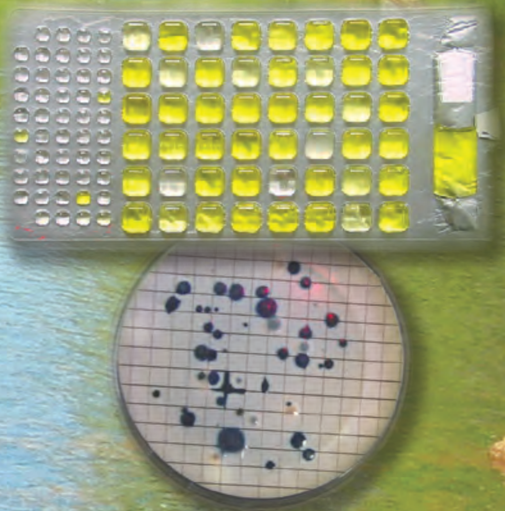
Prepared in cooperation with the National Park Service,
Upper Chattahoochee Riverkeeper, and Cobb County, Georgia

Escherichia coli Bacteria Density in Relation to Turbidity, Streamflow Characteristics, and Season in the Chattahoochee River near Atlanta, Georgia, October 2000 through September 2008—Description, Statistical Analysis, and Predictive Modeling



Scientific Investigations Report 2012–5037

U.S. Department of the Interior
U.S. Geological Survey

Cover. Background photograph: Chattahoochee River below Buford Dam, Forsyth County, Georgia. (Photo by Alan M. Cressler, USGS.)

Foreground photograph, left: Sewer pipes crossing streams are potential sources of raw sewage if damaged during large stormflows. This sewer pipe crosses Rottenwood Creek in the Chattahoochee River National Recreation Area. (Photo by Alan M. Cressler, USGS.)

Foreground photograph, top right: Analytical tray under ambient light showing cells positive for total coliform bacteria. (Photo by Stephen J. Lawrence, USGS.)

Foreground photograph, bottom right: Typical m-FC agar plate with the blue-green colonies of the fecal coliform group of indicator bacteria. (Photo by Stephen J. Lawrence, USGS.)

Escherichia coli Bacteria Density in Relation to Turbidity, Streamflow Characteristics, and Season in the Chattahoochee River near Atlanta, Georgia, October 2000 through September 2008—Description, Statistical Analysis, and Predictive Modeling

By Stephen J. Lawrence

Prepared in cooperation with the National Park Service,
Upper Chattahoochee Riverkeeper, and Cobb County, Georgia

Scientific Investigations Report 2012–5037

U.S. Department of the Interior
U.S. Geological Survey

U.S. Department of the Interior
KEN SALAZAR, Secretary

U.S. Geological Survey
Marcia K. McNutt, Director

U.S. Geological Survey, Reston, Virginia: 2012

For more information on the USGS—the Federal source for science about the Earth, its natural and living resources, natural hazards, and the environment, visit *http://www.usgs.gov* or call 1-888-ASK-USGS

For an overview of USGS information products, including maps, imagery, and publications, visit *http://www.usgs.gov/pubprod*

To order this and other USGS information products, visit *http://store.usgs.gov*

Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Suggested citation:
Lawrence, S.J., 2012, *Escherichia coli* bacteria density in relation to turbidity, streamflow characteristics, and season in the Chattahoochee River near Atlanta, Georgia, October 2000 through September 2008—Description, statistical analysis, and predictive modeling: U.S. Geological Survey Scientific Investigations Report 2012–5037, 81 p., available online at *http://pubs.usgs.gov/sir/2012/5037/.*

Contents

Figures

Appendix Figures

Tables

Appendix Tables

Conversion Factors and Datums

SI to Inch/Pound

Multiply	By	To obtain
	Length	
centimeter (cm)	0.3937	inch
	Volume	
liter (L)	33.82	ounce, fluid (fl. oz)
liter (L)	0.2642	gallon (gal)
liter (L)	61.02	cubic inch (in^3)
liter (L)	0.03531	cubic feet (ft^3)
milliliter (mL)	0.0338	ounce, fluid (fl. oz)
milliliter (mL)	0.06102	cubic inch (in^3)

Temperature in degrees Celsius (°C) may be converted to degrees Fahrenheit (°F) as follows:

$$°F=(1.8\times°C)+32$$

Inch/Pound to SI

Multiply	By	To obtain
	Length	
inch	2.54	centimeter (cm)
foot (ft)	0.3048	meter (m)
mile (mi)	0.6214	kilometer (km)
	Area	
acre	4,047	square meter (m^2)
square mile (mi^2)	0.003861	hectare (ha)
square mile (mi^2)	0.3861	square kilometer (km^2)
	Volume	
gallon (gal)	3.785	liter (L)
cubic foot (ft^3)	28.31687	liters (l)
cubic foot (ft^3)	0.02832	cubic meter (m^3)
acre-foot (acre-ft)	1,233	cubic meter (m^3)
	Flow rate	
foot per second (ft/s)	0.3048	meter per second (m/s)
cubic foot per second (ft^3/s)	28.317	liters per second (L/s)

Vertical coordinate information is referenced to the North American Vertical Datum of 1988 (NAVD 88).

Horizontal coordinate information is referenced to the North American Datum of 1983 (NAD 83).

Altitude, as used in this report, refers to distance above the vertical datum.

Acronyms and Abbreviations Used in This Report

AIC	Akaike Information Criterion
ANOVA	analysis of variance
Atlanta site	USGS station number 02336000
β_0	intercept term in regression
β_n	slope term for each n explanatory variable in regression
c	probability of concordance
CI	confidence interval
Colilert	Colilert®-18
colonies/100 mL	colony-forming units per 100 milliliters of water
COV	coefficient of variation
Cp	Mallow’s Cp statistic
CRNRA	Chattahoochee River National Recreation Area
DFFITS	measure of the degree that one data point influences the regression
DI	de-ionized
Dxy	Somer’s rank correlation
E. coli	*Escherichia coli*
EVENT	streamflow event (dry-weather flow or stormflow)
FallingQ	falling stage
FNU	formazin nephelometric unit
GAEPD	Georgia Environmental Protection Division
GAWSC	USGS Georgia Water Science Center
gCOV	coefficient of geometric variation
HCOND	streamflow condition
IQR	interquartile range
Lake Lanier	Lake Sidney Lanier
LOC	line of organic correlation
log$_{10}$	log base 10
log10Ecoli	log$_{10}$ transformed *E. coli*
log10Flow	log$_{10}$ transformed streamflow
log10FNU	log$_{10}$ transformed turbidity
LOGR	logistic regression

MLR	multiple linear regression
MOVE	Method of Variance Extension
MPN/100 mL	most probable number of colonies per 100 milliliters of water
MSS_f	mean sum of squared residuals
MSS_p	mean sum of squared prediction errors
MUG	4-methyllumbelliferyl-β-D-glucuronide
Norcross site	USGS station number 02335000
nm	nanometer
NPS	National Park Service
NTRU	nephelometric turbidity ratio unit
NURP	National Urban Runoff Program
NWIS	USGS National Water Information System
OLS	ordinary least-squares regression
PeakQ	peak stage
psi	pounds per square inch
QA/QC	quality assurance/quality control
Quanti-Tray	Quanti-Tray®/2000
R^2	coefficient of determination
RisingQ	rising stage
RPD	relative percent difference
SLR	simple linear regression
StableLow	stable flow, low stage
StableNorm	stable flow, normal stage
StableHigh	stable flow, high stage
USEPA	U.S. Environmental Protection Agency
USGS	U.S. Geological Survey
USPHS	U.S. Public Health Service
VIF	variance inflation factor
WTEMP	water temperature variable

Acknowledgments

The BacteriALERT program would not have been possible without the support and partnerships involved in the planning and implementation of the program. In particular, many thanks to the Cobb County–Marietta Water Authority; Stephen D. McCullers of the Cobb County Water System; Dan Brown, Rick Slade, Allyson Read, Kevin Cherie, Bill Carroll, David Ek, Michele Mitchell, and Lauren Theodore of the National Park Service, Chattahoochee River National Recreation Area; Sally Bethea, Jason Ulseth, Birgit Bolton, Kristi Rose, Bill Crawford, Alice Champagne, and others of the Upper Chattahoochee Riverkeeper; Alan Hallum of the Water Protection Branch, Georgia Environmental Protection Division (retired); John Sibley and others of the Georgia Conservancy; the Trust for Public Land; and Jennifer Derby and Hudson Slay of Region 4, U.S. Environmental Protection Agency. Also, thanks to Morgan Pecora-Saipe and Audrey Biggerstaff who, as students completing a high school science project at the Lovett School, Vinings, Georgia, in 2002–2003, collected water samples along the channel cross section at the Chattahoochee River at Atlanta and analyzed them for *Escherichia coli* bacteria and turbidity.

Escherichia coli Bacteria Density in Relation to Turbidity, Streamflow Characteristics, and Season in the Chattahoochee River near Atlanta, Georgia, October 2000 through September 2008—Description, Statistical Analysis, and Predictive Modeling

By Stephen J. Lawrence

Abstract

Water-based recreation—such as rafting, canoeing, and fishing—is popular among visitors to the Chattahoochee River National Recreation Area (CRNRA) in north Georgia. The CRNRA is a 48-mile reach of the Chattahoochee River upstream from Atlanta, Georgia, managed by the National Park Service (NPS). Historically, high densities of fecal-indicator bacteria have been documented in the Chattahoochee River and its tributaries at levels that commonly exceeded Georgia water-quality standards. In October 2000, the NPS partnered with the U.S. Geological Survey (USGS), State and local agencies, and non-governmental organizations to monitor *Escherichia coli* bacteria (*E. coli)* density and develop a system to alert river users when *E. coli* densities exceeded the U.S. Environmental Protection Agency (USEPA) single-sample beach criterion of 235 colonies (most probable number) per 100 milliliters (MPN/100 mL) of water. This program, called BacteriALERT, monitors *E. coli* density, turbidity, and water temperature at two sites on the Chattahoochee River upstream from Atlanta, Georgia. This report summarizes *E. coli* bacteria density and turbidity values in water samples collected between 2000 and 2008 as part of the BacteriALERT program; describes the relations between *E. coli* density and turbidity, streamflow characteristics, and season; and describes the regression analyses used to develop predictive models that estimate *E. coli* density in real time at both sampling sites.

Between October 23, 2000, and September 30, 2008, about 1,400 water samples were collected and turbidity was measured at each of the two USGS streamgaging stations in the CRNRA near the cities of Norcross and Atlanta, Georgia. At both sites, water samples were collected at frequencies ranging from daily to twice per week and analyzed in the laboratory for *E. coli* bacteria, using the Colilert-18® and Quanti-tray-2000® defined substrate method, and turbidity. Beginning in mid-2002, turbidity and water temperature were measured in real time at both sites. Streamflow at both sites is affected by the operation of two hydroelectric facilities upstream that release water in response to daily peak power demands in the area. During dry weather, offpeak water released from both dams ranges from about 600 to 1,500 cubic feet per second.

During dry weather, 98 and 93 percent of water samples from Norcross and Atlanta sites, respectively, contained *E. coli* densities below the USEPA single-sample beach criterion (235 MPN/100 mL). Conversely during stormflow, only 26 percent of the samples from Norcross and 10 percent of the samples from Atlanta contained *E. coli* densities below the USEPA beach criterion. At both sites, median *E. coli* density and turbidity were statistically greater in stormflow samples than dry-weather samples. Furthermore, median *E. coli* density and turbidity were statistically lower at Norcross than at Atlanta during dry weather. During stormflow, median turbidity values were statistically similar at the two sites (36 and 35 formazin nephelometric units at Norcross and Atlanta, respectively); whereas the median *E. coli* density was statistically higher at Atlanta (810 MPN/100 mL) than at Norcross (530 MPN/100 mL). During dry weather, the maximum *E. coli* density was 1,200 MPN/100 mL at Norcross and 9,800 MPN/100 mL at Atlanta. During stormflow, the maximum *E. coli* density was 18,000 MPN/100 mL at Norcross and 28,000 MPN/100 mL at Atlanta.

Regression analyses show that *E. coli* density in samples was strongly related to turbidity, streamflow characteristics, and season at both sites. The regression equation chosen for

the Norcross data showed that 78 percent of the variability in *E. coli* density (in log base 10 units) was explained by the variability in turbidity values (in log base 10 units), streamflow event (dry-weather flow or stormflow), season (cool or warm), and an interaction term that is the cross product of streamflow event and turbidity. The regression equation chosen for the Atlanta data showed that 76 percent of the variability in *E. coli* density (in log base 10 units) was explained by the variability in turbidity values (in log base 10 units), water temperature, streamflow event, and an interaction term that is the cross product of streamflow event and turbidity. Residual analysis and model confirmation using new data indicated the regression equations selected at both sites predicted *E. coli* density within the 90 percent prediction intervals of the equations and could be used to predict *E. coli* density in real time at both sites.

Introduction

In 1914, the U.S. Public Health Service (USPHS) created a coliform index using the density of total coliform bacteria as the indicator of ambient water quality in the United States (Maier and others, 2000, p. 491). During the 1940s and 1950s, the USPHS used the coliform index in epidemiological studies at beaches in Chicago, Illinois, the Ohio River in Kentucky, and Long Island Sound in New York. Citing comparisons between total coliform bacteria and the more specific fecal coliform bacteria densities measured in studies during the 1960s, the Department of the Interior National Technical Advisory Committee recommended that fecal coliform bacteria replace total coliform bacteria as the indicator of ambient water quality in the United States (U.S. Department of the Interior, 1968). In 1986, however, the U.S. Environmental Protection Agency (USEPA) recommended *Escherichia coli* (*E. coli)* bacteria as the preferred indicator bacteria for identifying fecal contamination in ambient freshwater (U.S. Environmental Protection Agency, 1986). Using the results from epidemiological studies on the relation between swimming-associated gastroenteritis and indicator bacteria density at beaches, the USEPA published a list of single-sample maximum allowable *E. coli* bacteria densities for various levels of body-contact recreation in surface waters of the United States (U.S. Environmental Protection Agency, 1986, table 4). Table 1 lists these criteria for *E. coli* bacteria and the Georgia standards for fecal coliform bacteria.

Table 1. Georgia water-quality standards for fecal coliform bacteria and the U.S. Environmental Protection Agency *Escherichia coli (E. coli)* bacteria criteria in ambient freshwater for primary and secondary body-contact recreation.

[All values are in colonies per 100 milliliters of water; —, not applicable; USEPA, U.S. Environmental Protection Agency]

Illness rate (per 1,000 swimmers)	Primary contact recreation[b]					Secondary contact recreation[c]
	Geometric mean[a]	Single-sample maximum allowable density				Single-sample maximum allowable density
		Designated beach area	Moderate full-body contact	Lightly used full-body contact	Infrequently used full-body contact	
Fecal coliform (Georgia Environmental Protection Division, 2009)						
8	200[d]/1,000	—	—	—	—	4,000
USEPA *E. coli* criteria (U.S. Environmental Protection Agency, 1986, 2002)						
8	126	235	298	406	576	—
9	160	300	381	524	736	—
10	206	383	487	669	941	—
11	263	490	622	855	1,202	—
12	336	626	795	1,092	1,536	—
13	429	799	1,016	1,396	1,962	—
14	548	1,021	1,298	1,783	2,507	—

[a] Geometric mean of at least five dry-weather samples collected during separate 24-hour periods within a 30-day period.

[b] Recreation, except fishing, during which the body is immersed in a body of water such that water may be ingested inadvertently.

[c] Recreation such as fishing or wading such that the ingestion of water from a body of water is unlikely. In effect from November 1 to April 30.

[d] Georgia standards require four samples during separate 24-hour periods within a 30-day period, between May and October; 1,000 colonies per 100 milliliters during rest of year.

In the mid- to late-1970s, the USEPA National Urban Runoff Program (NURP; U.S. Environmental Protection Agency, 1983) established that high densities of fecal coliform bacteria were widespread in rivers and streams within and downstream from major urban centers throughout the United States. The high fecal coliform bacteria densities also were reported in the Atlanta, Georgia, area during the NURP study, especially in streams tributary to the Chattahoochee River (McConnell, 1980). More recent studies in a 48-mile reach of the Chattahoochee River upstream from Atlanta showed that the densities of fecal coliform and *E. coli* bacteria increased in a downstream direction, especially as the river approached the high density urban core of Atlanta (Gregory and Frick, 2000, 2001). Because of the consistent fecal coliform bacteria densities that exceed the water-quality standards for Georgia, the Georgia Environmental Protection Division (GAEPD) has listed the 12-mile reach of the Chattahoochee River between Morgan Falls Dam and Peachtree Creek in Atlanta as partially impaired and unable to fully support its designated uses for drinking water and recreation (Georgia Environmental Protection Division, 2010).

The 48-mile reach of the Chattahoochee River upstream from Atlanta is managed by the National Park Service (NPS) as the Chattahoochee River National Recreation Area (CRNRA; fig. 1; National Park Service, 2009). Because a large number of visitors to the CRNRA use the river for recreation, primarily rafting, canoeing, and fishing, the historically high levels of fecal coliform bacteria in the Chattahoochee River concerned NPS staff because of the potential health effects to park visitors. The NPS expressed a desire to alert river users when fecal indicator bacteria exceeded Georgia water-quality standards. To explore the feasibility of an alert system in the CRNRA, the NPS along with the U.S. Geological Survey (USGS), the Upper Chattahoochee Riverkeeper, the Georgia Conservancy, the GAEPD, Cobb County Water System, and Cobb-Marietta Water Authority proposed a program for monitoring *E. coli* bacteria densities in the Chattahoochee River upstream from Atlanta.

This program, named BacteriALERT, was designed and implemented by the USGS beginning in October 2000. The program was designed to collect and analyze water samples for *E. coli* bacteria and alert park visitors when *E. coli* density in the river exceeded the USEPA single-sample beach criterion of 235 colony-forming units per 100 milliliters of water (colonies/100 mL; U.S. Environmental Protection Agency, 1986). The *E. coli* bacteria was chosen as the indicator bacteria for BacteriALERT because (1) the USEPA had selected *E. coli* as their preferred indicator bacteria for ambient freshwater and (2) methods for analyzing *E. coli* in water samples were available that made enumeration easier and quicker than the membrane filter methods commonly used to quantify fecal coliform bacteria. Although the USEPA has strongly encouraged State and local entities to adopt the *E. coli* criterion (U.S. Environmental Protection Agency, 1986), the State of Georgia, as of late 2011, continues to use fecal coliform as the indicator bacteria for ambient water in the State.

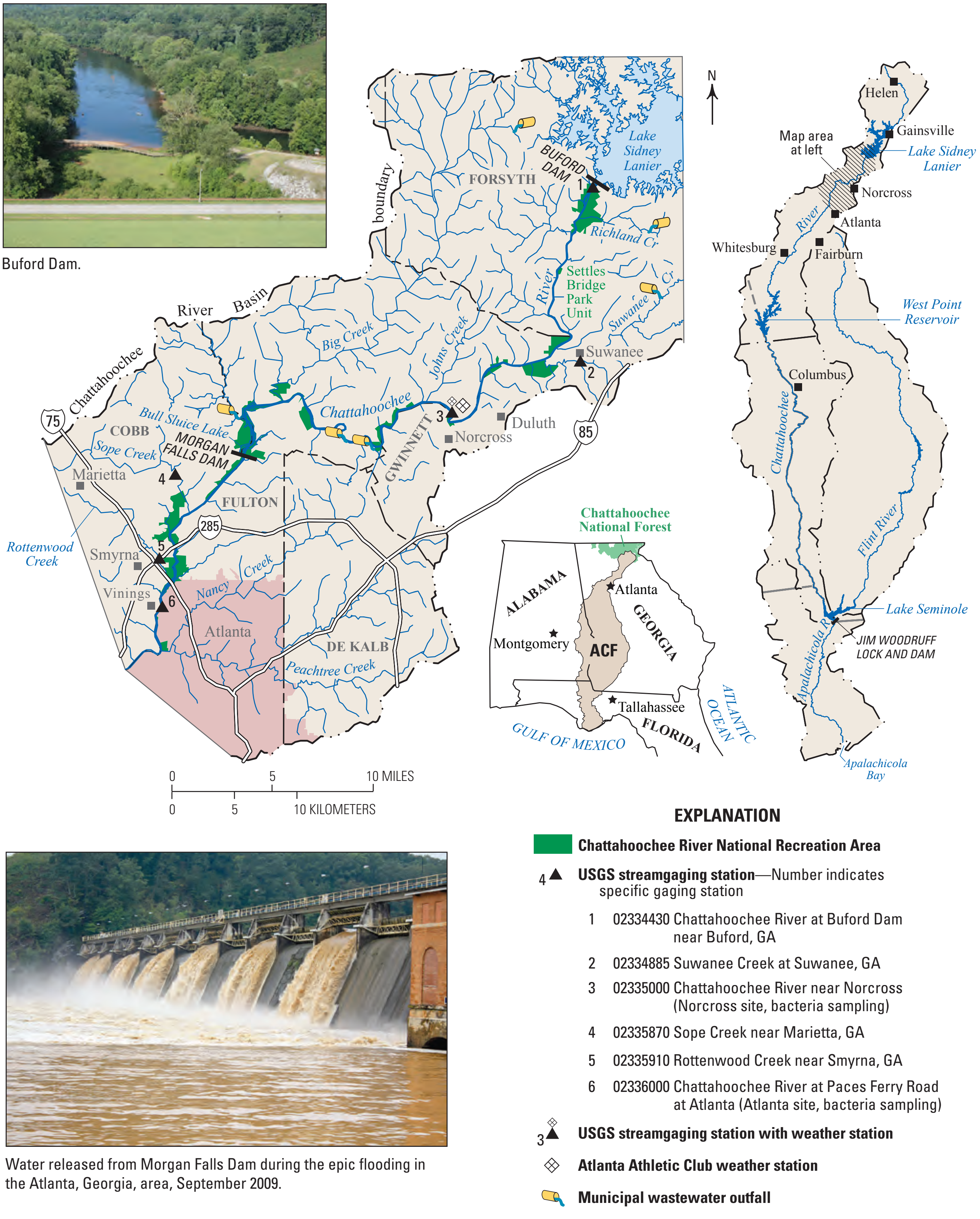
Purpose and Scope

The purpose of this report is fourfold: (1) describe *E. coli* density and turbidity at two sampling sites on the Chattahoochee River during different seasons and for various streamflow characteristics; (2) describe the development of regression models to predict *E. coli* density; (3) document, in detail, the methods used to collect water samples during the study period, analyze those samples for *E. coli* bacteria, and to measure turbidity, water temperature, and streamflow; and (4) describe the methods used for quality assurance, data and statistical analyses, and regression analyses. The report presents the results of more than 1,400 water samples collected and analyzed for *E. coli* density and turbidity between October 23, 2000, and September 30, 2008, at each of the two sites on the Chattahoochee River upstream from Atlanta. Also presented in the report are comparisons among three different laboratory methods for enumerating *E. coli* densities and between turbidity measured in the laboratory and instream at each site. In addition, the report presents conceptual models for predicting *E. coli* density in real time at both sampling sites.

Study Area and Site Descriptions

The Chattahoochee River begins as a small first-order stream within the Chattahoochee National Forest, northwest of Helen, Ga. (fig. 1). The river flows for approximately 540 river miles, in a southwesterly direction, then southerly through Georgia along the Georgia-Alabama border and into Lake Seminole, the water body impounded by the Jim Woodruff Lock and Dam at the Georgia-Florida-Alabama border. Water is released from Lake Seminole into the Apalachicola River in Florida and flows into the Gulf of Mexico at Apalachicola Bay, Florida. Before reaching Lake Seminole, the Chattahoochee River is impounded at several locations along the Georgia-Alabama border. The river flows through or near several Georgia cities, such as Helen, Gainesville, Norcross, Atlanta, and Columbus.

At a location 348 miles upstream from Apalachicola Bay in Florida (river mile 348; U.S. Army Corps of Engineers, 1985), the Chattahoochee River is impounded by Buford Dam to form Lake Sidney Lanier (Lake Lanier; fig. 1), a multipurpose reservoir in the upper Chattahoochee River Basin. Buford Dam was completed in 1956 (U.S. Army Corps of Engineers, 2006). Lake Lanier is a large reservoir with 692 miles of shoreline that inundates about 39,000 acres at a power pool altitude of 1,071 feet above North American Vertical Datum of 1988 (NAVD 88). At the power pool altitude, Lake Lanier has a storage capacity of 1.05 million acre-feet. Typically, Buford Dam releases water when the demand for electric power is greatest, usually mid- to late-afternoon on most days. During periods of low power demand, Buford Dam releases water at a minimum rate of at 600 to 1,500 cubic feet per second (ft^3/s); however, at peak demand that rate commonly increases to a maximum between 5,000 and 6,000 ft^3/s, but can be as high as 10,000 ft^3/s (Georgia Power, 2004a, b).


Buford Dam.
FORSYTH
BUFORD DAM
1
Lake Sidney Lanier
Richland Cr
Settles Bridge Park Unit
River
Suwanee Cr
Suwanee
2
boundary
Chattahoochee River Basin
Big Creek
Johns Creek
Chattahoochee
3
Duluth
Norcross
GWINNETT
85
75
COBB
Bull Sluice Lake
MORGAN FALLS DAM
Sope Creek
Marietta
4
FULTON
285
Rottenwood Creek
Smyrna
5
Vinings
6
Nancy Creek
Atlanta
DE KALB
Peachtree Creek
0 5 10 MILES
0 5 10 KILOMETERS
N
Map area at left
Helen
Gainsville
Lake Sidney Lanier
Norcross
Atlanta
Whitesburg
Fairburn
River
West Point Reservoir
Columbus
Chattahoochee
Flint River
Lake Seminole
JIM WOODRUFF LOCK AND DAM
Apalachicola R
Apalachicola Bay
Chattahoochee National Forest
Atlanta
ALABAMA
GEORGIA
Montgomery
ACF
Tallahassee
FLORIDA
GULF OF MEXICO
ATLANTIC OCEAN
EXPLANATION
🟩 Chattahoochee River National Recreation Area
4 ▲ USGS streamgaging station—Number indicates specific gaging station
1 02334430 Chattahoochee River at Buford Dam near Buford, GA
2 02334885 Suwanee Creek at Suwanee, GA
3 02335000 Chattahoochee River near Norcross (Norcross site, bacteria sampling)
4 02335870 Sope Creek near Marietta, GA
5 02335910 Rottenwood Creek near Smyrna, GA
6 02336000 Chattahoochee River at Paces Ferry Road at Atlanta (Atlanta site, bacteria sampling)
3 ▲ USGS streamgaging station with weather station
◇ Atlanta Athletic Club weather station
Municipal wastewater outfall

Buford Dam.

Water released from Morgan Falls Dam during the epic flooding in the Atlanta, Georgia, area, September 2009.

Figure 1. The Chattahoochee River corridor from Buford Dam to Atlanta, Georgia, showing the Chattahoochee River National Recreation Area, bacteria sampling sites, and USGS streamgages used in the study. (ACF, Apalachicola–Chattahoochee–Flint River basin; photographs by Alan M. Cressler, USGS.)

The Chattahoochee River flows south out of Lake Lanier and through the northernmost extent of the Atlanta metropolitan area as it moves southwest toward the Alabama border (fig. 1). The study area of the BacteriALERT program is the mainstem Chattahoochee River within the CRNRA. The CRNRA, which consists of 17 management units, contains about 75 percent of all public green space in a 10-county area of Metropolitan Atlanta (National Park Service, 2009). Between 1991 and 2000, the recreation area attracted about 1.7 to 3.5 million visitors, with nearly 30 percent of those participating in water-based recreation (Kunkle and Vana-Miller, 2000). The number of visitors to the CRNRA peaked in 1996 at 3.5 million, but since then visitation has steadily declined to about 2.7 million in 2000. Fifteen of those management units are within the study area of this report. In 1999, the release of 26 million gallons of raw or partially treated sewage effluent was documented by the GAEPD within the CRNRA (National Park Service, 2009).

Two USGS streamgaging stations were selected as sampling sites to represent the middle and lower reaches of the CRNRA: Chattahoochee River near Norcross, GA (USGS station number 02335000; Norcross site), about 17 river miles downstream from Buford Dam and Chattahoochee River at Atlanta, GA (USGS station number 02336000; Atlanta site), about 44 miles downstream from Buford Dam and 9 miles downstream from Morgan Falls Dam. In addition, streamflow or stream stage (gage height, water-surface altitude from an established datum) data from four USGS streamgaging stations in the Chattahoochee River basin (fig. 1) were used during data analysis: Chattahoochee River at Buford Dam near Buford, GA (USGS station number 02334430), Suwanee Creek at Suwanee, GA (USGS station number 02334885), Rottenwood Creek near Smyrna, GA (USGS station number 02335910), and Sope Creek near Marietta, GA (USGS station number 02335870). All four streamgaging stations are part of the USGS Georgia Stream-Discharge Measurement Network.

Streamflow in the Chattahoochee River at the Norcross site is affected by water releases from Buford Dam. During dry weather, nearly 90 percent of the streamflow at Norcross is water released by Buford Dam; however, that percentage (depending on the timing of peak discharges from Buford Dam) decreases during wet weather as a result of storm runoff from the numerous tributaries between Buford Dam and Norcross. The Chattahoochee River watershed between Buford Dam and the Norcross site encompasses 130 square miles (mi^2; U.S. Geological Survey, 2011a). In 2001, land use in the Chattahoochee River watershed between Buford Dam and the Norcross site consisted of low to high intensity urban (about 28 percent, primarily residential), open space (39 percent), and mixed forest (18 percent). The remaining 15 percent was a mixture of wetland, grass, scrub, and pasture (U.S. Geological Survey, 2011b). Within this drainage, three wastewater outfalls exist on tributaries upstream from the Norcross site.

In contrast to the Norcross site, the hydrology of the Chattahoochee River is more complicated at the Atlanta site because streamflow at Atlanta is not only influenced by water releases from Buford Dam, but also by water releases from Bull Sluice Lake behind Morgan Falls Dam and by the numerous tributaries between the Atlanta and Norcross sites. Morgan Falls Dam was completed in 1904 as one of the first hydroelectric powerplants in the United States. Currently the dam and hydroelectric facilities are owned and operated by the electric power subsidiary of the Southern Company. The dam impounds Bull Sluice Lake, which at full pool has a surface area of 673 acres and 2,250 acre-feet of usable storage (Georgia Power, 2004b).

During dry weather and with minimum inflows to Bull Sluice Lake (weekly inflow average of about 956 ft^3/s), the lake altitude fluctuates about 2 feet (ft) in response to the magnitude and duration (2 to 3 hours) of water released from Buford and Morgan Falls Dams. Moreover, during dry weather and with average inflows to Bull Sluice Lake (weekly inflow average of about 2,381 ft^3/s), the lake altitude fluctuates by about 4 ft in response to the magnitude and duration (10 to 12 hours) of water released from Buford and Morgan Falls Dams. When inflows to Bull Sluice Lake exceed 6,000 ft^3/s, the lack of storage capacity in the lake requires that Morgan Falls Dam operate as a run-of-the-river dam, whereby inflows equal outflows (Georgia Power, 2004a). The estimated hydraulic residence time in Bull Sluice Lake ranges from 6 hours at an inflow of 3,000 ft^3/s to 2.5 days at an inflow of 500 ft^3/s (Georgia Power, 2004a).

The Chattahoochee River watershed between the Norcross site and the Atlanta site encompasses about 410 mi^2 (U.S. Geological Survey, 2011a). In 2001, land use within this drainage area consisted of low to high intensity urban (about 44 percent, primarily residential), open space (30 percent), and mixed forest (22 percent). The remaining 6 percent was a mixture of wetland, grass, scrub, and barren land (U.S. Geological Survey, 2011b). Within this drainage area, three wastewater outfalls exist on tributaries upstream from the Atlanta site.

Climate and Streamflow Characteristics During the Study Period

During the study period, drought was the dominant weather pattern in north-central Georgia. Sixty percent of the time between October 23, 2000, and September 30, 2008, total monthly precipitation for the study area was below the 77-year monthly average (normal rainfall; fig. 2*A*). In May 1998, a moderate to severe drought began in Georgia and parts of the Southeastern United States and continued until late September 2002 (David Stooksbury, Georgia State Climatologist, written commun., December 2002). Between October 2000 and late September 2002, the cumulative monthly rainfall deficit was 44 inches. During the study period, only two extended periods of above average precipitation were measured in north-central Georgia (figs. 2*A*, *B*). The first began in late September 2002 and lasted for about 9 months. During this 9-month period, the cumulative monthly rainfall surplus was slightly more than 10 inches. The second period of above average precipitation began in September 2004 and lasted until the end of August 2005. During this 12-month period, the cumulative monthly rainfall surplus was slightly more than 17 inches. Beginning in September 2005, north-central Georgia was mired in a severe, multiyear drought that continued beyond the end of the study period. During this period, the cumulative monthly rainfall deficit was 21 inches. These drought conditions, which limited the number of storms in the study area, coincided with sample collection and resulted in a relatively small number of storm samples from both sites.

The climatic variability during the study period affected the amount and duration of water released from Buford Dam. Water releases from Buford Dam were minimal (between 600 and 750 ft^3/s) during two 7-month periods—October 2000 to April 2001 and November 2001 to March 2002 (figs. 3*A*, 4*A*). During these periods, water releases averaged less than 2 hours per day. These minimal releases affected streamflow in the study area, resulting in minimal flows at the Norcross and Atlanta sites. In contrast,

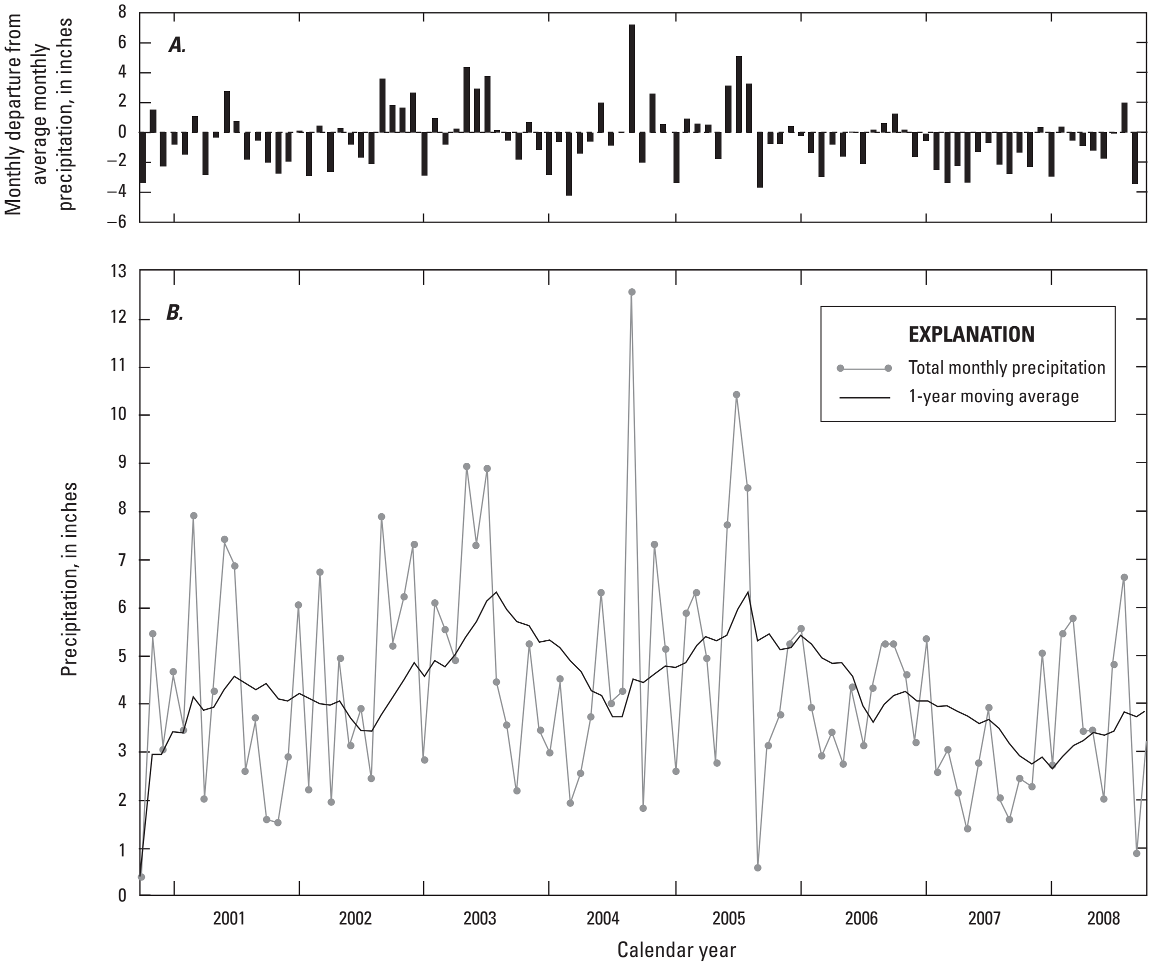


Figure 2. Precipitation trends from October 2000 through September 2008 for 19 aggregated climate stations in north-central Georgia. *(A)* Average monthly departures from the average monthly precipitation for period of record (1931–2008). *(B)* Monthly precipitation trends. (Data from National Oceanic and Atmospheric Administration, 2011).

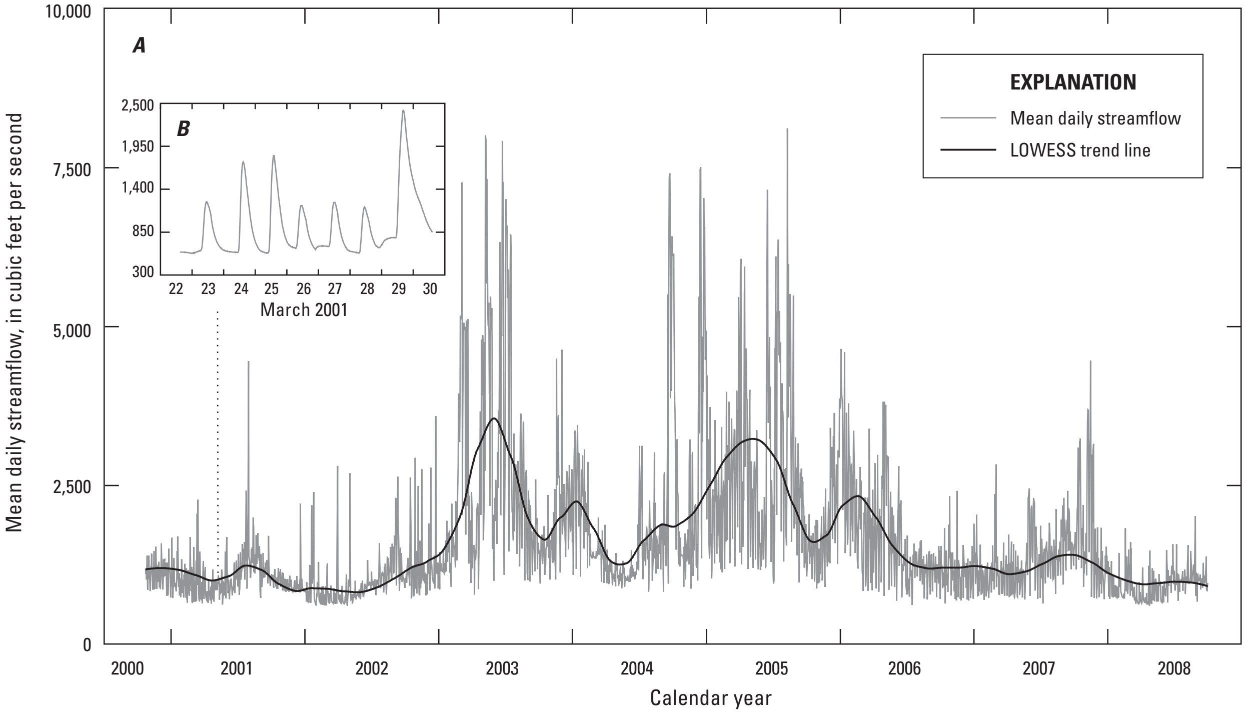


Figure 3. *(A)* Streamflow regime for the Chattahoochee River near Norcross, Georgia (USGS station number 02335000), from October 2000 through September 30, 2008. *(B)* Inset shows streamflow over a typical 9-day period, May 22 to May 30, 2001, due to water releases from Buford Dam.

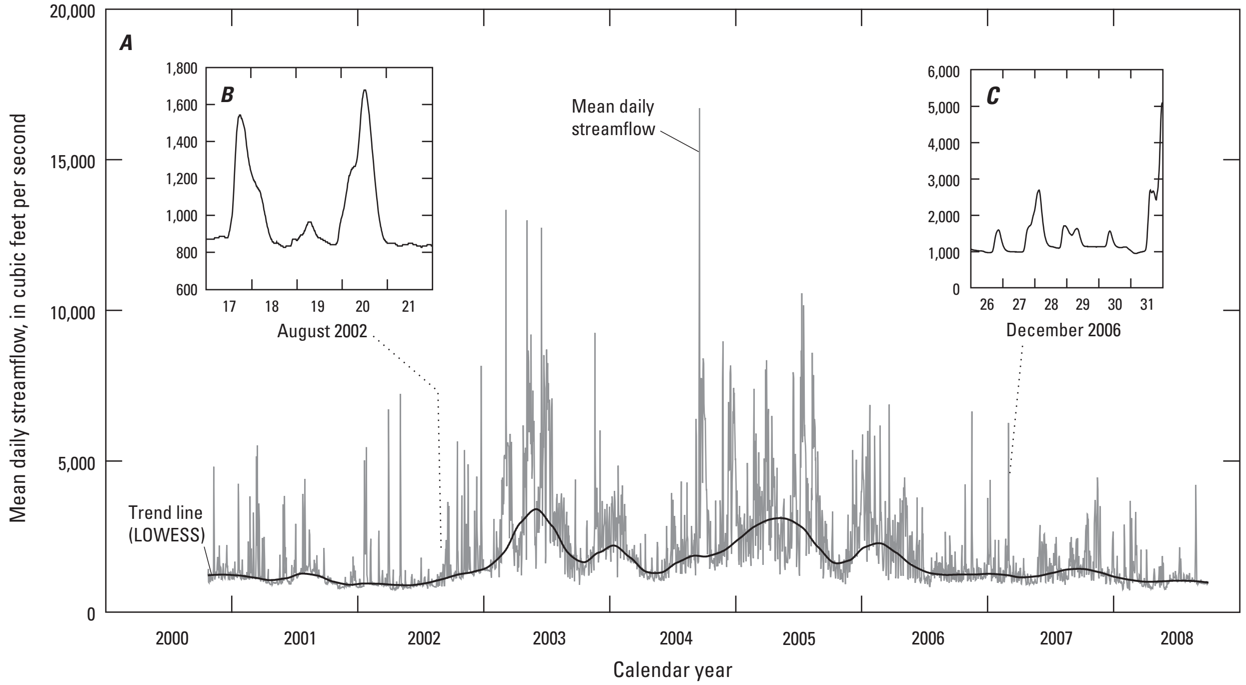


Figure 4. *(A)* Streamflow regime for the Chattahoochee River at Atlanta, Georgia (USGS station number 02336000), from October 2000 through September 30, 2008. *(B)* Streamflow over a typical 5-day period, August 17–21, 2002, due to water releases from Morgan Falls Dam. *(C)* Streamflow over a 6-day period, December 26–31, 2006, due to storm runoff and water releases from Morgan Falls Dam.

several consecutive months of above average rainfall occurred during the fourth quarter of 2002 and the first quarter of 2003, resulting in the release of large amounts of water from Buford Dam that, beginning in January 2003, continued for more than 24 consecutive hours. From January 2003 to April 2003, the average daily duration of water released from Buford Dam ranged from slightly more than 2 hours to nearly 16 hours per day at discharges as high as 7,000 ft^3/s.

The streamflow events at the Norcross and Atlanta sites during the study period are shown in figures 3*A* and 4*A*, respectively. The median streamflow at the Norcross and Atlanta sites was lowest between March and May 2001, between February and June 2002, and between mid-November 2007 and September 30, 2008, the end of the study period. These low streamflows correspond to the below average rainfall measured during those dates (fig. 2). The median streamflow at Norcross and Atlanta increased substantially during the first quarter of 2003 in response to drought-breaking rainfall and the subsequent increase in water released from Buford and Morgan Falls Dams (figs. 3*A* and 4*A*).

Streamflow at both sites varies greatly on a daily basis. Water released daily from Buford Dam can change from 600 to 6,000 ft^3/s within a 60-minute time span. Typically during dry weather, peak discharges from Buford Dam range from 4,500 to 5,500 ft^3/s. At these streamflows, the peak will reach the Norcross site within 6 or 7 hours, a distance of 17 river miles from Buford Dam. The streamflow response at the Norcross site to those releases from Buford Dam is shown in figure 3*B*. In addition, those peak discharges from Buford Dam reach Bull Sluice Lake in about 12 hours, a distance of 35 river miles. During dry weather, peak discharges from Morgan Falls Dam ranged between 1,400 to 1,500 ft^3/s and reached the Atlanta site in about 4.5 hours, a distance of 9 river miles. A typical pattern of dry-weather streamflow at the Atlanta site is shown in figure 4*B* and is the result of water releases from Morgan Falls Dam. Figure 4*C* shows streamflow at the Atlanta site between December 26, 2006, and January 1, 2007, in response to stormflow coinciding with water releases from Morgan Falls Dam.

At the Atlanta site, the effect of water releases from Buford Dam is attenuated by the distance from the dam (about 45 river miles) and by Bull Sluice Lake, the impoundment behind Morgan Falls Dam. As mentioned previously, Bull Sluice Lake has minimal storage capacity at low to average streamflows, but has enough storage to dampen the daily streamflow peaks from Buford Dam and affect the water temperature, sediment load, and turbidity of water released by Morgan Falls Dam. The vagaries of streamflow may add substantial variability to the measured turbidity values and *E. coli* densities at both sites. In addition, algae and aquatic macrophyte communities in Bull Sluice Lake probably add to the turbidity levels observed downstream from Morgan Falls Dam, especially during high flows. Between 2002 and 2005, 35 species of aquatic macrophytes and 2 algae species were identified in Bull Sluice Lake during aquatic plant surveys by NPS and Georgia Power (report online at *http://www.georgiapower.com/lakes/hydro/pdfs/StudyReport_Wetlands.pdf*, accessed May 31, 2011).

Previous Studies

Fecal coliform and *E. coli* bacteria densities in the Chattahoochee River and tributaries within and upstream from Atlanta have been studied by several researchers in years past. In a study for the U.S. Army Corps of Engineers (Robert N. Morris, Black, Crow, and Eidsness, Inc., 1975), stormwater samples from four urban tributaries to the Chattahoochee River within the city of Atlanta contained mean fecal coliform bacteria densities ranging from 2,100 to 11,000 colonies/100 mL and maximum densities ranging from 30,000 to 100,000 colonies/100 mL. In addition, a report by McConnell (1980) for the USEPA NURP described fecal coliform bacteria densities at the four tributary sites studied by Morris (1975) and at five additional urban sites on streams tributary to the Chattahoochee River within the city of Atlanta. In the McConnell (1980) report, mean fecal coliform densities ranged from 300 to 130,000 colonies/100 mL during dry-weather streamflow and from 4,500 to 260,000 colonies/100 mL during storm runoff. The maximum fecal coliform bacteria density reported in the McConnell (1980) report was 800,000 colonies/100 mL in a storm runoff (combined sewer system) sample from Peachtree Creek near its confluence with the Chattahoochee River (fig. 1).

More recent studies by Gregory and Frick (2000, 2001) focused on fecal coliform bacteria densities in the Chattahoochee River and its tributaries upstream from Atlanta. In water samples collected in 1994 and 1995, streams tributary to the Chattahoochee River upstream from Atlanta, such as Suwanee Creek, Richland Creek, Johns Creek, Big Creek, and Rottenwood Creek (fig. 1), contained fecal coliform densities that were 10 to 15 times higher than densities in the mainstem Chattahoochee River (Gregory and Frick, 2000). Seventy-four to 96 percent of water samples from those five tributaries contained fecal coliform densities that exceeded the USEPA review criterion of 400 colonies/100 mL. In addition, fecal coliform densities in those five tributaries were 6 to 10 times higher during stormflow than dry-weather flow.

The studies by Gregory and Frick (2000, 2001) showed that fecal coliform density increased in a downstream direction in the mainstem Chattahoochee River. In 1994 and 1995, fecal coliform densities in the Chattahoochee River upstream from Suwanee, Ga., typically were less than 20 colonies/100 mL, well below the Georgia single-sample (4,000 colonies/100 mL) criterion for secondary body contact recreation and below the Georgia primary contact (beach) recreation standard of 200 colonies/100 mL (table 1). Between the cities of Suwanee and Atlanta, however, fecal coliform bacteria densities increased markedly from a median density of 20 colonies/100 mL to 800 colonies/100 mL (Gregory and Frick, 2000).

Furthermore, during 1999 and 2000, fecal coliform densities exceeded the Georgia geometric mean water-quality standard (200 colonies/100 mL) in 12 percent of water samples, and *E. coli* densities exceeded the USEPA single-sample beach criterion (235 colonies/100 mL) in 13 percent of water samples collected from the Chattahoochee River at Settles Bridge, 4.5 river miles below Buford Dam (Gregory and Frick, 2001). In contrast, fecal coliform bacteria densities exceeded the Georgia primary recreation standard (200 colonies/100 mL) in 67 percent of water samples, and *E. coli* densities exceeded the USEPA single-sample beach criterion in 81 percent of water samples from the Chattahoochee River at Atlanta (referred to as the Atlanta site in this report). Gregory and Frick (2001) also noted that fecal coliform and *E. coli* densities fluctuated on a 12-hour cycle with the highest densities occurring between 10:00 p.m. and 2:00 a.m. and the lowest between 2:00 p.m. and 5:00 p.m. These time periods correspond to the expected periods of the daily minimum and maximum levels of ultraviolet light.

A number of studies show a strong relation between turbidity measurements and indicator bacteria densities (McSwain, 1977; Christensen, 2001; Rasmussen and Ziegler, 2003). Fries and others (2006) and Krometis and others (2007) reported that 34 to 42 percent of *E. coli* in surface-water samples were attached to particles in the water column. Gregory and Frick (2000) noted that fecal coliform bacteria densities in the Chattahoochee River were highest after rainstorms when the river was turbid. In addition, *E. coli* densities in water have been correlated to water temperature (Darakas, 2002). Darakas (2002) showed that *E. coli* survival consisted of a maintenance period (indicated by relatively stable densities over time) and decay phases (indicated by rapidly declining densities over time) that were temperature dependent. During the Darakas (2002) study, the period of maintenance was longest (13.6 days) at a water temperature of 10 degrees Celsius (°C) and shortest (0.5 days) at 37 °C. He and others (2007) showed that in southern California ponds, bottom sediments contained higher densities and greater survival of fecal indicator bacteria than did flowing water, a finding they attributed to higher water temperatures in the ponds. The findings of He and others (2007) may be relevant to *E. coli* densities at the Atlanta site in the current study, because the site is downstream from Bull Sluice Lake. In June 2005, a water temperature study by Georgia Power showed that water exiting Bull Sluice Lake is 4–6 °C higher than the water temperature of the Chattahoochee River entering the lake (Georgia Power, 2006).

Identifying anthropogenic sources of fecal coliform bacteria in surface water is hampered by natural sources of bacteria shed by warm-blooded wildlife. For example, Fujioka and others (1998) reported that the natural soil environment in subtropical areas, such as Guam, contained high densities of *E. coli* bacteria that entered streams during storm runoff. Streams that received this storm runoff consistently had *E. coli* densities that exceeded the USEPA single-sample beach criterion of 235 colonies/100 mL (table 1), even though anthropogenic activity was nonexistent in the upstream watersheds.

Methods of Study

The data collection methods used in this study followed the procedures and protocols published by the USGS in Wilde and others (2004), U.S. Geological Survey (2006), Wagner and others (2006), Anderson (2005), and Myers and others (2007); procedures published by the USEPA in U.S. Environmental Protection Agency (2000; 2003); and the procedures published by the American Public Health Association in Bordner (2005), Hall (2005), Meckes and Rice (2005), and Palmer (2005).

Collection, Processing, and Analysis of Water Samples

During the period of study, water samples were collected with varying frequency at the Norcross and Atlanta sites: 4 days per week (Monday–Thursday) from October 23, 2000, to October 1, 2001; daily from October 1, 2001, to February 7, 2002; and 3 days per week from February 10, 2002, to September 30, 2008. Methods approved by the USGS were used to collect water samples at each site (Myers, 2004). Employees and volunteers of the Upper Chattahoochee Riverkeeper organization collected water samples at the Atlanta site, and employees of the CRNRA collected water samples at the Norcross site. These samples were collected with a weighted yoke from bridges that spanned the river at each site. The yoke (made of polyvinyl chloride pipe) was designed to hold a sterile, narrow-mouth, 1-liter (L) polypropylene bottle. The goal while sampling at each site was to collect a single, vertically integrated sample at midchannel. Nitrile gloves were worn while collecting and handling each water sample.

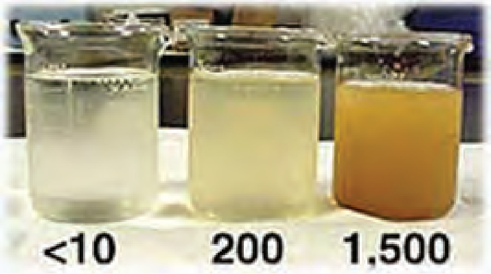
After collection, the samples were labeled with the station number and sample date and time, placed on ice, and transported to the microbiology laboratory at the USGS Georgia Water Science Center (GAWSC) in Atlanta, Georgia. The time between sample collection and the start of incubation was less than 8 hours and typically less than 4 hours. The USEPA requires that the time between sample collection and the start of incubation is not greater than 8 hours, and the laboratory cannot hold samples for more than 6 hours before the start of incubation (U.S. Environmental Protection Agency, 2000).

Water samples were analyzed for E. coli bacteria using the Colilert®-18 (Colilert) and Quanti-Tray®/2000 (Quanti-Tray) system manufactured by the IDEXX Corporation (IDEXX Laboratories, Inc., 2002a, b). The American Public Health Association (Palmer, 2005) and the USEPA (U.S. Environmental Protection Agency, 2003) have formally approved the Colilert method of analysis for quantifying total coliform and *E. coli* bacteria in drinking water and ambient water. The Colilert method is conceptually similar to the commonly used multiple tube method (Meckes and Rice, 2005) in which bacteria densities are determined statistically and expressed as a most probable number of colonies per 100 milliliters of water (MPN/100 mL). In the laboratory, 2 to 3 measured aliquots of sample were added to sterile de-ionized (DI) water to produce 100 milliliters (mL) of liquid. Aliquot volumes depended upon the turbidity of the sample (table 2). A nutrient packet containing nutrients and a chromagen was added to each dilution. The chromagen, which contains the compound 4-methyllumbelliferyl-β-D-glucuronide (MUG), reacts with enzymes released by *E. coli* bacteria causing cells in the incubation tray to fluoresce under ultraviolet light. Volume-weighted mean *E. coli* densities (MPN/100 mL) for the Colilert method were computed using equation 1–1 in appendix 1. Appendix 1 describes in detail the Colilert method of analyzing water samples for total coliform and *E. coli* bacteria in this study.

A number of researchers have concluded that the Colilert method is a better alternative to membrane filter methods because it has a slightly shorter incubation time, is more accurate (fewer false positives and false negatives), and is considered easier to use—especially by those untrained in microbiology—than membrane filter methods (Clark and others, 1991; Olson and others, 1991; Cowburn and others, 1994; Buckalew and others, 2006). In addition, these researchers concluded that *E. coli* densities determined using the Colilert method correlate well with fecal coliform densities determined using m-FC membrane filter methods and that sample handling time is shorter, thus reducing the potential for contamination. Aulenbach (2009) reports that the analytical precision for *E. coli* densities determined by the Colilert method ranged from 14 to 70 percent, slightly lower than the theoretical precision reported by IDEXX Laboratories of 17 to 94 percent (IDEXX Laboratories, Inc., 2002a, b).

During the first 6 months of the study, water samples from the Norcross and Atlanta sites were analyzed concurrently for *E. coli* bacteria using the Colilert method and HACH Corporation’s m-Coliblue24® membrane-filter method, and for fecal coliform bacteria using the membrane filter with m-FC agar method. These additional analyses were done to determine method comparability between Colilert and the two membrane filter methods. These samples were analyzed at a frequency such that a reasonable number of bacteria densities from multiple methods were distributed throughout streamflows that were typical of the wet and dry season in the Atlanta area. A detailed description of the membrane filter analyses is given in appendix 1.

The quality-assurance and quality-control (QA/QC) methods used during bacteria analyses were those recommended by the American Public Health Association for microbiological analysis (Bordner, 2005). Because most surfaces, including the human body, contain a broad spectrum of bacterial fauna, a number of steps were taken during sample preparation, collection, and processing to prevent or minimize contamination by foreign bacteria. A detailed description of the QA/QC methods used during the study is given in appendix 1.



Laboratory beakers showing three turbidity levels. (Photograph by Howard A. Perlman, USGS.)

Table 2. Volumes of river and sterile de-ionized water needed for dilutions at various river turbidity levels.

[FNU, formazin nephelometric units; mL, milliliter; X, always dilute at indicated turbidity level]

Turbidity (FNU)	Dilution ratio									
	1:2		1:10		1:100		1:1,000[a]		1:10,000[a]	
	Dilution volume									
	50 mL	50 mL	10 mL	90 mL	1 mL	99 mL	0.1 mL	99.9	0.01 mL	99.99
	River water	Sterile water	River water	Sterile water	River water	Sterile water	River water	Sterile water	River water	Sterile water
Less than 11	X		X				Not applicable at these turbidity levels			
11 to 40	X		X		X					
41 to 100			X		X		X			
Greater than 100	Not applicable at these turbidity levels				X		X		X	

[a] High turbidity samples require 1:1,000 and 1:10,000 dilutions. For the 1:1,000 dilution, 10 mL of river water is added to 90 mL of sterile water (subsample A) and 1 mL of subsample A is added to 99 mL of sterile water (equal to 0.1 mL of river water), which becomes the sample to be analyzed. For the 1:10,000 dilution, 10 mL of subsample A is added to 90 mL of sterile water (subsample B) and 1 mL of subsample B is added to 99 mL of sterile water, which then becomes the sample (equal to 0.01 mL of river water) to be analyzed.

Collection of Streamflow, Turbidity, and Meteorological Data

Streamflow was measured and processed in accordance with methods and techniques approved by the USGS Office of Surface Water and published in Buchanan and Somers (1969), Rantz and others (1982a, b), and Kennedy (1984). Streamflow and gage height data for the six streamgaging stations were obtained from the USGS National Water Information System (NWIS) database (online at *http://waterdata.usgs.gov/nwis*).

Turbidity data were collected and processed following the protocols published in Letterman (2005), Wagner and others (2006), and Anderson (2005). Water samples were measured for turbidity in the laboratory with a HACH 2100P turbidimeter using the procedures outlined in Letterman (2005). Beginning on May 24, 2002, at Norcross, and July 26, 2002, at Atlanta, water temperature and turbidity were continuously measured instream with YSI 6820 series water-quality sondes. Turbidity was measured with the YSI model 6136 turbidity probe. The water-quality sondes were serviced bi-weekly or as needed within that time period using protocols outlined in Wagner and others (2006). Data from these YSI sondes were uploaded to the NWIS database at the GAWSC in Atlanta on an hourly basis from the Norcross site and every 4 hours from the Atlanta site.

Meteorological data were obtained from three different sources. Average monthly rainfall for the study period and long-term (1971–2001) average monthly rainfall data for north-central Georgia were supplied by the National Climatologic Data Center (U.S. Department of Commerce at *http://www7.ncdc.noaa.gov/*, accessed March 9, 2011). Daily rainfall totals and daily maximum and minimum air temperature used in this report were from the meteorological station at the Atlanta Athletic Club, Johns Creek, Fulton County, Ga., which is operated by the University of Georgia as part of the Georgia Automated Environmental Monitoring Network (*http://www.georgiaweather.net/*, accessed February 23, 2008), and the meteorological station at the Norcross site.

Data and Statistical Analysis

The *E. coli* bacteria densities, and turbidity and streamflow measurements from the two sites on the Chattahoochee River are described using summary statistics, exploratory methods, and qualitative groupings. Regression analyses are described in an effort to find regression equations to estimate *E. coli* bacteria density in real time. The Method of Variance Extension (MOVE, Helsel and Hirsch, 1992), which computes a line of organic correlation (LOC) was used in this report to estimate missing data. For example, turbidity was not continuously measured instream at both sites until the spring/summer of 2002; however, laboratory turbidity was measured in samples collected since the beginning of the BacteriALERT program (October 2000). Therefore, the instream turbidity record was extended back to the start of BacteriALERT by computing the LOC with the instream turbidity measurements as the response variable and laboratory-measured turbidity as the explanatory variable.

Qualitative data describing season and streamflow characteristics such as streamflow event (EVENT) and streamflow condition (HCOND) were developed from data collected at each site. Water samples were grouped by the season in which they were collected. In the Atlanta area, the warm season corresponds to the time between April 16 and October 15 and the cool season corresponds to the time between October 16 and April 15. Streamflow event (EVENT) is defined as either dry-weather flow or stormflow. Dry-weather flow is streamflow generated by water releases from a dam during dry weather. Because of minimum flow requirements downstream from Atlanta, at least 600 ft^3/s of water is released from Buford Dam and 750 ft^3/s of water is released from Morgan Falls Dam at all times (Georgia Power, 2004b). Stormflow is streamflow generated by surface runoff during rainfall.

Streamflow condition (HCOND) describes the stream stage within a streamflow event in relation to the altitude of the stream surface above a known datum, commonly called gage height or stream stage. The streamflow conditions used in this report are: (1) stable flow, low stage (StableLow); (2) stable flow, normal stage (StableNorm); (3) stable flow, high stage (StableHigh); (4) rising stage (RisingQ); (5) falling stage (FallingQ); and (6) peak stage (PeakQ; table 3). Stable flow is streamflow that is relatively constant over a specified time period and is defined for this report as streamflow that is neither increasing nor decreasing by more than 15 percent within the 60 minutes before a water sample is collected. Figure 5 shows a hypothetical hydrograph identifying the six streamflow conditions used in the report. Details on how the HCOND parameters were computed are given in appendix 2.

Streamflow measurements in 15-minute intervals from the six gaging stations were used to assign EVENT and HCOND values to streamflow measurements at the Norcross and Atlanta sampling sites (table 3). Streamflow immediately downstream from Buford Dam is generated only by water released from Lake Lanier. Because Lake Lanier is so large, the water released does not reflect stormflow in a manner analogous to stormflow from tributaries; therefore, those water releases established the reference for nonstorm-related streamflow (dry-weather flow) in the Chattahoochee River. At times, stormflow from tributaries to the Chattahoochee River coincided and mixed with water released from Buford Dam for power generation and made it difficult to assign a streamflow event to samples and measurements at both sites.

Statistical analyses attempt to estimate an unknown and immeasurable parameter from an identified population by taking a sample from the population. The sample, if random and unbiased, is assumed to mirror the statistical properties of the population such that any statistical measure of the sample is also the statistical measure of the population (Ott, 1988). The equations used in this report for statistical summaries are those published in Ott (1988) or Helsel and Hirsch (1992).

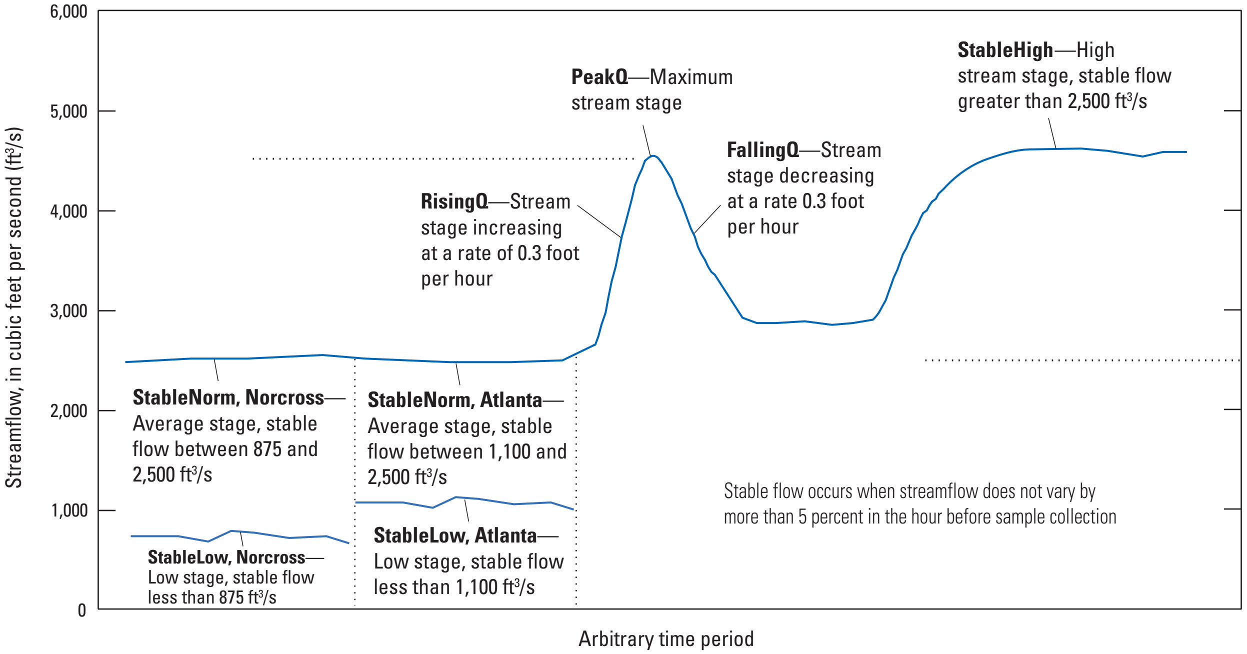


Figure 5. Hypothetical hydrograph showing the six streamflow conditions assigned to water samples collected from the Chattahoochee River during the study period, October 23, 2000, through September 30, 2008 (table 3).

Table 3. List of indicator variables (categorical or qualitative parameters) used in the multiple regression analyses of data from the Chattahoochee River near Norcross, Georgia (USGS station number 02335000), and at Atlanta, Georgia (USGS station number 02336000).

[ft^3/s, cubic foot per second]

Variable	Variable description	Value	Definition	
			Chattahoochee River near Norcross	Chattahoochee River at Atlanta
Streamflow regime (EVENT)	Dry-weather flow	0	Water releases from Buford Dam or Morgan Falls Dam	
	Stormflow	1	0.75 inches of rain within previous 48 hours Increasing[a] stream stage at Suwanee Creek at Suwanee, GA (USGS station 02334885)	Increasing[a] stream stage at Rottenwood Creek near Smyrna, GA (USGS station 02335910) or Sope Creek near Marietta, GA (USGS station 02335870)
Streamflow condition (HCOND; fig. 3)	StableLow	0	Low stage, stable[b] streamflow less than or equal to 875 ft^3/s	Low stage, stable streamflow less than or equal to 1,100 ft^3/s
	StableNorm	1	Average stage, stable streamflow between 875 and 2,500 ft^3/s	Average stage, stable streamflow between 1,100 and 2,500 ft^3/s
	StableHigh	2	High stage, stable streamflow greater than 2,500 ft^3/s	High stage, stable streamflow greater than 2,500 ft^3/s
	RisingQ	3	Stream stage increasing at a rate greater than 5 percent per hour	
	FallingQ	4	Stream stage decreasing at a rate greater than 5 percent per hour	
	PeakQ	5	Maximum stream stage between rising and falling stages	
Season	Warm	1	Warm season: April 16 to October 15	
	Cool	2	Cool season: October 16 to April 15	

[a] Absolute gage height greater than 0.3 foot above datum.

[b] Streamflow that varies by less than 5 percent in the 1-hour period before sample collection.

Exceedance probabilities, which are commonly used in hydrology to determine streamflow duration curves, were calculated for *E. coli* bacteria density and turbidity measurements. These curves, however, are presented in this report as non-exceedance probabilities (1-exceedance probability). The probabilities were calculated with an S-PLUS function using Cunnane's formula (Cunnane, 1978; Helsel and Hirsch, 1992; TIBCO Software, Inc., 2008).

Statistical inferences (hypothesis testing) using nonparametric methods are used to identify statistically significant differences in *E. coli* bacteria and turbidity measurements between the Norcross and Atlanta sites and among seasons, and stream characteristics at each site. The two primary tests for statistical inference used in this report are the Wilcoxon Signed-Ranks test (nonparametric two-sample t-test), which is used to test the similarity in the distribution of data between two groups of samples and Spearman's rank correlation, the nonparametric analog to Pearson's correlation analysis. Kendall's tau and the Sen slope estimate were used to determine if statistically significant time series trends existed in *E. coli* density during the study period (Helsel and Hirsch, 1992, p. 266).

Regression analysis is a statistical method for identifying and modeling the relations between two or more variables (Montgomery and others, 2006, p. 1). Ordinary least-squares regression or OLS, which includes simple linear and multiple linear regression, is probably the regression method most commonly used in water-resources studies. Although a regression model does not infer a cause and effect relation between variables, it can help to confirm a cause and effect relation and should not be the only basis for inferring that relation (Montgomery and others, 2006, p. 39–40). The methods used for basic linear regression analyses are described in most basic statistics textbooks (Ott, 1988). The linear and logistic regression methods used specifically for this report are described by Helsel and Hirsch (1992), Harrell (2001), and Montgomery and others (2006), and computed using S-PLUS® software (TIBCO Software, Inc., 2008).

In this report, the relations among streamflow characteristics, climate and meteorological measurements, and stream properties such as turbidity, water temperature, and *E. coli* bacteria densities in water samples were investigated using simple and multiple linear regression, and logistic regression. The resulting regression equations were evaluated for their predictive power using a variety of diagnostic tools and residual analyses, and validated with data collected between October 1, 2008, and September 30, 2009. The methods and steps used to develop and validate the regression equations are described in detail in appendix 2 to provide future researchers with enough information to maintain continuity with the data generated during the current study period and make meaningful comparisons with the predictive equations developed in this report.

Escherichia coli Bacteria Density in Relation to Turbidity, Streamflow Characteristics, and Season

Knowledge of *E. coli* densities and turbidity levels under different streamflow characteristics and seasons is important in selecting and validating the regression equations chosen to predict *E. coli* densities at both sampling sites. This section, therefore, describes study period and seasonal *E. coli* densities and turbidity values under varying streamflow characteristics such as stream stage during dry-weather flow or stormflow in the Chattahoochee River at the Norcross and Atlanta sites.

Description and Statistical Analysis of *Escherichia coli* Bacteria Density, Turbidity, and Streamflow Characteristics

Between October 23, 2000, and September 30, 2008, 1,417 water samples were collected at the Norcross site and 1,407 water samples were collected at the Atlanta site and analyzed in the laboratory for *E. coli* density and turbidity. Analytical precision and 95-percent confidence intervals (CI) for *E. coli* bacteria densities were calculated for a subset of the water samples analyzed from both sites. Also presented in this section is a comparison between laboratory and instream measured turbidity values. This comparison was used to impute instream turbidity values for samples collected before instream turbidity was measured at both sites. In addition, summary statistics provide useful descriptions of *E. coli* densities and turbidity values by season and under a variety of streamflow characteristics.

Quality Control Results for *Escherichia coli* Bacteria Analyses

At both sampling sites during the study period, two to three dilutions from each water sample were analyzed for *E. coli* density by Colilert in order to compute volume-weighted mean *E. coli* densities (eq. 1–1, appendix 1). The *E. coli* densities for the individual dilutions also were useful for calculating precision and confidence intervals for the Colilert method. Analytical precision and the 95-percent CI around the geometric mean *E. coli* density for each sample were computed using three types of laboratory replicates: (1) dilutions were treated as replicates by normalizing the *E. coli* densities in each dilution to 100 mL; (2) duplicates that were analyzed at only one dilution (1:2 dilution consisting of 50 mL sample and 50 mL of sterile DI water); and (3) duplicates in which two or three dilutions were analyzed for each subsample split from a sample.

The relative percent difference (RPD), precision, and 95-percent CI for laboratory replicates that represent the breadth of volume-weighted mean *E. coli* densities measured at both sites are presented in table 4. Among the dilution replicates, analytical precision ranged from 1.3 to 17 percent. The greatest precisions typically were seen in samples with mean *E. coli* densities greater than 3,000 MPN/100 mL. In addition, the 95-percent CIs, as a percentage of the geometric mean densities, ranged from 2.7 to 103 percent. About 50 percent of the Colilert analyses listed in table 4 are less than the lowest 95-percent CIs (10 percent) reported by Aulenbach (2009), and 72 percent are lower than the smallest theoretical 95-percent CIs (14 percent) provided by the Colilert manufacturer (IDEXX Laboratories, Inc., 2002a). One analysis with the smallest volume-weighted mean *E. coli* density (14 MPN/100 mL) listed in table 4, exceeded the 95-percent CI reported by Aulenbach (2009) and the Colilert manufacturer. The RPD for duplicates analyzed at a 1:2 dilution ranged from 3.6 to 35 percent.

Table 4. *Escherichia coli* bacteria density in duplicate analyses and various analytical dilutions of water samples collected from the Chattahoochee River near Norcross, Georgia (USGS station 02335000), and at Atlanta, Georgia (USGS station 02336000), October 23, 2000, through September 2008.—Continued

[MPN/100 mL, most probable number of colonies per 100 milliliters of water; RPD, relative percent difference; relative precision, sometimes called the coefficient of variation, is computed as the standard deviation divided by the mean ×100, in log base 10 units; *E. coli, Escherichia coli* bacteria; —, not applicable]

Site	Date	Duplicates[a]				Dilutions[f]						
		Density as MPN/100 mL				Density as MPN/100 mL					Percent	
		1[b]	2[b]	Mean[c,d]	RPD[e]	1[c]	2[c]	3[c]	Geometric mean	95-percent confidence interval[g]	Relative precision[h]	95-percent confidence interval[i] (±)
Norcross	3/20/2003	17	15	*20*	13	—	—	—	—	—	—	—
Norcross	3/26/2003	116	122	*120*	5.0	—	—	—	—	—	—	—
Norcross	4/16/2003	39	29	*30*	29	—	—	—	—	—	—	—
Norcross	7/9/2003	109	80	*90*	31	—	—	—	—	—	—	—
Norcross	3/3/2004	17	24	*20*	35	—	—	—	—	—	—	—
Norcross	3/28/2007	58	36	*50*	44	—	—	—	—	—	—	—
Norcross	12/20/2007	355	243	*300*	37	—	—	—	—	—	—	—
Atlanta	10/1/2003	126	102	*110*	22	—	—	—	—	—	—	—
Atlanta	1/7/2004	380	370	*370*	2.7	—	—	—	—	—	—	—
Atlanta	4/28/2004	158	182	*170*	14	—	—	—	—	—	—	—
Norcross	6/6/2001	90	70	80	25	80	60	50	80	50–110	8.2	8.7
Atlanta	1/22/2001	580	422	500	32	501	528	—	500	320–790	3	7.4
Atlanta	5/31/2001	135	140	140	3.6	140	110	—	130	90–170	3.9	6.3
Atlanta	6/14/2001	76	84	80	10	80	40	—	80	60–120	8.2	8.7
Norcross	11/27/2000	—	—	40	—	43	31	100	40	10–160	17	42
Norcross	1/11/2001	—	—	14	—	15	10	—	10	<1–160	11	103
Norcross	8/16/2001	—	—	170	—	158	243	100	190	110–310	9.1	9.6
Norcross	6/7/2002	—	—	2,000	—	1,935	2,780	6,300	3,200	720–15,000	7.5	19
Norcross	11/19/2003	—	—	9,500	—	9,804	6,830	10,000	8,700	5,000–15,000	2.4	5.9
Norcross	4/14/2004	—	—	840	—	821	958	500	730	320–1,700	5.2	13
Atlanta	3/15/2001	—	—	3,000	—	3,106	2,700	3,300	3,000	2,400–3,900	1.3	3.1

Table 4. *Escherichia coli* bacteria density in duplicate analyses and various analytical dilutions of water samples collected from the Chattahoochee River near Norcross, Georgia (USGS station 02335000), and at Atlanta, Georgia (USGS station 02336000), October 23, 2000, through September 2008.—Continued

[MPN/100 mL, most probable number of colonies per 100 milliliters of water; RPD, relative percent difference; relative precision, sometimes called the coefficient of variation, is computed as the standard deviation divided by the mean ×100, in log base 10 units; *E. coli, Escherichia coli* bacteria; —, not applicable]

Site	Date	Duplicates[a]				Dilutions[f]						
		Density as MPN/100 mL				Density as MPN/100 mL					Percent	
		1[b]	2[b]	Mean[c,d]	RPD[e]	1[c]	2[c]	3[c]	Geo-metric mean	95-percent confidence interval[g]	Relative preci-sion[h]	95-percent confidence interval[i] (±)
Atlanta	1/6/2002	—	—	2,200	—	2,240	1,951	1,210	1,700	780–3,900	4.3	11
Atlanta	5/4/2002	—	—	8,600	—	8,664	8,200	10,900	9,200	6,300–13,000	1.7	4.1
Atlanta	1/22/2003	—	—	40	—	46	30	100	50	20–150	16	27
Atlanta	11/19/2003	—	—	5,900	—	5,790	6,570	—	6,400	5,100–8,200	1.1	2.7
Atlanta	5/19/2004	—	—	1,800	—	1,633	2,851	3,000	2,400	1,000–5,600	4.3	11
Atlanta	11/30/2006	—	—	2,600	—	2,827	1,378	—	1,800	720–4,700	5	13
Atlanta	2/1/2007	—	—	1,000	—	977	1,246	—	1,100	240–5,200	2.5	22

[a] These are not duplicate samples from the sampling site, but duplicate aliquots from the sample bottle.

[b] Volume-weighted mean for all dilutions. The mean is weighted by the aliquot volumes (see equation 1).

[c] Volume-weighted mean for the duplicate densities (see equation 1).

[d] Italized means indicate duplicates were only analyzed at one dilution (50 milliliters of sample added to 50 mL of sterile de-ionized water).

[e] Relative percent difference is calculated as the absolute value of the difference in log base 10 density between the two duplicates divided by the geometric mean density of the duplicates ×100.

[f] Two to three dilutions were typically used to compute the mean *E. coli* density for each sample collected. Each dilution consisted of an aliquot of water from the sample bottle (0.1 to 75 milliliters) that was added to an amount of sterile de-ionized water for a total volume of 100 milliliters.

[g] $\overline{X} \pm \left(t \times s\right) / \sqrt{n}$

where,

$\overline{X}$ is the geometric mean;

t is the 95 percent t score from t distribution for n;

s is the geometric standard deviation;

n is the number of duplicates or dilution (Baker, 2005)

[h] $s / \overline{X} \times 100$

where,

$\overline{X}$ is the geometric mean

s is the geometric standard deviation (Baker, 2005)

[i] $\pm \left(\left(t \times s\right) / \sqrt{n}\right) / \overline{X} \times 100$

where,

$\overline{X}$ is the geometric mean;

t is the 95 percent t score from t distribution for n;

s is the geometric standard deviation;

n is the number of duplicates or dilution (Baker, 2005)

Escherichia coli Bacteria Density and Turbidity in Relation to Streamflow Characteristics

At both sites, turbidity and *E. coli* density were greater during stormflow than during dry-weather flow. At Norcross, median *E. coli* density was about 10 times greater and turbidity was nearly 8 times greater in stormflow than in dry-weather-flow samples (table 5). At the Atlanta site, the median *E. coli* density was about 10 times greater and median turbidity was 4 times greater in stormflow than in dry-weather-flow samples from Atlanta. During dry weather, the daily peak discharges from Buford Dam and Morgan Falls Dam (commonly about 5,000 and 2,000 ft^3/s, respectively) were substantially greater than streamflow in tributaries upstream from the Norcross and Atlanta sites. These water releases may remove sediment from the streambank through active bank erosion or, more likely, from the erosion of bank material where the streambank has collapsed or slipped into the river channel (Leopold, 1994).

In contrast, streams that are tributary to the Chattahoochee River contribute a large amount of water during stormflow because of urbanization, resulting in concomitant increases in suspended solids to the river and high turbidity measurements (Landers and others, 2007). Therefore, the major source of high turbidity and high *E. coli* densities to the Chattahoochee River, apart from sewage spills or releases, is stormflow from Chattahoochee River tributaries. In many instances, however, when the amount of water released by Buford Dam exceeds about 5,000 ft^3/s, the turbidity and *E. coli* bacteria contribution

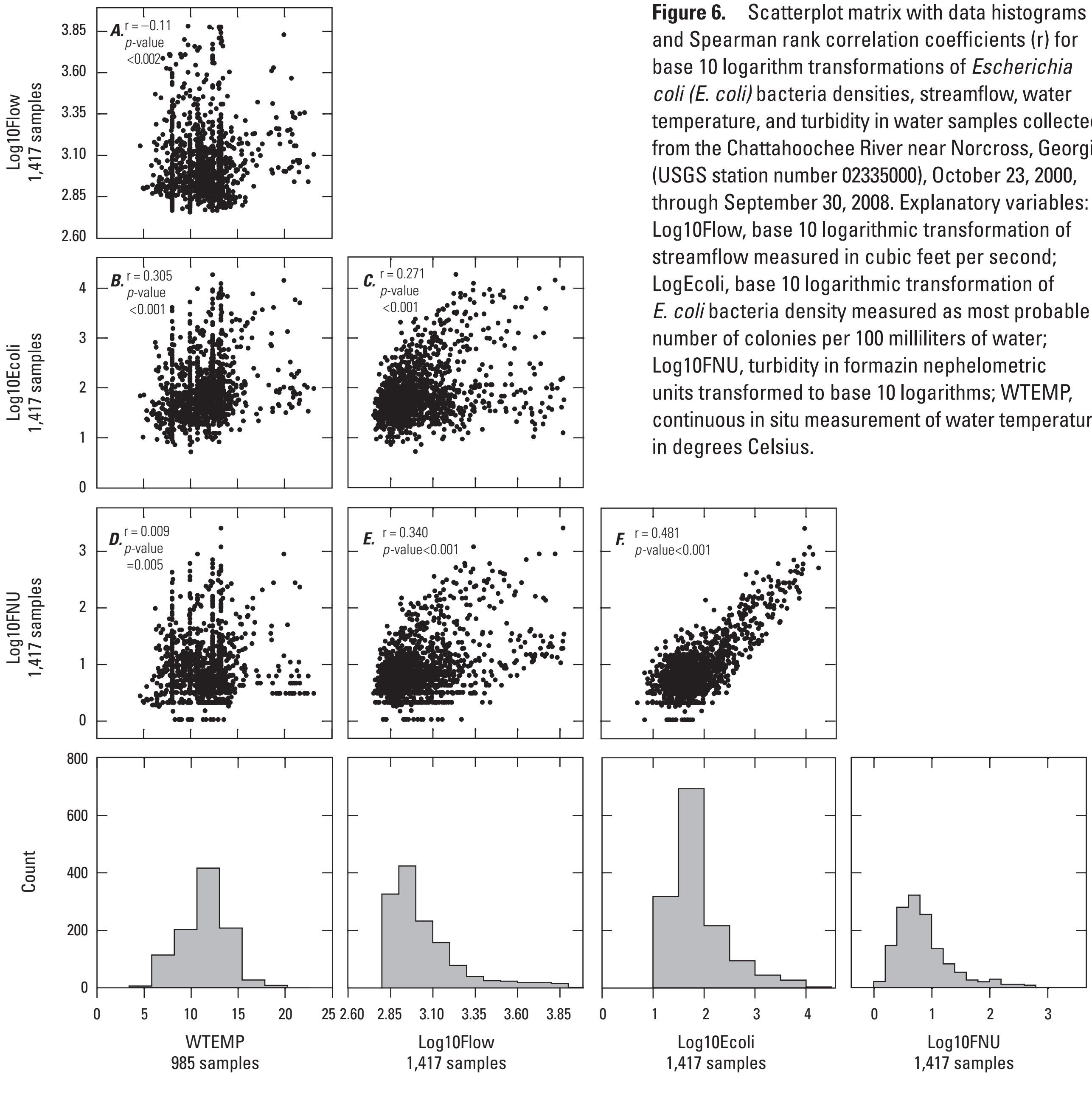


Figure 6. Scatterplot matrix with data histograms and Spearman rank correlation coefficients (r) for base 10 logarithm transformations of *Escherichia coli (E. coli)* bacteria densities, streamflow, water temperature, and turbidity in water samples collected from the Chattahoochee River near Norcross, Georgia (USGS station number 02335000), October 23, 2000, through September 30, 2008. Explanatory variables: Log10Flow, base 10 logarithmic transformation of streamflow measured in cubic feet per second; LogEcoli, base 10 logarithmic transformation of *E. coli* bacteria density measured as most probable number of colonies per 100 milliliters of water; Log10FNU, turbidity in formazin nephelometric units transformed to base 10 logarithms; WTEMP, continuous in situ measurement of water temperature, in degrees Celsius.

by tributaries during small storms is tempered at both sampling sites by the dilution effect of the water released by the dam. This dilution, however, may affect the Atlanta site to a lesser extant than the Norcross site because the drainage area between the Norcross and Atlanta sites is three times greater than the drainage area between Buford Dam and the Norcross site.

Matrix scatterplots and histograms show relations among the log base 10 ($\log_{10}$) transformed *E. coli* (log10Ecoli) in water samples, $\log_{10}$ transformed turbidity (log10FNU), water temperature, and $\log_{10}$ transformed streamflow (log10Flow) measurements at the Norcross (fig. 6) and Atlanta (fig. 7) sites. Figure 6*E* shows a strong relation between log10FNU and log10Ecoli. The scatterplots in figures 6*F* and 6*C*, respectively, show bimodal relations when log10Flow is plotted against log10FNU and log10Ecoli density measurements. The bimodal character in those relations indicate the distributions of turbidity and *E. coli* measurements were different during dry-weather flow and stormflow. In addition, linear relations between water temperature and log10Flow, log10FNU, or log10Ecoli density were not apparent at Norcross (figs. 6*A*, *B*, *D*). In contrast, turbidity and *E. coli* density were related to streamflow at the Atlanta site, but without the bimodal response seen at the Norcross site (figs. 7*C*, *F*). The *E. coli* density and turbidity at the Atlanta site were linearly related (fig. 7*E*). Although water temperature was clearly not related to streamflow and turbidity (figs. 7*B*, *D*), it was related to *E. coli* density at Atlanta (fig. 7*A*).

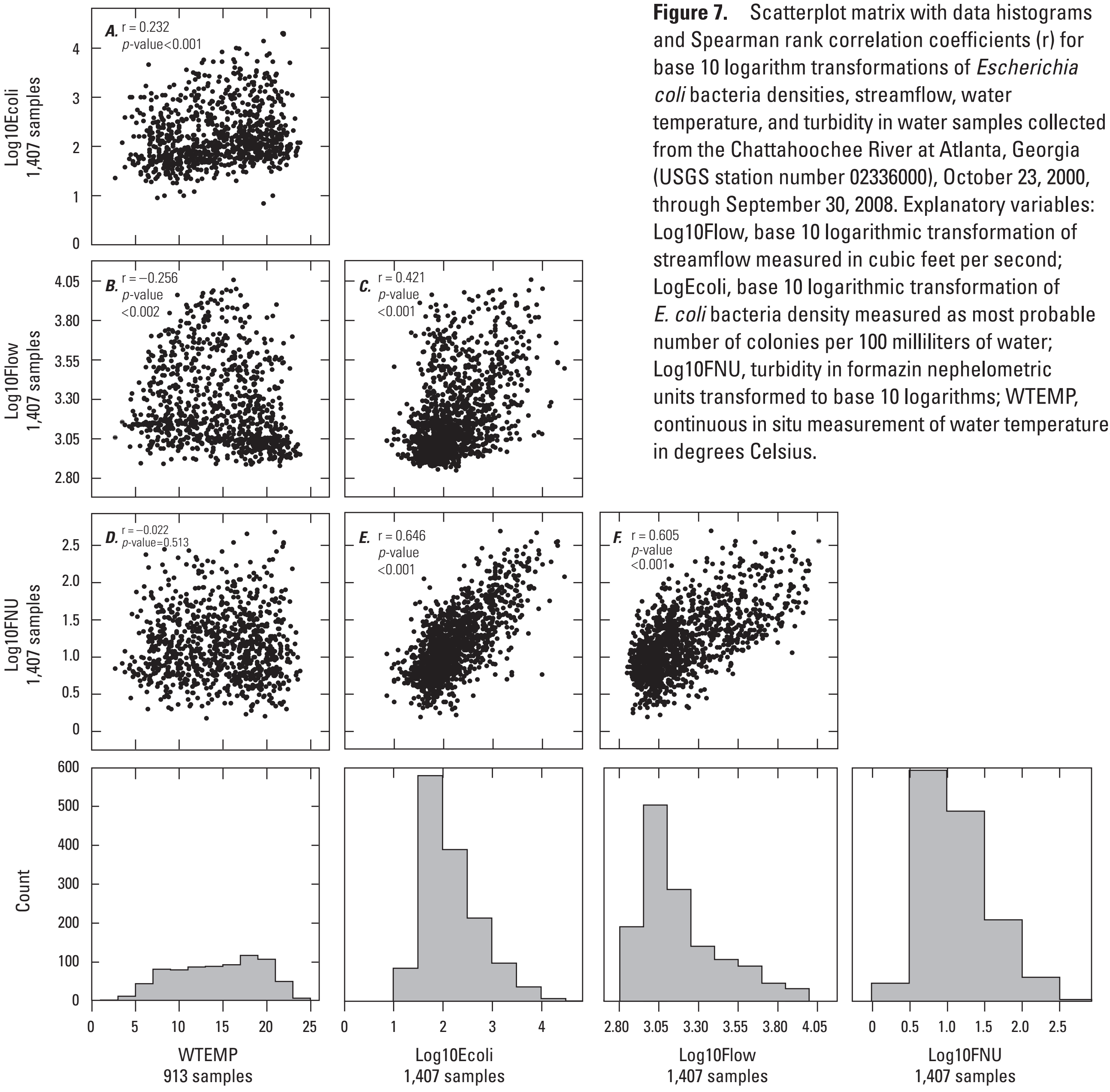


Figure 7. Scatterplot matrix with data histograms and Spearman rank correlation coefficients (r) for base 10 logarithm transformations of *Escherichia coli* bacteria densities, streamflow, water temperature, and turbidity in water samples collected from the Chattahoochee River at Atlanta, Georgia (USGS station number 02336000), October 23, 2000, through September 30, 2008. Explanatory variables: Log10Flow, base 10 logarithmic transformation of streamflow measured in cubic feet per second; LogEcoli, base 10 logarithmic transformation of *E. coli* bacteria density measured as most probable number of colonies per 100 milliliters of water; Log10FNU, turbidity in formazin nephelometric units transformed to base 10 logarithms; WTEMP, continuous in situ measurement of water temperature, in degrees Celsius.

Table 5. Summary statistics by streamflow regime for mean *Escherichia coli* bacteria density and turbidity measurements at sampling sites on the Chattahoochee River near Norcross, Georgia (USGS station number 02335000), and at Atlanta, Georgia (USGS station number 02336000), October 23, 2000, through September 30, 2008.

[IQR, interquartile range, defined as the difference between the 75th and 25th percentile values; dry-weather flow, streamflow generated by water released from Buford Dam or Morgan Falls Dam during dry weather; stormflow, streamflow from runoff associated with rainfall as defined in table 3; *E. coli, Escherichia coli* bacteria; FNU, formazin nephelometric unit; MPN/100 mL, most probable number of colonies per 100 milliliters of water]

Statistic	Chattahoochee River near Norcross						Chattahoochee River at Atlanta					
	All data		Streamflow regime, EVENT				All data		Streamflow regime, EVENT			
			Dry-weather flow		Stormflow				Dry-weather flow		Stormflow	
	Turbidity (FNU)	Mean *E. coli* density (MPN/100 mL)	Turbidity (FNU)	Mean *E. coli* density (MPN/100 mL)	Turbidity (FNU)	Mean *E. coli* density (MPN/100 mL)	Turbidity (FNU)	Mean *E. coli* density (MPN/100 mL)	Turbidity (FNU)	Mean *E. coli* density (MPN/100 mL)	Turbidity (FNU)	Mean *E. coli* density (MPN/100 mL)
Number of samples	1,417	1,417	1,172	1,172	245	245	1,407	1,407	1,070	1,070	339	339
Maximum	2,690	18,000	39	1,200	2,700 [a]	18,000	480	28,000	480	9,800	450	28,000
Minimum	.9	<10	.9	<10	5.0	60	1.5	<10	1.5	<10	3.7	450
Mean	23	300	6.3	65	104	1,420	26	515	15	150	60	1,670
Geometric mean	7.3	70	5	50	45	600	14	150	10	90	37	830
Median	5.7	60	4.6	50	36	530	12	110	9.1	80	35	810
Geometric standard deviation	3.1	3.7	1.9	2.2	3.3	3.5	2.7	3.7	2.2	2.3	2.6	3
Coefficient of geometric variation, percent (gCOV) [b]	57	31	40	20	31	20	38	26	34	19	26	16
Interquartile range (IQR)	7.5	80	4.8	50	92	1,100	18	250	10	90	52	1,350

[a] Sample diluted 1:3 in the laboratory.

[b] The geometric standard deviation, in log base 10 units, divided by geometric mean, in log base 10 units, ×100.

Statistical Analysis of *Escherichia coli* Density and Turbidity by Streamflow Characteristics and Season

Statistical profiles by streamflow event were developed for *E. coli* density and turbidity at both sites (table 5). Statistical profiles were developed for *E. coli* density and turbidity by dry-weather flow and stormflow in both seasons at the Norcross (table 6) and the Atlanta (table 7) sites. In addition, *E. coli* density and turbidity values at the Norcross site were compared by streamflow characteristic and season to those at the Atlanta site. Appendix 2 describes methods used to develop statistical profiles at the two sites.

The monthly distribution of *E. coli* bacteria densities and turbidity at the Norcross and Atlanta sites during the study period shows that the median *E. coli* density (fig. 8*A*) and median turbidity (fig. 8*B*) follow a semiannual cycle indicating seasonal trends at both sites. Turbidity-adjusted median *E. coli* densities and streamflow-adjusted median turbidity measurements aggregated by month for both sites were analyzed for seasonal trends and found to have statistically significant seasonality (Seasonal Kendall test, *p*-values less than 0.001; Helsel and Hirsch, 1992; TIBCO Software, Inc., 2008). At both sites, the highest median *E. coli* densities were typically seen during June and July, and were lowest during February and March at Norcross and January through April at Atlanta (fig. 8*A*). At the Norcross site, median turbidity values were lowest in April through June and August through September, and highest in November and December during the study period. In contrast, the median turbidity measurements at Atlanta were highest in July and August and lowest in November and December. Typically, the median monthly turbidity values and *E. coli* density at Atlanta were about twice those at Norcross (fig. 8*B*). Because a seasonal trend was apparent at both sites, water samples were parsed into two distinct seasons typical of the Atlanta metropolitan area: a warm season between April 16 and October 15 and a cool season between October 16 and April 15.

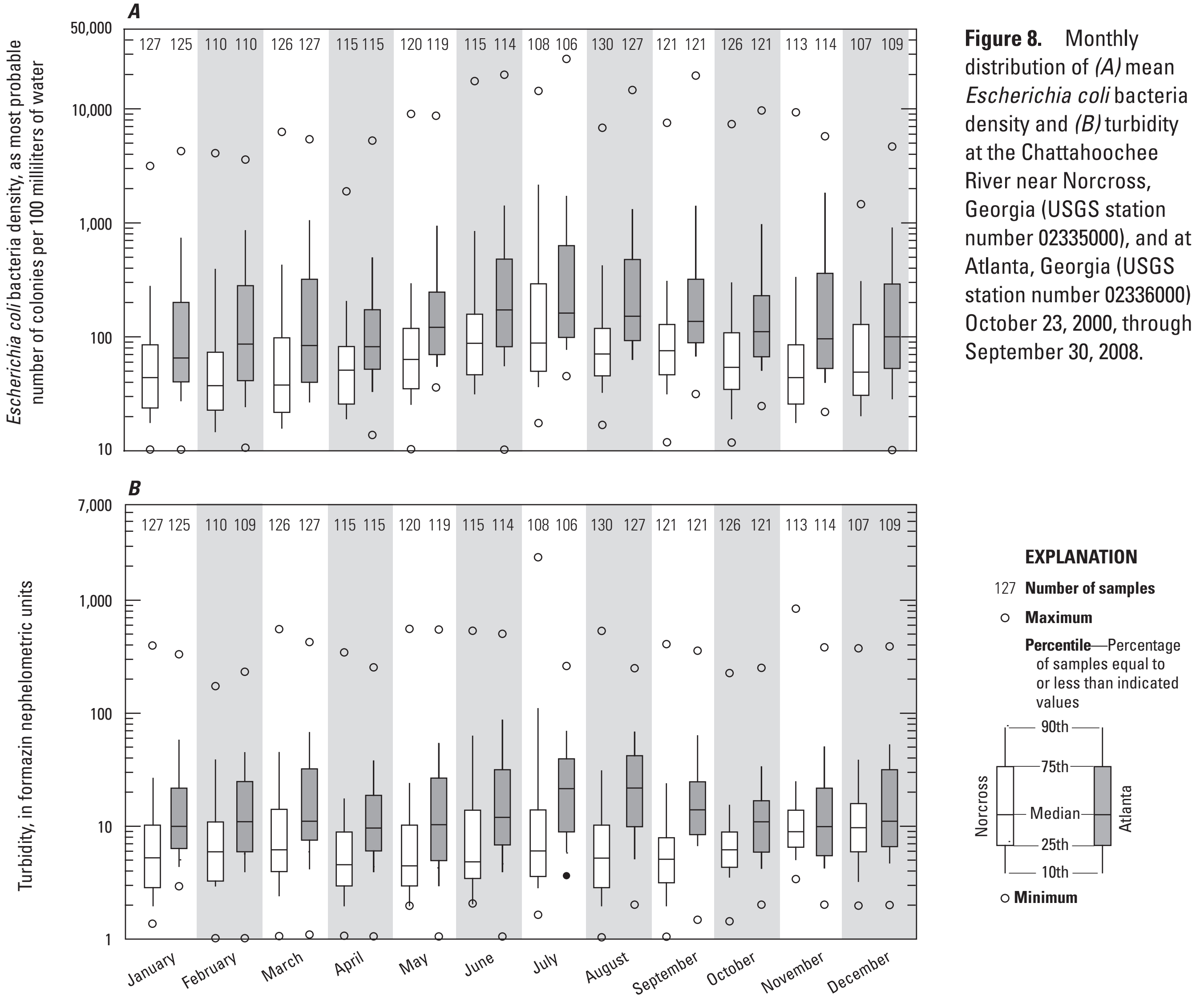


Figure 8. Monthly distribution of *(A)* mean *Escherichia coli* bacteria density and *(B)* turbidity at the Chattahoochee River near Norcross, Georgia (USGS station number 02335000), and at Atlanta, Georgia (USGS station number 02336000), October 23, 2000, through September 30, 2008.

Table 6. Summary statistics for seasonal mean *Escherichia coli* bacteria density and turbidity measurements by streamflow regime for sampling sites on the Chattahoochee River near Norcross, Georgia (USGS station number 02335000), October 23, 2000, through September 30, 2008.

[Dry-weather flow, streamflow generated by water released from Buford Dam during dry weather; stormflow, streamflow from runoff associated with rainfall as defined in table 3; IQR, interquartile range, defined as the difference between the 75th and 25th percentile values; *E. coli, Escherichia coli* bacteria; FNU, formazin nephelometric unit; MPN/100 mL, most probable number of colonies per 100 milliliters of water]

Statistic	Warm season (April 16 to October 15)						Cool season (October 16 to April 15)					
	All data		Streamflow regime, EVENT				All data		Streamflow regime, EVENT			
			Dry-weather flow		Stormflow				Dry-weather flow		Stormflow	
	Turbidity (FNU)	Mean *E. coli* density (MPN/100 mL)	Turbidity (FNU)	Mean *E. coli* density (MPN/100 mL)	Turbidity (FNU)	Mean *E. coli* density (MPN/100 mL)	Turbidity (FNU)	Mean *E. coli* density (MPN/100 mL)	Turbidity (FNU)	Mean *E. coli* density (MPN/100 mL)	Turbidity (FNU)	Mean *E. coli* density (MPN/100 mL)
Number of samples	712	712	587	587	125	125	705	705	585	585	120	120
Maximum	2,690	18,000	36	1,200	2,690	18,000	840	9,500	39	940	840	9,500
Minimum	1.0	1	1.0	1	5.0	60	.9	1	.9	1	5.0	60
Mean	25	390	5.3	80	116	1,850	22	209	7.4	50	90	970
Geometric mean	6.4	90	4.3	60	43	730	8.4	60	5.9	40	47	490
Median	4.5	70	4.0	60	32	640	6.9	40	5.8	40	38	430
Geometric standard deviation	3.2	3.6	1.8	2	3.6	3.9	3.0	3.6	1.9	2.1	3.1	3.0
Coefficient of geometric variation (gCOV)[a]	63	28	40	17	34	21	52	31	36	20	29	18
Interquartile range (IQR)	6.7	90	2.9	55	93	1,780	8.1	62	5.2	33	90	710

[a] The geometric standard deviation, in log base 10 units, divided by the geometric mean, in log base 10 units ×100.

Table 7. Summary statistics for seasonal mean *Escherichia coli* bacteria density and turbidity measurements by streamflow regime for the Chattahoochee River at Atlanta, Georgia (USGS station number 02336000), October 23, 2000, through September 30, 2008.

[Warm season, the period between April 16 and October 15; cool season, period between October 16 and April 15; dry-weather flow, streamflow generated by water released from Morgan Falls Dam during dry weather; stormflow, streamflow associated with runoff associated with rainfall as defined in table 3; E. coli, *Escherichia coli* bacteria; FNU, formazin nephelometric unit; MPN/100 mL, most probable number of colonies per 100 milliliters of water; IQR, interquartile range defined as the difference between the 75th and 25th percentile value]

Statistic	Warm season (April 16 to October 15)						Cool season (October 16 to April 15)					
	All data		Streamflow regime, EVENT				All data		Streamflow regime, EVENT			
			Dry-weather flow		Stormflow				Dry-weather flow		Stormflow	
	Turbidity (FNU)	Mean *E. coli* density (MPN/100 mL)	Turbidity (FNU)	Mean *E. coli* density (MPN/100 mL)	Turbidity (FNU)	Mean *E. coli* density (MPN/100 mL)	Turbidity (FNU)	Mean E. coli density (MPN/100 mL)	Turbidity (FNU)	Mean *E. coli* density (MPN/100 mL)	Turbidity (FNU)	Mean *E. coli* density (MPN/100 mL)
Number of samples	706	706	530	530	176	176	701	701	538	538	163	163
Maximum	480	28,000	480	9,800	450	28,000	350	8,400	280	2,400	350	8,400
Minimum	1.6	7	1.6	7	3.7	120	1.5	7	1.5	10	4.8	110
Mean	27	670	17	190	58	2,120	24	360	13	110	60	1,190
Geometric mean	15	200	11	120	35	950	13	120	9.4	70	39	720
Median	13	140	9.8	110	32	900	10	80	8.2	60	41	670
Geometric standard deviation	2.7	3.5	2.3	2.2	2.6	3.3	2.7	3.8	2.1	2.3	2.6	2.7
Coefficient of geometric variation, percent (gCOV)[a]	37	24	35	16	27	17	39	28	33	20	26	15
Interquartile range (IQR)	20	280	13	100	50	1,620	16	220	8.3	60	55	1,170

[a] The geometric standard deviation, in log base 10 units, divided by geometric mean, in log base 10 units, ×100.

Because *E. coli* bacteria are thermotolerant (tolerant of relatively high temperatures), it is likely that colder river temperatures during the cool season inhibit the growth and survival of *E. coli* in the river especially at the Norcross site (Darakas, 2002). The *E.coli* densities tend to be higher during the warm season because *E. coli* bacteria have greater vitality, and wildlife and human activity increase during the warm season in the Chattahoochee River Basin. Unpublished *E. coli* densities measured in samples collected in 2001 from Panola Mountain State Park, southeast of Atlanta, Ga., were higher during the warm than cool season, even though the watershed was undisturbed by human activity (Brent T. Aulenbach, U.S. Geological Survey, written commun., 2002). Reasons for the greater *E. coli* densities at the Atlanta site than at the Norcross site during the cool season is not known, but may be related to warmer temperatures in water discharged from Bull Sluice Lake than the river temperature at the Norcross site.

Chattahoochee River near Norcross, Georgia

The variability in *E. coli* bacteria density and turbidity values is notable and consistently large in ambient waters throughout many parts of the United States (Myers and others, 1998; Maluk, 2000; Morace and McKenzie, 2002), and the Chattahoochee River in Metropolitan Atlanta is no exception (Gregory and Frick, 2000, 2001). At Norcross, *E. coli* densities exceeded the USEPA single-sample beach criterion of 235 *E. coli* colonies/100 mL infrequently during dry-weather-flow conditions (fig. 9*A*). About 98 percent of dry-weather flow and 24 percent of stormflow samples from Norcross had *E. coli* densities that were less than the USEPA beach criterion (figs. 9*A*, *B*). In addition, about 85 percent of dry-weather flow and 6 percent of stormflow samples had turbidity values equal to or less than 10 formazin nephelometric unit (FNU). About 17 percent of stormflow samples had *E. coli* densities greater than 2,507 MPN/100 mL (the USEPA single-sample criterion for infrequently used, full-body contact recreation). These values are substantially higher than those observed for the dry-weather-flow samples.

Descriptions and Statistical Analysis of *Escherichia coli* Density and Turbidity in Dry-Weather Flow and Stormflow

At the Norcross site, the median values for turbidity and *E. coli* density were statistically higher (*p*-value less than 0.001) during stormflow than during dry-weather flow for the study period (table 5). The maximum turbidity value during stormflow at Norcross was 2,700 FNU, nearly two orders of magnitude greater than the maximum turbidity measured during dry-weather flow. The median turbidity value during stormflow, 36 FNU, was nearly nine times greater than the median during dry-weather flow (4.6 FNU). Although *E. coli* density and turbidity were higher in stormflow than dry-weather flow, *E. coli* density and turbidity varied more during dry-weather flow. During dry weather, the variation in turbidity at Norcross was 6 percent greater (coefficient of geometric variation, gCOV, is 40 percent) than during stormflow (gCOV is 34 percent; table 5).

In stormflow samples at Norcross, the maximum *E. coli* density was 18,000 MPN/100 mL, 15 times greater than the maximum density during dry-weather flow (1,200 MPN/100 mL; table 5). The median *E. coli* density at Norcross was nearly 11 times greater in stormflow samples (530 MPN/100 mL) than in dry-weather samples (50 MPN/100 mL). In addition, the variation in *E. coli* density was similar among dry-weather and stormflow samples (gCOV at both sites is 20 percent).

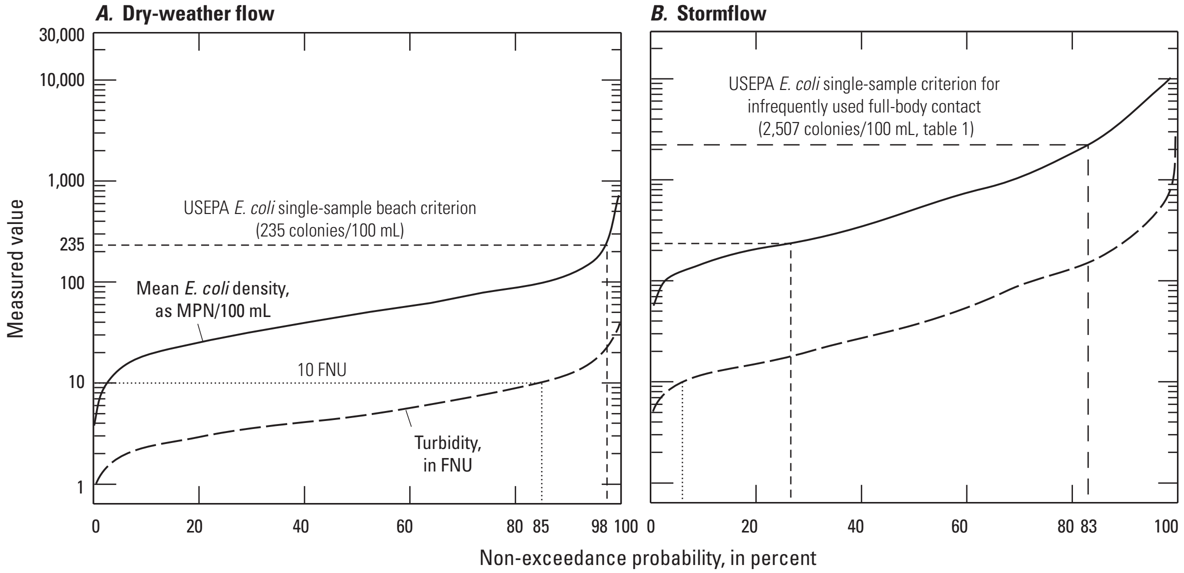


Figure 9. Non-exceedance probability distributions of turbidity and *Escherichia coli (E. coli)* bacteria measurements for *(A)* dry-weather flow and *(B)* stormflow samples collected from the Chattahoochee River near Norcross, Georgia (USGS station number 02335000), October 23, 2000, through September 30, 2008. [FNU, formazin nephelometric unit; /100 mL, per 100 milliliters of water]

Descriptions and Statistical Analysis of *Escherichia coli* Density and Turbidity by Season

The *E. coli* densities and turbidity values at Norcross were summarized by various combinations of season, dry-weather flow, and stormflow. Identifying how *E. coli* density and turbidity varied among those conditions was an important prelude to regression analysis because such knowledge aids in selecting explanatory variables. For samples and measurements from Norcross, median *E. coli* densities and turbidity were statistically greater during the warm than cool season (Wilcoxon Rank Sum test, *p*-value less than 0.001; table 6). Furthermore, median *E. coli* densities were highest in June and July, but median turbidity was highest in March, July, and August (figs. 8*A*, *B*).

In order to describe and compare *E. coli* densities under a variety of streamflow and seasonal conditions, water samples were split into groups that represented 24 combinations of season, streamflow event, and streamflow condition. In dry-weather samples collected during the six streamflow conditions in the warm season, median *E. coli* densities ranged from 40 to 70 MPN/100 mL (fig. 10*A*). During both seasons, median *E. coli* densities were statistically highest in dry-weather samples when streamflow was stable between 875 and 2,500 ft^3/s (StableNorm) and when stream stage was falling (fig. 10). In addition, median *E. coli* densities were statistically similar in samples collected when streamflow was stable and less than 875 ft^3/s (StableLow) or greater than 2,500 ft^3/s (StableHigh) and when stream stage was rising during peak water releases (fig. 10*A*, Wilcoxon Rank Sum test, *p*-value greater than 0.05).

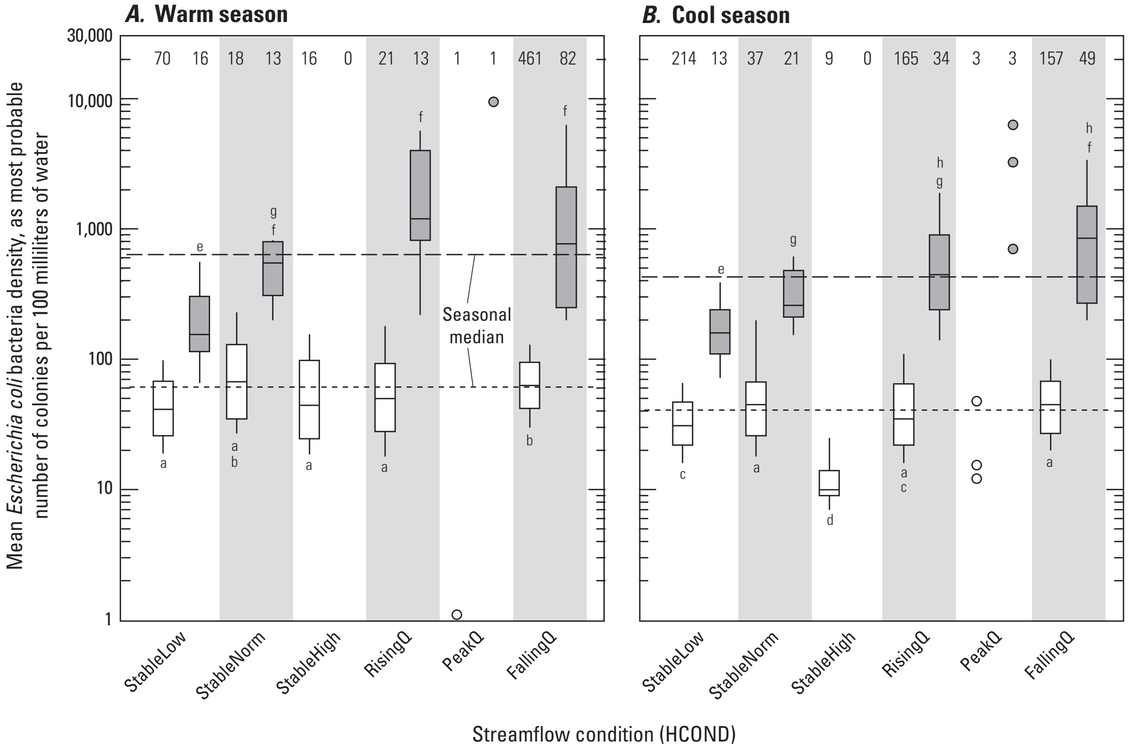


Figure 10. Distribution of *Escherichia coli* densities in dry-weather flow and stormflow samples collected during six flow conditions (table 3 and fig. 3) within *(A)* warm and *(B)* cool seasons, Chattahoochee River near Norcross, Georgia (USGS station number 02335000), October 23, 2000, through September 30, 2008.

In cool, dry-weather samples collected during the six streamflow conditions, median *E. coli* densities ranged from 10 to 45 MPN/100 mL (fig. 10*B*). Among these samples, the statistically lowest median *E. coli* density (10 MPN/100 mL) was measured when streamflow was stable and greater than 2,500 ft^3/s (StableHigh); however, within the StableHigh group, the lowest *E. coli* densities were measured in samples collected when streamflow was greater than 5,000 ft^3/s. This flow condition existed when large amounts of water were released from Buford Dam in response to several days of heavy rain. These events commonly occurred in February and March. Although these large releases were the response to large amounts of storm runoff entering Lake Lanier from upstream tributaries, the long duration of these releases commonly spanned both wet and dry weather and served to dilute streamflows (and high turbidity and *E. coli* densities) from tributaries upstream from the Norcross and Atlanta sites.

In warm-season stormflow samples collected at Norcross, median *E. coli* densities ranged from about 150 to 1,200 MPN/100 mL (fig. 10*A*). The highest turbidity value (2,690 FNU) and the highest *E. coli* density (18,000 MPN/100 mL) seen during the study period were measured in those samples (table 6). During the warm season, the median turbidity (32 FNU) and *E. coli* density (640 MPN/100 mL) were 8 to nearly 11 times higher in stormflow samples than dry-weather samples. During both seasons, the variation in turbidity was 6 to 7 percent higher in dry-weather than stormflow samples; whereas the variation among *E. coli* densities in stormflow and dry-weather samples was within 3 percent.

Chattahoochee River at Atlanta, Georgia

At the Atlanta site, *E. coli* densities exceeded the USEPA single-sample beach criterion (235 MPN/100 mL) infrequently during dry-weather-flow conditions. About 93 percent of dry-weather (fig. 11*A*) and 8 percent of stormflow (fig. 11*B*) samples had *E. coli* densities that were below the USEPA beach criterion. Moreover, about 60 percent of dry-weather (fig. 11*A*) and 7 percent of stormflow (fig. 11*B*) samples had turbidity values that were less than 10 FNU. In addition, about 17 percent of samples collected during stormflow at Atlanta contained *E. coli* densities greater than 2,507 MPN/100 mL (USEPA single-sample criterion for infrequently-used full-body contact recreation, table 1).

Descriptions and Statistical Analysis of *Escherichia coli* Density and Turbidity in Dry-Weather Flow and Stormflow

At the Atlanta site, *E. coli* density and turbidity measurements were substantially higher during stormflow than during dry-weather flow for the study period (table 6). The median values for *E. coli* density and turbidity were

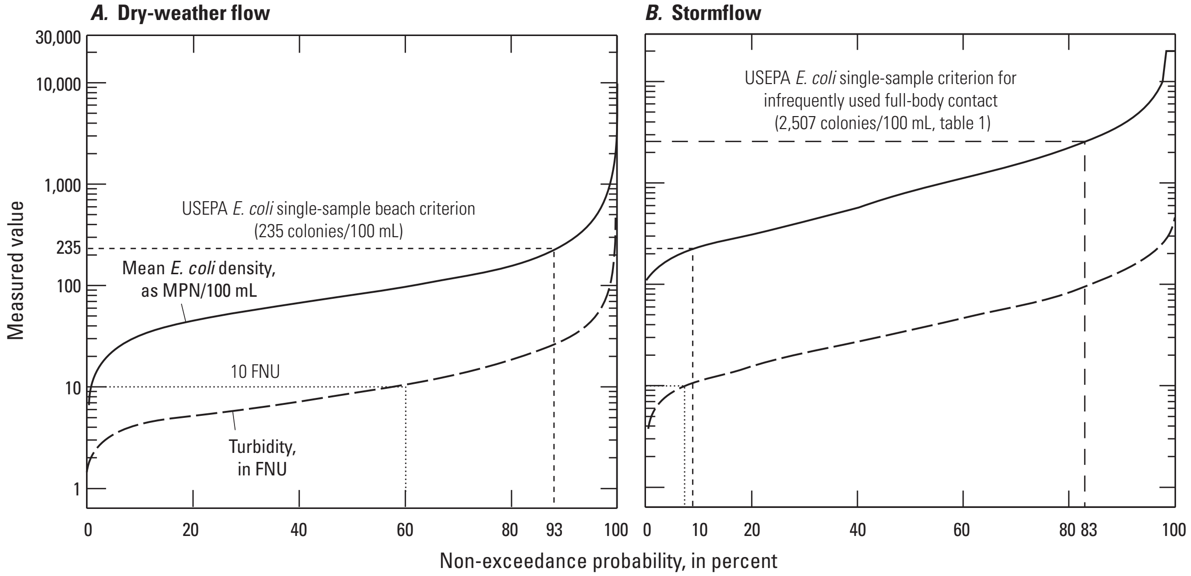


Figure 11. Non-exceedance probability distributions of turbidity and *Escherichia coli* bacteria measurements for samples collected during *(A)* dry-weather flow and *(B)* stormflow conditions at the Chattahoochee River at Atlanta, Georgia (USGS station number 02336000), October 23, 2000, through September 30, 2008.

statistically higher during stormflow than during dry-weather flow (Wilcoxon Rank Sum test, *p*-value less than 0.001). The maximum stormflow value for *E. coli* density at Atlanta was 28,000 MPN/100 mL; whereas the maximum during dry-weather flow was 9,800 MPN/100 mL. The median *E. coli* density in stormflow was 810 MPN/100 mL, an order of magnitude greater than the median during dry-weather flow. The maximum turbidity value during stormflow at Atlanta was 450 FNU, slightly lower than the maximum turbidity measured during dry-weather flow (480 FNU). In addition, the median turbidity value during stormflow, 35 FNU, was nearly four times greater than the median during dry-weather flow (9.1 FNU). At Atlanta, the variation in turbidity measurements was 8 percent higher during dry-weather flow than during stormflow, but the variation in *E. coli* density was similar (within 3 percent) in dry-weather and stormflow samples (table 6). This similarity may indicate that the analytical and sampling uncertainties were consistent regardless of season and did not bias *E. coli* densities.

One possible reason for the greater variation in turbidities during dry-weather flow at Atlanta may be the differences in the amount of water released from Bull Sluice Lake during the two different streamflow events. As mentioned previously, Morgan Falls Dam is operated as a run-of-the-river dam at the high streamflows common during storm runoff and extended periods of high discharges (greater than 5,000 ft^3/s) from Buford Dam; whereas during dry weather, at least 750 ft^3/s of water is released from Morgan Falls Dam because of instream flow requirements downstream; the magnitude and duration of additional releases depends on upstream releases from Buford Dam and the amount of storage available in Bull Sluice Lake.

Another source of variation in turbidity values and *E. coli* densities is cross-sectional differences at the sampling site. To determine the cross-sectional variability of *E. coli* density at the Atlanta site, water samples were collected at six locations in the channel cross section at different times, streamflow event, and streamflow condition during part of the study period. Figure 12 shows that *E. coli* density varied greatly even when streamflows were similar during two different streamflow events. At a high streamflow during dry weather (5,500 ft^3/s), the variation in *E. coli* density across the stream channel was low, indicating the river was well mixed (fig. 12*E*). In contrast, *E. coli* density was highest in samples collected near the left bank (looking downstream; figs. 12*C*, *D*) at a stable, normal stage (streamflow was 1,370 ft^3/s) during dry weather and falling stage of stormflow (streamflow was 2,600 ft^3/s). The thalweg of the river (that part of the stream channel that carries most of the streamflow during low flow) is present near the left bank at the Atlanta site. Water samples collected for the BacteriALERT project were routinely collected near mid-channel, away from the thalweg. The *E. coli* densities at mid-channel approximate the cross-sectional median when the river was well mixed during high flows. Conversely, during low flow, when most of the flow follows the thalweg, the mid-channel *E. coli* density underestimated the cross-sectional median.

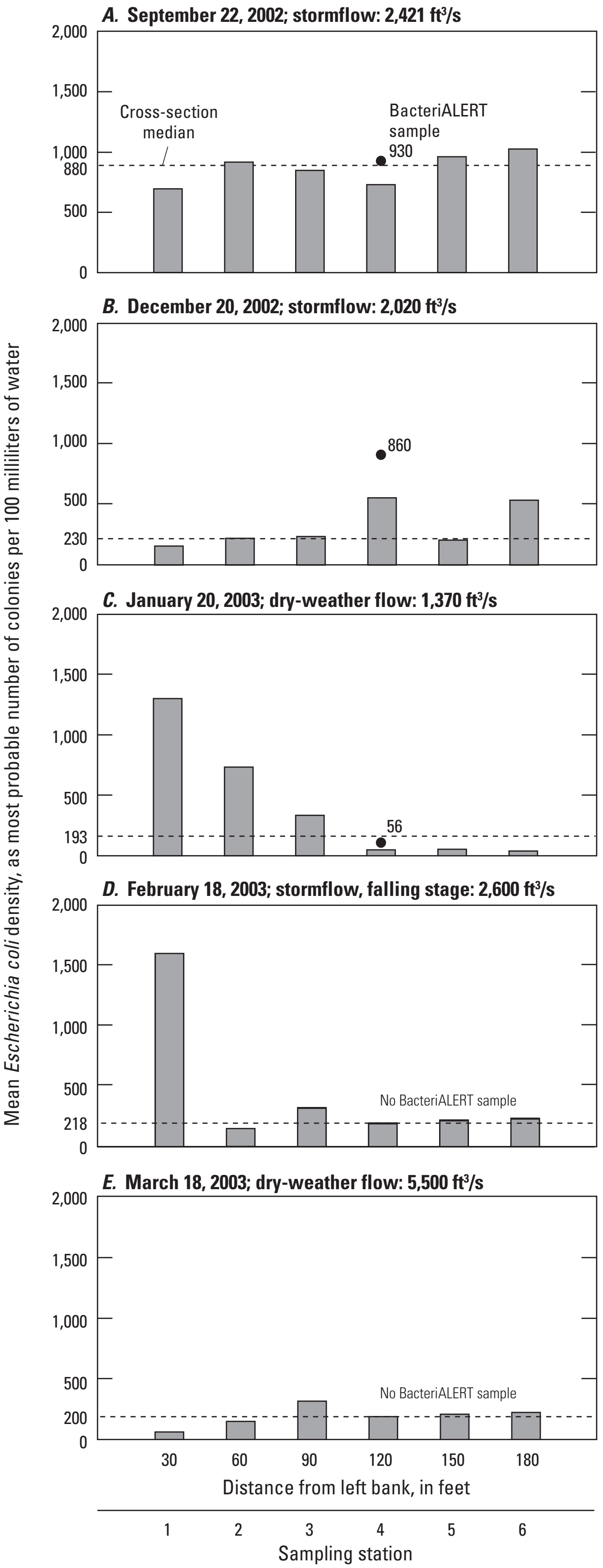


Figure 12. *Escherichia coli* density in water samples collected at six locations in the stream-channel cross section on five dates in the study period, Chattahoochee River at Atlanta, Georgia, September 22, 2002, to May 18, 2003.

Descriptions and Statistical Analysis of *Escherichia coli* Density and Turbidity by Season

Among warm-season samples from the Atlanta site, those collected in June and July had the highest median *E. coli* densities (fig. 8*A*); whereas the median turbidity was highest in July and August (fig. 8*B*). One source of turbidity at the Atlanta site may be algae or detritus from aquatic macrophytes growing in Bull Sluice Lake. The warm season is typically a time of peak growth for aquatic macrophytes in lakes, especially those that are small and shallow (Mitsch and Gosselink, 1986, p. 119). Between 2002 and 2005, an aquatic plant survey by the NPS and Georgia Power cataloged two algal species and 35 species of aquatic macrophytes in Bull Sluice Lake (report online at *http://www.georgiapower.com/lakes/hydro/pdfs/StudyReport_Wetlands.pdf*, accessed May 31, 2011).

Typically at the Atlanta site, *E. coli* density and turbidity were statistically greater during the warm season than during the cool season regardless of the streamflow event or streamflow conditions at the Atlanta site (Wilcoxon Rank Sum test, *p*-values less than 0.001; table 7; fig. 13). During stormflow, however, the median turbidity was 28 percent higher during the cool season than during the warm season, but the difference was not statistically significant (Wilcoxon Rank Sum

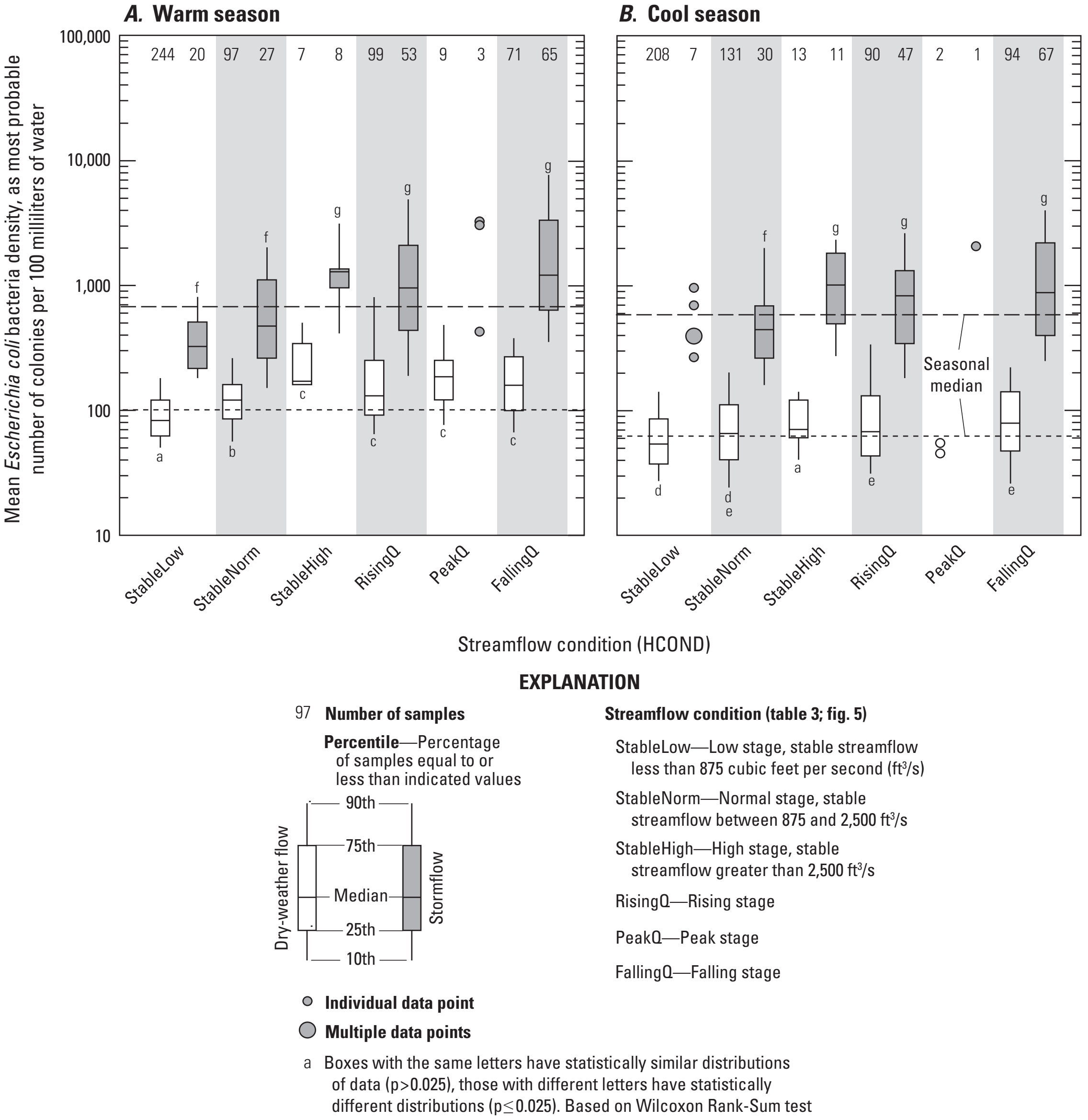


Figure 13. Distribution of *Escherichia coli* density grouped by hydrologic condition for dry-weather flow and stormflow samples collected during *(A)* the warm season and *(B)* the cool season from the Chattahoochee River at Atlanta, Georgia, October 23, 2000, through September 30, 2008.

test, *p*-value equals 0.302). In contrast, during dry-weather flows, the median turbidity value and median *E. coli* density, respectively, were 19 and 83 percent higher in warm-season than cool-season samples, differences that were statistically significant (Wilcoxon Rank Sum test, *p*-values less than 0.002; table 7). In addition, the dry-weather variation in turbidity values was similar during the warm and cool seasons (difference in gCOV is 2 percent), but the dry-weather variation in *E. coli* densities was slightly higher during the cool season (difference in gCOV is 4 percent; table 7). The stormflow variations in turbidity and *E. coli* density were similar during both seasons.

Regardless of season, *E. coli* densities were statistically similar among the six streamflow conditions during dry-weather flow and stormflow (fig. 13). In the warm season, the median *E. coli* densities ranged from 85 MPN/100 mL when dry-weather flow was stable and less than 1,100 ft^3/s (StableLow) to 190 MPN/100 mL when the daily maximum water releases from Morgan Falls Dam had peaked. The median *E. coli* densities were statistically similar among dry-weather samples collected during the rising, peak, and falling stream stages (fig. 13*A*, *p*-value greater than 0.025). During stormflow, the median *E. coli* densities in warm-season samples ranged from about 310 MPN/100 mL when streamflow was stable and less than 1,100 ft^3/s to about 1,300 MPN/100 mL when stormflow was receding (falling stage). In addition, median *E. coli* densities in stormflow samples were statistically similar in warm-season and cool-season samples collected under stable-low and StableNorm conditions (fig. 13).

In the cool season, the median *E. coli* densities ranged from 55 MPN/100 mL under stable-low conditions to about 80 MPN/100 mL at falling stages during dry-weather flows. During dry-weather flow, the statistically lowest median *E. coli* density (55 MPN/100 mL) was seen in cool-season samples collected under stable-low conditions. In contrast, dry-weather samples collected during the cool season under rising and falling stages had statistically similar median *E. coli* densities (70 to 80 MPN/100 mL; Wilcoxon Rank Sum test, *p*-value less than 0.025; fig. 13*B*). High stream stages with stable streamflow occurred primarily in late February and March during the study period as a result of heavy rain and corresponding large water releases (greater than 5,000 ft^3/s) from Buford and Morgan Falls Dams. These water releases were probably effective in scouring and suspending bottom sediments in Bull Sluice Lake.

Comparisons Between the Norcross and Atlanta Study Sites

The number of samples collected at both sites during dry-weather flow was nearly three to five times greater than the number of samples collected during stormflow. Among all data from Norcross, the study period median turbidity (5.7 FNU) and median *E. coli* density (60 MPN/100 mL) were nearly two times smaller than the medians for all data measured at Atlanta (table 5). In addition, the overall variability in turbidity measurements was 19 percent higher at Norcross (gCOV is 57 percent) than at Atlanta (gCOV is 38 percent); whereas the variability in *E. coli* density was about 5 percent higher at Norcross than Atlanta (gCOV at both sites is 31 and 26 percent, respectively). The turbidity and *E. coli* density may vary less at Atlanta, especially at low to moderate streamflows, because sediment transport is retarded by Bull Sluice Lake upstream from the Atlanta site. Nevertheless, *E. coli* bacteria can thrive in bottom sediments in slack water such as ponds and lakes (He and others, 2007) and can be transported out of the lake during large stormflows.

The *E. coli* densities and turbidity values were substantially higher in dry-weather-flow samples from the Atlanta than the Norcross site (table 5). The median turbidity value and *E. coli* density (9.1 FNU and 80 MPN/100 mL, respectively) in dry-weather-flow samples from Atlanta were 49 and 37 percent higher, respectively, than in dry-weather-flow samples from the Norcross site. The variability in turbidity, however, was 6 percent higher at Norcross than Atlanta (gCOV is 40 and 34 percent, respectively), but the variability in *E. coli* densities was similar at both sites. During dry-weather flow, *E. coli* densities exceeded the USEPA single-sample beach criterion three times more often in water samples from Atlanta (7 percent) than from the Norcross site (2 percent; figs. 11*A*, 9A). Similarly, turbidity values exceeded 10 FNU nearly three times more often at the Atlanta site (39 percent) than at the Norcross site (14 percent).

During stormflow, median turbidity values were statistically similar at the Norcross and Atlanta sites, but variability was 6 percent higher at Norcross (table 5). Although the median *E. coli* density was about 53 percent higher at Atlanta than at Norcross, the variability in *E. coli* density was similar at both sites (gCOV for both sites differed by 4 percent). During the study period, *E. coli* density exceeded the USEPA single-sample beach criterion in 16 percent more samples from Atlanta than Norcross. The percentage of samples with *E. coli* density above 4,000 MPN/100 mL was similar at both sites (figs. 9, 11). About 11 percent more turbidity measurements exceeded 100 FNU at Atlanta than at Norcross.

Typically, the monthly median *E. coli* density and turbidity were markedly higher among all samples collected from the Atlanta site than from the Norcross site (figs. 8*A*, *B*). During both seasons, the median *E. coli* density and turbidity values were twice as large at Atlanta than at Norcross (tables 6, 7). The variability in turbidity values was 26 percent higher for warm-season measurements and 13 percent higher for cool-season measurements at Norcross than at Atlanta. In addition, the variation in *E. coli* density was about 4 percent higher at Norcross than at Atlanta.

Regression Analysis of *Escherichia coli* Bacteria Density

This section describes the results from linear and nonlinear (logistic) regression analyses to determine the best regression equation for predicting median *E. coli* densities at the Norcross and Atlanta sampling sites. See appendix 2 for a complete description of regression analysis techniques. For this report, the "best" set of parameters maximize the R^2 and minimize the Mallow's Cp statistic. The variables used during the initial leaps and bounds procedure are listed in table 2–1 in appendix 2. The "best" one-variable regression equation is commonly the equation that explains most of the variability in the response variable. This equation was investigated initially for the Norcross and Atlanta sites to identify outliers and residual patterns using data collected during the study period. In addition, 90-percent prediction intervals for new observations were computed. Logistic regression analysis was used to develop the probability that *E. coli* density exceeds the USEPA single-sample beach criterion of 235 colonies/100 mL at a given turbidity measurement.

Chattahoochee River near Norcross, Georgia

Eight linear regression analyses were completed using the Norcross data (table 8). These eight equations range from a simple linear regression to multiple linear regressions. Regression plots and statistics for regressions 1–3, 5, and 7 are shown in appendix 5 to this report (figs. 5–1 to 5–5). Variable inflation factors among the variables within each equation typically did not exceed 3.6, a value below the threshold of 5.0 that indicates multicollinearity among explanatory variables.

Development of Linear Regression Equations

Initially, *E. coli* density as log10Ecoli was regressed against turbidity as log10FNU. This regression (regression-1, table 8) uses the full set of samples, sorted by collection date, from the Norcross site during the study period. This initial regression analysis indicates a statistically significant linear relation between log10Ecoli and log10FNU (*p*-value less than 0.001) and explains about 51 percent of the variability in the log10Ecoli densities. Table 8 summarizes the results of diagnostic tests on this initial equation. Regression and residual plots, and regression statistics are shown in appendix 5 (fig. 5–1).

The residual analysis for regression-1 indicates that the residuals are normally distributed (more than 90 percent are within ±2 standard deviation of the mean), but about 8 percent of those residuals depart substantially from the normal distribution, primarily in the upper and lower tails of the distribution. These residuals indicate that regression-1 underpredicts *E. coli* density, especially at higher turbidity values typical of stormflow (appendix 5, fig. 5–1*A*). The DurbinWatson statistic (table 8) indicates strong autocorrelation in the Norcross dataset. The correlation of residuals lagged by sample date confirms that one cause of the autocorrelation in this dataset is the presence of redundant samples (appendix 5, fig. 5–1). The autocorrelation coefficient (0.11) for the Norcross data fell below the critical value of 0.20 when the dataset was lagged by two samples.

Autocorrelation can indicate two phenomena: seasonality in the data or data redundancy, in which too many samples are collected within a short period of time (Helsel and Hirsch, 1992; Montgomery and others, 2006). The Norcross dataset is affected by both phenomena: seasonality is apparent in the Norcross data as discussed in previous sections of this report and redundant samples are present in the dataset because of the near daily sample collection between October 23, 2000, and September 30, 2001, and the daily sampling from October 1, 2001, to September 30, 2002. As indicated by Durbin-Watson statistics that range between 1.91 and 1.94, autocorrelation in the Norcross data is eliminated when the data are redistributed randomly through the study period before regression analysis (regressions-2 and 5, table 8). Because time-series forecasting of *E. coli* density into the future is not the intended use of the regression equations developed for this study, autocorrelation in the data does not prohibit these equations from being used to predict *E. coli* density on the basis of turbidity measurements.

Table 8 lists the "best" 2-, 3-, and 4-variable regression equations identified by the leaps and bounds procedure and shows the regression differences when the data are sorted by collection date, randomized, and with outliers removed. Regression-2 represents the "best" 2-variable equation and regression-5 the "best" 3-variable equation on the randomized dataset. With the addition of the indicator variable for streamflow event (EVENT: dry-weather flow or stormflow) in regression-2 and the addition of EVENT and Season in regression-5, the amount of variability in log10Ecoli density that was explained by regressions-2 and 5, was 13 and 17 percent, respectively, greater than that explained by regression-1. Furthermore, the residual standard error and the Akaike Information Criterion (AIC) for regressions-2 and 5 were substantially improved. Nevertheless, the residuals show data clusters that correspond to the value of the EVENT variable (appendix 5, figs. 5–2*A*, 5–4*A*) and an upward trend among the residuals, indicating the equations underpredict *E. coli* density, especially during stormflow (appendix 5, figs. 5–2*A*, *B*; figs. 5–4*A*, *B*).

Table 8. Regression statistics for the best fit regression analysis on datasets containing mean *Escherichia coli* bacteria density and turbidity measured at the Chattahoochee River near Norcross, Georgia (USGS station number 02335000), October 23, 2000, through September 30, 2008.

[Explanatory variables in the table are selected based on the lowest Cp statistic in each group of explanatory variables using the S-PLUS leaps and bounds command; no outliers, original dataset minus outliers; randomized, dataset in which samples were randomly selected with replacement from the full dataset; *Escherichia coli* bacteria density as most probable number per 100 milliliters of water; adjusted R^2, the coefficient of determination adjusted using the mean square error to negate the tendency for the R^2 to increase as the number of explanatory variables increases; VIF, variance inflation factor; AIC, Akaike Information Criterion, a measure of the goodness of fit; —, not applicable]

Regression number	Samples in dataset	Log10Ecoli in relation to significant explanatory variables	Number of samples	Adjusted R^2	Residual standard error	VIF[a]	Durbin-Watson statistic[b]	AIC	Figure number
1	All, sorted	Log10(FNU)	1,417	0.512	0.394	—	1.180	1,389	5–1
2	All, randomized	Log10(FNU), EVENT[c]	1,417	.636	.351	2.28	2.024	1,059	5–2
3	All, sorted	Log10(FNU), EVENT[c], [EVENT×Log10(FNU)][d]	1,417	.650	.334	3.63	1.195	920	5–3
4	Outliers removed, sorted	Log10(FNU), EVENT[c], [EVENT×Log10(FNU)][d]	1,328	.744	.268	1.73	1.466	278	5–4
5	All, randomized	Log10(FNU), EVENT[c], Season[e]	1,417	.684	.327	2.36	2.010	859	5–5
6	Outliers removed, sorted	Log10(FNU), EVENT[c], Season[e], [EVENT×Log10(FNU)][d]	1,328	.779	.249	1.79	1.650	82	15
7	Outliers removed, sorted	Log10(FNU), EVENT[c], Season[e], HCOND[f]	1,328	.763	.258	2.23	1.593	177	5–6
8	Outliers removed, sorted	Log10(FNU), EVENT[c], Season[e], HCOND[f], [EVENT×Log10(FNU)][d]	1,328	.791	.242	1.79	1.685	8	16

[a] Variance inflation factor, a measure of a variable's influence on the variance of the regression. Only the largest value among the explanatory variables is given. A VIF below 5 is considered insignificant (Montgomery and others, 2006). Interaction terms are not considered "variables" so VIF statistics are not computed.

[b] Durban-Watson statistic is a measure of first-order autocorrelation among observations in the dataset. The farther the absolute value is from 2, the greater the autocorrelation.

[c] Variable indicating streamflow regime in which water samples were collected (dry-weather flow or stormflow; table 3).

[d] Interaction variable (Montgomery and others, 2006).

[e] Variable indicating season in which samples were collected: warm (mid-April to mid-October); cool season (mid-October to mid-April; table 3).

[f] Variable indicating the flow condition in which samples were collected: low stable streamflow; normal, stable streamflow; high, stable streamflow; rising stage, peak stage; or falling stage of the hydrograph (table 3).

Exploratory data analysis showed that the linear relations between log10Ecoli and log10FNU for dry-weather-flow and stormflow samples at Norcross had different y-intercept and slope coefficients (fig. 14). These differences are important for regression equations that include the indicator variable EVENT because to counter the bias introduced by the differences in the slope or intercepts between dry-weather and stormflow, an interaction term is needed (Montgomery and others, 2006). An interaction term that is the cross product of EVENT and log10FNU was added to the explanatory variables in regression-2 to create regression-3 (Montgomery and others, 2006; table 8). The differences in the y-intercept and slope coefficients for dry-weather flow and stormflow samples probably indicate that *E. coli* bacteria densities are different depending on the source of the suspended solids in the Chattahoochee River upstream from Norcross as described earlier. Therefore, two populations of *E. coli* bacteria exist in the study area: (1) low densities of bacteria associated with soil and sediment from channel erosion and river-bed scouring in the Chattahoochee River that is transported during water releases from Buford Dam and (2) higher densities of bacteria associated with suspended solids transported in storm runoff from the heavily urbanized streams that are tributary to the Chattahoochee River in the study area. Regression statistics show that adding the interaction term in regression-3 produced a slightly stronger regression equation than regression-2 (table 8; appendix 5, fig. 5–3).

Regressions-4 and 6–8 represent the regression analyses with outliers removed. Outliers are samples with log10Ecoli or log10FNU values that exerted high leverage and influence on the regression. Eighty-nine samples (6.3 percent) were considered outliers because they had absolute studentized residuals that exceeded 1.9 standard deviations or had DFFITS values that exceeded 0.106, the computed critical value. The DFFITS statistic determines the degree to which each data point influences the regression statistics. Data points are removed one at a time and the regression statistics are re-computed minus that data point. The *E. coli* density and turbidity values for these samples were reviewed for measurement or computational errors and corrected if needed. By removing the outliers, the adjusted R^2 for regressions-4 and 6–8 were 9 to 28 percent higher than the regressions using all samples collected during the study period (table 8).

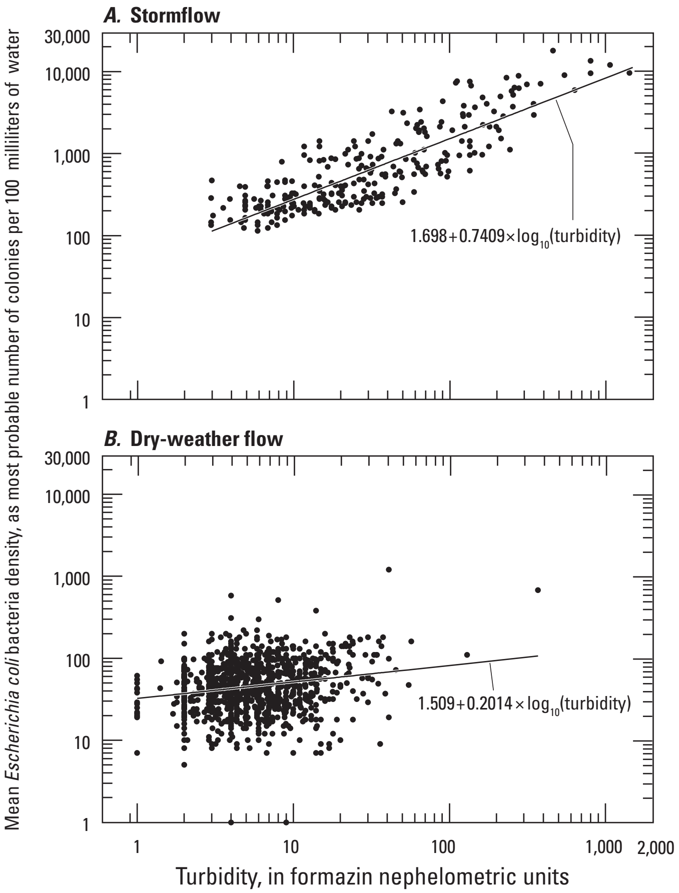


Figure 14. Linear relation between turbidity measurements and mean *Escherichia coli* bacteria density in dry-weather flow and stormflow samples from the Chattahoochee River near Norcross, Georgia, October 23, 2000, through September 30, 2008.

These outliers may reflect unidentified measurement errors, high *E. coli* densities associated with low turbidity, or low *E. coli* densities with high turbidity. Samples with high *E. coli* densities and low turbidity during both streamflow events were most likely caused by sewage leaking from broken sewer pipes, emergency releases of sewage effluent from wastewater-treatment facilities, illegal releases of raw or treated sewage, or sanitary-sewer overflows. Discharges of untreated or partially treated sewage effluent into the Chattahoochee River and its tributaries from broken or leaking sewer pipes, or sanitary sewer overflows are well known anecdotally, but were poorly documented during the study period. In addition, samples with abnormally low *E. coli* densities associated with high turbidity values can result from water releases from Buford Dam that are large (greater than 6,000 ft^3/s) and of long duration (longer than 24 hours), or the result of uncontrolled releases of wastewater effluent that are disinfected to avoid violating conditions of a discharge permit or fecal coliform standards in a receiving stream. Furthermore, soil disturbed during road and building construction projects, especially those that disrupt stream buffers, commonly enters stream channels in high amounts even with erosion control measures in place.

Regressions-6 and 8 are the statistically strongest regression equations computed from the Norcross data collected during the study period (table 8). The adjusted R^2 for regression-6 (0.779) and regression-8 (0.791) indicates that both equations explain 3.5 to 4.5 percent more of the variability in *E. coli* densities than regression-4. Furthermore, the residual standard errors for regressions-6 and 8 are 7 to 10 percent lower than those for regression-4; whereas, AICs are 4 to 35 times lower. In addition, residual analysis for regressions-6 and 8 show that the estimated *E. coli* densities correlate well with the observed *E. coli* densities (figs. 15*A*, 16*A*), and the residuals are normally distributed, appear to

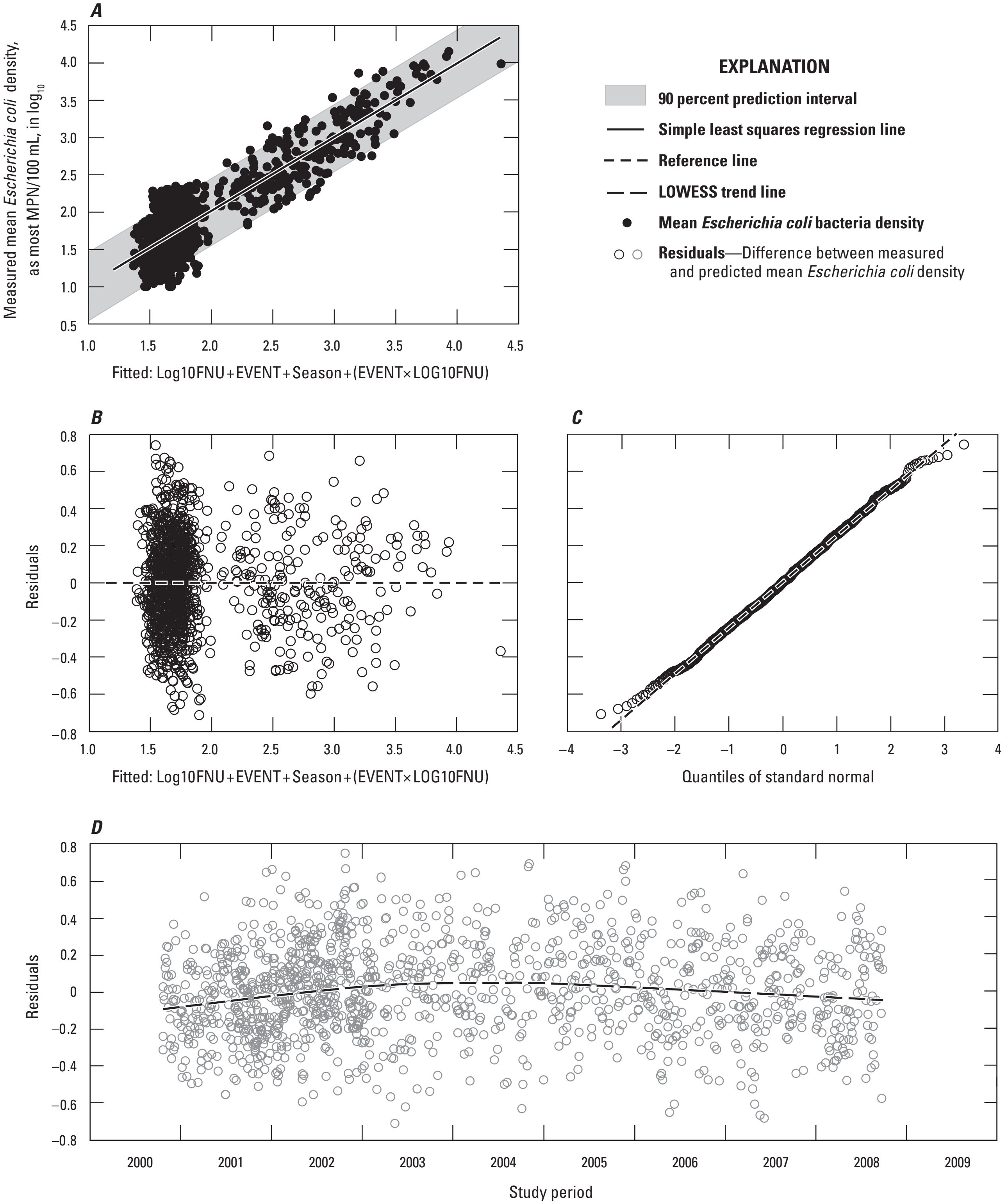


Figure 15. Diagnostic plots for regression-6 (outliers removed, table 8) using data collected at the Chattahoochee River near Norcross, Georgia (USGS station number 02335000), from October 23, 2000, through September 30, 2008. *(A)* Relation between measured and estimated mean *Escherichia coli (E. coli)* bacteria densities. *(B)* Relation between residuals and the estimated mean *E. coli* densities. *(C)* Distribution of the residuals compared to a standard normal distribution. *(D)* The trend in residuals. Explanatory variables: Log10FNU, turbidity in formazin nephelometric units transformed to base 10 logarithms; EVENT (dry-weather flow or stormflow); Season (warm or cool); and the interaction term (EVENT×Log10FNU); MPN/100 mL, most probable number of colonies per 100 milliliters of water.

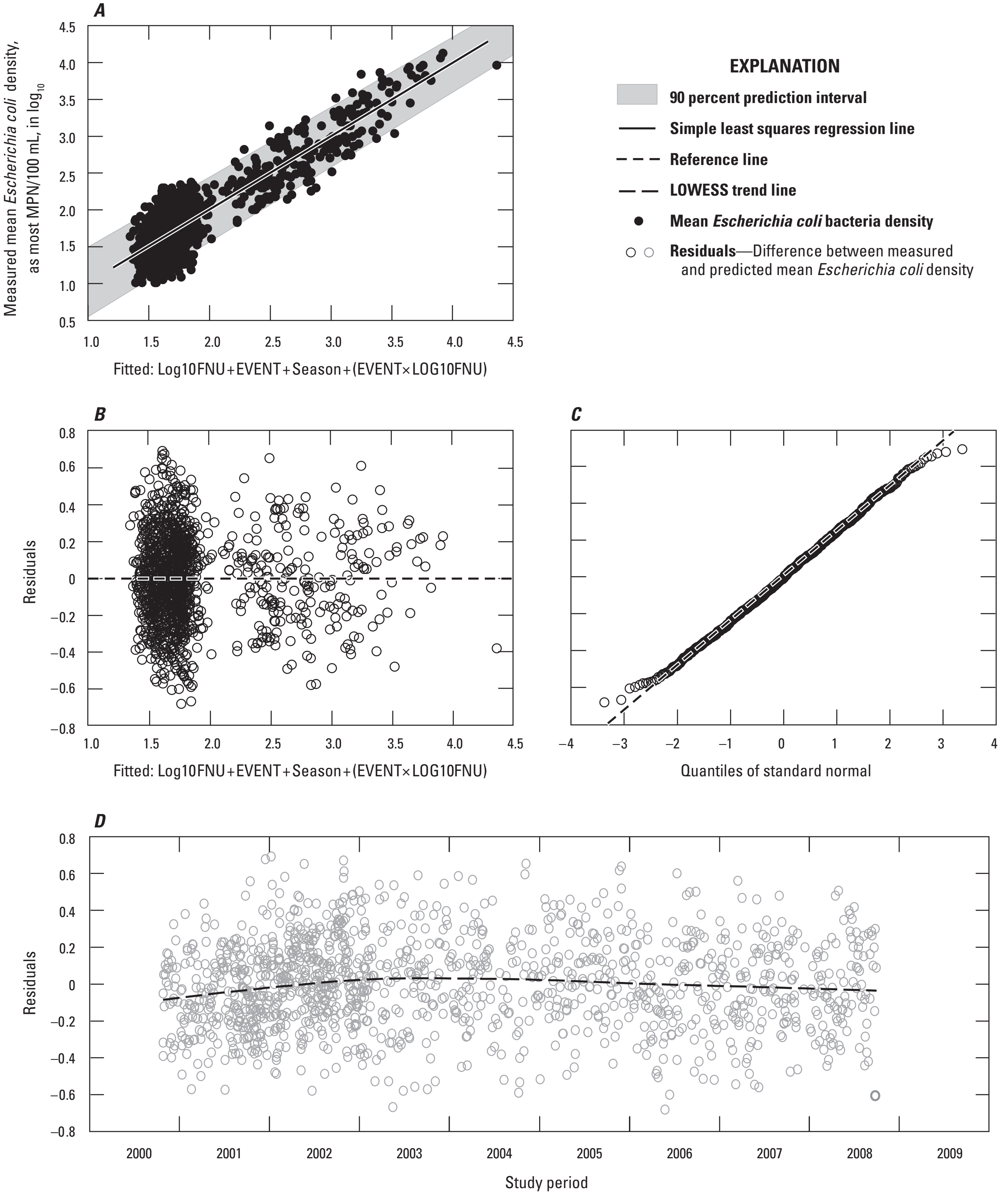


Figure 16. Diagnostic plots for regression-8 (outliers removed, table 8) using data collected at the Chattahoochee River near Norcross, Georgia (USGS station number 02335000), from October 1, 2008, through September 30, 2009. *(A)* Relation between measured and estimated mean *Escherichia coli* density in base 10 logarithm units (Log10Ecoli). *(B)* Relation between residuals and estimated Log10Ecoli). *(C)* Distribution of the residuals compared to a standard normal distribution. *(D)* The study-period trend in residuals. Explanatory variables: Log10FNU, turbidity in formazin nephelometric units transformed to base 10 logarithms; EVENT (streamflow regime: dry-weather flow or stormflow); Season (warm or cool), streamflow condition (HCOND); and the interaction term (EVENT×Log10FNU); MPN/100 mL, most probable number of colonies per 100 milliliters of water.

Table 9. Results, analysis of variance, and autocorrelation coefficients for the multiple linear regression analysis of mean *Escherichia coli* bacteria densities for stormflow samples (regression-6, outliers removed, table 8) from the Chattahoochee River near Norcross, Georgia (USGS station number 02335000), October 23, 2000, through September 30, 2008.

[Adjusted R^2, the coefficient of determination adjusted using the mean square error to negate the tendency for the R^2 to increase as the number of explanatory variables increases; *Escherichia coli* bacteria density, as most probable number of colonies (MPN) per 100 milliliters of water; SE, standard error of regression; t-statistic, used to determine if the coefficient is statistically equal to zero; residuals, the difference between the actual and predicted value of the response variable; —, not applicable]

Residual standard error is 0.2490, adjusted R^2 is 0.779, F statistic is 1,170, and *p*-value is <0.001 Response variable: Log10Ecoli				
Terms	**Coefficient**	**Standard error (SE)**	**t-statistic**	***p*-value[a]**
Intercept	1.798	0.026	69.3	<0.001
Log10(FNU)	.251	.027	9.2	<.001
EVENT[b]	–.084	.058	–1.4	.150
Season[c]	–.205	.014	–14.6	<.001
[Log10(FNU)×EVENT][d]	.577	.042	13.8	<.001

Analysis of variance					
Terms	**Degrees of freedom[e]**	**Sum of squares (SS)**	**Mean SS**	**F statistic[f]**	***p*-value (F)**
Log10(FNU)	1	218.93	218.93	3,535	<0.001
EVENT[b]	1	42.13	42.13	680	<.001
Season[c]	1	17.35	17.35	280	<.001
[Log10(FNU)×EVENT][d]	1	11.88	11.88	192	<.001
Residuals	1,323	81.94	.06	—	—

Autocorrelation coefficients	
Number of samples lagged	**Correlation coefficient[g]**
0	1.00
1	.17
2	.13

[a] *p*-value, the probability that the parameter is not important to the regression.

[b] Variable indicating streamflow regime in which water samples were collected (dry-weather flow or stormflow; table 3).

[c] Variable indicating cool or warm season, table 3.

[d] Interaction variable.

[e] Defined as the number of independent pieces of information used to calculate the statistics.

[f] F statistic, used to determine if there is a significant linear relation between the response variable and the explanatory variables.

[g] The critical value is 0.20. Coefficients greater than 0.20 are significant at alpha equal to 0.10.

Table 10. Results, analysis of variance, and autocorrelation coefficients for the multiple linear regression analysis of mean *Escherichia coli* bacteria densities for dry-weather flow samples (regression-8, outliers removed, table 8) from the Chattahoochee River near Norcross, Georgia (USGS station number 02335000), October 23, 2000, through September 30, 2008.

[Adjusted R^2, the coefficient of determination adjusted using the mean square error to negate the tendency for the R^2 to increase as the number of explanatory variables increases; *Escherichia coli* bacteria density, as most probable number of colonies (MPN) per 100 milliliters of water; SE, standard error of regression; t-statistic, used to determine if the coefficient is statistically equal to zero; residual, the difference between the actual and predicted value of the response variable; —, not applicable]

Residual standard error is 0.2420, adjusted R^2 is 0.791, F statistic is 1,010, and *p*-value is <0.001
Response variable: Log10Ecoli

Terms	Coefficient	Standard error (SE)	t-statistic	*p*-value[a]
Intercept	1.630	0.032	51.5	<0.001
Log10(FNU)	.255	.027	9.6	<.001
EVENT[b]	–.050	.057	–.9	.375
Season[c]	–.162	.014	–11.2	<.001
HCOND[d]	.038	.004	8.8	<.001
[Log10(FNU)×EVENT][e]	.548	.041	13.5	<.001

Analysis of variance

Terms	Degrees of freedom[f]	Sum of squares (SS)	Mean SS	F statistic[g]	*p*-value (F)
Log10(FNU)	1	218.93	218.93	3,739	<0.001
EVENT[b]	1	42.13	42.13	720	<.001
Season[c]	1	17.35	17.35	296	<.001
HCOND[d]	1	5.77	5.77	99	<.001
[Log10(FNU)×EVENT][e]	1	10.64	10.64	182	<.001
Residuals	1,322	77.41	.06	—	—

Autocorrelation coefficients

Number of samples lagged	Correlation coefficient[h]
0	1.00
1	.16
2	.12

[a] *p*-value, the probability that the parameter is not important to the regression.

[b] Variable indicating streamflow regime in which water samples were collected (dry-weather flow or stormflow; table 3).

[c] Variable indicating cool or warm season, table 3.

[d] Variable indicating flow condition, such as rising or falling stream stage, table 3.

[e] Interaction variable.

[f] Defined as the number of independent pieces of information used to calculate the statistics.

[g] F statistic, used to determine if there is a significant linear relation between the response variable and the explanatory variables.

[h] The critical value is 0.20. Coefficients greater than 0.20 are significant at alpha equal to 0.10.

have constant variance, and show no obvious trend over the study period (figs. 15*B*–*D*, 16*B*–*D*). Table 9 lists the summary statistics and analysis of variance (ANOVA) for regression-6 and those for regression-8 are shown in table 10. On the basis of the Kendall tau and Sen slope estimate test, statistically significant trends in *E. coli* density were not evident during the study period (*p*-value equals 0.138).

Although regression-8 has the highest adjusted R^2 (0.791) of any regression analyses on the Norcross data, it only differs from regression-6 by about 1 percent. In addition, the residual standard errors for regressions-8 and 6 are within 3 percent of each other. Nevertheless, regression-8 may be a better predictor of log10Ecoli densities than regression-6 because the AIC for regression-8 is 90 percent lower than for regression-6. Regression-6, however, was chosen for validation because it is statistically similar to regression-8 and is the simpler of the two equations. The HCOND variable in regression-8 adds a greater degree of computational complexity to the real-time model; therefore, regression-6 is computationally easier to use for the real-time prediction of *E. coli* density.

Validation of the Selected Linear Regression Equation

The "best" regression equation is one that meets the objectives of the intended use of the equation. In most cases, the objective of a regression equation is to accurately predict a response when measured values of the explanatory variables are given. To this end, regression validation determines how well a regression equation estimates the response variable either with data not used to develop the equation or with a subset of the estimation dataset (Montgomery and others, 2006, p. 424). For this report, the validation dataset consisted of *E. coli* densities in water samples collected and instream turbidity measured at the Norcross site from October 1, 2008, to September 30, 2009. Validation has two goals: to determine if the regression equation violates the basic assumptions of regression analysis and to determine if the equation can accurately (within the prediction interval of the regression) predict the response variable using the explanatory variables from a new dataset (Montgomery and others, 2006). The assumptions mentioned in the first goal above are:

- The relation between the response and explanatory variables is at least approximately linear.
- The error term in the regression has a mean of zero and constant variance.
- The errors are uncorrelated and normally distributed.

The validity of regression-6 is determined by computing diagnostic indices. These indices are based on the residuals (difference between the measured and estimated *E. coli* density) generated using regression-6 with the new dataset from Norcross. Diagnostic indices include Spearman's rank correlation coefficients between measured and predicted *E. coli* densities, mean sum of squared prediction errors (MSS_p), and prediction R^2.

Measured *E. coli* densities in water samples from the Norcross validation dataset were compared to those predicted using regression-6. The diagnostic indices indicate that regression-6 adequately predicts *E. coli* densities in the validation dataset (*p*-values are less than 0.001). The prediction R^2 was 0.637 and Spearman's rank correlation coefficient was 0.814 for regression-6. The MSS_p (0.1336) was twice the mean sum of squared residuals (MSS_f, 0.06) for regression-6. According to Montgomery and others (2006, p. 309), the prediction R^2 commonly is smaller and the MSS_p is larger than those for the estimation R^2 and MSS_f because regression equations typically do not predict new data as well as they fit the original data. Furthermore, the new dataset is much smaller than the dataset used to develop regression-6 and the smaller the number of samples in a statistical analysis the higher the computed variation. Nevertheless, the measured and predicted *E. coli* densities in the validation dataset plot within the 90 percent prediction interval for the estimation dataset using regression-6 (fig. 17*A*). Prediction residuals were generally normally distributed, but indicated the presence of a few outliers; the distribution of the prediction residuals differed slightly from the results of regression-6, but this was expected because of the small size of the validation dataset. The outliers were probably the result of random occurrences of broken sewer pipes or sanitary-sewer or combined-sewer overflows (figs. 17*B*, *C*).

Development and Validation of Logistic Regression Equations

Logistic regression analysis was used to develop an equation for estimating the probability that *E. coli* density will exceed the USEPA single-sample beach criterion of 235 colonies/100 mL for a given turbidity measurement at Norcross. Turbidity values and the binary variable EVENT (stormflow or dry-weather flow) are explanatory variables in the logistic regression analysis to predict the binary outcome of *E. coli* density, that is whether *E. coli* density is below (0) or above (1) the USEPA beach criterion. Logistic regression was completed on two different datasets from the Norcross site: the full 1,417 sample dataset and the 1,328 sample dataset with outliers removed. Five indices of regression strength indicate that turbidity and EVENT have a strong statistical relation to the probability of *E. coli* density exceeding 235 colonies/100 mL (table 11).

The five indices include the deviance statistic, probability of concordance (c), Somer's D_{xy} rank correlation, R^2, and Brier's score (Harrell, 2001, p. 249). The deviance statistic is a log-likelihood goodness-of-fit function that identifies how close the logistic model fits the observed probabilities of *E. coli* density exceeding 235 MPN/100 mL. The deviance is calculated by dividing the regression deviance by the degrees of freedom for the dataset. As the deviance statistic increases from 1.0, the greater the disparity between the observed and estimated probabilities and the poorer the model (Montgomery and others, 2006, p. 437). The c and D_{xy} indices measure the rank correlation between the predicted and observed probabilities of the response variable; in other words, the difference between concordance and discordance probabilities. Values of 0 for the c and D_{xy} indices indicate the equation is making random predictions; whereas a value of 1.0 indicates the predictions have perfect concordance with the observed probabilities (Harrell, 2001). The R^2 is a measure of the predictive strength of the logistic regression equation; in other words, how much of the variation in the response is explained by the variation in the explanatory variables. The Brier's score sums the difference between the predicted probabilities and the observed responses. The smaller the Brier's score, the smaller the differences between the predicted probabilities and the observed responses, and the stronger the predictive capability of the model.

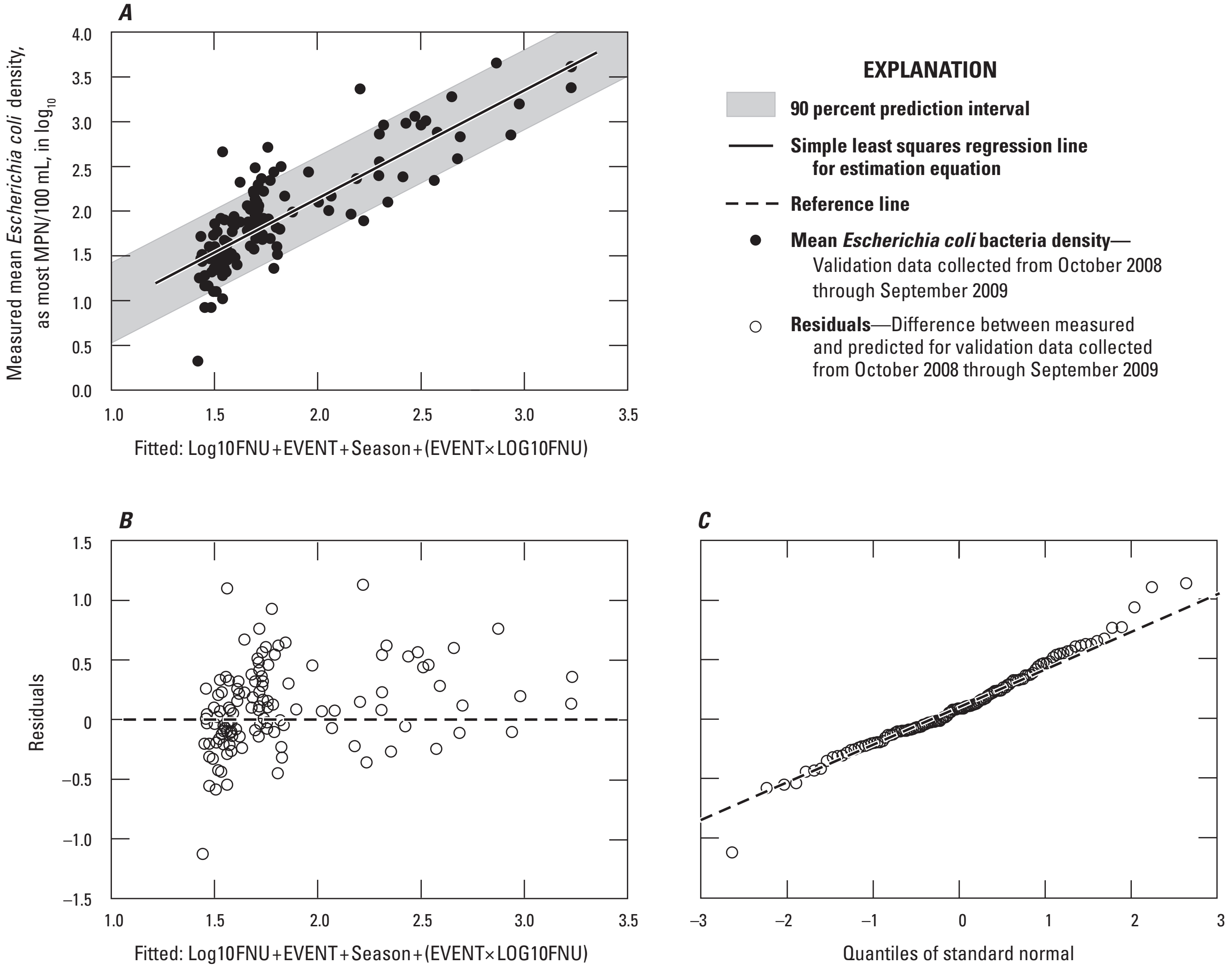


Figure 17. Diagnostic plots for the validation analysis of regression-6 (outliers removed, table 8) using data collected at the Chattahoochee River near Norcross, Georgia (USGS station number 02335000), from October 1, 2008, through September 30, 2009. *(A)* Relation between measured and predicted *Escherichia coli (E. coli)* bacteria densities. *(B)* Relation between residuals and the predicted *E. coli* densities. *(C)* Distribution of the residuals compared to a standard normal distribution. Explanatory variables: Log10FNU, turbidity in formazin nephelometric units transformed to base 10 logarithms; EVENT (streamflow regime: dry-weather flow or stormflow); Season (warm or cool); and the interaction term (EVENT×Log10FNU); MPN/100 mL, most probable number of colonies per 100 milliliters of water.

Although the five indices show that the logistic equations for both Norcross datasets are statistically significant, the computed indices for the full dataset were markedly different than the indices computed on the dataset with outliers removed (table 11). For both datasets, the c index ranged from 0.945 to 0.992 and the D_{xy} indices ranged from 0.890 to 0.983, indicating a high probability of concordance between observed and predicted probabilities. The R^2 of 0.884 indicates the logistic equation for the data without outliers has a stronger predictive ability than the equation for the full dataset. The Brier's score for both datasets ranged from 0.02 to 0.045, indicating the differences between the observed and predicted probabilities were small. A graphical depiction of both logistic equations for the Norcross data is shown in figures 18*A* and 18*C*. The observed proportions of samples with *E. coli* densities that exceeded the USEPA single-sample beach criterion, on the basis of turbidity classes, are listed in table 12.

Although the logistic regression equation was a better fit for the dataset without outliers than for the full dataset, the validation data was a better fit to the equation for the full dataset (fig. 18*B*). The results of the logistic regression on the full dataset show that when turbidity is less than 5 FNU during stormflow, there is on average a 23 to 45 percent chance that the *E. coli* density in a sample will exceed 235 colonies/100 mL (fig. 18*A*). Furthermore, when turbidity exceeds 30 FNU during stormflow, there is on average at least a 68 percent chance that water samples will contain *E. coli* densities that exceed the beach criterion. Conversely, when turbidity is less than 5 FNU during dry-weather flow, there is on average a 2 percent chance that the *E. coli* density in a sample will exceed 235 colonies/100 mL. Moreover, when turbidity exceeds 30 FNU during dry-weather flow, there is on average a 9 percent chance that the *E. coli* density in a sample will exceed 235 colonies/100 mL.

Table 11. Summary statistics for the binary logistic regression of turbidity against mean *Escherichia coli (E. coli)* densities above or below the U.S. Environmental Protection Agency's single-sample beach criterion of 235 *E. coli* colonies per 100 milliliters of water (table 1) for water samples collected at the Chattahoochee River near Norcross, Georgia (USGS station number 02335000), October 23, 2000, through September 30, 2008.

[LR, likelihood ratio; c, probability of concordance; Dxy, Somers' Dxy rank correlation coefficient; R^2, coefficient of determination; SE, standard error; FNU, formazin nephelometric unit]

	Regression indices						
Observations	LR[a]	Deviance[b]	c[c]	Dxy[d]	R^{2}[e]	Brier[f]	*p*-value[g]
1,417	714	0.848	0.945	0.890	0.693	0.045	<0.001
1,328[h]	856	.772	.992	.983	.884	.020	<.001

Regression coefficients, all samples				
Terms	Coefficient	SE	Wald Z	*p*-value
Intercept	–4.000	0.201	–19.91	<0.001
EVENT[i]	3.011	.292	10.30	<.001
Turbidity (FNU)	.059	.011	5.51	<.001

Regression coefficients, outliers removed				
Terms	Coefficient	SE	Wald Z	*p*-value
Intercept	–7.981	1.041	–7.67	<0.001
EVENT[i]	6.044	1.028	5.22	<.001
Turbidity (FNU)	.111	.021	5.88	<.001

[a] Global log likelihood ratio statistic, used to test the importance of all predictor variables in the model.

[b] Deviance statistic is the regression deviance divided by the degrees of freedom.

[c] This index measures the rank correlation between predicted probabilities of response and the observed response. Derived from the Wilcoxon-Mann-Whitney two-sample rank test (Harrell, 2001).

[d] Measures the rank correlation between predicted probabilities and observed responses; in other words, the difference between concordance and discordance probabilities. A value of 0 indicates the model is making random predictions and a value of 1.0 indicates the model predictions have perfect concordance (Harrell, 2001).

[e] A measure of the predictive strength of the regression model.

[f] Brier's score, sums the difference between the predicted probabilities and the observed responses. The smaller the value, the smaller the differences between the predicted probabilities and the observed responses.

[g] Probability that the model can not explain the variability of the response variable as a function of the variability in the independent variable.

[h] Eighty-nine samples removed as outliers.

[i] Variable indicating streamflow regime in which water samples were collected dry-weather flow or stormflow; table 3.

Table 12. Proportion of ambient water samples with mean *Escherichia coli (E. coli)* bacteria densities exceeding the U.S. Environmental Protection Agency's single-sample beach criterion of 235 *E. coli* colonies per 100 milliliters of water by turbidity range for the Chattahoochee River near Norcross, Georgia (USGS station number 02335000), October 23, 2000, through September 30, 2008, and October 1, 2008, through September 30, 2009.

[FNU, formazin nephelometric unit; USEPA, U.S. Environmental Protection Agency; —, not applicable]

		October 23, 2000, through September 30, 2008						October 1, 2008 through September 30, 2009		
		Full dataset (1,417 samples)			Dataset with outliers removed (1,328 samples)			Validation dataset		
Turbidity range (FNU)	Streamflow regime, EVENT	Number of samples	Number of samples exceeding USEPA beach criterion	Proportion of samples exceeding criterion (percent)	Number of samples	Number of samples exceeding USEPA beach criterion	Proportion of samples exceeding criterion (percent)	Number of samples	Number of samples exceeding USEPA beach criterion	Proportion of samples exceeding criterion (percent)
0 to 5	Dry-weather flow	625	7	1	601	0	0	73	1	1
	Stormflow	0	—	—	0	0	—	4	2	50
	All	625	7	1	601	0	0	77	5	7
>5 to 10	Dry-weather flow	379	13	3	352	0	0	17	1	6
	Stormflow	16	5	31	15	4	27	6	3	50
	All	395	18	5	367	4	1	23	4	17
>10 to 20	Dry-weather flow	137	8	6	121	1	1	0	—	—
	Stormflow	58	25	43	57	24	42	8	4	50
	All	195	33	17	178	25	14	8	4	50
>20 to 30	Dry-weather flow	19	2	11	17	0	0	0	—	—
	Stormflow	32	20	63	28	19	68	6	5	83
	All	51	22	43	45	19	42	6	5	83
>30 to 40	Dry-weather flow	9	2	22	6	0	0	0	—	—
	Stormflow	25	21	84	24	21	88	1	1	100
	All	34	23	68	30	21	70	1	1	100
>40 to 50	Dry-weather flow	0	—	—	2	0	0	0	—	—
	Stormflow	13	11	85	12	11	92	1	1	100
	All	13	11	85	14	11	79	1	1	100
>50 to 60	Dry-weather flow	1	0	0	0	0	—	0	—	—
	Stormflow	9	8	89	6	6	100	1	1	100
	All	10	8	80	6	6	100	1	1	100
>60 to 70	Dry-weather flow	0	—	—	0	—	—	0	—	—
	Stormflow	7	7	100	7	7	100	0	—	—
	All	7	7	100	7	7	100	0	—	—

Table 12. Proportion of ambient water samples with mean *Escherichia coli (E. coli)* bacteria densities exceeding the U.S. Environmental Protection Agency's single-sample beach criterion of 235 *E. coli* colonies per 100 milliliters of water by turbidity range for the Chattahoochee River near Norcross, Georgia (USGS station number 02335000), October 23, 2000, through September 30, 2008, and October 1, 2008, through September 30, 2009.—Continued

[FNU, formazin nephelometric unit; USEPA, U.S. Environmental Protection Agency; —, not applicable]

Turbidity range (FNU)	Streamflow regime, EVENT	October 23, 2000, through September 30, 2008						October 1, 2008 through September 30, 2009		
		Full dataset (1,417 samples)			Dataset with outliers removed (1,328 samples)			Validation dataset		
		Number of samples	Number of samples exceeding USEPA beach criterion	Proportion of samples exceeding criterion (percent)	Number of samples	Number of samples exceeding USEPA beach criterion	Proportion of samples exceeding criterion (percent)	Number of samples	Number of samples exceeding USEPA beach criterion	Proportion of samples exceeding criterion (percent)
>70 to 80	Dry-weather flow	0	—	—	0	—	—	0	—	—
	Stormflow	9	9	100	9	9	100	0	—	—
	All	9	9	100	9	9	100	0	—	—
>80 to 90	Dry-weather flow	0	—	—	0	—	—	0	—	—
	Stormflow	2	2	100	2	2	100	0	—	—
	All	2	2	100	2	2	100	0	—	—
>90 to 100	Dry-weather flow	0	—	—	0	—	—	0	—	—
	Stormflow	9	9	100	8	8	100	0	—	—
	All	9	9	100	8	8	100	0	—	—
>100 to 150	Dry-weather flow	0	—	—	0	—	—	0	—	—
	Stormflow	23	22	96	21	21	100	1	1	100
	All	23	22	96	21	21	100	1	1	100
>150 to 200	Dry-weather flow	0	—	—	0	—	—	0	—	—
	Stormflow	11	11	100	11	11	100	0	—	—
	All	11	11	100	11	11	100	0	—	—
>200 to 250	Dry-weather flow	0	—	—	0	—	—	0	—	—
	Stormflow	4	4	100	3	3	100	0	—	—
	All	4	4	100	3	3	100	0	—	—
>250 to 300	Dry-weather flow	0	—	—	0	—	—	0	—	—
	Stormflow	8	8	100	8	8	100	0	—	—
	All	8	8	100	8	8	100	0	—	—
>300	Dry-weather flow	0	—	—	0	—	—	0	—	—
	Stormflow	19	19	100	17	17	100	0	—	—
	All	19	19	100	17	17	100	0	—	—

The logistic regression equation in table 11 was validated by comparing the exceedance probabilities predicted by the equation with the computed proportion of validation samples containing *E. coli* densities that exceeded 235 MPN/100 mL (table 12). During dry-weather flow, the exceedance probabilities predicted from both logistic equations typically corresponded to the computed proportion of validation samples with *E. coli* densities that exceeded 235 MPN/100 mL (figs. 18*B*, *D*). During stormflow, however, the exceedance probabilities predicted from both logistic equations did not fit the computed proportion of validation samples with *E. coli* density that exceeded 235 MPN/100. The lack of fit between the predicted and computed probabilities in stormflow samples from the validation dataset is probably the result of the small number of samples in the validation dataset and in particular the relatively few stormflow samples collected because of the severe drought conditions that existed between October 1, 2008, and September 30, 2009 (fig. 3).

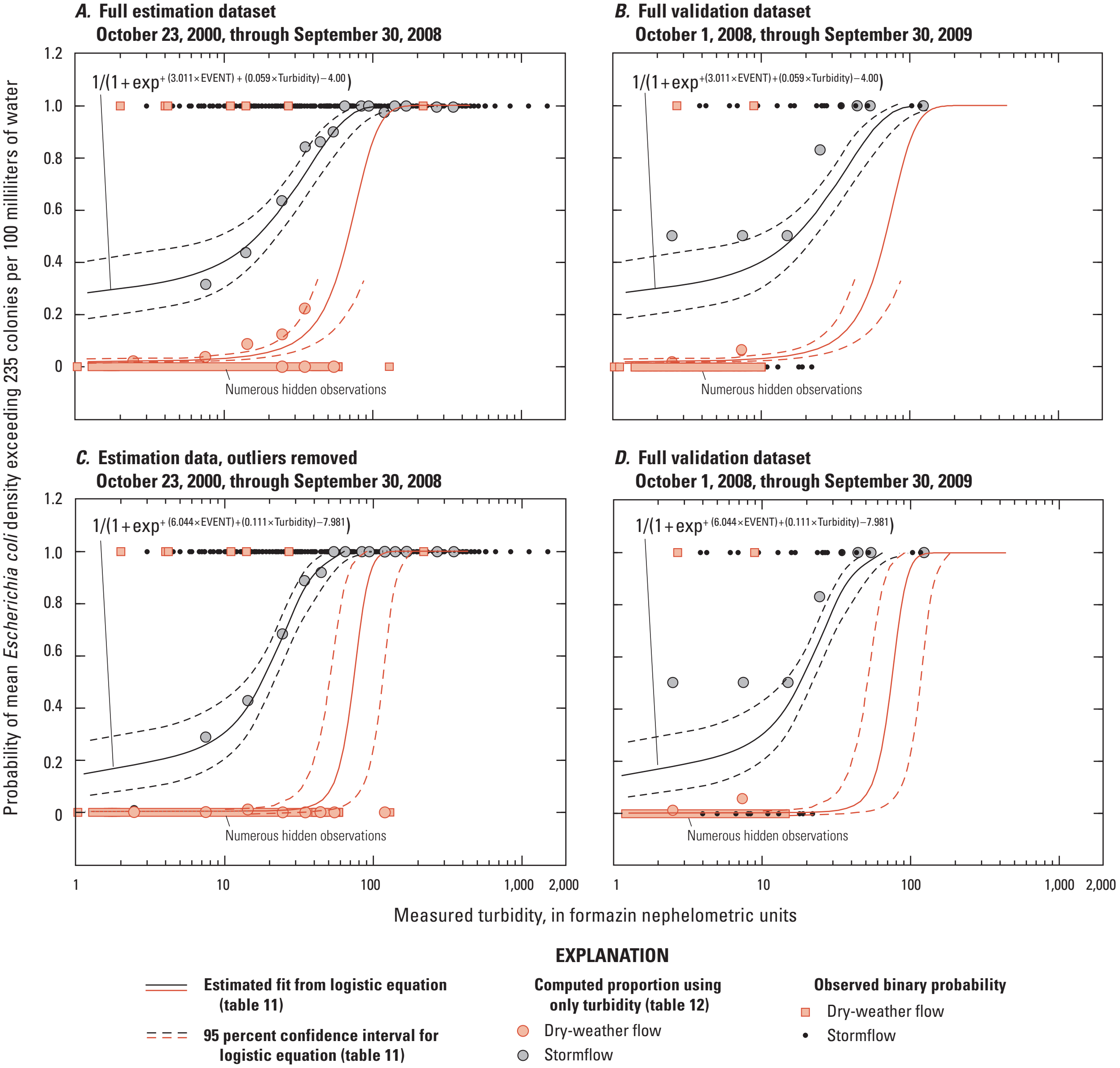


Figure 18. Logistic regression plots showing the probability that the mean *Escherichia coli (E. coli)* density in a sample from the Chattahoochee River near Norcross, Georgia (USGS station number 02335000), exceeds the U.S. Environmental Protection Agency's single-sample beach criterion of 235 *E. coli* colonies per 100 milliliters of water at specific turbidity values during regulated-flow or stormflow conditions. *(A)* Full estimation dataset, October 23, 2000, through September 30, 2008. *(B)* Full validation dataset with the logistic equation used for plot *(A)*, October 1, 2008, throughSeptember 30, 2009. *(C)* Estimation dataset without outliers, October 23, 2000, through September 30, 2008. (D) Full validation dataset with logistic regression equation used for plot *(C)*, October 1, 2008, through September 30, 2009.

Chattahoochee River at Atlanta, Georgia

The "best" one- and three-variable equations from the S-PLUS leaps and bounds procedure (USGS S-PLUS library) on the Atlanta data are given in table 13, including one equation with an interaction term. The interaction term is the cross product of log10FNU and EVENT that accounts for differences in the intercept and slope that exist in the relations between log10Ecoli and log10FNU at the two values for the indicator variable EVENT. For example, figure 19 shows that the y-intercept and slope coefficients for the relation between *E. coli* density and turbidity are markedly different between samples collected at Atlanta during dry-weather flow and stormflow. Without the interaction variable, these differences would not be captured by the regression and would be grouped into the overall error term for the regression.

Development and Validation of Linear Regression Equations

As with the Norcross data, the "best" one-variable regression equation developed for the Atlanta data consisted of turbidity in base 10 logarithms (log10FNU). This regression, regression-9, used the full Atlanta dataset with log10Ecoli as the response and log10FNU as the explanatory variables (table 13; appendix 6, fig. 6–1). Although regression-9 indicates a statistically significant linear relation between log10Ecoli and log10FNU (*p*-value less than 0.001), the adjusted R^2 was only 0.496, meaning that about 50 percent of the variability in log10Ecoli was explained by the variability in log10FNU, a low percentage for any equation expected to accurately predict the response variable. As with the Norcross data, autocorrelation due to seasonal

Table 13. Regression statistics for the best fit regression analysis on datasets containing mean *Escherichia coli* bacteria density, turbidity, and water temperature measured during stormflow and dry-weather flow at the Chattahoochee River at Atlanta, Georgia (USGS station number 02336000), October 23, 2000, through September 30, 2008.

[The explanatory variables in the table are selected based on the lowest Cp statistic in each group of explanatory variables using the S-PLUS leaps and bounds function; no outliers, original dataset minus outliers; randomized, dataset in which samples were randomly selected with replacement from the no outliers dataset; *Escherichia coli* bacteria density as most probable number per 100 milliliters of water; adjusted R^2, the coefficient of determination adjusted using the mean square error to negate the tendency for the R^2 to increase as the number of explanatory variables increases; VIF, variance inflation factor; AIC, Akaike Information Criterion, a measure of the goodness of fit; —, not applicable]

Regression number	Samples in dataset	Log10Ecoli in relation to significant explanatory variables	Number of samples	Adjusted R^2	Residual standard error	VIF[a]	Durbin-Watson statistic[b]	AIC	Figure number
9	All, October 23, 2000, through September 2008	Log10(FNU)	1,407	0.496	0.408	—	1.25	1,474	6–1
10	All, July 26, 2002, through September 2008	Log10(FNU), WTEMP[c], EVENT[d]	913	.669	.332	1.50	1.54	582	6–2
11	Outliers removed, July 26, 2002, through September 2008	Log10(FNU), WTEMP[c], EVENT[d]	855	.747	.260	1.51	1.53	127	6–3
12	Outliers removed, July 26, 2002, through September 2008	Log10(FNU), WTEMP[c], EVENT[d], [Log_{10}(FNU)×EVENT][e]	855	.758	.254	1.93	1.53	93	20

[a] Variance inflation factor, a measure of a variable's influence on the variance of the regression. A VIF below 5 is considered insignificant (Montgomery and others, 2006).

[b] Durban-Watson statistic is a measure of first-order autocorrelation among observations in the dataset. The farther the absolute value is from 2, the greater the autocorrelation.

[c] WTEMP, water temperature in degrees Celsius, in situ measurements not available before July 26, 2002.

[d] EVENT, indicator variable for the streamflow regime in which water samples were collected (dry-weather flow or stormflow; table 3).

[e] Interaction term.

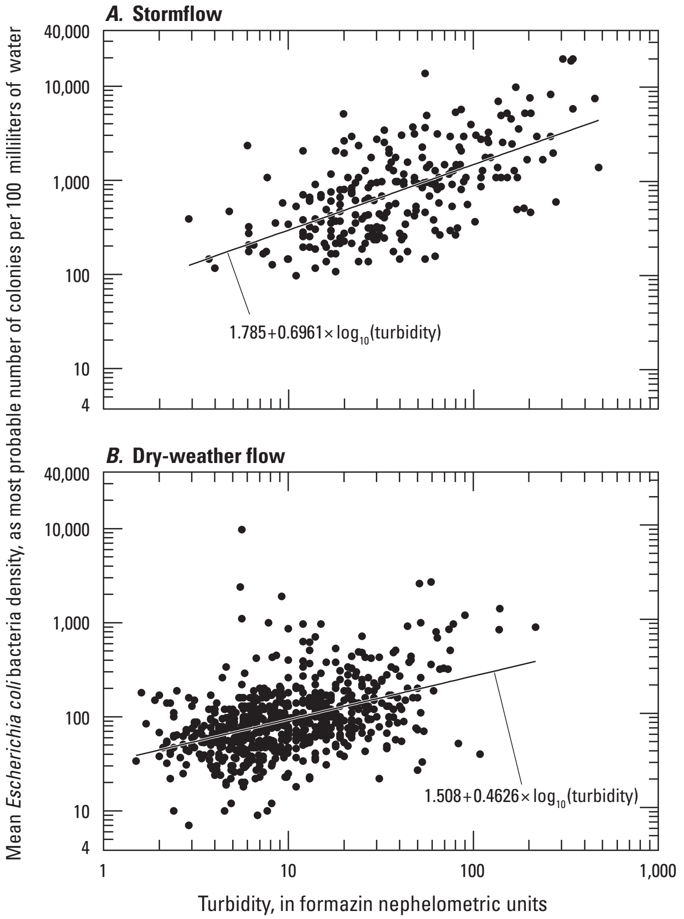


Figure 19. Linear relation between turbidity measurements and mean *Escherichia coli* bacteria density in dry-weather and stormflow samples from the Chattahoochee River at Atlanta, Georgia (USGS station number 02336000), October 23, 2000, through September 30, 2008.

effects and data redundancy was statistically significant when the Atlanta data were sorted by collection date and should not negatively affect the predicted *E. coli* densities using the chosen regression equation. The regression and residual plots, and regression statistics are shown in appendix 6 (fig. 6–1). The residual analyses show that regression-9 underestimates *E. coli* density, particularly at high turbidity values, and has numerous outliers.

Following the steps used during the regression analyses of the Norcross data, the “best” two-, three-, and four-variable equations were developed for the Atlanta data. The adjusted R^2 was less than 0.600 for these regressions on the full Atlanta dataset and the dataset with outliers removed (table 13). Because Georgia Power published water temperature data for Bull Sluice Lake that showed higher water temperatures in the lake than in the river (Georgia Power, 2004b), water temperature at Atlanta was included in the regression analyses as a potentially important explanatory variable. Unfortunately, only 913 of the 1,407 samples in the Atlanta dataset could be used for regressions involving water temperature because the measurement of continuous, instream water temperature only began on July 26, 2002. Using the leaps and bounds procedure in S-PLUS, water temperature was shown to account for a statistically significant amount of the variability in log10Ecoli values. A regression equation, regression-10, using log10FNU, water temperature, and EVENT as explanatory variables, explained about 67 percent of the variability in log10Ecoli values (table 13; appendix 6, fig. 6–2). Regression-10 represents a marked improvement in the ability to estimate *E. coli* density. The adjusted R^2 (0.669) shows that regression-10 explained about 17 percent more of the variability in log10Ecoli values than did regression-9. In addition, the residual standard error and AIC for the regression-10 analysis decreased by about 19 and 60 percent, respectively.

The residuals analysis for regression-10 revealed a substantial number of outliers in the Atlanta data (appendix 6, fig. 6–2). In addition, the axis of stormflow residuals does not parallel the axis of dry-weather flow residuals, which indicates an interaction term may improve the estimation power of regression-10 (appendix 6, fig. 6–2*B*). The diagnostic functions in S-PLUS were used to identify samples that showed high influence or high leverage in the regression. Samples were tagged as outliers if the absolute value of the studentized residual exceeded 1.9 and the absolute DFFITS value exceeded the computed critical value of 0.168 (Montgomery and others, 2006, p. 195). A total of 58 samples representing about 6 percent of the data were tagged as outliers and not used in subsequent regression analyses.

With outliers removed from the dataset, regressions-11 and 12 were statistically stronger equations than regressions-9 and 10, but regression-12 was a slightly stronger equation than regression-11 (table 13). Figure 20*A* shows that the estimated log10Ecoli values were strongly associated with the measured log10Ecoli values, that the residuals were randomly distributed with constant variance (fig. 20*B*), corresponded to a standard normal distribution (fig. 20*C*), and showed no apparent trend over the study period (fig. 20*D*). Table 14 gives the regression statistics and ANOVA for regression-12. Regression-12 appears to be the “best” equation for estimating median *E. coli* densities at the Atlanta site.

The validation dataset for the Atlanta site consisted of measured *E. coli* densities in water samples and continuous turbidity and water temperature measured instream from October 1, 2008, to September 30, 2009. During validation, diagnostic indices were used to determine if the regression equation could adequately predict median *E. coli* densities for the validation data from the Atlanta site. Diagnostic indices include Spearman’s rank correlation coefficients between measured and predicted *E. coli* densities, mean sum of squared prediction residuals (MSS_p), and prediction R^2.

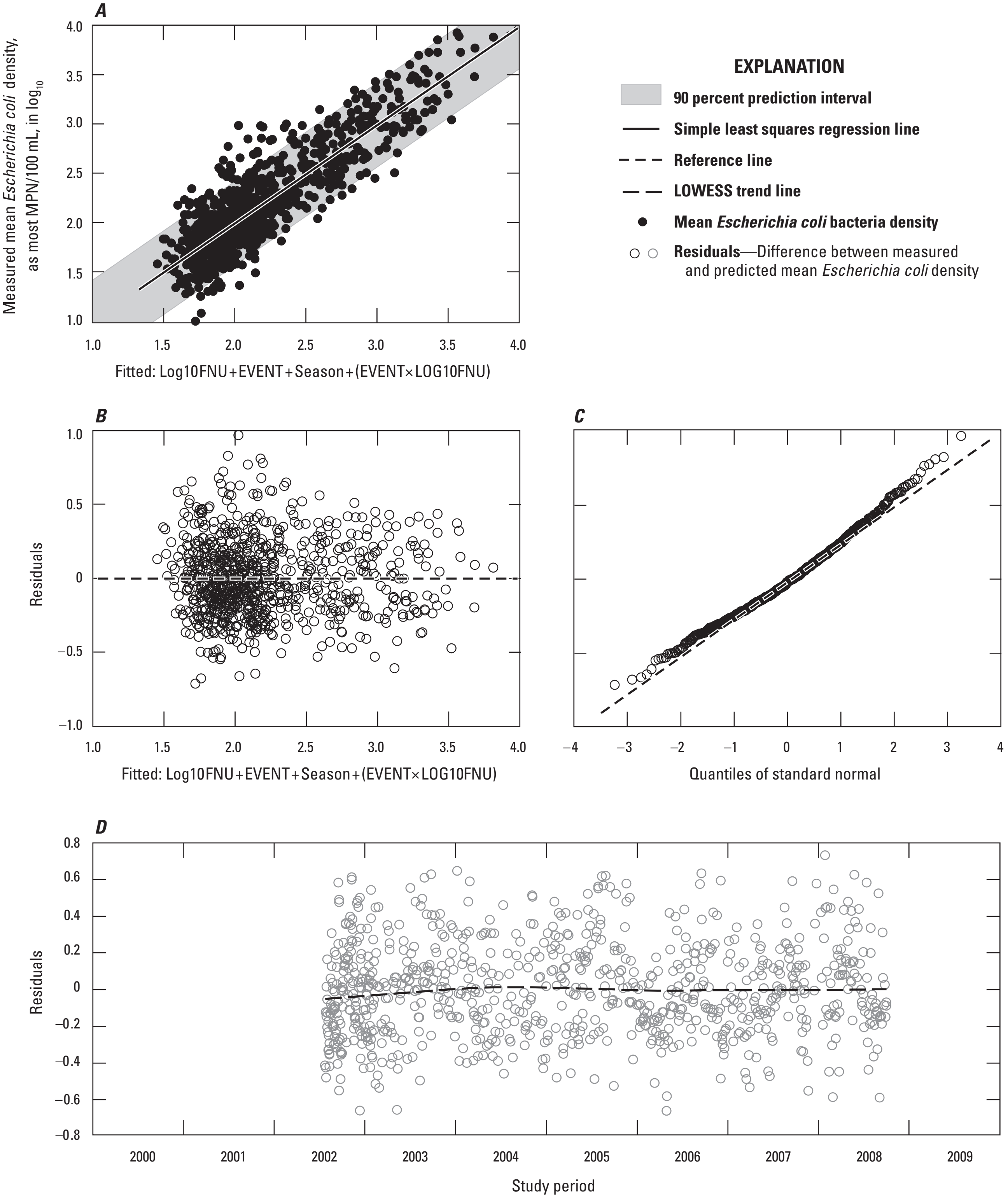


Figure 20. Diagnostic plots for regression-12 (outliers removed, table 13) using data collected at the Chattahoochee River at Atlanta, Georgia (USGS station number 02336000), July 26, 2002, through September 30, 2008. *(A)* Relation between measured and estimated mean *Escherichia coli* density in base 10 logarithm units (Log10*Ecoli*). *(B)* Relation between residuals and estimated Log10Ecoli). *(C)* Distribution of the residuals compared to a standard normal distribution. *(D)* Rhe study-period trend in residuals. Explanatory variables: Log10FNU, turbidity in formazin nephelometric units transformed to base 10 logarithms; EVENT, streamflow regime as dry-weather flow or stormflow; WTEMP, water temperature in degrees Celsius; and the interaction term (EVENT×Log10FNU); MPN/100 mL, most probable number of colonies per 100 milliliters of water.

Table 14. Results and analysis of variance for the linear regression analysis (regression-12, outliers removed, table 13) on *Escherichia coli* bacteria densities for samples from the Chattahoochee River at Atlanta, Georgia (USGS station number 02336000), July 26, 2002, through September 30, 2008.

[Adjusted R^2, the coefficient of determination adjusted using the mean square error to negate the tendency for the R^2 to increase as the number of explanatory variables increases; *Escherichia coli* bacteria density, as most probable number of colonies (MPN) per 100 milliliters of water; SE, standard error of regression; t-statistic, used to determine if the coefficient is statistically equal to zero; WTEMP, water temperature in degrees celsius; residual, the difference between the actual and predicted value of the response variable; —, not applicable]

Residual standard error is 0.254 and the adjusted R^2 is 0.758 Response variable: Log10Ecoli				
Terms	**Coefficient**	**Standard error (SE)**	**t-statistic**[a]	***p*-value**[b]
Intercept	1.166	0.040	29.47	<0.001
Log10(FNU)	.521	.027	19.44	.001
WTEMP	.020	.002	10.94	<.001
EVENT[c]	.069	.079	.88	.380
[Log10(FNU)×EVENT][d]	.316	.052	6.06	<.001

Analysis of Variance					
Terms	**df**[e]	**Sum of squares (SS)**	**Mean SS**	**F statistic**[f]	***p*-value (F)**
Log10(FNU)	1	131.0	131.0	2,024	<0.001
WTEMP	1	10.0	10.0	154	<.001
EVENT[c]	1	29.6	29.6	457	<.001
[Log10(FNU)×EVENT][d]	1	2.4	2.4	37	<.001
Residuals	850	55.0	.06	—	—

Autocorrelation coefficients	
Number of samples lagged	**Correlation coefficient**[g]
0	1.00
1	.23
2	.15
3	.16

[a] t-statistic, used to determine if the coefficient is statistically equal to zero.

[b] *p*-value, the probability that the parameter is not important to the regression.

[c] Variable indicating streamflow regime in which water samples were collected (dry-weather flow or stormflow; table 3).

[d] Interaction term.

[e] Defined as the number of independent pieces of information used to calculate the statistics.

[f] F statistic, used to determine if there is a significant linear relation between the response variable and the explanatory variables.

[g] The critical value is 0.20. Coefficients greater than 0.20 are significant at alpha equal to 0.10.

Measured *E. coli* densities in water samples from the Atlanta validation dataset were compared to those computed using regression-12. The diagnostic indices indicate that regression-12 is a good predictor of *E. coli* densities in the validation dataset. The prediction R^2 was 0.704 (*p*-value less than 0.001), Spearman's rank correlation coefficient was 0.588 (*p*-value less than 0.001), and the MSS_p (0.117) was nearly twice the MSS_f for the estimation dataset (table 14; 0.06). The graph of measured and predicted *E. coli* densities shows that the validation data plots within the 90 percent prediction interval computed from regression-12 (fig. 21*A*). Prediction residuals were generally normally distributed with constant variance, but indicated that a few outliers were present (fig. 21*B*); the distribution of the prediction residuals shows a negative skew that indicates log10Ecoli was underestimated when measured turbidities were low during dry-weather flow (fig. 21*C*). Because of the small size of the validation dataset from Atlanta, predicted *E. coli* densities were not as robust as they probably would be with a much larger dataset.

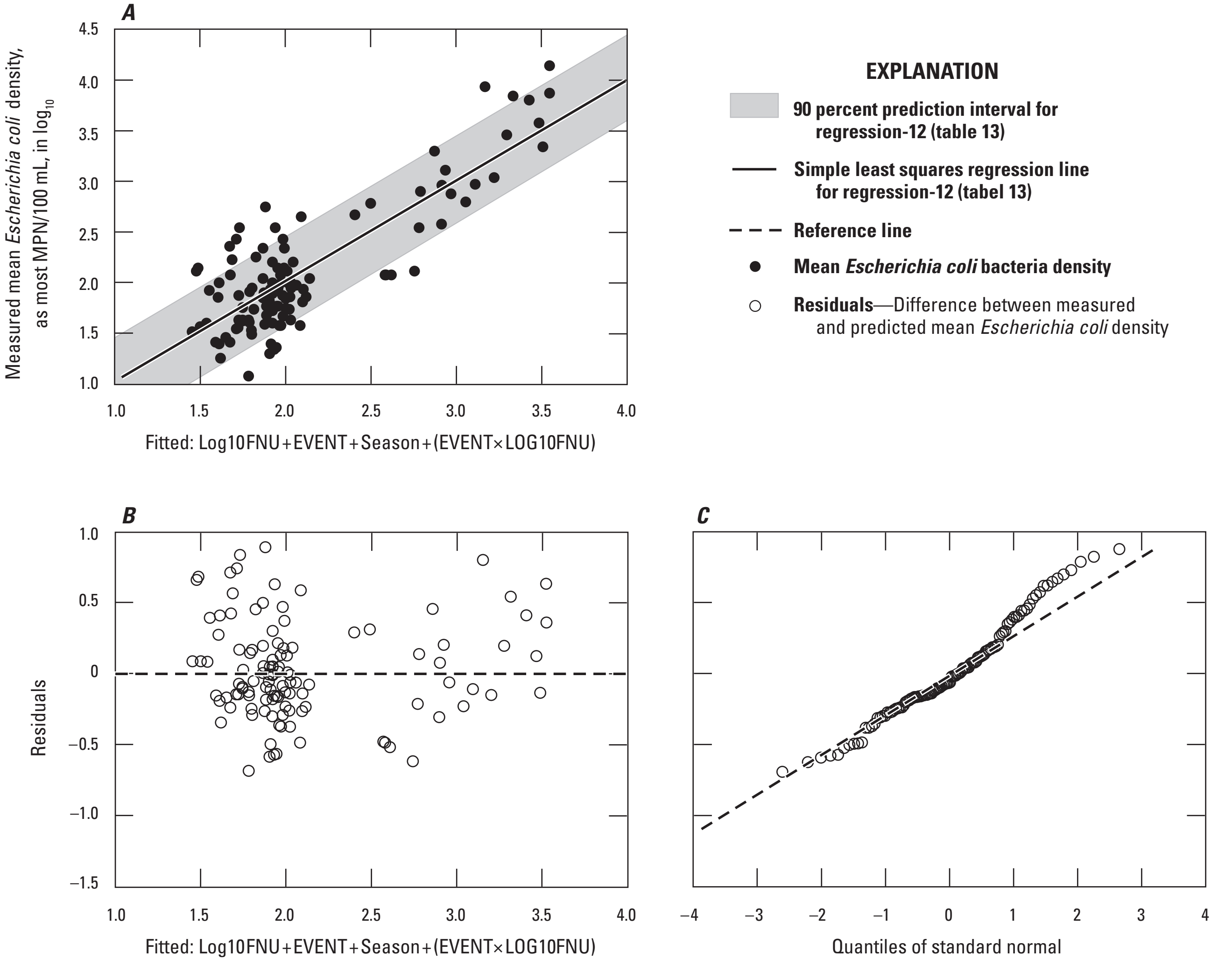


Figure 21. Diagnostic plots for the validation analysis of regression-12 (outliers removed, table 13) using data collected at the Chattahoochee River at Atlanta, Georgia (USGS station number 02336000), October 1, 2008, through September 30, 2009. *(A)* Relation between measured and predicted mean *Escherichia coli* bacteria *(E. coli)* densities. *(B)* Relation between residuals and the predicted mean *E. coli* densities. *(C)* Distribution of the residuals compared to a standard normal distribution. Explanatory variables: Log10FNU, turbidity in formazin nephelometric units transformed to base 10 logarithms; EVENT, streamflow regime as dry-weather flow or stormflow; WTEMP, water temperature in degrees Celsius; and the interaction term (EVENT×Log10FNU); MPN/100 mL, most probable number of colonies per 100 milliliters of water.

Development and Validation of Logistic Regression Equations

Logistic regression analysis was used to estimate the probability that *E. coli* density at the Atlanta site will exceed the USEPA single-sample beach criterion of 235 colonies/100 mL for a given turbidity measurement. Turbidity values and the binary variable EVENT were explanatory variables used in the logistic regression analysis to predict the binary outcome of *E. coli* density, that is, whether *E. coli* density is below (0) or above (1) 235 colonies/100 mL. The variable for water temperature that was included in regression-12 is omitted for the logistic regression analysis because measured water temperatures were associated with only 65 percent of the 1,407 samples collected at Atlanta. The full complement of water samples collected during the study was needed to produce a logistic regression that had the greatest predictive power.

The values of five regression indices indicate that turbidity and EVENT have a strong statistical relation to the probability of *E. coli* density exceeding 235 colonies/100 mL in water samples from Atlanta (table 15). The five indices include the deviance statistic, probability of concordance (c), Somer's D_{xy} rank correlation, R^2, and Brier's score and are defined as previously given. These indices are markedly different from the indices for the full dataset from Norcross. The c and D_{xy} indices (0.916 and 0.831, respectively) indicate a high probability of concordance between observed and predicted probabilities for the Atlanta data (table 15); however, the c index is smaller than that for the Norcross data, indicating slightly more discordance in the Atlanta data (tables 11 and 15). The R^2 of 0.659 indicates the logistic equation for the Atlanta data has a moderate predictive ability that is slightly weaker than the predictive ability for the full dataset from Norcross. The Brier's score for the Atlanta data (0.087)

Table 15. Summary statistics for the binary logistic regression of turbidity against mean *Escherichia coli (E. coli)* densities above or below the U.S. Environmental Protection Agency's single-sample beach criterion of 235 *E. coli* colonies per 100 milliliters of water for water samples collected at the Chattahoochee River at Atlanta, Georgia (USGS station number 02336000), October 23, 2000, through September 30, 2008.

[LR, likelihood ratio; c, probability of concordance; Dxy, Somers' Dxy rank correlation coefficient; R^2, coefficient of determination; event, indicator variable identifying non-storm regulated flow and stormflow; SE, standard error]

Regression indices						
Observations	**LR**[a]	**c**[b]	**Dxy**[c]	**$R^{2/d}$**	**Brier**[e]	***p*-value**[f]
1,407	882	0.916	0.831	0.659	0.087	<0.001

Regression coefficients				
Terms	**Coefficient**	**SE**	**Wald Z**	***p*-value**
Intercept	–3.098	0.155	–19.93	<0.001
Turbidity, FNU	.060	.006	10.19	<.001
EVENT[g]	3.332	.210	15.90	<.001

[a] Global log likelihood ratio statistic, used to test the importance of all predictor variables in the model.

[b] This index measures the rank correlation between predicted probabilities of response and the observed response. Derived from the Wilcoxon-Mann-Whitney two-sample rank test (Harrell, 2001).

[c] Measures the rank correlation between predicted probabilities and observed responses; in other words, the difference between concordance and discordance probabilities. A value of 0 indicates the model is making random predictions and a value of 1.0 indicates the model predictions have perfect concordance Harrell, 2001).

[d] A measure of the predictive strength of the regression model.

[e] Brier's score, sums the difference between the predicted probabilities and the observed responses. The smaller the value, the smaller the differences between the predicted probabilities and the observed responses.

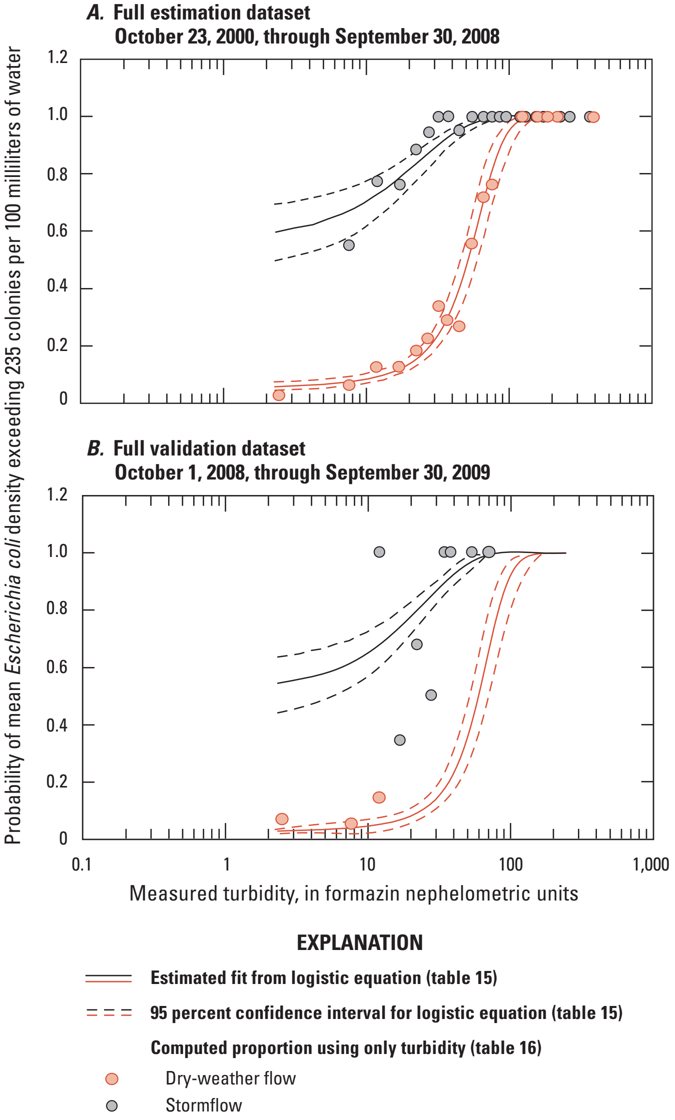
[f] Probability that the model can not explain the variability of the response variable as a function of the variability in the independent variable.

[g] EVENT, streamflow regime under which samples were collected (dry-weather flow or stormflow).

indicates the differences between the observed and predicted probabilities were small, but larger than the differences for the full dataset from Norcross.

The results of the logistic regression for the Atlanta site show that when turbidity is less than 5 FNU during stormflow, there is on average about a 60 percent chance that the *E. coli* density in a sample will exceed 235 colonies/100 mL (fig. 22*A*). Furthermore, when turbidity exceeds 30 FNU during stormflow, there is on average about a 95-percent chance that water samples will contain *E. coli* densities that exceed the criterion. In contrast, when turbidity is less than 5 FNU during dry-weather flow, there is on average less than a 3 percent chance that the *E. coli* density in a water sample will exceed 235 colonies/100 mL. Moreover, during dry-weather flow there is on average a 60 percent chance of exceeding the USEPA beach criterion when turbidity reaches about 60 FNU at the Atlanta site.

The logistic regression equation was validated by comparing the computed proportion of validation samples from Atlanta with *E. coli* densities that exceeded 235 colonies/100 mL (table 16; fig. 22*B*) to the probabilities predicted by the logistic equation described in table 15 (Harrell, 2001, p. 231). Among dry-weather-flow samples, the computed proportions align closely with the logistic regression line and plot within the 95-percent confidence interval shown in figure 22*B*; however, because of the small size of the validation dataset, the computed proportions only validate turbidity values that were less than 18 FNU. In contrast, the computed proportions for stormflow samples in the validation dataset were poorly predicted with the logistic regression equation. This disparity is probably the result of the weaker logistic equation for stormflow samples and the small number of stormflow samples (24) collected between October 1, 2008, and September 30, 2009. The small number of samples collected was due to severe drought in the study area during that time period (figs. 2 and 4). Although water samples were collected from the Atlanta site and analyzed for *E. coli* bacteria, real-time measurements of turbidity and water temperature were discontinued during the 2010 water year[1] because the bridge upon which the water-quality sonde was deployed needed rebuilding and repair. When additional data are available to increase the size of the validation dataset and the number of computed probabilities, the logistic regression equation can again be compared with those data.


***A.* Full estimation dataset**
October 23, 2000, through September 30, 2008

***B.* Full validation dataset**
October 1, 2008, through September 30, 2009

Figure 22. Logistic regression plots for the probability that the mean *Escherichia coli (E. coli)* density in a water sample from the Chattahoochee River at Atlanta, Georgia (USGS station number 02336000), exceeds the U.S. Environmental Protection Agency's single-sample beach criterion of 235 *E. coli* colonies per 100 milliliters of water at various turbidity values during dry-weather flow and stormflow. *(A)* Full estimation data, October 23, 2000, through September 30, 2008. *(B)* Full validation dataset, October 1, 2008, through September 30, 2009.

[1] Water year is the period October 1 through September 30 and is identified by the year in which the period ends.

Table 16. Proportion of ambient water samples with *Escherichia coli (E. coli)* bacteria densities exceeding the U.S. Environmental Protection Agency's single-sample beach criterion of 235 *E. coli* colonies per 100 milliliters of water by turbidity range for the Chattahoochee River at Atlanta, Georgia (USGS station number 02336000), for the periods October 23, 2000, through September 30, 2008, and October 1, 2008, through September 30, 2009.

[FNU, formazin nephelometric unit; USEPA, U.S. Environmental Protection Agency; —, not applicable]

Turbidity range (FNU)	Streamflow regime EVENT	October 23, 2000, through September 30, 2008			October 1, 2008, through September 30, 2009		
		Number of samples	Number of samples exceeding USEPA beach criterion	Proportion of samples exceeding criterion (percent)	Number of samples	Number of samples exceeding USEPA beach criterion	Proportion of samples exceeding criterion (percent)
0 to 5	Dry-weather flow	180	6	3	41	3	7
	Stormflow	0	—	—	0	0	0
	All	180	6	3	41	3	7
>5 to 10	Dry-weather flow	446	26	6	38	2	5
	Stormflow	15	8	53	0	0	0
	All	471	34	7	38	2	5
>10 to 15	Dry-weather flow	176	23	13	7	1	14
	Stormflow	38	29	76	1	1	100
	All	214	52	24	8	2	25
>15 to 20	Dry-weather flow	88	11	13	3	0	0
	Stormflow	36	29	81	3	1	33
	All	124	40	32	6	1	17
>20 to 25	Dry-weather flow	54	10	19	1	0	0
	Stormflow	26	23	88	3	2	67
	All	80	33	41	4	2	50
>25 to 30	Dry-weather flow	36	8	22	0	0	0
	Stormflow	24	23	96	2	1	50
	All	60	31	52	2	1	50
>30 to 35	Dry-weather flow	26	9	35	0	0	0
	Stormflow	22	22	100	1	1	100
	All	48	31	65	1	1	100
>35 to 40	Dry-weather flow	18	5	28	0	0	0
	Stormflow	11	11	100	3	3	100
	All	29	16	55	3	3	100
>40 to 50	Dry-weather flow	22	5	23	0	0	0
	Stormflow	28	27	96	1	1	100
	All	50	32	64	1	1	100
>50 to 60	Dry-weather flow	9	5	56	0	0	0
	Stormflow	23	23	100	0	0	0
	All	32	28	88	0	0	0
>60 to 70	Dry-weather flow	7	5	71	0	0	0
	Stormflow	18	18	100	0	0	0
	All	25	23	92	0	0	0
>70 to 80	Dry-weather flow	8	6	75	0	0	0
	Stormflow	12	12	100	1	1	100
	All	18	17	94	1	1	100

Table 16. Proportion of ambient water samples with *Escherichia coli (E. coli)* bacteria densities exceeding the U.S. Environmental Protection Agency's single-sample beach criterion of 235 *E. coli* colonies per 100 milliliters of water by turbidity range for the Chattahoochee River at Atlanta, Georgia (USGS station number 02336000), for the periods October 23, 2000, through September 30, 2008, and October 1, 2008, through September 30, 2009.—Continued

[FNU, formazin nephelometric unit; USEPA, U.S. Environmental Protection Agency; —, not applicable]

Turbidity range (FNU)	Streamflow regime EVENT	October 23, 2000, through September 30, 2008			October 1, 2008, through September 30, 2009		
		Number of samples	Number of samples exceeding USEPA beach criterion	Proportion of samples exceeding criterion (percent)	Number of samples	Number of samples exceeding USEPA beach criterion	Proportion of samples exceeding criterion (percent)
>80 to 90	Dry-weather flow	0	—	—	0	0	0
	Stormflow	13	13	100	2	2	100
	All	15	14	93	2	2	100
>90 to 100	Dry-weather flow	0	0	0	0	0	0
	Stormflow	4	4	100	1	1	100
	All	4	4	100	1	1	100
>100 to 125	Dry-weather flow	0	0	0	0	0	0
	Stormflow	16	14	88	2	2	100
	All	16	14	88	2	2	100
>125 to 150	Dry-weather flow	1	1	100	0	0	0
	Stormflow	11	11	100	2	2	100
	All	12	12	100	2	2	100
>150 to 175	Dry-weather flow	2	2	100	0	0	0
	Stormflow	10	10	100	2	2	100
	All	12	12	100	2	2	100
>175 to 200	Dry-weather flow	1	1	100	0	0	0
	Stormflow	7	7	100	0	0	0
	All	8	8	100	0	0	0
>200 to 250	Dry-weather flow	2	2	100	0	0	0
	Stormflow	6	6	100	0	0	0
	All	8	8	100	0	0	0
>250 to 300	Dry-weather flow	1	1	100	0	0	0
	Stormflow	4	4	100	0	0	0
	All	5	5	100	0	0	0
>300	Dry-weather flow	1	1	100	0	0	0
	Stormflow	5	5	100	0	0	0
	All	6	6	100	0	0	0

Predictive Modeling of *Escherichia coli* Density

The regression equations developed for the Norcross and Atlanta sites to predict *E. coli* density in real time require that several variables must either be available from instream water-quality sondes or computed from those variables in real time. The indicator variable EVENT used in the regression equations for both sites must be calculated by using measurements of streamflow or stream stage, or with an estimation equation. In addition, the value of the indicator variable Season is determined using the instream sonde's date stamp within the sonde to identify the period of cool weather (October 16 to April 15) or warm weather (April 16 to October 15).

Chattahoochee River near Norcross, Georgia

At the Norcross site, regression-6 (tables 9, 11) is the preferred equation with which to model *E. coli* densities because it is computationally simpler than and statistically similar to regression-8. The value for the indicator variable EVENT in regression-6 is computed with either a logistic regression equation using turbidity and streamflow at the Norcross site (fig. 23) or by comparing the gage height at the USGS streamgaging station on Suwanee Creek at Suwanee, GA (USGS station number 02334885), upstream from the Norcross site. The differences in the measured streamflow and turbidity during dry-weather flow and stormflow (fig. 9) can be exploited using logistic regression to determine whether dry-weather flow or stormflow conditions exist. This logistic

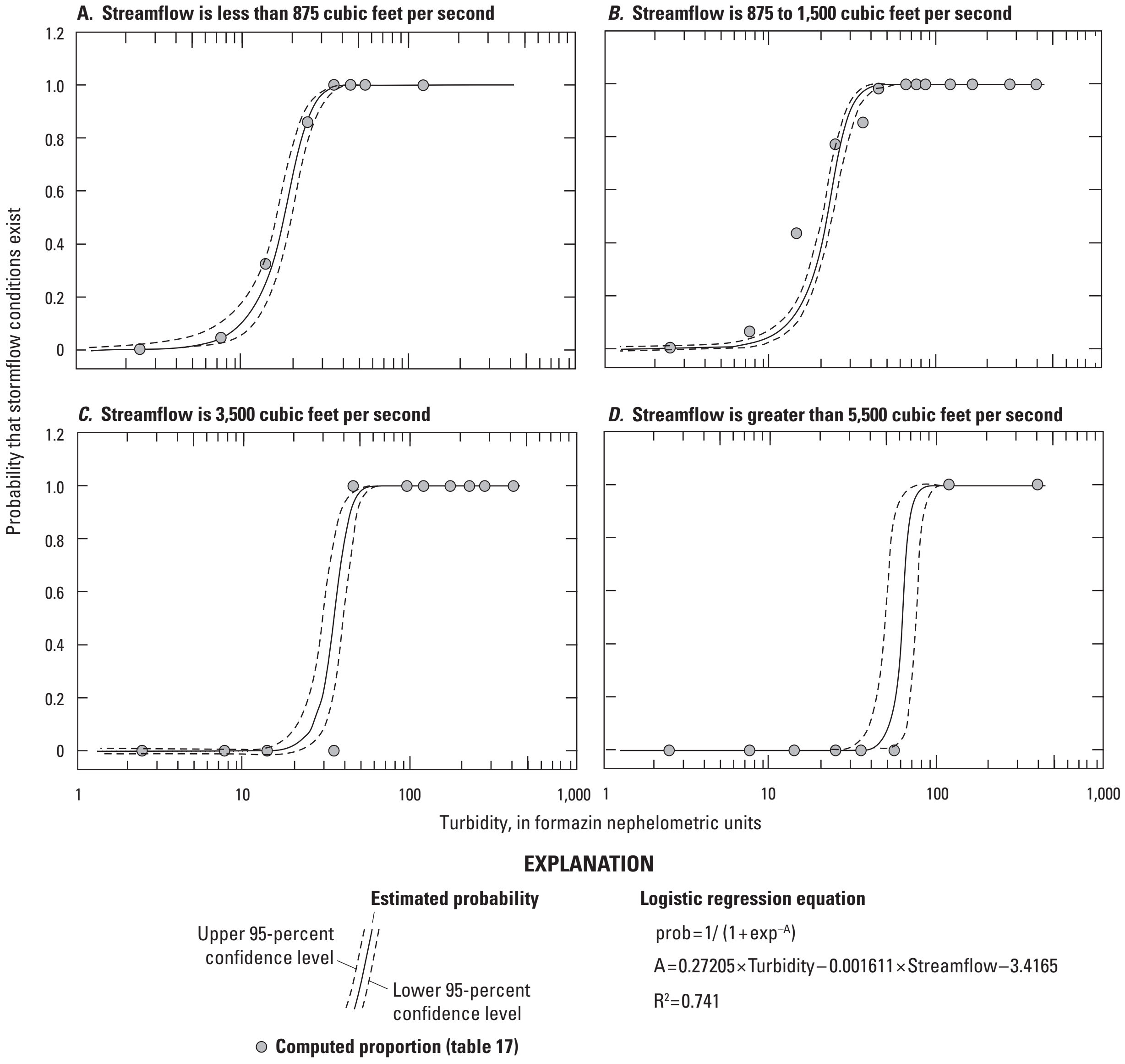


Figure 23. Logistic regression results estimating the probability that a sample from the Chattahoochee River near Norcross, Georgia (USGS station number 02335000), was collected during stormflow using only turbidity and streamflow measurements, October 23, 2000, through September 30, 2008.

regression is shown in figure 23, and computed proportions are given in table 17. The c score (0.971), Sommer's D_{xy} (0.941), R^2 (0.741), Brier's score (0.048), and *p*-value (less than 0.001) indicate this logistic regression equation can predict the probability of stormflow conditions using turbidity and streamflow measurements at the Norcross site. In addition, computed probabilities for various turbidity and streamflow ranges (table 17) agree closely with predicted EVENT values from the logistic regression (fig. 23). Results from the logistic regression will be supplemented by gage height (stage) measurements from the streamgaging station at Suwanee Creek. For example, if the probability of stormflow is between 30 and 70 percent, then the real-time stage of Suwanee Creek will be used to determine if there is storm runoff. Gage height from the Suwanee Creek gaging station will be read into a variable array at 15-minute intervals to determine if streamflow in Suwanee Creek is increasing due to storm runoff.

Table 17. Percentage of stormflow samples by turbidity and streamflow ranges for the Chattahoochee River near Norcross, Georgia (USGS station number 02335000), October 23, 2000, through September 30, 2008, and October 1, 2008, through September 30, 2009.

[FNU, formazin nephelometric unit; ft^3/s, cubic foot per second; —, category not represented by samples]

Turbidity range (FNU)	Streamflow range (ft^3/s)						
	<875	875–1,500	1,501–2,500	2,501–3,500	3,501–4,500	4,501–5,500	>5,500
	Percentage of water samples (number of water samples)						
0 to 5	0 (331)	0 (246)	0 (40)	0 (8)	—	—	—
>5 to 10	4.0 (173)	4.6 (174)	3.1 (32)	0 (9)	0 (4)	0 (1)	0 (2)
>10 to 20	33 (64)	43 (75)	21 (24)	0 (4)	0 (4)	0 (12)	0 (12)
>20 to 30	85 (13)	79 (24)	40 (5)	—	0 (4)	0 (1)	0 (4)
>30 to 40	100 (8)	83 (12)	87 (8)	0 (1)	0 (1)	0 (3)	0 (1)
>40 to 50	100 (3)	89 (9)	100 (1)	100 (1)	—	—	0 (1)
>50 to 60	100 (1)	100 (5)	100 (3)	—	—	—	0 (1)
>60 to 70	—	100 (4)	100 (3)	—	—	—	—
>70 to 80	—	100 (9)	—	—	—	—	—
>80 to 90	—	100 (2)	—	—	—	—	—
>90 to 100	—	100 (9)	100 (1)	100 (2)	—	—	—
>100 to 150	100 (2)	100 (10)	100 (6)	100 (2)	—	100 (1)	100 (2)
>150 to 200	—	100 (4)	100 (3)	100 (2)	100 (2)	—	—
>200 to 250	—	—	100 (2)	100 (1)	100 (1)	—	—
>250 to 300	—	100 (1)	100 (4)	100 (3)	—	—	—
>300	—	100 (1)	100 (7)	100 (6)	100 (1)	100 (2)	100 (2)

Figure 24 shows a graphical depiction of the proposed model to predict *E. coli* density in real time at the Norcross site. In this model, turbidity and streamflow measurements are collected every 15 minutes from the instream water-quality sonde at the Norcross site and entered into the logistic regression equation in figure 23 to determine the EVENT value. If the probability that EVENT equals stormflow is greater than 70 percent, then the EVENT variable will equal 1; however, if the probability is between 30 and 70 percent, then the current gage height at Suwanee Creek will be compared to previous gage height measurements, and if the absolute difference between those measurements is greater than 0.3 ft, then the EVENT variable will equal 1 to indicate stormflow. Furthermore, if the absolute difference between the current gage height and previous 15-minute measurements at Suwanee Creek is less than 0.3 ft, then the EVENT variable will equal 0 to indicate nonstorm, dry-weather flow conditions.

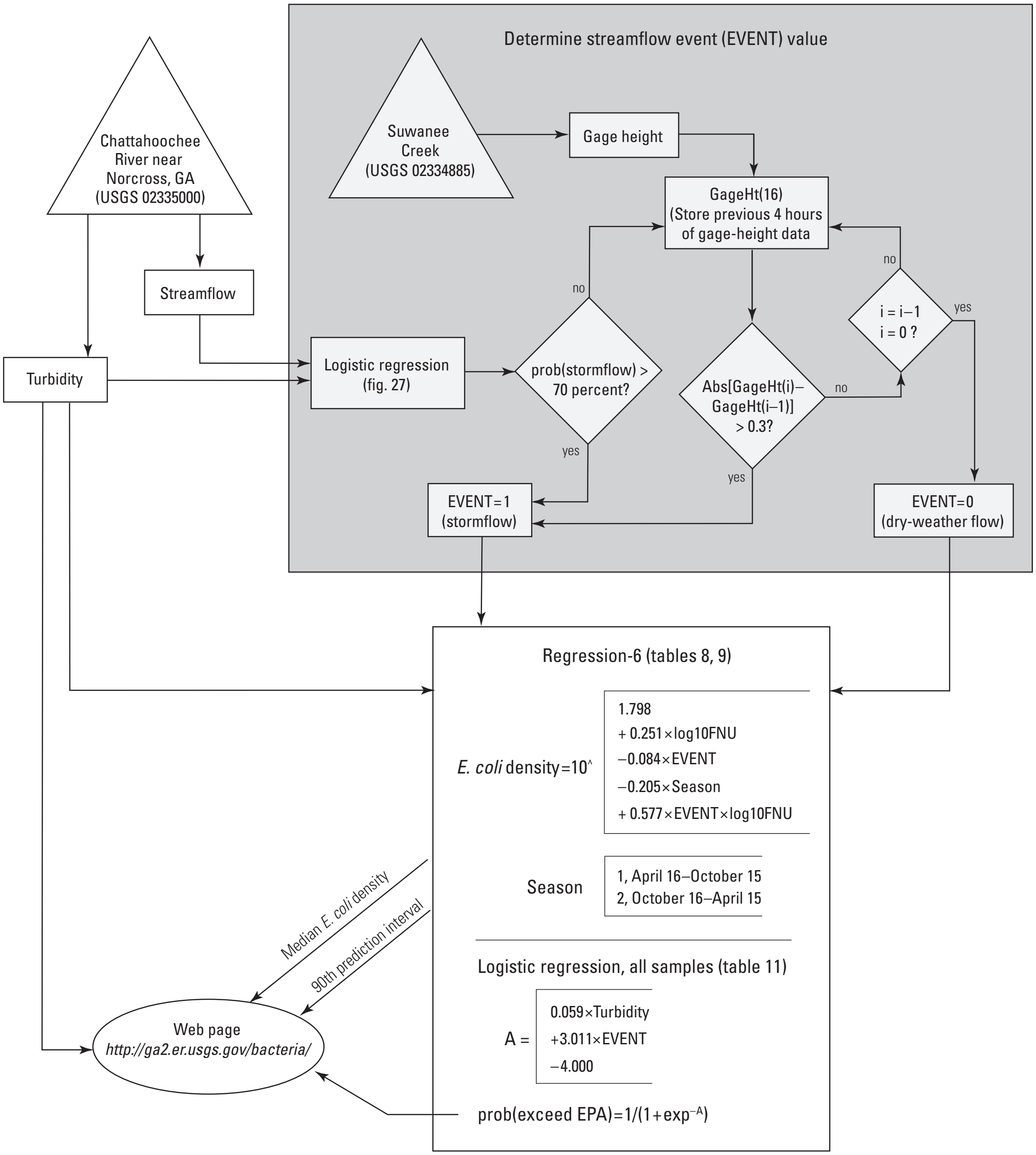
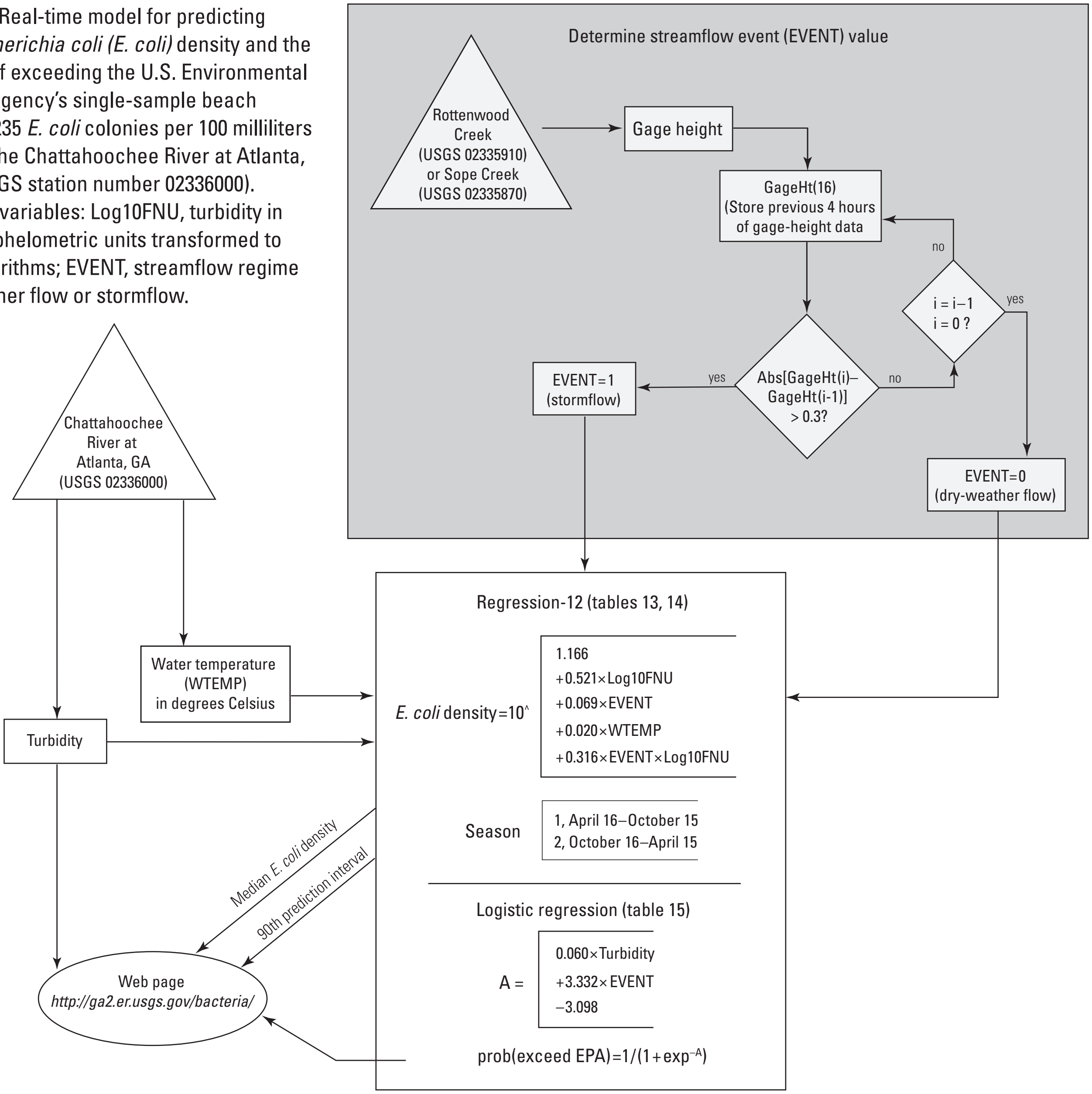

Determine streamflow event (EVENT) value
Chattahoochee River near Norcross, GA (USGS 02335000)
Suwanee Creek (USGS 02334885)
Gage height
GageHt(16) (Store previous 4 hours of gage-height data
no
Streamflow
no
i = i–1
i = 0 ?
yes
Turbidity
Logistic regression (fig. 27)
prob(stormflow) > 70 percent?
Abs[GageHt(i)–GageHt(i–1)] > 0.3?
no
yes
yes
EVENT=1 (stormflow)
EVENT=0 (dry-weather flow)
Regression-6 (tables 8, 9)
E. coli density=10^
1.798
+ 0.251×log10FNU
−0.084×EVENT
−0.205×Season
+ 0.577×EVENT×log10FNU
Season
1, April 16–October 15
2, October 16–April 15
Median E. coli density
90th prediction interval
Logistic regression, all samples (table 11)
A =
0.059×Turbidity
+3.011×EVENT
−4.000
Web page
http://ga2.er.usgs.gov/bacteria/
prob(exceed EPA)=1/(1+exp^−A)

Figure 24. Real-time estimation model for predicting median *Escherichia coli (E. coli)* density and the probability that median *E. coli* density exceeds the U.S. Environmental Protection Agency's single-sample criterion of 235 colonies per 100 milliliters of water at the Chattahoochee River near Norcross, Georgia (USGS station number 02335000). Explanatory variables: Log10FNU, turbidity in formazin nephelometric units transformed to base 10 logarithms; EVENT, streamflow regime as dry-weather flow or stormflow.

Chattahoochee River at Atlanta, Georgia

From a statistical and practical perspective, regression-12 is the equation chosen for predicting *E. coli* density at the Atlanta site. Unlike the Norcross model, stormflow or dry-weather-flow conditions are not clearly discernible based solely on the relation between streamflow and turbidity measurements at the Atlanta site. As a result, logistic regression analysis cannot be used to predict whether or not stormflow conditions exist using streamflow and turbidity measurements. Therefore, the EVENT variable must be computed using real-time data from a tributary upstream from the Atlanta site. Real-time gage height measurements from Rottenwood Creek near Smyrna, GA (USGS station number 02335910), or Sope Creek near Marietta, GA (USGS station number 02335870), would be used to determine if stormflow conditions exist at the Atlanta site. Gage heights at the Sope Creek gage would be used if gage heights are not available for the Rottenwood Creek gage. Every 15 minutes, the current gage height at Rottenwood or Sope Creeks would be compared to the gage height measured at previous 15-minute intervals and if the absolute change in stage is greater than 0.3 ft, then the EVENT variable equals 1 to indicate stormflow; if the absolute change in stage is less than 0.3 ft, then the EVENT variable equals 0 to indicate dry-weather flow conditions. Instream water temperature and turbidity would be measured every 15 minutes in real time at the Atlanta site. These data along with the EVENT value and the value of the interaction variable would be used in regression-12 to estimate *E. coli* density in real time. Figure 25 is a graphical depiction of the Atlanta prediction model.

Figure 25. Real-time model for predicting median *Escherichia coli (E. coli)* density and the probability of exceeding the U.S. Environmental Protection Agency's single-sample beach criterion of 235 *E. coli* colonies per 100 milliliters of water at the Chattahoochee River at Atlanta, Georgia (USGS station number 02336000). Explanatory variables: Log10FNU, turbidity in formazin nephelometric units transformed to base 10 logarithms; EVENT, streamflow regime as dry-weather flow or stormflow.


Determine streamflow event (EVENT) value
Rottenwood Creek (USGS 02335910) or Sope Creek (USGS 02335870)
Gage height
GageHt(16) (Store previous 4 hours of gage-height data
no
i = i–1 i = 0 ?
yes
EVENT=1 (stormflow)
yes
Abs[GageHt(i)– GageHt(i-1)] > 0.3?
no
EVENT=0 (dry-weather flow)
Chattahoochee River at Atlanta, GA (USGS 02336000)
Water temperature (WTEMP) in degrees Celsius
Turbidity
Regression-12 (tables 13, 14)
E. coli density=10^
1.166
+0.521×Log10FNU
+0.069×EVENT
+0.020×WTEMP
+0.316×EVENT×Log10FNU
Season
1, April 16–October 15
2, October 16–April 15
Logistic regression (table 15)
A =
0.060×Turbidity
+3.332×EVENT
–3.098
prob(exceed EPA)=1/(1+exp^–A)
Median E. coli density
90th prediction interval
Web page
http://ga2.er.usgs.gov/bacteria/

Summary and Conclusions

A 48-mile length of the Chattahoochee River upstream from Atlanta, Georgia, is managed by the National Park Service (NPS) as the Chattahoochee River National Recreation Area (CRNRA). Water-based recreation—such as rafting, canoeing, and fishing—is popular among visitors to the CRNRA. Historically, high densities of fecal-indicator bacteria have been documented in the Chattahoochee River and at levels that commonly exceeded Georgia water-quality standards. In order to maximize the recreational opportunities in the river and minimize potential health issues caused by high indicator bacteria densities, the NPS partnered with the U.S. Geological Survey (USGS), State and local agencies, and non-governmental organizations to monitor *Escherichia coli* (*E. coli)* bacteria density and develop a system to alert river users when *E. coli* densities exceeded the U.S. Environmental Protection Agency single-sample beach criterion of 235 colonies per 100 milliliters of water (colonies/100 mL). This system, called BacteriALERT, has been operating since October 23, 2000.

Between October 23, 2000, and October 1, 2008, about 1,400 water samples were collected at each of two sites on the Chattahoochee River upstream from the city of Atlanta, Ga. These sites are located at streamflow-gaging stations operated by the USGS near Norcross (USGS station number 02335000) and at Atlanta (USGS station number 02336000). Water samples were collected at fixed frequencies ranging from daily to twice per week and analyzed for *E. coli* bacteria density and turbidity in the laboratory. Beginning in mid-2002, turbidity, specific conductance, and water temperature were measured in real time at both sites using instream water-quality sondes. Minimum water releases to the Chattahoochee River from Buford Dam at the upper boundary of the study area and Morgan Falls Dam, upstream from the Atlanta site, maintain a base discharge of 600 to 750 cubic feet per second in the Chattahoochee River throughout the study period. During dry weather, water releases as high as 6,000 cubic feet per second from Buford Dam occur at least once daily to generate electricity when the power demand in the Atlanta metropolitan area is greatest.

Water releases from Buford Dam markedly affect the hydrology, turbidity, and *E. coli* density at Norcross; whereas Bull Sluice Lake and Morgan Falls Dam markedly affect the hydrology, turbidity, water temperature, and *E. coli* density at Atlanta. During dry-weather flow, *E. coli* density at Norcross seldom exceeded the U.S. Environmental Protection Agency single-sample beach criterion of 235 *E. coli* colonies/100 mL (only 2 percent of samples). In contrast, the criterion was exceeded about three times more often at Atlanta (7 percent of samples) than at Norcross during dry-weather flow. The median density of *E. coli* bacteria during the study period was typically greater at the Atlanta site (110 most probably number of colonies per 100 milliliters of water [MPN/100 mL]) than at the Norcross site (60 MPN/100 mL). At both sites, turbidity was the most statistically significant determinant of *E. coli* density, followed by streamflow event (dry-weather flow or stormflow). The study-period median turbidity was 5.7 formazin nephelometric units (FNU) at Norcross and 12 FNU at Atlanta. Seasonally, median *E. coli* density was statistically higher in warm-season samples than in cool-season samples from the Norcross and Atlanta sites. Although median turbidity values at Norcross were 53 percent higher during the cool than warm season, they were statistically similar during both seasons because of the high variability during streamflow. Similarly, although median turbidity values at Atlanta were 30 percent higher during the warm than cool season, statistically they were similar during both seasons.

At Norcross and Atlanta, median turbidity and *E. coli* density were statistically higher during stormflow than dry-weather flow. During stormflow at Norcross, the median turbidity was 36 FNU, about 7 times greater than during dry-weather flow; whereas median *E. coli* density was 530 MPN/100 mL, which is about 10 times greater than during dry-weather flow. Median turbidity values were statistically similar at Norcross and Atlanta (36 and 35 FNU, respectively) during stormflow, but the median *E. coli* density was statistically higher at Atlanta (810 MPN/100 mL) than at Norcross (530 MPN/100 mL). During dry weather, the median turbidity at Atlanta (9.1 FNU) was double the median value at Norcross, and the median *E. coli* density was 60 percent higher at Atlanta than Norcross. The maximum *E. coli* densities in dry-weather samples were 1,200 MPN/100 mL at Norcross and 9,800 MPN/100 mL at Atlanta. Moreover in stormflow samples, the maximum *E. coli* density was 18,000 MPN/100 mL at Norcross and 28,000 MPN/100 mL at Atlanta. The maximum turbidity value during dry-weather flow at Atlanta (480 FNU) was 12 times greater than the maximum turbidity measured at Norcross. In contrast during stormflow at Norcross, the maximum turbidity was measured at 2,700 FNU (by dilution in the laboratory) while the maximum at Atlanta was 450 FNU. This difference is probably the result of Bull Sluice Lake behind Morgan Falls Dam

Regression analyses show that *E. coli* density in samples was strongly related to turbidity and streamflow event (dry-weather flow or stormflow) at both sites. The regression equations chosen for this report are those that have the highest coefficient of determination (R^2), lowest residual standard error, lowest Akaike Information Criterion, and were computationally simple. The regression equation chosen for the Norcross data (regression-6) showed that 78 percent of the variability in *E. coli* density (in log base 10 units, $\log_{10}$) was explained by the variability in $\log_{10}$ turbidity values, streamflow event (dry-weather flow or stormflow), season (cool or warm), and an interaction term that is the cross product of streamflow event and turbidity.

The regression equation chosen for the Atlanta data (regression-12) showed that 76 percent of the variability in $\log_{10}$ *E. coli* density was explained by the variability in $\log_{10}$ turbidity values, water temperature, streamflow event, and an interaction term that is the cross product of streamflow event and turbidity. The importance of water temperature and the insignificance of season in estimating *E. coli* density at Atlanta are probably caused by the influence of Bull Sluice Lake, the small, shallow impoundment behind Morgan Falls Dam. Residual analysis and model confirmation using new data indicated the regression equations selected at both sites predicted *E. coli* density within the computed 90 percent prediction intervals and could be used to predict *E. coli* density in real time at both sites.

By all diagnostic measures, the multiple regression equations for the Norcross (regression-6) and Atlanta (regression-12) data can adequately estimate median *E. coli* density at their respective sites. Prediction R^2 for the regression equations developed for both sites show that nearly 70 percent of the variability in measured $\log_{10}$ transformed *E. coli* densities is explained by variability in the predicted $\log_{10}$ *E. coli* densities. Residual analyses show that residuals from the regression equations at both sites are normally distributed and have constant variance. Using a new set of data collected between October 1, 2008, and September 30, 2009, at both sites, median *E. coli* density was estimated using regression-6 on the new Norcross data and regression-12 on the new Atlanta data. These estimates were strongly correlated with measured *E. coli* densities at both sites, indicating regression-6 and regression-12 can accurately predict *E. coli* densities using new data at their respective sites.

References Cited

Alley, W.M., 1988, Using exogenous variables in testing for monotonic trends in hydrologic time series: Water Resources Research, v. 24, no. 11, p. 1955–1961.

Anderson, C.W., 2005, Turbidity: Techniques of Water-Resources Investigations, U.S. Geological Survey handbooks for water-resources investigations, book 9, chap. A6, section 6.7, 55 p. (Also available at *http://water.usgs.gov/owq/FieldManual/Chapter6/6.7_contents.html*.)

Aulenbach, B.T., 2009, Bacteria holding times for fecal coliform by mFC agar method and total coliform and *Escherichia coli* by Colilert®-18 Quanti-Tray® method: Environmental Monitoring and Assessment, accessed December 29, 2009, at *http://dx.doi.org/10.1007/s10661-008-0734-3*.

Baker, L.M., ed., 2005, Introduction—Statistics, in Franson, M.H., ed., Standard methods for the examination of water and wastewater (21st ed.): Washington D.C., American Public Health Association, chap. 1010, p. 1–1 to 1–2.

Bordner, R.H., ed., 2005, Microbiological examination—Quality assurance/quality control, in Eaton, A.D., Clesceri, L.S., Rice, E.W., and Greenberg, A.E., eds., Standard methods for the examination of water and wastewater (21st ed.): Washington D.C., American Public Health Association, chap. 9020, p. 9.02–9.13.

Buchanan, T.J., and Somers, W.P., 1969, Discharge measurements at gaging stations: U.S. Geological Survey Techniques of Water-Resources Investigations, book 3, chap. A8, 65 p. (Also available at *http://pubs.usgs.gov/twri/twri3a8/*.)

Buckalew, D.W., Hartman, L.J., Grimsley, G.A., Martin, A.E., and Register, K.M., 2006, A long-term study comparing membrane filtration with Colilert® defined substrates in detecting fecal coliforms and *Escherichia coli* in natural waters: Journal of Environmental Management, v. 80, no. 3, p. 191–197.

Chao, K.-K., Chao, C.-C., and Chao, W.-L., 2004, Evaluation of Colilert-18® for detection of coliforms and *Escherichia coli* in subtropical freshwater: Applied and Environmental Microbiology, v. 70, no. 2, p. 1242–1244.

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: U.S. Geological Survey Water-Resources Investigations Report 01–4248, p. 28.

Clark, D.L., Milner, B.B., Stewart, M.H., Wolfe, R.L., and Olson, B.H., 1991, Comparative study of commercial 4-methylumbelliferyl-beta-D-glucuronide preparations with the standard methods membrane filtration fecal coliform test for the detection of *E. coli* in water samples: Applied and Environmental Microbiology, v. 57, no. 5, p. 1528–1534.

Conover, W.J., 1980, Practical nonparametric statistics (2d ed.): New York, John Wiley & Sons, 493 p.

Covert, T.C., Rice, E.W., Johnson, S.A., Berman, Donald, Johnson, C.H., and Mason, P.J., 1992, Comparing defined-substrate coliform tests for the detection of *Escherichia coli* in water: American Water Works Association, v. 84, no. 5, p. 98–104.

Cowburn, J.K., Goodall, T., Fricker, E.J., Walter, K.S., and Fricker, C.R., 1994, A preliminary study of the use of Colilert for water quality monitoring: Letters in Applied Microbiology, v. 19, no. 1, p. 50–52.

Cunnane, C., 1978, Unbiased plotting positions—A review: Journal of Hydrology, v. 37, p. 205–222.

Darakas, Efthymios, 2002, *E. coli* kinetics—Effect of temperature on the maintenance and respectively the decay phase: Environmental Monitoring and Assessment, v. 78, no. 2, p. 101–110.

Eckner, K.F., 1998, Comparison of membrane filtration and multiple-tube fermentation by the Colilert and Enterolert methods for detection of waterborne coliform bacteria, *Escherichia coli*, and enterococci used in drinking and bathing water quality monitoring in southern Sweden: Applied and Environmental Microbiology, v. 64, no. 8, p. 3079–3083.

Fries, J.S., Characklis, G.W., and Noble, R.T., 2006, Attachment of fecal indicator bacteria to particles in the Neuse River Estuary, North Carolina: Journal of Environmental Engineering, v. 132, no. 10, p. 1338–1345.

Fujioka, R., Sian-Denton, C., Borja, M., Castro, J., and Morphew, K., 1998, Soil—The environmental source of *Escherichia coli* and *enterococci* in Guam's streams: Journal of Applied Microbiology, v. 85, Supplement 1, p. 83S–89S.

Georgia Environmental Protection Division, 2009, Water quality standards—Water use classifications and water quality standards: Atlanta, Georgia Department of Natural Resources, Environmental Rule 391–3–6–.03, accessed March 16, 2011, at *http://rules.sos.state.ga.us/docs/391/3/6/03.pdf*.

Georgia Environmental Protection Division, 2010, Georgia 305(b)/303(d) list documents: Accessed November 15, 2010 [or update to July 27, 2011], at *http://www.georgiaepd.org/Documents/305b.html*.

Georgia Power, 2004a, A Morgan Falls operations primer: Accessed April 29, 2011, at *http://www.georgiapower.com/lakes/hydro/pdfs/OperationsPrimer.pdf*.

Georgia Power, 2004b, Morgan Falls Dam pre-application document: Accessed June 6, 2011, at *http://www.georgiapower.com/lakes/hydro/pdfs/MorganFallsPAD.pdf*.

Georgia Power, 2006, Continuous water temperature monitoring locations, 2005: Accessed April 29, 2011, at *http://www.georgiapower.com/lakes/hydro/pdfs/SRMHandWtrRes.pdf*.

Gregory, M.B., and Frick, E.A., 2000, Fecal-coliform bacteria densities in streams of the Chattahoochee River National Recreation Area, Metropolitan Atlanta, Georgia, May–October 1994 and 1995: U.S. Geological Survey Water-Resources Investigations Report 00–4139, p. 8.

Gregory, M.B., and Frick, E.A., 2001, Indicator-bacteria concentrations in streams of the Chattahoochee River National Recreation Area, March 1999–April 2000, *in* Hatcher, K.J., ed., Proceedings of the 2001 Georgia Water Resources Conference, March 26–27, 2001: Athens, Ga., Institute of Ecology, The University of Georgia, p. 510–513.

Hall, N.H., ed., 2005, Microbiological examination—Fecal coliform membrane filter procedure, *in* Eaton, A.D., Clesceri, L.S., Rice, E.W., and Greenberg, A.E., eds., Standard methods for the examination of water and wastewater (21st ed.): Washington D.C., American Public Health Association, chap. 9222, p. 9.59–9.71.

Harrell, F.E., Jr., 2001, Regression modeling strategies: New York, Springer Science+Business Media, Inc., 568 p.

He, L.-M., Lu, Jun, and Shi, Weiyong, 2007, Variability of fecal indicator bacteria in flowing and ponded waters in southern California—Implications for bacterial TMDL development and implementation: Water Research, v. 41, no. 14, p. 3132–3140.

Helsel, D.R., and Hirsch, R.M., 1992, Statistical methods in water resources—Studies in Environmental Sciences 49: New York, Elsevier, 529 p.

Hosmer, D.W., and Lemeshow, Stanley, 2000, Applied logistic regression (2nd ed.)—Wiley series in probability and statistics: New York, John Wiley & Sons, 375 p.

Hurley, M.A. and Roscoe, M.E., 1983, Automated statistical analysis of microbial enumeration by dilution series: Journal of Applied Microbiology, v. 55, no. 1, p. 159–164.

IDEXX Laboratories, Inc., 2002a, Colilert®-18 test kit: Acessed February 5, 2003, at *http://www.idexx.com/water/refs/060202711C18.pdf*.

IDEXX Laboratories, Inc., 2002b, Quanti-Tray/2000®: Accessed February 5, 2003, at *http://www.idexx.com/water/refs/0602320.pdf*.

Kennedy, E.J., 1984, Discharge ratings at gaging stations: U.S. Geological Survey Techniques of Water-Resources Investigations, book 3, chap. A10, 59 p. (Also available at *http://pubs.usgs.gov/twri/twri3-a10/*.)

Krometis, L.H., Characklis, G.W., Dilts, M.J., Simmons, O.D., III, Likirduplos, C.A., and Sobsey, M.D., 2007, Intra-storm variability in microbial partitioning and microbial loading rates: Water Research, v. 41, no. 2, p. 506–516.

Kunkle, Sam and Vana-Miller, David, 2000, Water resources management plan—Chattahoochee River National Recreation Area, Georgia: National Park Service, NPS D–48, 244 p.

Landers, M.N., Ankcorn, P.D., and McFadden, K.W., 2007, Watershed effects on streamflow quantity and quality in six watersheds of Gwinnett County, Georgia: U.S. Geological Survey Scientific Investigations Report 2007–5132, 62 p., available online at *http://pubs.usgs.gov/sir/2007/5132/*.

Leopold, L.B., 1994, A view of the river: Cambridge, Mass., Harvard University Press, 298 p.

Letterman, R.D., ed., 2005, Turbidity—Nephelometric method, *in* Eaton, A.D., Clesceri, L.S., Rice, E.W., and Greenberg, A.E., eds., Standard methods for the examination of water and wastewater (21st ed.): Washington D.C., American Public Health Association, chap. 2130B, p. 2.9–2.10.

Maier, R.M., Pepper, I.L., and Gerba, C.P., 2000, Environmental microbiology: San Diego, Calif., Academic Press, 585 p.

Maluk, T.L., 2000, Spatial and seasonal variability of nutrients, pesticides, bacteria, and suspended sediment in the Santee River Basin and coastal drainages, North and South Carolina, 1995–97: U.S. Geological Survey Water-Resources Investigations Report 00–4076, 46 p.

McConnell, J.B., 1980, Impact of urban storm runoff on stream quality near Atlanta, Georgia: Cincinnati, Ohio, Wastewater Research Division, U.S. Environmental Protection Agency, EPA–600/2–80–094, 64 p.

McFeters, G.A., Pyle, B.H., Gillis, S.J., Acomb, C.J., and Ferrazza, D., 1993, Chlorine injury and the comparative performance of Colisure®, Colilert®, and ColiQuik® for the enumeration of coliform bacteria and *E. coli* in drinking water: Water Science and Technology, v. 27, no. 3/4, p. 261–265.

McSwain, M.R., 1977, Baseline levels and seasonal variations of enteric bacteria in oligotrophic streams, *in* Correll, D.L., ed., Watershed research in Eastern North America—A workshop to compare results: Smithsonian Institute, Edgewater, Md., Chesapeake Bay Center for Environmental Studies, p. 555–574.

Meckes, M.C. and Rice, E.W., eds., 2005, Microbiological examination—Multiple-tube fermentation technique for members of the coliform group, *in* Eaton, A.D., Clesceri, L.S., Rice, E.W., and Greenberg, A.E., eds., Standard methods for the examination of water and wastewater (21st ed.): Washington D.C., American Public Health Association, chap. 9221, p. 9.48–9.59.

Mirkin, Boris, 2005, Clustering for data mining—A data recovery approach: Boca Raton, Fla., Chapman and Hall/CRC Press, 266 p.

Mitsch, W.J. and Gosselink, J.G., 1986, Wetlands: New York, Van Nostrand Reinhold, 539 p.

Montgomery, D.C., Peck, E.A., and Vining, G.G., 2006, Introduction to linear regression analysis—Wiley series in probability and statistics: New York, John Wiley & Sons, 612 p.

Morace, J.L., and McKenzie, S.W., 2002, Fecal-indicator bacteria in the Yakima River Basin, Washington—An examination of the 1999 and 2000 synoptic-sampling data and their relation to historical data: U.S. Geological Survey Water-Resources Investigations Report 02–4054, 32 p.

Myers, D.N., 2004, Fecal indicator bacteria (ver. 1.2): U.S. Geological Survey Techniques of Water-Resources Investigations, book 9, chap. A7, section 7.1, November, accessed December 2004 from *http://water.usgs.gov/owq/FieldManual/Chapter7/Archive/7.1_ver1.2.pdf*.

Myers, D.N., Koltun, G.F., and Francy, D.S., 1998, Effects of hydrologic, biological, and environmental processes on sources and densities of fecal bacteria in the Cuyahoga River, with implications for management of recreational waters in Summit and Cuyahoga Counties, Ohio: U.S. Geological Survey Water-Resources Investigations Report 98–4089, 45 p.

Myers, D.N., Stoeckel, D.M., Bushon, R.N., Francy, D.S., and Brady, A.M.G., 2007, Fecal indicator bacteria (ver. 2.0): U.S. Geological Survey Techniques of Water-Resources Investigations, book 9, chap. A7, section 7.1, February, accessed July 29, 2011, at *http://pubs.water.usgs.gov/twri9A/*.

National Oceanic and Atmospheric Agency, 2011, North-central Georgia precipitation summary, October 1, 2000 to September 30, 2008: National Climatic Data Center, accessed March 9, 2011, at *http://www.ncdc.noaa.gov/oa/ncdc.html*.

National Park Service, 2009, Chattahoochee River National Recreation Area, final general management plan and environmental impact statement: Atlanta, Ga., U.S. Department of the Interior, National Park Service, 426 p., accessed August 3, 2011, at *http://www.nps.gov/chat/parkmgmt/general-management-plan.htm*.

Niemela, S.I., Lee, J.V., and Fricker, C.R., 2003, A comparison of the International Standards Organisation reference method for the detection of coliforms and *Escherichia coli* in water with a defined substrate procedure: Journal of Applied Microbiology v. 95, v. 6, p. 1285–1292.

Olson, B.H., Clark, D.L., Milner, B.B., Stewart, M.H., and Wolfe, R.L., 1991, Total coliform detection in drinking water—Comparison of membrane filtration with Colilert® and Coliquik®: Applied and Environmental Microbiology, v. 57, no. 5, p. 1535–1539.

Ott, Lyman, 1988, An introduction to statistical methods and data analysis: Boston, Mass., PWS-Kent Publishing Company, 835 p.

Palmer, Carol, ed., 2005, Microbiological examination—Enzyme substrate coliform test, *in* Eaton, A.D., Clesceri, L.S., Rice, E.W., and Greenberg, A.E., eds., Standard methods for the examination of water and wastewater (21st ed.): Washington D.C., American Public Health Association, chap. 9223, p. 9.72–9.74.

Rantz, S.E., and others, 1982a, Measurement and computation of streamflow, Volume 1, Measurement of stage and discharge: U.S. Geological Survey Water-Supply Paper 2175, 284 p.

Rantz, S.E., and others, 1982b, Measurement and computation of streamflow, Volume 2, Computation of discharge: U.S. Geological Survey Water-Supply Paper 2175, 356 p.

Rasmussen, P.P., and Ziegler, A.C., 2003, Comparison and continuous estimates of fecal coliform and *Escherichia coli* bacteria in selected Kansas streams, May 1999 through April 2002: U.S. Geological Survey Water-Resources Investigations Report 03–4056, 87 p.

Romesburg, H.C., 1984, Cluster analysis for researchers: Belmont, Calif., Wadsworth, Inc., 334 p.

Taylor, J.K., 1987, Quality assurance of chemical measurements: Boca Raton, Fla., Lewis Publishers, CRC Press, 328 p.

TIBCO Software, Inc., 2008, TIBCO Spotfire S+® 8.1–Guide to statistics, volume 1: TIBCO Software, Inc., 718 p.

U.S. Army Corps of Engineers, 1985, Florida–Georgia stream mileage tables: Mobile, Ala., U.S. Army Corps of Engineers, Mobile District, 191 p.

U.S. Army Corps of Engineers, 2006, Lake Sidney Lanier: Accessed April 29, 2011, at *http://www.sam.usace.army.mil/lanier/*.

U.S. Department of the Interior, 1968, Water quality criteria—Report of the National Technical Advisory Committee to the Secretary of the Interior: Washington, D.C., Federal Water Pollution Control Administration, U.S. Government Printing Office, p. 12.

U.S. Environmental Protection Agency, 1983, Results of the nationwide urban runoff program, Volume I—Final Report: Washington, D.C., Water Planning Division, U.S. Environmental Protection Agency, National Technical Information Service, accession number PB84–185552, 194 p.

U.S. Environmental Protection Agency, 1986, Ambient water quality criteria for bacteria—1986: Washington, D.C., Office of Water Regulations and Standards, Criteria and Standards Division, U.S. Environmental Protection Agency, EPA440/5–84–002, 18 p.

U.S. Environmental Protection Agency, 2000, Improved enumeration methods for the recreational water quality indicators—*Enterococci* and *Escherichia coli*: Washington, D.C., U.S. Environmental Protection Agency, EPA/821/R–97/004, 40 p.

U.S. Environmental Protection Agency, 2002, Implementation guidance for ambient water quality criteria for bacteria, May 2002 draft: U.S. Environmental Protection Agency, Office of Water, EPA–823–B–02–003, accessed March 9, 2011, at *http://nepis.epa.gov/Exe/ZyPURL.cgi?Dockey=20003PSB.txt*.

U.S. Environmental Protection Agency, 2003, Guidelines establishing test procedures for the analysis of pollutants; Analytical methods for biological pollutants in ambient water; Final rule, (part 136): U.S. Code of Federal Regulations, Title 40, parts 100-149, revised July 21, 2003, p. 43272–43283.

U.S. Geological Survey, 2006, Collection of water samples (ver. 2.0): U.S. Geological Survey Techniques of Water-Resources Investigations, book 9, chap. A4, September, accessed July 29, 2011, at *http://pubs.water.usgs.gov/twri9A4/*.

U.S. Geological Survey, 2011a, National Water Information System—Web interface, Site inventory for the Nation: Accessed April 26, 2011, at *http://waterdata.usgs.gov/nwis/inventory.*

U.S. Geological Survey, 2011b, Land Cover Analysis Tool: Accessed April 26, 2011, at *http://lcat.usgs.gov/lcat2*.

Wagner, R.J., Boulger, R.W., Jr., Oblinger, C.J., and Smith, B.A., 2006, Guidelines and standard procedures for continuous water-quality monitors—Station operation, record computation, and data reporting: U.S. Geological Survey Techniques and Methods 1–D3, 51 p. (Also available at *http://pubs.usgs.gov/tm/2006/tm1D3/*).

Wilde, F.D., Radtke, D.B., Gibs, Jacob, and Iwatsubo, R.T., eds., 2004 with updates through 2009, Processing of water samples (ver. 2.2): U.S. Geological Survey Techniques of Water-Resources Investigations, book 9, chap. A5, April, accessed July 29, 2011, at *http://pubs.water.usgs.gov/twri9A5/*.

Appendixes 1–6

Appendix 1. Methods of Sample Collection and Laboratory Analysis

At each site, water samples were collected at midchannel using a weighted-bottle sampler. The yoke was constructed of 4-inch-diameter, polyvinyl chloride pipe and couplings filled with steel pellets and permanently sealed. The total weight of the yoke was between 8 and 10 pounds. The yoke with sample bottle was lowered into the river and raised at a constant rate through the water column. The bottle fills with water as it travels within the water column and approximates a vertically integrated sample. An acceptable sample was one in which a constant rate is maintained as the yoke travels through the water column and fills at least one-half but no more than three-quarters of the bottle. The magnitude of streamflow determines whether samples traverse the entire water column. During stormflows, for example, the river was commonly 10 to 12 feet (ft) deep, and at those depths the weighted yoke was too light to travel through the entire water column.

Analysis of Water Samples

At the laboratory, water samples were analyzed for *Escherichia coli* (*E. coli)* bacteria density and turbidity. Water samples were also analyzed for fecal coliform bacteria during the early months of the BacteriALERT program as a quality control effort to establish the association between *E. coli* and fecal coliform densities measured by three different methods. The methods used in these analyses were approved for drinking water, ambient water, or both by the U.S. Geological Survey (USGS) (Myers and others, 2007), the American Public Health Association (Hall, 2005; Meckes and Rice, 2005), and the USEPA (U.S. Environmental Protection Agency, 2000, 2003).

Determination of *Escherichia coli* Bacteria Density

Water samples were analyzed for *E. coli* bacteria using the Colilert®-18 (Colilert) and Quanti-Tray®/2000 (Quanti-Tray) method manufactured by the IDEXX Corporation (IDEXX Laboratories, Inc., 2002a, b). The bacteria analysis used for *E. coli* in this study is an enzyme substrate or defined substrate method. The American Public Health Association (Palmer, 2005) and the U.S. Environmental Protection agency (USEPA; U.S. Environmental Protection Agency, 2000) have formally approved the Colilert method for drinking water and for ambient water (U.S. Environmental Protection Agency, 2003). The Colilert method is conceptually similar to the commonly used multiple tube method (Meckes and Rice, 2005) in which bacteria densities are determined statistically and expressed as a most probable number of colonies per 100 milliliters of water (MPN/100 mL).

Figure 1–1 shows in schematic how water samples for BacteriALERT were prepared for bacteria analysis using the Colilert method. Two or three different dilutions were prepared by collecting an aliquot of sample with a sterile pipet calibrated to deliver or a sterile, 60-mL polypropylene syringe. The aliquot volume depended on the turbidity of the sample (table 1, main body of this report) and was added to sterile de-ionized (DI) water to produce 100 mL of liquid. For water samples with low turbidity (less than 10 formazin nephelometric units, FNU), sample aliquots of 50, 10, and 1 mL, respectively, were added to 50, 90, and 99 mL of sterile DI water. For samples with moderate turbidity (11–40 FNU), water-sample aliquots of 10, 1, and 0.1 mL, respectively, were added to 90, 99, and 99.9 mL of sterile DI water (table 2). Pre-packaged dry reagent was added to the dilution bottles containing sample and sterile DI water, the bottles were shaken, and the bottles were allowed to sit until bubbles dispersed. After ensuring that all the powder had dissolved in the bottle, the contents of the dilution bottle were poured into a sterile Quanti-Tray, and the tray was sealed in a thermal tray sealer. As the edges of the tray were sealed, the sample was dispersed among 97 wells in the tray. Each tray was labeled on its foil side and incubated for 20 hours at 35 ±0.5 degrees Celsius (°C). The Quanti-Tray consists of three groups of different size wells: 48 small wells, 48 medium wells, and 1 large well. Using this tray and 100 mL of sample, the analyst can estimate the number of colonies from 1 to 2,419 MPN/100 mL without dilution or higher depending on the amount of dilution.

The Colilert method uses a proprietary medium from IDEXX Laboratories, Inc., that contains two chemicals that react with enzymes produced by the total coliform and *E. coli* bacteria. After incubating the samples for 20 hours, wells that were positive for total coliform bacteria yielded a yellow color; if these yellow-colored wells contained *E. coli* bacteria, the *E. coli* enzymes reacted with a fluorogen causing the wells to fluoresce under long-wave ultraviolet light at 366 nanometers (nm). The proprietary medium used in the Colilert method suppresses other noncoliforms that may either interfere with *E. coli* bacteria growth or produce false positives. The density of bacteria as MPN/100 mL for each dilution was determined by the ratio of positive small wells to the sum of positive medium and large wells taken from a statistical table provided by the Colilert manufacturer. The *E. coli* density values described throughout this report refer to the volume-weighted mean *E. coli* density for two to three dilutions. Volume-weighted mean *E. coli* densities (MPN/100 mL) for the Colilert method were computed using equation 1–1.

$$\left(\frac{\sum_{3}^{1} D_x}{\sum_{3}^{1} V_x}\right) \times 100 \qquad (1\text{–}1)$$

where

D_x = bacteria density as the most probable number of colonies in dilution X

V_x = volume of sample used for dilution X

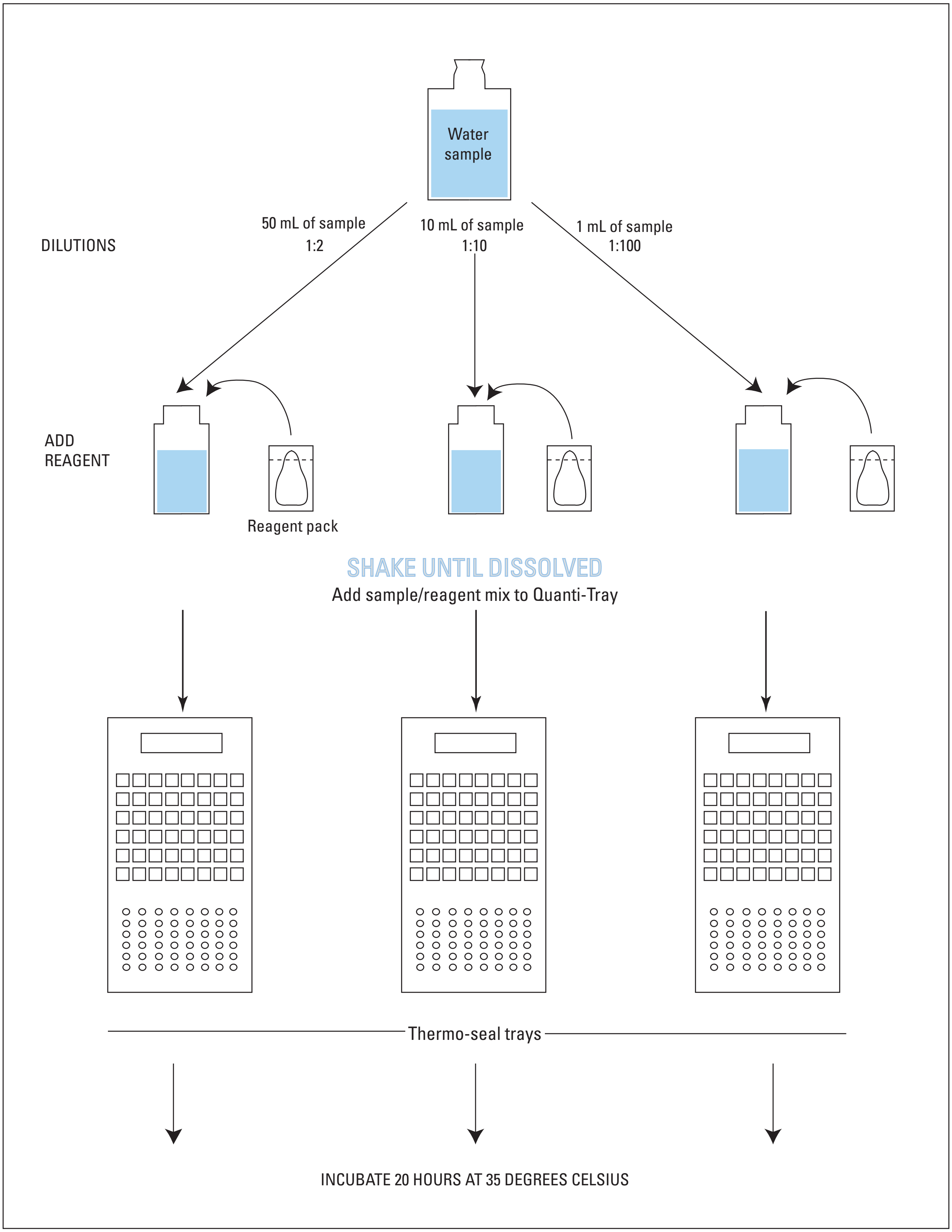


Figure 1–1. Flowchart for processing bacteria samples using the Colilert®-18 Quanti-Tray®/2000 method. [mL, milliliter]

Water samples also were analyzed for *E. coli* using a membrane filter technique. At the laboratory, three to four sample dilutions were prepared to obtain an ideal colony count (20–80 colonies/100 mL) before samples were passed through the membrane filters (Myers, 2004; Myers and others, 2007). During the membrane filter analysis of *E. coli* bacteria, water samples were passed through a 0.45-micrometer (µm) membrane filter on a stainless-steel filter manifold under 5 to 6 pounds per square inch (psi) of vacuum. The filter was then placed on a sterile pad saturated with HACH's m-Coliblue24® broth, and incubated at 35 ± 0.5 °C for 24 hours. The mean number of colonies/100 mL for the membrane filter samples was calculated in the following manner: If one plate had an ideal colony count (20−80 colonies), then equation 1–2 was used. If more than one plate had an ideal count or if all plates had non-ideal densities, then equation 1–1 was used.

$$\left(\frac{D}{V}\right)\times 100 \qquad (1\text{–}2)$$

where

D = colonies/100 mL

V = volume of sample

Quality Assurance/Quality Control Methods

The quality-control methods used during bacteria analyses were those recommended by the American Public Health Association for microbiological analysis (Bordner, 2005). Because most surfaces, including the human body, contain a broad spectrum of bacterial fauna, sample collection and sample processing methods were used that prevented foreign bacteria from contaminating the water sample.

In order to prevent sample contamination before, during, and after collection the following procedures were used:

- All bottles and utensils involved in sample collection or sample processing were either purchased pre-sterilized or sterilized at the USGS Georgia Water Science Center (GAWSC) by autoclave for 15 minutes at a pressure of 18 psi and a temperature of 121 °C.

- Latex or nitrile gloves were used by the sample collector and sample analyst when handling sample bottles or processing samples. In addition, a gel antiseptic containing at least 60 percent ethanol was applied to the gloved hands before sample handling.

- After sterilization of the 1-liter (L) sample bottles, a bottle blank was produced by pouring 500 mL of sterile water into a 1-L sample bottle, which was then shaken, and the blank was analyzed using the same analytical methods that were used for a regular water sample.

- Field blanks were not collected because such an exercise does not truly represent field conditions and sample handling and, therefore, would show little benefit for the effort expended. Producing a field blank would entail more handling than actually occurred with the stream sample.

Sterile dilution water was made in the GAWSC because the sterile, buffered dilution water commonly used in bacteria analyses by membrane filtration interferes with the fluorescence of the *E. coli* determination. Dilution water was sterilized by autoclaving DI water for 15 minutes at 121 °C and a pressure of 18 psi. After cooling, the water was pre-measured (50, 90, and 99 mL) into 125 mL polypropylene bottles. Each bottle was labeled with a lot number and the volume of sterile DI water it contained. The lot number consisted of the Julian day and the year in which the water was sterilized (for example, 00348 for December 14, 2000). At least one 100-mL bottle of this sterile water was processed as a sterile-water blank in the same manner as a regular sample to ensure the water was sterile. The *E. coli* results for this sterile-water blank were recorded in a quality assurance/quality control data book in the laboratory. If the sterile-water blank showed an *E. coli* density at or above 1 MPN/100 mL, then all water was discarded and additional water was sterilized. The pH and specific conductance of sterile water was recorded for each lot produced, and the DI water was analyzed for major ions, trace metals, and nutrients twice per year at the USGS National Water Quality Laboratory in Denver, Colorado. In addition, a reagent blank was analyzed before a new lot of powdered reagent was used for sample analysis.

In this report, quality-control samples were analyzed for three reasons: (1) to monitor the ability of laboratory personnel to maintain sterile conditions during sample processing, (2) to ensure that the Colilert method was able to produce results comparable to other analytical methods currently accepted for quantifying *E. coli,* and (3) to ensure that the analytical precision of the Colilert results was within the theoretical limits of the method. In order to satisfy item (1) above, positive and negative control materials were purchased from the Colilert manufacturer and analyzed intermittently to ensure that sample handling and processing were not contaminating the water samples (negative control) and (or) killing off bacteria (positive control). To satisfy item (2) above, *E. coli* density using the defined substrate procedures of Colilert/Quanti-Tray and the HACH Corporation's m-coliblue24® membrane filter method were compared to fecal coliform densities using the m-FC membrane filter method. These comparisons were made to determine if *E. coli* and fecal coliform densities were correlated at the Norcross and Atlanta sites. In order to satisfy item (3) above, duplicate samples were analyzed, and individual dilutions were treated as duplicates (after normalizing to 100 mL) to calculate confidence intervals and precision of the Colilert method.

Because bacteria densities commonly have log-normal distributions (Hurley and Roscoe, 1983), the precision and

95-percent confidence intervals for this report were computed using Colilert-derived *E. coli* densities transformed to base 10 logarithms. The *E. coli* density for each Colilert dilution was normalized to 100 mL then log transformed, the mean density was computed (geometric mean), and the geometric standard deviation computed using equation 1–3 (Taylor, 1987). The precision of Colilert-derived *E. coli* densities or relative standard deviation (sometimes called the coefficient of variation) is computed as the percentage of the mean. Equation 1–4 shows the precision computation. Confidence intervals were calculated using equation 1–5.

$$s = \sqrt{\frac{\sum_{i=1}^{k}\left(d_{2i} - d_{1i}\right)^2}{2k}} \qquad (1–3)$$

where

s = geometric standard deviation
d_1 and d_2 are measured densities in base 10 log units
i = counter for dilution pairs
k = number of pairs

$$\text{Precision} = \frac{s}{\overline{X}} \times 100 \qquad (1–4)$$

where

s = geometric standard deviation from equation 1–2
$\overline{X}$ = geometric mean *E.coli* density as most probable number of colonies per 100 mL

$$95\,\text{percent confidence interval} = \overline{X} \pm \frac{t \times s}{\sqrt{n}} \qquad (1–5)$$

where

$\overline{X}$ = mean *E.coli* density as MPN/100 mL
t = value from tables of the t-statistic
s = standard deviation from equation 1–2
n = number of dilutions

Determination of Fecal Coliform Bacteria Density

At the laboratory, three to four sample dilutions were prepared to obtain an ideal colony count (20–80 colonies/100 mL) before samples were passed through the membrane filters (Myers, 2004; Myers and others, 2007). Water samples were passed through a 0.7-µm membrane filter on a stainless-steel filter manifold at 5 to 6 psi of vacuum. The filter was plated on m-FC agar or broth and incubated at 44.5 ±0.5 °C for 18–20 hours. The nonstandard incubation period was needed to prevent vigorous colony growth from overgrowing adjacent colonies and to prevent colony die-off as nutrients in the broth or agar were depleted. The mean number of colonies/100 mL for the membrane filter samples was calculated in the following manner: If one plate had an ideal colony count (20–80 colonies), then equation 1–2 was used. If more than one plate had an ideal count or if all plates had non- ideal densities, then equation 1–1 was used.

The fecal coliform determinations with m-FC agar and membrane filtration were used to determine if a relation existed with the Colilert-determined *E. coli* rather than for regulatory purposes. Confirmatory tests for *E. coli* using the Colilert method and the m-Coliblue24® broth were unnecessary because the Colilert and the m-Coliblue24® broth are explicitly confirmatory for *E. coli* (Niemela and others, 2003; Palmer, 2005). Both methods use the reaction of the enzyme β-glucuronidase produced by *E. coli* (and *Shigella* spp.) with MUG (4-methyllumbelliferyl-β-D-glucuronide) to produce a blue fluorescence under ultraviolet light (Colilert) or a blue-green colony (m-Coliblue24®). According to 40 CFR 141.74, revision July 2000, and Standard Methods for the Analysis of Water and Wastewater, 21st edition (Palmer, 2005), confirmatory tests are not needed using Colilert and the m-Coliblue24® broth for determining and enumerating *E. coli* bacteria in drinking water. Confirmatory tests are not needed because the Colilert and the m-Coliblue24® broth confirm *E. coli* in one step; whereas the confirmatory test for *E. coli* after the formation of total coliform or fecal coliform colonies on membrane filters using m-ENDO or m-FC agars, respectively, requires a second step after incubation. The second step involves transferring the filter containing colonies to a broth or agar containing the chemical MUG, incubating for several hours, then illuminating the colonies with ultraviolet light to observe and count colonies emitting a blue fluorescent outer ring, which indicates *E. coli* colonies (Hall, 2005).

Turbidity Measurement

Turbidity data were collected and processed following the protocols published in Letterman (2005), Wagner and others (2006), and Anderson (2005). Water samples were measured for turbidity in the laboratory with a HACH 2100P turbidimeter using the procedures outlined in Letterman (2005). This turbidimeter uses a white or broadband (400–680 nm) light source with a 90-degree detection angle and gives turbidity values in nephelometric turbidity ratio units (NTRU). The meter was calibrated as needed but checked against certified standards before turbidity was measured. Three turbidity readings were taken, and the median value was recorded for each sample. Beginning on May 24, 2002, at Norcross, and July 26, 2002, at Atlanta, water temperature and turbidity were continuously measured instream with YSI 6820 series water-quality sondes. Turbidity was measured with the YSI model 6136 turbidity probe. This turbidity probe uses near-infrared (780–900 nm) or a monochrome light source with a 90-degree detection angle and gives turbidity values in FNU. The water-quality sondes were serviced bi-weekly or as needed within that time period using protocols outlined in Wagner and others (2006). Data from these YSI sondes were uploaded to the NWIS database at the USGS GAWSC in Atlanta on an hourly basis from the Norcross site and every 4 hours from the Atlanta site.

Appendix 2. Methods of Data and Statistical Analysis

For this report, data and statistical analyses consisted of methods and computations used to identify, summarize, and compare patterns, distributions, and outliers in the Norcross and Atlanta datasets. Streamflow measurements in 15-minute intervals from the six gaging stations were used to assign EVENT (streamflow event [dry-weather flow or stormflow]) and HCOND (streamflow condition) values to streamflow measurements at the Norcross and Atlanta sampling sites (table 3). Streamflow immediately downstream from Buford Dam is generated only by water released from Lake Sidney Lanier (Lake Lanier). Because Lake Lanier is so large, the water released does not reflect stormflow in a manner analogous to stormflow from tributaries; therefore, those water releases established the reference for nonstorm-related streamflow in the Chattahoochee River. At times, stormflow from tributaries to the Chattahoochee River coincided and mixed with water released from Buford Dam for power generation and made it difficult to assign a streamflow event to samples and measurements at both sites.

Because of the large influence of water releases from Buford and Morgan Falls Dams, gage heights at the U.S. Geological Survey (USGS) streamgaging stations on Suwanee Creek at Suwanee, GA, Rottenwood Creek near Smyrna, GA, and Sope Creek near Marietta, GA (streams closest to the Norcross and Atlanta sites), were used to determine when a stormflow event should be assigned to water samples collected at the Norcross and Atlanta sites. Rainfall data were of limited use in assigning a stormflow event to samples collected from both sites because early in the study few rain gages were present in the watersheds upstream from the sites. Moreover, the available rainfall data were difficult to reconcile with the timing of storm runoff in the Chattahoochee River. Rainfall-runoff relations are complex because of the interactions among rainfall amounts and intensity, antecedent rainfall period, degree of urbanization in tributary watersheds, and areal extent of rainfall. These interactions can be especially troublesome during small, isolated thunderstorms that may be confined to specific watersheds such as those tributary to the Chattahoochee River upstream from both sites. In addition, turbidity and *Escherichia coli* (*E. coli)* densities in stormflow were frequently masked when storm runoff from small storms was diluted by high volumes of water released by Buford Dam and Morgan Falls Dams.

An agglomerative, hierarchical cluster analysis using Ward's method (Romesburg, 1984; Mirkin, 2005) with *E. coli* density, turbidity and streamflow measurements, and streamflow event was used to identify sample clusters. This analysis produced 24 groups that corresponded to six streamflow conditions (HCOND, location on the hydrograph) during a given season (cool or warm) and streamflow event (dry-weather flow or stormflow). Table 3 lists these groups and their properties, and figure 5 shows a hypothetical hydrograph identifying the six streamflow conditions used in the report. A Visual Basic for Applications function was written within the Microsoft Access® database software to identify and parse streamflow measurements into one of the six HCOND categories. The input data were 15-minute streamflow estimates from rating curves computed for the USGS streamgaging stations at the Norcross and Atlanta sites.

Summary Statistics

Summary statistics were computed for *E. coli* bacteria density and turbidity values measured at both sites during the study period. Measures of mean, geometric mean (mean of log base 10 transformed data), median, and variability such as geometric standard deviation (standard deviation of log base 10 transformed data), interquartile range (IQR), and coefficient of geometric variation (as a percentage, gCOV, geometric standard deviation divided by the geometric mean times 100) were computed for measurements collected during dry weather and stormflow and by season. The equations used in this report for statistical summaries are those published in Ott (1988) or Helsel and Hirsch (1992). Exceedance probabilities, which are commonly used in hydrology to determine streamflow duration curves, were calculated for *E. coli* bacteria density and turbidity measurements. These curves, however, are presented in this report as non-exceedance probabilities (1-exceedance probability). The probabilities were calculated with an S-PLUS function using Cunnane's formula (Cunnane, 1978; Helsel and Hirsch, 1992; TIBCO Software, Inc., 2008).

Regression Analysis–Theory

Regression analysis is a statistical method for identifying and modeling the relations between two or more variables (Montgomery and others, 2006, p. 1). As in most statistical methods, regression analysis attempts to estimate an unknown and immeasurable parameter in a population with a subset or subsample from the population. The subsample (or sample), if random and unbiased, is assumed to mirror the statistical properties of the population. Thus, a regression line is the sample estimate of the true, but unknown, linear relation of two or more variables in a population. Commonly, regression analysis is used to fulfill three objectives: (1) identify relations between measurements in two or more sets of data; (2) remove variation due to the influence of an exogenous measurement in order to better understand the variation in a different measurement of interest (Alley, 1988; for example, remove variation in turbidity measurements due to variation in streamflow measurements so that long-term trends in turbidity can be assessed apart from trends in streamflow); or (3) predict the value of one measurement given the value of another measurement. A regression model does not infer a cause and effect relation between variables. Although a regression model can help to confirm a cause and effect relation, it cannot be the only basis for inferring that relation (Montgomery and others, 2006, p. 39–40).

In order to identify the best equation for predicting *E. coli* density, several regression methods were investigated. These methods include simple linear (SLR) and multiple linear regression (MLR) using ordinary least squares (OLS), line of organic correlation (LOC), and logistic regression (LOGR). In regression analysis, the term variable refers to a quantity that consists of measurements obtained from a quantifiable or qualifiable entity, such as streamflow or turbidity measurements, or a binary indicator variable (typically with a value of 0 or 1). A variable commonly called the explanatory, independent, or X variable is the set of measurements used to predict the mean response in another variable, commonly called the response, dependent, or Y variable (Helsel and Hirsch, 1992). Explanatory variables can be qualitative to represent categorical entities that describe nonquantifiable variables within a dataset, such as male/female or the presence/absence of a condition that may influence the response variable. These variables, commonly called indicator or dummy variables, are typically binary, having values of 0 or 1, although any arbitrary integer could be used (Montgomery and others, 2006, p. 237). Using indicator variables in a regression analysis enables the researcher to simplify data analysis and develop an equation with more predictive power and greater robustness than if equations were developed for each condition represented by the indicator variable (Montgomery and others, 2006, p. 237). If the slope or intercept is different under different values of the indicator variable, then an interaction term is added to the regression analysis. This interaction term is typically the cross product of the indicator variable and an explanatory variable that may vary under different categorical conditions (Montgomery and others, 2006, p. 64).

The variables used for regression analysis in this report included quantitative measurements of *E. coli* bacteria density, stream turbidity, total 72-hour rainfall, antecedent rainfall, stream temperature, and streamflow measurements at the Norcross and Atlanta sites and qualitative computations of season, streamflow event, and six streamflow conditions. Several data transformations of *E. coli* density, turbidity, streamflow, and sample date were included in the initial exploratory data analyses (table 2–1).

Table 2–1. List and description of water-quality and climate variables in addition to those in table 3 used to develop regression-based estimation equations for predicting *Escherichia coli* bacteria density at the Chattahoochee River near Norcross, Georgia (USGS station number 02335000), and at Atlanta, Georgia (USGS station number 02336000), October 23, 2000, through September 30, 2008.

Variable	Description
log10Ecoli	Base 10 logarithmic transformation of *Escherichia coli* bacteria density measured as most probable number of colonies per 100 milliliters of water
log10FNU	Base 10 logarithmic transformation of in situ turbidity measurements in formazin nephelometric units
log10Flow	Base 10 logarithmic transformation of streamflow measured in cubic feet per second
sqrt(Flow)	Square root transformation of streamflow measured in cubic feet per second
Flow-1	Inverse transformation of streamflow measured in cubic feet per second
sin(Flow)	Sine transformation of streamflow (for example, 1.0472×StreamFlow)
Season	Binary variable indicating a warm-weather period (1, April 16 to October 15) or a cool-weather period (2, October 16 to April 15)
WTEMP	Continuous in situ measurement of water temperature, in degrees Celsius
sqrt(WTEMP)	Square root transformation of water temperature
sin(biyear)	Sine transformation of 2.5×3.1416×((month/12)+0.2)
sin(year)	Sine transformation of decimal year taken from the sample date (for example, 6.283×month/12)
cos(year)	Cosine transformation of decimal year taken from the sample date (for example, 6.283×month/12)
cos(month)	Cosine transformation of decimal month (for example, 0.5236×month/12)
julian	Julian day as the day of the year beginning on January 1 as 1 and ending on December 31 as 365 in a non-leap year
sin(julian)	Sine transformation of Julian day
cos(julian)	Cosine transformation of Julian day

Linear Regression

The simple linear regression model, commonly called the equation of a straight line, is given in equation 2– 6.

$$\text{Conditional mean of Y, given } X = \left(Y_x\right) = \beta_0 + \beta_1 X \tag{2–6}$$

where

Y = response variable
β_0 = Y intercept parameter
β_1 = slope of the regression line
X = explanatory variable

Simple linear regression (SLR) analysis uses data containing paired variables—one variable is the explanatory variable and the second variable is the response variable. Regression using SLR attempts to produce a line with a slope coefficient and a y-axis intercept coefficient that minimizes the sum of the squared differences in the response variable (y-axis variable); the errors in the explanatory variable (x-axis variable) are not minimized because it is assumed that this variable is measured without error (Montgomery and others, 2006). The subsequent regression line represents the mean response to a given explanatory variable. One assumption in SLR, which conflicts with its use in water resources studies, is that the explanatory variable is measured without error (essentially a constant) and the corresponding response variable is measured with error (Helsel and Hirsch, 1992; Montgomery and others, 2006, p. 49). An explanatory variable without measurement error is a condition that is rarely observed in water resources studies; rather, most measurements in water resources are random in the statistical sense. Nevertheless, Montgomery and others (2006, p. 49–50) state that measurement error in the explanatory variable does not negate a regression analysis as long as the following assumptions are true: (1) the variables used for the response and explanatory data have similar joint normal distributions about the conditional mean of the regression (determined by statistically significant correlation) and (2) the value of the explanatory variable is independent and random without association with the β_0, β_1 or conditional variance of the regression.

Multiple linear regression analysis (MLR) is an OLS method that extends a simple linear regression analysis from one explanatory variable to multiple explanatory variables (Helsel and Hirsch, 1992; Montgomery and others, 2006, p. 63). The MLR is commonly used when knowledge of the system suggests that two or more variables act in concert to give the observed response or when residuals are so large that some unknown explanatory variable is affecting the response (Helsel and Hirsch, 1992). The multiple linear regression model is shown as equation 2– 7.

$$\text{Conditional mean of Y, given } X = \left(Y_x\right) = \beta_0 + \beta_1 X_1 .... + \beta_n X_n \tag{2–7}$$

where

Y = response variable
β_0 = Y−intercept parameter
β_1 = partial regression (slope) coefficient for variable X_1
X_1 = 1st explanatory variable
β_n = partial regression (slope) coefficient for variable X_n
X_n = the nth explanatory variable

The parameter β_n represents the change in the response variable for a unit change in its explanatory variable (X_n) when all other explanatory variables in the model are held constant (Montgomery and others, 2006, p. 64). When considering MLR, it is important to maintain parsimony of the final model. Parsimony is the concept that the best regression model is the simplest model consistent with the data and knowledge of the problem or process being modeled. Parsimony is maintained when explanatory variables are first transformed before adding more explanatory variables to the regression analysis (Montgomery and others, 2006, p. 202).

Several assumptions are inherent in linear regression analyses (Helsel and Hirsch, 1992; Montgomery and others, 2006, p. 122): (1) residuals (difference between measured and predicted values) are normally distributed, (2) residuals have zero mean and constant variance, (3) residuals do not show trends over time, and (4) residuals are not correlated.

The coefficient of determination (R^2), analysis of variance, and residual analyses are used to determine the veracity of the regression. The R^2 indicates the proportion of variability in the response variable that is explained by the variability in the explanatory variables (Montgomery and others, 2006, p. 35). The term regression equation is used to identify different sets of explanatory variables that are used to estimate the intercept (β_0), slope (β_1), and the mean response in a regression analysis. Also, the term linear regression means that the response or Y-variable is a linear function of the regression coefficients (β_0, β_1) rather than the linearity of the data (Montgomery and others, 2006, p. 63). As long as linearity in β_0 and β_1 is maintained, a polynomial linear regression analysis can be used to describe curvilinear data.

In contrast to OLS, which minimizes the squared errors in the response variable, LOC minimizes the squared errors in both the response and explanatory variables (Helsel and Hirsch, 1992). This is important in extending a hydrologic record or imputing missing data because the variance structure of the actual data is imparted to the estimated data. In this report, the LOC is used to develop equations for relations among *E. coli* and fecal-coliform bacteria determined by membrane filter methods and *E. coli* density determined by the Colilert method.

Logistic Regression

In contrast to SLR and MLR, LOGR is a nonlinear regression method. The LOGR is commonly used to develop a model that estimates the probability or chance that the value of one variable is above or below a threshold at a given value of a second variable. Two important differences between SLR and LOGR relate to the conditional mean and the conditional distribution of the response variable. In SLR, the conditional mean of Y is continuous and linearly related to a continuous explanatory variable as in equation 2–6, where the regression parameters (β_0 and β_1) are linear, and the errors are normally distributed. In LOGR, the mean response variable is binary or dichotomous (only assumes two values, usually 0 and 1) and nonlinear with respect to the explanatory variable and the regression parameters (β_0 and β_1), and errors have a binomial rather than a normal distribution (Hosmer and Lemeshow, 2000). The most common binomial distribution used to develop a regression model with binary data is the logistic distribution. The logistic regression model, which uses the logistic distribution, produces a conditional mean of the response variable (Y) bound by 0 and 1. The mathematical definition is given in equation 2–8:

$$\text{Conditional mean of } Y\text{, given } X = \left(\Pi_x\right) = \frac{1}{1+\exp\left(\beta_0+\beta_1 X_1\right)} \tag{2–8}$$

where

Y = response variable

β_0 = Y−intercept parameter

β_1 = slope parameter of the regression line

X_1 = explanatory variable

exp = 2.718282

In the logistic model, either the upper or lower binary value is approached asymptotically and must be transformed to create a linear equation and to develop a linear regression equation (Hosmer and Lemeshow, 2000). In order to create a linear equation from equation 2–8, a logit transformation is performed and shown as equation 2–9.

$$g_x = \ln\left[\frac{\Pi_x}{1+\Pi_x}\right] = \beta_0+\beta_1 X_1 \tag{2–9}$$

where

g_x = logit transformation

Π_x = equation 2

ln = natural logarithm

β_0, β_1, X_1 = as defined in equation 2–8

The logit transformation gives g_x many of the properties of a linear regression (such as linearity in β_0 and β_1). The logistic regression is fit to a binary dataset by maximum-likelihood estimation using statistical computer software (Hosmer and Lemeshow, 2000). Maximum-likelihood estimation computes the least squares functions, which estimate the values for the unknown regression parameters, β_0 and β_1, and maximize the probability of obtaining the observed conditional mean response for the given explanatory values in the original dataset.

Regression Analysis–Methods Used

Regression analysis is an iterative process of trial and error that ultimately may provide a usable predictive equation. For this report, regression analysis was divided into four phases: (1) exploratory analysis and data reduction, (2) variable selection, (3) equation selection, and (4) equation validation. During the first phase, *E. coli,* turbidity, and streamflow values were transformed to new variables using various transformations such as base 10 logarithms, and inverse, square root, square, and cubic functions (table 2–1). In addition, sample dates were transformed to produce new variables that described seasons, Julian day, months, and monthly and annual periodicity using sine and cosine functions. The S-PLUS® leaps and bounds function was used to calculate all possible regressions from the original and transformed variables (TIBCO Software, Inc., 2008).

The leaps and bounds function is an iterative process that shuffles and combines explanatory variables into various permutation sets and regresses the response variable against those sets of explanatory variables. The function then sorts each subset of explanatory variables into ascending order starting with the subset with the lowest number of variables and lowest C_p statistic for that group of variables. The C_p statistic (Mallow's C_p statistic, Montgomery and others, 2006, p. 268) is one of several diagnostic tools used to determine the strength of the regression equation. In the leaps and bounds procedure, the C_p statistic and the R^2 were used to identify the set of explanatory variables that had the lowest amount of bias with a given response variable. The C_p statistic balances the need to maximize the R^2 with the need to minimize the regression mean square error (Montgomery and others, 2006, p. 268).

This initial leaps and bounds procedure was completed on all Norcross and Atlanta data to identify the "best" one-variable equation. Simple linear regression analyses were completed on full datasets from the Norcross and Atlanta sites using *E. coli* density as the response variable and the "best" explanatory variable ($\log_{10}$ turbidity) identified during leaps and bounds. The primary purpose of this initial regression was to identify highly influential or highly leveraged values (outliers) in the dataset.

Measures of leverage and influence were used to detect measurements that lie far outside the linear relation implied by the rest of the data (commonly called outliers). Data with high leverage and influence can exert a strong, negative influence on the regression equations and bias the predicted response variable (Helsel and Hirsch, 1992; Montgomery and others, 2006, p. 143–144). Studentized residuals and the

DFFITS statistic are commonly used for identifying outliers in the data (Harrell, 2001; Montgomery and others, 2006, p. 125–126, 195). Outliers may represent a measurement error or other anomaly in one or more explanatory variables. The DFFITS statistic measures the influence that a value of the explanatory variable has on the slope of the regression line. The DFFITS is one of several statistics commonly used to identify values that have a large influence on the regression coefficients (β_0, β_1; Helsel and Hirsch, 1992). Samples are considered to have high influence when the calculated DFFITS statistic is greater than the critical value computed using equation 2–10 (Montgomery and others, 2006, p. 196). Any sample with a DFFITS statistic greater than the critical value or a studentized residual with an absolute value greater than 1.9 was considered an outlier and evaluated for measurement or transcription errors and corrected or removed if errors were not correctable.

$$|\text{DFFITS}| \geq 2\left(\sqrt{\frac{p}{n}}\right) \qquad (2\text{–}10)$$

where

p = number of parameters in regression equation
n = number of samples used in the regression

Regression analysis using this new dataset was completed to determine if removing outliers improved the regression statistics. Using this new equation, additional variables were added to the leaps and bounds results to determine other characteristics that accounted for a statistically significant amount of the variability in *E. coli* density. In addition, a logistic regression model was developed to predict the probability of *E. coli* bacteria densities exceeding the U.S. Environmental Protection Agency beach criterion at a specific turbidity measurement.

Evaluating the Regression Equation

Commonly a regression equation is intended to predict a response for quality control, management decisions, or clinical health studies; thus, it is imperative that the equation predict a response with minimal errors. In every regression equation there is an error term (often not shown but implied by definition) that consists of measurement errors in the response variable and regression errors (bias). The total error in any regression can never be less than the measurement error of the response variable. Furthermore, if statistical inference (hypothesis testing) is used to compare or validate equations, then the regression coefficients must be robust and unbiased estimates of the population and have minimum variance. Common measures, and those used for this report, for evaluating the regression analysis include identifying: (1) the statistical significance of the regression (is the slope statistically different from zero?); (2) the statistical significance of each explanatory variable; (3) the normality and variance character of the residuals; (4) measures of leverage and influence (outliers) for each value in the dataset; and (5) correlations among residuals, among explanatory variables, and over time.

Tables that summarize the results and analysis of variance for the regressions developed for each sampling site are presented in appendixes 5 and 6. These tables list several values that are diagnostic for the regression analysis; that is, these values help one determine the strength of the regression and the importance of each explanatory variable to the regression. Most statistical software, including S-PLUS, provides those statistics. One diagnostic value from the analysis of variance, the F-value and its associated *p*-value, indicates the overall statistical significance of the regression and for each explanatory variables. The significance of the regression rests on the hypothesis that there is no linear relation between the response variable and any of the explanatory variables; in other words, the slope of the regression or of the individual explanatory variable is zero (Montgomery and others, 2006, p. 80). Associated with the F-value for each regression analysis is a *p*-value, which indicates the probability of a higher F-value and thus the probability or chance that there is no linear relation between the response and explanatory variables. In this report, computed *p*-values that are less than 0.05 (the alpha value) indicate significant linearity between the response variable and explanatory variables (Montgomery and others, 2006, p. 84, 86). Only those regression analyses having a *p*-value less than 0.05 were kept for further development.

Another diagnostic value that is typically presented with the analysis of variance is the *t*-value. The *t*-value is calculated for the intercept and each explanatory variable added to the regression using a t-test with the null hypothesis that the intercept is zero. If the *t*-value calculated by the regression analysis shows that the computed *p*-value for the variable is less than 0.05, then the hypothesis is rejected, the intercept of the variable is not equal to zero, and the variable is statistically important to the regression (Montgomery and others, 2006, p. 84). Typically, when computed *t*-values result in a *p*-value greater than 0.05, that variable is removed from the analysis and the regression is re-computed. Including such a variable in the equation may increase the variance around the intercept and may be detrimental when statistically comparing the regression equation at some later date. Nevertheless, if adding an interaction term to the regression causes the *p*-value of an explanatory variable to exceed 0.05, that explanatory variable is not removed because it is interacting with one of the other explanatory variables.

Figures showing diagnostic graphs for the regression analyses considered but not chosen for the Norcross and Atlanta predictive models are presented in appendixes 5 and 6, respectively. These plots show the normality and variance character of the regression residuals, which are important measures of robustness in the regression. In the first graph (labeled A), the measured *E. coli* density was plotted against the predicted *E. coli* density from the regression equation and provides a view of the association between the two datasets. In the second graph (labeled B), residuals were plotted against the predicted *E. coli* density from the regression equation. This type of graph can show linear or monotonic trends in the residuals that may result from nonconstant variance. In the third graph (labeled C), quantile-quantile plots of the residuals were used to determine the normality of the residuals; if the residuals plot along the standard normal distribution line within ±2 standard deviations, then about 95 percent of the residuals are normally distributed (Helsel and Hirsch, 1992; Montgomery and others, 2006, p. 129).

If the assumptions of residual normality and constant variance were met with any regression analysis, then the residuals were analyzed for unwanted correlation. Unwanted correlation includes correlation among explanatory variables (multicollinearity), which is the linear correlation between two or more explanatory variables in a regression equation (Montgomery and others, 2006, p. 323). Multicollinearity can negatively affect an equation's ability to predict future observations by inflating the variances of the regression coefficients and introducing bias. The variance inflation factor (VIF) is the statistic used in this report to identify multicollinearity (Harrell, 2001; Montgomery and others, 2006, p. 334). If explanatory variables in the regression equation had a VIF greater than 5 or 10, then multicollinearity existed and one of the variables was removed from the dataset, and the dataset was re-analyzed.

In addition to multicollinearity, serial correlation may exist in a dataset when samples are collected sequentially within a short time period. This correlation, called autocorrelation, is determined in two ways in this report: (1) by the Durban-Watson statistic (Montgomery and others, 2006, p. 476–477; TIBCO Software, Inc., 2008), and (2) by a correlation analysis of regression residuals with another set of those residuals lagged by one to *n* number of observations. The closer the Durban-Watson statistic is to 2, the smaller the chance that the *E. coli* density in each sample was influenced by previously collected samples during the study period. In addition, autocorrelation coefficients were computed using an S-PLUS function. This function creates temporary copies of the dataset with regression residuals lagged by one to *n* number of observations and computes the correlation of the regression residuals between the original dataset and the copies. In other words, residuals for a given sample were compared sequentially to the regression residuals in the previous one to six samples. If autocorrelation did exist, then the analysis would indicate the sampling interval at which samples were no longer correlated in time. The sample interval at which autocorrelation no longer exists could be used to subset the original dataset or the dataset could be randomized to remove the autocorrelation. For example, if the autocorrelation coefficient falls below the critical value for the fourth lagged sample, then selecting every fifth sample for regression analysis should remove the autocorrelation seen in the original data. For this report, however, the autocorrelation analyses were informational only because the regression analyses did not incorporate a time series component and were not intended to forecast *E. coli* densities on future dates. Nevertheless, time series plots of residuals were used to identify study period trends in *E. coli* density.

Another diagnostic tool used in this report is the Akaike Information Criterion (AIC). This statistic is a log-likelihood function that can help determine how closely the regression-estimated data fit the measured data. The AIC is especially useful for validating regression equations developed by nonlinear regression, such as logistic regression (Harrell, 2001, p. 234). The smaller the AIC, the better the correspondence between the estimated and measured data. For each plausible equation identified by the leaps and bound procedure for both study sites, the adjusted R^2, the residual standard error, the variance inflation factor, the Durbin-Watson statistic, the first-order autocorrelation coefficient, and the AIC statistic were calculated. These measures, along with residual and quantile-quantile (q-q) plots, were then used to select the final or "best" regression equation for each sampling site.

Methods of Validating the Regression Equation

Regression equations chosen as the "best" for each site were considered estimation or calibration equations. The estimation equations were used with a validation dataset (sometimes called a confirmation dataset) to determine how well the equation predicts values of the response variable. In this report, the validation dataset consisted of data collected at the Norcross and Atlanta sites from October 1, 2008, to September 30, 2009. The *E. coli* bacteria density was predicted using the calibration equation on the validation dataset and compared with *E. coli* measurements in the validation dataset. Prediction residuals were computed as the difference between the measured and estimated *E. coli* densities in the validation dataset.

Prediction indices were calculated on the prediction residuals and used to determine how well predicted *E. coli* densities fit the measured *E. coli* densities in the validation dataset. These indices include the prediction R^2, prediction mean square error, and prediction AIC. Furthermore, scatterplots were constructed showing the relation of prediction residuals to the predicted *E. coli* density, the relation of measured *E. coli* density to the predicted *E. coli* density, residual q-q plots, and a time series plot of residuals for the study period.

Appendix 3. Comparisons Between *Escherichia coli* and Fecal Coliform Bacteria Density

The *Escherichia coli* (*E. coli)* bacteria densities measured by the Colilert method at both the Norcross and Atlanta sites were compared to *E. coli* bacteria densities and fecal coliform bacteria densities determined by membrane filter methods. These comparisons were made to ensure that the Colilert method could determine *E. coli* densities that were comparable to fecal coliform densities from historically accepted membrane filter methods. As of early 2012, fecal coliform bacteria were the indicator bacteria used for regulatory purposes by the State of Georgia; thus, documenting the relation between fecal coliform and *E. coli* bacteria provides a tool that may benefit Georgia as it evaluates the move from a fecal coliform bacteria standard to the U.S. Environmental Protection Agency supported *E. coli* standard.

At both sites, nonparametric statistical analysis (Wilcoxon Signed Ranks test; Conover, 1980) showed that *E. coli* densities using the Colilert method were statistically similar to fecal coliform densities using membrane filtration (*p*-values were 0.371 at Norcross and 0.316 at Atlanta; table 3–1). A 3-year study by Buckalew and others (2006) showed a strong correlation between *E. coli* densities using Colilert and fecal coliform densities using membrane filtration (Pearson's correlation coefficient, r, was 0.957). Figure 3–1*A* shows the strong correlation (Spearman's rho is 0.941) in *E. coli* density between the Colilert method and membrane filtration using HACH's m-coliblue24 method. In addition, figure 3–1*B* shows the moderate correlation (Spearman's rho is 0.737) between the mean fecal coliform bacteria densities with the membrane filtration using m-FC agar and mean *E. coli* bacteria densities using the Colilert method. Figure 3–1*C* shows the moderate correlation (Spearman's rho is 0.751) between mean fecal coliform bacteria densities and *E. coli* density using the membrane filtration methods. These relations indicate that the Colilert method of enumerating coliform bacteria is comparable to the results using membrane filter methods.

Although the mean *E. coli* densities from the Colilert and HACH m-coliblue24 methods were statistically similar to fecal coliform densities (table 3–1), the *E. coli* densities tended to be slightly greater than the fecal coliform density even though *E. coli* bacteria are a subset of the fecal coliform group (figs. 3–1*B*, *C*; Maier and others, 2000, p. 491). This discrepancy was probably caused by the greater efficiency of the Colilert method in growing *E. coli* bacteria and easier enumeration of the colonies. Studies have shown that the Colilert method has a lower incidence of false positive and false negative results than the fecal coliform method (Chao and others, 2004). In addition, the fecal coliform method commonly is influenced by the growth of thermotolerant coliforms other than *E. coli,* such as *Klebsiella* strains. With the fecal coliform method, as much as 15 percent of the plated colonies can be *Klebsiella* strains (false positives) that are ubiquitous in the environment, are not present in fecal matter, and are not associated with the occurrence of human disease (Chao and others, 2004). False negatives are common with the fecal coliform method because non-gas-producing strains of *E. coli* that either do not grow or are not counted as fecal coliform colonies may account for as much as 10 percent of the *E. coli* population (Chao and others, 2004). The Colilert method is accurate at low bacteria densities (McFeters and others, 1993; Niemela and others, 2003) and is able to recover stressed cells from a variety of environments (Covert and others, 1992; Eckner, 1998). Using the MPN estimate, Colilert has a reported precision of 1 cell/100 mL because the MUG substrate is highly sensitive to the presence of *E. coli* (IDEXX Laboratories, Inc., 2002a, b). This precision is corroborated in studies reported by Covert and others (1992) and Eckner (1998).

Table 3–1. Statistical comparisons between fecal coliform and *Escherichia coli* bacteria densities using the Colilert and membrane filter methods for water samples collected from the Chattahoochee River near Norcross, Georgia (USGS station number 02335000), and at Atlanta, Georgia (USGS station number 02336000) sampling sites.

[Wilcoxan Signed Ranks test was used with a null hypothesis that the statistical distributions and medians of the two datasets are similar, compared to the alternate hypothesis that the median bacteria density at the Norcross site is smaller than at the Atlanta site; *E. coli, Escherichia coli* bacteria; *p*-value, the probability that the two distributions are not different]

Indicator bacteria	Data sets	Method	Result	*p*-value
Fecal coliform density	Norcross compared to Atlanta	Membrane filter with m-FC agar	Norcross less than Atlanta	0.011
E. coli density	Norcross compared to Atlanta	Colilert$^{®}$-18/Quanti-Tray$^{®}$/2000	Norcross less than Atlanta	.002
Fecal coliform density versus *E. coli* density	Norcross	Colilert$^{®}$-18/Quanti-Tray$^{®}$/2001; membrane filter with m-FC agar	*E. coli* density equal fecal coliform density	.371
Fecal coliform density versus *E. coli* density	Atlanta	Colilert-18$^{®}$/Quanti-Tray$^{®}$/2002; membrane filter with m-FC agar	*E. coli* density equal fecal coliform density	.316

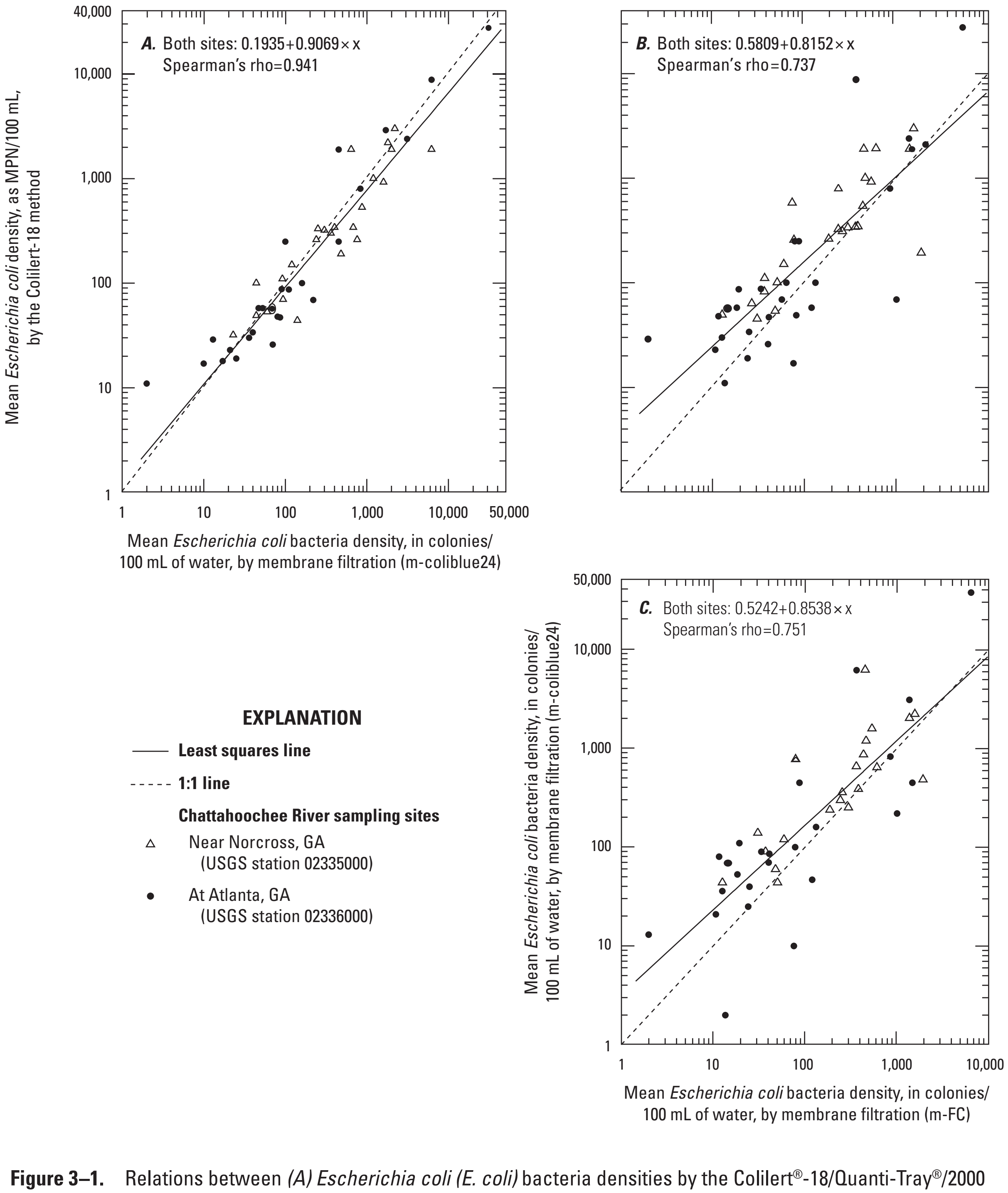


Figure 3–1. Relations between *(A) Escherichia coli (E. coli)* bacteria densities by the Colilert®-18/Quanti-Tray®/2000 method and HACH m-coliblue24 membrane filter method; *(B) E. coli* bacteria densities by the Colilert®-18 Quanti-Tray®/2000 method and the fecal coliform bacteria densities by the membrane filter method (m-FC); and *(C) E. coli* bacteria densities by the HACH m-coliblue24 membrane filter method and fecal coliform bacteria densities by the membrane filter method (m-FC) in water samples collected between November 6, 2000, and August 8, 2001, from the Chattahoochee River near Norcross, Georgia (USGS station number 02335000), and at Atlanta, Georgia (USGS station number 02336000). [MPN/100 mL, most probable number of colonies per 100 milliliters of water]

Appendix 4. Relation Between Laboratory and Field Turbidity Measurements

Throughout the study period, turbidity was measured in water samples collected for *Escherichia coli* (*E. coli)* analysis. Instream turbidity measurements did not begin until May (Norcross) or July (Atlanta) 2002, nearly 20 months after the start of the study. Because the goal of the BacteriALERT program is to predict *E. coli* density in real time using instream turbidity measurements at each sampling site, the turbidity data used in the regression analyses had to be equivalent to the instream measurements. The laboratory turbidity measurements were not equivalent to the instream measurements because different turbidimeters were used, the turbidity units were different, and the environmental conditions during measurement were different. In order to convert the laboratory turbidities measured to equivalent instream turbidities, a relation between laboratory and instream measured turbidities was constructed using the Method of Variance Extension (MOVE; Helsel and Hirsch, 1992). The MOVE analysis computed the line of organic correlation (LOC) and the equation of that LOC was used to convert laboratory turbidity values to equivalent instream turbidity values for the nearly 500 samples that were collected before instream turbidity measurements began.

The LOC from the MOVE calculation is plotted in figure 4–1 and indicates a strong correlation (Pearson's correlation coefficient, r, equals 0.90) between laboratory- and instream-measured turbidity. Most of the variation in this relation occurs at turbidity values less than about 20 formazin nephelometric units and probably is related to the different locations used to collect water samples and measure instream turbidity. The differences would be most pronounced at the lowest flows, which typically have the lowest turbidity. In addition, when a water sample is collected, the sample bottle is filled as the sampler descends through the water column, resulting in a composite that—depending on stream stage—may not fully integrate the water column. During high flows the weighted-bottle sampler probably does not descend through the full vertical depth of the water column, perhaps only sampling the upper third or upper fourth of the water column; whereas the turbidity probe is set at a depth below the water surface that varies with stream stage. The difference between the two turbidity measurements at the highest turbidity values is probably the result of inaccurate laboratory measurements. Laboratory-measured turbidities are inaccurate at high turbidity levels because heavier particles do not remain suspended in sample cuvettes long enough to be measured.

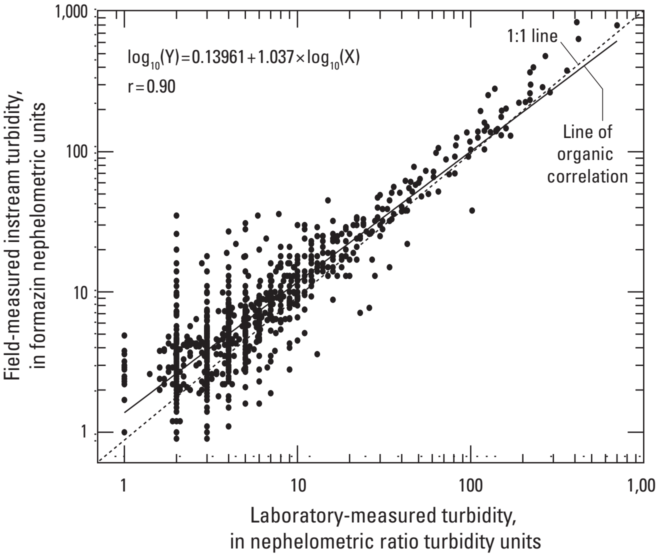
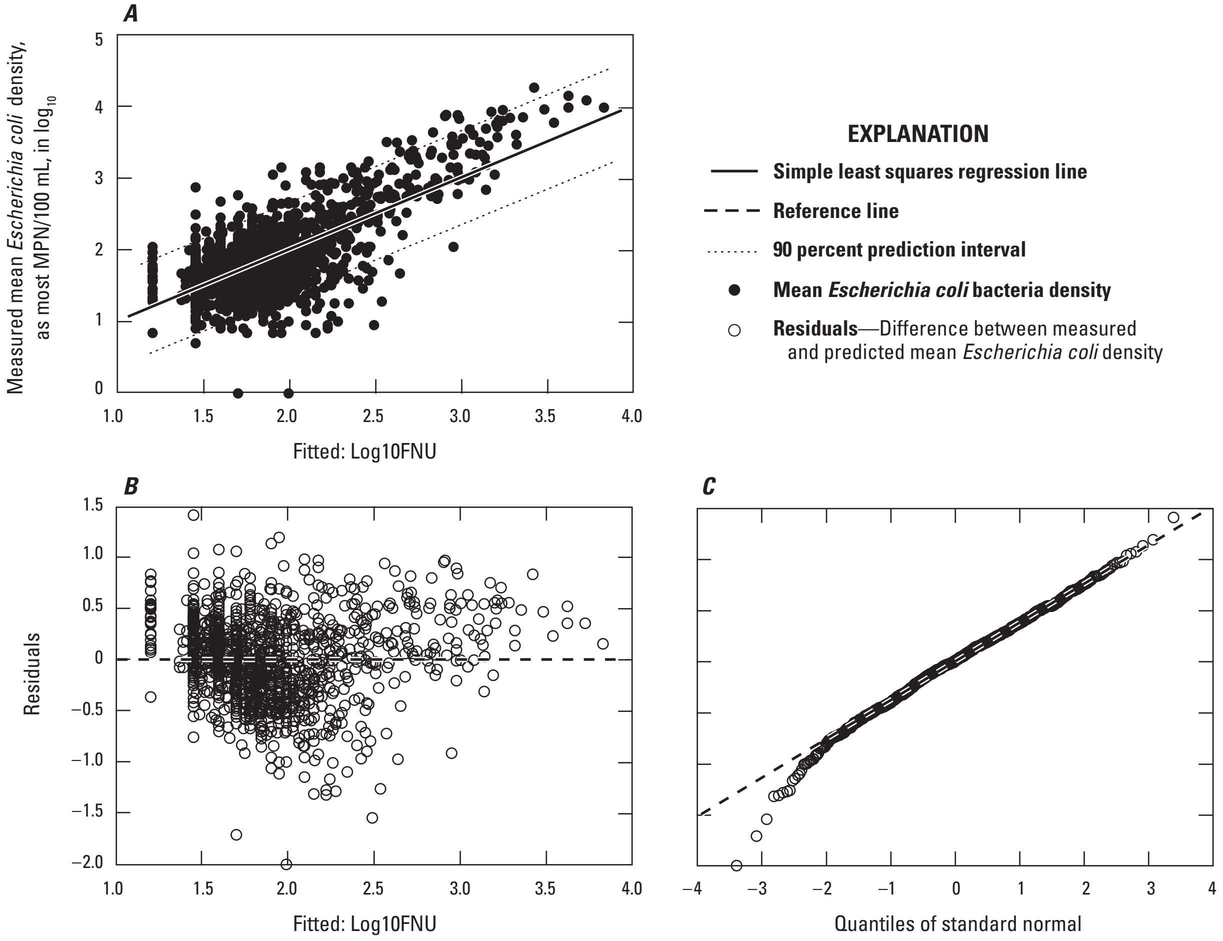


Figure 4–1. Relation between laboratory measured turbidity and turbidity measured instream at the Chattahoochee River near Norcross and at Atlanta, Georgia, October 23, 2000, through September 30, 2008.

Appendix 5. Regression Statistics and Residual Plots for the *Escherichia coli* Models at the Norcross Site



Residual standard error 0.3940, adjusted R^2 is 0.512, F statistic is 1,490, and *p*-value is less than 0.001

Response variable: Log10Ecoli

Terms	Coefficient	Standard error (SE)	*t*-statistic	*p*-value[a]
Intercept	1.204	0.020	59.82	<0.001
Log10FNU	.827	.021	38.59	< .001

Analysis of variance

Terms	Degrees of freedom[b]	Sum of squares (SS)	Mean SS	F statistic[c]	*p*-value (F)
Log10FNU	1	231.7	231.7	1,489	<0.001
Residuals	1,415	220.1	.2		

Autocorrelation coefficients

Number of samples lagged	Correlation coefficient[d]
0	1.0
1	.24
2	.11

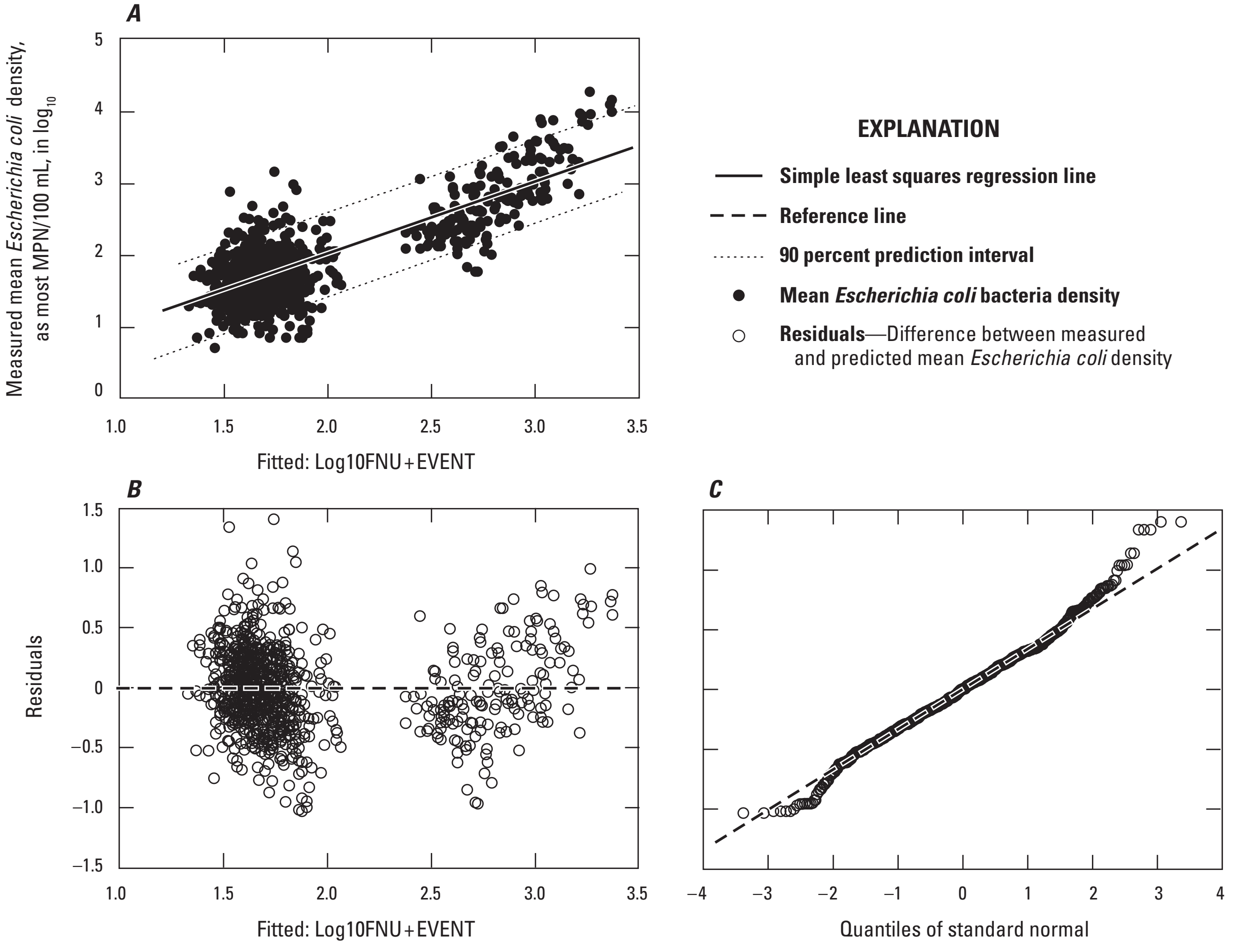
[a] *p*-value, the probability that the parameter is not important to the regression.

[b] Defined as the number of independent pieces of information used to calculate the statistics.

[c] F statistic, used to determine if there is a significant linear relation between the response variable and the explanatory variables.

[d] The critical value is 0.20. Coefficients greater than 0.20 are significant at alpha equal to 0.10.

Figure 5–1. Regression statistics for regression-1 (table 8) on water samples collected from the Chattahoochee River near Norcross, Georgia (USGS station number 02335000), October 23, 2000, through September 30, 2008. *(A)* Relation between measured and estimated mean *Escherichia coli (E. coli)* density. *(B)* Relation between residuals and estimated mean *E. coli* density. *(C)* Quantile-quantile plot of residuals. Explanatory variable: Log10FNU, turbidity in formazin nephelometric units, transformed to base 10 logarithms; MPN/100 mL, most probable number of colonies per 100 milliliters of water.



Residual standard error 0.3510, adjusted R^2 is 0.636, F statistic is 1,240, and *p*-value is less than (<) 0.001

Response variable: Log10Ecoli

Terms	Coefficient	Standard error (SE)	*t*-statistic	*p*-value[b]
Intercept	1.351	0.022	60.44	<0.001
Log10FNU	.448	.028	15.84	< .001
EVENT[a]	.712	.037	19.33	< .001

Autocorrelation coefficients

Number of samples lagged	Correlation coefficient[e]
0	1.0
1	< 0.0

Analysis of variance

Terms	Degrees of freedom[c]	Sum of squares (SS)	Mean SS	F statistic[d]	*p*-value (F)
Log10FNU	1	258.5	258.5	1,489	<0.001
EVENT	1	46.0	46.0	374	< .001
Residuals	1,414	174.2	.1		

[a] Indicator variable for streamflow regime as dry-weather flow (value of 0) or stormflow (value of 1).

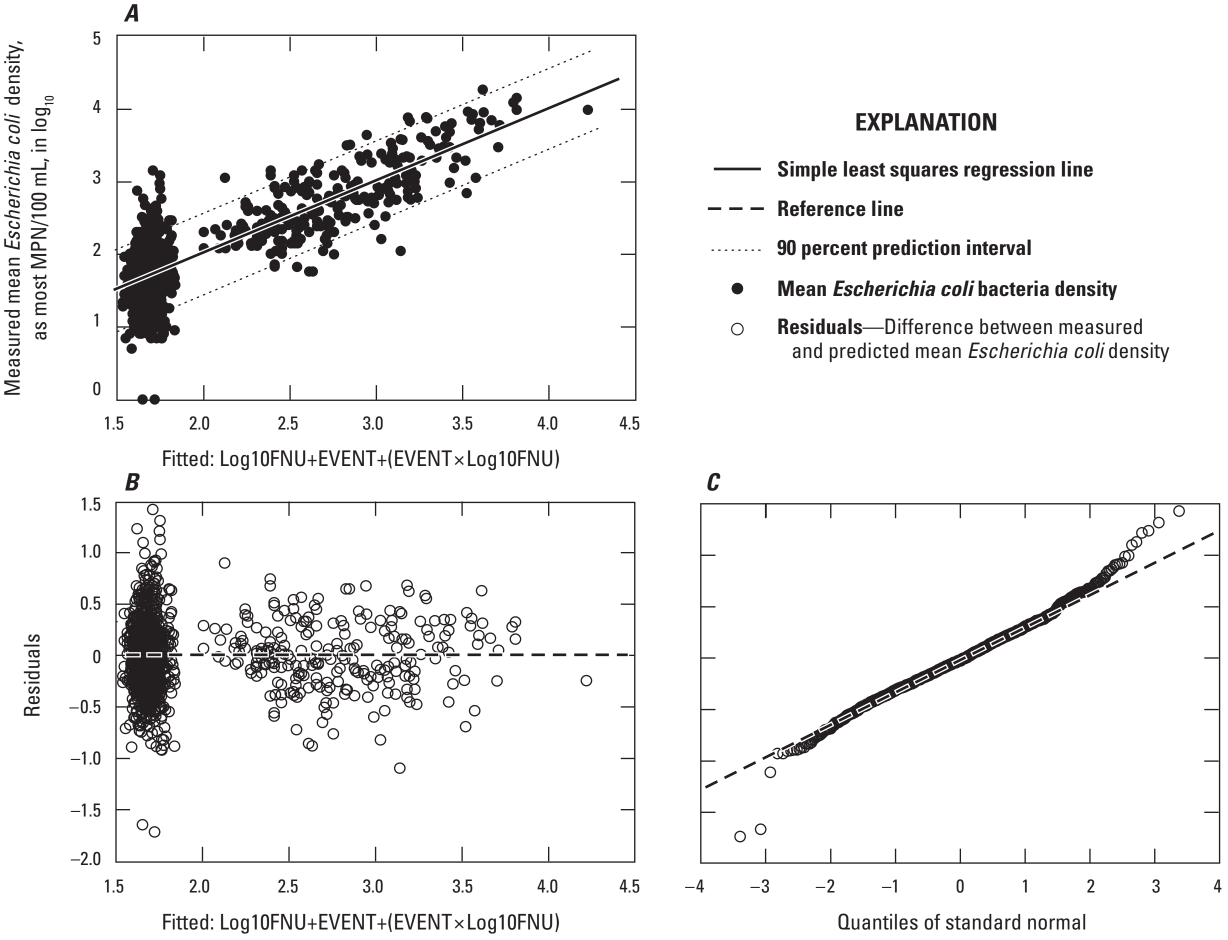
[b] *p*-value, the probability that the parameter is not important to the regression.

[c] Defined as the number of independent pieces of information used to calculate the statistics.

[d] F statistic, used to determine if there is a significant linear relation between the response variable and the explanatory variables.

[e] The critical value is 0.20. Coefficients greater than 0.20 are significant at alpha equal to 0.10.

Figure 5–2. Regression statistics for regression-2 (table 9) on water samples collected from the Chattahoochee River near Norcross, Georgia (USGS station number 02335000), October 23, 2000, through September 30, 2008. *(A)* Relation between measured and estimated mean *Escherichia coli (E. coli)* density. *(B)* Relation between residuals and estimated mean *E. coli* density. *(C)* Quantile-quantile plot of residuals. Explanatory variables: Log10FNU, turbidity in formazin nephelometric units, transformed to base 10 logarithms; EVENT, streamflow regime (dry-weather flow or stormflow, table 3); MPN/100 mL, most probable number of colonies per 100 milliliters of water.



Residual standard error 0.3340, adjusted R^2 is 0.650, F statistic is 878, and *p*-value less than 0.001

Response variable: Log10Ecoli

Terms	Coefficient	Standard error (SE)	*t*-statistic	*p*-value[b]
Intercept	1.546	0.026	59.80	<0.001
Log10FNU	.186	.034	5.43	< .001
EVENT[a]	–.112	.076	–1.48	.139
(EVENT×Log10FNU)	.627	.053	11.76	< .001

Autocorrelation coefficients

Number of samples lagged	Correlation coefficient[e]
0	1.00
1	.40
2	.35
3	.32

Analysis of variance

Terms	Degrees of freedom[c]	Sum of squares (SS)	Mean SS	F statistic[d]	*p*-value (F)
Log10FNU	1	234.8	234.8	2,103	<0.001
EVENT	1	43.9	43.9	393	< .001
(EVENT×Log10FNU)	1	15.4	15.4	138	< .001
Residuals	1,413	157.7	.1		

[a] Indicator variable for streamflow regime as dry-weather flow (value of 0) or stormflow (value of 1).

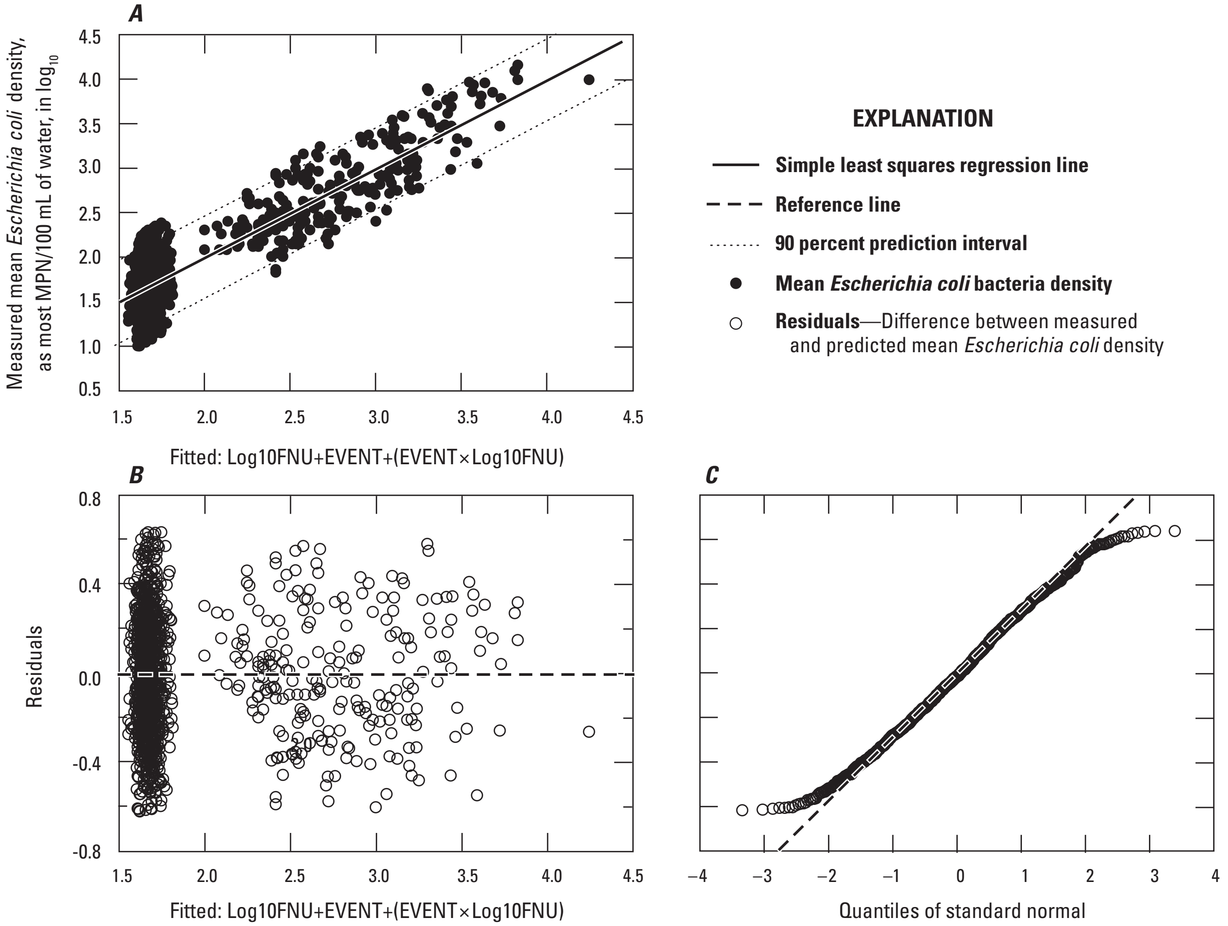
[b] *p*-value, the probability that the parameter is not important to the regression.

[c] Defined as the number of independent pieces of information used to calculate the statistics.

[d] F statistic, used to determine if there is a significant linear relation between the response variable and the explanatory variables.

[e] The critical value is 0.20. Coefficients greater than 0.20 are significant at alpha equal to 0.10.

Figure 5–3. Regression statistics for regression-3 (table 8) on water samples collected from the Chattahoochee River near Norcross, Georgia (USGS station number 02335000), October 23, 2000, through September 30, 2008. *(A)* Relation between measured and estimated mean *Escherichia coli (E. coli)* density. *(B)* Relation between residuals and estimated mean *E. coli* density. *(C)* Quantile-quantile plot of residuals. Explanatory variables: Log10FNU, turbidity in formazin nephelometric units, transformed to base 10 logarithms; EVENT, indicator variable for streamflow regime (dry-weather flow or stormflow, table 3); and an interaction term that is the cross-product of EVENT and Log10FNU; MPN/100 mL, most probable number of colonies per 100 milliliters of water.



Residual standard error 0.2680, adjusted R^2 is 0.744, F statistic is 1,290, and p-value is < 0.001

Response variable: Log10Ecoli

Parameters	Coefficient	Standard error (SE)	t-statistic	p-value[c]
Intercept	1.556	0.021	72.6	<0.001
Log10(FNU)	.159	.029	5.5	< .001
EVENT[a]	–.137	.062	–2.2	.029
EVENT×Log10(FNU)[b]	.664	.044	14.9	< .001

Analysis of variance

Terms	Degrees of freedom[d]	Sum of squares (SS)	Mean SS	F statistic[e]	p-value (F)
Log10(FNU)	1	218.9	218.9	3,047	<0.001
EVENT[a]	1	42.1	42.1	586	< .001
EVENT×Log10(FNU)	1	16.1	16.1	223	< .001
Residuals	1,324	95.1	.1		

Autocorrelation coefficients

Number of samples lagged	Correlation coefficient[f]
0	1.00
1	.27
2	.22

[a] Variable indicating streamflow regime in which samples were collected: dry-weather flow or stormflow (table 3).

[b] Interaction variable.

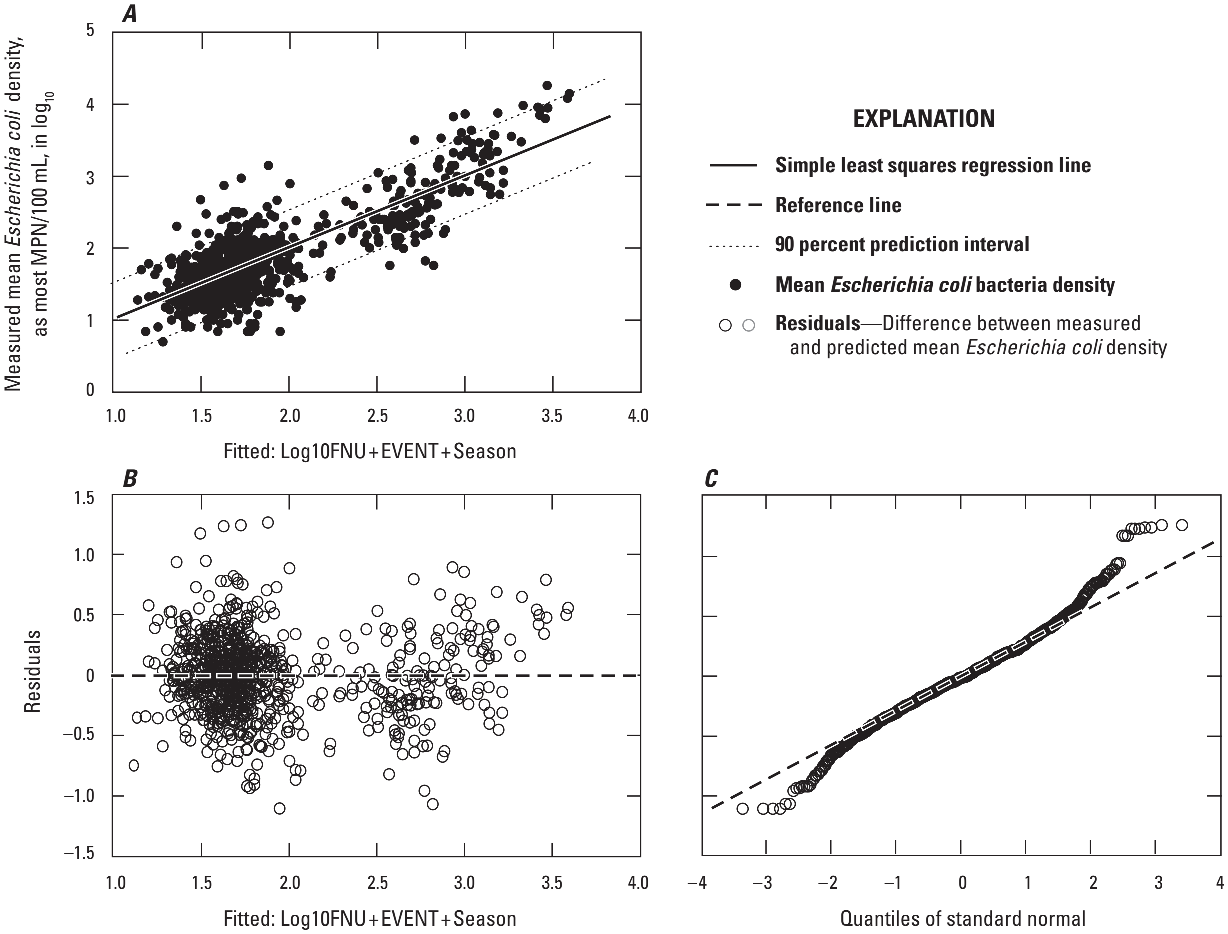
[c] *p*-value, the probability that the parameter is not important to the regression.

[d] Defined as the number of independent pieces of information used to calculate the statistics.

[e] F statistic, used to determine if there is a significant linear relation between the response variable and the explanatory variables.

[f] The critical value is 0.20. Coefficients greater than 0.20 are significant at alpha equal to 0.10.

Figure 5–4. Diagnostic plots for regression-4 (outliers removed, table 8) in residuals from October 23, 2000, through September 30, 2008, Chattahoochee River near Norcross, Georgia (USGS station number 02336000). *(A)* Relation between measured and estimated *Escherichia coli (E. coli)* bacteria densities. *(B)* Relation between regression residuals and the estimated *E. coli* densities. *(C)* Distribution of the residuals compared to a standard normal distribution. Explanatory variables: Log10FNU, turbidity in formazin nephelometric units transformed to base 10 logarithms; EVENT (dry-weather flow or stormflow); and the interaction term (EVENT×Log10FNU); MPN/100 mL, most probable number of colonies per 100 milliliters of water.



Residual standard error 0.3270, adjusted R^2 is 0.684, F statistic is 1,020, and *p*-value less than (<) 0.001

Response variable: Log10Ecoli

Terms	Coefficient	Standard error (SE)	*t*-statistic	*p*-value[a]
Intercept	1.687	0.031	54.55	<0.001
Log10FNU	.523	.027	19.49	< .001
EVENT	.637	.035	18.35	< .001
Season	–.260	.018	–14.71	< .001

Autocorrelation coefficients

Number of samples lagged	Correlation coefficient[d]
0	1.0
1	<0.0

Analysis of variance

Terms	Degrees of freedom[b]	Sum of squares (SS)	Mean SS	F statistic[c]	*p*-value (F)
Log10FNU	1	258.5	258.5	2,418	<0.001
EVENT	1	46.0	46.0	430	< .001
Season	1	23.1	23.1	216	< .001
Residuals	1,413	151.0	.1		

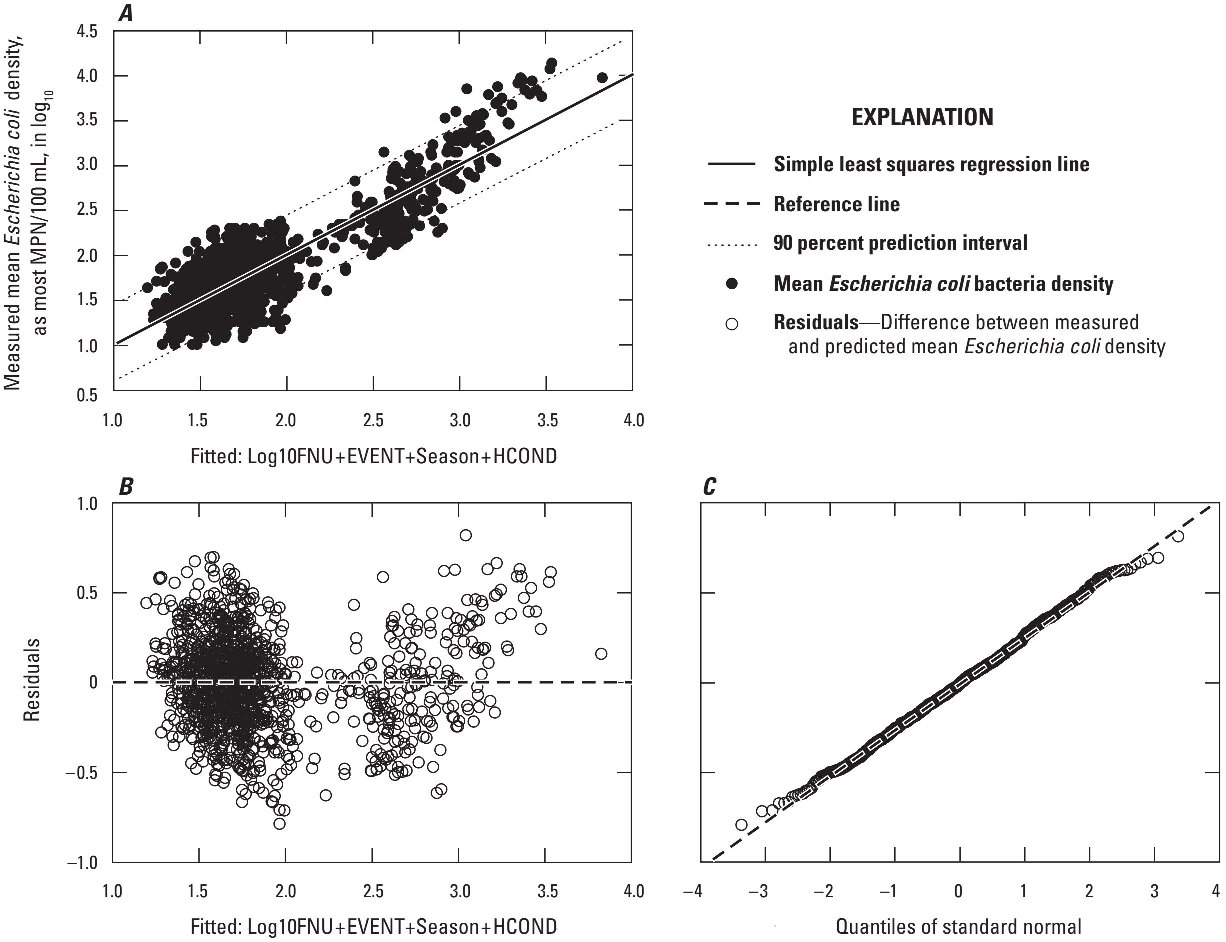
[a] *p*-value, the probability that the parameter is not important to the regression.

[b] Defined as the number of independent pieces of information used to calculate the statistics.

[c] F statistic, used to determine if there is a significant linear relation between the response variable and the explanatory variables.

[d] The critical value is 0.20. Coefficients greater than 0.20 are significant at alpha equal to 0.10.

Figure 5–5. Regression statistics for regression-5 (table 8) on water samples collected from the Chattahoochee River near Norcross, Georgia (USGS station number 02335000), October 23, 2000, through September 30, 2008. *(A)* Relation between measured and estimated mean *Escherichia coli (E. coli)* density. *(B)* Relation between residuals and estimated mean *E. coli* density. *(C)* Quantile-quantile plot of residuals. Explanatory variables: Log10FNU, turbidity in formazin nephelometric units, transformed to base 10 logarithms; EVENT, indicator variable for streamflow regime (dry-weather flow or stormflow, table 3); Season, indicator variable for season (cool, October 16 to April 15 or warm, April 16 to October 15; table 3); MPN/100 mL, most probable number of colonies per 100 milliliters of water.



Residual standard error 0.2580, adjusted R^2 is 0.763, F statistic is 1,070, and *p*-value is less than 0.001

Response variable: Log10Ecoli

Terms	Coefficient	Standard error (SE)	*t*-statistic	*p*-value[a]
Intercept	1.486	0.032	46.81	<0.001
Log10FNU	.490	.021	22.81	< .001
EVENT	.628	.028	22.55	< .001
Season	- .183	.015	-11.90	< .001
HCOND	.042	.005	9.31	< .001

Autocorrelation coefficients

Number of samples lagged	Correlation coefficient[d]
0	1.00
1	.20
2	.16

Analysis of variance

Terms	Degrees of freedom[b]	Sum of squares (SS)	Mean SS	F statistic[c]	*p*-value (F)
Log10FNU	1	218.9	218.9	3,289	<0.001
EVENT	1	42.1	42.1	633	< .001
Season	1	17.3	17.3	261	< .001
HCOND	1	5.8	5.8	87	< .001
Residuals	1,323	88.0	.07		

[a] *p*-value, the probability that the parameter is not important to the regression.

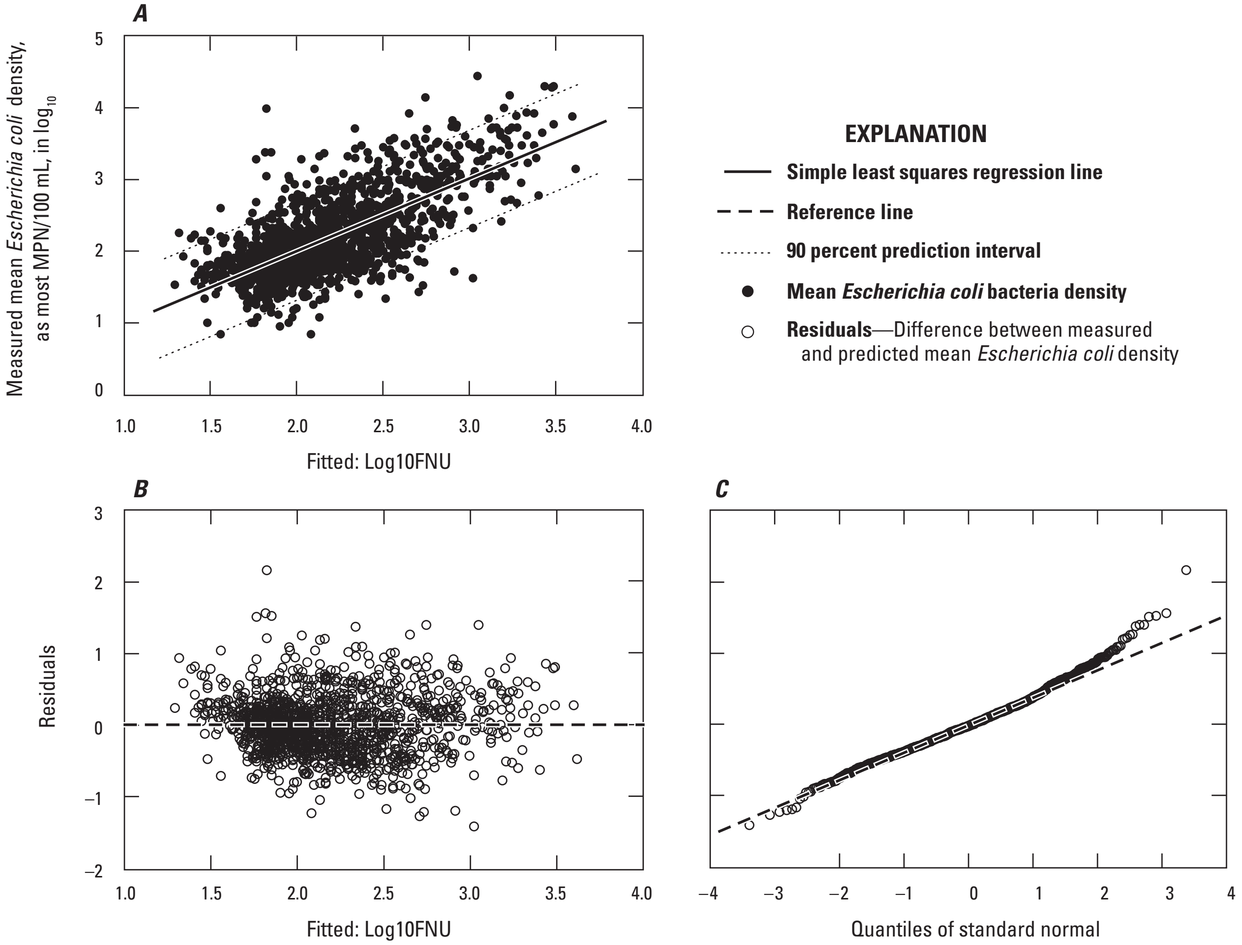
[b] Defined as the number of independent pieces of information used to calculate the statistics.

[c] F statistic, used to determine if there is a significant linear relation between the response variable and the explanatory variables.

[d] The critical value is 0.20. Coefficients greater than 0.20 are significant at alpha equal to 0.10.

Figure 5–6. Regression statistics for regression-7 (table 8) on water samples collected from the Chattahoochee River near Norcross, Georgia (USGS station number 02335000), October 23, 2000, through September 30, 2008. *(A)* Relation between measured and estimated mean *Escherichia coli (E. coli)* density. *(B)* Relation between residuals and estimated mean *E. coli* density. *(C)* Quantile-quantile plot of residuals. Explanatory variables: Log10FNU, turbidity in formazin nephelometric units transformed to base 10 logarithms; EVENT, indicator variable for streamflow regime (dry-weather flow or stormflow, table 3); Season, indicator variable for season (cool, October 16 to April 15 or warm (April 16 to October 15, table 3); HCOND, indicator variable for streamflow condition such as rising or falling stage, table 3; MPN/100 mL, most probable number of colonies per 100 milliliters of water.

Appendix 6. Regression Statistics and Residual Plots for the *Escherichia coli* Models at the Atlanta Site



Residual standard error 0.408, adjusted R^2 is 0.496, F statistic is 1,380, and *p*-value less than 0.001

Response variable: Log10Ecoli

Terms	Coefficient	Standard error (SE)	*t*-statistic	*p*-value[a]
Intercept	1.126	0.031	36.848	<0.001
Log10FNU	.931	.025	37.174	< .001

Analysis of variance

Terms	Degrees of freedom[b]	Sum of squares (SS)	Mean SS	F statistic[c]	*p*-value (F)
Log10FNU	1	230.1	230.1	1,382	<0.001
Residuals	1,405	233.9	.2		

Autocorrelation coefficients

Number of samples lagged	Correlation coefficient[d]
0	1.00
1	.37
2	.25
3	.20

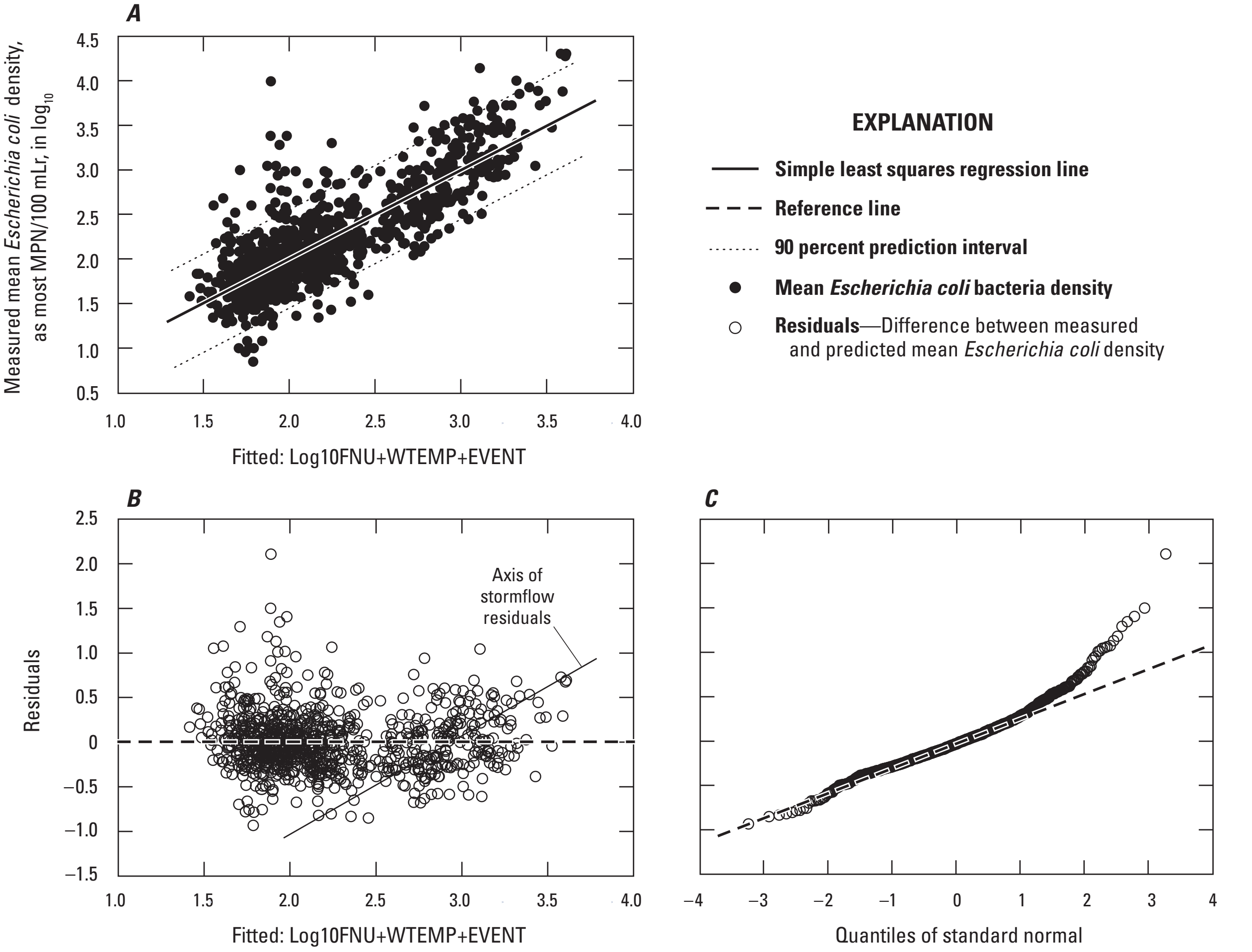
[a] *p*-value, the probability that the parameter is not important to the regression.

[b] Defined as the number of independent pieces of information used to calculate the statistics.

[c] F statistic, used to determine if there is a significant linear relation between the response vari.able and the explanatory variables.

[d] The critical value is 0.20. Coefficients greater than 0.20 are significant at alpha equal to 0.10

Figure 6–1. Regression statistics for regression-9 (table 13) on water samples collected from the Chattahoochee River at Atlanta, Georgia (USGS station number 02336000), October 23, 2000, through September 30, 2008. *(A)* Relation between measured and estimated mean *Escherichia coli (E. coli)* density. *(B)* Relation between residuals and estimated mean *E. coli* density. *(C)* Quantile-quantile plots of the residuals. Explanatory variable: Log10FNU, turbidity in formazin nephelometric units transformed to base 10 logarithms; MPN/100 mL, most probable number of colonies per 100 milliliters of water.



Residual standard error 0.332, adjusted R^2 is 0.669, F statistic is 616, and *p*-value less than < 0.001
Response variable: Log10Ecoli

Terms	Coefficient	Standard error (SE)	*t*-statistic	*p*-value[a]
Intercept	1.107	0.046	23.816	<0.001
Log10FNU	.572	.029	19.997	< .001
WTEMP	.021	.002	9.342	< .001
EVENT	.590	.031	19.297	< .001

Autocorrelation coefficients

Number of samples lagged	Correlation coefficient[d]
0	1.00
1	.23
2	.13
3	.15

Analysis of variance

Terms	Degrees of freedom[b]	Sum of squares (SS)	Mean SS	F statistic[c]	*p*-value (F)
Log10FNU	1	147.7	147.7	1,342	<0.001
WTEMP	1	14.6	14.6	132	< .001
EVENT	1	41.0	41.0	372	< .001
Residuals	909	102.9	.1		

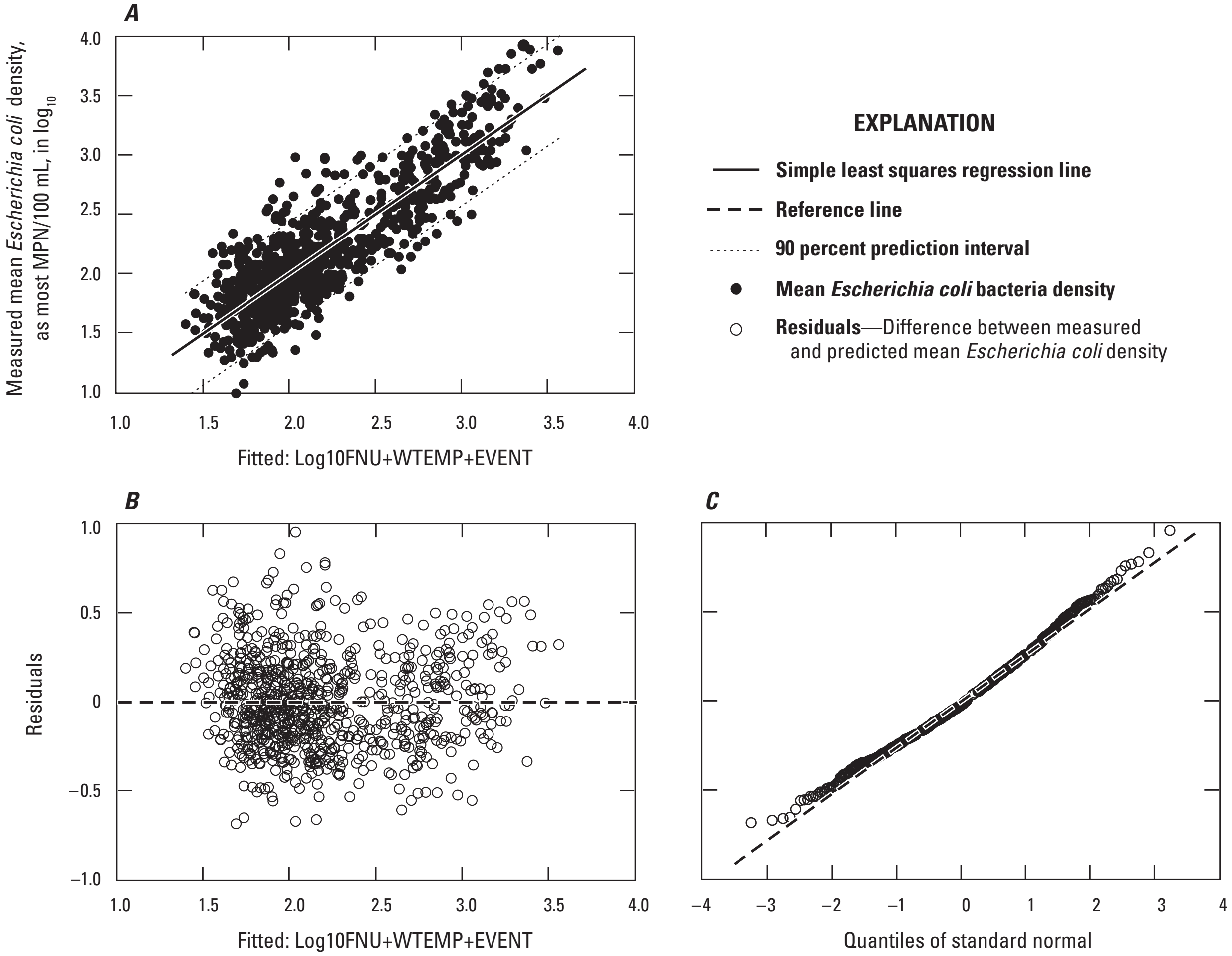
[a] *p*-value, the probability that the parameter is not important to the regression.

[b] Defined as the number of independent pieces of information used to calculate the statistics.

[c] F statistic, used to determine if there is a significant linear relation between the response variable and the explanatory variables.

[d] The critical value is 0.20. Coefficients greater than 0.20 are significant at alpha equal to 0.10.

Figure 6–2. Regression statistics for regression-10 (table 13) on water samples collected from the Chattahoochee River at Atlanta, Georgia (USGS station number 02336000), July 26, 2002, through September 30, 2008. *(A)* Relation between measured and estimated mean *Escherichia coli (E. coli)* density. *(B)* Relation between residuals and estimated mean *E. coli* density. *(C)* Quantile-quantile plot of residuals. Explanatory variables: Log10FNU, turbidity in formazin nephelometric units transformed to base 10 logarithms; WTEMP, water temperature in degrees Celsius; and EVENT, indicator variable for streamflow regime (dry-weather flow or stormflow, table 3); MPN/100 mL, most probable number of colonies per 100 milliliters of water.



Residual standard error 0.260, adjusted R^2 is 0.747, F statistic is 843, and *p*-value less than 0.001

Response variable: Log10Ecoli

Terms	Coefficient	Standard error (SE)	t-statistic	*p*-value[a]
Intercept	1.082	0.038	28.597	<0.001
Log10FNU	.604	.023	25.737	< .001
WTEMP	.020	.002	10.712	< .001
EVENT	.521	.025	20.936	< .001

Analysis of variance

Terms	Degrees of freedom[b]	Sum of squares (SS)	Mean SS	F statistic[c]	*p*-value (F)
Log10FNU	1	131.0	131.0	1,942	<0.001
WTEMP	1	10.0	10.0	148	< .001
EVENT	1	29.6	29.6	438	< .001
Residuals	851	57.4	.07		

Autocorrelation coefficients

Number of samples lagged	Correlation coefficient[d]
0	1.00
1	.24
2	.16
3	.10

[a] *p*-value, the probability that the parameter is not important to the regression

[b] Defined as the number of independent pieces of information used to calculate the statistics

[c] F statistic, used to determine if there is a significant linear relation between the response variable and the explanatory variables

[d] The critical value is 0.20. Coefficients greater than 0.20 are significant at alpha equal to 0.10

Figure 6–3. Regression statistics for regression-11 (table 13) on water samples collected from the Chattahoochee River at Atlanta, Georgia (USGS station number 02336000), July 26, 2002, through September 30, 2008. *(A)* Relation between measured and estimated mean *Escherichia coli (E. coli)* density. *(B)* Relation between residuals and estimated mean *E. coli* density. *(C)* Quantile-quantile plot of residuals. Explanatory variable: Log10FNU, turbidity in formazin nephelometric units, transformed to base 10 logarithms; WTEMP, water temperature in degrees Celsius; EVENT, indicator variable for streamflow regime (dry-weather flow or stormflow, table 3); MPN/100 mL, most probable number of colonies per 100 milliliters of water.

Manuscript approved on February 14, 2012

Edited by Kay P. (Hedrick) Naugle
Illustrations and layout by Caryl J. Wipperfurth
Science Publishing Network, Raleigh PSC

For more information about this publication, contact:
USGS Georgia Water Science Center
3039 Amwiler Road
Atlanta, Georgia 30360
telephone: 770-903-9100

http://ga.water.usgs.gov

http://ga2.er.usgs.gov/bacteria/

Lawrence—***Escherichia coli* bacteria density in the Chattahoochee River near Atlanta, Georgia, October 2000 through September 2008**—Scientific Investigations Report 2012–5037