl Environ Microbiol 84:e01876-17. https://doi.org/10.1128/AEM.01876-17
166. Kim S, Paul M, Negahban-Azar M, Micallef SA, Rosenberg Goldstein RE, Hashem F, Parveen S, Sapkota A, Kniel K, Sapkota AR, Pachepsky Y, Sharma M. 2022. Persistent spatial patterns of *Listeria monocytogenes* and *Salmonella* enterica concentrations in surface waters: empirical orthogonal function analysis of data from Maryland. Applied Sciences 12:7526. https://doi.org/10.3390/app12157526
167. Badgley BD, Steele MK, Cappellin C, Burger J, Jian J, Neher TP, Orentas M, Wagner R. 2019. Fecal indicator dynamics at the watershed scale: variable relationships with land use, season, and water chemistry. Sci Total Environ 697:134113. https://doi.org/10.1016/j.scitotenv.2019.134113
168. Dean K, Mitchell J. 2022. Identifying water quality and environmental factors that influence indicator and pathogen decay in natural surface waters. Water Res 211:118051. https://doi.org/10.1016/j.watres.2022.118051
169. Gu G, Luo Z, Cevallos-Cevallos JM, Adams P, Vellidis G, Wright A, van Bruggen AHC. 2013. Factors affecting the occurrence of *Escherichia coli* O157 contamination in irrigation ponds on produce farms in the

Suwannee river watershed. Can J Microbiol 59:175–182. https://doi.org/10.1139/cjm-2012-0599
170. Haley BJ, Cole DJ, Lipp EK. 2009. Distribution, diversity, and seasonality of waterborne *Salmonella* in a rural watershed. Appl Environ Microbiol 75:1248–1255. https://doi.org/10.1128/AEM.01648-08
171. Wilkes G, Edge TA, Gannon VPJ, Jokinen C, Lyautey E, Neumann NF, Ruecker N, Scott A, Sunohara M, Topp E, Lapen DR. 2011. Associations among pathogenic bacteria, parasites, and environmental and land use factors in multiple mixed-use watersheds. Water Res 45:5807–5825. https://doi.org/10.1016/j.watres.2011.06.021
172. Zhang X, Zhi X, Chen L, Shen Z. 2020. Spatiotemporal variability and key influencing factors of river fecal coliform within a typical complex watershed. Water Research 178:115835. https://doi.org/10.1016/j.watres.2020.115835
173. United States Department of Agriculture. 2019. 20017 census of Agriculture watersheds. Available from: https://www.nass.usda.gov/Publications/AgCensus/2017/Online_Resources/Watersheds/wtrsheds.pdf. Retrieved 5 Aug 2023.
174. Bates D, Maechler M, Bolker B, Walker S. 2014. lme4: linear mixed-effects models using eigen and S4. R package version 1.1-7
175. Robette N. 2023. Moreparty: a toolbox for conditional inference trees and random forests. Available from: https://cran.r-project.org/web/packages/moreparty/moreparty.pdf. Retrieved 04 May 2023.
176. Dixon P. 2003. VEGAN, a package of R functions for community ecology. J Vegetation Sci 14:927–930. https://doi.org/10.1111/j.1654-1103.2003.tb02228.x
177. Hothorn T, Bretz F, Westfall P. 2008. Simultaneous inference in general parametric models. Biom J 50:346–363. https://doi.org/10.1002/bimj.200810425