Modeling the Health Risk of Stream E. coli with Qualitative Microbial Risk Assessment and Assessing its Relationship with Social Vulnerability Index

by

Yiyi Liu

A thesis submitted
in partial fulfillment of the requirements
for the degree of
Master of Science (Geospatial Data Science)
(Environment and Sustainability)
in the University of Michigan
04 2024

Thesis Committee:
Assistant Professor Runzi Wang, Chair
Assistant Professor Derek Van Berkel

Abstract

Escherichia coli contamination poses a significant water quality challenge, especially in recreational water settings, with adverse health effects particularly pronounced in socially vulnerable communities. Despite regional studies, a comprehensive national understanding of spatial-temporal variations in *E. coli* health risks remains a challenge, and the association between social vulnerability and *E. coli* exposure in recreational waters is understudied. Utilizing a continuous *E. coli* time series from national stream data spanning 1987-2022, we conducted a quantitative microbial risk assessment (QMRA) to evaluate *E. coli* exposures during water-related recreational activities across the U.S. We employed the Mann-Kendall trend test to analyze monotonic trend patterns in *E. coli* infection probabilities. Our findings pinpoint E. coli risk exposure hotspots, predominantly in the Midwest and West regions, notably in Wisconsin, Minnesota, Iowa, Missouri, and Oregon. Regional disparities in social vulnerability are closely associated with elevated *E. coli* health risks across various dimensions. Of the 406 sampling sites assessed, 76 showed an increasing trend in *E. coli* infectious probability, while 107 exhibited a decreasing trend. Regions with increasing health burdens are predominantly vulnerable in the Southeast and Southwest, while decreasing health burdens are observed in the Northeast. This research offers novel insights into *E. coli* health risk dynamics in U.S. streams, informing the formulation of targeted public health interventions and environmental management strategies to enhance water safety during recreational activities.

Table of Contents

Chapter1: Introduction 3

1.1 *E. coli* Exposure 4

1.2 Quantitative microbial risk assessment (QMRA) 5

1.3 Social Vulnerability with Waterborne Diseases Health Burden 6

Chapeter2: Method 8

2.1 *E. coli* Sampling Sites Selection 9

2.2 Generating Continuous Time Series of *E. coli* Concentration 9

2.3 Quantitative Microbial Risk Assessment 10

2.4 Trend Analysis 12

2.5 Social Vulnerability Index 13

Chapeter3: Result and Discussion 14

3.1 *E. coli* Risk Level by Concentration and by Infectious Rate 15

3.2 The Association between Social Vulnerability and Health Burden by Regions 18

3.3 Trend of *E. coli* Infectious Rate 20

3.4 The Association between Trend of *E. coli* Infectious Rate and Social Vulnerability by Regions 22

Chapeter4: Study Limitations 24

Chapeter5: Conclusion 28

Chapter 1: Introduction

1. Introduction

1.1 E. coli Exposure

As one of the most important fecal indicator bacteria (FIB), *Escherichia coli* (*E. coli*) contamination is an essential water quality issue. *E. coli* commonly inhabits the gastrointestinal tract of warm-blooded animals, including humans and livestock such as cattle and poultry (Hart et al., 1993). *E. coli* in surface water primarily emanates from point source pollution, e.g., direct discharge of human domestic sewage, or from non-point source pollution, e.g., combined sewer overflow pollution or stormwater runoff from livestock pastures (Tousi et al., 2021). While most *E. coli* strains are harmless, certain strains can engender health risks when introduced into contaminated water, thereby amplifying the potential hazards associated with irrigated food, drinking water, and water-based recreational activities (Rodrigues et al., 2016). Given the close association between health risks and the presence of *E. coli*, an epidemiological study by the United States Environmental Protection Agency (1986) recommended a maximum surface water *E. coli* threshold of 126 CFU/100 mL, determined as the geometric mean of a minimum of five samples, with no individual sample exceeding 235 CFU/100 mL (EPA, 2012). Recreational water contact is a major exposure pathway of *E. coli*; and the above study reported swimmers experienced higher rates of gastrointestinal risk when compared to non-swimmers as the exposure of FIB.

Numerous studies have highlighted the significant health risks associated with *E. coli* in water environments, including those related to sewage treatment systems and the global burden of disease caused by intermittent water supplies (Bivins et al., 2017; Boehm et al., 2018). Also, Machdar analyzed the relationship between poor drinking water quality in a low-income community with health burden in Ghana (Machdar et al., 2013). Goh assessed the additional health burden of *E. coli* from surface water with land use, including residential, urban, parkland and agricultural regions from 2014 to 2016 in Singapore (Goh et al., 2023). However, most literature focused on isolated geographical areas or time frames, providing only a limited understanding of the broader trends and long-term impacts of *E. coli* across the U.S. (Jacob et al., 2015; O'Flaherty et al., 2019). Critical research has revealed links between *E. coli* and health risks in U.S. water recreation environments, particularly in the Great Lakes area (Corsi et al., 2016), and have investigated the relationship between water quality and gastrointestinal diseases following exposure to recreational water (Dorevitch et al., 2015). Similar regional studies have highlighted levels of *E. coli* exceeding safe thresholds for recreational waters in the Upper Oconee Watershed in Northeast Georgia (Cho et al., 2018), and have reported on significant challenges in predicting recreational water safety using *E. coli* as an indicator in New Jersey's Passaic and Pompton rivers (Rossi et al., 2020). But to the best of our knowledge, there is no comprehensive national research on the spatial and temporal variations of *E. coli* health risk.

1.2 Quantitative microbial risk assessment (QMRA)

Quantitative microbial risk assessment (QMRA) is a mathematical framework for quantifying the illness risk due to microbial pathogens (Goh et al., 2023; Pasalari et al., 2022). In the U.S., QMRA has increasingly been applied to address water quality-related health issues. This includes assessing public safety in recreational water areas, the effectiveness of wastewater treatment, and risks associated with reuse of water for agricultural purposes, and further proposing strategies to protect water safety (Brouwer et al., 2018; Goh et al., 2023; Machdar et al., 2013; Masciopinto et al., 2020). Specifically, QMRA has been widely used in *E. coli* risk analysis, such as how *E. coli* in recreational waters, wastewater systems, and drinking water leads to public health issues (Beaudequin et al., 2015; Brouwer et al., 2018; Machdar et al., 2013). Increasing research also attempts to combine QMRA with other risk assessment methods like Bayesian networks to better understand the dynamics of *E. coli* in micro-water ecosystems (Beaudequin et al., 2015). A recent study also estimated stream *E. coli* health risk via QMRA across Nepal to better manage sporadic nationwide diarrheal outbreaks by analyzing the spatial differences of infectious probability in 2017 (Uprety et al., 2020). The waterborne disease through drinking water systems in rural Colombia has been analyzed through QMRA by using secondary water quality data and survey approach in 2015 to 2017 (Barragán et al., 2021). However, current applications of QMRA for analyzing *E. coli* risk are either primarily based on smaller spatial scales of water environment or considering broader geographic areas without long term changes of waterborne diseases (Abia et al., 2016; Bivins et al., 2017). There is limited understanding regarding the dynamic health burden of stream *E. coli* on large time scales and on broad spatial scales in the recreational water environment by integrating QMRA. Also, no study uses trend analysis for displaying the long-term changes of *E. coli* exposure burden nationwide. As a result, finding the increasing or decreasing patterns of *E. coli* health burden could help health policy makers and water quality management to formulate and evaluate mitigation plans for targeted regions.

1.3 Social Vulnerability with Waterborne Diseases Health Burden

Social vulnerability is a systematic way to assess potential damage to communities from natural hazards and anthropogenic events (Perles Roselló et al., 2009). In contaminated aquifers, social vulnerability can measure how the waterborne diseases outbreak has adverse health effects to the community (Uzcategui-Salazar & Lillo, 2023). Studies have shown the social vulnerability of polluted groundwater environment by arsenic in Ganga-Brahmaputra-Meghna Basin, indicating the importance of poverty, unemployment, social instability, and disabilities in adaptive capacity to mitigate the effects caused by contaminated indicators (Biswas et al., 2022). Social vulnerability of diarrheal diseases led by E. *coli* suggested the distances to water bodies and local finances of communities are key factors to alleviate the negative disease outbreaks effects in tropical West Africa (Robert et al., 2021). Social vulnerability has also been used in different environmental hazards. The approaches incorporate factors of social vulnerability by assigning weighting and rating values to describe and map the groundwater vulnerability (Orellana-Macías & Perles Roselló, 2022; Perles Roselló et al., 2009). Ho combined the subjective and objective environmental factor measurements together to map the environmental vulnerability in Hong Kong (Ho et al., 2019). The same social vulnerability index system compared to the Centers for

Disease Control Social Vulnerability Index (CDC SVI, simply SVI) for managing the epidemic in India showed the relationship between social vulnerability and Covid19 positive cases (Acharya & Porwal, 2020).

Reviewing the revealed works in this field shows that lack of research has considered the long-term trend of E. coli health burden and its association with social vulnerability at national scale. So, the first goal of this study is to use QMRA as a health risk evaluation method to investigate the spatial differences of dynamic *E. coli* exposures in local rivers and streams for recreational use across the U.S. The second goal is to conduct a trend analysis to evaluate potential increasing or decreasing patterns of stream *E. coli* waterborne disease burden across the U.S. The third goal aims to find regional differences of *E. coli* infectious probability by Social Vulnerability Index. Specifically, we implemented four analytical steps(1) obtain national stream *E. coli* data from a public database for modeling the continuous *E. coli* concentration from 1987 to 2022; (2) measure the annual risk levels of *E. coli* exposure during water-related recreational activities; (3) analyze the trend changes of annual E. coli health burden within 5 geographical regions, including Northeast, Southeast, Midwest, Southwest, and West; and (4) identify which regions' social vulnerability is low by *E. coli* waterborne diseases risk level. Based on this, we then suggest public health interventions and environmental management strategies to the high risk and high vulnerability regions to better mitigate the adverse effects on residents' health.

Chapter 2: Method

2.Method

2.1 E. coli Sampling Sites Selection

We acquired a comprehensive dataset for *E. coli* sites and discharge stations from the U.S. Water Quality Portal. The discharge data was obtained together with *E. coli* concentration data to interpolate a complete time series from sparse *E. coli* concentration data. Initially, the acquisition of *E. coli* sites and discharge stations was implemented by the 'dataRetrieval' package in R. The inclusion criterion for *E. coli* sites was a minimum of 100 data records since 1987 (Hirsch et al., 2010), resulting in 2,903 sites. Concurrently, discharge stations were obtained from the GAGES II dataset on the USGS website, assembling a dataset of 9,322 discharge stations.

Then, we matched the *E. coli* and discharge stations to facilitate data interpolation. The initial pairing was executed based on unique USGS_IDs, resulting in 173 matched pairs. Further refinement was implemented by assigning HUC12_IDs and creating 1 km buffer zones around sites. If the station is located within the buffer of the site, shares the same HUC12_ID as the site, and is the closest station to the site, then the discharge station is considered a match for this *E. coli* site, matching an additional 529 pairs. Combining both matching approaches, a total of 624 pairs of *E. coli* sites and discharge stations were identified.

Based on the matched pairs of discharge stations and *E. coli* sites, we retrieved data using the 'EGRET' package in R. Rigorous cleaning procedures were applied to ensure the quality and integrity of the dataset. Specifically, we included pairs where the *E. coli* site possesses a minimum of 100 data records spanning a period of a decade or more, resulting in a refined dataset comprising 506 pairs. Subsequent filtering ensured that the time range of the *E. coli* data at a given site was entirely covered by the corresponding station's discharge data, ultimately yielding a final dataset comprising 452 pairs of *E. coli* sites and discharge stations in 31 states in the U.S.

2.2 Generating Continuous Time Series of E. coli Concentration

We used Weighted Regressions on Time, Discharge, and Season (WRTDS), a method for estimating daily water-quality data sets, to generate continuous *E. coli* time series (Hirsch et al., 2010). WRTDS is commonly used to characterize the status and trends in concentration and flux of pollutant indicators. This method is distinguished by its ability to construct a highly flexible statistical representation of expected concentration values for each day within the period of record. Utilizing this representation, the WRTDS method generates four essential daily time series for the given period. These include daily concentration, daily flux, flow-normalized daily concentration, and flow-normalized daily flux. The outcome is a comprehensive set of time-dependent metrics that facilitates a nuanced understanding of *E. coli* dynamics, allowing for the characterization of concentration and flux trends over the designated time frame. The WRTDS weighted regression model is estimated as follows (Eq. 1):

$$ln(c) = \beta_0 + \beta_1 q + \beta_2 T + \beta_3 sin(2\pi T) + \beta_4 cos(2\pi T) + \varepsilon \quad (1)$$

where

c is concentration, in m/l,

β are the regression coefficients,

q is $ln(Q)$ where Q is daily mean discharge, in m^3/s,

T is time, in decimal years, and

ε is the error (unexplained variation).

2.3 Quantitative Microbial Risk Assessment

We used Quantitative Microbial Risk Assessment (QMRA) to estimate the annual health risk associated with stream water *E. coli* exposure by evaluating the likelihood of *E. coli* infections due to recreational water activity exposure. Hazard identification is the first step in QMRA. This procedure aims to determine the pathogen of interest. In this study, we used *E. coli* as the infectious agent. Since not all *E. coli* strains are harmful, we used the pathogenic strain *E. coli O157:H7* to be the indicator in QMRA analysis. *E. coli O157:H7* is a strain of *E. coli* bacteria that can cause severe foodborne illness in humans (Pang et al., 2017). It produces a toxin called Shiga toxin, which can lead to complications such as bloody diarrhea, kidney failure, and even death in some cases. The ratio to estimate the overall *E. coli O157:H7* from the observed *E. coli* concentration is 0.08 (Machdar et al., 2013). The concentration of *E. coli* in the water environment is often right-skewed (Corradini et al., 2001). Using a log-normal distribution for *E. coli* can accommodate the skewness and ensure the output is positive in the process of QMRA (Goh et al., 2023). We used the annual continuous *E. coli* data generated from WRTDS to represent long-term bacteria concentrations from different locations and approximate it with log-normal distribution.

Exposure assessment is the second step in QMRA, which estimates the amount of exposure between humans and contamination. It was presumed that exposure to the harmful strains in recreational water predominantly occurred through ingesting water while swimming. We estimated the volume of water ingested by adults during a swimming session using a triangular distribution, with the highest likelihood at 16 mL, a minimum of 5 mL, and a maximum of 53 mL (Dufour et al., 2006). Then, given a known dose of a pathogen, we could estimate the infectious probability of the responses in dose-response analysis. For *E. coli*, the commonly used dose-response model is beta-Poisson model. By combining this ingestion rate with the bacteria concentrations in the water surfaces, we calculated the distribution of exposed doses (Eq.2):

$$D_{oral} = I_{oral} \times C \qquad (2)$$

Here, D_{oral} represents the ingested dose of *E. coli*, where I_{oral} represents the volume of freshwater consumed in milliliters (mL) by individuals and C is the concentration of *E. coli* in the freshwater in colony-forming units per milliliter (CFU/mL). Subsequently, the likelihood of *E. coli*-related infections was calculated by integrating dose-response analysis and exposure assessment (Eq.3):

$$P_{inf(E.coli)} = 1 - [1 + (\frac{dose}{N_{50}})(2^{\frac{1}{\alpha}} - 1)]^{-\alpha} \qquad (3)$$

In this equation, $P_{inf(E.coli)}$ represents the probability of infection stemming from *E. coli* exposure; $dose$ is the quantity of ingested *E. coli* organisms in CFU; the median infective dose. We set the median infective dose N_{50} at 2.11 × 10^6 CFU, representing the threshold at which half of the population becomes infected (L et al., 1971). Additionally, the slope parameter α was established at 0.155 (Haas et al., 2014). It is important to note that this equation assumes all infections eventually result in a detected illness.

Monte Carlo simulation was employed to predict the likelihood of *E. coli*-related infections. This method involves exploring various combinations of variables within defined ranges or distributions, allowing for consideration of both variability and uncertainty effects (Bivins et al., 2017). In this study, a Monte Carlo simulation consisting of 10,000 trials was implemented using the mc2d package in RStudio software (version 4.2.1). Mean values of infection likelihood from each of the 10,000 simulations were used for further trend analysis.

2.4 Trend Analysis

To analyze the non-normally distributed time series health burden for temporal tendencies, we employ Mann-Kendall trend test. The purpose of this approach is to find if there exists a potential monotonic increasing or decreasing pattern of the health burden of each sampling site. This test is used because of its characteristic that no assumptions are needed about the sampling dataset (Kendall, 2015). The null hypothesis of this test is that there is no trend in the population from the sampling dataset and the records are independent and identically distributed. The MK statistics quantifies any trend which is present by testing whether the time series lies in the confidence interval defined for the null hypothesis of the significance level (Kumar et al., 2023).

The likelihood of *E. coli* infection for 452 sampling sites are respectively included in the trend analysis. As the constraints of our data, there are missing values in the recorded years from

1987 to 2022, which lead to the different start time and end time of each site for trend analysis. In this study, “trend” package is employed in RStudio software (version 4.2.1).

2.5 Social Vulnerability Index

Social Vulnerability Index is an open access data from Centers for Disease Control and Prevention and Agency for Toxic Substances and Disease Registry. It contains 4 themes of social vulnerability, including socioeconomic status, household characteristics, racial and minority status, and housing type and transportation. Socioeconomic status include population below 150% poverty, unemployed population, housing cost burden, population with no high school diploma, and population with no health insurance. Household characteristics contain a population of aged 65 older, population of aged 17 younger, civilian with a disability, single-parent households, and English language proficiency. Racial and minority status include all the population of different races. Housing type and transportation describes the multi-unit structures, mobile homes, crowding, population of no vehicle, and group quarters. Here, we use the SVI 2020 product to see how different themes perform on the health burden of *E. coli* by region because SVI 2020 includes more comprehensive variables of minority and household characteristics, compared to previous products. Also, due to the different start and end year of our health burden records, selecting SVI 2020 could include more effective records of infectious probability in our limited data. To best preserve the health burden availability in time series, we selected the average annual risk from 2015 to 2019. With the infectious likelihood of *E. coli* in 452 sampling sites, we selected the census tract where the site is located through spatial selection approach in ArcGIS Pro 2.7. If a census tract has more than two sites, we calculate the average health burden to represent the health risk for that tract. After relocating the sampling sites to the census tract, 406 census tracts are selected.

Chapter 3: Results and Discussion

3. Results and Discussion

3.1 E. coli Risk Level by Concentration and by Infectious Rate

Figure 1 delineates the dispersion of *E. coli* health risks level across the U.S. based on EPA recommended concentration threshold, revealing significant regional variances. We define below 126 CFU 100/ml as the low risk level, below 236 CFU 100/ml as the moderate risk level, and above 236 CFU 100/ml as the high-risk level. Not all states have valid recorded water quality data of *E. coli* available, which is why some states were not considered in our health risk assessment procedure. For instance, California, Nevada, Michigan lacked data. High *E. coli* risk level regions were mainly located in the Midwest, especially Minnesota Iowa, Wisconsin, and Missouri. Meanwhile, in the West region, Oregon and Colorado also have high risk level sampling sites. Most sites were presented in medium risk level by *E. coli* concentration, except for the Midwest region, Virginia and Massachusetts have dense moderate risk level sampling sites. Low risk regions are mainly located in the Southeast and the Southwest, particularly in Florida, Georgia and Texas. The distributions present unbalanced risk level patterns across the U.S., with notable hotspots in the Midwest and the South for higher stream *E. coli* concentration. Figure 2 displays the dispersion of health risks estimation results from 1987 to 2022 via QMRA across the U.S. Figure 1(a) shows the mean value of each site from the 10,000 stimulations in the Monte Carlo process, while Figure 1(b) represents the average infection likelihood of all sample sites within the state. High *E. coli* exposure regions were mainly located in the Midwest, South, and Northwest. Most sites were safe since the infectious probability was low enough, between 0.000001 to 0.000088 overall. The low-risk states were Oregon, Arkansas, and Oklahoma. The medium-high infectious probability regions were in the Midwest and South. For the Midwest, the states of Iowa, Kansas, Missouri, and South Dakota had a medium-high risk tendency. Iowa had a dense medium-high infectious probability towards E. coli, which centered around 0.000089 to 0.000681 in value. Missouri showed an instate North-South disparity, displaying medium-high risk near Iowa. In the South, Texas, Maryland, Tennessee, and Georgia had medium-high health burden sample sites too. In Texas, most high-risk regions were in the South, while others were scattered in low-risk regions. Though the total number of sample sites in Maryland, Georgia, and Tennessee were not as high as in other states, the calculated health burden was displayed as medium high. The distributions present unbalanced risk patterns across the U.S., with notable hotspots in the Midwest and the South for higher health burdens. These two risk maps show two different approaches to measure *E. coli* health burden nationwide. There exists a high-risk level by the bacteria concentration but low infectious probability or vice versa. This is potentially because the health burden from 1987 to 2022 is generated through Monte Carlo simulations and we use the mean from the percentiles of this output. Though the concentration is the mean from this time span, the concentration of each site is a specific value to categorize the risk levels rather than an interval. Meanwhile, the infectious probability is generated based on the simulated concentration of *E. coli* by WRTDS. The function of infectious rate is different from the concentration. Acknowledging these differences, the discernible trends emphasize a potential interplay between geographical, environmental, and socio-economic factors, as certain regions exhibit consistently different risk levels across the varying measurements.

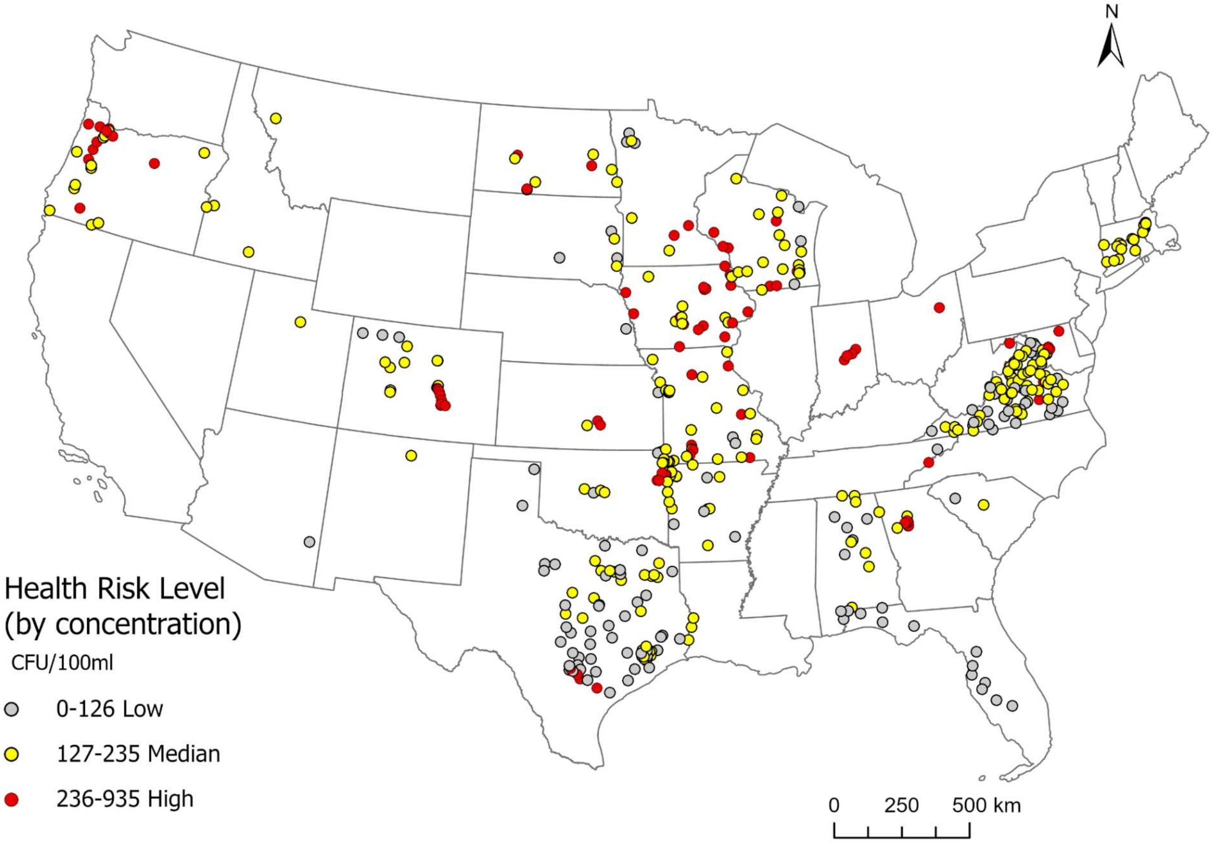

Figure 1. *Spatial distribution of the E. coli risk level on a granular level for sample sites in the United States*

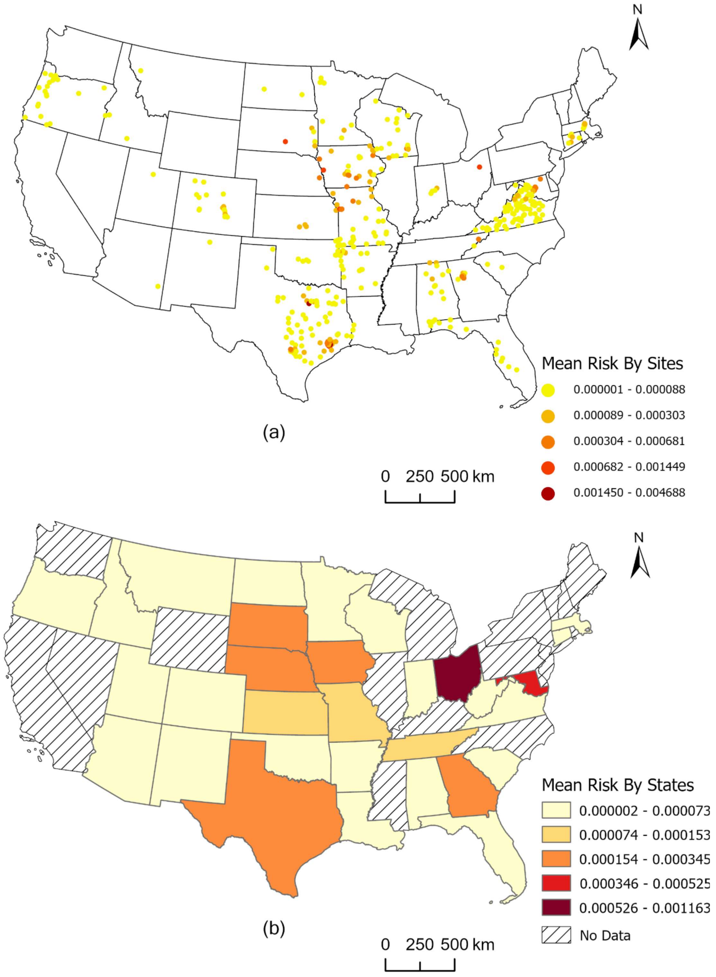

Figure 2. *Spatial distribution of the E. coli infection likelihood (a) on a granular level for sample sites, and (b) on a state-level average for the United States.*

3.2 The Association between Social Vulnerability and Health Burden by Regions

The analysis of regional disparities in social vulnerability reveals a clear association between elevated health risks due to *E. coli* contamination and heightened levels of social vulnerability across various dimensions, emphasizing the need for targeted mitigation strategies in both median and high-risk regions. Figure 3 delineates the regional differences in health risk levels across four dimensions of social vulnerability. A discernible trend is present in the regions with elevated health risks correspondingly exhibit heightened levels of social vulnerability.

Meanwhile, in some regions with median health risk levels, social vulnerability remains pronounced, such as the Northeast.

Among the various dimensions of social vulnerability assessed, socioeconomic status and household characteristics stand out for their evident association with elevated health burdens and heightened social vulnerability. This trend is particularly prominent in the Southwest, West, and Midwest regions, highlighting the critical need for interventions tailored to address these vulnerabilities. Furthermore, when examining racial status, housing type and transportation vulnerability, the Southwest region emerges as particularly vulnerable. The data suggest that *E. coli* burden issues may disproportionately affect local minority populations and poor housing status in this region, pointing towards environmental challenges that warrant focused attention and remedial actions.

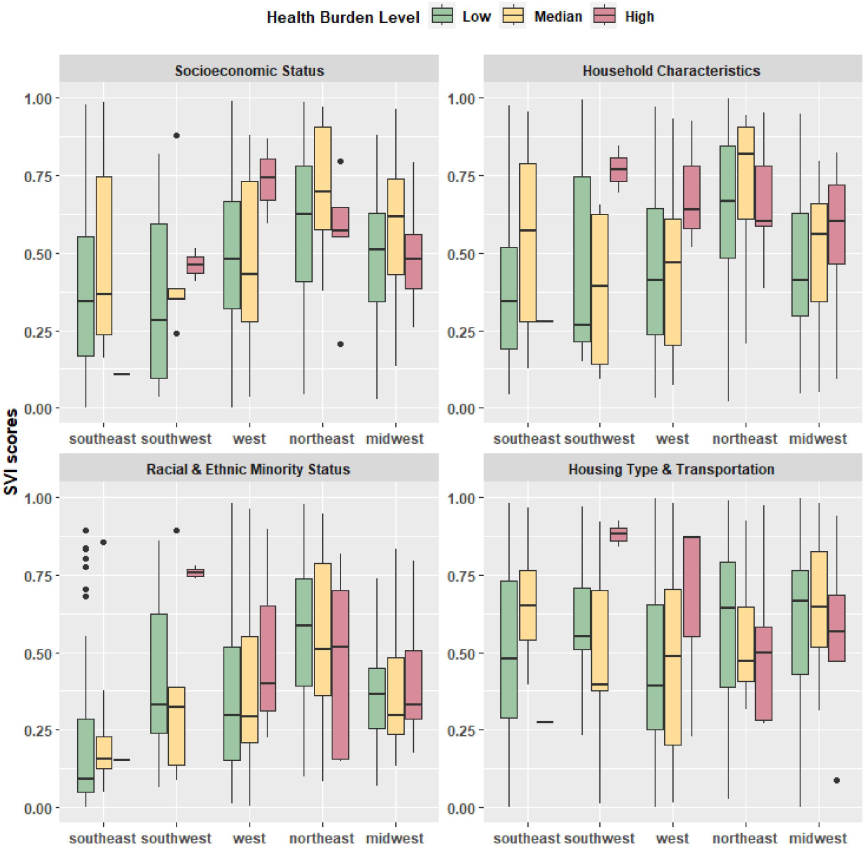

Figure 3. *Social vulnerability by health risk levels in 5 regions in the U.S. From upper left to lower right there are four different themes of SVI including socioeconomic status, household characteristics, racial and ethnic minority status, and housing type and transportation.*

3.3 Trend of E. coli Infectious Rate

Based on the results of the Mann-Kendall trend test, it was observed that out of the total 406 sampling sites assessed, 76 exhibited an increasing trend in *E. coli* levels, while 107 demonstrated a decreasing trend. The remaining sites did not show any significant temporal trend. On a national scale, the health risks associated with *E. coli* contamination generally manifest a decreasing trend. However, it is noteworthy that certain regions within the country display an opposing trend, indicating localized variations in water quality. Figure 4 displays the spatial distribution of the trend statuses across the sampling sites. Sampling sites exhibiting an increasing trend are predominantly situated in the Midwest region. Additionally, a subset of these sites is dispersed across several states, notably including Florida. In contrast, the sites with a decreasing trend are primarily located in the Southwest, and Oregon. States that encompass a notable number of sampling sites demonstrating both increasing and decreasing trends include Texas and Virginia. This dual trend suggests spatial heterogeneity in *E. coli* contamination levels within these states, warranting further investigation into the underlying factors influencing water quality variations.

Our analysis reveals an overlap between high-risk E. coli communities characterized by high social vulnerability. A notable observation from our study is the lag effect associated with health exposure (Martin et al., 2017). Even in areas with relatively low vulnerability, there exists a potential risk of increased health burden from *E. coli*. This delay in awareness and response mechanisms implies that communities may be unaware of escalating health risks until it reaches a critical threshold. Conversely, heightened awareness of *E. coli* contamination in an area does not necessarily correlate with an immediate increase in vulnerability. This interaction between awareness and vulnerability emphasizes the complexity of addressing health risks and necessitates proactive surveillance and public health interventions to address potential outbreaks in advance. Furthermore, the risk of *E. coli* contamination in stream water is also influenced by other indicators, which is shaped by a myriad of factors including anthropogenic activities, upstream conditions, and climatic variables. While anthropogenic factors contribute significantly to water quality degradation, natural factors such as upstream conditions and climate play a pivotal role in determining water quality. Importantly, these natural factors affect communities uniformly and do not disproportionately impact minority groups.

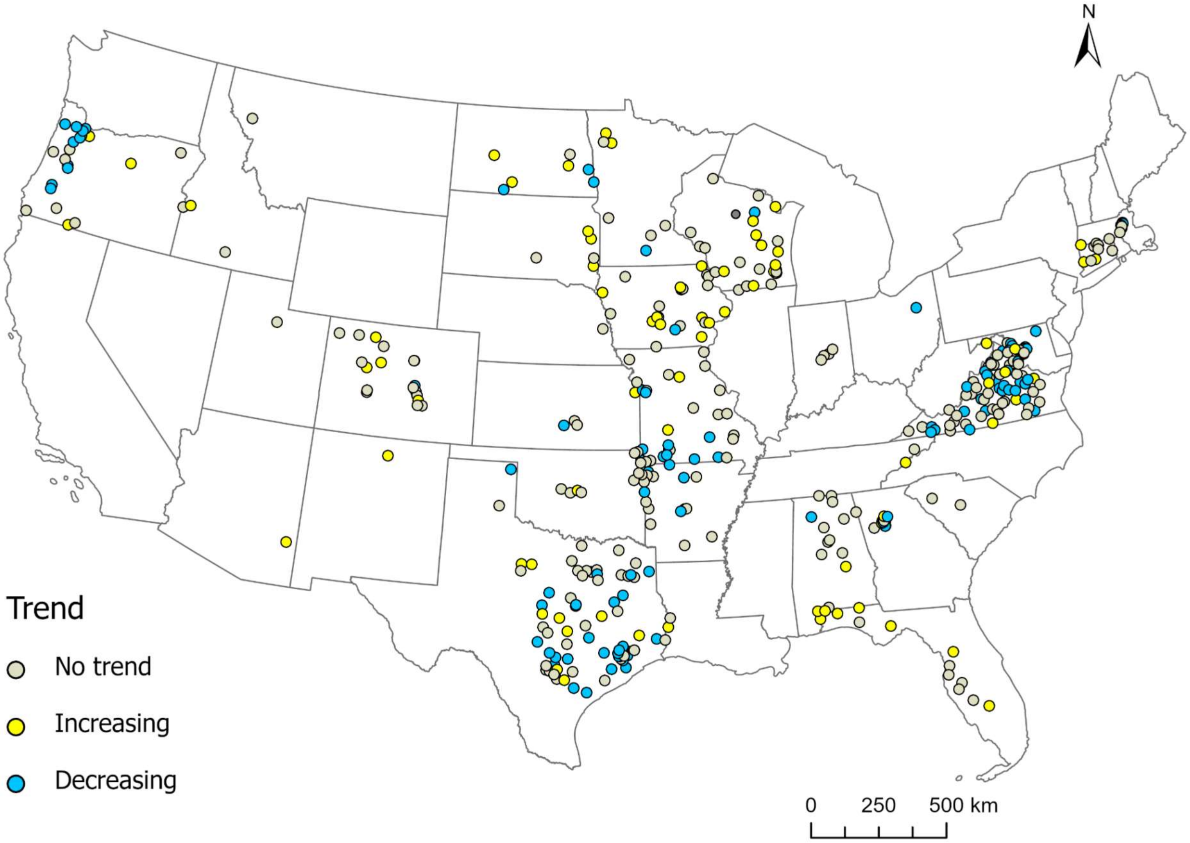

Figure 4. *Spatial distribution of the E. coli health burden trend status on a granular level for sample sites in the United States*

3.4 The Association between Trend of E. coli Infectious Rate and Social Vulnerability by Regions

The analysis of *E. coli* health burden trends reveals varying degrees of social vulnerability across regions, with distinct patterns in how vulnerability manifests across different themes of SVI. To explore potential relationships between trend and SVI, we use box plots to delineate the increasing and decreasing patterns in four themes of SVI by 5 regions. For the Midwest region, the increasing trend group is more vulnerable than the decreasing trend group in household characteristics and housing type and transportation, whereas less vulnerable in socioeconomic status and minority status. For the Northeast region, the decreasing trend group is considerably more vulnerable than the increasing trend group in four aspects of social vulnerability. In the Southeast and Southwest, vulnerability is more severe in the increasing category compared to the decreasing category, except for minority status for the Southwest. In the West, the social vulnerability in four themes of the rising group and the falling group are basically the same.

These findings underscore the intricate relationships between *E. coli* health burden trends and social vulnerabilities across different regions. The disparities observed highlight the need for region-specific interventions that address both environmental and social determinants of health.

The heightened health risk trend observed in certain regions, particularly in the Midwest simultaneously emphasizes the multifaceted social vulnerability aspects. This may relate to the non-point source pollution from the dense agricultural practice. This indicates that it is urgent for local governments to mitigate the *E. coli* exposure, especially for those vulnerable populations, such as the younger and the older populations. Addressing these disparities requires a holistic approach that integrates environmental remediation with targeted social interventions to ensure equitable health outcomes across all communities.

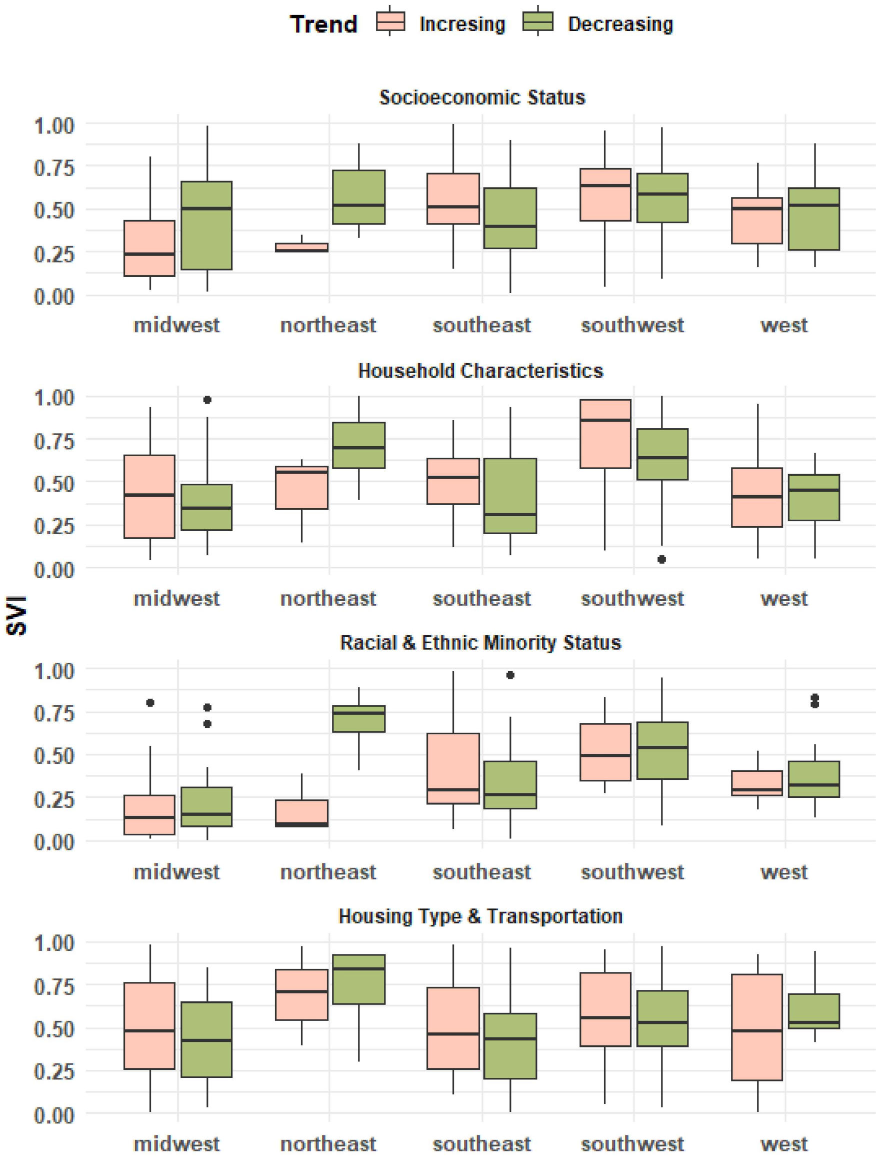

Figure 5. *Social vulnerability index by trend in 5 regions in the United States. From top to bottom, the themes of SVI are socioeconomic status, household characteristics, racial and ethnic status, and housing type and transportation.*

Chapter 4: Study Limitations

4. Study Limitations

The major limitation of this study is the absence of water quality data in some states. Insufficient number of sample sites within each state lead to the corresponding health risks being underreported. Besides, because the sample sites of *E. coli* data and sample sites of discharge data need to be matched in the pre-process of WRTDS, the total number of useful sites are reduced during any matching process. Hence, there is potential bias in *E. coli* health risk analysis for some states. For instance, Virginia and Texas had over 70 sample site collections to analyze the risk, while states such as Ohio only have one sampling site to be analyzed in QMRA. Therefore, we cannot cover the overall population health risks among all the state streams.

Our study is based on the assumption that people engaging in recreational water activity are therefore exposed to the infectious risk *E. coli*, predominantly through ingestion. This assumption may oversimplify the complexities of water-related recreational activities, as individuals may be exposed to microbial contaminants through multiple routes, such as inhalation and dermal contact. Also, for the QMRA process, the existing dose-response relationships may not adequately reflect the infectivity of *E. coli* as recently suggested by Goh (Goh et al., 2023). And we used a triangle distribution of ingestion water volume based on exposure scenarios articulated in the previous study (Dufour et al., 2006). This probability distribution is not likely to represent all water consumption behavior during the water-related recreational activity. Factors such as activity intensity, duration, and personal preference in water may vary differently among individuals, leading to potential inaccuracies in the estimating exposure levels (Boehm et al., 2018).

Chapter 5: Conclusion

5.Conclusion

In summary, this study used epidemiological and statistical methods to analyze the spatial distribution patterns of *E. coli* infectious probability of local surface water for recreational water activity use. We proposed a framework to integrate modeled *E. coli* concentration with a QMRA approach at national scale in the United States. Specifically, we used continuous *E. coli* modeled data to represent the concentration of pathogen distribution of each sample site with WRTDS. This allowed each site to have its own infectious probability associated with the concentration of a reference pathogen. Then, we integrated exposure assessment and existing dose-response relationship for *E. coli* by QMRA. Furthermore, our study analyzed the health burden trend of *E. coli* through Mann-Kendall test to find if there are significant increasing or decreasing patterns for each sampling site.

The results highlighted the long-term trend patterns from 1987 to 2022 in national scale. Different patterns of health burden by SVI were displayed across the Midwest, Southwest, Southeast, West and Northeast regions. The overall trend of infectious probability is decreasing across the states. However, a particular region like Midwest faces the increasing health burden by *E. coli* contamination with the high social vulnerability in local communities. In the specific aspects of vulnerability, socioeconomic status and household characteristics are evident in association with the heightened infectious risks in the Midwest, Southeast and Southwest. Ultimately, our results can provide the foundation for a comprehensive risk framework for policy makers and management to mitigate water health inequity problems. Further research should focus on health equity issues in hotspot regions by elucidating the relationship between high health burden and socio-economic factors.

References

Abia, A. L. K., Ubomba-Jaswa, E., Genthe, B., & Momba, M. N. B. (2016). Quantitative microbial risk assessment (QMRA) shows increased public health risk associated with exposure to river water under conditions of riverbed sediment resuspension. *Science of the Total Environment*, *566–567*, 1143–1151. https://doi.org/10.1016/j.scitotenv.2016.05.155

Acharya, R., & Porwal, A. (2020). A vulnerability index for the management of and response to the COVID-19 epidemic in India: an ecological study. *The Lancet Global Health*, *8*(9), e1142–e1151. https://doi.org/10.1016/S2214-109X(20)30300-4

Barragán, J. L. M., Cuesta, L. D. I., & Susa, M. S. R. (2021). Quantitative microbial risk assessment to estimate the public health risk from exposure to enterotoxigenic E. coli in drinking water in the rural area of Villapinzon, Colombia. *Microbial Risk Analysis*, *18*, 100173. https://doi.org/10.1016/J.MRAN.2021.100173

Beaudequin, D., Harden, F., Roiko, A., Stratton, H., Lemckert, C., & Mengersen, K. (2015). Beyond QMRA: Modelling microbial health risk as a complex system using Bayesian networks. *Environment International*, *80*, 8–18. https://doi.org/10.1016/J.ENVINT.2015.03.013

Biswas, S., Sahoo, S., & Debsarkar, A. (2022). Social Vulnerability of Arsenic Contaminated Groundwater in the Context of Ganga-Brahmaputra-Meghna Basin: A Critical Review. In P. K. Shit, H. R. Pourghasemi, G. S. Bhunia, P. Das, & A. Narsimha (Eds.), *Geospatial Technology for Environmental Hazards: Modeling and Management in Asian Countries* (pp. 39–61). Springer International Publishing. https://doi.org/10.1007/978-3-030-75197-5_3

Bivins, A. W., Sumner, T., Kumpel, E., Howard, G., Cumming, O., Ross, I., Nelson, K., & Brown, J. (2017). Estimating Infection Risks and the Global Burden of Diarrheal Disease Attributable to Intermittent Water Supply Using QMRA. *Environmental Science and Technology*, *51*(13), 7542–7551. https://doi.org/10.1021/acs.est.7b01014

Boehm, A. B., Graham, K. E., & Jennings, W. C. (2018). Can We Swim Yet? Systematic Review, Meta-Analysis, and Risk Assessment of Aging Sewage in Surface Waters. In *Environmental Science* *and Technology* (Vol. 52, Issue 17, pp. 9634–9645). American Chemical Society. https://doi.org/10.1021/acs.est.8b01948

Brouwer, A. F., Masters, N. B., & Eisenberg, J. N. S. (2018). Quantitative Microbial Risk Assessment and Infectious Disease Transmission Modeling of Waterborne Enteric Pathogens. *Current Environmental Health Reports*, *5*(2), 293–304. https://doi.org/10.1007/s40572-018-0196-x

Cho, S., Hiott, L. M., Barrett, J. B., McMillan, E. A., House, S. L., Humayoun, S. B., Adams, E. S., Jackson, C. R., & Frye, J. G. (2018). Prevalence and characterization of Escherichia coli isolated from the Upper Oconee Watershed in Northeast Georgia. *PLOS ONE*, *13*(5), e0197005-. https://doi.org/10.1371/journal.pone.0197005

Corradini, M. G., Horowitz, J., Normand, M. D., & Peleg, M. (2001). Analysis of the fluctuating pattern of E. coli counts in the rinse water of an industrial poultry plant. *Food Research International*, *34*(7), 565–572. https://doi.org/https://doi.org/10.1016/S0963-9969(01)00073-4

Corsi, S. R., Borchardt, M. A., Carvin, R. B., Burch, T. R., Spencer, S. K., Lutz, M. A., McDermott, C. M., Busse, K. M., Kleinheinz, G. T., Feng, X., & Zhu, J. (2016). Human and Bovine Viruses and Bacteria at Three Great Lakes Beaches: Environmental Variable Associations and Health Risk. *Environmental Science & Technology*, *50*(2), 987–995. https://doi.org/10.1021/acs.est.5b04372

Dorevitch, S., DeFlorio-Barker, S., Jones, R. M., & Liu, L. (2015). Water quality as a predictor of gastrointestinal illness following incidental contact water recreation. *Water Research*, *83*, 94–103. https://doi.org/10.1016/j.watres.2015.06.028

Dufour, A. P., Evans, O., Behymer, T. D., & Cantú, R. (2006). Water ingestion during swimming activities in a pool: A pilot study. *Journal of Water and Health*, *4*(4), 425–430. https://doi.org/10.2166/wh.2006.0026

EPA, 2012. Recreational Water Quality Criteria. US Environmental Protection Agency, Washington, DC, 820-F-12-058.

Goh, S. G., Haller, L., Ng, C., Charles, F. R., Jitxin, L., Chen, H., He, Y., & Gin, K. Y. H. (2023). Assessing the additional health burden of antibiotic resistant Enterobacteriaceae in surface waters through an integrated QMRA and DALY approach. *Journal of Hazardous Materials*, *458*, 132058. https://doi.org/10.1016/J.JHAZMAT.2023.132058

Haas, C. N., Rose, J. B., & Gerba, C. P. (2014). *Quantitative microbial risk assessment*. John Wiley & Sons.

Hart, C. A., Batt, R. M., & Saunders, J. R. (1993). Diarrhoea caused by Escherichia coli. *Annals of Tropical Paediatrics*, *13*(2), 121–131. https://doi.org/10.1080/02724936.1993.11747636

Hirsch, R. M., Archfield, S. A., & De Cicco, L. A. (2015). A bootstrap method for estimating uncertainty of water quality trends. Environmental Modelling & Software, 73, 148-166.

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 Inputs1. *JAWRA Journal of the American Water Resources Association*, *46*(5), 857–880. https://doi.org/https://doi.org/10.1111/j.1752-1688.2010.00482.x

Ho, H. C., Wong, M. S., Man, H. Y., Shi, Y., & Abbas, S. (2019). Neighborhood-based subjective environmental vulnerability index for community health assessment: Development, validation and evaluation. *Science of The Total Environment*, *654*, 1082–1090. https://doi.org/10.1016/J.SCITOTENV.2018.11.136

Jacob, P., Henry, A., Meheut, G., Charni-Ben-Tabassi, N., Ingr, V., & Helmi, K. (2015). Health risk assessment related to waterborne pathogens from the river to the tap. *International Journal of Environmental Research and Public Health*, *12*(3), 2967–2983. https://doi.org/10.3390/ijerph120302967

Kendall, M. (2015). Trend analysis of Pahang river using non-parametric analysis: Mann Kendall's trend test. *Malays. J. Anal. Sci*, *19*, 1327–1334.

Kumar, S., Ahmed, S. A., & Karkala, J. (2023). Time series data and rainfall pattern subjected to climate change using non-parametric tests over a vulnerable region of Karnataka, India. *Journal of Water and Climate Change*, *14*(5), 1532–1550. https://doi.org/10.2166/wcc.2023.441

L, D. H., B, F. S., B, H. R., J, S. M., P, L. J., G, S. D., H, L. E., & P, K. J. (1971). Pathogenesis of Escherichia coli Diarrhea. *New England Journal of Medicine*, *285*(1), 1–9. https://doi.org/10.1056/NEJM197107012850101

Machdar, E., van der Steen, N. P., Raschid-Sally, L., & Lens, P. N. L. (2013). Application of Quantitative Microbial Risk Assessment to analyze the public health risk from poor drinking water quality in a low income area in Accra, Ghana. *Science of The Total Environment*, *449*, 134–142. https://doi.org/10.1016/J.SCITOTENV.2013.01.048

Martin, S. L., Hayes, D. B., Kendall, A. D., & Hyndman, D. W. (2017). The land-use legacy effect: Towards a mechanistic understanding of time-lagged water quality responses to land use/cover. *Science of The Total Environment*, *579*, 1794–1803. https://doi.org/10.1016/J.SCITOTENV.2016.11.158

Masciopinto, C., Vurro, M., Lorusso, N., Santoro, D., & Haas, C. N. (2020). Application of QMRA to MAR operations for safe agricultural water reuses in coastal areas. *Water Research X*, *8*, 100062. https://doi.org/10.1016/J.WROA.2020.100062

O’Flaherty, E., Solimini, A., Pantanella, F., & Cummins, E. (2019). The potential human exposure to antibiotic resistant-Escherichia coli through recreational water. *Science of The Total Environment*, *650*, 786–795. https://doi.org/10.1016/J.SCITOTENV.2018.09.018

Orellana-Macías, J. M., & Perles Roselló, M. J. (2022). Assessment of Risk and Social Impact on Groundwater Pollution by Nitrates. Implementation in the Gallocanta Groundwater Body (NE Spain). *Water*, *14*(2). https://doi.org/10.3390/w14020202

Pang, H., Lambertini, E., Buchanan, R. L., Schaffner, D. W., & Pradhan, A. K. (2017). Quantitative Microbial Risk Assessment for Escherichia coli O157:H7 in Fresh-Cut Lettuce. *Journal of Food Protection*, *80*(2), 302–311. https://doi.org/https://doi.org/10.4315/0362-028X.JFP-16-246

Pasalari, H., Akbari, H., Ataei-Pirkooh, A., Adibzadeh, A., & Akbari, H. (2022). Assessment of rotavirus and norovirus emitted from water spray park: QMRA, diseases burden and sensitivity analysis. *Heliyon*, *8*(10). https://doi.org/10.1016/j.heliyon.2022.e10957

Perles Roselló, M. J., Vías Martinez, J. M., & Andreo Navarro, B. (2009). Vulnerability of human environment to risk: Case of groundwater contamination risk. *Environment International*, *35*(2), 325–335. https://doi.org/10.1016/J.ENVINT.2008.08.005

Robert, E., Grippa, M., Nikiema, D. E., Kergoat, L., Koudougou, H., Auda, Y., & Rochelle-Newall, E. (2021). Environmental determinants of E. coli, link with the diarrheal diseases, and indication of vulnerability criteria in tropical West Africa (Kapore, Burkina Faso). *PLOS Neglected Tropical Diseases*, *15*(8), e0009634-. https://doi.org/10.1371/journal.pntd.0009634

Rodrigues, V. F. V., Rivera, I. N. G., Lim, K. Y., & Jiang, S. C. (2016). Detection and risk assessment of diarrheagenic E. coli in recreational beaches of Brazil. *Marine Pollution Bulletin*, *109*(1), 163–170. https://doi.org/10.1016/J.MARPOLBUL.2016.06.007

Rossi, A., Wolde, B. T., Lee, L. H., & Wu, M. (2020). Prediction of recreational water safety using Escherichia coli as an indicator: case study of the Passaic and Pompton rivers, New Jersey. *Science of The Total Environment*, *714*, 136814. https://doi.org/https://doi.org/10.1016/j.scitotenv.2020.136814

Tousi, E. G., Duan, J. G., Gundy, P. M., Bright, K. R., & Gerba, C. P. (2021). Evaluation of E. coli in sediment for assessing irrigation water quality using machine learning. *Science of The Total Environment*, *799*, 149286. https://doi.org/10.1016/J.SCITOTENV.2021.149286

Uprety, S., Dangol, B., Nakarmi, P., Dhakal, I., Sherchan, S. P., Shisler, J. L., Jutla, A., Amarasiri, M., Sano, D., & Nguyen, T. H. (2020). Assessment of microbial risks by characterization of Escherichia coli presence to analyze the public health risks from poor water quality in Nepal. *International Journal of Hygiene and Environmental Health*, *226*, 113484. https://doi.org/10.1016/J.IJHEH.2020.113484

Uzcategui-Salazar, M., & Lillo, J. (2023). Assessment of social vulnerability to groundwater pollution using K-means cluster analysis. *Environmental Science and Pollution Research*, *30*(6), 14975–14992. https://doi.org/10.1007/s11356-022-22810-6