USGS
science for a changing world

In cooperation with the Erie County Health Department

Modeling to Predict *Escherichia coli* at Presque Isle Beach 2, City of Erie, Erie County, Pennsylvania

Scientific Investigations Report 2008-5039

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

Modeling to Predict *Escherichia coli* at Presque Isle Beach 2, City of Erie, Erie County, Pennsylvania

By Tammy M. Zimmerman

In cooperation with the Erie County Health Department

Scientific Investigations Report 2008-5039

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

Suggested citation:
Zimmerman, T.M., 2008, Modeling to predict *Escherichia coli* at Presque Isle Beach 2, City of Erie, Erie County, Pennsylvania: U.S. Geological Survey Scientific Investigations Report 2008-5039, 13 p.

Contents

Figures

Tables

Conversion Factors

Multiply	By	To obtain
inch (in.)	25.4	millimeter (mm)
foot (ft)	0.3048	meter (m)
mile (mi)	1.609	kilometer (km)
cubic inch (in^3)	1.639	milliliter (mL)

Concentrations of bacteria in water are reported in colonies per 100 milliliters (col/100 mL)

Modeling to Predict *Escherichia coli* at Presque Isle Beach 2, City of Erie, Erie County, Pennsylvania

By Tammy M. Zimmerman

Abstract

The Lake Erie beaches in Pennsylvania are a valuable recreational resource for Erie County. Concentrations of *Escherichia coli* (*E. coli*) at monitored beaches in Presque Isle State Park in Erie, Pa., occasionally exceed the single-sample bathing-water standard of 235 colonies per 100 milliliters resulting in potentially unsafe swimming conditions and prompting beach managers to post public advisories or to close beaches to recreation.

To supplement the current method for assessing recreational water quality (*E. coli* concentrations from the previous day), a predictive regression model for *E. coli* concentrations at Presque Isle Beach 2 was developed from data collected during the 2004 and 2005 recreational seasons. Model output included predicted *E. coli* concentrations and exceedance probabilities—the probability that *E. coli* concentrations would exceed the standard. For this study, *E. coli* concentrations and other water-quality and environmental data were collected during the 2006 recreational season at Presque Isle Beach 2. The data from 2006, an independent year, were used to test (validate) the 2004–2005 predictive regression model and compare the model performance to the current method. Using 2006 data, the 2004–2005 model yielded more correct responses and better predicted exceedances of the standard than the use of *E. coli* concentrations from the previous day. The differences were not pronounced, however, and more data are needed. For example, the model correctly predicted exceedances of the standard 11 percent of the time (1 out of 9 exceedances that occurred in 2006) whereas using the *E. coli* concentrations from the previous day did not result in any correctly predicted exceedances.

After validation, new models were developed by adding the 2006 data to the 2004–2005 dataset and by analyzing the data in 2- and 3-year combinations. Results showed that excluding the 2004 data (using 2005 and 2006 data only) yielded the best model. Explanatory variables in the 2005–2006 model were $\log_{10}$ turbidity, bird count, and wave height. The 2005–2006 model correctly predicted when the standard would not be exceeded (specificity) with a response of 95.2 percent (178 out of 187 nonexceedances) and correctly predicted when the standard would be exceeded (sensitivity) with a response of 64.3 percent (9 out of 14 exceedances). In all cases, the results from predictive modeling produced higher percentages of correct predictions than using *E. coli* concentrations from the previous day. Additional data collected each year can be used to test and possibly improve the model. The results of this study will aid beach managers in more rapidly determining when waters are not safe for recreational use and, subsequently, when to close a beach or post an advisory.

Introduction

Approximately 7 mi of Lake Erie shoreline lies within Presque Isle State Park in Erie, Pa. Monitored public bathing beaches at Presque Isle make up less than 1 mi (Natural Resources Defense Council, 2007). Previous studies conducted by the Erie County Health Department (ECHD) and other agencies (Erie County Health Department, 2007) identified sources of fecal contamination at the Presque Isle beaches that included stormwater runoff and combined sewer overflows. These same types of sources likely continue to influence concentrations of fecal-indicator bacteria at Presque Isle beaches, especially during periods of heavy rainfall. According to the Natural Resources Defense Council (2007), the number of beach advisories or closings for permitted public bathing beaches at Presque Isle rose from 39 days in 2005 to 53 days in 2006. Nine of the 12 public bathing beaches that are permitted at Presque Isle State Park had at least one advisory or closing in 2006 (U.S. Environmental Protection Agency, 2007). Most of the closings or advisories in 2006 (51 of 53 days) were due to elevated bacteria concentrations; a small percentage of the advisories or closings (2 of 53 days) was related to a rainfall advisory system (Natural Resources Defense Council, 2007).

Beach advisories or closings are issued by beach managers on the basis of recreational water-quality standards for fecal-indicator bacteria. In 2004, Pennsylvania adopted bacteriological criteria recommended by the U.S. Environmental Protection Agency that use *Escherichia coli* (*E. coli)* as the indicator to assess the recreational water quality of all public bathing beaches in the Commonwealth (U.S. Environmental Protection Agency, 1986). The regulations estab-

lish the single-sample bathing-water standard for *E. coli* as 235 col/100 mL and also state that the geometric mean for *E. coli* concentrations in multiple samples collected during any 30-day period during the recreational season should not exceed 126 col/100 mL (Pennsylvania Code, 28 PaCode § 18.28). Concentrations of *E. coli* in recreational waters may change substantially from the time of sample collection until the sample is processed and analyzed and the results are reported (18-24 hours). This presents a delay in posting closings or issuing advisories for Pennsylvania beaches because the current practice is to use the *E. coli* concentrations from the previous day to determine if an advisory should be posted or a beach should be closed. Such a delay may cause recreational users of the beaches to be exposed to potential health risks when the risks are the greatest (bacteria concentrations are high and exceed the standard). Conversely, a delay may prohibit access to recreational areas or prohibit swimming when the health risk is reduced or no longer present (bacteria concentrations have declined from the value on the previous day and are below the standard by the time the beach is closed or advisories are posted). Because of the drawbacks of using the *E. coli* concentrations from the previous day as the only method to assess the recreational quality of Presque Isle beaches and the fact that concentrations of bacteria in beach waters continue to frequently exceed the single-sample bathing-water standard of 235 col/100 mL, other methods of assessing the water quality of the beaches are being initiated to reduce the potential health risks to recreational users of the beaches.

Predictive modeling can supplement the use of measured *E. coli* concentrations for indicating if fecal contamination has occurred at Presque Isle beaches. Environmental and water-quality variables that are easily measured can be used in regression models to predict, or forecast, when concentrations of fecal-indicator bacteria will exceed recreational standards. This regression approach to modeling is appropriate for Presque Isle beaches where dominant sources of bacterial contamination are nonpoint or unidentified. More complicated models (that incorporate hydrodynamic mixing and transport) are appropriate in cases where bacterial contamination is dominated by point sources. Examples of selected predictive-modeling studies that have used the above-mentioned approaches to determine exceedances of recreational water-quality standards are referenced in Francy and Darner (2006).

The U.S. Geological Survey (USGS), in cooperation with the ECHD, validated and refined beach-specific predictive regression models at Presque Isle Beach 2 near Erie, Pa. Data initially were collected during two recreational seasons (2004 and 2005), and a model was developed (Zimmerman, 2006). This report presents the results of testing (validating) the 2004–2005 regression model for *E. coli* using data collected during 2006. In addition, this report presents the results of new models developed by adding the 2006 data to the 2004–2005 dataset. The new models consisted of combinations of data (2004 and 2006 data combined; 2005 and 2006 data combined) in order to determine the "best" model for Presque Isle Beach 2. Finally, this report describes the results of comparing the performance of predictive models to the current method of assessing recreational water quality (using *E. coli* concentrations from the previous day) for determining beach closings/advisories.

Methods

This study was done at Presque Isle Beach 2 in Erie, Pa. (fig. 1). Data were collected during the summer of 2006 recreational swimming season. The following sections describe the methodology of data-collection efforts, quality control, model validation, data analysis, and model refinement.

Data Collection

Data were collected during the 2006 summer recreational season (May through September) at Presque Isle Beach 2. Sampling each week involved the collection of water samples daily from Sunday through Wednesday at two locations in the swimming area. Water samples were collected using grab-sample techniques described in Myers and others (2007). To maintain sterile conditions, grab samples were collected in sterile 125-mL polypropylene bottles in at least 3 ft of water at approximately 12 in. below the water surface. Water samples were kept on ice prior to processing and were analyzed for *E. coli* bacteria using modified mTEC membrane-filtration techniques (U.S. Environmental Protection Agency, 2002). Water samples were processed by ECHD staff in their laboratory within 6 hours of sample collection.

Ancillary data for explanatory variables used in the regression equations were collected at the time of water sampling or compiled from other sources as described in Zimmerman (2006). Data that were collected at the time of sampling include water temperature, specific conductance, pH, dissolved oxygen, and turbidity. Other data collected at the time of sampling included number of birds (bird count) on the beach and wave heights that were categorized into one of five groups with heights ranging from 0 to 2 ft, 1 to 3 ft, 2 to 4 ft, 3 to 5 ft, and 4 to 6 ft. Data compiled from other sources included streamflow data from the nearest USGS streamflow-gaging station, Brandy Run near Girard, Pa. (04213075), to determine instantaneous streamflow to the nearest 30 minutes of when water samples were collected (Siwicki, 2005, 2006; U.S. Geological Survey, 2007). Brandy Run is a tributary to Elk Creek, a creek that discharges to Lake Erie approximately 15 mi from the Presque Isle Beaches. Speed and direction of water currents (model-derived for the Presque Isle Beach 2 location) were obtained from the National Oceanic and Atmospheric Administration (NOAA) Great Lakes Coastal Forecasting System (Greg Lang, National Oceanic and Atmospheric Administration, written commun., 2005, 2006). Wind speed, wind direction, and rainfall data were obtained from

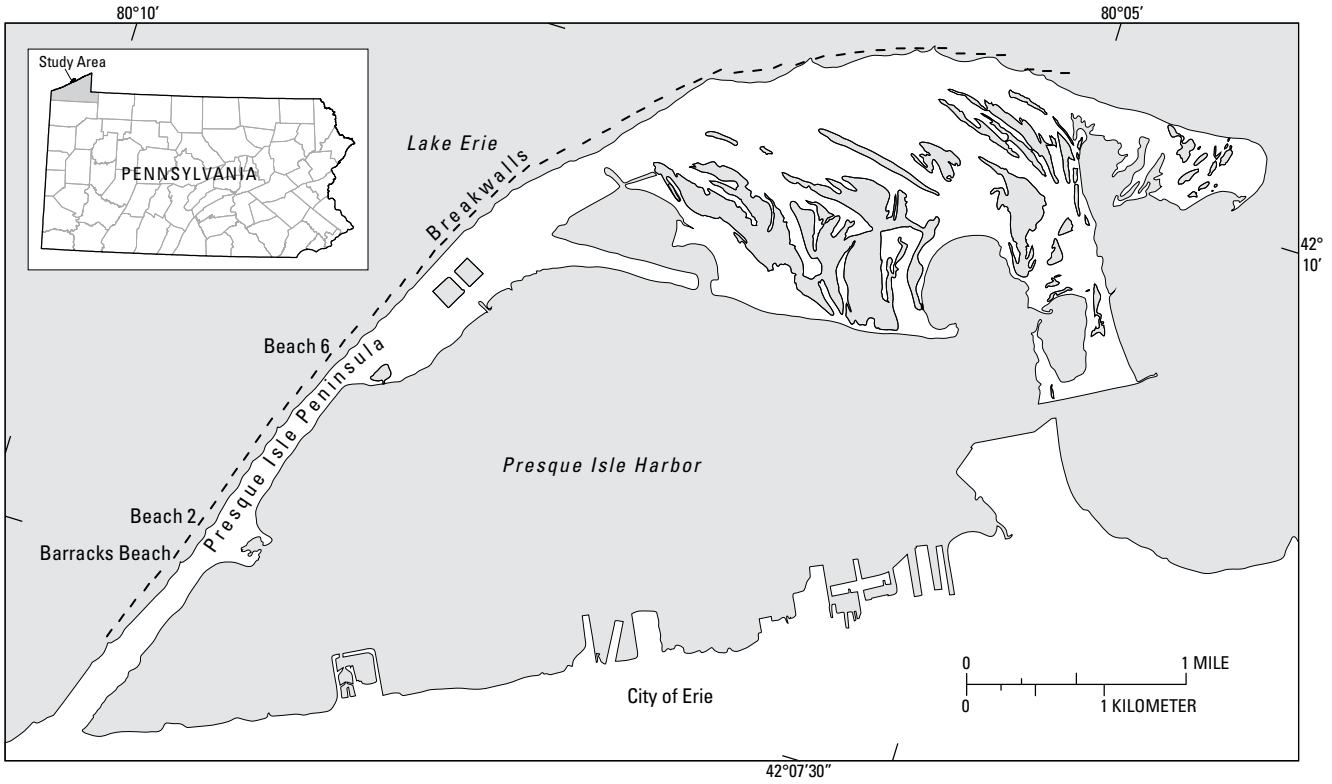

Figure 1. Location of Presque Isle beaches, City of Erie, Erie County, Pennsylvania (modified from Pennsylvania Department of Conservation and Natural Resources, 2006).

a weather station at the Erie International Airport, Erie, Pa. (Pennsylvania State Climatologist, 2007).

Quality Control

Quality-control measures were practiced in this study to ensure data quality met project objectives. These measures are the same as those described in detail in Zimmerman (2006) and include quality control in the field and quality control in the laboratory. Quality control in the field included the collection of a field blank approximately once each month and the collection of duplicate field measurements of turbidity, specific conductance, pH, dissolved oxygen, total dissolved solids, and water temperature with each sample. Measurements that did not agree within 10 percent were repeated. Quality-control measures in the laboratory included the processing of membrane-filtration equipment blanks and membrane-filtration procedure blanks with every sample. No bacteria colonies were detected in any of the equipment blanks or procedure blanks. Other quality-control measures in the laboratory included the processing of positive and negative control reference cultures.

Model Validation, Data Analysis, and Model Refinement

Model validation tests the performance of a model in predicting responses from data collected in an independent year that were not used in the development of the model. For this study, the 2004–2005 model (Zimmerman, 2006) was validated using data collected during 2006, and responses were compared to using the *E. coli* concentrations from the previous day—the current method of assessing recreational water quality. After validation testing, 2- and 3-year combinations of data were analyzed, and new models were developed to determine the "best" model for Presque Isle Beach 2.

Water-quality and ancillary data collected during 2006 were compiled and statistically analyzed in various combinations with data collected previously (2004 and 2005) using the methodology described in Zimmerman (2006). Scatterplots for continuous data and boxplots for categorical data were used as graphical screening tools to examine the relation between *E. coli* concentrations and ancillary variables measured during the study. Continuous data included water temperature, dissolved oxygen, turbidity, bird count, rain, streamflow, wind speed, and water-current speed. Categorical data were wind direction, water-current direction, and wave height. After graphical analysis of the data, statistical tests were used to determine the strength of the associations between *E. coli* concentrations and possible explanatory variables. The Spear-

man's rho statistical test was used to determine the relations between *E. coli* concentrations and continuous variables. Correlations at the 95-percent confidence interval (probabilites less than 0.05) were considered statistically significant. The Kruskal-Wallis test was used to determine the relations between *E. coli* concentrations and three or more groups of data (categorical variables). If results of the Kruskal-Wallis test showed statistically significant differences between groups of data, the Tukey-Kramer multiple-comparison test was used to determine which groups differed from each other. The groups with the highest medians were assigned a letter "A," the groups with the next highest medians were assigned a letter "B" or an "AB" combination, and so on. Any groups assigned the same letter (or combination of letters) designations were not statistically different from each other.

All variables found to be statistically related to *E. coli* concentrations were considered as potential explanatory variables in tobit regression models using 2- and 3-year combinations of data. Tobit regression was chosen because of the capability it has to handle multiple-censored data, and regression models were developed by methods documented in Zimmerman (2006) using the LIFEREG procedure using SAS statistical software (SAS Institute, 1990). Tobit regression models were evaluated using likelihood-ratio chi-square statistics and generalized R^2 values (generalized coefficient of determination). Models were chosen such that the combination and number of variables explaining the concentrations of *E. coli* in the model were significantly better, as determined using likelihood-ratio chi-square statistics, than simpler, nested models with fewer variables. Models also were selected to maximize the generalized R^2. The generalized R^2 for tobit regression models is a number between 0 and 1 that is larger when the explanatory variables are more strongly associated with the dependent variable (Allison, 1995). Evaluation criteria for the best models also included ensuring model-variable estimates were statistically significant at the 95-percent confidence level (probabilities = 0.05 or less) and graphically evaluating model fit by creating plots of the residuals from the tobit regression models.

The regression models produced two types of output. First, a predicted *E. coli* concentration was computed for each observation used in the regression. Then, the predicted values for *E. coli* were used to determine the probabilities (exceedance probabilities) that the predicted *E. coli* concentrations would exceed the *E. coli* standard of 235 col/100 mL. Using an equation modified from Allison (1995, p. 264), exceedance probabilities were calculated as 1 – Probnorm (log(235) - lp) / sep where probnorm is a SAS function (SAS Institute, 1990) that computes the probability that an observation from a standard normal distribution falls below the given value for x—in this case x = (log(235)-lp)/sep where lp is the predicted *E. coli* concentration and sep is the standard error of the prediction.

Once exceedance probabilities were computed for each observation, a threshold probability was determined. Computed exceedance probabilities above this threshold are considered to be associated with waters of poor bacteriological quality that present an unacceptable risk level for swimming. A threshold probability is selected by analyzing a scatterplot of actual *E. coli* concentrations and predicted exceedance probabilities for the corresponding predicted *E. coli* concentrations. The exceedance probability that maximizes the number of correct responses (correct exceedances and correct nonexceedances) and minimizes the number of incorrect responses (false positives and false negatives) is chosen as the threshold probability. Determinations of whether beach closings/advisories should be issued can be made if values meet or exceed the established threshold probability.

The responses of the best models using the threshold probability method and predicted *E. coli* concentrations were compared to responses of the current method that uses the *E. coli* concentrations from the previous day to assess recreational water quality. The responses of each method included specificities, sensitivities, and total correct predictions. Specificity is the proportion of nonexceedances that were predicted correctly and sensitivity is the proportion of exceedances (concentrations greater than 235 col/100 mL) that were predicted correctly.

Model Validation

The 2004-2005 model was used to predict responses during an independent period (2006) for model validation. Model responses were compared to use of *E. coli* concentrations from the previous day—the current method of assessing recreational water quality (table 1). Model output (predicted *E. coli* concentrations and predicted exceedance probabilities) produced slightly higher percentages of correct predictions than the use of *E. coli* concentrations from the previous day. The differences between model responses and responses using *E. coli* concentrations from the previous day were not pronounced. For example, the response of the model in correctly predicting when the standard would be exceeded and waters would not be safe for swimming (sensitivities) was only slightly better at 11.1 percent (1 out of 9 correct) than using *E. coli* concentrations from the previous day (zero correct responses).

Relations Between *Escherichia coli* Concentrations and Data Variables

The bacteriological quality of waters at Presque Isle Beach 2 in Erie, Pa., was generally good as determined by analyzing *E. coli* concentrations in 291 water samples collected during the 2004, 2005, and 2006 recreational seasons. *E. coli* was not detected in 52 of 291 samples (18 percent)—37 samples were reported at a detection limit of less than (<) 4 col/100 mL, and 15 samples were reported at a detection

limit of <10 col/100 mL. The maximum concentration of *E. coli* was 852 col/100 mL following a storm in July 2006. For the days sampled during this study at Presque Isle Beach 2, the single-sample bathing-water standard of 235 col/100 mL was exceeded 4 of 31 days in 2004, 5 of 58 days in 2005, and 7 of 57 days in 2006 (a total of 11 percent of the days sampled over the 3 years).

The Spearman's rho statistical test was used to determine if there were relations between *E. coli* concentrations and other continuous variables. The variables tested against *E. coli* concentrations are shown in table 2 in a summary of Spearman's rho correlations for this study.

Statistically significant correlations were found between $\log_{10}$ *E. coli* concentrations and several continuous variables in the Presque Isle Beach 2 datasets. Log_{10}turbidity, rain24, wind speed, and water-current speed were positively and significantly related to $\log_{10}$ *E. coli* for each dataset or combination of datasets analyzed (table 2). The strongest correlations were observed between $\log_{10}$ *E. coli* and $\log_{10}$turbidity with Spearman's rho values of 0.577 in the combined 2005-2006 dataset and 0.539 in the combined 2004–2006 dataset. Weaker correlations exist between $\log_{10}$ *E. coli* concentrations and other variables as indicated by smaller values of Spearman's rho (statistically significant at the 95-percent confidence level). The bird count variable was statistically correlated to $\log_{10}$ *E. coli* in all datasets analyzed for this study (Spearman's rho value of 0.342 in the 2006 dataset, for example) although it was not found to be statistically correlated to $\log_{10}$ *E. coli* in a previous study (Zimmerman, 2006). One streamflow variable was weakly correlated to $\log_{10}$*E. coli* in the combined 2004–2006 dataset with a Spearman's rho value of 0.142 ($\log_{10}$Q_inst). None of the streamflow variables were statistically correlated to $\log_{10}$ *E. coli* in the 2006 or combined 2005–2006 datasets. Variables that were statistically correlated to $\log_{10}$*E. coli* were used in model development.

The Kruskal-Wallis test and Tukey-Kramer multiple-comparison test were used to determine if there were relations between $\log_{10}$*E. coli* concentrations and categorical variables. The categorical variables were wind direction, water-current direction, and wave height. Statistical relations were found between $\log_{10}$ *E. coli* and wind direction. The $\log_{10}$ *E. coli* data were grouped according to wind direction with fewer than five observations for any particular wind direction being combined with a logically similar wind direction. Significantly higher median concentrations of $\log_{10}$*E. coli* were observed when winds were blowing from the north/northeast direction than when winds were from the east, southeast/south, or northwest directions (fig. 2). High concentrations of $\log_{10}$*E. coli* usually were observed when winds were parallel with the shoreline, keeping the contamination in the swimming area.

Statistically significant differences were observed between $\log_{10}$ *E. coli* concentrations and water-current direction categories using the Kruskal-Wallis test and Tukey-Kramer multiple-comparison test. Log_{10}*E. coli* data were grouped according to water-current direction with fewer than five observations for any particular water-current direction being combined with a logically similar water-current direction. Although most water currents (approximately 77 percent) were from the northeast or east, the median concentrations of $\log_{10}$*E. coli* when water currents were from those directions were similar to when they were from any other direction. Median $\log_{10}$ *E. coli* concentrations were significantly higher when water currents were from the southwest than when water currents were from southeast or northwest/north directions. Water currents were from the northeast/east or southwest directions when concentrations of $\log_{10}$*E. coli* exceeded the single-sample bathing-water standard (fig. 3). High concentrations of $\log_{10}$*E. coli* usually were observed when water currents were parallel to the shoreline (northeast and southwest), keeping the contamination in the swimming area.

Wave height was another categorical variable analyzed using the Kruskal-Wallis test and Tukey-Kramer multiple-comparison test. Wave heights were categorized as 0 to 2 ft, 1 to 3 ft, 2 to 4 ft, 3 to 5 ft, or 4 to 6 ft, and $\log_{10}$*E. coli* concentrations were grouped by wave-height category. Only 6 out of 291 observations had wave heights in the 4- to 6-ft category so the 3- to 5-ft and 4- to 6-ft categories were combined. The analyses showed that median $\log_{10}$*E. coli* concentrations increased as wave height increased (fig. 4) and all wave-height categories were statistically different from each other (as indicated by different Tukey-Kramer letter designations assigned to each category). Significantly higher median concentrations

Table 1. Comparison of responses of the Presque Isle (Beach 2) 2004–2005 model in 2006 and responses using current method of assessing bacteriological quality of recreational waters.

Assessment method		Number of samples	Response (percent)		
			Total correct	Specificity[1]	Sensitivity[2]
2004-2005 Model (using 2006 data)	Exceedance probability (threshold 27 percent)	111	90.1	97.0	11.1
	Predicted *E. coli*	111	92.8	100	11.1
Current	*E. coli* from the previous day	83	83.1	89.6	0.

[1] The specificity is the percentage of responses that were correctly predicted as not exceeding the standard (safe for swimming).

[2] The sensitivity is the percentage of responses that were correctly predicted as exceeding the standard (not safe for swimming).

Table 2. Summary of Spearman's rho correlations between log_{10} *Escherichia coli* concentrations in water and selected water-quality or environmental variables at Presque Isle Beach 2, Erie, Pennsylvania, 2004–2005.

[**bold type** denotes variables statistically correlated to log_{10} *Escherichia coli* at the 95-percent confidence level (probabilities of 0.05 or less); <, less than]

Variable	Spearman's rho correlation coefficient (probability) Number of samples		
	2006	2005–2006 combined	2004–2006 combined
Water temperature	0.117 (.2184) 113	0.111 (.0942) 229	0.108 (.0677) 287
Dissolved oxygen	.078 (.4125) 113	-.012 (.8635) 221	.003 (.9544) 279
Log_{10} turbidity	.514 **(<.0001)** 113	.577 **(<.0001)** 221	.539 **(<.0001)** 279
Bird count	.342 **(.0002)** 112	.186 **(.0063)** 214	.174 **(.0040)** 272
Rain24 [1]	.302 **(.0011)** 113	.378 **(<.0001)** 229	.344 **(<.0001)** 291
Rain48 [1]	-.023 (.8063) 113	.035 (.6016) 229	.025 (.6708) 291
Rain72 [1]	-0.034 (.7238) 113	.012 (.8580) 229	-.040 (.4986) 291
Rain weight [2]	.129 (.1734) 113	.258 **(<.0001)** 229	.203 **(.0005)** 291
Log_{10} Q_inst [3]	.000 (.9986) 113	.112 (.0904) 229	.142 **(.0153)** 291
Log_{10} Q_prev [4]	-.062 (.5149) 113	.072 (.2783) 229	.100 (.0879) 291
Wind speed [5]	.370 **(<.0001)** 113	.359 **(<.0001)** 229	.281 **(<.0001)** 291
Water-current speed [6]	.352 **(.0001)** 113	.213 **(.0012)** 229	.249 **(<.0001)** 291

[1]Rain24, rain48, and rain72 were the amounts of rain that fell at Erie International Airport, Erie, Pa., in the 24-, 48-, and 72-hour period, respectively, before the sample was collected at Presque Isle Beach 2.

[2]Rain weight is the sum of weighted rainfall amounts from the past 24-, 48-, and 72-hour period giving the most weight to amounts closest to sampling.

[3]Q_inst is the instantaneous streamflow measurement (nearest 30-minute measurement to time of sampling) from the U.S. Geological Survey streamflow-gaging station at Brandy Run near Girard, Pa. (04213075).

[4]Q_prev is the mean daily streamflow at Brandy Run near Girard station, Pa., the day prior to sampling.

[5]Wind speed was obtained from a weather station at the Erie International Airport, Erie, Pa. (Pennsylvania State Climatologist, 2007).

[6]Water-current speed was obtained from the National Oceanic and Atmospheric Administration (NOAA) Great Lakes Coastal Forecasting System (Greg Lang, National Oceanic and Atmospheric Administration, written commun., 2005, 2006).

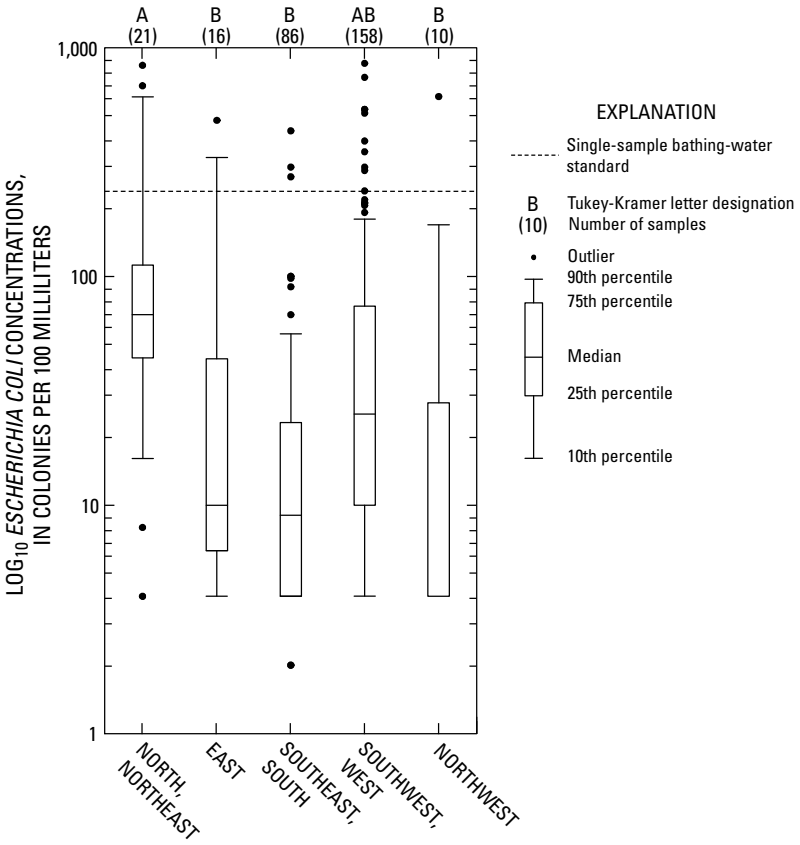

Figure 2. Log_{10} *Escherichia coli* concentrations by wind direction, Presque Isle Beach 2, City of Erie, Erie County, Pennsylvania.
[Results of the Tukey-Kramer test are designated with letters; categories with at least one letter in common are not significantly different]

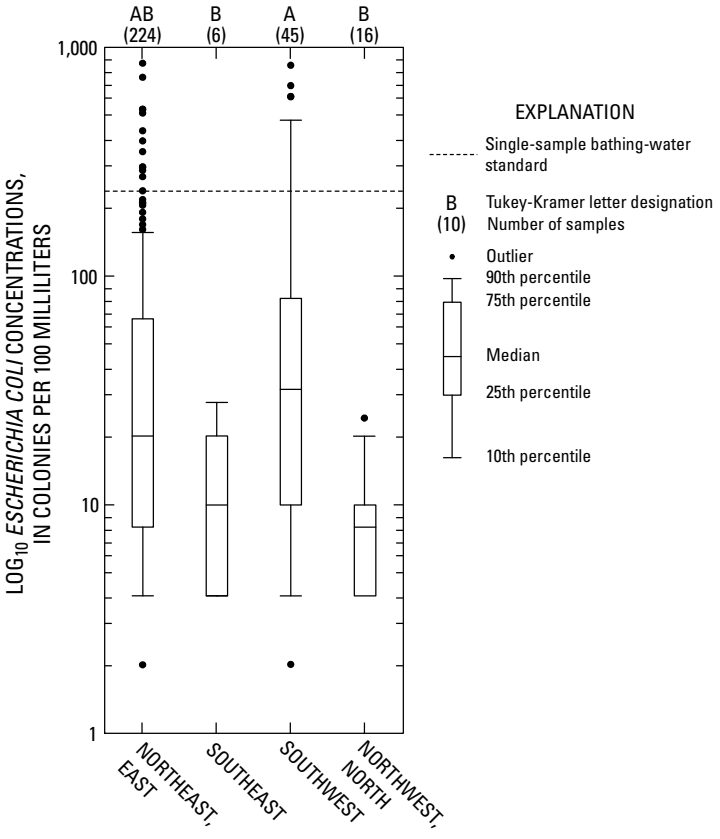

Figure 3. Log_{10} *Escherichia coli* concentrations by water-current direction, Presque Isle Beach 2, City of Erie, Erie County, Pennsylvania.
[Results of the Tukey-Kramer test are designated with letters; categories with at least one letter in common are not significantly different]

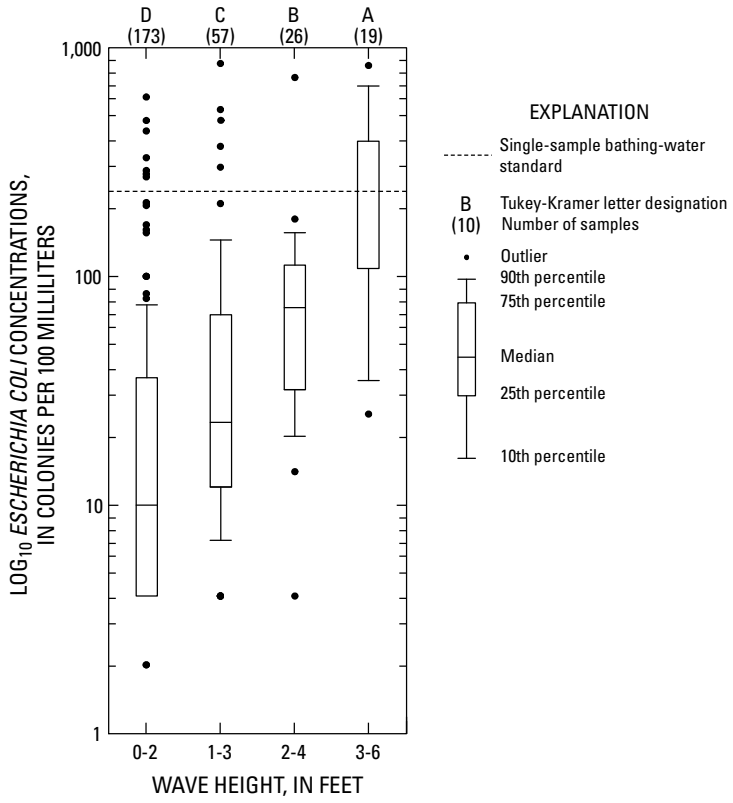

Figure 4. Log_{10} *Escherichia coli* concentrations by wave height, Presque Isle Beach 2, City of Erie, Erie County, Pennsylvania.
[Results of the Tukey-Kramer test are designated with letters; categories with at least one letter in common are not significantly different]

of log_{10} *E. coli* concentrations were observed when waves were from 3 to 6 ft than when wave height was in any other category. Wave heights from 0 to 2 ft had significantly lower median concentrations of log_{10} *E. coli* than observed in any other wave-height category. Concentrations of log_{10} *E. coli* exceeded the standard nearly 50 percent of the time when waves were from 3 to 6 ft.

Modeling to Predict *Escherichia coli*

Variables statistically related to *E. coli* as determined by Spearman's test or the Kruskal-Wallis and Tukey-Kramer tests were considered as explanatory variables in the development of models using tobit regression-analysis techniques. The models were used to predict concentrations of *E. coli* and to predict the probability that the *E. coli* single-sample bathing-water standard of 235 col/100 mL would be exceeded. Predicted exceedance probabilities as output by the best models produced from 2- and 3-year combinations of data analyzed were used to determine a threshold probability, above which the fewest number of false positives and false negatives and the greatest number of correct exceedances and correct nonexceedances for *E. coli* concentrations would be observed.

A new model was developed after model validation using the data collected during 2006 and adding it to the 2004–2005 dataset. The best 2004–2006 model (model with all 3 years of data combined) contained the variables log_{10} turbidity, bird count, and wave height. Combining the 2005 and 2006 datasets (omitting 2004 data) produced the best model overall with the same explanatory variables as when all 3 years of data were combined (log_{10} turbidity, bird count, and wave height) and the highest number of correct exceedances (9 out of 14 or 64.3 percent).

The summary of model output in table 3 shows variables and regression statistics for the "best" models produced from 2- and 3-year combinations of data. The threshold probabilities ranged from 27 in the 2004–2005 (2-year) model and 2004–2006 (3-year) model to 26 in the 2005–2006 (2-year) model; generalized R^2 values ranged from 0.66 in the 2005–2006 model to 0.62 in the 2004–2006 model. All the best models included the explanatory variables log_{10} turbidity and wave height. The best combined 2004–2005 model also included the explanatory variables rain weight and wind direction, whereas the best 2005–2006 and 2004–2006 models included the explanatory variable bird count.

The regression models were used to predict output values for each observation used to develop the models. First, models were used to predict *E. coli* concentrations, and then the predicted values for *E. coli* were used to determine the exceedance probabilities—the probability that the predicted *E. coli* concentrations would exceed the standard. A scatterplot of observed *E. coli* concentrations and predicted exceedance probabilities for the corresponding predicted *E. coli* concentrations helps determine a threshold probability for each model

Table 3. Summary of tobit regression model explanatory variables, statistics, and performance with selected threshold probabilities in predicting exceedances of the *Escherichia coli* single-sample bathing-water standard of 235 col/100 mL for Presque Isle Beach 2, 2004–2006.

[Generalized R^2, a number between 0 and 1 that is larger when the explanatory variables are more strongly associated with the dependent variable (Allison, 1995); threshold probability, exceedance probability chosen for a model on the basis of the *E. coli* standard of 235 col/100 mL being met or exceeded].

Summary	Combined 2004–2005 model (2-year dataset)	Combined 2005–2006 model (2-year dataset)	Combined 2004–2006 model (3-year dataset)
Number of observations	166	201	255
Generalized R^2	.64	.66	.62
Explanatory variables in model	$\log_{10}$ turbidity rain weight wave height wind direction	$\log_{10}$ turbidity bird count wave height	$\log_{10}$ turbidity bird count wave height
Threshold probability	27	26	27
Number of correct exceedances	6	9	9
Number of correct nonexceedances	150	178	226
Number of false positives	4	9	10
Number of false negatives	6	5	10

that is associated with too great a risk to allow swimming. The combined 2005–2006 model results are shown in figure 5. The plot is divided into four sections using the *E. coli* single-sample bathing-water standard to divide the plot vertically (2.37=$\log_{10}$ of 235 col/100 mL) and the chosen threshold probability of 26 to divide the plot horizontally. Moving the threshold-probability line up or down on the plot changes the number of observations in each of the four sections. The goal in establishing a threshold probability is to maximize the number of correct responses showing *E. coli* concentrations (1) above the *E. coli* standard of 235 col/100 mL and having a predicted probability of exceedance above the established threshold probability (correct exceedance—upper-right quadrant), and (2) meeting the standard by being less than 235 col/100 mL and having a predicted probability of exceedance below the established threshold probability (correct nonexceedance—lower-left quadrant). Minimizing the number of incorrect responses is another goal in establishing the threshold probability. False negatives (lower-right quadrant) result when the standard was, in fact, exceeded but the predicted probability of exceedance was below the established threshold probability. False positives (upper-left quadrant) result when the standard was not exceeded but the predicted probability of exceedance was above the established threshold probability. Determinations of whether to close the beach or post an advisory can be made if predicted exceedance probabilities exceed the established threshold probability.

The performance of each model was evaluated by comparing model responses in predicting *E. coli* concentrations and exceedance probabilities to the use of *E. coli* concentrations from the previous day (table 4). In all cases, responses from the models yielded higher percentages of correct predictions than *E. coli* concentrations from the previous day. The overall best model for Beach 2 was produced by excluding the 2004 dataset; it is unclear why omitting the 2004 data produced a better model. The 2005–2006 model correctly predicted when the standard would not be exceeded (specificity) with a response of 95.2 percent (178 out of 187 nonexceedances) and correctly predicted when the standard would be exceeded (sensitivity) with a response of 64.3 percent (9 out of 14 exceedances).

Summary and Conclusions

This report describes a study done during the 2006 recreational season by the U.S. Geological Survey, in cooperation with the Erie County Health Department, to validate and refine regression models designed to predict *E. coli* concentrations used to determine the probability of exceeding the *E. coli* single-sample bathing-water standard at Presque Isle Beach 2 in Erie, Pa. The results of this study will provide a supplemental method for determining the recreational water quality of Presque Isle Beach 2 that will aid beach managers in more rapidly determining when waters are not safe for recreational use and when to close a beach or post an advisory.

The 2004–2005 model was validated with data collected during 2006. Model output (predicted *E. coli* concentrations and predicted exceedance probabilities) produced higher percentages of correct predictions than the use of *E. coli* concentrations from the previous day. All assessment methods produced similar percentages of correctly predicting when the standard would not be exceeded (specificities); those percent-

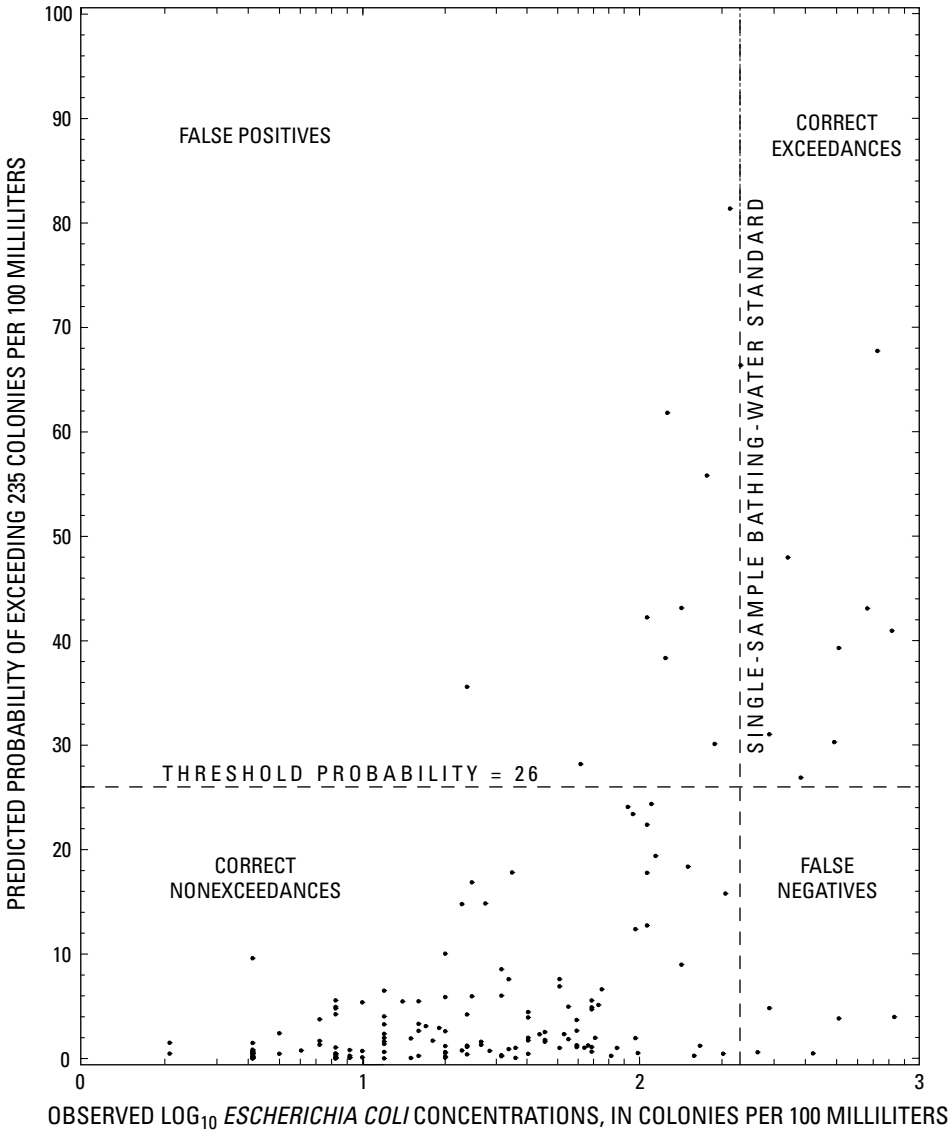

Figure 5. Observed $\log_{10}$ *Escherichia coli* concentrations and probability of exceeding 235 col/100 mL to achieve optimum threshold probability for the combined 2005–2006 model for Presque Isle Beach 2, City of Erie, Erie County, Pennsylvania.
[2005–2006 model explanatory variables for determining bacteria concentrations include $\log_{10}$ turbidity, bird count, and wave height]

Table 4. Comparison of responses of Presque Isle Beach 2 models and responses using current method of assessing bacteriological quality of recreational waters.

Dataset	Assessment method		Number of samples	Response (percent)		
				Total correct	Specificity[1]	Sensitivity[2]
2004–2005	Model	Exceedance probability (threshold 27 percent)	166	94.0	97.4	50.0
		Predicted *Escherichia coli*	166	92.8	98.7	16.7
	Current	*Escherichia coli* from the previous day	114	88.6	95.2	11.1
2005–2006	Model	Exceedance probability (threshold 26 percent)	201	93.0	95.2	64.3
		Predicted *Escherichia coli*	201	92.5	98.4	14.3
	Current	*Escherichia coli* from the previous day	161	87.6	94.0	0.
2004–2006	Model	Exceedance probability (threshold 27 percent)	255	92.2	95.8	47.4
		Predicted *Escherichia coli*	255	92.2	98.7	10.5
	Current	*Escherichia coli* from the previous day	197	86.3	92.8	6.67

[1] The specificity is the percentage of responses that were correctly predicted as not exceeding the standard (safe for swimming).

[2] The sensitivity is the percentage of responses that were correctly predicted as exceeding the standard (not safe for swimming).

ages were slightly higher in the responses from the model. There were differences between the model and the current assessment method in percentages of correctly predicting when the standard would be exceeded and waters would not be safe for swimming (sensitivities). The model responses were 11.1 percent for predicted *E. coli* concentrations and predicted exceedance probabilities. No correct responses (zero percent) were observed for sensitivity using *E. coli* concentrations from the previous day.

Correlation tests were conducted to determine the continuous variables that were related to $\log_{10}$ *E. coli* concentrations. In each of the datasets analyzed, $\log_{10}$ turbidity, rain24, wave height (calculated), wind speed, and water-current speed were found to be correlated to $\log_{10}$ *E. coli* concentrations. The strongest correlations were observed between $\log_{10}$ *E. coli* and $\log_{10}$ turbidity in the combined 2004–2005 dataset and the combined 2005–2006 dataset with Spearman's rho correlation coefficients 0.662 and 0.577, respectively. Weaker correlations exist between $\log_{10}$ *E. coli* concentrations and the streamflow variables, the rain variables, and number of birds (bird count) as indicated by smaller values of Spearman's rho (statistically significant at the 95-percent confidence level). Variables that were statistically correlated to *E. coli* were considered in model development.

The Kruskal-Wallis and Tukey-Kramer multiple-comparison tests were used to determine if there were relations between *E. coli* concentrations and categorical environmental variables. Statistical relations were observed between $\log_{10}$ *E. coli* and wind direction. Significantly higher median concentrations of $\log_{10}$ *E. coli* were observed when winds were blowing from the north/northeast direction than when winds were from the east, southeast/south, or northwest directions. Statistical relations also were observed between $\log_{10}$ *E. coli* and direction of water currents. The single-sample bathing-water standard was exceeded when water-current directions were northeast/east or southwest. Wave heights were categorized into one of five groups—0 to 2 feet, 1 to 3 feet, 2 to 4 feet, 3 to 5 feet, and 4 to 6 feet. All categories of wave height were significantly different from each other. Significantly higher $\log_{10}$ *E. coli* concentrations were observed when wave heights were from 3 to 6 feet than any other wave-height category. Statistically lower $\log_{10}$ *E. coli* concentrations were observed when wave heights were from 0 to 2 feet than any other wave height category.

Models were developed at Presque Isle Beach 2 using combinations of data collected from 2004 to 2006. All water-quality and environmental variables that were found to be related to $\log_{10}$ *E. coli* concentrations were considered in the tobit regression models. All the "best" models included the explanatory variables $\log_{10}$ turbidity and wave height. Rain weight and wind direction were explanatory variables in the 2004–2005 model; bird count was an explanatory variable in the 2005–2006 and 2004–2006 models. Combining the 2005 and 2006 datasets produced a model with the same explanatory variables as when all 3 years of data were combined (2004–2006). Using only 2005 and 2006 data produced a model with the highest number of correctly predicted exceedances (9 out of 14 or 64.3 percent). The explanatory

variables in the best predictive models for *E. coli* at Presque Isle Beach 2 could indicate that *E. coli* concentrations from birds are being washed into the lake during storms and then lakebed sediments harboring bacteria are being resuspended when winds produce high waves. In all cases, the results from predictive modeling produced higher percentages of correct predictions than using *E. coli* concentrations from the previous day.

Further study could focus on improving the predictive ability of the model by adding data from new weather stations and a streamflow-gaging station closer to the Presque Isle beaches. Two new weather stations have been installed—one on the tower of the Tom Ridge Environmental Center and one near Presque Isle Beach 2 in Presque Isle State Park. These weather stations collect various data including rainfall, wind speed, and wind direction. Wind and rainfall variables are among the explanatory variables that have been shown to explain the distribution in *E. coli* concentrations at Presque Isle Beach 2 and collecting data in close proximity to the beach might improve the model results. Similarly, streamflow is a variable that has been weakly correlated to *E. coli* concentrations at Beach 2. The installation of a steamflow-gaging station on Walnut Creek is planned for the fall of 2007. The new gaging station will be closer (about 20 miles) than the gaging station from which data were compiled for this study. Data from a gaging station closer to the Presque Isle beaches may improve the predictive ability of the model.

Acknowledgments

The author thanks Scott White from the Erie County Health Department for his asssistance throughout all phases of the project. The author also acknowledges the assistance that Linda Zarr of the U.S. Geological Survey provided in helping with data compilation and processing.

References Cited

Allison, P.D., 1995, Survival analysis using SAS—A practical guide: Cary, N.C., SAS Institute, Inc. 292 p.

Erie County Department of Health, 2007, Bacterial quality of waters near Presque Isle—A compilation of relevant publications 1987-2007: accessed October 1, 2007, at *http://www.ecdh.org/environmental/documents/BacterialWaterQuality1.pdf*.

Francy, D.S., and Darner, R.A., 2006, Procedures for developing models to predict exceedances of recreational water-quality standards at coastal beaches: U.S. Geological Survey Techniques and Methods 6-B5, accessed October 1, 2007, at *http://pubs.usgs.gov/tm/2006/tm6b5/pdf/tm6B5_web_rev120706.pdf*.

Myers, D.N., Stoeckel, D.M., Bushon, R.N., Francy, D.S., and Brady, A.M.G., 2007, Fecal indicator bacteria: U.S. Geological Survey Techniques of Water-Resources Investigations, book 9, chap. A7 (version 2.0), section 7.1, available from *http://pubs.water.usgs.gov/twri9A/*.

Natural Resources Defense Council, 2007, Testing the waters–A guide to water quality at vacation beaches, August 2007: accessed September 25, 2007, at *http://www.nrdc.org/water/oceans/ttw/sumpen.pdf*.

Pennsylvania Code, Title 28. Health and safety, Chapter 18. Public swimming and bathing places, (28PaCode § 18): accessed May 2, 2006, at *http://www.pacode.com/*.

Pennsylvania Department of Conservation and Natural Resources, 2006, State Parks: accessed May 2, 2006, at *http://www.dcnr.state.pa.us/stateparks/parks/maps/presqueisle_mini.pdf*.

Pennsylvania State Climatologist, 2007, Hourly Mesonet Data: accessed January 4, 2008, at *http://climate.psu.edu/data/mesonet/datainv.php*.

SAS Institute, 1990, SAS user's guide—The LIFEREG procedure: Cary, N.C., SAS Institute, Inc., Version 6, Fourth edition, v. 2, p. 997–1,025.

Siwicki, R.W., 2005, Water resources data, Pennsylvania, water year 2004, vol. 3, Ohio and St. Lawrence River Basins: U.S. Geological Survey Water-Data Report PA-04-3, 337 p.

Siwicki, R.W., 2006, Water resources data, Pennsylvania, water year 2005, vol. 3, Ohio and St. Lawrence River Basins: U.S. Geological Survey Water-Data Report PA-05-3, 355 p.

U.S. Environmental Protection Agency, 1986, Ambient water-quality criteria for bacteria—1986: Washington, D.C., Office of Water, EPA-440/5-84-002, 18 p.

U.S. Environmental Protection Agency, 2002, Method 1603: *Escherichia coli* (*E. coli*) in water by membrane filtration using modified membrane-thermotolerant *Escherichia coli* agar (modified mTEC): Washington, D.C., Office of Water, EPA-821-R-02-023, 9 p.

U.S. Environmental Protection Agency, 2007, EPA's BEACH report—Pennsylvania 2006 swimming season: accessed November 14, 2007, at *http://www.epa.gov/waterscience/beaches/seasons/2006/pdf/pa.pdf*.

U.S. Geological Survey, 2007, Water-resources data for the United States, Water Year 2006: U.S. Geological Survey Water-Data Report WDR-US-2006, accessed October 1, 2007, at *http://pubs.water.usgs.gov/wdr2006.*

Zimmerman, T.M., 2006, Monitoring and modeling to predict *Escherichia coli* at Presque Isle Beach 2, City of Erie, Erie County, Pennsylvania: U.S. Geological Survey Scientific Investigations Report 2006-5159, 15 p.