Combining Real-Time Bacteria Models and Uncertainty Analysis for Establishing Health Advisories for Recreational Waters

Matthew G. Heberger[1]; John L. Durant[1]; Kimberly A. Oriel[3]; Paul H. Kirshen[4]; and Lee Minardi[5]

Abstract: Tools are needed to allow accurate and timely prediction of water quality so that recreational users can make informed decisions about the safety of water, and beach managers can post updated advisories. In this paper, we describe the development of a health advisory system based on the probability that real-time estimates of sewage indicator bacteria levels exceed water quality standards. Real-time estimates of bacteria concentrations were made using multivariate linear regression models and real-time input data. Probability distribution functions based on the model results and their associated variances were used to determine the probability that predicted bacteria levels exceed water quality standards. The models were developed for the Mystic River watershed, an urban watershed near Boston, using *Enterococcus* bacteria data measured during the summers of 2002 and 2003. The linear regression models had adjusted-R^2 values of 0.55–0.82 for two river sites, and 0.42 for both a beach on a lake and a boathouse near a dam. Independent variables with predictive power included precipitation and the time since the last rainfall. The probabilistic models for the beach and the boathouse sites correctly predicted water quality exceedances and nonexceedances with >85% accuracy.

DOI: 10.1061/(ASCE)0733-9496(2008)134:1(73)

CE Database subject headings: Water quality; Bacteria; Recreation; Uncertainty principles; Public health.

Introduction

Motivation and Objectives

Recreational waters in urban areas frequently receive high loadings of fecal material from sanitary sewer overflows, combined sewer overflows (CSO), and stormwater runoff. As a result, urban water bodies often contain elevated levels of pathogenic enteric bacteria, requiring water quality managers to post swimming and boating advisories. The decision rule used by many managers is as follows: If indicator bacteria levels in a water sample exceed health standards for recreation, advisories are posted so as to dissuade recreational activities that could lead to significant exposure (USEPA 1999).

One disadvantage of this approach is that it relies on indicator bacteria measurements to drive the decision making process. Typically, quantitative bacteria tests take more than 28 h to complete from the time of sample collection to the time results are reported. Thus, if a water sample collected from a beach on a Friday morning is found to contain high levels of bacteria, the decision to close the beach is not made until Saturday afternoon. This means bathers may have been exposed to high levels of bacteria on Friday and Saturday before the beach was closed.

This example illustrates the need for tools to help make water quality information available to the public in a timely manner. Although modeling tools are available that allow prediction of water quality (e.g., BASINS, HSPF, GWLF, SWMM), the input data requirements for many models are significant, and often difficult or prohibitively expensive to obtain. For example, many municipalities do not have accurate maps of drainage systems, and the location of nonpoint sources of pollution are often unknown. In addition, many water quality modeling approaches do not yield models that can make accurate, real-time estimates of bacteria levels. Therefore, the goal of this research was to address this need by developing relatively simple, but accurate water quality models that can be used to estimate bacteria concentrations in a water body and the probability that concentrations will exceed recreation standards.

The models were developed for *Enterococcus* bacteria based on regression analysis techniques. *Enterococci* concentrations were routinely measured in several waterbodies during both dry and wet weather in the spring and summer of 2002 and 2003. Multivariate linear regression models were developed with precipitation, streamflow, time since last rainfall, and other hydrologic parameters as possible explanatory variables. Using the

[1] Camp Dresser and McKee, Cambridge, MA; formerly, Dept. of Civil and Environmental Engineering, Tufts Univ., Medford, MA 02155.

[2] Dept. of Civil and Environmental Engineering, Tufts Univ., Medford, MA 02155; and, Water Systems, Science, and Society (WSSS) Research and Graduate Education Program, Tufts Univ., Medford, MA 02155 (corresponding author). E-mail: john.durant@tufts.edu

[3] Camp Dresser and McKee, Milwaukee, WI; formerly, Dept. of Civil and Environmental Engineering, Tufts Univ., Medford, MA 02155.

[4] Dept. of Civil and Environmental Engineering, Tufts Univ., Medford, MA 02155; and Water Systems, Science, and Society (WSSS) Research and Graduate Education Program, Tufts Univ., Medford, MA 02155.

[5] Dept. of Civil and Environmental Engineering, Tufts Univ., Medford, MA 02155.

Note. Discussion open until June 1, 2008. Separate discussions must be submitted for individual papers. To extend the closing date by one month, a written request must be filed with the ASCE Managing Editor. The manuscript for this paper was submitted for review and possible publication on February 16, 2006; approved on October 16, 2006. This paper is part of the ***Journal of Water Resources Planning and Management***, Vol. 134, No. 1, January 1, 2008. ©ASCE, ISSN 0733-9496/2008/1-73–82/$25.00.

expected value and the variance of the predicted *Enterococci* concentrations, prediction interval probability distribution functions (PDFs) were constructed to estimate the probability that a predicted concentration was in excess of the swimming or boating standard.

Applications to Mystic Watershed

Models were developed for sites on the Mystic River watershed in eastern Massachusetts. The Mystic River is one of three major rivers that drain into Boston Harbor. Nearly 400,000 people live in the 17,000 hectare (ha) watershed, making the Mystic among the most densely populated watersheds in the state. Urbanization and industrial activities in the watershed have greatly influenced the hydraulics of the river as well as its water quality; mills, shipyards, automobile plants, tanneries, chemical manufactories, roadways, bridges, and sewage and stormwater infrastructure have all left their mark on the river and its banks. Contemporary water quality problems of concern in the Mystic include toxic chemicals leaching from waste disposal sites, excess nutrients, noxious aquatic plants, low dissolved oxygen, oil and grease, and sewage indicator bacteria (MyRWA 2004).

Bacteria inputs on the Mystic come from three main sources: stormwater runoff, combined sewer overflows, and illicit connections between sanitary waste lines and storm drains. Two tributaries to the Mystic, the Aberjona River and Alewife Brook, are significantly impacted by sewage inputs from CSOs. For example, in 2002 more than 15 CSO events occurred on the Alewife Brook, and in 2003 more than 19 occurred (Cambridge Department of Public Works 2004). Inputs from these tributaries as well as other sources may be impacting water quality at recreation areas such as Sandy Beach on Upper Mystic Lake and the Blessing of the Bay boathouse on the Mystic River (Fig. 1). At Sandy Beach, which is impacted by the Aberjona River, health advisories were posted 10 times (32 days) in 2002 due to elevated *Enterococci* levels and 7 times (20 days) in 2003 (USEPA 2003). Therefore, the problem of periodic high loadings of sewage bacteria to recreational waters in the Mystic makes this watershed a suitable candidate for study.

Methods

Enterococci Measurements

Enterococci measurements were made at four locations in the watershed: (1) Sandy Beach on Upper Mystic Lake, a public swimming beach operated by the Massachusetts Department of Conservation and Recreation; (2) Aberjona River at a United States Geological Survey (USGS) flow measurement station (#01102500) located ~0.8 km upstream of Sandy Beach; (3) Mystic River at the Blessing of the Bay boathouse, where recreational boating takes place; and (4) Alewife Brook, a CSO-impacted tributary to the Mystic River, located ~3 km upstream of Blessing of the Bay boathouse (Fig. 1). Additional information about the sites is provided in Table 1. Samples were collected between May and August in 2002 and 2003. Dry weather samples were collected on weekdays between 9:00 and 10:00 a.m., 2–5 times per week. Samples were also collected before, during, and after five wet-weather events (cumulative precipitation >1 cm during each event) to determine the response of the system to large, short-term inputs of polluted runoff. All samples were collected in autoclaved, 250 mL polypropylene bottles. Field duplicate samples were collected on each sampling day. Samples were stored on ice in the field and brought to a temperature of $1-4°\mathrm{C}$ before analysis. All samples were analyzed within 6 h of collection.

The samples were analyzed using the EPA Modified *Enterococci* Testing Method 1600 (USEPA 2000). Samples were vacuum filtered through 47 mm diameter, 0.45 μm pore-size, sterile membrane filters. The filters were set on top of mEI (membrane-Enterococcus Indoxyl-β-D-Glucoside) agar and incubated for 24 h at $41\pm0.5°\mathrm{C}$. Dilutions were chosen based on site-specific conditions, with the goal of obtaining 20–60 colonies on each filter. Colonies exhibiting blue halos were counted as *Enterococci* colonies. Colonies were manually counted by two different people and the counts were averaged. Laboratory blanks were run each day. To estimate laboratory error 175 sample dilutions were run in duplicate. The error was ~30%, which is relatively low for bacterial assays. Additional details of the analysis are described in Oriel (2003).

Explanatory Variables

In this section, the explanatory variables used in the models are described. The relative importance of each variable to the modeling results is then presented in the section entitled "Results and Discussion."

Precipitation and the time since the last rainfall event were investigated as explanatory variables in the regression models for all four sites. Continuous measurements of precipitation were obtained from the USGS gauging station (#01102500), located at the Aberjona River sampling site (USGS 2003). Flow data were also available for the Aberjona River site, and thus flow was evaluated as a possible explanatory variable in the Aberjona River regression models. Because Sandy Beach is only 0.8 km downstream of the Aberjona site, flows (measured at the Aberjona site) were also evaluated in the Sandy Beach models.

Several manipulations of the precipitation data were investigated. An algorithm was developed to sum the precipitation over a fixed number of hours prior to the time of sample collection. The distribution of precipitation data was found to be highly skewed, with a large number of zeroes and low values, and only a few high values. The data were thus log transformed, first adding 1 in. (2.54 cm) of rainfall to each observation. The new independent variable resulting from this transformation was $\log(P_X+1)$, where X=number of hours over which the precipitation data were summed.

Because there is a lag between when precipitation falls and runoff occurs at a sampling location, it was reasoned that the most recent precipitation should *not* be factored into the regression. Therefore, a second set of independent variables was created from the precipitation data that included a time lag. For example, the total precipitation in an 8 h period between 2 and 10 h prior to the time being evaluated is expressed as

$$P_8|_{t-2} = \sum_{t=-10}^{t=-2} P \tag{1}$$

The *Enterococci* concentration data for the Aberjona River showed a steady decrease following the end of a rainstorm. To capture this phenomenon, the variable T_F was created to represent the time (expressed in days) between sample collection and the end of the prior rainfall of greater than 0.05 in. (0.13 cm) in a 15 min period. T_F was highly sensitive to rainfall depth. To determine the optimum threshold for use in the regression, an ex-

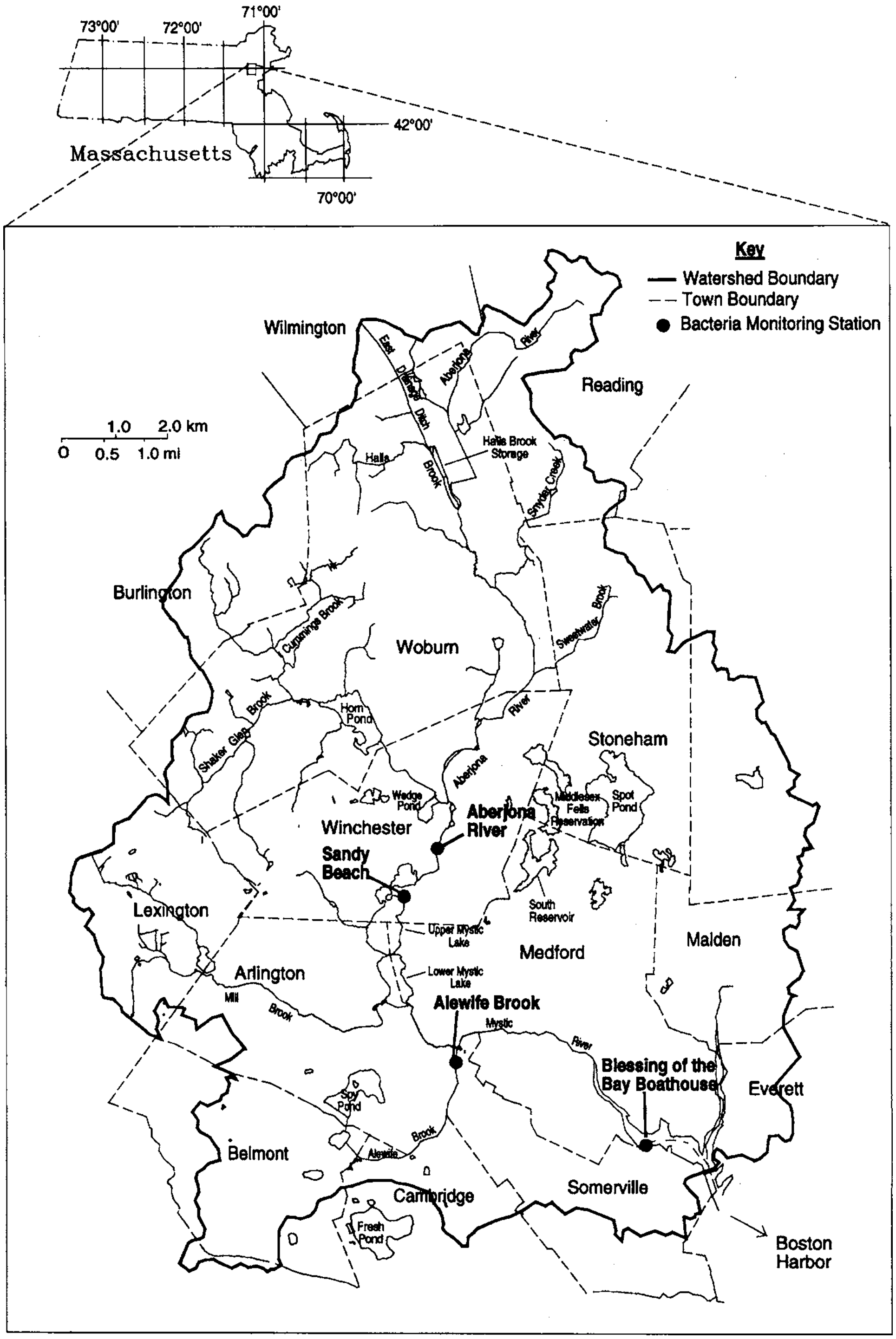

Fig. 1. Map of *Enterococci* measurement sites in the Mystic River watershed

periment was conducted in which T_F was calculated for precipitation thresholds from 0.025 to 0.25 cm and regressed against bacteria concentration. The distribution of T_F was skewed with a preponderance of values between 0 and 10 days, and only a few between 10 and 30 days. Thus, T_F was log transformed, first adding 1 day to each observation creating the new variable, $\log(T_F+1)$.

Several manipulations of the streamflow data were investigated including the flow at the time of water sample collection (Q_t), the time-lagged flow measured Δt hours before sampling ($Q_{t-\Delta t}$), and hydrograph slope. An experiment was conducted to determine whether the lagged flows would improve the regression model fit. A set of independent variables, $Q_{t-\Delta t}$, was created from the time series of streamflow at time $t-\Delta t$, and the data were then regressed against $\log(Q)$ and the correlation coefficient, R^2, was recorded. This experiment was repeated for Δt between 0 and 12 h.

It was hypothesized that flow averaged over X hours prior to sample collection ($\bar{Q}_{\mathrm{X\ h}}$) might also be an important variable. It was found that the average flow determined by a numerical integration method such as the trapezoidal rule gave an almost iden-

Table 1. Sampling Site Information

Location (lat × long)	Approximate drainage area[a]	Site features
Aberjona River (42°26′50″N × 71°08′22″W)	6,100 ha (23.6 mi^2)	Located at USGS flow gauging station, 0.3 km upstream of Upper Mystic Lake Inlet; impacted by CSO discharges
Alewife Brook (42°24′28″N × 71°07′60″W)	2,300 ha (8.9 mi^2)	Located 0.1 km upstream of Mystic River; affected by backwatering and CSO discharges
Upper Mystic Lake at Sandy Beach (42°26′22″N × 71°08′43″W)	7,400 ha (28.5 mi^2)	Located on northern end of lake; lake is 81×10^4 m^2 in area, with average depth of ∼12 m
Mystic River at Blessing of the Bay Boathouse (42°23′56″N × 71°05′21″W)	14,000 ha (53 mi^2)	Located upstream of Amelia Earhart Dam, which causes significant backwatering effects at the boathouse

[a] Watershed drainage area upstream of sampling site.

tical result to simply taking the arithmetic mean (e.g., for the 24 h average flow, the difference between the two methods less was <1%); therefore, the arithmetic mean was used. Experiments were also conducted to determine whether better correlations could be obtained using the average rather than instantaneously measured flow. Correlations were evaluated for $\bar{Q}$ averaged over 0–24 h.

Based on the work of Rudolph (2002), it was hypothesized that the slope of the hydrograph might help explain variations in *Enterococci* concentrations. A computer routine was used to fit a polynomial of order m to n observations bracketing the times at which the samples were taken. It was found that little was gained from using a second- or third-order polynomial; thus, the slope was determined by fitting a simple linear regression line through 10 points, or 2.5 h of streamflow data.

Model Development

Multivariate linear regression equations were developed with the 2002 *Enterococci* data by ordinary least squares in Minitab Statistical Software (State College, Pa.). The regression analyses were preformed using a stepwise approach that allowed linear combinations of the explanatory variables to be tested for significance, and yielded a set of possible models and performance statistics. The linear regression models were evaluated based on three criteria: the adjusted-R^2 value, the standard error of the model, S_e, and the prediction error sum of squares (PRESS) statistic. The model's coefficient of determination, the adjusted R^2, was used to give a measure of the percentage of the variation in the data that could be explained by the model. The adjusted R^2 takes into consideration the loss of degrees of freedom as additional variables are added to the model. The standard error of the model, S_e, is equivalent to the standard deviation, or the square root of the variance of the model residuals. The PRESS statistic was used to evaluate the errors associated with predictions made by the models (Helsel and Hirsch 2002). The selected regression models were those that had the highest adjusted-R^2 value and the lowest values for S_e and the PRESS statistic.

Probabilistic models were developed for Upper Mystic Lake at Sandy Beach and the Mystic River at Blessing of the Bay Boathouse—the two sites at which recreation activities take place—using the selected linear regression models. Predicted *Enterococci* concentrations and associated variances were used to generate PDFs based on the prediction interval of the regression model (Helsel and Hirsch 2002). For example, Fig. 2 shows a model's estimate of log *Enterococci* concentration for values of a given set of explanatory variables. The model's best estimate, $\log(C_{\text{entero}}) = 2.2$(∼160 cfu/100 mL), is shown in the center of the PDF of the regression estimate. The total area under the curve is equal to unity. The area of the shaded portion to the right of the line is the probability that the concentration will exceed 61 cfu/100 mL, the single-sample swimming standard (Commonwealth of Massachusetts 1997). In this case $p = 0.81$, which means there is an 81% chance that the water quality exceeds the swimming standard, and a 19% chance that the concentration is below the standard. The single-sample, *Enterococcus* swimming standard (61 cfu/100 mL) was used for Sandy Beach. Because an *Enterococcus* boating standard has yet to be established, we used a threshold of 305 cfu/100 mL for Blessing of the Bay Boathouse based on the ratio of the boating to swimming standards for fecal coliform bacteria (5).

Although this technique gives an estimate of the probability of exceeding water quality standards for a given set of conditions, it does not indicate the threshold probability at which water quality advisories should be posted. To estimate this threshold, Francy et al. (2003) recommended adjusting the threshold probability above which a beach is deemed unsafe so as to minimize the number of false negatives (predicting safe conditions when unsafe conditions prevail) and the number of false positives (unnecessary closures).

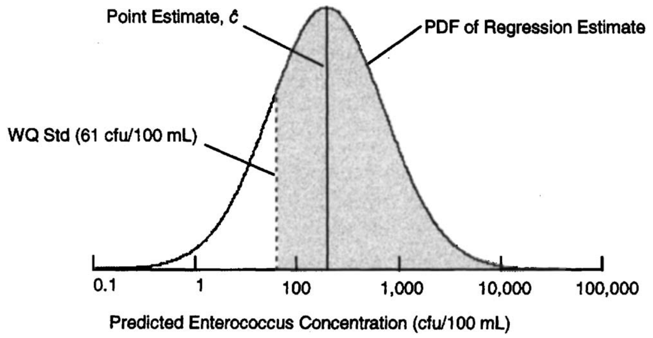

Fig. 2. Example illustrating how the PDF is used to calculate the probability that an *Enterococci* concentration (predicted with a linear regression model) exceeds the swimming standard of 61 cfu/100 mL. In this example, the C_{entero} estimate from the model is ∼160 cfu/100 mL; however, due to the uncertainty in the model, the probability that estimate actually exceeds the swimming standard is 81%, represented by the shaded area under the curve to the right of the hatched line.

Table 2. Summary of *Enterococci* Measurements Made in 2002 and 2003 at Each Sampling Site

			Concentration (cfu/100 mL)							
	Number of samples tested		Geometric mean		Median		Minimum		Maximum	
Sampling site	2002	2003	2002	2003	2002	2003	2002	2003	2002	2003
Aberjona River	62	22	500	360	300	270	46	110	21,000	3,700
Alewife Brook	80	22	990	320	460	260	35	7	55,000	12,000
Upper Mystic Lake at Sandy Beach	67	21	14	15	14	11	<1	<1	390	330
Mystic River at Blessing of the Bay	57	19	7	11	3	5	<1	1	>570	2,100

Results and Discussion

General

A summary of the *Enterococci* concentration data measured at the four sites in 2002 and 2003 is shown in Table 2. The highest *Enterococci* concentrations occurred at the Alewife Brook and Aberjona River sampling sites during both dry weather and wet weather. Geometric mean concentrations were 20 to 100 times higher at these river sites than at the more lacustrine sites—Upper Mystic Lake at Sandy Beach and Mystic River at the Blessing of the Bay boathouse. The Aberjona River and Alewife Brook sites reacted quickly to rainfall events, and had peak *Enterococci* concentrations shortly after a heavy rain had started. Peak concentrations were not seen at Sandy Beach and Blessing of the Bay boathouse until several hours after the start of a storm. All sites generally returned to prestorm concentrations within three days of the end of a storm.

The linear regression models that produced the best fits to the observed data at the four sites are shown in Table 3. These models explained between 42 and 82% of the variability in the *Enterococci* concentrations. Of the variables analyzed, precipitation was the most strongly correlated with bacteria concentrations. Weaker correlations were found for T_F (time between sample collection and the end of the most recent precipitation event) and Q_t (streamflow). Significant correlations were not observed for hydrograph slope. In general, the linear regression models for the two riverine sites, the Aberjona River and Alewife Brook, yielded better fits to the observed bacteria data than the models for the two open water sites, Upper Mystic Lake at Sandy Beach and the Mystic River at the Blessing of the Bay boathouse. These observations are consistent with the literature summarized in Table 4.

Estimates of Bacteria Concentrations

Plots of the measured *Enterococci* concentrations versus those predicted by the selected models are shown in Fig. 3. Time-series plots showing measured versus predicted bacteria concentrations are shown in Fig. 4. The model predictions in Fig. 4 represent the geometric mean of the distribution of possible predictions at any one time. Thus, the predictions tend to plot closer to the centers of the distributions than the very high and low observed values.

For the Aberjona River site a significant correlation was found between $\log(C_{\text{entero}})$ and P_{20}, precipitation summed over the 20 h period immediately prior to sample collection. The log-transformed variable $\log(P_{20}+1)$ yielded a slightly better linear fit. The regression was repeated for 1 h increments from 1 to 72 h. The results showed that the correlation was strongest for P_{20}; however, the results for a range of times from 20 to 36 h were nearly as good. In addition, the results showed that the correlation was strongest when there was no time lag ($\Delta t=0$) for the summed precipitation [Eq. (1)], suggesting the importance of highly localized bacteria inputs.

The relationship between $\log(C_{\text{entero}})$ and T_F (the time since last rainfall) was poor (adjusted $R^2=0.36$); however, when $\log(C_{\text{entero}})$ was regressed against $\log(T_F+1)$ the adjusted R^2 increased to 0.63. The rainfall depths at which $\log(C_{\text{entero}})$ versus T_F gave the highest adjusted R^2 ranged from 0.1 to 0.18 cm.

Table 3. Linear Regression Models for Log *Enterococci* Concentration at the Four Monitoring Sites. $\log(C_{\text{Entero}})=a+b_1\log(P_x+1)+b_2\log(T_{\text{F}}+1)+b_3Q_t+\varepsilon$.

Site[a]	a	b_1	b_2	b_3	S_e[b]	adj. R^2	PRESS
Aberjona River [$n=62$]	2.79 (26)	7.89 (7.7)	−0.563 (−4.9)		0.27	0.82	4.3
Alewife Brook [$n=80$]	2.88 (23)	11.6 (6.2)	−0.30 (−2.4)		0.45	0.55	14.9
Upper Mystic Lake at Sandy Beach [$n=67$]	2.12 (4.7)	7.1 (4.0)	−0.67 (−2.1)	−0.76 (−3.3)	0.50	0.42	16.9
Mystic River at Blessing of the Bay Boathouse [$n=57$]	1.12 (6.1)	4.7 (2.9)	−1.0 (−4.2)		0.45	0.61	11.3

Note: t-ratios are reported in parentheses. $\log(P_x+1)$=log of the precipitation in inches for x hours prior to time of sample collection. $x=20$ h for Aberjona, 12 h for Alewife, and 24 h for Sandy Beach and Blessing of the Bay Boathouse; T_F=time (days) between sample collection and last prior rainfall >0.05 in. in 15 min; Q_t=discharge at the time of sample collection; measured at Aberjona River discharge gauging station; a, b_1, b_2, and b_3 are constants; and random error, ε, is defined as follows: $\varepsilon=-\log(C_{\text{ave}})-\log(C_i)$.

[a] The number of C_{entero} measurements(2002 data) used to calibrate each model is shown in square brackets.

[b] Standard error of model residuals.

Table 4. Results of Studies That Have Used Linear Regression for Bacteria Modeling

Author(s)	Water body	Response variable(s)	Explanatory variables	Performance statistic
Ferguson et al. (1996)	Georges River (near Sydney, Australia)	Fecal coliform, fecal streptococci, *Clostridium perfringens*, F-RNA bacteriophage, *Aeromonas*, *Giardia*, *Cryptosporidium*	Precipitation and CSO activation	0.52–0.80 (R^2)
Serrano et al. (1998)	Beaches in San Sabastien (Basque Country)	Total coliform, fecal coliform, fecal streptococci, *E.coli*, *Salmonella*, somatic coliphages, F-RNA phages	Time of day, overcast skies, low and high tides, groundswell, turbidity, and presence of floating debris	0.17–0.42 (R^2)
Crowther et al. (2001)	Beaches on Fylde Coast (west coast of England)	Total coliform, fecal coliform, fecal streptococci	Precipitation, tide height at time of sampling, sunshine, and proportion of onshore winds	0.06–0.53 (adjusted R^2)
Christensen (2001)	Rattlesnake Ck. (Kansas)	Fecal coliform	Turbidity	0.04–0.62 (R^2)
Christensen et al. (2002)	Four rivers in Kansas	Fecal coliform	Turbidity, seasonality, streamflow, specific conductance	0.59–0.73 (R^2)
Eleria and Vogel (2005)	Charles River (Boston)	Fecal coliform	Storm characteristics, seasonality, net solar radiation, sky cover, wind speed, streamflow, hysteresis, and CSO activations	0.50–0.60 (adjusted R^2)
Rasmussen and Ziegler (2003)	KansasRiver and Little Arkansas River (Kansas)	Fecal coliform, *E. coli*	Turbidity	0.59–0.79 (R^2)
McLellan and Salmore (2003)	Beaches in Milwaukee Harbor	*E. coli*	Precipitation, time since last rainfall, wind speed, and direction	0.03–0.29 (R^2)
Francy et al. (2003)	Beaches on Lake Erie (Ohio)	*E. coli*	Streamflow, precipitation, wave height, birds	0.32–0.40 (R^2)
This study	Riverine and lacustrine sites on the Mystic River	*Enterococcus*	Streamflow, precipitation, time since last rainfall, hysteresis	0.42–0.82 (adjusted R^2)

A significant relationship was also found between $\log(C_{\text{entero}})$ and streamflow (Q_t) at the time of sampling for the Aberjona River site (adjusted $R^2=0.57$). The hydrograph slope, dQ/dt, proved not to be a useful independent variable. Based on data from two storm events, during which the Aberjona River was intensively sampled, it was hypothesized that the hydrograph slope would have explanatory power; however, it was found that bacteria concentrations were high throughout the runoff event, on both the rising and falling limbs of the hydrograph. This suggests that the magnitude of the slope, rather than its sign, was important. Taking the absolute value of the slope improved the correlation, but most of the dQ/dt values were very low, and thus the slope variable was further transformed by taking its logarithm. Because some slopes were zero, it was necessary to first add unity [1 cfs/h (28.3 L/s/h)]to each observation before taking its logarithm. The correlation between $\log(|dQ/dt|+1)$ alone and *Enterococci* bacteria had an adjusted R^2 of 0.70.

Results were developed for several models containing different combinations of explanatory variables. The model shown in Table 3 was selected because it had the highest adjusted R^2 and the lowest standard error and PRESS statistic. The model residuals were normally distributed, and that the sample slope would therefore obey a student's t-distribution. Slopes were accepted as significant for a probability $p<0.05$ or a t-score $|t|>2$. The t-scores of all coefficients in the selected model were large enough so that the slopes were statistically significant. A plot of the measured *Enterococci* concentrations versus those predicted by the selected model is shown in Fig. 3. Fig. 3 shows that at times the model may be in error by nearly an order of magnitude. A time-series plot showing measured versus predicted bacteria concentrations is shown in Fig. 4.

For the Alewife Brook site, significant correlations were found between $\log(C_{\text{entero}})$ and precipitation $(\log(P_X+1))$, as well as between $\log(C_{\text{entero}})$ and the time since the last rainfall $\log(T_F+1)$. It was found that precipitation summed over 12 h $(\log(P_{12}+1))$ produced the highest values of adjusted R^2. The two-variable model shown in Table 3 is considered the best one as it had the highest adjusted R^2, and lowest PRESS statistic and standard error. A single-variable regression model based on $\log(P_{12}+1)$, was nearly as good as the two-variable model shown in Table 3, demonstrating the effect of precipitation on water quality at this highly urban, CSO-impacted site.

Several single- and multivariable linear regression models were developed for Sandy Beach. It was found that precipitation summed over 24 h $(\log(P_{24}+1))$ regressed against $\log(C_{\text{entero}})$ produced the highest adjusted-R^2 values (0.33) of any of the single variables analyzed. The fit of the model to the data was slightly improved (adjusted $R^2=0.42$) by adding the time since the last rainfall (T_F) and Aberjona River discharge (Q_t).

There is some multicollinearity between the precipitation and streamflow variables ($R^2=0.35$), however, the three-variable model generally had the best fit (highest adjusted R^2, lowest

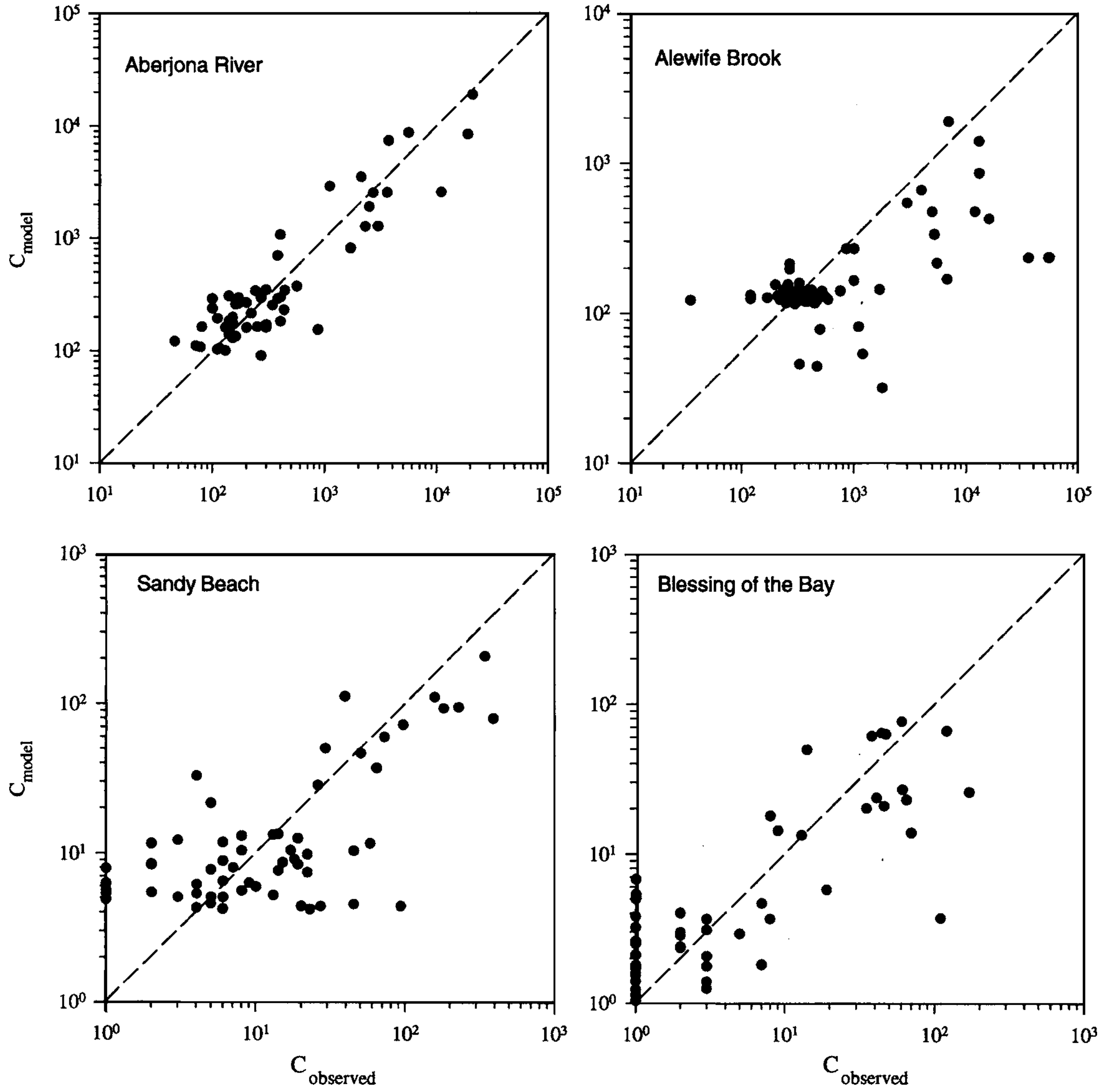

Fig. 3. Scatterplots of *Enterococci* concentrations predicted with the selected linear regression model versus observed concentrations (cfu/100 mL) for the Aberjona River, Alewife Brook, Upper Mystic Lake at Sandy Beach, and Mystic River at the Blessing of the Bay boathouse in Summer 2002. The 1:1 diagonal lines represent perfect correlation (theoretical) between observed and modeled results.

PRESS statistic). The variance inflation factors (VIF) for the variables were 3.1 and 1.6 for the flow and precipitation variables, respectively. VIF greater than 1 indicates multicollinearity, or relatedness, amongst the independent variables. In our context, this means that the prediction interval will be wider. Helsel and Hirsch (2002) advise that serious problems are avoided when the VIF are less than 10. We selected the three-variable model shown in Table 3.

Flow information was not available for the Alewife Brook site; therefore, only two independent variables (P and T_F) were evaluated. Both variables had predictive power. In the selected model (Table 3), the highest adjusted-R^2 value (0.61) occurred for precipitation summed over the last 24 h (P_{24}).

Probabilistic Model to Estimate Probability of Exceeding Water Quality Standards

Prediction interval PDFs were developed for Sandy Beach and the Blessing of the Bay sites to estimate the probabilities that real-time *Enterococci* concentrations (predicted using the selected linear regression models in Table 3) would exceed recreational water quality standards. The results for Sandy Beach in Fig. 5(a) show *Enterococci* concentrations plotted against the probability of exceeding the single-sample swimming standard (61 cfu/100 mL). The threshold probability at which a beach closure is announced was selected which gave the lowest number of false negatives (incorrect predictions that the standard is met), and the lowest number of false positives (incorrect predictions that the standard is exceeded). Fig. 5(a) shows that the optimum threshold probability is $p=0.20$. According to this framework, when the model predicts that the probability of exceeding the water quality standard is $\geq 20\%$, a water quality advisory is warranted; when the model predicts a probability of less $<20\%$, no advisory is needed. Using the 2002 dataset which was used to develop the model, the Sandy Beach probabilistic model correctly predicts 8 out of 9 (88%) exceedances (only 1 false negative), 46 out of 52 (88%) nonexceedances (6 false positives). Had this model been run each day during the summer of 2002 (May 1–August 31, 2002), it would have predicted exceedances 20% of the time (25 out of 123

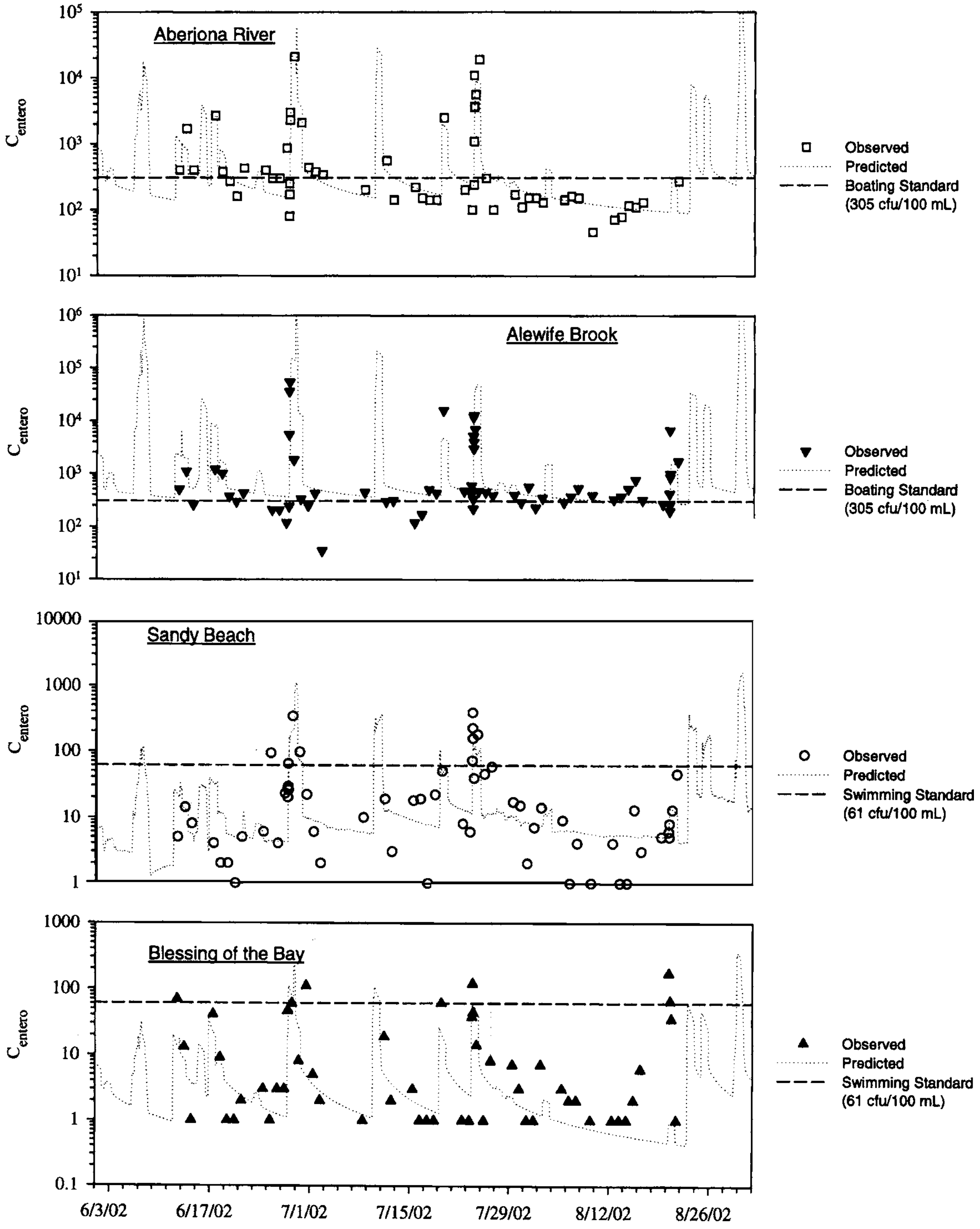

Fig. 4. Time series of observed *Enterococci* concentrations and those predicted with the selected linear regression model (cfu/100 mL) the Aberjona River, Alewife Brook, Upper Mystic Lake at Sandy Beach, and Mystic River at the Bay Boathouse in Summer 2002. The observed bacteria data are shown with error bars of ±30%, which represents the standard error of the measurements.

days). As an independed confirmation, 2003 *Enterococci* data were used to confirm the model. The results in Fig. 5(b) show that for 21 observations (May 26–August 11, 2003) the model correctly predicts 4 out of 5 exceedances and 15 out of 16 nonexceedances.

The results of the probabilistic model for the Blessing of the Bay Boathouse site indicate that $p=0.10$ is the optimum threshold for predicting exceedances of the single-sample boating standard [Fig. 6(a)]. As Massachusetts does not currently have a boating standard based on *Enterococcus*, it is assumed to be 5 times the swimming standard, or 305 cfu/100 mL. At a threshold of $p=0.10$, the probabilistic model correctly predicted 3 out of 3 (100%) exceedances (zero false negatives) and 58 out of 69 (84%) nonexceedances (11 false positives). Had this model been run for each day during the summer of 2002 (May 1–August 31, 2002), it would have predicted exceedances 5% of the time (6 out of 123 days). The results in Fig. 6(b) show that for 13 observations in 2003 (May 26–August 11, 2003), the model correctly predicts 3 out of 3 exceedances and 10 out of 10 nonexceedances.

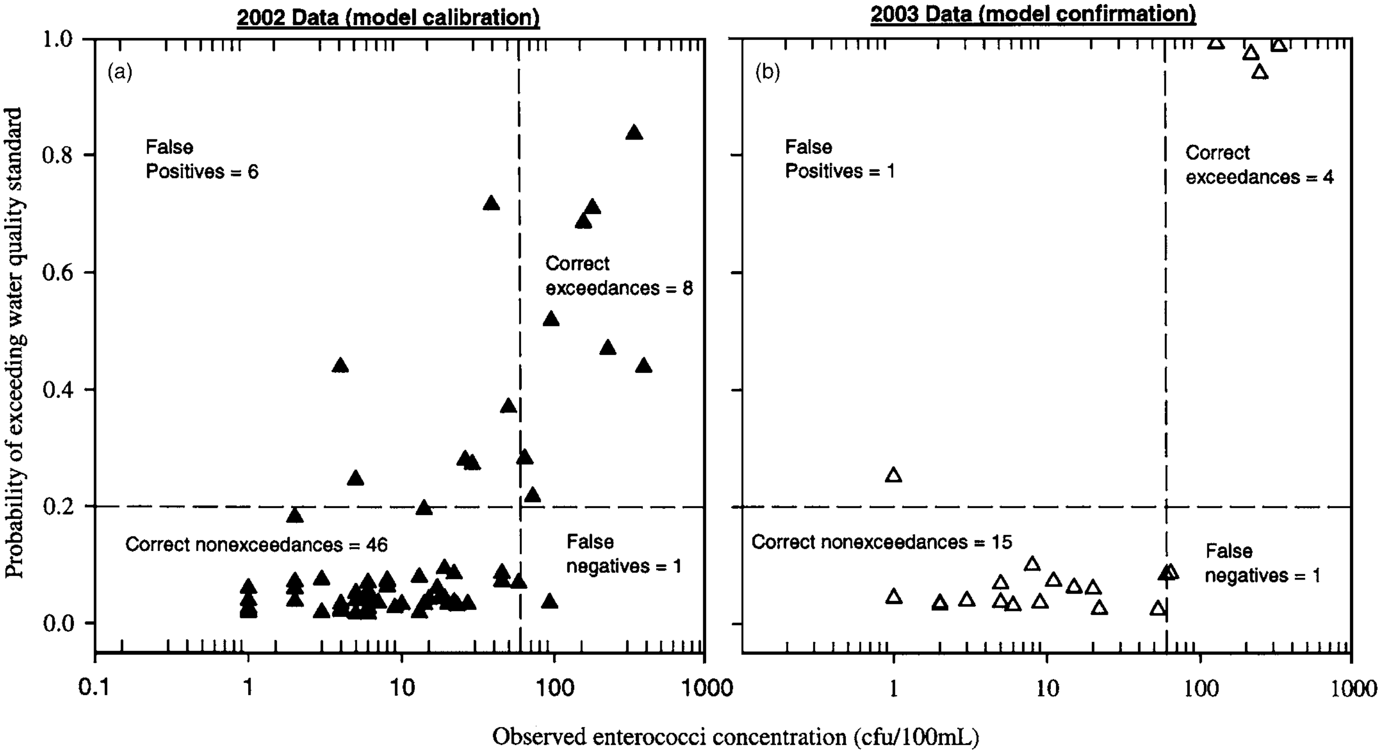

Fig. 5. Observed *Enterococci* concentrations in Upper Mystic Lake at Sandy Beach plotted against the probability of exceeding the water quality standard predicted by a PDF of the linear regression model results. The vertical line represents the single-sample *Enterococcus* swimming standard (61 cfu/100 mL); (a) Fit of the model to the 2002 data (with which the linear regression equation was developed); (b) fit of the model to confirmatory data collected in 2002.

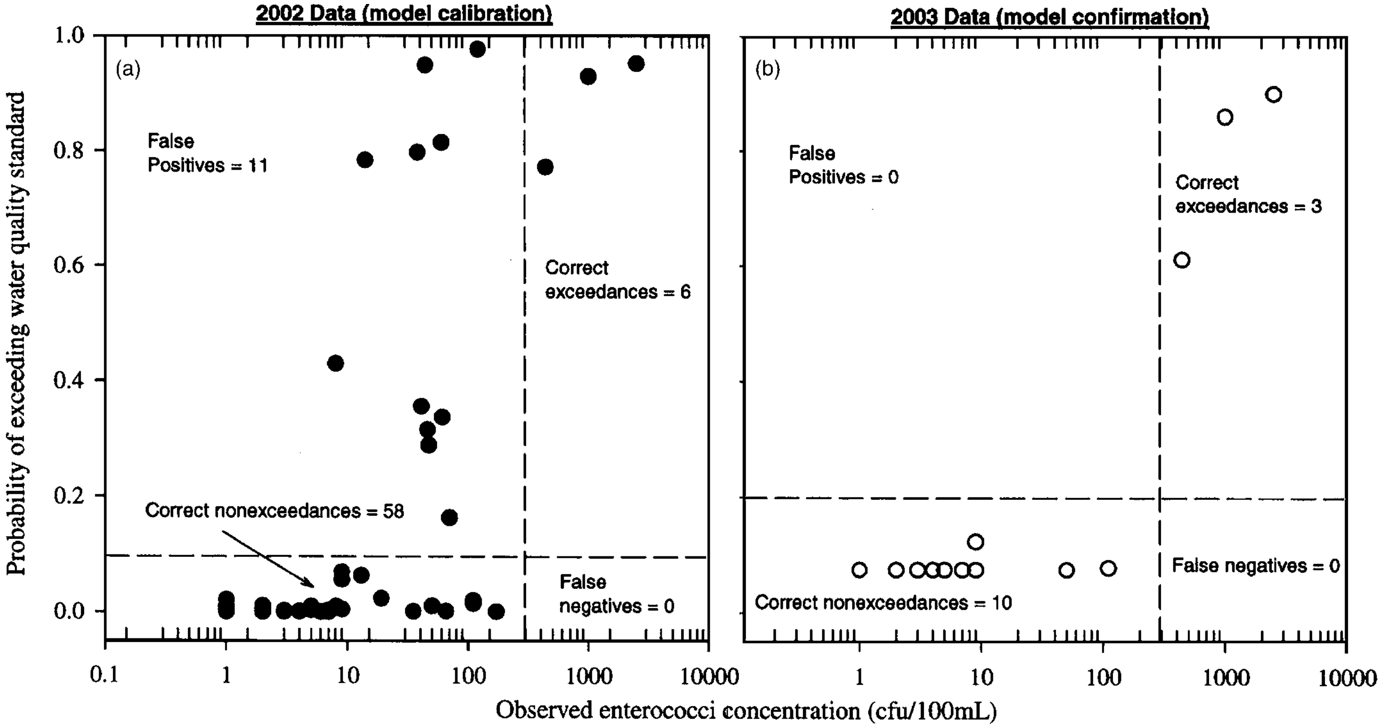

Fig. 6. Observed *Enterococci* concentrations in the Mystic River at Blessing of the Bay Boathouse plotted against the probability of exceeding the water quality standard predicted by a PDF of the linear regression model results. The vertical line represents the single-sample *Enterococcus* boating standard (assumed to be 305 cfu/100 mL); (a) Fit of the model to the 2002 data (with which the linear regression equation was developed); (b) fit of the model to confirmatory data collected in 2002.

Significance

In general, linear regression models are a practical alternative to distributed models because of the extensive data needs of the latter and the difficulty of obtaining the required data—particularly for complex urban drainage and sewerage systems. The linear regression bacteria models developed here show good agreement with similar models developed for other systems (Table 4). In addition, our model results are consistent with other studies that show linear regression models to be better at predicting bacteria levels in rivers than in lakes and impoundments. Another advantage of linear regression models is that they give users direct information about the uncertainty of the results. We showed that real-time estimates of bacteria concentrations and their associate variances can be used to generate prediction interval PDFs. PDFs can be used to give users the probability of exceeding a water quality standard, which can then be used as a basis for establishing water quality advisories. Using this approach, we found that for two lacustrine sites on an urban watershed the number of false negatives ranged from 14 to 36% and the number of false positives ranged from 10 to 11%.

One potential limitation of this approach is that if there are changes in hydraulic conditions (e.g., addition or removal of CSOs or other sources of sewage) or hydrologic conditions in the watershed, additional data must be collected and the models recalibrated. Linear regression models only partially explain the variability of observed bacteria concentrations. To help better explain this variability, and thereby improve the predictive powers of models, future research could include consideration of explanatory variables such as land use, drainage system density, and additional real-time water quality data (e.g., turbidity, dissolved oxygen, temperature, and conductivity). However, in urbanized catchments precipitation will likely remain the most significant variable for explaining rapid changes to bacteria levels following storm events.

Acknowledgments

The writers are grateful to Michele Cutrofello and Laurie Cormier for their help in collecting and analyzing water samples, and to Grace Perez, Vithal Deshpande, and Lisa Brukilacchio for assistance in the site selection process. In addition, they are grateful to Dr. Marc Doolittle and fellow staff at the M.A. Dept. of Conservation and Recreation and the Somerville Department of Parks and Recreation for permission to use their facilities for sampling. Funding for this work was provided by the U.S. Environmental Protection Agency (Grant No. R829338). The content of this paper does not necessarily represent the views of the agency and has not been subject to agency review.

References

Cambridge Department of Public Works. (2004). "Outfall monitoring—Combined sewer overflow (CSO) data." ⟨http://www.cambridgema.gov/%7ETheWorks/departments/swrMnt/cso.html⟩, Cambridge, Mass (May 24, 2005).

Christensen, V. G. (2001). "Characterization of surface-water quality based on real-time monitoring and regression analysis, Quivira National Wildlife Refuge, South-Central Kansas, December 1998 through June 2001." *WRIR01–4248*, U.S. Geological Survey, Lawrence, Kan.

Christensen, V. G., Rasmussen, P. P., and Ziegler, A. C. (2002). "Real-time water quality monitoring and regression analysis to estimate nutrient and bacteria concentrations in Kansas streams." *Water Sci. Technol.*, 45(9), 205–219.

Commonwealth of Massachusetts. (1997). "Code of Massachusetts Regulations." *105 CMR 445.00*: *Minimum Standards for Bathing Beaches-State Sanitary Code-Chapter VII*, Boston.

Crowther, J., Kay, D., and Wyer, M. D. (2001). "Relationships between microbial water quality and environmental conditions in coastal recreational waters: The Fylde Coast, UK." *Water Res.*, 35(17), 4029–4038.

Eleria, A., and Vogel, R. M. (2005). "Predicting Fecal Coliform Bacteria in the Charles River." *J. Am. Water Resour. Assoc.*, 41(5), 1195–1209.

Ferguson, C. M., Coote, B. G., Ashbolt, N. J., and Stevenson, I. M. (1996). "Relationships between indicators, pathogens and water quality in an estuarine system." *Water Res.*, 30(9), 2045–2054.

Francy, D. S., Gifford, A. M., and Darner, R. A. (2003). "Escherichia coli at Ohio bathing beaches—Distribution, sources, wastewater indicators, and predictive modeling." *WRIR02-4285*, U.S. Geological Survey, Columbus, Ohio.

Helsel, D. R., and Hirsch, R. M. (2002). "Statistical methods in water resources." *USGS techniques of water resources investigations*, Book 4, Chapter A3, ⟨pubs.usgs.gov/twri/twri4a3/⟩ (last date accessed, January 2006), USES, Washington, D.C.

McLellan, S. L., and Salmore, A. K. (2003). "Evidence for localized bacterial loading as the cause of chronic beach closings in a freshwater marina." *Water Res.*, 37, 2700–2708.

Mystic River Watershed Association (MyRWA). (2004). "Mystic watershed action plan." ⟨http://www.mysticriver.org/about_watershed/index.html⟩ (March 14, 2005).

Oriel, K. A. (2003). "Predictive bacteria models for a sewage-impacted urban watershed." MS thesis, Dept. of Civil and Environmental Engineering, Tufts Univ., Medford, Mass.

Rasmussen, P. P., and Ziegler, A. C. (2003). "Comparison and continuous estimates of fecal coliform and *Escherichia Coli* bacteria in selected Kansas streams, May 1999 through April 2002." *WRIR03-4056*, U.S. Geological Survey, Lawrence, Kan.

Rudolph, B. (2002). "Incorporating hysteresis into concentration-discharge models." MS thesis, Dept. of Civil and Environmental Engineering, Tufts Univ., Medford, Mass.

Serrano, E., Moreno, B., Solaun, M., Aurrekoetxea, J. J., and Ibarluzea, J. (1998). "The influence of environmental factors on microbiological indicators of coastal water pollution." *Water Sci. Technol.*, 38(12), 195–199.

United States Environmental Protection Agency (USEPA). (1999). "Review of potential modeling tools and approaches to support the BEACH program." *EPA823-R-99-002*, Washington, D.C.

United States Environmental Protection Agency (USEPA). (2000). "Improved enumeration methods for the recreational water quality indicators: *Enterococci* and *Escherichia Coli*." *EPA-821/R-97/004*, Office of Science and Technology, Washington, D.C.

United States Environmental Protection Agency (USEPA). (2003). "Bacterial water quality standards for recreational waters (freshwater and marine waters)." *EPA-823-R-03-008*, Office of Water, 4305T, ⟨http://www.epa.gov/OST/beaches/local/sum2.html⟩ (April 10, 2005).

United States Geological Survey (USGS). (2003). "Real-time data for USGS 01102500 Aberjona River at Winchester, MA." ⟨http://waterdata.usgs.gov/ma/nwis⟩, USGS, Reston, Va.