Modeling Nonpoint Source Bacteria Loading in Urban Rivers

A Thesis

submitted by

Matthew G. Heberger

In partial fulfillment of the requirements

for the degree of

Master of Science

in

Civil and Environmental Engineering

TUFTS UNIVERSITY

November 2003

Advisor: Steven C. Chapra

Abstract

Two methods have been developed to model bacteria loads and concentrations in the Mystic River in northeastern Massachusetts, where bacteria counts at recreational sites frequently exceed state standards for swimming and boating. The first, a stochastic method, involved fitting multivariate regression models to observed bacteria counts. Second, a deterministic lumped-parameter daily watershed model was developed. Based loosely on the Generalized Watershed Loading Functions model (GWLF), the watershed model represents a compromise between empirical statistical methods and complex simulation models such as HSPF or SWMM. Fecal coliform and *Enterococcus* bacteria concentrations were measured daily during the summers of 2002–2003. Simultaneous measurements of depth, temperature, specific conductivity, dissolved oxygen, pH, and turbidity were obtained from continuous in-stream monitoring equipment. Weak relationships were found between water quality parameters and bacteria. Independent variables with greater predictive power include streamflow, precipitation depth, and the time elapsed since the last rainfall. Multivariate linear regression models developed with 2002 data predicted bacteria concentrations with adjusted R^2 values of 0.55–0.82 for river sites. Regression models did not explain as much of the variability in bacteria concentration in a sampled lake or in the slack water of the lower basin, with typical adjusted-R^2 values of 0.39. The watershed model predicted daily average bacteria concentrations for river sites only with values of R^2 between 0.47 and 0.77. Comparing the statistical and simulation modeling approaches, the regression models yielded a slightly better fit to the calibration data set. In a split-year confirmation, however, the watershed model outperformed the regression equations in matching 2003 observations, with a lower root mean square error and a higher Nash-Sutcliffe model efficiency. The deterministic approach appears to have distinct advantages: a more realistic pollutograph, and better predictions of concentrations in the two to three days following runoff loads. Further, changes in the watershed would invalidate the ‘black-box’ regression model, while the simulation model may still perform well after one or more coefficients are adjusted.

Acknowledgements

This work was made possible by an EMPACT grant from the U.S. Environmental Protection agency.

I would like to thank my advisor Professor Steve Chapra ("Dr. C") for his guidance and intellectual contributions to the model development. Thanks to Professor Rich Vogel for assistance with statistics and useful background on rainfall-runoff modeling.

I would also like to thank the entire EMPACT team, especially Professor John Durant and fellow graduate students Cris Perez, Paul Lattanzi, and Christine Tabak for valuable assistance in the field.

Kim Oriel, Michele Cutrofello, and Laurie Cormier are due thanks for their diligent work in sampling and analyzing bacteria, more often than not at night and on weekends.

Table of Contents

List of Figures

List of Tables

1. General Introduction

The Mystic River, in eastern Massachusetts, is one of three major rivers which drain to Boston Harbor. It is estimated that nearly 500,000 people live in the river's 76 square mile watershed (Mystic River Watershed Association, 2002), making it among the most densely populated in the state. The watershed, which contains more than 40 lakes and ponds and many wetland areas, is also home to one of the largest urban river herring runs in the Northeast.

The region's industrial history has shaped the river; mills, shipyards, automobile plants, tanneries, and other industries have left their mark on the river and its banks. Water quality problems of greatest concern in the Mystic include toxic chemicals from contaminated waste sites, excess nutrients, noxious aquatic plants, organic enrichment and low dissolved oxygen, oil and grease, and pathogens (Massachusetts DEP, 1998, pages 58-59).

In spite of water quality problems, a great deal of recreation takes place in the Mystic watershed, including boating, fishing, and swimming. Potentially disease-causing bacteria and viruses, which are the focus of this thesis, come from three main sources: polluted runoff, combined and sanitary sewer overflows, and "illicit connections." An illicit connection is a sanitary waste line at a home or business that is connected to a storm drain, whether by accident or negligence. Pathogen loading to the river from other sources is largely driven by precipitation. Heavy rainfall can cause old combined sewers to overflow, sending untreated human waste into the river and its tributaries.

Runoff from land areas in the watershed is also a significant pathway by which bacteria enter surface waters. Urban runoff typically contains a variety of pollutants, including organic matter, oil and grease, nutrients, pesticides and herbicides, as well as bacteria and viruses (Horsley and Witten, 1999). Bacteria in runoff can come from waste from pets and wildlife or may be attached to soil particles. In an urban setting, where storm drains are designed to get water away from roads and buildings as quickly as possible, nonpoint source pollutants are quickly delivered to surface waters, with little time for settling or decay to occur.

In Massachusetts, water quality standards for swimming and boating are based on two kinds of indicator bacteria, fecal coliforms and *Enterococcus* bacteria (Commonwealth of Massachusetts, 1997a and 1997b). In both cases, the strains being tested do not cause disease, but their presence in the water is associated with other pathogenic bacteria and viruses (APHA, 1998).

Fecal coliforms are a category of bacteria that include several species that grow in the digestive system of warm-blooded animals. Fecal coliforms are widely used as an indicator for bacterial contamination. Massachusetts standards for contact recreation (i.e., swimming) state that single-sample concentrations of fecal coliform bacteria should be below 200 organisms per 100 mL. The use of bacteria of the genus Enterococcus as an indicator organism is relatively new. Epidemiological studies have shown that enterococci are better correlated with bacteria that cause gastroenteritis (Bowie et al., 1985). The standard states that, for bathing waters, there should be fewer than 61 organisms per 100 mL (Commonwealth of Massachusetts, 1997a).

Testing for bacteria is difficult, expensive, and slow. There is typically a 24–48 hour lag time between sample collection and when the results are available. For example, the most active swimming beach in the watershed is Sandy Beach, on Upper Mystic Lake. The Boston-area parks agency, the Massachusetts Department of Conservation and Recreation, Division of Urban Parks and Recreation (formerly the Metropolitan District Commission), measures Enterococcus bacteria levels at Sandy Beach once each week. However, bacteria counts at the beach increase rapidly following rainstorms; weekly sampling is insufficient to capture rapid changes in water quality at the beach.

The goal of this research was to develop an early-warning predictive model to fill in the gaps between samples and during the lag time when test results are pending. The use of the words "prediction" and "forecast" to refer to present conditions may be confusing, but make sense in context: present concentrations are unknown until a lab returns results 24 hours later.

In order to be useful in making timely decisions, a model must provide output at an appropriate time scale. Huber and Dickinson (1992) describe the temporal resolution required for a watershed loading model based on the receiving water and pollutant of interest, as reported in Table 1.1. Data show a fast response time for streams and rivers, and beaches in particular.

Table 1.1 Required temporal detail for receiving water analysis (from Huber and Dickinson, 1992)

Type of Receiving Water	Key Constituents	Response Time
Lakes, Bays	Nutrients	Weeks – Years
Estuaries	Nutrients, DO(?) [sic]	Days – Weeks
Large Rivers	DO, Nitrogen	Days
Streams	DO, Nitrogen, Bacteria	Hours – Days
Ponds	DO, Nutrients	Hours – Weeks
Beaches	Bacteria	Hours

1.1 Thesis Overview

This thesis describes two different approaches that were developed to model pathogen indicator bacteria concentrations in the Mystic River. First, the data collection efforts that went into this modeling study are described in section 2, General Methods. Sampling included bacteria enumeration and installation and maintenance of a network of real-time water quality monitoring equipment.

Following is a discussion of the development of the predictive models. The first of the modeling approaches is the stochastic, or statistical, approach described in section 3. In this approach, multivariate regression models were developed to predict bacteria levels based on measurements of climate and water quality variables. An innovative approach to incorporating water quality data into concentration-discharge models is explored which involves evaluating parameter rates of change with respect to time. Regression models were found to be useful at sites where the hydraulics are most river-like, with a model R^2 of 0.82 for the Aberjona River, and 0.55 for Alewife

Brook. At sites which more closely resemble a lake or basin, models had much less predictive ability; e.g., R^2 = 0.39 at Sandy Beach on Upper Mystic Lake.

In the first approach, mathematical equations were written to turn the model inputs (precipitation, streamflow, etc.) into outputs (bacteria concentration). No special significance is attached to the form of the equation or its coefficients, other than the fact that it produces the desired output.

By contrast, the second modeling approach may be called 'deterministic' because it seeks to describe and quantify actual physical processes, such as runoff, infiltration, pollutant washoff from land surfaces, mixing, and decay. This approach, described in section 4, focused on building a lumped-parameter watershed simulation model to predict streamflow and bacteria. The model, which was intended to be a good deal simpler than other complex simulation models currently in use, combines a lumped parameter hydrologic model with buildup and washoff loading functions for bacteria. The simulation model was found to produce results nearly as good as the regression models, with values of R^2 up to 0.76 for the Aberjona River, and 0.52 at the Alewife Brook.

Results from the two modeling techniques are compared in section 5, General Results. In a split-year confirmation, both the regression model and the simulation model performed relatively poorly. However, by some measures, the simulation model outperforms the regression model, with a higher Nash-Sutcliffe model efficiency (*E*), and a lower root-mean square error (RMSE).

2. General Methods

Bacteria concentrations and water quality data were collected at a number of locations in the Mystic watershed during the summers of 2002–03 as a part of the EPA-sponsored project, "Real-Time Water Quality Monitoring and Modeling for Equitable Recreation on the Mystic River," which was funded under the program entitled, "Environmental Monitoring for Public Access and Community Tracking (EMPACT)."

2.1 Bacteria Sampling

During the months of May through August in 2002 and 2003, fecal coliform and *Enterococcus* measurements were made as described in Oriel (2003). Briefly, measurements were made using standard methods (APHA, 1998), which involve collecting and filtering water samples and enumerating colonies after incubation. In general, water samples were collected and analyzed five days each week, including some weekends.

2.2 Sampling Locations

Samples were collected at locations where there is recreational activity (Sandy Beach and the Boys & Girls Club), as well as at other sites, including the Aberjona River at the USGS gaging station, and along Alewife Brook, where very high levels of bacterial contamination are frequently observed. Sample locations are reported in Table 2.1 and shown in Figure 2.1.

The Blessing of the Bay site in Somerville is home to the Boys & Girl's Club, where recreational boating takes place. Sandy Beach, on Upper Mystic Lake, is a popular swimming area during the summer months.

Alewife Brook is one of the most polluted tributaries of the Mystic. The drainage area of Alewife Brook consists mostly of residential, commercial, and industrial land uses, in addition to two large ponds, Spy Pond and Fresh Pond. The Alewife is the location of the majority of the combined sewer overflows (CSOs) in the watershed (MyRWA, 2002). Even when CSOs are not activated, runoff from the Alewife's urbanized watershed brings many pollutants into the Brook (Massachusetts DEP, 1998). Because the channel is poorly flushed and often stagnant, water quality is very poor; anoxic waters and foul odors are common.

The station furthest downstream is at the Amelia Earhart Dam. Built in 1969 as a flood-control dam, it prevents seawater in Boston Harbor from entering the upper Mystic basin. Further, the dam operators regulate the amount of freshwater stored in the river basin. When heavy rain is forecast, dam operators pump water out of the river basin to lower the water levels and help prevent upstream flooding. At other times, gates are opened daily at low tide to allow water to flow from the river basin out into the harbor.

Approximate drainage areas were determined for some sites by the use of geographic information system (GIS) software, making use of the *Sub-Basins* datalayer created by the state agency MassGIS. Because this determination of watershed boundaries was automatically derived based on digital elevation maps, inaccuracies may be present, especially as the Mystic is an urban basin, where exact drainage patterns are largely determined by the layout of storm drains.

Table 2.1 Location of real-time water quality monitoring stations

Location	Latitude	Longitude	Approximate Drainage Area
Aberjona River at USGS Gage*	71° 08’ 22" W	42° 26’ 50" N	61×10^6 m^2 (23.6 sq. mi.)
Sandy Beach	71° 8’ 43” W	42° 26’ 22”N	
High Street Bridge	71° 8’ 34” W	42° 25’ 12” N	
Alewife Brook	71° 7’ 60” W	42° 24’ 28” N	23×10^6 m^2 (8.9 sq. mi.)
Boys & Girls Club	71° 5’ 21” W	42° 23’ 56” N	140×10^6 m^2 (53 sq. mi.)
Amelia Earhart Dam	71° 4’ 33” W	42° 23’ 46” N	160×10^6 m^2 (63 sq. mi.)

*Only precipitation and streamflow measurements (USGS, 2002–03) are available at the Aberjona River site. Water quality (and not streamflow) was monitored at the other five sites.

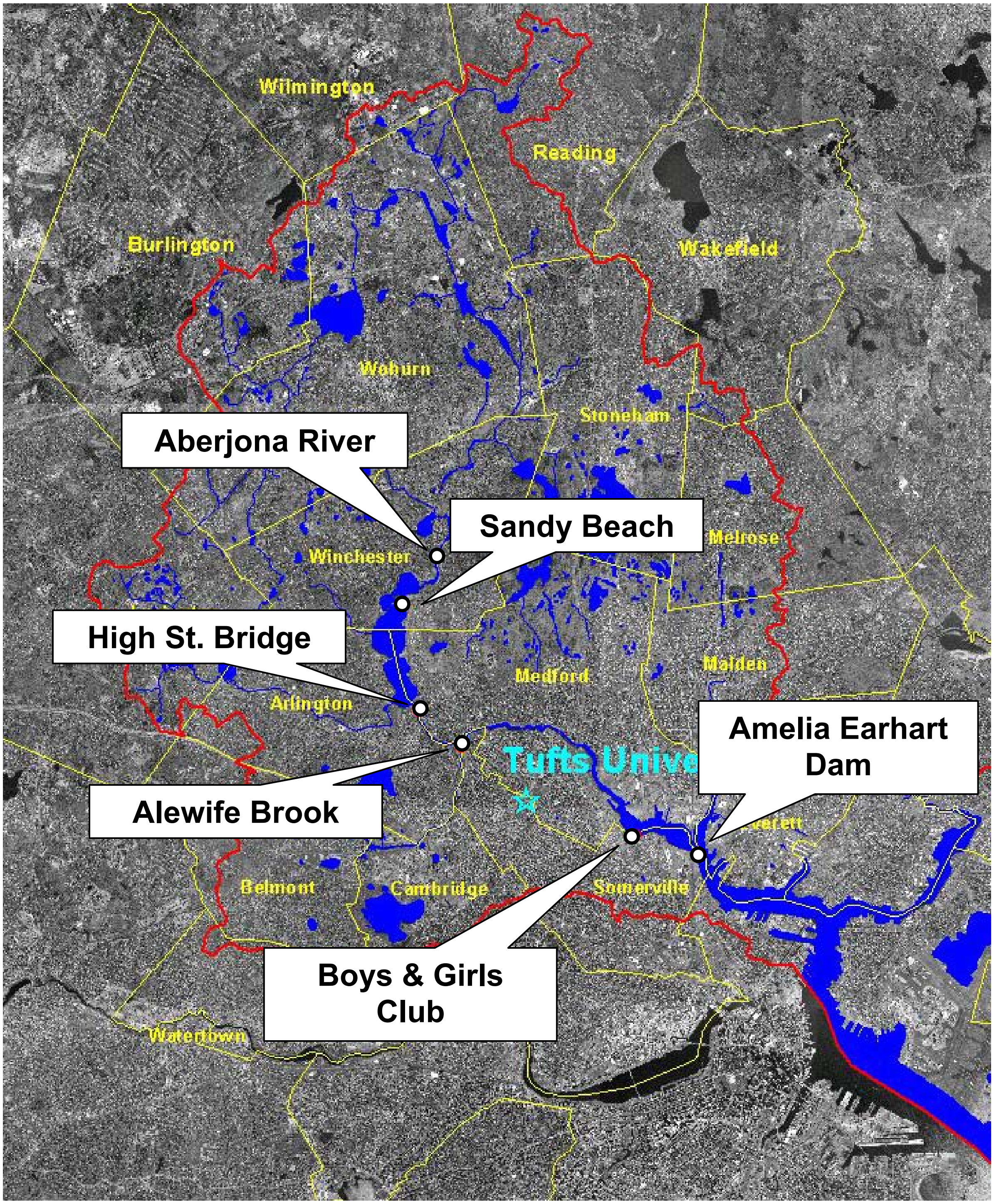


Figure 2.1 Map of bacteria sampling and water quality monitoring stations

2.3 Water Quality Monitoring

Continuous, real-time surface water quality monitoring stations were installed and maintained at five locations in the Mystic River watershed. Locations of the monitoring stations are reported in Table 2.1. Water quality was measured at 15-minute intervals using equipment from YSI, Inc. Probes that measure individual water quality parameters are housed in a ‘sonde’, pictured in Figure 2.2.

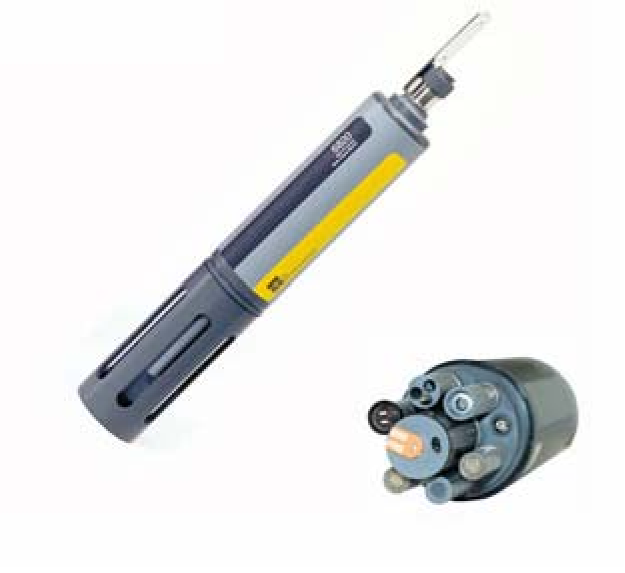


Figure 2.2 YSI water quality monitoring sonde used in the study

The sonde was equipped with a set of sensors to measure depth, temperature, conductivity, pH, dissolved oxygen, and turbidity. The reporting limit for each parameter (i.e., number of digits reported by the sensor) and the manufacturer’s stated accuracy are reported in Table 2.2 (YSI, 2001). Prior to use in the modeling study, all data passed through a stringent “Quality Assurance/Quality Control (QA/QC)” process to ensure that measurements are accurate and representative. Procedures for record computation given by Wagner et al. (2000) were generally followed.

Table 2.2 Reporting limits for the YSI sensor data

Parameter	Sensor Type	Reporting Limit	Manufacturer's Stated Accuracy
Depth	Stainless steel strain gauge	0.001 ft	± 0.06 ft*
Temperature	Thermistor	0.01 ºC	± 0.15 ºC
DO%	Rapid Pulse - Clark type, polarographic	0.1%	± 2%
DO mg/L	" "	0.01 mg/L	± 2% or 0.2 mg/L, whichever is greater
Turbidity	Optical, 90º scatter, with mechanical cleaning	0.1 NTU	± 5% or 2 NTU, whichever is greater
pH	Glass combination electrode	0.01 units	± 0.2 units
Conductivity	4 electrode cell with autoranging	4 digits (i.e.: 1.414 ms/cm or 14.14 ms/cm)	± 0.5%

*For unvented depth sensor, assuming constant atmospheric pressure, additional uncertainties created by barometric pressure correction

2.4 Modeling Approaches

Engineers use a variety of approaches for modeling nonpoint source pollution, such as loading rates (see for instance, Chandler, 1995), probabilistic methods (Hydroscience, 1979), or regression models (e.g., Tasker and Driver, 1990, Eleria, 2002). At the other extreme are highly detailed simulation models, such as SWMM, HSPF, STORM, CREAMS/GLEAMS, SWRRB, etc. (See for instance, Interstate Commission on the Potomac River Basin, (2002), for an example of bacteria loading simulated with HSPF). Simulation models often rely on assumptions such as default parameters, and may or may not be calibrated to fit a limited number of observations. Frequently, available data were collected for some other purpose by regulatory or public health agencies, and therefore may not be ideal for use in a modeling study (S. Chapra, personal communication, 2002).

Donigian et al. (1996) give an overview of these approaches, with an emphasis on comparing model capabilities and resource requirements. Huber and Dickinson (1992, page 127) caution simulation model users against placing too much confidence in their output: "Simulation of urban runoff quality is a very inexact science if it can even be called such. Very large uncertainties arise both in representation of the physical, chemical, and biological processes, and in the acquisition of data and parameters for model algorithms… Such uncertainties can be dealt with in two ways. The first option is to collect enough calibration and confirmation data to be able to calibrate the model equations used for quality simulation."

"The second option is to abandon the notion of detailed quality simulation altogether… statistical methods recognize the frustrations of physically-based modeling and move directly to a stochastic result […] but they are even more dependent on available data than methods such as those found in [simulation models such as] SWMM. That is, statistical parameters such as mean, median and variance must be available from other studies in order to use the statistical methods. Furthermore, it is harder to study the effect of controls and catchment modifications using statistical studies."

An example of a regression model for bacteria is shown in Figure 2.3 for a Total Maximum Daily Load (TMDL) study in Washington State (Pelletier and Seiders, 2000). The figure shows the time series of observed and predicted fecal coliform concentrations in the West Fork Hoquiam River, Washington, from 5/1/97 – 4/30/98; however, the authors do not report model fit statistics (e.g., R^2), or even whether the model is a significant improvement over assuming an average loading rate.

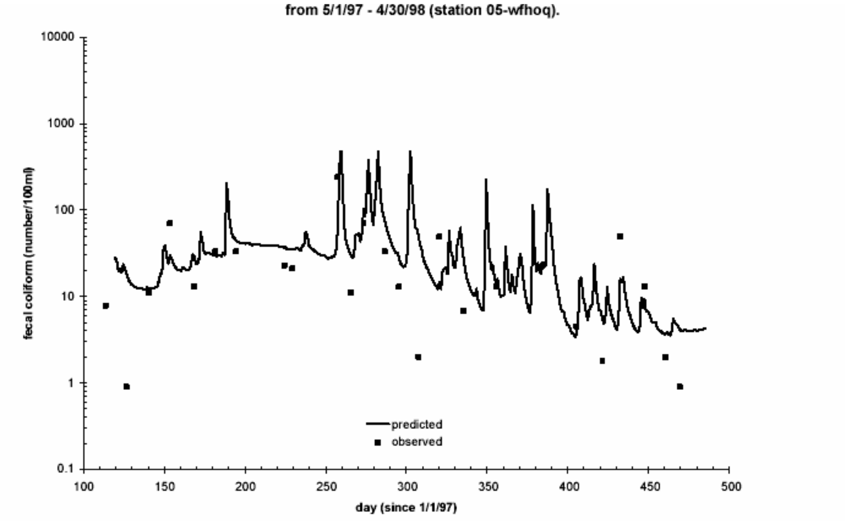


Figure 2.3 Regression model result from Grays Harbor, Washington TMDL study (from Pelletier and Seiders, 2000)

Figure 2.4 is an example of simulation model results (Interstate Commission on the Potomac River Basin, 2002). The HSPF simulation output developed for a fecal coliform bacteria TMDL study of Goose Creek, Virginia is plotted with only n = 22 bacteria samples over six years.

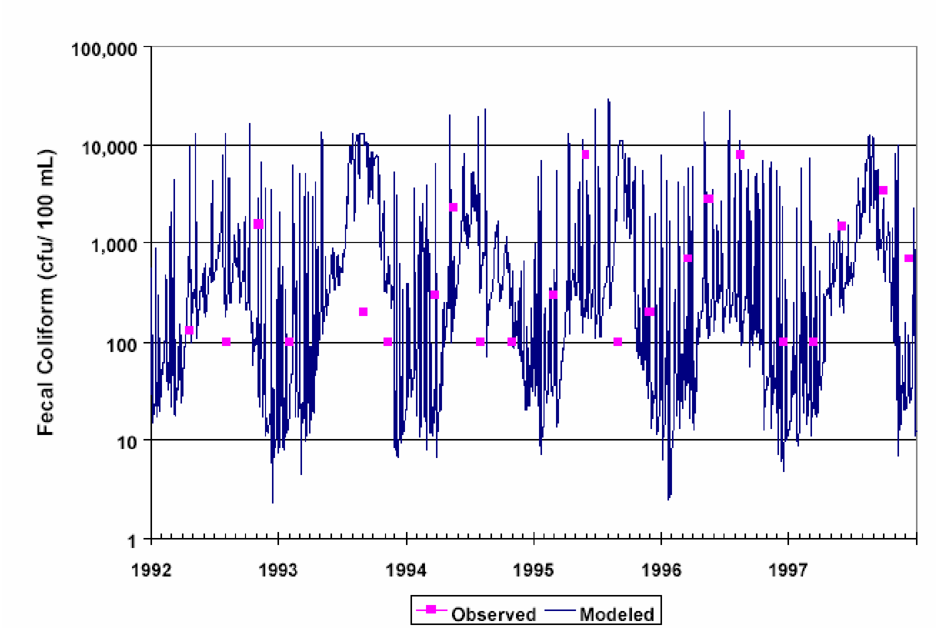


Figure 2.4 HSPF model output for fecal coliform concentration in Goose Creek, Virginia (from Interstate Commission on the Potomac River Basin, 2002)

It is worth asking whether the complexity of detailed simulation models is justifiable when faced with the sparse data typically available for calibration. One agency warned that model output should be viewed skeptically, as "highly complex, data-intensive watershed models such as HSPF [give only] the illusion of detail and accuracy" (York Watershed Council, 2002).

2.5 Model Evaluation Framework

The simplest qualitative methods for judging the fit of a model involve visually comparing plots of modeled and observed values. (Note that in the following discussion the terms *modeled*, *calculated*, and *predicted* are used interchangeably.) In addition to time series plots, it is useful to construct scatterplots of observed versus modeled values. A theoretical 1:1 line on the plot shows where observations are equal to the model. A plot where the points are tightly clustered about the 1:1 line indicates a strong fit.

In addition, it is useful to plot the cumulative distributions of model and observed values, also sometimes called quantile plots (see for example Maidment, ed., 1993, page 8.27, and 17.8-9). These figures plot cumulative frequency of discharge (or concentration) versus the percentage of time that discharge (or concentration) is exceeded. In creating these curves for this thesis, ranked data were plotted versus their empirical exceedance probability, P_{i}, based on the Weibull plotting position (Helsel and Hirsch, 1992, page 23):

$$P_{\mathrm{i}} = \frac{i}{n+1} \tag{2.1}$$

where i is the rank of the observation and n is the number of observations. An example of a concentration-duration curve is shown in Figure 2.5. In this example (created from a simulation model run for the Aberjona River data in the summer of 2002), the plot shows that the model does not accurately predict concentrations at the high end of the range.

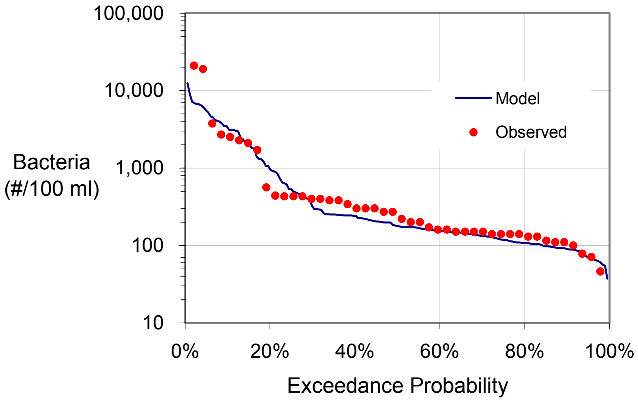


Figure 2.5 Example of a concentration-duration curve constructed for modeled and observed bacteria concentration

Summary statistics calculated for observed and predicted values are a useful check on the model. A model should reproduce the average of observations. In this thesis, the median and the geometric mean are most frequently reported. As these are 'resistant' measures of central tendency, they are less likely to be affected by outliers. Thus, they are appropriate for use with bacteria data, which often vary over several orders of magnitude. Further, a model should reproduce the variance of observed data, as measured by the standard deviation or interquartile range (the difference between the 75th percentile and the 25th percentile).

The error of a modeled value, or its residual, is calculated as follows:

$$e_i = y_i - \hat{y}_i \qquad (2.2)$$

where:

y_i = i^{th} observation

$\hat{y}_i$ = model prediction corresponding to the i^{th} observation

It is customary to square the residuals before adding them to prevent positive and negative errors from canceling each other out. The sum of squared errors, SSE, is:

$$SSE = \sum_{i=1}^{n} e_i^{\,2} \tag{2.3}$$

Further, this quantity may be normalized by the number of observations, and its square root taken. The advantage of the root mean square error (RMSE) is that it is in units of the observations (e.g., cfs for discharge data) and is more readily interpreted.

$$RMSE = \sqrt{\frac{SSE}{n}} \tag{2.4}$$

Plots of model residuals, e_i, versus time or the predicted variable are a valuable display of model fit. The goal is model residuals that are independent and randomly distributed. Helsel and Hirsch (1992, page 233) state that, "a good residuals pattern, one with no relation between residuals and time, will look similar to [Figure 2.6] – random noise. If on the other hand structure in the pattern over time is evident, seasonality, long-term trend, or correlation in the residuals may be the cause. If there is structure, or autocorrelation, present in the model residuals, that is evidence that the model does not describe the behavior of the data."

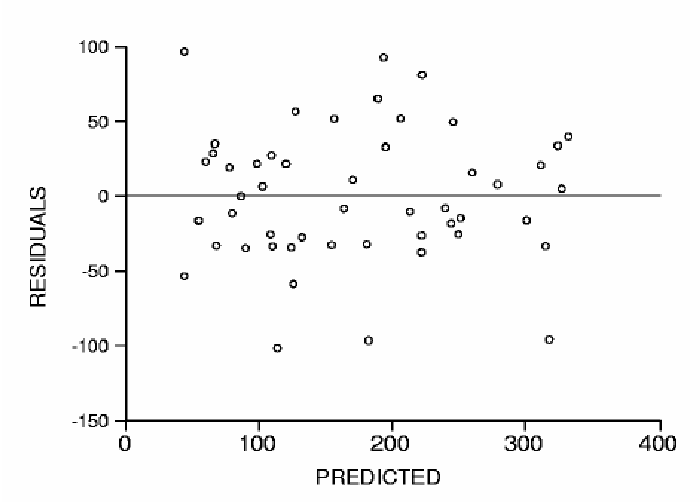


Figure 2.6 Example of a residuals plot for a good model (from Helsel and Hirsch, 1992)

Further quantitative measures for evaluating the models were applied, according to ideas put forth by Thomann (1982). While his discussion related to surface water quality models such as dissolved oxygen and eutrophication models, his ideas are directly applicable to evaluating a watershed loading model. Thomann suggested that the most useful way to evaluate model fit is to plot the observed and predicted values, and calculate the regression equation and its accompanying statistics. He warns against using the coefficient of determination (R^2) as the sole determinant of model fit, however. In a number of cases a high R^2 belies a biased model. The best model is one with a slope b =1 and an intercept a = 0. Figure 2.7 illustrates three cases where R^2 = 1, but the model has a significant bias, indicated by an intercept $a \neq 0$ or slope $b \neq 1$.

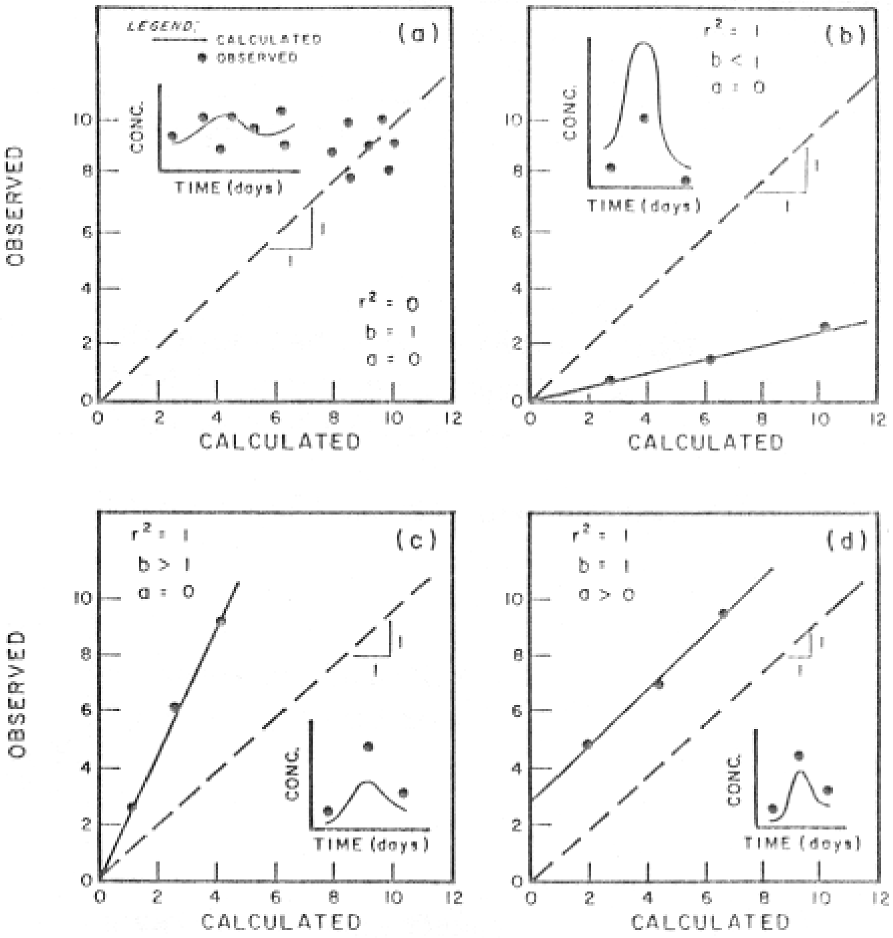


Figure 2.7 Possible cases in regression between calculated and observed values (from Thomann, 1982)

Nash and Sutcliffe (1970) proposed a measure of model efficiency in order to establish a framework for evaluating rainfall-runoff models. Nash and Sutcliffe defined the 'initial variance' as:

$$F_0^2 = \sum \left(Q - \bar{Q} \right)^2 \tag{2.5}$$

Further, their 'index of disagreement' is the same as the sum of squared errors:

$$F^2 = \sum \left(Q - \hat{Q} \right)^2 \tag{2.6}$$

where:

Q = observed flow

$\bar{Q}$ = average flow

$\hat{Q}$ = predicted flow

The Nash-Sutcliffe model efficiency is defined as:

$$E = \frac{F_0^2 - F^2}{F_0^2} \tag{2.7}$$

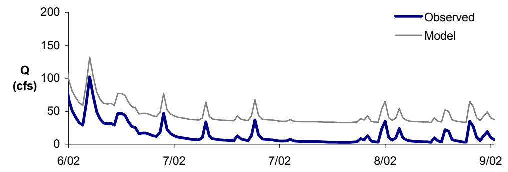
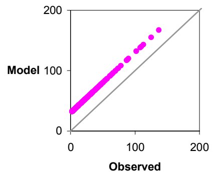
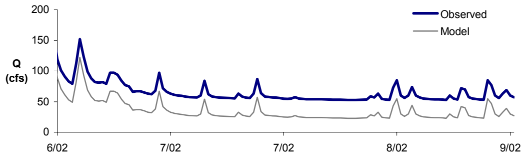
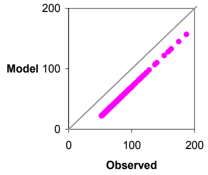
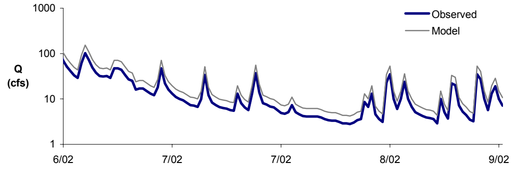
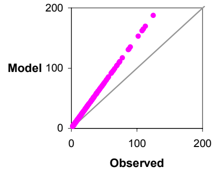
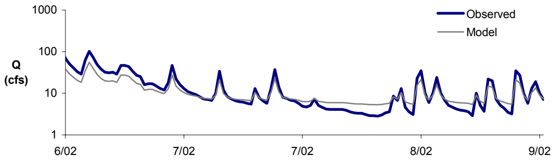
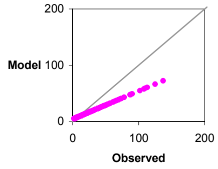
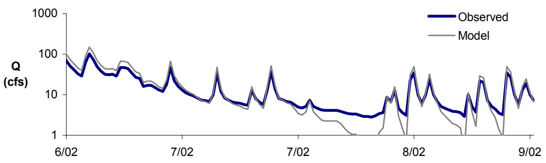
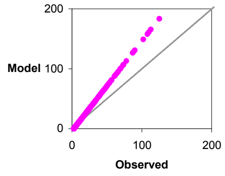
Some confusion was created by their original choice of the symbol R^2 for this statistic in their 1970 paper. In the literature, it is variously referred to as R^2, *E*, or *NS*. In this thesis, the letter *E* is used.

The Nash-Sutcliffe model efficiency is analogous to the coefficient of determination (Beven, 2001), but there are important differences, as shown in Table 2.3. Further, a number of theoretical cases are illustrated in Table 2.4 where *E* outperforms R^2 as a determinant of model fit.

Table 2.3 Comparison of the coefficient of determination (R^2) and the Nash-Sutcliffe model efficiency (E)

Nash-Sutcliffe Model Efficiency, E	Coefficient of Determination, R^2
$E = \dfrac{\sum\left(Q-\bar{Q}\right)^2 - \sum\left(Q-\hat{Q}\right)^2}{\sum\left(Q-\bar{Q}\right)^2}$	$R^2 = \dfrac{\sum\left(\hat{Q}-\bar{Q}\right)^2}{\sum\left(Q-\bar{Q}\right)^2}$
$-\infty < E < 1$	$0 < R^2 < 1$
$E = 1$ means model is perfect, i.e., $\hat{Q}_i = Q_i$ for all i	$R^2 = 1$ means model predictions are *perfectly correlated* with observations (but says nothing about bias)
$E < 0$ means model performs worse than assuming $\hat{Q} = \bar{Q}$	

Table 2.4 Comparison of model diagnostics *E* and R^2 ability to detect systematic model error

Description of Bias	Time Series Plot	Model vs. Observations	Coefficient of Determination R^2	Nash-Sutcliffe E	RMSE (cfs)
Model overpredicts all flows			1.00	-15.24	30.0
Model underpredicts all flows			1.00	-15.24	30.0
Model error proportional to flow $\hat{Q} = 1.5Q$			1.00	0.47	5.4
Systematic bias: model overpredicts low flows, underpredicts high flows			1.00	0.75	3.7
Systematic bias: model underpredicts low flows, overpredicts high flows			1.00	0.75	3.7

While the coefficient of determination (R^2) is useful in evaluating the fit of an ordinary least squares regression equation, the Nash-Sutcliffe model efficiency (E) is a better determinant of model fit when comparing observed and modeled time series. Nash and Sutcliffe originally proposed this statistic for evaluating rainfall-runoff models, and it is usually found in the hydraulic and hydrologic literature. However, there is no reason why it cannot be applied to concentration data, nor is there any difficulty in doing so.

In summary, a number of qualitative and quantitative measures were used to evaluate and compare the performance of predictive models for bacteria. Plots of observed and predicted concentrations are the first and most important check on a model. Care must be used in interpreting model fit statistics; those which are dependent on units of observations and sample size (e.g., sum of squared errors) cannot be readily compared from one data set to the next. Even using a normalized quantity such as the coefficient of determination (R^2) may obscure bias in the model. Finally, the Nash-Sutcliffe model efficiency (E) was shown to be most useful in comparing observed and modeled time series.

3. Multivariate Linear Regression Models for Bacteria

3.1 Introduction

Several investigators have developed regression models of the following form for estimating bacteria concentrations:

$$\log(C) = a + b_1x_1 + b_2x_2 + ...b_nx_n + \varepsilon \tag{3.8}$$

where:

$\log(C)$ = base-10 logarithm of bacteria concentration

$x_1, x_2, ..., x_n$ = independent variables (climate, hydrology, water quality)

$b_1, b_2, ...b_n$ = slope with respect to the independent variable x_i

a = intercept

ε = random error component

Eleria (2002) reviews several papers in which researchers predicted bacteria concentrations in water bodies using multivariate linear regression models. Some were more successful than others: R^2 values reported in the literature range from 0.01 to 0.80. Independent variables typically included precipitation and streamflow. Significant predictor variables in coastal waters included wind speed and direction; and at beaches, bird populations, number of bathers, and the presence of floatables. Eleria built multivariate models based on streamflow and precipitation to predict fecal coliform concentrations in the Charles River with adjusted R^2 values of 0.50 to 0.60.

Christensen (2001) developed a regression equation for Rattlesnake Creek in south-central Kansas to predict fecal coliform levels. The model, applicable for the summer months (April 1–October 31), was based on n = 18 samples and included the independent variables water temperature and turbidity. Christensen reported a coefficient of determination $R^2 = 0.66$.

Christensen et al. (2002) developed linear regression equations for the estimation of fecal coliform bacteria (as well as total nitrogen and total phosphorus) at four Kansas stream-gaging stations. The equations were of the following form:

$$\log_{10}\left(C_{FC}\right) = a_0 + a_1 \sin\left(\frac{4\pi J}{365}\right) + a_2 \cos\left(\frac{4\pi J}{365}\right) + a_3 \log_{10}\left(Q\right) + a_4 T + a_5 SC \tag{3.9}$$

where

$J =$ Julian Day (1–365)

$T =$ Water temperature (ºC)

$SC =$ Specific conductance (mS/cm)

The sine and cosine terms represent a seasonal trend in bacteria concentrations. The authors state that an independent variable was used only when there is a physical justification for its inclusion. They justify the use of a seasonality term based on the heavy bacteria load in spring runoff from cattle-producing areas. With sample sizes of n between 17 and 102, they reported values of R^2 between 0.40 and 0.73.

McLellan and Salmore (2003) studied the causes of fecal contamination, as indicated by the presence of *E. coli*, at beaches in Milwaukee Harbor near the confluence of the Milwaukee River, which is affected by several CSO outfalls. They found significant relationships with several parameters, including precipitation, hours since last rainfall, and wind speed and direction. They report adjusted R^2 values from 0.03 to 0.29.

Francy and Darner (2003) developed regression equations to forecast *E. coli* levels at bathing beaches in Ohio. They included the independent variables streamflow, rainfall, wave height, and number of birds on the beach at the time of sampling. The equations were developed with $n = 100$ samples, and have values of R^2 between 0.32 and 0.40.

Clark and Norris (2000) conducted research similar to that in this thesis on eight rivers in Wyoming. They sought to explain variability in fecal coliform bacteria through measurements of water quality measurements from a continuous monitoring program. Clark and Norris found significant correlations between bacteria concentration and streamflow, dissolved oxygen, specific conductance, pH, and water temperature, although they did not go on to develop regression equations.

Two patterns emerge from the literature. First, regression models for bacteria have been more successful in streams and rivers than in lakes or estuaries. Second, physical and chemical water quality parameters, which are controlled by complicated mechanisms, are imperfectly correlated with bacteria; nevertheless, useful site-specific relationships have been found.

3.2 Methods

A number of independent variables were collected for input to the model:

1. Climate data such as temperature and precipitation (USGS, 2002)
2. Flow data for the Aberjona River, an upstream tributary of the Mystic

3. Physical and chemical water quality parameters: depth, temperature, conductivity, pH, dissolved oxygen, and turbidity measured at 15-minute intervals

The incompleteness of the water quality records for summer 2002 made it difficult to include them in the analysis of each site. Researchers considering a continuous monitoring program are cautioned to allow several months to gain experience with operating and maintaining the equipment, and to develop protocols for data review and record computation. High-quality contemporaneous records of bacteria measurements and water quality exist for two of the five sites during 2002. At the High Street Bridge site, sensor data was available coinciding with n = 32 bacteria samples. In addition, coincident data for the Boys & Girls club comprised n = 37 samples. Nevertheless, a few problems limited the use of the complete data set at the latter, such as sensor failure, fouling, and inaccurate calibrations.

In initial trials, Enterococcus bacteria were more amenable to regression modeling; models for Enterococci had consistently higher values for R^2 and lower normalized errors. There is evidence that fecal coliform bacteria are less source-specific, originating from a variety of sources, and are able to survive in soil and sediments for extended periods (Bowie, 1989, chapter 8). In the Mystic River, observed fecal coliform concentrations in summer 2002 varied more widely, exhibiting much more scatter about their mean than Enterococcus. For these reasons, and because state standards for swimming are based on Enterococcus, this study focused on building models to predict their concentration.

3.2.1 Collecting and Processing the Input Data

In addition to water quality data, precipitation and streamflow were obvious candidates for inclusion in the regression model as independent variables. A description follows of how these variables were manipulated prior to inclusion in the regression models.

3.2.2 Streamflow

Continuous measurements of streamflow and precipitation are made by the USGS at 15-minute intervals, and are available at the website *http://water.usgs.gov* in near real-time; new information is typically available within four to six hours. Hence, a great deal of data is available, although the most useful way to assemble the data was not readily apparent. The goal was to extract as much information from the observed time series as possible in order to develop regressions model with the greatest predictive power.

For a bacteria concentration C_t, measured at time *t*, which measurement of streamflow has the most predictive power? The flow measured at the same time, Q_t? Or, if one supposes that antecedent streamflow is an important factor, perhaps the flow measured one hour prior to the sampling time, Q_{t-1}, is an important variable. Similarly, one may wish to look at the flow measured 2, 3, or 10 hours ago. The offset in hours or time lag of the flow measurement, is represented as Δt. Thus, the flow measured Δt hours before the sample is defined as $Q_{t-\Delta t}$.

An experiment was conducted to determine whether the lagged flows would improve to the regression model fit. A set of independent variables, $Q_{t-\Delta t}$, was created by looking up the streamflow at time $t-\Delta t$. The data were regressed against log(*Q*) and the correlation coefficient, R^2, was recorded. This experiment was repeated for Δt between 0 and 12 hours. A custom Excel/VBA routine was created to facilitate the repetitive and time-consuming work.

It was also hypothesized that the *average* flow over the last hour, determined over some interval might be an important variable. The flow averaged over X hours prior to the sample is represented as $\bar{Q}_{\text{X hr}}$. The flow was averaged over the period by taking a simple arithmetic mean[1]. For example, the 8-hour average flow is calculated as follows:

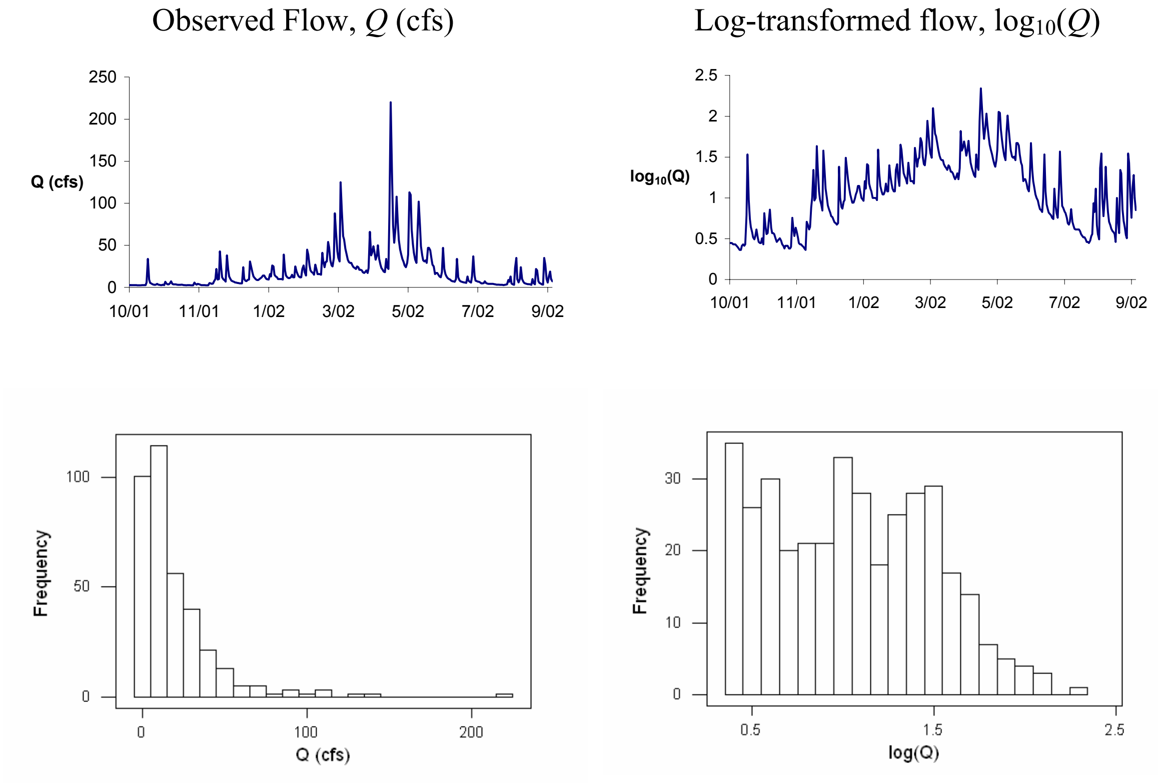
$$\bar{Q}_{\text{8 hr}} = \frac{\sum_{t-8}^{t} Q}{n_{\text{flow measurements}}} \tag{3.10}$$

An experiment was 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 hours. (The variable $\bar{Q}_{\text{0 hr}}$ is simply equal to instantaneous flow measurement at time t.)

Variables with a skewed distribution often benefit from being mathematically transformed or 'normalized' prior to analysis. In their text on hydrologic statistics, Helsel and Hirsch (1992) recommend "trying all sorts of transformations on […] variables to get a good and reasonable fit." They also state that looking at a graph of the variables is the best way to find an appropriate transformation. The effect of taking the logs of positively skewed data (such as streamflow) is to make the distribution more symmetric, expanding the distance between observations on the left side of the median, and contracting the distance between observations to the right of the median. Figure 3.1

[1] In an earlier experiment, it was found that determining the average flow using a numerical integration method such as the trapezoidal rule gives an almost identical result to simply taking the arithmetic mean. For the 24-hour average flow, the difference between the two methods less was usually less than 1%.

shows the effect of the log-transformation on discharge measurements for the Aberjona River during water year 2002. The time series is shown atop the histogram, which shows the empirical frequency distribution. It was investigated whether $\log_{10}(Q_t)$ was more highly correlated with bacteria.


Observed Flow, Q (cfs)

Log-transformed flow, $\log_{10}(Q)$

Figure 3.1 Log transformation of discharge data

3.2.3 Slope of the Hydrograph

Following promising work conducted by Rudolph (2002) investigating the feasibility of incorporating hysteresis into concentration-discharge models, it was supposed that the slope of the hydrograph may also make a useful independent variable.

The slope of the hydrograph was evaluated at the time of each sample as shown in Figure 3.2.

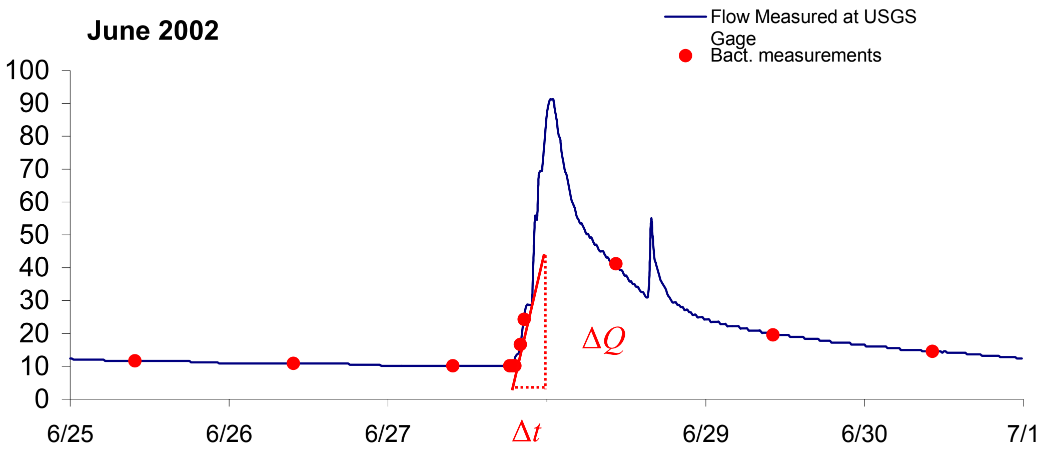


Figure 3.2 Example of hydrograph slope

Evaluating derivatives with environmental data often leads to inaccuracies (Chapra and Canale, 2002, page 638). A great deal of effort went into finding an acceptable technique for estimating the hydrograph slope that is both stable and accurate. Briefly, a computer routine was used to fit a polynomial of order *m* to *n* observations bracketing the time at which the sample was 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 hours worth of data. The slope is expressed as:

$$\frac{dQ}{dt} \approx \frac{\Delta Q}{\Delta t} \tag{3.11}$$

In this study, the derivative is evaluated by the slope of the best fit line (evaluated by ordinary least squares) through the preceding four hours of flow data. Streamflow measurements reported by USGS are in units of cubic feet per second (cfs or $\text{ft}^3 \cdot \text{s}^{-1}$). The units of the hydrograph slope in this thesis are cfs per day, or $\text{ft}^3 \cdot \text{s}^{-1} \cdot \text{day}^{-1}$.

Following the approach developed with other variables, it was hypothesized that the slope variable evaluated at a time in the past may have some additional predictive capability. The slope variable is evaluated at time $t - \Delta t$ and inserted in the regression equation. A computer experiment was conducted in order to determine the best time at which to evaluate the slope for inclusion in the regression model.

3.2.4 Time since Last Rainfall

The bacteria data for the Aberjona River show that there is a fairly steady decrease in concentration following the end of a rainstorm. This is demonstrated by the arrows in Figure 3.3. Within a few days after the end of the storm, concentrations decrease to 'background levels.' In an attempt to capture this phenomenon, a new variable, T_F, was created to represent the time since the end of the last rainfall. For each sampling time, t, the algorithm looks back in the precipitation time series to determine how long ago the last rain fell.

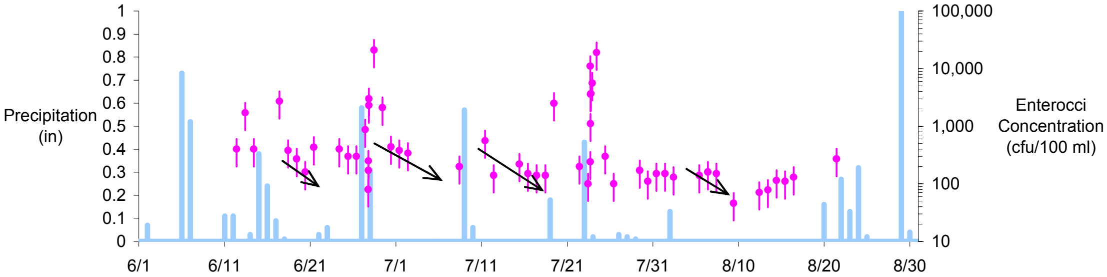


Figure 3.3 Observed precipitation and enterococci concentration in the Aberjona River, summer 2002

The units for the variable T_F are days, with the decimal expressing fractions of a day. Therefore, if the last rainfall was exactly 24 hours ago, $T_F = 1.00$. For an elapsed time of 1 day plus 1 hour and 15 minutes, $T_F = 1.052$.

In evaluating T_F, it was thought that trace amounts of precipitation should be ignored. A trace rainfall (0.01–0.04 inches) is unlikely to produce runoff, so it should not be enough to 'reset' T_F. Table 3.1 shows an example of T_F calculated with a threshold of 0.05 inches. Note that when the algorithm encounters a trace amount of rainfall less than 0.05 inches, the value of T_F continues to increase. However, when the rain gage has recorded a precipitation depth of greater than 0.05 inches, it 'trips' the algorithm, and T_F begins incrementing again from zero.

Table 3.1 Calculation of the variable T_F, time since last rainfall

Date & Time	Precipitation (inches)	Time since last rainfall, T_F (days hours:min)
…	…	…
5/9/02 6:45	–	06 22:15
5/9/02 7:00	–	06 22:30
5/9/02 7:15	0.01	06 22:45
5/9/02 7:30	–	06 23:00
5/9/02 7:45	–	06 23:15
…	…	…

$P <$ threshold 0.05 will *not* reset T_F

Date & Time	Precipitation (inches)	Time since last rainfall, T_F (days hours:min)
…	…	…
5/10/02 1:00	–	07 16:30
5/10/02 1:15	0.02	07 16:45
5/10/02 1:30	0.01	07 17:00
5/10/02 1:45	0.01	07 17:15
5/10/02 2:00	0.07	00 00:00
5/10/02 2:15	–	00 00:15
…	…	…

$P >$ threshold 0.05 *does* reset T_F

Changing the threshold for resetting T_F had a significant effect on the variable T_F. The time series plot in Figure 3.4 shows the effect of choosing a threshold of 0.02 inches versus 0.05 inches. In order to determine the best threshold for use in the regression, an experiment was conducted. The variable T_F was created with thresholds from 0.01 to 0.10 inches, regressed against bacteria concentration, and the coefficient of determination R^2 for each regression was recorded.

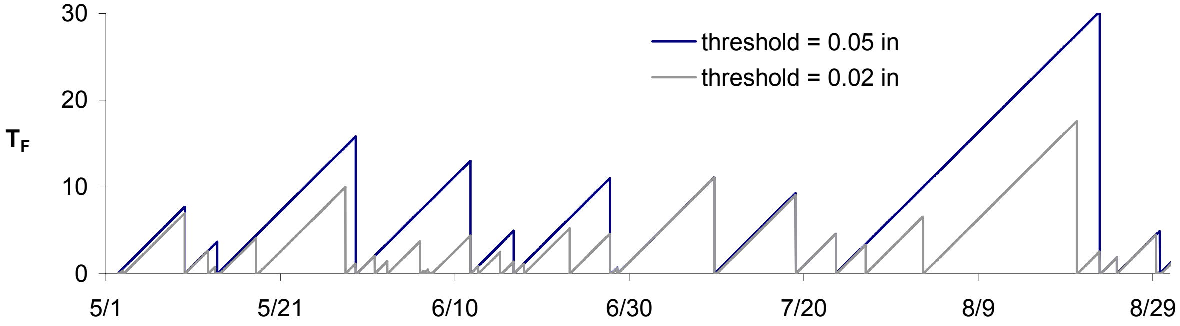


Figure 3.4 Effect of changing the threshold in the calculation of T_F, time since last rainfall

Initially, a threshold of 0.05 in was used to create T_F. After assembling the values of T_F corresponding to each of the bacteria samples, it was noted that the distribution of T_F was skewed to the right; with a preponderance of values between 0 and 10, and only a few between 10 and 30. As with discharge and precipitation, the variable was log-transformed, first adding 1 to each observation. Thus the new independent variable is $\log(T_F + 1)$.

3.2.5 Precipitation

As with the flow data, the best way to handle the precipitation data for input to the linear regression model was not immediately clear. As with the time-series of flow measurements, precipitation depth is available at 15-minute intervals. The idea of using a daily precipitation (i.e., the sum of precipitation from midnight to midnight) was immediately rejected. Because the majority of the bacteria samples were taken in the morning, it would be meaningless to take into consideration a rainstorm that occurred in the afternoon or the evening, hours after the sample was collected.

Therefore, an algorithm was developed to sum the precipitation over a given number of hours prior to the time the sample was collected. The number of hours is indicated by the following naming system: an 8-hour sum is P_8, a 12-hour sum P_{12}, etc. For a sample collected at time *t*, the eight-hour sum is:

$$P_8 = \sum_{t-8}^{t} P \tag{3.12}$$

The distribution of precipitation data was found to be highly skewed to the right, with a large number of low values and a few higher values. In fact, the majority of precipitation data are zeroes, which makes sense; most of the time it is not raining. In order to normalize the data, measurements were log-transformed, first adding 1 to each observation. The new independent variable resulting from this transformation is $\log(P_X + 1)$, where *X* is the number of hours in the sum.

It was hypothesized that the most recent precipitation should *not* to be factored into the regression. Stormwater runoff takes a certain amount of time to reach the basin outlet where the sample is collected. Therefore, it may make sense to exclude precipitation that has occurred in the last 30 minutes, 1 hour, or 2 hours. A second set of independent variables was created from the precipitation data including an ‘offset’ of Δt. An 8-hour sum, evaluated at t – 2 hours, is equivalent to:

$$P_8\big|_{t-2} = \sum_{t-10}^{t-2} P \tag{3.13}$$

An experiment was conducted to evaluate the regression model R^2 for P_{24} with offsets between 0 and 24 hours.

3.2.6 Rates of Change of Water Quality Parameters

Initial trials found that significant correlations did *not* exist between bacteria and water quality parameters measured with in-stream sensors described in section 2.3. The hypothesis was put forth that useful information resides in the *rate of change* of a variable. For example, following a runoff event, the water depth may quickly increase; hence, the rate of change of the depth, *dH/dt*, may be correlated with bacteria loading in runoff. In other words, a quickly rising water surface may coincide with elevated bacteria counts in the river.

This approach is seen as a way of overcoming non-stationarity in the parameters. Most parameters do not have a constant mean over time; a clear seasonal trend exists. For example, temperature increases as the summer continues. As temperature increases, the water’s ability to hold gases in concentration decreases. This causes a lowering of the saturation dissolved oxygen (DO) concentration. Hence, mean DO levels often decrease over the summer. By disaggregating rapid changes from seasonal trends in water quality, information may be gained which is correlated with bacteria counts. Therefore, rather than looking at the *magnitude* of a parameter, it may be useful to look at how it changes with respect to time.

One can imagine a number of similar scenarios to justify this approach. For instance, rainwater contains very little salt or other dissolved solids. Thus, measurements of conductivity in the stream frequently show a marked decline following heavy rain (at least in the summer, when there is no road salt in runoff). The 'signal' may not reside in the magnitude of the conductivity, but in whether the conductivity has significantly changed in the last few hours, minutes, or days. In Figure 3.5, the specific conductivity in Alewife Brook, a 'flashy' stream dominated by stormwater runoff, during the summer of 2003 is plotted versus time, with dips corresponding to rainstorms:

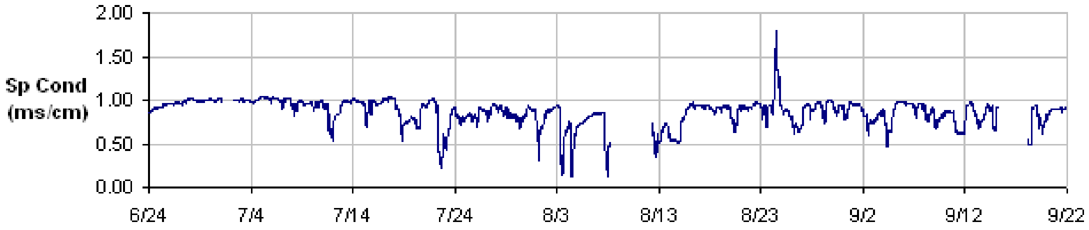


Figure 3.5 Specific conductivity measured in Alewife Brook, 2003

The method of estimating parameter derivatives is the same as that described for estimating the slope of the hydrograph in section 3.2.3 on page 30.

3.2.7 Evaluating Regression Model Fit

It was assumed that the best independent variables derived for single-variable regressions would also be the best choice for a multivariate model. The "best" multivariate regression model is determined by a number of factors, as there is more than one goodness of fit criterion. The model's coefficient of determination is an important determinant of model fit. This variable gives the percentage of the variation in the data that can be explained by the model, where *SSE* is the sum of squared errors, and SS_y is the total sum of squares:

$$R^2 = 1 - \frac{SSE}{SS_y} \tag{3.14}$$

For multivariate models, the *adjusted* R^2 should be reported (Helsel and Hirsch, 1992). The adjusted R^2 takes into consideration the loss of degrees of freedom as additional variables are added to the model. For a model with *p* variables fit to *n* variables:

$$R^2\text{- adjusted} = 1 - \frac{n-1}{n-p}\frac{SSE}{SS_y} \tag{3.15}$$

In addition, the standard error of the model, S_e, is reported. This is equivalent to the standard deviation, or the square root of the variance of the model residuals. The mean square error, or MSE, is given as:

$$s_e{}^2 = \frac{1}{n-(p+1)}\sum_{i=1}^{n} e_i{}^2 \tag{3.16}$$

The standard error of the regression or standard deviation of residuals is then:

$$S_e = \sqrt{S_e{}^2} \tag{3.17}$$

The standard error is an example of a model diagnostic that depends on the units of the observations. While it is useful for quantifying the fit of different models to a data set, it may be difficult to compare standard errors from different studies. Clearly, where different units are involved, the standard errors are not comparable. Even where the same units are employed, the standard errors may be incomparable because of the difference between the mean and variance of two populations.

The PRESS statistic is another important measure, particularly for multivariate models. Helsel and Hirsch (1992, page 247) declare it is "one of the best measures of the quality of a regression equation". PRESS stands for "PRediction Error Sum of Squares". For a dataset with n observations, the algorithm for calculating PRESS creates $n - 1$ regressions, each time omitting one observation, and using the equation to predict the omitted observation. It then repeats this for each observation, squaring and summing the prediction error each time. A regression equation with the lowest PRESS statistic produces the smallest errors when making new predictions.

3.3 Results

A detailed look at the regression modeling results for the Aberjona River is presented below. Following that, concise results are shown for the remainder of the sites.

3.3.1 Aberjona River

The Aberjona River was the furthest upstream of the six bacteria sampling sites in the project. Samples were collected at the site of the USGS gaging station on the Aberjona in Winchester, Massachusetts. The drainage area at this location is approximately 23.5 square miles (USGS, 2003), and the land use in the basin is predominantly residential, commercial, and industrial, with some forested and park land (MassGIS, 2002).

Because the Winchester site is not a place where contact recreation (e.g., swimming or boating) takes place, it was not included in the EMPACT real-time monitoring program, and water quality data is not available. Nevertheless, very good streamflow and precipitation records collected by the USGS are available for this site.

A significant relationship was found between log bacteria concentration and streamflow. The relationship between log(C) and Q was marginally better than for log(Q). In Figure 3.6, the regression equation and the coefficient of determination R^2 are reported on the plot for both cases.

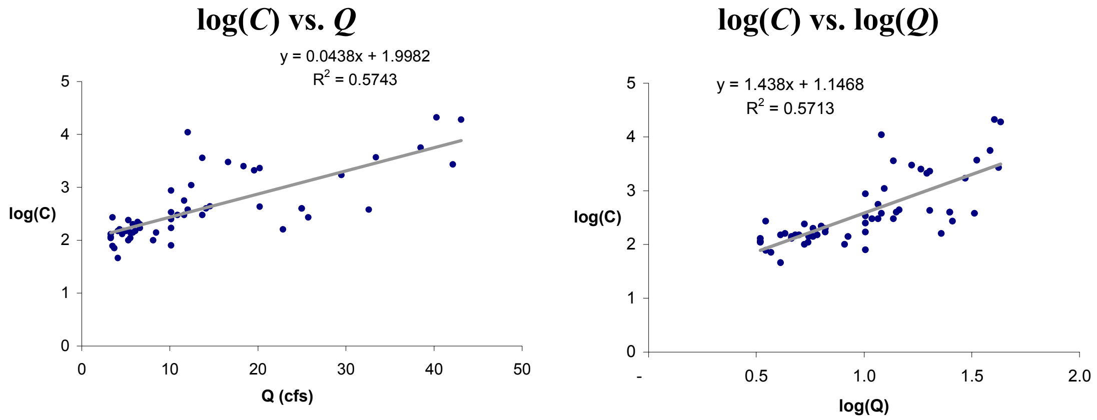


Figure 3.6 Enterococcus concentration regressed against discharge for the Aberjona River, summer 2002

An experiment showed that the correlations are no better for the time-lagged flows. Apparently, the best measurement for predicting the concentration C_t is the contemporaneous streamflow, Q_t, as shown in Figure 3.7.

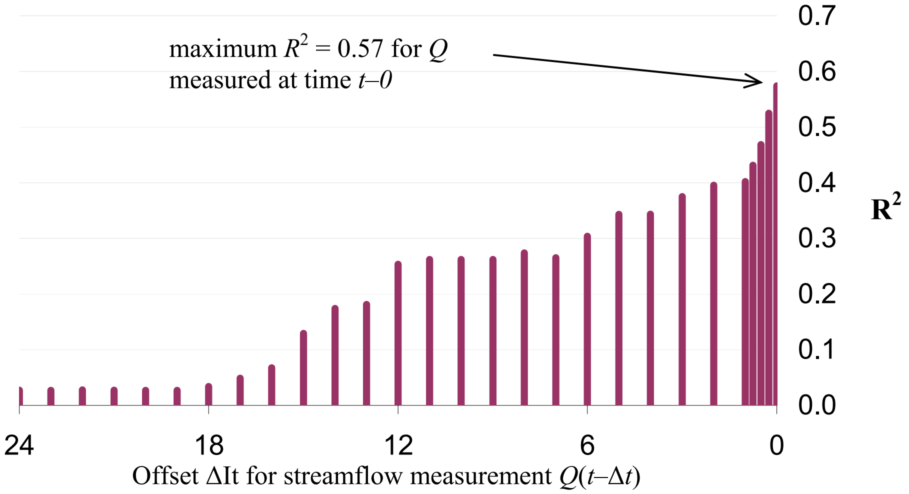


Figure 3.7 Strength of regression between Enterococcus bacteria and time-lagged streamflow, Aberjona River 2002

It was found that averaging the flow does not yield a better correlation, as shown in the Figure 3.8. In other words, it is the instantaneous flow that has the most predictive power.

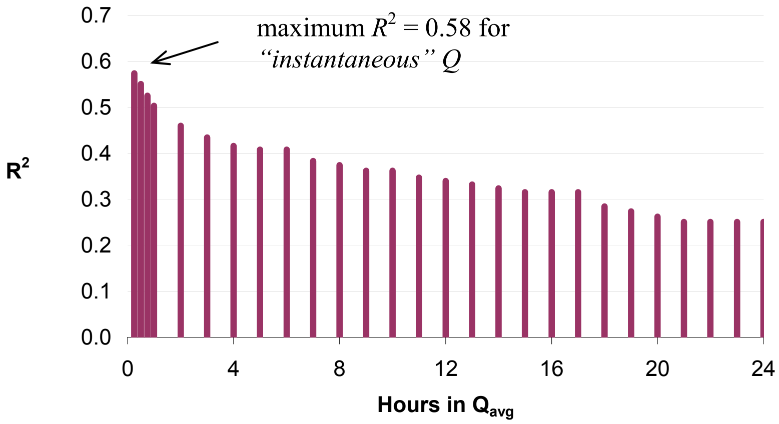


Figure 3.8 Relationship between Enterococcus and averaged flow, Aberjona River 2002

The hydrograph slope, *dQ/dt* proved not to be a useful independent variable, as evidenced by the lack of a discernable relationship in Figure 3.9.

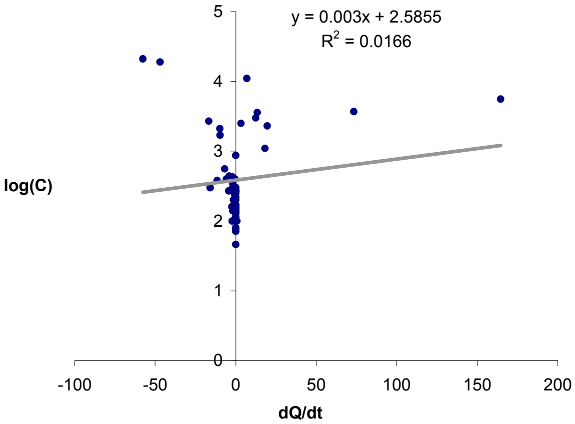


Figure 3.9 Log Enterococcus concentration regressed against the hydrograph slope, Aberjona River 2002

It was posited that the absolute value of the slope would also be a useful variable. This hypothesis grew out of examination of data from two storm events during which the Aberjona River was intensively sampled. It was found that bacteria concentrations were high *throughout* the runoff event, on both the rising and falling limbs of the hydrograph. This suggested that the magnitude of the slope, rather than its sign, is important. Taking the absolute value of the slope did yield an improvement, as can be seen in Figure 3.10.

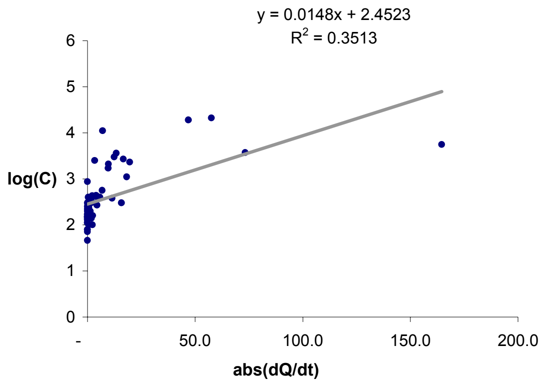


Figure 3.10 Relationship between log bacteria concentration and the absolute value of the hydrograph slope

Still, most of the values are very low, condensed on the left side of the plot. The slope variable was further transformed by taking its logarithm. Because some slopes are zero (and log(0) is undefined), it was necessary to first add a one to each observation. The relationship between the new variable with Enterococcus bacteria is shown Figure 3.11. The variable has now undergone two distinct transformations, and its predictive power has been greatly improved. In fact, the relationship is stronger than that for discharge, Q ($R^2 = 0.70$ vs. 0.57).

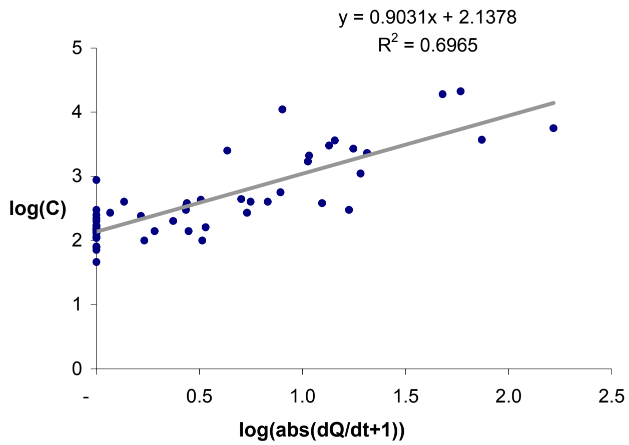


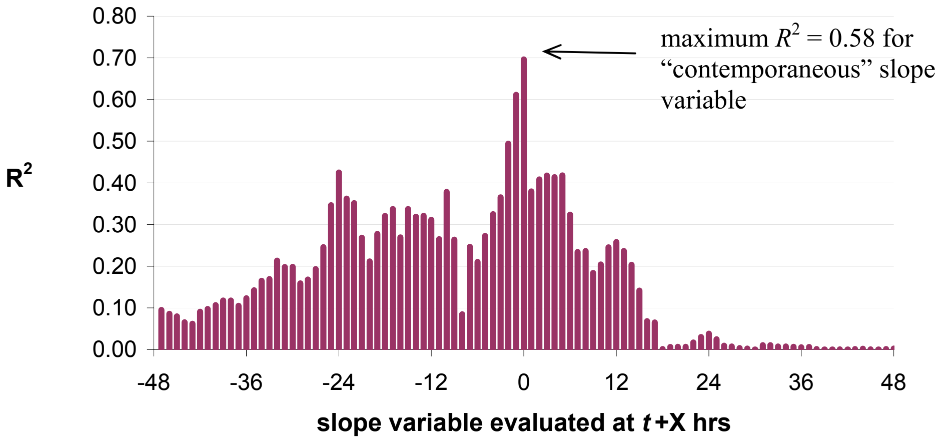
Figure 3.11 Log Enterocococcus concentration versus the log-transformed hydrograph slope variable

To summarize, two separate transformations of hydrograph slope greatly increased its predictive power, as reported in Table 3.2.

Table 3.2 Progressive transformation of hydrograph slope improves its value as a predictor of log bacteria concentration

Independent Variable:	$\frac{dQ}{dt}$	$\left\lvert\frac{dQ}{dt}\right\rvert$	$\log\left(\left\lvert\frac{dQ}{dt}\right\rvert+1\right)$
Regression R^2:	0.02	0.35	0.70

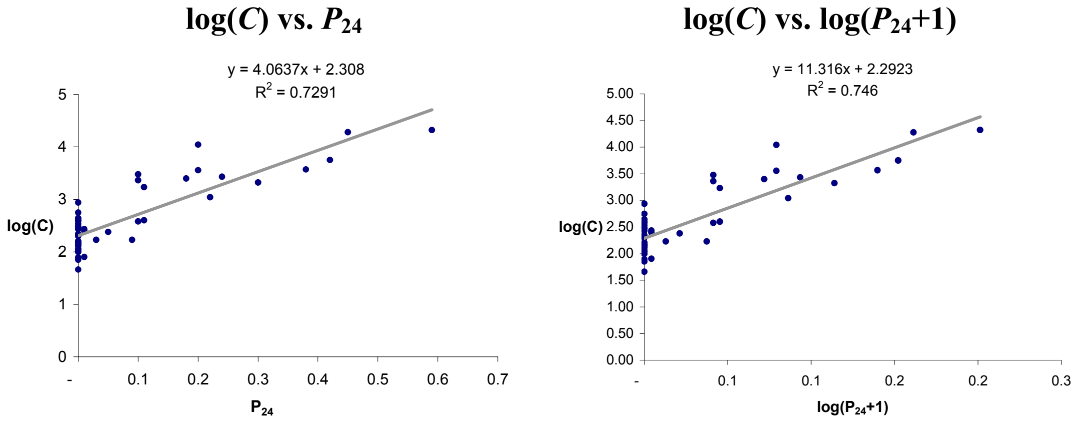
It was hypothesized that the hydrograph slope evaluated at a finite time before the sample was collected may also have predictive capabilities. The results of an experiment to test this hypothesis, summarized in Figure 3.12, showed that this was not the case.



maximum R^2 = 0.58 for “contemporaneous” slope variable

Figure 3.12 Strength of relationship between log bacteria concentration and the transformed slope variable evaluated at $t-\Delta t$

Significant correlations were also found with precipitation. Figure 3.13 shows log(*C*) regressed against the precipitation summed over the 24-hour period prior to sample collection, P_{24}. Note that the log-transformed variable $\log\left(P_{24}+1\right)$ yields a slightly stronger linear fit, as evidenced by a higher R^2:


log(*C*) vs. *P*$_{24}$

log(*C*) vs. log(*P*$_{24}$+1)

Figure 3.13 Log Enterococcus regressed against 24-hour precipitation sum

An experiment was conducted using Excel/VBA to find the best time period over which to sum the precipitation. The regression was repeated with the variables P_1, P_2, …, P_{72}, and the R^2 recorded. The results in Figure 3.14 show that the strongest correlation is with precipitation summed over 20 hours. Note, however, that the results for a range of times from 20-36 hours are nearly as good.

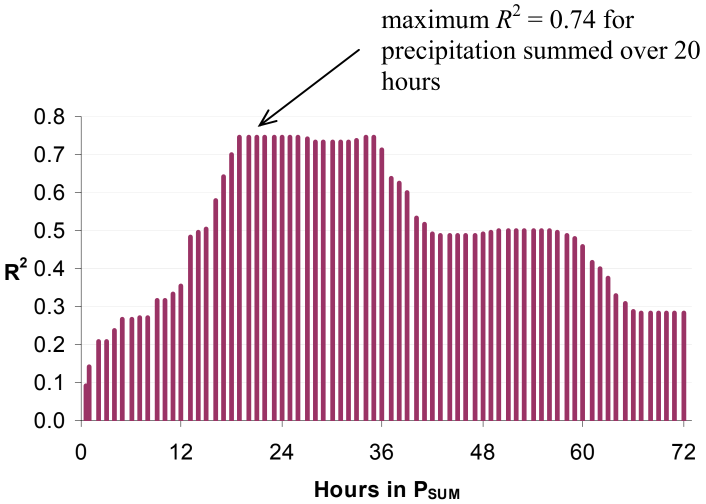


Figure 3.14 Strength of relationship between log bacteria and precipitation summed over 1–72 hours

Surprisingly, it was found that the best independent variable does not include any offset (i.e., $\Delta t = 0$). In other words, the best way to sum the precipitation is to include even the most recent rainfall. This suggests that highly localized runoff is important determinant of the bacteria concentration. The results of this experiment are shown in Figure 3.15.

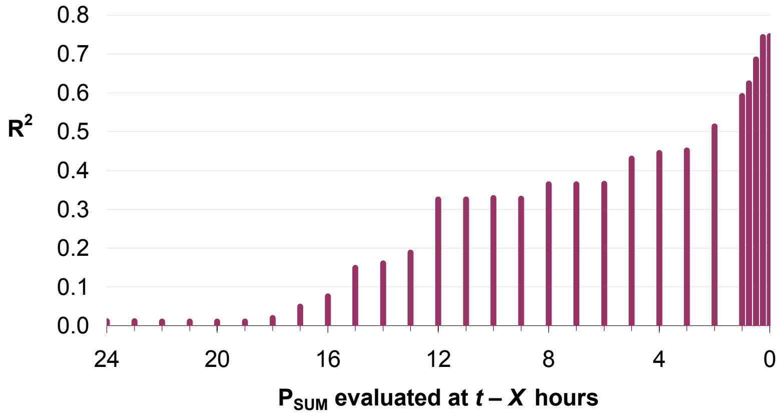


Figure 3.15 Strength of relationship between log bacteria and time-lagged precipitation sums

A significant relationship was also found with the variable T_F, or the time since last rainfall. Further, there is an advantage of using the log-transformed variable, as demonstrated in Figure 3.16.

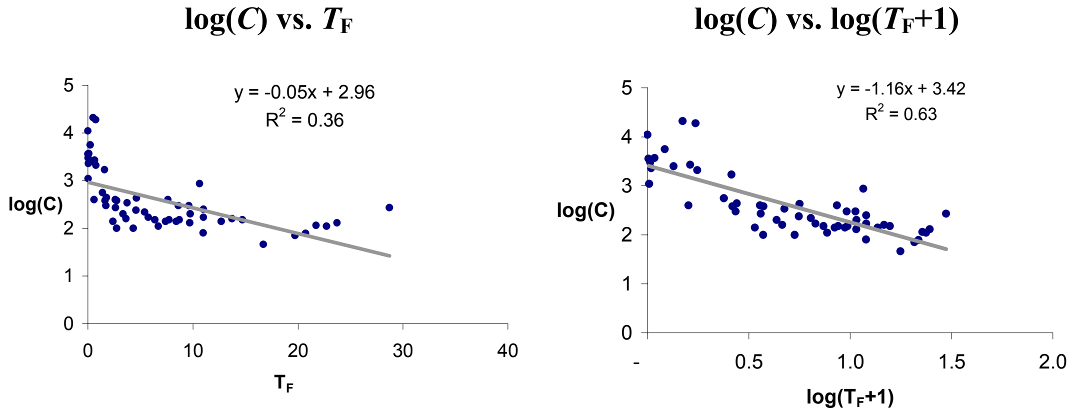


Figure 3.16 Log Enterococcus concentration regressed against T_F, the time since the last rainfall

Which is the “best” threshold for resetting T_F? Figure 3.16 was created for T_F evaluated with a threshold of 0.05 in. An experiment was conducted to determine the optimal threshold for calculating T_F. The threshold for resetting T_F was varied from 0.01–0.25 inches. A regression for log(C) vs. T_F was run for each variable, and the resulting R^2 was recorded. The experiment showed that the strongest linear relationship between log(C) and T_F is obtained using a threshold between 0.04 and 0.07 inches.

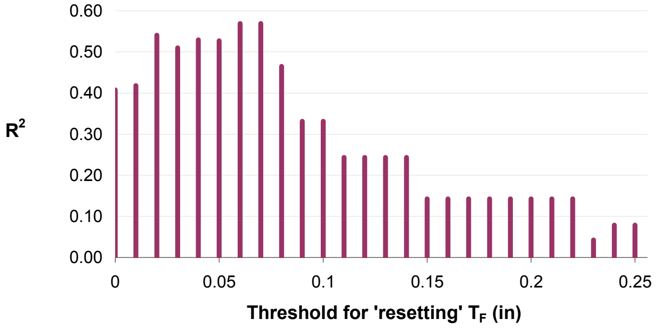


Figure 3.17 Strength of the regression of log bacteria concentration vs. T_F evaluated with varying "thresholds"

As with the other independent variables, it was hypothesized that T_F evaluated at some finite time before the sample was collected may be a better independent variable. Therefore, T_F was evaluated at times t–1, t–2, …, $t-\Delta t$.. The result is that T_F calculated *15 minutes before the sample* is the best variable to use in the regression model:

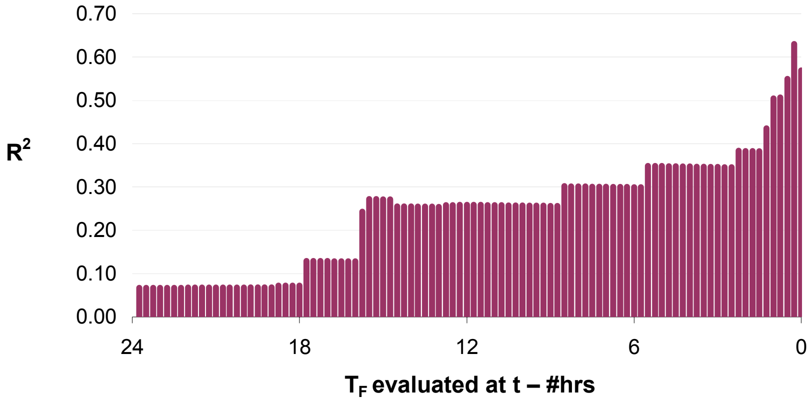


Figure 3.18 Effect of calculating T_F at times before sample collection

The single-variable linear regressions above indicate which variables have the most predictive power, reported in Table 3.3.

Table 3.3 Summary of simple linear regressions for Aberjona River Enterococcus concentration

Variable	Linear Regression Model R^2
$\log\left(P_{20}+1\right)$	0.75
$T_F\Big\vert_{t-15\text{ min}}$	0.63
Q_t	0.57
$\log\left(\left\lvert\frac{dQ}{dt}\right\rvert+1\right)$	0.70

The variables in Table 3.3 were used to develop multivariate regression equations. In a number of cases, when a multivariate equation was developed for a particular subset of variables, the slope calculated for one or more variables was not significantly different from zero. If one assumes that the model residuals are normally distributed, the sample slope will obey a student's *t* distribution. A rigorous hypothesis test on the sample slope can be conducted; alternatively, Helsel and Hirsch (1992, page 231) give a useful rule of thumb: "if $|t| > 2$, one can reject the null hypothesis that the sample slope $b = 0$ at α =0.05 for reasonably large sample sizes and therefore assert there is a statistically significant linear relationship." In general, slopes were accepted as significant for a probability $P > 0.05$ or a t-score, $|t| > 2$.

The results for a variety of 'good' multivariate models are reported in the Table 3.4. Model #6 is considered the best model by a variety of measures. It has the highest adjusted R^2, and the lowest standard error and PRESS statistic. The *t*-scores of all model coefficients are large enough that the slopes are considered statistically significant.

Model #7 is included for illustration; it contains a slope that is not statistically different from zero. In this case, the insertion of discharge (*Q*) as an additional explanatory variable did nothing to improve the model, and by some measures (e.g., the PRESS statistic), the model is *worse* than the two-variable model #6.

Table 3.4 Regression models for Enterococcus bacteria concentration at Aberjona River gaging station, *n* = 55 samples

$$\log\left(C_{Entero}\right) = a + b_1 \log\left(P_{20}+1\right) + b_2 T_{\mathrm{F}} + b_3 Q_{\mathrm{t}} + b_4 \log\left(\left|\frac{dQ}{dt}\right| + 1\right) + \varepsilon \tag{3.18}$$

Model No.	No. of variables	Independent Variables Included	a*	b_1	b_2	b_3	b_4	S_e†	R^2 adj.	PRESS
#1	1	$\log\left(P_{20}+1\right)$	2.29 (47)	11.3 (13)				0.32	0.74	6.0
#2	1	T_{F}	3.42 (34)		– 1.16 (– 9.6)			0.39	0.62	8.8
#3	1	Q_t	2.00 (22)			0.440 (8.5)		0.42	0.57	10.4
#4	1	$\log\left(\left\|\frac{dQ}{dt}\right\|+1\right)$	2.13 (34)				0.90 (11)	0.35	0.69	7.3
#5	2	$\log\left(P_{20}+1\right)$, Q_t	2.19 (42)	7.16 (5.1)		0.420 (3.6)		0.29	0.79	5.0
#6	2	$\log\left(P_{20}+1\right)$, T_{F}	2.79 (26)	7.89 (7.7)	– 0.563 (– 4.9)			0.27	0.82	4.3
#7	3	$\log\left(P_{20}+1\right)$, T_{F}, Q_t	2.70 (19)	7.29 (6.0)	– 0.521 (– 4.2)	0.005 (0.37)		0.27	0.82	4.5

*t-ratios reported in parentheses

†Standard Error of model residuals.

A plot of the observed and predicted values is an important visual indicator of the model fit. On the scatterplot, the theoretical 1:1 line represents *model = observations*. The more closely data are clustered about this line, the better the fit. Note that the scatterplot is constructed with log-log axes. Such a plot for the regression model #6 is reported in Figure 3.19. The figure shows that the model may be in error by nearly an order of magnitude at times. Nevertheless, these results are among the better regressions for bacteria reported in the literature.

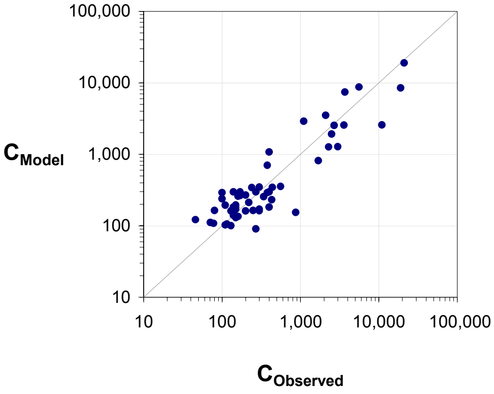


Figure 3.19 Modeled versus observed Enterococcus for the Aberjona River, summer 2002

When model #6 is used to calculate a continuous simulation of bacteria concentrations, it yields the result shown in Figure 3.20. In the plot, bacteria data have been plotted with error bars of ±30% to represent the variability reported in Oriel (2003).

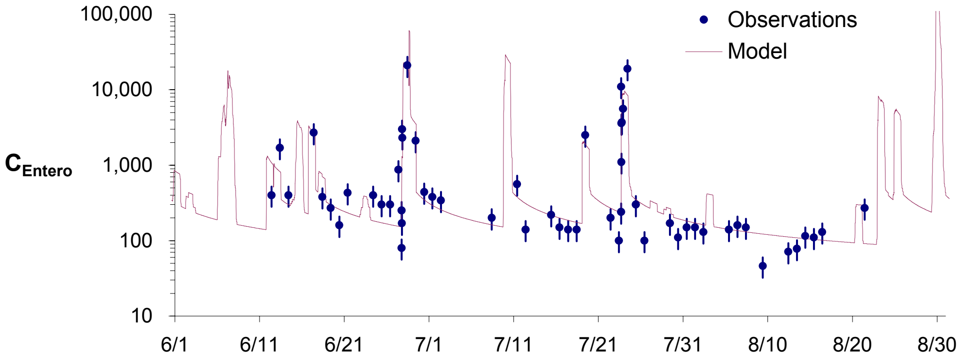


Figure 3.20 Regression model results for the Aberjona River for summer 2002.

3.3.2 Sandy Beach

Development of a predictive model for this site was a high priority, as it is the most widely used recreational area in the study. It was hypothesized that the analysis techniques developed for the Aberjona would yield useful regression models at Sandy Beach, as well as the other downstream sampling locations. Multivariate regression models developed for Sandy Beach are reported in Table 3.5.

The regression model reported in the last row in Table 3.5 has a negative slope with respect to the independent variable for streamflow, *Q*. This would imply that higher flows lead to lower bacteria concentrations. As this does not make sense, based on our understanding of watershed processes, this model was rejected. In general, regression models do not explain as much of the variability in the bacteria concentration data as those developed for the Aberjona River. It is most important to look at how well the model predicts violations of the bathing water quality standards; a framework for doing so is developed in section 3.3.6 below.

Table 3.5 Regression models for Enterococcus bacteria concentration at Sandy Beach, n = 61

$$\log\left(C_{Entero}\right) = a + b_1 \log\left(P_{24} + 1\right) + b_2 T_{\mathrm{F}} + b_3 Q_t + b_4 \log\left(\left|\frac{dQ}{dt}\right| + 1\right) + \varepsilon \qquad (3.19)$$

No.	No. of variables	Independent Variables Included	a*	b_1	b_2	b_3	b_4	$\mathrm{S_e}$†	R^2 $adj.$	PRESS
#1	1	$\log\left(P_{24}+1\right)$	0.85 (11)	8.30 (5.5)				0.54	0.33	18.2
#2	1	T_{F}	1.63 (12)		–0.74 (–4.8)			0.56	0.27	20.0
#3	1	Q_t	0.41 (1.8)			0.69 (3.1)		0.62	0.12	24.3
#4	1	$\log\left(\left\|\frac{dQ}{dt}\right\|+1\right)$	0.79 (8.3)				0.59 (4.5)	0.57	0.25	20.8
#5	2	$\log\left(P_{24}+1\right)$, T_{F}	1.24 (7.3)	5.9 (3.5)	–0.42 (–2.5)			0.52	0.39	16.9
#6	3	$\log\left(P_{24}+1\right)$, T_{F}, Q_{t}	2.2 (4.7)	7.1 (4.0)	–0.67 (–2.1)	–0.76 (–3.3)		0.50	0.42	16.2

*t-ratios reported in parentheses
†Standard Error of model residuals.

The observed and modeled bacteria concentrations are shown in Figure 3.21 and Figure 3.22:

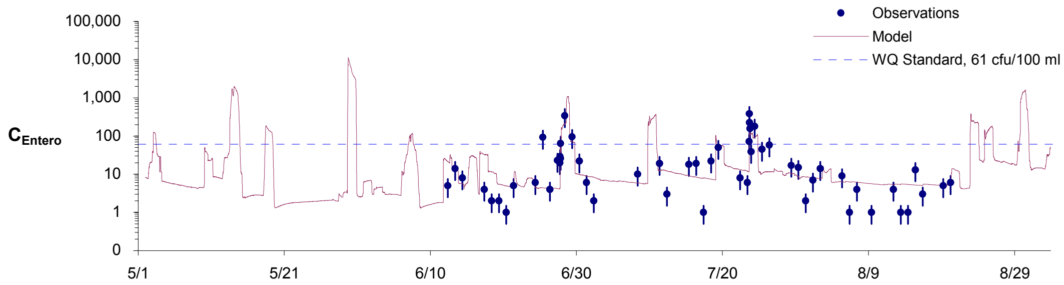


Figure 3.21 Time series of observed and modeled Enterococcus concentration (cfu/100 mL) at Sandy Beach in summer 2002

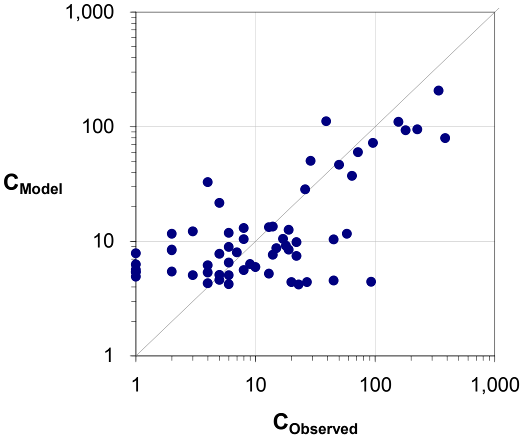


Figure 3.22 Scatterplot of observed and modeled Enterococcus concentration (cfu/100 mL) at Sandy Beach in summer 2002

3.3.3 Blessing of the Bay

The Boys & Girls Club at the Blessing of the Bay Boathouse is one of the more heavily-used recreational areas in the watershed. Therefore, it was a priority to develop a predictive bacteria model for this site. The site differs from Sandy Beach, however, in that the primary recreational use is boating rather than swimming. Streamflow was not included as an independent variable in the regression model for this site; flow measurements were only available for the Aberjona River, which is approximately 9 km (5.6 miles) upstream, and therefore has a smaller drainage area. The regression models developed for this site are summarized in Table 3.6. The best multivariate regression model (#3) explains about 61% of the variability in the Enterococcus bacteria data, as indicated by and adjusted $R^2 = 0.61$ for the two-variable model.

Table 3.6 Regression equations for Enterococcus at the Boys & Girls Club, sample size *n* = 57

$$\log\left(C_{Entero}\right) = a + b_1 \log\left(P_{24}+1\right) + b_2 \log\left(T_F+1\right) + \varepsilon \qquad (3.20)$$

No.	Variables included		a*	b_1	b_2	S$_e$†	R^2 *adj.*	PRESS
#1	1	P_{24}	0.42 (5.2)	9.4 (7.1)		0.52	0.49	14.7
#2	1	T_F	1.54 (13)		–1.5 (–8.1)	0.48	0.55	12.7
#3	2	P_{24}, T_F	1.12 (6.1)	4.7 (2.9)	–1.0 (–4.2)	0.45	0.61	11.3

*t-ratios reported in parentheses

†Standard error of model residuals

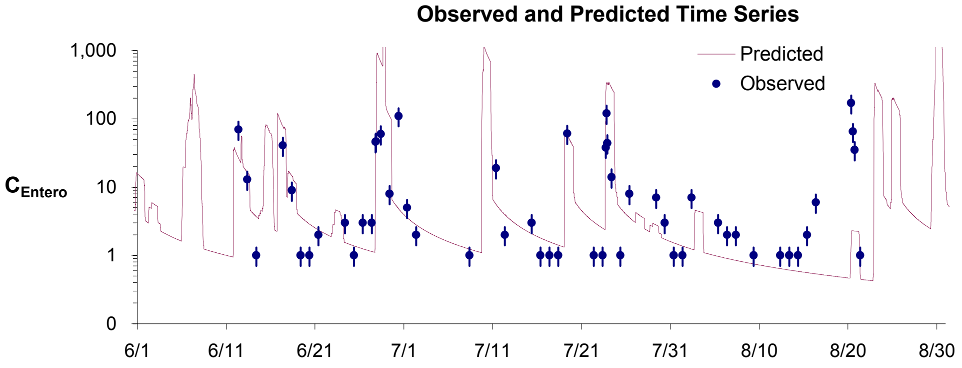


Figure 3.23 Time series of observed and modeled bacteria concentrations in the Mystic River at the Boys & Girls Club

3.3.4 Alewife Brook

Bacteria counts in this polluted tributary were the highest of those measured in the watershed. The geometric mean of measured Enterocci concentration is 740 cfu/100 mL, compared to 370 cfu/100 mL in the Aberjona River. Correlations were found between bacteria levels and precipitation, and the time since the last rainfall. An experiment to determine the best period over which to sum the precipitation led to the conclusion that 12 hours is the best; log Enteroccus concentration regressed against P_{12} had the highest R^2. Among the regression models reported in Table 3.7, the two-variable model (#3) is considered best; it has the lowest PRESS statistic and standard error. However, model #1, a single-variable regression with P_{12} is nearly as good.

Table 3.7 Multivariate regression equations for Enterococcus in Alewife Brook, with *n* = 70

$$\log\left(C_{Entero}\right) = a + b_1 \log\left(P_{12} + 1\right) + b_2 \log\left(T_F + 1\right) + \varepsilon \tag{3.21}$$

No.	Variables included		a*	b_1	b_2	S_e†	R^2 *adj.*	PRESS
#1	1	P_{12}	2.62 (42)	14.0 (8.6)		0.46	0.52	15.7
#2	1	T_F	3.40 (29)		–0.71 (–5.5)	0.56	0.30	22.5
#3	2	P_{12}, T_F	2.88 (23)	11.6 (6.2)	–0.30 (–2.4)	0.45	0.55	14.9

[*] t-ratios reported in parentheses

[†] Standard Error of model residuals

The model's fits versus observations are shown in Figure 3.24.

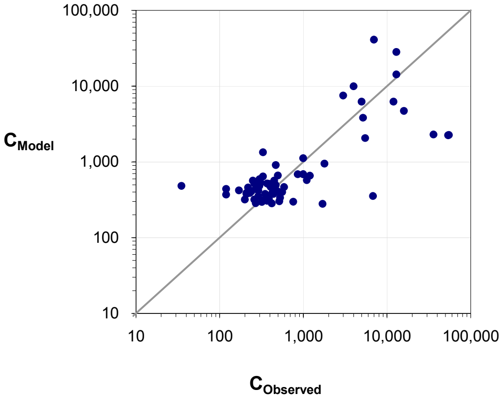


Figure 3.24 Scatterplot of observed versus predicted Enterococcus concentration in Alewife Brook, 2002

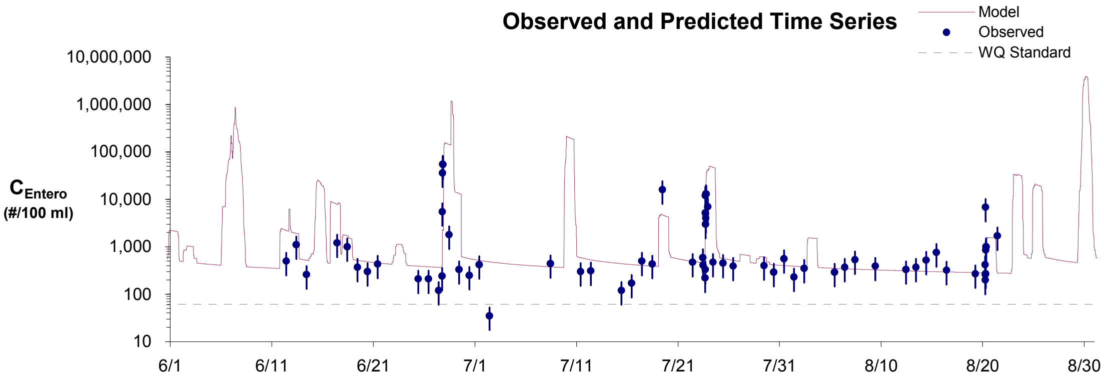


Figure 3.25 Time series plot of observed Enterococcus concentrations and predictions for Alewife Brook, 2002

The problem of extrapolation is foreseen in the applying the ‘black box’ regression model for this site. This can be seen clearly in Figure 3.26 where the regression model is applied with climate data from May 1, 2002 – September 30, 2002. Note the peak in May (prior to data collection), corresponding to a heavy rainfall. The regression model predicts a maximum concentration of 10^9 cfu/100 mL (1 billion organisms) far outside the range of the observed data. This arises because the model was developed with a finite dataset, and very large rainstorms (greater than 0.4 inches) were not encountered in the summer of 2002). This is clearly a case of garbage-in, garbage-out, and reinforces the point that regression models are only as good as the data used in their development.

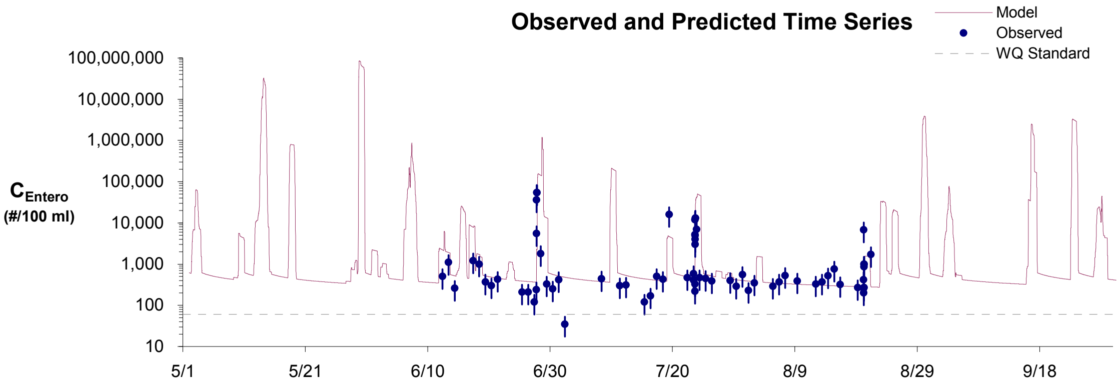


Figure 3.26 Time series demonstrating the extrapolation problem with the regression equation for Alewife Brook, May – September 2002

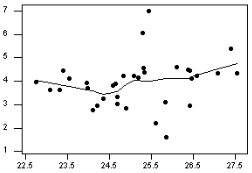
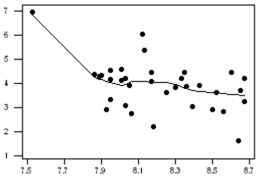
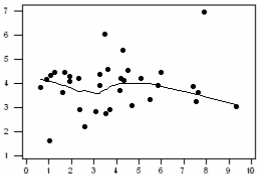
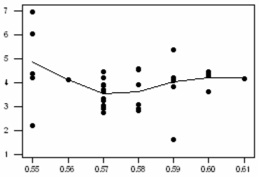
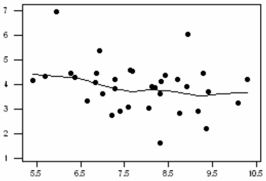
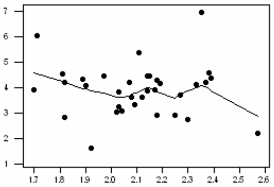
3.3.5 High Street Bridge

In general, bacteria concentrations at the High Street Bridge site were much lower than at the other sampling locations on the Mystic, and well below the state standards for swimming (Enterococcus < 61 cfu/100 mL) over 90% of the time. Pollutants enter the lake via runoff that directly enters the lake as well as streamflow in the Aberjona River and other small tributaries such as Mill Brook. The water here is relatively clean because it lies just downstream from the outlet of the Mystic Lakes. The residence time of the lakes is usually much greater than one week (S. Chapra, 2003 personal communication), during which time die-off occurs in bacteria populations. The lakes act as a large stilling basin where pollutants can settle out. Further, solar radiation may also contribute to reducing bacteria counts due to the large open area of the lake surfaces.

A significant relationship was not found between the streamflow or precipitation measured at the USGS gage and the bacteria measured at the High Street Bridge. Due to the physical distance between the two sites, it was hypothesized that the lagged flows (i.e., Q at $t-\Delta t$) might have some predictive capability. However, no matter how the variables were manipulated, no significant relationships were found.

Nevertheless, the most complete record of real-time water quality data existed for the High Street Bridge site. Hence, relationships were sought between bacteria and water quality parameters (depth, temperature, pH, dissolved oxygen, conductivity, turbidity.) The results for this investigation, with a sample size $n = 32$, are summarized in Table 3.8. A scatterplot for each parameter is shown versus the log *Enterococcus* concentration, with a Lowess trendline to give a visual sign of trend. Lowess stands for LOcally WEighted Scatterplot Smoothing, and is described in Helsel and Hirsch (1992, page 287). The fact that the Lowess trendlines are nearly horizontal indicates the absence of correlation. The coefficient of determination (R^2) is reported, along with the probability, which indicates the likelihood that the correlation would be observed by chance. In general, probability values greater than 0.05 were considered grounds to conclude the absence of significant correlation.

Table 3.8 Relationship between Enterococcus bacteria and water quality parameters measured in the Mystic River at the High Street Bridge

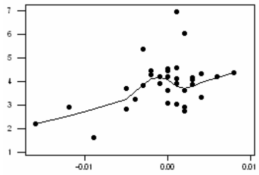
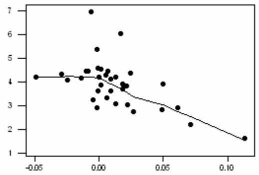
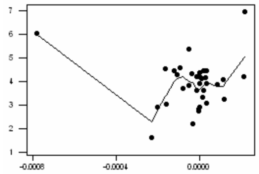
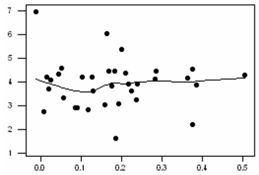
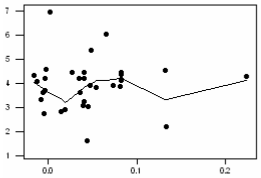
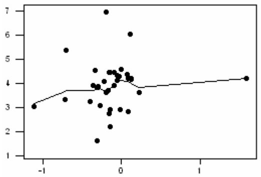
Parameter	Scatterplot with Lowess trendline	R^2	Probability
Temperature (ºC)		0.04	0.24
pH		0.21 (R^2-predicted =0.07)	0.007
Turbidity (NTU)		0.02	0.45
Specific Conductivity (ms/cm)		0.01	0.573
Dissolved Oxygen (mg/L)		0.09	0.08
Depth (ft)		0.003	0.75

The correlations reported above are not statistically significant. A slight correlation appears to exist at first sight with pH. However, the presence of a single outlier dominates the calculation. The *predicted* R^2 (reported in parentheses in Table 3.8) is an indication of the strength of the relationship when successive observations are removed. The fact that the predicted R^2 is very low (0.07) suggests that when one or more observations are discarded, a significant correlation no longer exists.

Next, the relationship was explored between bacteria and the parameter derivatives. Results of this analysis are shown in Table 3.9. Correlations with the time derivatives of depth and temperature are statistically significant at the 95% confidence level. Some influence is exerted by extreme observations, however, indicated by values of R^2-predicted which are a good deal lower than R^2-adjusted.

The parameter-derivative approach, which is believed to be novel, holds potential for incorporating continuous monitoring data into concentration-discharge models. Further work should be carried out (with a greater sampling frequency for bacteria and potentially other pollutants) to confirm this approach.

Table 3.9 Relationships between the water quality parameter derivatives and log Enterococcus bacteria concentration in the Mystic River at the High Street Bridge

Parameter Time Rate of Change		Scatterplot with Loess smooth	R^2 (R^2-predicted)	Prob-ability
Depth	$\frac{dH}{dt}\left(\frac{\text{ft}}{\text{hr}}\right)$		0.20 (0.12)	0.008
Temperature	$\frac{dT}{dt}\left(\frac{{}^{\circ}\text{C}}{\text{hr}}\right)$		0.33 (0.25)	<0.001
pH	$\frac{dSC}{dt}\left(\frac{\text{mS/cm}}{\text{hr}}\right)$		0.006	0.67
Dissolved Oxygen	$\frac{dO}{dt}\left(\frac{\text{mg/L}}{\text{hr}}\right)$		0.005	0.70
Specific Conductivity	$\frac{dpH}{dt}\left(\frac{\text{pH units}}{\text{hr}}\right)$		0.001	0.86
Turbidity	$\frac{dTurb}{dt}\left(\frac{\text{NTU}}{\text{hr}}\right)$		0.014	0.51

3.3.6 Evaluating Models with Respect to the Swimming Standard

Because the focus of this research is on recreational waters, the models should be evaluated in terms of their ability to predict exceedances of the swimming standard. Scatterplots of observed versus predicted bacteria concentrations can be modified to visualize how well the model predicts conditions causing beach closures. Figure 3.27 is similar to plots presented by Francy et al. (2003) in a USGS study of *E. coli* at Ohio beaches. Table 3.10 contains brief descriptions of the four quadrants on the plot.

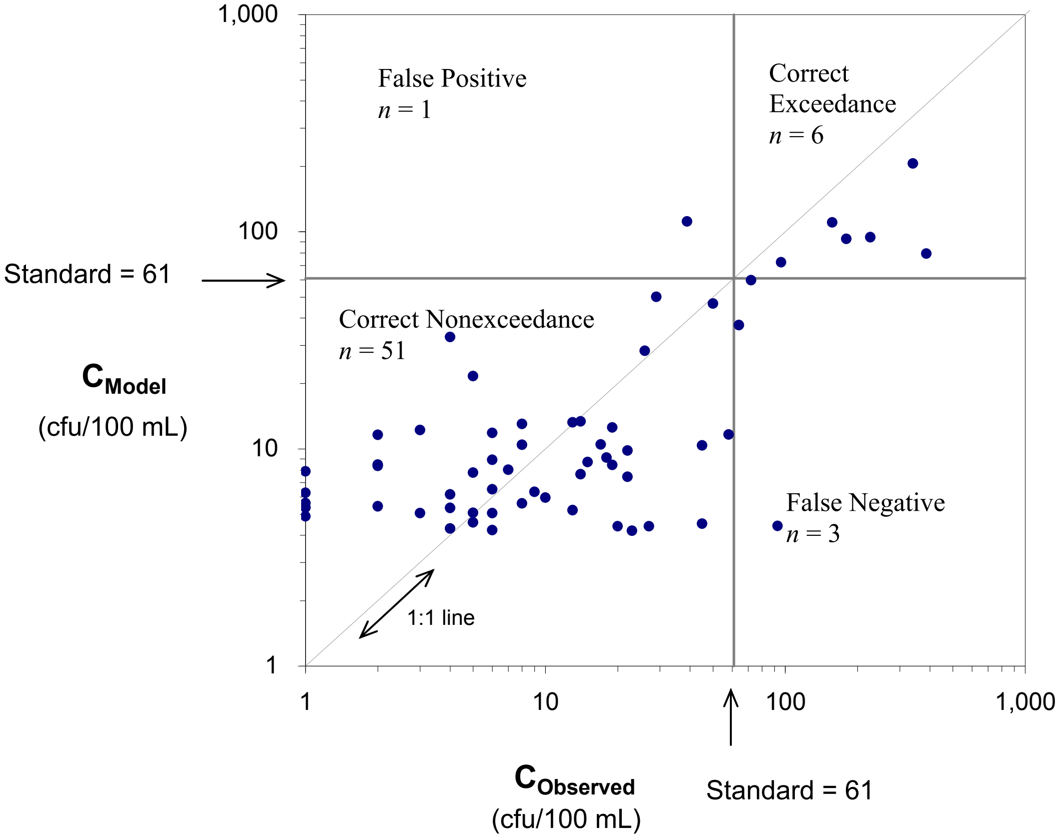


Figure 3.27 Performance of the regression model in predicting exceedances of the swimming standard at Sandy Beach, 2002

Table 3.10 Model evaluation framework

False Positive Points in the upper left quadrant falsely predict exceedances of the water quality standard, while observations are actually below the standard.	**Correct Exceedance** Points in the upper right correctly predict exceedances of the standard.
Correct Nonexccedance Points in the lower left correctly predict nonexceedances of the standard. In other words, the model correctly predicts safe water.	**False Negative** Points in the lower right falsely predict nonexceedance, while observations exceed the standard. In other words, the model falsely predicts safe water.

It should be evident that a false negative is more serious than a false positive. In deciding beach closures, a false positive would cause an unnecessary closure; an inconvenience, but a prediction that errs on the safe side. A false negative error, by contrast, would mean allowing the beach to stay open in unsafe conditions. The chance that the model accurately predicts an exceedance of the water quality standard given an observed exceedance can be expressed as a conditional probability. During the 2002 sampling season, there were 9 exceedances of the swimming standard. The regression model correctly predicts 6 out of 9, or 67% of the exceedances.

A similar plot for the Blessing of the Bay sampling location is shown in Figure 3.28.

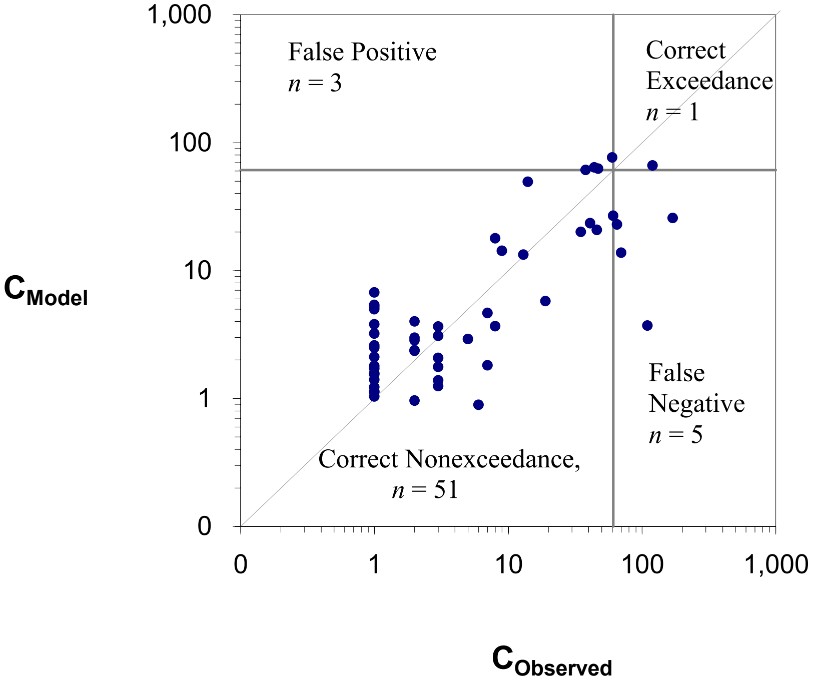


Figure 3.28 Efficiency of model at predicting exceedances of the water quality standard in the Mystic River at the Blessing of the Bay Boathouse, 2002

At the Blessing of the Bay, the water quality standard of 61 cfu/100 mL was exceeded 6 times during the summer of 2002. The model correctly predicted an exceedance only 1 out of 6 times, or a dismal 17%. However, it is worth noting that the exceedances may represent unusual observations. A number of the exceedances occurred during dry weather, and may represent bacteria loads from sources other than polluted runoff, as discussed in section 3.4 below.

Table 3.11 Final regression equation for Enterococcus bacteria concentration at each sampling location

Location	Equation	Standard Error	R^2-adjusted
Aberjona River	$\log\left(C\right)=2.79+7.89\log\left(P_{20}+1\right)-0.563T_F+\varepsilon$	0.27	0.82
Sandy Beach	$\log\left(C\right)=1.24+5.9\log\left(P_{24}+1\right)-0.42T_F+\varepsilon$	0.52	0.39
Alewife	$\log\left(C\right)=2.88+11.6\log\left(P_{12}+1\right)-0.30T_F+\varepsilon$	0.45	0.55
Boys & Girls Club	$\log\left(C\right)=1.12+4.7\log\left(P_{24}+1\right)-1.0T_F+\varepsilon$	0.45	0.61

3.4 Discussion

Regression modeling has shown that precipitation is the environmental variable which is most strongly correlated with bacteria concentrations in the Mystic River. Each sampling site has its own response behavior, requiring unique regression models, with different explanatory variables, and unique intercept and slopes. The final regression equations are reported in Table 3.11.

Significant correlations were also found with streamflow and the associated hydrograph slope. However, when both precipitation and streamflow are included, the streamflow variables tend not to have significant slopes and to ‘drop out’ of the equation. In general, regression models for most of the sampling locations in the Mystic River basin explain between 39% and 82% of the variability in the Enterococcus bacteria concentration.

It was found that regression models yield the best fit to the observed bacteria data at the most riverine sites, the Aberjona River and Alewife Brook. At sampling locations with more complex hydrodynamics, the regression models were not able to explain as much of the variability in the bacteria concentrations. Sites with more complex hydrodynamics include one site on a lake (Sandy Beach on Upper Mystic Lake), and one in the slack water of the lower river basin (the Blessing of the Bay Boathouse), where water is impounded behind the Amelia Earhart Dam.

A plausible hypothesis to explain the relatively poor results at these locations is that additional explanatory variables are missing from the regression. Perhaps additional mechanisms drive bacteria loading, for which data were not collected. For instance, at Sandy Beach, swimmers themselves may be a significant source of bacteria. “Bather load" is well documented, with individuals capable of shedding 10^6 or more fecal coliform bacteria in an hour of active bathing (Drew, 1971). Further, increased bacteria concentrations at Blessing of the Bay location are believed to be caused by the many birds that congregate around the dock (Oriel, 2003, personal communication). Fogarty et al. (2003) showed that gull feces contain Enterococcus concentrations of 10^4–10^8 per gram, and are a major contributor to bacterial contamination of surface waters in the Great Lakes area.

The fate and transport of bacteria in lakes are influenced by a number of complex mechanisms. Bacteria mortality rates are affected by solar radiation, salinity, and a number of other factors (Chapra, 1998, lecture 27). One may conclude that either: (a) regression equations are not the correct type of model to describe the behavior of bacteria in lakes; (b) more data and/or more explanatory variables are needed; or (c) the stochastic portion, or random error, of bacteria data is inherently too great to build satisfactory predictive models.

Climate during the first sampling season proved a further complication. The total precipitation during July and August 2002 was 3.55 inches, or 42% lower than the average 6.1 inches expected during those months (NCDC, 2003). Because the summer of 2002 was basically a drought, the data collected during that summer do not describe the full range of hydrologic conditions in the basin. The hydrograph for water year 2002 (November 1, 2001 – October 31, 2002) is shown in Figure 3.29 with circles representing sample times. Enterococcus bacteria data was collected from June 27-August 21, 2002. The long-term daily average flow of 31 cfs (USGS, 2003) is plotted on the figure for reference.

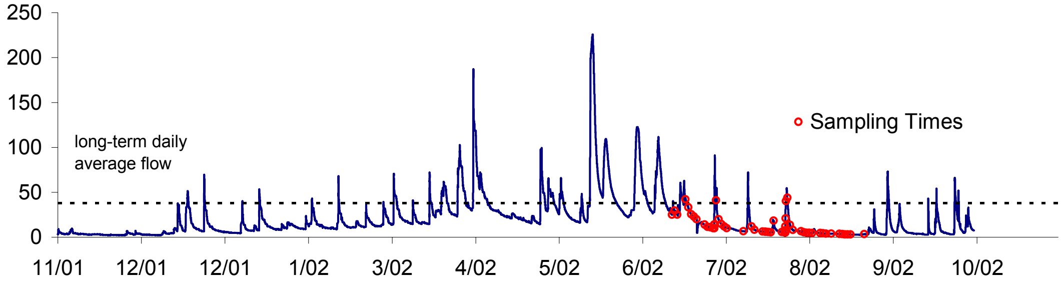


Figure 3.29 Aberjona River hydrograph with Enterococcus bacteria sampling times, summer 2002

A flow-duration curve for the Aberjona River is shown in Figure 3.30, created from daily flows for the water years 1991–2001. The minimum, mean, and maximum shown on the figure are for the set of flows observed at sampling times, and illustrate that the sampling program did not capture the full range of possible flow conditions. In other words, the available data was taken mostly at low flows; hence, sample times are not representative of the entire flow regime

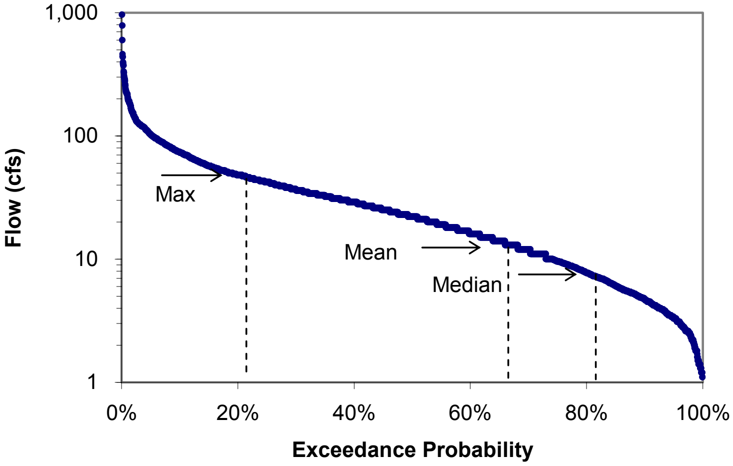


Figure 3.30 Flow duration curve for the Aberjona River, 1991-2001.

Water quality parameters from the real-time monitoring program tended to have limited predictive capability for bacteria. However, statistically significant relationships were found between log Enteroccocus concentration and the time rate of change of water depth and temperature. The parameter-derivative approach holds promise for future modeling efforts and should be explored further.

Finally, a framework for evaluating the regression model results with respect to the water quality standard was presented. As more data are collected, the model should be confirmed not only according to traditional goodness-of-fit statistics like the sum of squared errors, but also in terms of whether the model is able to correctly predict exceedances of the swimming or boating standards. Ultimately, this will decide the efficacy of the model as a decision-making tool.

4. Development of a Watershed Simulation Model for Bacteria Loading

4.1 Introduction

A computer model has been developed to predict daily average loads and concentrations of pathogen-indicator bacteria in surface water bodies. The program couples a rainfall-runoff model to predict streamflow with a pollutant buildup-washoff model to simulate bacteria loading. The inspiration for the simulation model is the relatively simple but useful Generalized Watershed Loading Functions model (GWLF) originally described by Haith and Tubbs (1981), and further developed by Haith and Shoemaker (1971) and Haith et al. (1996b). GWLF estimates monthly streamflow, sediment yield, and nutrient loadings given land use and a handful of parameters to describe the average hydrological characteristics of the basin. The overall goal was to create a planning-level model that will generate useful estimates of bacteria concentration while limiting the overall complexity of the model. The first step was to build a rainfall-runoff model to simulate streamflow in the basin.

First, a lumped-parameter, continuous soil-moisture accounting hydrologic model was developed. The hydrologic component in the original GWLF is very simplified; it overcomes the imprecision of its predictions by summing the output on a monthly basis. Figure 4.1 is an example of the inaccuracies in daily streamflow predicted by GWLF. For this example, the model's code was modified to output daily streamflow for the Aberjona River basin for the summer of 2002. Little effort was spent adjusting model parameters; the goal here was to show that the original, unmodified model does not accurately describe the behavior of daily streamflow data.

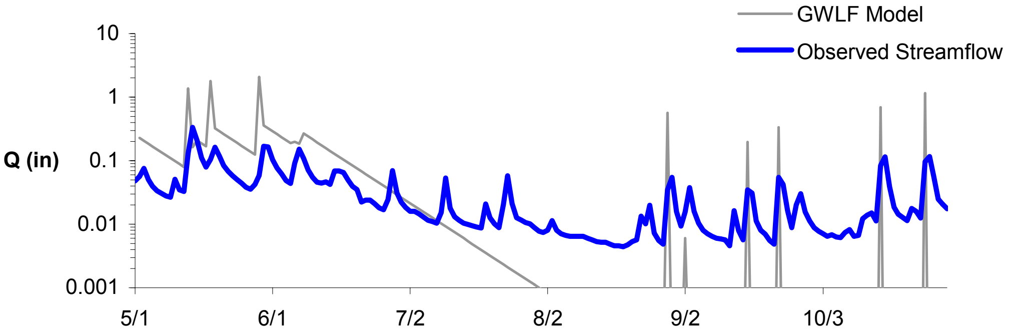


Figure 4.1 Daily output from the original, unmodified GWLF model

Inaccuracies in daily predictions tend to be unimportant when one looks at streamflow summed a monthly time scale. Monthly output is not appropriate for modeling bacteria, however, as instream concentrations respond quickly to bacteria loading, as illustrated by Table 1.1. Also, bacteria die off quickly, with typical survival times of less than 7 days. Hence, monthly output would be meaningless. The present study attempts to develop a reasonably accurate daily model without significantly increasing the number of model parameters.

The hydrology of the current model is more sophisticated than GWLF in two ways:

- Finer resolution of the output (daily rather than monthly)
- More sophisticated solution technique; the model is built as a system of differential equations, which are solved by Euler's method

When daily discharge measurements are available from a stream gage, the model user can take advantage of a number of statistical and graphical diagnostics to aid in calibration of the rainfall-runoff portion of the model.

Elsewhere in the literature, investigators have used empirical buildup and washoff functions to simulate pollutant loadings in stormwater runoff (e.g. Sartor and Boyd, 1972, Huber and Dickinson, 1992, Barbé et al., 1996). The work described here extends this method to modeling bacteria. The model was calibrated with field data collected during the summers of 2002 and 2003. *Enterococcus* and fecal coliform bacteria samples were collected on the Aberjona River in Winchester, MA, and Alewife Brook, in Somerville, MA, as described in section 2.1 on page 6.

The philosophy behind the model development was neatly summarized by Nash and Sutcliffe in 1970: it is "desirable that the model should reflect physical reality as closely as possible." Further, "there should be no unnecessary proliferation of parameters to be optimized… each additional part of a model must substantially extend the range of application of the whole model. In other words, we are prepared to accept additional parts and hence greater difficulty in determining parametric values only if the increased versatility of the model makes it much more likely to obtain a good fit between observed and computed output."

4.2 Methods

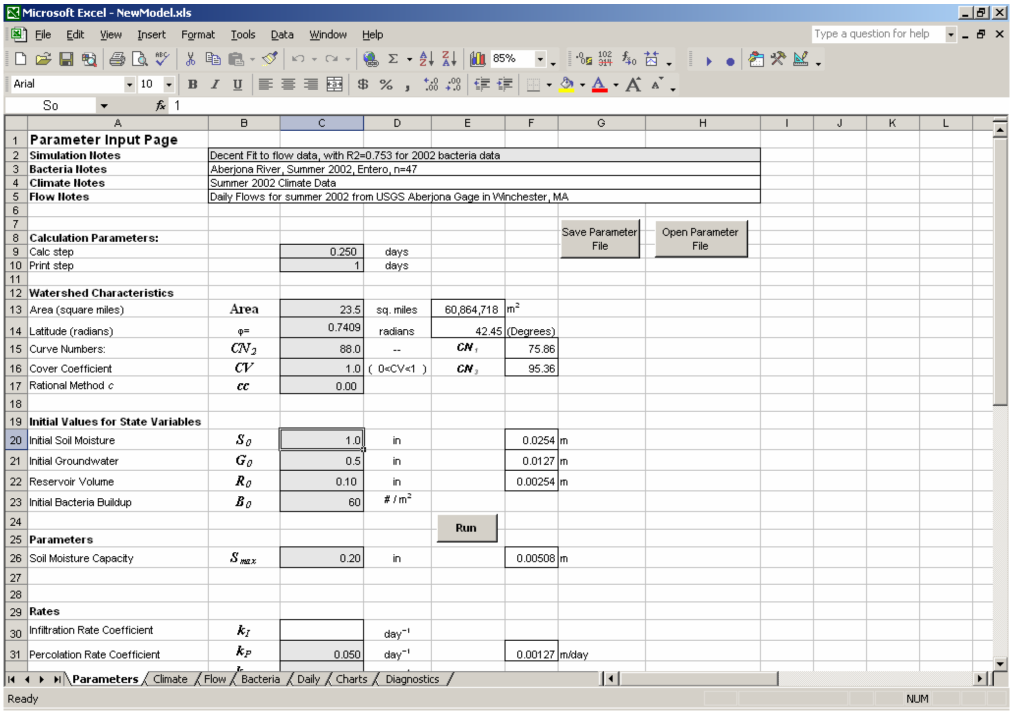
The model's computer code was written in Visual Basic for Applications (VBA) for Microsoft Excel. VBA's linkages with the spreadsheet application are exploited to facilitate data input and visualization of model output. The entire model is stored in an Excel workbook and can be distributed by copying a single file, without the need for a compiler or other special software. Because VBA is a standard part of Excel, anyone with the program can run the model.

Model input is entered directly on worksheets, in the familiar Excel computing environment, allowing the model user to easily import or cut and paste data from other sources. For instance, climate data from the web can be saved as a text file, imported into Excel, and then easily pasted onto the worksheet. When the program is run, VBA performs the model calculations in the background. The resulting output is written to another set of worksheets. When new data are written to a range of cells on the worksheet, plots are updated automatically. Therein lies the great advantage of model building in Excel/VBA; the developer can focus his or her efforts on developing the mathematical model, without expending effort to write utilities for charting, etc.

The simulation model workbook contains the input worksheets shown in Table 4.1. An example of the "Parameters" input sheet is shown in Figure 4.2.

Table 4.1 Simulation model input worksheets

Worksheet	Description
Parameters	Watershed characteristics and model coefficients
Climate	Meteorological Data, input in format similar to that which is downloaded from the NCDC website
Flow	Daily streamflow, if gage data is available
Bacteria	Observed daily average bacteria concentrations (calculate composite average if multiple samples are collected on a single day)

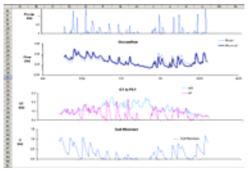
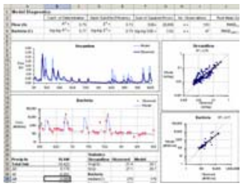


	A	B	C	D	E	F
1	**Parameter Input Page**					
2	**Simulation Notes**	Decent Fit to flow data, with R2=0.753 for 2002 bacteria data				
3	**Bacteria Notes**	Aberjona River, Summer 2002, Entero, n=47				
4	**Climate Notes**	Summer 2002 Climate Data				
5	**Flow Notes**	Daily Flows for summer 2002 from USGS Aberjona Gage in Winchester, MA				
6						
7						
8	**Calculation Parameters:**					
9	Calc step		0.250	days		
10	Print step		1	days		
11						
12	**Watershed Characteristics**					
13	Area (square miles)	**Area**	23.5	sq. miles	60,864,718	m^2
14	Latitude (radians)	φ=	0.7409	radians	42.45	(Degrees)
15	Curve Numbers:	CN_2	88.0	--	CN_1	75.86
16	Cover Coefficient	CV	1.0	(0<CV<1)	CN_3	95.36
17	Rational Method *c*	cc	0.00			
18						
19	**Initial Values for State Variables**					
20	Initial Soil Moisture	S_0	1.0	in		0.0254 m
21	Initial Groundwater	G_0	0.5	in		0.0127 m
22	Reservoir Volume	R_0	0.10	in		0.00254 m
23	Initial Bacteria Buildup	B_0	60	# / m^2		
24						
25	**Parameters**					
26	Soil Moisture Capacity	S_{max}	0.20	in		0.00508 m
27						
28						
29	**Rates**					
30	Infiltration Rate Coefficient	k_I		day^{-1}		
31	Percolation Rate Coefficient	k_P	0.050	day^{-1}		0.00127 m/day

Figure 4.2 Simulation model input worksheet 'Parameters'

The model output is written to the sheet "Daily", while two other sheets, shown in Table 4.2 display plots and statistics.

Table 4.2 Simulation model output worksheets

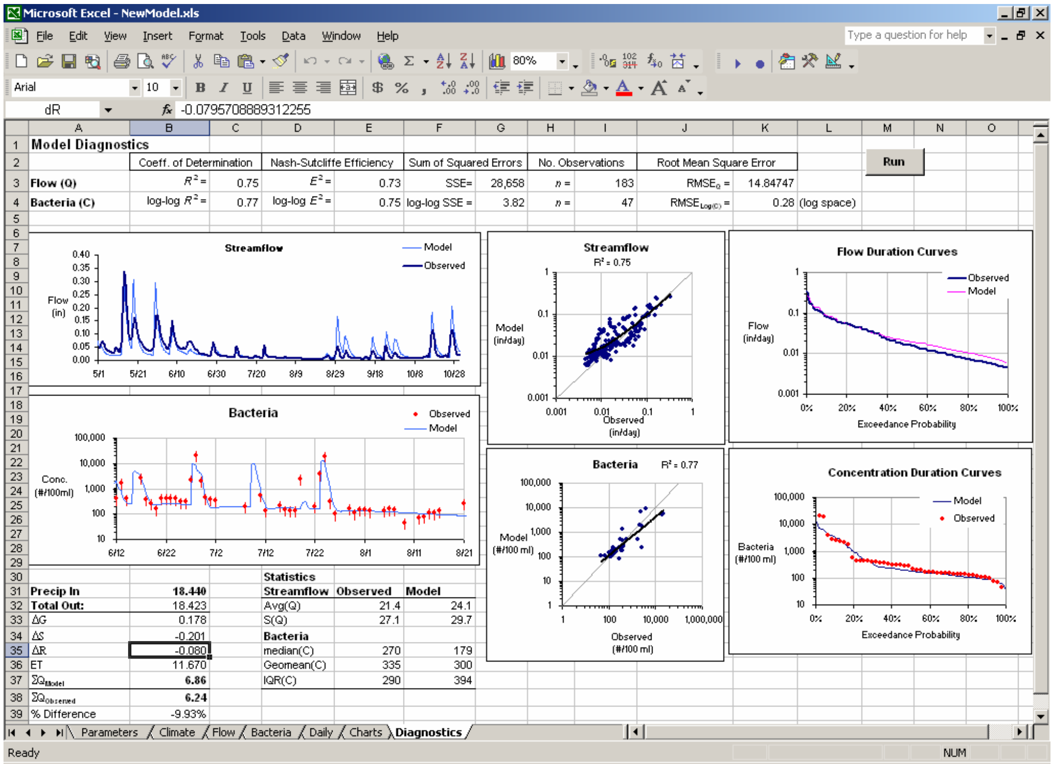
Worksheet		Description
	Daily	Calculated daily values of streamflow and bacteria
	*Charts*	Time series plots of model internal state variables and output
	*Diagnostics*	Information useful in evaluating and calibrating the model

	A	B	C	D	E
1	Daily Output				
2					
3	Day	P	ET	Q	C_{model}
28	5/25/2002	0.000	0.055	0.026	227
29	5/26/2002	0.000	0.033	0.023	244
30	5/27/2002	0.010	0.021	0.022	251
31	5/28/2002	0.070	0.027	0.021	253
32	5/29/2002	1.740	0.092	0.137	2116
33	5/30/2002	0.000	0.153	0.239	747
34	5/31/2002	0.130	0.163	0.139	484
35	6/1/2002	0.000	0.160	0.088	470
36	6/2/2002	0.070	0.135	0.057	384
37	6/3/2002	0.000	0.118	0.042	401
38	6/4/2002	0.000	0.107	0.032	210
39	6/5/2002	0.000	0.119	0.027	223
40	6/6/2002	0.730	0.106	0.044	4528
41	6/7/2002	0.520	0.095	0.094	3115

Graphical and statistical information useful in calibrating the model is summarized on the sheet "Diagnostics," shown in Figure 4.3. It allows the user to examine the model qualitatively, through a variety of plots, and quantitatively, through summary statistics and goodness-of-fit measures, reported in Table 4.3.

Table 4.3 Qualitative and quantitative indicators of model fit on the worksheet 'Diagnostics'

Charts
Time series plots
Observations vs. model scatterplots
Exceedance probability plots
Boxplots of observed and modeled bacteria
Residuals plots
Statistics (for observations and model)
Average
Standard deviation
25% quartile
Median
Geometric mean
75% quartile
Interquartile range
Indicators of model fit
Coefficient of determination
Nash-Sutcliffe model efficiency
Sum of squared errors
Root mean square error



Model Diagnostics

	Coeff. of Determination		Nash-Sutcliffe Efficiency		Sum of Squared Errors		No. Observations		Root Mean Square Error		
Flow (Q)	$R^2 =$	0.75	$E^2 =$	0.73	SSE=	28,658	$n =$	183	$\text{RMSE}_Q =$	14.84747	
Bacteria (C)	log-log $R^2 =$	0.77	log-log $E^2 =$	0.75	log-log SSE =	3.82	$n =$	47	$\text{RMSE}_{\text{Log(C)}} =$	0.28	(log space)

Precip In	**18.440**
Total Out:	18.423
ΔG	0.178
ΔS	-0.201
ΔR	-0.080
ET	11.670
ΣQ_{Model}	**6.86**
$\Sigma Q_{\text{Observed}}$	**6.24**
% Difference	-9.93%

Statistics		
Streamflow	**Observed**	**Model**
Avg(Q)	21.4	24.1
S(Q)	27.1	29.7
Bacteria		
median(C)	270	179
Geomean(C)	335	300
IQR(C)	290	394

Figure 4.3 Model diagnostic worksheet

4.2.1 Mathematical Model

The mathematical model was formulated as a set of state variables, which may be thought of as storage compartments. Bras (1990) and Eagleson (1970) give background on the 'engineering representation' of the hydrologic cycle. State variable units are entered in units of inches by the user, although the model performs all calculations in meters. State variables represent simplified parts of an actual watershed, namely soil moisture (*S*), groundwater (*G*), and a reservoir (*R*) which represents storage in the channels and is used to route the runoff flows.

Each of the state variables contains a measure of water that can be thought of either as depth or volume. For instance, a soil moisture $S = 0.01$ m (0.4 in) can be visualized as a thin layer of water spread evenly over the entire watershed area. For the

Aberjona River watershed, with an area of approximately 23.6 square miles, or 61×10^6 m^2:

$$\begin{aligned} S(\text{volume}) &= S(\text{depth})\times\text{Area} \\ &= (0.01\text{ m})(61{,}000{,}000\text{ m}^2) \\ &= 610{,}000\text{ m}^3 \end{aligned}$$

A mass balance approach is used; water can be neither created nor destroyed. Precipitation (*P*) is the only way that water can enter the model, and as such is the 'forcing function.' At the end of the simulation period, all of the water that has entered via precipitation ends up as either a change in storage (e.g., Δ*S*), streamflow (*Q*), or evapotranspiration (*ET*). This has been a useful check in verifying the model computations.

$$\sum inputs = \sum outputs \tag{4.1}$$

$$\sum P = \sum Q + \sum ET + \Delta S + \Delta G + \Delta R \tag{4.2}$$

Water moves between the various compartments via mechanisms such as infiltration, percolation, meant to mimic the way water moves in the environment. These rates of water movement are expressed in units of meters per day. All of the rates in the model change over time, usually as a function of the value of one or more state variables. In most cases, rates are written as linear functions of a state variable. The advantage of using linear functions is their simplicity, ease of calculation, and minimum number of parameters.

The overall model structure is shown in Figure 4.4. In general, state variables are represented as boxes, and rates have been drawn as arrows. Rates of change of the state variables are expressed as differential equations as follows:

Soil Moisture: $$\frac{dS}{dt} = Infiltration - ET - Percolation \tag{4.3}$$

Groundwater: $$\frac{dG}{dt} = Percolation - Recession \tag{4.4}$$

Routing Reservoir: $$\frac{dR}{dt} = Runoff - RoutedFlow \tag{4.5}$$

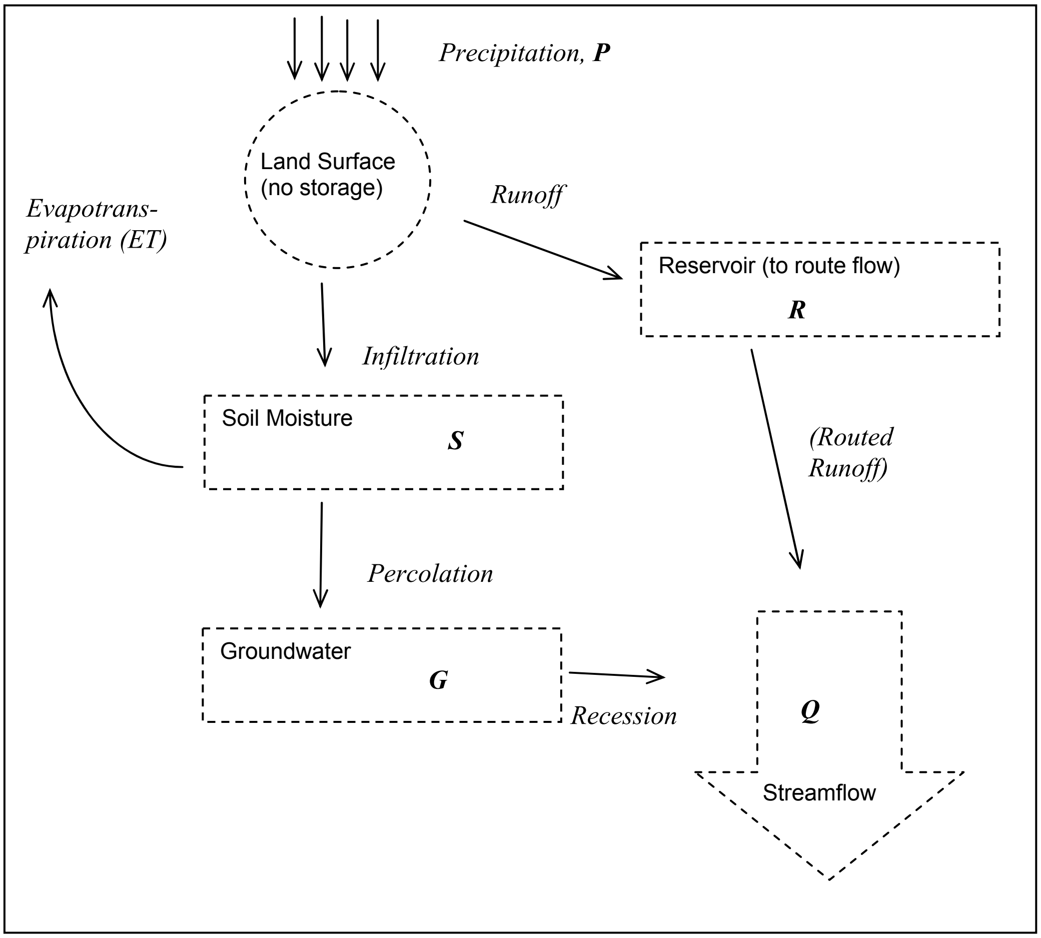
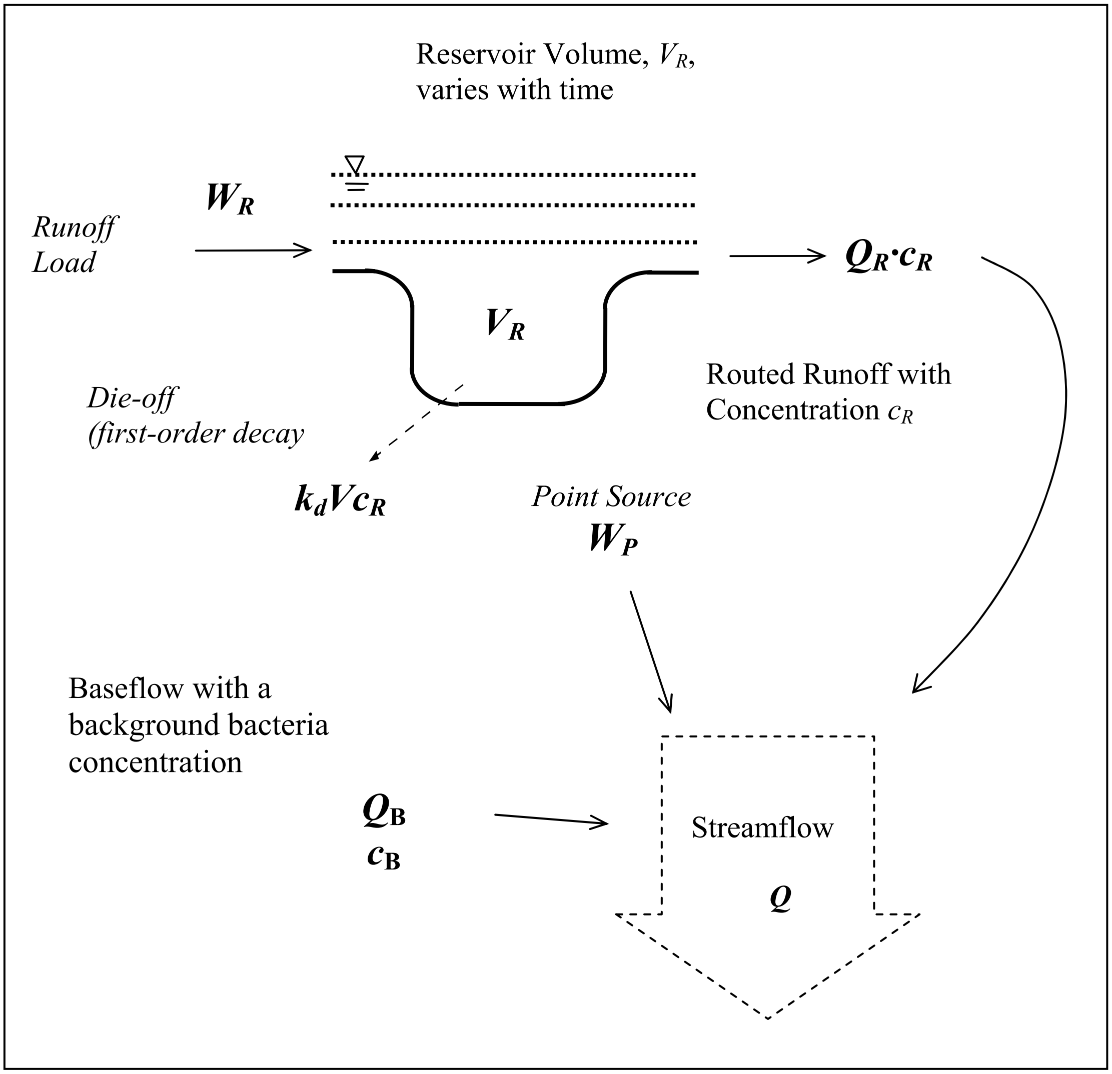


Figure 4.4 Simulation model structure

4.2.2 Bacteria Loading Model

Figure 4.5 shows the structure of the bacteria loading model. The loading model is similar to RUNQUAL (Whipple et al., 1983, page 112), but its structure and solution technique are significantly different. RUNQUAL, an event model developed by the Southeast Michigan Council of Governments in Detroit, is a combination of the SWMM runoff block, and the QUAL-II receiving water model.



Reservoir Volume, V_R, varies with time
W_R
Runoff Load
$Q_R \cdot c_R$
V_R
Routed Runoff with Concentration c_R
Die-off (first-order decay
$k_d V c_R$
Point Source
W_P
Baseflow with a background bacteria concentration
Q_B
c_B
Streamflow
Q

Figure 4.5 Polluted runoff model

4.2.3 Model Variables

A summary of all of the model's important variables, rates, and constants is given in Table 4.4. Note that not all parameters listed are required for every simulation, depending on which simulation options are chosen. For instance, several options are available for modeling buildup or washoff of bacteria from land surfaces, which are described in detail in section 4.2.13. Variables with an asterisk can be adjusted by the modeler during calibration to obtain the best model fit.

Table 4.4 Simulation model rates, variables, and constants

	Symbol	Units
Inputs		
Precipitation	$\boldsymbol{P}$	in/day
Temperature	$\boldsymbol{T}$	ºF
State Variables		
Soil Moisture	$\boldsymbol{S}$	in
Groundwater	$\boldsymbol{G}$	in
Routing Reservoir	$\boldsymbol{R}$	in
State Variable Initial Values		
Initial Soil Moisture*	$\boldsymbol{S_0}$	in
Initial Groundwater*	$\boldsymbol{G_0}$	in
Initial Reservoir Volume*	$\boldsymbol{R_0}$	in
Rates of Water Movement		
Runoff	***Runoff***	in/day
Routed Runoff	$\boldsymbol{Q_R}$	in/day
Infiltration	***Infiltration***	in/day
Percolation	***Percolation***	in/day
Recession	***Recession***	in/day
Parameters		
Curve Number*	$\boldsymbol{CN}$	–
Maximum Soil Moisture Capacity*	$\boldsymbol{S_{max}}$	in
Evapotranspiration Cover Coefficient*	$\boldsymbol{CV}$	–
Rate Constants		
Percolation Rate Coefficient*	$\boldsymbol{k_P}$	day^{-1}
Groundwater Recession Coefficient*	$\boldsymbol{k_G}$	day^{-1}
Reservoir Coefficient*	$\boldsymbol{k_R}$	day^{-1}
Bacteria Loading Model Parameters		
Initial Bacteria Buildup*	$\boldsymbol{B_0}$	# organisms / m^2
Upstream 'Background' Concentration*	$\boldsymbol{c_B}$	# organisms / 100 mL
Buildup Rate*	$\boldsymbol{k_B}$	(#/m^2·day)
Buildup Inhibition Factor*	$\boldsymbol{a}$	–
Washoff Rate*	$\boldsymbol{k_W}$	day^{-1}
Fixed Concentration in Runoff*	$\boldsymbol{C_R}$	# organisms / 100 mL
Washoff Parameter*	$\boldsymbol{a_W}$	–
Washoff Parameter*	$\boldsymbol{b_W}$	–
Bacteria Point Source*	$\boldsymbol{W_P}$	# organisms / day
Bacteria Decay Rate*	$\boldsymbol{k_d}$	day^{-1}
Upstream ‘Settling Reservoir’ Parameters		
Initial Bacteria Concentration*	$\boldsymbol{C_0}$	#/100 mL
Volume*	$\boldsymbol{V_S}$	m^3

*Parameters with an asterisk may be adjusted to fit model to observations

4.2.4 Units

Although some units are input by the user and output to the worksheet in customary US units, all calculations within the model use metric units:

Table 4.5 Units for variables inside the model

Variable	Units
Storage Compartments	m
Watershed Area:	m^2
Rates of water movement:	$\mathrm{m \cdot day^{-1}}$
Bacteria Loading	$\mathrm{organisms/day}$
Bacteria Concentrations	$\mathrm{organisms \cdot m^3}$

A number of conversions are carried out by the model so that quantities may be entered, and displayed on the sheet, in conventional units. For instance, bacteria concentrations are commonly measured in units of colony-forming units per 100 mL (cfu/100 mL). Until the 1990s, 'most probable number' was frequently used, with concentrations expressed as MPN/100 mL. In order to avoid ambiguity, this thesis refers to number of bacteria, or simply number. The model calculates bacteria concentration as number per cubic meter (#/m^3). This is converted to conventional units by multiplying by 10^{-4}.

$$\frac{\#\ \text{Bacteria}}{\text{m}^3}\times 10^{-4} = \frac{\#\ \text{Bacteria}}{100\ \text{ml}} \tag{4.6}$$

4.2.5 Model Solution Technique

The system of differential equations is solved by Euler's method, as described in Chapra and Canale (2002, chapter 25). The model user enters a time step for model calculation, t_c. Although it is not explicitly stated, the original GWLF model also uses Euler's method, stepping through the model solution with a time step of $t_c = 1$ day (Haith et al., 1996a). In order to reduce round-off errors, the time step is adjusted by the program so that:

$$t_c = \frac{1}{2^n} \qquad \text{in days, where } n \geq 0 \text{ is an integer} \tag{4.7}$$

Euler's method gives a numerical approximation to the true solution to the differential equations. With Euler's method, errors in the model solution decrease for smaller calculation time steps. This is not to say that the model fit is improved, but rather, the calculated solution to the model agrees more closely with the true mathematical solution to the equations. An experiment was performed to show that this is indeed the case, thereby verifying the model calculations. This was tested by verifying that both sides of equation (4.2) balance:

$$\sum P = \sum Q + \sum ET + \Delta S + \Delta G + \Delta R \tag{4.2}$$

For the calibration time period, summer 2002, the total input precipitation was 18.44 in. Using the right side of this equation, the "total water out" was evaluated for a number of time steps, t_c, as reported in Table 4.6 and plotted in Figure 4.6.

Table 4.6 Decreasing calculation time step lowers model error

# steps per day	t_c	Total water out (inches)	% Error
1	1	18.507	0.364%
2	0.5	18.475	0.192%
4	0.25	18.458	0.100%
8	0.125	18.449	0.051%
16	0.0625	18.445	0.027%
32	0.03125	18.443	0.014%
64	0.015625	18.441	0.006%
128	0.007813	18.441	0.003%

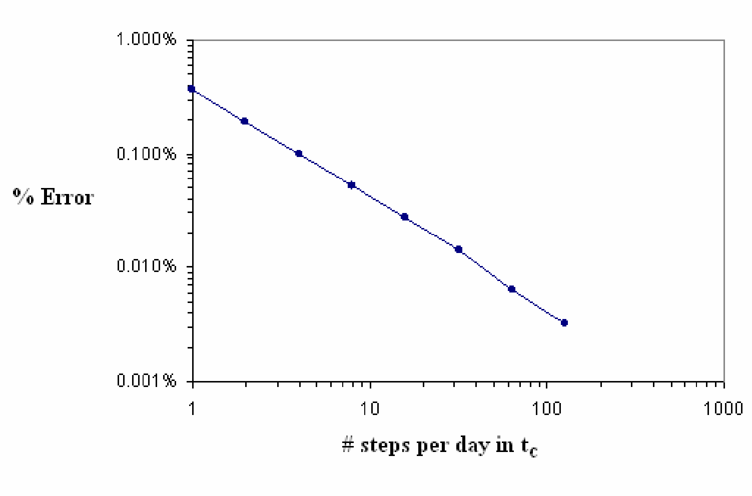


Figure 4.6 Example of model error decreasing with the calculation time step

Precipitation data may be entered in daily, hourly, or 15-minute intervals. The only requirement is that the calculation step, t_c, should be smaller than the interval of the input data. However, it was found that model fit is not necessarily improved by finer resolution input, and may not justify the additional computational time. For single computations, the added computational burden would be negligible. However, if the model were used for uncertainty analysis (such as Monte Carlo), it could prove important.

The model requires initial, or starting, values for the state variables, such as the initial soil moisture depth, S_0, and initial groundwater depth, G_0. Again, these are input by the user.

4.2.6 Runoff

Unlike more detailed simulation models such as HSPF, the model does not include a land-phase storage compartment. Typically, such a compartment is included to account for ponding, or depression storage. When rain falls, it goes in one of two directions, either downward into the soil as infiltration, or over land toward the stream as runoff. Thus the input precipitation, P, is split in two, with a portion running off, with the remainder entering soil moisture, S.

Runoff is calculated by the Soil Conservation Service (SCS) Curve Number method. This method was first described in the SCS Technical Release 55[2] (Cronshey et al., 1986). This empirical method was originally developed as an event model for estimating runoff volume from small urban watersheds. Runoff is determined by characteristics of the land surface (the curve number), precipitation, and antecedent moisture conditions. The SCS method calculates runoff as follows:

[2] The SCS is now the Natural Resources Conservation Service (NRCS). Some practitioners still refer to the model as TR55.

$$Q = \frac{\left(P - I_a\right)^2}{P - I_a + S_\mathrm{i}} \qquad \text{for} \qquad P \geq I_a \tag{4.8}$$

$$Q = 0 \qquad \text{for} \qquad P \leq I_a$$

where

$Q =$ Runoff depth (in)

$P =$ Event rainfall depth (in)

$I_a =$ Initial abstraction (in), or the amount of rainfall that is lost before runoff begins, including interception by vegetation, evaporation, and infiltration

$S_\mathrm{i} =$ Potential maximum retention after runoff begins, or the maximum possible difference between Q and P as $P \to \infty$

The familiar curve number is related to the storage index, S^*, by:

$$CN = \frac{1000}{10 + S_\mathrm{i}} \tag{4.9}$$

Throughout the history of the Curve Number method, the initial abstraction was calculated as $I_a = 0.2 S_\mathrm{i}$. Nevertheless, a recent study by Woodward (2003) concluded that the initial abstraction should be expressed as:

$$I_a = 0.05 S_\mathrm{i} \tag{4.10}$$

Indeed, this change was found to significantly improve the hydrologic model for the Aberjona River basin. The relationship between runoff, Q, and precipitation, P, is shown here for various curve numbers:

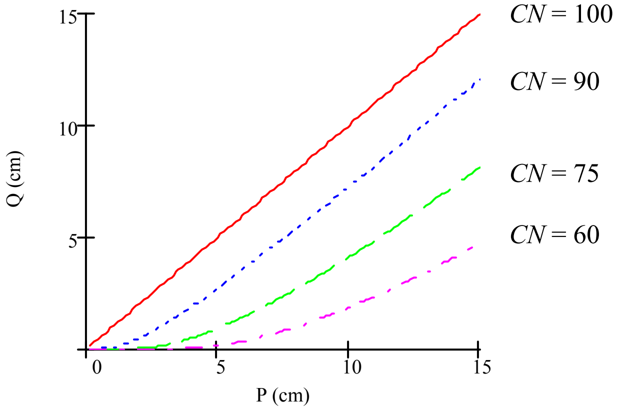


Figure 4.7 Rainfall-runoff curves for the SCS Curve Number method

Note that higher curve numbers indicate a more impermeable area. The maximum, *CN* = 100, means the land is completely impermeable and *Q* = *P*. The actual curve number for an area is based on three factors: (1) land use of the area; (2) antecedent moisture conditions; and (3) time of year. The antecedent moisture condition, *AMC*, is simply the sum of precipitation over the previous 5 days. The current "rule" for choosing the curve number based on *AMC* is presented in Figure 4.8. The breakpoints AMC_1 and AMC_2 in Figure 4.8 depend on the time of the year as reported in Table 4.7.

Table 4.7 Decision rule for determining breakpoints in the curve number calculation (from Haith et al., 1996b)

	Growing Season	Non-Growing Season
AMC_1 (in)	0.5	1.1
AMC_2 (in)	1.4	2.1

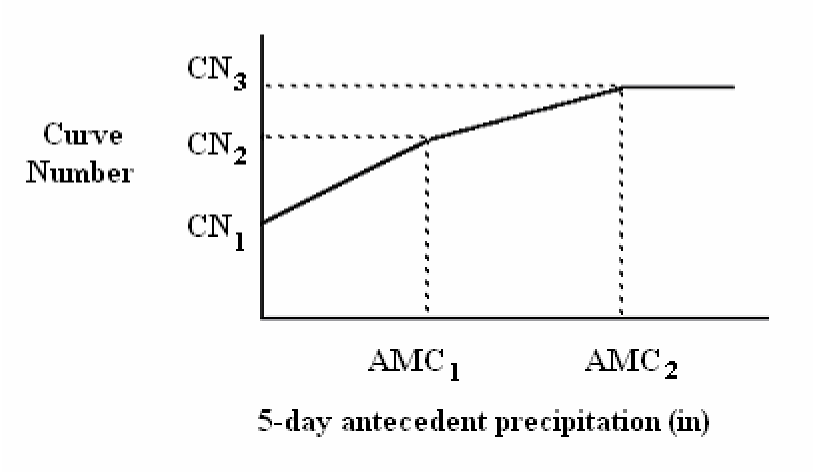


Figure 4.8 Curve number as a function of antecedent moisture condition (from Haith et al., 1996b)

It is important to remember that the SCS method was developed as an event model. It was easy for a worker to total up the last 5 days' rainfall, as it is easier than measuring soil moisture content. When this technique is applied to a continuous model, it often causes sudden jumps and discontinuities in the calculated CN, as in the example of Figure 4.9.

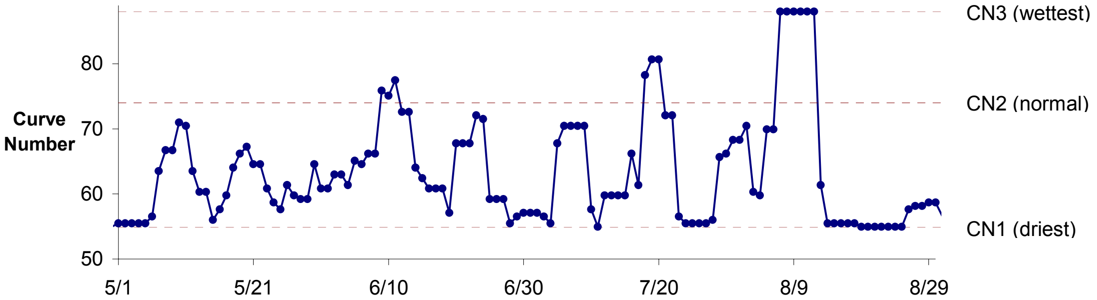


Figure 4.9 Example of curve numbers calculated for a daily simulation

Despite these shortcomings, it was decided to follow Haith et al. (1996a) in incorporating the method into the simulation model for two reasons. First, in the course of model development, it was found to perform as well or better than methods which relied on modeled soil moisture to determine runoff. Second, potential users of the model are likely to be familiar with this approach, and guidance is readily available on selecting curve numbers based on land use, soils, and other physical characteristics (e.g.: Cronshey et al., 1986, Haith et al., 1996b)

4.2.7 Flow Routing (Linear Pool Reservoir)

Flow routing was added in order to more accurately model the hydrograph resulting from a runoff event. Even in small watersheds, there is persistence to the streamflow that continues for hours or days after the end of a rainstorm. A variety of techniques have been developed to account for this, including the unit hydrograph (Singh, 1989). In the present case, the model makes use of a variable-depth linear pool reservoir. Figure 4.10 illustrates how the linear reservoir works. The bucket represents storm runoff from a rainstorm, and the bathtub represents the channel network of the watershed.

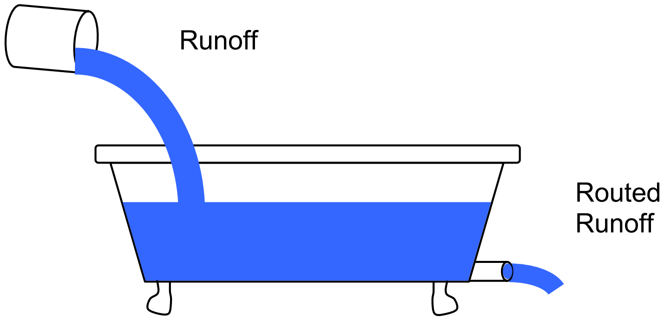


Figure 4.10 Linear pool routing reservoir

A mass-balance approach is used to describe the change in storage of the reservoir:

$$\frac{dR}{dt} = Q_{\text{in}} - Q_{\text{out}} \tag{4.11}$$

substituting,

$$\frac{dR}{dt} = Q_{\text{in}} - k_{\text{R}} R \tag{4.12}$$

Here, the inflow to the reservoir, Q_{in}, is a time series of the overland flow resulting from runoff, which is a proportion of the precipitation. A simple equation was written for outflow from the reservoir; at a given time, outflow is linearly proportional to the water level in the reservoir:

$$Q_{out} = k_{\text{R}} R \tag{4.13}$$

where:

$k_{\text{R}} = \text{reservoir constant } \left(\text{day}^{-1}\right)$

$R = \text{reservoir level } \left(\text{m}\right)$

Figure 4.11 is a demonstration of how the parameter k_{R} affects the outflow rate from the reservoir. Note that the time series of runoff events was artificially specified, in order to demonstrate the feasibility of this approach:

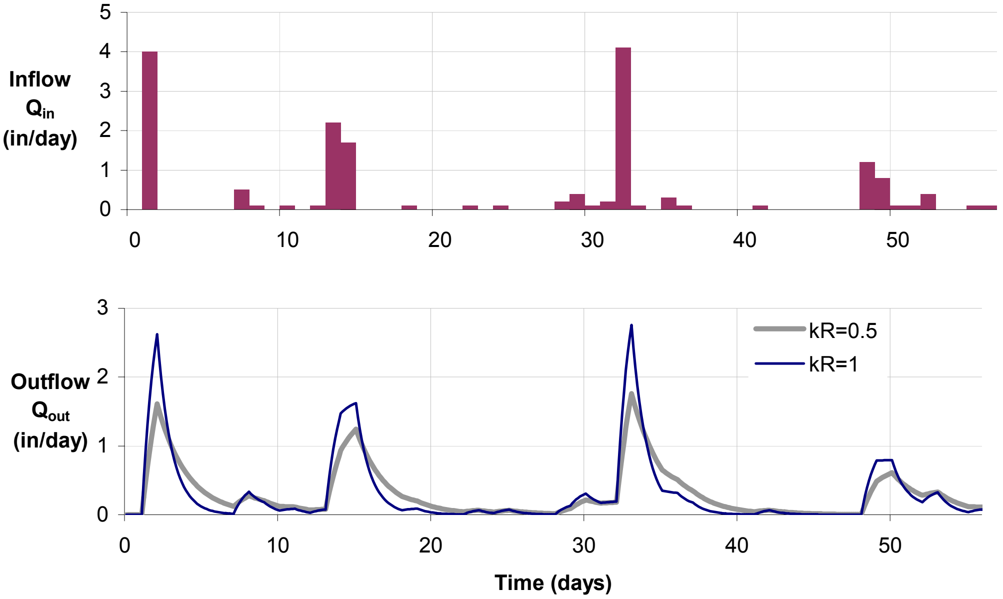


Figure 4.11 Linear routing reservoir performance; Q_{out} calculated by equation (4.13)

This simulation demonstrates a few important results. First, the value of the parameter k_R does not affect the time to peak of the outflow. In other words, changing k_R does not change the time to peak, which occurs at the end of the storm. Second, the outflow hydrograph has a peaked shape. Third, the recession is an exponential curve. An analytical solution for the level of the reservoir can be written for a unit input to the reservoir. For a level R_0 at time $t = 0$, the reservoir level is:

$$R = R_0 e^{-k_R t} \tag{4.14}$$

Finally, the reservoir never empties completely. The reservoir level decreases such that the reservoir volume R asymptotically approaches zero as $t \rightarrow \infty$. A useful rule of thumb is that reservoir's volume drains by 95% at a time of about $3/k_R$ days. This fact should give the model user some guidance in choosing a reservoir constant, k_R, especially where streamflow records are available.

Outflow from the routing reservoir can also be expressed as the volume of the reservoir raised to a power, such as 3/2, as in equation (4.15). This function more closely resembles traditional weir equations, and also the uniform flow equation. After some experimentation, it was found that this did not improve the model fit. Thus, it was decided that the extra complication did not confer a discernable advantage.

$$Q_{out} = k_{\mathrm{R}} \cdot R^{\frac{3}{2}} \tag{4.15}$$

Using two or more linked 'cascading reservoirs' was also investigated as a way to account for the time lag before the peak flow, i.e., the time of concentration of the watershed. The cascading reservoirs can be conceptualized as two bathtubs, as in Figure 4.12. Simulated output for a system of two cascading linear reservoirs, with $k_{\mathrm{R}1} = k_{\mathrm{R}2} = 1$ is shown in Figure 4.13.

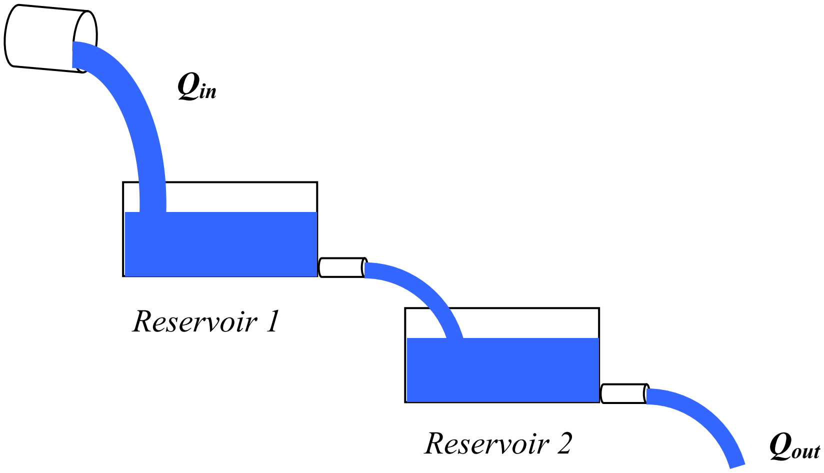


Figure 4.12 Cascading reservoirs

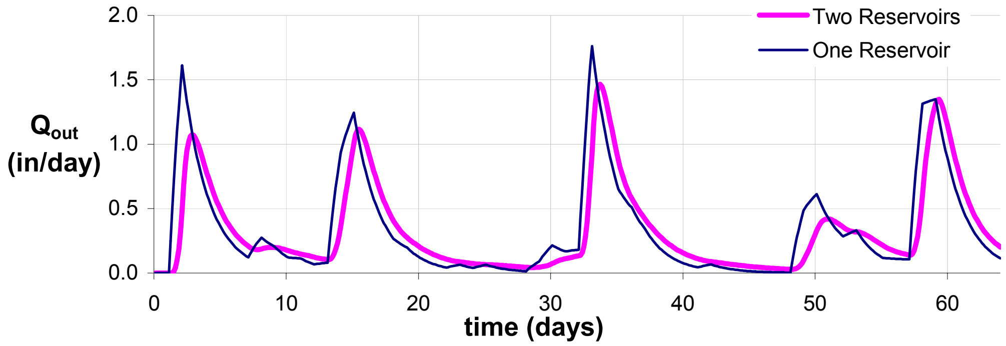


Figure 4.13 Performance of system of two cascading reservoirs

The results are promising, as the output signal of the two-reservoir model is beginning to more closely resemble a real hydrograph. In the end, however, it was decided that a simple, one-reservoir system was sufficient for the model. It was important to limit the complexity and number of parameters in the model. Additionally, as the model output is summed on a daily scale, any inaccuracies in the hydrograph time to peak should not be apparent, as long as the time of concentration of the modeled watershed has a time of concentration much less than one day. Hence, this approach is limited to relatively small watersheds. In larger watersheds, a more detailed channel runoff routing algorithm (Singh, 1989) would need to be incorporated into the model to account for a larger time of concentration.

4.2.8 Infiltration

Infiltration is the rate at which water moves into the soil moisture compartment of the model. Because the model does not include surface depression storage, infiltration is simply the difference between precipitation and runoff, expressed in inches per day:

$$\textit{Infiltration} = \textit{Precipitation} - \textit{Runoff} \qquad (4.16)$$

4.2.9 Percolation

Percolation is the rate at which water moves from the soil moisture (*S*) to groundwater (*G*). A simple function was adopted, where the percolation rate, is linearly proportional to the soil moisture (see, for instance, Singh, 1989):

$$Percolation = k_{\mathrm{P}} \cdot S \tag{4.17}$$

Even this simple function should be viewed as an improvement to GWLF. In GWLF, the daily percolation rate is quantified as the soil moisture in excess of the soil moisture capacity ($S_{\max}$):

$$Percolation = \left(S - S_{\max}\right) \cdot \left(1 \text{ day}^{-1}\right) \quad \text{for} \quad S > S_{\max} \tag{4.18}$$

$$Percolation = 0 \quad \text{for} \quad S \le S_{\max}$$

This relationship implicitly assumes a rate constant $k_{\mathrm{P}} = 1$. Having a constant $k_{\mathrm{P}} = 1$ can lead to wide swings in the soil moisture on a daily time scale. It was found that soil moisture should realistically be modeled with k_{P} □ 1. A model with k_{P} □ 1 incorporates a “resistance” where the percolation decreases as soil moisture nears zero. Again, because the GWLF model’s output is summarized monthly, these daily errors tend to be smoothed out. It was found, however, that substituting the linear function described above does significantly improve the daily model by producing a better fit to observed daily streamflow.

4.2.10 Recession

Recession is the rate at which water flows from the groundwater compartment, *G*, to the stream. It is fairly common to model recession as a simple linear function (Singh, 1989, page 172). In the following equation, k_{G} is the groundwater recession constant. Singh (1989), among others, gives guidance for choosing k_{G} based on inspection of streamflow records.

$$Recession = k_{\mathrm{G}} \cdot G \tag{4.19}$$

4.2.11 Evapotranspiration

The average actual evapotranspiration (ET) over the basin is expressed as a function of the potential evaporation (PET). In some areas pan evaporation records may be available which are good estimates of PET. However, such records are not available in many areas, such as our own watershed. Two empirical equations, the Hamon and Hargreaves methods, have been incorporated into the model to calculate the daily average PET from readily-available climate data, such as the daily average temperature.

In the Handbook of Hydrology, Shuttleworth (Maidment, ed., 1993, Chapter 4) recommends the Hargreaves equation (4.20) as the best of the temperature-based methods.

$$PET = 0.0023\, S_0 \left(T + 17.8\right) \cdot \sqrt{\delta_T} \qquad (4.20)$$

where:

S_0 = Water equivalent of extraterrestrial radiation in $\text{mm} \cdot \text{day}^{-1}$

$\bar{\delta}_T$ = Difference between the mean monthly maximum and mean monthly minimum temperatures

T = Average daily temperature in ºC

As the extraterrestrial radiation, S_0, depends on latitude, the model user must enter the latitude of the basin centroid when using this method. The model calculates S_0 internally based on standard formulas (Maidment, et al., 1993, Chapter 4).

Alternatively, the potential evapotranspiration may be calculated by the same function as GWLF, another empirical temperature-based method originally described by Hamon in 1963 (Maidment, ed. 1993).

$$PET = 29.8\, N \frac{e_S\left(T\right)}{T + 273.2} \qquad \text{mm} \cdot \text{day}^{-1} \qquad (4.21)$$

where

N = Maximum number of daylight hours

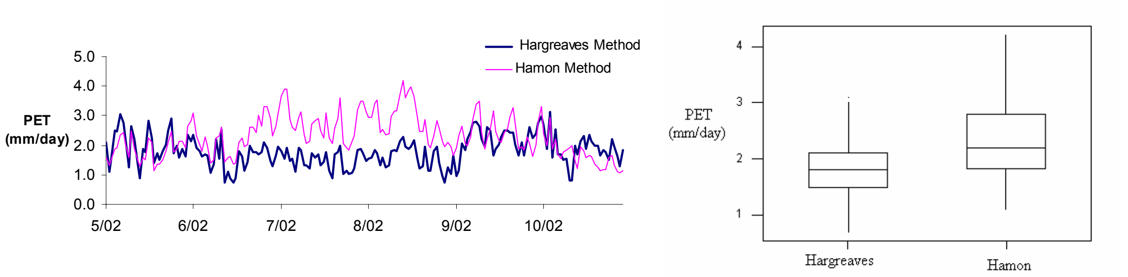
T = Average daily temperature in ºC

$e_S(T)$ = Saturation vapor pressure in kPa (or kN/m^2)

The saturation vapor pressure is the partial pressure of water vapor in a saturated air; a volume of air is considered saturated when it contains the maximum water vapor it can hold at a particular temperature. Saturation vapor pressure is approximated by the Clausius-Clapeyron equation (Maidment, ed., 1993):

$$e_S(T) \approx 0.6108 \exp\left(\frac{17.27\,T}{237.3 + T}\right) \qquad (4.22)$$

Comparing the two methods with climate data for the Mystic River basin in 2002, the Hamon method gives higher estimates of PET, as illustrated in Figure 4.14. Where actual measurements of pan evaporation or another surrogate of PET are not available, it is not possible to state *a priori* which of these empirical methods is a better estimate. However, inspection of annual records of annual climate and streamflow data may be useful. In general, long term evapotranspiration over a basin is equal to the precipitation minus the streamflow: $ET = P - Q$. Thus, the model user may wish to experiment with using both methods in order to obtain the best agreement with observations.



Mean ± St. Dev: 1.8 ± 0.5 2.3 ± 0.7 mm/day

Figure 4.14 Comparison of Hamon and Hargreaves estimates of PET

The calculated PET is the maximum possible ET that could occur given an abundant supply of water. Since the model does not have a land surface compartment, evapotranspiration is assumed to act only on water stored in the soil moisture compartment. Of course, in the real world, some water is lost from the surfaces of lakes and streams via direct evaporation. The simplifying assumption is made that open water represents a small fraction of the watershed, and therefore it is not worth the additional complexity to add a water surface evaporation function to the model. It stands to reason that the ratio of actual to potential evapotranspiration, ET:PET, will be affected by the soil moisture. In the simulation model, ET approaches PET as *S* approaches S_{max}. Similarly, as *S* approaches 0, ET approaches zero. This is expressed as a fraction,

$$\frac{ET}{PET} \propto \frac{S}{S_{max}} \tag{4.23}$$

An evaporation rate constant, *CV*, is introduced to the equation to give the following

$$ET = CV \cdot \frac{S}{S_{max}} \cdot PET \qquad \text{where } 0 < CV < 1 \tag{4.24}$$

This new function is a simplified version of that used in the Stanford Watershed Model (Singh, 1989, page 71) and HSPF (Aqua Terra Consultants, 1995). The coefficient *CV* is referred to as a “cover coefficient” in GWLF. *CV* varies as a function of the time of year, to capture the idea that, in a predominantly agricultural basin, growing crops will increase the evapotranspiration rate during the growing months. This formulation may still be appropriate in a basin with non-agricultural land uses (such as the Aberjona, which is predominantly suburban), because non-crop vegetation will increase ET during the spring and summer. Table 4.8 and Figure 4.15 show an example set of cover coefficients used in the calibration of GWLF for a basin in New York State (from Haith et al., 1996b).

Table 4.8 Example cover coefficients

Month	ET Cover Coefficient, CV
April	0.49
May	1.00
June	1.00
July	1.00
August	1.00
September	1.00
October	1.00
November	0.49
December	0.49
January	0.49
February	0.49
March	0.49

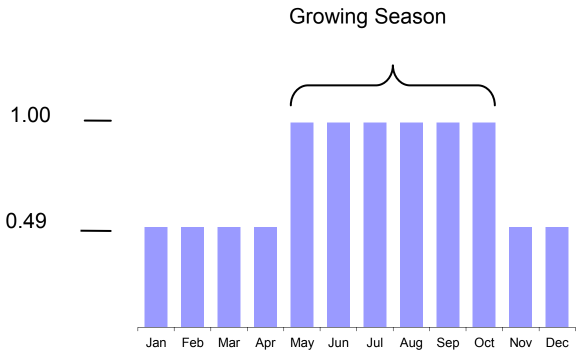


Figure 4.15 Example cover coefficients

Rather than entering a new cover coefficient for each month, it makes sense to write a function so that *CV* varies as a function of time. There are a number of functions that one can imagine to do this, for instance a sinusoid or a Gaussian curve. Initial trials with a sinusoidal function for CV showed that it was feasible. It was ultimately rejected, however, as it added to the model complexity without significantly improving the fit.

Figure 4.16 is a plot of the actual ET rate as a function of soil moisture, with two different PET rates. Note that ET continually increases, until ET = PET when the soils are saturated, i.e., $S = S_{max}$.

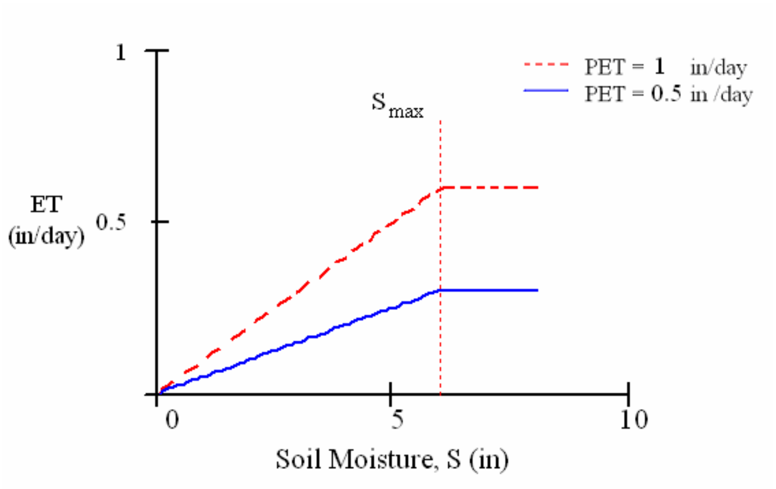


Figure 4.16 Actual ET rate, for different rates of PET

Figure 4.17 is a plot of the actual ET rate as a function of PET. If soil moisture is held constant, ET is a constant proportion of the PET. When $S = S_{max}$ there is 1:1 relationship of ET:PET.

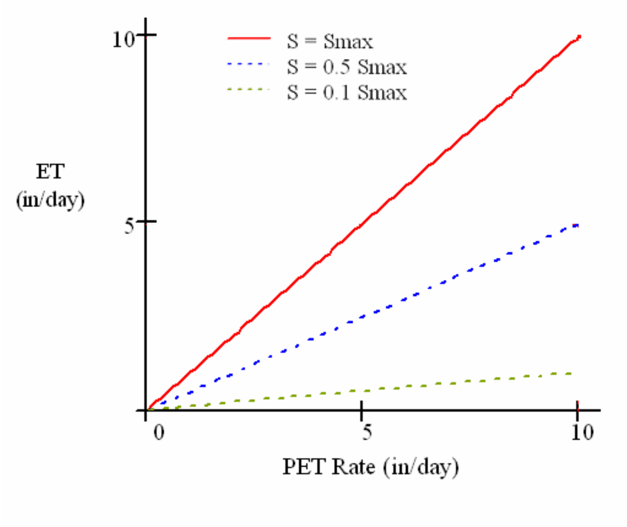


Figure 4.17 Actual ET rate, for varying soil moisture content, *S*

4.2.12 Evapotranspiration and Soil Moisture Accounting in GWLF

The formulation of ET as a function of *PET* and *S* described above should improve the daily output of the GWLF model. The functions affecting soil moisture accounting in GWLF are a likely contributor to inaccuracies in daily flow predictions. In the original GWLF model, ET is calculated as follows (note that Haith et al.'s nomenclature has been modified to make it consistent with variables in this thesis):

$$ET_t = \mathrm{Min}(CV_t \cdot PET_t;\ S_t + P_t - Q_t) \tag{4.25}$$

where

t = day

CV = Cover coefficient

PET = Potential evapotranspiration

S = Unsaturated soil zone moisture

P = Precipitation

Q = Runoff

There is a flaw in the function that only becomes apparent on close inspection. Assuming for the moment that CV =1, and that runoff and precipitation are zero, the function becomes:

$$\begin{aligned} &\text{If } \ PET < S \ \text{ then } \ ET = PET \\ &\text{If } \ PET > S \ \text{ then } \ ET = S \cdot \left(1\ \text{day}^{-1}\right) \end{aligned} \tag{4.26}$$

With an implicit rate constant equal to one, if the potential evapotranspiration is high, it may drive S to zero in one day. And on days when there is sufficient water in the soil moisture compartment, ET will equal the PET. This leads to significant overestimates and underestimates of daily ET values, and wide fluctuations in soil moisture. This can be seen in the simulation in Figure 4.18.

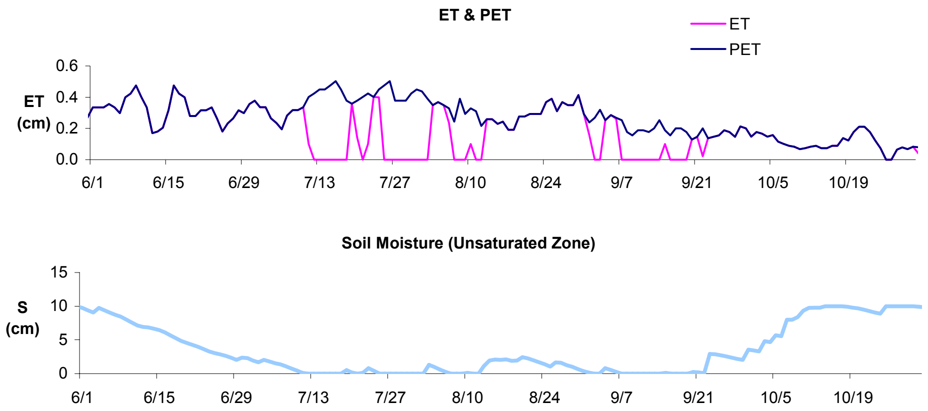


Figure 4.18 Example of calculated ET, PET, and soil moisture in a daily GWLF simulation

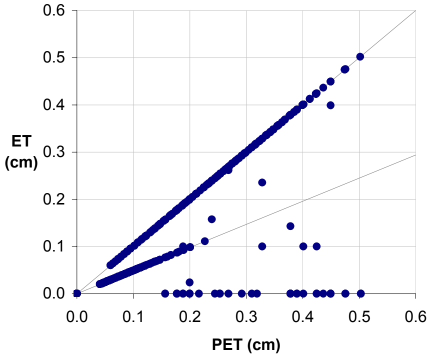


Figure 4.19 ET versus PET in a GWLF daily simulation

Note that almost all of the points are either on one of the lines corresponding to $CV \cdot PET$, where $CV = 0.49$ or $CV = 1.0$. In fact, only 8 out of 724 points take on values that are not equal to either $CV \cdot PET$ or zero. Again, the inaccuracies in GWLF described above are of little importance when dealing with monthly output. The new functions described above should allow the daily output to be interpreted with more confidence.

4.2.13 Bacteria Loading Model

In the conceptual model, there are *three* sources of bacteria load to the stream, summarized in Table 4.9.

Table 4.9 Sources of bacteria in the model

Source	Symbol	Units
Runoff Load	W_R	$\frac{\#\ \text{organisms}}{\text{day}}$
Background Concentration	$C_{Background}$	$\frac{\#\ \text{organisms}}{\text{m}^3}$
Point Source Load	W_P	$\frac{\#\ \text{organisms}}{\text{day}}$

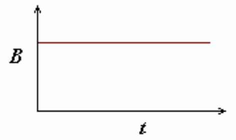
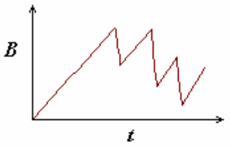
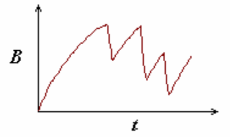
Several options for calculating the nonpoint source bacteria loading are available in the model. The simplest assumption is to assume that all polluted runoff is the same. That is to say, the bacteria concentration for all runoff is a constant, c_R. While stormwater concentrations can vary over several orders of magnitude in practice, assuming a constant concentration is equivalent to an event mean concentration (EMC) (see for instance, Whipple, 1983). When this option is chosen, a single parameter, c_R, is entered by the user; thus, there are a minimum of 'tuning factors' to calibrate the model.

A further refinement of this approach is to assume that as the runoff flow increases, the concentration also increases; thus, the runoff concentrations is proportional to the runoff flow. It stands to reason that a larger volumetric flow rate, which also has a higher velocity, will have more energy to pick up and transport sediment and other pollutants from land surfaces and in stormwater pipes. This is represented mathematically as a washoff function, described in section 4.2.15.

A further refinement of the bacteria loading model assumes that the availability of pollutants also varies with time. The use of a buildup function follows work by Sartor and Boyd (1972) to quantify nonpoint source sediment loads, which has subsequently been incorporated into well-known models such as SWMM and HSPF. It fits with our common sense that nonpoint source pollutants should build up over time. After a large rainstorm, land surfaces and pipes are ‘swept clean’, and, pollutants begin building up anew. A few engineers have extended this technique to bacteria loading, usually with fecal coliforms, and usually in the context of a TMDL study (see for instance Interstate Commission on the Potomac River Basin, 2002). The approach is attractive when one has a good dataset, as it contains a handful of calibration parameters that can be ‘tuned’ to fit the model to observed concentrations.

To maintain flexibility, a variety of buildup and washoff options were coded into the model, allowing the user to experiment and find the most appropriate. An overview of these is in Table 4.10 and Table 4.11. The buildup and washoff functions are described in detail in sections 4.2.14 and 4.2.15, respectively.

Table 4.10 Buildup options included in the model

Buildup Method	Number of Parameters	Equation	Example
Constant	1[*]	$\frac{dB}{dt} = 0$	
Linear	2	$\frac{dB}{dt} = k_B$	
Exponential	3	$\frac{dB}{dt} = k_B - \alpha B$	

[*] As an initial value, B_0, must be provided to solve for $B(t)$ in each case, each formulation should be considered to have one more parameter than those appearing in the equation.

Table 4.11 Washoff options in the model

Washoff Method	Number of Parameters	Equation
Constant concentration	1	$W = c_R \cdot Q$
Fraction of buildup washed off $\propto$ flow	1	$W = \left(1 - e^{-k_W Q}\right) B$
Washoff load carrying capacity $\propto$ flow	2	$W = a_{\mathrm{w}} Q^{b_{\mathrm{w}}}$

4.2.14 Buildup Function

Huber and Dickinson (1992, page 127), cautions that "it is naive to assume that empirical washoff equations truly represent the complex hydrodynamic (and chemical and biological) processes that occur while overland flow moves in random patterns over the land surface…the true mechanisms of buildup involve factors such as wind, traffic, atmospheric fallout, land surface activities, erosion, street cleaning and other imponderables."

In an oft-cited report, Sartor and Boyd (1972) derived relationships to quantify the amount of dust and dirt that accumulates on streets in areas under different land uses, as shown in Figure 4.20.

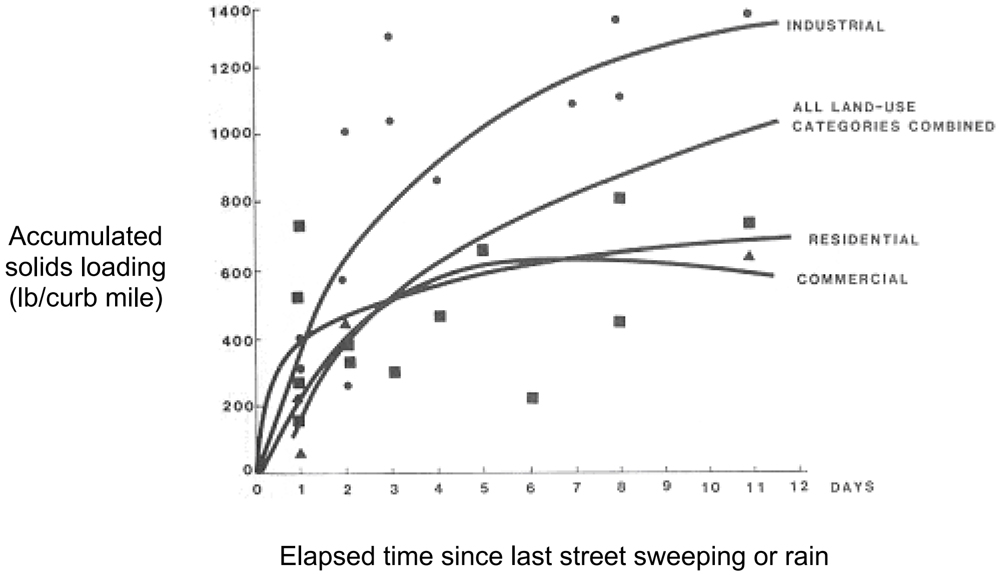


Figure 4.20 Pollutant loading for different land uses (from Whipple, 1983, redrawn from Sartor and Boyd, 1972)

For the sake of completeness, the model includes a ‘constant buildup’ option. With this option, the amount of available pollutant does not change with time. Rather, the initial pollutant level, B_0, persists throughout the simulation.

Two forms of time-variant buildup function are commonly used: linear and exponential. A linear buildup function has the advantage of being the simplest, and has the fewest parameters.

$$\frac{dB}{dt} = k_B \tag{4.27}$$

Solving with the initial condition, $B = B_0$ at time $t = 0$ gives the following solution

$$B = B_0 + k_B \cdot t \tag{4.28}$$

The time series of buildup and washoff typically looks like Figure 4.21. (In this example *B* on the ordinate represents the density of bacteria organisms built up on land surfaces, expressed as number of organisms per square meter.)

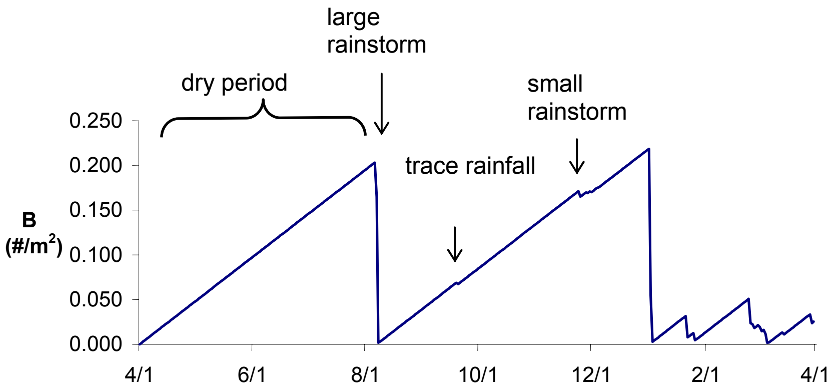


Figure 4.21 Linear buildup function, hypothetical example based on equation (4.30)

Exponential buildup models factor in the idea of a maximum buildup. As the buildup increases, inhibition begins to affect the rate of buildup. This formulation takes into account the idea that die-off may occur during the buildup. The mathematical formulation for this function is as follows:

$$\frac{dB}{dt} = k - \alpha B \tag{4.29}$$

where *α* is an inhibition factor that depends on the mass of accumulated pollutant. Integrating, the solution is:

$$B(t) = B_0 e^{-\alpha t} + \frac{k}{\alpha}\left(1 - e^{-\alpha t}\right) \tag{4.30}$$

This relationship is demonstrated in Figure 4.22.

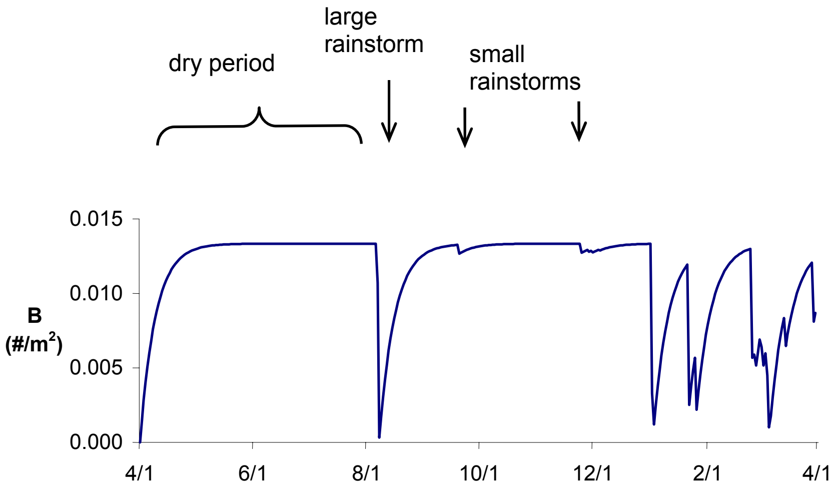


Figure 4.22 Exponential buildup function, hypothetical example based on equation (4.30)

While *t* is very small, the buildup rate is close to the rate k_B, as demonstrated in Figure 4.23.

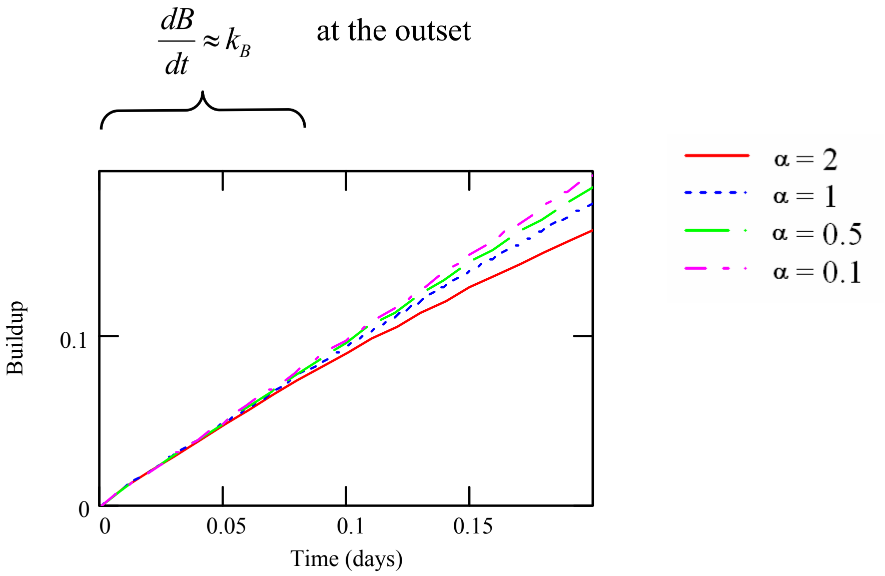


Figure 4.23 Effect of inhibition factor, *α*, on buildup rate

As the time t increases, the parameter α affects the buildup in two ways. It affects both the maximum buildup, $B_{\max}$, that occurs as $t \to \infty$, as well as how long it takes to get to a fixed percentage of the maximum, as demonstrated in Figure 4.24.

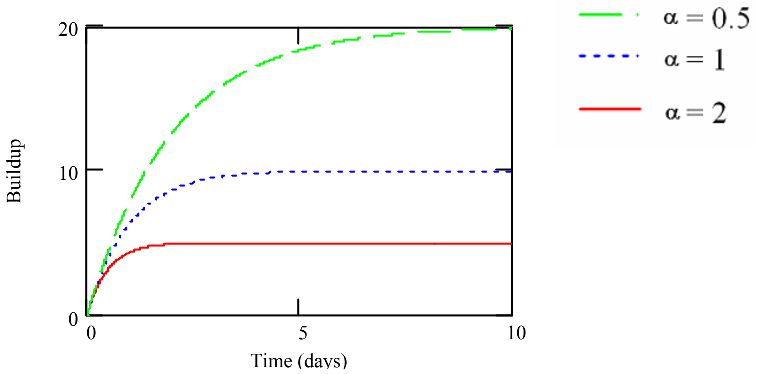


Figure 4.24 Effect of inhibition factor, *α*, on maximum buildup

It can be shown that the maximum buildup that can occur is:

$$\lim_{t \to \infty} B(t) = \frac{k_B}{\alpha} \tag{4.31}$$

How long will it take for the buildup to reach 90% of the maximum? Rewriting equation (4.30), and assuming that the buildup at time 0 is $B_0 = 0$:

$$0.90 \cdot \frac{k_B}{\alpha} = \frac{k_B}{\alpha}\left(1 - \exp\left[-\alpha \cdot t_{90}\right]\right) \tag{4.32}$$

Solving gives $t_{90} = 2.3/\alpha$. Similarly, it can be shown that 95% of the saturation buildup will accumulate at $t_{95} = 3.0/\alpha$. This simple relationship between the 90% buildup time and the parameter α is shown in Figure 4.25. As the concepts of a maximum 'saturation' buildup and 90% saturation time are more intuitive than to use than the parameters k_B and α, these relationships should be useful to the model user in choosing an appropriate set of parameters.

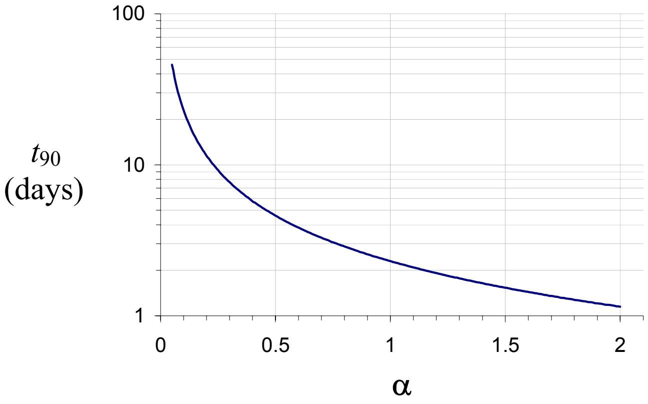


Figure 4.25 Time at which buildup reaches 90% saturation versus inhibition factor *α*

4.2.15 Washoff Function

In the simulation model, pollutants are removed from land surfaces by rainfall-induced runoff. In reality, it is not be this simple, as washoff will be influenced by other factors such as wind, street sweeping, etc. Not all runoff events have the same power to remove pollutants. It stands to reason that a runoff event with a larger flow rate will have more power to carry pollutants. This is analogous to sediment transport in a river bed, where the sediment transport capacity is expressed as the flow raised to a power.

$$Washoff \propto Q^{\beta} \tag{4.33}$$

Two formulations for washoff functions are commonly encountered in the literature:

Washoff = *mass* of pollutant (g)

Washoff = *fraction* of accumulated pollutant, $0 < W < 1$

Barbé et. al. (1996) used the following relationship in a statistical study of phosphorus loading in a suburban watershed in coastal Texas:

$$Washoff\,(\text{kg}) = \alpha V^{\beta} \tag{4.34}$$

where

$V =$ Runoff Volume (m^3)

α and $\beta =$ parameters

Barbé et. al. determined the optimal coefficients α and β by minimizing the sum of the squared errors.

The washoff function in GWLF expresses the pollutant load, *W,* as a *fraction* of the built up pollution that is washed off:

$$W = \left(1 - e^{-k_W Q}\right) \cdot B \qquad (0 < W < 1) \tag{4.35}$$

This approach helps avoid the following problem: if an inappropriate set of parameters is chosen and the buildup rate exceeds washoff, the pollutant buildup may increase unchecked, as in the example of Figure 4.26.

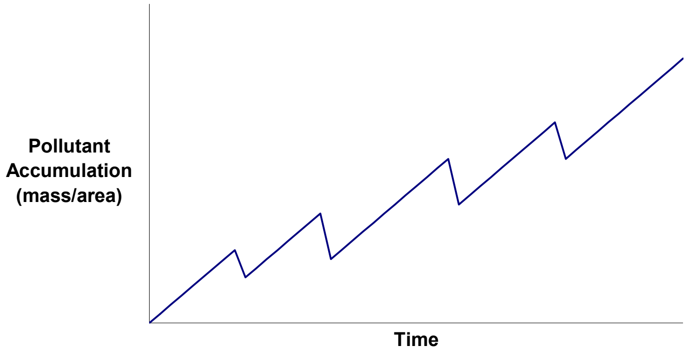


Figure 4.26 Unchecked pollutant accumulation arising from an inappropriate set of buildup and washoff parameters

Haith chose $k_W = 1.81\ \text{cm}^{-1}$ ($4.6\ \text{in}^{-1}$), because this ensures that 90% of the pollutant washes off during a 0.5 in (1.27 cm) runoff event. However, as the simulation model is intended to be calibrated to fit observed bacteria, this has not been hard-wired into the model, but reserved as a 'tuning factor'. Figure 4.27 shows the effect of changing this parameter, k_W, on the washoff ratio, W:

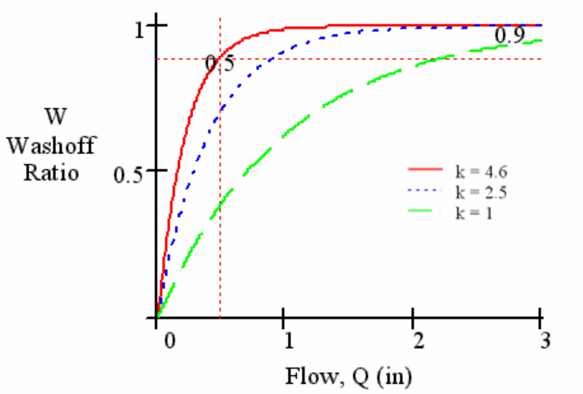


Figure 4.27 Effect of the flow volume on the washoff ratio.

It can be shown , as in equation (4.32), that 90% washoff of pollutants occurs for a runoff of $Q = 2.3/k_W$. Similarly, 95% of the accumulated buildup is washed off for $Q = 3.0/k_W$. Again, this rule of thumb should be a useful aid in choosing an appropriate set of parameters during calibration. A plot of the runoff depth required for 90% washoff is shown in Figure 4.28.

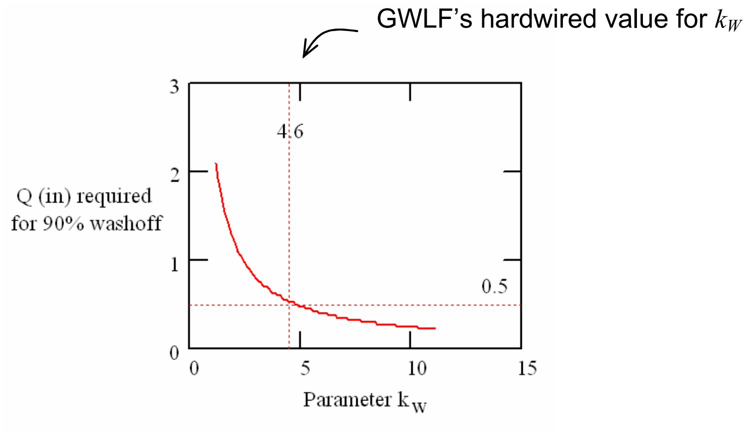

GWLF's hardwired value for k_W

Figure 4.28 Effect of the parameter k_W on the runoff volume required to wash off 90% of accumulated pollutants.

4.2.16 Calculating the bacteria load and concentration

The bacteria buildup and washoff are expressed in units of $\# \text{ bacteria}/\text{m}^2 \cdot \text{day}$. As this is a lumped-parameter model, we are assuming a homogeneous load for the entire watershed. Multiplying by the watershed area 'scales up' the load to the units of # organisms/day:

$$W_R = \text{Runoff Load}\left(\frac{\# \text{ bacteria}}{\text{day}}\right) = \mathit{Washoff}\left(\frac{\# \text{ bacteria}}{\text{m}^2 \cdot \text{day}}\right) \times \mathit{Area}\left(\text{m}^2\right) \quad (4.36)$$

4.2.17 Routing Bacteria with Runoff

As shown in Figure 4.5, bacteria washed off by runoff stays with that runoff flow as it passes through the routing reservoir. Bacteria in the reservoir die off over time according to a first-order decay rate. This is a common assumption in modeling bacteria in lakes and streams (Bowie et al., 1985, page 437; Chapra, 1997, lecture 27). Bacteria die off has been shown to be correlated with factors such as salinity, temperature, and solar radiation. Bowie et al. (1985) report a range of coliform decay rates from the literature, varying from 1–3.5 day^{-1}. Chapra (1997) describes models where the base mortality rate for coliform bacteria in freshwater is assumed as 0.8 day^{-1}. Decay rates for enterococci may be different (possibly lower) than the values reported for fecal coliform bacteria in the literature. Some models incorporate a temperature-dependent decay rate (e.g., HSPF), however this approach is not followed here as it is felt to be an unnecessary complication.

It is assumed that the reservoir is well-mixed, and therefore can be modeled as a continuously-stirred tank reactor, as shown in Figure 4.29 (modified from Chapra, 1997).

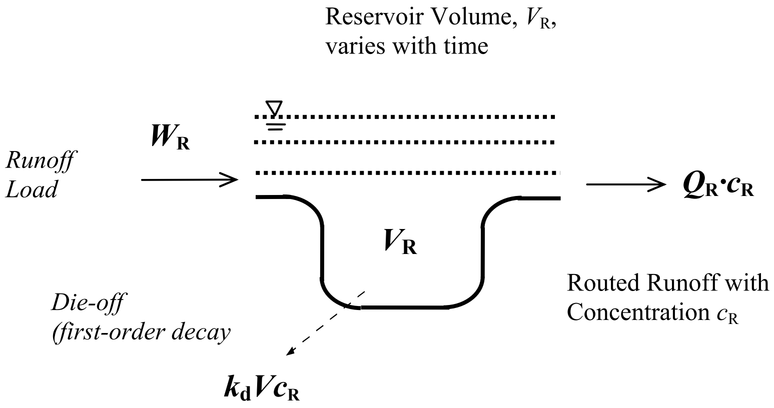

Reservoir Volume, V_R, varies with time
W_R
Runoff Load
$Q_R \cdot c_R$
V_R
Routed Runoff with Concentration c_R
Die-off (first-order decay
$k_d V c_R$

Figure 4.29 Routing and decay of bacteria in runoff

The differential equation governing the rate of change of number of bacteria in the reservoir is:

$$\frac{dm_R}{dt} = W_R - Q_R c_R - k_d V_R c_R \tag{4.37}$$

where:

m_R = number of bacteria (#)

V_R = Volume of the reservoir (m^3) $= R(\mathrm{m}) \cdot \mathrm{Area}\left(\mathrm{m}^2\right)$

c_R = concentration (#/m^3)

k_d = bacteria decay rate (day^{-1})

The concentration in the reservoir is:

$$c_R = \frac{m_R}{V_R} \tag{4.38}$$

4.2.18 Background Bacteria Source

In addition to adding a point source as a constant loading rate of W_P (# organisms per day), the modeler can also add a fixed 'background' concentration, c_B. The background load is calculated by:

$$W_B(t) = Q_B(t) \cdot c_B \tag{4.39}$$

where:

Q_B = Baseflow (portion of the streamflow that is from groundwater rather than runoff (m^3/day)

c_B =Background bacteria concentration (#/m^3)

A significant seasonal trend is present in the Aberjona River data bacteria data, which have guided the development of this model. Bacteria concentrations during dry weather, when baseflow predominates, are higher in June, a time of higher flows, than they are in August, shown in Figure 4.30.

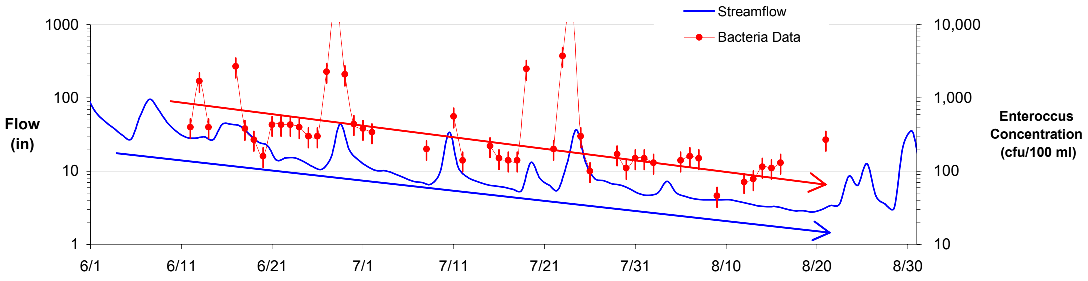


Figure 4.30 Seasonal trend in baseflow and bacteria concentration in the Aberjona River, summer 2002

A simple mechanism was postulated to explain this phenomenon. It was hypothesized that the baseflow has a constant bacterial concentration. At low flows water travels more slowly downstream, means there is a longer residence time in the river where more settling and decay may occur. This would tend to cause a lower concentration at low flows. The converse is true at high flows; water passes through the system quickly, allowing little time for bacteria to die off before reaching the basin outlet.

This phenomenon was captured by the creation of a 'tank model' with a first-order decay coefficient. This is similar to the routing reservoir; the key difference to the 'stream tank' is that it has a constant volume. Routing the baseflow through the routing reservoir described above would be inappropriate; it can be shown that the residence time in the reservoir is a constant, equal to the inverse of the reservoir outflow coefficient, $1/k_{\mathrm{R}}$. The routing reservoir's outflow is linearly proportional to the volume in the reservoir: $Q_{OUT} = k_{\mathrm{R}} \cdot V_{\mathrm{R}}$.

Therefore the residence time can be expressed as:

$$\tau = \frac{V_R}{Q_R} = \frac{V_R}{k_R \cdot V_R} = \frac{1}{k_R} \qquad (4.40)$$

Having a settling tank with a constant residence time does *not* yield the desired effect, increased bacteria die-off during periods of low flow. Therefore a new 'stream tank' with a constant volume V_S was created for the model. The flow through the tank is equivalent to the baseflow, calculated as groundwater recession by equation (4.19).

$$Q_{In} = Q_{Out} = Q_B \qquad (4.41)$$

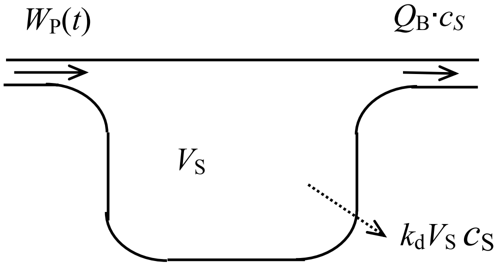


Figure 4.31 Stream tank for calculating the background bacteria concentration (after Chapra, 1997)

The differential equation for the rate of change of number of bacteria in the tank is:

$$\frac{dm_S}{dt} = Q_B \cdot c_B - Q_B \cdot c_S - k_d \cdot V_S \cdot c_S \tag{4.42}$$

where:

m_S = number of bacteria in the stream tank (#)

V_S = Volume of the instream tank (m^3)

c_S = concentration in the stream tank (#/m^3)

k_d = bacteria decay rate (day^{-1})

To help the model user choose an appropriate size for the tank, the concept of residence time, τ , is useful:

$$\tau(\text{days}) = \frac{V(\text{m}^3)}{Q(\text{m}^3 \cdot \text{day}^{-1})} \tag{4.43}$$

A flow-duration curve for the baseflow component of the hydrograph is shown in Figure 4.32. The plot was constructed for measured daily flow in the Aberjona River, May 1–October 31, 2002. To the right, residence time is plotted versus its exceedance probability. The data shown here are for V_S = 20,000 m^3.

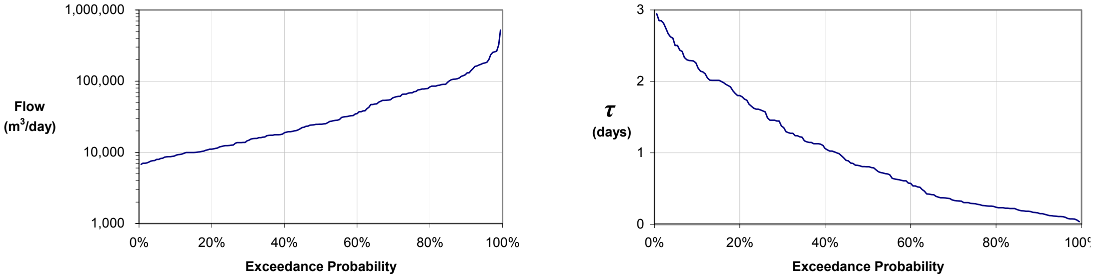


Figure 4.32 Exceedance probability plots for baseflow and stream tank residence time

At relatively high flows, such as those that are exceeded more than half the time, residence time $\tau < 1$ day, the flow will pass quickly through the reactor, not leaving much time for the die-off to occur. However, at the lower flows, the water will remain in the reactor for 2–3 days, during which time more decay will occur. At steady state (which is approximated when the inflow is steady for a few days in a row), the change in mass is zero.

$$\frac{dm_S}{dt} = 0 \tag{4.44}$$

Thus,

$$Q_B \cdot c_B - Q_B \cdot c_S - k_d \cdot V_S \cdot c_S = 0 \tag{4.45}$$

This can be rewritten as:

$$c_S = c_B \left(\frac{Q}{Q + k_d V_s} \right) = c_B \left(\frac{1}{1 + \frac{k_d V_S}{Q}} \right) = c_B \left(\frac{1}{1 + k_d \tau} \right) \tag{4.46}$$

For the low-flow case above with a decay rate k_d = 1 day^{-1} and a maximum residence time $\tau = 0.5$ days, the concentration in the stream equals one quarter the background concentration:

$$c_S = c_B\left(\frac{1}{1+1\cdot 3}\right) = \frac{1}{4}c_B \tag{4.47}$$

Similarly, at low flows, when $\tau = 3$ days, $c_S = 0.67c_B$. Thus, the tank effectively reduces the background concentration by approximately 75% at low flows and 23% at high flows. Changing one or both of the parameters, the tank volume, V_S, (and hence the residence time) or the background concentration, c_B, should allow for the simulation of a wide range of behaviors for background or 'dry weather' bacteria concentrations. The drawback, obviously, is the introduction of two new parameters, further complicating the model.

4.2.19 Calculating Instream Concentration

The concentration in the stream is simply the load divided by the flow:

$$C = \frac{\sum W}{Q} = \frac{W_{\text{Background}} + W_{\text{Point Source}} + W_{\text{Runoff}}}{Q} \tag{4.48}$$

Verifying the units:

$$\left(\frac{\text{organisms}}{\text{m}^3}\right) = \frac{\left(\text{organisms}\cdot\text{day}^{-1}\right)}{\left(\text{m}^3\cdot\text{day}^{-1}\right)}$$

The model’s code converts the concentration to the standard units of organisms/100 mL by multiplying by 10^4.

4.3 Results

4.3.1 Model Calibration

A two-step approach was employed for calibrating the simulation model. The first step was to calibrate the hydrologic portion of the model. The goal was to obtain the best fit to observed streamflow where measurements were available. Output from a 'manually-calibrated' hydrologic model for the Aberjona River in summer 2002 is shown in Figure 4.33; parameters were determined by trial and error, rather than using any sort of an optimization routine. The calibration parameter set is given in Table 4.12.

Table 4.12 Simulation model parameters for Aberjona River 2002 calibration

Hydrologic parameters		Bacteria loading parameters	
Watershed Area (square miles)	23.5	W_P	7×10^7
latitude (radians)	0.74	c_B	2,000
CN_2	88	c_R	50,000
CV	1.0	V_S	150,000
PET method	Hamon	k_d	1.4
S_0	1.0		
G_0	0.5		
R_0	0.1		
B_0	60		
S_{max}	0.2		
k_P	0.05		
k_G	0.025		
k_R	0.6		
Withdrawal	0		

An indication of the model fit is shown in the scatterplot in Figure 4.34 and cumulative frequency distribution of observed and modeled streamflow in Figure 4.35.

Note that the model fits well during certain times of the calibration period (e.g., very close agreement during the month of July. Note also that while some storm peaks are over predicted, others are under predicted. Sources of error include precipitation records; data from a single gage may not be representative of the average precipitation over the basin, especially during the summertime, when convective storms predominate (Hydroscience Inc., 1979, page 3-12). Nevertheless, the fit is reasonable for a simplified lumped-parameter model. The model does a fairly good job at recreating the mean and variance of observed flows, as shown in Table 4.13.

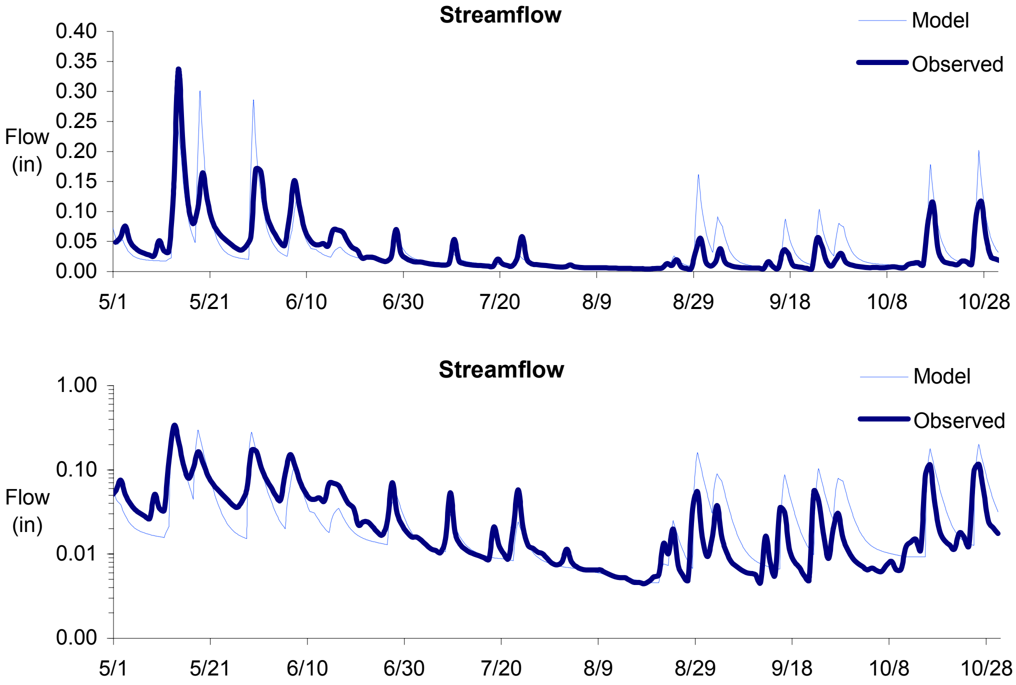


Figure 4.33 Observed and predicted streamflow for the Aberjona River, 5/1/02–10/31/02

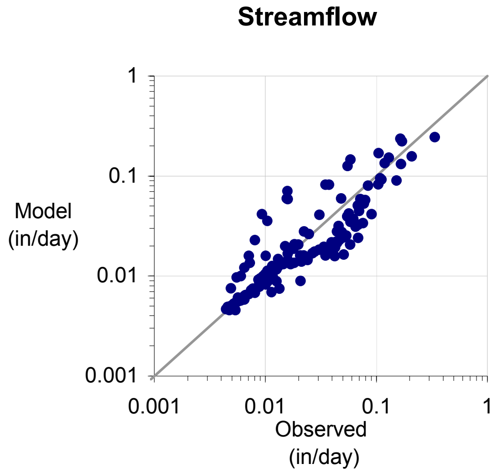


Figure 4.34 Observed versus predicted flows for the Aberjona River, 5/1/02–10/31/02

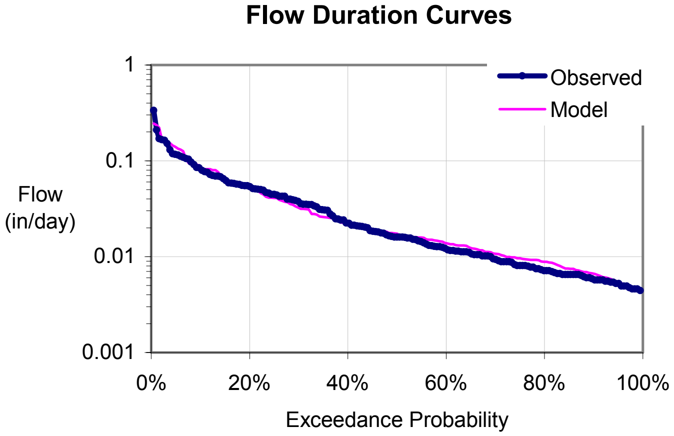


Figure 4.35 Exceedance probability plot for observed and predicted flows for the Aberjona River, 5/1/02–10/31/02

Table 4.13 Comparison of observed and modeled streamflow for the Aberjona River, summer, 5/1/02–10/31/02

	Observed	Model
Mean	21.4	21.9
Standard deviation	27.1	29.1

Once the hydrologic model was fit, the parameters of the bacteria loading submodel were adjusted to fit observed bacteria data. Parameters that govern background bacteria levels were adjusted, such as the background concentration, c_B, and the point sources, W_P. Further, parameters governing nonpoint source loading (i.e., buildup and washoff parameters) were adjusted.

Several assumptions were made in order to apply the model to the dataset. First, it was assumed that during dry weather, the sample collected around 10:00 a.m. each day represented the daily average concentration. When more than one sample was collected during and after a rainstorm, a daily average was evaluated by calculating a flow-weighted composite.

First, the model was run with the 'constant concentration in runoff' option (i.e., buildup and washoff relationships were not used). The results for the calibration simulation are plotted in Figure 4.36 and Figure 4.37. In this simulation, the runoff concentration is c_R = 50,000 organisms/100 mL, and the bacteria decay rate is k_d = 1.4 day^{-1}. The complete parameter set is given in Table 4.12.

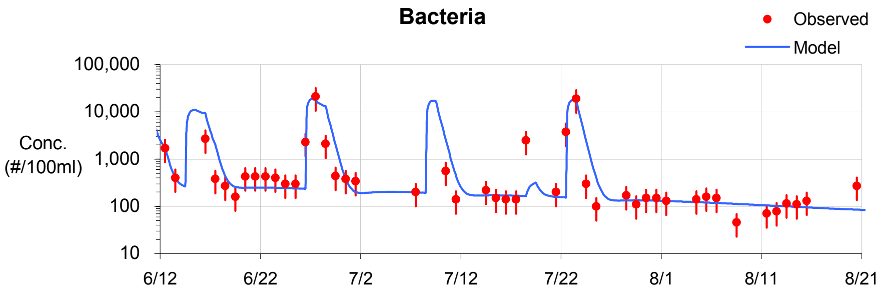


Figure 4.36 Time series plot of observed and predicted Enterococcus bacteria concentration in the Aberjona River, summer 2002

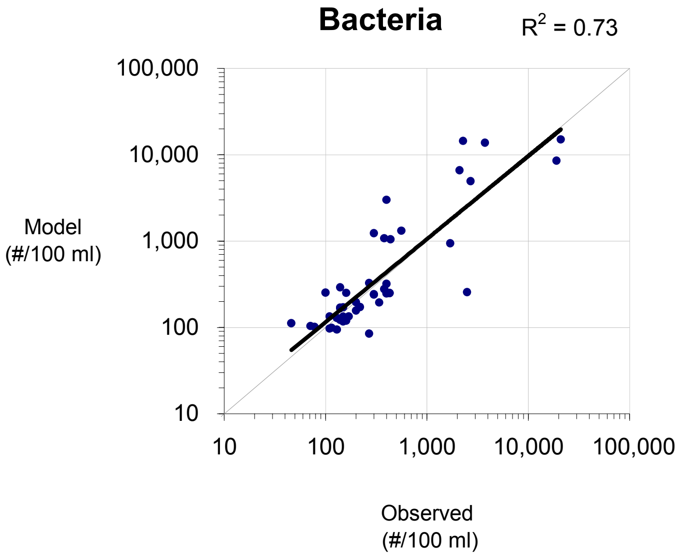


Figure 4.37 Scatterplot of observed and predicted Enterococcus bacteria concentration in the Aberjona River, summer 2002

A boxplot of the results, shown in Figure 4.38, demonstrates that the model is fairly good at reproducing the central tendency of the observations, but does not accurately describe the variance, over-predicting high values. Summary statistics of observed and modeled Enterococcus concentrations are reported in Table 4.14.

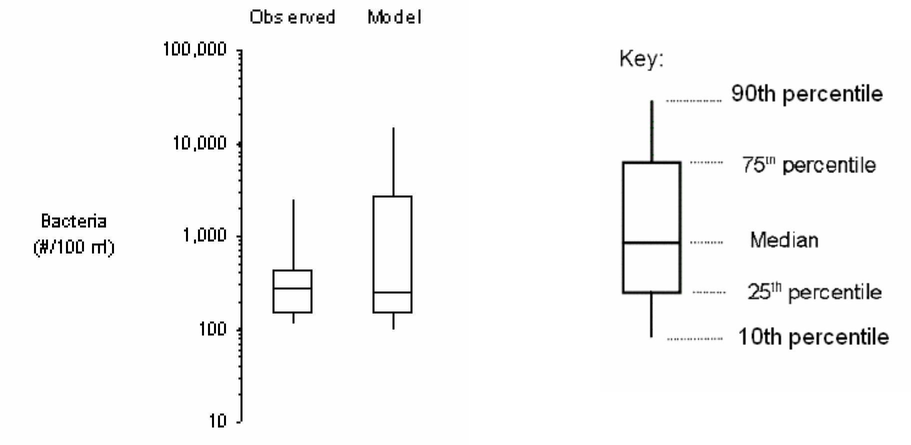


Figure 4.38 Boxplots of observed and predicted Enterococcus bacteria concentration in the Aberjona River, summer 2002

Table 4.14 Summary statistics for observed and predicted Enterococcus bacteria concentration in the Aberjona River, summer 2002

	Observed	Modeled
Median	270	245
Geometric mean	335	380
Interquartile Range	290	650

Finally, a plot of model residuals is shown in Figure 4.39. The residuals are calculated from the log bacteria concentration, or the i^{th} $\text{residual} = \log\left(C_{\text{i}}\right) - \log\left(\hat{C}_{\text{i}}\right)$. For the Aberjona River calibration year (2002) simulation, the residuals have a mean of –0.07 and a standard deviation of 0.25. Because the mean is close to zero, this is evidence that the model is not significantly biased. The variance of the residuals does not appear to be changing over time, meaning the residuals are approximately homoscedastic. However, there does appear to be some autocorrelation to the residuals indicating that there is some structure to the data the model did not capture. A hypothesis test was performed to determine whether or not there is a significant trend in the model residuals. When the residuals are regressed against time, the slope is not significantly different from zero (probability $P = 0.45$). This indicates that the model does not have a significant temporal or seasonal bias.

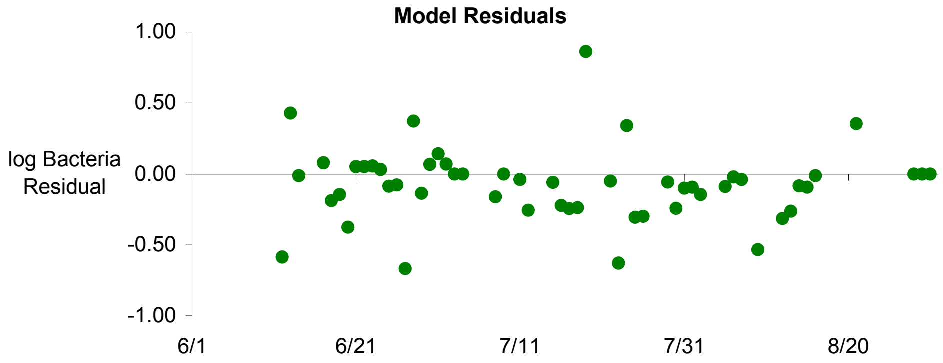


Figure 4.39 Simulation model residuals for Aberjona River Enterococcus bacteria data, summer 2002

As the model was programmed to include several options for simulating bacteria loading, it allows one to ask the question, "Does increasing the complexity in the model result in better predictions?" Results for the Aberjona River calibration year (2002) simulation, reported in order of increasing complexity in Table 4.15, indicate that this is indeed the case.

Table 4.15 Improved model accuracy with increasing model complexity for Aberjona River Enterococcus data, summer 2002

Bacteria Loading Model	Number of Parameters	Nash-Sutcliffe Efficiency, *E*	Root Mean Square Error (RMSE)
Constant Concentration	1	0.65	0.34
Linear Buildup	3	0.73	0.30
Exponential Buildup	4	0.75	0.28

Automatic Calibration

An attempt was made to find an optimal set of model parameters using an automatic optimization routine. An Excel Solver based on a genetic algorithm (Palisade

Corporation, 2001) was used with a variety of objective functions. Typically, the optimization routine was instructed to minimize the sum of squared errors (SSE) between the observed and modeled flows. In general, the Solver returned a solution with a lower SSE, but the fits were qualitatively less acceptable. Figure 4.40 shows a typical result.

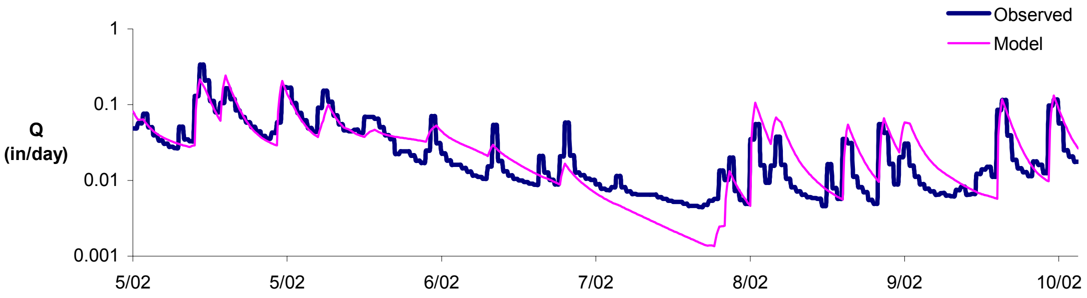


Figure 4.40 Hydrologic model output with optimal parameters from Solver

The reasons for stating that this result is "worse" than the manual calibration stem from an understanding of the modeling environment and the input data. It is known that there are inaccuracies present in the precipitation data, the forcing function for the model. Rainfall was measured at a single gage in the watershed. Especially in the summer months, convective thunderstorms cause highly localized rainfall events. The modeler, fully aware of the limitations of the input data, attempts to visually put the predicted line through as many observations as possible, accepting a few serious errors. For instance, the model may miss a blip in the hydrograph, simply because the rain gage did not 'see' that particular storm. The modeler accepts this error as unavoidable and moves on. The Solver gives every observation equal weight; in essence, it adjusts coefficients such as the infiltration rate to account for inaccurate input data.

4.3.2 Model Confirmation

A simple split-year model confirmation was performed. The model performance was evaluated by applying the model, with the same set of parameters, for data collected during the summer of 2003. The same assumption was made that the bacteria concentration measured at 10:00 a.m. represents the daily average. This assumption is more likely to have been violated during the summer 2003, due to the frequency of afternoon thunderstorms, which would cause higher bacteria levels in the afternoon and evening. This would mean that morning measurements are probably too low to be a reliable estimate of the daily average. However, these data were the best available, hence they were used in the confirmation model run. Results are shown in Figure 4.41 and Figure 4.42. Note that the hydrologic fit is not nearly as good as for the calibration year.

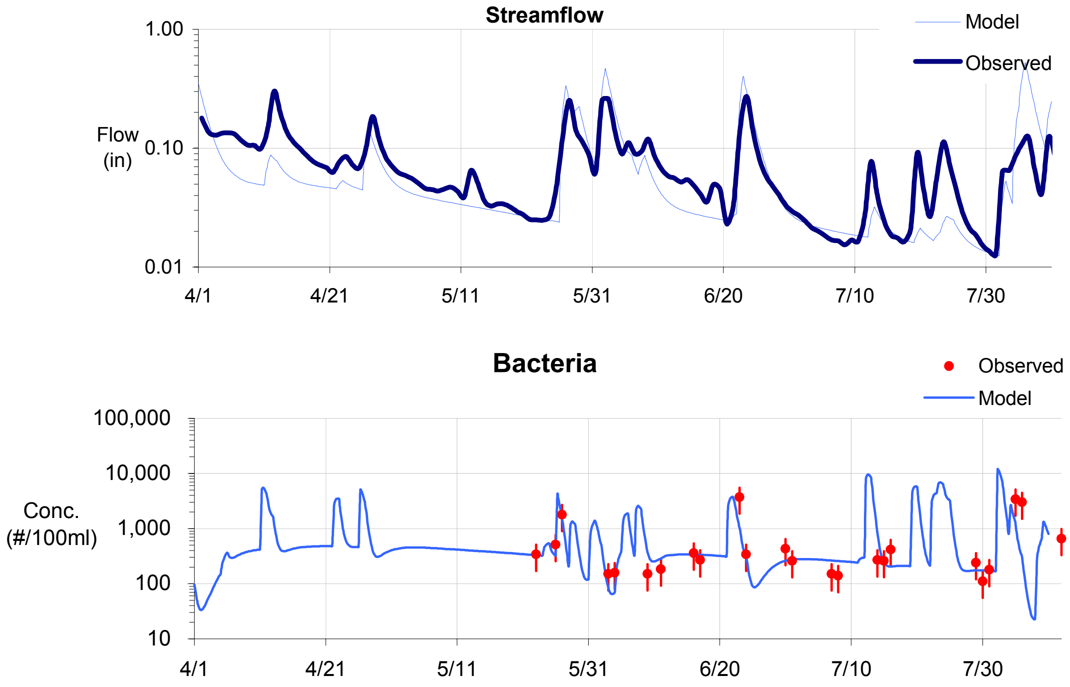


Figure 4.41 Model predictions for streamflow and Enterococcus in the Aberjona River for the confirmation year, 2003

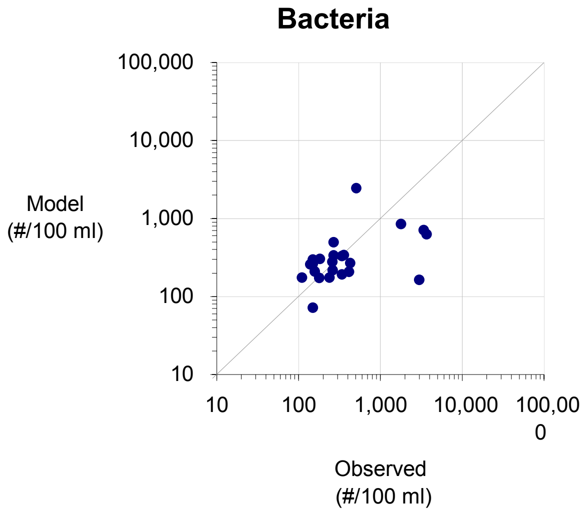


Figure 4.42 Model versus observed Enterococcus in the Aberjona River for the confirmation year, 2003

It was found that when the more complex options for bacteria buildup and washoff were used, with parameters derived from the 2002 calibration dataset, a slightly better fit to the 2003 observations was obtained, as shown in Table 4.16.

Table 4.16 Results of Aberjona River Enterococcus model, summer 2003

Bacteria Loading Model	Number of Parameters	*E*	RMSE
Constant Concentration	1	–1.1	0.65
Linear Buildup	3	0	0.45
Exponential Buildup	4	0.10	0.42

4.3.3 Extending the Model to an Ungaged Site

The simulation model was next used to predict Enterococcus bacteria concentrations at Alewife Brook in Somerville, Massachusetts. The brook's watershed includes 8.9 square miles of residential, commercial, and industrial land. Bacteria loading to the Alewife Brook includes polluted runoff and, occasionally during heavy rainstorms, combined sewer overflows. Calibrating the hydrologic model involved some guesswork here; because there are no streamflow measurements, we do not know 'the truth.' (It was initially hypothesized that the limited a few river stage observations might be a useful surrogate for streamflow; however stage and flow are not well correlated due to the backwater caused by the Amelia Earhart Dam.) Further, because the model does not include a module for predicting CSO activations, it may not accurately simulate bacteria loading at this site.

Nevertheless, bacteria data were available for this site, so the model was applied to evaluate its performance at an ungaged site. The starting point for calibration was to use the parameter set for the Aberjona developed previously. The curve number was increased from 88 to 92 to account for a greater proportion of impermeable area. This was determined qualitatively from the GIS datalayers for the watershed; Alewife Brook's watershed has a larger percentage of developed land than the Aberjona River watershed. Further, on numerous trips to the Brook, it was observed to be highly 'flashy', with high flood peaks, and minimal baseflow.

The buildup rate parameter k_B and the background concentration, c_B, were increased to obtain a better fit to the higher Enterococcus concentrations observed in the brook. The parameter set reported in Table 4.17 was used in the simulation. The results shown in Figure 4.43 and Figure 4.44 were obtained for the Alewife Brook in summer 2002. For log Enterococcus concentration, the simulation yielded a coefficient of determination, $R^2 = 0.47$ and a Nash-Sutcliffe Efficiency, $E = 0.34$. The fit is not as strong as that obtained for the Aberjona River above. However, the model has captured the essential patterns in the observations, and is usually well within an order of magnitude of the observed concentration.

Table 4.17 Simulation model parameters for Alewife Brook 2002 calibration

Hydrologic parameters		Bacteria loading parameters	
Watershed Area (square miles)	9.0	W_P	7×10^7
latitude (radians)	0.74	c_B	2,000
CN_2	92	k_B	30
CV	1.0	α	0.1
PET method	Hamon	k_W	23
S_0	1.0	V_S	10,000
G_0	0.5	k_d	1.4
R_0	0.1		
B_0	60		
S_{max}	0.2		
k_P	0.05		
k_G	0.025		
k_R	0.6		
Withdrawal	0		

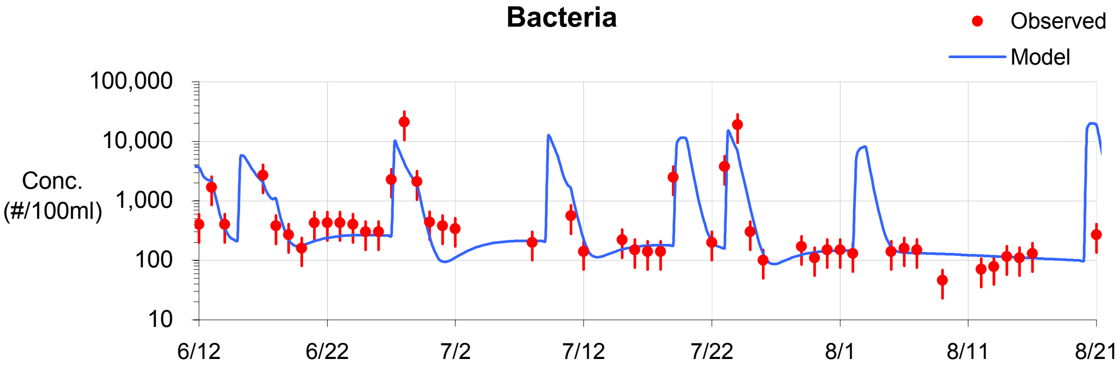


Figure 4.43 Time series of observed and modeled Enterococcus bacteria in Alewife Brook, summer 2002

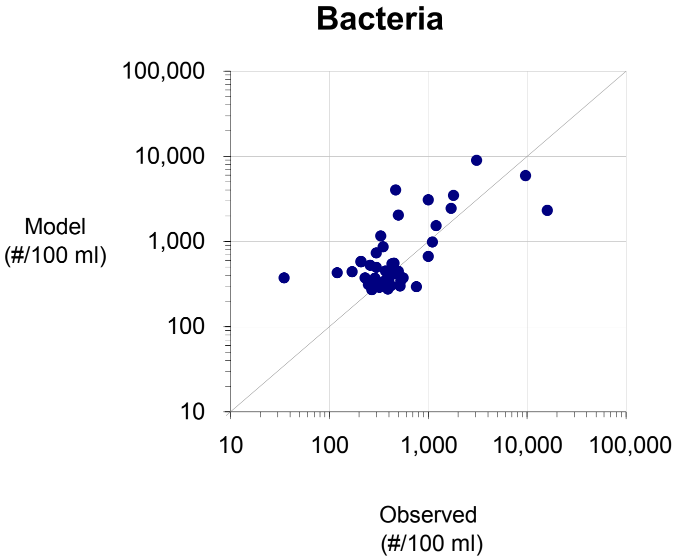


Figure 4.44 Observed versus simulated Enterococcus bacteria concentration in Alewife Brook, summer 2002

4.3.4 Sensitivity Analysis

The effect of changing model parameters on the output is briefly illustrated for the Aberjona River simulation presented above. For the buildup rate parameter, a calibrated value $k_B = 15\ \text{organisms}/\text{m}^2 \cdot \text{day}$ was used in the simulation. The effect of increasing or decreasing the parameter k_B on the modeled instream bacteria concentration is shown in Figure 4.45.

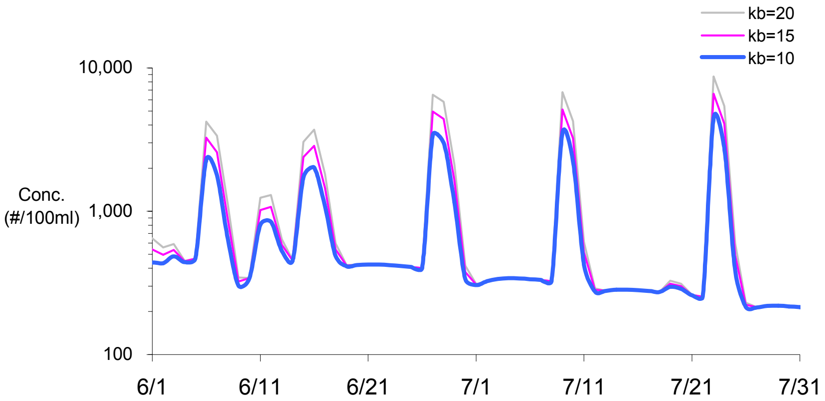


Figure 4.45 Modeled bacteria concentration sensitivity to parameter k_B

Increasing the model parameter k_B tends to increase the height of the peaks; doubling k_B tends to double the peak concentrations arising from runoff loads, as can be seen in Table 4.18. However, changes to k_B have very little effect on the median concentration. A short time after the runoff event, background conditions predominate, and k_B has little effect on the baseflow conditions. Thus the buildup rate k_B is found to be an important parameter in determining event loads, although it has minimal impact on loads during dry-weather flows.

Table 4.18 Effect of changing parameter k_B on modeled bacteria concentration

	Modeled bacteria concentration	
	Median	Maximum
$k_B = 10$	303	6,930
$k_B = 15$	320	10,400
$k_B = 20$	336	13,800

The bacteria mortality rate, k_d, has a significant effect on model output, as it affects both the upstream background source discussed in section 4.2.18, and the runoff source that passes through the routing reservoir. To perform a simple sensitivity analysis, the simulation model was run, decreasing the calibration value k_d from 1.4 to 0.8, and increasing k_d to 2.0. The results, in Figure 4.46, demonstrate that k_d is one of the most important model parameters, affecting: (1) the height of the peak concentration; (2) post-peak behavior of the pollutograph; and (3) the background concentration.

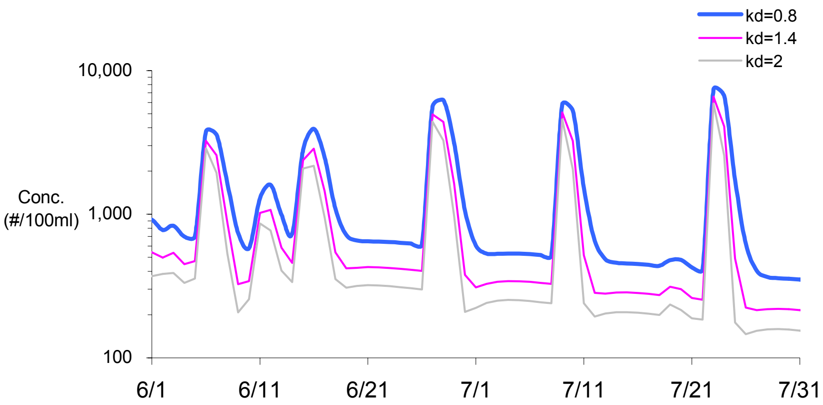


Figure 4.46 Modeled bacteria concentration sensitivity to parameter k_d

4.4 Discussion

Developing and applying the simulation model showed that a simple lumped-parameter model can yield useful estimates of bacteria concentration for rivers in the Mystic watershed. The simulation model is not a black box, and does require some understanding of hydrology and watershed modeling to calibrate and understand its output. Because some parameters are interdependent, there is probably no single 'optimal' parameter set. Efforts at using an optimization routine to calibrate the model proved unsatisfactory, and reinforced the idea that some expert knowledge and understanding of the watershed and the input data are essential to producing meaningful results.

The model performance in a split-year confirmation was not as strong as for the calibration year, although this may be attributed partly to error in bacteria measurements. When the model was extended to an ungaged site, Alewife Brook, some guesswork was involved in choosing appropriate hydrologic parameters. Nevertheless, the model produced order-of-magnitude estimates of Enterococcus bacteria concentration at this site.

The model represents a compromise between simple empirical methods and detailed simulation models and, as such, it should be considered a planning-level model. The model is straightforward to set up and run, uses readily available climate data, and gives a continuous simulation of streamflow and bacteria load. Beyond its use in predicting daily concentrations, the model may be useful in TMDL studies for modeling the bacteria inputs to receiving waters such as a lakes or estuaries. In the following section, the simulation model's performance is compared to that of a simple regression model, and its strengths and weaknesses further analyzed.

5. General Results and Discussion

Two modeling approaches were applied for modeling bacteria concentrations in the Mystic River watershed. Multivariate regression models predict bacteria concentrations based on climate variables such as precipitation. Next, a watershed simulation model was developed to predict daily average bacteria concentrations. The model simulates watershed processes, although a number of assumptions were made in order to create a parsimonious model.

5.1 Comparing the Models

Making direct comparisons between the two modeling techniques is complicated by the fact that the simulation model's output is a daily average, while the regression model makes predictions based on 15-minute data: 96 predictions each day. In order to make comparisons between the two modeling techniques, daily averages were calculated for the regression model output. A number of techniques for evaluating models were presented in section 2.5 on page 14. Here this framework is used to compare the two modeling techniques, using observed data from the Mystic River basin for the years 2002 and 2003.

Results indicate that the regression models outperform the simulation model in the Mystic River basin. Figure 5.1 shows the time series output from both of the modeling techniques. In Figure 5.2, scatterplots are shown for both models. Note that the regression model predictions are more tightly clustered about the 1:1 line, indicating a better fit. Further, the boxplots in figure Figure 5.3 indicate that the regression model is better at re-creating the median and the variance of the observed dataset.

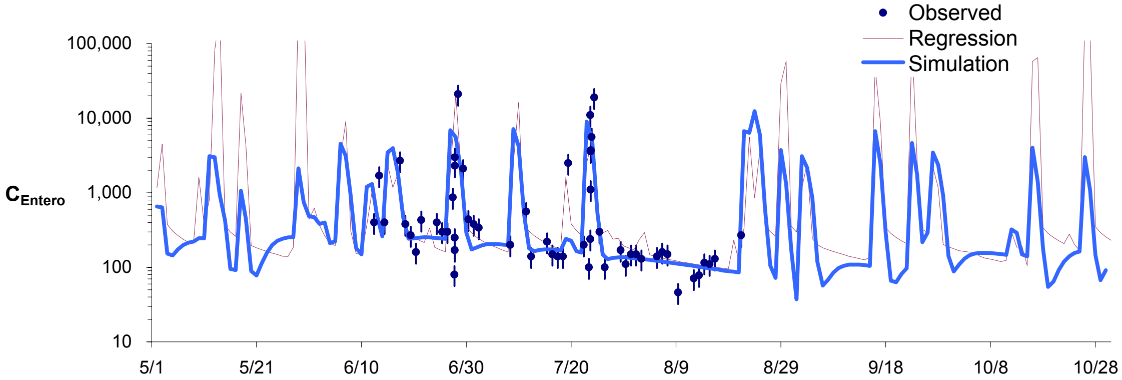


Figure 5.1 Enterococcus concentration for the Aberjona River, May – October 2002, predicted by regression and the simulation model

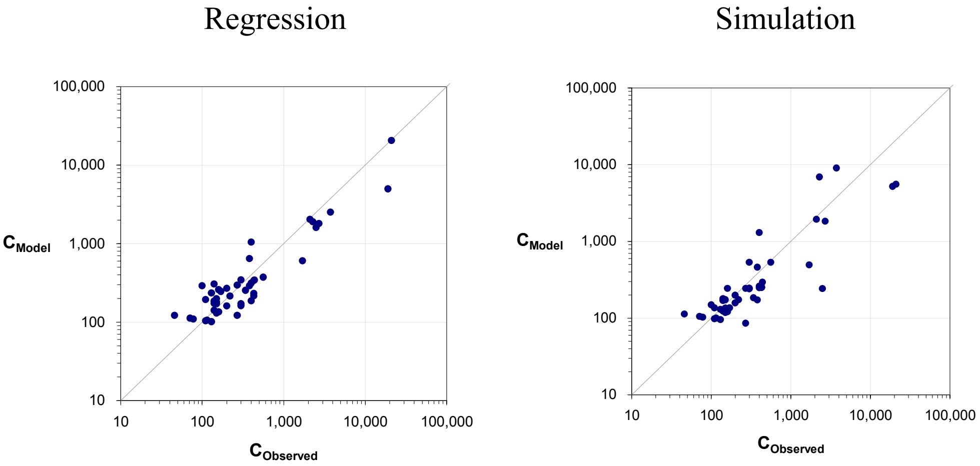


Figure 5.2 Model versus observations for Enteroccocus predicted by regression and simulation models, Aberjona River, May –October 2002

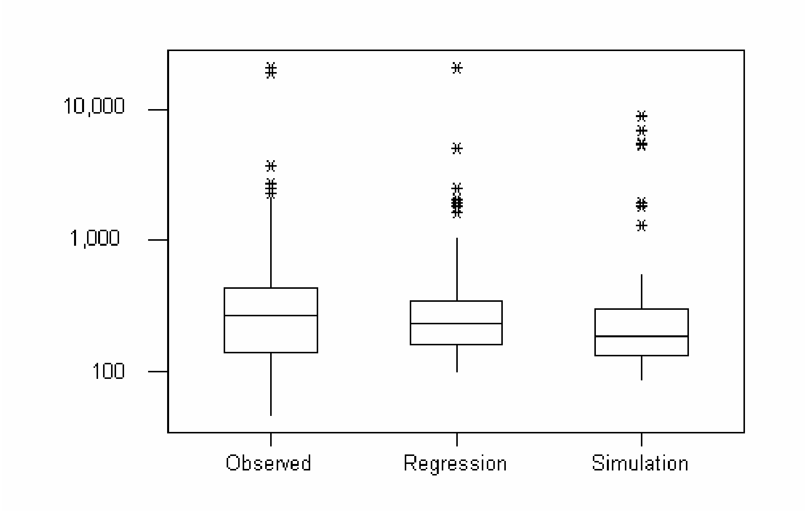


Figure 5.3 Boxplots comparing regression and simulation models for Aberjona River, calibration year 2002

Various quantitative criteria, reported in Table 5.1, demonstrate that the regression model yields a better fit to the confirmation year data. Note that the root mean square error (RMSE) was calculated using log Enterococcus concentrations, and thus the units are log(cfu/100 mL).

Table 5.1 Measures of fit of the regression and simulation models for Aberjona River Enterococcus bacteria concentration, summer 2002

Measure of Fit		Regression	Simulation
Coefficient of determination	R^2	0.85	0.77
Nash-Sutcliffe model efficiency	E	0.85	0.75
Root mean square error	$RMSE$	0.22	0.28

5.2 Model Robustness

Both the regression and the simulation models were run in a split-year confirmation to further evaluate how robust they are; i.e.: How well do models perform for a non-calibration year? Normally, when additional data is obtained, it could be added to previous data and used to amend the regression equations. However, it was desired to compare how the modeling techniques would work in a practical setting, where model calibration is necessarily limited by available time and resources.

Comparing output for the Aberjona River in summer 2003, it was found that the simulation model produces somewhat better predictions. Model outputs are compared in Figure 5.4. The boxplots in Figure 5.5 show that the regression model significantly over-predicts high concentration. This can be attributed to the problem of extrapolation, where the precipitation input to the regression model in 2003 is higher than any that occurred during the model calibration period.

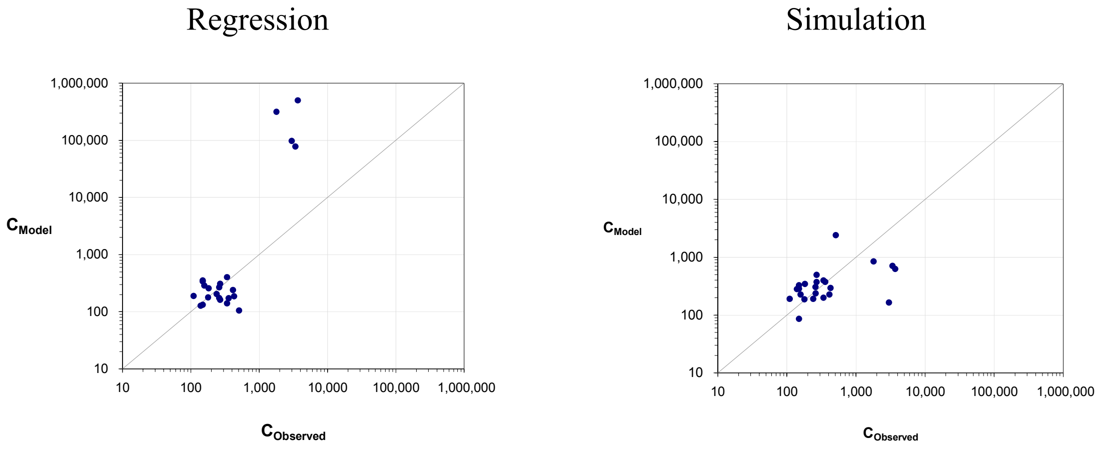


Figure 5.4 Comparison of the regression and the simulation models; model versus observations for Enterococcus bacteria in the Aberjona River, 5/1/02–10/31/02

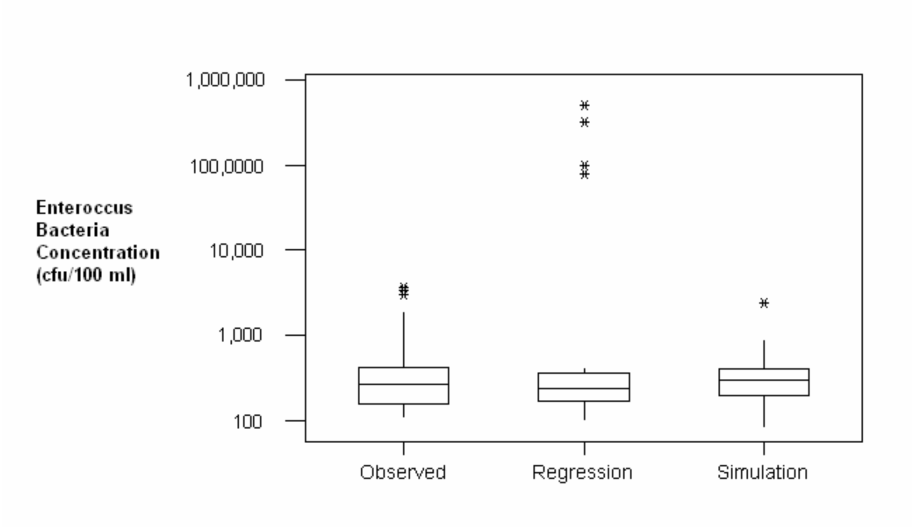


Figure 5.5 Boxplots comparing models for the Aberjona River, confirmation year 2003

When the regression equation developed for the 2002 data is used to predict bacteria concentrations in 2003, the model yields a very good coefficient of determination, $R^2 = 0.80$. Despite the strong correlation, the model is highly biased, significantly over-predicting high concentrations. By other measures, the simulation model yields a better fit to the confirmation year Enterococcus data, as reported in Table 5.2.

Table 5.2 Evaluation of regression and simulation models fit to Aberjona Enterococcus data for the confirmation year, 2003

Measure of Fit		Regression	Simulation
Coefficient of Determination	R^2	0.80	0.19
Nash-Sutcliffe Model Efficiency	E	-2.27	0.14
Root Mean Square Error	RMSE	0.81	0.42

The simulation model is slightly better at reproducing the mean and variance of the observations, as reported in Table 5.3. The cumulative frequency distributions in Figure 5.6 illustrate that the regression model does a far better at duplicating the frequency distribution at low concentrations, but produces very large errors at high concentrations.

Table 5.3 Summary statistics for the modeled Enterococcus concentration in the Aberjona River for the confirmation year, 2003

	Observed	Regression	Simulation
Minimum	110	105	86
Median	270	239	290
Geometric Mean	360	680	320
Maximum	3,700	500,000	2,400
Interquartile Range	250	170	170

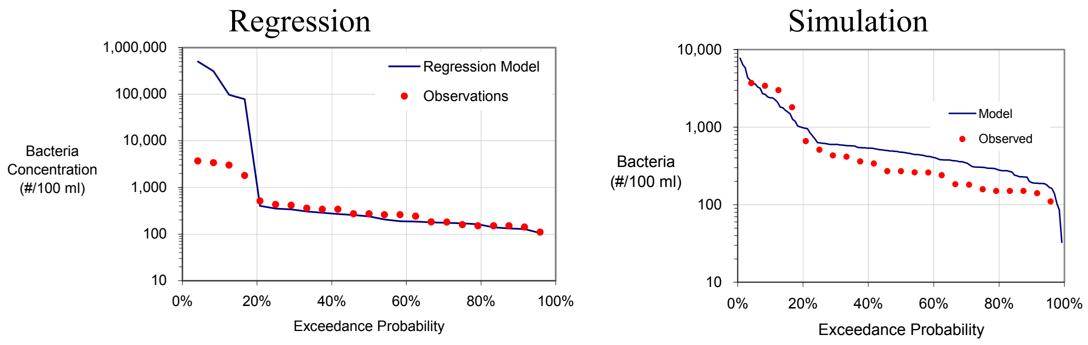


Figure 5.6 Concentration duration curves for observed and modeled Enterococcus concentration in the Aberjona River for the confirmation year, 2003

The problem of extrapolation is encountered with the regression model. The model should only be considered valid for the range of conditions observed during the calibration period. The largest rainstorm that occurred during the model calibration period during the summer of 2002 was 0.42 inches. Thus, it would be unrealistic to expect the model to accurately predict bacteria concentrations resulting from a 2-inch rainstorm. Evidence was presented, however, that the simulation model will still perform reasonably well, despite being calibrated with a limited data set.

6. General Summary and Conclusions

Two modeling techniques were developed for predicting pathogen indicator bacteria in the Mystic River watershed. Good data is essential for calibrating either of the models; to be representative, samples should cover the full range of climatic and flow conditions in the basin. A large investment of time, money, and skill is required to collecting this type of data. Further, modeling efforts are complicated by the spatial, temporal, and laboratory variability typical of bacteria data. The use of newer, more specific indicators, such as Enterococci appear to have substantially reduced these errors and may be more amenable to modeling.

In was shown that statistical regression models are relatively straightforward to build but rely on high-quality bacteria observations. Nevertheless, a thorough exploration of the input data, and appropriate manipulation of inputs was shown to increase their explanatory power. In applying regression models, care should be exercised to note the range of conditions for which the model is valid, as extrapolating the model outside this range may produce very unrealistic predictions.

It was found that water quality measurements (e.g., water depth, temperature, pH, etc.) are of limited use as predictors of bacteria in the Mystic watershed, based on data available for two sampling locations. However, a useful framework was developed for relating the time rate of change of water quality parameters to bacteria concentration. It is likely that this approach may be successfully incorporated into concentration-discharge models; further work should be done to verify this approach with larger data sets and, perhaps, pollutants other than bacteria.

A watershed simulation model was developed which outputs a continuous simulation of streamflow and bacteria concentration. The model, based loosely on GWLF, was developed as a middle ground between empirical statistical techniques and complex simulation models. Using the model requires more than just inputting some data and pressing a few buttons; it requires some modeling skill and an understanding of watershed hydrology and statistics.

The simulation model compared favorably to multivariate regression models, and outperformed them in a split-year confirmation, based on measures of fit such as the Nash-Sutcliffe model efficiency. Furthermore, the simulation model may also be useful in TMDL studies, as it gives a continuous simulation of streamflow and concentration, and there is evidence that it accurately reproduces the frequency distribution of observations. It would be highly useful to test the simulation model in other watersheds of various sizes and land use compositions.

While nonpoint source loading of bacteria is imperfectly understood, it has been shown that it is amenable to modeling. Accepted modeling techniques can successfully be applied to the problem, as long as the engineer is willing to accept the relatively large uncertainties inherent to working with bacteria.

Bibliography

American Public Health Association (APHA). (1998). *Standard methods for the examination of water and wastewater, 20th Edition.* APHA-AWWA-WEF, Washington, DC.

Aqua Terra Consultants. (1995). *HSPF version* 12 *user's manual*. Mountain View, California.

Barbé, D.E., Cruise, J.F., and Mo, X. (1996). "Modeling the buildup and washoff of pollutants on urban watersheds." *Water Resources Bulletin*, 32(3), 511–518.

Beven, K.J. (2001). *Rainfall-Runoff Modelling: The Primer.* John Wiley and Sons, Chichester, United Kingdom.

Bowie, G. L. et al. (1985). *Rates, constants, and kinetics formulations in surface water quality modeling*. EPN600/3-85/040, U.S. Environmental Protection Agency, Athens, Georgia.

Bras, R. L. (1990). *Hydrology: an introduction to hydrologic science*. Addison-Wesley, New York.

Chapra, S. C. (1997). *Surface water quality modeling*. McGraw-Hill, New York.

Chapra, S. C., and Canale, R. P. (2002). *Numerical methods for engineers*. McGraw-Hill, New York.

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*. WRIR 01–4248, US Geological Survey, Lawrence, Kansas.

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 Science and Technology*, 45(9), 205–219.

Clark, M. L. and Norris, J. R. (2000). *Occurrence of fecal coliform bacteria in selected streams in Wyoming, 1990-99*. WRIR 00-4198, U.S. Geological Survey, Cheyenne, Wyoming.

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

Commonwealth of Massachusetts. (1997b). Code of Massachusetts Regulations, 314 CMR 4.00: *Massachusetts Surface Water Quality Standards*.

Cronshey, R., et al. (1986). *Urban Hydrology for Small Watersheds.* TR-55, Natural Resources Conservation Service, Washington, DC.

Donigian, A.S., Huber, W. C., and Barnwell, T. O. Jr. (1996). "Models of nonpoint source water quality for watershed assessment and management." *Proceedings of Watershed 96*, Environmental Protection Agency, Washington, DC.

Drew, G. D. (1971). *The effect of bathers on the fecal coliform, total coliform and total bacteria density of water*. MS Thesis, Tufts University, Medford, Massachusetts.

Eagleson, P. S (1970). *Dynamic Hydrology*. McGraw-Hill, New York.

Eleria, A. (2002). *Forecasting fecal coliform bacteria in the Charles River basin*. Tufts University, Medford, Massachusetts.

Evolver, The genetic algorithm solver for Microsoft Excel, Windows version, release 4.0. (2001). Palisade Corporation, Newfield, New York.

Fogarty, L. R. et al. (2003). "Abundance and characteristics of the recreational water indicator bacteria *Escherichia coli* and enterococci in gull faeces." *Journal of Applied Microbiology*, 94, 865-878.

Francy, D. S., Gifford, A. M., and Darner, R. A. (2003). *Escherichia coli at Ohio bathing beaches—distribution, sources, wastewater indicators, and predictive modeling*. WRIR 02-4285, U.S. Geological Survey, Columbus, Ohio.

Haith, D. A., and Tubbs, L. J. (1981). "Watershed loading functions for nonpoint sources." *Journal of the Environmental Engineering Section, Proceedings of the American Society of Civil Engineers*, 107(EE1), 121–137.

Haith, D. A., and Shoemaker, L. L. (1987). "Generalized watershed loading functions for stream flow nutrients." *Water Resources Bulletin*, 23(3), 471–478.

Haith, D. A., Mandel, R., and Wu, R. S. (1996a). *Generalized watershed loading functions, version 2.0 Visual Basic computer program*. Cornell University, Ithaca, New York.

Haith, D. A., Mandel, R., and Wu, R. S. (1996b). *Generalized watershed loading functions, version 2.0 user's manual*. Cornell University, Ithaca, New York.

Helsel, D. R. and Hirsch, R.M. (1991). *Statistical Methods in Water Resources*. U.S. Geological Survey, Reston, Virginia.

Horsley & Witten, Inc. (1999). *Tools for watershed protection*. Sandwich, Massachusetts.

Huber, W. C. and Dickinson, R. E. (1992). *Storm water management model, version 4: user's manual*. U.S. Environmental Protection Agency, Athens, Georgia.

Hydroscience, Inc. (1979). *A statistical method for assessment of urban stormwater*. EPA-440/3-79-023. Environmental Protection Agency (EPA), Washington, DC.

Interstate Commission on the Potomac River Basin. (2002). *Bacteria TMDLs for the Goose Creek Watershed (DRAFT)*. http://www.deq.state.va.us/tmdl/pptpdf/gooseicprb1.pdf

Maidment, D. R. ed. (1993). *Handbook of Hydrology*. McGraw-Hill, New York.

Massachusetts Geographic Information System (MassGIS). (2002). *Geographic data on land use, hydrography, topography, etc.* < http://www.state.ma.us/mgis>

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 Research*, 37, 2700–2708.

Massachusetts Department of Environmental Protection (DEP). (1999). *Final Massachusetts Section 303(d) List of Waters, 1998.* Worcester, Massachusetts.

Mystic River Watershed Association (MyRWA). (2002). "About the Mystic River Watershed" http://www.mysticriver.org/about_watershed/index.html

Nash, J. E. and Sutcliffe, J.V. (1970). "River flow forecasting through conceptual models: Part 1. A discussion of principles*."* *Journal of Hydrology*, 10, 282–290.

National Climatic Data Center (NCDC), (2003). *Temperature and precipitation data measured at Boston Logan Airport climate station, station ID 20009288.* <http://www.ncdc.noaa.gov>

Oriel, K. (2003). *Predictive models for indicator bacteria in a sewage-impacted urban watershed.* MS Thesis, Tufts University, Medford, Massachusetts.

Pelletier, G. and Seiders, K. (2000). *Grays Harbor Fecal Coliform Total Maximum Daily Load Study*. Washington State Department of Ecology, Olympia, Washington.

Pelletier, G., and Seiders K. (2000). *Gray's Harbor Fecal Coliform Total Maximum Daily Load Study*. Washington State Department of Ecology, Olympia, Washington.

Rudolph, B. (2002). *Incorporating hysteresis into concentration-discharge models*. MS Thesis, Tufts University, Medford, Massachusetts.

Sartor, J. D., and Boyd, G. B. (1972). *Water pollution aspects of street surface contaminants*. EPA-R2/72-081. U.S. Environmental Protection Agency, Washington DC.

Singh, V. P. (1989). *Hydrologic systems, volume II: watershed modeling*. Prentice Hall, New Jersey.

Tasker, G. D. and Driver, N. E. (1998). "Nationwide regression models for predicting urban runoff water quality at unmonitored sites." Water Resources Bulletin, 24(5), 1091–1101.

Thomann, R. V. (1982). "Verification of water quality models." Journal of the Environmental Engineering Division, ASCE. 108, 923-940.

U.S. Geological Survey. (2003). "Real-time data for USGS 01102500 Aberjona River at Winchester, MA." <http://waterdata.usgs.gov/ma/nwis> USGS, Reston, Virginia.

Wagner, R. J. et al. (2000). *Guidelines and standard procedures for continuous water-quality monitors: site selection, field operation, calibration, record computation, and reporting*, 2000. WRIR 00–4252, U.S. Geological Survey, Reston, Virginia.

Walkenbach, J. (1999). *Microsoft Excel 2000 power programming with VBA*. Hungry Minds Inc. New York.

Whipple W. et al. (1983). Stormwater management in urbanizing areas. Prentice-Hall, Englewood Cliffs, New Jersey.

Woodward, D. E. et al. (2003). "Runoff number method: examination of the initial abstraction ratio." *Proceedings of World water & environmental resources congress 2003*. American Society of Civil Engineers, Philadelphia, Pennsylvania.

York Watershed Council. (2003). "Virginia's Use of Bacteria Source Tracking to Develop 'Cheaper, Better, Faster' Bacteria TMDLs." <http://www.yorkwatershed.org/yis/bst_whitepaper9_13.doc>

YSI Inc. (2001). *Water Quality Operations Manual*. Yellow Springs, Ohio.

Appendix I Notation

The following symbols are used in this thesis:

a_0	=	intercept of regression equation
AMC	=	antecedent moisture condition
a_W	=	bacteria washoff parameter
B	=	bacteria buildup
B_0	=	initial bacteria buildup
$b_1, b_2, \ldots, b_n$	=	slopes in regression equation
B_{max}	=	maximum bacteria buildup
b_W	=	washoff parameter
C	=	bacteria concentration
C_0	=	initial bacteria concentration in stream tank
c_B	=	background bacteria concentration
CN	=	curve number
c_R	=	routing reservoir bacteria concentration
c_S	=	stream tank bacteria concentration
CV	=	cover coefficient
DO	=	dissolved oxygen
E	=	Nash-Sutcliffe model efficiency
e_i	=	model residual for the i^{th} prediction
ET	=	evapotranspiration
F_0^2	=	initial variance
F^2	=	index of disagreement
G	=	groundwater depth
H	=	water depth
I_a	=	initial abstraction in SCS Curve Number method runoff calculations
J	=	Julian Day
k_B	=	bacteria buildup rate
k_B	=	buildup rate
k_d	=	bacteria decay rate
k_G	=	groundwater recession coefficient
k_P	=	percolation rate coefficient
k_R	=	reservoir coefficient

k_{W}	=	washoff rate
m_{R}	=	number of bacteria in routing reservoir
m_{S}	=	number of bacteria in stream tank
n	=	sample size
P	=	precipitation
P_{24}, P_{48}	=	precipitation depth in a 24, 48 hour period
PET	=	potential evapotranspiration
Q	=	streamflow
Q_{B}	=	baseflow
Q_{R}	=	routed runoff flow
$\bar{Q}$	=	average flow
$\hat{Q}$	=	predicted or modeled flow
R	=	routing reservoir depth
R_0	=	initial routing reservoir depth
R^2	=	coefficient of determination
S	=	soil moisture depth
S_{i}	=	storage index in the SCS Curve Number method runoff calculations
S_0	=	initial soil moisture depth
SC	=	specific conductance
S_e	=	standard error of model predictions
$S_{\max}$	=	maximum soil moisture depth
t	=	time
T	=	temperature
t_{90}, t_{95}	=	time at which bacteria buildup reaches 90% or 95% of saturation
t_{c}	=	calculation time step
T_{F}	=	time since last rainfall
V_{R}	=	routing reservoir volume
V_{S}	=	stream tank volume
W	=	washoff of bacteria from land surfaces
W_{P}	=	point source bacteria loading rate
W_{R}	=	runoff bacteria loading rate
α	=	bacteria buildup inhibition factor
ε	=	stochastic (random error) component of a predictive model
τ	=	residence time

Appendix II Abbreviations

EMPACT	Environmental Monitoring for Public Access and Community Tracking (EPA grant program under which thesis research was funded)
cfu	colony forming unit (bacteria organism)
cfs	cubic feet per second
CSO	combined sewer overflow
EPA	U.S. Environmental Protection Agency
GWLF	Generalized Watershed Loading Functions
MDC	Metropolitan District Commission (Boston-area parks agency)
HSPF	Hydrologic Simulation Program Fortran
Lowess	LOcally WEighted Scatterplot Smoothing
MWRA	Massachusetts Water Resources Authority
MyRWA	Mystic River Watershed Association
NRCS	Natural Resources Conservation Service
PRESS	Prediction error sum of squares
HSPF	Hydrologic Simulation Program Fortran
SCS	Soil Conservation Service
RMSE	Root mean square error
SSE	Sum of squared errors
SWMM	Stormwater Management Model
TMDL	Total Maximum Daily Load

Appendix III   Simulation Model Excel/VBA Code

```
Option Explicit
Option Base 0

' Define Global Variables
' Note that I have the model set up to accept 365 input values, although this can be
easily expanded

' Input Time Series
Public n As Integer 'n is the number of days of input data
Public dDate(365) As Date  'dDate is the date (in Excel Date/Time code)
Public Precip(365) As Double  'Daily Precipitation in meters
Public Tmin(365) As Double   'Temperatures
Public Tmax(365) As Double
Public Tavg(365) As Double
Public AntMoist(365) As Double  '5-day antecedent moisture (for the TR55 runoff
calculations)

' Calculated Daily Climate Variables
Public Trange(365) As Double  'Tmax - Tmin on Day i
Public PETdaily(365) As Double 'Potential Evapotranspiration in m on Day i

' Geographic Parameters
Public Area As Double  'Watershed Area in square miles
Public lat As Double   'latitude of the watershed centroid (in radians)
Public CN1 As Double, CN2 As Double, CN3 As Double  'Curve Numbers for the land surfaces

' State Variable Initial Values at t=0
Public G0 As Double 'Initial Groundwater level (in)
Public S0 As Double   'Initial Soil Moisture level (in)
Public R0 As Double  'Initial Reservoir Volume (in)
Public B0 As Double  'Initial Bacteria buildup

' Parameters
Public Smax As Double 'Maximum Soil Moisture capacity (in)
Public CV As Double  'Cover coefficient (affects ET as f(PET, Soil Moisture)
Public Withdrawal As Double 'Daily Groundwater Withdrawal (in/day)
Public Wp As Double  'Point Source = Constant Bacteria Load = Illicit Connections!
Public cp As Double    'Concentration due to Point source
Public cB As Double   'Background bact. concentration
Public cc As Double    'cc is the coeff. for the rational method, you know Q=cIA

' Mixing Reservoir Parameters
Public C0 As Double  'initial conc
Public Vs As Double  ' "Instream Tank" volume
Public kd As Double  'decay rate

' Rates
Public kI As Double    ' infiltration rate
Public kP As Double   ' percolation rate
Public kG As Double   ' groundwater recession constant
Public kR As Double   ' reservoir outflow rate constant
Public kB As Double   ' buildup rate
Public alpha As Double  ' buildup inhibition factor
Public kW As Double  ' washoff rate

' Calculation parameters
Public tp As Double 'Print step for use in solving set of diff equations
Public tc As Double 'Calculation step

' Output (Daily)
Public Pout(365) As Double
Public ETout(365) As Double
Public Qout(365) As Double
Public Cout(365) As Double
Public Loadout(365) As Double
Public Runoffout(365) As Double
Public np As Integer   'number of rows of output to print
Public deltaG As Double  'Change in Groundwater Storage over modeled time period
Public deltaS As Double  'Change in Soil Moisture Storage over modeled time period
Public deltaR As Double  'Change in Reservoir Volume
Public Crunoff As Double  'To model a *constant* bacteria conc. in polluted runoff
Public Buildup_method As Integer
Public Washoff_method As Integer
Public PET_method As Integer
Public aw As Double
Public bw As Double

'Bacteria Observations
Public bDate(365) As Date    'Date time stamp for bacteria sample
Public Cobs(365) As Double  'Observed bacteria concentration
Public nb As Integer  'number of bacteria observations
```

```
Sub ReadParameters()
'Reads parameters from worksheet "Parameters"
Workbooks("NewModel.xls").Sheets("Parameters").Select

'Read watershed area from sheet and convert from sq. miles to sq. meters
Area = Range("Area").Value * 2589988
lat = Range("lat").Value

'Calculate Curve numbers 1 & 3 from CN2
CN2 = Range("CN_2").Value
CN1 = CN2 / (2.334 - 0.01334 * CN2)
CN3 = CN2 / (0.4036 + 0.0059 * CN2)

Smax = Range("Smax").Value / 39.37
S0 = Range("So").Value / 39.37
G0 = Range("Go").Value / 39.37
R0 = Range("Ro").Value / 39.37
B0 = Range("Bo").Value
CV = Range("CV").Value  'cover coefficient
cc = Range("cc").Value    'rational method c
tp = Range("tp").Value
tc = Range("tc").Value
kP = Range("kP").Value
kG = Range("kG").Value
kR = Range("kR").Value
Withdrawal = Range("Withdrawal").Value / 39.37
Wp = Range("Wp").Value  'point source
cB = Range("cB").Value * 10000# 'background bacteria conc.
kB = Range("kB").Value * Area / 10000 'buildup rate over watershed
alpha = Range("alpha").Value  'buildup inhibition factor
kW = Range("kW").Value  'washoff parameter
C0 = Range("C0").Value * 10000#   'initial value for concentration in reservoir
Vs = Range("V").Value  'Instream Tank volume
kd = Range("kd").Value  'bacteria decay rate in the reservoir
Crunoff = Range("Crunoff").Value * 10000#  'const bact conc in runoff
aw = Range("aw").Value
bw = Range("bw").Value

Call MethodChooser

End Sub
```

```
Sub WatershedModel()
' Main Program
Application.Calculation = xlManual
Application.ScreenUpdating = False

' Local Variable Declarations
Dim PercentDone As Double
Dim i As Integer
Dim T As Double   'modeled time
Dim tf As Double  'end time
Dim dt As Double  'time step
Dim td(365) As Double 'input time (to avoid adding and subtracting dates in Excel
time/date code)
Dim P As Double  'Precipitation
Dim AMC5 As Double '5-day antecedent moisture condition (for the TR55 runoff calculation)
Dim PET As Double  'daily potential evapotranspiration in m
Dim ET As Double
Dim Q As Double

' State Variables
Dim S As Double
Dim G As Double
Dim R As Double
Dim B As Double
'Bacteria Load and Conc
Dim Load As Double  '  Bacteria Load (#/day)
Dim Cr As Double  '  Bacteria Conc in the Reservoir compartment
Dim VR As Double  'VR is the volume of the reservoir (VR=R*Area)
Dim Mr As Double   'm is the number of bacteria in the reservoir
Dim Ms As Double
Dim C As Double  '  Bacteria Conc in Stream (#/100 ml)
Dim Cs As Double
Dim cp As Double ' Bacteria conc in stream due to the point source

' Slopes = Rates of Change
Dim dSdt As Double
Dim dGdt As Double
Dim dRdt As Double
Dim dBdt As Double
Dim dMdt As Double
Dim dMsdt As Double

' Functions
Dim Runoff As Double
Dim Qr As Double
Dim Infiltration As Double
Dim Percolation As Double
Dim Recession As Double
Dim Buildup As Double
Dim Wr As Double

' Sums of ET over the calculated time step
Dim ETsum As Double
Dim Qsum As Double
Dim Psum As Double
Dim Csum As Double
Dim LoadSum As Double
Dim RunoffSum As Double

' START
Call ReadParameters
Call ReadClimateData
Call ReadBactData
Call PET_calc

Worksheets("Output").Visible = True
Sheets("Output").Select
Range("A6:z65536").ClearContents
Range("begin_out").Select

'change the calculation time step entered by user to 2^n
dt = 2 ^ (Int(Log(tc) / Log(2)))

'Set initial conditions for state variables
S = S0
G = G0
R = R0
B = B0
Cr = C0
Cs = C0
VR = R * Area
Mr = VR * Cr
```

```
For i = 0 To n + 1
  td(i) = i
Next i
tf = n  'final time

'Calculate water balance for each day
'using Euler's method to solve equations
T = 0
np = 0

Do
  'Do the calculations
  Call Selector(td(), Precip(), n, T, P)
  Call Selector(td(), PETdaily(), n, T, PET)
  Call Selector(td(), AntMoist(), n, T, AMC5)

  'Calculate rates of water movement (in/day)
  Runoff = Runoff_calc(P, S, AMC5)
  Qr = Route_calc(R)
  Infiltration = P - Runoff
  ET = ET_calc(PET, S)
  Percolation = Percolation_calc(S)
  Recession = Recession_calc(G)
  Buildup = Buildup_calc(B)
  Wr = Washoff_calc(B, Runoff)
  If Washoff_method <> 1 Then If Wr > B Then Wr = B

  'Calculate slopes (rates of change of state variables (m/day)
  dSdt = Infiltration - Percolation - ET
  dGdt = Percolation - Recession - Withdrawal

  If Buildup_method = 1 Then
    dBdt = 0
  Else
    dBdt = Buildup - Wr
  End If

  dRdt = Runoff - Qr

  'Overall Streamflow = Runoff + Baseflow

  Q = Qr + Recession

  'Background bacteria concentration calculations
  Wp = cB * Recession * Area  'constant background conc.
  Cs = Ms / Vs
  dMsdt = Wp - (Recession * Area * Cs) - kd * Ms

  'Calculate nonpoint source bacteria load
  ' Recall that all flows are expressed as depths, and so must
  'be scaled up by multiplying by the watershed area to
  'convert units from m to m3

  Load = Wr * Area
  VR = R * Area
  Cr = Mr / VR

  dMdt = Load - (Qr * Area * Cr) - kd * Mr

  'Conc in the stream
  C = (Qr * Cr * Area + Recession * Area * Cs) / (Q * Area)

  cp = (Recession * Area * Cs) / (Q * Area)
  'cp = Wp / (Q * Area)

  'Write to the sheet (during the program testing phase)
  'Report most state variables and rates in inches, and Conc. as #/m3
  ActiveCell.Offset(0, 0).Value = T + dDate(0)
  ActiveCell.Offset(0, 1).Value = S * 39.37
  ActiveCell.Offset(0, 2).Value = G * 39.37
  ActiveCell.Offset(0, 3).Value = R * 39.37
  ActiveCell.Offset(0, 4).Value = P * 39.37
  ActiveCell.Offset(0, 5).Value = PET * 39.37
  ActiveCell.Offset(0, 6).Value = ET * 39.37
  ActiveCell.Offset(0, 7).Value = Runoff * 39.37
  ActiveCell.Offset(0, 8).Value = Qr * 39.37
  ActiveCell.Offset(0, 9).Value = Infiltration * 39.37
  ActiveCell.Offset(0, 10).Value = Percolation * 39.37
  ActiveCell.Offset(0, 11).Value = Recession * 39.37
  ActiveCell.Offset(0, 12).Value = dSdt
  ActiveCell.Offset(0, 13).Value = dGdt
  ActiveCell.Offset(0, 14).Value = dRdt
  ActiveCell.Offset(0, 15).Value = Q * 39.37
  ActiveCell.Offset(0, 16).Value = B
  ActiveCell.Offset(0, 17).Value = Wr
  ActiveCell.Offset(0, 18).Value = Load
  ActiveCell.Offset(0, 19).Value = Cr / 10000
  ActiveCell.Offset(0, 20).Value = C / 10000
  ActiveCell.Offset(0, 21).Value = Runoff * Area
  ActiveCell.Offset(0, 22).Value = cp

  ActiveCell.Offset(1, 0).Select

  'Integrate with trapezoidal rule, as we want to know the *average* over the interval
  Psum = Psum + P * dt
  ETsum = ETsum + ET * dt
  Qsum = Qsum + Q * dt
  Csum = Csum + C * dt
  LoadSum = LoadSum + Load * dt
  RunoffSum = RunoffSum + Runoff * dt

  T = T + dt
  S = S + dSdt * dt
  G = G + dGdt * dt
  R = R + dRdt * dt
  B = B + dBdt * dt
  Mr = Mr + dMdt * dt
  Ms = Ms + dMsdt * dt

' Report the average daily value (put into an array for printing later)
  If Abs(Int(T) - T) < 0.00001 Then
    np = np + 1
    Pout(np) = Psum * 39.37
    Psum = 0
    ETout(np) = ETsum * 39.37
    ETsum = 0
    Qout(np) = Qsum * 39.37
    Qsum = 0
    Cout(np) = Csum / 10000#
    Csum = 0
    Loadout(np) = LoadSum
    LoadSum = 0
    Runoffout(np) = RunoffSum * Area
    RunoffSum = 0
  End If

  'Display the progress of the calculations in Excel' status bar
  PercentDone = Round((T / tf) * 100, 0)
  Application.StatusBar = "Calculating: " & PercentDone & "%"
  If T > tf Then Exit Do
Loop
Application.StatusBar = "Writing Output to Sheet"
deltaG = (G - G0) * 39.37
deltaS = (S - S0) * 39.37
deltaR = (R - R0) * 39.37

Worksheets("Output").Visible = False
Call WriteOutput
Call MakeFDC
Call MakeCDC
Application.StatusBar = "Updating Charts"
Application.Calculation = xlAutomatic
Application.StatusBar = False
Sheets("Diagnostics").Activate
Application.ScreenUpdating = True
End Sub
```

```
Sub ReadClimateData()
'Reads in the climate data from the Sheet "Input"
Dim i As Integer

Sheets("Climate").Select
Range("A4").Select
i = 0

Do While Not IsEmpty(ActiveCell)
  dDate(i) = ActiveCell.Offset(0, 0).Value
  Precip(i) = ActiveCell.Offset(0, 1).Value
  Precip(i) = Precip(i) / 39.36996 'Convert inches to m
  Tmax(i) = ActiveCell.Offset(0, 2).Value
  Tmin(i) = ActiveCell.Offset(0, 3).Value
  Tavg(i) = ActiveCell.Offset(0, 4).Value
  AntMoist(i) = ActiveCell.Offset(0, 5).Value  'units are in
  ActiveCell.Offset(1, 0).Select
  i = i + 1
Loop
n = i - 1

'Note that since the Temperature input by the user is in °F,
'it needs to be converted to °C
Call ConvertTemp

Range("A4").Select
End Sub
```

```
Sub ReadBactData()
'Get the erved bacteria data from the Sheet "Bacteria"
'so that it can be plotted in the right position alongside the modeled data
Dim i As Integer
i = 0
Sheets("Bacteria").Select
Range("A4").Select

Do While Not IsEmpty(ActiveCell)
  bDate(i) = ActiveCell.Value
  Cobs(i) = ActiveCell.Offset(0, 1).Value
  ActiveCell.Offset(1, 0).Select
  i = i + 1
Loop
nb = i - 1

End Sub
```

```
Sub ConvertTemp()
'Converts Temperature data from °F to °C
Dim i As Integer
For i = 0 To n
  Tmin(i) = (Tmin(i) - 32) * 5 / 9
  Tmax(i) = (Tmax(i) - 32) * 5 / 9
  Tavg(i) = (Tavg(i) - 32) * 5 / 9
  Trange(i) = Tmax(i) - Tmin(i)
Next i
End Sub
```

```
Sub MethodChooser()
Dim text As String
```

```
text = Range("PET_method").Value
Select Case text
  Case Is = "Hargreaves"
     PET_method = 1
  Case Is = "Hamon"
    PET_method = 2
End Select

text = Range("Buildup_method").Value
Select Case text
  Case Is = "constant"
     Buildup_method = 1
  Case Is = "linear"
    Buildup_method = 2
  Case Is = "exponential"
    Buildup_method = 3
End Select

text = Range("Washoff_method").Value
Select Case text
  Case Is = "constant concentration"
     Washoff_method = 1
  Case Is = "calculate load"
    Washoff_method = 2
  Case Is = "calculate fraction"
    Washoff_method = 3
End Select

End Sub
```

```
Function Runoff_calc(P, S, AMC5)
  Runoff_calc = TR55(P, AMC5) / 39.37
End Function
```

```
Function Route_calc(R)
  Route_calc = kR * R
End Function
```

```
Function Infiltration_calc(P, Runoff)
  Infiltration_calc = P - Runoff
End Function
```

```
Function Recession_calc(G)
 Recession_calc = kG * G
End Function
```

```
Function ET_calc(PET, S)
  If S >= Smax Then
    ET_calc = CV * PET
  Else
    ET_calc = CV * (S / Smax) * PET
  End If
End Function

Function Percolation_calc(S)
    Percolation_calc = kP * S
End Function

Function Buildup_calc(B)
  Select Case Buildup_method
    Case Is = 2
        Buildup_calc = kB
    Case Is = 3
        Buildup_calc = kB - alpha * B
    End Select
End Function
```

```
Function Washoff_calc(B, Runoff)
  Select Case Washoff_method
    Case Is = 1
        Washoff_calc = Crunoff * Runoff
    Case Is = 2
        Washoff_calc = aw * (Runoff * 39.97) ^ bw
    Case Is = 3
       Washoff_calc = (1 - Exp(-kW * Runoff * 39.37)) * B
    End Select
End Function
```

```
Sub PET_calc()
'This sub computes the Potential Evapotranspiration for the i_th day
'via either the Hamon or Hargreaves Method
'One of the global parameters that is required is the latitude in radians

'Variable declarations:
Const Pi = 3.1415926
Dim dr As Double  'Relative Distance, sun to earth
Dim j As Integer 'Julian Day
Dim k As Integer
Dim dec As Double  'Solar declination
Dim ws As Double  'Sunset hour angle
Dim So As Double  'Extraterrestrial Solar radiation
Dim Hrs As Double 'Number of possible daylight hours

For k = 0 To n
  'Calculate the Julian Date
  j = Julian(dDate(k))
  'Calculate the relative distance from the sun to the earth, dr
  dr = 1 + 0.033 * Cos(2 * Pi * j / 365)
  'Calculate the solar declination in radians
  dec = 0.4093 * Sin(2 * Pi * j / 365 - 1.405)
  'Calculate the sunset hour angle in radians
  ws = Acos(-Tan(lat) * Tan(dec))
  'Max daylight Hours
  Hrs = 24 * ws / Pi
  'Extraterrestrial solar radiation
  So = 15.329 * dr * (ws * Sin(lat) * Sin(dec) + Cos(lat) * Cos(dec) * Sin(ws))
  'Convert PET units from mm to m

  If PET_method = 1 Then  'PET_method:  (1) Hamon, (2) Hargreaves
    PETdaily(k) = Hamon(Tavg(k), Hrs) / 1000
  Else
    PETdaily(k) = Hargreaves(Tavg(k), Trange(k), So) / 1000
  End If

Next k

End Sub
```

```
Function Hargreaves(Tavg, Trange, So)
'Returns Potential Evapotranspiration in mm
  If Tavg < 0 Then
    Hargreaves = 0
  Else
    Hargreaves = 0.0023 * So * (Tavg + 17.8) * Sqr(Trange)
  End If
End Function
```

```
Function Hamon(Tavg, Hrs)
'Returns Potential Evapotranspiration in mm
Dim es As Double  'Saturation Vapor Pressure = f(Tavg)
es = 0.6108 * Exp(17.27 * Tavg / (237.3 + Tavg))
Hamon = 29.8 * Hrs * es / (Tavg + 273.2)
```

```
End Function
```

```
Function Julian(dDate)
'Input the Excel Time date code and get the Julian Date out
  Julian = dDate - DateSerial(year(dDate), 1, 0)
End Function
```

```
Function Acos(x)
'The MacLaurin Series approximation of the inverse cosine
Const Pi = 3.1415926
  Acos = 0.5 * Pi - x - 0.166666666666667 * x ^ 3 - 0.075 * x ^ 5 - 4.46428571428571E-02
* x ^ 7 - 3.03819444444444E-02 * x ^ 9
End Function
```

```
Function TR55(P, AMC5)
'Computes runoff using the SCS Curve Number Method
'described in their publication Technical Release 55
'Note that this was originally developed as an event model
'but has been incorporated into the GWLF model

Dim CNum As Double      'curve number for current day
Dim Retention As Double 'detention parameter
Dim Melt As Double
Dim Grow As Boolean  '1 during the growing season, 0 for non-growing season

'Want to maintain the flexibility of using these variables later on
'although for now, we are only modeling summer months

'Convert precip data from m to inches for this method to work properly (easier than
recoding)
P = P * 39.36996

Grow = False
Melt = 0

If CN2 > 0 Then

  If Melt <= 0 Then

    If Grow Then
      'growing season
      Select Case AMC5
        Case Is >= 2.1  'wettest
          CNum = CN3
        Case Is < 1.4   'driest
          CNum = CN1 + (CN2 - CN1) * AMC5 / 1.4
        Case Else        'average
         CNum = CN2 + (CN3 - CN3) * (AMC5 - 1.4) / 0.7
      End Select

    Else
      'dormant season
      Select Case AMC5
        Case Is >= 1.1
          CNum = CN3
        Case Is < 0.5
          CNum = CN1 + (CN2 - CN1) * AMC5 / 0.5
        Case Else
          CNum = CN2 + (CN3 - CN2) * (AMC5 - 0.5) / 0.6
    End Select
  End If

   Else 'Melt>0
     CNum = CN3
End If

Retention = 1000 / CNum - 10
```

```
If Retention < 0 Then Retention = 0
  If P >= 0.05 * Retention Then
    TR55 = (P - 0.05 * Retention) ^ 2 / (P + 0.95 * Retention)
  End If
End If
P = P / 39.36996
End Function
```

```
Function MyGeoMean(x)
'calculates the geometric mean of an array of numbers
'I wrote this because Excel's function seems to break
'down with large arrays and reports error messages.
Dim i As Long
Dim n As Long
Dim total As Long
Dim sum As Double
sum = 0
total = 0
n = Int(x.Count)
For i = 1 To n
  If x(i) <> 0 Then
    sum = sum + Log(x(i))
    total = total + 1
  End If
Next i
MyGeoMean = Exp(sum / total)
End Function
```

```
Function SigFigs(Num, figs)
'Returns a number to the specified number of significant figures
  Dim temp As Double
  If figs < 1 Then
    SigFigs = "Error"
  Else
  temp = Log10(Abs(Num))
  SigFigs = Application.WorksheetFunction.Round(Num, figs - 1 - Int(temp))
  End If
End Function

Static Function Log10(x)
'Calculates the base-10 logarithm of a number
'(VBA does not have this function built-in)
    Log10 = Log(x) / Log(10#)
End Function
```

```
Function SSE(x, y)
'Calculates the Sum of Squared Errors between the values in array X and array Y
'Arguments can be passed from either inside a VBA program or from the worksheet
Dim i As Integer
Dim n As Integer
Dim sum As Double
n = Int(x.Count)
For i = 1 To n
  If Not IsEmpty(x(i)) And Not IsEmpty(y(i)) Then
    sum = sum + (y(i) - x(i)) ^ 2
  End If
Next i
SSE = sum
End Function
```

```
Function RMSE(x, y)
'Calculates the Root Mean Square Error between arrays x and y
Dim SSE As Double
Dim i As Integer
Dim n As Integer
Dim nrev As Integer
Dim sum As Double
nrev = 0
n = Int(x.Count)
For i = 1 To n
  If Not IsEmpty(x(i)) And Not IsEmpty(y(i)) Then
    sum = sum + (y(i) - x(i)) ^ 2
    nrev = nrev + 1
  End If
Next i
SSE = sum
RMSE = Sqr(SSE / nrev)
End Function
```

```
Function NSE(x, xm)
' Arguments are arrays of type Double
' x = observed values
' xm = modeled values

'Calculates the Nash-Sutcliffe Model Efficiency, E-squared
'which is analagous to, but not equal to the coefficient of determination
'This is a useful criterion for evaluating the fit of water quality models

'Reference: Nash, J.E., and J.V. Sutcliffe (1970), "River flow forecasting through
conceptual models,
'Part I - A discussion of Principles".  Journal of Hydrology 10(1970), pages 282-290.

'Note that in the original paper, their measure of efficiency is R-squared, which is
confusing,
'as this is the same notation for the coefficient of determination, which is NOT the
same.
'I have taken the liberty of calling the efficiency "E-squared" to make it clear that it
is different.

Dim i As Long, n As Long, nrev As Long
Dim xbar As Double, sum As Double
Dim F As Double, F0 As Double

n = Int(x.Count)

'Calculate the average of observations, x-bar, for all coincident data
For i = 1 To n
  If Not IsEmpty(x(i)) And Not IsEmpty(xm(i)) Then
    sum = sum + x(i)
    nrev = nrev + 1
  End If
Next i
xbar = sum / nrev

'Calculate F0, initial variance and F, index of disagreement
 For i = 1 To n
  If Not IsEmpty(x(i)) And Not IsEmpty(xm(i)) Then
    F0 = F0 + (x(i) - xbar) ^ 2
    F = F + (x(i) - xm(i)) ^ 2
  End If
Next i

'Calculate the Model Efficiency, E-squared
NSE = (F0 - F) / F0

End Function
```

```
Sub Interp(x, y, n, x1, y1)
'Interp does simple linear interpolation for an array
'the OUTPUT of this subroutine is y1
'in this example, x is time, and y is either P(t) or Q(t)
'n is the number of elements in the array
'x1 is the value of t at which you interpolate
'If you ask this routine to interpolate within 0.00001 of a known value
'it will tell you interp out of range
```

```
Dim i As Integer
Dim tol As Single
i = 0
tol = 0.00001
If x1 < x(0) - tol Or x1 > x(n) + tol Then
  MsgBox "Interp outside range"
  End
Else
  Do
    If x1 <= x(i + 1) + tol Then
      y1 = y(i) + (y(i + 1) - y(i)) / (x(i + 1) - x(i)) * (x1 - x(i))
      Exit Do
    Else
      i = i + 1
    End If
  Loop
End If

End Sub
```

```
Sub Selector(x, y, n, x1, y1)
'Instead of doing a linear interpolation, this sub will just select a value
'based on the day.
'This avoids the problem of apportioning some rainfall to the day before!

Dim tol As Double
Dim i As Integer
tol = 0.000001
i = 0
If x1 < x(0) - tol Or x1 > x(n) + tol Then
  MsgBox "Selector outside range"
  End
Else
  Do
    If x1 < x(i + 1) - tol Then
      y1 = y(i)
      Exit Do
    Else
      i = i + 1
    End If
  Loop
End If

End Sub
```

```
Function TrapUn(x, y) As Double
'Uses the Trapezoidal Rule to Numerically Integrate unequally spaced data
Dim i As Integer, sum As Double

sum = 0
For i = 2 To Int(x.Count)
  sum = sum + (x(i) - x(i - 1)) * (y(i) + y(i - 1)) / 2
Next i
TrapUn = sum
End Function
```

```
Function Lastincolumn(rng As Range)
Dim WorkRange As Range
Dim i As Integer
Dim CellCount As Integer
'Application.Volatile
Set WorkRange = rng.Columns(1).EntireColumn
Set WorkRange = Intersect(WorkRange.Parent.UsedRange, WorkRange)
CellCount = WorkRange.Count
For i = CellCount To 1 Step -1
  If Not IsEmpty(WorkRange(i)) Then
    Lastincolumn = WorkRange(i).Value
    Exit Function
  End If
Next i
End Function
```

```
Function ArrayMax(arr)
'Finds the subscript of the last non-empty entry in an array
Dim i As Long
Dim max As Long
max = UBound(arr)
i = 1
Do
  If i = max Then Exit Do
  If arr(i) = "" Then Exit Do
  i = i + 1
Loop
ArrayMax = i - 1
End Function
```

```
Sub WriteOutput()
Dim i As Integer
Dim ib As Integer  'separate counter for the bacteria data
Dim D As Date
Sheets("Daily").Select
Range("A4:J65536").ClearContents

Range("A4").Select
ib = 0
For i = 1 To np
  D = i + dDate(0) - 1
  ActiveCell.Offset(0, 0) = D
  ActiveCell.Offset(0, 1).Value = Pout(i)
  ActiveCell.Offset(0, 2).Value = ETout(i)
  ActiveCell.Offset(0, 3).Value = Qout(i)
  ActiveCell.Offset(0, 4).Value = Cout(i)
  ActiveCell.Offset(0, 5).Value = Runoffout(i)
  ActiveCell.Offset(0, 6).Value = Loadout(i)
  If D = bDate(ib) Then
    ActiveCell.Offset(0, 7).Value = Cobs(ib)
    ActiveCell.Offset(0, 8).FormulaR1C1 = "=LOG(RC[-1])"
    ActiveCell.Offset(0, 9).FormulaR1C1 = "=LOG(RC[-5])"
    ActiveCell.Offset(0, 10).FormulaR1C1 = "=RC[-2]-RC[-1]"
    ib = ib + 1
  End If
  ActiveCell.Offset(1, 0).Select
Next i

Sheets("Diagnostics").Select
Range("dG").Select
ActiveCell.Value = deltaG
Range("dS").Select
ActiveCell.Value = deltaS
Range("dR").Select
ActiveCell.Value = deltaR

End Sub
```

```
Option Explicit

Private MyFileName As String
Private Cells(100) As String
Private i As Integer
Private n As Integer

Sub NameCells()
  Cells(1) = "Notes"
  Cells(2) = "tc"
  Cells(3) = "tp"
  Cells(4) = "Area"
  Cells(5) = "lat"
  Cells(6) = "CN_2"
  Cells(7) = "CV"
  Cells(8) = "cc"
  Cells(9) = "So"
  Cells(10) = "Go"
  Cells(11) = "Ro"
  Cells(12) = "Bo"
  Cells(13) = "Smax"
  Cells(14) = "kI"
  Cells(15) = "kP"
  Cells(16) = "kG"
  Cells(17) = "kR"
  Cells(18) = "Withdrawal"
  Cells(19) = "Wp"
  Cells(20) = "cB"
  Cells(21) = "kB"
  Cells(22) = "alpha"
  Cells(23) = "kW"
  Cells(24) = "CRunoff"
  Cells(25) = "aw"
  Cells(26) = "bw"
  Cells(27) = "C0"
  Cells(28) = "V"
  Cells(29) = "kd"
  Cells(30) = "PET_method"
  Cells(31) = "Buildup_method"
  Cells(32) = "Washoff_method"
End Sub
```

```
Sub SaveFile()
'Saves the set of parameters used in the model
'The names of the cells with data to save
Call NameCells
n = ArrayMax(Cells)
'Query the user for the file name and location
Application.ScreenUpdating = False
MyFileName = Application.GetSaveAsFilename( _
    InitialFileName:="FileName", _
    fileFilter:="My Files (*.mgh), *.mgh," & _
                "Text Files (*.txt), *.txt,")
If MyFileName = "False" Then
  End
End If
Open MyFileName For Output As #1
'Read in Name of Simulation
Sheets("Parameters").Select
'Write the parameters to the file
For i = 1 To n
  Write #1, Cells(i) & "=", Range(Cells(i)).Value
Next i
'Write a blank line at the end of the file
Write #1, , , ,
Close #1
End Sub
```

```
Sub Opener()
Dim dummy As Variant
Application.ScreenUpdating = False
Call NameCells
n = ArrayMax(Cells)
MyFileName = Application.GetOpenFilename( _
    fileFilter:="My Files (*.mgh), *.mgh," & _
                "Text Files (*.txt), *.txt,")
If MyFileName = "False" Then
  End
End If
Sheets("Parameters").Select
```

```
Open MyFileName For Input As #1
For i = 1 To n
  Input #1, dummy, dummy
  Range(Cells(i)).Value2 = dummy
Next i
Close #1
End Sub
```

```
Sub SaveClimate()
'Saves the climate data used by the model
Dim D As Double, P As Double, Tmin As Double, Tmax As Double, Tavg As Double
Sheets("Climate").Select
Range("A4").Select

'Query the user for the file name and location
Application.ScreenUpdating = False
MyFileName = Application.GetSaveAsFilename( _
    InitialFileName:="ClimateData", _
    fileFilter:="My Files (*.clm), *.clm," & _
                "Text Files (*.txt), *.txt,")
If MyFileName = "False" Then
  End
End If
Open MyFileName For Output As #1
Write #1, Range("ClimateNotes").Value
Write #1, "xlDate", "P (in)", "Tmax °F", "Tmin °F", "Tavg °F"
'Write the climate data to a file

Do While Not IsEmpty(ActiveCell)
  D = ActiveCell.Offset(0, 0).Value
  P = ActiveCell.Offset(0, 1).Value
  Tmin = ActiveCell.Offset(0, 2).Value
  Tmax = ActiveCell.Offset(0, 3).Value
  Tavg = ActiveCell.Offset(0, 4).Value
  Write #1, D, P, Tmin, Tmax, Tavg
  ActiveCell.Offset(1, 0).Select
Loop
'Write a blank line at the end of the file
Write #1, , , , , ,
Close #1
End Sub
```

```
Sub OpenClimate()
'Saves the climate data used by the model
Dim D As Double, P As Double, Tmin As Double, Tmax As Double, Tavg As Double
Dim dummy As String
Dim MyRow As Integer
Dim MyRange As String

'Query the user for the file name and location
Application.ScreenUpdating = False
Application.Calculation = xlCalculationManual
MyFileName = Application.GetOpenFilename( _
    fileFilter:="My Files (*.clm), *.clm," & _
                "Text Files (*.txt), *.txt,")
If MyFileName = "False" Then End

'Out with the old
Sheets("Climate").Select
Range("A4:E4").Select
Range(Selection, Selection.End(xlDown)).Select
Selection.ClearContents

Range("F10").Select
Range(Selection, Selection.End(xlDown)).Select
Selection.ClearContents

Open MyFileName For Input As #1
Input #1, dummy
```

```
Range("ClimateNotes").Value = dummy
Line Input #1, dummy
'The second line contains a header
'Write the climate data to a file
Range("A4").Select

Do
  Input #1, D, P, Tmin, Tmax, Tavg
  If D = 0 Then Exit Do
  ActiveCell.Offset(0, 0).Value = D
  ActiveCell.Offset(0, 1).Value = P
  ActiveCell.Offset(0, 2).Value = Tmin
  ActiveCell.Offset(0, 3).Value = Tmax
  ActiveCell.Offset(0, 4).Value = Tavg
  ActiveCell.Offset(1, 0).Select
Loop
'Write a blank line at the end of the file
MyRow = ActiveCell.row - 1
MyRange = "F9:F" & MyRow

'Fill in the AMC5 formulas
Range("F9").Select
Selection.AutoFill Destination:=Range(MyRange)
Range(MyRange).Select
Calculate

Range("ClimateNotes").Select
Close #1
End Sub
```

```
Sub SaveFlowData()
'Saves the climate data used by the model
Dim D As Double, Q As Double
Sheets("Flow").Select
Range("A4").Select

'Query the user for the file name and location
Application.ScreenUpdating = False
MyFileName = Application.GetSaveAsFilename( _
    InitialFileName:="FlowData", _
    fileFilter:="My Files (*.flow), *.flow," & _
                "Text Files (*.txt), *.txt,")
If MyFileName = "False" Then
  End
End If
Open MyFileName For Output As #1
Write #1, Range("FlowNotes").Value
Write #1, "xlDate", "Q (cfs)"
'Write the climate data to a file

Do While Not IsEmpty(ActiveCell)
  D = ActiveCell.Offset(0, 0).Value
  Q = ActiveCell.Offset(0, 1).Value
  Write #1, D, Q
  ActiveCell.Offset(1, 0).Select
Loop
'Write a blank line at the end of the file
Write #1, , , ,
Close #1
End Sub
```

```
Sub OpenFlowData()
'Saves the climate data used by the model
Dim D As Double, Q As Double
Dim dummy As String
Dim MyRow As Integer
Dim MyRange As String

'Query the user for the file name and location
```

```
Application.ScreenUpdating = False
MyFileName = Application.GetOpenFilename( _
    fileFilter:="My Files (*.flow), *.flow," & _
                "Text Files (*.txt), *.txt,")
If MyFileName = "False" Then
  End
End If

'Out with the old, before in with the new
Sheets("Flow").Select
Range("A4:B4").Select
Range(Selection, Selection.End(xlDown)).Select
Selection.ClearContents


'Clear the formulas, but leave the first row so we don't have to replace them
Range("C5:E5").Select
Range(Selection, Selection.End(xlDown)).Select
Selection.ClearContents
Range("A4").Select

Open MyFileName For Input As #1
Input #1, dummy
Range("FlowNotes").Value = dummy
Input #1, dummy, dummy
'The second line contains a header
'Write the climate data to a file

Do
  Input #1, D, Q
  If D = 0 Then Exit Do
  ActiveCell.Offset(0, 0).Value = D
  ActiveCell.Offset(0, 1).Value = Q
  ActiveCell.Offset(1, 0).Select
Loop

'Fill down the formulas in Colums C, D, and E
  MyRow = ActiveCell.row - 1
  Range("C4:E4").Select
  MyRange = "C4:E" & MyRow
  Selection.AutoFill Destination:=Range(MyRange)
  Range(MyRange).Select
  Calculate
Close #1
Range("A4").Select
End Sub
```

```
Sub SaveBactData()
'Saves the climate data used by the model
Dim D As Double, C As Double
Sheets("Bacteria").Select
Range("A4").Select

'Query the user for the file name and location
Application.ScreenUpdating = False
MyFileName = Application.GetSaveAsFilename( _
    InitialFileName:="BactData", _
    fileFilter:="My Files (*.bact), *.bact," & _
                "Text Files (*.txt), *.txt,")
If MyFileName = "False" Then
  End
End If
Open MyFileName For Output As #1
Write #1, Range("BactNotes").Value
Write #1, "xlDate", "C (#/100 ml)"
'Write the climate data to a file

Do While Not IsEmpty(ActiveCell)
  D = ActiveCell.Offset(0, 0).Value
  C = ActiveCell.Offset(0, 1).Value
  Write #1, D, C
  ActiveCell.Offset(1, 0).Select
Loop
'Write a blank line at the end of the file
Write #1, , , ,
Close #1
End Sub
```

```
Sub OpenBactData()
'Saves the climate data used by the model
Dim D As Double, C As Double
Dim dummy As String

'Query the user for the file name and location
Application.ScreenUpdating = False
MyFileName = Application.GetOpenFilename( _
    fileFilter:="My Files (*.bact), *.bact," & _
                "Text Files (*.txt), *.txt,")
If MyFileName = "False" Then
  End
End If

'Out with the old, before in with the new
Sheets("Bacteria").Select
Range("A4:B4").Select
Range(Selection, Selection.End(xlDown)).Select
Selection.ClearContents
Range("A4").Select

Open MyFileName For Input As #1
Input #1, dummy
Range("BactNotes").Value = dummy
Input #1, dummy, dummy
'The second line contains a header

'Write the climate data to a file
Do
  Input #1, D, C
  If D = 0 Then Exit Do
  ActiveCell.Offset(0, 0).Value = D
  ActiveCell.Offset(0, 1).Value = C
  ActiveCell.Offset(1, 0).Select
Loop
Close #1

Range("A4").Select
End Sub
```

```
Sub MakeFDC()
'Builds the Flow-Duration Curve
'or a plot of ranked flows vs. their exceedance probability
'based on the Pearson plotting position
'Reference: Handbook of Hydrology, 1992, Maidment, ed., page 8-27
'Specifically, in the context of my program
'Takes the daily flows, Qout(i), ranks them, and writes them to the sheet "FDC"
'Sort the values in the array Qout()
Dim Qobs(365) As Double
Dim i As Integer
'Out with the old, in with the new
Worksheets("FDC").Visible = True
Sheets("FDC").Select
Range("A4:D4").Select
Range(Selection, Selection.End(xlDown)).Select
Selection.ClearContents
Sheets("Flow").Select
Range("C4").Select
For i = 1 To n
  Qobs(i) = ActiveCell.Offset(i, 0).Value
Next i
```

```
Call Quicksort(Qout(), 1, n)
Call Quicksort(Qobs(), 1, n)
Sheets("FDC").Select
Range("a3").Select
For i = 1 To n
  ActiveCell.Offset(i, 0).Value = Qobs(n - i + 1)
  ActiveCell.Offset(i, 1).Value = Qout(n - i + 1)
  ActiveCell.Offset(i, 2).Value = i
  ActiveCell.Offset(i, 3).FormulaR1C1 = "=RC[-1]/(" & n & "+1)"
Next i
Worksheets("FDC").Visible = False
Range("B4").Select
End Sub
```

```
Sub MakeCDC()
'Builds the Concentration-Duration Curve

'Takes the daily avg bacteria conc, Cout(i), ranks them, and writes them to the sheet
"CDC"
'Note that the observed data and the model predictions are handled separately, as there
are
'different numbers of observations in each set.

'Sort the values in the array Cout()
Dim i As Integer
'Out with the old, in with the new
Worksheets("CDC").Visible = True
Sheets("CDC").Select
Range("A4:G4").Select
Range(Selection, Selection.End(xlDown)).Select
Selection.ClearContents

Call Quicksort(Cout(), 1, n)
Call Quicksort(Cobs(), 1, nb)
Sheets("CDC").Select

'Write the MODEL values to the sheet first
Range("A3").Select
For i = 1 To n
  ActiveCell.Offset(i, 0).Value = i
  ActiveCell.Offset(i, 1).FormulaR1C1 = "=RC[-1]/(" & n & "+1)"
    ActiveCell.Offset(i, 2).Value = Cout(n - i + 1)
Next i

'Now write the OBSERVED data to the sheet
Range("E3").Select
For i = 1 To nb
  ActiveCell.Offset(i, 0).Value = i
  ActiveCell.Offset(i, 1).FormulaR1C1 = "=RC[-1]/(" & nb & "+1)"
  ActiveCell.Offset(i, 2).Value = Cobs(nb - i + 1)
Next i
Range("B4").Select
Worksheets("CDC").Visible = False
End Sub
```

```
Sub Quicksort(values() As Double, ByVal min As Long, ByVal max As Long)
'From Walkenbach, 1999

Dim med_value As String
Dim hi As Long
Dim lo As Long
Dim i As Long

' If the list has only 1 item, it's sorted.
If min >= max Then Exit Sub

' Pick a dividing item randomly.
i = min + Int(Rnd(max - min + 1))
med_value = values(i)
```

```
' Swap the dividing item to the front of the list.
values(i) = values(min)

' Separate the list into sublists.
lo = min
hi = max
Do
  ' Look down from hi for a value < med_value.
  Do While values(hi) >= med_value
    hi = hi - 1
    If hi <= lo Then Exit Do
  Loop

  If hi <= lo Then
    ' The list is separated.
    values(lo) = med_value
    Exit Do
  End If

  ' Swap the lo and hi values.
  values(lo) = values(hi)

  ' Look up from lo for a value >= med_value.
  lo = lo + 1
  Do While values(lo) < med_value
    lo = lo + 1
    If lo >= hi Then Exit Do
  Loop

  If lo >= hi Then
    ' The list is separated.
    lo = hi
    values(hi) = med_value
    Exit Do
  End If

  ' Swap the lo and hi values.
  values(hi) = values(lo)
Loop ' Loop until the list is separated.

' Recursively sort the sublists.
Quicksort values, min, lo - 1
Quicksort values, lo + 1, max

End Sub
```