POLLUTANT SOURCE AND RAINFALL EFFECTS ON *E. COLI* LEVELS IN 247 MICHIGAN WATERSHEDS

by

Yifanzi Zhu

A thesis submitted

in partial fulfillment of the requirements

for the degree of

Master of Landscape Architecture

Master of Science

(Environment and Sustainability)

in the University of Michigan

04 2026

Thesis Committee:

Professor Allen Burton, Chair

Associate Professor Stan Jones

Assistant Professor Runzi Wang

ABSTRACT

Microbial contamination represents a critical threat to freshwater ecosystems and public health, particularly in regions where agricultural activities, urban infrastructure, and hydrologic variability converge. Despite extensive research on individual pollution sources, limited understanding remains regarding how these factors collectively and spatially regulate Escherichia coli (*E. coli*) dynamics and how rainfall influences their transport across heterogeneous watersheds. This study examined the spatiotemporal distribution of stream *E. coli* concentrations and loadings across 247 watersheds in Michigan from 2018 to 2022, quantified the relative contributions of multiple pollution sources and watershed-scale environmental variables, and assessed the role of rainfall in shaping microbial loading. The results reveal substantial spatial heterogeneity in *E. coli* contamination, with persistent hotspots concentrated in southern and southeastern regions and seasonal peaks observed in late summer and early autumn. Anthropogenic pressures, particularly point-source discharges from municipal separate storm sewer systems and broader land use associated with agriculture, were linked to higher *E. coli* concentrations. Soil characteristics such as sand and clay content, variability in topography, and watershed area further influence microbial loading by affecting infiltration, runoff, and contaminant transport. Rainfall functioned as a critical modulator that enhanced microbial mobilization and delivery, although the magnitude of its effect varied among watersheds depending on land use, infrastructure connectivity, and physical landscape features. These findings highlight the complex interactions among hydrologic processes, landscape characteristics, and human activities in shaping *E. coli* contamination and emphasize the need for targeted watershed management strategies, including improvements to stormwater systems, optimization of agricultural practices, and soil conservation measures to protect freshwater quality under changing precipitation regimes.

TABLE OF CONTENTS

Abstract ..... **ii**

Chapter 1 Introduction **1**

Chapter 2 Materials and methods **3**

2.1. Study Site 3

2.2. Data sources, quality screening, and preprocessing..... 4

2.2.1. *E. coli* concentration and loading 4

2.2.2. Potential Drivers of *E. coli* Levels 4

2.2.3. Data quality screening 7

2.3. Data Analysis 7

2.3.1. Variable selection 7

2.3.2. Linear Mixed-Effects Regression 8

Chapter 3 Results **9**

3.1. *E. coli* Concentrations, Distributions, and Seasonal Patterns..... 9

3.2. The relationship between rainfall and *E. coli* relationships under varying land use and hydrological conditions..... 10

3.3. Influencing factors of *E. coli* concentration 12

3.4 The interaction effects between pollutant source and rainfall effects on *E. coli* 15

Chapter 4 Discussion **17**

4.1. Pollutant sources and watershed characteristics affecting E. coli..... 17

4.2. Rainfall effects on E. coli transport under different watershed conditions..... 17

4.2.1. Implications for watersheds under lower rainfall conditions..... 18

4.2.2. Implications for watersheds under heavy rainfall conditions..... 19

4.3. Limitations and future research..... 19

Chapter 5 Conclusion **20**

Bibliography **21**

Appendices **25**

Chapter 1 Introduction

Microbial contamination of surface waters threatens both freshwater ecosystems and human health, particularly in regions where agricultural practices, urban development, and hydrologic variability intersect (Wilkes et al., 2009). In the Great Lakes region, fecal contamination remains a widespread water-quality concern, and *Escherichia coli* (*E. coli*) is routinely used as a primary microbial indicator for evaluating recreational water quality and potential public health risk. (Verhougstraete et al., 2015; Edge & Hill, 2007). Previous studies have shown that *E. coli* contamination is influenced by multiple factors, including land use, pollution sources, hydrologic conditions, and precipitation (Hong et al., 2021; Dila et al., 2018; Sinclair et al., 2009). However, these relationships often vary across systems, and much of the available evidence comes from individual watersheds or limited sets of waterways, making broader regional patterns less clear. This knowledge gap is particularly important in the Midwest, where changing rainfall patterns may enhance the mobilization and transport of fecal contaminants from both urban and agricultural landscapes. (Patz et al., 2008; Liu et al., 2025). Within this context, Michigan offers an important setting for investigation, as its mix of intensive agriculture, expanding urban areas, and direct hydrologic connection to the Great Lakes make the region particularly vulnerable to rainfall-driven *E. coli* contamination (Verhougstraete et al., 2015; Nevers & Whitman, 2011; Scavia et al., 2014).

Escherichia coli (*E. coli*) is a common fecal indicator bacteria used to assess microbial contamination and public health risk in freshwater systems (Ishii & Sadowsky, 2008; Jang et al., 2017). In Michigan, point sources such as wastewater treatment plants, municipal storm sewer systems (MS4s), combined and sanitary sewer overflows (CSOs), industrial discharges (IndStorm), and concentrated animal feeding operations (CAFOs) are regulated under the National Pollutant Discharge Elimination System (NPDES) (EGLE, Statewide *E. coli* TMDL). Nonpoint sources, including failing septic systems (Septic), unregulated storm water (Groundwater), livestock manure applications (Manure), and wildlife, also contribute to widespread contamination (EGLE, *E. coli* Pollution and Solution Mapper; Pandey et al., 2014). These diverse sources, together with watershed characteristics such as land use, soil properties, slope, and impervious cover, produce uneven patterns of *E. coli* pollution across the state (Oliver et al., 2016; Garzio-Hadzick et al., 2010). Understanding the role of both regulated and diffuse sources within Michigan's watersheds is essential for identifying vulnerable areas and developing targeted management strategies (Cho et al., 2016; Oliver et al., 2016).

Precipitation is a key driver of *Escherichia coli* (*E. coli*) dynamics in freshwater systems because rainfall and snowmelt can mobilize fecal material from both regulated and diffuse sources into rivers, streams, and lakes (Boithias et al., 2016; Buckerfield et al., 2019). Heavy rainfall increases *E. coli* concentrations by enhancing surface runoff and triggering sewer overflows, leading to elevated microbial loads in receiving waters (Carlton et al., 2014). Frequent rainfall can have a dilution effect, where continuous flushing reduces pathogen accumulation in the environment (Levy et al., 2009). Dry periods followed by intense storms can increase *E. coli* loads by allowing pathogens to accumulate on land surfaces and then flush into waterbodies during runoff (Kraay et al., 2020; Maphanga et al., 2022). Studies in the United States have shown that waterborne disease outbreaks occur with extreme precipitation, indicating the health risks associated with

rainfall variability. (Curriero et al., 2001). In the Midwest, projected increases in rainfall intensity under climate change raise concerns about more frequent movement of microbial contaminants during storms. Existing studies have documented that precipitation, storm events, and watershed land use affect microbial water quality, but evidence remains fragmented across study scales and hydrologic frameworks (Wu et al., 2018; Flood et al., 2022). Many studies focus on single watersheds, individual storm events, or seasonal contrasts, whereas fewer studies compare how multiple watershed characteristics and contamination-related variables are associated with *E. coli* across rainfall classes in mixed-source Michigan watersheds.(Boithias et al., 2016; Buckerfield et al., 2019). This gap limits understanding of how hydrologic variability interacts with multiple sources of contamination in Michigan's watersheds.

Progress in understanding rainfall-driven *E. coli* contamination is limited by the data and methods available. Routine monitoring provides valuable long-term records for characterizing baseline conditions and broad temporal trends, but its sampling frequency and site density are often insufficient to resolve short-term storm responses and event-scale variability (Whitman & Nevers, 2004; Nevers & Whitman, 2011; U.S. EPA, 2012). Hydrologic and water quality models have been used to simulate contaminant transport, but they depend on detailed data for land use, sources, and hydrology that are often incomplete at the watershed scale (Cho et al., 2016; Oliver et al., 2016). Statistical analyses also tend to use global regression models, which assume uniform relationships across space and fail to reflect local differences in the drivers of *E. coli* (Fotheringham et al., 2017).

To capture the spatiotemporal dynamics of rainfall-driven *E. coli* contamination, this study employs a framework that combines statistical and spatial approaches. Weighted Regression on Time, Discharge, and Season (WRTDS) generates daily estimates of *E. coli* concentrations and loads, providing higher temporal resolution than routine monitoring data (Hirsch et al., 2010). Geographic Information System (GIS) methods delineate watersheds and map potential pollution sources, supporting the identification of contamination hotspots. These outputs are linked with regression models that relate rainfall, land use, point and nonpoint sources, and microbial outcomes. Scenario-based risk assessment defines rainfall and flood conditions, quantifies exceedance probabilities, and identifies high-risk watersheds. While each of these tools has been used in water quality studies, their integration to examine *E. coli* dynamics under extreme rainfall remains limited, offering a systematic way to evaluate interactions between hydrologic variability and multiple drivers across Michigan's watersheds.

To address these research gaps, this study focused on Michigan watersheds between 2018 and 2022 to (1) characterize the spatiotemporal distribution of stream *Escherichia coli* (*E. coli*) concentrations and loadings, (2) assess how multiple pollution sources (e.g., municipal separate storm sewer systems, manure application, septic systems, and agriculture) together with watershed-scale environmental factors shape *E. coli* variability, and (3) evaluate how rainfall conditions alter the probability of *E. coli* exceedance of recreational water-quality thresholds and modify the influence of watershed drivers. While this research is centered on Michigan, the framework is designed to provide insights that can be applied in other regions facing rainfall-driven microbial contamination.

Chapter 2 Materials and methods

2.1. Study site

This study was conducted across 247 monitoring stations in Michigan, U.S.A., covering watersheds that drain into inland rivers, lakes, and the Great Lakes. We initially delineated 275 candidate watersheds; to maintain comparable spatial scales, very large basins were excluded, yielding 247 study watersheds.The state receives about 850-1,000 mm of annual precipitation, shaped by continental weather systems and lake-effect dynamics (NOAA, 2022). Land cover is a mix of agriculture, forest, and urban development, creating variability in hydrologic response and fecal pollution sources.

Michigan provides a strong foundation for watershed-scale analysis because *E. coli* monitoring and pollution source inventories have been systematically compiled by the Department of Environment, Great Lakes, and Energy (EGLE). The Statewide *E. coli* Total Maximum Daily Load (TMDL) and supporting datasets document regulated point sources (e.g., wastewater treatment plants, storm sewer systems, animal feeding operations) and nonpoint sources (e.g., septic systems, manure application, stormwater) (EGLE, Statewide *E. coli* TMDL; EGLE, *E. coli* Pollution and Solution Mapper). From these 247 sites, 28 stations with sufficient *E. coli* measurements between 2018 and 2022 were selected for cross-validation, representing diverse watershed sizes, land uses, and pollution contexts (Fig. 1).

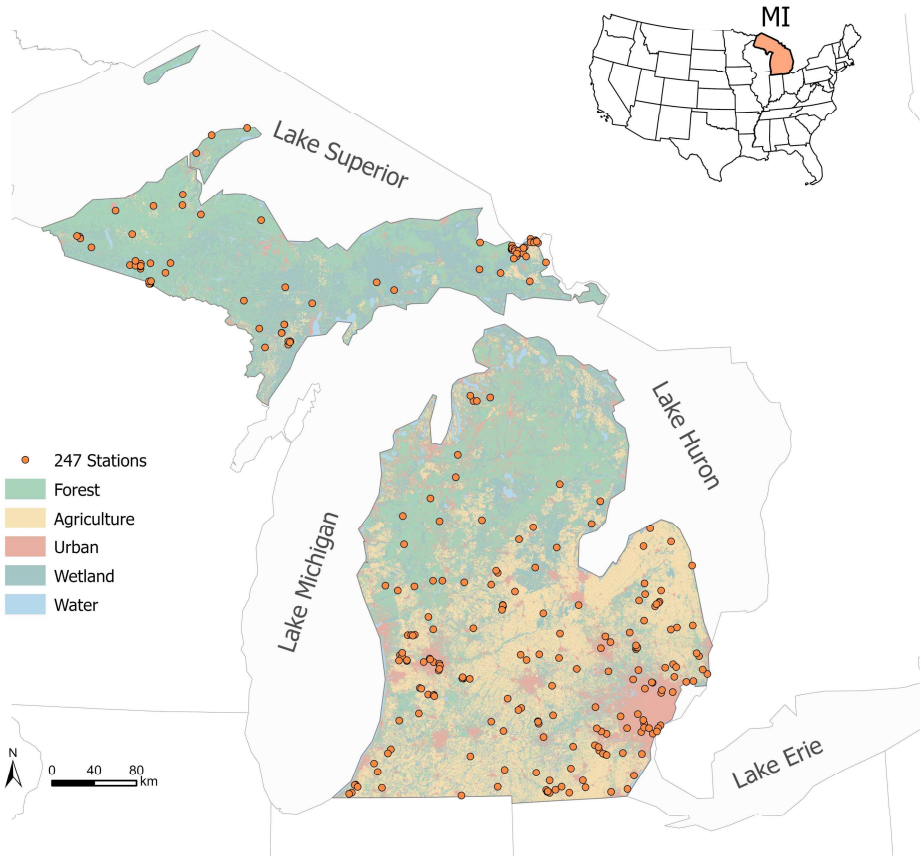


Figure 1. Distribution of *E. coli* monitoring stations 247 stations

2.2. Data sources, quality screening, and preprocessing

2.2.1. E. coli concentration and loading

We obtained *E. coli* concentration from the Michigan Department of Environment, Great Lakes, and Energy (EGLE). Outliers were removed only when attributable to measurement or data entry errors, and concentrations were log10-transformed prior to analysis. For stations with multiple measurements on the same day, we computed the geometric mean of these samples as daily average concentration. Then the geometric mean of daily average concentration was calculated for each watershed as monthly average concentration. Watersheds were delineated for each monitoring station using a 30 m digital elevation model from the U.S. Geological Survey (USGS) to derive land use and pollutant source variables.

To estimate *E. coli* loading, we paired these concentration records with daily streamflow data derived from the National Water Model (Cosgrove et al., 2024). Daily *E. coli* loading is calculated as the multiplication of *E. coli* concentration and discharge, and then aggregated monthly as well.

Throughout the analysis, we used the health-related *E. coli* threshold defined by EGLE. It specifies that *E. coli* levels must not exceed 300 CFU per 100 mL at any time in waters designated for total body contact recreation (Michigan Department of Environment, Great Lakes, and Energy, 2019). Additionally, a 30-day geometric mean standard of 130 CFU per 100 mL is applied to assess longer-term compliance with recreational water quality criteria.

2.2.2. Potential Drivers of E. coli Levels

- ***Pollutant Sources***

The point source data attributes include facility location, permit type, discharge category, and in some cases reported effluent quality or annual loads, which are compiled from EGLE and the U.S. EPA's National Pollutant Discharge Elimination System (NPDES) records. Permit records and compliance data are available from 2007-2024, with historical facilities archived in the MI Waters permitting system. Under NPDES, surface water discharges are regulated through individual permits, general permits, and permits by rules, which specify discharge limitations, monitoring requirements, and compliance schedules. Facilities include municipal separate storm sewer systems (MS4s), combined sewer overflows (CSOs), wastewater treatment plants (WWTPs), wastewater stabilization lagoons (WWSLs), concentrated animal feeding operations (CAFOs), industrial stormwater discharges, and biosolids land-application sites. In this study, we used MS4 facility counts, watershed area, mean slope, mean elevation, sand fraction, clay fraction, agricultural land percentage, and manure application area as predictors of *E. coli* concentration. Other source variables, including approved septage land application (SeptageApp), CAFOs, CSOs, septic system metrics (Septic), permitted groundwater discharges (Groundwater), industrial stormwater outfalls (IndStorm), and wastewater treatment plant variables (WWTP), were not included because their data quality and spatial completeness were insufficient for consistent watershed scale analysis. Additionally, when variables exhibited strong collinearity, we kept the variable with the stronger relationship to *E.coli* concentration.

Variables showing weak associations with *E. coli* concentration in preliminary bivariate regressions were also excluded.

Nonpoint source data were also assembled from EGLE's statewide inventories. Variables used in this study included septic system density, areas of tiled agricultural land, manure application areas, approved septage land-application sites, and permitted groundwater discharges. We overlaid all spatial datasets with subwatershed boundaries and summarized them as areas, counts, or densities within each subwatershed to match the hydrologic units used in our analysis. These variables represent diffuse fecal contamination sources that are transported to surface waters through precipitation-driven runoff, shallow subsurface flow, and erosion processes, particularly during high-flow events.

Table 1: Pollutant sources, land use, environmental, and rainfall variables used in this study

Category	Abbreviation	Variables	Description	Unit
Point source	MS4	MS4s	Municipal separate storm sewer systems within watershed	count (facilities)
	IndStorm	Industrial storm water (EPA)	Industrial stormwater outfalls within watershed	count (outfalls)
	WWTP	WWTPs & WWSLs	Wastewater treatment plants and stabilization lagoons	count (facilities)
	CAFOs	CAFOs	Concentrated animal feeding operations	count (facilities)
	CSOs	CSOs	Combined sewer overflow reaches intersecting watershed	km (line length)
Nonpoint source	Septic	Estimated septic / agricultural land / manure	Areas with septic systems or ag manure influence	% of watershed (area)
	SeptageApp	Approved septage land application	Permitted septage land-application polygons	% of watershed (area)
	Groundwater	Groundwater discharges	Permitted sanitary discharges to groundwater	count (points)

	Manure	Manure Land-Application	Areas receiving agricultural manure land-application	% of watershed (area)
Weather Variables	Tmean	Daily temperature (2018-2022)	Average temperature at the subwatershed level	°C
	Ppt	Daily rainfall (2018-2022)	Average daily precipitation at the subwatershed level	mm day^{-1}
Land Cover	LandCover	Land Cover	Land cover classes (urban, agriculture, forest, wetland, grassland) at the subwatershed level	% of watershed (by class)
Environmental Variables	Slope	Slope	Mean slope derived from DEM	% (or degrees)
	Elevation	Elevation	Mean elevation derived from DEM	m
	PopDensity	Population density	Residents per unit area	persons km^{-2}
	Soil	Soil type	Soil texture	categorical
	Area	Watershed area	Delineated drainage area for each monitoring station	km^2

- ***Land Cover and Environmental Variables***

Land cover and environmental variables used as control variables in this study, which are related with *E. coli* generation, transport, and attenuation processes. Land cover including the area percentage of urban, agriculture, forest, wetland, and grassland are included. Land cover including the area percentage of urban, agriculture, forest, wetland, and grassland are included. We focused on urban and agricultural cover because these land uses are more directly linked to human activities and are commonly associated with higher *E. coli* concentrations. Forest, wetland, and grassland cover were also included because they can influence hydrologic transport and provide conditions that promote retention and attenuation of microbial contaminants. We additionally include population density to represent human sources of *E. coli*. Topography including elevation and slope indicates runoff and erosion potential. Soils texture affects infiltration and temporary storage of *E. coli*. Watershed area is included as a control variable too because it represents drainage size and can influence flow volume, dilution capacity, and the integration of upstream sources affecting *E. coli* concentrations.

- ***Weather variables***

Weather (daily mean temperature and precipitation, aggregated for 2018-2022) drives mobilization, dilution, and decay; Daily precipitation and mean temperature data were obtained from the PRISM Daily Spatial Climate Dataset (AN81d). We used a threshold of 2.54 mm per day to define effective rainfall (Densmore et al., 2020). We define "Dry" conditions as less than 2.54 mm of rain on the sampling day and the three days before it. "Wet" conditions were identified if rainfall exceeded 2.54 mm on the sampling day or the day before it. All other instances were placed in an "Other" category.

We also investigated how short-term rain accumulation relates to *E. coli* levels on different land uses. To do this, we calculated total rainfall over various time frames from a single day up to ten days before each sample was taken. We then fit linear models to identify which time window best explained the changes in *E. coli* levels, and lastly plotted the relationship between *E. coli* levels and the accumulated rainfall in that time period for different land uses to understand varying rainfall-driven *E. coli* risks.

2.2.3. Data quality screening

After compiling the data, we performed a comprehensive check of data quality to ensure the reliability and comparability of datasets collected from different sources, spatial scales, and time periods. Specifically, we investigated five dimensions, including comparability, representativeness, completeness, precision, and documentation [cite]. Comparability confirms that all units and scales were consistent across the different watersheds. Representativeness checks if the data reflected the geographic areas and time periods we were studying. We measured completeness by looking at how much data was actually available, while precision focused on the recorded accuracy of the measurements. Finally, the documentation check ensured that all data came with clear quality-assurance records.

We used a three-tier grading system to rank each variable: high pass, adequate pass, or fail. These rankings were applied to every dimension. Variables that received a fail (F) rating in any dimension were excluded from further analysis. Variables with no fail ratings and with no more than two adequate pass (AP) ratings were retained as candidate explanatory variables.

2.3. Data Analysis

2.3.1. Variable selection

Bi-variate correlation analyses were first used to explore relationships among all variables. For regression, we selected important predictors of *E. coli* among all the potential drivers as input to the regression models (See Section 2.3.2) by three steps: (1) data quality checks to remove variables with insufficient quality or coverage; (2) tests for multicollinearity, in which we calculated Pearson correlation coefficients among the remaining variables and removed variables with an absolute correlation coefficient greater than 0.7, while retaining source

related variables when a choice was required; and (3) bivariate regressions between each remaining variable and *E. coli* concentration, and we kept variables with stronger relationships for the final models.

2.3.2. Linear Mixed-Effects Regression

We applied a linear mixed-effects model (LMM) to 1,252 observations. We used log-transformed values for concentration and loading as the dependent variables. We standardized predictors using z scores by subtracting the mean and dividing by the standard deviation, and we kept watershed area on its original scale for interpretation. We included a random intercept for watershed ID. The watershed random intercept accounts for unobserved heterogeneity among watersheds, such as differences in land use, infrastructure, or monitoring conditions, which may influence baseline *E. coli* levels. The LMM model is specified as:

$$y_{ij} = \beta_0 + \sum_{k=1}^{K} \beta_k x_{kij} + u_j + \epsilon_{ijt}$$

In this equation, y_{ijt} is the log-transformed *E. coli* concentration or loading for a specific observation. The term β_0 represents the fixed intercept or the baseline level of the data. Each β_k acts as a fixed effect coefficient for predictor k (e.g., MS4 or Manure). The variable x_{kij} is the specific value of that predictor at a given site. ϵ_{ij} represents the residual error for the model. The term u_j represents the random intercept associated with watershed j, capturing unexplained variability among watersheds.

This statewide dataset covers 247 monitoring stations, among which 28 sites have long-term records during 2018-2022. These stations were used in an additional analysis based on Weighted Regressions on Time, Discharge, and Season (WRTDS) to verify the robustness of relationships identified in the statewide LMM analysis.

To better understand the varying influence of predictors in different rainfall conditions, we also used rainfall categories as interaction terms in the LMM. We grouped daily rainfall into four levels including Dry (under 2.54 mm), Light (2.54-5 mm), Moderate (5-12.7 mm), and Heavy (12.7 mm or more). Rainfall was grouped into categories to represent different runoff conditions that affect pollutant transport. The categories of rainfall are used as an interaction term with the predictors, where dry conditions were kept for reference. The model form as:

$$y_{ij} = \beta_{0,c(ij)} + \sum_{k=1}^{K} \beta_{k,c(ij)} x_{kij} + u_j + \epsilon_{ijt}$$

In this formula, $c(ij)$ represents the rainfall class associated with observation i in watershed j. The term $\beta_{0,c(ij)}$ is the rainfall-class-specific intercept, while $\beta_{k,c(ij)}$ allows the coefficient of predictor k to vary across rainfall categories. The variable x_{kij} is the value of predictor k for observation i in watershed j. The term u_j is a random intercept that captures watershed-specific baseline differences. The term ϵ_{ijt} represents the residual error for the model.

Chapter 3 Results

3.1. *E. coli* Concentrations, Distributions, and Seasonal Patterns

High *E. coli* concentrations appear mostly in southern and southeastern Michigan (Figure 2). These hotspots align with areas of dense urban centers and the associated MS4 . Sites with geometric mean concentrations above 130 CFU/100 mL were concentrated in southeast Michigan, and additional stations above this threshold were distributed across south-central and southwestern Lower Michigan. Sites with concentrations above 300 CFU/100 mL were found in southeastern Michigan and in several southern Lower Michigan watersheds. These hotspots occurred in urban areas as well as in mixed agricultural and urban landscapes (Figure 9). Most stations in northern Michigan and the Upper Peninsula were below 130 CFU/100 mL. These areas have sparse populations and lower farming activity. Dense forest cover in northern Michigan may help reduce *E. coli* concentrations by increasing soil infiltration, reducing surface runoff, and filtering fecal contaminants before they reach streams.

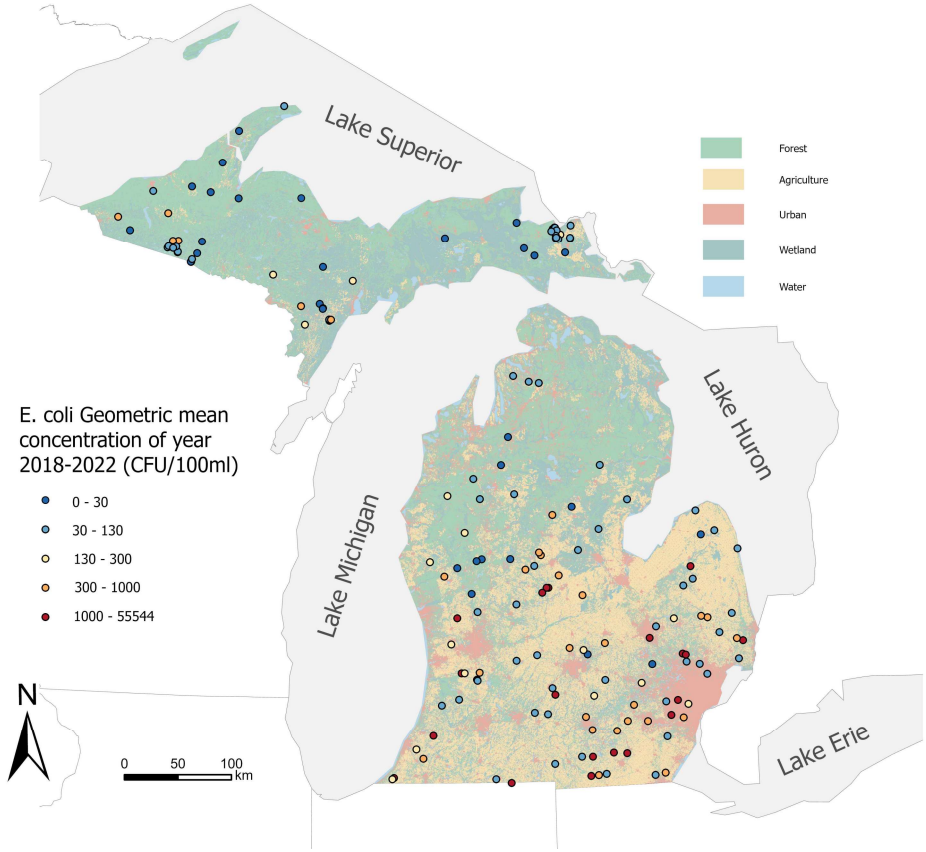


Figure 2. Average *E. coli* concentrations for all the monitoring stations in 2018–2012.

Temporal analysis reveals a clear seasonal pattern for *E. coli* levels (Figure 3). *E. coli* are usually lowest from February to April, when precipitation is lowest during the year. *E. coli* concentrations rise and stay stable in summer months, when they typically reach the highest values between August and October. These seasonal increases were observed in each year from

2018 to 2022 (Appendix **Figure 10**). Typically, average concentrations during late summer months approached or exceeded the commonly used recreational water quality threshold of 130 CFU/100 mL, whereas concentrations during colder months were below this threshold. Also, we observed no consistent long-term trend throughout the study period.

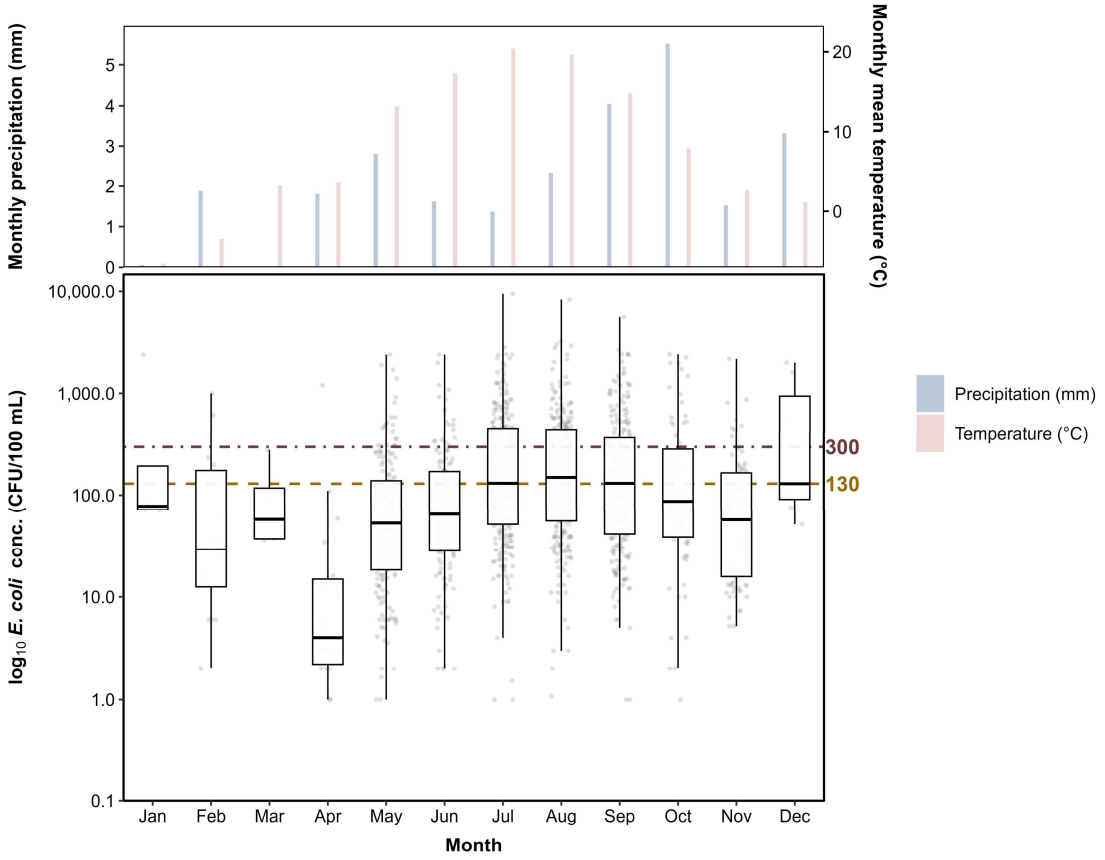


Figure 3. Monthly *E. coli* concentrations and precipitation, temperature distributions (2018–2022). The horizontal red and yellow dashed lines indicate the Michigan Water Quality Standards for total body contact recreation: the 30-day geometric mean threshold (130 CFU/100 mL) and the daily maximum threshold (300 CFU/100 mL).

3.2. The relationship between rainfall and *E. coli* relationships under varying land use and hydrological conditions

E. coli concentrations significantly differed between dry and wet periods. Wet conditions, which are defined as precipitation happening during the sample date or 1-day before it, were associated with higher and more variable *E. coli* levels than dry conditions, which have at least 3-day antecedent dry period (Figure 4a). Wet conditions also had several extremely high *E. coli* concentrations above 1,000 CFU/100 mL, whereas dry-period concentrations were generally much lower than 130 CFU/100 ml. The non-exceedance probability plots indicated that at the 130 CFU/100 mL threshold, the exceedance probability during wet periods was 43.3%,

compared with 17.5% during dry periods (Figure 4b). At the 300 CFU/100 mL threshold, the exceedance probability was 29.9% in wet periods and 5.3% in dry periods.

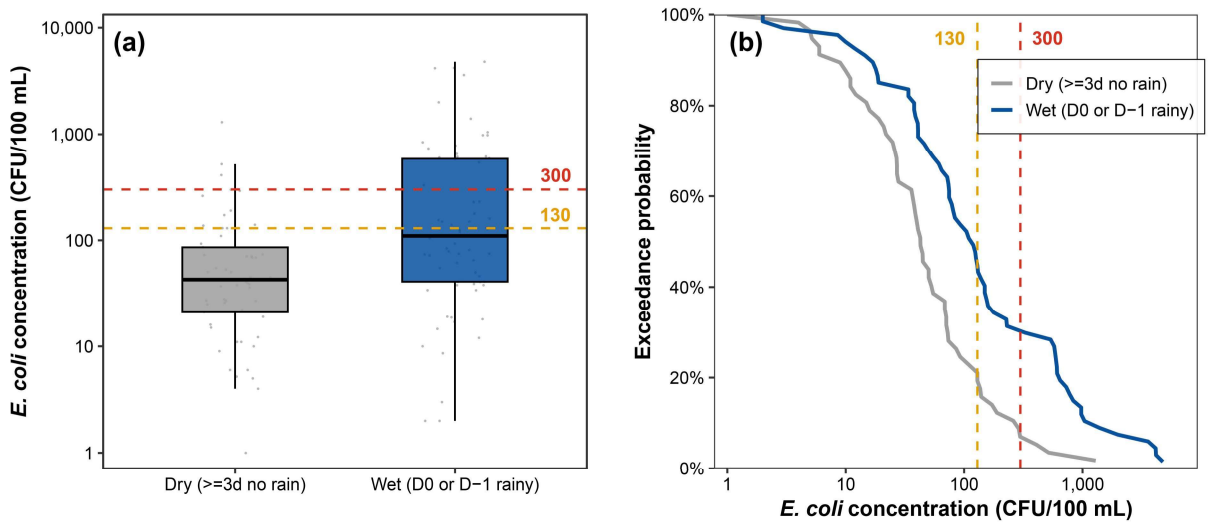


Figure 4. *E. coli* concentration under dry and wet conditions. (a) Boxplots of *E. coli* concentrations during dry (at least 3-day antecedent dry period) vs. wet (precipitation on or 1-day before sampling date) conditions. (b) Cumulative distribution functions (CDF) showing the non-exceedance probability for dry and wet conditions. Horizontal red dashed lines indicate Michigan Water Quality Standards for total body contact recreation: the 30-day geometric mean threshold (130 CFU/100 mL) and the daily maximum threshold (300 CFU/100 mL).

The influence of rainfall on *E. coli* concentration depends on the land use and hydrological contexts. While both urban and agricultural watersheds show positive nonlinear relationships between *E. coli* levels and precipitation, the specific nonlinear relationship varies. In urban watersheds where urban area percent is larger than 20%, *E. coli* concentrations rise initially when 2-day accumulated rainfall is smaller than 25mm. When 2-day accumulated rainfall is 10 mm, *E. coli* concentration starts to reach 130 CFU/100 ml in urban areas. Then, *E. coli* remains stable and starts to decrease as accumulated rainfall increases due to dilution (Figure 5a). Agricultural watersheds with agricultural area percent larger than 50% show a higher positive correlation between *E. coli* and accumulated rainfall before 2-day accumulated rainfall reaches 30mm. The average *E. coli* becomes higher than 300 CFU/100 ml at a smaller amount of rainfall with 2-day accumulated rainfall larger than 10mm. The dilution effect happens at a higher rainfall amount when 3-day accumulated rainfall larger than 35mm.

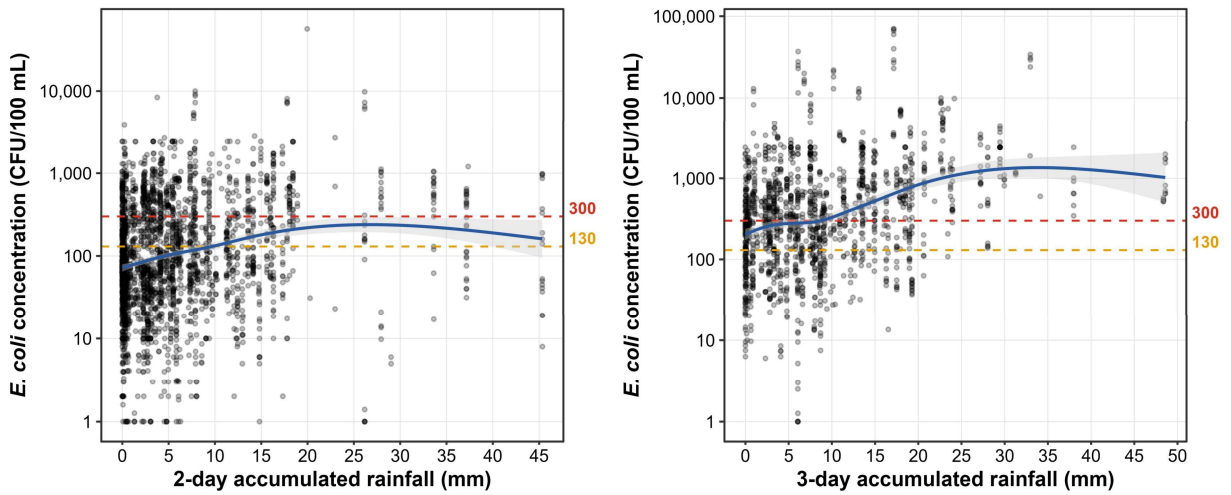


Figure 5. Relationships between accumulated rainfall and *E. coli* concentration in urban and agricultural watersheds. Points represent individual *E. coli* observations, and solid blue curves show fitted nonlinear trends from generalized additive models (GAMs) with shaded 95% confidence intervals. Horizontal dashed lines indicate the Michigan water-quality thresholds of 130 and 300 CFU/100 mL. For each land use group, rainfall accumulation windows from 1 to 10 days were evaluated, and the panel displays the window with the best model fit.

3.3. Influencing factors of *E. coli* concentration

Our exploratory bivariate analysis reveals that the relationships between *E. coli* concentration and predictor variables differed across different categories of predictors. Among point-source variables, MS4, industrial stormwater, WWTPs, CAFOs, and CSOs all showed positive associations with *E. coli* concentration. For nonpoint-source variables, septage application and groundwater showed positive relationships, while manure was only weakly associated with *E. coli* concentration and septic systems did not show a clear relationship. Among weather variables, mean temperature showed a relatively strong positive relationship with E. coli, whereas precipitation had a weaker positive association. For land cover, urban land, agriculture, and impervious surfaces were positively related to *E. coli* concentration, while forest cover showed a clear negative relationship; wetlands and grasslands also tended to be negatively associated, although their effects appeared weaker. Among environmental variables, slope and elevation were negatively associated with *E. coli* concentration. Soil texture variables showed mixed patterns: sand was negatively related to *E. coli* concentration, whereas silt and clay were positively associated. Watershed area showed little obvious relationship.

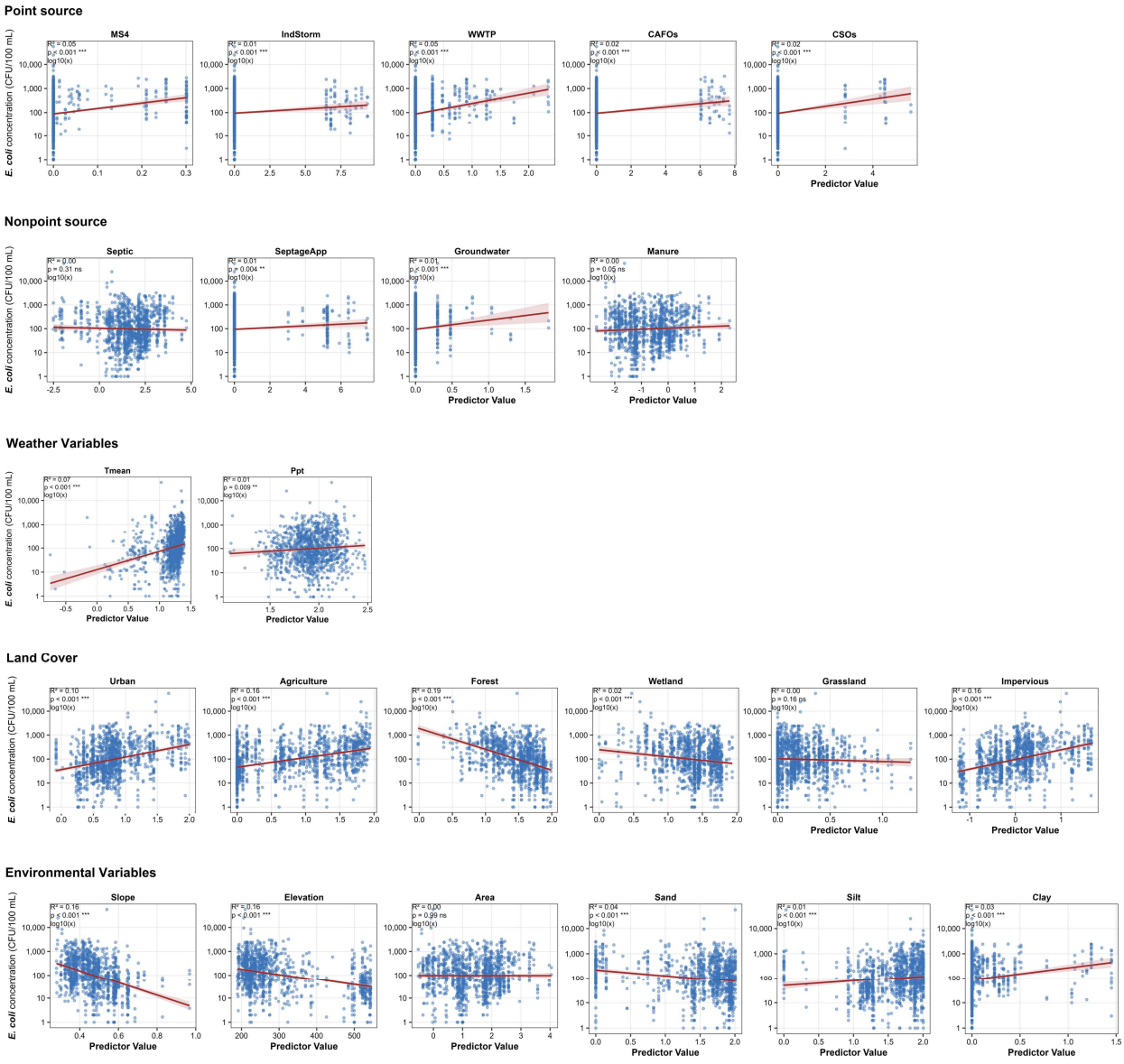


Figure 6. Bivariate correlations between 16 environmental predictors and *E. coli* geometric mean concentrations

The fixed effects regression (**Table 1**) indicates urban infrastructure represented by MS4 as the most significant variable associated with bacterial loading (β= 0.812, p< 0.001), Agricultural land is also positively associated with *E. coli* concentration (β= 0.526, p< 0.001), while mean elevation is negatively associated with *E. coli* concentration (β= −0.368, p<0.001). Other factors, including sand content, slope, watershed area, clay, and manure application, are not associated with concentration.

The loading results are consistent with the concentration results. MS4 remains the most important variable associated with bacterial loading (β=0.812, p<0.001). Some differences are that sand content is positively associated with *E. coli* loading (β= 0.392, p= 0.006), while steeper

slopes are negatively associated with *E. coli* loading (β= −0.316, p= 0.029). Watershed area is associated with *E. coli* loading (p= 0.002) but is not associated with concentration (p= 0.223). In contrast, agricultural land and mean elevation are associated with concentration but are not associated with loading. Clay shows an association with loading (β= 0.276, p=0.067), while manure application is not associated with either *E. coli* loading or concentration. In addition, the random intercept for watersheds indicates that baseline differences among watersheds are important in explaining variation in bacterial loading and concentration.

Factors such as manure application and mean elevation did not have statistical significance for *E. coli* loading. In addition, the random intercept for watersheds indicates that baseline differences among watersheds are important in explaining variation in bacterial loading and concentration.

Table 1: Linear Mixed-Effects Model Results

Variable		Estimate	Std. Error	df	t value	p value	Sig.
(Intercept)	L	3.398	0.144	159.27	23.643	< 0.001	***
	C	4.727	0.071	126.09	66.19	<0.001	***
MS4	L	0.812	0.135	172.91	6.019	< 0.001	***
	C	0.409	0.071	158.79	5.755	<0.001	***
Sand	L	0.392	0.142	163.63	2.771	0.006	**
	C	0.080	0.072	134.3	1.121	0.264	
Watershed Area	L	0.001	0	169.57	3.141	0.002	**
	C	0	0	188.16	-1.223	0.223	
Slope	L	-0.316	0.144	163.71	-2.201	0.029	*
	C	-0.113	0.074	153.8	-1.54	0.126	
Clay	L	0.276	0.149	162.02	1.845	0.067	.
	C	-0.059	0.075	125.74	-0.792	0.430	

Agricultural Land	L	0.238	0.135	166.32	1.772	0.078	.
	C	0.526	0.069	147.99	7.599	<0.001	***
Manure Application	L	-0.204	0.211	183.55	-0.965	0.336	
	C	0.078	0.12	259.45	0.653	0.514	
Elevation (Mean)	L	-0.001	0.192	153.77	-0.003	0.998	
	C	-0.368	0.092	106.93	-3.999	<0.001	***

Notes: n = 1252. Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1.

3.4 The interaction effects between pollutant source and rainfall effects on *E. coli*

The relationship between watershed characteristics and *E. coli* concentration differs by rainfall level (**Table 2**). As rainfall level changes from light to heavy, more variables are associated with *E. coli* concentration. Under heavy rainfall, MS4, agricultural land, slope, and mean elevation are associated with *E. coli* concentration. Under heavy rainfall, MS4 (β=0.204, p<0.001) and agricultural land (β=0.204, p=0.001) are positively associated with *E. coli* concentration, while slope (β=−3.897, p=0.023) is negatively associated and mean elevation (β=2.829, p=0.039) is positively associated. Under moderate rainfall, slope and manure are associated with *E. coli* concentration. Under moderate rainfall, slope (β=−3.476, p=0.002) and manure (β=−0.309, p=0.004) are negatively associated with *E. coli* concentration. Under light rainfall, agricultural land is associated with *E. coli* concentration (β=0.143, p=0.007). Sand, slope, mean elevation, clay, MS4, and manure are not associated with *E. coli* concentration under light rainfall.

Table 2. Effects of watershed variables on *E. coli* concentration by rainfall class

rain_class	variable	estimate	se	df	t_value	p_value	sig
Heavy	ms4	0.204	0.055	232.542	3.692	< 0.001	***
Heavy	ag	0.204	0.063	246.520	3.237	0.001	**
Heavy	slope	-3.897	1.708	247.598	-2.282	0.023	*
Heavy	elevMean	2.829	1.360	253.695	2.080	0.039	*
Heavy	manure	-0.332	0.171	250.765	-1.936	0.054	*

Heavy	sand	0.020	0.043	232.317	0.452	0.652	
Heavy	clay	-0.008	0.043	226.397	-0.186	0.853	
Light	ag	0.143	0.052	231.249	2.743	0.007	**
Light	sand	-0.045	0.040	241.192	-1.117	0.265	
Light	slope	-0.657	0.853	229.371	-0.771	0.442	
Light	elevMean	0.389	0.839	228.517	0.464	0.643	
Light	clay	-0.006	0.035	189.629	-0.182	0.856	
Light	ms4	-0.006	0.046	140.349	-0.138	0.891	
Light	manure	0.014	0.104	204.367	0.134	0.894	
Moderate	slope	-3.476	1.108	197.166	-3.138	0.002	**
Moderate	manure	-0.309	0.105	175.664	-2.938	0.004	**
Moderate	ag	0.105	0.059	220.167	1.786	0.076	*
Moderate	clay	0.057	0.035	149.729	1.655	0.100	.
Moderate	ms4	0.062	0.044	158.940	1.409	0.161	.
Moderate	elevMean	1.277	0.927	202.904	1.377	0.170	.
Moderate	sand	0.002	0.038	163.858	0.053	0.958	

Signif. codes: < 0.001 '***' 0.01 '**' 0.05 '*' 0.1 '.' 1 ' '

Chapter 4 Discussion

4.1. Pollutant sources and watershed characteristics affecting *E. coli*

In this study, we examined how pollutant sources and watershed characteristics affect *E. coli* concentration and loading across Michigan watersheds under different rainfall conditions. The results show that *E. coli* is not controlled by one source alone. Instead, its variation reflects the combined effects of urban infrastructure, agricultural land use, soil conditions, topography, and watershed size. This means that the distribution of fecal contamination across watersheds depends on both source availability and transport conditions.

Among the selected variables, MS4 was the most important factor in the model. This indicates that urban stormwater systems are an important pathway for *E. coli* transport. In watersheds with more MS4 facilities, rainfall can move fecal material from impervious surfaces and drainage systems to streams more efficiently. This pathway is different from agricultural transport, but both increase the delivery of microbial contaminants to receiving waters. Previous studies also showed that stormwater drainage, land use, and watershed setting affect microbial water quality at the watershed scale (Cho et al., 2016; Oliver et al., 2016; Hong et al., 2021).

Agricultural land was also important in the concentration model. This suggests that agricultural watersheds provide broad source areas for fecal contamination. Compared with urban systems, agricultural transport is less related to direct drainage structures and more related to runoff from fields, manure-affected land, and shallow surface flow. In this way, the positive effect of agricultural land should be understood as a land use effect on source generation and delivery at the same time.

The environmental variables in this study help explain why watersheds with similar source types do not always have similar *E. coli* responses. Soil texture affects infiltration and runoff, slope affects overland flow and erosion, and watershed area affects the integration of upstream sources and discharge volume. These variables are not direct pollutant sources, but they control how microbial contaminants are stored, mobilized, and transported. The difference between the concentration and loading models also supports this interpretation. Some variables were more closely related to concentration, while others were more important for loading. This indicates that *E. coli* concentration and loading describe different parts of the contamination process. Concentration reflects the level of contamination in water, while loading also reflects transport through flow.

These results suggest that *E. coli* contamination in Michigan watersheds should be treated as both a source problem and a transport problem. Source variables explain where contamination may originate, but watershed characteristics explain whether that contamination reaches stream channels and how far it is carried. For this reason, watershed management should not focus only on one source category. It should consider how source areas and transport pathways work together.

4.2. Rainfall effects on *E. coli* transport under different watershed conditions

The rainfall results help explain when these watershed drivers become important. In this study, rainfall did not simply increase *E. coli* levels. It changed the transport conditions under which different watershed characteristics became active. This is important because the same watershed may respond differently under dry, light, moderate, and heavy rainfall.

Under lower rainfall conditions, only some drivers were associated with *E. coli*. This suggests that under limited runoff, only a part of the watershed is hydrologically connected to stream channels. In this situation, some fecal material remains stored on land surfaces or in drainage areas and does not reach the stream efficiently. Under moderate and heavy rainfall, more variables became associated with *E. coli*. This indicates that rainfall increases hydrologic connection across the watershed and activates more transport pathways. Rainfall therefore changes not only the amount of contamination, but also the structure of transport.

This pattern also helps explain the difference between urban and agricultural watersheds in the rainfall analysis. In urban watersheds, the response to accumulated rainfall was faster. This is likely related to impervious surfaces and stormwater systems, which can move contaminants to streams quickly after rainfall. In agricultural watersheds, the rainfall response extended over a longer accumulation period. This pattern fits field runoff and land surface transport, where rainfall must accumulate before broader parts of the watershed contribute to stream contamination. Earlier studies also showed that rainfall, runoff, and land use interact in shaping microbial water quality, but the effect differs across watersheds and hydrologic settings (Boithias et al., 2016; Carlton et al., 2014; Cho et al., 2016).

The rainfall-class model further shows that the importance of watershed drivers changes with rainfall intensity. Under light rainfall, agricultural land remained important, while under moderate and heavy rainfall more variables entered the response. This means that rainfall intensity acts as a filter on watershed processes. Under weaker rainfall, transport is limited. Under stronger rainfall, runoff, drainage flow, and landscape connection increase together. In this way, heavy rainfall should be viewed as a condition in which multiple source and transport mechanisms operate at the same time, rather than as a simple increase in precipitation amount.

4.2.1. Implications for watersheds under lower rainfall conditions

Under lower rainfall conditions, management should focus on source areas that remain active even without strong runoff. In this study, agricultural land was still associated with *E. coli* under lighter rainfall classes. This suggests that some watersheds maintain contamination pathways even when rainfall is limited. In these settings, source reduction remains important. This includes improving manure management, reducing runoff from fields, and maintaining buffer areas that reduce delivery to streams.

The lower-rainfall results also show that routine conditions can still produce microbial risk in some watersheds. For this reason, management should not focus only on major storms. Watersheds with strong source pressure may continue to show elevated *E. coli* even when the hydrologic response is limited. This is particularly important in agricultural watersheds where contamination can accumulate on land surfaces and enter streams during repeated low to moderate rainfall events.

4.2.2. Implications for watersheds under heavy rainfall conditions

Under heavy rainfall, the results indicate that *E. coli* transport becomes a broader watershed process. Urban infrastructure, agricultural land, slope, and elevation all became relevant under these conditions. This means that heavy rainfall connects more source areas to streams and allows more parts of the watershed to contribute to microbial loading at the same time. In this situation, watersheds with both strong pollutant sources and efficient transport pathways are likely to face the highest risk.

For urban watersheds, this points to the role of stormwater systems during storms. Reducing direct drainage connection, improving stormwater treatment, and identifying areas with repeated runoff-driven contamination may help reduce *E. coli* delivery after rainfall. For agricultural watersheds, the results point to runoff control, soil conservation, and timing of land application before rainfall events. Because heavy rainfall activates multiple pathways at once, management during these conditions should focus on reducing rapid transport rather than only reducing background source levels.

These findings are also relevant for public health risk. Previous work showed that extreme precipitation is linked to higher microbial exposure and waterborne disease risk (Curriero et al., 2001; Maphanga et al., 2022). In this study, the rainfall results support the use of rainfall-aware monitoring and post-storm risk assessment, especially in watersheds with urban drainage systems and mixed agricultural land use.

4.3. Limitations and future research

This study has several limitations. First, some potential source variables were excluded from the final analysis because of data quality, spatial completeness, or collinearity. As a result, the final models do not represent all regulated and nonpoint sources equally. Second, the study used routine monitoring data and watershed-scale summaries. These data are useful for identifying broad statewide patterns, but they are less able to capture short-term storm peaks and within-event variation. Third, this study focused on rainfall classes and short-term rainfall accumulation. Other hydrologic factors, such as antecedent moisture, snowmelt, and seasonal land management, may also affect *E. coli* transport.

Future research could improve this framework by adding event-based sampling, improving source inventories, and including variables that better describe antecedent wetness and timing of source release. It would also be useful to test whether the rainfall-related changes in driver importance observed in Michigan occur in other regions with different land use and hydrologic settings. Even with these limitations, this study provides a statewide framework for linking pollutant sources, watershed characteristics, and rainfall conditions to *E. coli* contamination.

Chapter 5 Conclusion

This study examined how pollutant sources, watershed characteristics, and rainfall conditions affect *E. coli* contamination across 247 Michigan watersheds. It provides a watershed-scale framework for understanding the spatiotemporal distribution of E. coli, the main drivers associated with *E. coli* variability, and the role of rainfall in changing microbial transport and risk. The results show that *E. coli* contamination is shaped by both source distribution and watershed transport conditions. Urban stormwater systems and agricultural land use were important in explaining *E. coli* variability, while soil texture, slope, elevation, and watershed area affected the movement of contaminants through the watershed. This indicates that *E. coli* contamination is not only a source issue, but also a transport issue. The rainfall analysis further shows that the role of watershed drivers changes under different rainfall conditions. Under heavier rainfall, more transport pathways become active and more watershed characteristics are involved in *E. coli* delivery to receiving waters. This suggests that rainfall should be treated as an important part of microbial risk assessment and watershed management. In summary, this study shows that *E. coli* contamination across Michigan watersheds is controlled by the interaction between pollutant sources, watershed characteristics, and rainfall conditions. These findings support management strategies that combine source control, stormwater management, agricultural runoff reduction, and rainfall-aware monitoring under changing precipitation conditions.

Bibliography

Boithias, L., Choisy, M., Souliyaseng, N., Jourdren, M., Quet, F., Buisson, Y., ... & Ribolzi, O. (2016). Hydrological regime and water shortage as drivers of the seasonal incidence of diarrheal diseases in a tropical montane environment. *PLoS neglected tropical diseases*, *10*(12), e0005195.

Carlton, E. J., Eisenberg, J. N., Goldstick, J., Cevallos, W., Trostle, J., & Levy, K. (2014). Heavy rainfall events and diarrhea incidence: the role of social and environmental factors. *American journal of epidemiology*, *179*(3), 344-352.

Cho, K. H., Pachepsky, Y. A., Oliver, D. M., Muirhead, R. W., Park, Y., Quilliam, R. S., & Shelton, D. R. (2016). Modeling fate and transport of fecally-derived microorganisms at the watershed scale: State of the science and future opportunities. *Water research*, *100*, 38-56.\

Cosgrove, B., Gochis, D., Flowers, T., Dugger, A., Ogden, F., Graziano, T., ... & Zhang, Y. (2024). NOAA's National Water Model: Advancing operational hydrology through continental-scale modeling. *JAWRA Journal of the American Water Resources Association*, *60*(2), 247-272.

Curriero, F. C., Patz, J. A., Rose, J. B., & Lele, S. (2001). The association between extreme precipitation and waterborne disease outbreaks in the United States, 1948-1994. *American journal of public health*, *91*(8), 1194-1199.

Densmore, B. K., Hall, B. M., & Moser, M. T. (2020). Modeling Escherichia coli in the Missouri River near Omaha, Nebraska, 2012–16.

Dila, D. K., Corsi, S. R., Lenaker, P. L., Baldwin, A. K., Bootsma, M. J., & McLellan, S. L. (2018). Patterns of host-associated fecal indicators driven by hydrology, precipitation, and land use attributes in Great Lakes watersheds. *Environmental science & technology*, *52*(20), 11500-11509.

Dong, Y., Jiang, Z., Hu, Y., Jiang, Y., Tong, L., Yu, Y., ... & Wang, Y. (2024). Pathogen contamination of groundwater systems and health risks. *CritiCal reviews in environmental sCienCe and teChnology*, *54*(4), 267-289.

Edge, T. A., & Hill, S. (2007). Multiple lines of evidence to identify the sources of fecal pollution at a freshwater beach in Hamilton Harbour, Lake Ontario. *Water research*, *41*(16), 3585-3594.

EGLE (Michigan Department of Environment, Great Lakes, and Energy). (n.d.). *Statewide E. coli Total Maximum Daily Load (TMDL)*. Michigan Department of Environment, Great Lakes, and Energy. Retrieved from https://www.michigan.gov/egle

Flood, M. T., Hernandez-Suarez, J. S., Nejadhashemi, A. P., Martin, S. L., Hyndman, D., & Rose, J. B. (2022). Connecting microbial, nutrient, physiochemical, and land use variables for the evaluation of water quality within mixed use watersheds. *Water research*, *219*, 118526.

Fotheringham, A. S., Yang, W., & Kang, W. (2017). Multiscale geographically weighted regression (MGWR). *Annals of the American Association of Geographers*, *107*(6), 1247-1265.

Garzio-Hadzick, A., Shelton, D. R., Hill, R. L., Pachepsky, Y. A., Guber, A. K., & Rowland, R. (2010). Survival of manure-borne *E. coli* in streambed sediment: effects of temperature and sediment properties. *Water research*, *44*(9), 2753-2762.

Hall, B. M. (2020). Modeling Escherichia coli in the Missouri River near Omaha, Nebraska, 2012-16: Model Inputs and Outputs. *US Geological Survey (USGS) Data Release*, 325.

Hirsch, R. M., Moyer, D. L., & Archfield, S. A. (2010). Weighted regressions on time, discharge, and season (WRTDS), with an application to Chesapeake Bay river inputs 1. *JAWRA Journal of the American Water Resources Association*, *46*(5), 857-880.

Hong, Y., Soulignac, F., Roguet, A., Li, C., Lemaire, B. J., Martins, R. S., ... & Vinçon-Leite, B. (2021). Impact of *Escherichia coli* from stormwater drainage on recreational water quality: an integrated monitoring and modelling of urban catchment, pipes and lake. *Environmental Science and Pollution Research*, *28*(2), 2245-2259.

Ishii, S., & Sadowsky, M. J. (2008). *Escherichia coli* in the environment: implications for water quality and human health. *Microbes and environments*, *23*(2), 101-108.

Jang, J., Hur, H. G., Sadowsky, M. J., Byappanahalli, M. N., Yan, T., & Ishii, S. (2017). Environmental Escherichia coli: ecology and public health implications—a review. *Journal of applied microbiology*, *123*(3), 570-581.

Kraay, A. N., Man, O., Levy, M. C., Levy, K., Ionides, E., & Eisenberg, J. N. (2020). Understanding the impact of rainfall on diarrhea: testing the concentration-dilution hypothesis using a systematic review and meta-analysis. *Environmental health perspectives*, *128*(12), 126001.

Levy, K., Hubbard, A. E., & Eisenberg, J. N. (2009). Seasonality of rotavirus disease in the tropics: a systematic review and meta-analysis. *International journal of epidemiology*, *38*(6), 1487-1496.

Liu, X., Zuo, C., Guan, J., Ma, Y., Liu, Y., Zhao, G., & Wang, R. (2025). Extreme rainfall disproportionately impacts *E. coli* concentrations in Texas recreational waterbodies. *Science of The Total Environment*, *958*, 178062.

Liu, X., Zuo, C., Guan, J., Ma, Y., Liu, Y., Zhao, G., & Wang, R. (2025). Extreme rainfall disproportionately impacts *E. coli* concentrations in Texas recreational waterbodies. *Science of The Total Environment*, *958*, 178062.

Maphanga, B. R., Abia, A. L., Amoah, I. D., Bux, F., & Ubomba-Jaswa, E. (2022). Impacts of extreme precipitation and temperature on *E. coli* loads in riverine systems. *Science of The Total Environment, 828*, 154434. https://doi.org/10.1016/j.scitotenv.2022.154434

Nevers, M. B., & Whitman, R. L. (2011). Efficacy of monitoring and empirical predictive modeling at improving public health protection at Chicago beaches. *Water Research*, *45*(4), 1659-1668.

Oliver, D. M., Porter, K. D., Pachepsky, Y. A., Muirhead, R. W., Reaney, S. M., Coffey, R., ... & Quilliam, R. S. (2016). Predicting microbial water quality with models: over-arching questions for managing risk in agricultural catchments. *Science of the Total Environment*, *544*, 39-47.

Oliver, D. M., Porter, K. D., Pachepsky, Y. A., Muirhead, R. W., Reaney, S. M., Coffey, R., ... & Quilliam, R. S. (2016). Predicting microbial water quality with models: over-arching questions for managing risk in agricultural catchments. *Science of the Total Environment*, *544*, 39-47.

Pandey, P. K., Kass, P. H., Soupir, M. L., Biswas, S., & Singh, V. P. (2014). Contamination of water resources by pathogenic bacteria. *Amb Express*, *4*(1), 51.

Patz, J. A., Vavrus, S. J., Uejio, C. K., & McLellan, S. L. (2008). Climate change and waterborne disease risk in the Great Lakes region of the US. *American journal of preventive medicine*, *35*(5), 451-458.

Patz, J. A., Vavrus, S. J., Uejio, C. K., & McLellan, S. L. (2008). Climate change and waterborne disease risk in the Great Lakes region of the US. *American journal of preventive medicine*, *35*(5), 451-458.

Ramseier, M. K., von Gunten, U., Freihofer, P., & Hammes, F. (2011). Kinetics of membrane damage to high (HNA) and low (LNA) nucleic acid bacterial clusters in drinking water by ozone, chlorine, chlorine dioxide, monochloramine, ferrate (VI), and permanganate. *Water Research*, *45*(3), 1490-1500.

Scavia, D., Allan, J. D., Arend, K. K., Bartell, S., Beletsky, D., Bosch, N. S., ... & Zhou, Y. (2014). Assessing and addressing the re-eutrophication of Lake Erie: Central basin hypoxia. *Journal of Great Lakes Research*, *40*(2), 226-246.

Sinclair, A., Hebb, D., Jamieson, R., Gordon, R., Benedict, K., Fuller, K., ... & Madani, A. (2009). Growing season surface water loading of fecal indicator organisms within a rural watershed. *Water Research*, *43*(5), 1199-1206.

U.S. Environmental Protection Agency. (2012). *2012 Recreational water quality criteria*. U.S. Environmental Protection Agency, Office of Water, EPA 820-F-12-058. https://www.epa.gov/wqc/recreational-water-quality-criteria-and-methods

Verhougstraete, M. P., Martin, S. L., Kendall, A. D., Hyndman, D. W., & Rose, J. B. (2015). Linking fecal bacteria in rivers to landscape, geochemical, and hydrologic factors and sources at the basin scale. *Proceedings of the National Academy of Sciences*, *112*(33), 10419-10424.

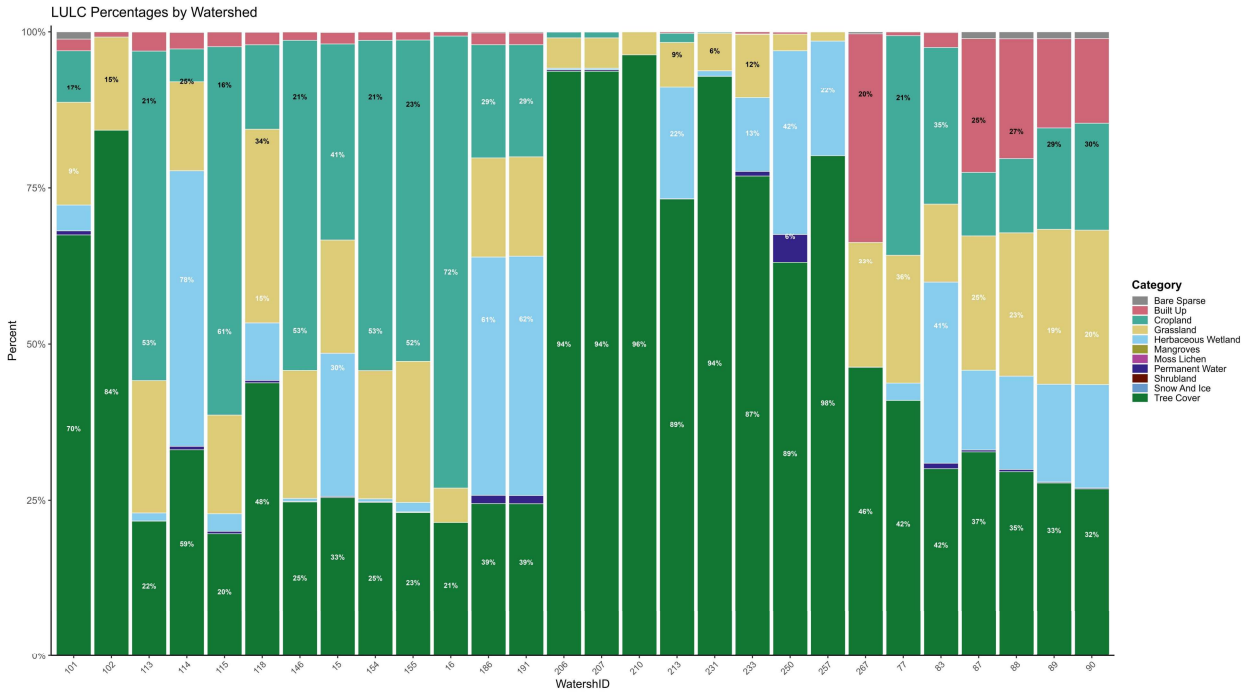
Wang, L., Chen, Y., Chen, S., Long, L., Bu, Y., Xu, H., ... & Krasner, S. (2019). A one-year long survey of temporal disinfection byproducts variations in a consumer's tap and their removals by a point-of-use facility. *Water research*, *159*, 203-213.

Whitman, R. L., & Nevers, M. B. (2004). *Escherichia coli* sampling reliability at a frequently closed Chicago beach: monitoring and management implications.

Wilkes, G., Edge, T., Gannon, V., Jokinen, C., Lyautey, E., Medeiros, D., ... & Lapen, D. R. (2009). Seasonal relationships among indicator bacteria, pathogenic bacteria, Cryptosporidium oocysts, Giardia cysts, and hydrological indices for surface waters within an agricultural landscape. *Water research*, *43*(8), 2209-2223.

Wu, H., Oun, A., Kline-Robach, R., & Xagoraraki, I. (2018). Microbial pollution characterization at a TMDL site in Michigan: effect of hydrological conditions on pollution loading. *Journal of Great Lakes Research*, *44*(3), 421-427.

Appendices



Appendix 1. 28 watershed land cover heterogeneity

We applied an ordinary least squares (OLS) regression model to explore the main drivers of *E. coli* loading. The predictors included manure application, rainfall (ms4), watershed area, clay content, agricultural land use (ag), mean elevation, elevation variability (elevStdDev), and sand content.

The model explained about 42% of the variance in *E. coli* loading (adjusted $R^2 = 0.42$, $F = 3.43$, $p = 0.013$). Among the predictors, rainfall (ms4) showed the strongest positive effect ($p < 0.01$), indicating that higher rainfall is associated with increased *E. coli* loading. Manure application ($p < 0.05$), agricultural land use ($p < 0.05$), and sand content ($p < 0.05$) were also significant contributors. Elevation variability (elevStdDev) was marginally significant ($p \approx 0.05$), suggesting that heterogeneous terrain may also influence loading. In contrast, watershed area, mean elevation, and clay content did not show significant effects.

These findings highlight the combined roles of hydrological conditions, land management, and soil texture in shaping *E. coli* contamination patterns across the study area.

Appendix 2. OLS regression results for *E. coli* loading (n = 28)

Variable	Estimate	Std. Error	t value	p value	Sig.
(Intercept)	−0.279	0.144	0.000	1.000	
Manure	0.603	0.289	2.095	0.049	*

Ms4	0.677	0.184	3.687	0.002	**
Area	−0.586	0.314	−1.867	0.077	·
Clay	−0.083	0.178	−0.463	0.649	
Ag	0.293	0.132	2.211	0.040	*
ElevMean	−0.061	0.233	−0.261	0.797	
ElevStdDev	0.299	0.145	2.048	0.053	·
Sand	0.592	0.236	2.516	0.021	*

Model fit: Residual SE = 0.76 (df = 19); Multiple R^2 = 0.59; Adjusted R^2 = 0.42; F = 3.43, p = 0.013.

Appendix 3. Datasets and preprocessing details

Data source / Provider	**Format**	**Native resolution**	**Time coverage**	**Variables derived (Table 1)**
USGS 3DEP DEM	Shapefile/DEM	30 m	2024 (by 247 stations)	Watershed area; Slope; Elevation
WQP: Water Quality Portal	CSV	Site stations	1996-2024	Stream *E. coli* concentration
National Water Model	CSV	Site stations	1979-2022	Stream discharge
EGLE: Potential MS4 Regulation Areas (Urbanized Areas)	Shapefile (polygon)	—	1899	MS4s
EPA: Industrial Storm Water Data	Shapefile (point)	—	2018-2022 (2007-2025)	Industrial storm water (EPA)
EGLE: NPDES Facilities	Shapefile (point)	—	—	WWTPs & WWSLs
EGLE: CAFOs	Shapefile (point)	—	2020	CAFOs
EGLE: CSOs—Waters Directly Affected	Shapefile (line)	—	2021	CSOs
EGLE: Subwatershed Summary	Shapefile (polygon)	—	—	Estimated septic / ag. land / manure

EGLE: Approved Septage Land Application Areas	Shapefile (polygon)	—	1899	Approved septage land application
EGLE: Groundwater Permitted Sanitary Discharges	Shapefile (point)	—	2023	Groundwater discharges
SoilGrids (ISRIC 250 m)	Raster	250 m	2020	Soil type
ERA5 (ECMWF Reanalysis v5)	Raster	0.25° (~31 km)	2018-2022 (daily)	Daily temperature; Rainfall
Copernicus ESA CGLS-LC100	Raster	300 m	2021	LULC
NASADEM (30 m DEM)	Raster	30 m	2000 / —	Slope; Elevation
GISD30: Global 30 m Impervious Surface	Raster	30 m	2021	Impervious density
HRSL: High-Resolution Population Density & Demographic Estimates	Raster	30 m	2019-2020	Population density

To identify the major drivers of *E. coli* variability across watersheds, we first applied an ordinary least squares (OLS) regression model using data from the 28 monitoring stations with sufficient long-term records between 2018 and 2022:

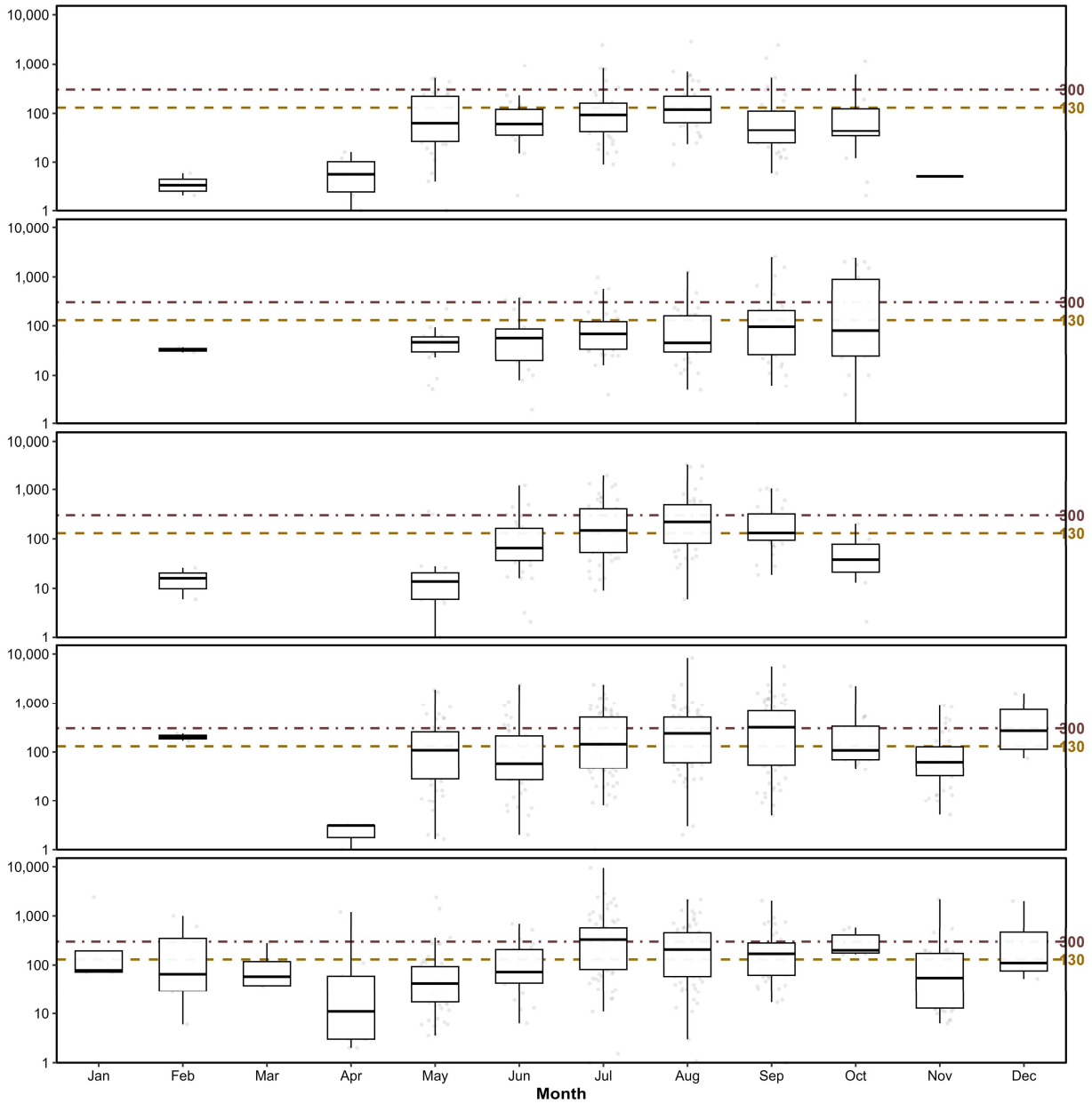
$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \epsilon_i$$

Independent variables included point and nonpoint sources (e.g., MS4s, manure application), land use characteristics (e.g., percentage of agricultural land), and environmental controls (e.g., soil texture, mean elevation, elevation variability). This model provided an initial assessment of the overall relationships between watershed characteristics and *E. coli* concentrations.

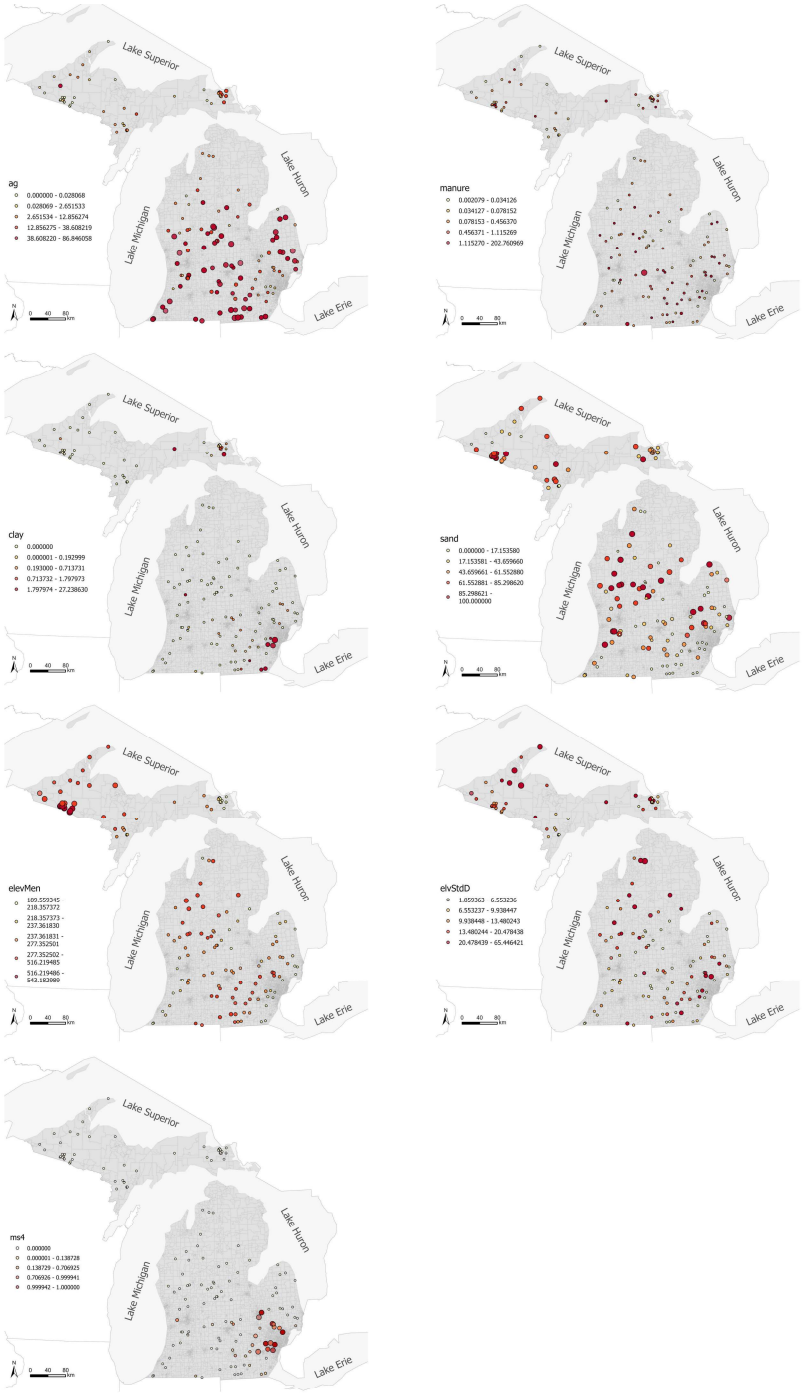
To verify results from some frequently monitored stations, Weighted Regressions on Time, Discharge, and Season (WRTDS) was then applied to model concentration-discharge relationships and to interpolate daily *E. coli* concentrations, which were combined with streamflow to calculate daily loadings:

$$\ln(c) = \beta_0 + \beta_1 t + \beta_2 \ln(Q) + \beta_3 \sin(2\pi t) + \beta_4 \cos(2\pi t) + \epsilon$$

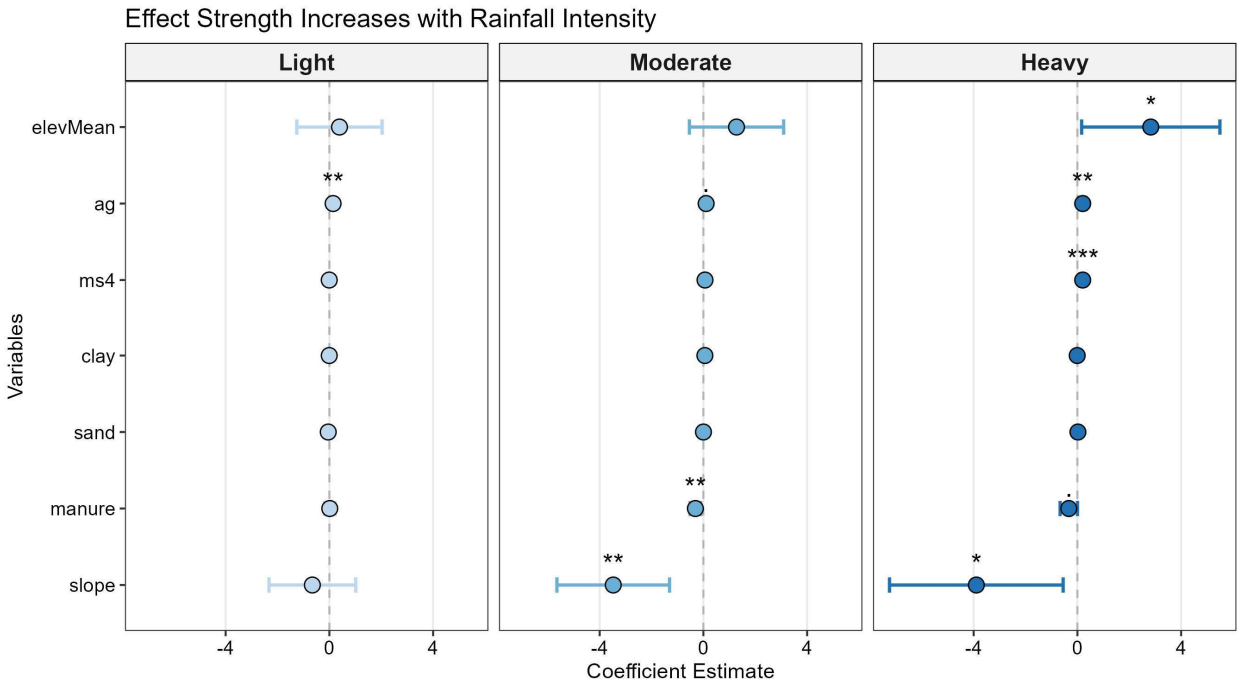
where c is concentration ($CFU\ 100\ mL^{-1}$), Q is discharge ($m^3\ s^{-1}$), and t is time in decimal years. We used tricube weights in (t, $\ln Q$, season) so estimates emphasize observations near the target day (Hirsch et al., 2010)



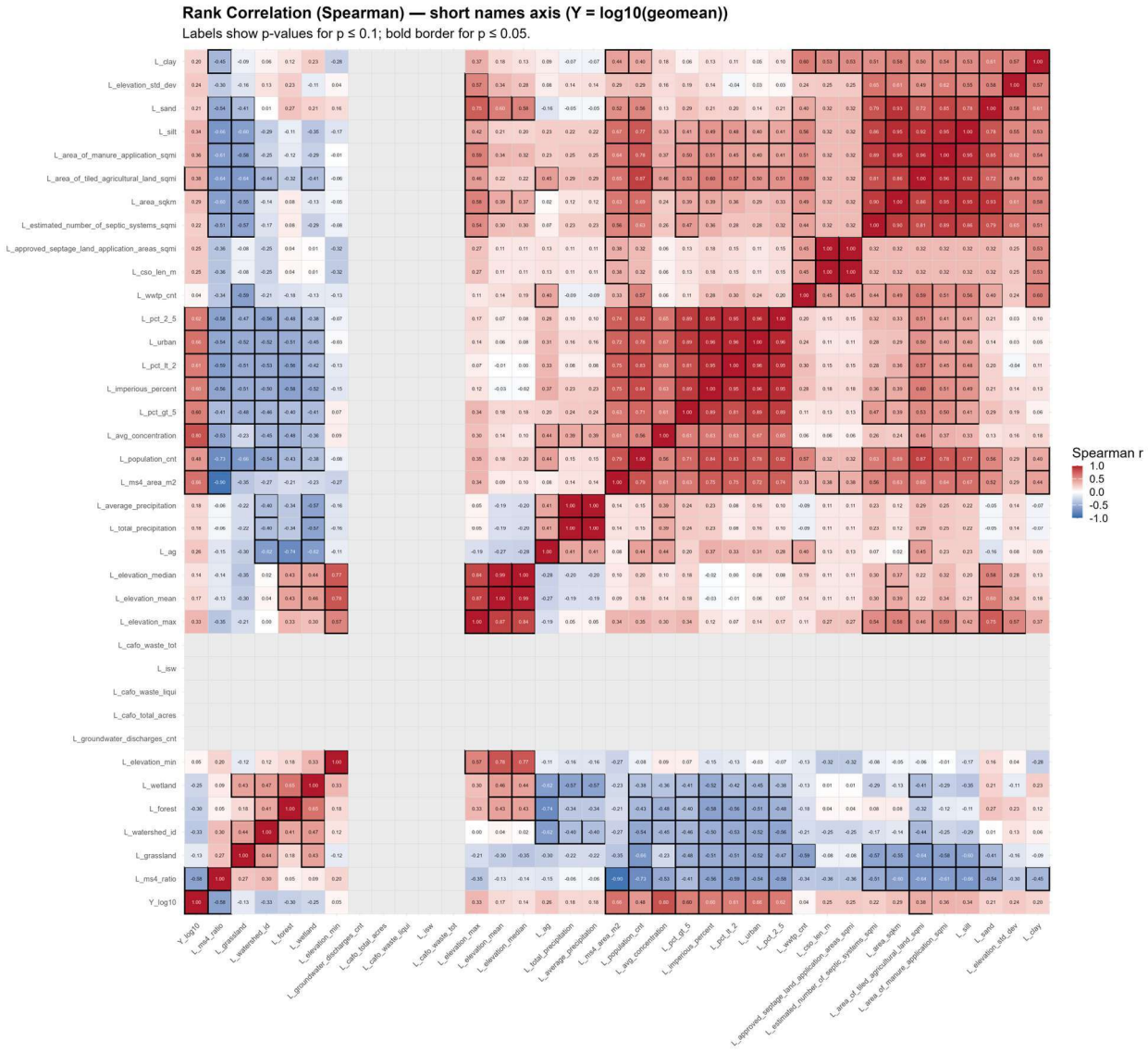
Appendix 4. Year-specific monthly distributions of *E. coli* concentrations from 2018 to 2022. Boxplots show the distribution of monthly *E. coli* concentrations for each year, and gray points represent individual observations. The dashed lines indicate the recreational water quality thresholds of 130 and 300 CFU/100 mL. Seasonal increases from summer to early fall were observed in each year.



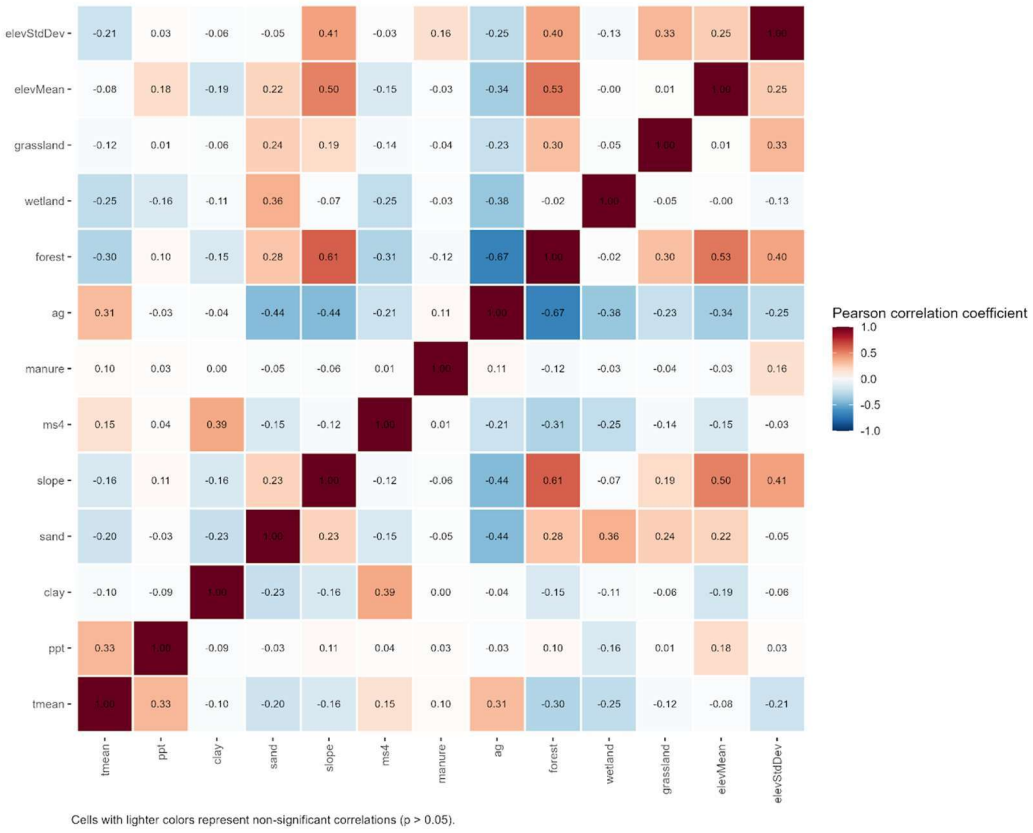
Appendix 5. Maps show the spatial patterns of (a) MS4 coverage, (b) elevation variability (*elevStdDev*), (c) mean elevation (*elevMean*), (d) percentage of agricultural land (*ag*), (e) clay content, (f) sand content, and (g) manure application.



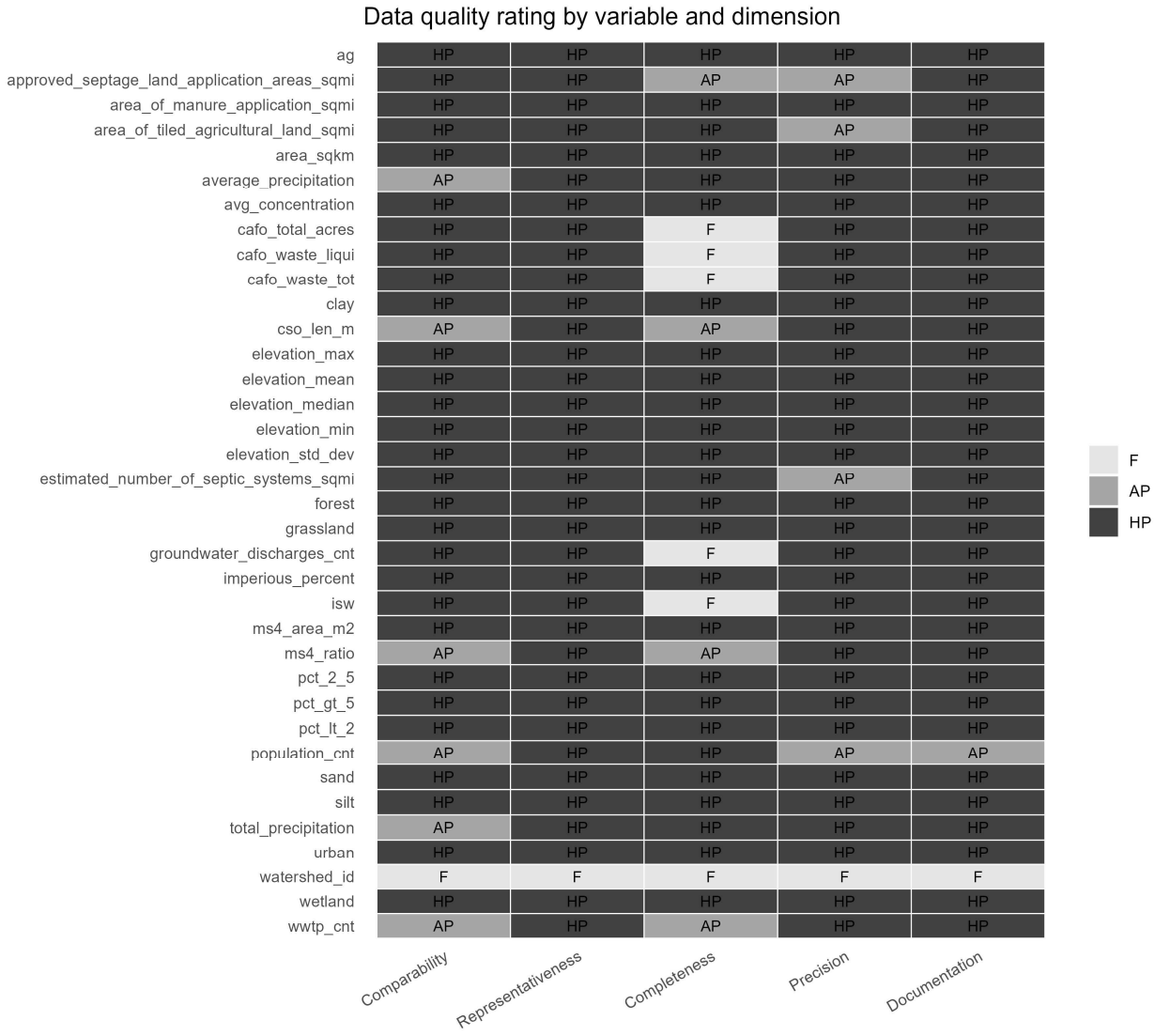
Appendix 6. Effects of waStershed variables on *E. coli* concentration across rainfall classes. Points show coefficient estimates and horizontal bars show 95% confidence intervals. Asterisks indicate statistical significance.



Appendix 7. Spearman rank correlation matrix for *E. coli* concentration and candidate explanatory variables. Cell values indicate correlation coefficients, and bold borders indicate p ≤ 0.05.



Appendix 8. Pearson correlation matrix of selected explanatory variables after collinearity screening. Cell values show Pearson correlation coefficients, and lighter colors indicate non-significant correlations ($p > 0.05$). No retained variable pair had an absolute correlation coefficient greater than 0.7.



Appendix 9. Data quality ratings of candidate variables across five dimensions: comparability, representativeness, completeness, precision, and documentation. HP indicates high pass, AP indicates adequate pass, and F indicates fail.