Linking fecal bacteria in rivers to landscape, geochemical, and hydrologic factors and sources at the basin scale

Marc P. Verhougstraetea,†, Sherry L. Martinb, Anthony D. Kendalld, David W. Hyndmanb, and Joan B. Rosear

*Department of Fisheries and Wildlife, Michigan State University, East Lansing, MI 48824; and †Department of Geological Sciences, Michigan State University, East Lansing, MI 48824

Edited by Rita R. Colwell, University of Maryland, College Park, MD, and approved June 29, 2015 (received for review August 15, 2014)

Linking fecal indicator bacteria concentrations in large mixed-use watersheds back to diffuse human sources, such as septic systems, has met limited success. In this study, 64 rivers that drain 84% of Michigan’s Lower Peninsula were sampled under baseflow conditions for Escherichia coli, Bacteroides thetaiotaomicron (a human source-tracking marker), landscape characteristics, and geochemical and hydrologic variables. E. coli and B. thetaiotaomicron were routinely detected in sampled rivers and an E. coli reference level was defined (1.4 log10 most probable number/100 mL−1). Using classification and regression tree analysis and demographic estimates of wastewater treatments per watershed, septic systems seem to be the primary driver of fecal bacteria levels. In particular, watersheds with more than 1,621 septic systems exhibited significantly higher concentrations of B. thetaiotaomicron. This information is vital for evaluating water quality and health implications, determining the impacts of septic systems on watersheds, and improving management decisions for locating, constructing, and maintaining on-site wastewater treatment systems.

Escherichia coli | Bacteroides thetaiotaomicron | baseflow | reference conditions | septic system

Water quality degradation influenced by diffuse sources at large watershed scales has been difficult to describe. Human modifications of natural landscapes can permanently alter hydrologic cycles and affect water quality (1, 2). Deforestation (3) and increased impervious surface area (4) have been linked with decreased infiltration and thus increased surface runoff. Overland flows concentrate pollutants and rapidly transport them down gradient where they eventually enter surface water systems and affect water quality (5, 6). A number of models have been developed to calculate overland and surface water flows (7, 8) and nutrient/chemical transport (9), but few studies have focused on microbial movement from land to water, particularly nontraditional fecal indicator bacteria that can be used to track human sources of pollution.

Microbial contamination poses one of the greatest health risks to swimming areas, drinking water intakes, and fishing/shellfish harvesting zones where human exposures are highest (10–12). These highly visible areas often receive more attention than sources of contamination because identifying the origin of pollution in complex watersheds requires costly comprehensive investigation of environmental and hydrologic conditions across temporal and spatial scales (13). Grayson et al. (14) suggest using a “snapshot” approach that captures water quality characteristics at a single point in time across broad areas to provide information frequently missed during routine monitoring. Compared with long-term comprehensive investigations, the snapshot approach reduces the number of samples, cost, and personnel required to examine pollution sources.

Escherichia coli concentrations are commonly used to describe the relative human health risk during water quality monitoring in lieu of pathogen detection. Studies attempting to trace pollution in water back to a specific land use with E. coli have rarely produced definitive conclusions (15, 16). Using molecular approaches, specific source targets can be isolated in complex systems and have recently been used to investigate land use and water quality impairments (17). Furtula et al. (18) demonstrated ruminant, pig, and dog fecal contamination in an agriculturally dominated watershed (Canada) using Bacteroides markers. The Bacteroides thetaiotaomicron α−1–6 mannanase (B. theta) gene has a high human specificity (19–22), but no studies to date have linked its presence to land use patterns.

Reference conditions have been established for minimally disturbed environments based on measurements of macroinvertebrates, fish, and diatoms (23–25), but microbial reference conditions have not been adequately explored or defined. Based on 15 unimpaired California streams, microbial reference conditions for E. coli [1.0 log10 most probably number (MPN)-100 mL−1] and enterococci (1.2 log10 MPN-100 mL−1) were defined as being below state water quality thresholds (26). In the Great Lakes, a human health threshold of 2.57 log10 E. coli MPN-100 mL−1 (27), or a level equally protective of human health, has been adopted by all state governments. However, this health-assOCIated reference level was derived from epidemiological studies undertaken at beaches throughout the United States (28, 29) with limited knowledge of local implications. In response to water quality degradation from human stressors and the poorly understood microbial conditions in large-scale fresh water systems such as the Great Lakes basin, this paper aims to (i) examine the spatial distribution of E. coli and a human specific source marker (B. theta) in 64 river systems that drain most of the state’s Lower Peninsula under baseflow conditions, (ii) identify baseflow reference levels of fecal contamination in rivers, and (iii) determine how key chemical, physical, and biological factors influence fecal bacteria concentrations.

Significance

New microbial source-tracking tools can be used to elucidate important nonpoint sources of water quality degradation and potential human health risks at large scales. Pollution arising from septic system discharges is likely more important than previously realized. Identifying these sources and providing reference levels for water quality provides a basis to assess water quality trends and ultimately remediate degraded areas.


The authors declare no conflict of interest.

†This Direct Submission article had a prearranged editor.

†To whom correspondence should be addressed. Email: mverhougstraete@email.arizona.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1415836112/-/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.1415836112

PNAS Early Edition | 1 of 6
Results and Discussion
To address fecal microbial water quality impairment, this study examined fecal bacteria source tracking across a large spatial scale with classification and regression tree (CART) statistical method to link fecal contamination in rivers to landscape, geochemical, and hydrologic factors as well as potential human fecal sources such as septic systems and sewage effluent at the basin scale. The B. theta results suggest human fecal contamination was affecting 100% of the studied river systems. These results have significant implications for water and environmental quality managers. Further details on hydrologic, geochemical, and land use characteristics, as well as a CART analysis of the reduced dataset, are described in SI Materials and Methods.

Microbial Water Quality and Reference Conditions. This project measured E. coli and B. theta concentrations in 64 rivers under baseflow conditions. Across all sites, E. coli concentrations ranged from 0.3 to 3.0 log_{10} MPN 100 mL^{-1} (geometric mean of 1.4 log_{10} MPN 100 mL^{-1}) and B. theta ranged from 4.2 to 5.9 log_{10} cell equivalents (CE) 100 mL^{-1} (geometric mean of 5.1 log_{10} CE 100 mL^{-1}). E. coli levels were below the detection limit (<1 MPN 100 mL^{-1}) in four rivers, whereas B. theta was detected in all samples (Table 1 and Table S1). Nine rivers (14% of sites) exceeded the US Environmental Protection Agency (USEPA) suggested E. coli criterion for safe contact (2.37 log_{10} MPN 100 mL^{-1}), ranging in concentrations from 2.4 to 3.0 log_{10} MPN 100 mL^{-1}. In these same nine rivers, B. theta concentrations ranged from 4.6 to 5.6 log_{10} CE 100 mL^{-1}. These nine E. coli values were significantly different (P < 0.001) from those of the other 55 sites, which had a geometric mean of 1.3 log_{10} MPN 100 mL^{-1}. In contrast, there was no statistically significant difference (P = 0.433) between B. theta concentrations from these two sets of sites.

E. coli concentrations (geometric mean of 1.4 log_{10} MPN 100 mL^{-1}) were generally below USEPA recreational water quality criteria and consistent with previously measured ranges in Great Lakes tributary rivers (30–32). A comprehensive review (33) found that E. coli levels in freshwater below 2.23 log_{10} MPN 100 mL^{-1} were associated with low relative risks of gastrointestinal illness for swimmers compared with nonswimmers. Because the E. coli geometric mean concentration observed in this study was below the safety level reported by Wade et al. (33), we suggest a reference condition for E. coli of 1.4 log_{10} MPN 100 mL^{-1} for Michigan’s Lower Peninsula rivers under baseflow conditions in the absence of recent storm runoff. Wade et al. (28) reported positive associations between occurrence of illness and molecularly detected Bacteroides at one Great Lakes beach with a geometric mean concentration of 3.08 log_{10} CE 100 mL^{-1}, while noting that the associations were statistically weak (P < 0.1). Yampara-Iquse et al. (19) reported B. theta levels ranged from 5.8 to 9.8 log_{10} copies 100 mL^{-1} in multiple urban, agricultural, and small-town creek systems that represented various levels of human impact. In the current study, B. theta concentrations (range = 4.2–5.9 log_{10} CE 100 mL^{-1}; geometric mean = 5.1 log_{10} CE 100 mL^{-1}) averaged 1.6 times higher than levels reported by Wade et al. (28) but slightly lower than those reported by Yampara-Iquse et al. (19). Establishing B. theta reference conditions for Michigan rivers under other flow conditions would require additional sample analysis and a greater understanding of the bacterial distributions because comparative B. theta datasets are relatively small relative to available E. coli data, a key aspect to defining reference levels (34). Reference levels are important for establishing acceptable levels of disturbances, defining long-term water quality changes, and supporting management decisions (34). Although the concept of a reference condition lies in the notion of minimal impact, it is recognized that few streams or rivers are truly unimpaired because most receive treated sewage effluent, and the current study supports this premise.

CART Analysis of Microbial Water Quality. A primary goal of this study was to address diffuse pollution sources, historically a significant challenge in managing water quality. Major sources of nutrient loads from point and nonpoint sources of contamination were previously examined for Michigan’s Lower Peninsula and shown to vary significantly between watersheds (35). The current study examines these drivers under baseflow conditions, where groundwater inputs dominate flows and wastewater effluent generally provides only a small fraction of total river discharges (Table S1). Effects of wastewater treatment plant (WWTP) effluent on microbial water quality were examined using multiple approaches (see Supporting Information for details), and it was ultimately determined that WWTP were not a driving factor of microbial water quality in the studied watersheds. Future analysis of the seasonal efficacy of WWTP could improve the understanding of wastewater impact on water quality by quantifying effluent discharge contributions in key urban areas.

The initial hypothesis of this research was that land use would best explain fecal bacterial concentrations in water. Instead, we found that land use characteristics such as septic systems and nutrients were the primary explanatory factors of microbial water quality. The influence of septic systems on microbial water quality, measured by E. coli, at a smaller watershed scale has also been reported in other regions (36, 37). In the current study, E. coli concentrations were linked primarily to total phosphorus and potassium. B. theta concentrations were primarily associated with the total number of septic systems in the watershed and within a 60-m buffer. Because WWTPs were not a driving factor of microbial water quality in the studied watersheds, these results indicate that under low flow conditions septic systems are a significant source of human fecal contamination to surface water in the studied watersheds.

CART analysis was used to evaluate the influence of the independent variables on E. coli and B. theta. Results from CART analyses for E. coli and B. theta concentrations at the full and reduced watersheds are summarized in Fig. 2 and Fig. S1, respectively. The CART outputs indicated complex causes of river water quality variability under baseflow conditions. For instance, E. coli concentrations at the full watershed scale were mainly related to total phosphorus (TP) concentrations, which is consistent with results by Carrillo et al. (38). TP concentrations accounted for 48% of E. coli variance with a threshold of 19.0 μg L^{-1}.

Fig. 1. (A) E. coli (log_{10} MPN 100 mL^{-1}) and (B) B. theta (log_{10} CE 100 mL^{-1}) concentrations measured in 64 rivers under baseflow conditions. Areas in black were not represented with samples.
Although TP is essential for bacterial growth, the authors acknowledge that treated wastewater effluent includes high levels of both E. coli and TP. However, as stated above, WWTPs were not a driving factor of microbial water quality in the studied watersheds. Phosphorus, like E. coli, may be derived from sediments in the rivers, soil, plants, animal wastes, or manure and thus, unlike the B. theta, is not exclusive to fecal pollution.

The full watershed CART outputs and correlation analysis indicated B. theta concentrations were strongly associated with total numbers of septic systems in the watershed (r = 0.364, P = 0.002) and in the 60-m buffer (r = 0.357, P = 0.004). B. theta concentrations were not correlated with septic system density in the watershed (P = 0.361) or in the 60-m buffer (P = 0.520). Interestingly, the total number of septic systems in the watershed accounted for 36% of the B. theta concentration variance with a threshold count of 1,622 systems per watershed, as shown in Figs. 2B and 3. The snapshot sampling strategy used in this study focused on a spatial composite of the watersheds near the drainage point toward the Great lakes. Thus, the total number of people on septic tanks equates to the level of feces entering each watershed, and these levels are potentially dominated by failing septic systems contributing high concentrations of bacteria to nearby water systems. A Michigan health department reported a 26% on-site wastewater failure rate during time of sale or transfer inspections that discharged an estimated 65,000 gallons of untreated fecal waste each year to nearby water bodies (39). Future watershed-based studies should include analysis of total septic systems in the watershed and septic density, because it would be possible to overlook failing septic systems if the sample size were small or the focus were only on septic density. Additional efforts aimed at the condition of septic systems, their ability to remove bacteria, and microbial transport to nearby surface waters are required.

The direct and significant correlation between estimated number of septic systems and the human-specific marker B. theta in water (Fig. 3) illustrates a major issue for water quality of Michigan’s streams and rivers, with an estimated 1.4 million on-site septic systems statewide (35, 40). In this study, the overall B. theta geometric mean was one log_{10} unit higher than secondary treated sewage effluent, whereas the highest measured concentrations were 1.5 log_{10} higher than biologically treated septage effluent (20). Interestingly, when the CART analysis considered the entire upstream drainage area, including lakes, 2.5 times fewer septic systems were required to produce B. theta levels similar to when these drainage areas were restricted to downstream of the nearest lake, potentially indicating increased failure rates of septic systems surrounding lakes compared with rivers (see Supporting Information for details). Habteselassie et al. (41) identified that surface water and groundwater near failing on-site wastewater treatment systems contained higher concentrations of E. coli and enterococci than water surrounding properly functioning on-site wastewater treatment systems (P < 0.001). Combined, these results illustrate the importance and need for responsible development and septic system maintenance along lake and river riparian zones to protect water quality. Future analysis should include incremental spatial assessment of B. theta with respect to septic systems in watersheds to assess the fate and transport of bacteria from septic systems and define their acute/chronic impacts on water quality.

E. coli and B. theta Z-scores [(observed – mean)/SD] were compared using CART, as shown in Fig. 4, to identify the characteristics that could differentiate between E. coli and B. theta concentrations. Positive values of the Z-score differences occur when E. coli concentrations are higher, relative to their population mean, than B. theta concentrations. Negative values imply the opposite, with relatively higher B. theta concentrations. In catchments with discharge <0.66 m s^{-1} and with fewer than 294 septic systems in the 60-m buffer, E. coli concentrations were much higher than those of B. theta. In contrast, B. theta concentrations were much higher than those of E. coli in rivers with discharge >0.66 m s^{-1}, particularly in catchments with dissolved organic carbon >5.4 μg L^{-1}. E. coli, which occurs in the feces of all warm-blooded mammals and birds, has been shown to persist and regrow in the environment under some conditions and has been associated with suspended particles that have low settling rates (42–45). Therefore, in watersheds with low discharge it is possible that E. coli can attach to particles and persist longer than B. theta, which is an anaerobic organism with a faster decay rate in rivers (46).

We compared the concentrations and loads of E. coli and B. theta across all sites (Fig. S2). No statistically significant relationship was identified between E. coli and B. theta concentrations (r = 0.18; P = 0.16). Bacterial entry to rivers during baseflow seems to be occurring from some of the same diffuse sources, including septic systems. The comparison of E. coli versus B. theta concentrations illustrated that each of these microorganisms was entering rivers from similar sources (i.e., diffuse sources such as septic systems) (Fig. 2). However, each organism was influenced by different environmental parameters as identified by the Z-score CART analysis (Fig. 4). E. coli was ubiquitous in most rivers and concentrations were primarily associated with TP and K levels. This study indicates that B. theta can be used as a source-tracking marker to investigate diffuse sources of human-derived contaminants from septic systems under baseflow hydrologic conditions at watershed scales.

![Figure 2](image-url)  
**Fig. 2.** CART analyses for (A) E. coli and (B) B. theta concentrations as dependent variables and land use, nutrient, chemical, hydrologic, and environmental parameters as independent variables in watersheds. PRE, proportion of reduction in error.

![Figure 3](image-url)  
**Fig. 3.** B. theta versus septic systems illustrating the CART output from the first split of Fig. 2B.
E. coli - B. theta Z-scores, log-transformed
Total PRE = 68.4%

\[ \begin{aligned}
\mu &= 0 \\
n &= 64
\end{aligned} \]
Stream Discharge < 0.66 m³/s  
\[ \text{PRE} = 57.3\% \]
\[ \begin{aligned}
\mu &= -0.76 \\
n &= 39
\end{aligned} \]
\[ \begin{aligned}
\mu &= 0.52 \\
n &= 15
\end{aligned} \]
\[ \begin{aligned}
\mu &= 1.19 \\
n &= 25
\end{aligned} \]
\[ \begin{aligned}
\mu &= -1.04 \\
n &= 24
\end{aligned} \]
Buffer Septic Count > 293.5  
\[ \text{PRE} = 6.3\% \]
\[ \begin{aligned}
\mu &= -0.32 \\
n &= 15
\end{aligned} \]
\[ \begin{aligned}
\mu &= 0.52 \\
n &= 9
\end{aligned} \]
\[ \begin{aligned}
\mu &= 1.57 \\
n &= 16
\end{aligned} \]
Increasing B. theta Increasing E. coli

Fig. 4. CART of E. coli and B. theta Z-scores illustrating conditions associated with different concentrations between these two microbes. PRE, proportion of reduction in error.

Conclusions

To address impaired waters and restore them to designated uses, the process for total maximum daily loads (TMDLs) has been developed under the Clean Water Act. According to Stiles (47) there are currently 65,000 TMDLs and 43,000 listings that need to be addressed. Many stretches of water systems are impaired due to fecal pollution and E. coli, but there have been no established approaches or tools to identify nonpoint sources. This study provides a path forward to assess and ultimately improve water quality at large scales. More importantly, this study provides reference conditions for a large number of watersheds that, in the event of major landscape disturbance, could be used to measure remediation progress. Using a synoptic sampling approach for regional water quality assessment, this study found that human fecal contamination was prevalent under baseflow conditions. Baseflow in the study watersheds was generally dominated by groundwater and not by wastewater treatment effluent. Results suggest a regional E. coli reference condition below the current USEPA freshwater recreational criterion could be established. However, identifying specific sources of fecal contamination in rivers cannot be achieved using ubiquitous bacteria, such as E. coli. Assessing water quality using solely E. coli may mislead water quality managers and severely limit the ability to remediate impaired waterways. However, microbial source-tracking markers, such as the human-specific B. theta marker, can provide a more refined tool to identify the impacts of nonpoint sources of human fecal pollution, which could help prioritize restoration activities that should be implemented at watershed scales. The high variability of water quality measurements illustrates complex relationships between bacteria and landscape, geochemical, and hydrologic properties. The influence of septic systems in riparian zones also indicates that additional localized control measures, including septic system maintenance and construction, should be implemented to protect water quality and human health.

Materials and Methods

Study Area. This study investigated 64 watersheds draining Michigan’s Lower Peninsula to the Great Lakes (Fig. 53). Watersheds were selected using the following criteria: (i) the 30 largest watersheds that represent >80% of Michigan’s Lower Peninsula land area and (ii) 34 smaller watersheds randomly selected across the state from locations near their outlets to the lake. All sampling sites were located at bridge crossings and selected on the criteria that each was reasonably accessible, had adequate flow, river water dominated discharge, and the maximum amount of upstream land use was captured while meeting the above criteria.

Water Sample Collection. A synoptic sampling scheme was used to capture water quality characteristics under a single flow condition (i.e., baseflow) across broad spatial areas (14). Compared with long-term comprehensive investigations, this approach reduces the number of samples, cost, and personnel resources required to address pollution sources while providing essential information missed during routine monitoring.

Grab samples were collected from each river sampling site between October 1–13, 2010, which was chosen as a groundwater-dominated baseflow period based on historical hydrographs and antecedent precipitation. Groundwater-driven baseflow is critical to the preservation of water quality and quantity in the Great Lakes and provides year-round support for aquatic habitats. Before sampling each watershed, meteorological conditions were monitored to ensure that no significant precipitation had occurred within several days and hydrographs from nearby US Geological Survey (USGS) stream gauges were inspected to check that sampled rivers were at baseflow. October was chosen for the sampling period because the late growing season baseflow period is least likely to have large variability in water quality because flows are dominated by groundwater in the region. There is variability in water quality between baseflow periods (i.e., fall versus summer), but this variability is small relative to the variability between baseflow and other periods due to overland flow and dilution effects (48, 49). Water temperature (degrees Celsius), specific conductance (microsiemens per centimeter), and dissolved oxygen (milligrams per liter) were measured using YSI 6000 Sonde (YSI Incorporated). Field samples were placed on ice in coolers and transported to Michigan State University for other analyses, including bacterial testing (described below) within 24 h.

Water Analysis. Each sample was assayed for water chemistry as summarized in Table S2. The methods for assaying chemicals and nutrients are described in Table S3. E. coli analyses were performed within 24 h of collection using the QIAamp DNA mini kit protocol. Quantitative PCR (qPCR) was performed on extracted DNA from 200 μL of the thawed pellet via QIAamp DNA mini kit protocol. Quantitative PCR (qPCR) was performed on extracted DNA following Yambara-Iisque et al. (19), specifically by filtering 900 mL of water through a 0.45-μm hydrophilic mixed cellulose esters filter. Each filter was placed into a 50-mL centrifuge tube containing 20 mL of sterile phosphate-buffered water, vortexed, and centrifuged (30 min, 4,000 × g, 21 °C). Eighteen microliters were decanted from the tube and the remaining sample and pellet were stored at −80 °C. DNA was extracted from 200 μL of the thawed pellet via QIAamp DNA mini kit protocol. Quantitative PCR (qPCR) was performed on extracted DNA following Yambara-Iisque et al. (19) with a probe modification (20) using a Roche Light-Cycler 2.0 Instrument (Roche Applied Sciences). Each B. theta assay was carried out with 10 μL of LightCycler 480 Probe Mastermix (Roche Applied Sciences), 0.4 μL forward and reverse primers, 0.2 μL probe 62 (5′FAM-ACCTGCTG-NFQ; Roche Applied Sciences Universal Probe Library), 1.0 μL BSA, 3.0 μL nuclease-free water, and 5.0 μL of extracted DNA and processed in triplicate. The qPCR analyses included a 15-min, 95 °C preincubation cycle, followed by 50 amplification cycles, and a 0.5-min 40 °C cooling cycle. A diluted plasmid standard was included during each qPCR run as a positive control and molecular-grade water was used in place of DNA template for negative controls. One copy of the targeted B. theta gene is assumed present per cell, and thus one gene copy number corresponded to one equivalent cell (19, 20). B. theta gene copies were converted to CE and reported as qPCR CE-100 mL⁻¹.

Climate and Hydrology. Hourly precipitation data were extracted from the Grand Rapids, Gaylord, and Detroit (Michigan) Next Generation Radar (NEXRAD) stations through the National Climate Data Center (www.ncdc.noaa.gov/nexradinv), with a base reflectivity of 0.50°, an elevation range of 124 nautical miles, and 16-km² cells. Hourly precipitation averages across each watershed were used to calculate total rainfall weighted by the
proportion of each NEXRAD cell within the sampled watershed. Precipitation was categorized into cumulative hourly totals (millimeters) before sample collection at intervals of 6, 12, 18, and 24 h and 2, 3, 4, 6, and 8 d, reported as millimeters per time before sample collection.

Real-time river discharge was measured at each site during sample collection using an Acoustic Doppler Current Profiler (53), collocated USGS stream gauges (waterwatch.usgs.gov), or current meter via wading following USGS protocol (54). River discharge is reported as cubic meters per second.

**Land Use.** Watersheds were delineated and then land use and septic system statistics were calculated for each watershed using Esri ArcMap GIS software (Table S4). The spatial analyst watershed tool was used to develop surface watersheds for each sampling point at 1 arc-second. Two watersheds were defined for each river site, referred to in this paper as full watersheds, which include the entire upstream drainage area (n = 64), and reduced watersheds, which only include drainage areas upstream of the sampling site to the nearest lake, reservoir, or pond (n = 52). The full watershed analysis (n = 64) included 12 sites that were at or near lake outlets, resulting in significantly smaller watersheds (average = 108 km²) than the other 52 watersheds (average = 366 km²). These 12 sites were removed in the reduced watershed analysis because it was originally hypothesized that longer retention time in the lentic water systems would likely reduce microbial concentrations owing to dilution and degradation by skimming. The lentic vs. lotic major river distinction was used to develop 30–33 river systems. Landstad imagery and the National Land Cover Database (NLCD 2006; mrlc.gov/nlcd2006.php) was used to define land use in each watershed and buffer. Land use was categorized using the NLCD classification system with 16 categories and seven categories using the Anderson Level 1 Land Cover System (55); Table S5 describes the Anderson classifications and equivalent NLCD categories. A 60-m riparian buffer was applied to streams in both full and reduced watersheds because land parcels are generally located adjacent to roads and require a buffer between surface waters and septic tanks. The average septic system setback from surface waters in Michigan is 15 m. Additionally, the 60-m riparian buffer ensured all riparian land uses were accounted for if the land use/river/septic system GIS layers were not completely matched under the 30-m resolution.

All land parcels that likely use on-site septic systems to treat wastewater were previously developed for this study region (35). Briefly, septic system totals and locations were estimated following the cumulative examination of WWTP infrastructure, incorporated municipality areas, household location according to 2010 census blocks, 2006 NLCD and road layers, and residential drinking water well information. Estimated septic system numbers (per watershed) and densities (per square kilometer) in each watershed and 60-m-wide buffer around surface water bodies were calculated for the 64 river systems.

Estimates of total population and population relying on WWTPs for water treatment were performed for each watershed and 60-m buffer. The total population in each watershed was estimated by multiplying the number of households (based on 2010 census data, described above during septic system estimates) by the average household size in each census block. The number of people relying on the WWTP that was estimated by overlaying census block fly-information and wastewater treatment plant service area boundaries. Additionally, the USEPA Discharge Monitoring Report (DMR) Pollutant Loading Tool (cfpub.epa.gov/dmr/ez_search.cfm) was used to estimate the ratio of average annual WWTP effluent to measured baseflow. A full description of this method is provided in Supporting Information.

**Statistical Analysis.** A constant value of 1 was added to *E. coli* and *B. theta* concentrations before log transformation and analysis. Soil hydraulic conductivity values were log-transformed before statistical analyses. Spearman correlation tests were used to examine relationships among physical, geochemical, and microbial measurements. Descriptive statistics were performed using IBM SPSS Statistics software (Version 19.0) with a significance threshold of (α) 0.01.

CART analysis was used to compare *E. coli* and *B. theta* (dependent variables) to the independent geographic, hydrologic, environmental, and land use variables. CART has been used to assess pathogenic bacteria and parasite relationships with environmental and land use factors (56), to classify lakes based on chemistry and clarity (57), and to predict the occurrence of fecal indicator bacteria with respect to physiochemical variables (58). CART was selected because it allows for robust nonlinear model development using multiple potentially interacting predictor variables (59) that splits dependent variables into categories based on the influence of independent variables. Following previously published methods (56, 57), CART recursively split dependent variables using a recursive partitioning algorithm (rpart) and a 10-fold cross-validation criterion. The 10-fold cross-validation approach breaks all data into 10 subsets and calculates the split based on 9 of the 10 subsets. This method is used for each group until reaching a minimum stopping criterion of five observations per subgroup. Fully developed CART outputs often required pruning to remove insignificant splits and significant variable associations were not missed due to the splitting and stopping criteria (60). We first pruned CART outputs using the 1-SE rule (61–63), and, if needed, a subsequent pruning step was performed if splits did not reduce error by 5% or more. This rule minimized the cross-validation error of the model, which has been shown to produce optimal sized trees that are stable across replications (61, 64).

To compare concentrations of the two organisms at each site relative to the average concentration of each organism, the Z-score of each sample was calculated. The z-scores were calculated as the sample concentration minus the mean of the concentrations before log transformation and analysis. Spearman correlations were calculated in R using the “spearman” function. For the development of this manuscript, partial funding for this project came from National Oceanic and Atmospheric Administration Great Lakes Environmental Research Laboratory Grant “Land Use Change and Agricultural Lands Indicators and Tipping Points.” Partial support was also provided by the Environmental Protection Agency Grants 112013 and 118539.

**ACKNOWLEDGMENTS.** We thank Steve Hamilton, Emily Luszcz, Bobby Chrisman, Rebecca Ives, Sarah AchiMoody, and Seth Hunt for vital support during this project and Drs. Shannon Briggs and Jon Bartholic for their technical support during the development of this manuscript. Partial funding for this project came from National Oceanic and Atmospheric Administration Great Lakes Environmental Research Laboratory Grant “Land Use Change and Agricultural Lands Indicators and Tipping Points.” Partial support was also provided by the Environmental Protection Agency Grants 112013 and 118539.


