Predictive models have been recognized as tools to improve assessments of recreational water quality at coastal beaches. Multiple linear regression (MLR) models use easily measured environmental and water-quality factors to estimate bacterial indicator concentrations or the probability of exceeding target concentrations. In Ohio, the USGS is working with local and state agencies to develop and test models to predict E. coli concentrations at five Lake Erie beaches—Lakewood (Loran), Huntington Reservation (Bay Village), Edgewater Park (Cleveland), Villa Angela (Cleveland), and Lakeshore Park (Ashtabula). For Huntington, where investigations are further along than at other beaches, 6 years of data (2000–2005) were used to compare the model’s performance to the currently used method for assessing recreational water quality.

**Predictive models are beach specific and can be developed by following the steps listed below.**

1. **Collect daily data on E. coli concentrations and environmental and water-quality variables.** Samples were collected at 1 m depth in the swimming area on weekdays between 8 a.m. and noon. Samples were collected at two or three locations at each beach, and resulting concentrations were averaged for analysis. Samples were analyzed for E. coli by means of the mTEC or modified mTEC membrane-filtration methods (USEPA, 2000 and 2002). Associated environmental and water-quality data were collected by field crews or compiled from a variety of sources. Because a high-quality dataset is important for model development and testing, strict quality-assurance and quality-control procedures were followed.

2. **Conduct exploratory data analysis and compute statistics to determine relations between E. coli concentrations and possible explanatory variables.** Correlation coefficients were calculated and scatterplots constructed to determine the strength of the association between E. coli and possible explanatory variables. Box plots, analysis of variance, and Tukey’s multiple comparison test (Helsel and Hirsch, 1992) were used to examine the distribution of E. coli concentrations as a function of variables grouped by categories.

3. **Select explanatory variables and generate a list of models.** The variables selected were used to generate a list of possible MLR models ordered to minimize the Mallows’ Cp (Mallows, 1973) and maximize the R² statistic. The Cp statistic is a measure of the error variance and bias introduced by not including important variables in a model. The R² is the fraction of the variation in E. coli that can be explained by a combination of explanatory variables.

4. **Assess the models.** Hypothesis tests and regression diagnostics were done to confirm that the model provided the best linear unbiased estimator of E. coli concentrations (Helsel and Hirsch, 1992). These included analysis of partial residual plots, plots of actual versus predicted concentrations, plots of residuals versus predicted concentrations or explanatory variables, and examination of outliers.

5. **Generate model output values, determine the threshold probability, and validate the model.** In earlier studies, prediction intervals were shown to be too wide to estimate E. coli concentrations with reasonable confidence (Francy and Barker, 1998; Francy and others, 2003). Consequently, the probability of exceeding the single-sample bathing-water standard of 235 col/100 mL was used as the model output variable. A threshold probability associated with too great a risk of exceeding the bathing-water standard was determined by maximizing the number of correct responses while minimizing the number of false negative responses. Data were collected during an independent year (2005) to compare the model’s performance with the current method for assessing water quality (using yesterday’s E. coli concentration).

6. **Model output values are then converted into the probability of exceedance.** Probability of exceedance results are then converted into predicted correct responses.

**Conclusions and next steps**

- MLR techniques can be used to generate beach-specific models for coastal recreational waters.
- A model for Huntington based on rainfall, wave height, and turbidity performed better than using yesterday’s E. coli during a validation period.
- An Internet-based “nowcasting” system will be developed for Huntington in 2008 to inform the public of water-quality conditions based on predictive models.
- For the other beaches, the models developed with data from 2004–05 will be validated in 2006.