Prepared in cooperation with the Michigan Department of Environmental Quality

Environmental Factors and Flow Paths Related to *Escherichia coli* Concentrations at Two Beaches on Lake St. Clair, Michigan, 2002–2005

Scientific Investigations Report 2008–5028

U.S. Department of the Interior
U.S. Geological Survey

Cover. Metropolitan Beach near Mt. Clemens, Michigan. (Photograph by Lisa R. Fogarty, U.S. Geological Survey)

Environmental Factors and Flow Paths Related to *Escherichia coli* Concentrations at Two Beaches on Lake St. Clair, Michigan, 2002–2005

By David J. Holtschlag, Dawn Shively, Richard L. Whitman, Sheridan K. Haack, and Lisa R. Fogarty

Prepared in cooperation with the Michigan Department of Environmental Quality

Scientific Investigations Report 2008–5028

U.S. Department of the Interior
U.S. Geological Survey

U.S. Department of the Interior
DIRK KEMPTHORNE, Secretary

U.S. Geological Survey
Mark D. Myers, Director

U.S. Geological Survey, Reston, Virginia: 2008

For product and ordering information:
World Wide Web: *http://www.usgs.gov/pubprod*
Telephone: 1-888-ASK-USGS

For more information on the USGS—the Federal source for science about the Earth, its natural and living resources, natural hazards, and the environment:
World Wide Web: *http://www.usgs.gov*
Telephone: 1-888-ASK-USGS

Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Although this report is in the public domain, permission must be secured from the individual copyright owners to reproduce any copyrighted materials contained within this report.

Suggested citation:
Holtschlag, D.J., Shively, Dawn, Whitman, R.L., Haack, S.K., and Fogarty, L.R., 2008, Environmental factors and flow paths related to ***Escherichia coli*** concentrations at two beaches on Lake St. Clair, Michigan, 2002–2005: U.S. Geological Survey Scientific Investigations Report 2008–5028, 38 p.

Contents

Figures

Tables

Conversion Factors and Abbreviations

Multiply	By	To obtain
	Length	
inch (in.)	2.54	centimeter (cm)
foot (ft)	0.3048	meter (m)
meter (m)	3.281	foot (ft)
mile (mi)	1.609	kilometer (km)
	Area	
square mile (mi^2)	259.0	hectare (ha)
square mile (mi^2)	2.590	square kilometer (km^2)
	Volume	
milliliter (mL)	0.06102	cubic inch (in^3)
	Velocity and flow rate	
mile per hour (mi/h)	1.609	kilometer per hour (km/h)
cubic foot per second (ft^3/s)	0.02832	cubic meter per second (m^3/s)

Temperature in degrees Celsius (°C) may be converted to degrees Fahrenheit (°F) as follows:

$°F=(1.8\times°C)+32$

Temperature in degrees Fahrenheit (°F) may be converted to degrees Celsius (°C) as follows:

$°C=(°F-32)/1.8$

Vertical coordinate information is referenced to the International Great Lakes Datum of 1985 (IGLD 85). This datum is a dynamic height system for measuring elevation, which varies with the local gravitational force, rather than an orthometric system, which provides an absolute distance above a fixed point. The primary reason for adopting a dynamic height system within the Great Lakes is to provide an accurate measurement of potential hydraulic head. The reference zero for IGLD 85 is a tide gage at Rimouski, Quebec, which is located near the outlet of the Great Lakes-St. Lawrence River system. The mean water level at the Rimouski, Quebec gage approximates mean sea level.

Elevation, as used in this report, refers to distance above the vertical datum.

Horizontal coordinate information is referenced to the Michigan South (2113) State Plane Coordinate System of 1983 (SPCS 83).

Concentrations of chemical constituents in water are given in milligrams per liter (mg/L). Concentrations of bacteria are given in colony-forming units per 100 milliliters (CFU/100 mL) or most probable number per 100 milliliters (MPN/100 mL).

Turbidity is given in Nephelometric Turbidity Units (NTU).

Abbreviations used in this report

ANGB	Air National Guard Base
AUROC	Area under the receiver operating characteristic
MDEQ	Michigan Department of Environmental Quality
MPN	Most probable number
NOAA	National Oceanic and Atmospheric Administration
NDBC	National Data Buoy Center
RMSE-E	Root-mean square error
RMSE-P	Root-mean square error prediction
ROC	Receiver operating characteristics
SMS	Surface-water modeling system
TMDL	Total maximum daily load
USGS	U.S. Geological Survey

Environmental Factors and Flow Paths Related to *Escherichia coli* Concentrations at Two Beaches on Lake St. Clair, Michigan, 2002–2005

David J. Holtschlag, Dawn Shively, Richard L. Whitman, Sheridan K. Haack, and Lisa R. Fogarty

Abstract

Regression analyses and hydrodynamic modeling were used to identify environmental factors and flow paths associated with *Escherichia coli* (*E. coli*) concentrations at Memorial and Metropolitan Beaches on Lake St. Clair in Macomb County, Mich. Lake St. Clair is part of the binational waterway between the United States and Canada that connects Lake Huron with Lake Erie in the Great Lakes Basin. Linear regression, regression-tree, and logistic regression models were developed from *E. coli* concentration and ancillary environmental data.

Linear regression models on $\log_{10}$ *E. coli* concentrations indicated that rainfall prior to sampling, water temperature, and turbidity were positively associated with bacteria concentrations at both beaches. Flow from Clinton River, changes in water levels, wind conditions, and $\log_{10}$ *E. coli* concentrations 2 days before or after the target bacteria concentrations were statistically significant at one or both beaches. In addition, various interaction terms were significant at Memorial Beach. Linear regression models for both beaches explained only about 30 percent of the variability in $\log_{10}$ *E. coli* concentrations.

Regression-tree models were developed from data from both Memorial and Metropolitan Beaches but were found to have limited predictive capability in this study. The results indicate that too few observations were available to develop reliable regression-tree models.

Linear logistic models were developed to estimate the probability of *E. coli* concentrations exceeding 300 most probable number (MPN) per 100 milliliters (mL). Rainfall amounts before bacteria sampling were positively associated with exceedance probabilities at both beaches. Flow of Clinton River, turbidity, and $\log_{10}$ *E. coli* concentrations measured before or after the target *E. coli* measurements were related to exceedances at one or both beaches. The linear logistic models were effective in estimating bacteria exceedances at both beaches. A receiver operating characteristic (ROC) analysis was used to determine cut points for maximizing the true positive rate prediction while minimizing the false positive rate.

A two-dimensional hydrodynamic model was developed to simulate horizontal current patterns on Lake St. Clair in response to wind, flow, and water-level conditions at model boundaries. Simulated velocity fields were used to track hypothetical massless particles backward in time from the beaches along flow paths toward source areas. Reverse particle tracking for idealized steady-state conditions shows changes in expected flow paths and traveltimes with wind speeds and directions from 24 sectors. The results indicate that three to four sets of contiguous wind sectors have similar effects on flow paths in the vicinity of the beaches. In addition, reverse particle tracking was used for transient conditions to identify expected flow paths for 10 *E. coli* sampling events in 2004. These results demonstrate the ability to track hypothetical particles from the beaches, backward in time, to likely source areas. This ability, coupled with a greater frequency of bacteria sampling, may provide insight into changes in bacteria concentrations between source and sink areas.

Introduction

The Macomb County Health Department (MCHD) began monitoring bathing beaches for bacteria in 1948 to help ensure a healthy recreational experience for people who visit area beaches (Macomb County Health Department, 2006). Michigan's water-quality standard (R323.1062 Rule 62) for total body contact recreation states that no site should have an *Escherichia coli (E. coli)* concentration that exceeds 130 colony forming units per 100 milliliters (CFU/100 mL) of water as a geometric mean of all samples collected over a 30-day period. According to this standard, the daily geometric mean of three or more samples should not exceed 300 CFU/100 mL of water. If beach waters exceed these standards, an advisory or closure is implemented by the MCHD until standards are met.

Public beaches along Lake St. Clair, in Macomb County (fig. 1), are intermittently closed when *E. coli* concentrations exceed the single-day Michigan recreational water-quality standard (referred to as the “recreational water-quality standard” hereafter). According to MCHD data from 1998 to 2005, this standard was exceeded at two beaches along the lake, Memorial Beach and Metropolitan Beach (fig. 1). It was exceeded 32 times out of 360 daily geometric mean samples (8.9 percent) at Memorial Beach and 23 times out of 367 samples (6.3 percent) at Metropolitan Beach. Because there is typically a 24-hour delay in obtaining test results, beach closings based on *E. coli* tests are not completely effective in preventing exposures of concern or allowing contact to resume when conditions are considered safe.

Uncertainties in source areas of microbial pathogens and indicator bacteria contribute to the difficulty of managing beach access to maximize public safety and to minimize the public inconvenience and economic losses associated with unnecessary or prolonged beach closings. The U.S. Geological Survey (USGS), in cooperation with the Michigan Department of Environmental Quality (MDEQ), began a study in 2006 to identify environmental factors and flow paths related to *E. coli* concentrations at Memorial and Metropolitan Beaches.

Purpose and Scope

This report presents environmental factors that are related to concentrations of *E. coli* at Memorial and Metropolitan Beaches on Lake St. Clair. Linear regression, regression-tree, and logistic regression models were used to identify these environmental factors. In addition, this report presents flow paths on Lake St. Clair that were developed using a hydrodynamic and particle-tracking model. These flow paths can be used to help identify potential source areas of *E. coli* to the two beaches. The statistical and hydrodynamic models can be used in combination by regulatory agencies for the development of Total Maximum Daily Loads (TMDLs) for *E. coli* at both beaches.

Site Description

Memorial and Metropolitan Beaches are on the west shore of Lake St. Clair in Macomb County, Mich. (fig. 1). Memorial Beach is about 2.5 mi north of St. Clair Shores, Mich., and about 2.6 mi south of the mouth of the Clinton River Cutoff Canal on Lake St. Clair. Metropolitan Beach is on the northern edge of L’anse Creuse Bay, along a peninsula that carries water from the Clinton River to Lake St. Clair. Metropolitan Beach is about 2.5 mi east of the Cutoff Canal and about 3.0 mi south of the mouth of Clinton River. The beaches provide recreational opportunities to Detroit-area residents and stimulate considerable economic activity. Changing current patterns on Lake St. Clair bring water and associated materials and organisms to the beaches from various source areas contributing to the lake.

Lake St. Clair is a binational body of water within the United States and Canada that is part of the Great Lakes Basin. The St. Clair River, which has a drainage area of about 222,400 mi^2, delivers water to Lake St. Clair at a rate of about 185,000 ft^3/s from Lake Huron. In comparison, local tributaries—including the Clinton River and the Cutoff Canal in the United States, and the Thames and Sydenham Rivers in Canada—and precipitation contribute a small fraction of the flow of the St. Clair River directly to Lake St. Clair. Water in the lake is lost by evaporation and withdrawals for water supply or discharges through the Detroit River into Lake Erie.

Lake St. Clair has a surface area of about 430 mi^2 and an average depth of 10 ft (Great Lakes Informational Network, 2006). The lake, which has a length of 26 mi and a breadth of 24 mi, contains about 1 mi^3 of water. An 800-ft-wide by 25-ft-deep channel is maintained through Lake St. Clair to facilitate navigation. Most of the navigational channel lies within the United States. Public and private beaches are located along the 130-mi-long shoreline of mainland areas.

Data Used in the Study

This report was developed on the basis of existing *E. coli* concentration and ancillary environmental data collected and analyzed by the Macomb County Public Health Department (MCHD), in cooperation with the Michigan Department of Environmental Quality (MDEQ). Additional meteorological and water-level data were obtained from the stations at Selfridge Air National Guard Base (ANGB) and the Lake St. Clair, Mich., weather station (LSCM4). Station LSCM4 is operated by the National Data Buoy Center (NDBC).

Escherichia coli Concentrations

The MCHD and MDEQ measured *E. coli* concentrations at Memorial and Metropolitan Beaches (figs. 2 and 3). *E. coli* sampling generally occurred on selected weekday mornings from April through September. Beginning in 1998, MCHD samples were generally obtained twice per week on Mondays and Wednesdays, although supplemental samples were obtained occasionally. MCHD sampling has been ongoing but MDEQ sampling, which generally occurred once per week on Tuesdays or Thursdays, began in 2003 and ended in 2004. Each *E. coli* concentration reported herein represents the daily geometric mean of six samples collected at about the same time.

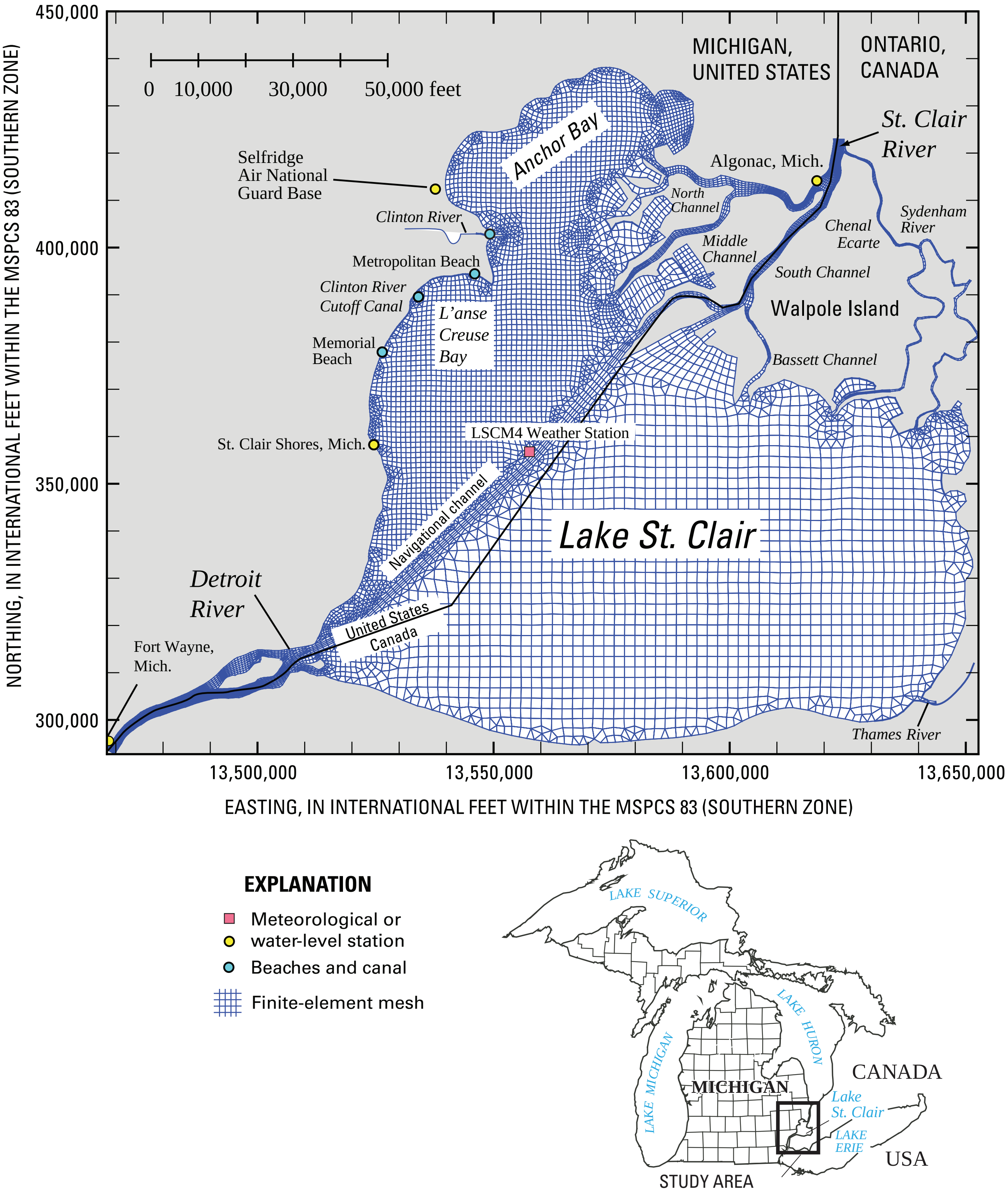


Figure 1. Lake St. Clair study area.

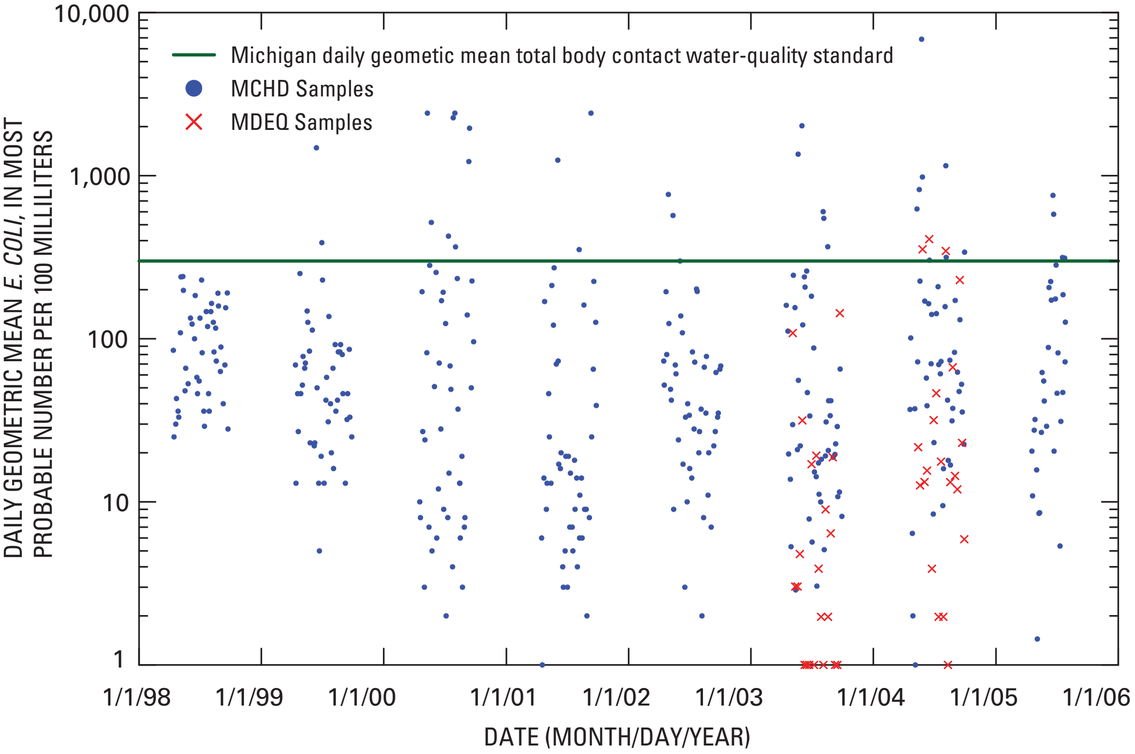


Figure 2. Time-series plot of the daily geometric mean *E. coli* concentrations at Memorial Beach on Lake St. Clair, Mich., 1998–2006. (MCHD, Macomb County Health Department; MDEQ, Michigan Department of Environmental Quality).

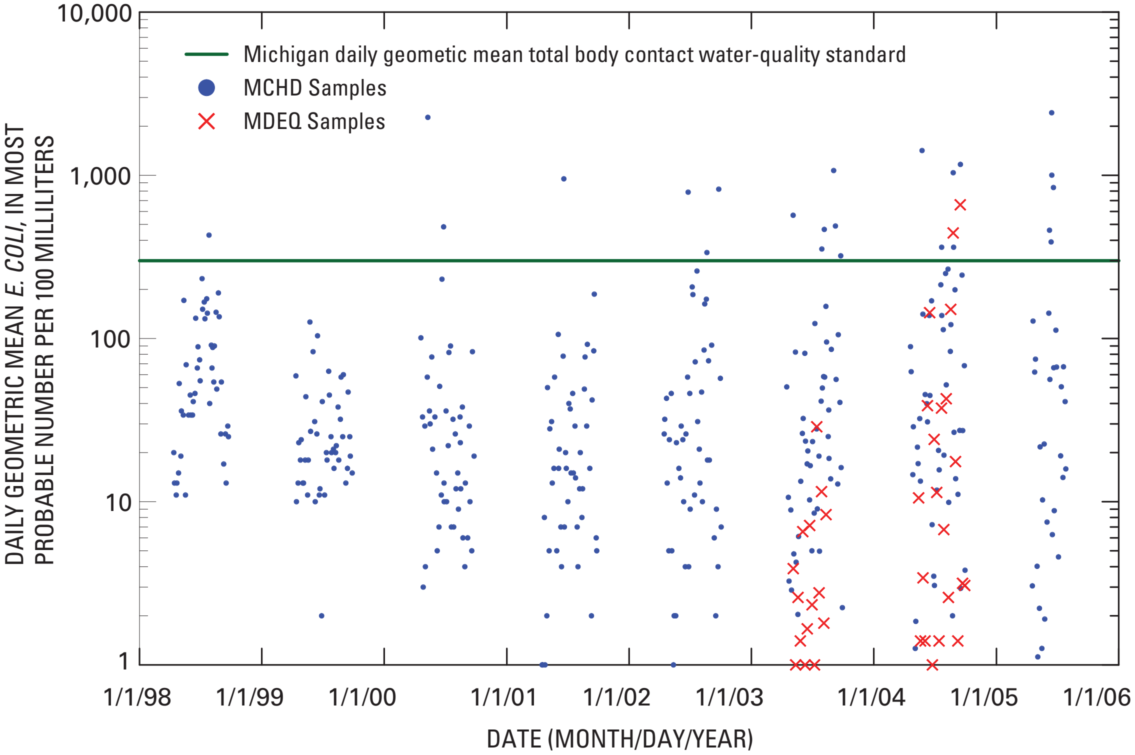


Figure 3. Time-series plot of the daily geometric mean of *E. coli* concentrations at Metropolitan Beach on Lake St. Clair, Mich., 1998–2006. (MCHD, Macomb County Health Department; MDEQ, Michigan Department of Environmental Quality).

In general, *E. coli* concentrations reported by MDEQ were lower than those reported by MCHD during the same periods. The reasons for these apparent differences are not known. Although sample locations were approximately the same, depth of sampling or sample-collection techniques may have varied by agency and person collecting the sample. In addition, MCHD and MDEQ use somewhat different analytical protocols to determine *E. coli* concentrations. MCHD samples are analyzed by means of the Colilert reagent system developed by IDEXX Laboratories (Westbrook, Maine), which quantifies *E. coli* concentrations by use of the most probable number (MPN) per 100 mL of sample water (Macomb County Health Department, 2007). In contrast, the MDEQ used membrane filtration (Standard Method 9222G in 2003, and U.S. Environmental Protection Agency method 1103.1 in 2004), a technique documented by the U.S. Environmental Protection Agency (2006), which quantifies *E. coli* concentrations by the number of colony-forming units (CFU) per 100 mL of water. Eckner (1998) concludes that MPN concentrations are comparable to CFU concentrations; therefore, for the purposes of this report, 300 MPN/100 mL is used as a criterion equivalent to the daily Michigan recreational water-quality standard.

Owing to concerns about possible systematic differences in *E. coli* concentrations measured by MCHD and MDEQ, the MCHD data were used in this report because they had been collected over a longer period and at more frequent intervals than the MDEQ data. Use of the MCHD dataset was restricted to April 2002 to September 2005 so that bacteria concentrations could be related to wind conditions on Lake St. Clair. In particular, the NOAA weather station LSCM4 on Lake St. Clair, which is near the middle of the lake just off the navigational channel at north latitude 42^{o}27’54” and west longitude 82^{o}45’18,” began reporting meteorological data in November 2001.

Ancillary Environmental Data

Numerous environmental factors may be associated with variations in *E. coli* concentrations. Environmental data considered in this study included simultaneous water-quality measurements made by MCHD with *E. coli* concentrations at the two beaches. These measurements included water temperature (reported in degrees Fahrenheit), pH, specific conductance (reported in millisiemens per centimeter), turbidity (reported in nephelometric turbidity units, NTU), dissolved oxygen (reported in milligrams per liter (mg/L)), and salinity, which is computed as a function of specific conductance and water temperature and reported in parts per thousand (PPT). MCHD also furnished data on *E. coli* concentrations for the Clinton River (station identifier U–41–Clinton River) and the Clinton River Spillway (station identifier Z–42–Spillway).

In addition to *E. coli* data and simultaneous water-quality measurements, continual measurements of meteorological, water-level, and flow data were available from nearby gaging stations operated by NOAA, Environment Canada, and the USGS. Hydrometeorological data were retrieved from Selfridge Air National Guard Base (ANGB) in Mount Clemens, Mich., which is operated by NOAA’s National Climatic Data Center as station 725377 (about 3 and 8 mi from Metropolitan and Memorial Beaches, respectively; fig. 1). Precipitation data were used to compute 24-, 48-, and 72-hour totals (in inches) preceding sampling for *E. coli* at the beaches. In addition, air temperature, wind velocity, atmospheric pressure, dewpoint, sky cover, and cloud ceiling data were obtained from NOAA. The National Data Buoy Center (NDBC) furnished wind speed and direction data for station LSCM4 on Lake St. Clair (approximately 7.5 mi offshore, fig. 1), which were used to compute the average 24-hour wind speeds and directions prior to *E. coli* sampling. In addition, sea-surface temperature (sometimes referred to in this report as SeaTemp), in degrees Celsius, barometric pressure, and dewpoint data were collected at station LSCM4. Environment Canada furnished data on lake-surface temperature, wave height, maximum wave height, and wave period from its buoy C45147 on Lake St. Clair. Missing values for the lake-surface temperature for NDBC station LSCM4 were estimated by linear regression, using lake-surface temperatures from Environment Canada Station C45147. The NOAA water-level gaging station at St. Clair Shores, Mich. (station 9034052) was the source of lake-level data. Water levels referenced to time t, WL_t, are measured at 6-minute intervals, $\Delta = 6$ minutes, at the NOAA gaging station on Lake St. Clair at St. Clair Shores, Mich. A smoothed water-level series, SWL_t was computed from the water-level data by use of a moving-average filter. The temporal averaging was used to dampen erratic fluctuations in water levels to help distinguish subtle changes in average water levels over short intervals. The smoothed water levels were computed as $SWL_t = \sum_{i=-4}^{4} WL_{t+i} \cdot \phi_i$ where the 9 ϕ coefficients are symmetrical, such that $\phi_{-i} = \phi_i$. Coefficients $\phi_{i=0,\pm1,\ldots\pm4} \approx$ {0.2042, 0.1802, 0.1238, 0.0663, 0.0276}. Changes in the smoothed water levels, ΔSWL_t , were computed as $\Delta SWL_t = SWL_t - SWL_{t-10\cdot\Delta t}$. The USGS provided streamflow (sometimes referred to in this report as ClintonFlow) information at the streamgage on Clinton River at Mount Clemens, Mich. (station 04165500, about 10 river miles from the mouth of the river at Lake St. Clair), in cubic feet per second.

Analytical Approach and Methods

An exploratory data analysis was done to evaluate the distributions of *E. coli* concentrations measured at Memorial and Metropolitan Beaches by MCHD, the relation between *E. coli* concentrations and wind velocity, and simple correlations between *E. coli* concentrations and other environmental factors. The results of this analysis were the basis for regression modeling, which attempted to identify environmental factors that are statistically associated with the *E. coli* concentrations or likelihoods of concentrations exceeding a specified threshold. Three types of regression models were developed and evaluated: linear regression, regression trees, and logistic regression models. The three types of models are based on different sets of statistical equations and constitute three unique ways to evaluate the data. Each is discussed individually in the following section. Briefly, linear regression models identify sets of environmental factors associated with different level of *E. coli* concentrations. With robust data sets, linear regression models have been used to statistically predict when *E. coli* concentrations are likely to exceed water-quality standards, given a set an explanatory variables (U.S. Environmental Protection Agency, 1999; Olyphant and Whitman, 2004; Nevers and Whitman, 2005; Francy and others, 2006). However, a required assumption for linear regression models is that the data are linear, which may not always be the case. Therefore, regression-tree analysis, like linear regression, was used to evaluate the data based a series of if-then statements that help identify a series of environmental variables associated with higher *E. coli* concentrations without the assumption of linearity. Finally, logistic regression models were used to estimate the probability of *E. coli* concentrations exceeding the recreational water-quality standard on the basis of selected environmental variables. Estimating probabilities is an alternative to estimating *E. coli* concentrations and comparing the estimates to the standard; the logistic model describes the odds of exceeding the standard with a unit change in the corresponding environmental variable, given that the other environmental variables do not change.

An existing two-dimensional (2D) hydrodynamic model (Holtschlag and Koschik, 2004) was modified to simulate flow patterns on Lake St. Clair near Memorial and Metropolitan Beach areas. Steady-state simulations were used to identify idealized current patterns near the beaches for selected wind conditions. Transient simulations were used to identify current patterns likely associated with specific bacteria-sampling events. Simulation results were the basis for particle-tracking analyses that were used to identify expected flow paths of water prior to bacteria sampling. The framework for the regression and hydrodynamic models is discussed in the following section of this report and the results of the models are discussed in the subsequent sections.

Regression Models

Regression models describe a statistical association between *E. coli* magnitudes and readily measurable environmental (explanatory) factors. These statistical associations do not ensure causal relations. Parameter values in regression equations, however, are commonly interpreted as indicators of possible physical relation between explanatory and response variables. In particular, positive parameters may indicate a direct relation between explanatory and response variables, whereas negative parameters may indicate an inverse relation. Any interpretation of parameter magnitudes is problematic because explanatory variables in these equations are often related or dependent on each other; therefore, distinguishing their individual contributions is difficult.

Linear Regression

Linear regression provides a statistical model for estimating the continuous variation of *E. coli* concentrations as a function of environmental factors (explanatory variables). General linear regression equations are described by Draper and Smith (1998), and further explanations on the equations used for these models can be found in sections following.

Statistics summarizing the regression help assess model performance. Key statistics include the following:

- Sum of squared errors, $SSE = (y - \hat{y})'(y - \hat{y}) = \varepsilon' \cdot \varepsilon$,
- Total sum of squares, $SST = (y - \overline{y})'(y - \overline{y})$, where $\overline{y}$ is the mean of the response variable, and
- Model sum of squares, $SSM = SST\text{-}SSE$.
- Degrees of freedom for the error, $df_e = n\text{-}p\text{-}1$, and
- Model degrees of freedom, $df_m = p$.
- Mean square error, $MSE = \mathrm{SSE}/df_e$,
- Model mean square, $MSM = \mathrm{SSM}/df_m$, and
- Root mean square error, $RMSE = \sqrt{MSE}$.

In addition, an F statistic, computed by dividing the MSM by the MSE, characterizes the statistical significance of the model. On the basis of the F probability distribution, a probability (p-value) is computed with the F statistic, and the degrees of freedom in the model and error components to assess the likelihood that the null hypothesis (that all model parameters are zero) is true. A small p-value, commonly less than 0.05, is used to reject the null hypothesis, thereby accepting the alternative hypothesis that the regression model is statistically significant.

The fraction of the variability of response variables described by the predicted values is determined by the multiple coefficient of determination, R^2, which tends to decrease monotonically with increasing p whether or not there is an improvement in prediction accuracy. The adjusted coefficient of determination, R_a^2 is a better measure of the prediction accuracy and of comparing equations with different numbers of explanatory variables.

Like the response estimates, estimated parameters associated with the individual explanatory variables are uncertain. As n becomes large, estimated parameters are assumed to be unbiased and normally distributed about their true values. A t statistic can be computed to assess the significance of individual parameters. On the basis of the t probability distribution with n-p-1 degrees of freedom, a p-value is computed with the t statistic to assess the likelihood that the null hypothesis of the model parameter being equal to zero is true. Again, a small *p*-value is used to reject the null hypothesis, thereby accepting that the parameter estimate as statistically different from zero.

A stepwise procedure was used to automatically select explanatory variables for inclusion in preliminary regression equations. Preliminary equations only included main effects (simple variables), without consideration of interaction effects (explanatory variables formed as the products of other explanatory variables), or dynamic effects associated with past or future measurements of *E. coli* concentrations. In the stepwise procedure, variables are added one at a time to the model from a candidate list. In this report, the included variables were required to have a *p*-value less than 0.05. Once included, they were required to retain a *p*-value of less than 0.10 as new variables were added. Stepwise methods are efficient for evaluating a large candidate list of explanatory variables. Inferential measures, such as R^2 and RMSE values, however, are biased towards indicating too much strength in the relation between response and explanatory variables.

Once the preliminary regression equation identified main-effect variables, first-order interaction variables were computed as the product of all main-effect variable pairs. Finally, both causal and noncausal dynamic components were evaluated. In this report, *E. coli* concentrations measured 2 days before to the time of the response *E. coli* concentration were treated as an explanatory variable in a causal dynamic regression formulation. The relation is considered causal because past *E. coli* concentrations might be used to predict current *E. coli* concentrations. *E. coli* concentrations, measured 2 days after the response *E. coli* concentration was measured, also were evaluated as a noncausal dynamic component. Noncausal implies that the future *E. coli* concentrations do not affect current concentration, but may be helpful in estimating them offline. Including both causal and noncausal components in the same equation provides a mechanism to investigate the symmetry of the dynamic component.

Regression Trees

Regression trees are an alternative to linear regression for estimating a response variable on the basis of one or more explanatory variables. Regression trees subdivide responses into a hierarchical system of nodes by a sequence of logical if-then statements based on selected threshold values of the explanatory variables. Threshold values are computed as part of the regression-tree analysis. In this study, *E. coli* measurements were grouped, divided, and subdivided on the basis of one of the environmental factors. Graphically, regression trees are depicted with the single root node at the top, which contains all observations (or *E. coli* measurements), and branches forming below. Starting at the root (parent) node, two terminal or internal (child) nodes are placed below the root. If the response to the logical if-then statement is true, the response is included with the observations in the child node below and to the left of the parent; otherwise the response is included in a child node to the lower right. For example, the if-then statement might be, if water temperature on day of sampling were less than 20ºC, then all *E. coli* measurements on those days would be grouped as “true” and be represented on the left side below the parent node; all other measurements would be grouped together on the right side below the parent node. As the level of the branching increases, child nodes become parent nodes. In the above example, the two groups of *E. coli* measurements based on water temperature (new parent nodes) could be subdivided in four groups (two additional child nodes under each new parent). The if-then statements might be different for each of the new parent groups. In the group defined by water temperatures less than 20ºC, the if-then statement could be if rainfall 24 hours prior to day of sampling were less than 0.1 in., then all *E. coli* samples that were previously collected on days with water temperature less than 20ºC and those also collected on days with less than 0.1 in. of rain would be “true” and subdivide a new child node under that parent node. Those that haven’t met this criterion would be “false” and grouped to the right as a new child node under that parent. The branching process continues until all child nodes are terminal nodes. In this report, branching was constrained so that every internal node had to have at least 10 observations for a split to occur. Terminal nodes can have a minimum of one observation, which is why regression trees are robust to the presence of outliers (extreme values of responses).

Regression trees are not based on a probability model, so there is no statistical criterion, analogous to a *p*-value, for determining the appropriate level of branching. Therefore, statistical resampling techniques were used to assess possible overfitting of regression trees to *E. coli* data and resulting effects on the degradation of prediction accuracy. In overfitting or parameterizing, the sample is fitted rather than the population. In particular, if we drew another random sample from the population, the model we would develop might not be similar to the one that was estimated from the sample we have. Consequently, we may not be able to infer how well we can predict new observations based on the overfitted model from the sample data. Resampling is based on the distinction between estimation errors—differences between measured and predicted values for observations used in the model building process—and prediction errors—differences between measured and predicted values for observations not used in the model building process.

To implement the resampling analysis, the Memorial and Metropolitan *E. coli* data sets (including the explanatory variables) were repetitively subdivided into two random subsets for model development and testing. The model development dataset contained about 75 percent of the measurements, and the testing dataset contained the remaining 25 percent. In each repetition, a new regression-tree model was created with the model development dataset. The root-mean-square error of estimation (RMSE-E) was computed with the development data, in which the number of terminal nodes was used to estimate the number of model parameters. Then the model-testing data set was applied to the model structure identified with the model-development data, and a root-mean-square error for prediction (RMSE-P) was computed without adjustment for the number of model parameters. After 500 repetitions, the paired distributions of RMSE for estimation and prediction were displayed as boxplots as a function of the number of terminal nodes.

Logistic Regression

As an alternative to estimating *E. coli* concentrations and comparing the estimates to thresholds of interest, one can directly estimate the probability that a specified threshold will be exceeded using classification methods. In this report, logistic linear regression models were developed to estimate probabilities of exceedance of the recreational water-quality standard. Application of the logistic model involves a transformation of *E. coli* concentrations into two classes: those that exceed the standard and those that do not. In this report, binary logistic models were developed to estimate the probability of *E. coli* concentrations exceeding 300 MPN/100 mL. Equations used to develop this model are further explained in a later section.

Logistic regression equations describe probabilities that *E. coli* concentrations will exceed a specified magnitude. A cut point is a fixed probability value for classifying these computed probabilities as exceedances or nonexceedances. Although a cut point of 0.5 might be taken as an arbitrary choice, a receiver operating characteristic (ROC) analysis can be used for exploring the sensitivity and specificity of the logistic equation for alternative cut points. Here, sensitivity refers to the probability that the model will indicate an exceedance when an *E. coli* concentration greater than 300 MPN/100 mL is measured. Sensitivity also is referred to as the true positive rate. Specificity is the probability that the model will indicate nonexceedance when the *E. coli* measurement is below the threshold concentration. One minus the specificity is the false positive rate.

An ROC analysis is depicted by use of an ROC curve, which depicts the tradeoff between sensitivity and specificity. An ROC curve for a given logistic regression equation and measured concentrations, typically plots the false positive rate on the *x*-axis and the true positive rate on the y-axis for a set of cut points ranging from 0 to 1. A diagonal line, which is drawn from the lower left-hand corner of the plot at [0,0] to the upper right hand corner at [1,1], is a reference line for evaluating alternative models. The area under the diagonal line is 0.5; the area under the receiver operating characteristic (AUROC) curve is a measure of the utility of the logistic model. An AUROC of 1 indicates a perfect model, and a value of 0.5 indicates a complete model failure. A rough guide classifies AUROC value from 0.9 to 1.0 as excellent and from 0.8 to 0.9 as good (Tape, 2007).

Hydrodynamic Modeling

Hydrodynamic simulations were used to describe the horizontal (vertically averaged) velocity components that are expected to occur for specified flow, water level, and wind boundary conditions, given the geometry and hydraulic characteristics of the waterway. Hydrodynamic simulations were developed by means of the generalized hydrodynamic code RMA2-WES version 4.5 (RMA2). The site-specific geometric and hydraulic characteristics within the Lake St. Clair area were based on the hydrodynamic model of the St. Clair-Detroit River Waterway developed for the MDEQ Source Water Assessment Program by Holtschlag and Koschik (2002).

RMA2 Hydrodynamic Code

RMA2 is a generalized computer code for two-dimensional hydrodynamic simulation of surface-water bodies. RMA2 implements a finite-element solution of the Reynolds form of the Navier-Stokes equations for turbulent flows. Donnell and others (2005) provide detailed documentation of the governing equations, their solution by the finite-element method, and the recommended uses and limitations of the RMA2 code. RMA2 is maintained and frequently updated with enhanced features by the U.S. Army Corps of Engineers Waterways Experiment Station in Vicksburg, Miss.

Following is a brief overview of those modeling aspects needed to help understand this application of the hydrodynamic model. To numerically solve the partial differential equations describing flow in two horizontal dimensions, the waterway area is discretized into a mesh of quadrilateral and triangular elements defined at points by nodes (fig. 1). The finite-element discretization provides a flexible representation of an irregularly shaped waterway. Corner nodes are placed at the vertices of elements, and midside nodes are placed approximately midway between the arcs (edges) connecting the corner nodes. Thus, eight nodes are defined per quadrilateral element, and six nodes are defined per triangular element. Contiguous elements forming branches or subreaches of the waterway are grouped into material zones to facilitate characterization of the hydraulic properties of the waterway. Channel-roughness coefficients and eddy viscosity characteristics are assigned to material zones, which in turn affect the velocities and water-surface elevations simulated at nodes.

Hydrodynamic simulations compute water levels and flow velocities at interior nodes on the basis of boundary conditions specified at exterior nodes, hydraulic characteristics of the waterway, and flow equations. Boundary conditions include flow (discharge across one or more elements), water level, and wind conditions. Quadratic interpolation may be used to determine water levels and velocities anywhere within the elemental areas on the basis of nodal values. Therefore, greater densities of nodes result in improved spatial resolution of velocity fields but also require additional computational resources.

RMA2 supports steady-state and transient simulations. When boundary conditions remain unchanged for a prolonged period, flow is considered at steady state (not changing with time). Although there is always some variability in the boundary conditions with time, these changes may be sufficiently minor that steady-state simulations are useful approximations of actual flow conditions. In other situations, steady-state simulations are idealized scenarios that facilitate generalizing the effect of various forces on flow and circulation patterns. When changes at the boundaries occur before flow has approximately equilibrated to previous changes, flow is considered transient. Although transient simulations require greater computational and storage resources, they can result in useful flow information for historical events of specific interest. In this report, transient flow was simulated at 1.0-hour time steps.

The horizontal water velocity field is characterized by 2D hydrodynamic simulation. Water velocities may decrease in shallow areas because aquatic vegetation effectively increases channel roughness, and adjacent points may have more or less similar velocities based on the effectiveness of turbulent exchange in mixing the water. Neither channel roughness nor turbulent exchange, which is the fluid momentum transfer due to chaotic motions of fluid particles (Donnell and others, 2005), can be measured directly in the field, but they can be inferred from measurements of flow, water level, and velocity within the waterway. The hydrodynamic model of the St. Clair-Detroit River developed for the MDEQ’s Surface Water Assessment Program (SWAP) was calibrated with extensive sets of velocity, water level, and discharge information by use of a nonlinear, parameter estimation technique developed by Poeter and others (2005).

The Surface-Water Modeling System (SMS) is a computer program for preprocessing and postprocessing RMA2 datasets (Environmental Modeling Systems, Inc., 2005). SMS was used to facilitate the development of the hydrodynamic model of the St. Clair-Detroit River Waterway by helping to implement and quality-assure the mesh design and by helping to visualize simulated water velocities and elevations. To produce particle tracks, simulated velocities and locations of corner nodes were output from SMS into an ASCII file format and read into a plotting program (Tecplot by Amtec Engineering, Inc., 2006), which was then used to track hypothetical, massless particles by use of linear interpolation across elements at a time step of 0.1 hour through both steady-state and transient flow fields.

Two-Dimensional Hydrodynamic Model of Lake St. Clair

The St. Clair-Detroit River Model is a structured finite-element data set in the RMA2 input format that describes the geometry, bathymetry, and hydraulic characteristics of the St. Clair-Detroit River Waterway. In particular, the input includes a connectivity table that indicates which nodes are associated with each element and which elements are associated with each material zone. In addition, the easting and northing specification of each node is referenced to International feet within the southern zone (2113) of the Michigan State Plane Coordinate System of 1983 (MSPCS 83), and the bottom elevation at each node is referenced to the International Great Lakes Datum of 1985 (IGLD 85).

The St. Clair-Detroit River hydrodynamic model extends from the head of the St. Clair River at the water-level gage operated by NOAA near Fort Gratiot, Mich. (Station ID: 9014098) to the Bar Point, Ontario, gage (station number 12005) at the mouth of the Detroit River, which is operated by the Canadian Hydrographic Survey. Inflows are specified to the St. Clair River from Lake Huron, and the Black, Pine, and Belle Rivers; to Lake St. Clair from the Clinton, Sydenham, and Thames Rivers; and to the Detroit River from the River Rouge. Water levels are specified only at the Bar Point gage. Wind conditions are specified for the entire mesh on the basis of data at the NOAA weather station LSCM4 on Lake St. Clair (fig. 1).

Several versions of the model were developed in response to changing data availability and application emphasis. The initial version discretized the waterway into 13,783 quadratic elements defined by 42,936 nodes and was used to estimate channel-roughness characteristics in 25 material zones. Data from velocity surveys (Holtschlag and Koschik, 2003a,b) were used to enhance the model calibration so as to describe the variation in channel roughness with depth of flow by material zones and to estimate the turbulent exchange coefficient or eddy viscosity. This version is referred to as the “standard version”. Following this, the “public water intake version” refined the mesh geometry in the vicinity of 13 U.S. public water intakes, thereby increasing the number of elements to 30,306 and the number of nodes to 90,386.

Particle Tracking

Particle tracking uses a spatial-temporal interpolation technique to move hypothetical massless particles across the lake with time and is based on simulated velocities of water movement at model nodes. Particle tracking is an inherently transient procedure even in steady-state velocity fields. This transient component indicates traveltimes along a flow path. In steady-state flow fields, the paths of particles started at different initial positions do not cross. In contrast, particle paths in unsteady flow fields may cross because of different arrival times at various points in the time-varying flow field.

Particles may be tracked forward or backward in time through simulated steady or transient flow fields. In steady flow, particles are tracked forward in time by specifying an initial position, the length of the time step, the duration of the tracking period, and a temporally constant but spatially varying flow-velocity field. Some experimentation is required to determine a tracking time step that is sufficiently short so that further decreases have little impact on ending locations. Similarly, particles may be tracked backward in time by reversing the sign of the simulated velocity components at each node and specifying ending, rather than starting, locations. Particles may be forward tracked through a transient flow field by specifying a particle-tracking time step and a series of transient flow solutions computed at specified time intervals that may differ from the particle-tracking time step. Finally, particles may be backtracked through a transient flow field by reversing the order of the transient-flow solutions and reversing the sign of simulated velocities.

Relation of Environmental Factors to *Escherichia coli* Concentrations

Statistical Distributions of *Escherichia coli* Concentration and Effects of Wind Direction

During 2002–05, the 139 selected *E. coli* concentrations at Memorial Beach ranged from 1 to 2,020 MPN/100 mL, with mean, standard deviation, and skewness characteristics of 137-, 268-, and 4.26 MPN/100 mL, respectively. In this report, the coefficient of skewness was computed as

$$Skewness = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^{3/2}} \tag{1}$$

where

x_i are the individual observations,

$\overline{x}$ is the average of the x_i values,

and

n is the number of observations.

Similarly, the 143 selected *E. coli* concentrations at Metropolitan Beach ranged from 1 to 1,270 MPN/100 mL, with mean, standard deviation, and skewness characteristics of 93.2, 189, and 3.72 MPN/100 mL, respectively. Thus, the statistical distributions of bacteria at both beaches were positively skewed (fig. 4). This skewness is a problem in regression analyses because it effectively weights the larger concentrations at both beaches too highly with respect to their frequency in the sample. In addition, direct use of the *E. coli* concentrations in a regression analysis would not result in a residual distribution that was approximately normally distributed, thereby violating a fundamental assumption of least-squares regression. To overcome this problem, a common logarithmic ($\log_{10}$) transformation was applied to *E. coli* concentrations. For Memorial Beach, the mean, standard deviation, and skewness of the transformed concentrations are 1.69, 0.63, and -0.017 $\log_{10}$ MPN/100 mL, respectively. Likewise, for Metropolitan Beach, the mean, standard deviation, and skewness of the transformed concentrations are 1.46, 0.67, and 0.10 $\log_{10}$ MPN/100 mL, respectively. Thus, the transformed concentrations show little skewness, and their distributions are closely approximated by normal distributions (fig. 4). All subsequent analyses used the transformed concentration values. For simplicity in discussing the statistical analyses, the $\log_{10}$ transformed *E. coli* concentration values are referred to as *E. coli* concentrations, unless ambiguity between the original and transformed metrics would result.

Wind conditions are thought to be the primary determinant of current patterns and source areas of water to beaches on Lake St. Clair. Thus, differences in 24-hour average wind conditions before a sampling event may be associated with different source areas. If the release of bacteria is constant and differs among sources, then a polar plot of bacteria concentrations with wind speed and direction may indicate possible source-area effects. In a polar plot, angular and radial distances used to plot bacteria concentrations are determined by use of 24-hour vector averages of wind speeds and directions from LSCM4, and the posted location of the sample is color coded by concentration. For the bacteria data from Memorial Beach, the polar plot indicates little apparent differences in bacteria concentrations with wind speed or direction (fig. 5).

Possible variations in median bacteria concentrations with wind direction were evaluated by subdividing the bacteria data into corresponding equiangular sectors of 45 degrees each. Analysis by use of a notched boxplot indicates no statistically significant differences in median *E. coli* concentrations at Memorial Beach among sectors (fig. 6). In a notched boxplot, a notch is formed in each boxplot that corresponds to the 95-percent confidence interval about the median. If the range in concentrations formed by notches from all sectors overlaps in some area, the null hypothesis of no differences in median concentrations among sectors is not rejected. A Kruskal-Wallis test (Conover, 1980), which is a nonparametric form of an analysis of variance, confirms the absence of statistical differences in *E. coli* concentrations among wind directions (*p*-value equals 0.3335).

In contrast to Memorial Beach, median concentrations at Metropolitan Beach differed significantly among wind sectors. In particular, the polar plot shows a higher percentage of large -magnitude *E. coli* concentrations when winds are from the east-northeast (ENE.) than from some of the other wind sectors (fig. 7). And unlike Memorial Beach, exceedance of the water-quality criteria did not occur in all wind sectors. In the ENE., ESE., SSE., SSW., and WSW. sectors, three to five sample exceeded the Michigan recreational water-quality standard. In the NNE. sector, however, there was only one exceedance, and there were none in the NNW. and WNW. sectors. The corresponding boxplot (fig. 8) shows that the ENE. sector has the highest median value, but the fewest number of samples. Based on the notched boxplot (fig. 8), the median concentration in the ENE. sector is significantly greater than the median concentrations in the WSW., WNW., and NNW. sectors. In addition, the median concentration in the SSW. sector is significantly greater than median concentrations in the WNW. and NNW. sectors. The Kruskal-Wallis test also rejects the equality of median *E. coli* concentrations in all sectors (*p*-value equals 0.0009).

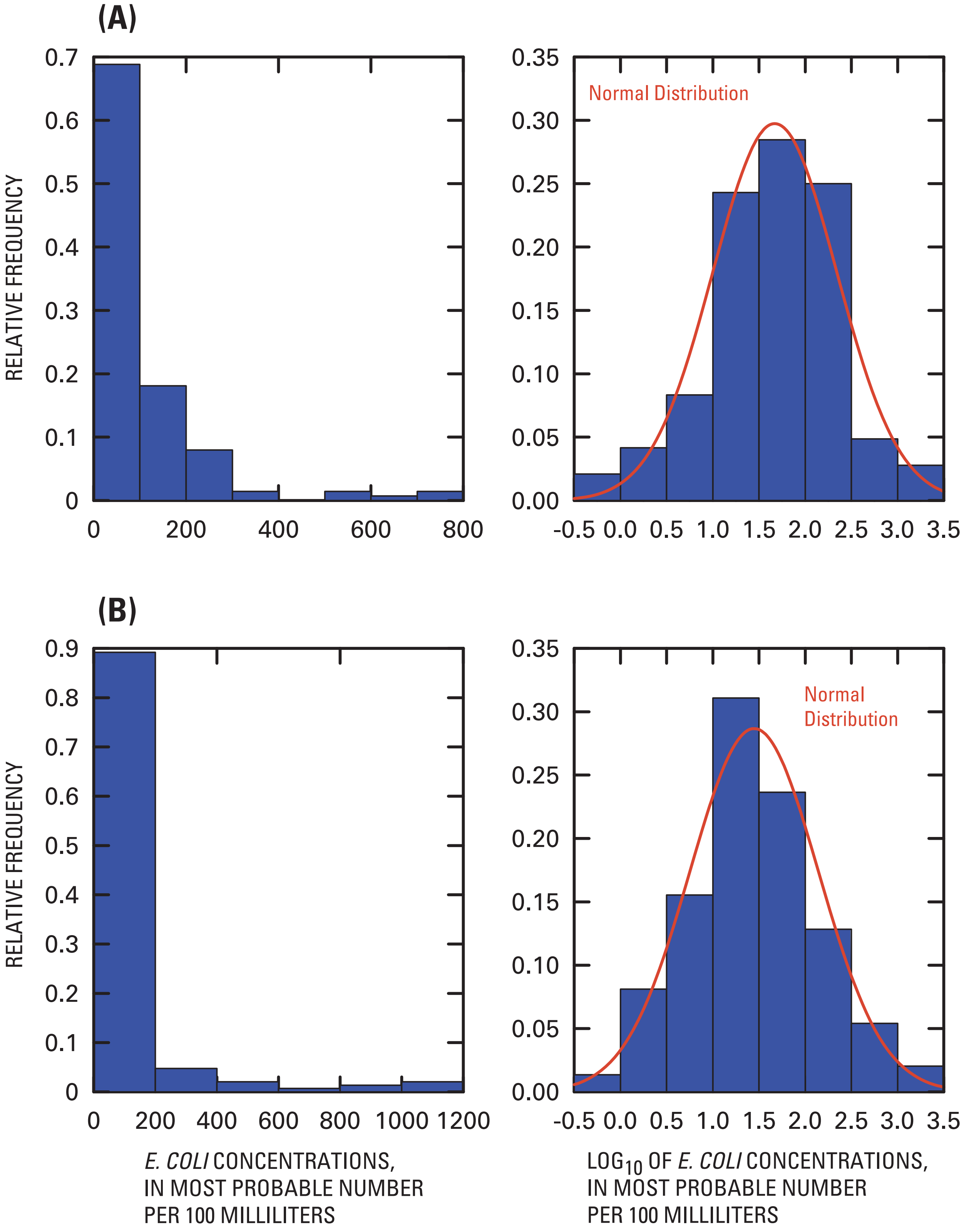
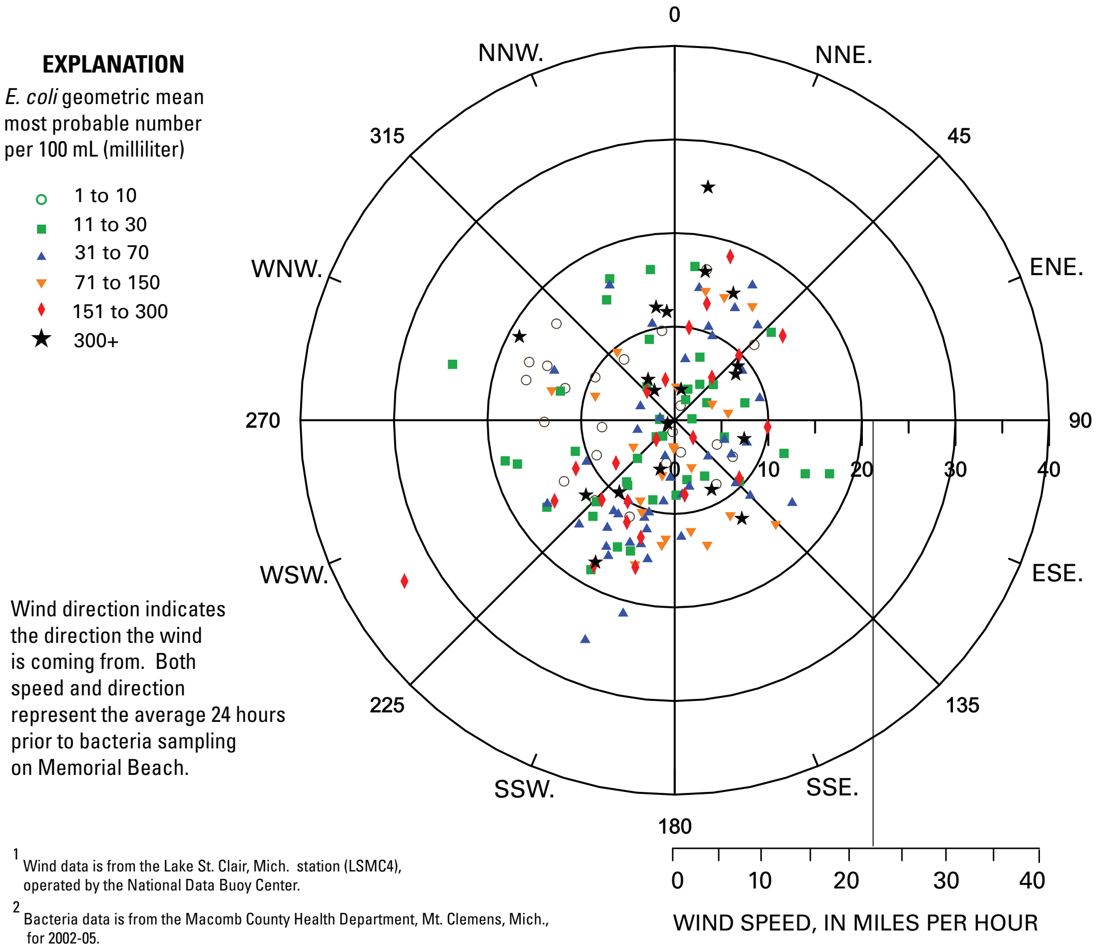


Figure 4. Statistical distributions of *E. coli* concentrations at (**A**) Memorial and (**B**) Metropolitan Beaches on Lake St. Clair, Mich., 2002–05.

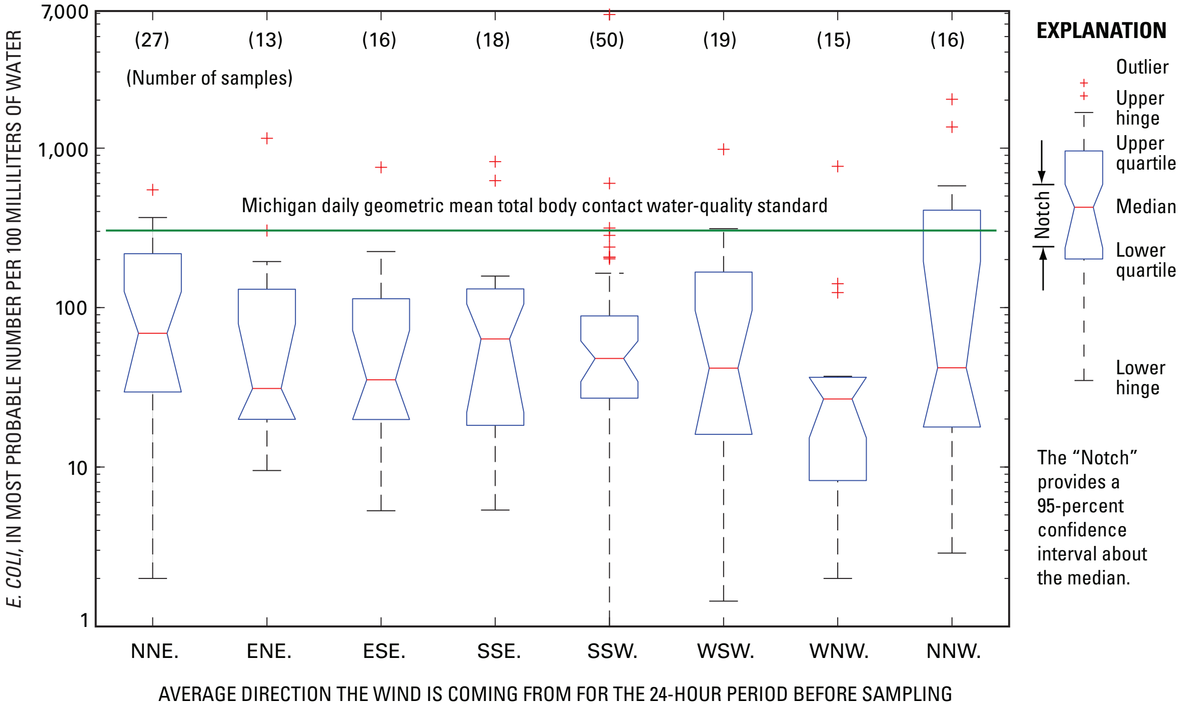
Figure 5. Relation between wind conditions and *E. coli* concentrations at Memorial Beach on Lake St. Clair, Mich., 2002–05.



Wind direction indicates the direction the wind is coming from. Both speed and direction represent the average 24 hours prior to bacteria sampling on Memorial Beach.

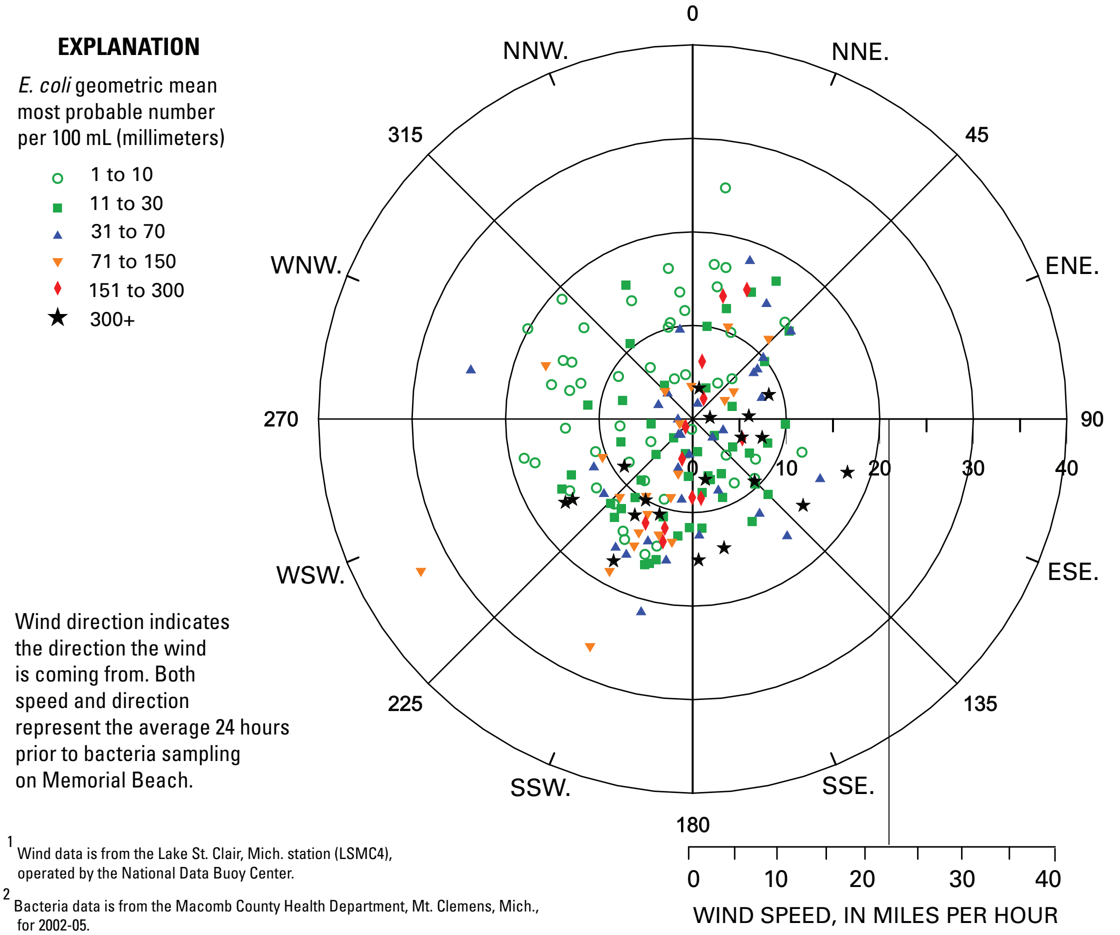
[1] Wind data is from the Lake St. Clair, Mich. station (LSMC4), operated by the National Data Buoy Center.

[2] Bacteria data is from the Macomb County Health Department, Mt. Clemens, Mich., for 2002-05.



The "Notch" provides a 95-percent confidence interval about the median.

Figure 6. Distributions of Macomb County Health Department *E. coli* concentrations at Memorial Beach by wind direction on Lake St. Clair, Mich., 2002–05.



Wind direction indicates the direction the wind is coming from. Both speed and direction represent the average 24 hours prior to bacteria sampling on Memorial Beach.

[1] Wind data is from the Lake St. Clair, Mich. station (LSMC4), operated by the National Data Buoy Center.

[2] Bacteria data is from the Macomb County Health Department, Mt. Clemens, Mich., for 2002-05.

Figure 7. Relation between wind conditions and *E. coli* bacteria concentrations at Metropolitan Beach on Lake St. Clair, Mich., 2002–05.

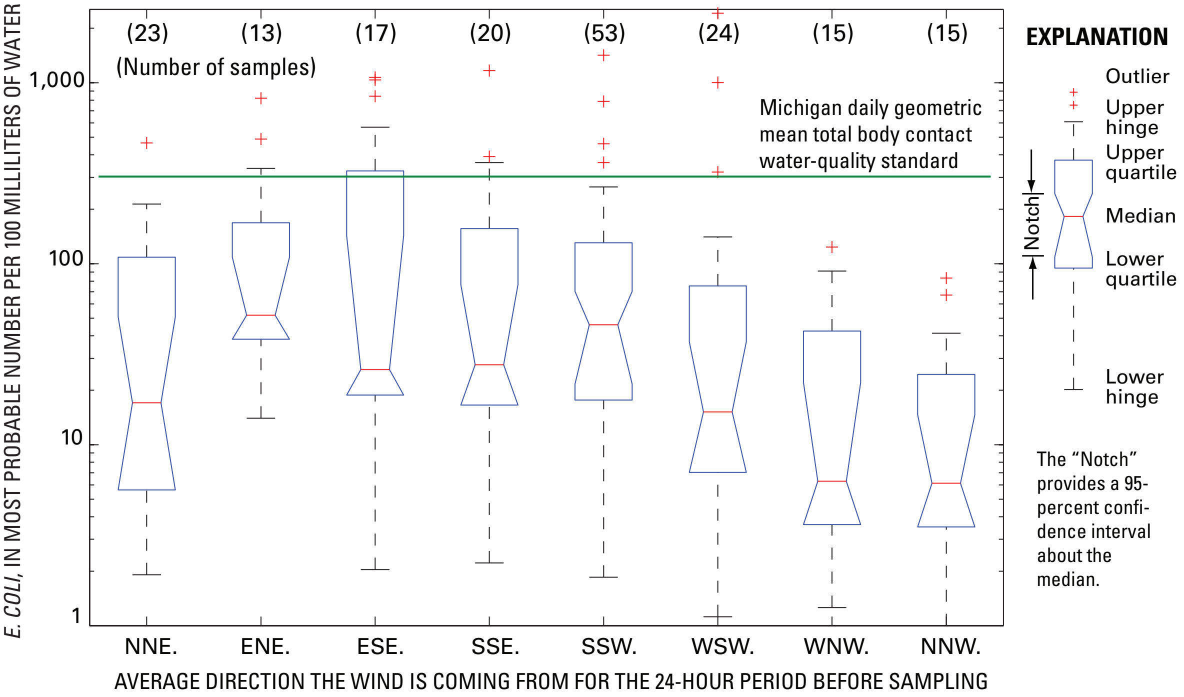


Figure 8. Distributions of Macomb County Health Department *E. coli* data, by wind direction for Metropolitan Beach on Lake St. Clair, Mich., 2002–05. (Overlapping notched regions indicate that differences in median concentrations are not statistically significant. The two boxplots labeled WNW and NNW have median concentrations that differ statistically from the median concentration at ENE.)

In summary, wind direction was not apparently associated with *E. coli* concentrations at Memorial Beach, but it did influence *E. coli* concentrations at Metropolitan Beach. Onshore winds are associated with higher waves, and higher waves are historically associated with higher *E. coli* concentrations. At Metropolitan Beach, onshore winds (from ESE. to SSW.) accounted for the largest number of exceedances of the recreational water-quality standard but not the largest median concentration of *E. coli*. During onshore winds, waves might resuspend or wash *E. coli*-laden sand, vegetative material (algae), or animal waste to nearshore beach water. Unfortunately, the wave data from Environment Canada proved to be problematic in characterizing this possibility; attempts to fit the wave data during regression analyses were not successful. Perhaps the Environment Canada buoy is too far offshore or the lake structure is too complex (shallow depth of 3-m with an 8-m navigation channel to the east of this buoy) to anticipate nearshore waves at the study beaches.

Linear Regression Results

Memorial Beach

In the linear regression analysis of $\log_{10}$ *E. coli* concentrations (LogEC) at Memorial Beach, the stepwise method identified total inches of rainfall at Selfridge ANGB 72 hours preceding the *E. coli* sample (Rainfall_72), sea-surface temperature of Lake St. Clair in degrees Celsius (SeaTemp), $\log_{10}$ of flow at Clinton River in cubic feet per second (ClintonFlow) 1 day before the bacteria sample, and turbidity in Nephelometric Turbidity Units (Turbidity) as significant explanatory variables in the main-effects model (table 1). From this model, all first-order interaction terms were formed and evaluated, again by stepwise analysis. Although Rainfall_72-ClintonFlow and ClintonFlow-Turbidity interaction terms contributed to the main effects model and resulted in the lowest RMSE and a higher adjusted R^2 value, neither interaction term was individually significant at the $p = 0.05$ level. Therefore, interactions were not included in the dynamic-effects model. Similarly, neither wind speed nor a north wind direction indicator variable contributed significantly to the regression model at the $p = 0.05$ level.

In a dynamic regression analysis of *E. coli* concentrations, 108 observations were formed from *E. coli* samples collected 2 days apart at Memorial Beach. Of these, 54 samples had explanatory *E. coli* concentrations measurements before the response *E. coli* concentrations and 54 samples had explanatory concentrations measurements afterwards. In addition to the presampling or postsampling concentrations (LagLeadLogEC), an indicator vector was formed to identify whether the explanatory *E. coli* concentrations were obtained before or after the response *E. coli* concentration. This type of model is used to help determine whether *E. coli* concentration in samples collected 2 days prior could be used as a predictor for *E. coli* concentrations on a particular day of interest. This would also help determine how frequently the beach should be sampled (Is sampling every other day for *E. coli* protective of human health?). Understanding the effects of *E. coli* 2 days after the response *E. coli* may help to understand the process that may transport bacteria to and from the beach area.

Inclusion of the explanatory *E. coli* concentration data and the indicator vector in the main-effects regression indicated that the explanatory *E. coli* concentrations were statistically significant, but the timing indicator was not statistically significant ($p = 0.4602$). These results imply that the correlation between *E. coli* concentrations 2 days apart are statistically significant given the other main effects in the model. In addition, there is no statistical evidence that the correlation function is not symmetrical because the magnitude of the parameter was not significantly affected by whether the explanatory *E. coli* concentrations were measured 2 days before or after the response *E. coli* concentrations. In other words, the variability in *E. coli* concentrations for a specific date is somewhat explained by *E. coli* concentrations 2 days prior and (or) 2 days later. The final form of dynamic regression equation includes the main-effect variables, even though the statistical significance of Clinton River discharge was somewhat greater than 0.05 in the subset model, after inclusion of the lag and lead explanatory $\log_{10}$ *E. coli* concentrations.

Although the preceding regression analyses identified explanatory variables that are statistically related to $\log_{10}$ *E. coli* concentrations at Memorial Beach, all the equations have limited explanatory ability as indicated by the adjusted R^2 values and the match between the measured $\log_{10}$ *E. coli* concentrations and those estimated by the main-effects model (fig. 9). In fact, 7 of the 11 measured concentrations that exceeded the Michigan recreational water-quality standard would not have resulted in exceedance based on the models estimation. Therefore, this model would not be useful as a predictive tool given the current set of data.

Table 1. Summary of linear regression models at Memorial Beach on Lake St. Clair, Mich., 2002–05.

[<, less than; NA, not applicable; NI, not included because factor was insignificant in the interaction-effects model]

Environmental factor	Parameter estimate (*p*-value) Main-effects model	 Interaction-effects model	 Dynamic-effects model
Parameter summary			
Intercept	0.88304 (<.001)	0.61808 (.0055)	0.32144 (.2431)
Rainfall_72	.20945 (.0004)	.31804 (.0030)	.21611 (.0101)
SeaTemp	.02536 (.0038)	.03206 (.0006)	.03162 (.0101)
ClintonFlow	.01200 (.0023)	.02209 (.0006)	.01620 (.0865)
Turbidity	.00174 (.0014)	.00352 (.0029)	.00236 (.0039)
Rainfall_72-ClintonFlow	NA	-.00434 (.1470)	NI
ClintonFlow-Turbidity	NA	-.00083 (.0866)	NI
LagLeadLogEC	NA	NA	.21648 (.0123)
Model summary			
Model F value	14.27 (<.0001)	10.48 (<.0001)	10.70 (<.0001)
Root mean square error	.53807	.53277	.54141
Adjusted R^2	.2778	.2919	.3119
Number of observations	139	139	108

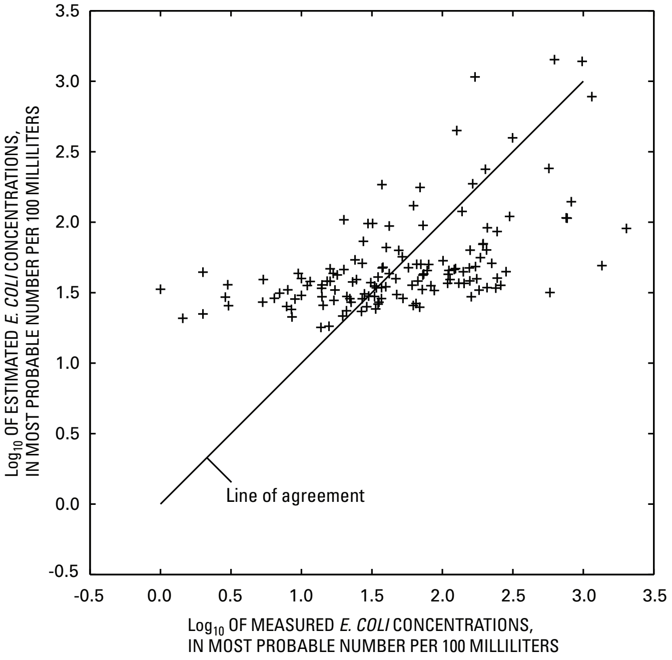


Figure 9. Relation between measured and main-effect model estimates of *E. coli* concentrations at Memorial Beach on Lake St. Clair, Mich., 2002–05.

Metropolitan Beach

A preliminary regression model for estimating $\log_{10}$ concentrations of *E. coli* at Metropolitan Beach was identified by use of the stepwise procedure. The main effects for this model included total rainfall at Selfridge ANGB during the 24 hours prior to bacteria sampling (Rainfall_24), sea-surface temperature (SeaTemp), turbidity, and the 1-hour change in smoothed water levels at the St. Clair Shores (DeltaWaterLevel) (table 2). In contrast to Memorial Beach, the Clinton River discharge was not significantly correlated with *E. coli* concentrations.

A stepwise evaluation of all first-order interaction terms from the main-effects model resulted in no additional variables at the 0.15 level of significance. Addition of wind speed (WindSpeed) and a north-wind indicator vector (NorthWindInd) marginally improved the accuracy of the model, although the individual terms were not statistically significant ($p = 0.05$). Finally, among the 143 *E. coli* measurements at Metropolitan Beach, 64 paired with explanatory *E. coli* concentrations in samples collected 2 days before the response variable, and 50 paired with explanatory *E. coli* concentrations 2 days after the response variables. In contrast to the Memorial Beach model, however, no dynamic components were significant at Metropolitan Beach when included with the main effects. Although the main-effects model was selected for the purpose of this report, it explains less than 30 percent of the variability in the $\log_{10}$ *E. coli* concentrations and so has limited utility for prediction of *E. coli* concentrations (fig. 10) with none of the 10 samples that exceeded the Michigan water-quality standard being predicted as exceedances by the model.

Table 2. Summary of linear-regression models at Metropolitan Beach on Lake St. Clair, Mich., 2002–05.

[<, less than; NA, not applicable; NI, not included because factor was insignificant in the interaction-effects model]

Environmental factor	Parameter estimate (p-value)		
	Main-effects model	Wind-effects model	Dynamic-effects model
Parameter summary			
Intercept	0.29630 (.0842)	0.11153 (.6055)	0.23413 (.2430)
Rainfall_24	.26538 (.0270)	.23488 (.0531)	.18185 (.1987)
SeaTemp	.05792 (<.0001)	.05810 (<.0001)	.05703 (<.0001)
Turbidity	.00477 (.0014)	.00399 (.0088)	.00438 (.0087)
DeltaWater-Level	-114.14097 (.0101)	-102.80 (.0237)134	-97.187 (.0829)
NorthWindInd	NA	.20155 (.0406)	NI
WindSpeed	NA	.01608 (.4283)	NI
LagLeadLogEC	NA	NA	.06563 (.4412)
Model summary			
Model F value	15.78 (<.0001)	11.55 (<.0001)	9.30 (<.0001)
Root mean square error	.56647	.56045	.58927
Adjusted R^2	.2935	.3084	.2685
Number of observations	143	143	114

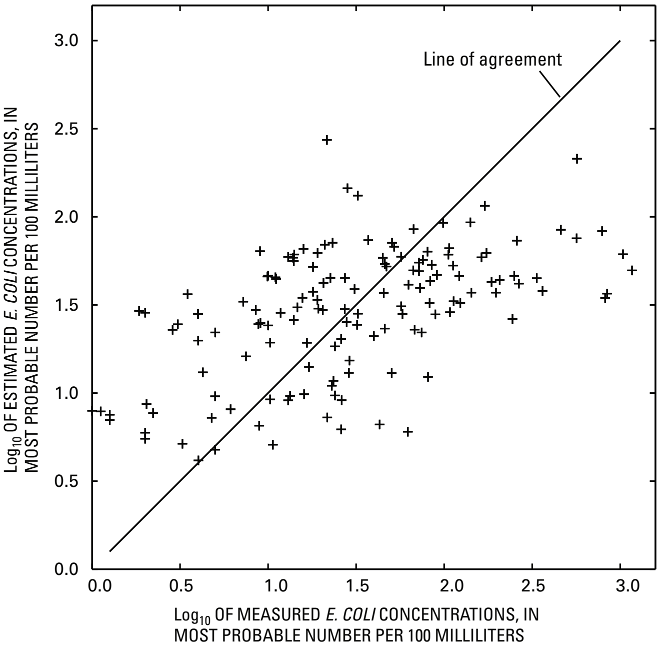


Figure 10. Relation between measured and linear regression estimates of *E. coli* concentrations at Metropolitan Beach on Lake St. Clair, Mich., 2002–05.

Regression-Tree Results

Regression trees were developed as an alternative to linear regression models for estimating *E. coli* concentrations. Like linear regression models, regression trees identify explanatory variables that are associated with different levels of *E. coli* concentrations. Regression trees are sometimes more effective than linear models when there is a nonlinear relation between environmental factors and *E. coli* concentrations or there are high-order interactions. The flexibility of the regression-tree structure, however, sometimes makes identification of relations in small data sets problematic.

Memorial Beach

A regression tree was constructed for the Memorial Beach $\log_{10}$ *E. coli* concentrations by use of the same explanatory variables identified for the main effects linear regression model (fig. 11). The resulting full tree contained 28 terminal nodes, 27 parent nodes (where explanatory variables are used for branching), and 23 levels of branching before pruning. In the model, the explanatory variable Rainfall_72 was used 3 times, SeaTemp 10 times, ClintonFlow 5 times, and Turbidity 9 times. *E. coli* concentrations were first partitioned by ClintonFlow. The second variable separating *E. coli* concentrations was Rainfall_72. A dynamic regression tree also was developed for the 108 observations at Memorial Beach where *E. coli* measurements 2 days apart could be paired. The RMSE of regression trees decrease monotonically with the number of terminal nodes.

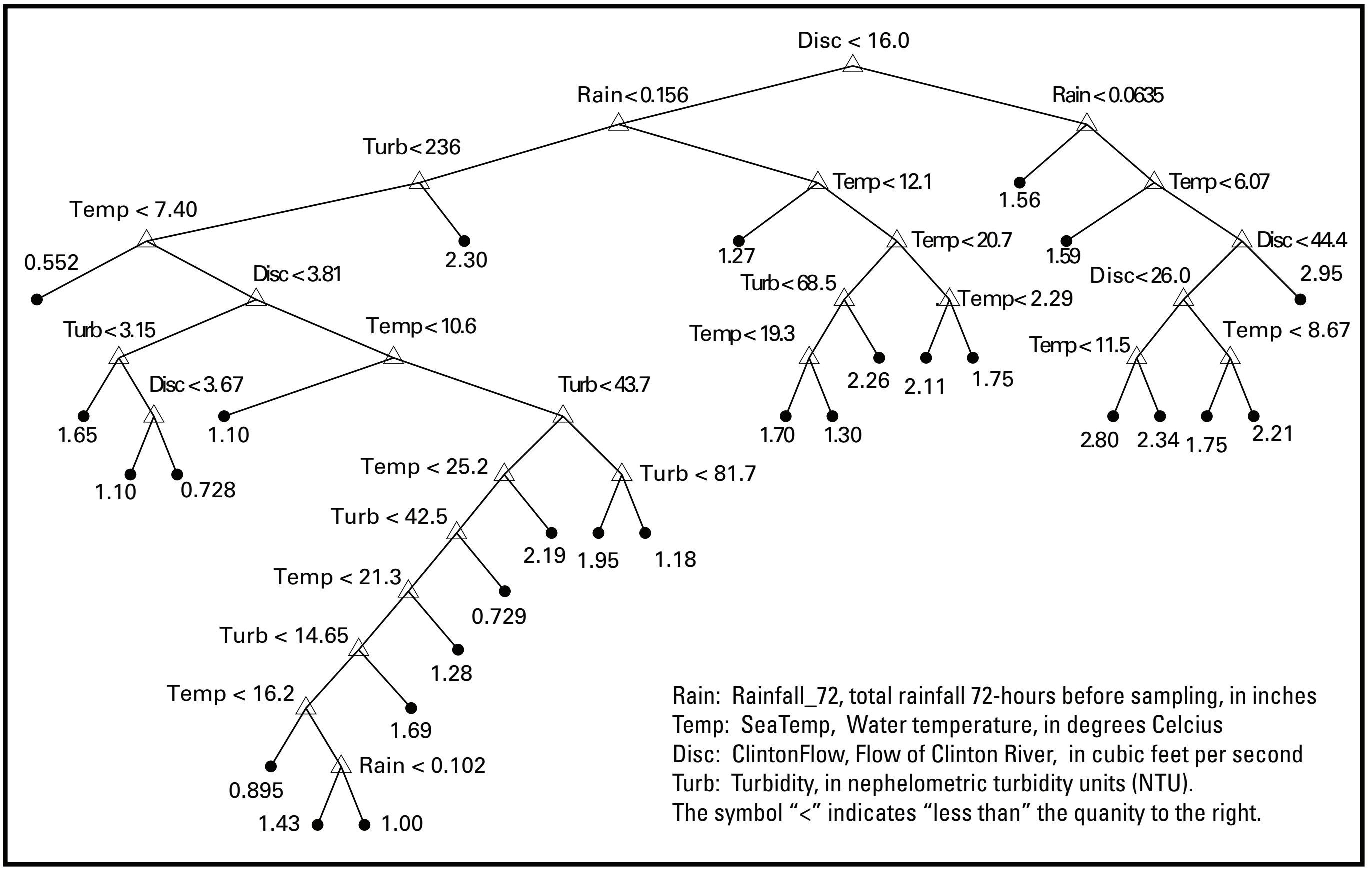


Figure 11. Regression tree for estimating $\log_{10}$ *E. coli* concentrations for Memorial Beach on Lake St. Clair, Mich., 2002–05.

As described in the approach, a statistical resampling analysis was used to evaluate the reproducibility of the regression trees. For data from Memorial Beach, the monotonic decrease in median RMSE of estimation (RMSE-E) with increasing number of terminal nodes (fig. 12) was not consistent with the nearly constant or slightly increasing median RMSE of prediction (RMSE-P). In addition, the interquartile ranges of RMSEs for prediction were nearly 3 times the RMSE for estimation in regression trees having 1 through 10 terminal nodes. This lack of reproducibility indicates that the regression trees may be overparameterized for the number of *E. coli* measurements available for model development and therefore not good predictors of *E. coli* concentrations.

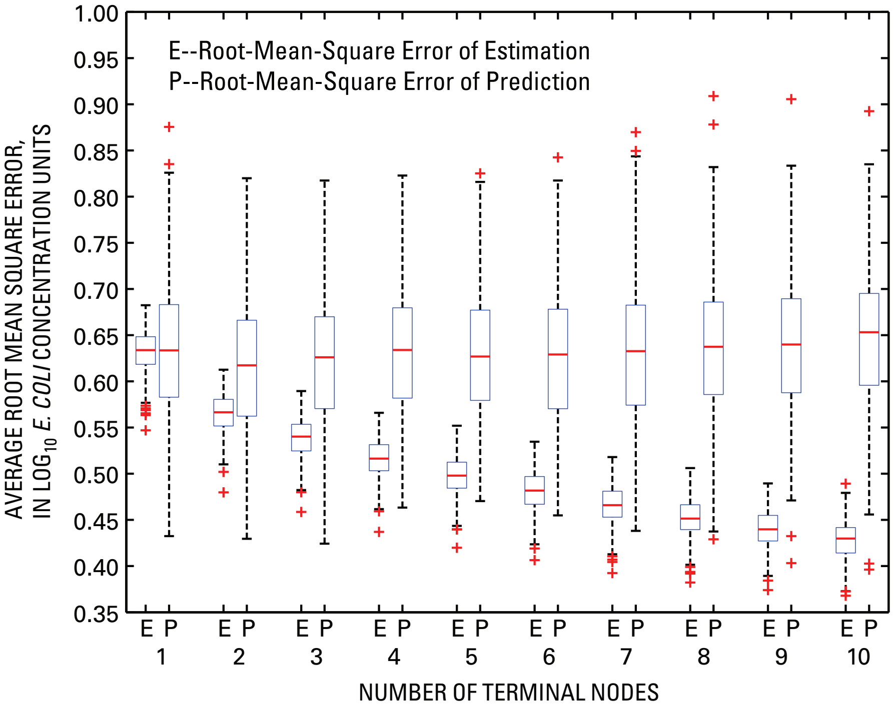


Figure 12. Regression tree root mean square error for estimation and prediction for Memorial Beach on Lake St. Clair, Mich.

Metropolitan Beach

A regression tree also was developed for the Metropolitan Beach $\log_{10}$ *E. coli* concentrations by use of the same three explanatory variables identified for the main-effects linear regression model. The resulting full tree for the variables in the main-effects model contained 33 terminal nodes, 32 parent nodes (where explanatory variables are used for branching), and 23 levels of branching. In this model, *E. coli* concentrations were first separated by SeaTemp, then by Turbidity.

In a resampling analysis similar to the one for Memorial Beach data, the median RMSE of estimation (RMSE-E) for the regression tree decreased monotonically with the number of terminal nodes, whereas the median RMSE-P remained constant or increased slightly. Again, the inconsistency between the RMSE-E and the RMSE-P indicates an overparameterization of the regression tree.

For Metropolitan Beach, where prior analyses had indicated a potential wind effect, regression trees were also constructed for two datasets partitioned by onshore or offshore winds. This analysis was not done for Memorial beach, where wind effects were not indicated. Concentrations of *E. coli* in the Clinton River Spillway, previously indicated to be correlated with *E. coli* concentrations at Metropolitan Beach (although not significant in the linear regression models) were also included in these regression-tree models. During onshore winds, *E. coli* concentrations were first separated by water temperature; if temperature was less than 17.1°C, the mean log *E. coli* was 1.2, and the branch was further subdivided by Clinton River Spillway discharge. If water temperature was greater than 17.1°C the mean $\log_{10}$ *E. coli* was 1.72, and this population was subdivided by turbidity. During offshore winds, *E. coli* concentrations were first split by Clinton River Spillway discharge, then by Clinton River Spillway *E. coli* concentration.

Logistic Regression Results

Logistic regression models were developed for Memorial and Metropolitan Beaches to provide a mechanism for estimating the probability of *E. coli* concentrations exceeding the water-quality standard as a function of selected environmental variables. Estimating probabilities is an alternative to estimating *E. coli* concentrations and comparing the estimates to the standard. Exponentiating parameters in the logistic model describe the odds of exceeding the standard with a unit change in the corresponding environmental variable, given that the other environmental variables do not change.

Memorial Beach

Of the 139 samples from Memorial Beach analyzed for *E. coli* concentration, 127 had concentrations less than Michigan's recreational water-quality standard, and 12 had concentrations that exceeded the standard. In a main-effects logistic model, Rainfall_72, ClintonFlow, and Turbidity were identified as significant explanatory variables (table 3). No interaction effects were significant. The parameters in the main-effects model indicate that all significant explanatory variables were positively associated with the odds of exceeding the beach-closure threshold. In particular, an increase in Rainfall_72 of 1 in. or more than doubles (2.36) the odds that *E. coli* concentrations will exceed the water-quality standard, assuming that the other explanatory variables remain unchanged.

Table 3. Summary of logistic regression models for *E. coli* concentrations at Memorial and Metropolitan Beaches on Lake St. Clair, Mich., 2002–05.

[<, less than; NA, not applicable]

Environmental factor	Memorial Beach (*p*-value)		Metropolitan Beach (*p*-value)
	Main-effects parameter	Dynamic-effects parameter	Main-effects parameter
Parameter summary			
Intercept	-4.5845 (<.0001)	-8.0100 (.0002)	-1.6317 (.0080)
Rainfall_72 exp(Rainfall_72)	.8588 2.360 (.0052)	1.4930 4.451 (.0052)	NA
Rainfall_24 exp(Rainfall_24)	NA	NA	3.3988 29.928 (.0013)
ClintonFlow exp(ClintonFlow)	.0387 1.039 (.0350)	NA	-.1918 .825 (.0205)
Turbidity exp(Turbidity)	.00688 1.007 (.0291)	.0133 1.013 (.0031)	NA
LagLeadLogEC exp(LagLeadLogEC)	NA	1.4343 4.197 (.0420)	NA
Intercept	-4.5845 (<.0001)	-8.0100 (.0002)	-1.6317 (.0080)
Rainfall_72 exp(Rainfall_72)	.8588 2.360 (.0052)	1.4930 4.451 (.0052)	NA
Model summary			
$-2LogL_{\alpha}$	81.722	61.957	72.489
$-2LogL_{\alpha+\beta' x}$	53.991	34.922	54.922
Chi-squared statistic	27.7313 (<.0001)	27.0343 (<.0001)	17.5667 (.0002)
Rank correlation c (area under the ROC)	.891	.933	.807
Number of observations	139	139	108

Similar to the linear regression model, a dynamic logistic regression also was developed for the 108 observations that could be formed by pairing an explanatory *E. coli* concentration (LagLeadLogEC) measured 2 days before or after the response *E. coli* concentration. Of these sample, 99 had concentrations less than 300 MPN/100 mL. An explanatory variable was initially included to indicate whether the explanatory *E. coli* concentration was sampled before or after the response. Results indicated that explanatory $\log_{10}$ *E. coli* concentrations were significant in estimating response *E. coli* exceedance probabilities, but the timing indicator was not (similar results to the linear regression model). Specifically, an increase of 1 $\log_{10}$ unit in explanatory *E. coli* concentrations would quadruple (4.197) the odds that the response *E. coli* concentration would exceed the beach-closure threshold, assuming the other explanatory variables did not change (table 3).

An ROC (receiver operating characteristics) analysis was done for the main-effects and the dynamic-effects logistic models for Memorial Beach (fig. 13). The area under the ROC for the main-effects model is estimated by *c* equal to 0.891, whereas the area under the ROC for the main and dynamic-effects model is slightly higher at c equal to 0.933. For both models, substantial gains in sensitivity are possible with modest loses in specificity above the arbitrary cut point of 0.5. For the main-effects logistic model of Memorial Beach, a cut point of 0.14 yields a sensitivity of about 0.83 and a specificity of 0.89 (false positive rate of 0.11). For the main- and dynamic-effects logistic model, a cut point of 0.23 yields a sensitivity of about 0.77 and a specificity of 0.92 (false positive rate of 0.08).

Metropolitan Beach

Of the 143 samples from Metropolitan Beach analyzed for *E. coli* concentration, 133 had concentrations less than Michigan recreational water-quality standard, and 10 had concentration that exceeded the standard. In a main-effects logistic model, Rainfall_24 and ClintonFlow were identified as significant explanatory variables. No interaction or dynamic effects were found to be significant. In the main-effects model, *E. coli* exceedance odds increase nearly thirtyfold (29.9) for each inch of increase in total rainfall during the 24-hour period preceding sampling. In contrast, an increase of 1 $\log_{10}$ flow of Clinton River slightly decreases the odds of *E. coli* exceedances.

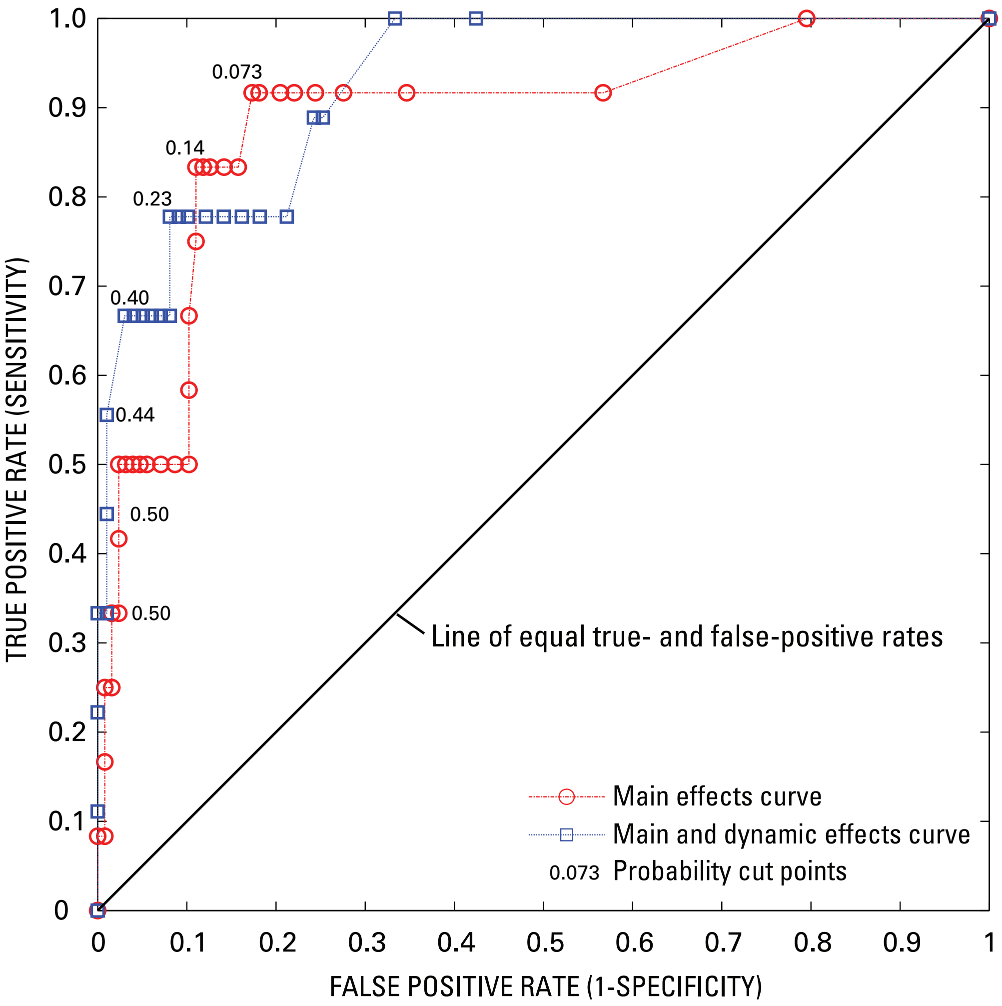


Figure 13. Receiver operating characteristics (ROC) curves for logistic equations predicting the probability of *E. coli* concentrations exceeding 300 MPN/100 mL at Memorial Beach on Lake St. Clair, Mich., 2002–05.

An ROC analysis was done for the main-effects logistic model for Metropolitan Beach (fig. 14). A cut point of 0.11 yields a sensitivity of about 0.70 and a specificity of 0.85 (false positive rate of 0.15). In contrast, a cut point of 0.5 yields a sensitivity of only 0.2, with a specificity of 0.99.

Relation of Flow Paths to *Escherichia coli* Concentrations

Because lakeshore beaches are part of a very dynamic system of hydrologic flow paths, it is important to understand these paths to identify potential source waters to the beach areas. Hydrodynamic modeling is a means of tracking the path particles may take under a given set of wind conditions. This study builds on the particle-tracking hydrodynamic model created specifically for Lake St. Clair.

Flow Simulation

The Lake St. Clair model referred to in this report is a subset of the 2D hydrodynamic model of the St. Clair-Detroit River Waterway model developed for the MDEQ SWAP. This version uses the same geometry, bathymetry, and hydraulic parameters as the standard model, except in areas excluded from the model domain. The Lake St. Clair version extends from a point near Algonac, Mich., on St. Clair River, through Lake St. Clair to the Fort Wayne water-level gaging station (station ID 9044036) operated by NOAA on the Detroit River.

With respect to the standard model version, the mesh of the Lake St. Clair version has been refined northwest of the boundary between the United States and Canada (fig. 1). Refinement implies that arcs (lines) connecting opposing midside nodes are formed and used to subdivide each quadrilateral and triangular element into four subelements. The subelements have the same number of sides as the parent element. Nodes are added at the intersection of crossing arcs in quadrilateral elements and to the midsides of all refined elements to maintain their quadratic characteristic. The Lake St. Clair version contains 9,784 elements delimited by 28,749 nodes.

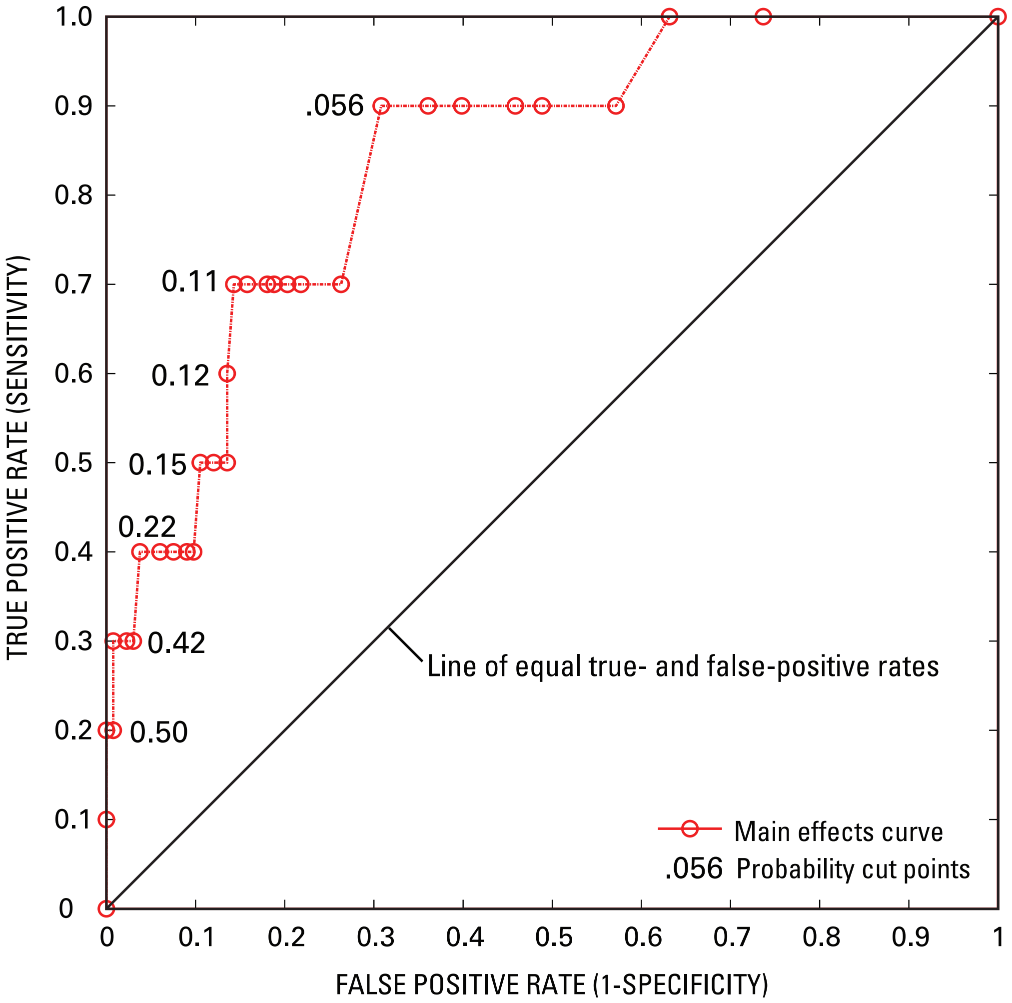


Figure 14. Receiver operating characteristics (ROC) curve for logistic equations predicting the probability of *E. coli* concentrations exceeding 300 MPN/100 mL at Metropolitan Beach on Lake St. Clair, Mich., 2002–05.

Particle Tracking

The paths that particles take may be important to the transport of *E. coli* to and from the beach. These paths may be highly variable throughout the year because they often depend on wind direction.

Steady-State Simulations

Steady-state hydrodynamic simulations were used to describe current patterns on Lake St. Clair for wind directions discretized into 24 sectors. Simulated current patterns on Lake St. Clair are primarily determined by wind conditions rather than flow magnitudes (Holtschlag and Koschik, 2004). Therefore, identification of steady-state flow paths to Memorial and Metropolitan Beaches were described as a function of wind conditions, while flow and water-level specifications were held constant. In particular, St. Clair River flow was specified as 185,000 ft^3/s, Clinton River flow was 637 ft^3/s, Sydenham River flow was 1,860 ft^3/s, and Thames River flow was 637 ft^3/s. The water level specified for the boundary at the Fort Wayne, Mich. was 573.19 ft (IGLD 85).

In this report, wind characteristics are summarized on the basis of LSCM4 data obtained from November 8, 2001 (the first available data), to October 23, 2003, when this information was initially compiled. Vector averages of data at 10-minute intervals were used to compute daily mean values, which were summarized in a wind chart (fig. 15). In this chart, wind directions were divided into 24 sectors, each having an angular width of 15 degrees. Sectors are numbered counterclockwise, starting at the sector from 90 to 75 degrees east of north. Within each sector, annuli describe the percentage of time that daily mean wind speeds were within specified 5-mi/h intervals. Average wind speeds and directions were computed for each sector. Average wind directions were transformed following the mathematical convention for input into RMA2, which describes the direction that winds are blowing toward and references angles counterclockwise from the positive *x*-axis (zero degrees east). Using the compass convention, a wind reported by NDBC from the southeast, at 130 degrees, for example, would be described in a mathematical convention as a wind blowing toward the northwest at 140 degrees for use in RMA2.

In this analysis, particle tracking scenarios were used to identify flow paths and traveltimes for idealized steady-state conditions, where simulated current patterns on Lake St. Clair varied only as a function of wind sector. In the 24 scenarios, hypothetical particles were initially placed near Memorial and Metropolitan Beaches corresponding to their ending locations. Then, reverse particle tracking was used to map expected flow paths backward in time towards source areas. Expected flow paths do not reflect the increasing uncertainty of particle locations with increasing simulated traveltime.

Subsets of contiguous wind sectors produced similar flow paths. For Memorial Beach, sector 1 simulation results (fig. 16) are similar to results from sectors 1–3, and 20–24, where particles generally maintain a southwestern flow direction along the northwestern shoreline of Lake St. Clair. In contrast, results from sector 9 and 14 simulations, which are representative of sectors 4–14, show particles that tend to initially flow in a southerly direction just east of L'anse Creuse Bay but revert to a northerly flow direction along the western shoreline of Lake St. Clair as they passed south of L'anse Creuse Bay. Sector 15 simulation resulted in a recirculation pattern within L'anse Creuse Bay. Finally, results from sector 17, which is representative of sectors 16–19 are similar to those for sector 1 with the exception that these sectors circumvent a circulation pattern south of Metropolitan Beach area. Markers are placed along flow paths at 12-hour increments for 168 hours from their destination near the beaches or until they were simulated out of the model domain.

For Metropolitan Beach, particle tracks resulting from simulations with sector 1 winds are similar to those for sectors 1–8 and 19–24 (fig. 17). In these simulations, flows are southerly past Clinton River and along the north shore of Anchor Bay from St. Clair River. Sector 10 is representative of sectors 9–11, where simulated particles flow are northerly past Cutoff Canal and Memorial Beach after flowing toward the south east of L'anse Creuse Bay. Finally, results for sector 15 are representative of those for sectors 12–18, which result in a simulated recirculation pattern within L'anse Creuse Bay south of Metropolitan Beach.

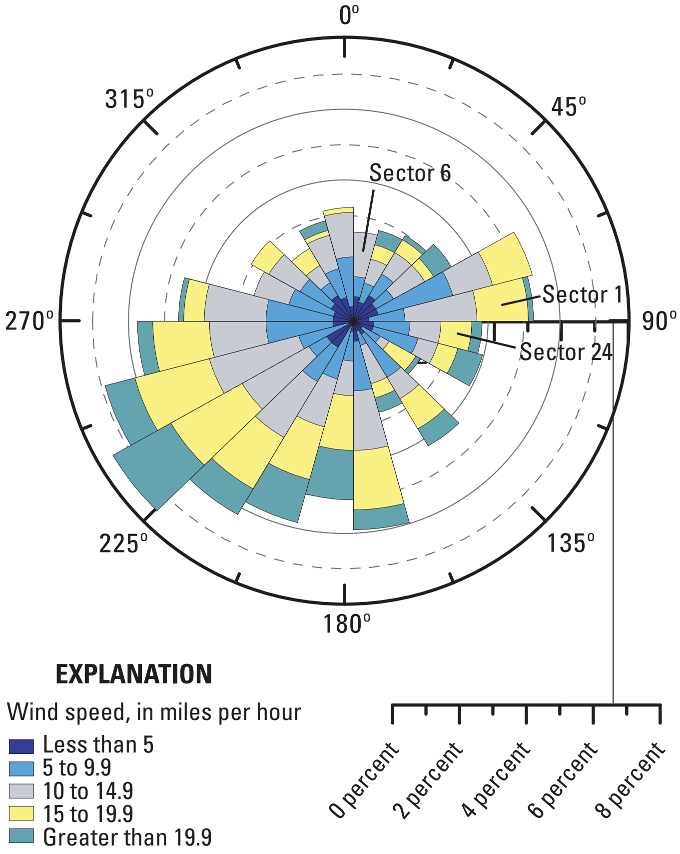
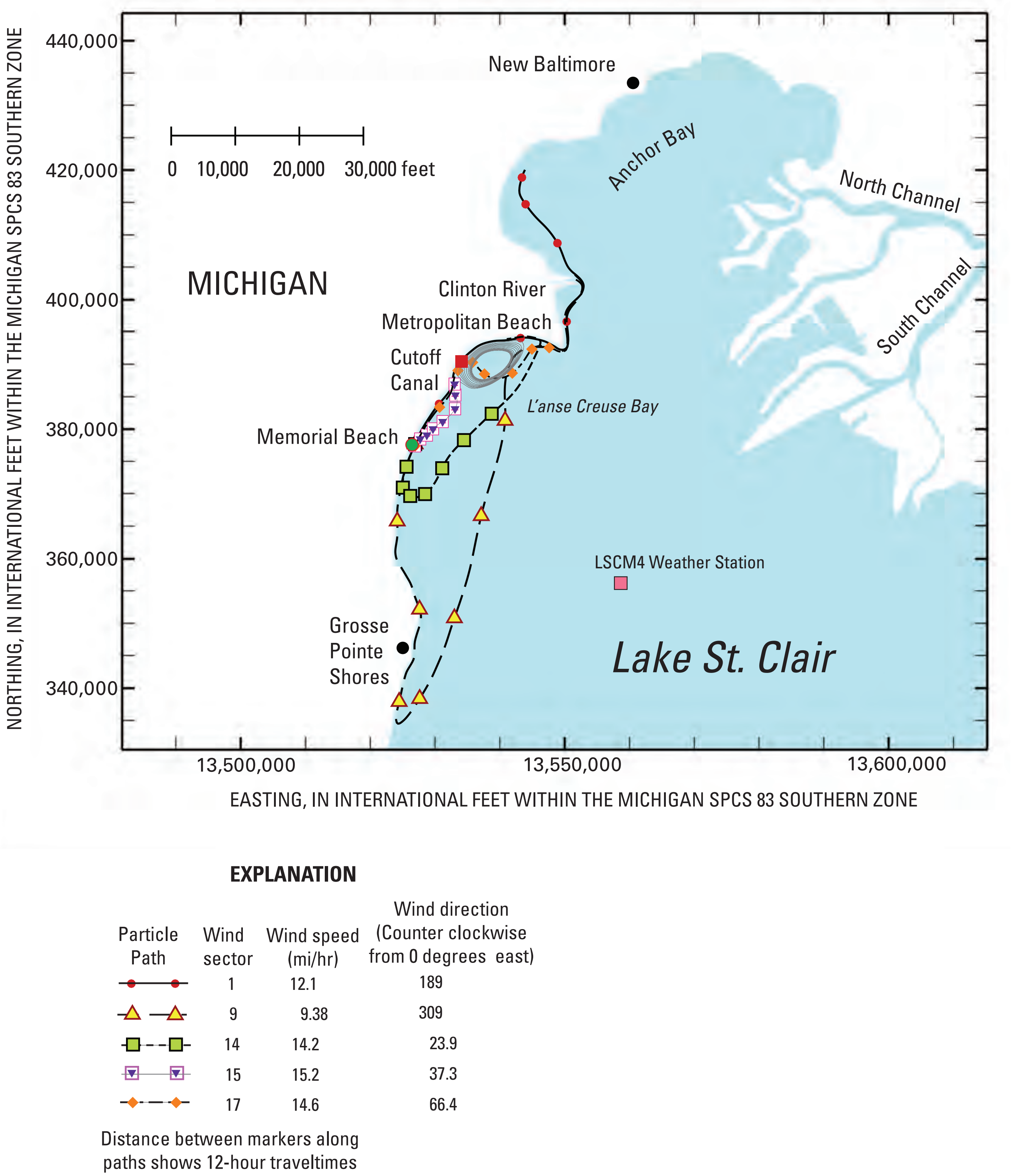


Figure 15. Wind characteristics, by sector, on Lake St. Clair based on data from Lake St. Clair Monitoring Station LSCM4 in Michigan. (The location of station LSCM4 is shown on fig. 1.)

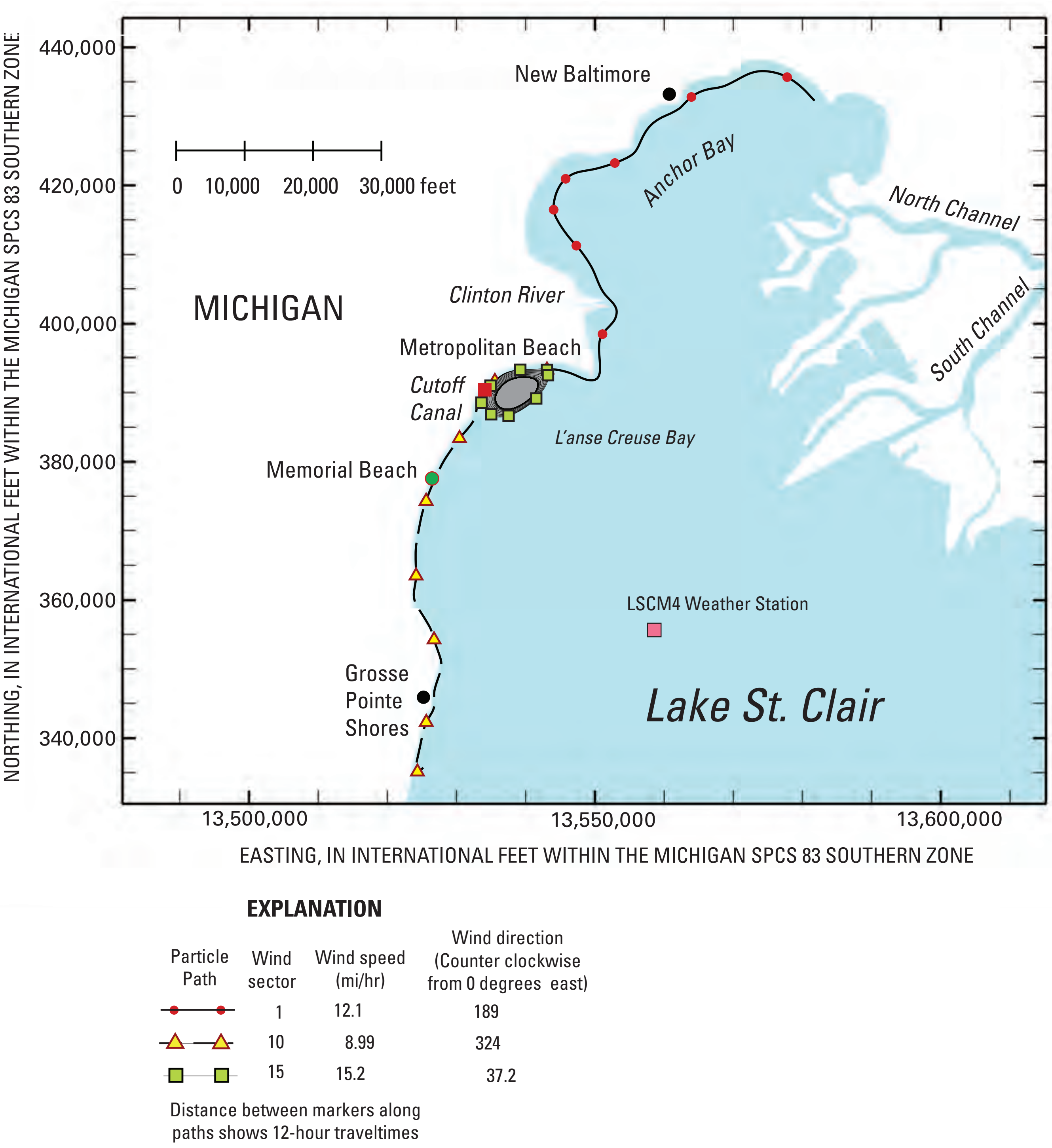


EXPLANATION

Particle Path	Wind sector	Wind speed (mi/hr)	Wind direction (Counter clockwise from 0 degrees east)
●—●	1	12.1	189
△—△	9	9.38	309
■--■	14	14.2	23.9
▼—▼	15	15.2	37.3
◆-·-◆	17	14.6	66.4

Distance between markers along paths shows 12-hour traveltimes

Figure 16. Simulated reverse particle paths from Memorial Beach for selected wind sectors on Lake St. Clair, Mich.



Particle Path	Wind sector	Wind speed (mi/hr)	Wind direction (Counter clockwise from 0 degrees east)
●—●	1	12.1	189
△—△	10	8.99	324
◻—◻	15	15.2	37.2

Distance between markers along paths shows 12-hour traveltimes

Figure 17. Simulated reverse particle paths from Metropolitan Beach for selected wind sectors on Lake St. Clair, Mich.

Transient Simulations

Transient flow simulation and reverse particle-tracking analyses were used to map flow paths from beach areas towards source areas for 10 selected *E. coli* sampling events. These events took place on Memorial and Metropolitan Beaches on Wednesday mornings (table 4). Data for both beaches included samples where *E. coli* concentrations were less than 20 and greater than 1,000 MPN/mL. There is no apparent serial correlation in the bacteria data, nor is there significant cross-correlation between *E. coli* concentrations sampled on the same day at the two beaches.

Transient boundary conditions were used to simulate hourly flows during the selected *E. coli* sampling events. Hourly flows of St. Clair River were computed on the basis of the standard stage-fall-discharge rating, which was developed by the U.S. Army Corps of Engineers (John Koschik, U.S. Army Corps of Engineers, written commun., 2005) and supported by hourly water-level information from NOAA. Hourly flows for Clinton River were based on data from the USGS streamflow-gaging station on Clinton River at Mount Clemens, Mich. (station 04165500). Hourly water-level data were from the NOAA gage at Fort Wayne, and wind conditions were from the Lake St. Clair, Mich., weather station (LSCM4). Constant flows were specified for the Sydenham and Thames Rivers in Canada.

Particle tracks at both Memorial and Metropolitan Beaches generally move laterally along the beach; but unlike steady-state simulations, particles can reverse directions along the same flow path and can cross flow paths from other events in response to transient wind and flow conditions. Visualizing the full results from a static display is problematic. On figures 18 and 19, simulated flow paths were color-coded by bacteria-concentration intervals. The complexity of the flow paths and the irregularity of the relation between flow path and bacteria concentrations confirms the need for model simulations and particle tracking, if mapping flow paths associated with individual sampling events is desired.

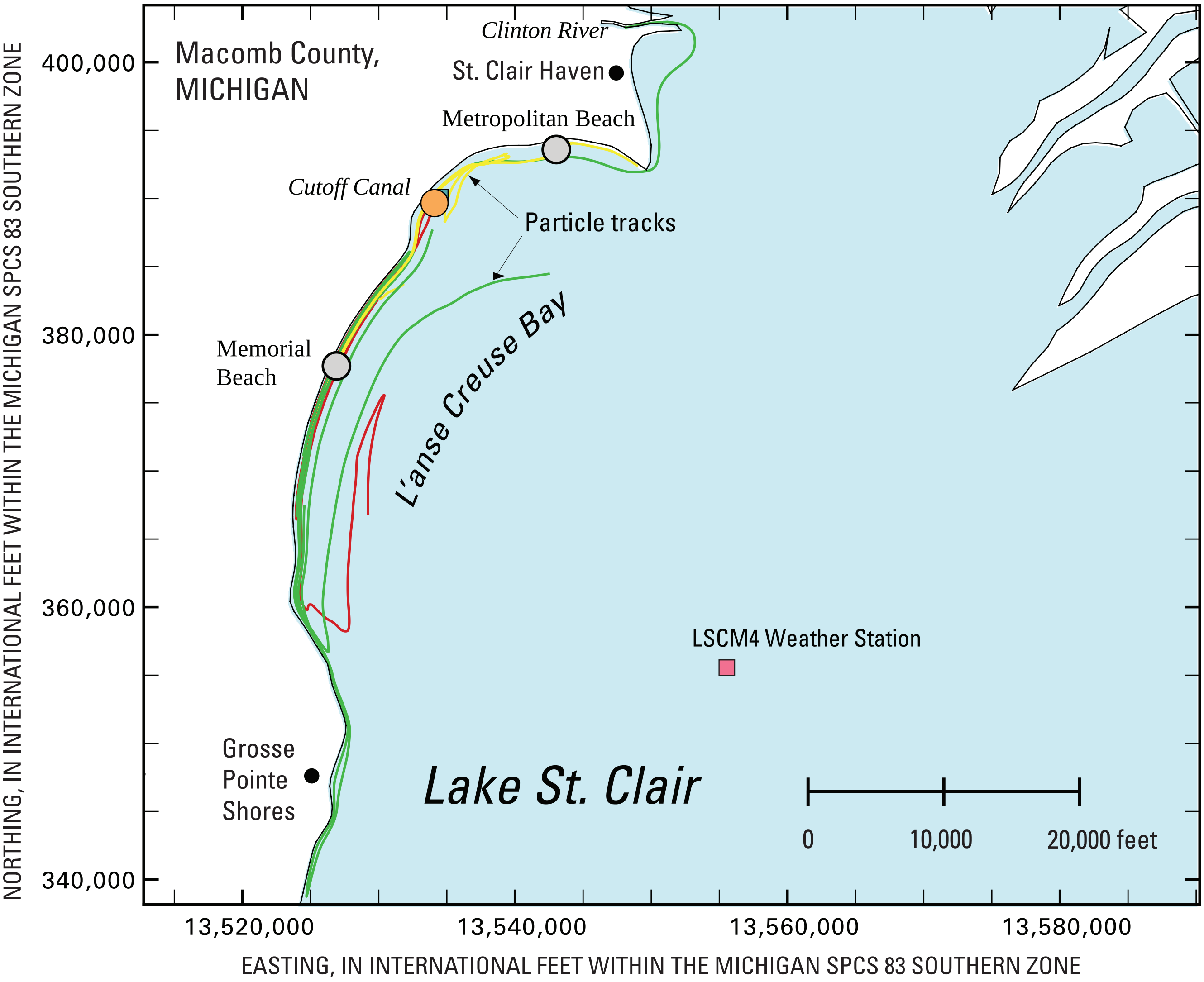
Numerous potential sources of *E. coli* bacteria to Memorial Beach are within the probable paths indicated by hydrodynamic simulations. Sources include the Clinton River Cutoff Canal (Spillway), for which Macomb County Health Department data show bacteria concentrations > 1,000 MPN/100 mL for most samplings that follow rainfall events. However, 16 drains, and 2 small rivers also deliver water to the shoreline within 3 mi north or south of Memorial Beach (Buzonik and Shoemaker, 2004). MCHD has evaluated nearshore bacteria concentrations (offshore of each of these drains) for several years. Annual geometric mean concentrations tend to be well below the recreational water-quality standard for the nearshore sampling sites (Buzonik and Shoemaker, 2004), but data for the water quality in the drains themselves is not available. Given that particle tracks indicate movement up and down this shoreline for most hydrodynamic conditions and that dynamic regression models indicate both *E. coli* concentrations 2 days presampling and postsampling are significant variables, it is likely that these numerous sources contribute to Memorial Beach water quality.

Steady-state hydrodynamic simulations indicated that when winds were from the WNW. or NNW. (sector 1), particles generally originated near the mouth Clinton River within 24 hours of their sampling at the beach. An exceedance of the recreational water-quality standard at Metropolitan Beach occurred on only 1 of the 10 dates for which transient hydrodynamic particle tracking was done August 25, 2004. On this date, the average wind direction for the preceding 24 hours was from the WNW., and there had been 0.38 in. of rain. Particle tracks for the August 25 event indicate particles originating from along the shoreline east of the beach, as well as from the Clinton River, within the preceding 24 hours. It appears likely that the source of contamination to Metropolitan Beach on this date was the Clinton River.

Table 4. Selected *E. coli* sampling events on Memorial and Metropolitan Beaches on Lake St. Clair, Mich.

[*E. coli, Escherichia coli*; CFU/100 mL, colony-forming units per 100 milliliters]

Date of *E. coli* sample	Memorial Beach		Metropolitan Beach		Simulation start time (EST) on date of sample	Number of simulated hours
	Time of sample (EST)	*E. coli* concentration (CFU/100 mL)	Time of sample (EST)	*E. coli* concentration (CFU/100 mL)		
May 26, 2004	8:34 a.m.	982	9:32 a.m.	141	9:00 a.m.	145
June 9, 2004	7:20 a.m.	39	7:55 a.m.	31	9:00 a.m.	141
June 21, 2004	7:16 a.m.	70	7:48 a.m.	170	9:00 a.m.	141
July 28, 2004	7:32 a.m.	16	8:22 a.m.	3	9:00 a.m.	141
Aug. 4, 2004	7:30 a.m.	1,150	8:24 a.m.	52	9:00 a.m.	141
Aug. 11, 2004	7:28 a.m.	18	8:18 a.m.	10	9:00 a.m.	141
Aug. 18, 2004	7:26 a.m.	17	8:18 a.m.	122	9:00 a.m.	141
Aug. 25, 2004	7:20 a.m.	37	8:02 a.m.	1,037	8:00 a.m.	137
Sept. 1, 2004	7:54 a.m.	172	8:42 a.m.	14	9:00 a.m.	141
Sept. 29, 2004	8:25 a.m.	340	9:16 a.m.	4	9:00 a.m.	141



EXPLANATION

Particle track	*E. coli* concentration (MPN / 100 mL)
🟥	> 500
🟨	<= 500 to > 50
🟩	<= 50 to > 10

Figure 18. Reverse particle tracks for selected sampling events in 2004 starting from Memorial Beach on Lake St. Clair, Mich.

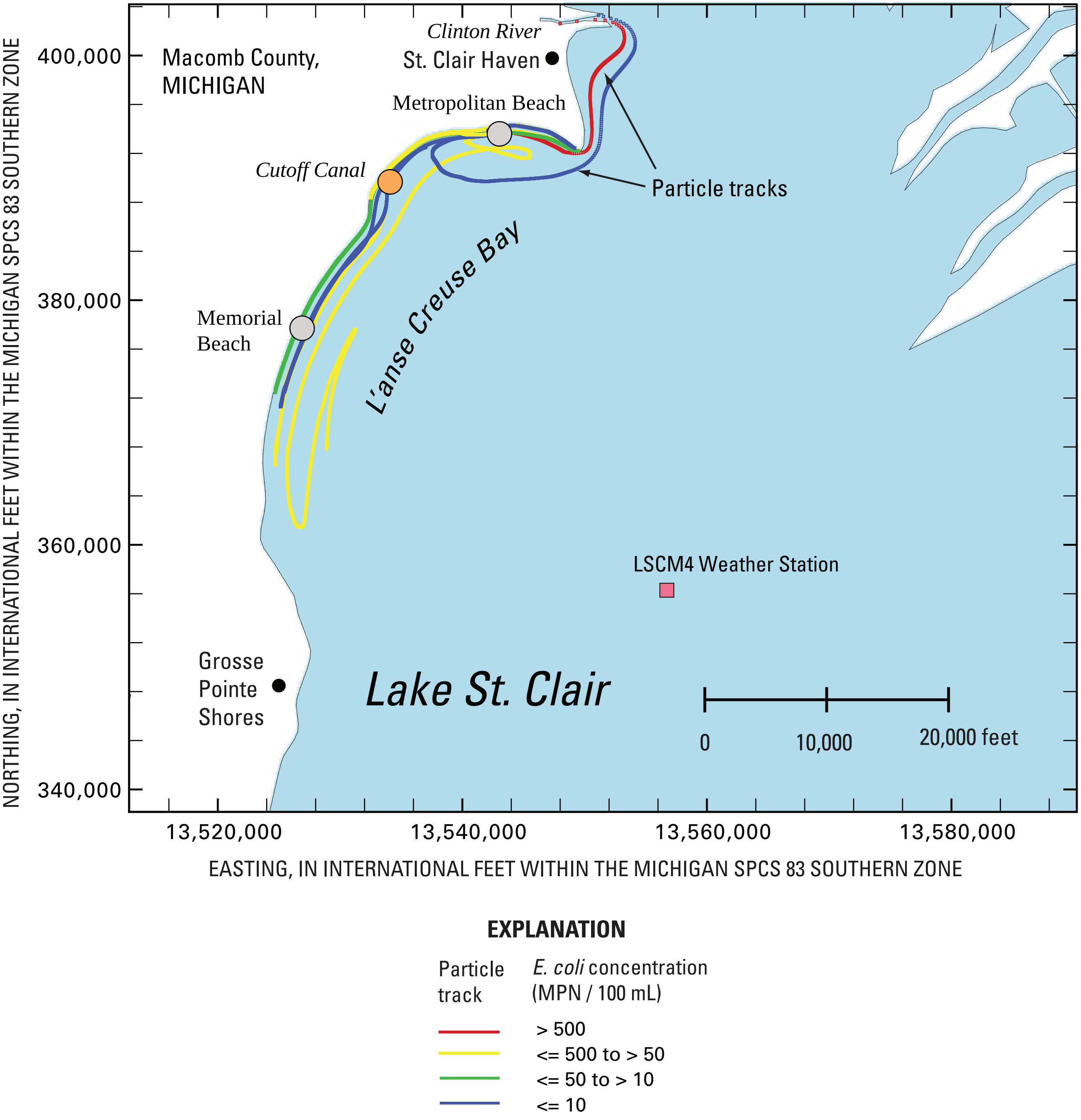


Figure 19. Reverse particle tracks for selected sampling events in 2004 starting from Metropolitan Beach on Lake St. Clair, Mich.

Additional Analysis of Period Covered by Transient Hydrodynamic Simulation

To further explore factors that may influence bacteria concentrations at Memorial and Metropolitan Beaches, the period from May 1 to September 30, 2004, was examined in more detail. This period encompasses the 10 transient particle-tracking simulations for the 2 beaches.

Sky cover (cloudiness or solar radiation) can influence bacteria concentrations in water (Davies-Colley and others, 1994). Between May 1 and September 30, 2004, 41 bacteria samplings were done at Memorial Beach and 46 at Metropolitan Beach. Table 5 lists the distribution of these samplings and the maximum *E. coli* concentrations for various sky-cover conditions. There were no exceedances of the recreational water-quality criterion when clear skies prevailed at the time of sampling. Preceding 24 and (or) 72 hour rainfall was statistically significant factor in several regression models. Therefore, weather conditions appear to affect the concentrations of *E. coli* at these beaches, but whether this is due to increased sunlight or effects of rainfall or a combination of both is uncertain.

Changes in water level induce dynamic flow patterns within beach sands and may wash material collected along the shoreline (algae, bird feces) into the nearshore waters. Several studies have shown that *E. coli* are retained in nearshore beach sand and may affect swimming waters (Alm and others, 2003; Haack and others, 2003; McLellan and Salmore, 2003; Whitman and others, 2003; Francy and others 2006). Changes in water level, coupled with wind patterns, may induce flow reversals in the Clinton River, sometimes for significant distances upstream (W.F. Baird & Associates, 2005). Figure 20 shows the relation between atmospheric pressure and water level at the St. Clair Shores recording station for May 2004. A shift from low water levels in winter/early spring to higher water levels in early summer is common in Lake St. Clair (Bolsenga and Herdendorf, 1993) and is evident in figure 20. The figure also shows a general pattern of decreasing atmospheric pressure corresponding to a general increase in water level towards the end of the month (May 2005), consistent with the passage of weather systems across the Great Lakes.

Figure 21 shows rainfall and *E. coli* concentrations at Memorial and Metropolitan beaches, along with water levels for the 2004 during this same period. *E. coli* concentrations at both beaches increased generally as water levels increased in May. This relation may result from washing of materials accumulated on the shoreline during previously low water levels. Superimposed on this increase were several substantial rain events. *E. coli* concentrations at Memorial Beach increased rapidly in response to rain, but *E. coli* concentrations at Metropolitan Beach increased less dramatically. This difference might result from the greater density of sources (drains and rivers) surrounding Memorial Beach as opposed to Metropolitan Beach. It appears that other factors affect beach waters at Memorial and Metropolitan beaches, as is evident by exceedances at both beaches during periods of no precipitation (May 17 at Memorial Beach and August 25 at Metropolitan Beach; figs. 21 and 23).

June and July water levels were more stable than in May (fig. 22). There was little rainfall in these months, and few there were exceedances of the recreational water-quality standard at either beach. In August and September, water levels typically start to decline as discharge from rivers begins to decrease (Bolsenga and Herdendorf, 1993). *E. coli* concentrations at both beaches generally followed this decline (fig. 23). The MCHD found that *E. coli* concentrations in foreshore sand at Metropolitan Beach were correlated with beach-water *E. coli* concentrations during sampling in 2004 (Buzonik and Shoemaker, 2004). No such correlation was reported for Memorial Beach. An increase in water level at this beach might have entrained *E. coli* in sand or materials (algae, gull feces) deposited on the beach during the preceding dry period (in this case, 11 days). Although particle tracking indicated the Clinton River as a probable source for the August 25 exceedance, beach sources also may have contributed to this event.

Table 5. Number of sampling events and maximum *E. coli* concentrations associated with various sky-cover conditions, May 1 to September 30, 2004, on Lake St. Clair, Mich.

Beach	Sky cover			
	Clear	Scattered (1/8–4/8 cover)	Broken (5/8–7/8 cover)	Overcast
Memorial Beach				
Number of samplings	8	15	12	6
Maximum *E. coli* concentration	245	1,167	1,418	363
Median *E. coli* concentration	73	71	57	90
Metropolitan Beach				
Number of samplings	8	14	13	11
Maximum *E. coli* concentration	172	6,881	982	1,150
Median *E. coli* concentration	42.5	130	42.5	21

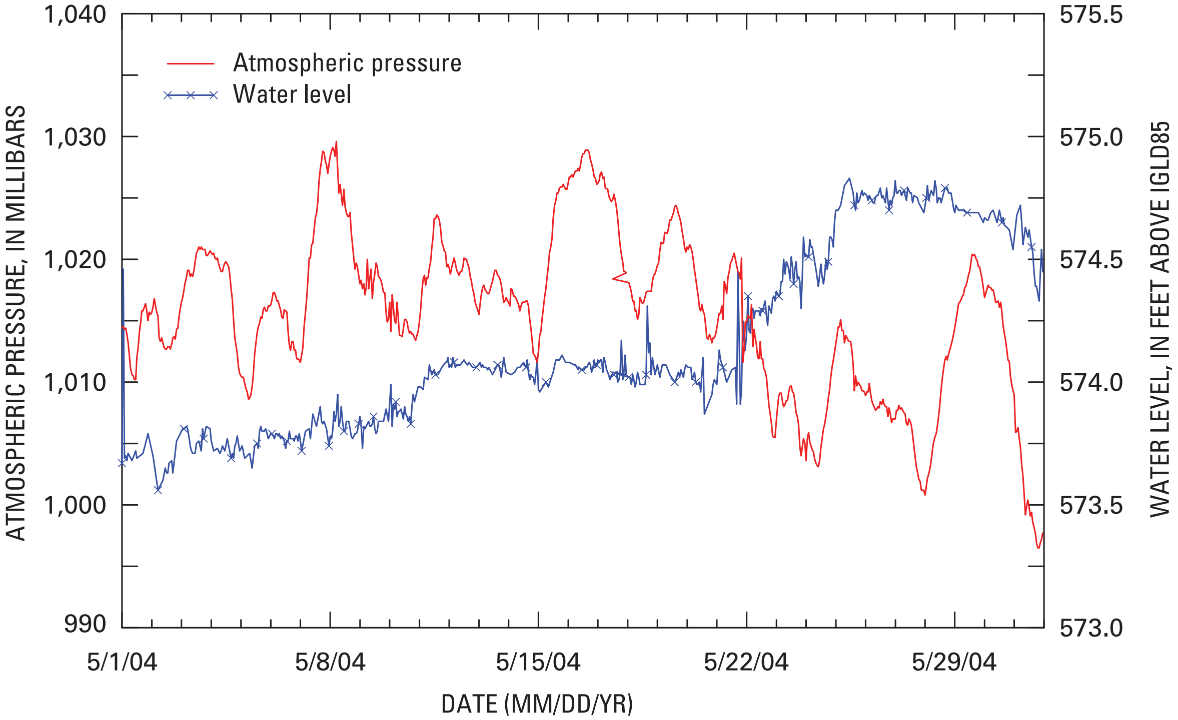


Figure 20. Atmospheric pressure and water levels in May 2004 at St. Clair Shores, Mich.

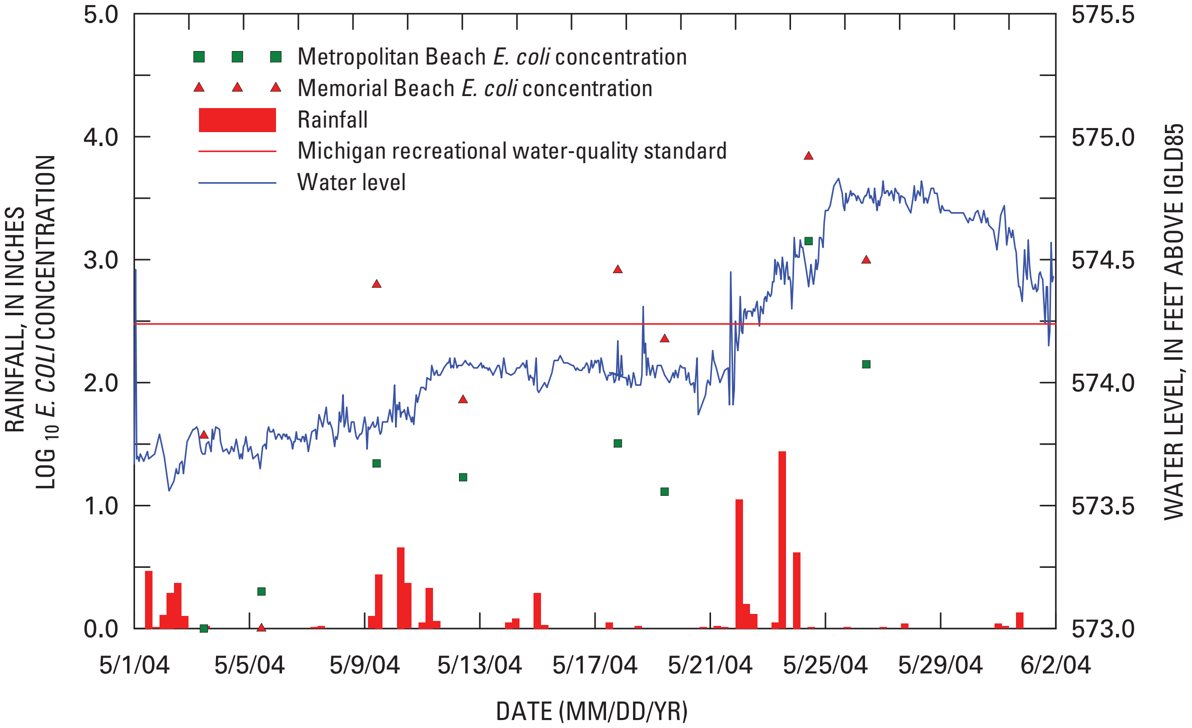


Figure 21. Rainfall, *E. coli* concentrations, and water levels in May 2004 on Lake St. Clair, Mich.

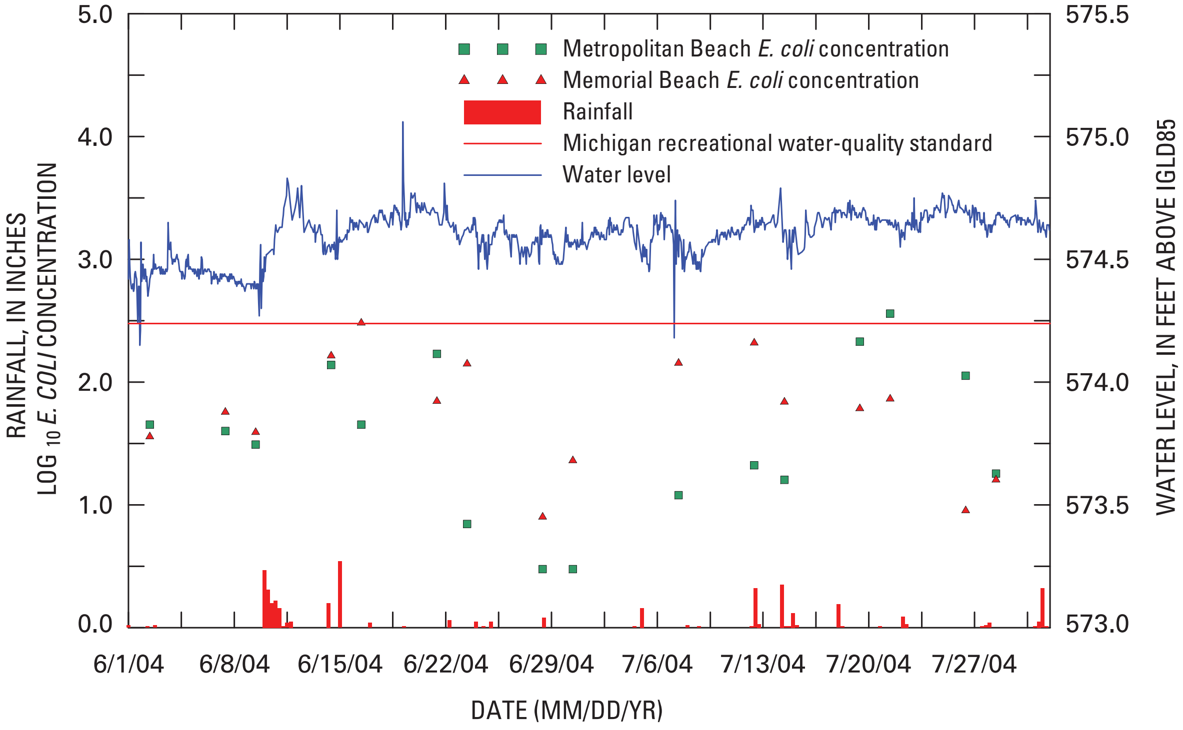


Figure 22. Rainfall and water levels in June and July 2004 on Lake St. Clair, Mich.

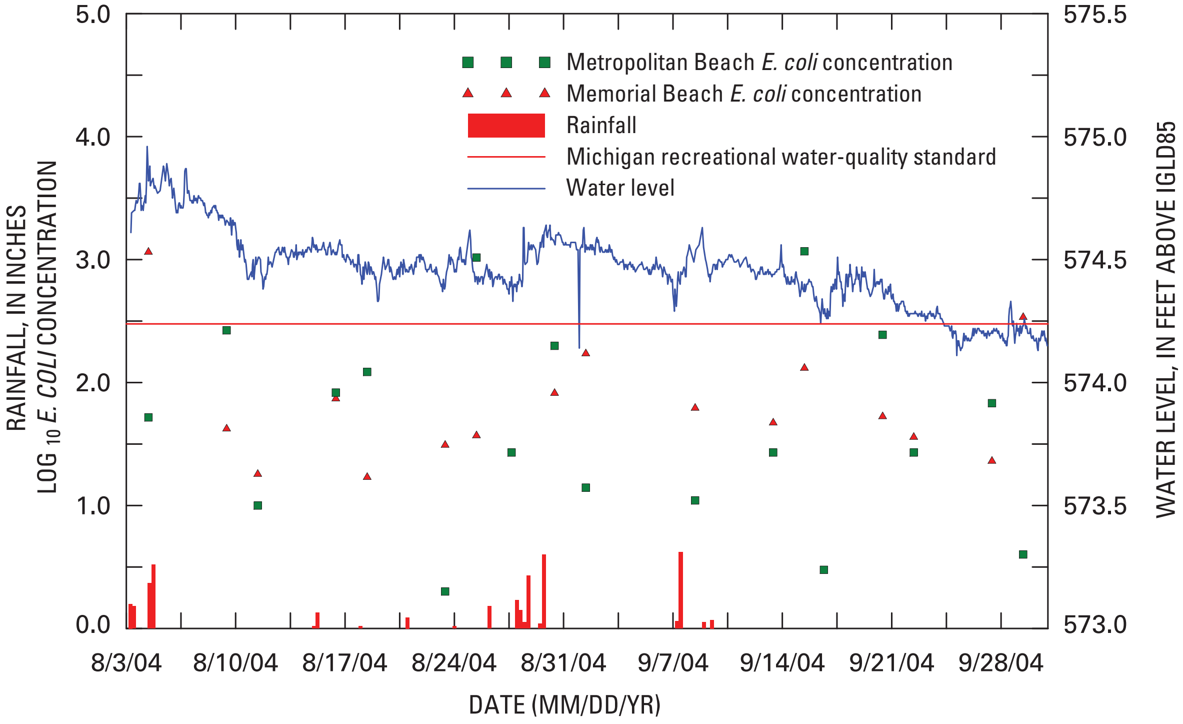


Figure 23. Rainfall and water levels in August and September 2004 on Lake St. Clair, Mich.

Suggestions for Future Studies

This study has demonstrated that circulation patterns and associated flow paths to Memorial and Metropolitan Beaches on Lake St. Clair are transient phenomena that are sensitive to variable wind conditions and uncertain dispersion characteristics. This sensitivity and uncertainty demonstrates the utility of hydrodynamic simulation for identifying flow paths, travel and exposure times, and source areas of bacteria to beaches. Changes in the flow paths alone may change the concentration or activation of *E. coli* bacteria during transit as a result of dieoff from solar radiation or other factors (Davies-Colley and others, 1993).

The massless hypothetical particles used in this report have limitations for identifying source areas of bacteria. Bacteria may be preferentially associated with particles having mass rather than moving strictly with the advective component of flow. Settling of these actual particles may significantly reduce bacteria concentrations with time and travel distances. Similarly, resuspension of these particles may be associated with wind and wave conditions on Lake St. Clair, which may be unrelated to bacteria concentrations in source areas, unless source areas are defined as anywhere along historic flow paths.

Despite these limitations, particle tracking yields information on the expected flow paths, traveltimes, and the integrative effects of transient hydrodynamic conditions on particle movement that is associated with the transport of bacteria. This information may be critical for identifying source areas, particularly if bacteria discharges in these areas are highly variable overtime. MCHD bacteria samples acquired at beaches and other selected areas at the rate of two per week have shown substantial variability and limited temporal correlation. Given the temporal variability of bacteria concentrations at the beaches, however, suspected source areas or beaches would need to be monitored more frequently than twice a week. Otherwise, the uncertainty in interpolating bacteria concentrations between samples would diminish the usefulness of particle tracking. Sampling only twice per week, therefore, it is difficult to relate specific samples of *E. coli* concentrations at beaches with samples at suspected source areas.

Future investigations of the temporal variation in bacteria concentrations would benefit from event sampling; that is, a series of frequent samplings (every few hours) at suspected source areas during two or more rainfall-runoff events when large variations in concentrations were expected over a short period of time. In addition, similar sampling would be needed at both beaches—in addition to suspected source areas—so that the dynamics of *E. coli* concentrations can be characterized. This type of data could help improve understanding of the uncertainty of the simulated flow paths.

Loading rate is the product of flow and concentration data. For measurements of bacteria concentrations near the USGS streamflow-gaging station on Clinton River at Mount Clemens, Mich. (station 04165500), supporting streamflow data are available at 15-minute intervals. At other sites, such as the spillway on the Cutoff Canal, it may be possible to compute unit streamflow values on the basis of hydraulic information and continuous water-level data. Supporting measurements of *E. coli* at nearshore and offshore sites on Lake St. Clair would help verify dispersion and mixing characteristics.

The hydrodynamic model of Lake St. Clair developed for this analysis could serve as a basis for future water-quality investigations. In particular, the velocity field simulated by an RMA2-based hydrodynamic model is the primary input to an RMA4-based water-quality model. RMA4 (Letter and others, 2005) is a finite-element water-quality transport code that can simulate a depth-averaged advection-diffusion process in an aquatic environment. Within this framework, RMA4 could be used to model constituent concentrations that are conservative or that have dynamics that can be described by a first-order decay process.

Future efforts to develop a statistical model that relates or predicts bacteria concentrations at beaches from environmental factors may benefit from these hydrodynamic simulations. To the extent that flow paths are associated with source areas, it may be helpful to develop separate statistical equations within defined sets of contiguous wind sectors. As in this report, the wind sector for a specific bacteria sampling event may be identified by use of wind data from the LSCM4 weather station for the 24-hour period preceding sample collection. As a starting point for model development at Memorial Beach, four distinct equations may be appropriate for winds for sectors 1–3, 20–24, 4–14, 15, and 16–19. At Metropolitan Beach, three distinct equations may be appropriate for sectors 1–8, 19–24, 9–11, and 12–18. Final determination of the number of equations and partitioning of sectors would be based on statistical considerations.

Additional factors may influence both beaches and could be considered in future studies. Bathers at the beach are a source of *E. coli* bacteria, both to beach sands and to the water itself (Calderon and others, 1991). Warmer temperatures likely result in an increase in the number of bathers, which may increase water turbidity. No data are available on the number of bathers associated with various samples. Likewise, there is no information on the numbers of birds on the beach or on aggravating factors such as algal mats along the shore. These factors have been implicated at other beaches as potential sources of *E. coli* (Whitman and Nevers, 2003; Whitman and others, 2003; Olapade and others, 2006). Beach-grooming practices may influence beach-water bacteria concentrations (Kinzelman and others, 2004). Francy and others (2006) used microbial source-tracking techniques and modeling to show the influence of temperature, turbidity, and fecal sources on selected Lake Erie bathing beaches. Any of these factors may account for occasional high concentrations of *E. coli* bacteria, and may be correlated with factors that were significant in regression models.

Summary and Conclusions

In cooperation with the Michigan Department of Environmental Quality, the USGS analyzed environmental data and circulation patterns on Lake St. Clair, which is part of the connecting channel between Lake Huron and Lake Erie in the Great Lakes Basin. The analysis identified factors associated with *E. coli* concentrations at Memorial and Metropolitan Beaches, in Macomb County, Mich. Data from 2002–2005 provided the basis for development of linear regression models, regression trees, and logistic regressions to help identify environmental factors that were statistically associated with *E. coli* concentrations. A hydrodynamic and particle-tracking model was used to identify flow paths in Lake St. Clair that may help identify potential source areas of *E. coli* to the study beaches. This study has demonstrated that circulation patterns and associated flow paths to Memorial and Metropolitan Beaches on Lake St. Clair are transient phenomena that are sensitive to variable wind conditions and uncertain dispersion characteristics. This sensitivity and uncertainty demonstrates the utility of hydrodynamic simulation for identifying flow paths, travel and exposure times, and source areas of bacteria to beaches. The hydrodynamic and statistical analyses can be used in combination by regulatory agencies for the development of Total Maximum Daily Loads (TMDLs) by helping to identify source areas and conditions that may influence exceedances of the daily Michigan recreational water-quality standard at these beaches.

Exploratory analysis indicated that *E. coli* samples obtained by MDEQ in 2003 and 2004 tended to have lower concentrations than samples obtained by MCHD during the same period. Differences in exact sampling locations and procedures, and in analytical techniques between agencies may have contributed to the differences between concentration characteristics. MCHD data were used in this report because they had been collected over a longer period and at more frequent intervals than the MDEQ data.

The statistical distributions of *E. coli* concentrations at both Memorial and Metropolitan Beaches are positively skewed, which is manifested by numerous low and medium concentrations interspersed with infrequent high values. For consistency with assumptions underlying statistical models used to associate environmental factors with concentration data, a $\log_{10}$ transformation was applied to the *E. coli* data before analysis. The distribution of the log-transformed *E. coli* data closely approximated a normal distribution.

Wind is a primary determinant of circulation patterns on Lake St. Clair, and may be associated with differences in *E. coli* concentrations because of changes in flow paths and potential source areas. To assess this potential effect, *E. coli* concentration data were grouped into eight 45-degree wind-direction categories. Results of statistical analyses indicated no differences in median *E. coli* concentrations among wind-direction categories at Memorial Beach, whereas significant differences were detected from data at Metropolitan Beach. In particular, the median *E. coli* concentration in the east-northeast (ENE.) sector is significantly greater than the median concentrations in the west-southwest (WSW.), west-northwest (WNW.), and north-northwest (NNW.) sectors. In addition, the median concentration in the south-southwest (SSW.) sector is significantly greater than median concentrations in the WNW. and NNW. sectors.

Results from linear regression analyses indicated that the amount of rainfall prior to *E. coli* sampling, the temperature of the water, and the turbidity of the water were positively associated with $\log_{10}$ *E. coli* concentrations at both beaches. Flow from Clinton River also was positively associated with concentrations at Memorial Beach and the hourly change in water level preceding sampling was negatively associated with concentrations at Metropolitan Beach. Some of the first-order interactions among “main-effects” model variables were marginally significant at Memorial Beach in the “interaction-effects” model. In addition, $\log_{10}$ *E. coli* concentrations measured 2 days before or after concentrations at Memorial Beach were statistically significant when added to the main-effects model. This result indicates that, for Memorial Beach, *E. coli* concentrations in conjunction with other environmental variables may help forecast *E. coli* concentrations at this beach. The fact that postsampling *E. coli* concentrations were also significant in the model suggests that complex processes may transport bacteria back and forth from the beach area. A north wind direction was identified as a significant factor for estimating $\log_{10}$ *E. coli* concentrations at Metropolitan Beach when added to the main-effects model, although wind speed was not a significant factor. Linear-regression models developed for Memorial and Metropolitan Beaches explained only about 30 percent of the variability in $\log_{10}$ *E. coli* concentrations. From this result, it is evident that several factors or a combination of events that were not accounted for in the models contributed to bacteria concentrations at these beaches.

Regarding the regression-tree models developed to estimate $\log_{10}$ *E. coli* concentrations from data collected at Memorial and Metropolitan Beaches, the apparent estimation error generally decreased with the number of terminal nodes in these models. A statistical resampling analysis indicated, however, that the apparent estimation error was not consistent with the prediction error, which tended to increase slightly for models with more than two terminal nodes. Models with two terminal nodes had little explanatory capability. These results are interpreted as indicating that too few observations of *E. coli* concentrations were available for developing regression-tree models in this study. Regression-tree analysis is commonly used to define variables associated with high *E. coli* concentrations. The results of this study show some limitations of this analysis and the need for robust datasets for the model to be useful.

Results from the logistic regression models, which were developed to estimate the probability that *E. coli* concentrations would exceed 300 MPN/100 mL at Memorial and Metropolitan Beaches, showed that rainfall preceding the sampling events was positively associated with the probabilities of bacteria exceedances at both beaches. Flow in the Clinton River was positively associated with bacteria exceedances at Memorial Beach but negatively associated with bacteria exceedances at Metropolitan Beach. For Memorial Beach, turbidity and $\log_{10}$ *E. coli* concentrations measured two days before or after concentrations at Memorial Beach were positively associated with bacteria exceedances. The logistic models were effective in estimating bacteria exceedances at both beaches. A receiver operation characteristics (ROC) analysis identified cut points for maximizing the true positive prediction rate while minimizing the false positive rate.

The hydrodynamic model of Lake St. Clair was developed from an existing 2D hydrodynamic model of the St. Clair-Detroit River Waterway. Steady-state simulations were used to depict velocity fields on Lake St. Clair as a function of wind direction. Reverse particle tracking was used with these fields to describe associated flow paths from the beaches towards source areas. Sets of contiguous wind sectors were associated with similar flow paths. Transient simulations were used to describe changing velocity fields on Lake St. Clair during periods preceding 10 selected bacteria sampling events in 2004. Reverse particle tracking was used with these velocity fields to identify expected flow paths of bacteria from beaches for specific *E. coli* sampling events.

The main focus of this study was to determine what factors may influence or explain *E. coli* concentrations at the study beaches. Some significant explanatory variables from the linear regression analysis for the two beaches were different, but the common significant explanatory variables for both beaches were rainfall, water temperature, and turbidity. Several other studies have documented similar factors affecting beaches (Nevers and Whitman, 2005; Francy and others, 2006). The study of beach watersheds is complex and needs to be associated with numerous water bodies (rivers and creeks) as well as an extensive and complex drainage system, such as outfalls that drain to Lake St. Clair directly or indirectly through the Clinton River. Rainfall as a significant variable may indicate that storm runoff or erosion is influencing the *E. coli* concentrations in nearshore beach water. Discharge was an explanatory variable for Memorial Beach; rainfall can increase stream discharge, which further corroborates this hypothesis. Turbidity also was a commonly included explanatory variable during model development for the study beaches; turbidity itself is influenced by rainfall, runoff, and high waves, and turbidity has historically been associated with increased *E. coli* concentrations. Water temperature may be an indication of seasonality; the highest *E. coli* concentrations were in August, with the highest water temperatures in July and August. Water temperature may also serve as a surrogate indicator for variables not addressed in the data set (such as algal accumulations, beach visitors, or waterfowl).

At Memorial Beach, little systematic variation was observed in bacteria concentrations with respect to wind direction or magnitude. At least one exceedance of the 300 MPN/100 mL criterion was observed for each 45^{o} wind sector, and the median concentration for each wind sector was not statistically different. Hydrodynamic models indicated that, under most wind conditions, particles sampled at Memorial Beach would generally have traveled along the shoreline to the north and south of the beach or recirculated within L'anse Creuse Bay within the preceding 24 to 48 hours.

Linear regression models indicated that rainfall in the preceding 72 hours, water temperature, turbidity, and discharge of the Clinton River were positively correlated with bacteria concentrations at Memorial Beach. Logistic regression models indicated that an increase of 1 in. of rain over the preceding 72-hour period more than doubles the probability that *E. coli* concentrations will exceed the 300 MPN/100 mL standard. *E. coli* concentrations 2 days before and 2 days after a sampling event were correlated at Memorial Beach. The limited spatial extent of particle tracks over 48- to 72-hour timeframes—and the several simulations indicating particle movement in first one direction, then in the opposite direction along the shoreline—may explain correlation between *E. coli* concentrations on successive dates at Memorial Beach. The tendency for water to move in different directions along the shoreline near Memorial Beach may result in a relatively constant and perhaps well-mixed source, regardless of wind direction.

Significant explanatory variables in the regressions may indicate a similar process influencing Memorial Beach. Clinton River discharge is likely a surrogate variable for the discharge from the drains and creeks in the area, and from the Clinton River Spillway after rainfall. Water draining from these sources after rainfall carries a relatively high load of suspended material. MCHD sampling data have shown a statistically significant relation between turbidity and *E. coli* concentrations nearshore to some of the drains and to the mouth of the Clinton River.

In contrast to Memorial Beach, bacteria concentrations at Metropolitan Beach varied with respect to wind direction. Exceedances of the 300 MPN/100 mL standard were noticeably absent when wind was from the WNW. or NNW. Samplings during times when winds were from the SSW. and ESE. accounted for 7 of 14 exceedances. The median *E. coli* concentration was greatest when winds were from the ENE. Steady-state hydrodynamic simulations indicated that when winds were from the WNW. or NNW. (sector 1), particles generally originated near the mouth of Clinton River within 24 hours of when they were sampled at the beach. In contrast, when winds were from the ESE., particles originated from along the shoreline to the west or north of Memorial Beach within 24 hours of when they were sampled. When winds were from the northeast, a recirculation pattern within L'Anse Creuse Bay developed also might include particles from the western shoreline of Lake St. Clair.

Linear regression models indicated that rainfall in the preceding 24 hours, water temperature and turbidity, but *not* Clinton River discharge, were positively correlated with bacteria concentrations at Metropolitan Beach. In addition, and in contrast to Memorial Beach, linear regression models indicated change in water level was inversely correlated with bacteria concentrations at Metropolitan Beach. Logistic regression models indicated that an increase of 1 in. of rain over the preceding 24-hour period increases the probability that *E. coli* concentrations will exceed the 300 MPN/100 mL standard more than 30 times. In contrast, an increase in discharge of the Clinton River was associated with a *decreased* probability that *E. coli* concentrations would exceed the standard. *E. coli* concentrations 2 days before and 2 days after a sampling event were not correlated at Metropolitan Beach. Metropolitan Beach appears to be subject to short-term, transient events of poor water quality that are not necessarily associated with the Clinton River as a source. Spikes in *E. coli* concentrations occurred at the beach without always being associated with rainfall. These may be occasions when wind directions are appropriate for entrainment of bacteria trapped within beach sands or when contaminants may be drawn from sources to the west and southwest of the beach. The hydrodynamic simulations were based on a 2D model and depict the movement of massless particles. As noted previously, these models may not capture the full complexity of water circulation patterns or of bacterial transport in the nearshore. A 3D model might show additional complexities or water-shoreline interactions.

The MCHD found that *E. coli* concentrations in foreshore sand at Metropolitan Beach were correlated with *E. coli* concentrations in water near the beach during sampling in 2004 (Buzonik and Shoemaker, 2004). No such correlation was reported for Memorial Beach. Winds from the south might force water onto the shoreline at Metropolitan Beach, perhaps entraining some of the bacteria from the foreshore sand. It is possible that this source might account for the association between winds from the southwest to southeast and exceedances of *E. coli* criteria. In addition, the median *E. coli* concentration was highest at Metropolitan Beach for winds from the ENE., which were associated with a recirculation pattern in the bay. This recirculation might entrain bacteria from some of the same sources that influence Memorial Beach. Even though most particle tracks in the hydrodynamic models eventually move from Metropolitan Beach to the Clinton River within 24 to 48 hours, it appears that circulation patterns bringing particles from the shoreline to the west (or perhaps the sands themselves) may have greater influence on bacteria concentrations.

Acknowledgments

The authors gratefully acknowledge the assistance of Mr. James Buzonic, Manager, Macomb County Health Department, Environmental Health Services Division, for providing access to the division’s *E. coli* and supporting data. Donna S. Francy, USGS Ohio Water Science Center (WSC), Norman G. Grannemann, USGS Michigan WSC, and David L. Lorenz, USGS Minnesota WSC, provided technical reviews. Michael Eberle, USGS Columbus Publishing Service Center, provided editorial review. Dorothy H. Tepper, USGS Northeastern Region, recommended that the report be approved for publication by the Director of the USGS.

References Cited

Alm, E.W., Burke, Janice, and Spain, Anne, 2003, Fecal indicator bacteria are abundant in wet sand at freshwater beaches: Water Research, v. 37, no. 16, p. 3978–3982.

Baird, W.F., & Associates, Ltd., 2005, Clinton River sediment transport modeling study: U.S. Army Corps of Engineers, Detroit District, accessed February 27, 2007, at *http://www.glc.org/tributary/models/documents/ClintonRiverModelingStudy_main1.pdf*

Bolsenga, S.J., and Herdendorf, C.E., eds., 1993, Lake Erie and Lake St. Clair handbook: Detroit, Mich., Wayne State University Press, 469 p.

Buzonik, Jim, and Shoemaker, Cole, 2004, 2004 Lake St. Clair water quality assessment: Macomb County Health Department, Environmental Health Services Division, accessed October 9, 2007, at *http://www.macombcountymi.gov/publichealth/EH/LSCA%20Report/LSCA2004Report.htm*

Calderon, R.L, Mood, E.W., and Dufour, A.P., 1991, Health effects of swimmers and nonpoint sources of contaminated water: International Journal of Environmental Health Research, v. 1, p. 21–31.

Conover, W.J., 1980, Practical nonparametric statistics (2d ed.): New York, John Wiley & Sons, 493 p.

Davies-Colley, R.J., Bell, R.G., and Donnison, A.M., 1994, Sunlight inactivation of enterococci and fecal coliforms in sewage effluent diluted in seawater: Applied and Environmental Microbiology, v. 60, no. 6, p. 2049–2058.

Donnell, B.P., Letter, J.V., McAnally, W. H., and others, 2005, Users guide for RMA2, v. 4.5: [Vicksburg, Miss.] U.S. Army Engineer Research and Development Center, Waterways Experiment Station, Coastal and Hydraulics Laboratory, accessed October 9, 2007, at *http://chl.erdc.usace.army.mil/Media/3/2/7/RMA2_v45_Users_Guide_01-20-2006.pdf*

Draper, N.R., and Smith, Harry, 1998, Applied regression analysis (3d ed.): New York, John Wiley, Wiley Series in Probability and Statistics, 706 p.

Eckner, K.F., 1998, Comparison of membrane filtration and multiple-tube fermentation by the Colilert and Enterolert methods for detection of waterborne coliform bacteria, *Escherichia coli*, and *enterococci* used in drinking and bathing water quality monitoring in southern Sweden: Applied and Environmental Microbiology, v. 64, no. 8, p. 3079–3083.

Environmental Modeling Systems, Inc., 2005, Surface-water Modeling System, v. 9.0: South Jordan, Utah.

Francy, D.S., Bertke, E.E., Finnegan, D.P., Kephart, C.M., Sheets, R.A., Rhoades, John, and Stumpe, Lester, 2006, Use of spatial sampling and microbial source-tracking tools for understanding fecal contamination at two Lake Erie beaches: U.S. Geological Survey Scientific Investigations Report 2006–5298, 29 p.

Francy, D.S., Darner, R.A. and Bertke, E.E., 2006, Models for predicting recreational water quality at Lake Erie beaches: U.S. Geological Survey Scientific Investigations Report 2006–5192, 13 p.

Great Lakes Information Network, 2006, Lake St. Clair facts and figures: accessed July 14, 2006, at *http://www.great-lakes.net/lakes/ref/stclfact.html*

Haack, S.K., Fogarty, L.R., and Wright, Christopher, 2003, *Escherichia coli* and enterococci at beaches in the Grand Traverse Bay, Lake Michigan—Sources, characteristics, and environmental pathways: Environmental Science & Technology, v. 37, no. 15, p. 3275–3282.

Holtschlag, D.J., and Koschik, J.A., 2002, A two-dimensional hydrodynamic model of the St. Clair-Detroit River Waterway in the Great Lakes Basin: U.S. Geological Survey Water-Resources Investigations Report 01–4236, 63 p.

Holtschlag, D.J., and Koschik, J.A., 2003a, An acoustic Doppler current profiler survey of flow velocities in St. Clair River, a connecting channel of the Great Lakes: U.S. Geological Survey Open-File Report 03–119 [online only], accessed October 9, 2007, at *http://mi.water.usgs.gov/pubs/OF/OF03-119/index.php*

Holtschlag, D.J., and Koschik, J.A., 2003b, An acoustic Doppler current profiler survey of flow velocities in Detroit River, a connecting channel of the Great Lakes: U.S. Geological Survey Open-File Report 03–219 [online only], accessed October 9, 2007, at *http://mi.water.usgs.gov/pubs/OF/OF03-219/index.php*

Holtschlag, D.J., and Koschik, J.A., 2004, Hydrodynamic simulation and particle-tracking techniques for identification of source areas to public-water intakes on the St. Clair-Detroit River Waterway in the Great Lakes Basin: U.S. Geological Survey Scientific Investigations Report 2004–5072, 29 p.

Kinzelman, J.L., Pond, K.R., Longmaid, K.D., and Bagley, R.C., 2004, The effect of two mechanical beach grooming strategies on *Escherichia coli* density in beach sand at a southwestern Lake Michigan beach: Aquatic Ecosystem Health & Management, v. 7, no. 3, p. 425–432.

Letter, J.V., Donnell, B.P., and others, 2005, Users guide to RMA4, v. 4.5: U.S. Army Engineer Research and Development Center, Waterways Experiment Station, Coastal and Hydraulics Laboratory, accessed April 14, 2008, at *http://chl.erdc.usace.army.mil/Media/3/2/9/RMA4_Users_Guide_07-08-2005.pdf*

Macomb County Health Department, 2006, Developing a bathing beach monitoring program: Environmental Health Services Division, Recreational Water Quality Monitoring Project, 18 p, accessed August 3, 2006, at *http://www.deq.state.mi.us/documents/deq-swq-beach-beach.doc*

Macomb County Health Department, 2007, *E. coli*, sampling procedures and total body contact standards: accessed March 7, 2007, at *http://macombcountymi.gov/publichealth/EH/beachmethodology.htm*

McLellan, S.L., and Salmore, A.K., 2003, Evidence for localized bacterial loading as the cause of chronic beach closings in a freshwater marina: Water Research, v. 37, no. 11, p. 2700–2708.

Nevers M.B., and Whitman, R.L., 2005, Nowcast modeling of *Escherichia coli* concentrations at multiple urban beaches of southern Lake Michigan: Water Research, v. 39, no. 20, p. 5250–5260.

Olapade, O.A., Depas, M.M., Jensen, E.T., and McLellan, S.L., 2006, Microbial communities and fecal indicator bacteria associated with *Cladophora* mats on beach sites along Lake Michigan shores: Applied and Environmental Microbiology, v. 72, no. 3, p. 1932–1938.

Olyphant, G.A., and Whitman, R.L., 2004, Elements of a predictive model for determining beach closures on a real time basis—The case of 63rd Street Beach Chicago: Environmental Monitoring and Assessment, v. 98, p. 175–190.

Poeter, E.P., Hill, M.C., Banta, E.R., Mehl, Steffen, and Christensen, Steen, 2005, UCODE_2005 and six other computer codes for universal sensitivity analysis, calibration, and uncertainty evaluation: U.S. Geological Survey Techniques and Methods 6–A11, 283 p.

Tape, T.G., [n.d.], The area under an ROC curve: University of Nebraska Medical Center: accessed December 2006 at *http://gim.unmc.edu/dxtests/roc3.htm*

Tecplot, Inc., 2006, Tecplot 360 user's manual: Bellevue, Wash., 755 p.

U.S. Environmental Protection Agency, 1999, Review of potential modeling tools and approaches to support the BEACH program: Washington, D.C., Office of Science and Technology, 823–R–99–002 [variously paged].

U.S. Environmental Protection Agency, 2006, Method 1603—*Escherichia coli* (*E. coli*) in water by membrane filtration using modified membrane-thermotolerant *Escherichia coli* agar (Modified mTEC): accessed April 14, 2008, at *http://epa.gov/waterscience/methods/method/biological/1603.pdf*

Whitman, R.L., and Nevers, M.B., 2003, Foreshore sand as a source of *Escherichia coli* in nearshore water of a Lake Michigan beach: Applied and Environmental Microbiology, v. 69, no. 9, p. 5555–5562.

Whitman, R.L., Shively, D.A., Pawlik, Heather, Nevers, M.B., and Byappanahalli, M.N., 2003, Occurrence of *Escherichia coli* and enterococci in *Cladophora* (Chlorophyta) in near-shore water and beach sand of Lake Michigan: Applied and Environmental Microbiology, v. 69, no. 8, p. 4714–4719.

This page is intentionally blank.

Appendix 1. Equations and Detail Used in the Linear and Logistic Regression Models

Linear Regression

Linear regression provides a statistical model for estimating the continuous variation of *E. coli* concentrations as a function of environmental factors (explanatory variables). The general form of a linear regression model is

$$y = X \cdot \beta + \varepsilon \tag{1}$$

where

y is a column *n*-vector of response variables,

X is a *n* by (*p*+1*)* design matrix that is composed of *p n*-vectors, x_1,x_2,. . .,x_p, of explanatory variables augmented with a leading column of ones,

β is a *p*-vector of coefficients, $\beta_0, \beta_1, \ldots, \beta_p$ with ordinary least-squares estimator $\beta_{ols} = (X' \cdot X)^{-1} \cdot X' \cdot y$,

and

ε is an *n*-vector of residuals that is assumed to be normally distributed and independent with mean zero and constant variance σ^2, which is commonly written $\varepsilon \sim NI(0, \sigma^2)$. In addition, the regression model is based on the assumption that the covariance between ε and X, $Cov(\varepsilon, X)$, equals zero. Finally, the design matrix *X* must be of full rank (*p*+*1*) < *n*, so that $\left(X' \cdot X\right)^{-1}$ exists. If these conditions are met, the predicted values, $\hat{y} = X \cdot \beta_{ols}$, are the unbiased, minimum variance linear estimates of *y*.

The multiple coefficient of determination, R^2, describes the fraction of the variability of *y* described by $\hat{y}$, where the adjusted value is computed as $R_a^2 = 1 - [SSE \cdot (n-p)] \,/\, [SST \cdot (n-1)]$.

Like the response estimates, estimated parameters β_{ols} associated with the individual explanatory variables are uncertain. As *n* becomes large, estimated parameters are assumed to be unbiased and normally distributed about their true values with covariance $Cov(\beta_{ols}) = \sigma^2 (X' \cdot X)^{-1}$, which is commonly written $\beta_{ols} \sim N_{p+1}(\beta, Cov(\beta_{ols}))$. Diagonal elements of the covariance matrix describe the variance of the corresponding estimated parameters. Therefore, a *t* statistic can be computed to assess the significance of individual parameters as $t = \beta_{ols,i} \,/\, \sqrt{Cov(\beta_{ols})_{i,i}}$. On the basis of the *t* probability distribution with *n*-*p*-1 degrees of freedom, a *p*-value is computed with the *t* statistic to assess the likelihood that the null hypothesis of the model parameter being equal to zero is true. Again, a small *p*-value is used to reject the null hypothesis, thereby accepting that the parameter estimate $\beta_{ols,i}$ is statistically different from zero.

Logistic Regression

As an alternative to estimating *E. coli* concentrations and comparing the estimates to thresholds of interest, one can directly estimate the probability that a specified threshold will be exceeded using classification methods. In this report, linear logistic models were developed to estimate probabilities of exceedance. In accordance with Michigan's Water Quality Standard, a threshold of 300 CFU/100 mL was used.

Application of the logistic model involves a transformation of *E. coli* concentrations into two classes. In this report, binary logistic models were developed to estimate the probability of *E. coli* concentrations exceeding 300 CFU/100 mL by coding response variables y_i = 1 for samples exceeding this threshold and $y_i = 0$ otherwise. Then, the probability *p* that *y* equals 1 given a set of explanatory variables ***x*** can be written *p* = *Pr(y*=1|***x***).

The linear logistic model relates the logit transformation of probabilities *p* linearly to the explanatory variables. The logit transformation is the log of the odds, which is the ratio of the probability that an event will occur divided by the probability that it will not occur as

$$\mathrm{lo\hat{g}it}(\hat{p}) \equiv \log\left[\frac{\hat{p}}{1-\hat{p}}\right] = \alpha + \beta' x \quad .$$

As in linear regression, α is the intercept parameter and β' is a vector of parameters associated with explanatory *(x)* variables. Parameters are estimated by use of maximum likelihood. The inverse logit transformation is used to compute estimated probabilities as

$$\hat{p} = \left[\frac{\mathrm{e}^{\mathrm{lo\hat{g}it}(p)}}{1 + e^{\mathrm{lo\hat{g}it}(p)}}\right]$$

Parameters in the logistic regression are estimated by the minimizing -2 times the log likelihood (-2 Log L) function, which is computed as

$$-2LogL = -2\sum_i y_i \ln(\hat{p}_i) + (1 - y_i)\ln(1 - \hat{p}_i)$$

The difference between the -2 Log L function minimized under the intercept-only model $-2LogL_\alpha$ and the linear model $-2LogL_{\alpha+\beta' x}$ is used to compute a chi-square statistic, which is then compared to the chi-squared statistical distribution to assess the overall statistical significance of the explanatory variables in the logistic equation.

Logistic models provide a measure of rank correlation for assessing the predictive ability of the models. This measure is based on the number of concordant (*nc*) and discordant (*nd*) pairs in dataset. Specifically, a pair of observations with different response values is said to be concordant (discordant) if the larger response has a lower (higher) predicted probability than the smaller response. For example, if $Pr(y_i=1|x) < Pr(y_j=0|x)$, then observations $\{y_i, y_j\}$ would be a concordant pair. If there are *Ni* measurements where *E. coli* concentrations are greater then 300 CFU/100 mL, and *Nj* measurements less than 300 in a set of size *Ni* + *Nj*, then there are $t = Ni \cdot Nj$ pairs. The selected measure of rank correlation was computed as $c = (nc + 0.5(t - nc - nd)) \,/\, t$.

The exponentiated values of the estimated parameters are commonly interpreted as the increase (or decrease) in the odds of an event associated with a unit increase in the explanatory variable, given that there are no changes in the other explanatory variables. For example, an estimated parameter value of 0.2 would have an exponentiated value of 1.2214. Thus, one would expect that a unit increase in the associated explanatory variable (by itself) would be associated with a 22.14 percent increase in the odds of an event such as *E. coli* concentration exceeding 300 CFU/100 mL. Estimated parameters less than zero are associated with a decrease in the odds. The estimated standard deviation of the parameter can be used in a similar way to compute an approximate confidence interval about the increase in the odds.

Holtschlag and others—***Escherichia coli* Concentrations at Two Beaches on Lake St. Clair, Michigan, 2002–2005**—Scientific Investigations Report 2008–5028