PREDICTING FECAL COLIFORM BACTERIA LEVELS IN THE CHARLES RIVER, MASSACHUSETTS, USA[1]

Anna Eleria and Richard M. Vogel[2]

ABSTRACT: In Massachusetts, the Charles River Watershed Association conducts a regular water quality monitoring and public notification program in the Charles River Basin during the recreational season to inform users of the river's health. This program has relied on laboratory analyses of river samples for fecal coliform bacteria levels, however, results are not available until at least 24 hours after sampling. To avoid the need for laboratory analyses, ordinary least squares (OLS) and logistic regression models were developed to predict fecal coliform bacteria concentrations and the probabilities of exceeding the Massachusetts secondary contact recreation standard for bacteria based on meteorological conditions and streamflow. The OLS models resulted in adjusted R^2s ranging from 50 to 60 percent. An uncertainty analysis reveals that of the total variability of fecal coliform bacteria concentrations, 45 percent is explained by the OLS regression model, 15 percent is explained by both measurement and space sampling error, and 40 percent is explained by time sampling error. Higher accuracy in future bacteria forecasting models would likely result from reductions in laboratory measurement errors and improved sampling designs.
(KEY TERMS: rivers and streams; nonpoint source pollution; statistical analysis; water quality; recreational management; fecal coliform; bacteria.)

INTRODUCTION

The Charles River Basin is one of the most heavily used recreational areas in the country. Upwards of 20,000 people a day visit the river and the parkland along both banks of the nine-mile (14.5 km) section of river (Figure 1). Historically known for its polluted waters, water quality in the river has improved tremendously over the past 15 years as point sources of pollution from combined sewer overflows and industrial plants have been reduced or treated prior to discharging to the river. Despite these efforts, the health of the river is impaired after a rainstorm because stormwater discharges pollutants, such as pathogens from untreated combined sewage, waterfowl feces, wildlife feces, and domestic pet waste, that have collected on parking lots, streets, driveways, and other impervious surfaces. Pathogens are the pollutant of greatest concern to human health.

Because of the enormous popularity of the river for recreation, there is a need to inform the public of the potential health risks involved with boating on the river. In 1998, the Charles River Watershed Association (CRWA), one of the first watershed organizations in the country, established the Flagging Program, a water quality monitoring and public notification program during the high use recreational season. On a routine basis from June through October, CRWA staff has collected river samples at four sites in the basin. A heuristic approach based primarily on the previous day's fecal coliform bacteria levels, antecedent rainfall conditions and combined sewer overflow activation, enabled CRWA to qualitatively determine river water quality and a color coded flag was hoisted at numerous boating centers located on the banks of the basin. A blue flag implies the river is safe for secondary contact recreation (i.e., boating, kayaking, canoeing) and meets Massachusetts (MA) standard for bacteria while a red flag signifies elevated bacteria levels and the associated potential health risks. Unfortunately, reporting of water quality conditions is often untimely
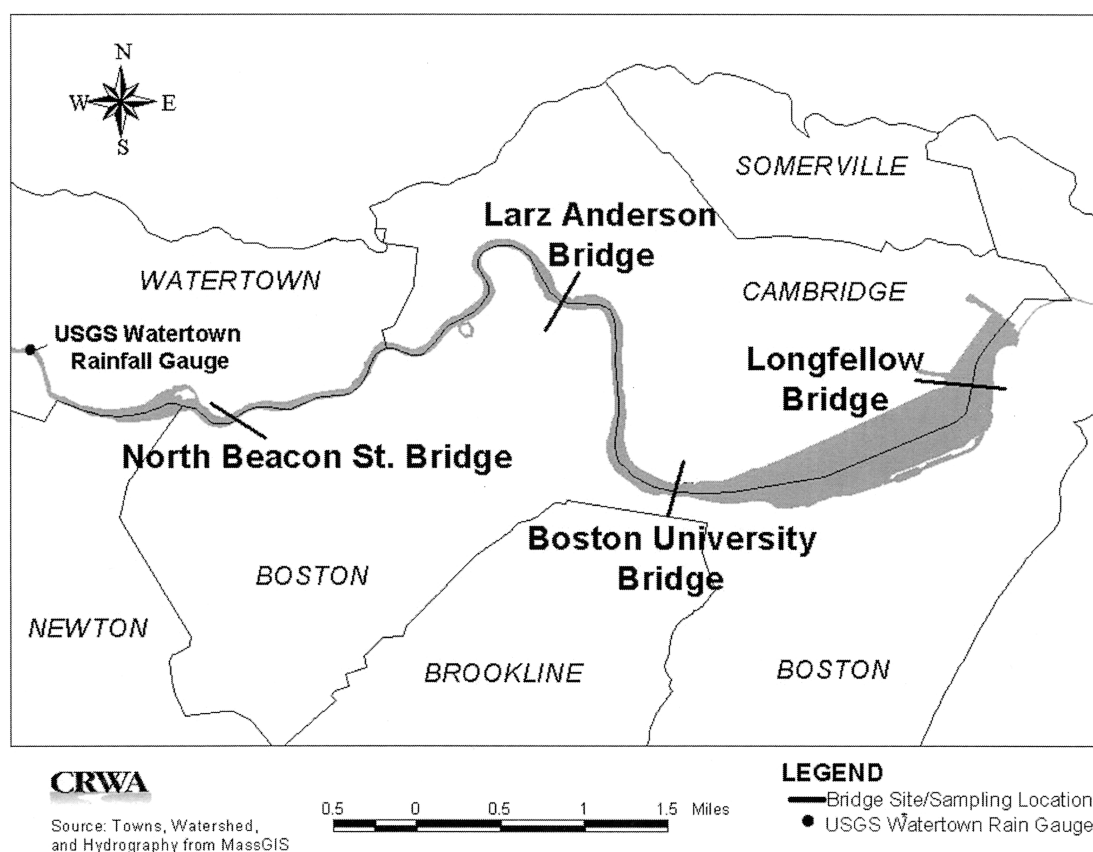

Figure 1. Charles River Basin, Massachusetts, USA.

and inaccurate because the CRWA program cannot monitor the river seven days a week owing to financial and staffing constraints. In addition, time constraints are imposed by the fact that laboratory analysis of fecal coliform bacteria requires a 24-hour incubation period.

With the advent of regular water quality monitoring and public notification programs at water related recreational areas throughout the country, there has been increased interest in developing models to predict water quality conditions without relying on bacteria data and instead correlating precipitation or other easily measured surrogate explanatory variables to bacteria concentrations. The goal was to create prediction models for bacteria at various locations in the Charles River Basin and to eliminate the dependence on water quality sampling. The objectives of the project were to predict instantaneous bacteria levels from meteorological and hydrological conditions using multivariate regression, and to estimate the probability of exceeding the secondary contact recreation standard for bacteria using multivariate logistic regression.

LITERATURE REVIEW

Previous Statistical Studies of Bacteria

The following section reviews some studies that have sought to develop multivariate statistical models to predict bacteria concentrations in rivers. Table 1 summarizes the results of studies by Ferguson et al. (1996), Christensen et al. (2000), Clark and Norris, (2000), Francy et al. (2000, 2002), Crowther et al. (2001), and Rasmussen and Ziegler (2003), all of whom developed multiple linear regression models to relate bacteria concentrations to explanatory variables. Also listed in Table 1 are the explanatory variables as well as the overall goodness of fit associated with the regressions.

Logistic regression models are useful when one's interest is in predicting the probability of the river water quality exceeding a threshold. Smith et al. (2001) employed logistic regression to show that watersheds with large proportions of urban land cover or agriculture on steep slopes had a very high probability of being impaired by pathogens.

TABLE 1. Previous Research on Linear Regression Modeling of Coliform Bacteria.

Citation	Independent Variable	Explanatory Variables	Goodness-of-Fit Statistic
Christensen et al. (2000)	Fecal coliform bacteria	Turbidity and time (month)	0.55 to 0.60 (adjusted R^2)
Clark and Norris (2000)	Fecal coliform bacteria	Discharge, specific conductance, pH, water temperature, dissolved oxygen	0.012 to 0.775 (correlation coefficient)
Crowther et al. (2001)	Fecal coliform bacteria	Daily rainfall	0.10 to 0.50 (R^2)
Ferguson et al. (1996)	Geometric mean fecal coliform bacteria	Rainfall and sewage overflows	0.80 (adjusted R^2)
Francy et al. (2000)	Total coliform bacteria	Dissolved organic carbon, ammonia and organic nitrogen, total phosphorus, nitrate and nitrite, chloride, suspended sediment and specific conductance	0.20 to 0.40 (Spearman's correlation coefficient)
Francy et al. (2002)	E. coli bacteria	Wave height, lake-current direction, turbidity, streamflow of nearby river, rainfall, number of birds on the beach at time of sampling	0.17 to 0.58 (adjusted R^2)
Rasmussen and Ziegler, (2003)	Fecal coliform bacteria	Turbidity	0.16 to 0.79 (R^2)

There is a growing literature that has explored the relationship between land use and bacterial concentrations in coastal estuaries (as opposed to rivers which is the focus here). For example, Mallin et al. (2000) found that a simple regression using percent impervious cover explained 95 percent of the variability in the geometric mean fecal coliform density across watersheds. Similarly, Kelsey et al. (2004) found that stormwater runoff from urban land uses were the primary source of fecal pollution. They found that proximity to areas with septic tanks and rainfall prior to the sampling date are good predictors of fecal pollution.

Factors That Influence Fecal Coliform Bacteria Concentrations

Bacteria levels in a river are a function of initial loading and the disappearance rate which, in turn, is a function of the time or the distance of travel from the source and of other factors including: temperature, salinity, and light intensity. Auer and Niehaus (1993) found that the fecal coliform bacteria death rate is impacted by both solar radiance and water temperature.

Myers et al. (1998) found that the bacteria decay rate was a measure of the die-off of bacteria resulting from ultraviolet light and temperature stress, cell starvation, predation by other bacteria and protozoans, and removal by filter feeders. They also determined that transport, dilution, dispersion, and concentrations of fecal coliform are strongly influenced by the timing, spatial distribution, and amount of rainfall, runoff, and streamflow and that light penetration, which is reduced by turbidity, is the most important factor in determining decay rates. Young and Thackston (1999) found that fecal bacteria counts in urban tributaries were much higher in sewered basins than in nonsewered basins and in general were related to housing density, population, development, percent impervious area, and domestic animal density. Mallin et al. (2000) found that fecal coliform densities were strongly correlated with turbidity (positively) and salinity (negatively).

FECAL COLIFORM BACTERIA: INDICATOR OF HEALTH RISKS

Total coliform bacteria, present in the intestines of warm blooded animals, are excreted in the feces of animals and humans. Fecal coliform bacteria, a subset of the total coliform group, are a more specific indicator of warm blooded animal origin. Since the federal Clean Water Act of 1972, fecal coliform bacteria have been established as an indicator of other

disease causing organisms that may pose a health risk to the public.

Massachusetts has established surface water quality criteria for fecal coliform bacteria that sustain the designated uses of the waterbody (MADEP, 1997). The Charles River, classified as Class B Warm Water, is designated as a habitat for fish, other aquatic life, and wildlife, and is suitable for primary contact recreation, such as swimming and fishing, and secondary contact recreation, such as boating. The MA primary contact recreation standard for fecal coliform bacteria is a geometric mean threshold of 200 colony forming units per 100 milliliters (cfu/100 mL) in any representative sample set. No more than 10 percent of the samples may exceed 400 cfu/100 mL. The MA secondary contact recreation standard is equal to or less than a geometric mean of 1,000 cfu/100 mL in any representative sample set and 10 percent of the samples shall not exceed 2,000 cfu/100 mL.

STUDY DESCRIPTION

Description of Fecal Coliform Bacteria Data

Charles River Watershed Association staff collected river samples at four sites in the Charles River Basin during the high use recreational season; however, the Larz Anderson Bridge site (Figure 1) is the only site considered here. For further information on other monitoring locations, see Eleria (2002). At the Larz Anderson Bridge site, 141 samples were collected over the two-year sampling period from 2000 to 2001 with each sampling period occurring from mid-June through mid-October. Between the hours of 7:00 a.m. and 8:30 a.m., instantaneous grab samples were collected mid-stream between the riverbanks and six inches below the water surface via decontaminated and sterilized buckets. Samples were then transferred into sterile, opaque 125 mL plastic containers. The CRWA followed strict water quality control and assurance measures outlined in the Flagging Program Quality Assurance Project Plan (QAPP) approved by the U.S. Environmental Protection Agency-New England (CRWA, 1999). Duplicate samples were collected for at least 10 percent of the total samples and equipment blanks collected for at least 5 percent of the samples. Immediately after collection, samples were placed on ice and cooled to a temperature of at least 4°C. Finally, samples were delivered within the six-hour holding time for bacteria to a State-certified laboratory for bacteria analyses. Samples were analyzed for fecal coliform bacteria using the membrane filtration method (Method No. 9222-D), described in

Standard Methods for the Examination of Water and Wastewater (APHA, 1998).

Potential Explanatory Variables

The selection of explanatory variables to predict bacteria concentrations is based on several factors: prior knowledge of the explanatory variables' relationships with fecal coliform bacteria in the Charles River Basin, previous findings in the literature concerning factors that influence microbiological organisms, and the ease of obtaining explanatory variables on a daily basis. Table 2 lists the explanatory variables considered. The following subsections outline the explanatory variables and their rationale for inclusion.

Meteorologic Variables.

Rainfall – Stormwater runoff is a significant source of pollutants to the river, which can include bacteria, viruses, and sediment, to which the substrate pollutants attach. Storm rainfall characteristics and conditions prior to the storm are significant factors in the transport and concentration of pollutants in the river. The U. S. Geological Survey (USGS) collected hourly rainfall from June 2000 through October 2001 at Watertown Dam, located several miles upstream of the site. Total volume (inches), duration (hours), and intensity (inches/hour) of rainfall in a storm event were considered. In addition, antecedent storm characteristics, such as time (hours) since storms greater than 0.01 inches (0.25 mm), 0.10 inches (2.5 mm), 0.25 inches (6.4 mm), 0.50 inches (13 mm), and 1.0 inches (25 mm) of rainfall and amount of rainfall (inches) that fell in the previous 24 hours, 48 hours, 72 hours, and 168 hours, were extracted from the hourly precipitation data sets using an unofficial USGS computer program called METCOMP (A.M. Lumb and J.L. Kittle, Jr., 1995, unpublished report).

Seasonality – Due to the flushing effect mechanism associated with bacteria transport (see McDonald and Kay, 1981; and Kelsey *et al.,* 2004), one expects bacteria to vary seasonally. To accommodate the influence of seasonality, the following term was introduced

$$\text{Seasonality term} = \beta_1 \sin(\omega t) + \beta_2 \cos(\omega t) \qquad (1)$$

where ω is $2\pi/365$, t is the Julian day, and β_1 and β_2 are model coefficients to be estimated using multivariate regression.

TABLE 2. Explanatory Variables.

Explanatory Variable	Notation	Range	Units
Volume of rainfall	vol (in)	0 to 1.8	inches
Duration of storm event	dur (hr)	0 to 40	hours
Intensity of storm event	int (in/hr)	0 to 0.17	inches per hour
Time since storm greater than 0.01 inches	>0.01 in	0 to 231	hours
Time since storm greater than 0.10 inches	>0.10 in	0 to 393	hours
Time since storm greater than 0.25 inches	>0.25 in	0 to 535	hours
Time since storm greater than 0.50 inches	>0.50 in	0 to 1099	hours
Time since storm greater than 1.0 inches	>1.0 in	0 to 1282	hours
Amount of rainfall in previous 24 hours	24 hr	0 to 3.29	inches
Amount of rainfall in previous 48 hours	48 hr	0 to 3.29	inches
Amount of rainfall in previous 72 hours	72 hr	0 to3.29	inches
Amount of rainfall in previous 168 hours	168 hr	0 to 3.96	inches
Seasonality	Sine+Cosine	-1 to 0.32	radians
Average daily net radiation	net rad	3.18 to 23.11	langleys
Average daily sky cover	sky cov	0 to 1.0	percent
Average daily wind speed	win spd	5.3 to 17.6	miles per hour
Discharge at time t (day)	Q (t)	24.0 to 744.0	cfs
Discharge at time (day) t-1, t-2, t-3, t-4, t-5	Q(t-1, t-2, t-3, t-4, t-5)		cfs
Natural log discharge (cfs) at time t	LN Q(t)	3.18 to 6.61	NA
Natural log discharge (cfs) at time t-1, t-2, t-3, t-4, t-5	LN Q(t-1, t-2, t-3, t-4, t-5)		NS
Hydrograph dummy variable	HYDRO	0 or 1	NA
Interaction term – bacteria concentration and rainfall over the previous 24 hours	C(t-1)*24 hr	0 to 9.91	NA
Combined sewage overflow dummy variable	COM	0 or 1	NA

Notes: cfs = cubic feet per second; NA = Not applicable.

Net Radiation – Average daily net solar radiation, expressed in langleys, from the National Climate Data Center (NCDC) was considered as an explanatory variable because light intensity is known to affect the die-off rate of fecal coliform bacteria.

Sky Cover – As a measure of light intensity, average daily sky cover was considered. The National Weather Service (NWS) at Logan Airport in Boston records average daily sky cover.

Wind Speed – Average daily wind speed (miles per hour), also measured by NWS, reflects a transport mechanism for bacteria at the water surface.

Hydrologic Variables.

Streamflow – River flow is the primary transport medium of fecal coliform bacteria. Daily streamflow (discharge) measurements at 7:00 a.m. from the USGS Waltham gauge were employed. Bacteria concentrations in the river tend to increase during the

hydrograph rise and decrease during the hydrograph recession due to watershed washoff processes. To account for this phenomenon, known as hysteresis, a dummy variable set to either 1 or 0 was employed to signify either the hydrograph rise or recession, respectively. Bacteria concentrations exhibit persistence from one day to the next, hence lagged bacteria and streamflow data were considered as predictor variables. Because samples were not collected on weekends and holidays, the data set reduced to 78 observations with inclusion of this explanatory variable. In addition, because it was postulated that there is a strong relationship between bacteria data and previous rainfall, interaction terms between lagged bacteria data and rainfall in the previous 24, 48, and 72 hours were considered by multiplying the lagged bacteria data with each antecedent rainfall characteristic.

Combined Sewer Overflow Activation – Boston and Cambridge are served by combined sewer systems, where both wastewater and stormwater flow in the

same conveyance pipes to the nearby wastewater treatment plant. When it rains heavily, the hydraulic capacities of the combined sewer pipes are exceeded and a portion of the untreated combined sewage discharges to the Charles River Basin, raising bacteria levels in the river. Combined sewer overflow activation was included using a dummy variable.

EXPLORATORY DATA ANALYSES

In this section the stochastic and probabilistic structure of the daily streamflow and bacteria data are examined. The measures of central tendency of both bacteria and streamflow vary dramatically as is shown in Table 3. Figures 2(a) and 2(b) illustrate lognormal probability plots for the fecal coliform bacteria and discharge (Helsel and Hirsch, 1992). In both cases, a lognomal distribution provides a first approximation to the probability distribution of both bacteria and streamflow and as a result, logarithmic transformations are employed in all future analyses.

TABLE 3. Statistics of Fecal Coliform Bacteria at Larz Anderson Bridge and Discharge at USGS Waltham Gauge.

	Fecal Coliform Bacteria (cfu/100 mL)	Discharge (cfs)
Sample Size	141	254
Mean	910	147
Median	100	100
Geometric Mean	110	101
Harmonic Mean	38	77
Standard Deviation	6,640	146
Minimum	2.5	24
First Quartile	38.8	55
Third Quartile	265	171
Maximum	79,000	747

The decay rates of bacteria, combined with the natural persistence associated with streamflow, result in bacteria concentrations that exhibit memory or autocorrelation. Because of gaps in the bacteria data set, only one-day lags were considered. Figure 3 illustrates a plot of bacteria concentrations at time t versus concentrations the previous day (t-1). With a few exceptions, high concentrations of bacteria on one day tend to be followed by high concentrations the next day. The lag-1 serial correlation for bacteria concentrations is 0.51.
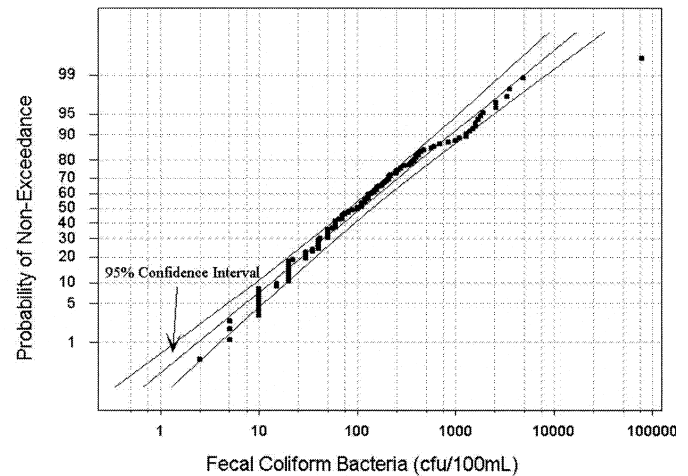
(a)
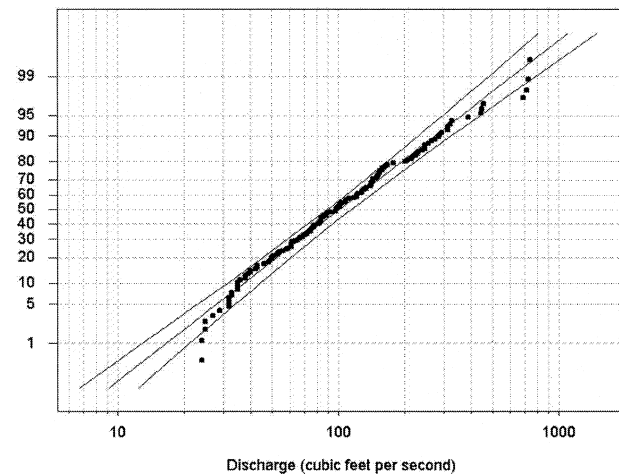
(b)

Figure 2. Lognormal Probability Plot of (a) Fecal Coliform Bacteria and (b) Discharge at USGS Waltham Gauge.
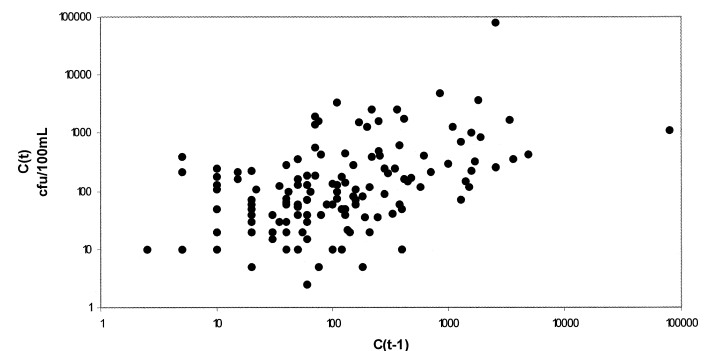

Figure 3. Persistence of Fecal Coliform Bacteria.

Figure 4 compares the autocorrelation function of the observed daily discharges with a Markov process. The autocorrelation coefficient measures the strength of association between the daily streamflow at time t and t-1. The lag-1 and lag-2 correlation coefficients of observed discharge are $r_1 = 0.96$ and $r_2 = 0.90$. For a Markov or first-order autoregressive process, one expects $r_2 = r_1^2 = 0.96^2 = 0.92^2$, which is close to the observed lag-2 discharge value of $r_2 = 0.90$. This demonstrates that the discharges are roughly Markov, hence one only needs to consider yesterday's streamflow to approximate the complete memory of daily streamflow.
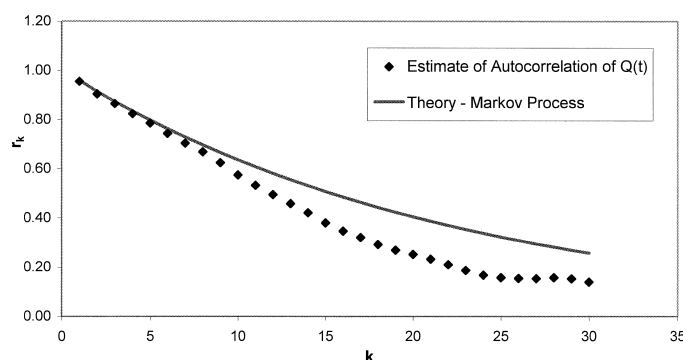

Figure 4. Correlogram of Daily Discharge at USGS Waltham Gauge.

REGRESSION METHODS

Multivariate Linear Regression

Multivariate linear regression models of the following form were fit to the daily bacteria concentrations

$$y_j = \beta_0 + \beta_{11}x_{11} + \beta_{12}x_{12} + \dots + \beta_{i,j}x_{i,j} + \varepsilon_i \qquad (2)$$

where y_j is the natural logarithm of daily fecal coliform bacteria concentration (cfu/100 ml) on day j; $\beta_{i,j}$ is the slope coefficient of explanatory variable x_i; $x_{i,j}$ is the ith explanatory variable on day j; and ε_i is the model error or residual on day j.

The OLS method in Minitab® (Minitab, Inc., 2000), was used to estimate the model coefficients and statistical tests were performed to ensure that the model residuals were approximately normally distributed and that each explanatory variable increased the goodness-of-fit in a statistically significant fashion. In addition, variance inflation factors (VIF) were estimated to ensure against excess multicollinearity among the explanatory variables. A total of six model types were considered and ranged from simple to complex. The first model considered only antecedent

rainfall characteristics. For Model 2, the hydrological variables were added to Model 1, while for Model 3, the other meteorological explanatory variables, including seasonality, net radiation, cloud cover, and wind speed, and the combined sewer overflow (CSO) activation variable were added to Model 2. Models 4 through 6, similar to Models 1 through 3, respectively, also included lagged bacteria concentrations. As a result, all explanatory variables were considered in Model 6. Unusual combinations of explanatory variables can strongly influence the regression model and reduce its predictive power, hence they were identified using standard influence statistics, such as Cooks' D (see Helsel and Hirsch, 1992), and subsequently removed. Stepwise multivariate linear regression was applied to select combinations of the independent variables. The resulting models were evaluated using prediction type goodness-of-fit metrics, such as prediction R^2, and the predicted residual sums of squares (PRESS) statistic as well as various graphical diagnostic evaluations of the behavior of the model residuals outlined by Helsel and Hirsch, (1992).

Multivariate Logistic Regression

Multivariate logistic regression is useful for determining the relationship between categorical or discrete model responses, in this case, the probability of bacteria levels exceeding the State secondary contact recreation (boating) standard, and a variety of predictor variables (see Chapter 12 in Helsel, 2005). The values of the bacteria response above or below the threshold are designated using binary variables. Logistic regression transforms estimated probabilities into a continuous response variable. The transformed response is predicted from one or more explanatory variables, and subsequently retransformed back to a value between 1 and 0. For this project, a value of 1 signifies that there is a greater than 50 percent probability of the river exceeding the secondary contact recreation standard for bacteria, while a 0 implies there is a less than 50 percent chance of exceeding the secondary contact recreation standard.

The odds ratio is defined as the ratio of the probability of obtaining a 1 divided by the probability of obtaining a 0.

$$\text{odds ratio} = \frac{p}{1-p} \qquad (3)$$

where p is the probability of a response of 1.

The natural log of odds ratio (termed the logit) transforms a variable constrained between 0 and 1 into a continuous and unbounded variable. To

estimate logistic regression, the logit is modeled as a linear function of one or more explanatory variables so that

$$\log\left[\frac{p}{1-p}\right] = \beta_o + \beta X \qquad (4)$$

where β_0 is the intercept and β is the slope coefficients for each explanatory variable.

Exponentiation of Equation (4) leads to

$$\left[\frac{p}{1-p}\right] = \exp[\beta_o + \beta X] \qquad (5)$$

which can be rewritten as

$$p = \frac{\exp[\beta_o + \beta X]}{1 + \exp[\beta_o + \beta X]} \qquad (6)$$

where p is the probability of the river exceeding the secondary contact recreation standard for bacteria. Maximum likelihood estimates of the model parameters in Equation (6) were obtained using Minitab®. Akaike's Information Criteria was employed to evaluate the goodness-of-fit of alternative logistic models.

Unlike multivariate linear regression methods, a stepwise approach for selecting explanatory variables based on goodness-of-fit is currently not available for logistic regression. Instead, the stepwise approach of multivariate linear regression models served as a screening tool for suitable explanatory variables for the logistic regression models.

MODELING RESULTS

Multivariate Linear Regression Results

Table 4 summarizes the results of the multivariate linear regression models and lists the explanatory variables, their corresponding coefficient values, and goodness-of-fit statistics. In terms of overall goodness-of-fit, the 'best' model at Larz Anderson Bridge was Model 6, which had an adjusted R^2 of 60.4 percent and four significant predictor variables: lag-1 bacteria concentration, the interaction term between the lag-1 bacteria concentration and the amount of rainfall in the previous 24 hours, time since rainfall greater than 0.10 inches (2.5 mm), and average daily wind speed. The standard error and PRESS statistic for Model 6 were the lowest among all the models evaluated. Figure 5 illustrates the modest linear relationship

between observations and predictions of bacteria concentrations at Larz Anderson Bridge.

The best model that did not include lag-1 bacteria concentrations was Model 3, the model category that considered all meteorologic and hydrologic variables. This model had the second highest adjusted R^2, 56 percent; yet, it also required eight explanatory variables. The significant variables included average rainfall intensity, amount of rainfall in the previous 168 hours, time since rainfall greater than 0.10 inches (2.5 mm), natural log of discharge, natural log of lag-4 discharge, hydrograph, average wind speed, and CSO activation in the previous 24 hours.

Split Sample Validation Experiment. A split sample validation experiment was performed to test the predictive power of Model 4, which included lag-1 bacteria data as an explanatory variable. Although this model did not have the highest adjusted R^2, this model was selected for validation because of the ease of applying this model by CRWA. The procedures for the split sample validation experiment were: (1) the multivariate regression model was fit to the first half of the data; (2) the fitted multivariate regression model from Step 1 was used with the second half of the data to compute predicted values; and (3) the performances of the model over the calibration and validation portions of the dataset were compared in Table 5 and Figure 6.

The multivariate regression equation for Model 4 based on 78 observations was

$$\ln(C_t) = 2.16 + 0.39C_{t-1} + 2.89*24 \text{ hr} + 0.61*168 \text{ hr} \qquad (7)$$

where $\ln(C_t)$ is the natural log of fecal coliform bacteria concentration; C_{t-1} is the previous day's bacteria concentration (cfu/100 mL); 24 hr is the amount of rainfall in the previous 24 hours (in); and 168 hr is the amount of rainfall in the previous 168 hours (in).

Water quality at Larz Anderson Bridge during this study period was fairly good; only eight out of 78 observations exceeded the secondary contact recreation standard for bacteria. Table 5 summarizes the number of correct and incorrect predictions of meeting or exceeding the standard at Larz Anderson Bridge over the validation period. The number in the parentheses equals the percent of the time the model correctly or incorrectly predicted when the river was safe or unsafe for boating. The predictive power of the model was very good when the river was safe for boating, leading to correct predictions 97 percent of the time. On the other hand, when observed bacteria concentrations were greater than 1,000 cfu/100 mL, the model predicted those violations with less accuracy (64 percent of the time). Figure 6 compares the time

TABLE 4. Larz Anderson Bridge Multivariate Linear Regression Models.

Model	Sample Size of Bacteria Data Set	Explanatory Variables	b_0 Constant	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	Std. Error	Adj. R^2	PRESS
1	141	Antecedent Rainfall Data	4.21[1] (0.000)	48 hr^2 0.01 (0.001)	168 hr 0.85 (0.000)	>0.01 in -0.0052 (0.009)						1.21	46.11	222.32
2	141	Antecedent Rainfall Data and Hydrologic Variables	2.23 (0.000)	48 hr 0.72 (0.022)	168 hr 0.42 (0.026)	>0.10 in -0.032 (0.032)	LN Q(t) 1.23 (0.001)	LN Q(t-4) -0.72 (0.018)				1.16	50.55	207.22
3	141	Antecedent Rainfall Data, Hydrologic Data, and Remaining Meteorologic Variables	1.23 (0.000)	168 hr 0.38 (0.036)	avgint -9.0 (0.007)	>0.10 in -0.0033 (0.004)	LN Q(t) 1.49 (0.000)	LN Q(t-4) -0.92 (0.002)	Hydrograph 0.47 (0.000)	Wind 0.084 (0.027)	CSO <24 hr 1.11 (0.024)	1.10	55.58	183.74
4	78	Antecedent Rainfall Data and Lag-1 Bacteria	2.16 (0.000)	C(t-1) 0.39 (0.001)	24 hr 2.89 (0.000)	168 hr 0.61 (0.023)						1.13	53.84	106.66
5	78	Antecedent Rainfall Data, Hydrologic Variables and Lag-1 Bacteria	2.16 (0.000)	C(t-1) 0.39 (0.001)	24 hr 2.89 (0.000)	168 hr 0.61 (0.023)						1.13	53.84	106.66
6	78	Antecedent Rainfall Data, Hydrologic Data, Remaining Meteorologic Variables, and Lag-1 Bacteria	1.53 (0.022)	C(t-1) 0.39 (0.000)	C(t-1) *24 hr 0.527 (0.000)	>0.10 in -0.0067 (0.001)	Wind 0.160 (0.002)					1.05	60.40	89.57

[1]The first value is the coefficient value for the constant and the value in the parentheses is the p-value of the constant.
[2]The explanatory variables are presented with their corresponding coefficient value and p-value in parentheses

series of observed bacteria concentrations (diamonds) with the time series of bacteria concentrations predicted by Model 4 (the dotted black line) for 78 observations. In general, the model performed well in predicting concentrations between 100 cfu/100 mL and 1,000 cfu/100 mL but tended to overestimate the low concentrations and underestimate the high concentrations.
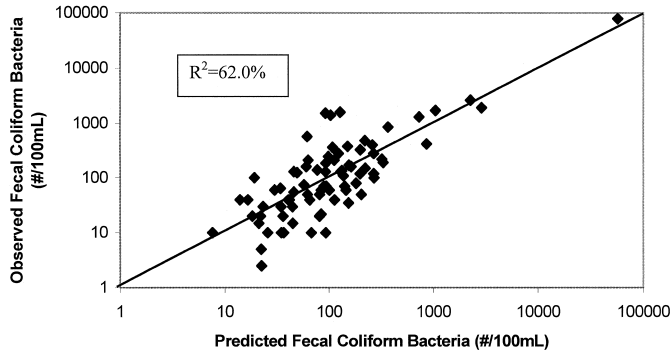

Figure 5. Observed Versus Predicted Fecal Coliform Bacteria Concentrations at Larz Anderson Bridge.

Additional Validation Experiments. Models 4, 5, and 6 required observations of lagged bacteria concentrations, yet such information is not always available in practice. These models are not nearly as accurate as they appear because if the program is unable to collect bacteria samples, modeled estimates of lagged concentrations are needed for their application and such modeled estimates contain significant additional model error. Therefore, an additional experiment was conducted to verify the accuracy of these models when lagged bacteria concentrations must also be estimated from a regression model. First, a single bacteria observation was used in Model 4 to obtain a regression estimate of the next day's bacteria concentration. From that day on, the values of lagged bacteria became the regression estimates obtained from Model 4. The results of this experiment in Table 6 showed that the use of estimates of bacteria concentrations in Model 4 led to lower prediction accuracy. The model was 98 percent accurate at predicting when the river met the secondary contact recreation standard, while the model was only accurate 44 percent of the time, for predicting violations to the standard (Table 6), which is worse than a simple guess.

TABLE 5. Split Sample Experiment:
Model 4 of Larz Anderson Bridge.

		Observations	
		Met Boating Standard	**Exceeded Boating Standard**
Predictions	Met Boating Standard	68 (97%)	3 (37%)
	Exceeded Boating Standard	2 (3%)	5 (63%)

In addition, recognizing that the original hope was to eliminate dependence on indicator bacteriological
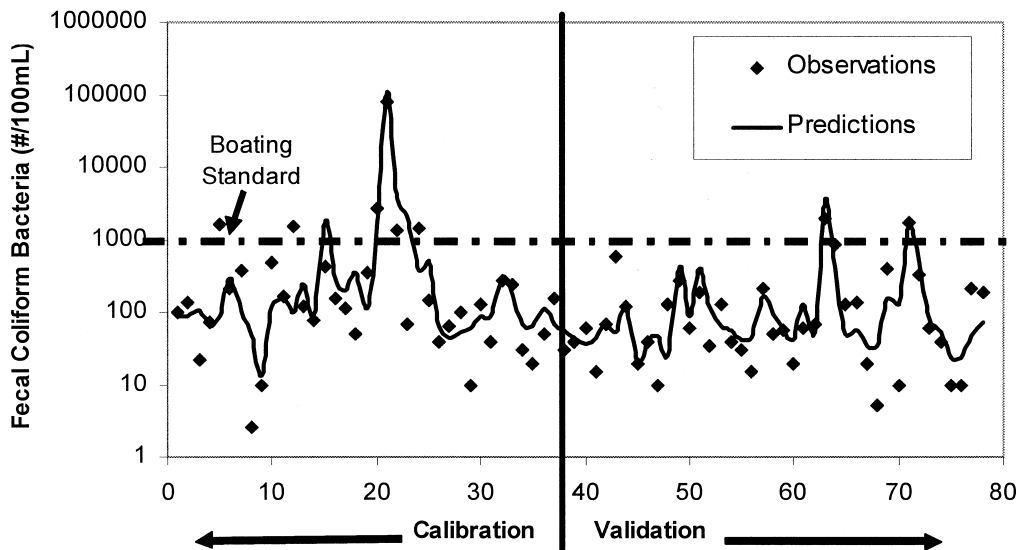

Figure 6. Larz Anderson Bridge Split Sample Model Experiment: Observed Bacteria Concentrations Versus Predicted Bacteria Concentrations.

monitoring and that practitioners may need a model without a lagged bacteria explanatory variable, the accuracy of Model 3 was tested. This model had the second highest adjusted R^2 and did not rely on the persistence structure of bacteria to estimate bacteria levels. Model 3's predictions were similar to the previous experiment. Ninety-eight percent of the time, the model predicted when the river met the secondary contact recreation standard; however, its accuracy decreased to 44 percent for predicting when the river did not meet the standard.

TABLE 6. Model 4 Experiment With Modeled Estimates of Lagged Bacteria Concentrations.

		Observations	
		Met Boating Standard	Exceeded Boating Standard
Predictions	Met Boating Standard	123 (98%)	9 (56%)
	Exceeded Boating Standard	2 (2%)	7 (44%)

Multivariate Logistic Modeling Results

Similar to the 'best' linear regression model, the 'best' logistic model for Larz Anderson Bridge also included the lag-1 bacteria concentration variable. The probability of the river exceeding the secondary contact recreation standard at Larz Anderson Bridge was estimated from the following equation

$$P = \frac{\exp(\theta)}{1 + \exp(\theta)} \qquad (8)$$

where P is the probability of the river exceeding the secondary contact recreation standard at Larz Anderson Bridge; θ = -4.451 + 0.828* ln(C_{t-1}) * 24 hr + (1.592 * 168 hr); ln(C_{t-1}) * 24 hr is the interaction term between the natural log of lag-1 bacteria concentration and rainfall in the previous 24 hours; and 168 hr is the rainfall (inches) in the previous 168 hours.

The accuracy of the logistic model in predicting when the river meets or exceeds the secondary contact recreation standard is presented in Table 7. The model had very accurate predictions (97 percent) when the river met the secondary contact recreation standard, but the model was less successful (64 percent) when violations to the standard occurred.

TABLE 7. Number of Observations Versus Number of Predictions: 'Best' Logistic Regression Model at Larz Anderson Bridge.

		Observations	
		Met Boating Standard	Exceeded Boating Standard
Predictions	Met Boating Standard	68 (97%)	3 (36%)
	Exceeded Boating Standard	2 (3%)	5 (64%)

Discussion and Comparison of Linear Regression and Logistic Regression

Both regression approaches had excellent success in predicting when the river met the secondary contact recreation standard at Larz Anderson Bridge, yet only fair to poor success in predicting violations of the standard. The 'best' models of both the linear and logistic regression approaches predicted correctly at least 95 percent of the time when the river met the secondary contact recreation standard. The linear regression models correctly predicted violations between 44 and 63 percent of the time and logistic regression models correctly predicted violations 64 percent of the time. The explanatory powers of both models are much higher during dry weather events (e.g., rainfall less than 0.10 inches (2.5 mm) in the previous 72 hours) than wet weather events.

Lower accuracies in the models' predictive capabilities of the higher bacteria concentrations (>1,000 cfu/100 mL) were anticipated because of fewer observations of elevated bacteria concentrations than the lower bacteria concentrations. In addition, the observed bacteria concentrations may not be representative of bacteria concentrations in the river at the time of sampling because of the significant uncertainties discussed in the next section.

IMPACT OF UNCERTAINTY IN MODEL, MEASUREMENTS, AND REPRESENTATIVENESS OF DATA

To better understand the challenges of predicting instantaneous bacteria concentrations, this section explores the uncertainty associated with bacteria measurements and the ability of those measurements to reflect the true bacteria concentrations in the river at the time of sampling. Ideally, samples would be collected at several locations across a river cross section and in six-hour intervals over a 24-hour period

(absent of rain) to capture the spatial and temporal variability associated with bacteria measurements. Unfortunately, the available bacteria data only captured an instantaneous snapshot of bacteria concentrations at a specific location in the river, hence the data may not be representative of true water quality conditions of the entire river cross section during the day of sampling. In addition, bacteria measurements are known to contain significant laboratory measurement error. The overall bacteria modeling problem can be written as

$$y_i = f(x_i) + \varepsilon_i \qquad (9)$$

where y_i is the natural logarithm of the ith bacteria concentration observation; $f(x_i)$ is the multivariate linear regression model where x_i denotes logarithm of the ith independent variable; and ε_i is the ith error realization in log space.

The goal of the regression model is to describe most of the variability in y_i, with the errors describing only a small portion of that variability. However, in this application, the error terms play a central role. The overall error may be disaggregated into the following sources

$$\varepsilon_i = \text{model error} + \text{measurement error}$$
$$+ \text{spatial and temporal representation error}$$
$$(10)$$

Model error characterizes one's inability to select the correct form of the multivariate regression model and the correct explanatory variables to include in that model. Measurement error represents the error associated with laboratory measurements of bacteria levels. Spatial and temporal representation error reflects the inability of a single instantaneous bacteria measurement to represent the behavior of the bacteria concentrations over the full spatial (river cross section) and temporal range (daily) considered. In the following section, an attempt is made to quantify these three sources of error.

Error Sources and Their Variability

Since the regression model is fit in log space, it is instructive to compare the variability of the various terms in Equation (10) by simply comparing their variances. This is analogous, but not equivalent, to comparing the coefficient of variations of the various terms, because for a lognormal variable

$$C_v = \sqrt{\exp\left(\sigma_y^2\right) - 1} \qquad (11)$$

where $\sigma_y^2 = 2.74$ is the variance of the natural logarithms of the bacteria concentrations and $C_v = 3.8$ is the coefficient of variation of bacteria in log space.

Assuming that all sources of error are independent, Equation (9) yields

$$Var(y) = \sigma_y^2 = Var(f(x)) + Var(\varepsilon) \qquad (12)$$

where $Var(y)$ is the variance of logs of bacteria; $Var(f(x))$ is the variance of logs of model error; and $Var(\varepsilon)$ is the variance of log of three additional sources of variability resulted from laboratory measurement, spatial sampling, and temporal sampling errors. The overall variability, $Var(y) = 2.74$, results from at least three sources of error in addition to the variability explained by the model: laboratory measurement error, spatial sampling error, and temporal sampling error. Note that in most situations (when developing prediction models for other parameters), the primary source of error is model error, and even that error can be quite small. When modeling instantaneous bacteria concentrations, these additional error sources served to further confound the ability to reproduce observed bacteria observations.

The variance of the logarithm of the model error term is equal to the variance of the log space residuals for the regression model, which was equal to $Var(f(x)) = 1.23$. Clearly, Model 4 was able to explain a good portion of the original variability of the bacteria concentrations or $100(1.23/2.74)$ equals 44.9 percent of that variability. The question addressed here is whether the remaining 55.1 percent variability can be explained by the additional sources of error described above in Equation (10).

Assuming independence among the individual errors in Equation (10), the standard deviation of the total errors, ε_i, is

$$\sigma_\varepsilon = \sqrt{\sigma_m^2 + \sigma_s^2 + \sigma_t^2} \qquad (13)$$

where σ_ε is the standard error associated with bacteria observations, σ_m is the standard error of laboratory duplicate errors, σ_s is the standard error of bacteria data collected over space; and σ_t is the standard error of bacteria data over time.

Laboratory Measurement Error. Because of the known variability associated with laboratory measurement of water quality constituents in river samples, laboratories routinely conduct duplicate measurements to quantify this variability and ensure that an estimated precision criterion is met. The CRWA laboratory derives yearly the precision

criterion, the index for comparison of duplicate bacteriological data, according to the Standard Methods for the Examination of Water and Wastewater (APHA, 1998). The CRWA laboratory conducted duplicate analyses of 130 observations from March 2000 to December 2001. The duplicate error, or laboratory measurement error, is the difference between the first sample and the duplicate measurement. The mean, m_m, and standard deviation, s_m, of the laboratory measurement error, in real space, equaled 11.9 cfu/100 mL and 380, respectively. The standard error of the mean was 33.3, hence the estimated mean measurement error is not significantly different from zero as expected.

Spatial Variability of Fecal Coliform Bacteria. The spatial variability of bacteria concentrations in the vertical and horizontal cross sections of the river was not captured in this monitoring project. Samples were collected at the same location, at the same time of day, representing only an instantaneous picture of the river's health. As part of another study, the USGS quantified the spatial variability of bacteria in river cross sections during two separate storm events in July 2002, which was the only spatial data available for this area and time period. The USGS collected three bacteria samples each at several points: the middle, near the right bank and near the left bank, and across the horizontal cross section. One USGS site corresponded to the Larz Anderson Bridge monitoring location. The spatial error was estimated by calculating the difference between the middle of the river sample and the right or left bank sample. The mean and standard deviation of the spatial errors were equal to $\mu_s = 29.2$ cfu/100 mL and $\sigma_s = 64.3$, respectively. In addition, the standard error of the mean spatial error was equal to 18.6, hence the mean spatial error is not significantly different from zero.

Temporal Variability of Fecal Coliform Bacteria. Temporal conditions may vary substantially between daily bacteria samples due to the natural die-off of bacteria, additional inputs of bacteria to the river, and/or the transport of pollutants within a 24-hour period. High frequency data over time has only been collected in the basin during wet weather events to determine the response of the river to various storm volumes, durations, and intensities. Therefore, quantification of the temporal variability of bacteria measurements is not possible in the following analysis.

Summary Comparison of Sources of Uncertainty

Recall that the mean bacteria concentration is 910 cfu/100 mL. From Equation (13), $\sigma_\varepsilon = \sqrt{\sigma_m^2 + \sigma_s^2}$

$= \sqrt{380^2 + 64.3^2} = 385,$ is obtained. Hence the coefficient of variation of the measurement and space sampling errors in real space are $C_v(\varepsilon) = 385/910 = 0.423,$ which corresponds to a log space variance of $Var(\varepsilon) = 0.406$. Recall from Equation (10) that $Var(y) = Var(f(x)) + Var(\varepsilon),$ which becomes 2.74 = 1.23 + 0.406 + x, where $Var(y) = 2.74, Var(f(x)) = 1.23$ and $Var(\varepsilon) = 0.406,$ and x = 1.1 is the remaining variance explained by time sampling error, which was ignored in the above analysis. Thus, of the overall variability of the observed bacteria concentrations, 45 percent is explained by the model, 15 percent is explained by both measurement and space sampling error, and, apparently, 40 percent is explained by time sampling error.

CONCLUSIONS AND FUTURE WORK

This project was an effort to predict bacteria concentrations in the Charles River Basin using easily measured and readily available explanatory variables. Multivariate regression models were developed between fecal coliform concentrations and a variety of hydrologic, environmental, and meteorologic variables. The linear regression models employing meteorologic and hydrologic explanatory variables (Models 1 through 3) could only moderately explain the observed variance in bacteria (adjusted R^2 values ranged from 46 percent to 56 percent). Models that included the observed persistence structure of bacteria (Models 4 through 6) led to slight improvements with the highest adjusted R^2 equal to 60 percent for Model 6.

Although Model 6 had the highest adjusted R^2, it was not preferred for application in the CRWA Flagging Program because it requires a large number of explanatory variables that originate from different data sources. Model 4 was the preferred linear regression model because it requires explanatory variables from only two different data sources. During a split sample validation experiment, the predictive capability of Model 4 was excellent for concentrations below the secondary contact recreation standard of 1,000 cfu/100 mL but only fair for concentrations greater than 1,000 cfu/100 mL. The regression model accurately predicted when the river met the bacteria

secondary contact recreation standard over 90 percent of the time. However, the percentage decreased to 63 percent when predicting violations to the secondary contact recreation standard.

Logistic regression models were also developed to predict the probability of the river being safe or not safe for secondary contact recreation, which is of greater concern to recreational users than the actual bacteria levels of the river. The best logistic regression model showed no improvements to predictions. It accurately predicted when the river met the secondary contact recreation standard over 90 percent of the time and when the river exceeded the standard about 60 percent of the time.

Finally, the impact of additional sources of variability on the accuracy of model results was explored. Of the total variability of fecal coliform bacteria concentrations, 45 percent is explained by the ordinary least squares regression model, 15 percent is explained by both measurement and space sampling error, and apparently 40 percent is explained by time sampling error. Clearly, future improvements to such models are likely to come from reductions in both the time and space sampling errors, which are under the modeler's control.

The relationships developed here, although only modestly successful, are an improvement over models developed previously. For CRWA's Flagging Program, either model with or without lagged bacteria data as an explanatory variable is used on a daily basis, providing a useful *quantitative* tool for predicting the suitability of the river for secondary contact recreation. The application of this statistical approach with one or more of the same explanatory variables used here has already been tested in other freshwater recreational rivers in the country.

LITERATURE CITED

APHA (American Public Health Association), 1998. Standard Methods for the Examination of Water and Wastewater (20th Edition). American Public Health Association, American Water Works Association, and the Water Environment Federation, Washington, D.C.

Auer, M.T. and S.L. Niehaus, 1993. Modeling Fecal Coliform Bacteria. I. Field and Laboratory Determination of Loss Kinetics. Water Resources 27(4):693-701.

CRWA (Charles River Watershed Association), 1999. Charles River Flagging Program Quality Assurance Project Plan. Newton, Massachusetts.

Christensen, V., X. Jian, and A. Ziegler, 2000. Regression Analysis and Real-Time Water Quality Monitoring to Estimate Constituent Concentrations, Loads and Yields in the Little Arkansas River, South Central Kansas, 1995-1999. USGS Water Resources Investigations Report 00-4126, Lawrence, Kansas.

Clark, M.L. and J.R. Norris, 2000. Occurrence of Fecal Coliform Bacteria in Selected Streams in Wyoming, 1990-99. USGS Water Resources Investigations Report 00-4198, Cheyenne, Wyoming.

Crowther, J., D. Kay, and M. Wyer, 2001. Relationships Between Water Quality and Environmental Conditions in Coastal Recreational Waters: The Fylde Coast, United Kingdom. Water Research 35(17):4029-4038.

Eleria, A.L., 2002. Forecasting Fecal Coliform Bacteria in the Charles River Basin. Master's Thesis, Tufts University, Medford, Massachusetts.

Ferguson, C.M., B.G. Coote, N.J. Ashbolt, and I.M. Stevenson, 1996. Relationships Between Indicators, Pathogens and Water Quality in an Estuarine System. Water Research 30(9):2045-2054.

Francy, D.S., A.M. Gifford, and R.A. Darner, 2002. *Escherichia coli* at Ohio Bathing Beaches – Distribution, Sources, Wastewater Indicators, and Predictive Modeling. U.S. Geological Survey Water-Resources Investigations Report 02-4285, Columbus, Ohio.

Francy, D.S., D.R. Helsel, and R.A. Nally, 2000. Occurrence and Distribution of Microbiological Indicators in Groundwater and Streamwater. Water Environment Research 72(2):152-161.

Helsel, D.R., 2005. Nondetects and Data Analysis, John Wiley and Sons, Hoboken, New Jersey, 250 pp.

Helsel, D.R. and R.M. Hirsch. 1992. Statistical Methods in Water Resources. Elsevier, New York, New York.

Hirsch, R.M., 1988. Statistical Methods and Sampling Design for Estimating Step Trends in Surface-Water Quality. Water Resources Bulletin 24(3):493-503.

Kelsey, H., D.E. Porter, G. Scott, M. Neet, and D. White, 2004. Using Geographic Information Systems and Regression Analysis to Evaluate Relationships Between Land Use and Fecal Coliform Bacterial Pollution. Journal of Experimental Marine Biology and Ecology 298:197-209.

Mallin, M.A., K.E. Williams, E.G. Esham, and R.P. Low, 2000. Effect of Human Development on Bacteriological Water Quality in Coastal Watersheds. Ecological Applications 10(4):1047-1056.

MA DEP (Massachusetts Department of Environmental Protection), 1997. Massachusetts Surface Water Quality Standards. Massachusetts Department of Environmental Protection, Division of Water Pollution Control, Technical Services Branch, Westborough, Massachusetts (Revision of 314 CMR 4.00, effective May 30, 1997).

McDonald A. and D. Kay, 1981. Enteric Bacterial Concentrations in Reservoir Feeder Streams: Baseflow Characteristics and Response to Hydrograph Events. Water Research 15:961-968.

Minitab Inc., 2000. MINITAB® Statistical Software Release 13 Windows 95/98, NT. Minitab Inc., State College, Pennsylvania.

Myers, D.N., G.F. Koltun, and D.S. Francy, 1998. Effects of Hydrologic, Biological, and Environmental Processes on Sources and Concentrations of Fecal Bacteria in the Cuyahoga River, With Implications for Management of Recreational Waters in Summit and Cuyahoga Counties, Ohio. U.S. Geological Survey Water Resources Investigations Report 98-4089, 45 pp., Columbus, Ohio.

Rasmussen, P.P. and A.C. Ziegler, 2003. Comparison and Continuous Estimates of Fecal Coliform and *Escherichia Coli* Bacteria in Selected Kansas Streams, May 1999 Through April 2002. U.S. Geological Survey Water Resources Investigations Report 03-4056, 80 pp., Lawrence, Kansas.

Smith, J.H., J.D. Wickham, D. Norton, T.G. Wade, and K.B. Jones, 2001. Utilization of Landscape Indicators to Model Potential Pathogen Impaired Waters. Journal of the American Water Resources Association (JAWRA) 37(4):805-814.

Young, K.D. and E.L. Thackston, 1999, Housing Density and Bacterial Loading in Urban Streams, Journal of Environmental Engineering 125(12):1177-1180.