

## The Evolution of a Predictive Model at a Lake Erie Bathing Beach

Jill Lis, R.S.  
Cuyahoga County Board of Health  
5550 Venture Drive  
Parma, Ohio 44130

### Background

Ohio's state health department, the Ohio Department of Health (ODH), has been administering a Bathing Beach Monitoring Program since mid-1970. The program is a cooperative effort between the ODH, the Ohio Department of Natural Resources, public and private organizations along the Lake Erie shoreline, and local health departments with beaches in their jurisdictions. The ODH aims to encourage local health departments to develop their own monitoring program; however, it is not mandatory. The ODH program also provides incentive for the development of predictive models for assessing water quality, pre-emptive warning systems for public notification, and for identifying and eliminating potential sources of pollution.

In May of 1993, the Cuyahoga County Board of Health (CCBH) adopted regulations to conduct a Bathing Beach Water Quality Program in order to protect the public from potential health risks associated with swimming in bathing waters. Cuyahoga County is located along the Lake Erie shoreline in northeast Ohio. The CCBH, located in Parma, Ohio, and within the Cleveland Metropolitan Area, currently has 2 public and 16 private (community) beaches under its jurisdiction. The beaches are scattered throughout the County and consist of inland lakes and beaches located along the Lake Erie shoreline. The public beaches are sampled a minimum of 4 days a week and the community beaches are sampled at varying frequencies, ranging from once a week to once a month depending on its potential usage.

In Ohio, beach water samples are collected to determine the presence of *Escherichia coli* (*E. coli*) bacteria. *E. coli* is a group of bacteria that is commonly found in the intestinal tracts of man and other warm-blooded animals. *E. coli* concentrations have been shown to correlate more closely with incidences of swimming-associated gastrointestinal illness than fecal coliform, the former water quality standard, and it is therefore used as an indicator of contamination of water. Most strains of *E. coli* are harmless; however, the presence of this bacterium may indicate that disease-causing organisms are also present.

The standard used to assess bathing beach water quality in Ohio is the single-sample maximum level for *E. coli* of 235 colonies per 100 milliliters water. To determine concentrations of *E. coli*, however, the bacteria must be cultured for 18-24 hours. Water sample results are therefore not available until the day after a water sample is collected. Beach managers are consequently using the previous day's *E. coli* results to evaluate current beach water quality conditions. This lag time may result in an inaccurate evaluation of

current conditions because water quality may change overnight, as well as throughout the day, leaving the public at risk of exposure to potential waterborne pathogens.

The importance of beach water quality and its impact on human health resulted in the Beaches Environmental Assessment and Coastal Health (BEACH) Act, which assists governments and public health officials in reducing the risk of illness associated with using recreational waters. The BEACH Act mandates that states develop performance criteria for monitoring recreational water quality by enumerating the bacteria that indicate the presence of fecal waste in recreational waters and to subsequently notify the public in a timely manner when water quality standards are exceeded. In keeping consistent with BEACH Act requirements, the CCBH revised its original Bathing Beach Regulations in 2006 to include an enhanced emphasis on public notification and the development of a beach classification scheme based upon potential usage and water quality risk factors. The CCBH Program is now referred to as its Bathing Beach Water Quality and Public Notification Program, which is managed and operated utilizing a tiered monitoring approach. One means of further enhancing this plan was the implementation of the *Nowcasting System for Predicting Beach Advisories*, which will be the focus of this paper.

#### Collaboration with the U.S Geological Survey

Agencies that monitor bathing beaches need tools that can provide quick, reliable indicators of recreational water quality. Real-time forecasting using mathematical models may help resolve the delayed notification problems inherent with the current approach. Mathematical models use easily measured environmental and water-quality variables (“explanatory variables”), such as wave height and rainfall, to estimate *E. coli* concentrations or the probability of exceeding the water quality standard of 235 col/100 mL of *E. coli* (Francy and others, 2006).

Mathematical models, also known as predictive models, are endorsed by the BEACH Act as a means of providing estimations in water quality due to the lag time in obtaining actual water sample results. Predictive modeling, however, is not a replacement for bathing beach water sampling, but rather is meant to enhance and complement current monitoring efforts.

The CCBH began working with the U.S. Geological Survey (USGS), Ohio Water Science Center, during the 2000 recreation season. The USGS received a grant through the Ohio Lake Erie Commission – Lake Erie Protection Fund and the Ohio Water Development Authority to build upon earlier studies to develop and test multiple linear regression (MLR) models to predict *E. coli* concentrations using water quality and environmental factors as explanatory variables. MLR is a statistical tool that is used for identifying relationships between two or more explanatory variables. It is used to find a linear equation that best predicts the dependent variable from the independent variables.

Six beaches were studied by the USGS within the Cleveland Metropolitan Area: Edgewater Beach, Villa Angela Beach, Huntington Beach, Mentor Headlands, Fairport Harbor, and Mosquito Lake. Models were developed for each beach that resulted in an output variable of the probability that the single sample standard for *E. coli* would be exceeded. Threshold

values were established for each beach, whereby if the computed probability was determined to be less than the threshold value, bacteria levels would likely be acceptable. Computed probabilities that were equal to or greater than the threshold value would conversely indicate that bacteria levels were not acceptable and that a water quality advisory may be warranted (Francy and others, 2003).

The goal of a threshold value, or threshold probability, is to produce the most correct responses and the least false negative responses. A correct response by the model would mean that the model accurately calculated (predicted) that the water quality standard for *E. coli* either would or would not be exceeded. A false positive response would mean that the model predicted water quality conditions to be unacceptable, but they actually were acceptable. A false negative response, however, would mean that the model predicted water quality conditions to be acceptable, but they were actually unacceptable. The false negative responses must therefore be minimized, since under these conditions the public is erroneously being exposed to potential waterborne pathogens.

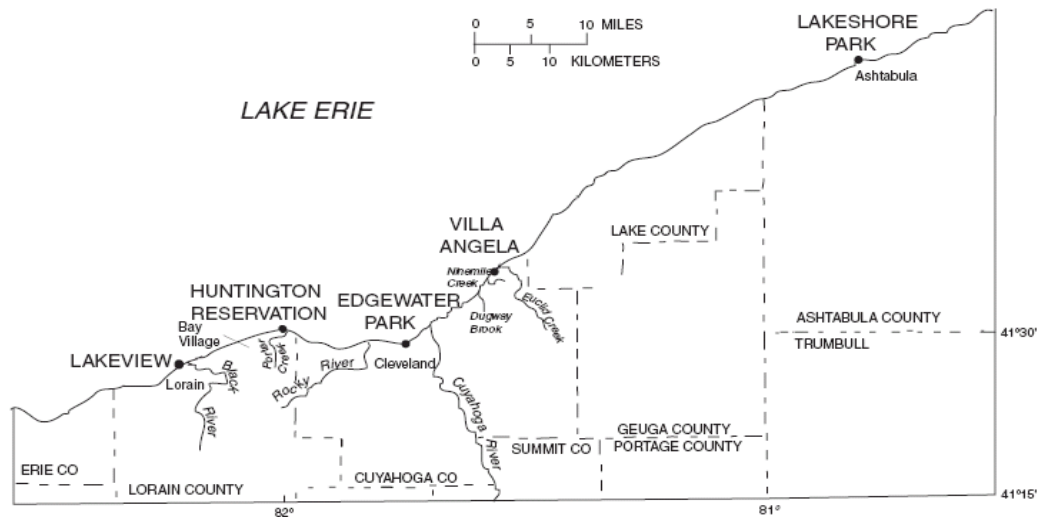
For the six beaches studied, a model was developed and tested for each beach based on its individual explanatory variables; different factors were shown to affect different beaches, resulting in beach-specific models. The exceptions to the modeling were the beaches at Mentor Headlands and Fairport Harbor. Due to historically acceptable water quality at these beaches, it was determined that these beaches would not be modeled (Francy and others, 2003).

These initial modeling studies led to further grant work between the USGS and the CCBH, specific to Huntington Beach. Huntington Beach is a Lake Erie beach located in Bay Village, Ohio. Although the preliminary model showed promise, testing and validation of the model by the USGS revealed that more research was needed to improve the model's performance. The USGS consequently received another Ohio Lake Erie Commission - Lake Erie Protection Fund Grant to continue to study Huntington Beach during the 2004 and 2005 recreational seasons, with a goal of developing a model that worked well when validated against actual water sample results. The final project outcome would be the implementation of the model utilizing a near-real-time Internet-based system for disseminating the data to the public. This concept was called a "Nowcast", which was put into operation during the 2006 recreational season via the creation of a dedicated website: [www.ohionowcast.info](http://www.ohionowcast.info).

### Model Development

All information pertaining to the *Nowcast* project can be found on the website, with the principal project details provided below to demonstrate how the work was carried out. For reference purposes, the website was a collaboration between the CCBH and the USGS and the information below was taken, in part, directly from the website. The information below was provided on the website during 2006, and will be updated shortly for 2007. It should be noted that the USGS has continually been studying beaches in addition to Huntington Beach; however those beaches are not under the jurisdiction of the CCBH. The other beaches include Lakeview Beach, Edgewater Beach, Villa Angela Beach, and Lakeshore

Park. Models are being developed for those beaches; however the *Nowcast* system has only been implemented at Huntington Beach since its research was further along than the other beaches and since its model had been shown through a validation process to be working well. The map below illustrates the USGS study area for the development of predictive models.



(USGS Study Area, Northeast Ohio)

### ***Data collection***

Data collection included analysis of daily water samples for *E. coli* and measurement of explanatory variables for model development and testing.

At Huntington, CCBH collected water samples Monday through Thursday during the recreational seasons (late May through Labor Day) of 2000–2005. Samples were collected between 7 and 10 a.m. where the water was 3 feet deep in areas of the beach used for swimming. All water-sample bottles were filled about 1 foot below the water surface using a grab-sampling technique (Myers and Wilde, 2003). Water samples were kept on ice and analyzed for concentrations of *E. coli* and turbidity at local laboratories within 6 hours of collection.

Although the most recent modeling project began in 2003, the USGS utilized data previously collected by the CCBH beginning with the 2000 recreational season as a means of testing and refining the various models. As part of its routine sampling regimen, the CCBH collects data at the time of sample collection pertaining to: water temperature, wave height, number of birds and bathers on the beach, weather conditions at the time of sampling, and information on whether or not it rained over the previous 24 hours. Turbidity measurements are also made on water samples at the time they are delivered to the lab that conducts the sample analysis. This routine environmental data was extremely valuable in the development of the model and is the primary reason that the research at Huntington Beach was further along than the other beaches.

During the study period, field personnel collected or compiled data for environmental and water-quality variables expected to affect *E. coli* concentrations.

- Bird counts. Manual counts were made of the number of birds on the beach upon arrival.
- Wave heights. At the time of sample collection, wave heights were estimated into four categories based on minimum and maximum heights in each wave train: (1) 0 to 2 feet, (2) 1 to 3 feet, (3) 2 to 4 feet, (4) > 3 to 5 feet.
- Water temperatures. Water temperature was measured at the sampling location using an alcohol-filled thermometer.
- Lake levels. Lake-level data were obtained from the National Oceanic and Atmospheric Administration (NOAA) station in Cleveland (NOAA ID 9063053) at URL <http://www.co-ops.nos.noaa.gov/>.
- Weather data. Rainfall and wind direction data were obtained from the National Weather Service station at Hopkins International Airport at URL <http://nndc.noaa.gov/>.

Several different rainfall variables were calculated and used in predictive model development.

- $R_{d-1}$  was the amount of rain, in inches, that fell in the 24-hour period (9 a.m. to 9 a.m.) preceding the morning sampling.
- $R_{d-2}$  and  $R_{d-3}$  were amounts of rain that fell in 24-hour periods 2 days and 3 days preceding the morning sampling, respectively.
- Rainfall weighted 48 hours (Rw48) is 48 hours of cumulative rainfall and gives more weight to the most recent rainfall amount as follows:

$$(Rw48) = (2 * R_{d-1} + R_{d-2})$$

“Wind direction 24” was calculated by summing hourly wind vectors for the 24-hour period preceding sampling and determining the direction of the resultant vector. A vector is a calculation based on both the direction and intensity (speed) of wind directions.

### ***Sample analysis***

Samples were analyzed by use of the mTEC membrane-filtration method (U.S. Environmental Protection Agency, 2000). Membrane-filtration equipment and supplies include a manifold and vacuum pump, filter funnel and base, membrane filters, a graduated cylinder and pipets to measure sample volumes, forceps to handle the filter, an alcohol lamp to flame forceps, sterile buffered water to rinse the filter funnel, and agar plates to grow the bacteria.

Several different volumes of sample water were filtered through a membrane filter with the goal of obtaining 20–80 colonies on at least one of the agar plates. Usually, sample volumes of 100, 30, 10, 3, and 1 mL were plated. If the water was suspected to have concentrations of *E. coli* in the thousands of colonies per 100 milliliters, serial dilutions of the sample were made.

The bacteria were concentrated on a filter using the manifold and vacuum pump and filter funnel and base. The filter was placed on an agar plate and incubated at 35°C for 2 hours and at 44.5°C for an additional 20–22 hours. After the prescribed incubation time, the plates were removed, and those membranes with yellow colonies were placed on pads saturated with urea-phenol solution. Those colonies that remain yellow (indicating that they are negative for the enzyme urease) after 20 minutes exposure were counted as positive for *E. coli*. Results were calculated to be reported as colonies per 100 milliliters (col/100 mL).

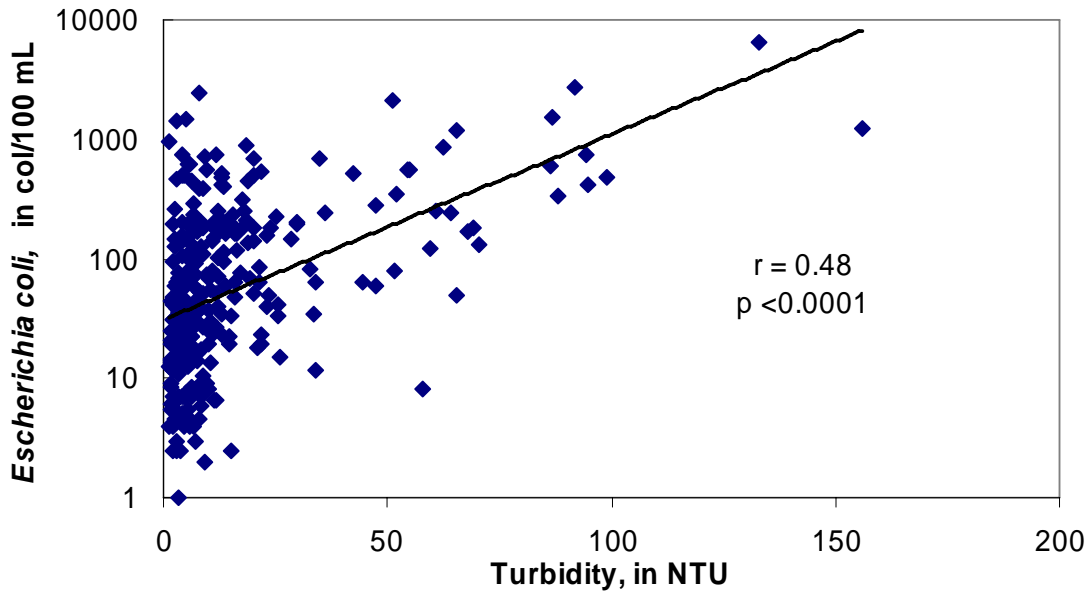
Turbidity measures the scattering effect that suspended solids have on light; the higher the intensity of scattered light, the higher the turbidity. Turbidity can make water look cloudy or muddy. Contributors to turbidity include clay, silt, finely divided organic matter, plankton, microscopic organisms, and dyes (Anderson, 2005). Turbidity was determined in water samples with a turbidimeter. Turbidity is reported in nephelometric turbidity units (NTUs).

## ***Study Results***

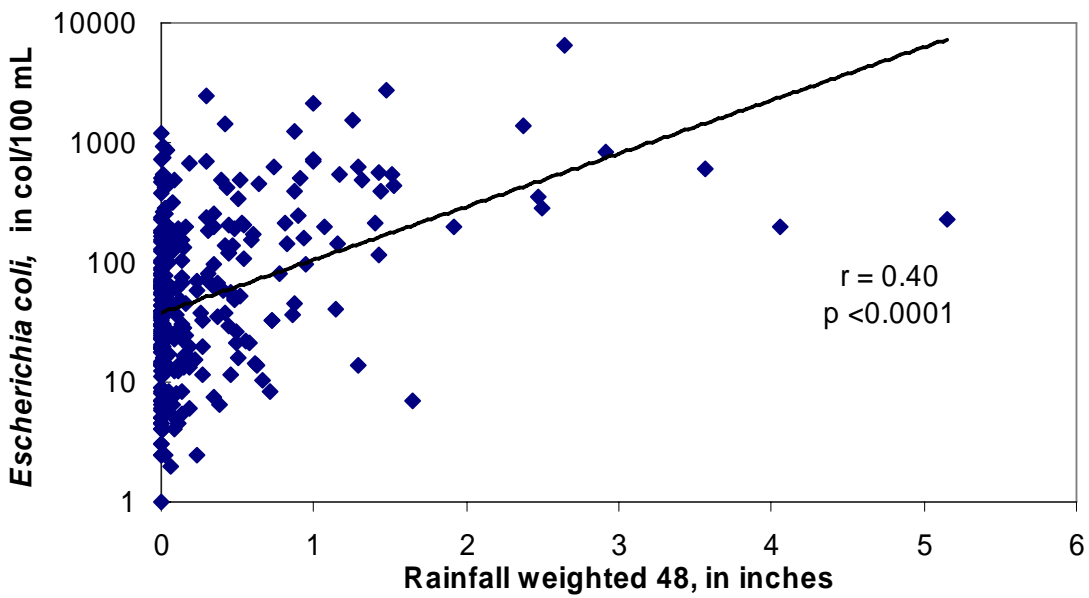
### *Explanatory variables*

The first step in development of predictive models was to identify explanatory variables related to *E. coli*. This was done by making scatterplots of the continuous data collected at Huntington Beach with the measured variable on the x-axis and level of *E. coli* on the y-axis. Continuous data are measured data that have an infinite range of values. A statistical test—correlation analysis—was done to provide a quantitative measure of the relation between the variable and *E. coli*. The result from correlation analysis is a Pearson's *r* correlation value (*r*), which measures the linear (straight line) association between the variable and *E. coli* concentrations. If the data lie exactly along a straight line with positive slope, then the *r* value is equal to 1 (Helsel and Hirsch, 1992, p. 209). The more the correlation coefficient deviates from 1 or -1 and approaches zero, the weaker the relation. Correlation coefficients were considered statistically significant if the *p*-value (level of significance) was < 0.05. When the *p*-value is <0.05, it means that there is less than a 5% chance that the results were statistically significant when they were not. An example of one of the related variables at Huntington Beach is turbidity. As *E. coli* concentrations increased, turbidity also increased ( $r = 0.48$ ,  $p < 0.0001$ ). However, the scatter of *E. coli* concentrations at a given level of turbidity was considerable. One reason for the scatter is that turbidity does not explain all of the variability in *E. coli* concentrations; other explanatory variables are needed to more fully explain this variability. Rainfall weighted 48 ( $r = 0.40$ ,  $p < 0.0001$ ) and day of the year ( $r =$

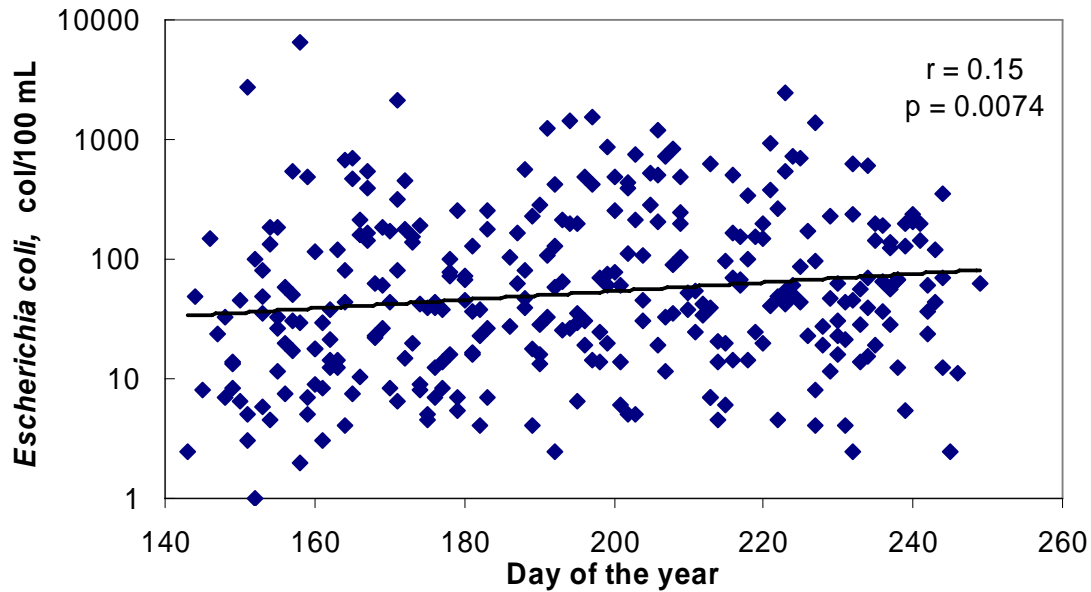
0.15,  $p = 0.0074$ ) were also significantly related to *E. coli* at Huntington Beach during 2000–2005, although day of the year was only weakly related to *E. coli*.



Scatterplot of turbidity versus *E. coli*, Huntington Beach, Bay Village, Ohio 2000–2005.

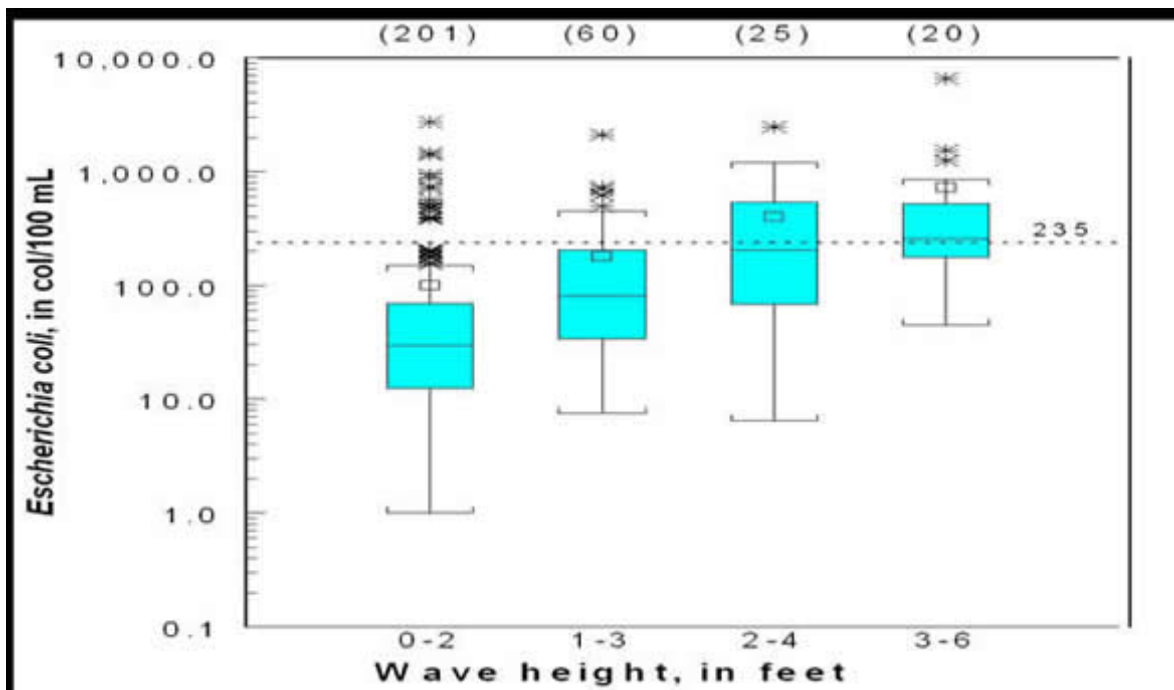


Scatterplot of Rainfall weighted 48 versus *E. coli*, Huntington Beach, Bay Village, Ohio 2000–2005.



Scatterplot of day of the year versus *E. coli*, Huntington Beach, Bay Village, Ohio 2000-2005.  $r$  is the Pearson's correlation coefficient and  $p$  is the significance of the correlation.

Box plots were used to understand the distribution of *E. coli* concentrations in variables that are not continuous but have a finite range of values and are grouped by categories, such as wave height and wind direction. At Huntington Beach during 2000–2005, median *E. coli* concentrations increased with increasing wave height.



Boxplots of wave height versus *E. coli*, Huntington Beach, Bay Village, Ohio 2000–2005. *E. coli* concentrations increased with increasing wave heights.

### *Predictive models*

Different combinations of variables related to *E. coli* were tested through the use of MLR. In MLR, as previously discussed, a unique set of variables is used to develop a model that best explains the variation in *E. coli* concentrations, leaving as little variation as possible to unexplained “noise”. At Huntington Beach, using data collected during 2000–2005, the best MLR model included the variables wave height, turbidity, Rw48, and day of the year. The model explained 42% of the variability in *E. coli* concentrations; this is called the  $R^2$  or coefficient of determination of the model.

### *Output from the model*

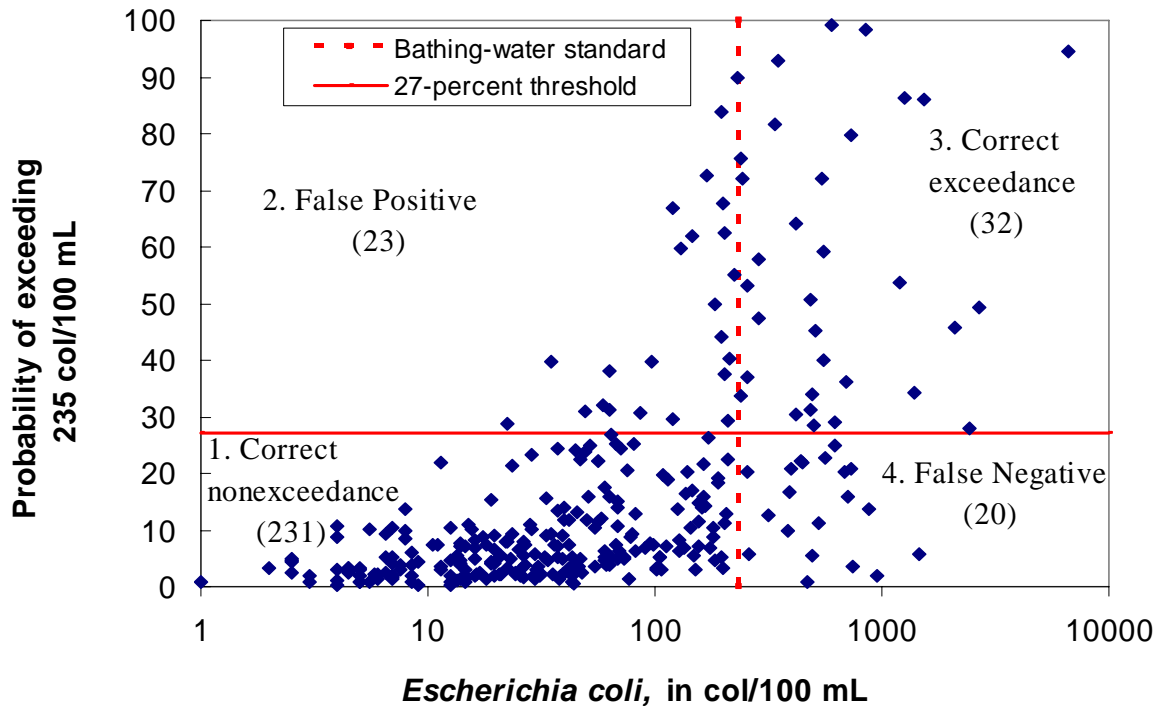
Two types of output values were produced by the models. The first, and simpler, output was the predicted *E. coli* concentration. Because the potential for error in the predicted *E. coli* concentration was shown to be fairly wide in earlier studies, (Francy and Darner, 1998; Francy and Darner, 2002; Francy and others, 2003), a second output variable was developed in the hope of providing a more accurate prediction of recreational water quality—the probability of exceeding the Ohio single-sample bathing water standard for *E. coli* of 235 colonies per 100 milliliters (col/100 mL). This approach results in estimated probabilities similar to those in a weather forecast. For example, if the following data are measured one morning at Huntington Beach: wave height 1-3 ft, Rw48 = 0.29 inches, turbidity = 15.6 NTU, day of the year = 232, the Huntington model predicts a 33.7% probability that the single-sample bathing-water standard will be equaled or exceeded.

The results from the model can be used daily by beach managers and the public. For the model to be useful, the probability that is associated with too great a risk to allow swimming needs to be determined. This is the “threshold probability.” Probabilities that are less than a threshold probability indicate that bacterial water quality that day is most likely acceptable; the beach manager would not issue an advisory and beachgoers would feel fairly confident that the water is safe for swimming. Probabilities equal to or above the threshold probability indicate that the water quality is most likely not acceptable and that a water-quality advisory may be needed.

The threshold probability for the Huntington model was established by determining the lowest probability that produced the most correct responses and fewest false negative responses. This was done by plotting the Huntington 2000–2005 data used to develop the model. Because these data have been examined retrospectively, each point on the graph represents the actual *E. coli* concentration determined by culturing the sample (x-axis) and the associated computed probability based on the model (y-axis). The plot is divided into four quadrants by a vertical line through 235 col/100 mL on the x-axis and a horizontal line through the threshold probability on the y-axis. The four quadrants are:

1. Correct below the standard. *E. coli* concentrations were less than 235 col/100 mL, and the predicted probabilities were **below** the threshold.
2. False positive. *E. coli* concentrations were less than 235 col/100 mL, but the predicted probabilities were **above** the threshold.

3. Correct above the standard. *E. coli* concentrations were equal to or greater than 235 col/100 mL, and the predicted probabilities were **above** the threshold.
4. False negative. *E. coli* concentrations were equal to or greater than 235 col/100 mL, but the predicted probabilities were **below** the threshold.



Scatterplot of threshold probability.

If one were to raise or lower the horizontal line, it would change the number of correct and incorrect responses. For example, a threshold of 45 would have produced the highest number of correct responses (265), but would also produce a high number of false negatives (31). False negative responses are especially troubling because the recreational water quality is determined to be acceptable when in fact the standard was exceeded. Thresholds between 40 and 44 do little to reduce the number of false negatives. Selecting a threshold of 27, however, still maintains a high number of correct responses (263), but yet reduces the false negatives to a more acceptable level and represents a compromise between false negative and false positive responses. In addition, setting the threshold at 27 rather than 28 enables the beach manager to err on the side of safety.

The example previously mentioned - wave height 1-3 ft, Rainfall weighted 48 = 0.29 inches, turbidity = 15.6 NTU, day of the year = 232—resulted in a 33.7% probability that the single-sample bathing-water standard will be equaled or exceeded. The actual *E. coli* concentration measured by culture was 237 col/100 mL. So this result would be placed in quadrant 3—correct exceedance above the standard.

### *Next steps*

The *Nowcast* system that uses the Huntington 2000–2005 model was a pilot study conducted during the 2006 recreational season. This was the first year that model output was provided to the public and the first year that it was used by beach managers to post water-quality advisories at Huntington. In 2006, the *Nowcast* system accurately predicted water quality conditions 80% of the time. The remaining 20% of the data were inaccurate, consisting of a combination of false positive and false negative results. The chart below demonstrates the 2006 data comparisons between the model and current methods of water sample analysis.

		Percentage of Responses		
Predictive Tool	Sample Size	Correct	False Positives	False Negatives
Huntington 2000-2005 Model	85	80%	10% (6/59)	42% (11/26)
Previous Day's <i>E. coli</i> (All days)	84	57%	30% (18/59)	72% (18/25)

- A total of 85 water samples were collected to validate the Model
- 59 of the samples were within the *E. coli* standard
  - 6 of these 59 samples were not predicted correctly by the Model
    - Predicted to be POOR, but actually GOOD (False Positive)
- 26 of the samples exceeded the *E. coli* standard
  - 11 of these 26 samples were not predicted correctly by the Model
    - Predicted to be GOOD, but actually POOR (False Negative)

Investigators and beach managers still need test the utility of the model and determine whether it performs as well as or better than using the current method during future recreational seasons. For the 2007 recreational season, the *Nowcast* system will be utilized again at Huntington Beach, however, refinements have been made to the model by the USGS. The model will include a new variable, referred to as Radar rainfall data, in order to obtain rainfall data from a more widespread area. This radar data is obtained from the National Weather Service and is provided for 4-kilometer grids for each hour of the day. For the area around Huntington Beach, data from 6 grids will be used to estimate rainfall amounts in the past 24 hours (“radar6cell”). In addition, the current threshold value for the model is being increased from 27% to 30% for 2007, based upon testing of the inclusion of the new variable to the model. It is anticipated that further data and model refinement will continue to occur with the *Nowcast* system, which will increase the prediction capability of the model, and ultimately improve protection of the public’s health.

## References

- Anderson, C.W., and Wilde, F.D., eds., September 2005, Turbidity (Version 2.1): U.S. Geological Survey Techniques of Water-Resources Investigations, book 9, chap. A6., section 6.7, available at [http://water.usgs.gov/owq/FieldManual/Chapter6/6.7\\_contents.html](http://water.usgs.gov/owq/FieldManual/Chapter6/6.7_contents.html).
- Francy, D.S., Darner, R.A., 1998, Factors affecting *Escherichia coli* concentrations at Lake Erie public bathing beaches: U.S. Geological Survey Water-Resources Investigations Report 98-4241, 41 p.
- Francy, D.S., Darner, R.A., 2002, Forecasting bacteria levels at bathing beaches in Ohio: U.S. Geological Survey Fact Sheet FS-132-02, 4 p., available at <http://oh.water.usgs.gov/reports/fs-132-02.pdf>
- Francy, D.S., Gifford, A.M., Darner, R.A., 2003, *Escherichia coli* at Ohio bathing beaches—distribution, sources, wastewater indicators, and predictive modeling: U.S. Geological Survey Water-Resources Investigations Report 02-4285, 120 p., available at <http://oh.water.usgs.gov/reports/Abstracts/wrir02-4285.html>
- 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., available at <http://pubs.usgs.gov/sir/2006/5192/>
- Helsel, D.R., and Hirsch, R.M., 2002, Statistical methods in water resources: U.S. Geological Survey Techniques of Water-Resource Investigation, book 4, chap. A3, accessed March 2006 at <http://pubs.er.usgs.gov/pubs/twri/twri04A3>.
- Myers, D.N., and Wilde, F.D., eds., 2003, Biological indicators (3d ed.): U.S. Geological Survey Techniques of Water-Resources Investigations, book 9, chap. A7, accessed March, 2006 at <http://pubs.water.usgs.gov/twri9A7/>
- U.S. Environmental Protection Agency, 2000, Improved enumeration methods for the recreational water quality indicators—Enterococci and *Escherichia coli*: Washington, D.C., Office of Science and Technology, EPA/821/R-97-004, 49 p.